Développement d'une approche markovienne pour l'analyse de l'organisation spatiale des génomes.

Christelle Melo de Lima

N° d'ordre 238-2995 Année 2005

THÈSE

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON-1

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 25 avril 2002)

présentée et soutenue publiquement le 28 novembre 2005

Par

Christelle Gonindard
Melo de Lima

TITRE :

Développement d'une approche markovienne pour l'analyse de l'organisation des génomes

Directeurs de thèse : Piau Didier et Rechenmann François

JURY :		
	Florence Forbes	Examinatrice
	Christian Gautier	Président
	Didier Piau	Directeur
	François Rechenmann	Directeur
	Sophie Schbath	Rapporteur
	Claude Thermes	Rapporteur

N° d'ordre 238-2995 Année 2005

THÈSE

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON-1

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 25 avril 2002)

présentée et soutenue publiquement le 28 novembre 2005

Par

Christelle Gonindard
Melo de Lima

TITRE :

Développement d'une approche markovienne pour l'analyse de l'organisation des génomes

Directeurs de thèse : Piau Didier et Rechenmann François

JURY :		
	Florence Forbes	Examinatrice
	Christian Gautier	Président
	Didier Piau	Directeur
	François Rechenmann	Directeur
	Sophie Schbath	Rapporteur
	Claude Thermes	Rapporteur

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le professeur D. DEBOUZIE**
Vice-Président du Conseil Scientifique	M. le professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le professeur R. GARRONE
Vice-Président des Etudes et de la Vie Universitaire	M. le professeur G. ANNAT
Secrétaire Général	M. J. P. BONHOTAL

SECTEUR SCIENCES

Composantes	Directeur
UFR de Physique	M. le Professeur J. L. VIALLE
UFR de Biologie	M. le Professeur H. PINON
UFR de Mécanique	M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	M. le Professeur A. BRIGUET
UFR de Sciences de la Terre	M. le Professeur P. HANTZPERGUE
UFR de Mathématique	M. le Professeur M. CHAMARIE
UFR d'Informatique	M. le Professeur M. EGEA
UFR de Chimie Biochimie	M. le Professeur J. P. SCHARFF
UFR de STAPS	M. le Professeur R. MASSARELLI
Observatoire de Lyon	M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	M. le Professeur J. P. PUAUX
Département de premier cycle Sciences	M. J. C. DUPLAN Maître de conférences
IUT A	M. le Professeur M. ODIN
IUT B	M. le Professeur G. MAREST
Institut de Science Financière et d'Assurances	M. le Professeur J. C. AUGROS

SECTEUR SANTE

Composantes	Directeur
UFR de Médecine Lyon R.T.H Laënnec	M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	M. le Professeur F.N. GILLY
UFR d'Odontologie	M. le Professeur J.DOURY
Institut des Sciences Pharmacieutiques Biologies	M. le Professeur F. LOCHER
Institut techniques de Réadaptation	Mme le Professeur D. BOISSON
Département de Formation et Centre Recherche en Biologie Humaine	M. le Professeur P. FARGE
Département de Formation à la Recherche et à l'Evaluation Pédagogiques	M. le Professeur M. LAVILLE

Remerciements

L'usage conduit le doctorant, travaillant en équipe, à signer sa thèse de son seul nom. Heureusement, la page des remerciements permet de compenser un peu cet individualisme, tout en respectant l'usage.
Je désire donc remercier tout particulièrement Christian Gautier de m'avoir ouvert les portes de son laboratoire et confier mon sujet de DEA qui c'est prolongé par une thèse. Je tiens également à remercier Didier Piau et François Rechenmann. Tout au long de ces trois années de travail, leur expérience ainsi que nos nombreuses conversations m'ont permis de beaucoup apprendre et développer une pluri-discininarité. Je les remercie pour leur confiance et leur soutient à chaque instant de ce travail.
Je souhaite également remercier Marie-France Sagot et Laurent Guéguen pour le soutien qu'ils m'ont apportés tout au long de ma thèse. Leurs conseils et nos discussions m'ont aidés à de nombreuses reprises.
Je remercie l'ensemble des membres du jury (Florence Forbes, Sophie Schbath et Claude Thermes) d'avoir accepté de venir examiner mon travail. En particulier Sophie Schbath et Claude Thermes qui ont acceptés les rôles de rapporteurs de mon mémoire.
Je remercie sincèrement l'ensemble des membres des équipes BAOBAB et BGE pour toute l'aide qu'ils m'ont apportés. Leur disponibilité et leurs conseils m'ont beaucoup aidé tout au long de ma thèse. Je tiens à remercier également Marie-Joseph Pieri et Dominique Vauday pour leur aide administrative.
J'adresse un remerciement très particulier à tous les étudiants que j'ai côtoyés durant ma thèse pour les excellents moments que nous avons passé ensemble.
Pour finir, je souhaite dédier cette thèse à Marina et surtout à David que j'aimerais remercier du fond du cœur, qu'il reçoive tout mon amour pour sa patience, sa gentillesse et surtout pour sa présence.

Table des matières

Résumé

Le séquençage à grande échelle a permis d'accéder aux génomes complets de nombreux organismes. Les modèles de Markov cachés sont une des méthodes probabilistes les plus utilisées pour l'analyse des séquences. L'objectif de ces travaux est de participer à l'analyse de l'organisation et à la compréhension de l'évolution des génomes. Une méthode de prédiction des isochores adaptée au génome humain a été développée. L'originalité de cette approche consiste à identifier des ruptures d'homogénéité des séquences mais surtout leurs causes biologiques, comme l'influence des régions UTRs lors du classement des gènes dans un isochore. Cette méthode a ensuite été appliquée au génome du tetraodon, pour lequel l'existence d'une structure en mosaïque nouvelle le long de son génome a été mise en évidence.

Introduction

Au cours des dix dernières années, le séquençage d'ADN à grande échelle a permis d'augmenter considérablement la quantité des données disponibles. Leur arrivée permet aujourd'hui d'accéder aux génomes complets de nombreux organismes vivants, c'est-à-dire à l'ensemble de leur patrimoine génétique. Actuellement, 261 génomes complets ont été séquencés dont 228 génomes procaryotes et 33 génomes eucaryotes. L'interprétation de ces séquences est naturellement devenue un enjeu central de la biologie. La recherche de l'ensemble des gènes protéiques dans cette masse de données, l'analyse de la structure des génomes ou encore la comparaison des différents génomes ne peuvent être envisagées de manière expérimentale. Par exemple, le séquençage complet du génome humain représente à lui seul la connaissance d'environ trois milliards de paires de bases, les parties codantes ne constituant que 1 à 3% de ces données. Les analyses mathématique et informatique des séquences d'ADN ont été des voies de recherche privilégiée et sont devenues une première étape nécessaire pour la compréhension des séquences d'ADN. Parmi les modèles probabilistes utilisés, les modèles de Markov cachés (HMM pour Hidden Markov Model) ont pris une place particulière car ils constituent une solution simple et efficace pour permettre une description probabiliste de l'alternance de différents types d'éléments le long des séquences génomiques.

L'analyse des génomes complets représente une source exceptionnelle d'informations concernant l'évolution des êtres vivants. L'objectif de ce travail de thèse est de participer à l'analyse de l'organisation des génomes et de la compréhension de leur évolution. Le long des chromosomes de mammifères, il existe, sur une échelle de quelques centaines de kilobases, une forte variabilité de la composition en bases $G + C$. Les régions d'un génome dont la longueur est supérieure à 300 kb et dont la composition en bases $G+C$ est homogène sont appelées isochores. Ces régions ont été mises

en évidence expérimentalement pour la première fois par des techniques de centrifugation en gradient de densité. La structure en isochores est corrélée avec d'importantes caractéristiques de l'organisation des génomes, comme par exemple la densité en gènes, la taille des gènes, les distributions des éléments transposables ou le taux de recombinaison. La localisation et l'analyse de telles régions présentent donc un grand intérêt pour la compréhension des génomes.

Au cours de ce travail de thèse, deux domaines d'application des HMMs pour l'analyse des génomes d'eucaryotes ont donc été abordés. Le premier est la détection de la structure en isochores le long des génomes. Le deuxième est la description et l'interprétation des zones d'hétérogénéités détectées, d'abord à l'échelle du gène, et ensuite à l'échelle des isochores. Dans les deux cas, la méthodologie employée repose essentiellement sur des sélections de modèles HMMs. L'originalité de la méthode proposée consiste à utiliser des modèles simples qui permettent facilement une analyse de l'aspect biologique tout particulièrement en cas d'échec des prédictions de la méthode afin de détecter des propriétés biologiques nouvelles. Cette capacité d'interprétation est originale, dans la mesure où les méthodes classiques sont très opaques dans l'explication des succès comme des échecs.

Le manuscrit est organisé en cinq chapitres. Le premier présente l'utilisation des HMMs dans le cadre de l'analyse des séquences biologiques. Il détaille également les principaux aspects mathématiques et algorithmiques de l'utilisation des chaînes de Markov cachées. Le deuxième chapitre décrit la mise en place des modèles pour l'exploration de données génomiques et particulièrement dans le cadre de l'analyse de la structure des gènes. Le troisième chapitre présente une méthode de prédiction des isochores adaptée au génome humain. Le quatrième chapitre utilise la méthode de prédiction développée pour vérifier les hypothèses émises par les analyses par centrifugation en gradient de densité (Bernardi 2000). L'étude de Bernardi montre que les génomes des mammifères et des oiseaux sont fortement hétérogènes. En revanche, cela n'est pas le cas des génomes des amphibiens et des poissons, qui sont, eux, relativement homogènes, et en général dépourvus d'isochores très riches en $G + C$. Afin de vérifier cette hypothèse, une analyse de l'organisation des génomes du *Tetraodon nigroviridis* et du fugu est donc conduite en utilisant les outils développés au cours des chapitres précédents. Enfin, le cinquième chapitre généralise cette méthode aux génomes de la souris, du

chimpanzé et du poulet pour permettre une analyse ultérieure de l'évolution de la structure en isochores chez les vertébrés.

Chapitre 1

De la biologie à la modélisation mathématique

Ce chapitre présente l'utilisation des chaînes de Markov cachées dans le cadre de l'analyse des séquences biologiques que sont l'ADN, l'ARN et les protéines. La première partie décrit ces molécules et leurs principaux rôles biologiques et met en évidence la nécessité du développement de méthodes mathématiques et informatiques pour l'analyse et l'interprétation des données issues des séquençages des génomes entiers. Ces problèmes sont au coeur des préoccupations de la biologie actuelle. La deuxième partie se décompose en deux sections. La première est une courte introduction qui présente l'analyse des séquences, basée sur leur modélisation probabiliste à partir des HMMs. La seconde section donne une présentation mathématique des différents modèles de Markov couramment utilisés et leurs algorithmes associés.

1.1 Présentation des séquences biologiques

Certains organismes vivants sont multicellulaires, d'autres sont unicellulaires, dans tous les cas, la cellule est l'unité vivante élémentaire. Les cellules ont la capacité de se reproduire. Selon leur organisation, deux grands types de cellules vivantes se distinguent : les cellules procaryotes, relativement simples et constituées essentiellement d'un unique compartiment ; et les cellules eucaryotes, compartimentées, et souvent beaucoup plus grandes que les cellules procaryotes. Alors que les organismes multicellulaires sont tous des organismes eucaryotes (végétaux, champignons, animaux), il existe des organismes unicellulaires aussi bien eucaryotes (algues unicellulaires, levures) que bactériens. Cette thèse traitera uniquement le cas des organismes eucaryotes.

Les cellules vivantes contiennent trois principaux types de macromolécules : l'acide désoxyribonucléique (ADN), l'acide ribonucléique (ARN) et les protéines. La description des relations entre les séquences d'ADN, d'ARN et des protéines est au cœur de la théorie de la biologie moléculaire. Les molécules d'ADN sont le support de l'information génétique. Cette information se perpétue grâce au mécanisme de réplication de l'ADN. L'expression de l'information génétique repose sur la transcription de l'ADN en ARN et sur la traduction de l'ARN en protéines.

1.1.1 Présentation des macromolécules biologiques

La double hélice d'ADN

La découverte de la structure en double hélice de l'ADN (Figure 1.1) par Watson et Crick en 1953 a révolutionné la biologie. L'information génétique d'une cellule est contenue dans un petit nombre de molécules d'ADN : les chromosomes. Par exemple, le noyau d'une cellule humaine contient 46 chromosomes. La molécule d'ADN est formée d'une chaîne linéaire constituée d'une combinaison ordonnée de quatre nucléotides distincts A, T, C, G pour Adénine, Thymine, Cytosine et Guanine respectivement. Trois paires successives de bases nucléotidiques forment un « triplet » ou codon qui code dans une région codante pour un acide aminé. La règle d'appariement des nucléotides permet de déduire la séquence d'un brin en fonction de l'autre : A fait face à T et G à C. Cette observation suggère le mécanisme de réplication de la séquence : chaque brin sert de matrice pour la synthèse d'un

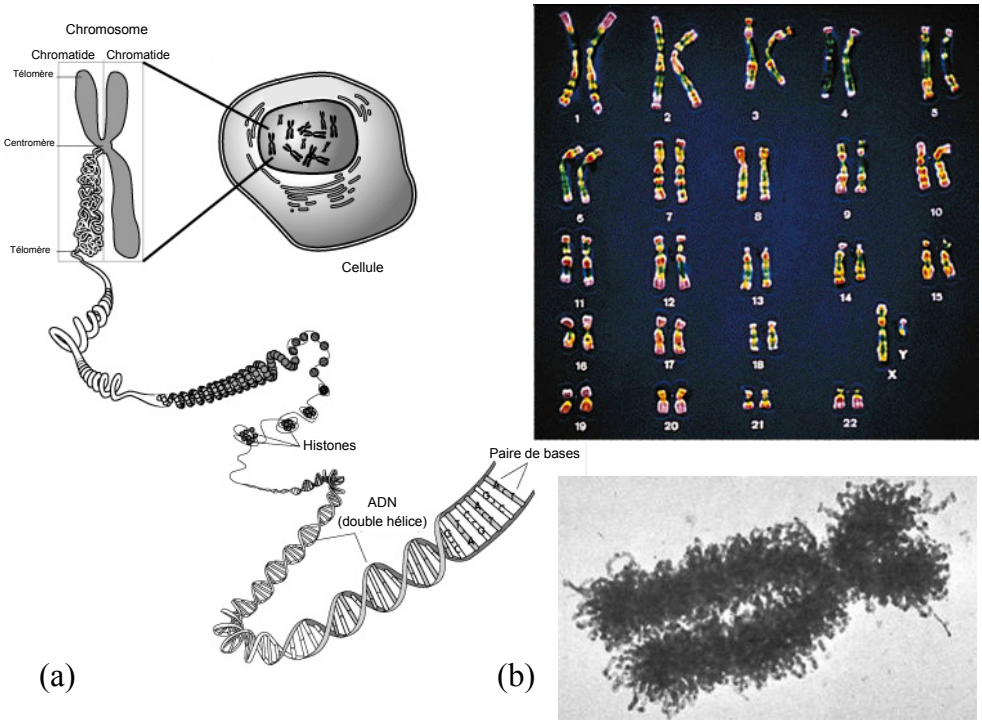


Fig. 1.1 : *(a)Représentation schématique d'une cellule et de la structure de la double hélice d'ADN composant les chromosomes, (b) Le caryotype correspond à l'établissement de la carte chromosomique des chromosomes humains. Il représente donc les 46 chromosomes présents dans chacune des cellules humaines, vus au microscope électronique après une préparation et une coloration par le laboratoire (Génoscope).*

nouveau brin.

Les séquences d'ADN sont constituées de différentes régions, notamment des gènes. Un gène est un segment d'ADN qui assure une fonction cellulaire précise en codant pour la synthèse d'une protéine particulière. Le génome humain contient environ 24000 gènes. Chez les eucaryotes, les gènes sont composés de régions codantes (les exons) et non codantes (les introns). La structure d'un gène, constituée de plusieurs exons, est la suivante (Figure 1.2) : il commence par une région promotrice (point de départ de la transcription), qui est suivie d'une région transcrite mais non traduite (5'UTR). Il existe ensuite une série alternant exons et introns qui se termine par une nouvelle région transcrite mais non traduite (3'UTR). Le gène se termine, en général, par un site de polyadénylation qui est une répétition de nucléotides

A. Les régions intergéniques séparent les gènes mais sont très mal connues. Au cours des chapitres suivants, nous nommerons CDS la partie codante constituée d'une juxtaposition d'exons, les introns étant excisés. Les CDS débutent par un codon méthionine (codon "start") présent dans ce que l'on nommera premier exon codant et se termine par un codon "stop" qui est contenu dans ce que l'on nommera l'exon terminal. Cette partie codante ne représente que 1 à 3% du génome humain (Duret *et al*. 1995).

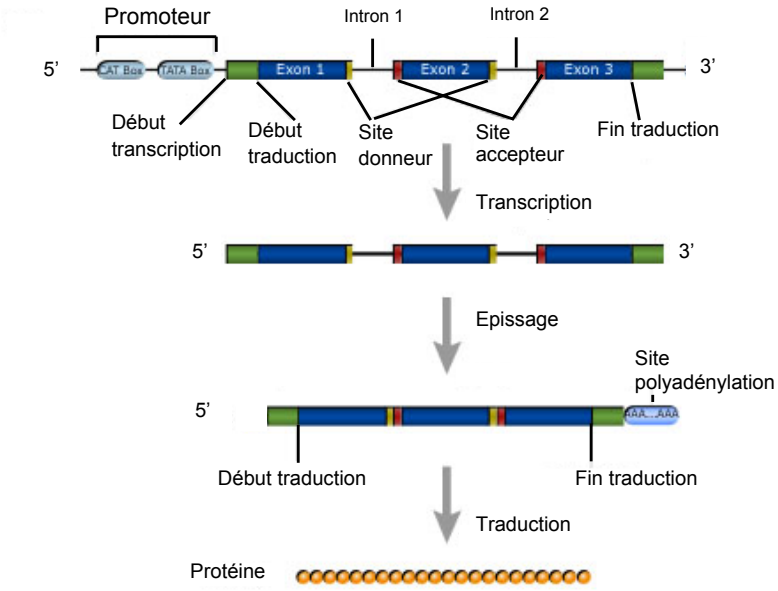


Fig. 1.2 : *Expression de l'information génétique*

La taille des génomes des eucaryotes est généralement plus grande et plus variable que celle des procaryotes. Il semblerait que la majorité du génome des eucaryotes n'ait pas de fonction spécifique et qu'il soit apparu indépendamment des besoins de l'organisme. Une proportion très importante de génome de la plupart des eucaryotes est constituée d'ADN répété. Cette proportion d'ADN répété est très variable selon les lignages : 20% chez les levures, jusqu'à 60% chez les mammifères et jusqu'à 80% chez les plantes. Il existe deux types de séquences répétées. Le premier type est représenté par les séquences répétées localement. Il s'agit de répétitions de séquences d'ADN les unes à la suite des autres (en tandem), elles sont nommées «

tandem repeats » ou ADN satellite. Il en existe différentes familles : les minisatellites (blocs de 100 à 2000 bp) et les microsatellites (blocs inférieur à 100 bp). Ces séquences sont généralement trouvées sur tous les chromosomes et leurs rôles fonctionnels sont mal connus. Les séquences répétées dispersées sur l'ensemble du génome, constituent le deuxième type de séquences. Elles se retrouvent dans les régions géniques, intergéniques et même dans les introns. Elles sont regroupées principalement en deux classes : les SINEs (short interspersed repeated sequences) et les LINEs (long interspersed repeated sequences). L'ADN non-génique constituant la majorité de l'ADN chez les eucaryotes, il est légitime de se demander quel est son rôle. Plusieurs explications sont proposées :

- il aurait certaines fonctions essentielles et serait notamment nécessaire pour la régulation globale de l'expression des gènes ;
- il serait inutile, il serait alors porté passivement par les chromosomes ;
- il serait un parasite sans fonction, un ADN égoïste, qui se multiplierait sans se soucier de son hôte.

L'ARN

L'ARN conserve la totalité de l'information contenue dans la séquence d'ADN à partir de laquelle il a été copié. Tout comme l'ADN, c'est une molécule linéaire. Cependant, elle ne comporte qu'un seul brin et ne s'enroule pas en double hélice, de plus, les nucléotides de l'ARN diffèrent de ceux de l'ADN par le remplacement d'une base (thymine) par une autre (uracile). Les molécules d'ARN sont relativement courtes comparées aux molécules d'ADN, puisqu'elles sont copiées à partir d'une région limitée de l'ADN, suffisante, par exemple, pour fabriquer une ou plusieurs protéines. L'ARN joue donc un rôle essentiel dans la synthèse des protéines. Il existe différents types d'ARN dont les principaux sont l'ARN messager (ARNm) et l'ARN de transfert (ARNt). Les ARNm transportent l'information portée par l'ADN pour les protéines. Les ARNt quant à eux jouent un rôle majeur lors de la traduction : chaque type d'ARNt porte un acide aminé et reconnaît un ou des codons qui lui correspondent. La reconnaissance a lieu grâce à l'appariement du codon de l'ARNm avec trois bases successives de l'ARNt (anticodon). Les ARNt ont une longueur d'environ 80 nucléotides.

Les protéines

Les protéines constituent l'une des plus importantes classes de molécules présentes dans tous les organismes vivants et les virus. Elles assurent l'essentiel des fonctions de la cellule (architecture cellulaire, effecteurs au niveau du fonctionnement). Elles se retrouvent sous différentes formes : enzymes, hormones, récepteurs, neurotransmetteurs... Les protéines sont formées de l'enchaînement de molécules beaucoup plus petites, les acides aminés. Il existe vingt acides aminés différents. Certaines protéines en comptent quelques dizaines, d'autres plusieurs milliers. Chaque protéine a une structure particulière, qui dépend de l'ordre des acides aminés et de la façon dont leur chaîne se replie dans l'espace. La forme des protéines est étroitement liée à leur fonction : fonction de structure (kératine des ongles et des cheveux, collagène de la peau...), fonction de transport (hémoglobine qui transporte l'oxygène dans le sang...), fonction de défense contre des micro-organismes (anticorps), fonction hormonale (insuline, adrénaline...), fonction d'activation des réactions chimiques des cellules (enzymes)... La protéine est donc la résultante du message génétique contenu dans un gène.

1.1.2 Mécanismes liés à l'information génétique

Réplication de l'ADN et variabilité de l'information génétique

Durant la réplication de l'information génétique, la séquence de chaque brin d'ADN sert de modèle à la synthèse d'un brin fils de séquence complémentaire inverse. À partir de son origine, la réplication progresse en séparant les deux brins modèles tout en synthétisant les deux brins fils. Ce mécanisme permet aux brins de rester peu de temps non appariés, mais implique une asymétrie de leur traitement.

De nombreux mécanisme sont à l'origine de la variabilité de l'information génétique nécessaire à l'évolution. Ainsi, bien que la réplication de l'ADN soit très fidèle, des erreurs, ou mutations ponctuelles, peuvent survenir. Des mutations peuvent également se produire entre les réplications et sont alors transmises à la descendance lors de la réplication. De même, les réarrangements jouent un rôle essentiel parmi les mécanismes à l'origine de la variabilité génétique en permettent le déplacement de fragments d'ADN. Ces réarrangements peuvent notamment entraîner des insertions, des délétions, des duplications de grandes ampleurs et d'autres types de remaniements.

Transcription de l'ADN en ARN

L'expression de l'information génétique passe par la transcription de la séquence d'ADN en séquence d'ARN (Figure 1.3 (a)), puis par la traduction de la séquence d'ARN en séquence protéique. Le principe de la transcription est proche de celui de la réplication. Comme la synthèse de l'ADN, la synthèse de l'ARN se fait dans le sens 5' vers 3' et se fonde sur l'appariement des nucléotides de l'ARN synthétisé avec ceux de l'ADN modèle, mais en remplaçant la thymine par l'uracile. La plupart des ARN sont traduits, au moins partiellement, en protéines : ce sont les ARN messagers (ARNm). La transcription des ARNm peut en partie être interprétée comme une amplification de l'information génétique. Ainsi, une même séquence d'ADN est généralement transcrite en de multiples molécules d'ARNm qui seront elles-mêmes traduites en de multiples séquences protéiques. La quantité d'ARN fabriquée à partir d'une région particulière de l'ADN est contrôlée par des protéines régulatrices de gènes qui se lient à des sites spécifiques de l'ADN proches des séquences codantes d'un gène.

Traduction de l'ARN en protéine

La traduction (Figure 1.3 (b)) utilise un code, dit code génétique (Figure 1.3 (c)). Celui-ci permet de déterminer la séquence protéique, écrit dans un alphabet à 20 lettres (acides aminés), en fonction de la séquence nucléique, écrit dans un alphabet à 4 lettres (nucléotides). Ce code associe à chacun des 64 codons, l'un des acides aminés ou un signal d'arrêt de la traduction (codons stop). Le code génétique est donc dégénéré puisqu'il fait correspondre 20 acides aminés à 61 codons. La plupart des codons synonymes diffèrent uniquement par le nucléotide en troisième position. La séquence d'ADN est lue comme une suite de codons non-chevauchants dans le sens 5' vers 3' et la séquence protéique est synthétisée. Pour déterminer la séquence protéique correspondant à une séquence d'ARN, il faut être capable de retrouver la phase de lecture. Il existe donc des codons start et des codons stop qui permettent respectivement de débuter et de terminer la traduction. Chez les eucaryotes, l'ARN transcrit subit un épissage au cours duquel les introns sont excisés de manière très précise pour mettre bout à bout les exons avant la traduction.

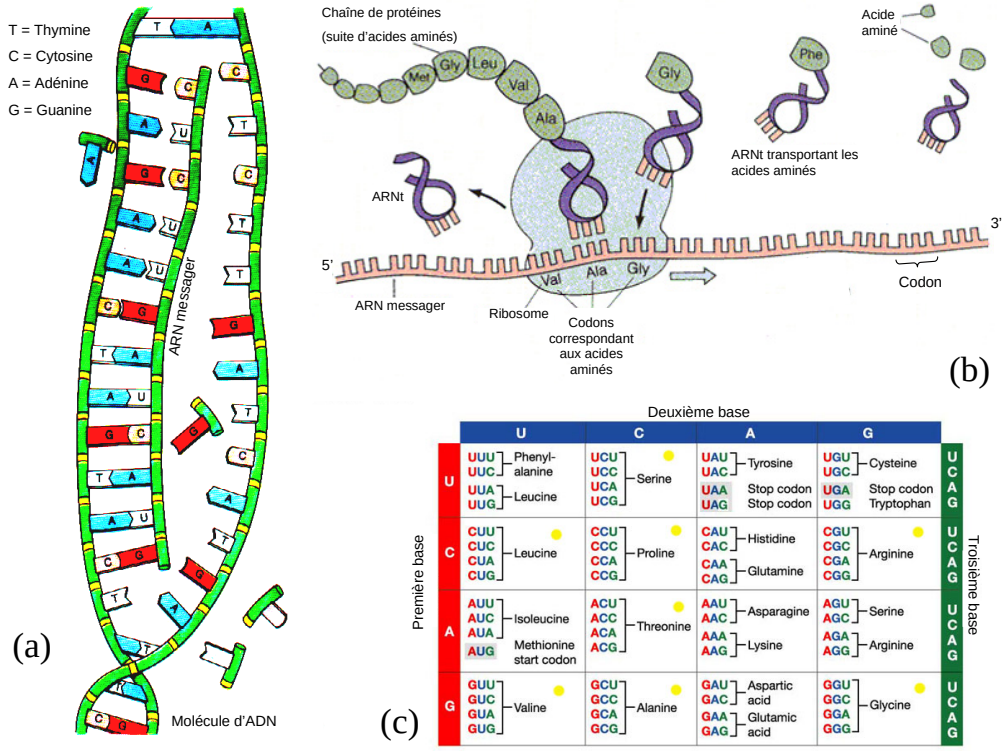


Fig. 1.3 : *(a)Transcription, (b) Traduction, (c) Code génétique*

Signaux

La présence de signaux le long des séquences nucléiques et protéiques est essentielle à l'expression de l'information génétique. On peut citer, par exemple, les signaux sur les séquences d'ADN qui permettent le choix des segments transcrits ou les signaux sur l'ARN qui déterminent les limites des introns et le choix du site de début de traduction. Des signaux existent aussi sur les protéines. Les signaux peuvent être décrits comme l'occurrence de séquences particulières. Bien souvent, de nombreuses séquences peuvent représenter un même signal, qui doit donc être décrit comme un motif plutôt que comme une séquence particulière. La situation est encore plus compliquée par le fait que des séquences interprétées comme un signal dans certains contextes peuvent apparaître « par hasard » le long de la séquence sans pour autant jouer le rôle d'un signal.

La première partie de ce chapitre a permis de mettre en évidence de façon sommaire les liens entre les différentes macromolécules ainsi qu'entre les mécanismes liés à l'information génétique qui seront utilisés par la suite. Les travaux de modélisation et d'analyse développés au cours de cette thèse seront uniquement consacrés à la molécule d'ADN, qui est à la base de l'information génétique.

1.1.3 Séquençage et analyse des génomes

Au cours des années 1990, de nouvelles technologies ont permis d'augmenter considérablement la quantité des données disponibles concernant les informations génétiques. Ainsi, le séquençage de l'ADN peut être considéré comme la première des technologies de la biologie à grande échelle, aussi bien par son importance actuelle que d'un point de vue chronologique. L'arrivée de ces données a eu un profond impact sur la biologie moléculaire, elles permettent aujourd'hui d'accéder au génome complet de nombreux organismes vivants, c'est-à-dire à l'ensemble de leur patrimoine génétique. Fin juin 2005, 261 génomes complets ont été séquencés dont 228 génomes procaryotes et 33 génomes eucaryotes. La masse de données à analyser est donc considérable. Par exemple, le génome humain est constitué d'environ 3 milliards de paires de bases, la partie codante ne représentant que 3% du génome total. La recherche des gènes protéiques dans cette masse de données, l'analyse de la structure des génomes ou encore la comparaison des différents génomes ne peuvent donc pas être envisagées manuellement.

Bien avant l'essor de la biologie à grande échelle, des algorithmes étaient déjà nécessaires pour comparer les rares séquences disponibles. Ainsi, dans les années 1970, Needleman et Wunsch décrivaient un algorithme de comparaison de deux séquences protéiques. C'est cependant dans les années 1990, avec l'avènement des technologies de la biologie à grande échelle et des ordinateurs, que l'utilisation de l'informatique s'est généralisée afin de stocker, intégrer et tenter d'interpréter les énormes quantités de données produites.

Parallèlement à l'informatique, les statistiques se sont révélées très utiles pour répondre à certaines questions, notamment pour quantifier l'incertitude liée à l'interprétation des données récoltées, par exemple, pour déterminer si une similitude trouvée en comparant une séquence à un ensemble de séquences n'est pas dû au hasard. En amont de ces problèmes, qui relèvent principalement du calcul de significativité, les statistiques ont aussi

largement guidé la conception des méthodes d'extraction d'information. Par exemple, la fonction de score à maximiser lors d'un alignement de deux séquences utilise des méthodes statistiques.

L'analyse des génomes complets est d'un grand intérêt, elle représente une source exceptionnelle d'informations sur l'évolution des organismes vivants. En effet, un génome contient la trace de certains processus évolutifs, à travers les répétitions des séquences, les hétérogénéités de compositions, ou la présence de gènes fossiles non fonctionnels. La comparaison de génomes constitue une source d'information encore plus riche car elle confère une vision globale des fondements génétiques des différences entre organismes. C'est dans ce cadre que va s'inscrire le travail présenté aux cours des chapitres suivants. Une approche markovienne pour l'analyse de la structure spatiale de génomes, et notamment l'analyse de la structure en isochores des génomes est développée.

1.2 Aspects mathématiques des modèles de Markov pour l'analyse de séquences

Ce paragraphe décrit les principaux aspects mathématiques de la mise en œuvre des modèles de Markov et leurs principales utilisations dans le cadre de l'analyse de séquences génomiques.

1.2.1 Les chaînes de Markov

1.2.1.1 Modélisation par des chaînes de Markov

Une chaîne de Markov est un modèle probabiliste caractérisant la dépendance entre les observations successives d'une variable aléatoire. Dans le cas d'une séquence d'ADN, cette variable représente la nature de la base nucléique (c'est-à-dire A, C, G, ou T) et les observations successives correspondent aux positions successives sur la séquence. Ces modèles de Markov modélisent des séquences statistiquement homogènes. Une chaîne de Markov est couramment représentée par la collection de ses états (A, C, G, T) reliés par des flèches qui indiquent les probabilités de transitions d'un état vers un autre (Figure 1.4). À chaque flèche est associée la probabilité de transition pour qu'un nucléotide suive un autre nucléotide dans la séquence. Il est d'usage de ne pas faire apparaître de flèche lorsque la probabilité de

transition est nulle c'est à dire que la transition est "impossible".

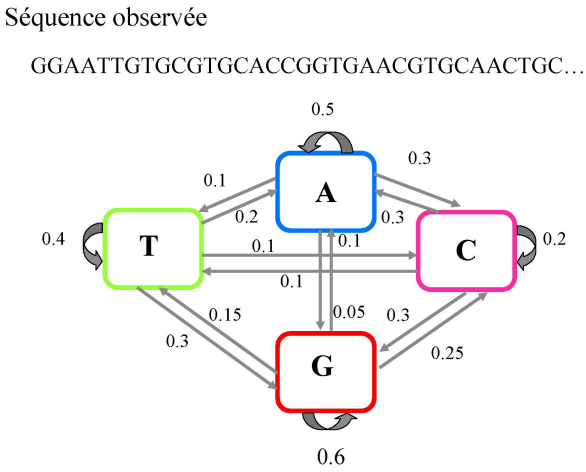


Fig. 1.4 : *Graphe représentant une chaîne de Markov. Chaque état représente une des 4 lettres (A, C, G ou T) qui constituent les séquences d'ADN.*

1.2.1.2 Expression mathématique d'une chaîne de Markov

Une chaîne de Markov homogène et d'ordre 1 sur un ensemble Q de N états est un modèle stochastique S à valeur dans Q. Une chaîne de Markov homogène possède la propriété suivante : la loi de S_{t+1} à la position $t+1$ dépend seulement de la valeur de S_t occupé à la position t. Les paramètres d'un modèle de Markov sont :
la matrice stochastique a décrivant les probabilités de transitions entre états :

$$a(u,v) = P(S_{t+1} = v \mid S_t = u), \qquad u, v \in Q^2, \qquad \sum_{v \in Q} a(u,v) = 1,$$

la distribution initiale (π) :

$$\pi : Q \to [0,1],$$

avec

$$\sum_{Q} \pi(u) = 1.$$

Les coefficients de la matrice stochastique ont les propriétés suivantes :

$$\begin{cases} a(u,v) \geq 0 \\ \sum_{v \in Q} a(u,v) = 1. \end{cases}$$

La chaîne de Markov (Q, π, a) homogène et d'ordre 1 est alors définie par la loi conjointe de chaque vecteur $(S_0, \ldots, S_N)$:

$$P(S_1 = s_1, ..., S_n = s_n) = \pi(s_1) \times a(s_1, s_2) \times ... \times a(s_{n-1}, s_n),$$

ainsi,

$$P(S_1 = s_1) = \pi(s_1)$$

$$P(S_{t+1} = s_{t+1} \mid S_t = s_t) = a(s_t, s_{t+1}).$$

Nous considérons uniquement les chaînes de Markov irréductibles, c'est-à-dire telles que n'importe quel état peut être atteint par la chaîne à partir de n'importe quel état, soit directement, soit en passant par des états intermédiaires.

Une chaîne de Markov d'ordre k est une généralisation de ce modèle, qui consiste à demander que la distribution de l'état S_t ne dépende que des k états précédents. On se donnera donc les probabilités de transitions :

$$P(S_t = s_t \mid S_{t-1} = s_{t-1}, ..., S_{t-k} = s_{t-k}) \qquad \text{pour} \quad \text{tous} \quad s_{t-k}, ..., s_t.$$

1.2.1.3 Application des chaînes de Markov à l'analyse de séquences

En 1993, Borodovsky propose un modèle de Markov ("Genmark") destiné à déterminer la nature codante ou intergénique d'une séquence x_1^n définie par $x_1^n = (x_1, x_2, ..., x_n)$ de longueur n avec $x_i \in \chi = \{A, C, G, T\}$. Pour cette séquence, deux modélisations sont proposées : l'une correspondant aux régions intergéniques, et l'autre aux régions codantes.

Modélisation des régions intergéniques

Sous le modèle qui tient compte de la composition en mono-nucléotides, la probabilité d'apparition de la séquence s'écrit :

$$P(x_1^n | inter) = b_0(x_1) \times b_0(x_2) \times ... \times b_0(x_n).$$

Modélisation des régions codantes

Dans les régions codantes, la fréquence d'apparition d'un nucléotide varie aussi selon la position de ce nucléotide dans le codon. Il est intéressant, notamment en vue de retrouver la phase de lecture, de prendre en compte la périodicité. La modélisation de la périodicité de composition des séquences

codantes nécessite d'introduire une loi d'apparition des nucléotides différente pour chacune des trois positions des codons. Les paramètres de ce modèle se séparent en trois jeux b_{+0}, b_{+1},b_{+2}, où b_{+i} décrit la fréquence d'apparition des nucléotides en position $+i$ des codons. Sous ce modèle, dit périodique, la probabilité d'apparition de la séquence x_1^n s'écrit de trois façons différentes selon la phase +0, +1, +2 dans laquelle la séquence code :

$$P(x_1^n|+0) = b_{+0}(x_1)b_{+1}(x_2)b_{+2}(x_3)b_{+0}(x_4)...$$

$$P(x_1^n|+1) = b_{+1}(x_1)b_{+2}(x_2)b_{+0}(x_3)b_{+1}(x_4)...$$

$$P(x_1^n|+2) = b_{+2}(x_1)b_{+0}(x_2)b_{+1}(x_3)b_{+2}(x_4)...$$

Ainsi, une séquence observée est supposée provenir soit d'une région intergénique, soit d'une région codante sur le brin étudié (+0, +1, +2), soit d'une région codante sur le brin complémentaire (−0, −1, −2). La formule de Bayes permet d'évaluer localement la nature codante d'une séquence à partir des fréquences attendues $P(i)$ de chacune des structures de la séquence :

$$P(i|x_1^n) = \frac{P(x_1^n|i)P(i)}{\sum_{j=-2}^{2} P(x_1^n|j)P(j)}.$$

Cet algorithme a été repris pour prédire les gènes chez Arabidopsis thaliana (plante de la famille des Brassicacées (Crucifères) à laquelle appartiennent de nombreuses espèces cultivées utilisées dans l'alimentation comme le chou, le navet, le radis, la moutarde...) (Mathé *et al*. 2000). Cette méthode a aussi été utilisée pour la recherche de régions promotrices chez les eucaryotes. Dans ce cas, une chaîne de Markov est adaptée à chacune de ces régions (Audic *et al*. 1997).

1.2.1.4 Généralisation : les semi-chaînes de Markov

Dans un modèle markovien classique, la durée de temps de séjour dans un état u est modélisée implicitement par une loi géométrique de paramètre $1-a(u,u)$ et d'espérance $1/(1-a(u,u))$. La probabilité $t_u(k)$ de rester dans

l'état u pendant exactement k pas de temps vaut

$$\begin{aligned} t_u(k) &= P(S_{t_0+1}^{t_0+k-1} = (u,...,u), S_{t_0+k} \neq u \mid S_{t_0} = u, S_{t_0-1} \neq u) \\ t_u(k) &= P(S_{t_0+1}^{t_0+k-1} = u^{k-1}, S_{t_0+k} \neq u \mid S_{t_0} = u) \\ t_u(k) &= P(S_{t_0+k} \neq u \mid S_{t_0+k-1} = u) \times \prod_{t=t_0+1}^{t_0+k-1} P(S_t = u \mid S_{t-1} = u) \\ t_u(k) &= (1 - a(u,u))\, a(u,u)^k, \end{aligned}$$

où $S_{t_0+1}^{t_0+k-1}$ représente la suite des états entre les temps t_0+1 et t_0+k-1.

Le modèle markovien implique donc, une distribution des temps de séjour dont l'histogramme est décroissant. En revanche dans un modèle semi-markovien la dépendance entre états n'est plus traduite explicitement dans la définition du modèle mais implicitement dans la définition des lois d'occupation des états. Les lois d'occupation des états sont par exemple des lois discrètes élémentaires (loi binomiale, loi de Poisson ou loi binomiale négative). Le mécanisme d'une semi-chaîne de Markov peut être interprété comme suit : à un instant donné t, on passe de l'état u à l'état v selon une loi de transition de l'état u puis on reste dans l'état v un temps t qui suit la loi d'occupation de l'état v. Enfin, on effectue une nouvelle transition conformément à la loi de transition de l'état v.

La modélisation d'une séquence par un modèle de Markov ou semi-markovien suppose l'homogénéité tout le long de cette séquence. Ces modèles ne prennent donc pas compte une éventuelle hétérogénéité pouvant exister le long de la séquence comme par exemple, la variation de composition en bases, ainsi que les différences de fonctions et de structures à l'intérieur du génome. Il est donc nécessaire de pouvoir disposer de modèles répondant à ce problème, c'est le cas des modèles de Markov cachés.

1.2.2 Les chaînes de Markov cachées

1.2.2.1 Modélisation par des chaînes de Markov cachées

Les modèles de chaînes de Markov cachées (HMM pour "Hidden Markov Model") sont utilisés pour rendre compte de l'existence de séquences distinctes et de natures différentes. Dans ces modèles, une variable qui correspond à la nature de la séquence est introduite. Cette variable est cachée,

au sens où elle n'est observée que par ses effets sur les propriétés de la séquence observée. La nature d'un segment de séquence dépend de la nature des fragments adjacents. C'est en prenant en compte ce contexte que l'information sur une position peut être enrichie.

Les HMMs modélisent ces dépendances sous la forme d'une chaîne de Markov, il s'agit d'un modèle très simple qui décrit uniquement les probabilités de transition « d'une nature à une autre ». Les différentes natures de séquences correspondent aux états cachés du modèle. À chacun des états cachés est associée une loi d'apparition des lettres de la séquence, dite aussi loi d'émission. Celle-ci peut être choisie avec une grande liberté, ainsi, les modèles markoviens sont souvent d'un grand intérêt car ils permettent de faire dépendre des quelques nucléotides précédents la probabilité d'apparition d'un nucléotide (c'est-à-dire de prendre en compte la composition en mots plutôt qu'en lettres).

Généralement, seules certaines transitions entre états cachés sont autorisées. Il est possible de les représenter sous forme d'une figure. La figure 1.5 correspond à un modèle simple qui décrit l'alternance des exons, introns et régions intergéniques le long d'une séquence d'ADN. Les HMMs permettent de modéliser très librement l'alternance de textures et de signaux le long de séquences. Ils ne se contentent donc pas de modéliser l'apparition de séquences conditionnellement à leur nature, mais modélisent aussi la nature elle-même de la séquence. Il peut paraître difficile de voir la position des régions codantes sur une séquence comme un événement aléatoire. Il faut néanmoins bien comprendre qu'il s'agit essentiellement de résumer des caractéristiques des séquences, comme la distribution des longueurs des régions codantes et intergéniques.

Les problèmes qui se posent, une fois la structure du modèle choisie, sont le choix des paramètres du modèle et la reconstruction de la suite d'états cachés associée à une séquence observée. Quelques méthodes statistiques ont été développées pour faciliter le choix de la structure du modèle. Ce problème demeure beaucoup plus difficile que celui de l'estimation des paramètres, mais la simplicité d'interprétation de la structure HMM rend souvent assez facile la traduction des connaissances biologiques en une architecture de modèles appropriée à l'usage envisagé. On appelle ici architecture du modèle ce qui relève du choix de la forme du modèle : choix du nombre d'états cachés, des transitions autorisées et du modèle d'émission des observations.

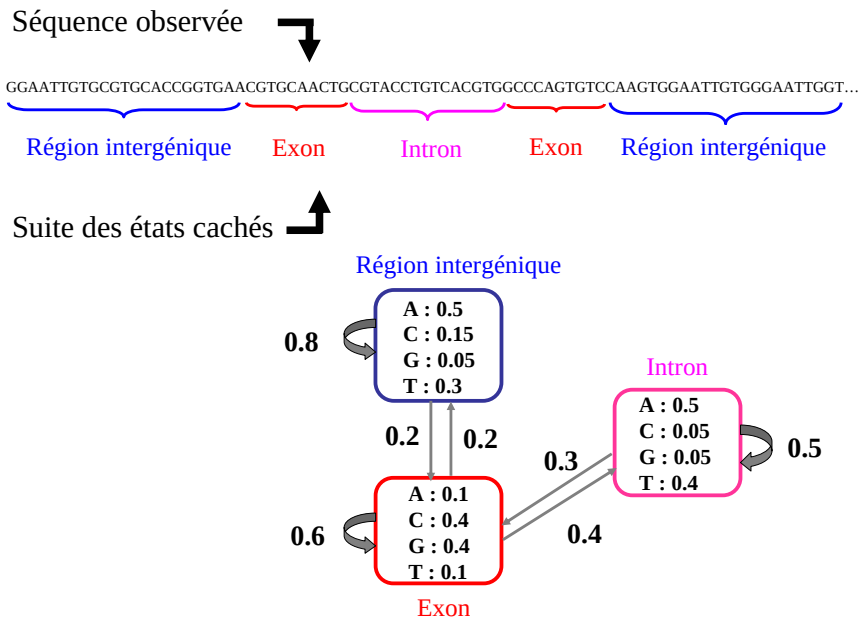


Fig. 1.5 : *États cachés d'un HMM simple modélisant l'alternance des gènes (exons et introns) avec la région intergénique. S={exon, intron,régions intergéniques} est l'ensemble des états représentés par des carrés.* $\chi = \{A, C, G, T\}$ *est l'ensemble des lettres que peuvent émettre les états. Les flèches représentent les transitions autorisées entre les états.*

1.2.2.2 Expression mathématique d'une chaîne de Markov cachée

Mathématiquement, les chaînes de Markov cachées permettent de modéliser deux processus emboîtés (Rabiner 1989, MacDonald et Zucchini 1997). Le processus observé correspond à la séquence observée. Le processus caché représente la structure sous-jacente de la séquence, que l'on cherchera à reconstruire.

Le processus caché

Le processus caché, noté $(S_t)_{t\in\{1,2...\}}$, correspond à la suite des états cachés. Les états cachés S_t prennent leurs valeurs dans un espace discret Q de cardinal fini $|Q|$. Ce processus est une chaîne de Markov homogène d'ordre 1 dont les paramètres, a, sont notés de la manière suivante :
la matrice de transition d'éléments :

$$a(u,v) = P(S_{t+1} = v \mid S_t = u), \qquad u,v \in Q^2, \qquad \sum_{v\in Q} a(u,v) = 1,$$

la distribution initiale :

$$a(u) = P(S_1 = u), \qquad u \in Q, \qquad \sum_{u\in Q} a(u) = 1.$$

Sous ce modèle, la probabilité d'apparition d'une suite de n états cachés particuliers $s_1^n = (s_1, \ldots, s_n)$ s'écrit

$$\begin{aligned} P(S_1^n &= s_1^n) = P(S_1 = s_1) \times P(S_2 = s_2 \mid S_1 = s_1) \\ &\quad \times ... \times P(S_n = s_n \mid S_{n-1} = s_{n-1}) \\ P(S_1^n &= s_1^n) = a(s_1) \prod_{t=1}^{n-1} a(s_t, s_{t+1}). \end{aligned}$$

Le modèle de chaîne de Markov est l'un des plus simples qui puisse être envisagé. Le processus a une mémoire très limitée puisque

$$P(S_t^n = s_t^n \mid S_1^{t-1} = s_1^{t-1}) = P(S_t^n = s_t^n \mid S_{t-1} = s_{t-1}).$$

L'information qu'apporte la connaissance du passé $(S_1^{t-1} = s_1^{t-1})$ sur le futur $(S_t^n = s_t^n)$ se résume uniquement à la connaissance du dernier état du passé $(S_{t-1} = s_{t-1})$ sur le présent. Symétriquement, cette propriété implique que

toute l'information apportée par la connaissance du futur $(S_{t+1}^n = s_{t+1}^n)$ se résume à celle du premier état du futur $(S_{t+1} = s_{t+1})$:

$$\begin{aligned} G &= P(S_1^t = s_1^t \mid S_{t+1}^n = s_{t+1}^n) \\ G &= \frac{P(S_1^t = s_t^t, S_{t+1} = s_{t+1}, S_{t+2}^n = s_{t+2}^n)}{P(S_{t+1} = s_{t+1}, S_{t+2}^n = s_{t+2}^n)} \\ G &= \frac{P(S_1^t = s_t^t, S_{t+1} = s_{t+1})}{P(S_{t+1} = s_{t+1})} \times \frac{P(S_{t+2}^n = s_{t+2}^n \mid S_{t+1} = s_{t+1})}{P(S_{t+2}^n = s_{t+2}^n \mid S_{t+1} = s_{t+1})} \\ G &= P(S_1^t = s_1^t \mid S_{t+1} = s_{t+1}). \end{aligned}$$

Par ailleurs, l'efficacité de la mise en œuvre des modèles de chaînes de Markov cachées est liée à ces propriétés d'indépendance conditionnelle. Bien souvent, on souhaite interdire certaines transitions entre états cachés, il suffit alors de fixer à zéro les probabilités de ces transitions $(a(u, v) = 0)$.

Le processus observé

Le processus observé, noté $(X_t)_{t\in\{1,2,\ldots\}}$, correspond à la séquence observée. Dans le cadre de la modélisation de séquences de macromolécules biologiques, il est à valeurs discrètes dans l'alphabet χ de taille 4 pour les acides nucléiques et 20 pour les acides aminés. L'apparition, dite aussi émission, des observations peut être modélisée comme uniquement dépendante de l'état caché sous-jacent. La séquence observée X_1^n est alors modélisée conditionnellement à la suite des états cachés (Figure 1.6) selon un modèle défini par des paramètres b, représentant la probabilité d'émettre la lettre x dans l'état u, de la forme :

$$b_u(x) = P(X_t = x \mid S_t = u), \qquad u \in Q, \quad x \in \chi \quad et \quad \sum_{x\in\chi} b_u(x) = 1.$$

Cependant, il peut être intéressant de prendre également en compte quelques observations précédentes grâce à un modèle markovien d'émission des observations (Figure 1.7) dont on notera b les paramètres

$$pour\ t > r_u, \quad b_u(x; w) = P(X_t = x \mid S_t = u, X_{t-r_u}^{t-1} = w)$$
$$u \in Q,\ x \in \chi,\ w \in \chi^{r_u}\ avec \sum_{x\in\chi} b_u(x; w) = 1$$

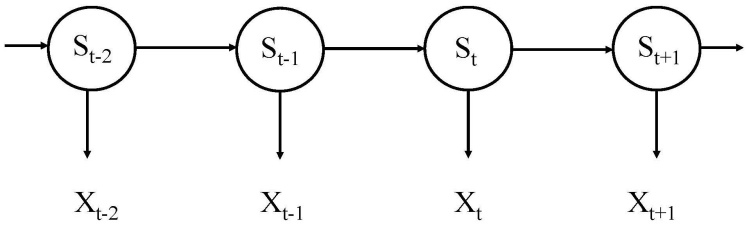


Fig. 1.6 : *Modèle d'ordre 0*

$$pour\ t \leq r_u, \quad b_u(x;w) = P(X_t = x \mid S_t = u, X_1^{t-1} = w)$$
$$u \in Q,\ x \in \chi,\ w \in \chi^{t-1}\ avec\ \sum_{x \in \chi} b_u(x;w) = 1,$$

où r_u correspond à l'ordre du modèle markovien d'apparition des observations conditionnellement à l'état caché u. Ce modèle d'émission des observations peut être étendu aux modèles markoviens à mémoire variable (où r_u est une fonction variable de w), voir par exemple (Salzberg *et al.* 1998), ou à n'importe quelle autre loi d'émission qui dépend du contexte ayant précédé dans la séquence observée. Dans la suite de ce travail, le modèle d'émission des observations sera considéré comme un modèle markovien dont l'ordre r_u est égal à r pour tous les états cachés, on parlera alors d'un modèle $M1-Mr$ (Churchill 1989, Muri 1998, Boys *et al.* 2000).
Cependant, il faut remarquer que si les S sont markoviens d'ordre supérieur à zéro, le début de chaque zone dépend d'une base d'une autre zone ce qui peut biologiquement être délicat.

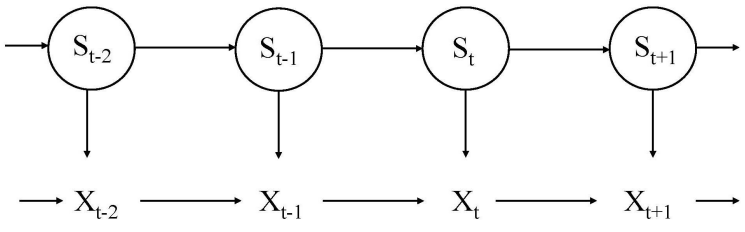


Fig. 1.7 : *Modèle d'ordre 1*

1.2.2.3 Analyse de séquences génomiques par des chaînes de Markov cachées

Les chaînes de Markov cachées ont largement été utilisées depuis le début des années 1990, pour l'analyse de séquences génomiques. Elles ont été développées pour des organismes aussi différents que les procaryotes et les eucaryotes. Deux grands thèmes ont été abordés : l'analyse de l'hétérogénéité des séquences mais aussi et surtout la prédiction de régions caractéristiques.

Churchill fut le premier à proposer d'utiliser les HMMs pour modéliser l'hétérogénéité des séquences d'ADN, en considérant ces dernières comme des segments homogènes où chaque type de segments est représenté par un état (Churchill 1989, Churchill 1992). Dans le même ordre d'idées, l'approche de Nicolas *et al*. (2002) développée pour la recherche des transferts horizontaux a permis l'identification de régions de composition atypique chez B. subtilis (bactérie).
Cependant, l'utilisation principale des chaînes de Markov cachées concerne la prédiction de gènes. Plusieurs logiciels ont été développés comme par exemple, Veil (Hendersen *et al*. 1997) ou HMMgene qui fut régulièrement réactualisé et adapté à différents organismes tel que E. Coli (Krogh *et al*. 1994), l'homme (Krogh *et al*. 1997) ou encore la Drosophile (Krogh *et al*. 2000). Les modèles HMMs sont très compétitifs lors de la recherche de régions promotrices (Pedersen *et al*. 1996, Nicolas *et al*. 2004), la détection des éléments transposables (Andrieu *et al*. 2004) qui repose principalement sur la sélection de modèles HMMs, la recherche de signaux particuliers tel que les peptides signaux (Nielsen *et al*. 1998, Barash *et al*. 2002, Zhang *et al*. 2003) ou encore lors de la modélisation protéique (Krogh *et al*. 1994).

Des modèles markoviens à mémoire variable ont été développés pour résoudre les problèmes d'estimations des paramètres liés aux ordres de plus en plus élevés des HMMs. Ces modèles ont été développés dans le cadre de la prédiction de gènes dans les génomes microbiens (Salzberg *et al*. 1998, Delcher *et al*. 1999), mais aussi pour la détection des régions promotrices chez les eucaryotes (Ohler *et al*. 1999).

1.2.2.4 Généralisation : Les chaînes semi-markoviennes cachées

Le cadre HMM restreint la modélisation du processus caché au modèle de chaîne de Markov. Cette restriction est à l'origine des propriétés d'indépendance conditionnelle du modèle de chaînes de Markov cachées présentées

dans la section précédente. Dans un modèle markovien classique, la durée du temps de séjour dans un état u est modélisée implicitement par une loi géométrique de paramètre $1-a(u,v)$ et d'espérance $1/(1-a(u,v))$ explicité lors de la section 1.2.1.4. Ce modèle implique notamment une distribution des temps de séjours dont l'histogramme est décroissant. Parmi les modèles alternatifs, le modèle semi-markovien est sans doute le plus utilisé en biologie, notamment pour la prédiction de gènes (Kulp *et al.* 1996, Burge *et al.* 1997, Besemer *et al.* 1999, Reese *et al.* 2000, pour ne citer que certains des algorithmes les plus connus). Le modèle semi-markovien (Kulkarni *et al.* 1997) généralise le modèle de chaînes de Markov en permettant la modélisation explicite des durées des séjours dans les états (voir section 1.2.1.4), les transitions entre états aux instants de sauts restant markoviennes. Le processus semi-markovien proprement dit est :

$$\underbrace{S_{T_1},\cdots,S_{T_2-1}}\,,\underbrace{S_{T_2},\cdots,S_{T_3-1}}\,,\underbrace{S_{T_3},\cdots}\,,$$

où les T_i sont les instants de sauts et $S_t = S_{T_i}$ pour tout t tel que $T_i \leq t < T_{i+1}$.

Il peut être défini comme l'emboîtement de deux processus : le processus des instants de sauts $\{T_i\}_{i\geq 1}$ à valeurs strictement croissantes dans $\mathbb{N}_+^*$; et le processus des transitions entre états $\{S_{T_i}\}_{i\geq 1}$ défini aux instants de sauts et à valeurs dans Q qui est une chaîne de Markov. Les paramètres (a,d) d'un modèle de chaînes semi-markoviennes sont de la forme :
pour la distribution initiale

$$\begin{aligned} T_1 &= 1 \\ a(u) &= P(S_{T_1} = u) \end{aligned}$$

puis

$$\begin{aligned} d_u(k) &= P(T_{i+1} = t_i + k \mid T_i = t_i, S_{T_i} = u) \\ a(u,v) &= P(S_{T_{i+1}} = v \mid S_{T_i} = u), \end{aligned}$$

où a correspond aux paramètres de la chaîne de Markov des transitions entre états définie aux instants de saut et d correspond aux lois de temps de séjour ($d_u(k)$ est la probabilité de rester un temps k dans l'état u).

Dans un modèle semi-markovien caché, le processus caché est une chaîne semi-markovienne (Rabiner 1989, Guédon *et al.* 2003). Les états cachés sont donc modélisés comme apparaissant par plages dont la longueur est égale au temps de séjour. Dans ce cadre, l'émission des observations conditionnellement aux états cachés peut bien sûr être modélisée comme dans un

HMM mais il est aussi assez naturel d'introduire des lois qui génèrent les observations par plages correspondant aux plages d'états cachés :

$$P(X_t^{t+k-1} = x_t^{t+k-1} \mid X_1^{t-1}, T_i = t, S_{T_i} = u, T_{i+1} = t + k).$$

Il est ainsi, par exemple, possible d'introduire des modèles d'émission périodiques pour la modélisation des séquences des régions traduites et des modèles hétérogènes pour celle des signaux (Burge *et al.* 1997, Lukashin *et al.* 1998).

1.2.2.5 Conclusion

L'utilisation des chaînes de Markov n'est pas appropriée à la modélisation des séquences biologiques constituées de plusieurs régions dans la mesure où elles modélisent uniquement les séquences statistiquement homogènes. Les modèles semi-markoviens cachés surmontent cette difficulté grâce à l'introduction d'un état pour chaque région. Ces modèles sont donc largement utilisés lors de la prédiction de gènes mais possèdent eux aussi certains inconvénients. Le coût de reconstruction du chemin optimal des états cachés dans un modèle semi-markovien caché associé à une séquence observée croît théoriquement comme le carré de la longueur de la séquence (éventuellement plus rapidement lorsque certaines lois d'émission des observations sont utilisées). En pratique, ces modèles ne sont utilisés pour de longues séquences que lorsque des simplifications, ou même des approximations sont possibles. On peut notamment citer la majoration des distributions de temps de séjour dont la loi n'est pas géométrique et, surtout, l'utilisation de pré-traitements pour trouver les positions possibles des changements d'états (Burge 1997). Une autre complication des modèles semi-markoviens cachés porte sur la multiplication du nombre de paramètres décrivant la distribution empirique des temps de séjour dans les états, qui doivent être estimés en plus des paramètres habituellement utilisés dans les HMMs (Rabiner 1989, Guédon *et al.* 2003).

À l'évidence chacune de ces classes de modèles (semi-HMM et HMM) présente des avantages et des inconvénients, le choix ne peut se faire qu'en fonction des données à modéliser et du type d'analyse recherchée. Nous avons choisi d'utiliser les chaînes de Markov cachées car elles permettent de représenter l'ensemble des régions caractérisant notre séquence grâce aux différents états cachés. De plus, les HMMs permettent la reconstruction du

chemin optimal des états cachés et le calcul de la vraisemblance par des algorithmes simples dont le coût croît linéairement avec la longueur de la séquence. Cet aspect est important car les études des chapitres 3, 4 et 5 portent sur des génomes entiers. Le chapitre 2 présente une méthode qui permet d'utiliser les HMMs sur des séquences statistiquement non homogènes, il présente également une méthode permettant de contourner la contrainte du temps de séjour dans les HMMs.

1.3 Algorithmes associés aux HMMs

Cette section présente les principaux algorithmes utilisés lors de la modélisation par HMMs.

1.3.1 Reconstruction d'un chemin optimal des états cachés

L'une des questions majeures lorsque les paramètres $\theta = (a, b)$ d'un HMM sont connus et la séquence observée donnée est de savoir comment reconstruire le chemin des états cachés associé à une séquence observée, qui représente les successions de différentes régions constituant la séquence. Le problème réside dans le nombre de chemins possibles qui peut atteindre $|S|^n$ si toutes les transitions sont autorisées, ce qui rend difficile voire impossible l'évaluation des chemins un à un. Il est nécessaire d'utiliser des algorithmes de reconstruction du chemin optimum d'états associés à une séquence observée. La difficulté réside dans le fait qu'il existe plusieurs critères optimaux possibles. Les algorithmes les plus souvent utilisés sont l'algorithme de Viterbi et l'algorithme de Forward-Backward (Rabiner 1989, Durbin *et al.* 1998). Ils envisagent la reconstruction du chemin caché sous des angles différents. L'algorithme de Viterbi reconstruit la séquence d'états qui a la probabilité maximale de générer la séquence. L'algorithme de Forward-Backward choisit pour chaque base de la séquence, l'état qui a la plus forte probabilité d'être obtenu connaissant la séquence indépendamment des états qui le précèdent ou le succèdent. Ce critère optimal maximise donc le nombre d'états individuellement corrects.

1.3.1.1 L'algorithme de Viterbi

Étant donnés une séquence observée $x_1 \dots x_n$ et un HMM de paramètres $\theta = (a, b)$, l'algorithme de Viterbi (Viterbi 1967) permet de trouver le chemin

le plus probable, conditionnellement à l'émission de x_1^n, $s^* = s_1^* \cdots s_n^*$:

$$s^* = \arg\max_{s_1^n} P(S_1^n = s_1^n \mid X_1^n = x_1^n)$$
$$s^* = \arg\max_{s_1^n} P(S_1^n = s_1^n, X_1^n = x_1^n),$$

ces deux formulations sont équivalentes car :

$$P(S_1^n = s_1^n \mid X_1^n = x_1^n) = P(S_1^n = s_1^n, X_1^n = x_1^n)/P(X_1^n = x_1^n)$$

avec x_1^n fixé.
L'algorithme de Viterbi est une méthode de programmation dynamique. Soit

$$v_t(u) = \max_{s_1^{t-1}} P(X_1^t = x_1^t,\, S_1^{t-1} = s_{t-1},\, S_t = u)$$

la probabilité du chemin le plus probable permettant de générer $x_1, \ldots, x_t$ qui se termine au site t dans l'état caché s_t. La formule de récurrence sur t est :

$$v_{t+1}(u) = P(X_{t+1} = x_{t+1} \mid S_{t+1} = u) \max_{w \in Q}[P(S_{t+1} = u \mid S_t = w) v_t(w)]$$
$$v_{t+1}(u) = b_u(x_{t+1}) \max_{w \in Q}[a(w,u) v_t(w)].$$

À chaque étape, le meilleur état est mémorisé dans la variable ψ :

$$\psi_t(u) = \arg\max_{w \in Q}[P(S_t = u \mid S_{t-1} = w) v_{t-1}(w)]$$
$$\psi_t(u) = \arg\max_{w \in Q}[a(w,u) v_{t-1}(w)].$$

La procédure complète pour trouver le chemin des états le plus probable peut maintenant être définie comme suit :

1. Initialisation
$$v_1(u) = a(u) b_u(x_1) \quad u \in Q$$
$$\psi_1(u) = 0,$$

2. Récurrence
$$v_t(u) = b_u(x_t) \max_{w \in Q}[a(w,u) v_{t-1}(w)] \quad 2 \le t \le T \quad u \in Q$$
$$\psi_t(u) = \arg\max_{w \in Q}[a(w,u) v_{t-1}(w)] \quad 2 \le t \le T \quad u \in Q,$$

3. Déroulement
$$P^* = \max_{w \in Q}(v_n(w))$$
$$s_n^* = \arg\max_{w \in Q}[v_n(w)].$$

4. La récurrence arrière retrace $s^* = (s_1^*, ..., s_n^*)$ en parcourant en arrière le meilleur chemin

$$s_t^* = \psi_{t+1}(s_{t+1}^*) \quad t = n-1, n-2, ...1.$$

La complexité de l'algorithme de Viterbi, au temps t, est proportionnelle à $m \times n$ où m est le nombre de transitions autorisées entre les états cachés et n est la longueur de la séquence. D'un point de vue pratique, les probabilités calculées étant très petites, il est préférable de travailler avec le logarithme des probabilités pour réduire les problèmes d'approximation numériques car les probabilités se multiplient et sont très petites.

1.3.1.2 L'algorithme de Forward-Backward

L'algorithme de Forward-Backward propose de choisir l'état s_t qui est individuellement le plus vraisemblable. Ce critère optimal maximise le nombre d'états individuellement correct attendu pour un modèle HMM. L'algorithme de Forward-Backward (Baum *et al.* 1970) calcule donc à chaque position t de la séquence, la probabilité de chaque état $(S_t = u)$ conditionnellement à la séquence observée :

$$\delta_t(u) = P(S_t = u \mid X_1^n = x_1^n) \quad u \in Q.$$

L'algorithme de Forward-Backward est un algorithme de programmation dynamique qui réalise une récurrence " avant et arrière " pour en déduire ensuite la valeur de $\delta_t(u)$.

La récurrence avant calcule, pour t croissant de 1 à n, la probabilité de la séquence partielle $x_1, \ldots, x_t$, et l'état $S_t = u$ au temps t conditionnellement :

$$\alpha_t(u) = P(x_1^t, S_t = u).$$

Cette probabilité est calculée par induction :

1. Initialisation

$$\alpha_1(u) = a(u) b_u(x_1) \quad u \in Q.$$

2. Induction

$$\alpha_{t+1}(v) = b_v(x_{t+1}) \sum_{u \in Q} \alpha_t(u) a(u, v) \quad 1 \leq t \leq n-1, \quad v \in Q.$$

3. Reconstruction

$$P(x_1^n) = \sum_{u \in Q} \alpha_n(u).$$

L'étape d'induction représente le cœur du calcul de la variable Forward, la figure 1.8 illustre cette étape, elle montre comment à l'état $S_t + 1 = v$, la probabilité $\alpha_{t+1}(v)$ est associée.

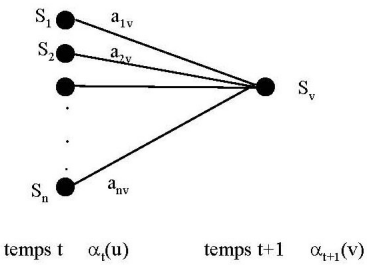


Fig. 1.8 : *Illustration de la récurrence de l'algorithme de Forward pour calculer la variable* $\alpha_{t+1}(v)$.

La récurrence arrière calcule la probabilité de la séquence partielle $x_{t+1}, \ldots, x_n$ conditionnellement à l'état $S_t = u$ au temps t :

$$\beta_t(u) = P_\theta(x_{t+1}^n \mid S_t = u).$$

Cette probabilité est calculée par induction (Figure 1.9) :

1. Initialisation

$$\beta_n(u) = 1 \quad u \in Q.$$

2. Induction

$$\beta_t(u) = \beta_{t+1}(v) \times \sum_{v \in Q} b_v(x_{t+1}) a(u, v) \quad t = n-1, ..., 1 \quad u \in Q.$$

La figure 1.9 illustre le calcul de la probabilité $\beta_t(u)$. Ainsi, il est possible d'obtenir la variable de Forward-Backward en fonction des variables $\alpha_t(u)$, $\beta_t(u)$:

$$\delta_t(u) = \frac{\alpha_t(u)\beta_t(u)}{P(x_1^n)} = \frac{\alpha_t(u)\beta_t(u)}{\sum_{u \in Q} \alpha_t(u)\beta_t(u)}.$$

La normalisation par $P(x_1^n)$ permet d'obtenir une mesure de probabilité $\sum_{u \in Q} \delta_t(u) = 1$. En utilisant $\delta_t(u)$, il est possible de trouver l'état s_t, au temps t, le plus vraisemblable :

$$s_t^{**} = \arg\max_{u \in Q} [\delta_t(u)] \quad \text{avec } t = 1 \ldots n.$$

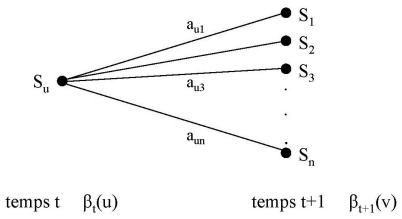


Fig. 1.9 : *Illustration de la récurrence de l'algorithme Backward pour calculer la variable* $\beta_t(u)$.

Il faut toutefois remarquer que lorsque certaines transitions entre états cachés sont interdites, les probabilités calculées grâce à l'algorithme de Forward-Backward peuvent ne pas permettre de proposer facilement un chemin unique compatible avec la structure du modèle. En effet, $a[s_t^{**}, s_{t+1}^{**}]$ peut être égal à 0.

La réalisation matérielle de cet algorithme est complexe car il faut faire coïncider dans le temps des données générées à des instants différents. Ce qui pose des problèmes de gestion de mémoire importants surtout dans le cas de l'analyse de génomes entiers et nécessite des efforts d'organisation des calculs (Dawid 1995).

1.3.2 L'estimation des paramètres d'un HMM

L'estimation des paramètres est généralement une étape incontournable lors de l'utilisation d'un modèle probabiliste pour analyser les données. Un estimateur est une fonction des observations vers l'espace de paramètres dont la valeur, si les données observées ont effectivement été générées par un modèle fixé, est "proche" du paramètre dérivant le modèle. Une fois la structure du modèle HMM définie, l'estimation des paramètres suit deux stratégies : soit la nature des séquences est connue, alors les matrices de transitions des émissions sont estimées séparément par maximum de vraisemblance (MV) ; soit la nature des séquences est inconnue, dans ce cas il est nécessaire d'employer la méthode EM (Expectation-Maximisation), celle-ci utilisant également le maximum de vraisemblance. Ces deux méthodes sont décrites dans ce paragraphe.
L'estimateur du maximum de vraisemblance défini sur les données complètes

est θ^{ML} :

$$\theta^{ML}(x_1^n, s_1^n) = \arg\max_\theta P_\theta(X_1^n = x_1^n, S_1^n = s_1^n),$$

ou sur les données incomplètes par

$$\theta^{ML}(x_1^n) = \arg\max_\theta P_\theta(X_1^n = x_1^n).$$

L'estimateur du maximum de vraisemblance est l'un des estimateurs les plus couramment utilisés car il possède de bonnes propriétés quand le jeu de données est assez grand. Il est notamment consistant, c'est-à-dire que les paramètres estimés tendent vers les paramètres réels lorsque la quantité d'observations augmente suffisamment (Baum *et al.* 1966, Gassiat *et al.* 2000).

1.3.2.1 Estimation par le maximum de vraisemblance

La vraisemblance sur les données complètes s'écrit :

$$\begin{aligned} P_\theta(X_1^n, S_1^n) &= a(S_1)\prod_{t=2}^{n} a(S_{t-1}, S_t)\prod_{t=1}^{n} b_{S_t}(X_t; X_{t-r}^{t-1}) \\ &= a(S_1)\prod_{t=1}^{r} b_{S_t}(X_t; X_1^{t-1})\prod_{u\in Q}\prod_{v\in Q} a(u,v)^{N(uv)} \\ &\qquad \prod_{u\in Q}\prod_{w\in\chi^r}\prod_{x\in\chi} b_u(x;w)^{N(wx;u)}, \end{aligned}$$

où le terme $a(S_1)\prod_{t=1}^{r} b_{S_t}(X_t; X_1^{t-1})$ correspond à la contribution des paramètres de lois initiales. Dans cette formule, $N(uv)$ désigne le comptage du nombre de transitions de u vers v dans S_1^n et $N(wx;u)$ le nombre d'occurrences du mot wx (de longueur $r+1$) dans X_1^n dont la dernière lettre x est dans l'état u.

$$\begin{aligned} N(uv) &= \sum_{t=1}^{n-1} 1[S_t = u, S_{t+1} = v] \\ N(wx;u) &= \sum_{t=r+1}^{n} 1[X_{t-r}^t = wx, S_t = u], \end{aligned}$$

où 1 est la fonction indicatrice : $1[S_t = u, S_{t+1} = v]$ vaut 1 si $S_t = u$ et $S_{t+1} = v$, 0 sinon.

Le logarithme de la vraisemblance est généralement utilisé :

$$\log(P_\theta(X_1^n, S_1^n)) = \log[a(S_1)\prod_{t=1}^{r} b_{S_t}(X_t; X_1^{t-1})] + A(s_1^n) + B(x_1^n, s_1^n)$$

où

$$\begin{aligned} A(s_1^n) &= \underbrace{\sum_{u\in Q}\sum_{v\in Q} N(uv)\log[a(u,v)]}_{(1)} \\ B(x_1^n, s_1^n) &= \sum_{u\in Q}\sum_{w\in\chi^r}\underbrace{\sum_{x\in\chi} N(wx;u)\log[b_u(x;w)]}_{(2)}. \end{aligned}$$

L'estimateur au maximum de vraisemblance est obtenu par maximisation séparée de chacun des termes désignés par une accolade sous la contrainte $\sum_{v\in Q} a(u,v) = 1$ pour les termes (1) et $\sum_{x\in\chi} b_u(x;w) = 1$ pour les termes (2). L'estimateur $\theta^{ML} = (a^{ML}, b^{ML})$ est donc :

$$a^{ML}(u,v) = \frac{N(uv)}{\sum_{v\prime\in Q} N_{uv'}} \tag{1.1}$$

$$b_u^{ML}(x;w) = \frac{N(wx;u)}{\sum_{x\prime\in\chi} N_{wx';u}}. \tag{1.2}$$

Il s'agit d'un estimateur très intuitif puisque, par exemple, la probabilité de transition de l'état u à l'état v $(a(u,v))$ est estimée par la fréquence empirique de v parmi les états apparaissant juste après u.

1.3.2.2 L'algorithme EM

L'algorithme EM est un algorithme itératif de maximisation locale de la vraisemblance des données incomplètes $(P_\theta(x_1^n))$ introduit par Demspter *et al.* (1977) dans les modèles à données manquantes (s_1^n). Il est aussi appelé algorithme de Baum-Welch (Baum *et al.* 1970) dans le cadre des HMMs. À partir d'un jeu initial de paramètres θ, chaque itération permet de remplacer le jeu de paramètres θ par un jeu de paramètres θ' qui augmente la vraisemblance :

$$P_{\theta'}(x_1^n) \geq P_\theta(x_1^n).$$

Définissons la log-vraisemblance conditionnelle du modèle θ par rapport au modèle θ', comme

$$L(\theta \mid \theta') = E_{\theta'}(\log P_\theta(x_1^n,\, s_1^n) \mid x_1^n),$$

alors l'étape E basée sur la valeur $\theta^{(m)}$ consiste à calculer

$$L(\theta|\theta^{(m)}).$$

L'étape M consiste à proposer $\theta^{(m+1)}$ comme

$$\theta^{(m+1)} = \arg\max_{\theta} L(\theta \mid \theta^{(m)}).$$

La justification de l'algorithme se fonde sur l'expression du logarithme de la vraisemblance des données incomplètes $\log(P_\theta(x_1^n))$ comme une différence d'espérances conditionnelles, d'après la formule de Bayes :

$$\log P_\theta(x_1^n) = \log P_\theta(x_1^n, s_1^n) - \log P_\theta(s_1^n \mid x_1^n).$$

En prenant l'espérance de chacun des termes sous la loi conditionnelle $P_{\theta'}(s_1^n \mid x_1^n)$, on obtient :

$$\log P_\theta(x_1^n) = L(\theta \mid \theta') - H(\theta \mid \theta')$$

où H est un terme d'entropie relative conditionnelle, défini par :

$$H(\theta \mid \theta') = E_{\theta'}[\log P_\theta(s_1^n \mid x_1^n) \mid x_1^n]$$

car $E_{\theta'}[\log P_\theta(x_1^n) \mid x_1^n] = \log[P_\theta(x_1^n)]$.
Ce choix de θ' assure la croissance de la vraisemblance des données incomplètes :

$$\log P_{\theta'}(x_1^n) - \log P_\theta(x_1^n) = L(\theta' \mid \theta) - L(\theta \mid \theta) - H(\theta' \mid \theta) + H(\theta \mid \theta) \ \geq 0,$$

puisque $L(\theta' \mid \theta) > L(\theta \mid \theta)$ par choix de θ' et $H(\theta' \mid \theta) < H(\theta \mid \theta)$ d'après l'inégalité de Jensen. Le calcul de $L(\theta' \mid \theta)$ repose dans le cade HMM sur l'algorithme de Forward-Backward :

$$\begin{aligned} L(\theta' \mid \theta) &= E_\theta[\log P_{\theta'}(x_1^n, s_1^n) \mid x_1^n] \\ &= E_\theta[\log P(a(s_1) \prod_{t=1}^{r} b_{s_t}(x_t; x_t^{t-1})) \\ &+ \sum_{u,v} N(u,v) \log a(u,v) + \sum_{u,w,x} N(wx;u) \log b_u(x;w) \mid x_1^n] \end{aligned} \tag{1.3}$$

Définissions maintenant, $n(uv, \theta)$ et $n(wx; u, \theta)$ par :

$$\begin{aligned} n(uv,\theta) &= \sum_t P_\theta(S_t = u, S_{t+1} = v \mid x_1^n) \\ n(wx;u,\theta) &= \sum_t 1[X_{t-r_u}^t = wx] P_\theta(S_t = u \mid x_1^n). \end{aligned}$$

À l'aide des probabilités de l'algorithme de Forward-Backward $\alpha_t(u) = P_\theta(X_1^{t-1}, S_t = u)$ et $\beta_t(u) = P_\theta(X_{t+1}^n | S_t = u)$, il est possible d'exprimer l'équation précédente (1.3) de la manière suivante :

$$L(\theta'|\theta) = E_\theta[\log(a(s_1)\prod_{t=1}^{r} b_{s_t}(x_t; x_t^{t-1})) \mid x_1^n] + \sum_{u,v} n(uv, \theta)\log a(u, v) + \sum_{u,w,x} n(wx; u, \theta)\log b_u(x; w). \quad (1.4)$$

Finalement, la nouvelle valeur $\theta^{(m)} = argmax_\theta L(\theta \mid \theta^{(m-1)})$ des paramètres s'obtient par une maximisation similaire à celle de la vraisemblance des données connues qui conduit à :

$$a(u, v, \theta^{(m)}) = \frac{n(uv, \theta^{(m-1)})}{\sum_v n(uv, \theta^{(m-1)})} \quad (1.5)$$

$$b_u(x; w, \theta^{(m)}) = \frac{n(wx; u, \theta^{(m-1)})}{\sum_w n(wx; u, \theta^{(m-1)})}. \quad (1.6)$$

Ces formules sont similaires à celle de la maximisation de la vraisemblance sur les données complètes (équations 1-1 et 1-2) en remplaçant les comptages par leur espérance conditionnellement à x_1^n et aux paramètres $\theta^{(m-1)}$ (ce qui correspond à l'exploration de tous les chemins cachés possibles puisque le chemin est inconnu).

En résumé, l'algorithme EM alterne une reconstruction du chemin caché grâce à l'algorithme de Forward-Backward avec une mise à jour des paramètres selon les équations 1.5 et 1.6. Il permet d'obtenir une suite $\theta^{(0)}, \ldots, \theta^{(m)} \ldots$ assurant la croissance de la vraisemblance. Cette suite converge vers un maximum local. Toutefois, si le point de départ $\theta^{(0)}$ est proche du maximum global alors la suite convergera vers le maximum global (Baum *et al*. 1970, Wu 1983, Muri 1997). En pratique, deux problèmes se posent lors de l'utilisation de cet algorithme. Le premier est celui du choix du critère d'arrêt. Une solution simple et raisonnable consiste à arrêter l'algorithme à l'itération M telle que $\log P_{\theta^{(M)}}(x_1^n) - \log P_{\theta^{(M-1)}}(x_1^n) < \epsilon$ où ϵ est choisi à l'avance. Le second, et le plus difficile à résoudre, consiste à savoir si la valeur $\theta^{(m)}$ obtenue est proche du maximum global. Il n'y a pas de réelle solution à ce problème. Lorsque cela est possible, une première solution pragmatique est de choisir un point de départ que l'on pense pas trop éloigné du maximum global. Une autre solution qui peut être complémentaire de la première, consiste à réaliser plusieurs fois algorithme EM avec des initialisations

différentes, par exemple aléatoire, puis à sélectionner après convergence les paramètres associés à la plus forte vraisemblance.

Conclusion

Ces vingt dernières années, la priorité a été donnée à la prédiction des gènes, difficilement identifiable de manière expérimentale en raison du grand nombre de données mises à disposition mais aussi du fait de leur faible représentation dans les génomes. Les processus de Markov sont largement utilisés pour aider à la localisation des gènes, leur efficacité n'étant plus à démontrer.

Aujourd'hui, avec le séquençage de nombreux génomes entiers, de nouvelles problématiques sont apparues. La principale problématique concerne des aspects évolutifs car si un génome est une source exceptionnelle d'information sur l'évolution des organismes vivants, la comparaison de génomes constitue une source d'information encore plus riche. Il s'agit en effet d'une approche qui confère une vision globale sur les fondements génétiques des différences entre organismes. Ainsi, l'objectif de cette thèse est de développer une approche markovienne pour l'analyse de l'organisation spatiale des génomes de plusieurs espèces, et particulièrement, pour la prédiction et l'analyse des isochores le long de différents génomes.

Chapitre 2

Exploration de la structure des gènes par modèles de Markov cachés

Les modèles de Markov ont, jusqu'à présent, été utilisés pour prédire la position des gènes (Burge 1997, Borodovsky 1993) ou d'autres régions telles que les régions promotrices (Pedersen *et al*. 1996, Nicolas *et al*. 2004) et les éléments transposables (Andrieu *et al*. 2004). L'objectif de ce chapitre est de développer des modèles de Markov cachés pour l'analyse de séquences biologiques. La stratégie développée au cours de cette section consiste à utiliser les erreurs de prédiction des modèles pour essayer de comprendre, d'analyser et de détecter de nouvelles propriétés biologiques des séquences d'ADN du génome humain. C'est donc une méthode d'exploration des génomes à grande échelle qui est proposée pour l'analyse de séquences biologiques, et en particulier des gènes. Lors du premier chapitre, le rôle capital des gènes a été mis en évidence. Ils assurent une fonction cellulaire précise, chacun d'eux portant les informations nécessaires pour la synthèse d'une ou plusieurs protéines. Dans un premier temps, une méthode destinée à contourner l'utilisation des modèles semi-markoviens cachés est décrite. Certains problèmes qui peuvent être générés par l'algorithme de Viterbi sont mis en évidence et une solution alternative est proposée. Dans un deuxième temps, ce chapitre présente une étude de la structure des gènes humains à partir d'une sélection de modèles HMMs qui conduit à la mise en évidence de nouvelles propriétés biologiques. La notion clef de ce chapitre est donc l'utilisation des HMMs comme un outil

d'exploration et d'analyse des données et non de prédiction.

2.1 Modélisation des distributions de longueurs des régions composant le gène

2.1.1 Introduction

Le séquençage complet du génome humain a aboutit à la connaissance d'une séquence d'environ trois milliards de paires de bases (International Human Genome Sequencing Consortium, 2001). La recherche expérimentale et l'analyse de l'ensemble des gènes protéiques dans cette masse de données n'est envisageable que de manière automatisée. Ces vingt dernières années, l'analyse mathématique et informatique des séquences d'ADN a donc été une voie de recherche privilégiée (Stormo *et al.* 2000). Cependant, la modélisation des gènes chez les eucaryotes est plus complexe que chez les procaryotes. Les difficultés de représentation sont principalement liées à l'alternance des introns et des exons, qui constituent respectivement les parties non codantes et codantes du gène. Cette dernière ne représente que 1 à 3% du génome humain. La présence de nombreux signaux rend leur prédiction encore plus difficile (Burge et Karlin 1997). De plus, la fréquence en $G + C$ influence la structure de la région (Duret *et al.* 1995, Chen *et al.* 2001). Ainsi, dans les régions riches en $G + C$ la densité en gènes est plus forte, les introns sont plus petits et la longueur de la partie codante du gène (CDS) est plus grande. Tous ces facteurs rendent plus complexe la modélisation des gènes.

Trois types d'approches markoviennes ont largement été développés (cf. chapitre 1) : les algorithmes utilisant les chaînes de Markov cachées (HMMgene de Krogh 1997, VEIL de Henderson *et al.* 1997), ceux employant les interpolés de Markov (GlimmerM de Salzberg *et al.* 1998) et enfin ceux utilisant les chaînes semi-markoviennes cachées (Genscan de Burge et Karlin 1997, GeneMark.hmm de Borodovsky et Lukashin 1998). Comme il a été décrit au chapitre précédent, le temps de séjour dans un état de ces modèles représente la longueur de la région. Lors de l'utilisation d'un HMM, le temps de séjour dans un état est décrit par une loi géométrique. Cette contrainte pose des difficultés lors de la modélisation des gènes car, contrairement à celles des régions intergéniques et des introns, la longueur des exons ne suit pas une loi géométrique, comme le montre l'allure en cloche de leur histogramme (Burge et Karlin 1997, Berget 1995, Hawkins 1988). Jusqu'à maintenant, la solution employée pour résoudre cette difficulté a consisté à utiliser des modèles semi-markoviens cachés. Dans ce cas, la durée du temps

de séjour dans un état est fonction de la distribution empirique de la longueur de la région modélisée (cf. chapitre 1). Ces modèles représentent donc de façon précise les gènes, mais leur utilisation est plus contraignante que les chaînes de Markov cachées pour deux raisons. D'une part, le nombre de paramètres est significativement plus important dans le cas des modèles semi-markoviens à cause de l'utilisation des distributions empiriques des longueurs. D'autre part, l'utilisation d'algorithmes classiques comme ceux de Baum-Welch (1970) ou de Viterbi (1967) est rendue plus complexe pour des modèles semi-markoviens cachés que pour les HMMs et nécessite des optimisations (Burge 1997). Ainsi, par exemple, la complexité des principaux algorithmes utilisés par les modèles semi-markoviens cachés (Forward-Backward et Viterbi) peut être quadratique par rapport à la longueur de la séquence. Ce facteur est limitant notamment lorsque l'analyse des données porte sur des génomes entiers. Il est alors nécessaire de fournir un gros effort d'optimisation. Un des avantages d'utilisation des chaînes de Markov cachées est la complexité des algorithmes qui est alors linéaire (Burge *et al*. 1998, Rabiner 1989, Guédon 2003).

L'objectif de cette étude consiste à employer les algorithmes et les modèles les plus simples possibles. Notre choix s'est donc porté sur les HMMs. Pour modéliser l'allure en cloche de la distribution empirique de la longueur des exons, nous proposons d'utiliser des sommes de plusieurs lois géométriques de paramètres égaux ou différents.

2.1.2 Matériel

Les données utilisées dans ce travail concernent uniquement le génome humain et sont extraites de la banque HOVERGEN (Duret *et al*. 1994)(mise à jour Mars 2003 release n° 43). HOVERGEN est une banque de données dédiée à l'analyse comparative des gènes homologues de vertébrés, elle est basée sur le système ACNUC. Elle contient l'ensemble des séquences nucléiques de vertébrés publiées dans GenBank. En se basant sur des critères de similitude entre séquences protéiques, les séquences sont classées par familles de gènes homologues. Pour chaque famille, les alignements multiples sont calculés et les arbres phylogénétiques optimaux sont construits.

Pour s'assurer de l'exactitude des données concernant l'organisation en introns et exons dans les gènes, nous avons restreint notre analyse aux gènes dont le transcrit d'ARN a été séquencé. De plus, pour éviter des biais lors

de nos analyses statistiques, les redondances expérimentales et celles dues à la présence de grandes familles de gènes ont été écartées. Cette procédure nous a fourni un jeu de données constitué de 5034 gènes multi-exons et 817 gènes sans introns. Au cours de cette étude, les introns sont localisés uniquement entre les exons codants car les régions UTR n'ont pas été séparées de la région intergénique. Les caractéristiques statistiques des régions codantes varient fortement suivant le contenu dans le gène de la fréquence en $G + C$ en position trois dans le codon ($G + C_3$) (Mouchiroud *et al.* 1991, Duret 1995). Trois classes sont définies, en fonction de la composition en $G + C_3$ des gènes. Notre jeu de données est donc divisé en trois classes contenant un nombre égal de gènes. Cette séparation est nécessaire afin de disposer de suffisamment de données dans chaque classe pour une bonne estimation des paramètres des modèles. Cette procédure a fourni une classe H (high) pour les gènes ayant une fréquence en $G + C_3$ supérieure à 72%, une classe L (low) pour les gènes dont le $G + C_3$ est inférieur à 56% et une classe M (medium) pour les gènes intermédiaires. Ces limites correspondent à celles employées couramment dans la littérature (Mouchiroud *et al.* 1991, Zouback *et al.* 1996) et seront utilisées au cours du chapitre 3 pour définir les limites des classes d'isochores. Chaque classe est divisée aléatoirement en deux parties égales, constituant le jeu d'entraînement et le jeu test des modèles. L'ensemble de ces calculs est réalisé avec le logiciel R.

2.1.3 Méthode

Une loi géométrique ne peut modéliser la distribution de la longueur des exons. L'originalité de la méthode proposée dans cette section consiste à ajuster à cette distribution empirique de la longueur des exons une convolution de plusieurs lois géométriques de paramètres égaux ou différents. Bien que cette approche ait déjà été suggérée (Durbin *et al.* 1998 page 69), elle n'a jamais été utilisée pour la modélisation des génomes.

Les distributions des longueurs des exons et des introns sont estimées à partir des échantillons $x_1 \ldots x_n$ des séquences des jeux d'entraînement. Les x_i sont considérés comme des réalisations de variables indépendantes de même loi. Les lois testées sont les suivantes :

– une somme de $m \geq 1$ lois géométriques de paramètre p (*i.e.* une loi binomiale négative) :

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

– une somme de deux lois géométriques de paramètres différents avec $p_1 > p_2$ (annexe B) :

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

– une somme de trois lois géométriques de paramètres différents avec $p_1 < p_2 < p_3$ (annexe B) :

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

L'étape suivante consiste à résoudre le problème d'ajustement. Parmi une famille de lois de probabilité, il est nécessaire d'établir la loi qui se rapproche au mieux de la distribution empirique observée sur un échantillon. L'estimation au maximum de vraisemblance est une méthode souvent proposée dans la littérature mais n'a pas été retenue car elle s'adapte assez mal à notre étude. Elle tend à surestimer les queues des distributions des longueurs des exons en délaissant de nombreux exons autour de la médiane de l'histogramme (Figure 2.1). Que ce soit pour une loi géométrique ou une convolution de lois géométriques, l'espérance est estimée par l'inverse du paramètre p ($E[X] = 1/p$) par la méthode du maximum de vraisemblance. Les valeurs extrêmes tendent donc à étirer la distribution vers les grands exons. Pour éviter cette erreur, notre choix s'est porté sur l'estimation à partir de la distance de Kolmogorov-Smirnov afin de modéliser plus précisément une majorité d'exons. L'éloignement de la distribution empirique par rapport à une loi théorique est quantifié en utilisant les distances entre lois de probabilité. La distance de Kolmogorov-Smirnov correspond à la distance de la norme uniforme entre fonctions de répartitions. Pour deux fonctions de lois de probabilité F et G, elle est notée $D_{KS}(F, G)$:

$$D_{KS}(F, G) = sup_{x \in R} |F(x) - G(x)|.$$

En pratique, cette distance est utilisée dans le cas où F est la fonction de répartition de la loi théorique, et G est la fonction de répartition de la loi empirique.

Ainsi, pour estimer les paramètres des différentes lois, la distance de Kolmogorov-Smirnov est minimisée pour chaque loi. La loi théorique considérée comme s'ajustant au mieux à la distribution empirique est celle qui

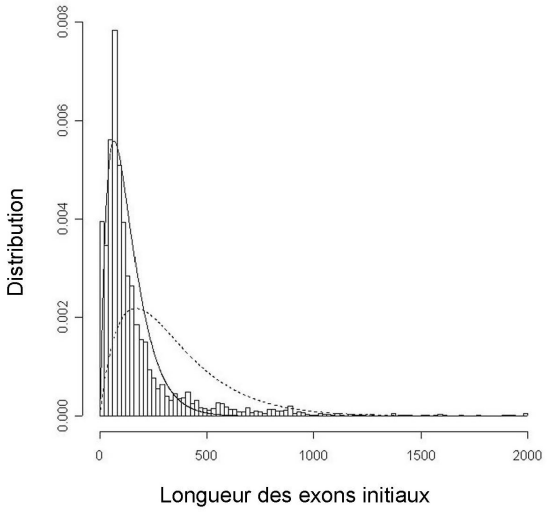


Fig. 2.1 : *L'histogramme représente la distribution empirique des longueurs des exons initiaux. La courbe pleine caractérise la distribution théorique obtenue à partir de la méthode de la distance de Kolmogorov-Smirnov et la courbe pointillée décrit la loi théorique obtenue par la méthode du maximum de vraisemblance.*

possède la plus petite distance de Kolmogorov-Smirnov. Toutefois, cette distance ne peut pas être minimisée par les algorithmes classiques de Newton ou du gradient car elle n'est pas dérivable. La distance a donc été minimisée graphiquement pour chacune des lois en utilisant une grille d'un pas de 10^{-5}. Les résultats des minimisations sont présentés dans le tableau 2.2 et l'annexe A.

En utilisant la distance de Kolmogorov-Smirnov, si la distribution de la longueur de la région est ajustée par une somme de n lois géométriques, l'état représentant la région est remplacé par une juxtaposition de n états qui ont les mêmes probabilités d'émission. Ce regroupement d'états est nommé macro-état (Figure 2.2). Le temps de séjour est caractérisé par les paramètres de la convolution de ces lois géométriques. Dans le but de favoriser des modèles simples mais efficaces, les transitions égales entre états d'un macro-état sont favorisées. Les distributions sommant le moins de lois géométriques sont favorisées afin de limiter le nombre de paramètres supplémentaires introduits dans le modèle. Ainsi, nous avons considéré que les distributions de longueurs théoriques sont correctement ajustées aux distri-

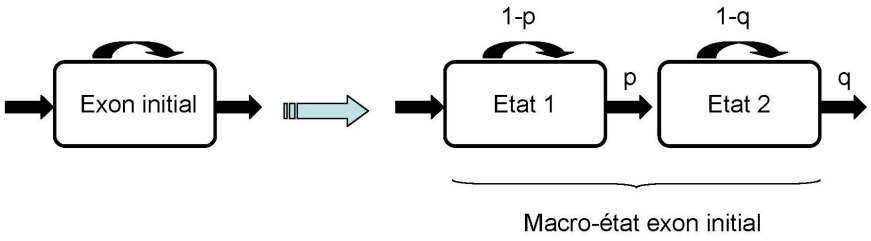


Fig. 2.2 : *Représentation du macro-état exon initial.*

butions empiriques des longueurs des exons et introns dès lors que la distance de Kolmogorov-Smirnov est inférieure à 2.10^{-4}. En effet, en dessous de cette valeur, le gain en précision dû à l'ajustement des distributions de longueurs est négligeable par rapport au gain du nombre de paramètres et donc en temps de calculs, comme le montre les figures 2.3 et 2.4.

De nombreuses études (Burge *et al*. 1997, Rogic *et al*. 2001, Chen *et al*. 2002) ont montré que la longueur des exons varie suivant leur position dans le gène. Notre étude a donc porté sur les quatre types d'exons : les premiers exons codants, les exons internes, les derniers exons et les exons uniques (gènes sans intron). Des comparaisons des longueurs entre les différentes régions ont été réalisées en utilisant le test non paramétrique de Wilcoxon.

2.1.4 Résultats

La longueur des exons dépend de leur position dans le gène. Ainsi, les exons initiaux et finaux tendent à être plus longs que les exons internes (Tableau 2.1). La longueur des introns varie également en fonction de leur position dans le gène. Les distributions de la longueur des introns internes et terminaux sont relativement similaires, de plus ces deux types d'introns sont tous deux beaucoup plus petits que les introns initiaux (Tableau 2.1). Comme le suggèrent certaines études, notamment Chen *et al*. 2002, la longueur des exons et des introns dépend du contenu en $G+C$ du gène . Le tableau 2.1 montre que la fréquence en $G+C$ en troisième position dans le codon est négativement corrélée avec la longueur des introns, *i.e.*, une forte fréquence en $G+C_3$ correspond à des introns courts et réciproquement. Les exons initiaux et finaux sont plus longs dans les régions riches en $G+C$. La longueur des exons internes ne varie pas en fonction du contenu en $G+C_3$.

Régions étudiées	Longueur en bp dans Classe H		Longueur en bp dans Classe M		Longueur en bp dans Classe L	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
exon initial codant	223	123	176	102	160	87
exon interne	144	126	143	125	144	120
exon final	244	165	237	145	218	138
intron initial	4027	3189	4139	3540	5315	4857
intron interne	1461	958	1767	1310	2850	2433
intron final	1394	884	1764	1282	2819	2415

Tab. 2.1 : *Longueur moyenne des exons et des introns suivant leur position dans le gène et la fréquence de $G+C$ en position 3 du codon.*

Les distributions de longueurs des exons montrent clairement des courbes en cloche pour les trois classes d'isochores (Figures 2.1 et 2.5, Annexe 1 (homme)). Les distributions théoriques des longueurs obtenues par minimisation de la distance de Kolmog- orov-Smirnov s'ajustent bien aux distributions empiriques des longueurs des exons et des introns (Figures 2.3 et 2.4 et Tableau 2.2).

Nous définissons $G_n(D_1, ..., D_n)$ comme la distribution d'une somme de n variables de lois géométriques, chacune d'espérance D_i et de paramètres $p_i = 1/D_i$. Ainsi, l'espérance de $G_n(D_1, ..., D_n)$ est $D_1 + D_2 + \cdots + D_n$. Quand $D_i = D$ pour tout i, cette loi correspond à une loi binomiale négative de paramètres (n,$1/D$), qui sera notée $G_n(D)$. Enfin, $G_1(D)$ est une loi géométrique d'espérance D et de paramètres $p = 1/D$, qui sera notée $G(D)$. Par souci de clarté, seuls les résultats des modélisations des distributions des longueurs des gènes appartenant à la classe H sont représentés au cours de cette section. Les distributions des longueurs des gènes des classes M et L ont également été modélisées par des sommes de lois géométriques et sont présentées en annexe A. Les estimations des distributions sont les suivantes : G_2(58.8, 74.7) pour les exons initiaux, G_3(86.2, 181.8,10) pour les exons finaux, G_5(26.3) pour les exons internes, G_2(1075.3, 106.4) pour les gènes sans introns et G(111.1) pour les introns initiaux. Les autres types d'introns sont également modélisés par des lois géométriques.

La distribution de la longueur des gènes sans intron met clairement en évidence une bimodalité (Figure 2.6). Par l'intermédiaire du logiciel Blast (Altschul *et al*. 1990) de recherche de similarité, les régions similaires à nos gènes annotés sans intron ont été recherchées le long du génome humain. Cette recherche a montré que de nombreux petits gènes annotés sans introns

Lois	Paramètres p	Distance K-S
$G_2(p)$	66,7	0.08775
$G_3(p)$	40,0	0.12161
$G(p_1, p_2)$	58,8 - 74,7	0.08621
$G(p_1, p_2, p_3)$	100,0 - 9,5 - 30,3	0.08611

(a) exon initial

Lois	Paramètres p	Distance K-S
$G_2(p)$	90,9	0.06592
$G_3(p)$	59,2	0.05569
$G(p_1, p_2)$	90,1 - 94,8	0.06991
$G(p_1, p_2, p_3)$	86,2 - 181, 8 - 10	0.04029

(b) exon final

Lois	Paramètres p	Distance K-S
$G_4(p)$	33,3	0.03220
$G_5(p)$	26,3	0.02242

(c) exon interne

Lois	Paramètres p	Distance K-S
$G_2(p)$	623,0	0.17466
$G_3(p)$	416,6	0,19721
$G(p_1, p_2)$	1075,3 - 106,4	0,13527
$G(p_1, p_2, p_3)$	754 - 446 - 198	0,13510

(b) exon unique

Tab. 2.2 : *Résultats des estimations des paramètres des différentes lois obtenus en minimisant la distance de Kolmogorov-Smirnov pour les différents types d'exons. (a) exon initial, (b) exon final, (c) exon interne, (d) exon unique. K-S est l'abbréviation de Kolmogorov-Smirnov.*

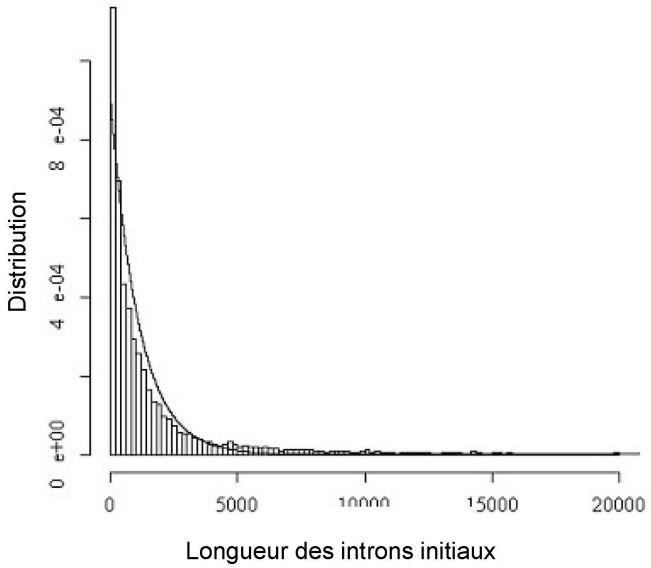


Fig. 2.3 : *Histogramme représentant la distribution empirique de longueurs des introns initiaux. La courbe pleine décrit la loi théorique obtenue par la distance de Kolmogorov-Smirnov.*

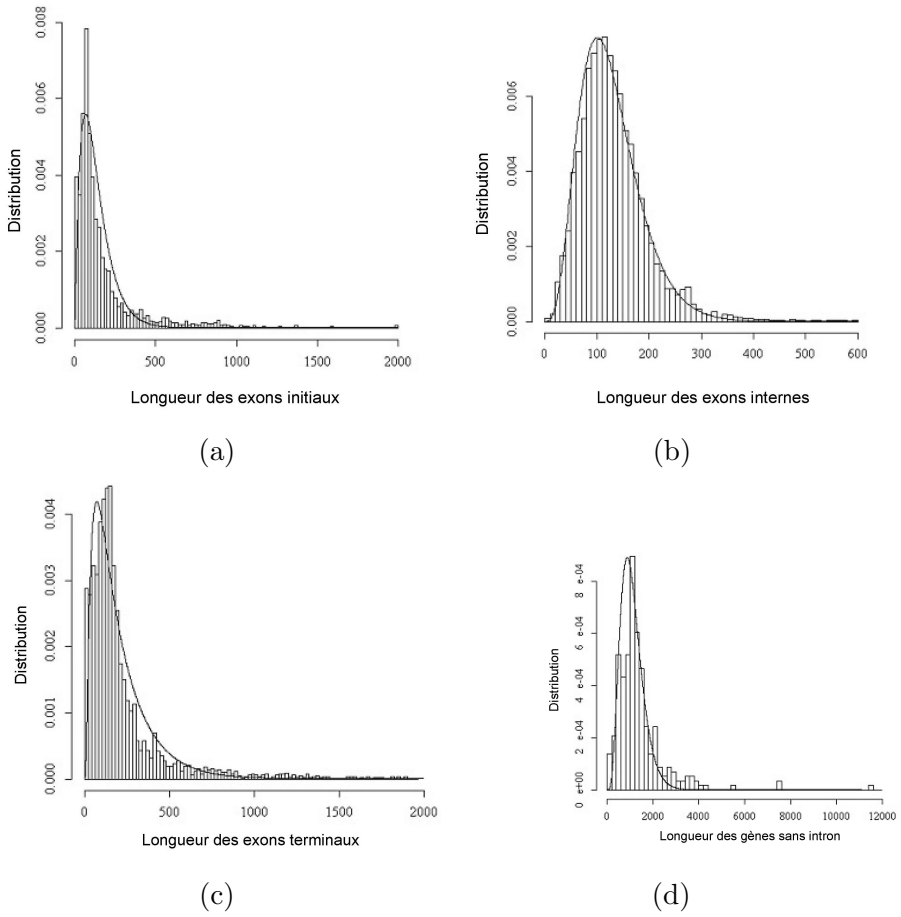
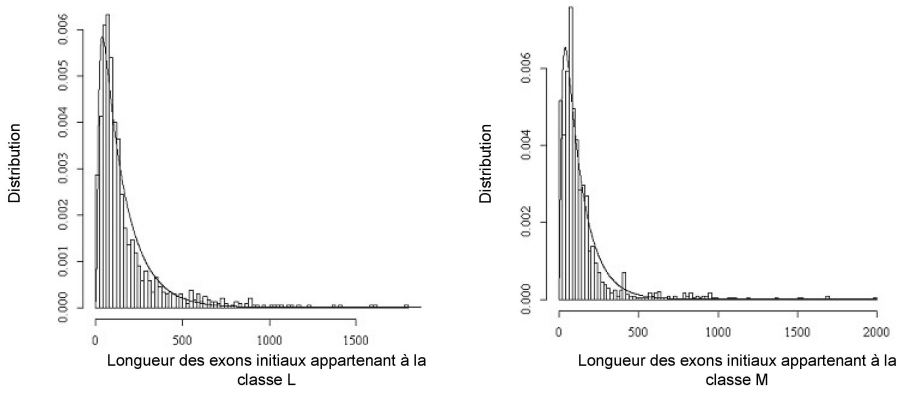


Fig. 2.4 : *Distribution empirique des longueurs des exons appartenant à la classe H suivant leur positions dans le gène. (a) exons initiaux, (b) exons internes, (c) exons finaux et (d) gènes sans intron. L'histogramme représente la distribution empirique des longueurs. La courbe pleine décrit la loi théorique obtenue par la distance de Kolmogorov-Smirnov.*



(a) (b)

Fig. 2.5 : *Histogramme représentant la distribution empirique de longueurs des exons initiaux : (a) appartenant aux classes L. (b) appartenant aux classes M. La courbe pleine décrit la loi théorique obtenue par la distance de Kolmogorov-Smirnov.*

se trouvaient fréquemment répétés le long du génome humain. Puis, ces petits gènes répétés ont été comparés aux pseudogènes contenus dans la banque Hopsygene (Khelifi *et al.* 2005). Cette comparaison a révélé que la plupart des petits gènes annotés sans introns (65%) correspondent à des pseudogènes, c'est-à-dire à des gènes qui ont perdu leur fonction. Après avoir éliminé ces faux gènes du jeu de données, la distribution de la longueur des gènes sans introns a l'allure d'une courbe en cloche tout comme les autres types d'exons (Figure 2.4 (d)). Elle a ainsi été modélisée par une somme de deux lois géométriques de paramètres différents.

2.1.5 Discussion

La première partie de ce chapitre a montré que la mise en oeuvre de modèles permettant une analyse à grande échelle du génome humain nécessite d'apporter une attention particulière à la modélisation des distributions des longueurs des différentes régions constituant un gène. Ainsi, il a été montré très clairement qu'il est possible, à partir de familles de lois constituées de convolutions de lois géométriques, de s'ajuster correctement aux distributions empiriques des exons. Alors qu'un HMM classique représente la région exon initiale par un seul état, la méthode proposée dans notre étude représente la région exon initiale par deux états. Le temps de séjour

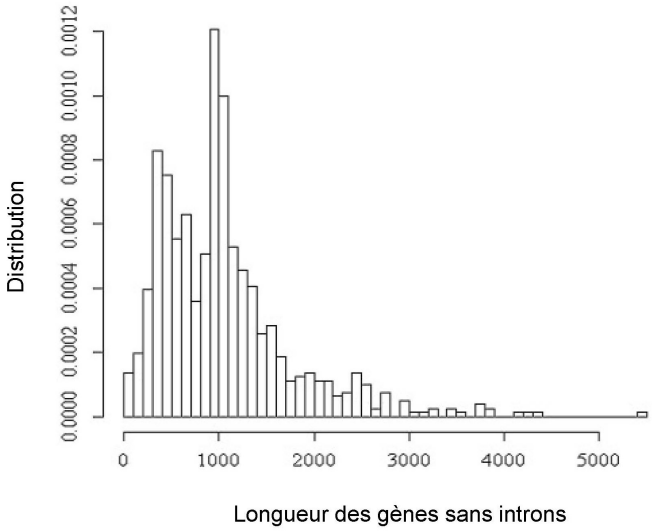


Fig. 2.6 : *Distribution empirique de la longueur des gènes sans introns de la classe H*

dans chacun des deux états correspond aux paramètres de la convolution de deux lois géométriques modélisant la longueur de l'exon. De plus, les probabilités d'émission sont identiques dans chacun des états. Cette méthode n'augmente pas le nombre de paramètres d'émission à estimer dans le modèle car les matrices juxtaposées représentent la même région. En revanche, elle augmente le nombre de paramètres correspondant aux transitions entre les états, mais celui-ci reste raisonnable. Ainsi, le nombre de paramètres rajoutés à la suite des juxtapositions correspond au maximum à trois paramètres différents pour les exons finaux et à cinq paramètres identiques pour les exons internes.

Cette étude a également permis de mettre en évidence les limites de la méthode du maximum de vraisemblance lors de l'estimation des paramètres. Pour la plupart des lois de probabilité usuelles, l'estimateur du maximum de vraisemblance est défini de façon unique, et se calcule explicitement. Sur le plan théorique, il présente de nombreux avantages. Sous des hypothèses vérifiées par de nombreux modèles courants, il est asymptotiquement sans biais et convergeant. De plus, sa variance est minimale. Cependant, dans le cas présent, cette méthode a tendance à surestimer les grands exons tout en sous estimant les petits exons. Ces erreurs d'estimation sont probablement

liées à l'estimation par la moyenne des paramètres pour les différentes lois proposées. Celle-ci est faussée par la présence de quelques longueurs d'exons extrêmes qui tendent à étirer la courbe vers ces valeurs. L'utilisation de la distance de Kolmogorov-Smirnov permet de minimiser l'écart entre les fonctions de répartition théorique et empirique ce qui constitue une meilleure approche du problème dans notre cas particulier.

Les prédictions des modèles markoviens et semi-markoviens obtenues par des algorithmes récents (voir chapitre 1) sont efficaces mais de nombreux problèmes subsistent. En effet, il leur est toujours difficile de prédire les petits exons (<75 bp), les premiers et les derniers exons ainsi que les gènes avec de nombreux exons (Rogic *et al*. 2001). Cette étude a montré que l'estimation des longueurs des exons à partir du maximum de vraisemblance délaisse les petits exons. Ainsi, les difficultés rencontrées par ces divers algorithmes lors de la prédiction des petits exons sont améliorés par une estimation de leur longueur à partir de la distance de Kolmogorov-Smirnov. La difficulté liée à la prédiction de gènes sans intron est due à la présence de nombreux pseudogènes, mal annotés dans les banques de données. La prise en compte de nombreuses propriétés biologiques est donc réalisée dans cette thèse pour améliorer les analyses. Si la différence de longueurs des exons selon leur position dans le gène est connue et utilisée, une généralisation aux introns permet d'augmenter la précision de prédiction des gènes avec plusieurs exons. Enfin, cette étude a mis en évidence une forte influence de la fréquence en $G + C$ sur la longueur des exons et des introns. La qualité des prédictions des premiers et derniers exons est améliorée en prenant en compte ces quelques remarques. Afin d'améliorer les méthodes existantes, il pourrait être intéressant de prendre en compte ces quelques propriétés qui ont été mises en valeurs au cours de cette étude. Toutefois, l'objectif de cette thèse n'étant pas la prédiction de gènes, nous n'avons pas poursuivi plus loin dans cette voie.

2.2 Comparaison des algorithmes de Viterbi et de Forward-Backward

2.2.1 Introduction

Au cours de la section précédente, l'importance de l'utilisation des macro-états lors de la modélisation des distributions de longueurs des différentes

régions a été mise en avant. Les algorithmes de Viterbi et de Forward-Backward, décrits au chapitre 1, sont largement employés aussi bien pour la sélection de modèles que pour la reconstruction des chemins optimaux. L'intérêt de chacun de ces deux algorithmes lorsque les modèles de Markov cachés sont constitués de macro-états est présenté dans la section suivante.

2.2.2 Sélection de modèles

Dans le but de mesurer l'adéquation de modèles avec une région génomique, la théorie des HMMs propose deux solutions. Il s'agit soit de calculer la probabilité d'observer une séquence conditionnée par la trajectoire optimale dans le modèle M (Viterbi), soit de calculer la probabilité de la séquence x sous le modèle M : $P[x \mid M]$ obtenu par l'algorithme de Forward.

2.2.2.1 Méthode

L'algorithme de Viterbi néglige le fait que plusieurs trajectoires peuvent être équivalentes. L'algorithme de Forward quant à lui somme des probabilités correspondant à des structures internes d'une séquence qui sont différentes. Ainsi, dans les deux cas, un modèle prédisant une mauvaise structure peut être associé à une forte probabilité. Pour illustrer simplement cette particularité, les deux techniques de sélection de modèles ont été comparées dans le contexte des HMMs pour un exemple simple.

Considérons un HMM de type $M1M0$, constitué de deux états, nommés A et B, et de deux observations, nommées 0 et 1. Supposons que les probabilités de transition de l'état A à l'état B et de l'état B à l'état A soient de t=9,53653.10^{-7}, et que l'état A et B émettent respectivement les lettres 0 et 1 avec la probabilité p. Nous noterons M_p le modèle HMM avec une probabilité p et nous comparerons les cas p=0,6 et p=0,9. Soit la séquence $x = 0^n1^n$ pour la valeur donnée n=10. L'objectif est de choisir entre le modèle $M_{0,9}$ et $M_{0,6}$.

2.2.2.2 Résultat

En utilisant l'algorithme de Viterbi, la probabilité de la séquence conditionnée par la trajectoire optimale du modèle HMM est calculée. Pour les deux modèles $M_{0,9}$ et $M_{0,6}$, la séquence optimale est composée de n états A puis n états B. En utilisant l'algorithme de Viterbi, le calcul de $P[x|s_{op}, M]$

pour chaque modèle donne :

$$P(x|s_{op}, M_{0,9}) = 0,1215 > P(x \mid s_{op}, M_{0,6}) = 3,65.10^{-5}.$$

Ainsi, le modèle $M_{0,9}$ est meilleur que le modèle $M_{0,6}$.
En utilisant l'algorithme de Forward-Backward, le calcul de $P(X|M)$ pour chaque modèle donne :

$$P(x \mid M_{0,9}) = 6,97.10^{-11} < P(x \mid M_{0,6}) = 1,27.10^{-6}.$$

Dans ce cas-là, le modèle $M_{0,6}$ est meilleur que le modèle $M_{0,9}$.
Cet exemple montre donc qu'à partir d'un modèle HMM simple ces deux techniques peuvent fournir des conclusions diamétralement opposées.

Dans un HMM représentant un macro-état, la situation est biologiquement plus simple. Toutes les trajectoires du macro-état sont biologiquement égales. Il apparaît donc clairement que la méthode des trajectoires optimales, correspondant à l'algorithme de Viterbi, n'est pas adaptée à ce problème. L'utilisation de algorithme de Forward-Backward, dans ce cas, est intéressante. Elle permet de décrire correctement la situation rencontrée par la probabilité de la séquence sous le modèle. En effet, les probabilités qui sont sommées correspondent à la même structure biologique.

2.2.3 Reconstruction du chemin optimal

Un des problèmes majeur lié à l'utilisation des modèles de Markov cachés est la reconstruction du chemin optimal des états. Les heuristiques employées pour résoudre ce problème ont été décrites précédemment au cours du chapitre 1 section 3. Ce paragraphe illustre, par un exemple concret, les limites de ces algorithmes lors de la mise en œuvre d'un HMM constitué de plusieurs macro-états (Figure 2.7).

L'algorithme de Viterbi permet de trouver le chemin caché le plus probable. Il a tendance à favoriser le passage à l'intérieur des macro-états possédant le moins d'états. L'exemple représenté sur la figure 2.7 (a) illustre ce problème. Il s'agit d'un HMM composé de 2 états simples et de 2 macro-états (exon interne et exon final). Dans un macro-état, chaque état possède les mêmes probabilités d'émission et de transition. Considérons une région de longueur lg. Pour modéliser cette région, le nombre de chemins possibles parcourant le macro-état exon interne est C^4_{lg-1}. Il est supérieur à celui du nombre de chemins possibles obtenus en parcourant le macro-état exon final

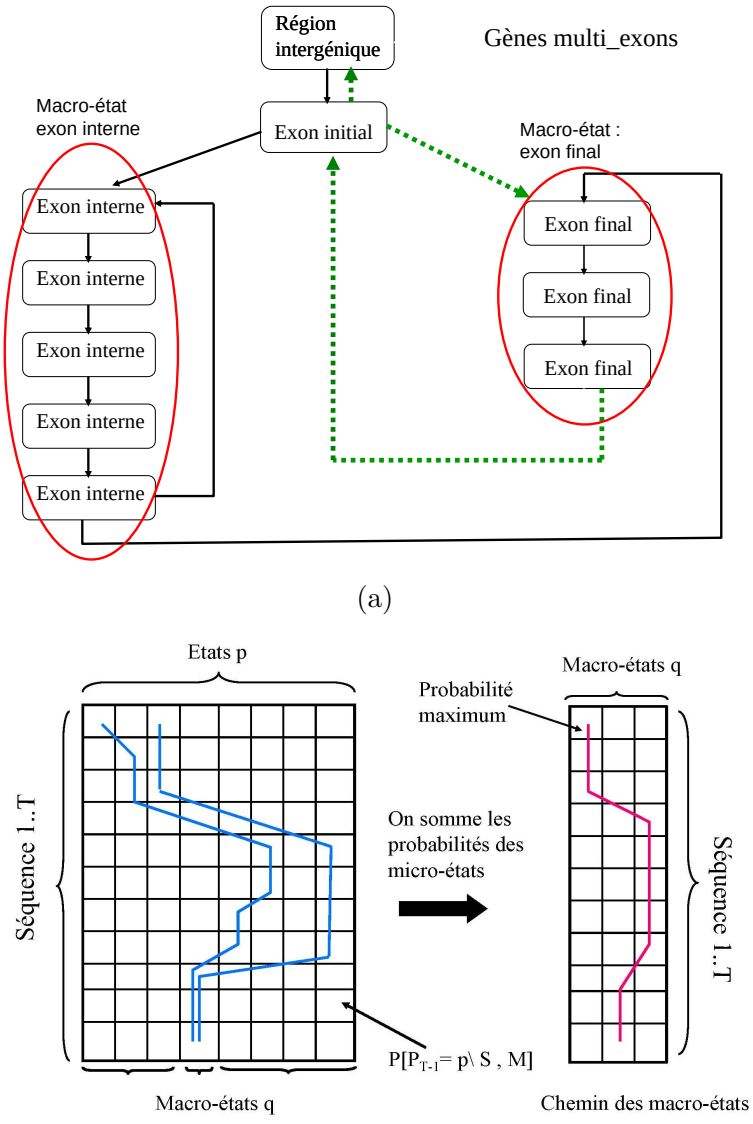


Fig. 2.7 : *Comportement des algorithmes de Viterbi (a) et Forward-Backward (b) avec un HMM constitué de macro-états.*

qui est C^2_{lg-1} . Le macro-état exon interne est fortement pénalisé. Dans le cas présent, l'algorithme de Viterbi emprunte préférentiellement le chemin vert au lieu du chemin noir (Figure 2.7 (a)). Le chemin optimal des " états " ne correspond pas au chemin optimal des macro-états. Les séquences optimales sont prédites sans prendre en compte la structure interne de l'HMM et donc les propriétés biologiques qu'elle contient.

L'algorithme de Forward-Backward présente également des inconvénients en raison de la présence de macro-états dans le modèle HMM. Cet algorithme permet de choisir l'état s_t qui est individuellement le plus vraisemblable. Il calcule donc à chaque position t de la séquence, la probabilité de chaque état $(S_t = u)$ conditionnellement à la séquence observée. Appliqué ainsi, l'algorithme de Forward-Backward tout comme l'algorithme de Viterbi renvoient un chemin qui ne correspond pas aux données biologiques représentées par la structure interne de l'HMM. La solution proposée dans le cadre de notre étude consiste à calculer la probabilité de chaque macro-état conditionnellement à la séquence observée. Ainsi, il est possible de reconstruire un chemin optimal qui prend en compte la structure en macro-états de l'HMM (Figure 2.7 (b)).

2.2.4 Discussion

Deux aspects de la modélisation ont été mis en avant. D'une part, lors de la sélection des modèles, il a été mis en évidence que différentes méthodes peuvent conduire à des résultats diamétralement opposés et ceci même sur des exemples simples. D'autre part, l'utilisation de macro-états permet une bonne modélisation des distributions de longueurs et ainsi la prise en compte de plusieurs propriétés biologiques dans notre modèle mais cet aspect implique l'utilisation d'algorithmes différents de ceux habituellement proposés pour les HMMs. Il a ainsi été montré que l'utilisation de l'algorithme de Viterbi, que ce soit pour la sélection de modèles ou pour la reconstruction des chemins cachés, ne permet pas de prendre en compte l'information biologique supplémentaire apportée par les macro-états. En revanche, l'algorithme de Forward-Backward peut aisément être adapté à cette situation. Nous avons donc opté pour l'utilisation de l'algorithme de Forward-Backward pour l'ensemble des calculs qui vont suivre.

2.3 Modélisation et analyse de la structure des gènes

2.3.1 Introduction

Les modèles de Markov sont couramment utilisés pour l'annotation complète des génomes, tel que Genscan par Ensembl, mais reste rarement employée pour l'exploration de données et l'interprétation biologique. Le séquençage de nombreux génomes entiers offre une étendue considérable de données mais requiert le développement de méthodes efficaces pour leur analyse et leur compréhension. L'objectif de cette section consiste à développer des modèles nécessitant peu de paramètres, une faible ressource de calculs, mais qui soient efficaces lors de l'analyse de génomes entiers.

La première partie de ce chapitre a montré que l'allure en cloche de la distribution des longueurs des exons peut être modélisée par des macro-états en utilisant des chaînes de Markov cachées. Ainsi, à chaque macro-état, qui représente une région du gène (exon initial, exon interne, exon terminal, intron...), correspond à un HMM. Au cours de cette section, une méthode destinée à l'analyse de la structure des gènes est développée. Elle repose sur une comparaison de modèles HMMs. L'originalité de cette méthode est d'utiliser des modèles très simples qui permettent facilement une analyse des aspects biologiques tout particulièrement en cas d'échec de la méthode afin de détecter des propriétés biologiques nouvelles. Cette capacité d'interprétation est originale, dans la mesure où les méthodes classiques fournissant des segmentations efficaces sont très opaques dans l'explication des raisons des succès comme des échecs (par exemple les réseaux neuronaux ou les méthodes de prédictions comme Genscan (Burge *et al*. 1997)).

2.3.2 Méthode

Une évaluation des différents modèles est nécessaire afin d'analyser des séquences biologiques au moyen de chaînes de Markov cachées. La méthode choisie consiste à diviser les séquences de nature connue en deux jeux de données, un jeu d'entraînement et un jeu test, afin de pouvoir comparer les prédictions des modèles entre eux.

Bien que la séquence d'ADN soit hétérogène le long du génome humain, elle peut être considérée comme une succession de régions homogènes, telles que les régions codantes et non codantes. Lors du premier chapitre, le rôle

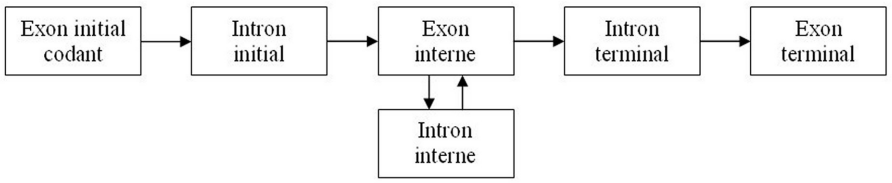


Fig. 2.8 : *Représentation de la définition d'un "gène" adoptée au cours de ce chapitre.*

capital des gènes a été rappelé. Ils assurent une fonction cellulaire précise, chacun d'eux codant pour la synthèse de protéines. Au cours de ce paragraphe, les efforts de modélisation et d'analyse sont portés sur l'unité " gène ". Dans cette étude, un gène est défini comme une alternance d'exons et d'introns. L'étude est restreinte aux exons codants, les régions UTRs ne sont donc pas modélisées dans cette partie (Figure 2.8).

Chaque région représentée par la figure 2.8 est modélisée par un HMM prenant en compte la distribution de sa longueur. Les exons sont constitués d'une succession de codons où chacune des trois positions du codon possède des caractéristiques statistiques spécifiques. Pour modéliser ces différences statistiques, l'état exon est séparé en trois sous-états caractérisant chacun la position du nucléotide correspondant dans le codon. Ainsi, la figure 2.9 représente l'HMM exon initial qui est constitué d'une juxtaposition de deux macro-états modélisant, par la somme de deux lois géométriques de paramètres différents, la distribution empirique de longueur des exons initiaux. À l'intérieur de chaque macro-état, les trois positions du nucléotide dans le codon sont représentées.

Les dépendances entre deux codons successifs sont modélisées à partir d'un modèle d'ordre 5 (Borodosvky *et al*. 1993, Burge *et al*. 1998). Ainsi, lors de l'émission d'une lettre par le modèle, les cinq lettres qui l'ont précédées sont prises en compte. Les jeux tests et d'entraînement sont les mêmes que ceux définis au cours de la première partie de ce chapitre. Les probabilités d'émission des HMMs sont estimées à partir des fréquences des mots de 6 lettres des séquences des jeux d'entraînement qui constituent le gène (intron, exon initial, ...). Un HMM est construit pour chaque région, les probabilités de transition et d'émission de ces HMMs étant estimées à partir des jeux d'entraînement.

Une comparaison systématique des HMMs a ensuite été mise en place

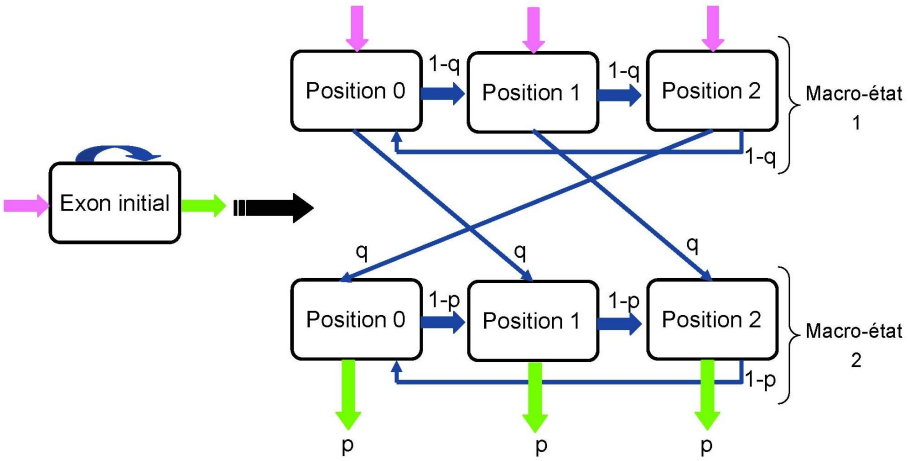


Fig. 2.9 : *Représentation de l'HMM exon initial qui modélise la distribution de longueur $G_2(1/q, 1/p)$ par deux macro-états ainsi que la position du nucléotide dans le codon pour chaque macro-état. Les flèches bleues représentent les transitions à l'intérieur du macro-état. Les flèches roses et vertes correspondent respectivement aux différentes entrées et sorties possibles dans le macro-état.*

pour analyser de façon plus précise la structure des exons et celle des introns. Dans le but d'identifier le modèle qui a la plus forte vraisemblance marginale, les HMMs sont comparés deux à deux pour chaque séquence d'une région donnée (intron, exon initial...) composant les jeux tests. La mesure de discrimination D est donnée par le rapport des vraisemblances :

$$D = \frac{P(S \mid HMM_1)}{P(S \mid HMM_2)},$$

où S, HMM_1 et HMM_2 désignent, respectivement, la séquence et les modèles testés. Une séquence est donc caractérisée par le modèle ayant la plus forte vraisemblance marginale. Chaque modèle est finalement défini par la fréquence avec laquelle il reconnaît les séquences. Cette approche originale permet de détecter et d'analyser la structure des séquences qui ne sont pas bien reconnues par le modèle sensé les représenter. L'objectif de cette section est de mettre en évidence des propriétés biologiques nouvelles.

Une analyse factorielle des correspondances est ensuite réalisée à partir des différentes régions afin de confirmer les résultats obtenus lors de la comparaison des modèles HMMs.

2.3.3 Résultats

2.3.3.1 Analyse du comportement des modèles exons

Ce paragraphe présente l'évaluation des HMMs modélisant les exons initiaux (M_EI), les exons internes (M_EInt) et les exons terminaux (M_ETer) de la classe H. L'ensemble des résultats, obtenus pour la classe H, a été vérifié pour les classes M et L. Il apparaît notamment que les modèles exon initial (M_EI) et exon interne (M_EInt) ne se différentient pas sur les séquences d'exons initiaux (Figure 2.10, colonne 1). Ce résultat implique que soit les deux modèles sont similaires, soit le modèle M_EI est mal ajusté aux séquences d'exons initiaux. La figure 2.10 résume les résultats des comparaisons obtenues pour la classe H. Plusieurs constations ressortent de ces calculs :

1. Les séquences d'exons internes et finaux ont les mêmes propriétés statistiques. L'évaluation des modèles M_EInt et M_ETer ne permet pas leur différentiation (Figure 2.10, colonnes 4 et 6).
2. Le modèle M_EI est significativement différent des modèles M_EInt et M_ETer. Le pourcentage de séquences d'exons internes en faveur

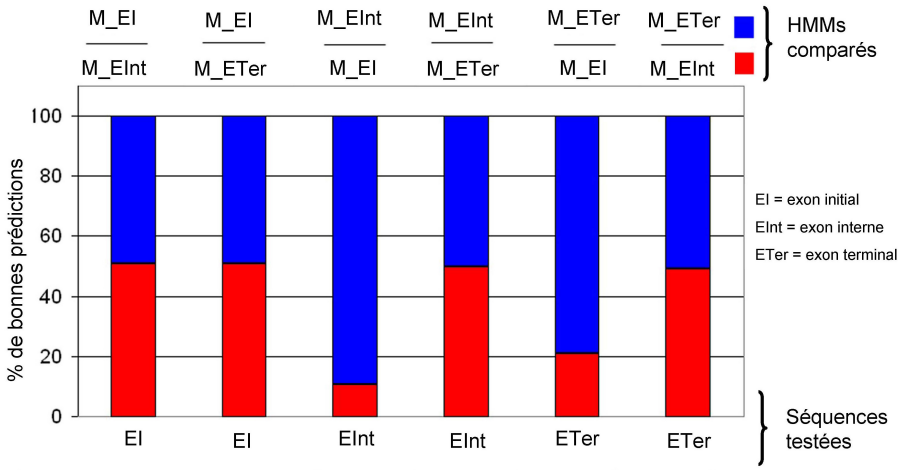


Fig. 2.10 : *Comparaison des HMMs sur les différents jeux de séquences test. Les modèles, appris sur les différentes séquences d'entraînement d'exons (exons initiaux, exons internes, exons terminaux), sont comparés deux à deux.*
Par exemple, sur la colonne 2, la vraisemblance marginale de chaque séquence d'exon initial est calculée par les modèles M_EI et M_ETer. La barre bleue représente le pourcentage de séquences d'exons initiaux ayant une plus forte vraisemblance marginale par le modèle M_EI. La barre rouge représente le pourcentage de séquences d'exons initiaux ayant une plus forte vraisemblance par le modèle M_ETer.

du modèle M_EInt et contre le modèle M_EI est de 92% (Figure 2.10, colonne 3). Des résultats similaires sont obtenus à partir des séquences d'exons finaux, le pourcentage de séquences d'exons finaux en faveur du modèle M_ETer et contre le modèle M_EI est de 79% (Figure 2.10, colonne 5).

3. Le modèle M_EI semble mal ajusté aux séquences d'exons initiaux, comme le montre le faible pourcentage (de l'ordre de 50%) de séquences d'exons initiaux qui sont préférentiellement reconnues par le modèle M_EI (Figure 2.10, colonnes 1 et 2).

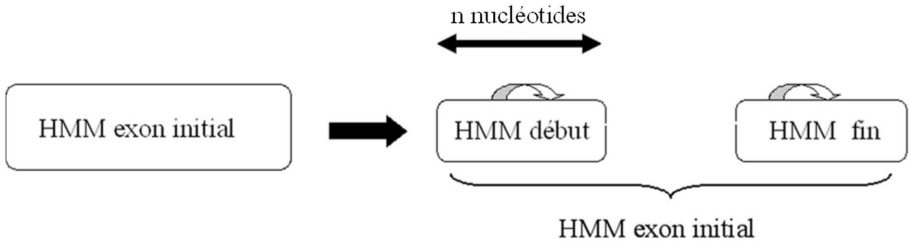


Fig. 2.11 : *Séparation du modèle HMM exon initial en deux HMMs . Ce modèle est appelé "M_EI_n". Le nombre de nucléotides contenu dans le premier HMM varie de 40 à 240 par pas de 40 nucléotides. Le second HMM contient le nombre de nucléotides restants.*

2.3.3.2 Étude des exons initiaux

La modélisation peu satisfaisante des séquences d'exons initiaux par le modèle M_EI peut être expliquée par l'existence de signaux localisés en début de gène. Ainsi, une rupture d'homogénéité à l'intérieur des premiers exons codants est susceptible d'avoir été créée par des signaux peptides. Afin de vérifier cette hypothèse, le modèle HMM représentant les exons initiaux (M_EI) est divisé en deux HMMs (figure 2.11). Le premier HMM caractérisant les n premiers nucléotides de l'exon initial, et le second caractérisant la fin de l'exon initial. Le nombre de nucléotides contenus dans le premier HMM varie de 40 à 240 par pas de 40 nucléotides afin de localiser le point de rupture à l'intérieur du premier exon codant. Le second HMM contient le nombre de nucléotides restants. Ce nouveau modèle est appelé M_EI_n. Des comparaisons deux par deux ont été effectuées entre les différents modèles M_EI_n (n=40, 80, 120, 160, 200, 240) et M_EI pour déterminer celui possédant le meilleur pourcentage de prédiction. La figure 2.12 montre que le modèle M_EI_{80} est celui qui caractérise le mieux les séquences des premiers exons codants. Ces résultats suggèrent que le point de rupture d'homogénéité à l'intérieur du premier exon codant se situe aux alentours du nucléotide 80.

Le modèle M_EI_{80} se différentie plus des modèles M_EInt et M_ETer que le modèle M_EI. La comparaison à partir des séquences d'exons initiaux passe de 49% (en faveur du modèle M_EI contre le modèle M_EInt) à 61% (en faveur du modèle M_EI_{80} contre le modèle M_EInt) (figures 2.9 et 2.10 colonne 1 et figure 2.12 colonne 7). De plus, la comparaison à

partir des séquences d'exons internes passe de 89% à 92% pour le modèle M_EInt contre les modèles respectifs M_EI et M_EI_{80}. Les mêmes différentiations entre M_ETer et M_EI_{80} sont obtenues que celles décrites entre M_EInt et M_EI_{80}.

Cette rupture d'homogénéité à l'intérieur du premier exon peut s'expliquer par la présence d'un peptide signal. Les protéines des eucaryotes doivent, après leur synthèse, atteindre leur cible finale qui peut être le cytoplasme, la membrane ou un organite quelconque. La plupart des protéines traversent des membranes, seules les protéines cytoplasmiques sont produites directement sur leur lieu d'utilisation. Le phénomène qui permet aux protéines d'atteindre leur cible finale est l'adressage. Toute protéine spécifique d'un organite débute par une séquence signal ou peptide signal spécifique de la cible. Il y a parfois plusieurs peptides signaux à la suite pour affiner la destination. Il s'agit d'une séquence consensus ; cela signifie que toutes les protéines ne débutent pas par cette même séquence, mais par une séquence qui lui ressemble fortement. Quelques acides aminés peuvent parfois être remplacés par d'autres acides aminés de la même famille. Cette séquence du peptide signal est alors éliminée très tôt ; en fin de synthèse elle n'existe déjà plus.

Pour vérifier l'hypothèse concernant la présence d'un peptide signal, les exons initiaux des jeux tests (définis en première section de ce chapitre) sont séparés suivant s'ils possèdent ou non un peptide signal, à partir du programme signalP (Nielsen *et al.* 1998). Deux jeux de 300 gènes chacun sont constitués : le jeu A contient les séquences d'exons initiaux possédant un peptide signal, et le jeu B contient les séquences d'exons initiaux sans peptide signal. Les modèles M_EI_{80} et M_EInt sont comparés sur les jeux A et B. Les exons initiaux qui possèdent un peptide signal sont reconnus dans 70% des cas par le modèle M_EI_{80} alors que ceux ne possédant pas de peptide signal ne différentient pas les modèles M_EI_{80} et M_EInt.

Une deuxième comparaison a été réalisée entre les modèles M_EI_{80} début et M_EI_{80} fin (voir figure 2.11) qui composent le modèle M_EI_{80}. Les séquences du jeu A sont reprises, mais cette fois seule la partie de l'exon contenant le peptide signal est extraite, ce nouveau jeu est nommé C. Les séquences du jeu C sont mieux reconnues par le modèle M_EI_{80} début que par le modèle M_EI_{80} fin dans 75% des cas. Le modèle M_EI_{80} fin reconnaît mieux les séquences du jeu test B dans 90% des cas par rapport

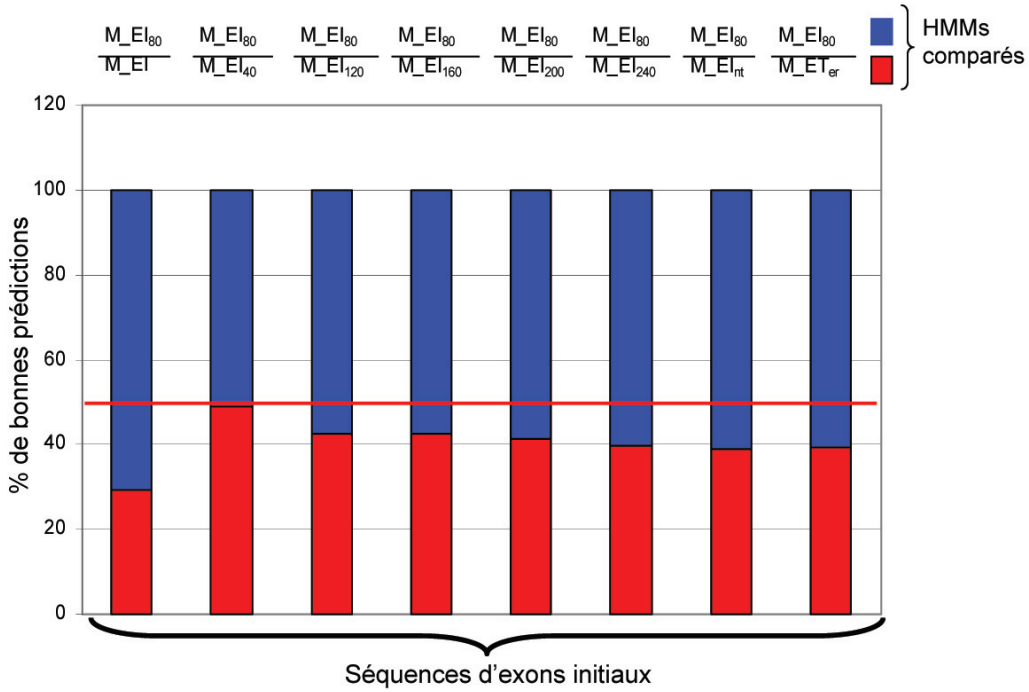


Fig. 2.12 : *Comparaisons des HMMs exons initiaux. Les modèles appris depuis les différents types de séquences (exons initiaux, internes et terminaux) sont comparés deux par deux sur les séquences d'exons initiaux pour identifier le point de rupture à l'intérieur du premier exon codant. Le modèle M_EI_{80} fourni la meilleur prédiction des séquences d'exons initiaux par rapport aux autres modèles testés.*

au modèle M_EI_{80} début. Enfin, les modèles M_EInt et M_EI_{80} fin ne se différentient pas sur le jeu B.

En conclusion, le modèle M_EI_{80} début caractérise le peptide signal. En revanche, le modèle M_EI_{80} fin représente la partie du premier exon ne contenant pas de peptide signal, celle-ci possédant les mêmes caractéristiques que les exons internes et les exons initiaux sans peptide signal.

2.3.3.3 Analyse du comportement des modèles suivant la classe d'isochore

L'importance de la prise en compte du contenu en $G + C$ lors de la modélisation des gènes a été décrite au cours de la première partie de ce chapitre où son influence sur la distribution de la longueur des exons et des introns a été montrée. Les résultats concernant l'influence du taux en $G + C$ sur les fréquences des mots de 6 lettres dans les différentes régions qui composent le gène, confirment le rôle important du contenu en $G + C$. Comme attendu, pour chaque type d'exons (initial, interne et terminal), le modèle entraîné sur une classe spécifique d'isochores est plus performant sur cette classe d'isochores que les modèles entraînés sur les autres classes d'isochores (Figure 2.13).

Les résultats concernant les introns sont différents. Les introns des classes H et M sont mieux prédits, respectivement, par les modèles HMMs Intron H et M (Figure 2.14, colonnes 1 à 4), alors que les trois modèles H, L et M intron sont plus ou moins équivalents sur les séquences d'introns appartenant à la classe L (Figure 2.14, colonnes 5 et 6). Cette analyse révèle clairement des différences statistiques majeures entre les trois classes d'isochores, et montre l'importance de prendre en compte cette hétérogénéité le long du génome humain dans le contexte de la prédiction d'isochores. La reconnaissance légèrement moins bonne des introns appartenant à la classe d'isochores L par le modèle intron L peut résulter de la présence d'éléments répétés, tels que les LINEs. En effet, ces derniers s'insèrent préférentiellement dans les isochores L.

De nombreuses méthodes d'exploration de données existent. L'analyse multivariée est l'une des méthodes les plus fréquemment employées. Ainsi, lorsque les séquences sont représentées par les fréquences des mots de 6 lettres, l'analyse factorielle des correspondances (AFC) prend en compte exactement les mêmes données que celles servant à l'estimation des para-

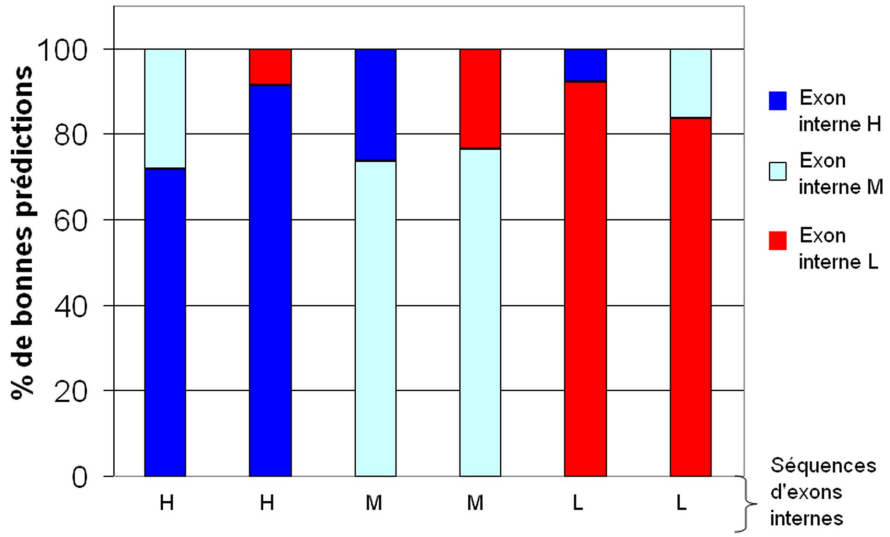


Fig. 2.13 : *Comparaison des modèles HMMs exons suivant la classe d'isochores. Les modèles appris depuis les séquences d'exons internes des classes H, L et M sont comparés deux par deux sur les jeux de séquences tests.*

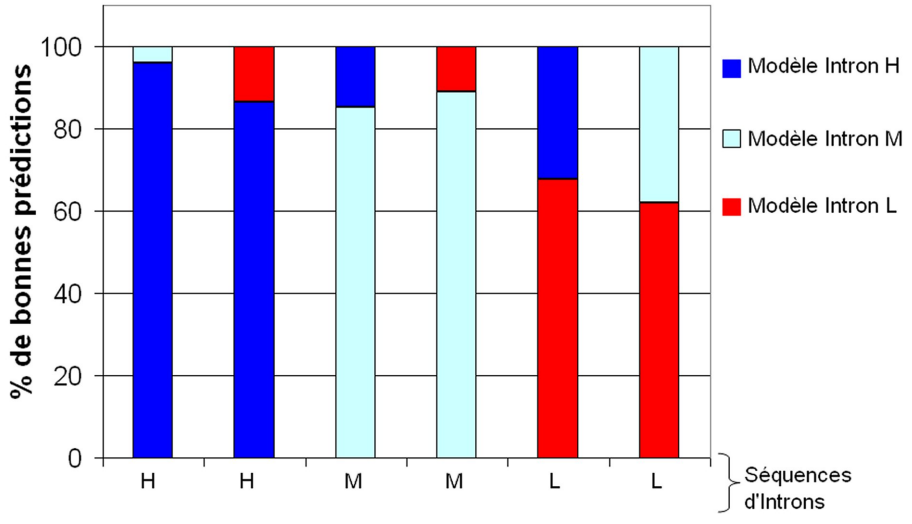


Fig. 2.14 : *Comparaison des modèles HMM introns suivant la classe d'isochores. Les modèles appris depuis les séquences d'introns des classes H, L et M sont comparés deux par deux sur les jeux de séquences testées.*

mètres des modèles HMMs. Les caractéristiques générales mises en évidence par l'AFC sont présentées dans les figures 2.15 et 2.16. Lors de la réalisation de l'AFC, les exons sont séparés selon leur phase de lecture. L'AFC divise nettement les exons et les introns en quatre groupes, les exons H, les exons L, les exons M et les introns (figure 2.15) et confirme ainsi l'importance du contenu en $G + C$ lors de la modélisation des gènes. Ces résultats ont été obtenus à partir des exons en phase 0. La même division a été obtenue pour les deux autres cadres de lecture des exons. De plus, lorsque les trois phases de lecture des exons sont prises en compte dans l'AFC (figure 2.16), le premier axe représente clairement la différence statistique qui existe entre les positions des nucléotides dans les codons.

2.3.4 Discussion

L'utilisation des modèles de Markov cachés semble avoir largement été sous estimée lors de l'analyse des génomes entiers, en effet, ces modèles sont principalement utilisés pour la prédiction. Ce chapitre a pu mettre en évidence la performance de tels modèles liée à leur simplicité et à la facilité de leur interprétation qui permet un retour à la biologie en cas d'échec de la méthode. Par exemple, la modélisation des premiers exons codants a permis de mettre en évidence la présence de peptides signaux. Ce modèle est facilement utilisable pour l'exploration des nouvelles propriétés biologiques. Cette approche a été appliquée aux gènes car ce sont les régions qui sont les plus étudiées et les mieux connues du fait de leur rôle lors de la synthèse des protéines. Cette démarche a ainsi permis de valider notre méthode. Les études des chapitres suivants vont utiliser les modèles développés dans cette partie pour analyser d'autres régions du génome dont les fonctions ne sont pas encore très bien connues. Cette approche repose sur une sélection de modèles HMM permettant la différentiation de plusieurs régions génomiques et l'analyse de leur structure interne pour la découverte de nouvelles organisations à l'intérieur de ces régions. Cette stratégie employée sur les régions codantes du génome humain a mis en évidence les particularités suivantes :

1. Une similarité entre les exons internes et finaux par l'intermédiaire des modèles M_EInt et M_ETer.
2. Les exons initiaux révèlent une organisation très spécifique, dû à la présence d'un peptide signal en début de leur séquence. Le temps moyen de séjour à l'intérieur du premier macro-état M_EI_{80} de 80 bases

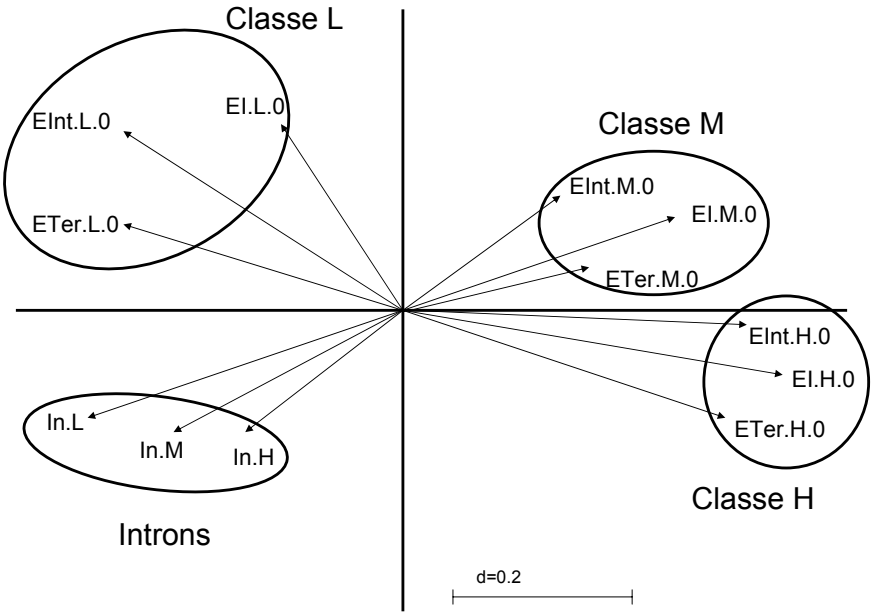


Fig. 2.15 : *Analyse factorielle des correspondances à partir des probabilités d'émissions des différents états des modèles qui constituent le gène (en phase 0 pour les modèles codants). L'axe des abscisses (36,2% de la variabilité totale) représente le gradient en $G+C$.*
EInt.H.0= modèle exon interne de la classe H en phase 0 ;
EInt.M.0= modèle exon interne de la classe M en phase 0 ;
EInt.L.0= modèle exon interne de la classe L en phase 0 ;
ETer.H.0= modèle exon terminal de la classe H en phase 0 ;
ETer.M.0= modèle exon terminal de la classe M en phase 0 ;
ETer.L.0= modèle exon terminal de la classe L en phase 0 ;
EI.H.0= modèle exon initial de la classe H en phase 0 ;
EI.M.0= modèle exon initial de la classe M en phase 0 ;
EI.L.0= modèle exon initial de la classe L en phase 0 ;
IN.H= modèle intron de la classe H ;
IN.M= modèle intron de la classe M ;
IN.L= modèle intron de la classe L

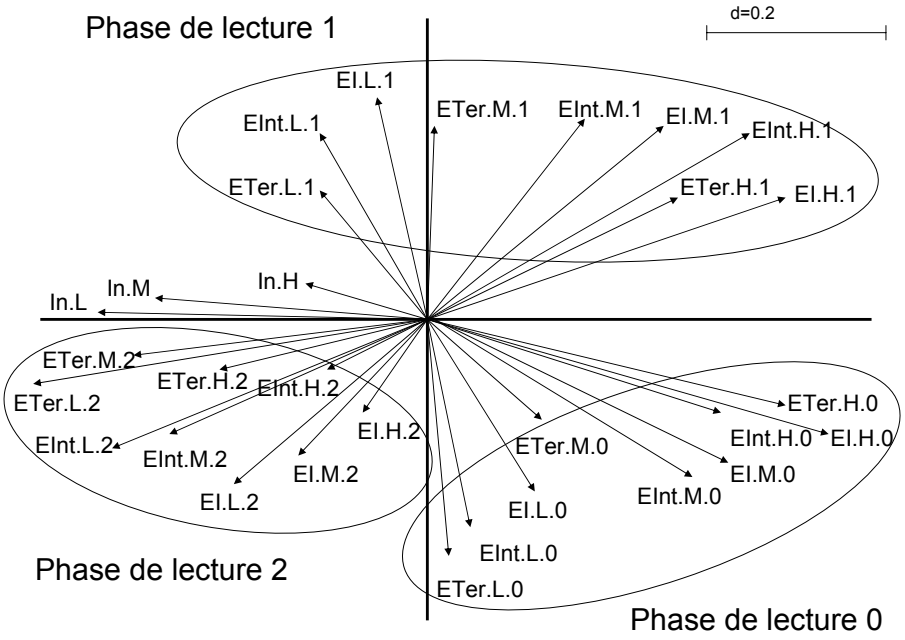


Fig. 2.16 : *Analyse factorielle des correspondances à partir des probabilités d'émission des différents états des modèles qui décrivent le gène. L'axe des abscisses (59,5% de la variabilité totale) correspond au gradient des phases.*

a été observé, et correspond à la longueur moyenne connue des peptides signaux (45 à 90 bp). Les exons initiaux sans peptide signal et la seconde partie des exons initiaux qui ont un peptide signal sont statistiquement similaires aux exons internes et terminaux comme le confirme le manque de différentiation entre les modèles M_EInt et M_EI_{80} fin.

3. Une différence statistique significative entre les trois classes d'isochores est mise en évidence par les modèles au niveau de la structure des exons mais aussi des introns. Cette hétérogénéité est confirmée par une AFC. Pour une modélisation efficace des génomes, il est donc indispensable de prendre en compte les informations concernant les différentes caractéristiques qui régissent l'organisation des isochores, ce que les modèles actuels ne font pas.
4. L'importance bien connue des phases de lecture dans les exons est confirmée par une AFC.

L'analyse multivariée est l'une des méthodes les plus anciennes et aussi l'une des plus utilisées lors de l'analyse de données. Elle emploie les mêmes données que celles servant à l'entraînement des modèles HMMs. Elle permet l'identification de motifs tout en étant en général beaucoup moins coûteuse en ressources CPU. C'est la méthode de prédilection employée par les anciennes méthodes de prédiction de gènes comme RECSTA (Fichant *et al.* 1987). L'approche markovienne possède toutefois de nombreux avantages. Tout d'abord, il n'est pas nécessaire de connaître les limites des régions avant leur analyse. Plus important encore, le modèle est moins versatile. Ainsi, de nouvelles hypothèses peuvent être introduites explicitement, comme nous l'avons montré avec la présence du peptide signal.

2.4 Conclusion

En résumé, trois résultats principaux ont été obtenus au cours de ce chapitre. D'une part, il a clairement été montré que la distribution de longueur de chacune des régions constituant les gènes peut être modélisée par des sommes de lois géométriques, ce qui offre une alternative à l'utilisation des algorithmes semi-markoviens. D'autre part, l'utilisation de l'algorithme de Forward-Backward est préférable à celui de Viterbi lors de l'introduction de macro-états dans un modèle HMM, dans la mesure où le chemin optimal

des macro-états ne correspond pas forcement au chemin optimal des états simples. Enfin, ce chapitre a mis en valeur la possibilité d'une exploration à grande échelle de génomes entiers à partir d'une sélection de modèles HMMs simples. L'originalité d'une telle approche réside dans la facilité d'interprétation des résultats ce qui permet un retour à la biologie en cas d'échec de la méthode, alors que les méthodes classiques fournissent des segmentations efficaces mais très opaques dans l'explication des succès comme des echecs.

La méthode développée au cours de ce chapitre peut avoir un double emploi : elle peut soit être utilisée dans un but de modélisation, soit dans un but d'analyse et de découverte du génome. Au cours des chapitres suivants de cette thèse, les modèles mis en place dans ce chapitre vont servir à la prédiction et à l'analyse de la structure en isochores de différents génomes. Ainsi, après avoir étudié une petite fraction seulement du génome humain (les parties codantes ne représentant que 1 à 3% du génome), notre approche est appliquée à grande échelle lors de l'exploration de génomes entiers de différentes espèces (homme, fugu , *Tetraodon nigroviridis*, poulet, chimpanzé et souris).

Chapitre 3

Prédiction et analyse des isochores du génome humain

Ce chapitre présente une méthode de prédiction et d'analyse des isochores du génome humain à partir de modèles de Markov cachés. La première partie décrit la notion d'isochore et les controverses qui lui sont liées. De plus, l'importance de cette structure au niveau de l'organisation du génome et la nécessité du développement de méthodes mathématiques pour sa localisation est mise en évidence. La deuxième partie détaille la mise en place d'un modèle HMM pour la prédiction des isochores le long d'un génome entier. La troisième partie se décompose en deux sections. La première est une étude qui présente la structure en isochores du génome humain obtenu par notre modèle HMM. La seconde section expose les propriétés biologiques liées aux isochores qui ont pu être mises en évidence par notre méthode.

3.1 Introduction

3.1.1 Définition des isochores

Le long des chromosomes de mammifères, il existe sur une échelle de quelques centaines de kilobases, une forte variabilité de la composition en bases $G + C$. Les régions d'un génome présentant cette structuration compositionnelle, décrite par le terme d'"isochores", ont été mises en évidence expérimentalement pour la première fois par des techniques de centrifugation en gradient de densité (Macaya 1976, Bernardi 1985). Ces techniques permettent de séparer des fragments d'ADN en fonction de leur composition en bases $G + C$. Ainsi, il a pu être montré que chez les mammifères et les oiseaux, le taux en $G+C$ de grands fragments génomiques (>300 kb) variait de 35% à 55% au sein d'un même génome. Cette forte variabilité en $G + C$ sur une échelle de quelques centaines de kilobases a été confirmée par les données de divers projets "génomes" (Fukagawa *et al*. 1995, Stephens *et al*. 1999, The MHC Sequencing Consortium 1999).

Le modèle proposé par Bernardi (2000) pour représenter la structure compositionnelle des génomes de mammifères repose sur l'idée qu'il existe de grandes régions de composition homogène en $G + C$ (les isochores), avec des limites bien marquées entre deux isochores successifs. Ainsi, les génomes de mammifères sont décrits comme une mosaïque d'isochores.

3.1.2 Controverses liées à l'existence des isochores

Les premières publications (Fukagawa *et al*. 1995, Stephens *et al*. 1999, The MHC Sequencing Consortium 1999) portant sur les grandes séquences du génome humain ont montré que certaines régions chromosomiques correspondent bien au modèle de Bernardi. Cependant, l'analyse de la séquence complète du génome humain a fait apparaître qu'en règle générale, il n'existe pas de frontière nette entre les régions de taux en $G + C$ différents, mais plutôt une variation continue de la composition en bases (Nekrutenko et Li 2000). Certains auteurs ont même considéré qu'il fallait rejeter le terme d'isochore, car il n'existe pas dans le génome humain de régions réellement homogènes telles qu'on en attendrait sous un modèle de séquences aléatoires (Lander *et al*. 2001, Häring et Kypr 2001). Il faut souligner qu'il existe tout de même bien une relative homogénéité locale de la composition en base. En effet, à l'échelle du gène, le taux en $G + C_3$ est fortement corrélé au

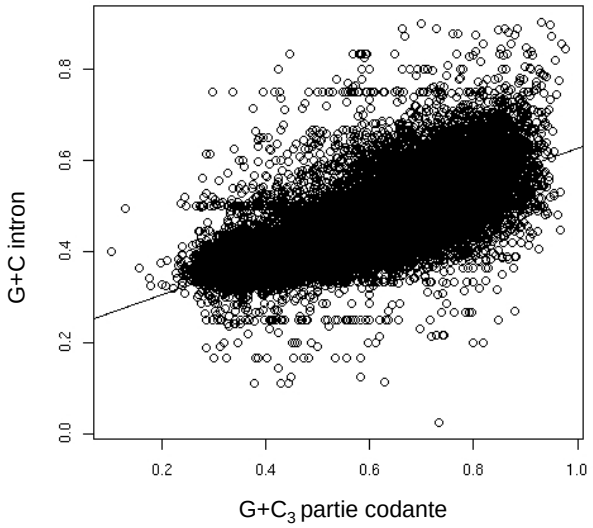


Fig. 3.1 : *Corrélation entre la composition en $G+C$ des introns et la composition en $G+C_3$ chez l'homme, R^2=0,56.*

taux en $G+C$ des introns (Figure 3.1) (Bernardi *et al.* 1985, Aïssani *et al.* 1991, Clay *et al.* 1996). Il existe donc une relative homogénéité locale de la composition en bases.

À l'échelle de 100 kb, cette forte corrélation entre le $G+C_3$ et le $G+C$ de la région dans laquelle le gène est localisé se retrouve également (Zoubak *et al.* 1996, Clay *et al.* 1996). Une telle corrélation ne pourrait pas exister si l'hétérogénéité locale était aussi forte que l'hétérogénéité globale à l'échelle des génomes. Ainsi, bien que les données ne correspondent pas exactement au modèle initialement proposé par Bernardi (2000), Eyre-Walker et Hurst 2001 suggèrent que le terme "isochore" est approprié pour désigner ces régions du génome de relative homogénéité compositionnelle. Dès 1981, lorsque Cunny avait quantifié l'homogénéité des isochores, il avait effectivement montré que leur homogénéité était relative.

3.1.3 Propriétés biologiques liées aux isochores

La structure en isochores est corrélée avec d'importantes caractéristiques de l'organisation des génomes comme par exemple la densité en gènes (Mouchiroud *et al.* 1991, Zouback *et al.* 1996), la taille des gènes (Duret *et al.* 1995), les distributions des éléments transposables (Soriano *et al.* 1983, Smit

1999), le taux de recombinaison (Eyre-Walker 1993, Fullerton *et al*. 2001) et les bandes chromosomiques (Saccone 1993). De plus, certains acides aminés (alanine, arginine...) sont plus fréquents (Aota 1986, D'Onofrio 1991, Clay 1996) dans les régions riches en $G + C$. Ces propriétés ont été confirmées par l'analyse complète du génome humain. Les isochores riches en $G + C$ correspondent à des régions riches en gènes (avec des introns relativement courts), denses en Alu, pauvres en répétitions LINEs, et possèdent un taux de recombinaison en moyenne plus élevé que les régions pauvres en $G + C$ (Jabbari 1998, Lander *et al*. 2001). Ainsi, la structure en isochores reflète un niveau d'organisation à grande échelle du génome.

3.1.4 L'origine des isochores

Comprendre comment s'est mise en place cette structuration en bases dans les génomes de vertébrés et comment elle évolue est primordial afin de comprendre l'organisation et le fonctionnement de l'information génétique à l'échelle des génomes complets. Cependant, les processus évolutifs à l'origine des isochores demeure relativement mal connu. Différents modèles ont donc été proposés pour expliquer l'émergence et l'évolution des isochores, chacun d'eux concordant plus ou moins bien avec les réalités biologiques. Cette section décrit les modèles les plus couramment cités, ils peuvent être classés en deux catégories : modèles neutres ou modèles sélectifs.

3.1.4.1 La sélection (modèle sélectionniste)

Selon Bernardi (2000), la présence d'isochores riches en $G + C$ dans les génomes des mammifères et des oiseaux résulterait d'une adaptation à l'homéothermie. En effet, la thermostabilité de l'ADN et des ARN structuraux est directement liée au taux en $G + C$. Cette hypothèse sélectionniste implique que les organismes eucaryotes à température interne variable ne sont pas soumis à une telle sélection (homéothermie) et ne devrait donc pas présenter une compartimentation en base $G + C$ aussi marquée que chez les vertébrés homéothermes (Bernardi 2000). Cependant, Hugues (1999) a montré que les tortues et les crocodiliens possèdent des isochores riches en $G + C$. L'origine des isochores riches en $G + C$ précède la divergence des mammifères et oiseaux et ne peut donc pas correspondre à une adaptation à l'homéothermie. Dans cette mesure, l'hypothèse de Bernardi paraît peu réaliste.

D'une manière générale, il est difficile d'imaginer pour quelle raison la sélection naturelle agirait sur des mégabases de séquences non-codantes. Ceci ne montre pas que les isochores ne sont pas dû à la sélection car il se peut que le $G+C$ affecte d'autres paramètres que l'on ne connaît pas encore. Cependant, il faudrait supposer que cette pression de sélection soit relativement forte. En conclusion ; il paraît peu probable que le taux en $G+C$ des séquences non-codantes ait un impact suffisamment fort sur la fitness pour que la sélection soit efficace.

3.1.4.2 Les biais de mutation (modèle neutraliste)

Selon ce modèle, la structure en isochores serait dû à une grande variabilité des patrons de mutation le long des chromosomes.

Biais mutationnel lié à la réplication

L'hypothèse repose sur l'existence d'un biais mutationnel lié à la réplication (Wolfe *et al*. 1989). Elle se base sur deux observations.

Premièrement, il existe des régions du génome humain à réplication précoce et d'autres à réplication tardive au cours du cycle cellulaire. Cette structuration est associée au banding chromatidien. Les bandes Giemsa de type R correspondent à des régions à réplication précoce, alors que les bandes G correspondent à des régions à réplication tardive (Saccone *et al*. 1993). Or la répartition des isochores suit en partie celle du banding chromatidien. Les bandes G contiennent en majorités des isochores pauvres en $G+C$ alors que les bandes R contiennent préférentiellement des isochores riches en $G+C$.

La deuxième observation, sur laquelle repose l'hypothèse de Wolfe (1989), a montré que la concentration de nucléotides libres dans le noyau, est variable au cours de la division cellulaire. D'après ce modèle, les isochores riches en bases $G+C$, à réplication précoce, disposerait du pool complet de nucléotides au début du cycle cellulaire. Les isochores pauvres en bases $G+C$, qui se répliquent en fin de cycle, disposeraient d'un pool de nucléotides appauvrit en G et C et donc riche en A et T. Cette hypothèse paraît séduisante, mais Woodfine (2004) a démontré que l'hétérochromatine est riche en bases $G+C$ alors qu'elle a une réplication tardive.

Biais mutationnel lié à la réparation

Filipski (1990) a proposé l'hypothèse concernant l'existence d'un biais de réparation le long de l'ADN pour expliquer la structuration en isochores. Comme les régions fortement transcrites sont localisées dans les parties décondensées de la chromatide, elles seraient plus facilement accessibles à la réparation. La réparation étant biasée en faveur des bases $G+C$, ces régions s'enrichiraient en bases $G+C$. Mais, l'absence de corrélation entre la composition en bases $G+C$ et le taux de substitution va à l'encontre de cette hypothèse. Elle n'est pas considérée aujourd'hui comme valable.

3.1.4.3 La recombinaison

Chez les organismes à reproduction sexuée, la recombinaison méiotique a pour rôle d'assurer la ségrégation correcte des chromosomes homologues au cours de la méiose mais également d'augmenter la diversité génétique. Récemment, il a été proposé que la recombinaison pourrait influencer l'évolution de la composition en bases G et C (Galtier *et al.* 2001). Chez les mammifères, plusieurs observations renforcent cette hypothèse. Tout d'abord, une corrélation positive a été établie entre la composition en bases G et C de l'ADN génomique et le taux de recombinaison local (Fullerton *et al.* 2001, Kong *et al.* 2002). De plus, les familles multi-géniques sujettes à l'évolution concertée (avec des évènements de conversion fréquents entre les gènes d'une même famille) sont riches en G et C (Galtier *et al.* 2001, Galtier 2003). Cette corrélation positive entre la composition en bases G et C et le taux de recombinaison local semble être un phénomène répandu, puisqu'une telle corrélation a été retrouvée chez la levure, la drosophile et le nématode (Gerton *et al.* 2000, Marais *et al.* 2001, Birdsell 2002, Piganeau et Marais 2002). Cependant une telle corrélation n'indique pas le lien de cause à effet entre ces deux variables. Selon certains auteurs, la corrélation observée chez la levure n'est pas due à l'action de la recombinaison sur le taux de substitution, mais plus tôt au fait qu'une composition riche en bases G et C influence positivement le taux de recombinaison local (Gerton *et al.* 2000, Blat *et al.* 2002, Petes et Merker *et al.* 2002). De plus, la corrélation entre le taux de recombinaison local et la composition en G et C est relativement faible chez l'homme (Kong *et al.* 2002) : seulement 15% de la variabilité de la compositoion en bases G et C sont expliquées par la recombinaison. À première vue, on pourrait conclure à une action mineure, du processus de

recombinaison sur la composition en bases G et C des génomes eucaryotes.

Toutefois, ces corrélations visant l'étude du lien entre recombinaison en bases G et C ont été effectuées entre deux variables qui opèrent sur des échelles différentes de temps. En effet, le taux de recombinaison est une variable qui reflète un processus " actuel ". En revanche, la composition en bases G et C résulte d'évènements de substitutions survenus durant une longue période évolutive (probablement plusieurs centaines de millions d'années). Ainsi, on peut envisager qu'un biais de temporalité masque une partie de la relation entre recombinaison et composition en G et C, et que celle-ci est plus forte que ce que suggèrent les corrélations mentionnées ci-dessus. Ce biais de temporalité est d'autant plus probable que les taux de recombinaison peuvent changer rapidement durant l'évolution, ceci dû à des inversions, mutations ponctuelles... (True *et al.* 1996, Depaulis *et al.* 2000, Jeffreys et Neumann 2002, Montoya-Burgos *et al.* 2003). Par exemple, la carte génétique de l'homme est 30% plus grande que celle du Babouin (Rogers *et al.* 2000).

Deux hypothèses ont été proposées pour expliquer l'influence de la recombinaison sur la composition en bases (Eyre-Walker et Hurst 2001) :

1. **hypothèse sélective :** il existe une pression de sélection en faveur d'une composition riche en bases G et C, le taux en $G+C$ augmente avec un fort taux de recombinaison car la sélection est alors plus efficace (Charlesworth 1994)
2. **hypothèse neutre :** la recombinaison augmente la probabilité de fixation des allèles G et C par rapport aux allèles A et T, par la conversion génique biaisée (ou BCG) (Galtier *et al.* 2001)

La sélection pourrait agir par l'intermédiaire du mécanisme de recombinaison. Ainsi, de nombreux modèles mettent en évidence une augmentation de l'efficacité de sélection avec la recombinaison par des effets de Hill-Robertson (Figure 3.2) (Charlesworth 1994). Une diminution de la fréquence des mutations délétères et une augmentation de l'efficacité de la sélection, liées à la recombinaison, ont été décrites chez l'homme et la drosophile (Williams et Hurst 2000, Bachtrog et Charlesworth 2002). Cependant, ce modèle en faveur d'une pression de sélection apparaît peu vraisemblable. En effet, l'enrichissement en bases G et C induit par une seule mutation AT vers GC dans les grandes séquences non codantes est extrêmement faible. Etant donné la taille limitée des populations chez les mammifères réduisant

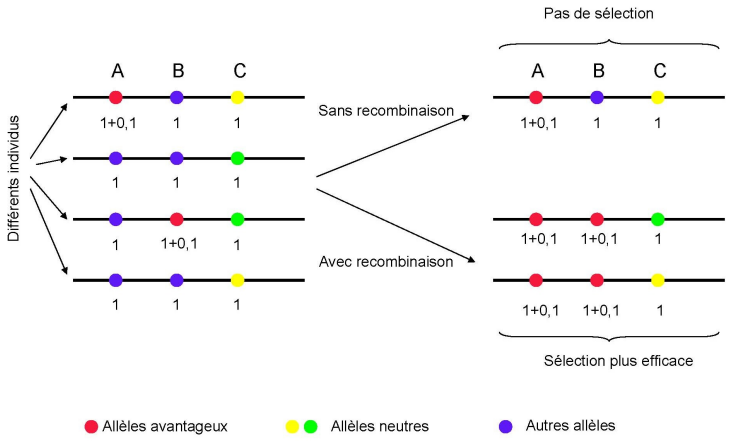


Fig. 3.2 : *Les effets de Hill-Robertson*
Prenons comme exemple trois loci A, B et C, génétiquement liés. Une mutation avantageuse, augmentant de 10% la fitness de l'allèle A, est apparue dans la population. Une mutation moins avantageuse, augmentant de 1% la fitness de l'allèle B, est apparue au même moment dans la population : (1) en l'absence de recombinaison et donc de brassage génétique, la sélection favorisera la première combinaison d'allèles car la fitness est plus grande. De plus, le polymorphisme au niveau de l'allèle neutre C est perdu, (2) en cas de recombinaison, la fitness maximale des individus sera plus grande. La sélection maximise la fitness pour les allèles A et B et permet la conservation du polymorphisme au niveau de l'allèle C.

l'efficacité de la sélection, il est peu réaliste que cette augmentation infime de composition en bases G et C puisse être sélectionnée.

En dehors des biais sélectifs, plusieurs modèles neutres ont été proposés pour expliquer l'évolution de la composition en bases $G + C$. Le modèle le plus répandu à l'heure actuelle met en avant l'effet de la conversion génique biaisée ou biais de conversion génique (Holmquist 1992, Eyre-Walker 1993). Au cours de la méiose, les chromosomes homologues paternels et maternels sont appariés et des échanges entre les deux molécules d'ADN ont lieu *via* une série de processus moléculaires qui peuvent ou non aboutir à un événement de recombinaison et au cours duquel une région d'un des chromosomes est recopiée (convertie) à partir de l'autre chromosome (Figure 3.3). Si la zone d'appariement entre l'ADN des chromosomes paternels et maternels recouvre un site polymorphe, alors cet hétéroduplex contiendra un mésap-

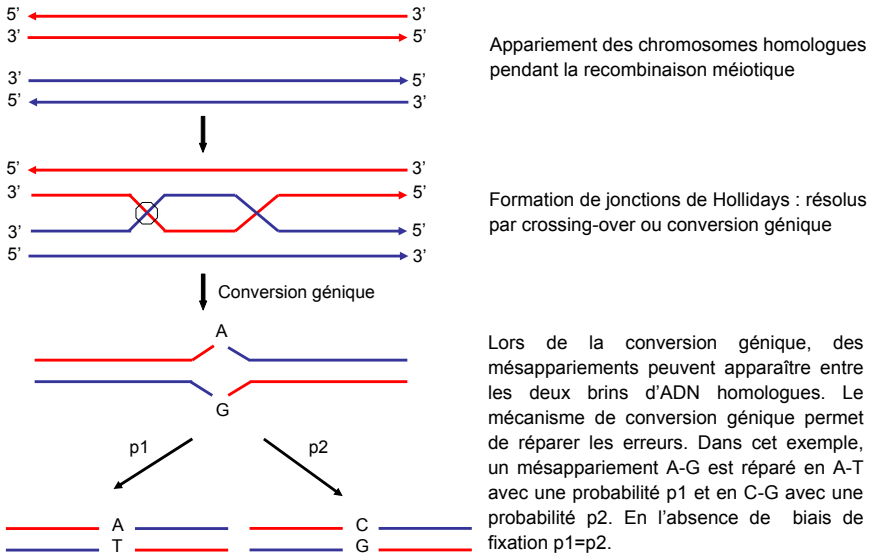


Fig. 3.3 : *Le mécanisme de conversion génique*

pariement qui pourra être reconnu et réparé par les systèmes de réparation de l'ADN. Cette réparation aboutira donc à la conversion d'un des deux allèles par l'autres. On parle de biais de conversion génique (BCG) lorsque la probabilité de conversion n'est pas identique pour les deux allèles. Ainsi, la conversion génique peut être biaisée vers les bases $G+C$ chez les mammifères (Bill *et al*. 1998, Galtier 2003, Kudla *et al*. 2004), la composition en $G+C$ des séquences soumises à la conversion génique aurait donc tendance à augmenter. Ce mécanisme pourrait expliquer la corrélation entre la composition en bases $G+C$ et le taux de recombinaison (Galtier *et al*. 2001). La conversion génique biaisée est un mécanisme neutre car elle n'est pas soumise aux effets de la sélection traductionnelle et/ou de la sélection transcriptionnelle qui pourraient la favoriser ou l'inhiber. Les effets attendus du BCG sur la composition locale en $G+C$ (rythme de fixation) sont les mêmes que pour un modèle de sélection de faible intensité (Marais *et al*. 2003), car la recombinaison se produit pendant la méiose. L'intensité de la sélection ou le BCG dépendent de la taille des populations qui y sont soumises et du coefficient de sélection et/ou de conversion génique.

Ce dernier modèle paraît actuellement le plus vraisemblable pour expliquer l'origine des isochores riches en bases $G + C$.

3.1.5 Méthodes de prédiction existantes

Le séquençage complet du génome humain offre la possibilité d'affirmer ou d'infirmer l'hypothèse concernant l'existence d'une structure en mosaïque le long du génome humain. Ce paragraphe présente rapidement les différentes méthodes qui ont été mises en œuvre afin de détecter la présence d'isochores au niveau de la séquence génomique et qui ont permis d'alimenter le débat relatif aux isochores.

Les premières études conduites sur des génomes entiers utilisent des fenêtres glissantes et montrent des graphiques représentant le contenu en $G+C$ le long des chromosomes. Par exemple, Hattori (2000) et Dunham (1999) ont ainsi respectivement détecté des isochores le long des chromosomes 21 et 22 de l'homme. Cependant, par cette méthode, la localisation des zones homogènes en $G + C$ se fait de façon visuelle. Les fluctuations plus ou moins importantes du $G + C$ le long de la séquence posent alors des problèmes lors d'une identification précise des zones d'homogénéité. Des méthodes plus complexes ont ensuite été mises en oeuvre, toujours basées sur des fenêtres glissantes, mais utilisant des modèles aléatoires pour tester l'homogénéité des séquences. Nekrutenko et Li (2000) ont montré que ces approches ne permettent pas d'identifier les isochores. Häring et Kypr (2001) sont allés jusqu'à nier l'existence d'isochores le long des chromosomes 21 et 22 de l'homme et Lander (2001) en a conclu que les isochores ne méritent pas leur préfixe " iso " et que seules les séquences d'ADN répétées sont homogènes. Toutefois, ces affirmations sont à modérer. En effet, elles soulèvent un problème méthodologique dans la mesure où elles reposent principalement sur l'utilisation de modèles aléatoires pour lesquels les nucléotides sont indépendants, ce qui n'est pas le cas le long de la séquence d'ADN comme le montre notamment la présence de codons dans les exons. Ainsi, ces méthodes obtiennent des résultats contradictoires, par exemple, Li (2002 et 2003) montre que la notion d'isochores peut être rejetée lorsque la distribution en $G + C$ le long des fenêtres est considérée comme binomiale alors qu'une analyse de variance sur le même jeu de données met en évidence la présence d'isochores.

Des méthodes alternatives de segmentations ont été mises en œuvre pour l'analyse de l'hétérogénéité des génomes (notamment Bernaola-Galvan *et al*.

2001, Li *et al*. 2002, Cohen *et al*. 2005). Ces méthodes ont d'abord été développées pour calculer le contenu en $G + C$, puis elles ont été utilisées pour identifier les isochores. Parmi celles-ci deux types de méthodes se distinguent.

La première méthode, nommée **IsoFinder**, a été proposée par Oliver *et al*. en 2004. Il s'agit d'un algorithme de segmentation récursif destiné à localiser les isochores. La procédure est la suivante : considérons une séquence de longueur L à analyser, le long de laquelle se déplace un pointeur de gauche à droite. Lors de la première étape, à chaque position du pointeur est calculé le contenu en $G + C$ moyen à droite et à gauche du pointeur jusqu'au bout de la séquence. Pour mesurer la différence entre les valeurs moyennes de $G + C$ à droite et à gauche, le test de statistique de Student est utilisé (t-test). La position du pointeur pour laquelle t atteint sa valeur maximale est ainsi déterminée. Ce point est donc supposé séparer la séquence en deux sous séquences de longueurs L_{droite} et L_{gauche}. La seconde étape consiste à vérifier si cette séparation correspond bien à un point de cassure délimitant deux isochores, c'est-à-dire, à vérifier si la différence de composition entre les deux sous séquences candidates est significative. Pour cela, les deux sous séquences L_{droite} et L_{gauche} sont divisées en fenêtres non chevauchantes de longueur l_0 et leur contenu en $G + C$ moyen est calculé. Un test de Student est alors appliqué pour s'assurer que le $G + C$ moyen des fenêtres de L_{droite} est bien différent de celui des fenêtres de L_{gauche}. Si les deux sous séquences ne sont pas significativement différentes, la séquence n'est pas divisée, sinon la séquence est divisée et l'algorithme est réappliqué récursivement à L_{droite} et L_{gauche}. Il est intéressant de noter que dans le but d'éviter l'influence statistique des petites zones d'hétérogénéités, celles dont la longueur est inférieure à une valeur seuil l_0 sont filtrées. Le programme propose trois valeurs pour l_0 : 1, 2 ou 3 kb.

La seconde méthode, nommée "Z-Curve method", (Zhang *et al*. 2003) cumule des profils de $G+C$. Elle a été appliquée au génome humain en 2003 et à celui de la souris en 2004 (Zhang *et al*. 2004). Ainsi, pour chaque base n, considérée comme un noeud, cette méthode compte le nombre d'occurrences des lettres A, C, T, G entre le début de la séquence et la base n (nommé respectivement A_n, T_n, G_n et C_n). Les coordonnées suivantes sont ensuite

calculées pour chaque base n :

$$\begin{aligned} X_n &= (A_n + G_n) - (C_n + T_n) \\ Y_n &= (An + Cn) - (Tn + Gn) \\ Z_n &= (An + Tn) - (Cn + Gn) \end{aligned}$$

pour n=0, 1, ..., L où L est la longueur de la séquence.
Cette méthode décrit donc la séquence d'ADN par trois distributions différentes. Les composantes X_n et Y_n représentent respectivement les distributions des bases purines/pyrimidines et AC/TG le long des sous séquences. La composante Z_n représente la distribution des bases de types AT/GC le long de la séquence. Lorsque le nombre de A et T dans la sous séquence allant de la base 0 à la base n est supérieur à celui de C et G, $Z_n > 0$, sinon $Z_n < 0$. Cette dernière composante offre une approche intuitive de l'organisation en isochores le long d'une séquence par l'intermédiaire du tracé de Zn en fonction de n qui permet de détecter les changements globaux et locaux du contenu en $G+C$ le long de la séquence. Ainsi, les brusques changements d'inclinaison de la pente de la courbe permettent de détecter les limites entre les isochores.
Ces deux méthodes de segmentation s'accordent pour conclure que le concept d'homogénéité du contenu en $G + C$ est relatif et qu'une structure en isochores existe bien dans le génome humain. Que ce soit avec les méthodes par fenêtres glissantes ou avec les méthodes par segmentation, seule la composition en $G + C$ le long de la séquence d'ADN est utilisée pour localiser les isochores. Cependant, les caractéristiques statistiques du contenu en $G + C$ varient fortement entre les régions codantes et non codantes chez les vertébrés, ce qui peut poser quelques problèmes lors de la détection des isochores si ces variations ne sont pas prises en compte. Par exemple, le programme **IsoFinder** (Oliver *et al*. 2004) néglige les hétérogénéités locales comme la séparation entre exons et introns dans le gène ou la présence de séquences répétées. Ainsi, les fichiers de sorties de cet algorithme concernant le génome humain représentent fréquemment des petits segments inférieurs à 300kb. Des domaines de compositions différentes en $G + C$ sont donc caractérisés mais ce n'est pas une véritable segmentation en isochores du génome humain.

Le but de ce chapitre est de proposer une méthode de segmentation plus complète et plus rigoureuse à partir de modèles de Markov cachés.

La première application des HMMs à l'étude de données génomiques a été réalisée par Churchill (1989). Elle avait pour but d'analyser l'hétérogénéité de la composition des séquences d'ADN. Plus récemment, Peshkin (1999) a montré que les HMMs pouvaient être utilisés avec succès pour l'analyse des structures des génomes tout en permettant une interprétation biologique. L'utilisation de modèles de Markov cachés semble donc bien adaptée au problème de la prédiction des isochores sur des génomes entiers, où chaque état de l'HMM représente un type de segment homogène. L'originalité de notre méthode repose sur l'idée d'introduire des HMMs, qui prennent en compte la composition en $G+C$ de la séquence d'ADN, mais aussi et surtout, les propriétés biologiques associées à la structure en isochores des génomes (tels que la densité en gènes, la longueur des introns suivant la classe d'isochores...).

3.2 Matériels et Méthodes

3.2.1 Structure des modèles HMMs

3.2.1.1 Matériels

Le taux en $G+C_3$ est fortement corrélé au taux en $G+C$ de la région dans laquelle le gène est localisé (Bernardi 2000). Le taux de $G+C$ génomique est cependant beaucoup moins variable que le $G+C_3$, probablement en raison de l'insertion d'éléments transposables dans les régions non-codantes (Duret et Hurst 2001). Quel que soit les facteurs à l'origine des isochores, il est clair que celui-ci opère à la fois sur les régions codantes et non codantes. Le $G+C_3$ est donc un bon marqueur de la structuration en isochores. Par ailleurs, la composition du $G+C_3$ dans les gènes orthologues de différentes espèces permet d'étudier l'évolution de la structure en isochores (cf. chapitre 4 et 5). Ainsi, les modèles HMMs développés au cours de ce chapitre et ceux qui vont suivre seront principalement adaptés aux différences de caractéristiques des régions codantes et non codantes suivant le contenu en $G+C_3$ du gène. La méthodologie utilisée dans ce chapitre pour la mise en place des modèles HMMs est similaire à celle développée au chapitre précédent. Au cours du chapitre 2, les classes H, L et M ont été réalisées à partir de la composition en $G+C_3$ des gènes humains extraits de la banque HOVERGEN (Duret 1994). Ces classes ont servi à caractériser trois classes d'isochores nommées H, L et M utilisées dans ce chapitre.

Toutefois, il est important de bien comprendre la démarche employée au cours de cette étude, car elle sera reprise lors des chapitre quatre et cinq. Dans un premier temps, à partir des jeux d'entraînement correspondant aux trois classes fixées suivant le contenu en $G + C_3$ des gènes, trois modèles sont construits. Ainsi, ces modèles savent discriminer, le but recherché maintenant consiste à discriminer les gènes. Dans un premier temps, il est possible de reclasser les gènes, grâce à nos modèles, de manière plus précise, car par cette méthode d'apprentissage, les modèles prennent en compte plus d'informations que le simple contenu en $G + C_3$. Dans un deuxième temps, l'étude des "erreurs" de prédictions des modèles permet la réalisation d'analyses complémentaires comme lors de l'analyse des régions UTRs (paragraphe 3.4.5). Enfin, c'est une méthode générale d'analyse de données qui permet d'étudier les caractéristiques liées aux isochores au sein même des génomes supposés ne pas en posséder.

3.2.1.2 Méthode

À chaque classe (H, L et M) est ajustée un modèle HMM (respectivement H, L et M). Chaque modèle sera considéré comme caractéristique de la classe d'isochore à laquelle il a été adapté. La procédure d'ajustement d'un modèle est la suivante : des sommes de lois géométriques sont ajustées aux distributions des longueurs des régions pour contourner la contrainte du temps de séjour dans les HMMs (méthode décrite lors du chapitre 2 section 1). Ainsi, chaque région est représentée par un macro-état (intergénique, intronique, exonique... Figure 3.4). La structure en codons des exons nécessite de séparer l'état exon en trois sous états représentant chacun la position du nucléotide dans le codon (Figure 3.5). Afin de conserver la correspondance de phase entre deux exons successifs, l'état intron est dupliqué en trois états qui ont les mêmes probabilités d'émission et de transition. Cette duplication permet par exemple, lorsque l'on sort d'un exon en phase 1, de revenir après l'intron dans un exon en phase 2 (Figure 3.5). De plus, pour prendre en compte la dépendance entre deux codons, un modèle d'ordre 5 est utilisé (Borodovsky *et al*. 1993, Burge *et al*. 1998). Les probabilités d'émission des différents états sont estimées à partir des fréquences des mots de 6 lettres des régions d'entraînement (intron, exon initial...). Enfin, la structure en double brin de la séquence d'ADN est représentée par l'intermédiaire des macro-états exons qui sont séparés en deux catégories, l'une représentant

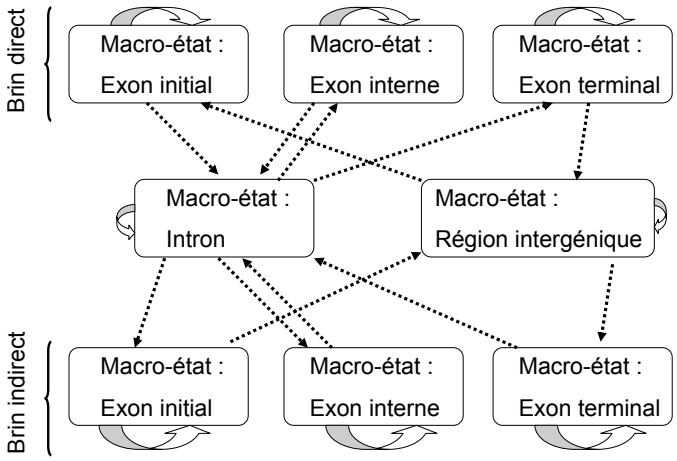


Fig. 3.4 : *Représentation simplifiée des différents macro-états qui constituent le modèle HMM H. Les doubles flèches représentent le temps de séjour dans un macro-état et les flèches en pointillés montrent les transitions autorisées entre les macro-états.*

les états codants du brin direct et l'autre ceux du brin indirect (Figure 3.4). Au final, les modèles H, L et M sont chacun constitués de 52 états.

3.3 Prédiction des isochores

3.3.1 Sélection de modèles : l'approche bayésienne

L'objectif de cette section consiste à déterminer le type d'isochore auquel appartient une séquence S donnée. Pour cela, le modèle le mieux adapté à la séquence doit être choisi dans l'ensemble de modèles $Model$=$\{H, L, M\}$ représentant respectivement les isochores H, L et M. Ainsi, le type d'isochore auquel appartient la séquence correspond au modèle ayant la plus forte probabilité $P(m \mid S)$, où $m \in Model$. La formule de Bayes permet alors d'obtenir la probabilité a posteriori de chaque modèle conditionnellement à la séquence observée S :

$$P(m \mid S) = \frac{P(S \mid m)P(m)}{\sum_{m' \in Model} P(S \mid m')P(m')}$$

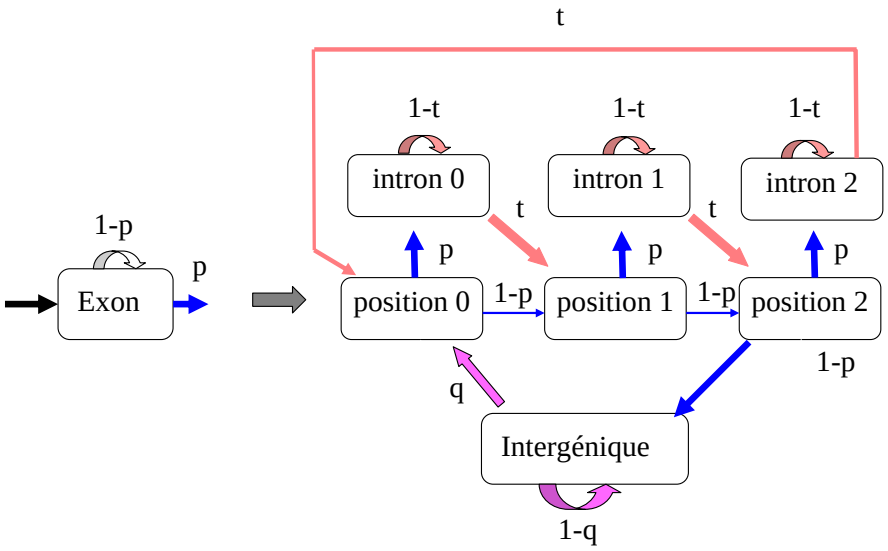


Fig. 3.5 : *Représentation de la séparation de l'état exon en trois sous états modélisant chacun la position du nucléotide dans le codon. Les transitions dans les macro-états introns, exons et régions intergéniques sont représentées respectivement par des flèches roses, bleues et violettes. Les doubles flèches représentent le temps de séjours dans un état.*

Dans cette formule, le terme $P(m)$ correspond à la probabilité *a priori* du modèle m et représente la contribution de l'*a priori* sur le choix du modèle m. Le terme $P(S \mid m)$ reflète quant à lui l'information apportée par l'observation S et correspond ainsi à la contribution de l'observation sur le choix du modèle m. Le terme $P(S \mid m)$, parfois appelé vraisemblance marginale, joue un rôle fondamental.

3.3.1.1 Procédure de prédiction des isochores

Matériel

L'assemblage du génome humain est réalisé à partir des données de la banque ENSEMBL (release de juin 2004). L'ensemble des calculs a été réalisé à l'IN2P3 (Institut National de Physique Nucléaire et de Physique de Particules). Ce centre dispose d'une exploitation de 1200 machines. Plus de 2000 tâches sont ainsi susceptibles de tourner simultanément sur ses serveurs. Dans notre cas, le découpage en fenêtres glissantes est avantageux car il permet une soumission en parallèle de tâches. Pour conduire l'étude sur l'ensemble du génome humain, 2943 tâches de 8 heures ont été parallélisées. Le temps de calcul est alors d'environ 10 jours au centre de l'IN2P3 pour l'obtention d'une carte d'isochores sur l'ensemble du génome humain. Les calculs de probabilités des modèles de Markov ont été réalisés grâce à une bibliothèque de modules en langage C++ interfacé en Python (SARMENT) regroupant des méthodes de partition maximalement prédictifs et de segmentation markovienne de séquences (Guéguen 2005), et disponible via le web (http ://pbil.univ-lyon1.fr/software/sarment).

Méthode

Notre méthode de prédiction des isochores le long du génome humain se décompose en trois étapes :

Première étape : Les chromosomes sont découpés en fenêtres glissantes de 100 kb avec un chevauchement de 50 kb, cette dimension étant couramment utilisée dans la littérature (Bernardi 1995, Clay *et al.* 2001, Pavlicek *et al.* 2002). De plus, elle laisse supposer la présence de gènes à l'intérieur des fenêtres. Cette condition est importante car la principale information discriminante lors des prédictions des modèles

de Markov entre eux est l'unité *gène*, la région intergénique étant modélisée simplement. Enfin, les isochores sont considérés comme des régions de plus de 300kb pouvant atteindre plusieurs mégabases, cette dimension procure donc une précision satisfaisante.

Deuxième étape : Sur chaque fenêtre, est calculée $P[Modele \mid fenetre]$ pour chaque modèle (H, L et M). Le modèle ayant la plus forte probabilité caractérise la fenêtre. Afin de rester cohérent avec la définition des isochores, un segment sera considéré comme étant un isochore s'il est constitué d'une succession de fenêtres associées à la même classe (H, L ou M) de longueur supérieure à 300kb. Par cette méthode, très peu de fenêtres isolées sont dénombrées. Moins de 5% des fenêtres ne sont pas regroupées dans un isochore. De plus, elles sont réparties aléatoirement le long des chromosomes.

Les probabilités *a priori* de chaque modèle ont été estimées à partir de la fréquences des gènes suivant leur contenu en $G + C_3$.

Troisième étape : À partir de ces résultats, les graphiques représentant la répartition des " isochores ", le contenu en $G + C$ et la densité en gènes sont tracés le long de chacun des 23 chromosomes humains à partir de fenêtre glissante de 100 kb.

3.3.2 Validation des modèles

Si les régions au faible contenu en $G + C$ sont réputées pour la relative homogénéité du contenu en $G + C_3$ des gènes qui s'y trouvent, ce n'est pas le cas des régions riches en $G + C$ qui sont connues pour leur grande variabilité. Le contenu en $G + C_3$ n'est pas toujours suffisant pour associer un gène à une classe d'isochore. Par exemple, un nombre non négligeable de gènes ayant un $G + C_3$ faible peut être présent à l'intérieur des isochores H (Nekrutenko *et al*. 2000) et poser des problèmes lors de leur identification. Cependant, notre modèle met clairement en évidence une relative "homogénéité" à l'intérieur des classes d'isochores tout en étant catégorique lors de l'identification des isochores dans les classes H et L. Par notre étude, la classe d'isochore associée à une fenêtre est déterminée par la probabilité $P[Modele \mid fenetre]$. Il est important de noter que pour chaque fenêtre, le modèle qui la représente a une probabilité $P[Modele \mid fenetre]$ très supérieure (valeur minimum 0,7546, moyenne 0,9885, écart type 0,0014) à celle obtenue par les deux autres modèles. Pour ces raisons, les régions prédites

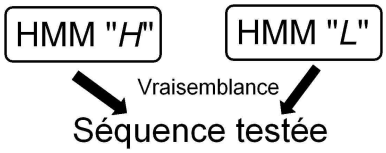


Fig. 3.6 : *Comparaison des vraisemblances de deux modèles.*

comme étant des isochores H ou L sont maintenant étudiées plus précisément. Ainsi, chaque macro-état (CDS, intron, gène, région intergénique) utilisé dans les modèles H et L a été isolé et analysé. Pour chaque isochore et pour chaque région, les prédictions des HMMs associés à chaque macro-état H et L ont été comparées au contenu en $G + C_3$ du gène.
La procédure de comparaison des prédictions de deux modèles, sur une séquence test, est la suivante :

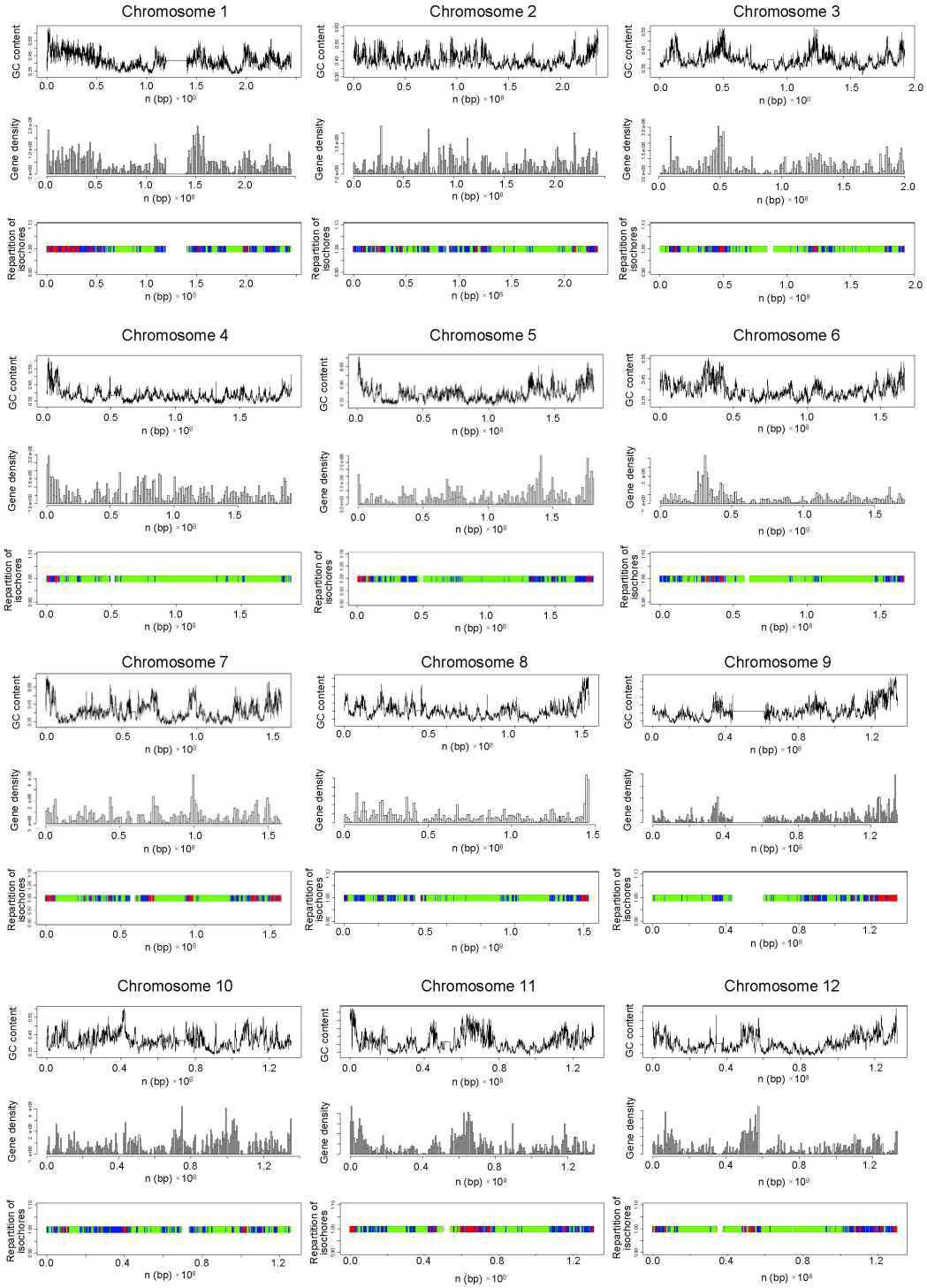
1. La vraisemblance marginale de la séquence $P[Sequence \mid Modele]$ est calculée pour chaque modèle.

2. Le modèle ayant la plus forte vraisemblance caractérisera la séquence, autrement dit, ce modèle prédira le "mieux" la séquence (cf. chapitre 2 section 2.3.2).

3.4 Résultats

3.4.1 La structure en isochores du génome humain

Les méthodes classiques de prédiction des isochores se trouvent confrontées aux problèmes liés à une définition peu précise de la notion d'isochore ainsi qu'à la variabilité du contenu en $G+C$ à l'intérieur même d'une classe (Nekrutenko *et al.* 2000). Cette constatation est, pour une grande part, à l'origine du débat sur l'existence des isochores. Cependant, au paragraphe précédent, il a été montré que nos modèles se distinguent nettement, il n'y a pas d'ambiguïté en ce qui concerne leurs prédictions. Cette approche permet de prendre en compte le phénomène des petites zones d'hétérogénéités à l'intérieur d'un isochore, dû à la présence de structures locales comme les gènes.



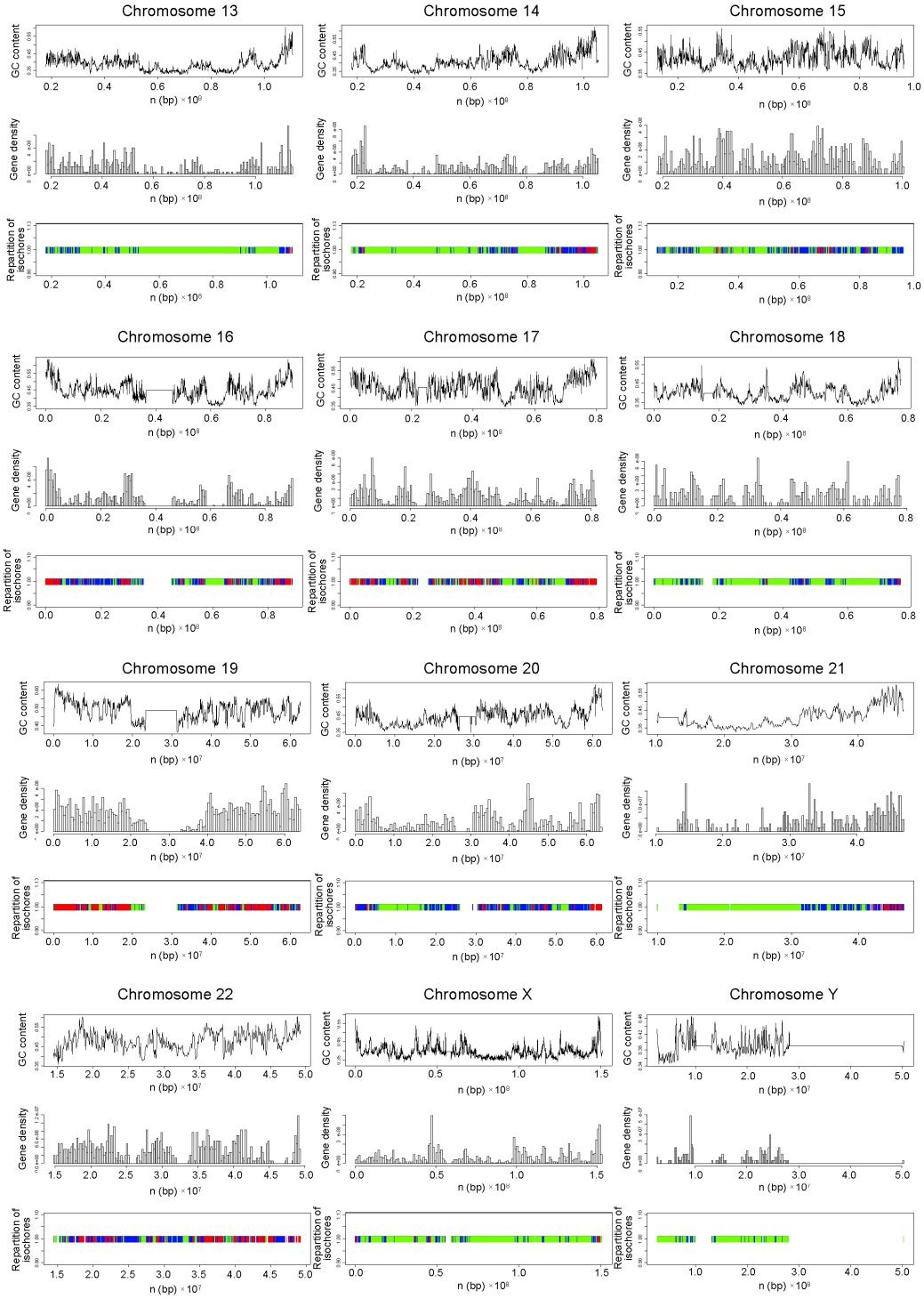


Fig. 3.7 : *L'évolution du taux en $G+C$, la répartition des gènes et la répartition des isochores prédite par notre modèle sont représentés le long de chaque chromosome humain. En rouge, sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.*

3.4.2 Carte des isochores du génome humain

Sur la figure 3.7, les isochores prédits par notre méthode sur l'ensemble du génome humain apparaissent en rouge, vert et bleu représentant, respectivement, les classes H, L et M. Ces cartes d'isochores mettent en évidence une organisation en mosaïque le long du génome humain (Bernardi *et al*. 1985, Bernardi 2001, Pavlicek *et al*. 2001) composée de plusieurs régions dont le contenu en $G+C$ est relativement homogène (Li *et al*. 2003). La distribution du contenu en $G+C$ le long des chromosomes humains s'ajuste bien avec l'organisation en isochores de chacun des chromosomes. Le contenu moyen en $G+C$ des différentes classes d'isochores H, M et L sur l'ensemble des chromosomes humains est respectivement de $51,5 \pm 3,5\%$, $45,2 \pm 1,2\%$ et $39,5 \pm 1,7\%$. Un test non paramétrique de Wilcoxon montre que le contenu moyen en $G+C$ des isochores H est significativement différent de celui des isochores L (p-valeur=10^{-4}). Le même test montre que le contenu moyen en $G+C$ des isochores M est également significativement différent de celui des isochores L (p-valeur=5.10^{-4}) et H (p-valeur=3.10^{-2}).

3.4.3 Variation de la taille des isochores en fonction de leur classe

Les différents types d'isochores (H, L et M) montrent clairement une variation de leur taille en fonction de leur contenu en $G+C$. Les isochores pauvres en $G+C$ (L) sont significativement plus grands que les isochores riches en $G+C$ (la p-valeur du test non paramétriqe de Wilcoxon étant de 9.10^{-10}). La longueur moyenne des isochores L est de 7,71 Mb, celle des isochores M étant de 5,74 Mb alors que la longueur moyenne d'un isochore H est de 2,93 Mb. Cette relation a précédemment été observée lorsque les isochores ont été détectés par des techniques de centrifugation en gradient de densité (Bettecken *et al*. 1992, Pilia *et al*. 1993 , De Sario *et al*. 1996).

3.4.4 Variation de la composition et de la densité en gènes en fonction de la classe d'isochores

Lors de la mise en évidence des isochores par des techniques de centrifugation, puis lors des études statistiques complémentaires, plusieurs auteurs (Bernardi *et al*. 1985, Mouchiroud *et al*. 1991, Zoubak *et al*. 1996) ont observé que la densité en gènes augmentait significativement en fonction du

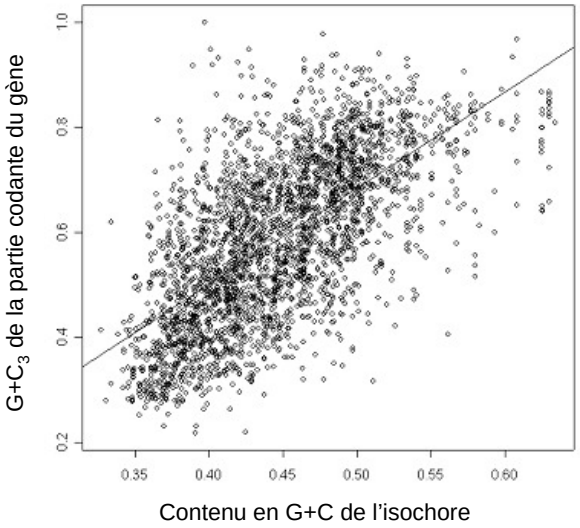


Fig. 3.8 : *Corrélation entre le taux en $G+C$ en position 3 dans les codons et le taux en $G+C$ de l'isochore dans lequel se situe le gène. La droite représente la droite de régression linéaire avec $R^2 = 0,43$.*

contenu en $G+C$. Cette densité en gènes peut être jusqu'à vingt fois supérieure dans les régions H à la densité dans les régions L. Les résultats obtenus au cours de cette étude confirment ces observations. Pour chacun des chromosomes, la structure en isochores s'ajuste visuellement bien à la distribution de densité en gènes. La densité des isochores H (15 gènes par Mégabase en moyenne) est supérieure à celle des isochores L (3,67 gènes par Mégabase en moyenne), ce qui est confirmé par un test significatif de Wilcoxon (p-valeur=4.10^{-5}). Le même test montre que la densité en gènes des isochores M (7,5 gènes par Mb en moyenne) est également significativement différente des isochores L (p-valeurs=2.10^{-3}) et H (p-valeur=6.10^{-8}). La figure 3.8 montre qu'il existe une corrélation entre la composition des régions codantes qui représentent 1 à 3% du génome humain et la composition des isochores prédits par notre méthode, qui sont principalement composées de régions intergéniques. Le taux en $G+C_3$ est fortement corrélé au taux de $G+C$ de l'isochore dans laquelle le gène est localisé (R^2=0,43). Le $G+C$ génomique est cependant beaucoup moins variable que le $G+C_3$ (de 35 à 60% au lieu de 20 à 90%).

3.4.5 Propriétés biologiques liées aux isochores

3.4.5.1 Analyse de la structure en isochores des régions H et L

Analyse des isochores H

Les régions riches en $G+C$ ont une grande variabilité, elles sont donc généralement difficiles à localiser par un graphique représentant le contenu en $G+C$ le long de la séquence. En effet, le long des séquences prédites comme appartenant à la classe d'isochore H, les gènes ayant un contenu en $G+C_3$ respectivement, supérieur à 72%, entre 56% et 72%, ou inférieur à 56% ont les fréquences relatives respectives suivantes : 48%, 35% et 17%. Une étude approfondie des séquences identifiées comme étant des isochores H a donc été menée afin de comprendre et d'analyser le comportement des modèles sur les gènes suivant leur contenu en $G+C$. L'objectif de cette étude consiste à trouver des caractéristiques communes aux gènes appartenant à l'isochore H. Ainsi, le comportement des modèles HMMs H et L sur les gènes des isochores H qui ont un contenu en $G+C$ supérieur à 72% et inférieur à 56% a été étudié selon la méthode décrite dans le paragraphe 3.2.3. Le tableau 3.1 décrit l'influence des différentes régions sur la prédiction de la classe d'isochore du gène à partir des prédictions des HMMs adaptés à chaque région. Il en ressort que 94% des gènes ayant un $G+C_3$ supérieur à 72% sont classés dans l'isochore H. Ainsi, le modèle HMM "gène" H décrit correctement les gènes avec un fort taux en $G+C_3$. De plus, 60% des gènes ayant un $G+C_3$ inférieur à 56% sont classés dans l'isochore H. Notre méthode met en évidence deux types de gènes possédant un $G+C_3$ inférieur à 56% : les gènes qui sont reconnus par le modèle HMM "gène" L (40%) et les gènes qui sont reconnus par le modèle HMM "gène" H (60%). Ceci montre la présence d'un facteur indépendant du contenu en $G+C_3$ permettant la caractérisation des gènes suivant la classe d'isochores à laquelle ils appartiennent. Cette classification des gènes ayant un $G+C_3$ inférieur à 56% ne semble pas dépendre uniquement des caractéristiques des CDS et des introns mais aussi, et surtout, des caractéristiques des régions 5'UTRs et 3'UTRs. Cette hypothèse est confirmée par la corrélation qui existe entre la prédiction des gènes et les prédictions des régions UTRs (5' et 3') par leurs modèles HMM respectifs (Tableau 3.2). Dans notre modèle, les régions UTRs ont été incluse dans le macro-état intergénique. C'est donc ce modèle

Région	Séquences des gènes classées dans l'isochore H par le modèle HMM représentant chaque région	
	Gènes ayant un $G+C_3 > 72\%$	Gènes ayant un $G+C_3 < 56\%$
Gène	94%	60%
CDS	96%	26%
Introns	93%	29%
5'UTR	86%	60%
3'UTR	86%	61%
intergénique	75%	57%

Tab. 3.1 : *Analyse des prédictions des macro-états sur les différentes régions qui composent le gène.*
Pour comparer les modèles H et L, la probabilité de chaque séquence d'un type donné (gènes, introns...) est calculée sous les deux modèles macro-états qui caractérisent la séquence. Ainsi, le modèle qui a la plus forte probabilité caractérise la séquence. Par exemple, les séquences de gènes ($G+C_3$>72%) sont caractérisées préférentiellement par le modèle HMM "gène" H. En effet, pour 82% des séquences de gènes ($G+C_3$>72%) la vraisemblance du modèle HMM "gène" H est plus grande que la vraisemblance du modèle "gène" L.

qui est a servi lors de la comparaison de des régions UTRs.

Dans 80% des cas (Tableau 3.2 : somme des ligne 1 et 4), la décision des modèles HMM " gène " et HMM " 5'UTR " sont similaires. Les mêmes résultats sont obtenus lorsque les modèles " gène " et " 3'UTR " sont comparés. Les régions UTRs prédites dans l'isochore H ont un contenu en $G+C$ $(51,3 \pm 4.10^{-1}\%)$ supérieur à celui des régions UTRs prédites dans l'isochore L $(42,9 \pm 0,1.10^{-1}\%)$. Ainsi, lorsque le modèle HMM identifie un isochore H dans le génome humain, deux facteurs permettent de classer les gènes ayant un contenu en $G+C_3$ inférieur à 56%. Premièrement, le contenu en $G+C$ des régions UTRs influence la prédiction des HMM (60% des cas). Deuxièmement, un effet de lissage se produit dans 40% des cas. Cet effet est dû à la structure de la fenêtre située aux alentours du gène qui influence le choix du modèle. L'influence de la région intergénique est particulièrement importante.

Prédiction HMM "gènes"	Prédiction HMM "5'UTR"	% de gènes ayant un $G+C_3 < 56\%$ dans cette configuration
H	H	50%
H	L	10%
L	H	10%
L	L	30%

Tab. 3.2 : *Comparaison des prédictions des modèles " gène " et "5'UTR " sur les gènes ayant un taux en $G+C_3$ inférieur à 56% dans les isochores H.*

Analyse des isochores L

Une analyse similaire a été conduite sur les isochores L. Les régions pauvres en $G+C$ sont beaucoup plus homogènes que les régions riches en $G+C$. Ainsi, les fréquences relatives des gènes ayant respectivement un contenu en $G+C_3$ supérieur à 72%, compris entre 56% et 72% et inférieur à 56% sont 6%, 19% et 75%. Par notre méthode, 92% des gènes ayant un $G+C_3$ inférieur à 56% sont classés dans l'isochore L (Tableau 3.3). Le modèle HMM " gène " L décrit donc correctement les gènes avec un faible taux en $G+C_3$. De plus, le tableau 3.3 montre que la classification des gènes ayant un contenu en $G+C_3$ supérieur à 72% ne dépend pas uniquement de la composition en bases du CDS et des introns mais aussi des caractéristiques des régions 5'UTRs et 3'UTRs comme c'est le cas dans les isochores H. Cette hypothèse est confirmée par la correspondance entre les prédictions des modèles " gène ", " 5'UTR " et " 3'UTR " dans 67% des cas (Tableau 3.4).

3.4.5.2 Etude des caractéristiques des différentes régions composant les gènes suivant leur classe d'isochore

L'étude de la longueur et du contenu en $G+C$ des différentes régions qui caractérisent les gènes dans les différents types d'isochores, confirme également des propriétés biologiques connues. Par exemple, dans les régions L, les introns et UTRs sont plus longs et leur contenu en $G+C$ est plus faible que ceux présents dans les isochores H. Cependant, l'étude de ces régions suivant les prédictions des modèles HMMs et leur classe d'isochores

Région	Séquences des gènes classées dans l'isochore L par le modèle HMM représentant chaque région	
	Gènes ayant un $G+C_3$<56%	Gènes ayant un $G+C_3$> 72%
Gène	92%	83%
CDS	72%	33%
Introns	93%	24%
5'UTR	83%	84%
3'UTR	88%	83%
intergenique	91%	91%

Tab. 3.3 : *Analyse des prédictions des macro-états sur les différentes régions qui composent le gène.*

Prédiction HMM "gènes"	Prédiction HMM "5'UTR"	% de gènes ayant un un $G+C_3 > 72\%$ dans cette configuration
H	H	0%
H	L	17%
L	H	16%
L	H	67%

Tab. 3.4 : *Comparaison des prédictions des modèles " gène " et "5'UTR " sur les gènes ayant un taux en $G+C_3$ supérieur à 72% dans les isochores L.*

	Isochore H				Isochore L			
	UTR associé gène H		UTR associé gène L		UTR associé gène H		UTR associé gène L	
	Décision modèle *H*	Décision modèle *L*	Décision modèle *H*	Décision modèle *L*	Décision modèle *H*	Décision modèle *L*	Décision modèle *H*	Décision modèle *L*
Longueur (kb)	28	32	38	54	93	94	103	131
$G+C$ (%)	52	50	50	46	42	41	40	39
nb gènes	966	216	301	191	23	117	235	1527

Tab. 3.5 : *Résumé des longueurs et du contenu en $G+C$ des régions 5'UTRs suivant le type d'isochore, le contenu en $G+C_3$ du gène et la décision du modèle.*

met en évidence le fait que la longueur et le contenu en $G+C$ influence le choix des modèles. Ainsi, les introns et les régions UTRs associés à un gène ayant un faible contenu en $G+C_3$ et prédits en H par nos modèles sont significativement plus courts que ceux des gènes prédits en L situés dans les isochores L. De plus, leur contenu en $G+C$ est significativement plus fort (Tableau 3.5). Des résultats similaires sont obtenus lors de l'étude des isochores L.

Pour étudier plus précisément les régions UTRs, une analyse en composantes principales a été réalisée sur les mots de six lettres constituant ces régions (Figure 3.9). Deux types de différences de structures sont clairement montrées par l'AFC (analyse factorielle des correspondances). D'une part, à l'intérieur même d'un isochore, une différence compositionnelle est observée suivant les prédictions des modèles, celle-ci étant représentée par le premier axe. D'autre part, le deuxième axe caractérise la différentiation des régions UTRs suivant le type d'isochores auquel elles appartiennent. La même étude a été menée en masquant les éléments répétés à partir du logiciel RepeatMasker, l'AFC produit des résultats similaires à la figure 3.9. La différence de structures observée dans la composition en mots de six lettres ne semble pas provenir *a priori* de la présence d'éléments répétés. Chez les mammifères, les dinucléotides symétriques CpG ont été progressivement perdus, sauf dans des séquences courtes (0,2 à 200 kb) enrichies en CpG, appelés îlots CpG. Ceux-ci ne sont pas répartis au hasard dans les génomes mais se trouvent préférentiellement dans les régions promotrices en amont des gènes de vertébrés. Cependant, dans notre cas, la distribution des îlots CpG ne semble pas être à l'origine de la différence de structure observée puisque leur présence n'est pas significativement différente dans les isochores H entre les

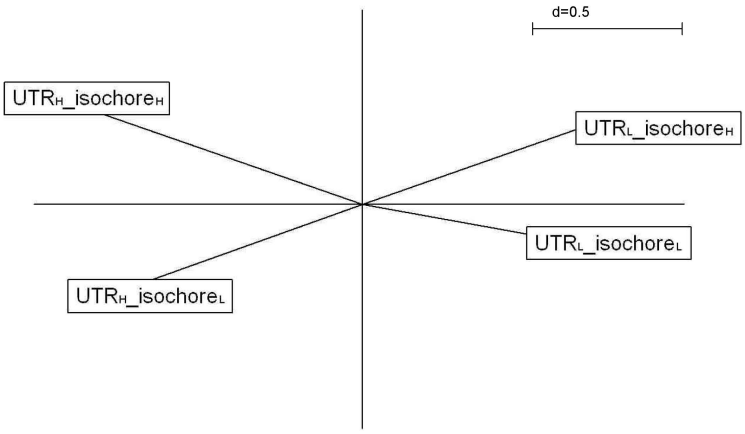


Fig. 3.9 : *Analyse en composante principale réalisée sur les mots de 6 lettres constituants les régions 5' UTRs.*
Les notations sont les suivantes :

UTRH _isochoreH caractérise les régions 5' UTRs dont le gène possède un taux en $G + C_3$ *élevé et qui appartiennent à un isochore H*
UTRH _isochoreL caractérise les régions 5' UTRs dont le gène possède un taux en $G + C_3$ *élevé et qui appartiennent à un isochore L*
UTRL_ isochoreH caractérise les régions 5' UTRs dont le gène possède un taux en $G + C_3$ *faible et qui appartiennent à un isochore H*
UTRL _isochoreL caractérise les régions 5' UTRs dont le gène possède un taux en $G + C_3$ *faible et qui appartiennent à un isochore L*

UTRs associées aux gènes possédant un fort taux en $G + C_3$ et ceux ayant un taux en $G + C_3$ faible (χ^2, p-valeur=8.10^{-1}) et réciproquement dans les isochores L (χ^2, p-valeur=4.10^{-1}).

3.5 Discussion

La clarification de la notion d'isochores est importante pour la compréhension de l'organisation du génome humain et des fonctions biologiques qui leur sont liées. Ce chapitre montre qu'il est possible, à partir de modèles de Markov cachés, d'analyser l'organisation spatiale de la molécule d'ADN

le long du génome humain. Seuls les gènes protéiques et les régions intergéniques sont modélisés pour obtenir des modèles simples mais efficaces. Ainsi, l'analyse d'un génome entier est rendue possible tout en permettant un retour aux données lors de l'échec des prédictions dans l'objectif de mettre en évidence de nouvelles propriétés dans le génome.

Les caractéristiques statistiques des régions codantes chez les vertébrés varient significativement entre les différentes classes d'isochores (Thierry *et al.* 1976), les efforts de modélisation se sont donc principalement portés sur ces régions. Notre méthode a consisté à adapter à chaque type d'isochores un modèle de Markov caché simple caractérisant au mieux les gènes. Nous avons préféré cette stratégie à celle qui consisterait à adapter un modèle global représentant les trois HMMs (H, M et L) par trois " macro-états " car un changement d'isochores peut se produire à l'intérieur même d'un gène. Dans ce cas, pour conserver la correspondance de phase entre les exons, il serait nécessaire d'introduire de nombreux états " intermédiaires " (comme dans le cas des successions d'exons) et de nombreuses transitions supplémentaires ce qui aurait pour répercussion une forte augmentation du temps de calcul. De plus, les transitions entre les macro-états représentant les isochores sont inconnues. L'emploi de trois modèles distincts a pour avantage d'éviter l'application de l'algorithme de Viterbi lors de la reconstruction du chemin des macro-états isochores et les problèmes qui lui sont associés lorsque les macro-états n'ont pas le même nombre de sous états (cf. chapitre 2). Enfin, l'usage de trois modèles permet la parallélisation des calculs et une dizaine de jours de calcul suffisent alors à obtenir le découpage en isochores sur l'ensemble du génome humain. Les modèles HMMs présentés dans ce chapitre ont ainsi pu mettre clairement en évidence la présence d'une structure en isochores le long du génome humain. La distribution des gènes le long des chromosomes s'ajuste parfaitement à l'organisation en isochores prédite par notre méthode. Les régions denses en gènes correspondent aux régions caractérisées comme étant des isochores H et possèdent un fort contenu en $G+C$. La relation entre classe d'isochores et structure des gènes est clairement montrée pour notre approche HMM. Par exemple, dans les isochores H, les gènes sont plus compacts, plus denses que dans les isochores L. Cette mise en évidence d'une structure en mosaïque généralisée à l'ensemble du génome humain contredit la suggestion d'Eyre-Walker et Hurst (2001) selon laquelle la structure en isochores serait présente sur seulement quelques parties du

génome humain et confirme les résultats obtenus par Oliver (2002).

Le résultat majeur mis en relief par cette étude concerne les régions UTRs. La variabilité du contenu en $G + C_3$ des gènes appartenant aux isochores H ne se reflète pas dans les régions UTRs. Il semblerait que ces régions possèdent une structure particulière qui dépend du type d'isochores dans lequel elles se situent. Ainsi, il apparaît que dans les isochores H, la structure des régions UTRs appartenant à des gènes ayant un faible $G + C_3$ s'apparente aux UTRs des gènes ayant un $G + C_3$ fort et réciproquement dans les isochores L. Ceci souligne le fait que le contenu en $G + C_3$ des gènes n'est pas la seule propriété biologique qui résulte de la structure des isochores. Les mécanismes biologiques impliqués dans ces différences de " structures " entre les régions codantes et les régions UTRs peuvent être soit neutres soit adaptatifs. Ainsi, de telles différences dans les régions UTRs peuvent résulter soit de mutations spécifiques qui s'accumulent plus rapidement dans les régions UTRs car la contrainte fonctionnelle est moins importante dans ces régions (hypothèse neutraliste), soit d'une contrainte de sélection exercée sur les régions UTRs, probablement associée à l'expression des gènes (hypothèse adaptative). A l'heure actuelle, il nous est difficile de trancher entre ces deux hypothèses. Notre modèle permet donc d'identifier les isochores mais aussi de reconnaître des gènes appartenant à un isochore bien que leur contenu en $G + C_3$ ne correspond pas au type d'isochore dans lequel ils se situent.

Le débat sur l'existence des isochores nous semble être en grande partie dû à un manque de rigueur dans la définition même de la notion d'isochores. Le concept d'isochore est relié au concept de domaines homogènes sur une grande échelle (plusieurs centaines de kilobases) dans les génomes et dans lesquels la variation du contenu en $G + C$ peut être considéré comme étant faible. Les variations du contenu en $G + C$ sont parfois plus grandes que les fluctuations statistiques et les isochores sont difficiles à déterminer par des méthodes classiques. Ainsi, lorsque l'assemblage du génome humain a été mis à disposition, Lander (2001) a échoué dans sa tentative de détection des isochores. L'utilisation de chaînes de Markov cachées permet de s'affranchir de cette définition des isochores. En effet, la structure en isochores est prédite par des HMMs qui ne prennent pas en compte seulement le contenu en $G+C$ des régions. Par rapport aux approches classiques, notre méthode améliore les prédictions des isochores. Ainsi, trois modèles HMMs sont ajustés à chaque classe d'isochores pour prendre en compte les principales proprié-

tés biologiques associées aux classes H, L ou M telles que les différences de longueurs des introns, la densité en gènes, etc. Ces propriétés sont négligées par les différentes méthodes existantes. De plus, par cette méthode d'apprentissage, les modèles prennent en compte plus d'informations que le simple contenu en $G + C_3$. Ainsi, l'étude des "erreurs" de classement des modèles a permis d'analyser de plus près la structure des gènes appartenant aux classes d'isochores H, indépendamment de leur contenu en $G + C_3$.

Ces dernières années, de nombreux génomes entiers ont été séquencés. La quantité de données que cela représente est considérable et rend impossible l'analyse manuelle des structures qui proviennent des propriétés biologiques. De nombreuses méthodes mathématiques et informatiques ont été développées. Notre approche, utilisant des HMMs, semble prometteuse et particulièrement adaptée à l'analyse de l'organisation des génomes. Une étude comparative à grande échelle de l'organisation en isochores de différents génomes en adaptant un modèle HMM à chaque génome peut permettre d'étudier la mise en place et l'évolution des isochores entre les espèces au cours du temps. C'est pourquoi, nous présentons dans les deux chapitres suivants un début d'analyse des génomes de divers organismes.

Chapitre 4

Analyse de la structure en " isochores " des génomes du *Tetraodon nigroviridis* et du fugu

Au cours du chapitre précédent, une méthode efficace pour la prédiction des isochores le long du génome humain a été présentée. Le résultat majeur a consisté à montrer qu'il est possible de reconnaître des gènes appartenant à un isochore bien que leurs contenus en $G+C_3$ ne correspondent pas au type d'isochores dans lequel ils se situent. Les poissons sont connus pour la forte homogénéité de leur contenu en $G+C$, cette observation a donc conduit à supposer que ces espèces ne possèdent pas d'isochores. Une courte introduction présente l'hypothèse couramment admise concernant l'évolution des isochores chez les vertébrés. L'objectif de ce chapitre consiste à vérifier l'existence ou l'inexistence d'une structure en isochores chez les poissons à partir de notre méthode de prédiction décrite au chapitre 3. Plus particulièrement, nous nous intéresserons aux génomes *Tetraodon nigroviridis* et fugu qui sont entièrement séquencés. Le but étant de mettre en évidence une structure liée aux isochores ne dépendantpas uniquement du contenu en $G+C$ qui représente le facteur limitant lors de la prédiction d'isochores lorsque le génome étudié est très homogène.

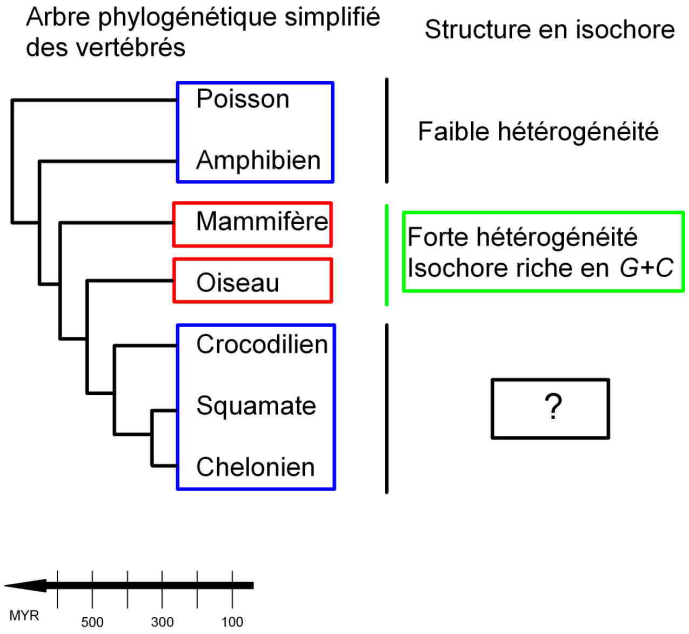


Fig. 4.1 : *Structure en isochores des génomes de vertébrés.*

4.1 Introduction

L'existence d'une compartimentation en bases $G+C$ dans les génomes est typique des vertébrés bien qu'elle ait été observée ponctuellement dans les génomes de plantes supérieurs, comme *Arabidopsis thaliana* (Zhang et Zhang 2004). La structuration en isochores s'est mise en place dans les génomes de vertébrés, il y a 300 Millions d'année (Bernardi 2000). Pour comprendre l'évolution des isochores chez les vertébrés, il faut au préalable déterminer leur distribution phylogénétique. Les analyses par centrifugation en gradient de densité ont montré que les génomes des mammifères et des oiseaux sont fortement hétérogènes (Bernardi 2000, Clay 2003). Les génomes des amphibiens et des poissons sont eux relativement homogènes et sont en général dépourvus d'isochores très riches en $G + C$.

Les données expérimentales suggèraient que tout comme ceux des amphibiens et des poissons, la plupart des génomes de reptiles étaient faiblement hétérogènes (Bernardi et Bernardi 1990). Ainsi, selon Bernardi, la structure en isochores serait apparue deux fois au cours de l'évolution, chez les mammifères et chez les oiseaux. Cette apparition coïncide avec l'acquisition de l'homéothermie par ces deux groupes taxonomiques. Toutefois, l'analyse des

séquences d'ADN a montré qu'il existe une forte hétérogénéité du $G + C_3$ chez certains reptiles. De plus, la comparaison du $G + C_3$ de gènes orthologues entre oiseaux et reptiles (par exemple crocodiles ou tortues) montre une forte conservation de la structure en isochores au sein des sauropsidés. Enfin, les comparaisons entre mammifères et oiseaux indiquent une forte corrélation du $G + C_3$ entre leurs gènes orthologues. L'hypothèse généralement admise consiste à dire que la structure en isochore ne s'est mise en place qu'une seule fois au cours de l'évolution, après la divergence entre amphibiens et poissons, mais avant la divergence entre oiseaux et mammifères (Hugues *et al*. 1999).

Le séquençage des génomes du fugu et du *Tetraodon nigroviridis* ainsi que l'assemblage de ce dernier permettent enfin la comparaison à grande échelle des génomes de poissons à ceux des autres génomes de vertébrés, et notamment celui de l'homme. Au cours de ce chapitre, les génomes de ces deux poissons vont être analysés à partir de notre méthode de prédiction des isochores développée au chapitre 3. L'objectif est de vérifier si la présence d'isochores peut être mise en évidence par notre approche, ce qui permettrait d'apporter des éléments de réponses aux questions concernant la mise en place des isochores chez les vertébrés au cours de l'évolution.

4.2 Analyse de l'organisation compositionnelle du génome du *Tetraodon nigroviridis*

4.2.1 Introduction

Le *Tetraodon nigroviridis* est un poisson qui vit en Asie du sud-est, le long des estuaires, dans des eaux dont la température est comprise entre 22 et 26 degrés celsius. Il atteint une taille moyenne de 14 cm à l'âge adulte. L'étude de la répartition des gènes le long des chromosomes du *Tetraodon* a permis pour la première fois d'étudier l'organisation chromosomique des génomes de mammifères et de poissons (Jaillon *et al*. 2004). L'ancêtre commun à ces deux lignées vivait à l'ère Paléozoïque, il y a environ 450 millions d'années. La comparaison (Jaillon *et al*. 2004) des gènes du *Tetraodon* avec les gènes humains indique que les poissons à nageoires rayonnées ont dupliqués leur génome, peu après la séparation des deux lignées qui ont abouti aux poissons et aux mammifères. Ce mécanisme est bien connu chez les plantes mais très rare chez les animaux. Ses conséquences sont très importantes car ce

Fig. 4.2 : *Tetraodon nigroviridis.*

mécanisme confère à une espèce un plus important potentiel de création de nouvelles fonctions biologiques.

Le *Tetraodon* présente la particularité de posséder le plus petit génome connu parmi les vertébrés. Par certains aspects, bien que le génome humain soit huit fois plus grand, le génome du *Tetraodon* est proche de celui de l'homme (Tableau 4.1), notamment du point de vue du nombre de gènes et de chromosomes. Bien que les gènes de l'homme et du *Tetraodon* présentent de grandes similarités de séquences (Jaillon *et al*. 2004), la petitesse du génome du *Tetraodon* le rend particulièrement compact. Ainsi les régions codantes représentent 11% de son génome contre seulement 1 à 3% chez l'homme. Une conséquence de cette remarquable compacité chez le *Tetraodon* est que son contenu en $G+C$ est plus élevé que dans la grande majorité des génomes de mammifères. Toutefois, Jaillon constate la présence de zones hétérogènes, certes moins marquées que chez l'homme. Chez les mammifères, les régions riches en $G+C$ possèdent une plus forte densité en gènes que celles pauvres en $G+C$ (cf. chapitre 3). Cette structuration semble être conservée chez le *Tetraodon* (Jaillon *et al*. 2004). À cette forte densité en gènes, s'ajoute chez le *Tetraodon* une grande compacité des gènes eux mêmes, les introns étant de très petites tailles, pouvant aller jusqu'à 50 ou 60 bp.

La rareté des éléments transposables (Tableau 4.1) dans le génome du *Tetraodon* différencie nettement celui-ci du génome humain où leur quantité est remarquable. Chez l'homme, la distribution des SINEs n'est pas uniforme, leur insertion est favorisée dans les régions riches en $G+C$. Chez le *Tetraodon*, cette préférence est inversée : les LINEs sont présents préférentiellement dans les régions $G+C$ riches alors que les SINEs se retrouvent majoritairement dans les régions $G+C$ pauvres. Les raisons de ces différences ne sont pas encore clairement expliquées (Jaillon *et al*. 2004).

La compacité du génome du *Tetraodon*, la rareté des éléments transpo-

Descriptif	*Tetraodon*	Humain
Chromosomes	21	23
Gènes	environ 28000	environ 24000
Nombre d'exons par gènes	6,9	8,7
Taille du génome	environ 340 MB	environ 28000 MB
$G+C$	homogène et riche	hétérogène
Éléments transposables	rares	Abondants
Introns Longueur moyenne	600 bp	environ 2067 bp

Tab. 4.1 : *Comparaison de quelques propriétés biologiques des génomes de l'homme et du Tetraodon.*

sables et un contenu en $G+C$ très élevé tendent à confirmer l'hypothèse selon laquelle il n'existe pas d'isochores le long du génome du *Tetraodon* (Bernardi 2000). La notion actuelle d'isochores, qui consiste à les définir comme des séquences supérieures à 300 kb dont le contenu en $G+C$ est plus ou moins homogène, reste vague. Au cours du chapitre précédent, nous avons montré que certaines propriétés biologiques intervenaient pour classer un gène dans un certain type d'isochores, plus ou moins indépendamment du contenu en $G+C$ de la séquence. Notre modèle de Markov caché a ainsi pu localiser précisément les isochores le long du génome humain. L'objectif de cette section est de montrer qu'il est possible de mettre en évidence chez le *Tetraodon*, à partir de notre méthode de prédiction, une structure en mosaïques liée aux isochores qui n'aurait pas pu être identifiée par leur définition classique.

Lors de la construction des modèles HMMs pour le génome humain, la distribution en $G+C_3$ était étalée de manière plus ou moins uniforme (Figure 4.3 a). À partir de notre jeu de données, il a été aisé d'obtenir trois classes (H, L et M) dont le contenu en $G+C_3$ est bien distinct. Ces classes correspondent aux classes d'isochores décrites dans la littérature. Le long du génome du *Tetraodon*, la distribution en $G+C_3$ est très homogène et très concentrée comme le montre la figure 4.3 (a). Ainsi, la séparation en trois classes aboutirait, comme au chapitre 3, à l'entraînement de trois modèles HMMs qui se distingueraient difficilement les uns des autres. De plus, aucune propriété biologique ne peut justifier l'existence de ces classes. La construction de trois classes à partir de cette distribution en $G+C_3$ n'est

donc pas envisageable.

Toutefois, une approche nouvelle a été proposée afin de vérifier s'il n'existait pas d'autres facteurs que le contenu en $G + C$ qui pourraient mettre en évidence une structure liée à la notion d'isochores chez le *Tetraodon*. Ainsi, les gènes du *Tetraodon* orthologues aux gènes humains appartenant aux trois classes d'isochores prédites dans le chapitre précédent sont utilisés pour construire les trois classes de gènes nécessaires à l'entraînement des modèles chez le *Tetraodon*. Cette démarche repose sur l'hypothèse que les propriétés qui lient les gènes aux isochores chez l'homme ont été conservées chez le *Tetraodon*. Si elles existent, ces propriétés permettront d'identifier les isochores sans prendre en compte le contenu en $G + C$ qui représente le facteur limitant chez le *Tetraodon* du fait de sa forte homogénéité.

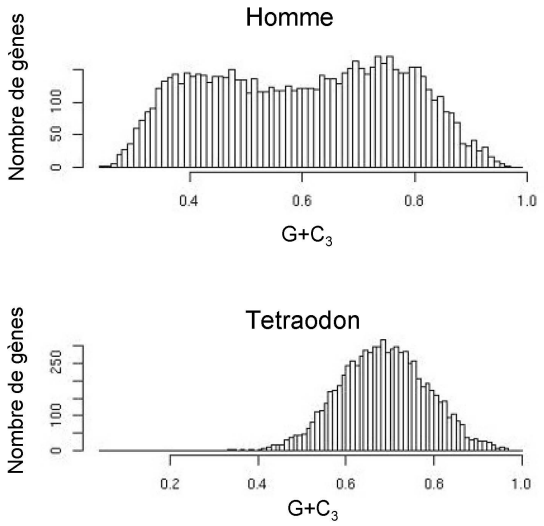
4.2.2 Matériel et Méthodes

4.2.2.1 Notion de gènes orthologues

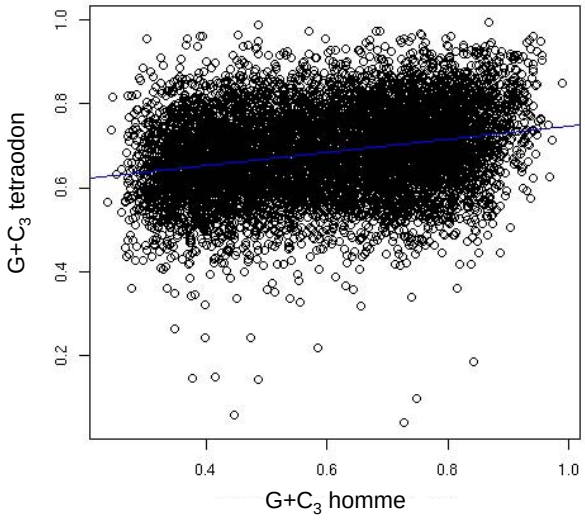
Deux séquences sont dites homologues si elles ont un ancêtre commun. On distingue parmi les gènes homologues ceux qui sont orthologues, c'est-à-dire ceux qui ont divergé à la suite d'un évènement de spéciation, et ceux qui sont paralogues, c'est-à-dire qui découlent d'un évènement de duplication génique au sein d'un génome (Figure 4.4). Cette distinction est essentielle pour établir une phylogénie, puisqu'il est impératif de travailler avec des gènes orthologues pour reconstituer l'histoire des espèces. Cette distinction est également importante pour les études fonctionnelles puisque deux gènes paralogues, même très proches, peuvent avoir des fonctions et des modes de régulation différents et ne sont donc pas strictement comparables. Dans le cas présent, l'hypothèse de conservation des propriétés des isochores entre les deux espèces recommande l'utilisation des gènes orthologues.

4.2.2.2 Données

Les gènes orthologues entre l'homme et le *Tetraodon* sont extraits de la banque de données GemCore. Elle contient 9 génomes de vertébrés (*Homo sapiens*, *Pan troglodytis*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*,*Gallus gallus*, *Danio rerio*, fugu rubripes, *Tetraodon nigroviridis*) extraits de la banque ENSEMBL. Les relations d'orthologie entre ces 9 organismes sont obtenues dans GemCore à partir de la méthode du " Reciprocal Best



(a)



(b)

Fig. 4.3 : *(a) Répartition des gènes suivant leur $G+C_3$ chez l'homme et le Tetraodon. (b) Corrélation du taux en $G+C_3$ des gènes orthologues entre le tertaodon et l'homme (R2=0,07).*

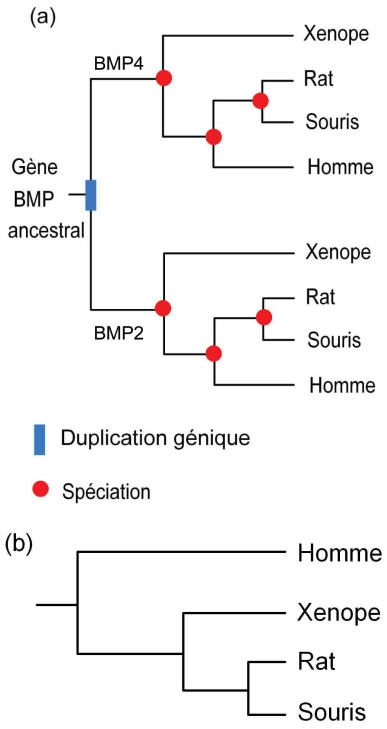


Fig. 4.4 : *a - Les gènes BMP2 du rat, de la souris, de l'homme et du xenope sont orthologues puisqu'ils ont divergé à la suite d'événements de spéciation. Les gènes BMP2 et BMP4 sont paralogues puisqu'ils ont divergé à la suite d'une duplication génique. b - La position du gène du xenope démontre que les gènes du rat et de l'homme sont paralogues.*

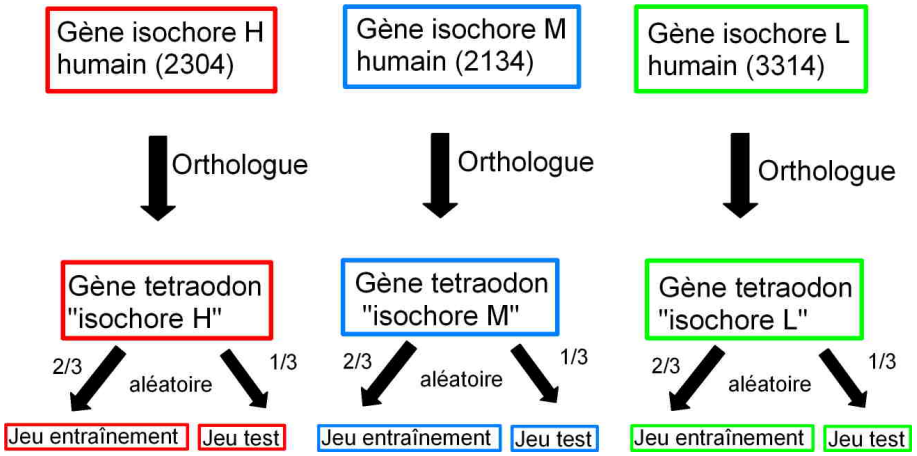


Fig. 4.5 : *Répartition des données nécessaires à l'entraînement des modèles " H ", " L " et " M " chez le Tetraodon.*

Hit ". Ainsi, le jeu de données est constitué de 7753 gènes, soit 27% des gènes qui sont annotés dans Ensembl.

4.2.2.3 Construction des modèles

Première étape : La figure 4.5 décrit la répartition des gènes du *Tetraodon* orthologues aux gènes humains en trois classes qui seront supposées représenter les trois classes d' " isochores " du poisson. Elles sont nommées " H ", " M " et " L " et contiennent respectivement 2304, 2134 et 3314 gènes. Les gènes de chaque classe sont séparés aléatoirement afin de constitué un jeu d'entraînement et un jeu test contenant respectivement 2/3 et 1/3 des gènes. Un modèle est ajusté sur chacune d'elle (annexe 1).

Deuxième étape : Les chromosomes du *Tetraodon* sont découpés en fenêtres glissantes de 14 kb avec un chevauchement de 7 kb. Dans le cas présent, les fenêtres sont moins grandes que chez l'homme car le génome du *Tetraodon* est plus petit. La taille des fenêtres est définie proportionnellement aux longueurs des deux génomes. Leur dimension n'est pas aberrante, car la compacité du génome chez le *Tetraodon* laisse supposer la présence de gènes dans les fenêtres. Cette condition est importante car la principale information discriminante lors des prédictions des modèles de Markov est l'unité *gène*.

Troisième étape : Sur chaque fenêtre est calculée la probabilité

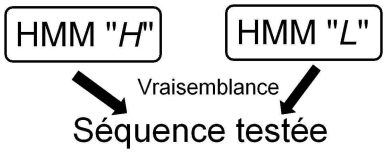


Fig. 4.6 : *Comparaison des vraisemblances de deux modèles.*

$P[Modele \mid fenetre]$ pour chaque modèle ("H", "L" et "M") (voir détail du calcul dans le chapitre 3. Le modèle ayant la plus forte probabilité caractérise la fenêtre.

Quatrième étape : À partir de ces résultats, les graphiques représentant respectivement la répartition des " isochores ", le contenu en $G+C$ et la densité en gènes, sont tracés pour chacun des 21 chromosomes du *Tetraodon* à partir de fenêtre glissante de 14 kb.

4.2.2.4 Évaluation des modèles

La justification de l'utilisation des HMMs décrite dans la section précédente repose sur l'existence de la conservation d'une différence de structure entre les gènes du *Tetraodon* orthologues aux gènes des isochores H, L et M de l'homme. Pour vérifier l'exactitude de cette supposition, les prédictions des modèles " H " et " L " du *Tetraodon* ont été comparées aux prédictions de différents modèles entraînés sur des jeux de gènes orthologues séparés aléatoirement. La procédure de comparaison des prédictions des deux modèles (figure 4.6), sur une séquence test, est la suivante :

1. La vraisemblance marginale $P[Sequence \mid Modele]$ de la séquence est calculée pour chaque modèle.
2. Le modèle ayant la plus forte vraisemblance caractérisera la séquence.

4.2.3 Résultats

Deux études distinctes ont été réalisées lors de l'analyse du génome du *Tetraodon*. La première étude met en évidence l'existence d'une structure dépendante de l'organisation en isochores au sein du génome du *Tetraodon*. La deuxième étude présente les cartes d'"isochores" obtenues à partir de modèles de Markov cachés.

Tetraodon	Gènes classés dans **l'isochore "H"**	Gènes classés dans **l'isochore "L"**
Prédictions par le **modèle *"H"***	62,5 %	41 %
Prédictions par le **modèle *"L"***	37,5 %	59 %

Tab. 4.2 : *Prédictions des HMMs du Tetraodon " H " et " L " sur les gènes orthologues des jeux tests du Tetraodon " H " et " L ".*

4.2.3.1 Étude de la structure des gènes du *Tetraodon* suivant leur classe d'"isochore"

Validation des modèles

A) Comparaison des modèles " H " et " L "

La première étape de validation des modèles a permis de vérifier qu'il existait une différence de structure entre les classes " H " et " L " obtenues chez le *Tetraodon*. Dans le cas contraire, l'hypothèse d'existence des " isochores " chez le *Tetraodon* aurait été écartée. Les prédictions des modèles " *H* " et " *L* " du *Tetraodon* sont comparées selon la procédure décrite dans la section 4.2.2.4 de ce chapitre. Les gènes appartenant à la classe " H " sont préférentiellement reconnus par le modèle " *H* " (62,5%) ; réciproquement les gènes des isochores " L " sont majoritairement préférés par le modèle "*L* " (59%) (Tableau 4.2). Cette différentiation montre l'existence d'une différence de structure entre les deux classes " H " et " L " du *Tetraodon*, le résultat du test du χ^2 étant significatif (p-valeur = 3.10^{-27}). La même analyse a été conduite sur les gènes humains à partir des modèles HMMs de l'homme. Cette fois la différence est plus grande avec une p-valeur = 2.10^{-57} pour le test du χ^2 (Tableau 4.3). La différence de qualité des prédictions entre l'homme et le *Tetraodon* peut aisément être expliquée. Lorsque les classes " H " et " L " du *Tetraodon* ont été construites, il a été supposé que chaque gène du *Tetraodon* a gardé une caractéristique liée à la classe d'isochores du gène orthologue humain. Cependant, il est probable que certains gènes aient perdu cette caractéristique en évoluant différemment dans les deux espèces. La différence de prédiction entre l'homme et le *Tetraodon* peut ainsi s'expliquer par la présence de ces gènes qui n'appartiennent plus à la bonne classe d'isochores chez le *Tetraodon*.

Homme	Gènes classés dans **l'isochore H**	Gènes classés dans **l'isochore L**
Prédictions par le **modèle *H***	91 %	5 %
Prédictions par le **modèle *L***	9 %	95 %

Tab. 4.3 : *Prédictions des modèles humains H et L sur les gènes orthologues humains H et L.*

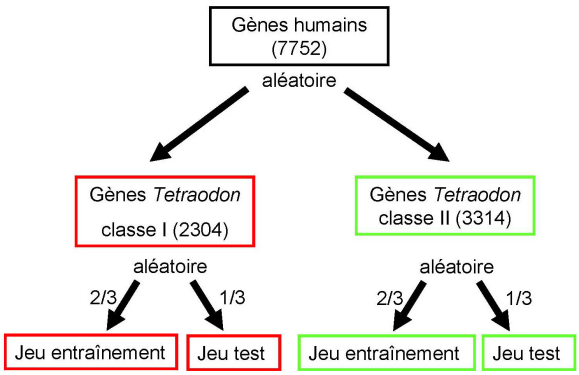


Fig. 4.7 : *Répartition des données nécessaires à l'entraînement des modèles I et II chez le Tetraodon.*

B) Comparaison de modèles aléatoires

Pour confirmer ces premiers résultats, le jeu de gènes orthologues a cette fois été séparé aléatoirement en deux jeux de données, qui correspondent aux classes nommées I et II(Figure 4.7). Puis, chaque classe est divisée aléatoirement en un jeu d'entraînement et un jeu test constitués respectivement de 2/3 et 1/3 des gènes. Deux modèles sont construits à partir des jeux d'entraînements des classes I et II, puis leurs prédictions sont étudiées sur les jeux tests I et II. Les comparaisons des prédictions des modèles *I* et *II* (tableau 4.4) montrent qu'il n'existe pas de différence significative entre les classes aléatoires I et II (p-valeur = 8.10^{-1} pour le test du χ^2). Ces résultats permettent d'éliminer l'hypothèse d'un artefact lié à notre méthode et confirment la différence de structure entre les classes " H " et " L " du *Tetraodon*.

Tetraodon	Gènes classés dans **l'isochore I**	Gènes classés dans **l'isochore II**
Prédictions par le **modèle *I***	54 %	47 %
Prédictions par le **modèle *II***	46 %	53 %

Tab. 4.4 : *Prédictions des modèles I et II sur les classes de gènes orthologues I et II.*

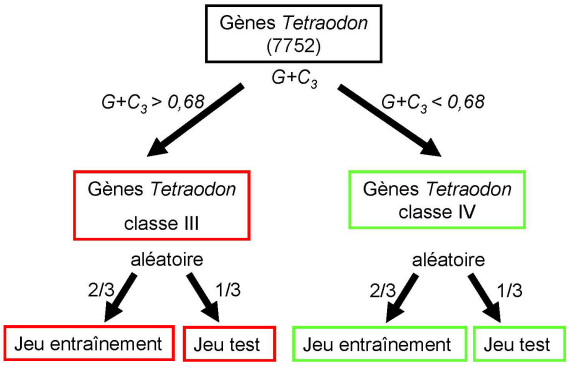


Fig. 4.8 : *Répartition des données nécessaires à l'entraînement des modèles III et IV chez le Tetraodon.*

C) Existe-t-il une influence de la composition en $G+C$ sur les modèles " H " et " L " ?

Il est intéressant de constater que les gènes du *Tetraodon* orthologues aux gènes humains présents dans les isochores H ont un contenu en $G+C_3$ moyen (0,69) supérieur à celui des gènes qui sont orthologues aux gènes humains appartenant aux isochores L (0,66). La différence est sensiblement significative, la p-valeur du test de Wilcoxon étant de 2.10^{-3}. Il semble que la différentiation du contenu en $G+C_3$ se soit conservée entre les deux espèces, bien que celle-ci soit nettement plus faible chez le *Tetraodon* que chez l'homme. Afin de déterminer si le contenu en $G+C_3$ influence le classement des gènes dans les deux types d'isochores " H " et " L ", deux classes (III et IV) sont constituées en séparant les gènes orthologues du *Tetraodon* suivant leur contenu en $G+C_3$ (Figure 4.8). La médiane de la distribution du contenu en $G+C_3$ (0,68) sert de limite pour classer les gènes, les classes

Tetraodon	Gènes classés dans **l'isochore III**	Gènes classés dans **l'isochore IV**
Prédictions par le **modèle *III***	56,5 %	41,5 %
Prédictions par le **modèle *IV***	43,5 %	58,5 %

Tab. 4.5 : *Prédictions des modèles III et IV sur les classes de gènes orthologues III et IV.*

III et IV correspondent respectivement aux classes dont les gènes ont un fort et faible taux en $G + C_3$. Puis, chaque classe est divisée aléatoirement en un jeu d'entraînement et un jeu test constitués respectivement de 2/3 et 1/3 des gènes. Deux modèles sont entraînés et comparés à partir des classes III et IV (Tableau 4.5). Une différence significative est observée entre les deux classes (p-valeur= 5.10^{-8} pour le test du χ^2). Toutefois, la différentiation des classes III et IV est moins importante que celle des classes "H" et "L", les p-valeurs valant respectivement 6.10^{-8} et $< 2.10^{-16}$. Le contenu en $G + C_3$ des gènes semble donc être à l'origine d'une des caractéristiques qui permet la différenciation des classes " H " et " L " obtenue à partir des gènes orthologues humains des isochores H et L (cas A). Mais d'autres facteurs accentuent également cette différenciations.

Analyse des gènes prédits dans les isochores " H " et " L "

L'objectif de ce paragraphe est d'identifier les caractéristiques des classes " H " et " L " qui ont permis la différenciation de leurs gènes par les modèles de Markov. Pour cela, quelques propriétés biologiques des gènes constituant les classes " H " et " L " ont été étudiées, tel que leur contenu en $G + C$, la longueur des exons et des introns ainsi que le nombre d'exons par gènes. Pour cela, seul les gènes orthologues dont les annotations sont complètes ont été utilisés. Les résultats sont présentés dans le tableau 4.6. Ils mettent en évidence deux caractéristiques. Dans un premier temps, il est possible d'observer une certaine conservation des propriétés qui caractérisent les gènes humains classés dans les isochores H et L chez leur orthologue *Tetraodon*. Le paragraphe précédent a montré une différentiation suivant le $G + C_3$, en revanche, le contenu en $G + C$ des régions UTRs ne varie pas. Chez le *Tetraodon* comme chez l'homme, le nombre d'exons, la longueur des introns et du premier exon sont supérieurs pour les gènes contenus dans la classe H

Propriétés biologiques	Isochore H humain			Isochore L humain		
	Tetraodon gène prédit "H"	*Tetraodon* gène prédit "L"	Wilcoxon p-valeur	*Tetraodon* gène prédit "H"	*Tetraodon* gène prédit "L"	Wilcoxon p-valeur
nb gènes	443	265		452	652	
GC_3 CDS	73%	68%	1.10^{-6}	66%	65%	2.10^{-3}
GC Introns	47%	44%	3.10^{-2}	45%	44%	4.10^{-1}
GC UTR	43%	43%	7.10^{-1}	43%	44%	1.10^{-1}
Nombre exons	7,3	10,2	2.10^{-16}	10,1	14,2	2.10^{-16}
Longueur exon initiaux (bp)	225	146	1.10^{-9}	187	156	4.10^{-2}
Longueur exon internes (bp)	170	151	2.10^{-1}	153	152	4.10^{-1}
Longueur exon terminaux (bp)	291	275	9.10^{-1}	212	224	4.10^{-1}
Longueur intron (bp)	741	605	5.10^{-4}	462	452	4.10^{-1}
Notation	hum.H_ tetra.H	hum.H_ tetra.L		hum.L_ tetra.H	hum.L_ tetra.L	

Tab. 4.6 : *Caractéristiques des différentes régions prédites. Les valeurs données dans ce tableau correspondent aux valeurs moyennes obtenues.*
Les classes suivantes sont introduites :
hum.H _ tetra.H correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore H et qui est prédit en H par les modèles HMMs.
hum.H _ tetra.L correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore H et qui est prédit en L par les modèles HMMs.
hum.L _ tetra.L correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore L et qui est prédit en L par les modèles HMMs.
hum.L _ tetra.H correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore L et qui est prédit en H par les modèles HMMs.

par rapport à ceux contenus dans la classe L, les p-valeurs valant respectivement : $<2.10^{-16}$, 4.10^{-2} et 8.10^{-4}. Dans un deuxième temps, notre modèle a reclassé certains gènes du *Tetraodon* (colonnes 3 et 5 du tableau 4.6), ce qui suggère que ces gènes ont probablement changé d'isochores au cours de l'évolution des deux espèces. Les propriétés qui ont permis le reclassement de ces gènes sont : le contenu en $G + C_3$, le $G + C$ des introns, le nombre d'exons ainsi que la longueur des premiers exons et des introns (colonnes 4 et 7 du tableau 4.6). Ces caractéristiques reflètent la même organisation chez les mammifères et chez le *Tetraodon*.

Une seconde étude a consisté à réaliser une analyse factorielle des correspondances à partir des fréquences des mots de six lettres utilisés par les HMMs. Plus précisément, cette AFC permet de comparer les différentes classes de gènes (tableau 4.6 : hum.H _ tetra.H...) pour les régions CDS, introns et UTRs (Figure 4.9). Deux conclusions ressortent de cette analyse. D'une part, il existe une différentiation des mots de 6 lettres entre les gènes du *Tetraodon* qui appartiennent à la classe " H " et ceux qui appartiennent à la classe " L ". Les ellipses rouges et bleues représentées sur la figure 4.9 mettent en évidence ces deux groupes distincts aussi bien dans les régions codantes que dans les régions introniques et les régions UTRs. Cette conservation des mots de six lettres entre le *Tetraodon* et l'homme est plus forte dans les CDS (représentée par le premier axe) que dans les introns ou les régions UTRs (représentée par le deuxième axe). Ceci est probablement dû aux contraintes plus fortes exercées sur la partie codante du gène. D'autre part, au sein même des gènes du *Tetraodon* dont l'orthologue humain appartient à l'isochore H, une différentiation entre les gènes du *Tetraodon* qui sont prédits en "H " et ceux prédits en " L " peut être observée, délimitée respectivement par le deuxième axe chez les CDS et par le premier axe chez les introns et régions UTRs. Cette différentiation est moins nette dans les isochores L.

4.2.3.2 Carte des " isochores " le long du génome du *Tetraodon*

Les cartes des " isochores " obtenues à partir des modèles HMMs du *Tetraodon* sont montrées sur la figure 4.10. Pour chaque chromosome la répartition du $G + C$, la densité en gènes et la répartition des isochores obtenue par notre approche sont représentées. Ces cartes mettent nettement en évidence la présence d'une structure en mosaïque le long du génome du

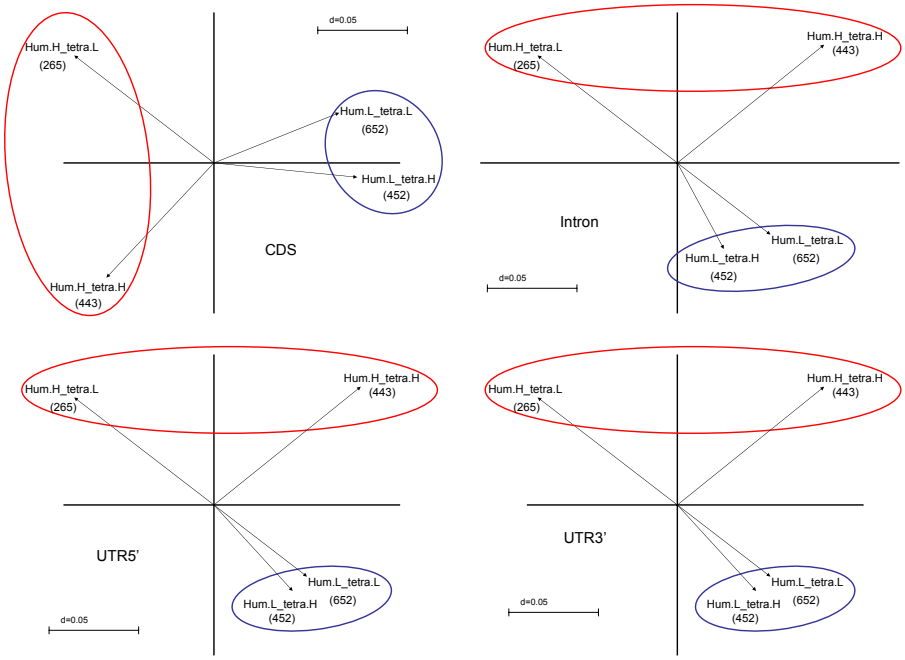


Fig. 4.9 : *Analyse des correspondances à partir des fréquences des mots de 6 lettres des CDS, introns et régions UTRs suivant les prédictions des modèles " H " et " L ". L'ellipse rouge regroupe les gènes du Tetraodon dont l'orthologue humain est situé dans un isochore H. L'ellipse bleue regroupe les gènes du Tetraodon dont l'orthologue humain est situé dans un isochore L.*

Les classes suivantes sont introduites :

hum.H _ tetra.H correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore H et qui est prédit en H par les modèles HMMs.

hum.H _ tetra.L correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore H et qui est prédit en L par les modèles HMMs.

hum.L _ tetra.L correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore L et qui est prédit en L par les modèles HMMs.

hum.L _ tetra.H correspond aux gènes du Tetraodon dont l'orthologue humain appartient à un isochore L et qui est prédit en H par les modèles HMMs.

Tetraodon.

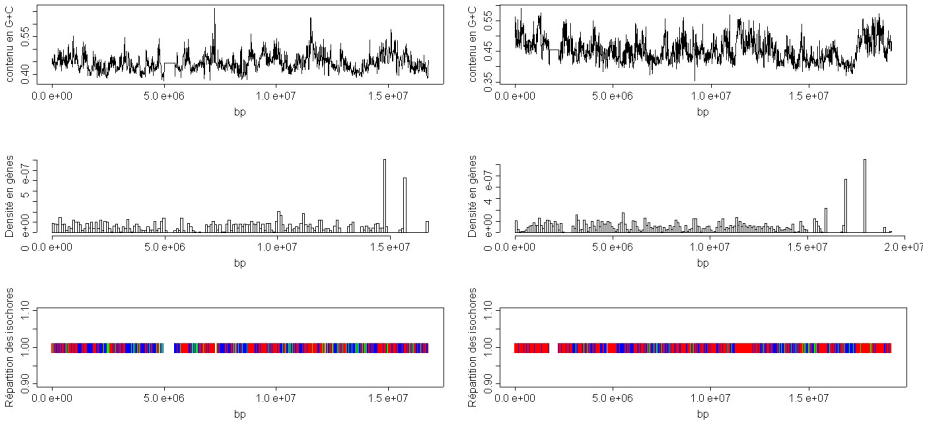
4.2.4 Discussion

Notre méthode a permis d'obtenir une nette segmentation du génome du *Tetraodon* suivant trois classes nommées " H ", " L " et " M " (figure 4.10). Jaillon *et al*. 2004 ont montré que le long du génome du *Tetraodon* il existe une alternance entre des zones relativement riches en gènes et d'autres plus pauvres en gènes. Notre étude permet de mettre en relation l'existence de telles régions et le découpage en " isochores " obtenu pour le génome du *Tetraodon*. La représentation de la densité en gènes le long des chromosomes est à prendre avec précaution dans la mesure où les données ont été extraites des premières annotations obtenues par la banque Ensembl (annotations de Genscan pour beaucoup) qui comptent 27918 gènes alors que Jaillon et *al*. estiment qu'il n'y aurait que 22400 gènes environ. De nombreuses zones riches en $G+C$ sont présentes le long des différents chromosomes et correspondent fréquemment à des régions identifiées comme appartenant à " un isochore H " par nos modèles. Cependant, quelques zones identifiées comme étant des " isochores L " sont observables, ces régions sont généralement plus petites que celles décrites comme étant des "isochores H". Il semblerait que l'organisation en " isochores " soit inversée par rapport à celle du génome humain où ce sont les régions L qui sont les plus longues.

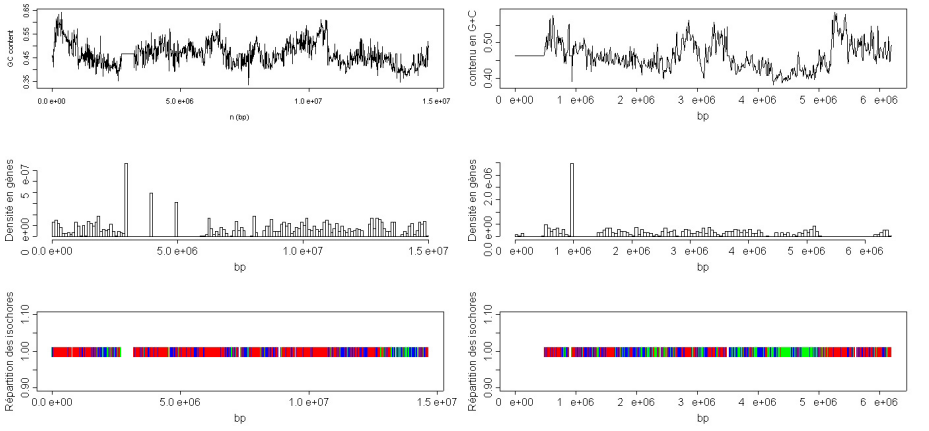
4.3 Analyse de l'organisation compositionnelle du génome du fugu

4.3.1 Introduction

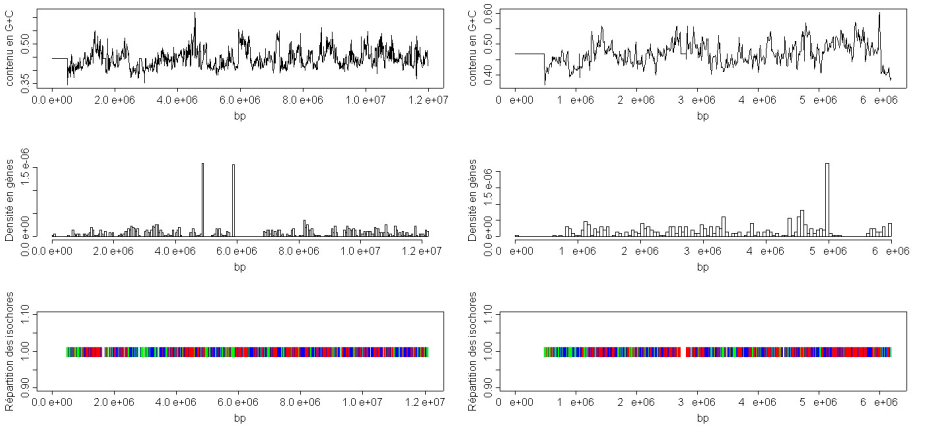
Le fugu appartient, comme le *Tetraodon nigroviridis*, à la famille des tétraodontidés dont il existe une vingtaine de variétés différentes. Il est surnommé poisson-ballon car son estomac se gonfle d'eau et peut doubler de volume quand il a peur. Le foie et les organes sexuels de l'animal sécrètent un poison toxique, la tétrodotoxine, contre lequel il n'existe aucun antidote et dont l'ingestion peut être mortelle. Son génome vient d'être séquencé mais n'est pas encore entièrement assemblé. Tout comme le *Tetraodon*, le génome du fugu présente une taille modeste (huit fois plus petite que le génome humain) (Tableau 4.7) et contient de nombreux points communs



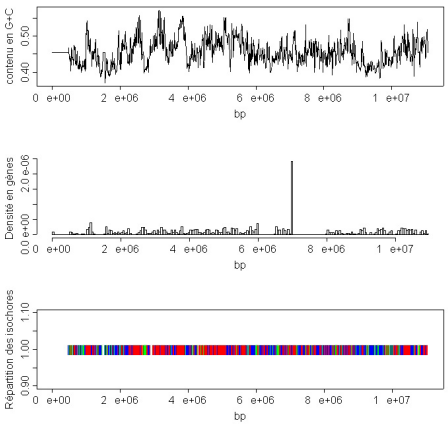
Chromosome 1 Chromosome 2



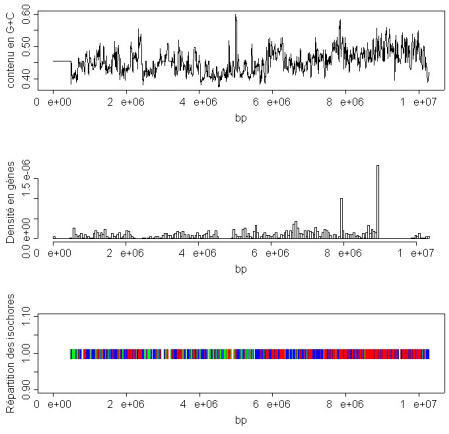
Chromosome 3 Chromosome 4



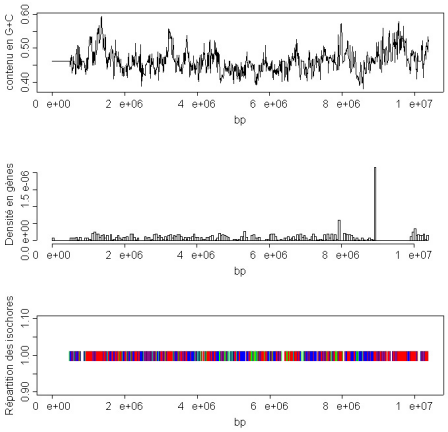
Chromosome 5 Chromosome 6



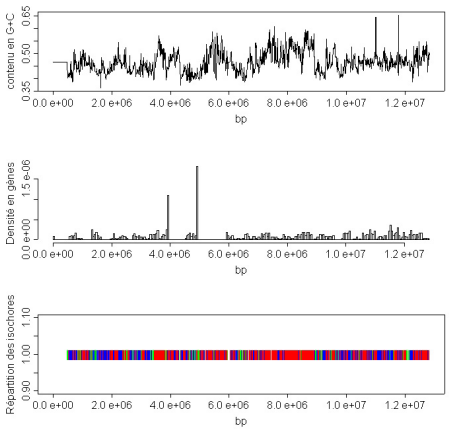
Chromosome 7



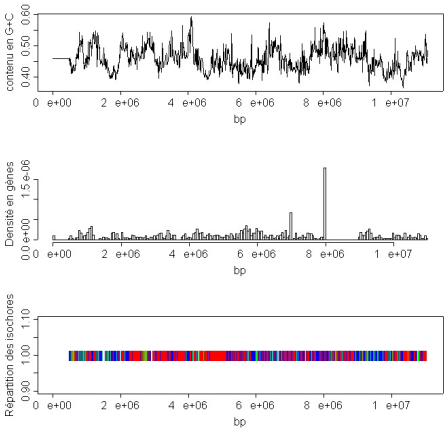
Chromosome 8



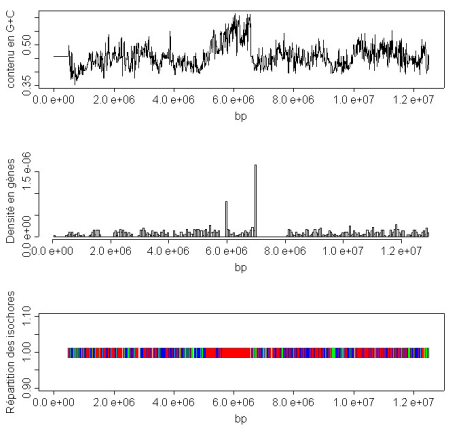
Chromosome 9



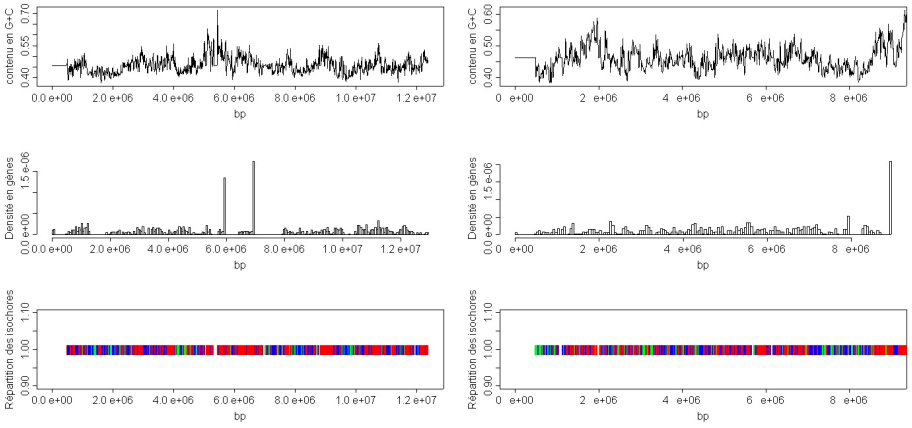
Chromosome 10



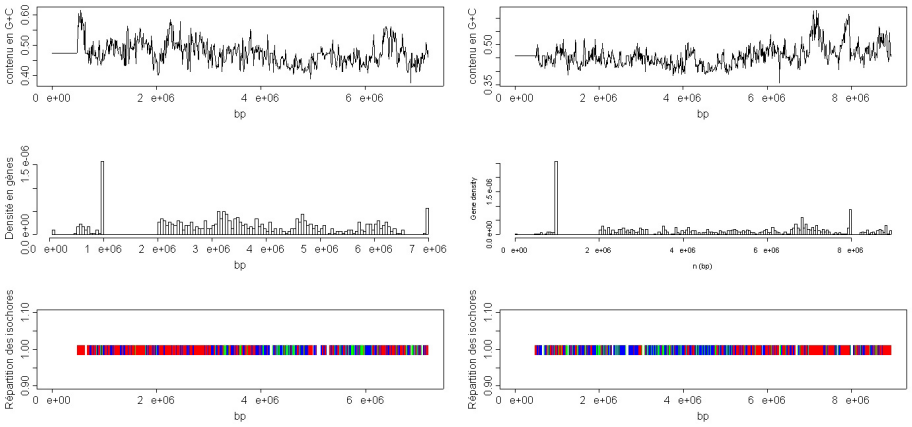
Chromosome 11



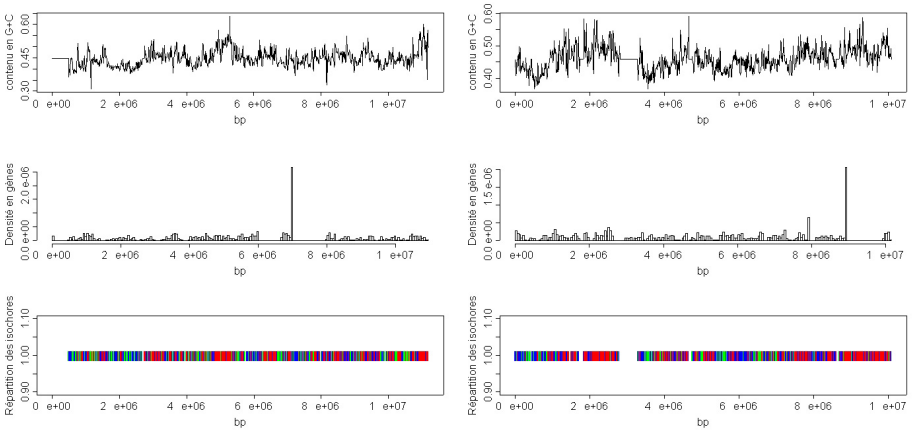
Chromosome 12



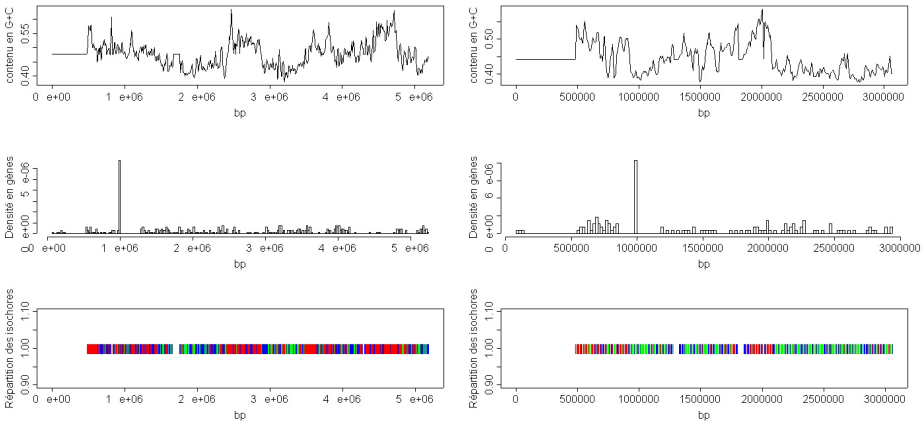
Chromosome 13 Chromosome 14



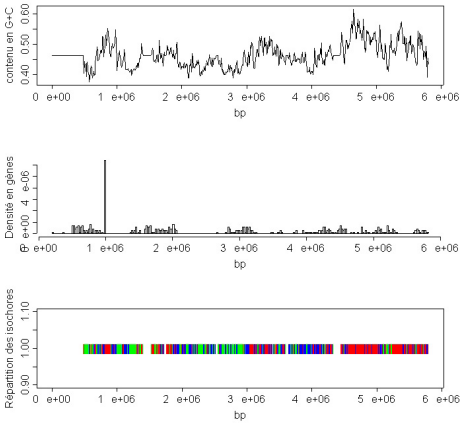
Chromosome 15 Chromosome 16



Chromosome 17 Chromosome 18



Chromosome 19 Chromosome 20



Chromosome 21

Fig. 4.10 : *L'évolution du taux en $G+C$, la répartition des gènes et la répartition des isochores prédite par notre modèle sont représentés le long de chaque chromosome du Tetraodon nigroviridis. En rouge, sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.*

Fig. 4.11 : *fugu ou poisson ballon.*

Descriptif	Fugu	Humain
Gènes	environ 22000	environ 24000
Nombre d'exons par gènes	10,5	8,7
Taille du génome	environ 329 Mb	environ 28000 Mb
$G+C$	homogène et riche	hétérogène
Longueur moyenne des introns	690 bp	environ 2067 bp

Tab. 4.7 : *Comparaison de quelques propriétés biologiques des génomes de l'homme et du fugu.*

avec l'homme. Ainsi, la comparaison des deux génomes a permis de prédire l'existence d'environ mille gènes humains qui étaient encore inconnus.

Cette étude complémentaire du génome du fugu, proche de celui du *Tetraodon nigroviridis*, a été conduite dans le but de s'assurer que les résultats encourageants concernant la possibilité de l'existence de classes de gènes liées à la notion d'isochore, pouvait être étendu à d'autres génomes de poissons. Ainsi, la même démarche que celle présentée dans la section précédente a été appliquée au génome du fugu.

4.3.2 Matériel et méthodes

Tout comme chez le *Tetraodon nigroviridis*, la distribution en $G+C_3$ est très homogène (Figure 4.12 a et b). Le génome du fugu est riche en $G+C$ et très compact, les régions intergéniques et les régions introniques sont donc très courtes. La constitution des classes d'isochores " H ", " L " et

" M ", et l'estimation des modèles HMMs qui leurs sont associés, ont donc été réalisées à partir des gènes du fugu orthologues aux gènes humains.

À partir de la banque de données GemCore, 10757 gènes orthologues entre l'homme et le fugu ont été extraits. Les gènes orthologues du fugu ont été partagés en trois classes (nommées " H ", " M " et " L ") correspondant à la répartition en isochores des gènes humains obtenue au chapitre 3. Les classes d'isochores " H ", " M " et " L " contiennent respectivement 3185, 2918 et 4654 gènes, soit 48% du nombre total des gènes chez le fugu. Les gènes de chaque classe sont séparés aléatoirement afin de constituer un jeu d'entraînement et un jeu test contenant respectivement de 2/3 et 1/3 des gènes. Un modèle (" H ", " L " et " M ") est ajusté à chacune de ces trois classes. Les prédictions des modèles " H " et " L " du fugu ont été comparées aux prédictions de différents modèles sur les jeux tests de gènes orthologues. La procédure de comparaison des prédictions de deux modèles est la même que celle utilisée dans la section précédente.

4.3.3 Résultats

L'étude conduite dans cette section met en évidence l'existence d'une structure dépendante de l'organisation en isochores au sein du génome du fugu similaire à celle du *Tetraodon nigroviridis*.

4.3.3.1 Comparaison des modèles " H " et " L "

La première étape de validation des modèles a permis de vérifier qu'il existe une différence de structure entre les classes " H " et " L " obtenues chez le fugu tout comme chez le *Tetraodon*. Les gènes appartenant à la classe " H " sont préférentiellement reconnus par le modèle " H " (60%). Réciproquement les gènes des isochores " L " sont majoritairement préférés par le modèle " L " (65,3%) (Tableau 4.8). Il existe donc bien une différence de structure entre les deux classes " H " et " L " du fugu, celle-ci est confirmée par un test du χ^2 significatif (p-valeur = 2.10^{-7}).

4.3.3.2 Comparaison des modèles aléatoires

Afin de confirmer ces résultats, comme chez le *Tetraodon*, le jeu de gènes orthologues a été séparé aléatoirement en deux jeux de données, qui correspondent aux classes nommées I et II. Puis, chaque classe est divisée aléa-

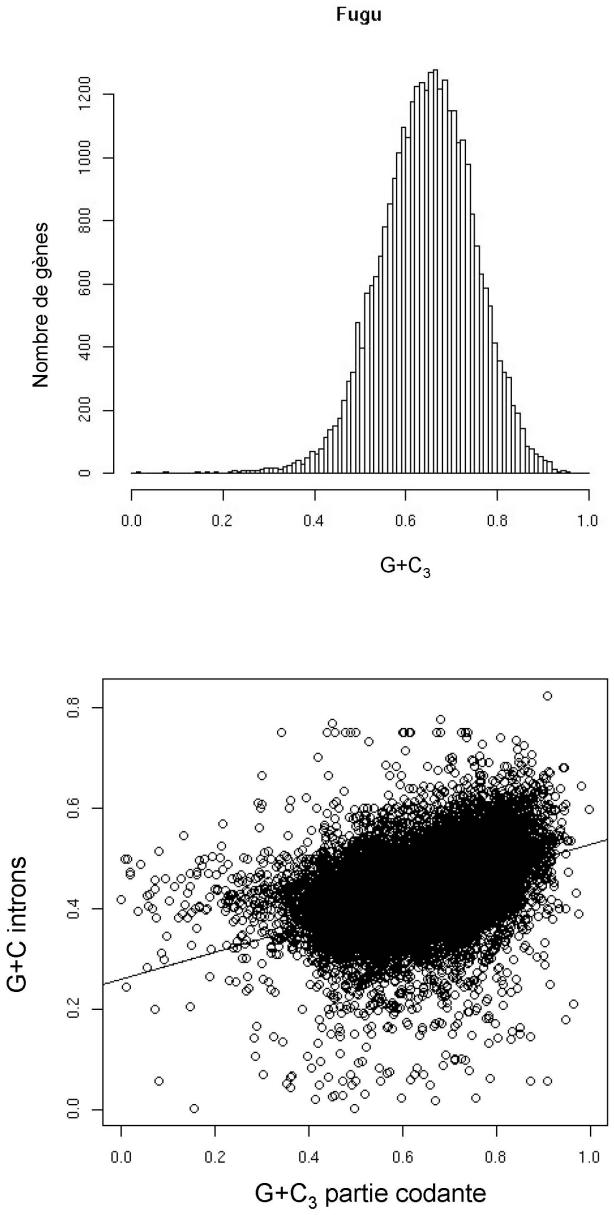


Fig. 4.12 : *(a) Répartition des gènes suivant leur $G+C_3$ chez le fugu. (b) Corrélation entre le taux de $G+C_3$ des CDS et le taux en $G+C$ des introns, R^2=0.22.*

fugu	Gènes classés dans **l'isochore "H"**	Gènes classés dans **l'isochore "L"**
Prédictions par le **modèle *"H"***	60 %	34,7 %
Prédictions par le **modèle *"L"***	40 %	65,3 %

Tab. 4.8 : *Prédiction des HMMs du fugu " H " et " L " sur les gènes orthologues des jeux tests du fugu " H " et " L ".*

	Gènes classés dans **l'isochore I**	Gènes classés dans **l'isochore II**
Prédictions par le **modèle *I***	52 %	46 %
Prédictions par le **modèle *II***	48 %	54 %

Tab. 4.9 : *Prédiction des modèles I et II sur les classes de gènes orthologues I et II.*

toirement en un jeu d'entraînement et un jeu test constitués respectivement de 2/3 et 1/3 des gènes. Deux modèles sont construits à partir des jeux d'entraînement des classes I et II, puis leurs prédictions sont étudiées sur les jeux tests I et II. Les comparaisons des prédictions des modèles I et II (Tableau 4.9) montrent qu'il n'existe pas de différence significative entre les classes aléatoires I et II (p-valeur = 7.10^{-1} pour le test du χ^2). Ces résultats permettent d'éliminer l'hypothèse d'un artefact lié à notre méthode et confirment la différence de structure entre les classes " H " et " L " du fugu.

4.3.3.3 Comparaison de quelques caractéristiques biologiques des gènes de l'homme, du *Tetraodon* et du fugu

La très faible corrélation entre le taux en $G + C_3$ des gènes humains et celui des gènes des deux poissons considérés est mise en évidence par les figures 4.13. Ces résultats confirment notre hypothèse suivant laquelle il existerait une organisation particulière chez les génomes des poissons, liée à la structure en isochores du génome humain, qui ne peut être expliquée uniquement par le contenu en $G + C_3$ des gènes.

Une analyse factorielle des correspondances (AFC) est réalisée à partir des fréquences des mots de six lettres présents dans les CDS et les introns chez l'homme, le *Tetraodon* et le fugu. Cette analyse met en évidence quatre caractéristiques biologiques importantes. Dans un premier temps, la figure

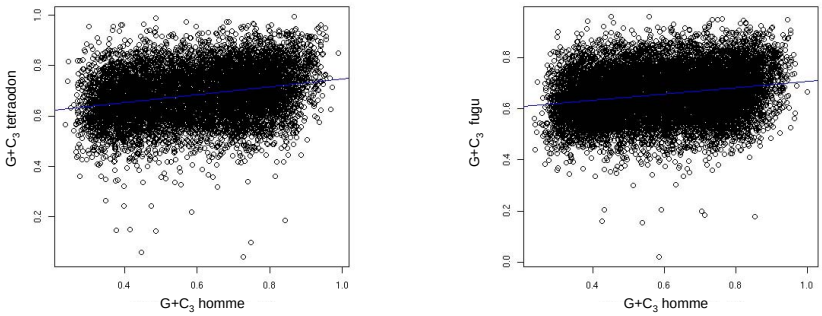


Fig. 4.13 : *(a) Corrélation du taux en $G+C_3$ des gènes orthologues entre le Tetraodon et l'homme (R^2=0,07). (b) Corrélation du taux en $G+C_3$ des gènes orthologues entre le fugu et l'homme (R^2=0,047).*

4.14(a) montre une différence de structure au niveau de la composition des mots de six lettres entre la lignée des mammifères et celle des poissons, celle-ci étant caractérisée par le deuxième axe (ellipse bleue et rouge). Dans un deuxième temps, la forte homogénéité du contenu en $G+C$ des gènes chez le *Tetraodon* et le fugu est caractérisée par le regroupement le long du premier axe de leurs exons et de leurs introns appartenant aux classes H et L. En revanche, la forte diversité du contenu en $G+C$ des gènes chez l'homme est représentée par la séparation le long du premier axe entre les exons qui appartiennent à l'isochore H et ceux situés dans l'isochore L. Dans un troisième temps, la figure 4.14 (b) met en valeur une différentiation compositionnelle entre les deux espèces de poissons décrite par leur séparation le long du deuxième axe. Enfin, cette figure permet également de mettre en évidence une forte différentiation entre les exons et les introns chez les poissons, probablement dû à la structure en codons à l'intérieur des exons, celle-ci étant caractérisée par le premier axe de l'AFC.

Les résultats des comparaisons des caractéristiques des gènes des trois espèces entre les classes H, M et L des isochores prédits chez l'homme par notre approche sont présentés dans les tableaux 4.10 à 4.12. En ce qui concerne les exons, leur longueur est sensiblement égale dans les trois classes d'isochores pour les trois espèces, à l'exception des exons initiaux qui ont tendance à être plus longs dans les classes H de chacune des trois espèces. Les introns sont quant à eux plus longs chez l'homme et leur longueur dépend de la classe d'isochores, alors que cette relation n'est pas observable chez les poissons.

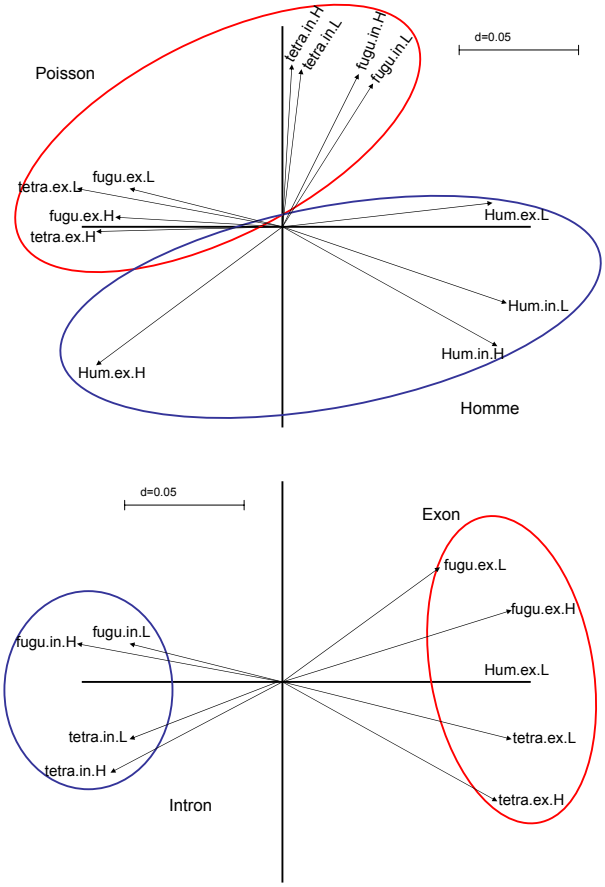


Fig. 4.14 : *Analyse des correspondances à partir des fréquences des mots de six lettres des exons et des introns suivant le taux en $G + C_3$ du gène chez l'homme, le Tetraodon et le fugu. (a) L'ellipse bleue représente l'homme. L'ellipse rouge représente les poissons. (b) l'ellipse rouge regroupe les exons et la bleue regroupe les introns.*
Les classes suivantes sont introduites :
hum.ex.H représente les exons humain des isochores H
hum.ex.L représente les exons humain des isochores L
hum.in.H représente les introns humain des isochores H
hum.in.L représente les introns humain des isochores L
tetra.ex.H correspond aux exons du Tetraodon des isochores H
tetra.ex.L correspond aux exons du Tetraodon des isochores L
tetra.in.H représente les introns du Tetraodon des isochores H
tetra.in.L représente les introns du Tetraodon des isochores L
fugu.ex.H représente les exons du fugu des isochores H
fugu.ex.L représente les exons du fugu des isochores L
fugu.in.H représente les introns du fugu des isochores H
fugu.in.L représente les introns du fugu des isochores L

Régions du gène	Longueur dans l'isochore H (moyenne bp)			Longueur dans l'isochore M (moyenne bp)			Longueur dans l'isochore L (moyenne bp)		
	Homme	*Tetraodon*	fugu	Homme	*Tetraodon*	fugu	Homme	*Tetraodon*	fugu
Exon initial	233	205	207	176	198	190	160	167	166
Exon interne	144	156	146	143	159	155	144	151	142
Exon terminal	244	253	229	237	253	240	218	233	198
Intron	1275	594	654	1809	597	611	3117	606	729

Tab. 4.10 : *Comparaison des longueurs des régions exons et introns entre l'homme, le Tetraodon et le fugu.*

Classe d'isochores	Région du gène	Contenu moyen en $G+C$ (%)		
		Homme	*Tetraodon*	fugu
Isochore H	CDS	80	72	69
	Introns	59	46	44
Isochore M	CDS	64	69	68
	Introns	51	45	44
Isochore L	CDS	43	66	64
	Introns	38	45	43

Tab. 4.11 : *Comparaison du contenu en $G+C$ entre l'homme, le Tetraodon et le fugu.*

Le nombre d'exons par gènes est similaire dans les trois espèces et une corrélation entre nombre d'exons par gène et classe d'isochore est observée pour chacune d'elles. Enfin, le contenu en $G+C$ varie fortement chez l'homme suivant la classe d'isochores, aussi bien dans les introns que dans les exons. Chez le fugu et le *Tetraodon*, cette variation est nettement plus faible au niveau des CDS, voire inexistante dans les introns.

Classe d'isochores	Nombre moyen d'exons		
	Homme	*Tetraodon*	fugu
H	8,93	8,31	9,26
M	10,76	9,94	10,02
L	12,31	11,03	11,75

Tab. 4.12 : *Comparaison du nombre de gènes entre l'homme, le Tetraodon et le fugu.*

4.4 Discussion

Les génomes du *Tetraodon nigroviridis* et du fugu sont très compacts, avec une faible représentation des éléments transposables, une forte densité en gènes et une composition en $G+C$ homogène et plus riche que chez la plupart des mammifères. Toutes ces caractéristiques ont contribué à supposer que ces deux génomes ne possèdent *a priori* pas de structure en isochores tout comme chez l'ensemble des populations de poissons et d'amphibiens. Cependant, les résultats obtenus par notre étude démontrent la présence d'une structure nouvelle chez le *Tetraodon* liée à la notion d'"isochores" (figure 4.10)

Dans un premier temps, les travaux réalisés au cours de ce chapitre ont confirmé que la définition classique des isochores ne s'applique pas aux génomes du fugu et du *Tetraodon*, comme indiqué dans la littérature. Cependant, une approche nouvelle a permis de démontrer que certaines propriétés, liées à la notion d'isochores chez les mammifères, pouvaient se retrouver chez les poissons et servir à segmenter leur génome. Ces segmentations n'étant pas uniquement liées au contenu en $G+C$, elles sont difficilement identifiables par les méthodes classiques de prédiction d'isochores. L'originalité de la méthode proposée repose sur l'hypothèse suivante : les caractéristiques des gènes contenus dans les isochores du génome humain pourraient se retrouver au sein des gènes orthologues des espèces supposées dépourvues d'isochores. Le chapitre 3 ayant montré que nos modèles HMMs étaient capables de classer les gènes dans la bonne catégorie d'isochores de manière plus ou moins indépendante du contenu en $G+C_3$, la même approche a donc été appliquée avec succès à ces deux poissons à partir des gènes orthologues à l'homme.

Les résultats obtenus sont encourageants, les cartes d'"isochores" réalisées à partir de l'assemblage du génome du tétrodon et de nos modèles révèlent une segmentation de l'ensemble des chromosomes du *Tetraodon*. Lorsque l'assemblage complet du génome du fugu sera disponible, il serait intéressant de comparer les cartes d'isochores entre les deux espèces de poissons. De plus, une structure spécifique des gènes du fugu et du *Tetraodon* associée à la structure des gènes orthologues humains a été nettement mise en évidence grâce aux comparaisons entres les modèles HMMs. La présence de cette structure dans les deux espèces laisse penser que cette organisation pourrait être généralisée à d'autres génomes et n'est pas dû à une coïncidence. Un prolongement de notre étude consisterait à vérifier si une telle

organisation peut être observée chez d'autres espèces dont le contenu en $G + C$ est homogène et très pauvre, comme par exemple le *xenope*. Notre étude démontre donc qu'il est possible de retrouver, chez une espèce, à partir de gènes orthologues, des traces d'isochores qui sont difficilement identifiables à partir d'une définition classique. Ces gènes orthologues conservent des caractéristiques liées aux isochores. Cette étude montre également que les limites du contenu en $G+C$ des classes d'isochores ne sont pas fixes, mais varient d'une espèce à l'autre suivant la teneur et l'homogénéité en $G+C$ du génome étudié. Des régions associées aux isochores humains peuvent ainsi être caractérisées et prédites grâce à des facteurs différents du contenu en $G + C$, comme par exemple la longueur des exons, des introns et la densité en gènes.

Il est toutefois important de souligner que la comparaison des performances des modèles de Markov cachés adaptés aux classes " H ", " L " et " M " à ceux adaptés à des classes aléatoires, a permis d'éliminer l'idée d'un artefact dû à notre méthode, et de renforcer l'idée d'une structure particulière des gènes liés aux isochores. Plus qu'une simple cartographie des " isochores " chez le *Tetraodon*, l'entraînement des modèles HMMs (" H", "M", "L ") et leur comparaison sur les jeux tests a mis en évidence des caractéristiques différentes entre les gènes du *Tetraodon* qui sont classés en " H " et appartiennent à " l'isochore " prédit " H " et ceux qui sont classés en " L " et appartiennent à un isochore prédit en " L " : densité en gènes, longueurs des premiers exons, fréquences des mots de 6 lettres dans les CDS, les introns, les régions 5' et 3' UTRs. Notre méthode permet donc l'identification des isochores ainsi que leur analyse grâce à un retour aux données particulièrement en cas d'échecs des modèles.

Si la présence de la structure en " isochores " est confirmée expérimentalement chez le *Tetraodon* et le fugu, alors notre étude aura montré que l'hypothèse expliquant que la structure en isochores s'est mise en place une seule fois au cours de l'évolution, après la divergence entre amphibiens et amniotes, mais avant la divergence entre oiseaux et mammifères (Hugues *et al*. 1999), serait érronée. Cette structure serait présente chez la plupart des vertébrés de manière plus ou moins visible.

Un prolongement intéressant, à notre sens, de ce travail consisterait à comparer des segments d' " isochores " qui pourraient être conservés entre les trois espèces (homme, *Tetraodon* et fugu) afin d'obtenir des renseigne-

ments complémentaires sur l'évolution des trois génomes, et notamment sur les causes de l'amplification des différences de composition en $G + C$ entre l'homme, le *Tetraodon* et le fugu.

Chapitre 5

Prédiction des isochores des génomes du chimpanzé, de la souris et du poulet

Ce chapitre présente les cartes d'isochores du chimpanzé, de la souris et du poulet qui ont été obtenues à partir de notre méthode de prédiction. Le but de ce chapitre étant de fournir le matériel nécessaire pour une étude ultérieure de l'évolution des isochores entre ces différentes espèces, seule une analyse succincte a été pratiquée. Ce chapitre est constitué de deux parties distinctes. La première est une introduction qui rappelle l'importance de l'analyse comparative des génomes entre différentes espèces et présente les génomes des organismes étudiés. La deuxième partie débute par une comparaison succincte de l'évolution du contenu en $G + C_3$ chez les trois espèces étudiées par rapport à celui de l'homme présenté au chapitre 3. Les cartes d'isochores obtenues par notre méthode pour chacune des trois espèces sont ensuite détaillées.

5.1 Introduction

L'objectif de l'analyse comparative des génomes est la compréhension de leur organisation, de leur évolution et de leur fonctionnement. Ainsi, l'analyse comparative des séquences est une approche très efficace pour repérer les caractéristiques fonctionnelles du génome, même celles qui sont soumises à de faibles pressions de sélection. L'analyse comparative s'avère donc être un outil très utile en complément de l'approche expérimentale. C'est pour cette raison que le séquençage des génomes entiers de différents organismes a été entrepris parallèlement au séquençage du génome humain. Comprendre le fonctionnement du génome nécessite d'identifier non seulement les gènes mais également toutes les autres régions qui sont soumises à une pression de sélection. Il faut noter que ce n'est pas nécessairement la séquence elle-même (*i.e.* l'enchaînement des nucléotides) qui est contrainte, certaines caractéristiques plus globales de l'organisation des chromosomes (propriété compositionnelle, configuration spatiale, répartition des gènes...) peuvent également être importantes pour le fonctionnement du génome. L'analyse comparative joue également un rôle capital lors de l'étude de l'évolution des génomes entre espèces plus ou moins proches. Ainsi, notre méthode de détection des isochores présentée au chapitre 3 a été étendue à différents organismes pour permettre une analyse ultérieure de l'influence des isochores sur le fonctionnement et l'évolution des génomes.

Notre choix s'est porté sur les génomes du chimpanzé, de la souris et du poulet pour plusieurs raisons. Premièrement, ils offrent la possibilité de comparer l'évolution des isochores entre les mammifères et les oiseaux à partir de génomes entièrement séquencés. De plus, le génome du chimpanzé et celui de l'homme possèdent un très fort taux de similarité (99% du génome) ; il est donc intéressant de vérifier si ces similitudes sont conservées au niveau de l'organisation en isochores des deux génomes. De plus, le génome de la souris est connu pour sa faible hétérogénéité, celui du poulet est quant à lui réputé pour être très riche en $G + C$ et constitué de nombreux petits chromosomes. L'utilisation de notre méthode sur de tels génomes permet de vérifier sa sensibilité et sa précision. Le tableau 5.1 présente quelques caractéristiques de ces quatre génomes.

	Homme	Chimpanzé	Souris	Poulet
Nombre de chromosomes	24	25	21	30
Nombre de gènes	24194	22524	17784	28069
Contenu en $G+C$	Hétérogène	Hétérogénéité Faible	Hétérogène	Riche en $G+C$
Taille du génome (MB)	3272	2733	2932	1054

Tab. 5.1 : *Description des génomes de l'homme, du chimpanzé, de la souris et du poulet.*

5.2 Matériel

L'ensemble des données, utilisées dans de ce chapitre, est extrait de la banque Ensembl (mise à jour d'octobre 2004). Au total, chez le chimpanzé, la souris et le poulet 22524, 13932 et 28491 gènes ont respectivement été obtenus. Les génomes de ces espèces étant supposés posséder une hétérogénéité liée à la structure en isochores comme chez la plupart des mammifères et des oiseaux (Bernardi 2000), la procédure de mise en place des modèles est la même que celle développée pour l'analyse du génome humain (cf. chapitre 3). Pour chacune de ces espèces, les gènes ont été séparés en trois classes H, L et M de manière à posséder 1/3 des gènes dans chacune des classes, ce qui garantit suffisamment de données pour l'entraînement de modèles d'ordre 5. Puis, pour chacune de ces classes, les gènes ont été répartis aléatoirement et de manière égale en un jeu test et un jeu d'entraînement. Les limites des classes H, L et M obtenues pour ces trois espèces sont sensiblement différentes des limites des classes d'isochores obtenues pour l'homme qui sont : H = taux supérieur à 72%, M = taux compris entre 56% et 72% et L = taux inférieur à 56%. Chez le chimpanzé, la souris et le poulet, la classe H contient les gènes dont le taux en $G+C_3$ est supérieur respectivement à 66%, 63%, 58%. La classe L contient les gènes dont le $G+C_3$ est respectivement inférieur à 49%, 52%, 45%. Et enfin, la classe M est constituée des gènes dont le contenu en $G+C_3$ est intermédiaire. Pour chacune des classes, un modèle HMM est ajusté à chaque espèce. Ces différences peuvent s'expliquer simplement. Les gènes extraits d'Ensembl sont encore en court d'annotation, alors que pour le modèle concernant le génome humain, seuls les gènes dont le transcrit d'ARN a été séquencé ont été extraits de la banque Hovergen.

Pour comparer l'évolution du contenu en $G+C_3$ des CDS entre les gènes des organismes étudiés et ceux de l'homme, il a été extrait respectivement 20252, 14913 et 6838 gènes orthologues entre l'homme et les trois espèces :

chimpanzé, souris et poulet à partir de la banque GeMCore. L'ensemble des calculs a été réalisé au centre de calculs de l'IN2P3. Les graphiques et analyses statistiques sont obtenus à partir du logiciel R.

5.3 Méthode

À partir de ces données, un modèle est adapté à chacune des trois espèces et pour chaque classe d'isochore. Les distributions des longueurs des exons et des introns sont ajustées en minimisant la distance de Kolmogorov-Smirnov, les résultats sont présentés en annexe. Les modèles de Markov cachés prennent en compte les mêmes caractéristiques que chez l'homme, structure en codons des exons, structure en double brin de l'ADN ainsi que les différences de longueur des exons et des introns suivant leur position dans le gène.

Différentes études statistiques ont été réalisées, d'une part des régressions linéaires pour analyser la répartition du contenu en $G + C_3$ entre les gènes orthologues des différentes espèces, et d'autre part des analyses factorielles des correspondances pour comparer les probabilités d'émission des différentes régions entre les diverses espèces étudiées.

Enfin, la même méthode que celle décrite dans les chapitres précédents est appliquée pour la prédiction d'isochores. Ces derniers sont localisés pour chacune des trois espèces. Les chromosomes sont assemblés à partir des données de la banque Ensembl, puis découpés en fenêtres glissantes de 100kb avec un chevauchement de 50kb. Sur chaque fenêtre est calculée $P[Modele \mid fenetre]$ pour chaque modèle H, L et M. Le modèle ayant la plus forte probabilité caractérise la fenêtre. Afin de rester cohérent avec la définition des isochores, un segment est considéré comme étant un isochore s'il est constitué d'une succession de fenêtres associées à la même classe H, L ou M et de longueur supérieure à 300kb. À partir de ces résultats, les graphiques représentant la répartition des " isochores ", le contenu en $G+C$ et la densité en gènes sont tracés le long de chacun des chromosomes du chimpanzé, de la souris et du poulet.

Régions étudiées	Longueur en bp dans Classe H		Longueur en bp dans Classe M		Longueur en bp dans Classe L	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
Exon initial	184	123	167	114	162	112
exon interne	140	121	138	118	139	117
exon final	193	138	201	130	202	127
intron	3564	1161	4726	1495	4474	1575

Tab. 5.2 : *Longueurs moyennes et médianes des exons et des introns suivant leur position dans le gène et la fréquence de $G+C$ en position 3 du codon pour le chimpanzé.*

Régions étudiées	Longueur en bp dans Classe H		Longueur en bp dans Classe M		Longueur en bp dans Classe L	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
Exon initial	205	135	195	119	178	113
exon interne	144	126	145	123	145	121
exon final	239	152	226	147	263	144
intron	3036	889	3950	1139	4367	1299

Tab. 5.3 : *Longueurs moyennes et médianes des exons et des introns suivant leur position dans le gène et la fréquence de $G+C$ en position 3 du codon pour la souris.*

5.4 Résultats

5.4.1 Caractérisation des gènes

Les tableaux 5.2, 5.3 et 5.4 présentent pour chacune des trois espèces étudiées un résumé statistique concernant les longueurs des différentes régions constituant les gènes qui ont été prises en compte dans nos modèles.

Des caractéristiques similaires concernant les longueurs des exons et des introns peuvent être observées chez les quatre espèces. Ainsi, les introns sont plus courts dans les classes H que dans les classes L. Toutefois, les introns du poulet sont plus courts que ceux de la souris et du chimpanzé alors que les longueurs des exons sont les mêmes entre les trois espèces. Les exons initiaux et finaux sont plus longs que les exons internes. Si la longueur des exons internes ne varie pas en fonction des classes de $G+C_3$, ce n'est pas le cas des premiers exons codants et finaux dont la taille est plus petite dans la

Régions étudiées	Longueur en bp dans Classe H		Longueur en bp dans Classe M		Longueur en bp dans Classe L	
	Moyenne	Médiane	Moyenne	Médiane	Moyenne	Médiane
exon initial	206	135	182	127	169	116
exon interne	148	124	144	126	142	125
exon final	213	145	201	136	191	133
intron	1972	696	2285	810	2536	888

Tab. 5.4 : *Longueurs moyennes et médianes des exons et des introns suivant leur position dans le gène et la fréquence de $G + C$ en position 3 du codon pour le poulet.*

Espèces	Nombre moyen d'exons		
	Classe H	Classe M	Classe L
Homme	8,9	10,7	12,3
Chimpanzé	7,1	8,2	8,7
Souris	9	9,36	9,63
Poulet	9	9,7	12,1

Tab. 5.5 : *Comparaison du nombre de gènes entre homme, chimpanzé, souris, poulet.*

classe L. Enfin, le nombre d'exons est supérieur dans la classe L par rapport à la classe H. Ainsi, lorsque les gènes ont un contenu en $G+C_3$ fort, ils sont plus compacts avec moins d'exons et des introns plus courts. Les propriétés bien connues reliant la classe d'isochores aux propriétés biologiques des gènes sont donc confirmées par notre étude et modélisées par nos HMMs.
Il est intéressant de constater que la distribution de la longueur des introns chez certaines espèces n'est pas géométrique. Cependant, elle peut être modélisée aisément par des sommes de lois géométriques selon la méthode décrite au chapitre 2. C'est par exemple le cas pour le génome du poulet, où la distribution de longueur des introns a été ajustée à une somme de deux lois géométriques de paramètres différents (Figure 5.1).

5.4.2 Analyse du contenu en $G + C_3$ des gènes orthologues

Bien avant d'avoir à disposition les données du séquençage de génomes entiers, la présence d'une structure compositionnelle à l'intérieur des génomes de mammifères a pu être mise en évidence. Une des premières méthodes a consisté à comparer statistiquement les compositions en $G + C_3$ des gènes orthologues entre différentes espèces (Perrin *etal*. 1987, Bernardi 2000). La figure 5.2 représente les résultats obtenus par cette méthode lors de la comparaison du contenu en $G+C_3$ des gènes orthologues entre l'homme et chacune des trois espèces étudiées. Un fort degré de similarité au niveau de l'organisation compositionnelle des gènes orthologues peut être observée (figure 5.2) entre l'homme et le chimpanzé (R^2=0,92). De même, la corrélation entre les gènes orthologues de l'homme et de la souris suivant leur contenu en $G+C_3$ (R^2=0,73) reste très élevée et confirme l'existence chez les mammifères d'une organisation en isochores. Cette corrélation est beaucoup plus faible entre l'homme et le poulet (R^2=0,34). Toutefois, elle montre que de nombreux gènes possèdent les mêmes caractéristiques compositionnelles dans les deux lignées indépendantes : mammifères et oiseaux.

La figure 5.3 met en évidence l'existence d'une relative homogénéité locale de la composition en bases. À l'échelle du gène, le taux en $G + C_3$ est corrélé au taux en $G + C$ des introns dans les trois espèces de manière plus ou moins forte. Le génome du poulet est connu pour posséder un contenu riche en $G+C$ et relativement homogène le long de son génome, ce qui se reflète également au niveau de la structure des gènes par une forte corrélation entre le $G + C_3$ des gènes et le $G + C$ des introns (R^2=0,62). En revanche,

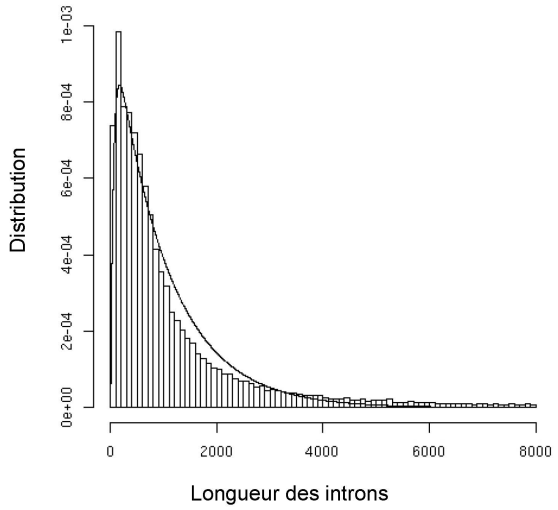


Fig. 5.1 : *Distribution de longueurs des introns de la classe H chez le poulet, ajustée par une somme de deux lois géométriques de moyennes 555 et 100.*

chez le chimpanzé et la souris, cette corrélation est plus faible, R^2 valant respectivement 0,33 et 0,39.

5.4.3 Analyse des fréquences des mots de 6 lettres chez les différentes espèces.

Des AFC ont été réalisées à partir des probabilités d'émission des états exons et introns appartenant aux modèles H et L pour chacune des espèces étudiées. Dans un premier temps, les différentes structures compositionnelles liées au $G + C$ existant chez les mammifères sont mises en évidence. Une nette séparation entre les régions H et L est observable pour chacune des quatre espèces (Figure 5.4 a) et représentée par le premier axe (abscisse). De plus, la différentiation entre les exons et les introns, probablement dû à la présence des codons, est caractéristique pour chacune des quatre espèces. Cette franche distinction est représentée par le premier axe (figures 5.4 b et c). Enfin, la répartition sur la figure 5.4 des probabilités d'émission représente l'aspect évolutif des génomes. Par exemple, le chimpanzé et l'homme sont représentés côte à côte, ce qui correspond à deux ancêtres proches. En revanche, le groupe des mammifères (homme, chimpanzé et souris) se retrouve éloigné graphiquement du groupe des oiseaux (poulet).

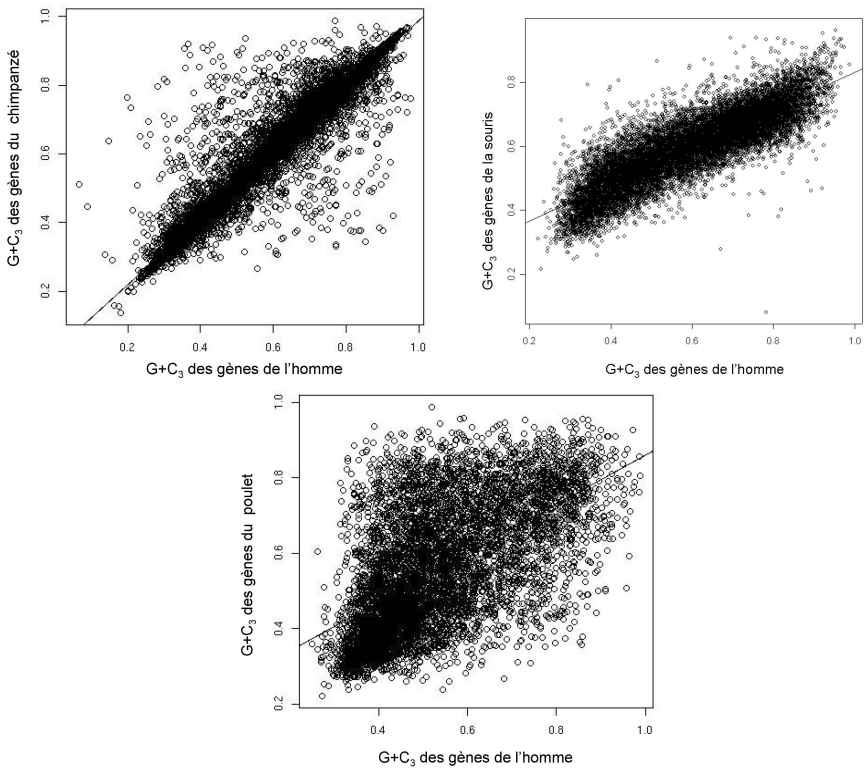


Fig. 5.2 : *Corrélation entre le $G + C_3$ des gènes orthologues de l'homme et des trois espèces étudiées : (a) de l'homme et du chimpanzé (R^2= 0,92), (b) de l'homme et de la souris (R^2=0,73) et (c) de l'homme et du poulet. (R^2=0,34). La droite de régression linéaire est tracée pour chaque figure.*

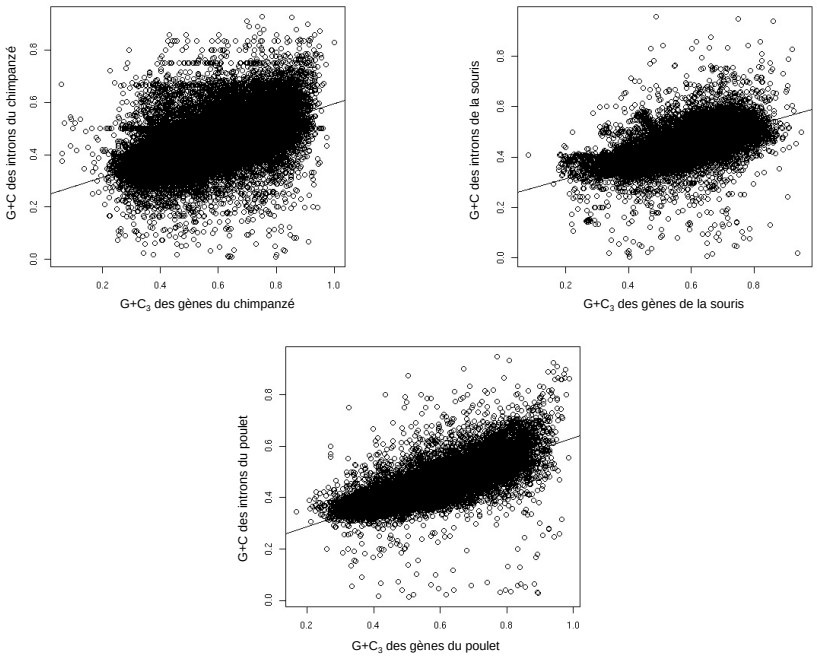


Fig. 5.3 : *Corrélations entre la composition en base des introns et des troisièmes positions des codons chez les trois espèces : (a) chez le chimpanzé (R^2= 0,33), (b) chez la souris (R^2=0,39), (c) chez le poulet (R^2=0,62).*

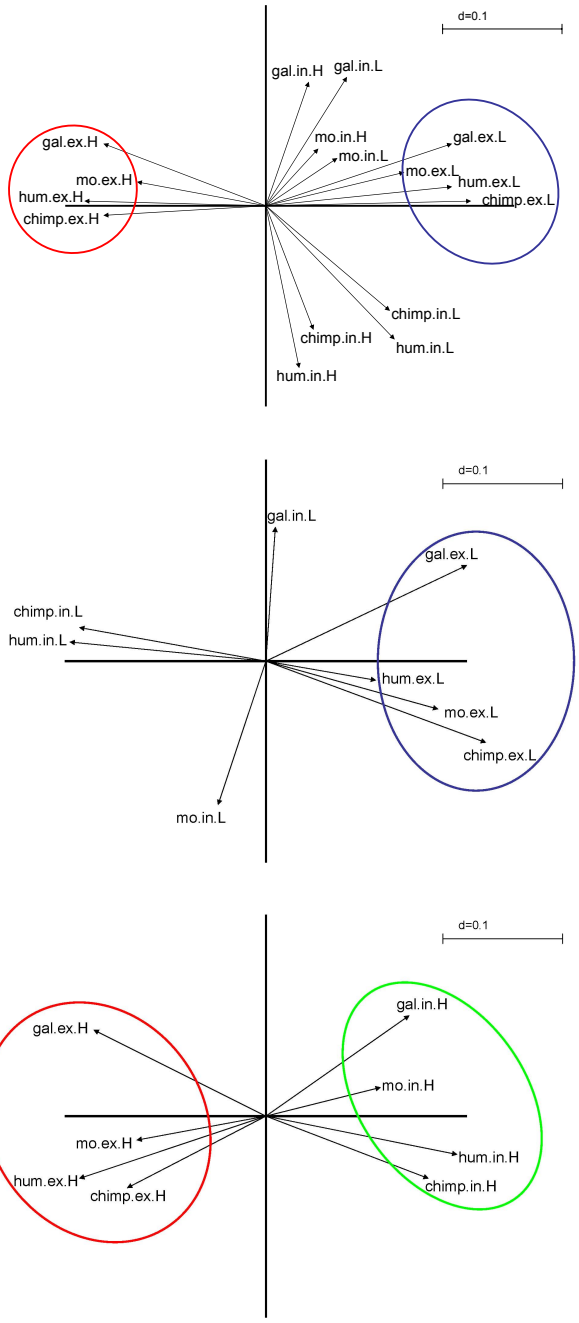


Fig. 5.4 : *AFC réalisées à partir des probabilités d'émissions des états exons et introns des isochores H et L constituant les différents modèles des espèces. Les notations utilisées sont les suivantes :*
Hum = humain, Chimp = chimpanzé, Mo = souris, Gal = poulet,
ex = exon, in = intron,
H = modèle H, L = modèle L.

5.4.4 Cartes d'isochores

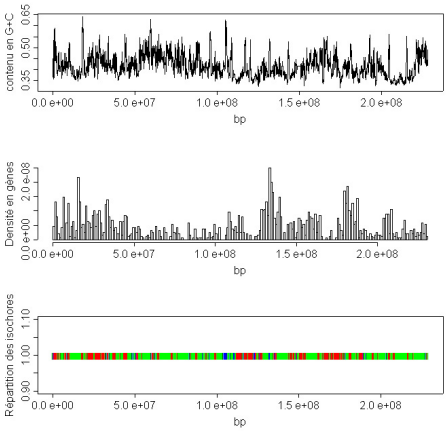
Les cartes des isochores réalisées à partir des modèles HMMs du chimpanzé, de la souris et du poulet, sont représentées de la manière suivante : pour chaque espèce et pour chaque chromosome, la répartition du $G+C$, la densité en gènes et la répartition des isochores obtenue par notre approche sont représentées. Ces cartes mettent en évidence pour chacun des trois génomes une nette segmentation suivant les trois classes d'isochores H, L et M. L'obtention de ces cartes laisse envisager une comparaison future à grande échelle des isochores entres ces différents génomes grâce à la connaissance des gènes orthologues. Ainsi, par cette méthode, une étude approfondie de l'évolution des isochores chez les vertébrés pourra être réalisée. La section qui suit est un préambule, elle décrit simplement les premières observations visuelles obtenues à partir des cartes.

5.4.4.1 Chimpanzé

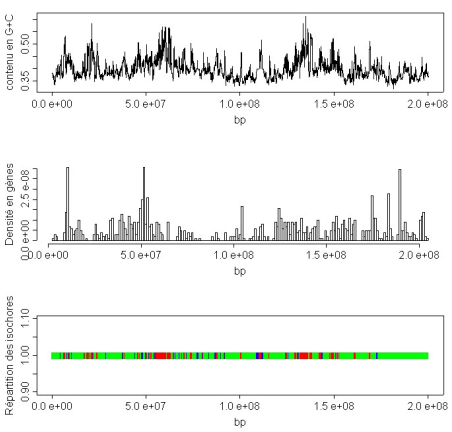
Tout comme chez l'homme, le génome du chimpanzé est très hétérogène (Figure 5.5). Il possède des zones riches en $G+C$ qui alternent avec des zones plus faibles en $G+C$, ces successions correspondant, respectivement, à l'alternance des régions riches en gènes et plus faible en gènes. La segmentation en isochores obtenue par notre approche s'ajuste bien à ces successions comme le suggèrent les propriétés biologiques liées aux isochores. Différents types de segmentations sont obtenues. Certains chromosomes sont constitués principalement d'isochores prédits en H, comme c'est le cas des chromosomes 1, 18, 19, 20, 21 et 23. À l'opposé, d'autres chromosomes sont riches en isochores L (chromosomes 2, 3, 4, 5, 6, 13, 14, 17 et 22). Les chromosomes 7, 8, 9, 10, 11, 12, 15 et 16 alternent les zones prédites en H et celles prédites en L. Les distributions des isochores prédits le long des chromosomes sexuels sont très différentes. Le chromosome X est très riche en isochores L alors que le chromosome Y est très riche en isochores M.

5.4.4.2 Souris

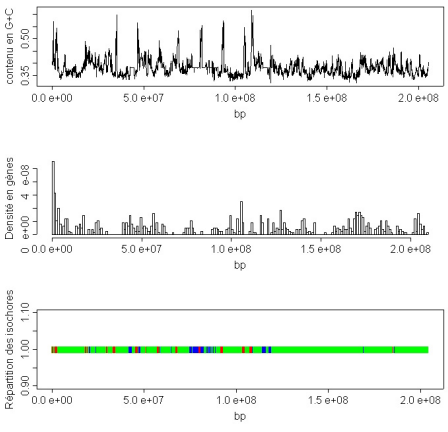
Le génome de la souris possède moins de chromosomes (19 plus les deux chromosomes sexuels) que le génome humain ou celui du chimpanzé. Une nette segmentation du génome peut être observée (Figure 5.6). La distribution des isochores prédits par notre approche est plus homogène que chez



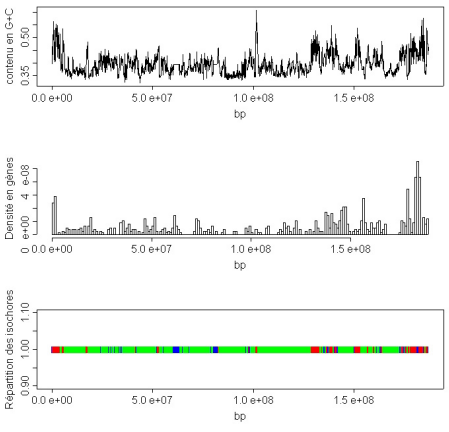
Chromosome 1



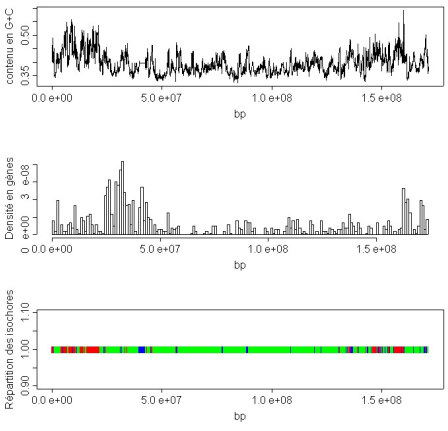
Chromosome 2



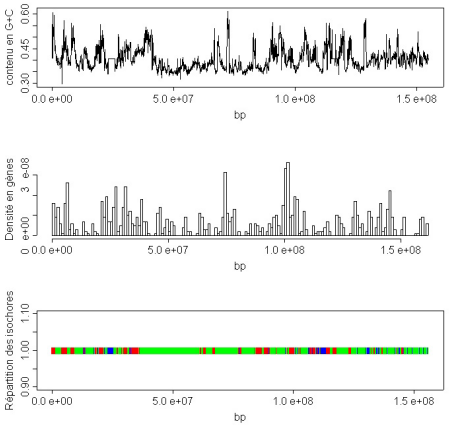
Chromosome 3



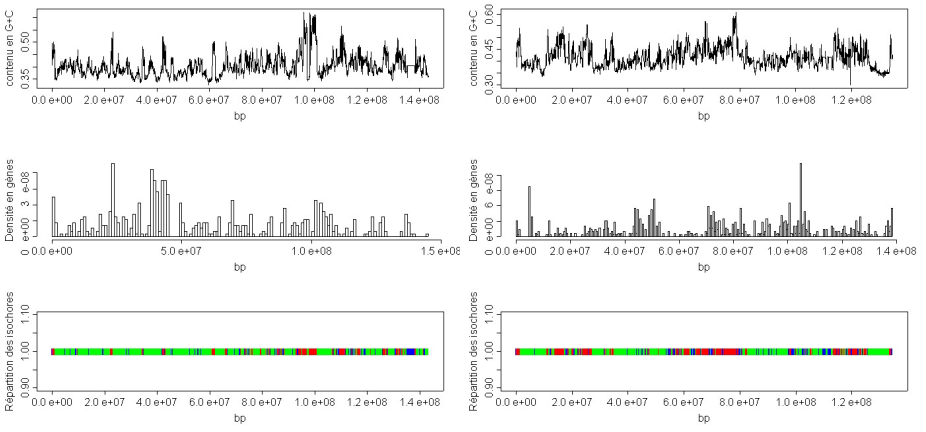
Chromosome 4



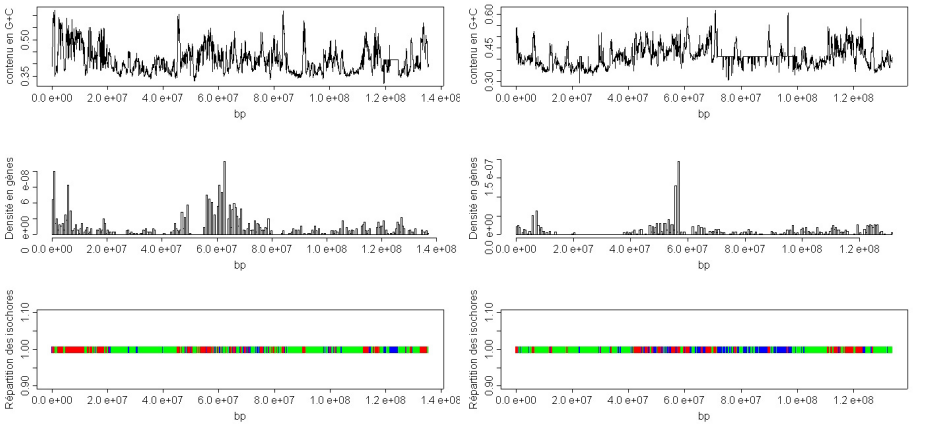
Chromosome 5



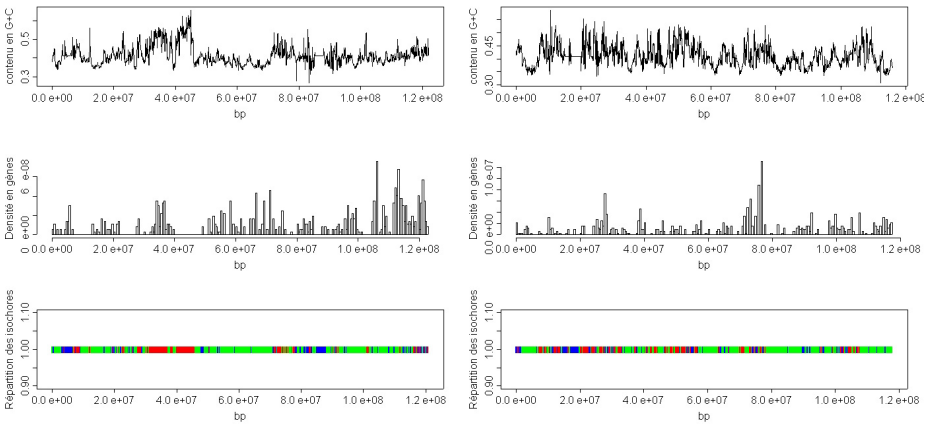
Chromosome 6



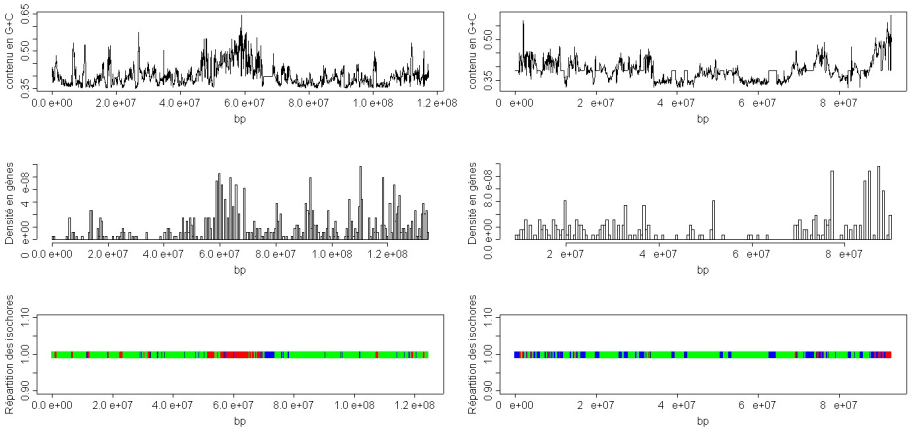
Chromosome 7 Chromosome 8



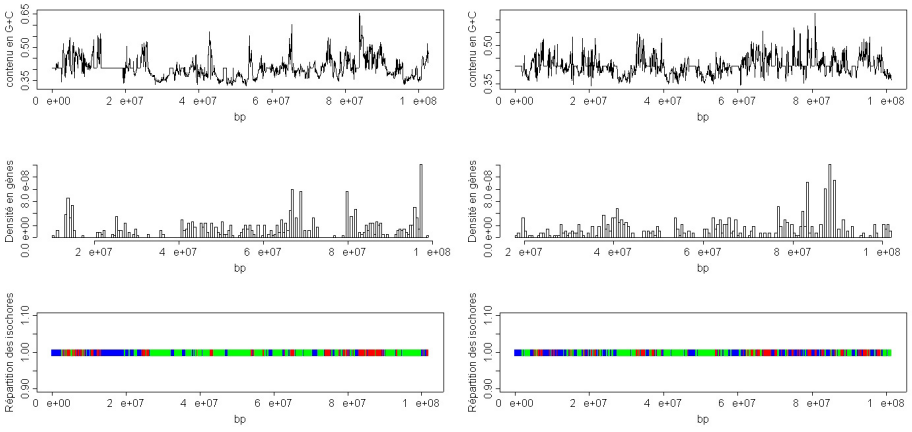
Chromosome 9 Chromosome 10



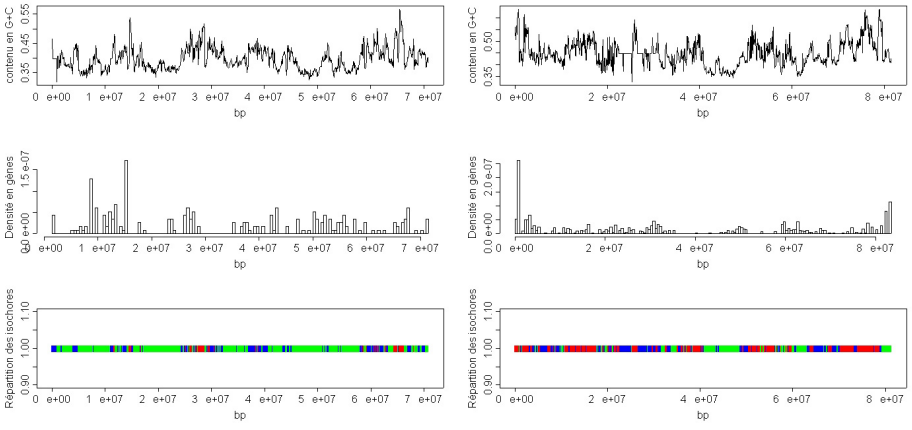
Chromosome 11 Chromosome 12



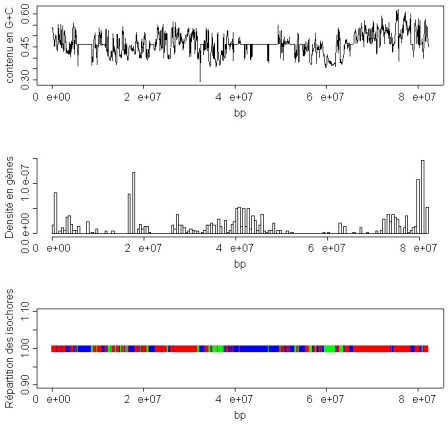
Chromosome 13 Chromosome 14



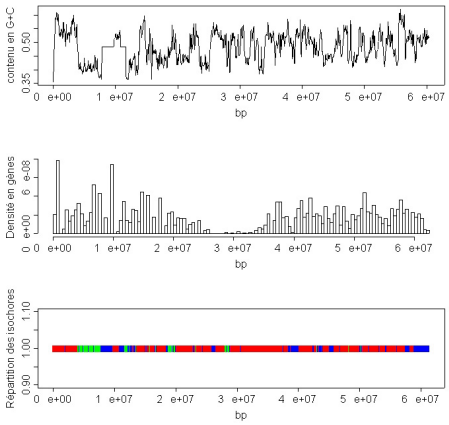
Chromosome 15 Chromosome 16



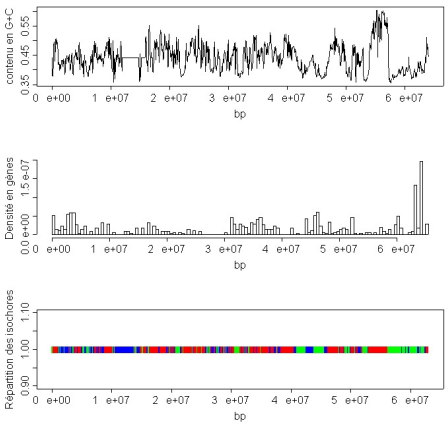
Chromosome 17 Chromosome 18



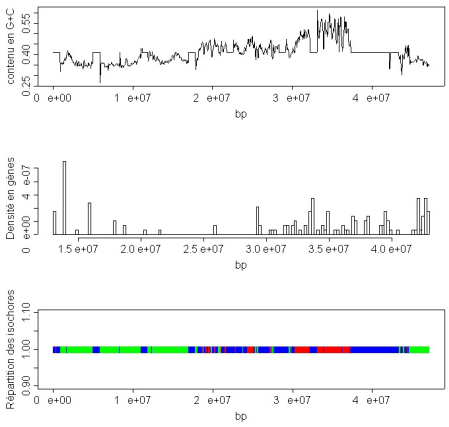
Chromosome 19



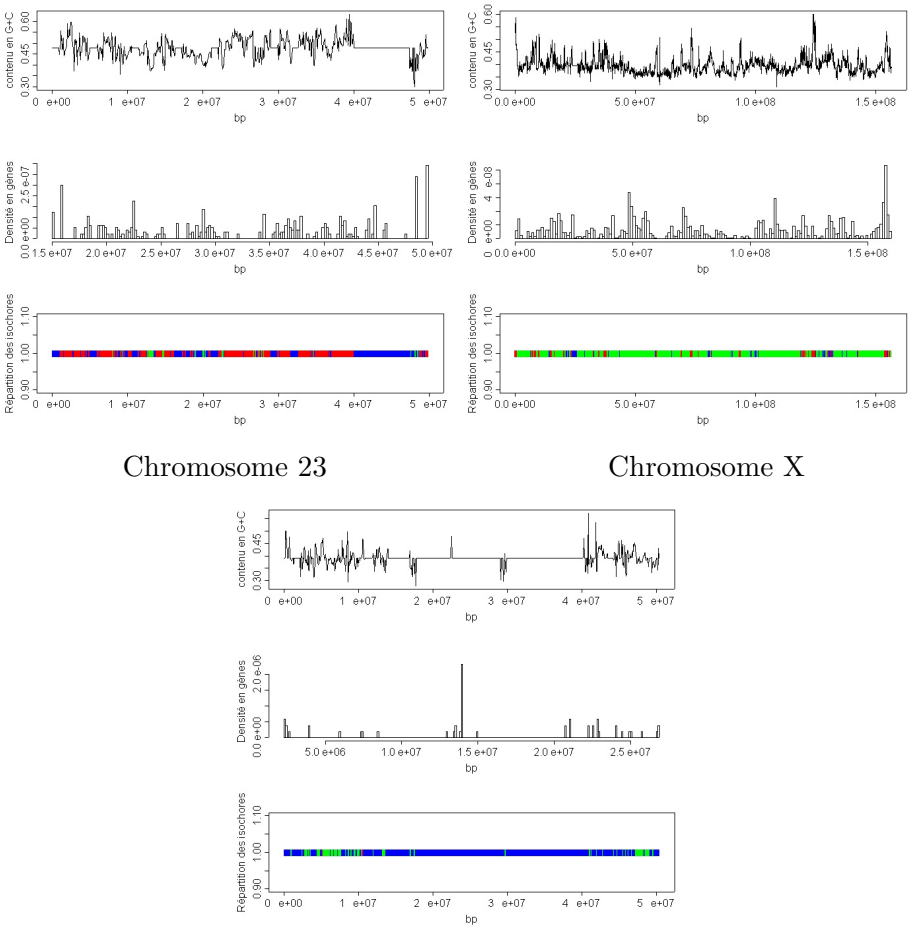
Chromosome 20



Chromosome 21



Chromosome 22



Chromosome 23 Chromosome X

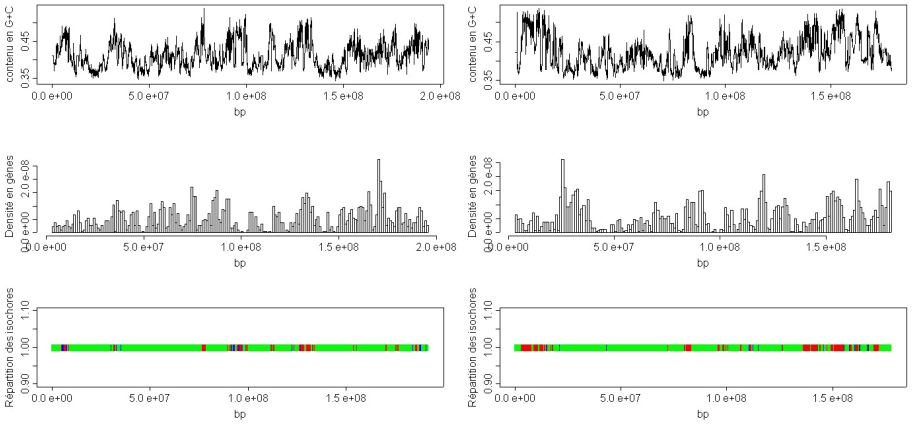
Chromosome Y

Fig. 5.5 : *L'évolution du taux en $G+C$, la répartition des gènes et la répartition des isochores prédite par notre modèle sont représentés le long de chaque chromosome du chimpanzée. En rouge, sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.*

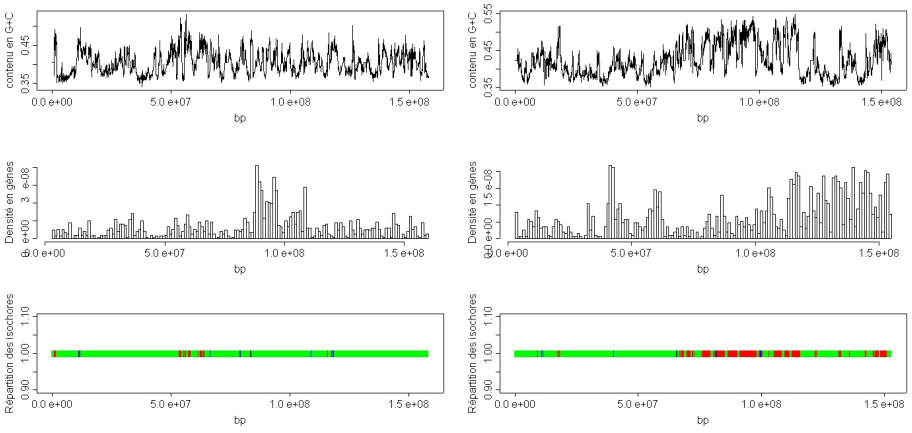
les deux autres espèces de mammifères étudiées précédemment. En effet, sur la grande majorité des chromosomes s'alternent les régions prédites en H et en L. Ainsi, les chromosomes constitués majoritairement d'un seul type d'isochores sont rares. Seul le chromosome 11 est principalement composé d'isochores H. Comme chez l'homme, la distribution des isochores prédits le long des chromosomes sexuels est similaire, les chromosomes X et Y étant très riches en régions L.

5.4.4.3 Poulet

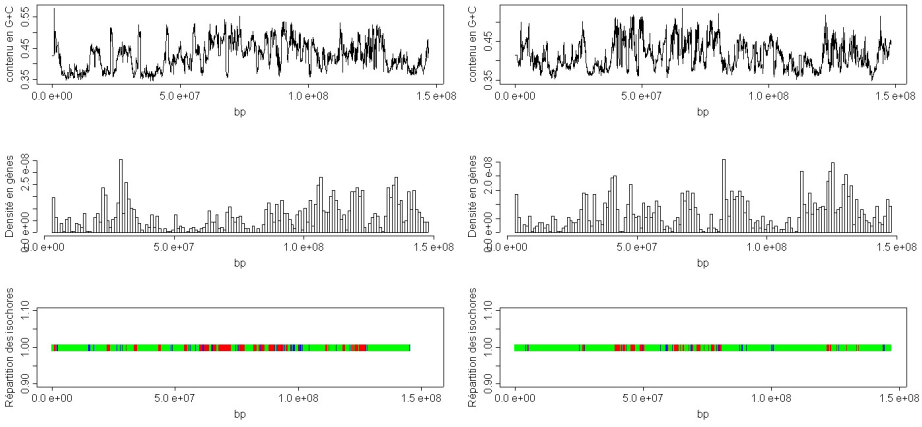
Seule l'analyse des chromosomes entièrement assemblés lors de la mise à jour d'octobre 2004 de la banque Ensembl est présentée. Les chromosomes du poulet sont généralement plus petits et plus nombreux que dans les génomes du chimpanzé, de l'homme et de la souris. Ces cartes d'isochores mettent en évidence une mosaïque de régions le long du génome du poulet, alternant entre des zones relativement riches en $G + C$ et d'autres plus pauvres en $G + C$. Les zones riches en $G + C$ sont majoritairement représentées sur les différents chromosomes et particulièrement le long des plus petits. C'est le cas par exemple pour les chromosomes 17, 19 et 23 à 32. Ces zones correspondent fréquemment à des régions identifiées comme appartenant à un isochore H par nos modèles. Le long des plus grands chromosomes, les isochores prédits en H alternent avec des isochores M et L, par exemple le long des chromosomes 1, 2, 3, 5 ou 6. Il est intéressant de constater que, bien que ce génome soit assez homogène et riche en $G + C$, de nombreuses zones identifiées comme étant des isochores L peuvent être observées et qu'elles recouvrent la majorité de la surface de certains chromosomes, notamment les chromosomes 11, W et X. Tout comme chez le *tetraodon nigroviridis* qui possède un petit génome compact et riche en $G + C$, les régions prédites en H sont plus longues que les régions prédites en L à l'inverse de ce qui est obtenu chez les trois autres mammifères étudiés.



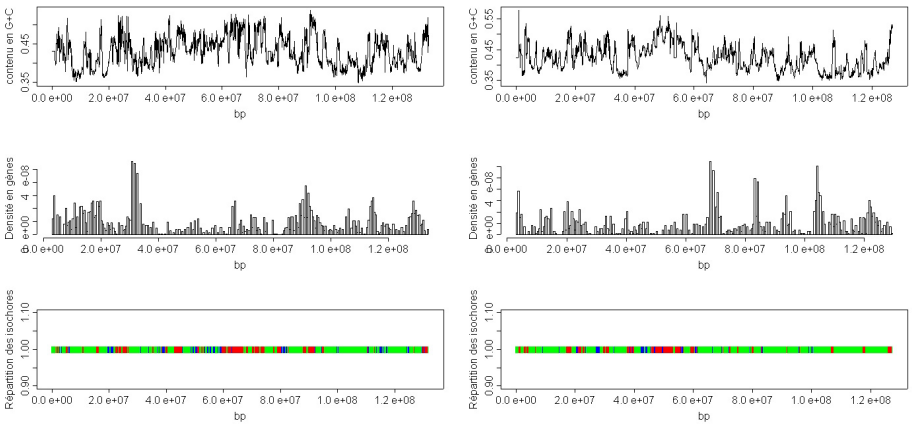
Chromosome 1 Chromosome 2



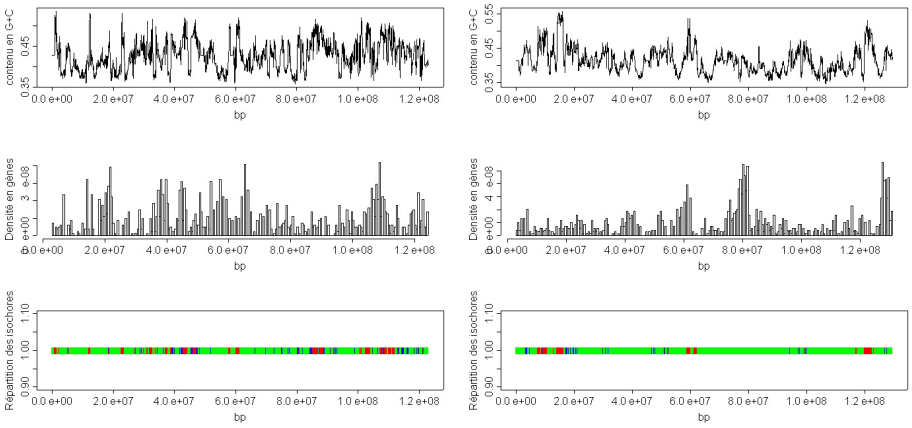
Chromosome 3 Chromosome 4



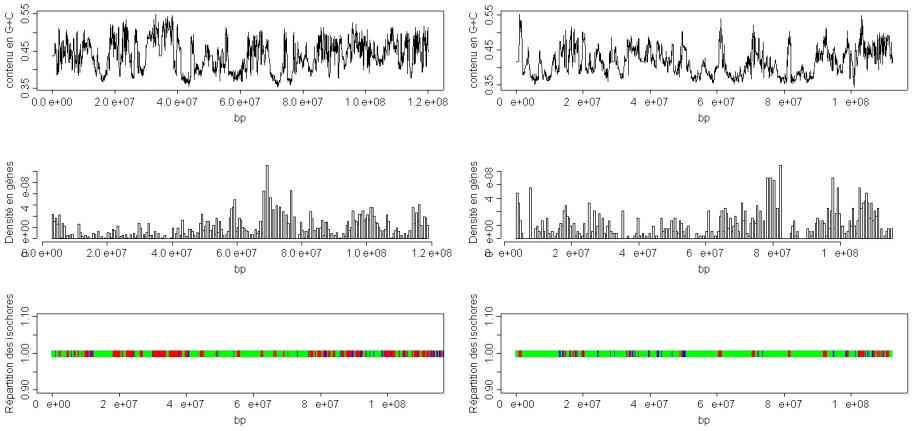
Chromosome 5 Chromosome 6



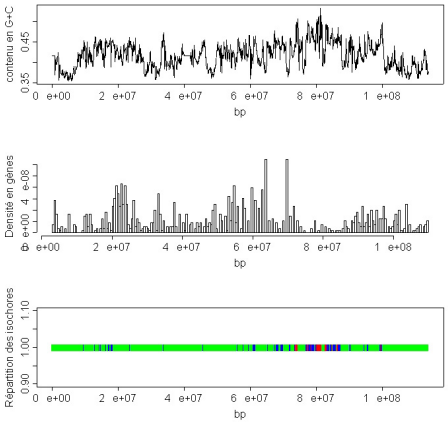
Chromosome 7 Chromosome 8



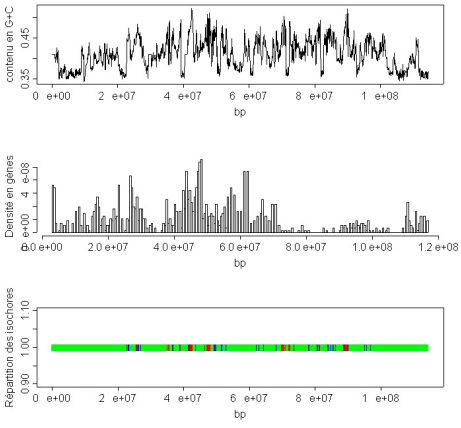
Chromosome 9 Chromosome 10



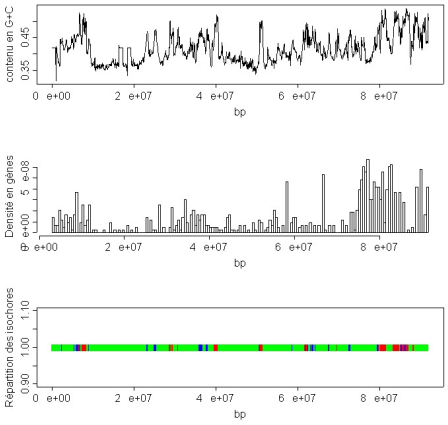
Chromosome 11 Chromosome 12



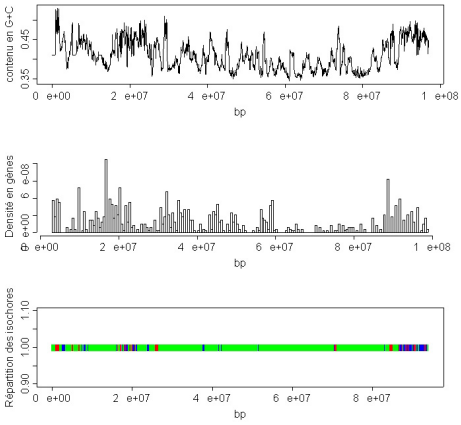
Chromosome 13



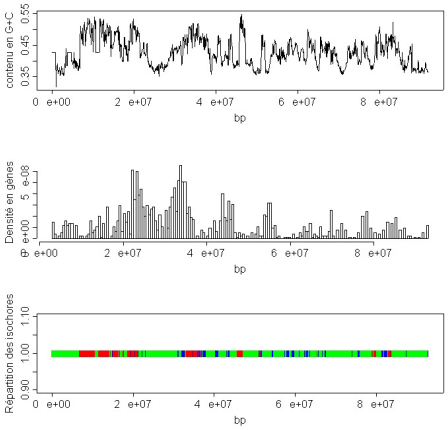
Chromosome 14



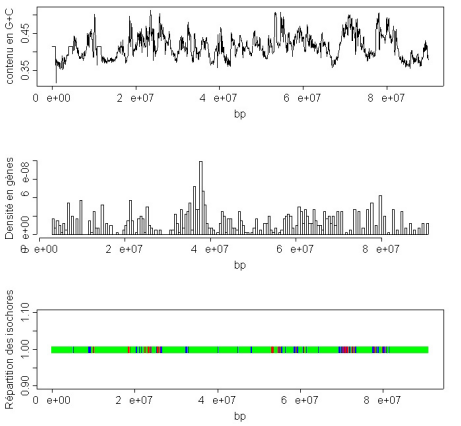
Chromosome 15



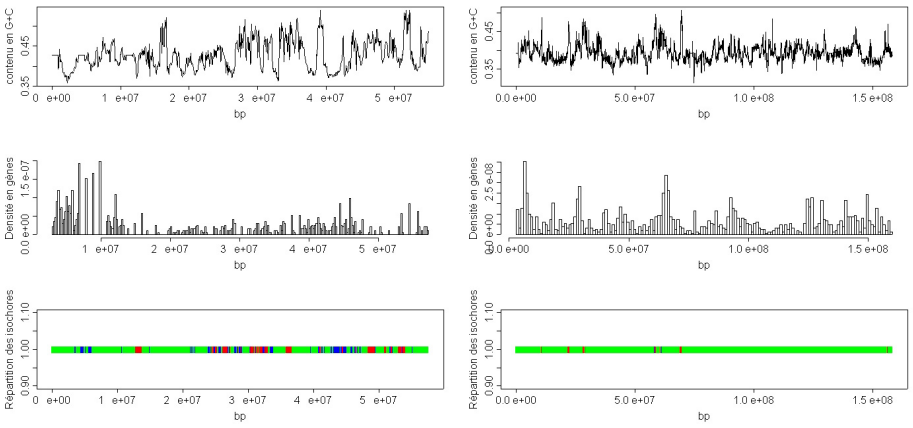
Chromosome 16



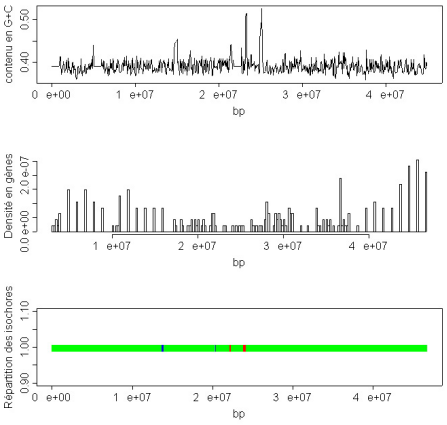
Chromosome 17



Chromosome 18

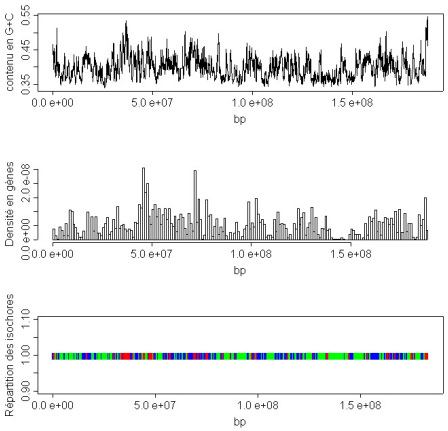


Chromosome 19 Chromosome X

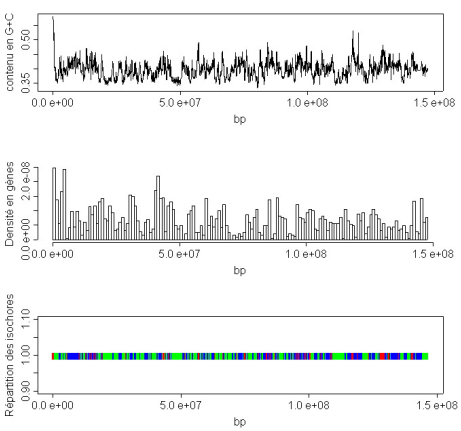


Chromosome Y

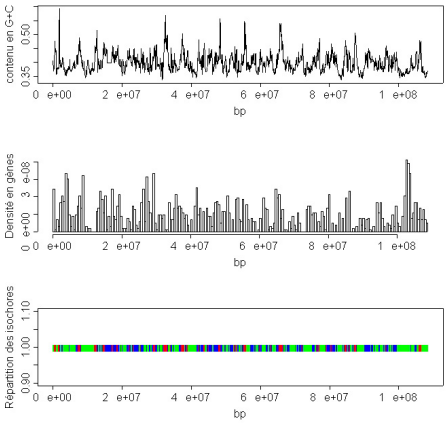
Fig. 5.6 : *L'évolution du taux en $G+C$, la répartition des gènes et la répartition des isochores prédite par notre modèle sont représentés le long de chaque chromosome de la souris. En rouge, sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.*



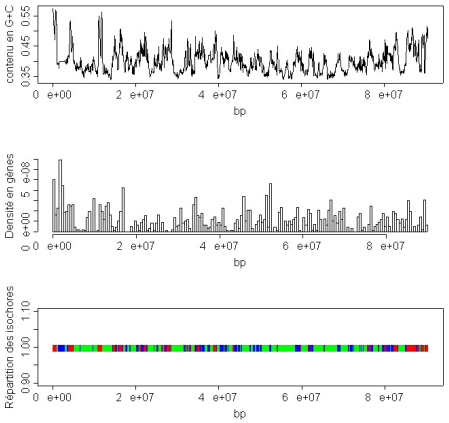
Chromosome 1



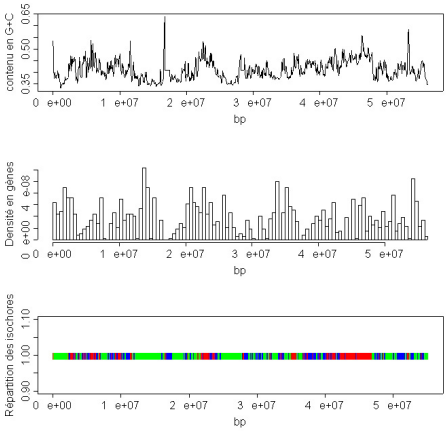
Chromosome 2



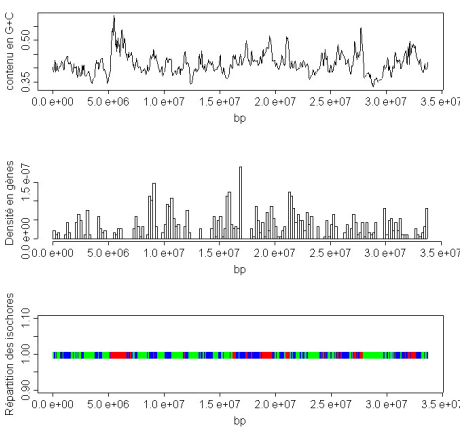
Chromosome 3



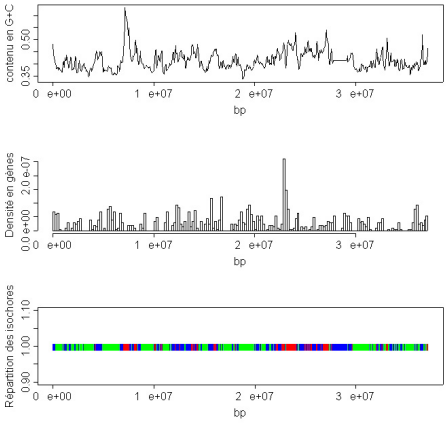
Chromosome 4



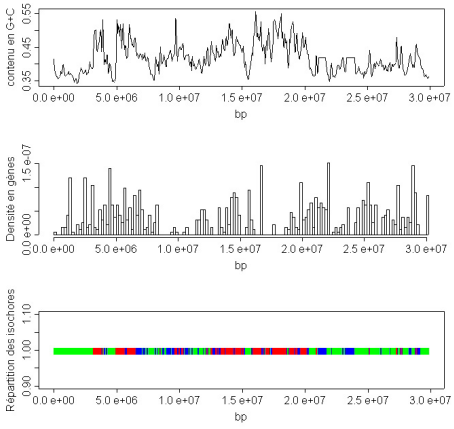
Chromosome 5



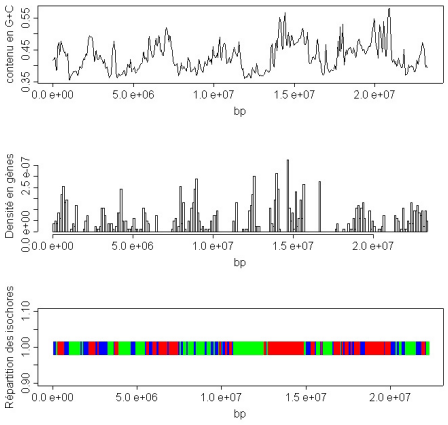
Chromosome 6



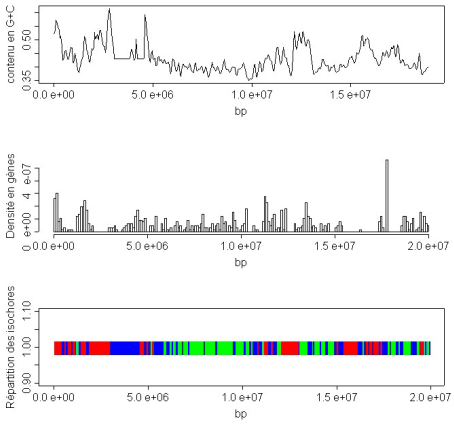
Chromosome 7



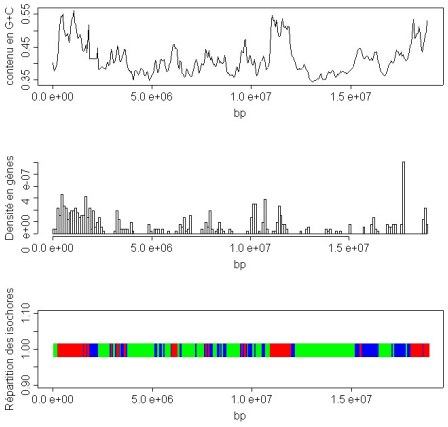
Chromosome 8



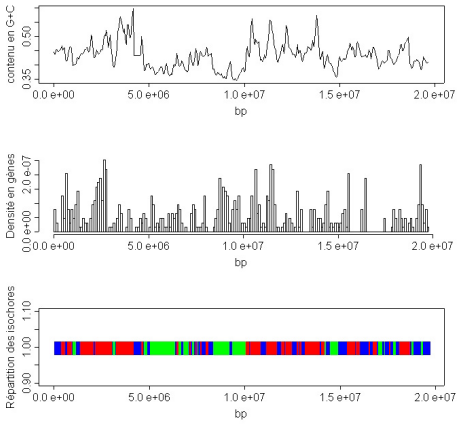
Chromosome 9



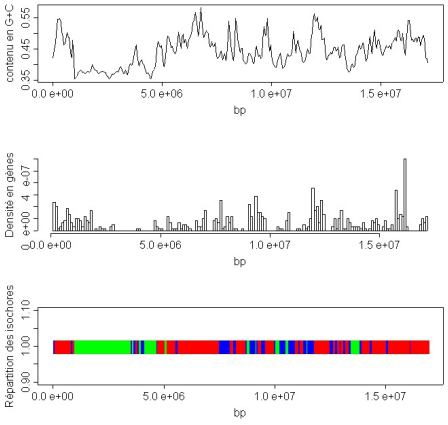
Chromosome 10



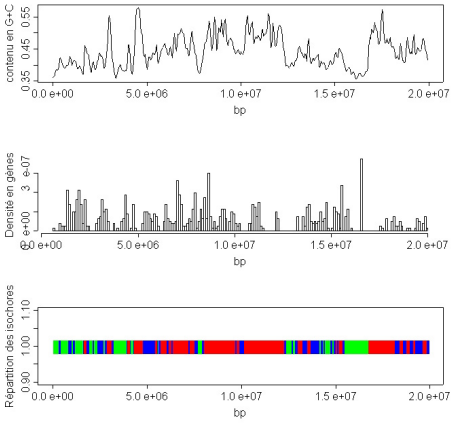
Chromosome 11



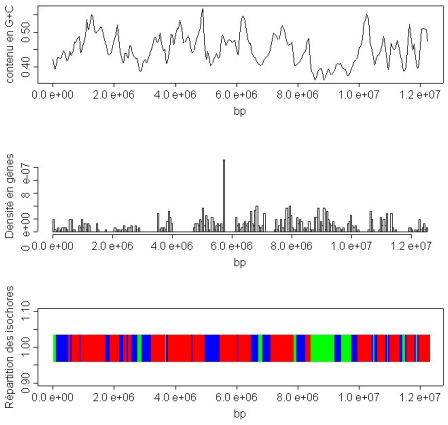
Chromosome 12



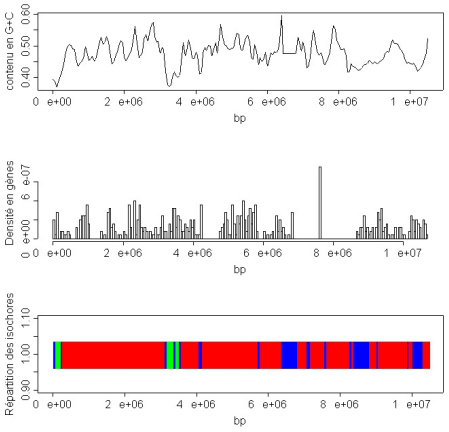
Chromosome 13



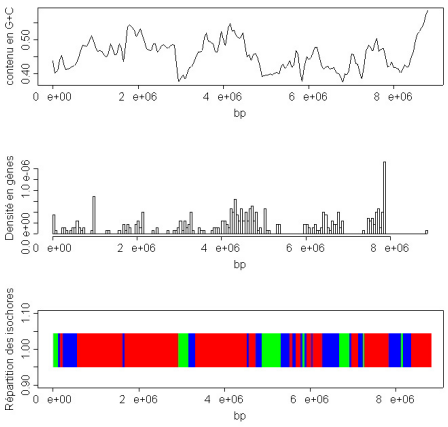
Chromosome 14



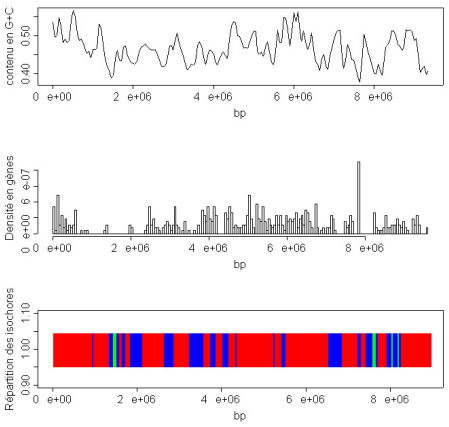
Chromosome 15



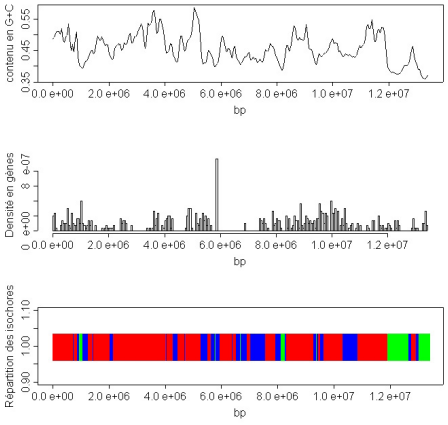
Chromosome 17



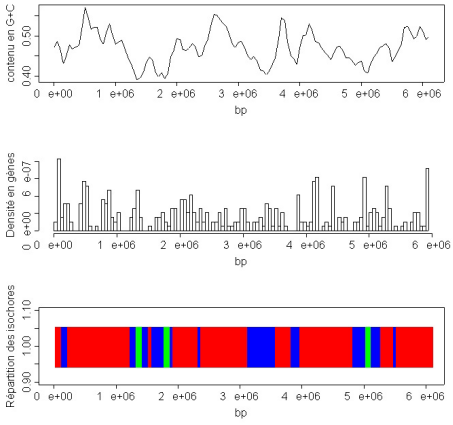
Chromosome 18



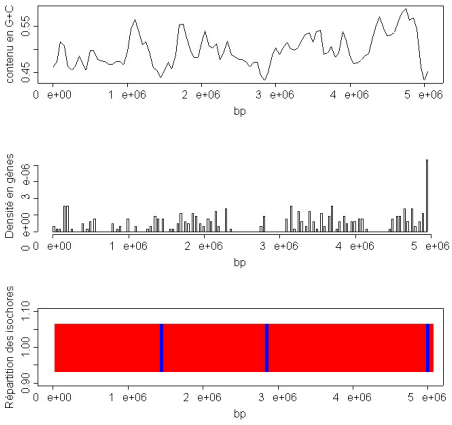
Chromosome 19



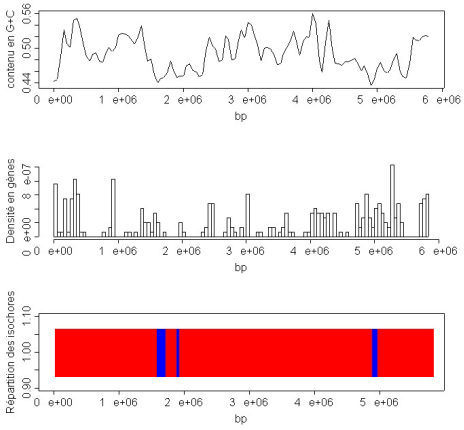
Chromosome 20



Chromosome 21



Chromosome 23



Chromosome 24

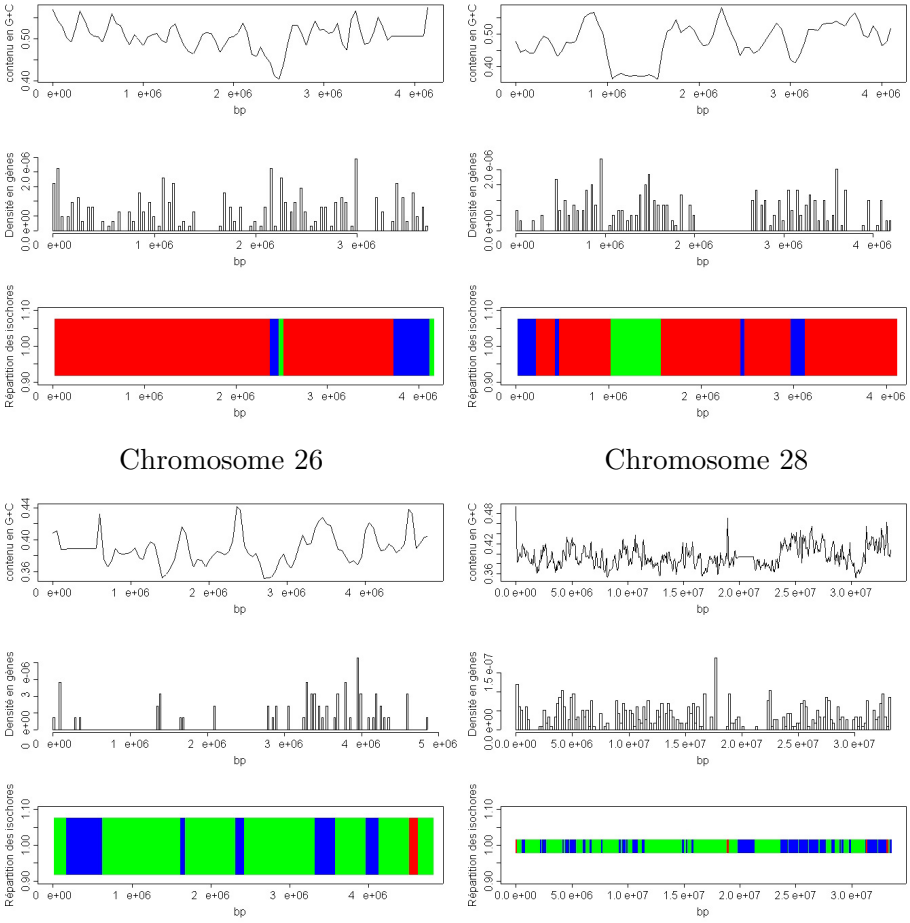


Fig. 5.7 : *L'évolution du taux en $G+C$, la répartition des gènes et la répartition des isochores prédite par notre modèle sont représentés le long de chaque chromosome du poulet. En rouge, sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.*

5.5 Discussion

L'évolution des isochores chez les vertébrés est à l'origine de nombreuses études, et ceci dès 1976, lorsque Macaya, au moyen d'analyses par centrifugation en gradient de densité, a mis en évidence que les génomes des mammifères et des oiseaux étaient fortement hétérogènes. Cependant, jusqu'à présent, aucune étude de comparaison à grande échelle de la structure en isochores entre les génomes de vertébrés entièrement séquencés n'a été réalisée. Ce chapitre a donc permis de fournir des cartes détaillées de la structure en isochores le long des génomes du chimpanzé, de la souris et du poulet afin de permettre une analyse ultérieure de la compréhension et de l'évolution de l'organisation en isochores de ces génomes.

L'étude des paramètres utilisés par nos modèles de Markov cachés pour chaque espèce a permis de généraliser au génome humain, du chimpanzé, de la souris et du poulet de nombreux résultats attendus d'après la littérature. Cela concerne par exemple, des différences de compacité entre les gènes des régions riches en $G+C$ et pauvres en $G+C$, un nombre plus élevé d'exons dans les régions faibles en $G+C$. Une relation linéaire entre la composition en $G+C_3$ des gènes et le contenu en $G+C$ des introns qui les composent peut être observée. Les analyses factorielles des correspondances à partir des mots de six lettres qui constituent les différentes régions des gènes suivant leur contenu en $G+C_3$ ont montré une structuration compositionnelle commune aux quatre espèces. De même, les analyses factorielles des correspondances ont permis de retracer simplement l'évolution des organismes au cours du temps et de retrouver les liens de parentés entre espèces classiquement établis. Ainsi, l'homme est fortement apparenté au chimpanzé, ces deux espèces sont plus éloignées de la souris et encore plus du poulet. De plus, lors des chapitres 3 et 4, il a été vérifié que le $G+C_3$ est un bon marqueur de la structuration en isochores. La comparaison du contenu en $G+C_3$ des gènes orthologues de différentes espèces a mis en évidence une organisation compositionnelle liée à la notion d'isochores chez les différents génomes.

Les cartes obtenues par notre méthode ont permis de situer les isochores de manière plus précise que par une simple étude du $G+C_3$ des gènes orthologues. De plus, elles montrent une relation entre l'organisation en isochores, la répartition et la structure des gènes chez les mammifères et l'oiseau étudiés. Une comparaison succincte a mis en évidence trois caractéristiques

principales.
Premièrement, les chromosomes sexuels sont très majoritairement composés de régions L dans les quatre espèces.
Deuxièmement, la segmentation en isochores obtenue par notre méthode chez la souris met en évidence une homogénéisation de son génome par rapport à celui de l'homme et du chimpanzé. Ces résultats confirment des études ultérieurs. Si en 1999, Hugues a montré qu'au sein des mammifères, la structure en isochores est globalement très conservée. Certains changements significatifs peuvent toutefois être remarqués, notamment chez les rongeurs où une diminution de l'hétérogénéité compositionnelle par rapport à la plupart des autres mammifères est observée. Les gènes riches en $G+C$ chez la souris sont moins riches que chez les autres mammifères (et inversement pour les gènes pauvres en $G+C$) (Mouchiroud et Gautier 1988, Mouchiroud et Gautier 1990, Galtier et Mouchiroud 1998). Cette modification de la structure se traduit donc par une homogénéisation globale de la composition en bases $G+C$, par rapport à la structure ancestrale observée chez l'homme et les autres mammifères (Galtier et Mouchiroud 1998, Belle *et al.* 2004) : moins de régions pauvres en $G+C$ et moins de régions riches en $G+C$.
Troisièmement, il est intéressant de noter que plus les chromosomes sont petits plus ils ont tendance à être constitués de régions prédites en H. Une interprétation biologique possible à ce phénomène d'attraction des zones riches en $G+C$ le long des petits chromosomes pourrait être fournie par une étude récente qui suggère qu'il existe un pré-requis d'un événement de crossing-over par bras chromosomique et par méiose (Pardo-Manuel de Villena et Sapienza 2001). En effet, Meunier (2005) a montré au cours de sa thèse qu'il existe une corrélation négative entre le taux de crossing-over et la longueur des bras chromosomiques humain. Tandis que leur longueur augmente, le taux de crossing-over converge vers une valeur basale. Ainsi, cette corrélation n'est valable que pour les bras chromosomiques courts. Le caryotype de la souris comprend 20 bras chromosomiques (tous ses chromosomes sont acrocentriques), dont la longueur fluctue entre 58 et 192 mégabases. Les bras chromosomiques de la souris sont donc tous relativement longs et il y a moins de variabilité inter-chromosomique dans les taux de crossing-over chez la souris que chez l'homme (Meunier 2005). Ainsi, le taux de recombinaison relativement homogène pourrait très bien expliquer l'homogénéisation observée de la composition en bases chez les rongeurs (voir carte d'isochores de

la souris). De plus, le génome du poulet, qui présente un caryotype très hétérogène (9 macro-chromosomes et 30 micro-chromosomes) et donc une très forte variabilité inter-chromosomique de taux de crossing-over (Burt 2002), est aussi caractérisé par une hétérogénéité extrême dans sa composition en bases $G+C$ (exemple : petits chromosomes du poulet). Ainsi, comme le suggère Meunier (2005), l'évolution du caryotype (nombre et longueur des bras chromosomiqes) est probablement un déterminant important de l'évolution de la composition en bases G et C à l'échelle du génome.

Cette étude pourrait permettre d'améliorer notre compréhension de la mise en place des isochores entre la lignée des mammifères et celle des oiseaux et les liens éventuels qui existeraient avec la lignée des poissons. En effet, au chapitre 4, il a été vérifié l'existence d'un lien entre l'organisation en isochores chez l'homme et les propriétés qui lui sont liés chez les poissons. Il a été possible d'obtenir une segmentation du génome du *Tetraodon nigroviridis*. Ainsi, cette étude concernant l'évolution des isochores peut conduire à une meilleure connaissance de leur influence sur le fonctionnement des génomes.

Conclusion

L'objectif de ce travail de thèse était de développer une approche markovienne pour l'analyse de l'organisation en isochores le long des génomes afin d'aider à la compréhension de l'évolution des génomes. Dans un premier temps, une étude bibliographique a permis de recenser et de comparer les différentes méthodes existantes. Les modèles de Markov en général sont généralement reconnus comme étant l'une des techniques les plus simples et les plus efficaces lors de la prédiction de gènes ou la modélisation des génomes. Parmi les différentes approches possibles, les chaînes de Markov cachées ont été préférées. En effet, le nombre de paramètres est significativement plus important dans le cas des modèles semi-markoviens à cause de l'utilisation des distributions empiriques des longueurs. D'autre part, l'utilisation d'algorithmes classiques comme Baum-Welch ou Viterbi est rendue plus complexe pour des modèles semi-markoviens cachés que pour les HMMs et nécessite des optimisations. Ainsi, par exemple, la complexité des principaux algorithmes utilisés par les modèles semi-markoviens cachés (Forward-Backward et Viterbi) peut être quadratique par rapport à la longueur de la séquence. Ce facteur est limitant, notamment lorsque l'analyse des données porte sur des génomes entiers. Un des avantages d'utilisation des chaînes de Markov cachées est la complexité des algorithmes qui est alors linéaire.

Dans un premier temps, des HMMs, intégrant aux mieux les propriétés biologiques les plus influentes (composition en bases, structure en codon des exons, longueurs des exons et introns...), ont été développés afin de permettre une exploration à grande échelle des génomes. Pour cela, une approche reposant sur une classification de modèles HMMs gérant les longueurs des états par la construction d'états complexes a été mise en œuvre pour permettre la discrimination de certaines régions d'un génome. Il est intéressant de constater que l'utilisation de l'algorithme de Viterbi, que ce soit pour la sélection de modèles ou pour la reconstruction de chemin caché, ne permet

pas de prendre en compte l'information biologique supplémentaire apportée par les macro-états. La solution proposée a consisté à adapter l'algorithme de Forward-Backward.

La deuxième étude, utilisant les modèles mis en place précédemment, a été menée dans le but de détecter les isochores le long du génome humain car de nombreuses propriétés biologiques des génomes de mammifères sont liées à leur structure en isochores (la densité en gènes, la taille des gènes, les distributions des éléments transposables, le taux de recombinaison...). Les modèles développés durant ce travail de thèse prennent en compte ces propriétés biologiques et permettent une nette segmentation des chromosomes en isochores. Notre approche se distingue des méthodes existantes sur deux aspects. Dans un premier temps, elle permet une meilleur localisation des isochores que les méthodes classiques car elle combine différentes informations biologiques. Le deuxième aspect est la possibilité de réaliser une analyse complémentaires de la structure des génomes. Ainsi, une étude basée sur une sélection de modèles de Markov a montré que certaines propriétés biologiques intervenaient pour classer un gène dans un isochore indépendamment du contenu en $G+C$ de la séquence. Il est donc possible de déterminer des ruptures d'homogénéité le long des séquences et d'identifier leurs causes biologiques. Le résultat majeur mis en relief dans cette étude concerne les régions UTRs. La variabilité du contenu $G+C_3$ des gènes appartenant aux isochores dont le taux en $G+C$ est fort est moins marquée dans les régions UTRs. Il apparaît donc que les régions UTRs possèdent une structure particulière qui dépend du type d'isochore dans lequel elles se situent. Ceci souligne le fait que le contenu en $G+C_3$ des gènes n'est pas la seule propriété biologique qui résulte de la structure des isochores. De telles différences dans les régions UTRs peuvent résulter, soit de mutations spécifiques qui s'accumulent plus rapidement dans les régions UTRs car la contrainte fonctionnelle est moins importante dans ces régions (hypothèse neutraliste), soit d'une contrainte de sélection exercée sur les régions UTRs, probablement associée à l'expression des gènes (hypothèse adaptative). À l'heure actuelle, il est impossible de trancher entre ces deux hypothèses.

Lors d'une troisième étude, notre méthode de prédiction a été appliquée à l'analyse de l'organisation des génomes du *Tetraodon nigroviridis* et du fugu. Les génomes des amphibiens et des poissons sont relativement homogènes et en général dépourvus d'isochores très riches en $G+C$. L'hypothèse

actuelle tend à dire qu'il n'existe pas d'isochore chez les poissons. Pourtant, notre étude a clairement mis en évidence l'existence d'une structure en mosaïques nouvelle le long du génome du *Tetraodon*. Ces résultats ont été obtenus par une approche originale reposant sur l'hypothèse suivante : les caractéristiques des gènes contenus dans les isochores du génome humain pourraient se retrouver au sein des gènes orthologues des espèces supposées dépourvues d'isochores. Cette organisation est liée aux différences de structures entre les gènes du *Tetraodon* orthologues aux gènes humains selon leur classe d'isochores. Notre étude démontre de façon claire qu'il est possible de retrouver, chez une espèce, à partir de gènes orthologues, des traces d'isochores qui ne sont pas repérables à partir d'une définition classique. Ces gènes orthologues conservent des caractéristiques liées aux isochores. Cette étude montre également que les limites du contenu en $G + C$ des classes d'isochores ne sont pas fixes, mais varient d'une espèce à l'autre suivant la teneur et l'homogénéité en $G + C$ du génome étudié. Notre méthode permet donc l'identification des isochores mais aussi l'interprétation et la compréhension biologiques des résultats obtenus au moyen d'un retour aux données et d'une analyse des échecs et des succès des modèles.

L'évolution des isochores chez les vertébrés est à l'origine de nombreuses études. Cependant, jusqu'à présent, aucune étude de comparaison à grande échelle de la structure en isochores entre les génomes de vertébrés entièrement séquencés n'a été réalisée. Les données concernant de nombreux génomes entièrement séquencés étant à présent disponibles, les ressources informatiques suffisantes grâce aux centres de calculs, il a donc été possible au cours du chapitre cinq de fournir des cartes détaillées de la structure en isochore le long des génomes du chimpanzé, de la souris et du poulet. Ainsi une analyse ultérieure de la compréhension et de l'évolution de l'organisation en isochores de ces génomes est envisageable. Les cartes obtenues ont permis de situer les isochores de manière plus précise que par une simple étude du $G+C_3$ des gènes orthologues. Elles mettent nettement en évidence une relation entre l'organisation en isochores, la répartition et la structure des gènes et le contenu en $G + C$ chez le chimpanzé, la souris et le poulet.

Le modèle actuellement le plus vraisemblable concernant l'origine des isochores met en avant l'effet des biais de conversion génique vers les bases G et C chez les mammifères (Bill *et al.* 1998, Galtier 2003, Kudla *et al.* 2004). La composition en $G + C$ des séquences soumises à la conversion gé-

nique aurait donc tendance à augmenter. Ce mécanisme pourrait expliquer la corrélation entre la composition en bases $G+C$ et le taux de recombinaison (Galtier *et al*. 2001). Suite à ces observations, Meunier (2005) a montré qu'il existe une corrélation négative entre le taux de crossing-over et la longueur des bras chromosomiques humain. Tandis que leur longueur augmente, le taux de crossing-over converge vers une valeur basale. Ainsi, cette corrélation n'est valable que pour les bras chromosomiques courts. Les données obtenues par notre méthode le long des chromosomes des différentes espèces étudiées vont dans ce sens. En effet, la souris apparaît sur nos cartes comme étant plus homogène que les génomes de l'homme et du chimpanzé. Ceci pourrait d'après l'hypothèse de Meunier s'expliquer par le fait que les bras chromosomiques de la souris sont tous relativement longs et il y a moins de variabilité inter-chromosomique dans les taux de crossing-over chez la souris que chez l'homme. Les résultats obtenus chez le poulet confortent le modèle proposé. En effet, le génome du poulet, qui présente un caryotype très hétérogène (9 macro-chromosomes et 30 micro-chromosomes) et donc une très forte variabilité inter-chromosomique de taux de crossing-over (Burt 2002), est aussi caractérisé par une hétérogénéité extrême dans sa composition en bases $G+C$. Ainsi, les petits chromosomes sont majoritairement représentés par les régions H alors que les grands chromosomes sont très hétérogènes. Ce phénomène se retrouve également chez l'homme et chez le chimpanzé où les régions H sont majoritairement représentées respectivement le long des chromosomes 19 et 22, et le long des chromosomes 18 à 21. Cette hypothèse s'applique également le long des chromosomes du *Tetraodon Nigroviridis*. En effet, celui-ci possède de très petits chromosomes et bien qu'une segmentation puisse être observée le long de ces chromosomes, celle-ci est majoritairement biaisées en faveur des régions H. Ainsi, l'évolution du caryotype (nombre et longueur des bras chromosomiqes) est probablement un déterminant important de l'évolution de la composition en bases G et C à l'échelle du génome.

Différents prolongements de l'utilisation des modèles HMMs mis en place pendant cette thèse peuvent donc être envisagés. D'un point de vue méthodologique, il serait intéressant de développer plus en profondeur le côté sélection de modèles sur lequel repose la méthodologie mise en place au cours de ce travail. Cela est d'autant plus important qu'aucune méthode vraiment satisfaisante n'est recensée : que ce soit les méthodes de mesures de distances

entre HMMs, le critère de sélection " bayésien ", le critère AIC... D'un point de vue biologique, l'analyse comparative des génomes ouvre un grand champ d'investigation : compréhension de leur organisation, de leur évolution et de leur fonctionnement. Ainsi, l'analyse comparative des séquences est une approche très efficace pour repérer les caractéristiques fonctionnelles du génome, même celles qui sont soumises à de faibles pressions de sélection. Elle s'avère donc être un outil très utile en complément de l'approche expérimentale. C'est pour cette raison que le séquençage des génomes entiers de différents organismes a été entrepris parallèlement au séquençage du génome humain. Comprendre le fonctionnement du génome nécessite d'identifier non seulement les gènes mais également toutes les autres régions qui sont soumises à une pression de sélection. L'analyse de ces différentes régions par notre méthode d'exploration des génomes mise en place au cours du second chapitre constituerait une voie de recherche intéressante. La mise en œuvre d'une telle méthode basée sur une sélection de modèles HMMs simples permettrait la mise en évidence de propriétés biologiques nouvelles grâce à un retour aux données. Par ailleurs, il faut noter que ce n'est pas nécessairement la séquence elle-même (*i.e.* l'enchaînement des nucléotides) qui est contrainte, certaines caractéristiques plus globales de l'organisation des chromosomes (propriété compositionnelle, configuration spatiale, répartition des gènes...) peuvent également être importantes pour le fonctionnement du génome. L'analyse comparative joue également un rôle capital lors de l'étude de l'évolution des génomes entre espèces plus ou moins proches. Pour toutes ces raisons, le développement de méthodes pour l'analyse à grande échelle des génomes semble incontournable. Ainsi, grâce aux données fournies au cours du cinquième chapitre, il serait intéressant d'étudier les isochores entre les différents génomes pour mieux comprendre leur mise en place entre la lignée des mammifères, celles des oiseaux et les liens éventuels qui existeraient avec la lignée des poissons. Cette étude concernant l'évolution des isochores est susceptible de conduire à une meilleure connaissance de leur influence sur le fonctionnement des génomes et doit permettre d'obtenir des renseignements complémentaires sur l'évolution de ces génomes, notamment au niveau des causes de l'amplification des différences de composition en $G + C$ entre les espèces. De plus, les modèles étant simples et efficaces, il serait facile d'intégrer les résultats des futurs génomes entiers séquencés. Si l'objectif premier de cette thèse était la mise en place de modèles HMMs pour la localisation

des isochores, la méthodologie développée peut aisément être appliquée à toute autre organisation structurelle des génomes dans le but d'une interprétation biologique. Ainsi, une idée sous-jacente consisterait à essayer au travers de l'analyse des structures statistiques des génomes d'associer à ces structures des processus agissant sur les génomes.

Bibliographie

Aïssani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C, Bernardi G. (1991) The compositional properties of human genes. *J. Mol. Evol.*, 32, 497-503.

Altschul S.F., Gish W, Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215-410.

Andrieu O, Fiston A.S., Anxolabehere D, Quesneville H. (2004) Detection of transposable elements by their compositional bias. *BMC Bioinformatics*, 5(1), 94.

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.

Audic S, Claverie J.M. (1997) Detection of eukaryotic promoters using Markov transition matrices. *Computers Cem.*, 21(4), 223-227.

Bachtrog D, Charlesworth B. (2002) Reduced adaptation of a non-recombining neo-Y chromosome. *Nature*, 416-6878, 323-326.

Barash S, Wang W, Shi Y. (2002) Human secretory signal peptide description by hidden Markov model and generation of a strong artificial signal peptide for secreted protein expression. *Biochem. Biophys. Res. Commun.*, 294(4), 835-842.

Baum L.E., Petrie T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37, 1554-1563.

Baum L.E., Petrie T, Soules G, Weiss N. (1970) A maximization technique occuring in statistical analysis of probabilistic functions in Markov chains. *The Annals of Mathematical Statistics*, 41, 164-171.

Belle E, Duret l, Galtier N, Eyre-Walker A. (2004) The decline of isochores in mammals : an assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.*, 58, 653-660.

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

Bernaola-Galvan P, Carpena P, Roman-Roldon R, Oliver JL. (2001) Mapping isochores by entropic segmentation of long genome sequences. *Proceedings of the Fifth Annual International Conference on Computational Biology*, Montreal, Canada, ACM Press, New York, 217-218.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, 228(4702), 953-958.

Bernardi G. (1995) The human genome : organization and evolutionary history. *Annu. Rev. Genet.*, 29, 445-476.

Bernardi G, Bernardi G. (1990) Compositional patterns in the nuclear genome of could-blooded vertebrates. *J. Mol. Evol.*, 31, 265-281.

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

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

Besemer J, Borodovsky M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Research*, 27(19), 3911-3920.

Bettecken T, Aissani B, Muller CR, Bernardi G. (1992) Compositional map-

ping of the human dystrophin-encoding gene. *Gene*, 122(2), 329-335.

Bill C.A., Duran W.A., Miselis N.R., Nickoloff J.A. (1998) Efficient repair of all types of single base mismatches in recombination intermediates in Chines Hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics*, 147, 1997-1999.

Blat Y, Protacio R.U., Hunter N, Kleckner L. (2002) Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell*, 111, 791-802.

Borodovsky M, McIninch J. (1993) Genmark : parallel gene recognition for both DNA strands. *Computers Chem.*, 17(2), 123-133.

Boys R.J., Henderson D.A., Wilkinson D.J. (2000) Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Appl. Stat-J. Roy. St.C*, 49, 269-285.

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

Burt D.W. (2002) Origin and evolution of avian microchromosomes. *Cytogenet. Genome Res.*, 96, 97-112.

Charlesworth B. (1994) Genetic recombination : patterns in the genome. *Curr. Biol.*, 4, 182-184.

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.

Churcill G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1), 79-94.

Churcill G.A. (1992) Hidden Markov chains and the analysis of genome structure. *Computer chem.*, 16, 107-115.

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.

Clay O, Bernardi G. (2001) The isochores in human chromosomes 21 and 22. *Biochem. Biophys. Res. Commun.*, 285(4), 855-856.

Cohen N, Dagan T, Stone L, Graur D. (2005) GC Composition of the Human Genome : In Search of Isochores. *Mol. Biol. Evol.*, 22(5), 1260-1272.

Cuny G, Soriano P, Macaya G, Bernardi G. (1981) The major components of the mouse and human genomes. Preparation, basic properties and compositional heterogeneity. *Eur. J. Biochem.*, 115(2), 227-233.

Dawid H, Meyr H. (1995) Real-time algorithms and VLSI architectures for soft output MAP convolutional decoding. *PIMRC'95*, 1, 193-197.

Delcher A.L., Harmon D, Kasif S, White O, Salzberg S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27(23), 4636-4641.

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical society*, 39, 1-38.

Depaulis F, Brazier L, Mousset S, Turbe A, Veuille M. (2000) Selective sweep near the in(2L)t inversion breakpoint in an African population of Drosophila melanogaster. *Genet. Res.*, 76, 149-158.

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. USA*, 93(3), 1298-1302.

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*. (1999) The DNA sequence of human chromosome 22. *Nature*, 402(6761), 489-495.

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

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

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.

Duret L, Hurst L.D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.*, 18(5), 757-762.

Eyre-Walker A, Hurst L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.*, 2(7), 549-555. Review.

Eyre-Walker A. (1993) Recombinaison and mammalian genome evolution. *Proc. R. soc. Lond. B-Biol. Sci.*, 25, 237-243.

Fichant G, Gautier C. (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. *CABIOS*, 3, 287-295.

Filipski J. (1987) Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *Febs. Lett.*, 217, 184-196.

Fullerton S.M, Bernardo Carvalho A, Clark A.G, (2001) Local rates of recombinaison are positivel correlated with GC content in the human genome. *Mol. Biol. Evol.*, 18, 1139-142.

Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, Ikemura T. (1995) A boundary of long-range G + C% mosaic domains in the human MHC locus : pseudoautosomal boundary-like sequence exists near the boundary. *Genomics*, 25(1), 184-191.

Gassiat E, Keribin C. (2000) The likelihood ratio test for the number of components in a mixture with observations of Markov regime. *ESAIM Prob. And Stat.*, 4, 25-52.

Galtier N, Mouchiroud D. (1998) Isochore evolution in mammals : a human-like ancestral structure. *Genetics*, 150(4), 1577-1584.

Galtier N, Piganeau G, Mouchiroud D, Duret L. (2001) GC-content evolution in mammals : the biased gene conversion hypothesis. *Genetics*, 159, 907-911.

Galtier N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18, 866-873.

Galtier N. (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.*, 19, 65-68.

Gerton J.L., Derisi J.L., Shroff M, Litchen M, Brown P.O, Petes T.D. (2000) Global mapping of meiotic recombinaison hotspots and coldspots in the yeast *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci.*, 97, 11383-11390.

Guédon Y. (2003) Estimating hidden semi-Markov chains from discrete sequences. *J. Comput. Stat.*, 12(3), 604-639.

Guéguen L. (2005) Sarment : Python modules for hmm analysis and partitioning of sequences. *Bioinformatics*, accepted with revision.

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.

Haring D, Kypr J. (2001) No isochores in the human chromosomes 21 and 22 ? *Biochem. Biophys. Res. Commun.*, 280(2), 567-573.

Hattori M, *et al.* (2000) The DNA sequence of human chromosome 21. *Nature*, 405(6784), 311-319.

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

Henderson J, Salzberg S, Fasman K.H. (1997) Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.*, 4(2), 127-141.

Holmquist G.P. (1992) Chromosome Bands, their chromatin flavor, and their functional features. *Am. J. Hum. Genet.*, 51, 17-37.

Hugues S, Zelus D, Mouchiroud D. (1999) Warm-blooded isochore structure in Nile crocodile and turtle. *Mol. Biol. Evol.*, 16, 151-157.

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

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), 123-127.

Jaillon O *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertabrate proto-karyotype. *Nature*, 431, 946-957.

Jeffreys A.J., Neumann R. (2002) Reciprocal crossover asymetry and meiotic drive in a human recombinaison hot spot. *Nat. Genet.*, 31, 267-271.

Kass R.E., Raftery A.E. (1995) Bayes factors. *J.AM.Stat.Assoc.*, 90, 773-795.

Khelifi A, Duret L, Mouchiroud D (2005) HOPPSIGEN : a database of

human and mouse processed pseudogenes. *Nucleic Acids Research*, 3 Database Issue : D59-66.

Kong A *et al*. (2002) A high-resolution recombinaison map of the human genome. *Nat. Genet.*, 31, 241-247.

Krogh A, Brown M, Mian I.S., Sjölander K, Haussler D. (1994) Hidden Markov models in computational Biology. Application to protein modelling. *J. Mol. Biol.*, 235, 1501-1531.

Krogh A, Mian IS, Haussler D. (1994) A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22(22), 4768-4778.

Krogh A. (2000) Using database matches with for HMMGene for automated gene detection in Drosophila. *Genome Res.*, 10(4), 523-528.

Krogh A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *In proceeding of the fifth international conference on intelligent systems for molecular biology*, 179-186.

Kuddla G, Helwak A, Lipinsk L. (2004) Gene conversion and GC-content evolution in mammalian Hsp70. *Mol. Biol. Evol.*, 21, 1438-1444.

Kulkarni V.G. (1997) Modeling and analysis of stochastic systems. Chapman and Hall, London.

Kulp D, Haussler D, Reese M.G. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 4, 134-142.

Lander E.S., Linton L.M., 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, 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. (2002) Are isochore sequences homogeneous ? *Gene*, 300(1-2), 129-139.

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

Lukashin V.A, Borodovsky M. (1998) Gene-Mark.hmm : new solutions for gene finding. *Nucleic Acids Research*, 26, 1107-1115.

Macaya G, Thiery J.P., Bernardi G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.*, 108(1), 237-254.

MacDonald I.L., Zucchini W. (1997) Hidden Markov and other models for discrete-valued times series. Chapman and Hall, London.

Marais G. (2003) Biaised gene conversion : implication for genome and sex evolution. *Trends Genet.*, 19, 330-338.

Mathe C, Dehais P, Pavy N, Rombauts S, Van Montagu M, Rouze P. (2000) Gene prediction and gene classes in Arabidopsis thaliana. *J. Biotechnol.*, 78(3), 293-9.

Meunier J. (2005) Variabilité intr-génomique des taux de substitutions chez les primates, Phd thesis, Université C. Bernard, Lyon 1.

Montoya-Burgos J.I., Boursot P, Galtier N. (2003) Recombination explains isochores in mammalian genomes. *Trends Genet.*, 19, 128-130.

Mouchioroud D, Gautier C. (1988) High codon-usage changes in mammalian genes. *Mol. Biol. Evol.*, 5(2), 192-194.

Mouchioroud D, Gautier C. (1990) Codon usage changes and sequence dissimilarity between human and rat. *J. Mol. Evol.*, 31(2), 81-91.

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. (1991) The distribution of genes in the human genome. *Gene*, 100, 181-187.

Muri F. (1997) Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN. PhD thesis, Université R. Descartes, Paris V.

Muri F. (1998) Modelling bacterial genomes using hidden Markov models. In Physica-Verlag, editor, *Compstat'98 Proc. in Comput. Stat.*, 98-100.

Needleman S.B., Wunsch C.D. (1970) A general method applicate to the search for similarities in the amino acid sequence two proteins. *J. Mol. Biol.*, 48, 443-453.

Nekrutenko A, Li WH. (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.*, 10(12), 1986-1995.

Nicolas P, Bize L, Muri F, Hoebeke M, Rodolphe F, Ehrlich S, Prum B, Bessières P. (2002). Mining bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research*, 30, 1418-1426.

Nicolas P, Tocquet A.S., Miel V and Muri F. (2004). A reversible jump Monte-Carlo Markov chain algorithm for bacterial promoter discovery.

Nielsen H, Krogh A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 6, 122-130.

Ohler U, Harbeck S, Niemann H, Noth E, Reese MG. (1999) Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics*, 15(5), 362-369.

Oliver J.L., Bernaola-Galvan P, Carpena P, Roman-Roldan R.(2001) Isochore chromosome maps of eukaryotic genomes. *Gene*, 276(1-2), 47-56.

Oliver J.L., 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), 117-127.

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

Pardo-Manuel de Villena F, Sapienza C. (2001) Female meiosis drives karyotypic evolution in mammals. *Genetics*, 159, 1179-1189.

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

Pavlicek A, Paces J, Clay O, Bernardi G. (2002) A compact view of isochores in the draft human genome sequence. *FEBS Lett.*, 511(1-3),165-169.

Pedersen A.G., Baldi P, Brunak S, Chauvin Y. (1996) Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 4, 182-191.

Perrin P, Bernardi G. (1987) Directional fixation of mutations in vertebrate evolution. *J. Mol. Evol.*, 26, 301-310.

Peshkin L, Gelfand MS. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatic*, 15(12), 980-986.

Petes T.D., Merker J.D. (2002) Context dependence of meiotic recombination hotspots in yeast. The relationship between recombination activity of a reporter construct and base composition. *Genetics*, 162, 2049-2052.

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

Rabiner L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 2.

Reese M.G., Kulp D, Tammana H, Haussler D. (2000) Genie–gene finding in Drosophila melanogaster. Genome Research, 10(4), 529-538.

Rogers J, Mahaney M.C. *et al*. (2000) A genetic linkage map of the baboon (papio hamadryas) genome based on human microsatellite polymorphisms. *Genomics*, 67, 237-247.

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

Saccone S, de Sario A, Wiegant J, Raap A.K., Della Valle G, Bernardi G. (1993) Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA*, 90, 11929-11933.

Salzberg S.L., Delcher A. L., Kasif S, White O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26, 544-548

Smit A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, 9, 657-663.

Soriano P, Meunier-Rotival M, Bernardi G. (1983) The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA*, 80, 1816-1820.

Stephens R, Horton R, Humphray S, Rowen L, Trowsdale J, Beck S. (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.*, 291(4), 789-799.

Stormo G.D. (2000) Gene-Finding approaches for eukaryotes. *Genome Research*, 10, 394-397.

The MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401, 921-993.

Thiery J.P., Macaya G, Bernardi G. (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.*, 108(1), 219-235.

True J.R., Mercer J.M, Laurie C.C. (1996) Differences in cross-over frequency and distribution among three sibling species of Drosophila. *Genetics*, 142, 507-523.

Viterbi A.J. (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory*, IT-13, 260-269.

Williams E.J., Hurst L.D. (2000) The proteins of linked genes evolve at similar rates. *Nature*, 407(6806), 900-903.

Woodfine K, Fiegler H, Beare DM, Collins J.E., McCann O.T., Young B.D., Debernardi S, Mott R, Dunham I, Carter N.P. (2004) Replication timing of the human genome. *Hum. Mol. Genet.*, 13(2), 191-202.

Wolfe K.H, Sharp P.M, Li W.H. (1989) Mutation rates differ among regions of mammalian genome. *Nature*, 337, 283-285.

Wu C.F. (1983) On the convergence properties of the EM algorithm. *Ann. Stat.*, 11, 95-103.

Zhang Z, Wood WI. (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, 19(2), 307-308.

Zhang C.T., Zhang R. (2003) An isochore map of the human genome based on the Z curve method. *Gene*, 317(2), 127-135.

Zhang C.T., Zhang R. (2004) Isochore structures in the mouse genome. *Genomics*, 83(3), 384-394.

Zhang R. and C.T. Zhang (2004) Isochore structures in the genome of the plant Arabidopsis thaliana. *J. Mol. Evol.* 59, 227-238.

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

Annexes

5.6 Estimations des distributions de longueurs des exons et introns chez les différentes espèces

Cette annexe présente les résultats des estimations des paramètres obtenus à partir de l'estimation des distributions de longueurs par la minimisation de la distance de Kolmogorov-Smirnov pour chacune des espèces étudiées au cours de cette thèse et pour chaque classe d'isochore.

Les notations abordées sont les suivantes : $G_n(D_1, ..., D_n)$ définie la distribution d'une somme de n variables de lois géométriques, chacune d'espérance D_i et de paramètres $p_i = 1/D_i$. Ainsi, l'espérance de $G_n(D_1, ..., D_n)$ est $D_1 + D_2 + \cdots + D_n$. Quand $D_i = D$ pour tout i, cette loi correspond à une loi binomiale négative de paramètres (n,$1/D$), qui sera noté $G_n(D)$. Enfin, $G_1(D)$ est une loi géométrique d'espérance D et de paramètres $p = 1/D$, qui sera noté $G(D)$.

Les résultats des estimations des paramètres des différentes lois obtenues en minimisant la distance de Kolmogorov-Smirnov pour les différents types d'exons et d'introns sont représentés dans les tableaux qui suivent où K-S est l'abréviation utilisée pour Kolmogorov-Smirnov

HOMME

Lois	Paramètres	Distance de K-S
$G_2(p)$	66,7	0.08775
$G_3(p)$	40	0.12161
$G(p_1, p_2)$	58,8-74,7	0.08621
$G(p_1, p_2, p_3)$	100-9,5-30,3	0.08611

Tab. 5.6 : *Humain : Exon initial classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	184,9	0,09752
$G_3(p)$	47,6	0,08009
$G_4(p)$	33,3	0.03220
$G_5(p)$	26,3	0.02242

Tab. 5.7 : *Humain : Exon interne classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	90,9	0.06592
$G_3(p)$	59,2	0.05569
$G(p_1, p_2)$	90,1-94,8	0.06991
$G(p_1, p_2, p_3)$	86,2-181,8-10	0.04029

Tab. 5.8 : *Humain : Exon terminal classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	623	0.17466
$G_3(p)$	416,6	0,19721
$G(p_1, p_2)$	1075,3-106,4	0,13527
$G(p_1, p_2, p_3)$	754,2-446,1-198,3	0,13510

Tab. 5.9 : *Humain : Gène sans intron classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	69,4	0.08692
$G_3(p)$	43,4	0.11834
$G_4(p)$	31,25	0,14672
$G(p_1, p_2)$	10,8-142,9	0.05251
$G(p_1, p_2, p_3)$	3,9-13,4-132,9	0.04706

Tab. 5.10 : *Humain : Exon initial classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	166,7	0,06297
$G_3(p)$	45,5	0,05932
$G_4(p)$	33,3	0.02531
$G_5(p)$	27,1	0.02845

Tab. 5.11 : *Humain : Exon interne classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	90,9	0.07090
$G_3(p)$	60,6	0.09421
$G-4$	43,5	0,12394
$G(p_1,p_2)$	162,2-32,3	0.04242
$G(p_1,p_2,p_3)$	3,9-63,2-120,5	0.04237

Tab. 5.12 : *Humain : Exon terminal classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	58,1	0.06435
$G_3(p)$	35,7	0.09157
$G_4(p)$	25,839	0.04008
$G(p_1,p_2)$	16,2- 111,1	0.04023
$G(p_1,p_2,p_3)$	2,6 - 113,2 - 113,2	0,04011

Tab. 5.13 : *Humain : Exon initial classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	166,7 0,53892	
$G_3(p)$	43,5 0,06161	
$G_4(p)$	32,3 0.02316	
$G_5(p)$	25,6 0.03425	

Tab. 5.14 : *Humain : Exon interne classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	94,3	0,09111
$G_3(p)$	61,1	0,11111
$G-4$	43,9	0,13355
$G(p_1,p_2)$	17,2 - 181,8	0,06566
$G(p_1,p_2,p_3)$	3,9 - 63,1 - 119,0	0,06313

Tab. 5.15 : *Humain : Exon terminal classe L*

CHIMPANZE

Lois	Paramètres	Distance de K-S
$G_2(p)$	79,4	0,05761
$G_3(p)$	50,8	0,08783
$G_4(p)$	44,1	0,11243
$G_5(p)$	27,9	0,13299
$G(p_1, p_2)$	52,6 - 106,4	0,05023

Tab. 5.16 : *Chimpanzé : Exon initial classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	70,7	0,04458
$G_3(p)$	46,5	0,06281
$G_4(p)$	32,7	0,08581
$G_5(p)$	25,1	0,10506
$G(p_1, p_2)$	71,4 - 69,9	0,04457

Tab. 5.17 : *Chimpanzé : Exon initial classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	67,0	0,04011
$G_3(p)$	44,6	0,05339
$G_4(p)$	31,3	0,07321
$G_5(p)$	24,1	0,09453
$G(p_1, p_2)$	67,4 - 66,6	0,04007

Tab. 5.18 : *Chimpanzé : Exon initial classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	69,7	0,09315
$G_3(p)$	44,2	0,04409
$G_4(p)$	31,4	0,03794
$G_5(p)$	24,7	0,05123

Tab. 5.19 : *Chimpanzé : Exon interne classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	70,7	0,08895
$G_3(p)$	46,5	0,04009
$G_4(p)$	32,7	0,03399
$G_5(p)$	25,1	0,04263

Tab. 5.20 : *Chimpanzé : Exon interne classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	69,3	0,08992
$G_3(p)$	44,6	0,04193
$G_4(p)$	43,7	0,03571
$G_5(p)$	24,5	0,04417

Tab. 5.21 : *Chimpanzé : Exon interne classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	84,2	0,05053
$G_3(p)$	50,8	0,06631
$G_4(p)$	54,5	0,08764
$G_5(p)$	39,4	0,1095
$G(p_1, p_2)$	58,8 - 108,7	0,04577

Tab. 5.22 : *Chimpanzé : Exon terminal classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	81,4	0,05574
$G_3(p)$	53,8	0,07375
$G_4(p)$	29,0	0,11009
$G(p_1, p_2)$	58,8 - 103,1	0,05289

Tab. 5.23 : *Chimpanzé : Exon terminal classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	79,7	0,05903
$G_3(p)$	52,7	0,07396
$G_4(p)$	36,7	0,08971
$G_5(p)$	27,8	0,10718
$G(p_1, p_2)$	93,2 - 66,2	0,05768

Tab. 5.24 : *Chimpanzé : Exon terminal classe L*

Lois	Paramètres	Distance de K-S
$G(p)$	1923,1	0,08420
$G_2(p)$	740,7	0,16956
$G_3(p)$	450,5	0,22118
$G_4(p)$	327,9	0,24122
$G_5(p)$	256,4	0,26227
$G(p_1, p_2)$	1428,6 - 1111,1	0,13673

Tab. 5.25 : *Chimpanzé : Intron classe H*

Lois	Paramètres	Distance de K-S
$G(p)$	2564,1	0,07751
$G_2(p)$	1000	0,16054
$G_3(p)$	625	0,21012
$G_4(p)$	434,9	0,23737
$G_5(p)$	340,1	0,25736
$G(p_1, p_2)$	2000 - 33,3	0,12259

Tab. 5.26 : *Chimpanzé : Intron classe M*

Lois	Paramètres	Distance de K-S
$G(p)$	2631	0,07342
$G_2(p)$	1111	0,16481
$G_3(p)$	662	0,20565
$G_4(p)$	476	0,25505
$G_5(p)$	362	0,25889
$G(p_1, p_2)$	1666 - 1204	0,01543

Tab. 5.27 : *Chimpanzé : Intron classe M*

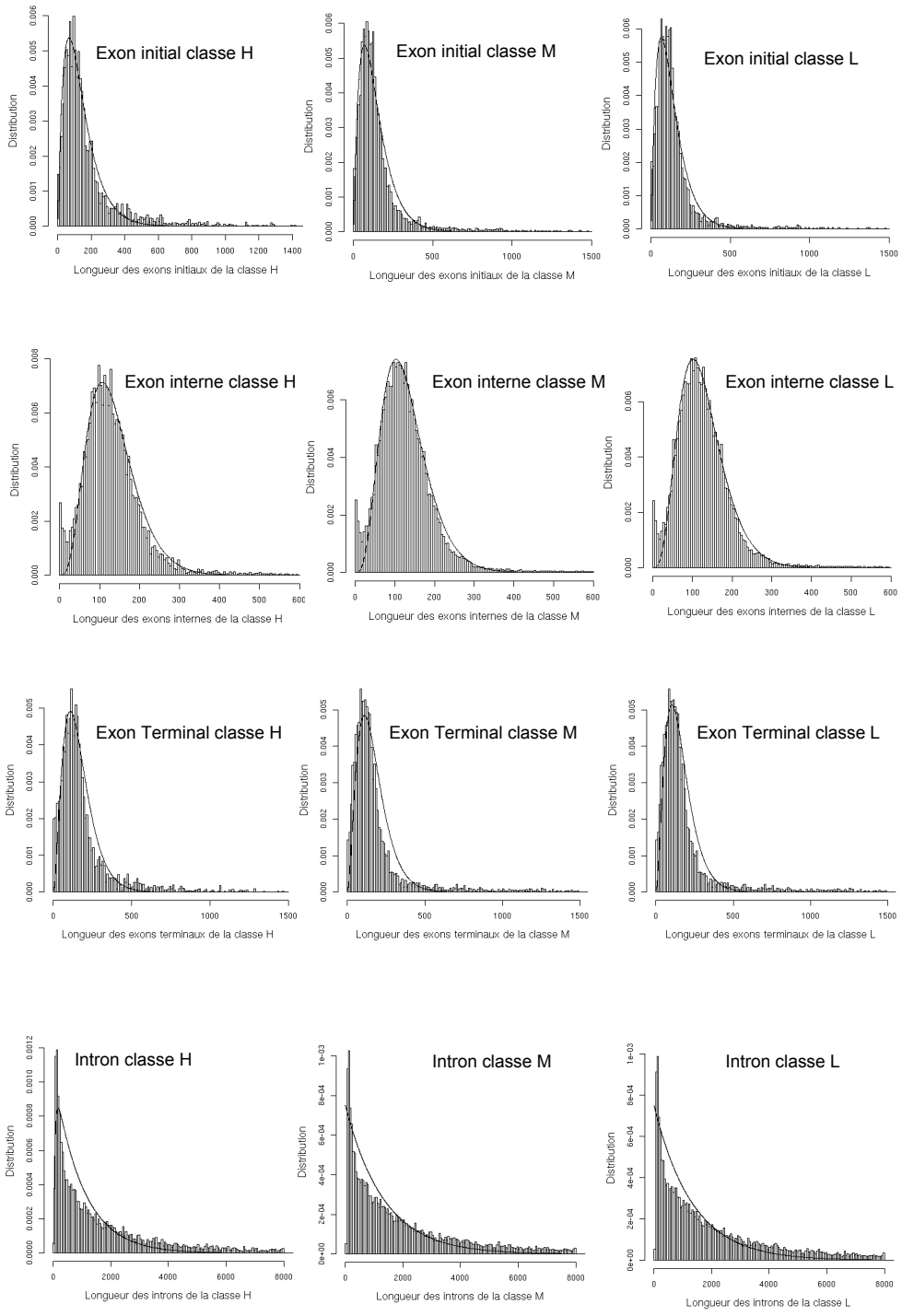


Fig. 5.8 : *Les histogrammes représentent les distributions empiriques de longueurs des différentes régions et les courbes pleines décrivent les lois théoriques qui minimisent dans chaque cas la distance de Kolmogorov-Smirnov.*

SOURIS

Lois	Paramètres	Distance de K-S
$G_2(p)$	86,5	0,06402
$G_3(p)$	56,7	0,10644
$G_4(p)$	41,1	0,13798
$G_5(p)$	31,9	0,01619
$G(p_1, p_2)$	55,6 - 117,6	0,05139

Tab. 5.28 : *Souris : Exon initial classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	77,6	0,06311
$G_3(p)$	51,3	0,10446
$G_4(p)$	36,7	0,13798
$G_5(p)$	28,4	0,16188
$G(p_1, p_2)$	71,4 - 84,0	0,05342

Tab. 5.29 : *Souris : Exon initial classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	70,1	0,05786
$G_3(p)$	46,4	0,07347
$G_4(p)$	33,3	0,09658
$G_5(p)$	25,4	0,11526
$G(p_1, p_2)$	25,4 - 74,1	0,056294

Tab. 5.30 : *Souris : Exon initial classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	75,9	0,11871
$G_3(p)$	48,3	0,06700
$G_4(p)$	35,1	0,02933
$G_5(p)$	27,9	0,02137

Tab. 5.31 : *Souris : Exon interne classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	74,1	0,18455
$G_3(p)$	46,6	0,06762
$G_4(p)$	33,9	0,02932
$G_5(p)$	26,8	0,02134

Tab. 5.32 : *Souris : Exon interne classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	71,4	0,11285
$G_3(p)$	46,1	0,06006
$G_4(p)$	33,5	0,02286
$G_5(p)$	26,5	0,02486

Tab. 5.33 : *Souris : Exon interne classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	93,5	0,06781
$G_3(p)$	19,4	0,09036
$G_4(p)$	44,8	0,10844
$G_5(p)$	35,1	0,12497
$G(p_1, p_2)$	58,8 - 83,3	0,05535

Tab. 5.34 : *Souris : Exon terminal classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	83,4	0,07024
$G_3(p)$	56,4	0,09786
$G_4(p)$	41,2	0,11583
$G_5(p)$	31,5	0,13221
$G(p_1, p_2)$	55,5 - 11,4	0,06611

Tab. 5.35 : *Souris : Exon terminal classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	80,6	0,09662
$G_3(p)$	53,2	0,11811
$G_4(p)$	39,4	0,13151
$G_5(p)$	30,4	0,14821
$G(p_1, p_2)$	66,6 - 95,4	0,08426

Tab. 5.36 : *Souris : Exon terminal classe L*

Lois	Paramètres	Distance de K-S
$G(p)$	1176,5	0,08451
$G_2(p)$	486,4	0,17540
$G_3(p)$	294,1	0,22634
$G_4(p)$	207,5	0,34512
$G_5(p)$	160,5	0,45361
$G(p_1, p_2)$	55,6 - 1000	0,15423

Tab. 5.37 : *Souris : Intron classe H*

Lois	Paramètres	Distance de K-S
$G(p)$	1333,3	0,09473
$G_2(p)$	588,2	0,15967
$G_3(p)$	400	0,22165
$G_4(p)$	253,8	0,28971
$G_5(p)$	195,3	0,38164
$G(p_1, p_2)$	66,7 - 1111,1	0,12647

Tab. 5.38 : *Souris : Intron classe M*

Lois	Paramètres	Distance de K-S
$G(p)$	1333,3	0,08737
$G_2(p)$	584,8	0,15642
$G_3(p)$	392,1	0,20061
$G_4(p)$	277,8	0,24168
$G_5(p)$	212,8	0,29781
$G(p_1, p_2)$	55,6 - 1250	0,13481

Tab. 5.39 : *Souris : Intron classe L*

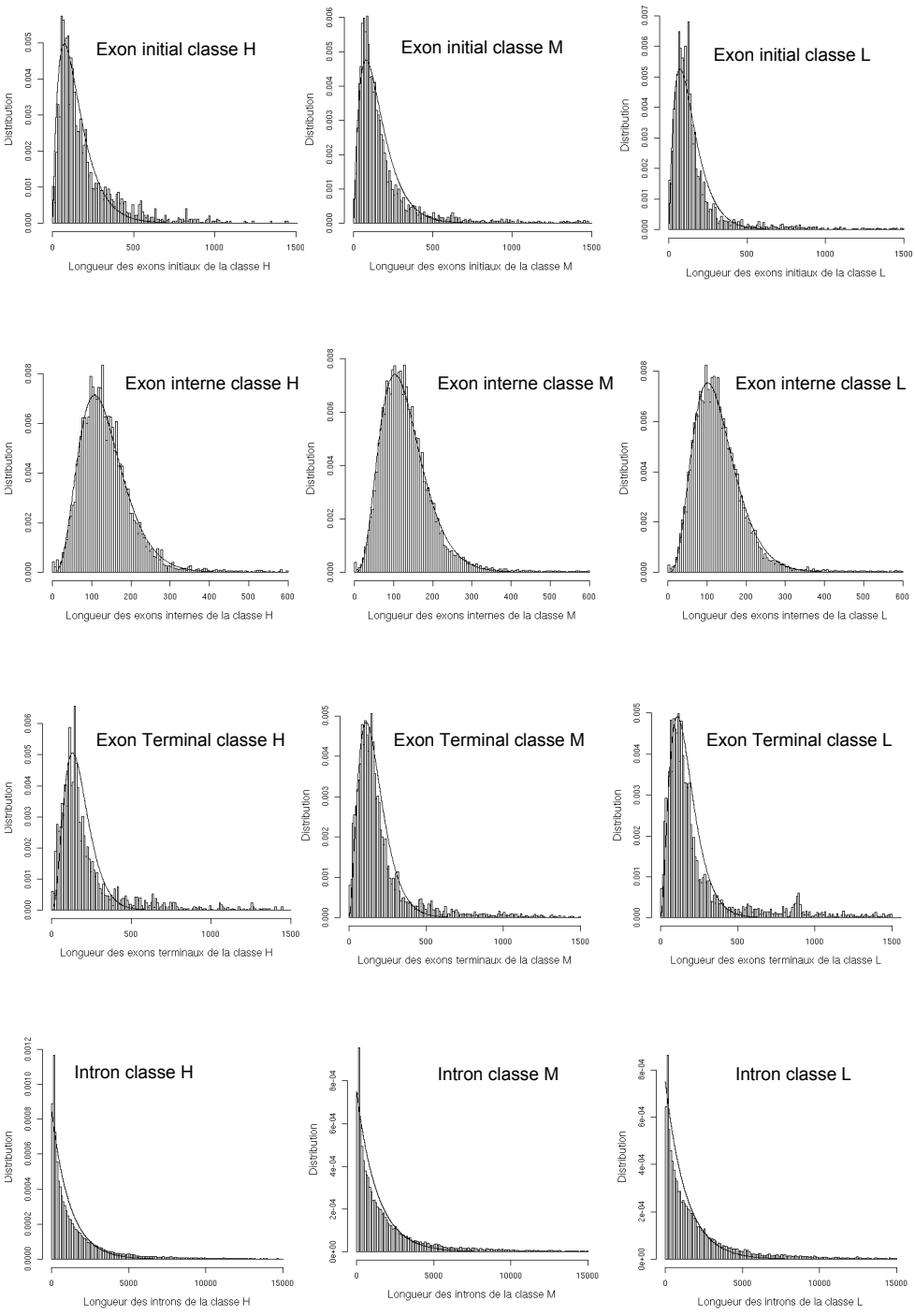


Fig. 5.9 : *Les histogrammes représentent les distributions empiriques de longueurs des différentes régions et les courbes pleines décrivent les lois théoriques qui minimisent dans chaque cas la distance de Kolmogorov-Smirnov.*

POULET

Lois	Paramètres	Distance de K-S
$G_2(p)$	86,5	0,05786
$G_3(p)$	56,7	0,08058
$G_4(p)$	41	0,10351
$G_5(p)$	31,8	0,12235
$G(p_1, p_2)$	55,6 - 117,7	0,051301

Tab. 5.40 : *Poulet : Exon initial classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	80,7	0,04594
$G_3(p)$	51,3	0,06227
$G_4(p)$	36,7	0,08355
$G_5(p)$	28,4	0,10192
$G(p_1, p_2)$	71,4 - 84	0,04547

Tab. 5.41 : *Poulet : Exon initial classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	70,1	0,05540
$G_3(p)$	46,4	0,05146
$G_4(p)$	33,3	0,06442
$G_5(p)$	25,4	0,08072
$G(p_1, p_2)$	66,7 - 74,1	0,05447

Tab. 5.42 : *Poulet : Exon initial classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	76,8	0,12323
$G_3(p)$	49,5	0,07499
$G_4(p)$	35,1	0,04108
$G_5(p)$	27,9	0,02709

Tab. 5.43 : *Poulet : Exon interne classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	74,7	0,12688
$G_3(p)$	46,5	0,07346
$G_4(p)$	33,8	0,03857
$G_5(p)$	26,9	0,02454

Tab. 5.44 : *Poulet : Exon interne classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	71,4	0,13535
$G_3(p)$	46,1	0,07418
$G_4(p)$	33,5	0,03879
$G_5(p)$	26,5	0,01841

Tab. 5.45 : *Poulet : Exon interne classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	93,4	0,06904
$G_3(p)$	60,8	0,08754
$G_4(p)$	44,8	0,10236
$G_5(p)$	47,6	0,11665
$G(p_1, p_2)$	58,8 - 83,3	0,06898

Tab. 5.46 : *Poulet : Exon terminal classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	83,5	0,06223
$G_3(p)$	56,4	0,07630
$G_4(p)$	41,0	0,08932
$G_5(p)$	31,5	0,09991
$G(p_1, p_2)$	58,6 - 83,3	0,05615

Tab. 5.47 : *Poulet : Exon terminal classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	80,6	0,06289
$G_3(p)$	53,2	0,05676
$G_4(p)$	39,4	0,07191
$G_5(p)$	30,4	0,08619
$G(p_1, p_2)$	95,2 - 66,6	0,06357

Tab. 5.48 : *Poulet : Exon terminal classe L*

Lois	Paramètres	Distance de K-S
$G(p)$	1176,5	0,06699
$G_2(p)$	486,4	0,12428
$G_3(p)$	294,1	0,17016
$G_4(p)$	207,5	0,19552
$G_5(p)$	160,5	0,21766
$G(p_1, p_2)$	55,6 - 1000	0,06556

Tab. 5.49 : *Poulet : Intron classe H*

Lois	Paramètres	Distance de K-S
$G(p)$	1333,3	0,07361
$G_2(p)$	584,8	0,13089
$G_3(p)$	292,1	0,16299
$G_4(p)$	277,7	0,19151
$G_5(p)$	212,8	0,21387
$G(p_1, p_2)$	55,6 - 1250	0,071428

Tab. 5.50 : *Poulet : Intron classe M*

Lois	Paramètres	Distance de K-S
$G(p)$	1388,9	0,07262
$G_2(p)$	588,2	0,13146
$G_3(p)$	400	0,19577
$G_4(p)$	253,8	0,20366
$G_5(p)$	185,3	0,21636
$G(p_1, p_2)$	66,7 - 111,1	0,06607

Tab. 5.51 : *Poulet : Intron classe L*

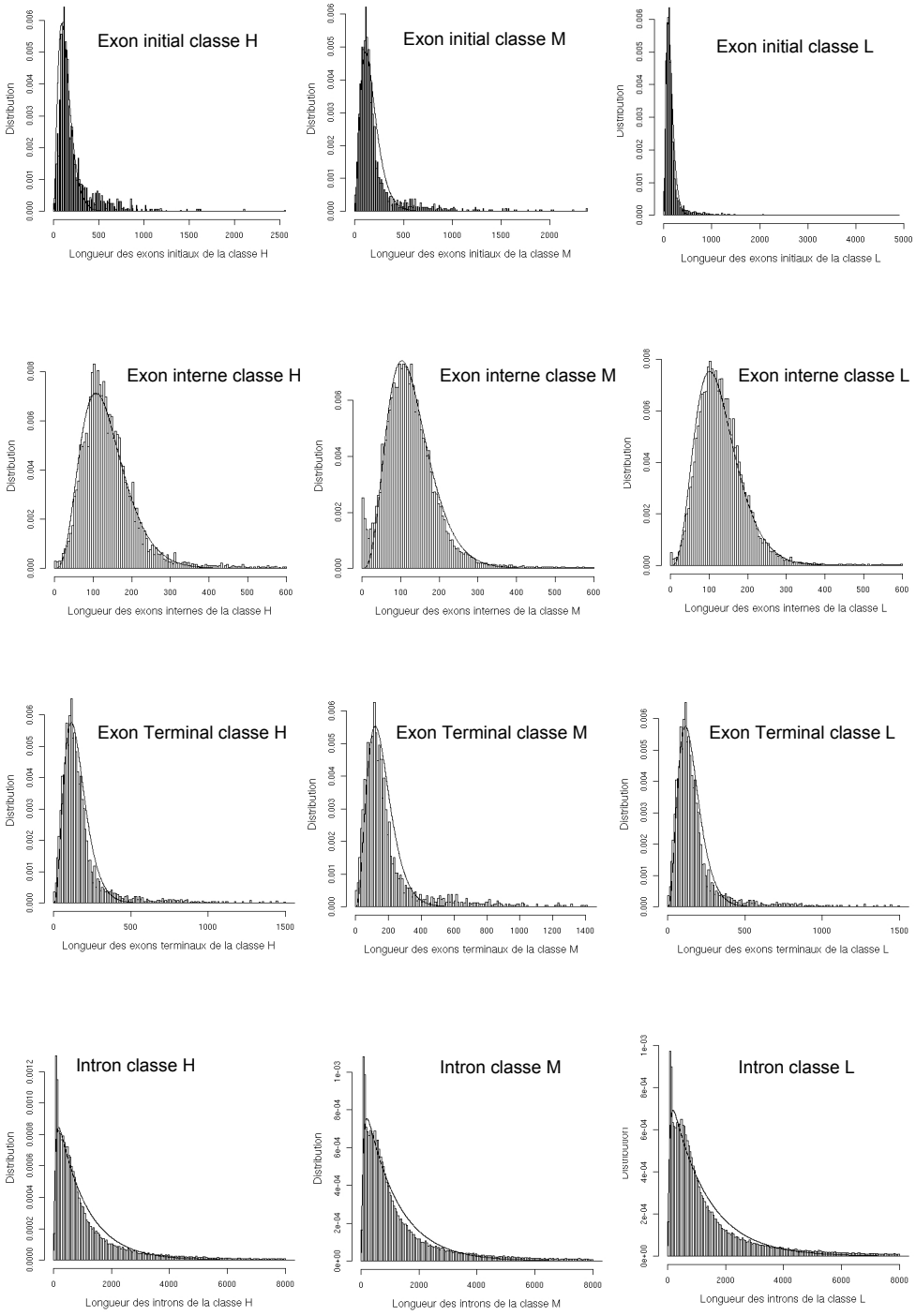


Fig. 5.10 : *Les histogrammes représentent les distributions empiriques de longueurs des différentes régions et les courbes pleines décrivent les lois théoriques qui minimisent dans chaque cas la distance de Kolmogorov-Smirnov.*

TETRAODON

Lois	Paramètres	Distance de K-S
$G_2(p)$	91,2	0,05296
$G_3(p)$	57,6	0,08992
$G_4(p)$	40,9	0,12098
$G_5(p)$	31,25	0,14514
$G(p_1, p_2)$	55,5 - 125	0,04692

Tab. 5.52 : *Tetraodon : Exon initial classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	82,3	0,05943
$G_3(p)$	53,3	0,08530
$G_4(p)$	38,1	0,01100
$G_5(p)$	25	0,10506
$G(p_1, p_2)$	57,1 - 108,7	0,054199

Tab. 5.53 : *Tetraodon : Exon initial classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	68,4	0,05360
$G_3(p)$	43,9	0,07424
$G_4(p)$	31,2	0,09674
$G_5(p)$	22,2	0,01424
$G(p_1, p_2)$	55,5- 110,4	0,04535

Tab. 5.54 : *Tetraodon : Exon initial classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	78,4	0,10955
$G_3(p)$	48,8	0,05900
$G_4(p)$	36,1	0,03959
$G_5(p)$	28,9	0,0476

Tab. 5.55 : *Tetraodon : Exon interne classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	78,4	0,11040
$G_3(p)$	48,6	0,05857
$G_4(p)$	35,9	0,03813
$G_5(p)$	28,7	0,046544

Tab. 5.56 : *Tetraodon : Exon interne classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	75,5	0,10512
$G_3(p)$	45,7	0,05382
$G_4(p)$	34,7	0,03431
$G_5(p)$	27,9	0,02555

Tab. 5.57 : *Tetraodon : Exon interne classe L*

Lois	Paramètres	Distance de K-S
$G_2(p)$	110,4	0,05724
$G_3(p)$	73,2	0,08456
$G_4(p)$	52,6	0,10819
$G_5(p)$	40	0,33554
$G(p_1, p_2)$	57,1 - 172,4	0,056627

Tab. 5.58 : *Tetraodon : Exon terminal classe H*

Lois	Paramètres	Distance de K-S
$G_2(p)$	104,3	0,06765
$G_3(p)$	67,9	0,09261
$G_4(p)$	45,4	0,12879
$G_5(p)$	38,4	0,12922
$G(p_1, p_2)$	55,1 - 153,8	0,05555

Tab. 5.59 : *Tetraodon : Exon terminal classe M*

Lois	Paramètres	Distance de K-S
$G_2(p)$	93,8	0,05540
$G_3(p)$	60,6	0,08277
$G_4(p)$	44,6	0,10707
$G_5(p)$	34,9	0,128056
$G(p_1, p_2)$	53,8- 161,3	0,054312

Tab. 5.60 : *Tetraodon : Exon terminal classe L*

Lois	Paramètres	Distance de K-S
$G(p)$	316,5	0,15874
$G_2(p)$	112,1	0,18581
$G_3(p)$	68,2	0,19461
$G_4(p)$	47,9	0,21656
$G(p_1, p_2)$	112,4 - 130,0	0,174652

Tab. 5.61 : *Tetraodon : Intron classe H*

Lois	Paramètres	Distance de K-S
$G(p)$	344,8	0,14069
$G_2(p)$	125,9	0,18704
$G_3(p)$	76,3	0,19834
$G_4(p)$	54,1	0,20648
$G(p_1, p_2)$	125,9 - 117,6	0,17584

Tab. 5.62 : *Tetraodon : Intron classe M*

Lois	Paramètres	Distance de K-S
$G(p)$	375,9	0,17621
$G_2(p)$	132,6	0,20985
$G_3(p)$	79,6	0,22431
$G_4(p)$	56,2	0,25505
$G(p_1, p_2)$	133,3 - 100	0,18744

Tab. 5.63 : *Tetraodon : Intron classe L*

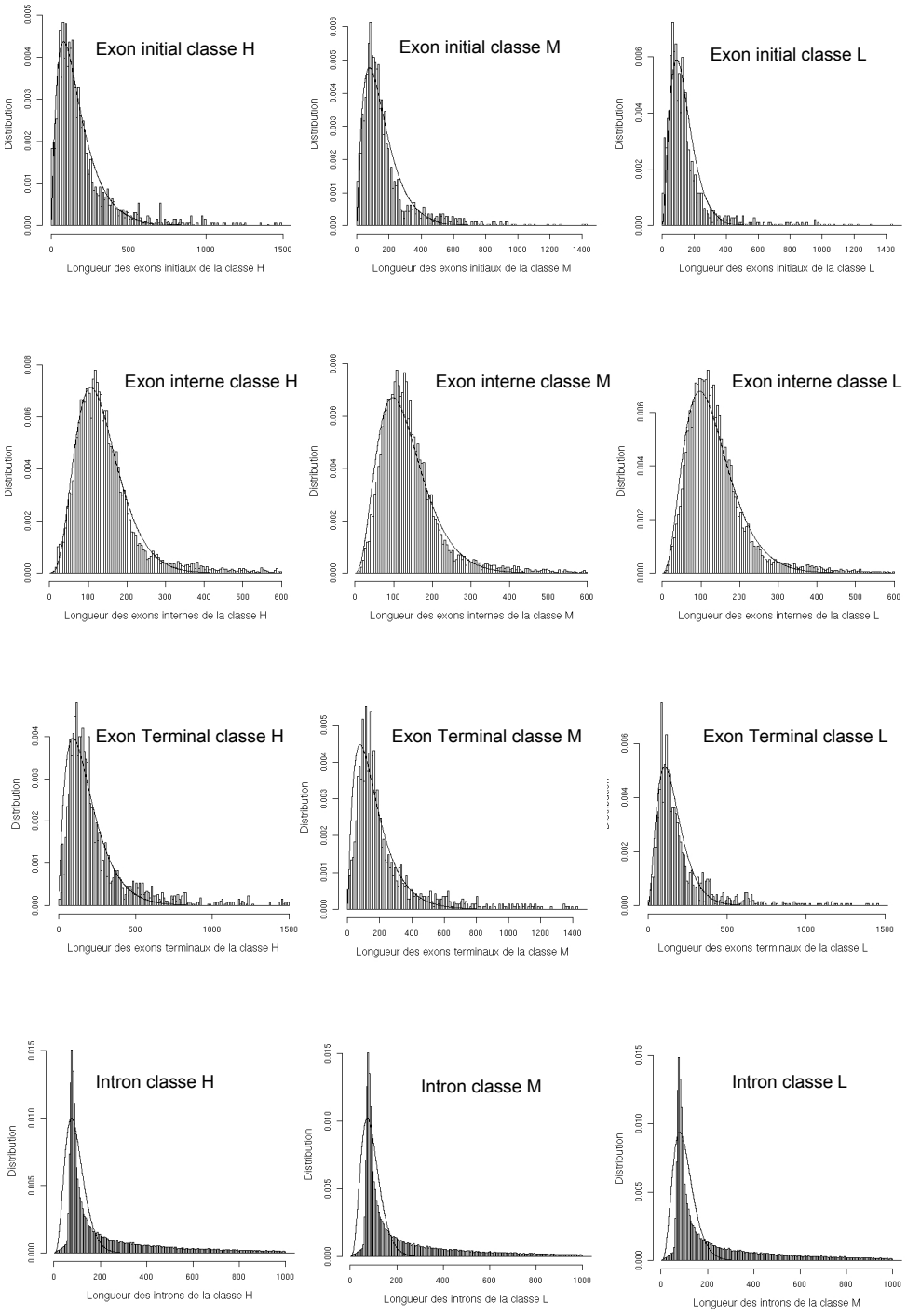


Fig. 5.11 : *Les histogrammes représentent les distributions empiriques de longueurs des différentes régions et les courbes pleines décrivent les lois théoriques qui minimisent dans chaque cas la distance de Kolmogorov-Smirnov.*

5.7 Convolution de lois géométriques

Une loi géométrique ne peut modéliser la distribution de la longueur des exons. L'originalité de la méthode proposée dans cette section consiste à ajuster à cette distribution empirique de la longueur des exons une convolution de plusieurs lois géométriques de paramètres égaux ou différents. Bien que cette approche ait déjà été suggérée (Durbin *et al.* 1998 page 69), elle n'a jamais été utilisée pour la modélisation des génomes.

Les distributions des longueurs des exons et des introns sont estimées à partir des échantillons $x_1 \ldots x_n$ des séquences des jeux d'entraînement. Les x_i sont considérés comme des réalisations de variables indépendantes de même loi. Les lois testées sont les suivantes :

– une somme de $m \geq 1$ lois géométriques de paramètre p (*i.e.* une loi binomiale négative) :

$$P[S = s] = C_{s-1}^{m-1} \times p^m \times (1-p)^{s-m},$$

– une somme de deux lois géométriques de paramètres différents avec $p_1 > p_2$ (annexe) :

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

– une somme de trois lois géométriques de paramètres différents avec $p_1 < p_2 < p_3$ (annexe) :

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

Somme de deux lois géométriques de paramètres différents

Soient X et Y deux variables aléatoires indépendantes de lois géométriques de paramètres respectifs p_1 et p_2. Soit S=X+Y

$$\begin{aligned}
P[S=s] &= \sum_{i\geq 1} P[Y=i, X+Y=s] \\
&= \sum_{i=1,\ s-i\geq 1} P[Y=i, X=s-i] \\
&= \sum_{i=1}^{s-1} P[Y=i] \times P[X=s-i] \\
&= \sum_{i=1}^{s-1} p_1(1-p_1)^{i-1} \times p_2(1-p_2)^{s-i-1} \\
&= p_1p_2(1-p_2)^s \sum_{i=1}^{s-1} (1-p_1)^{i-1} \times (1-p_2)^{-i-1} \\
&= p_1p_2(1-p_2)^s(1-p_2)^{-2} \sum_{i=1}^{s-1} \frac{(1-p_1)^{i-1}}{(1-p_2)^{i-1}} \\
&= p_1p_2(1-p_2)^{s-2} \sum_{i=0}^{s-2} \frac{(1-p_1)^i}{(1-p_2)^i} \\
&= p_1p_2(1-p_2)^{s-2} \frac{1-\frac{(1-p_1)^{s-1}}{(1-p_2)^{s-1}}}{1-\frac{(1-p_1)}{(1-p_2)}} \quad p_1 > p_2 \\
&= p_1 \times p_2 \frac{(1-p_2)^{s-1} - (1-p_1)^{s-1}}{p_1 - p_2}
\end{aligned}$$

Somme de trois lois géométriques de paramètres différents

Soient X, Y et Z trois variables aléatoires indépendantes de lois géométriques de paramètres respectifs p_1, p_2 et p_3. Soit S=X+Y+Z

$$
\begin{aligned}
P[S=s] &= \sum_{i\geq 1} P[Y=i, X+Y+Z=s] \\
&= \sum_{i=1,\ s-i\geq 1} P[Y=i, X+Z=s-i] \\
&= \sum_{i=1}^{s-1}\sum_{j\geq 1} P[Y=i, Z=j, X+Z=s-i] \\
&= \sum_{i=1}^{s-1}\sum_{j=1}^{s-i-1} P[Y=i, Z=j, X=s-i-j] \\
&= \sum_{i=1}^{s-1}\sum_{j=1}^{s-i-1} P[Y=i]P[Z=j]P[X=s-i-j] \\
&= \sum_{i=1}^{s-1}\sum_{j=1}^{s-i-1} p_2(1-p_2)^{i-1}\times p_3(1-p_3)^{j-1}\times p_1(1-p_1)^{s-i-j-1} \\
&= p_1\ p_2\ {}_3\ (1-p_1)^s \sum_{i=1}^{s-1}\frac{(1-p_2)^{i-1}}{(1-p_1)^i}\sum_{j=1}^{s-i-1}\frac{(1-p_3)^{j-1}}{(1-p_1)^{j+1}} \\
&= p_1\ p_2\ p_3\ (1-p_1)^{s-2} \sum_{i=1}^{s-1}\frac{(1-p_2)^{i-1}}{(1-p_1)^i}\sum_{j=1}^{s-i-1}\frac{(1-p_3)^{j-1}}{(1-p_1)^{j-1}} \\
&= p_1\ p_2\ p_3\ (1-p_1)^{s-2} \sum_{i=1}^{s-1}\frac{(1-p_2)^{i-1}}{(1-p_1)^i}\left\{\frac{1-\frac{(1-p_3)^{s-1}}{(1-p_1)^{s-i-1}}}{1-\frac{(1-p_3)}{(1-p_1)}}\right\} \quad p_3>p_1 \\
&= \frac{p_1p_2p_3}{p_3-p_1}\left\{\sum_{i=1}^{s-1}(1-p_2)^{i-1}(1-p_1)^{s-i-1}-\sum_{i=1}^{s-1}(1-p_2)^{i-1}(1-p_3)^{s-i-1}\right\} \\
&\qquad p_2>p_1 \quad p_2>p_3 \\
&= \frac{p_1\times p_2\times p_3}{p_3-p_1}\times\left\{\frac{(1-p_1)^{s-1}-(1-p_3)^{s-1}}{p_2-p_1}-\frac{(1-p_2)^{s-1}-(1-p_3)^{s-1}}{p_2-p_3}\right\}
\end{aligned}
$$

5.8 Publications

5.8.1 Article 1

Prediction of human isochores using a hidden Markov model.

Melodelima C., Guéguen L., Piau D. et Gautier G.

JOBIM, 2005 : 427-434.

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.

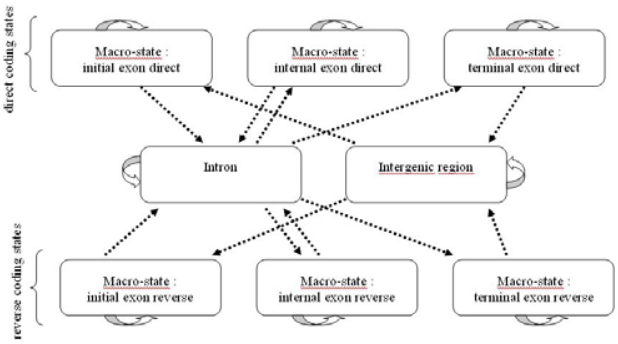
3 METHODS

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.

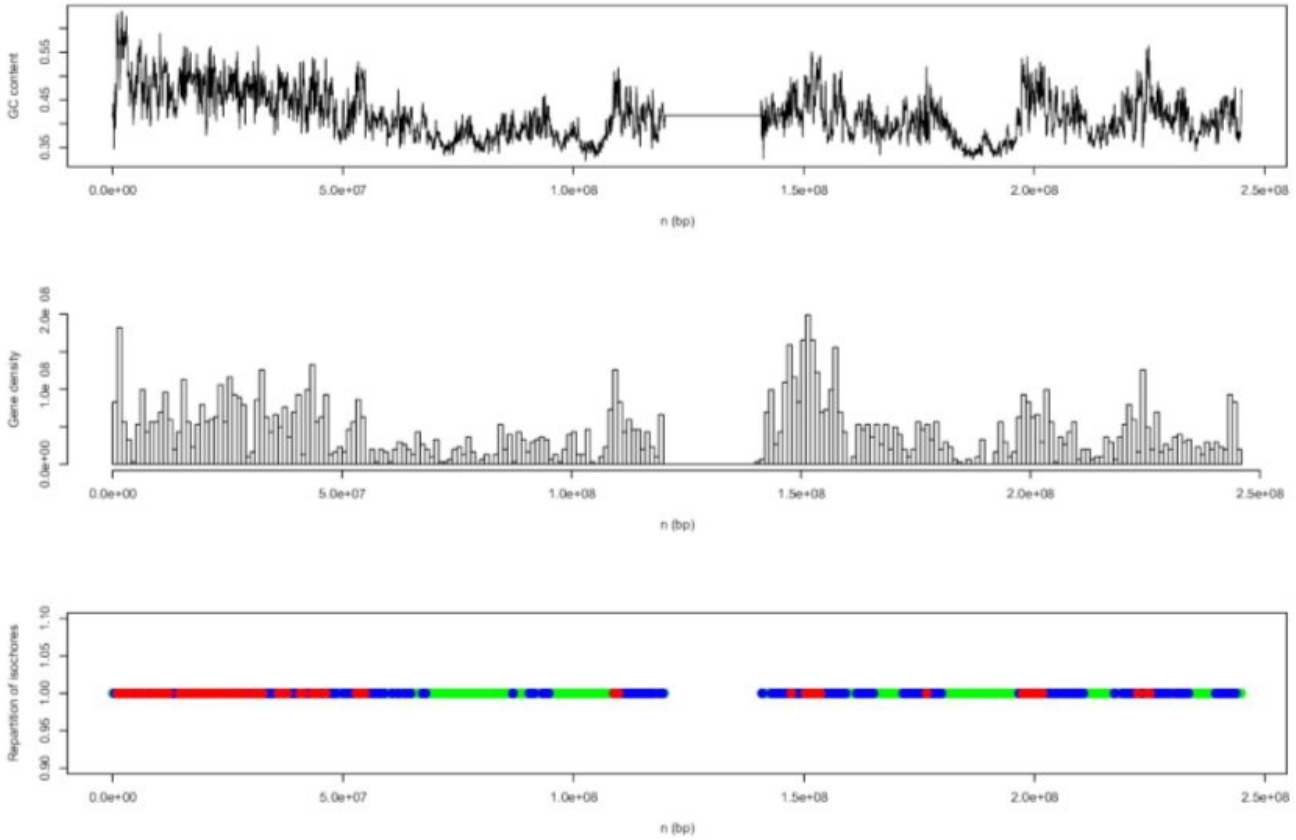


4 RESULTS

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.

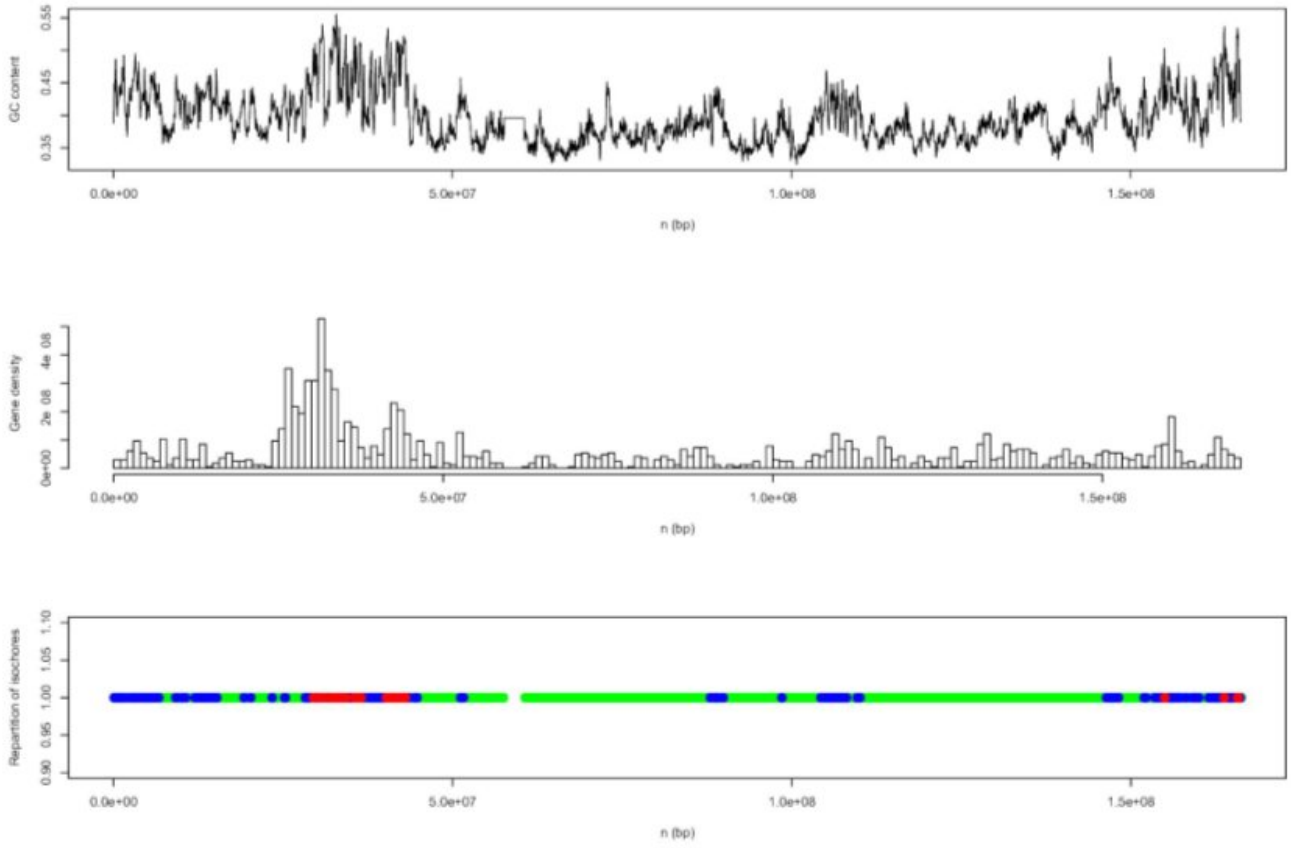


Table 1: Analysis of the prediction of the macro-states on the different regions.

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

5.8.2 Article 2

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

Melodelima C. , Guéguen L., Gautier C. et Piau D.

JOBIM, 2004 : 142-154.

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 $x_1 \dots x_n$ 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.

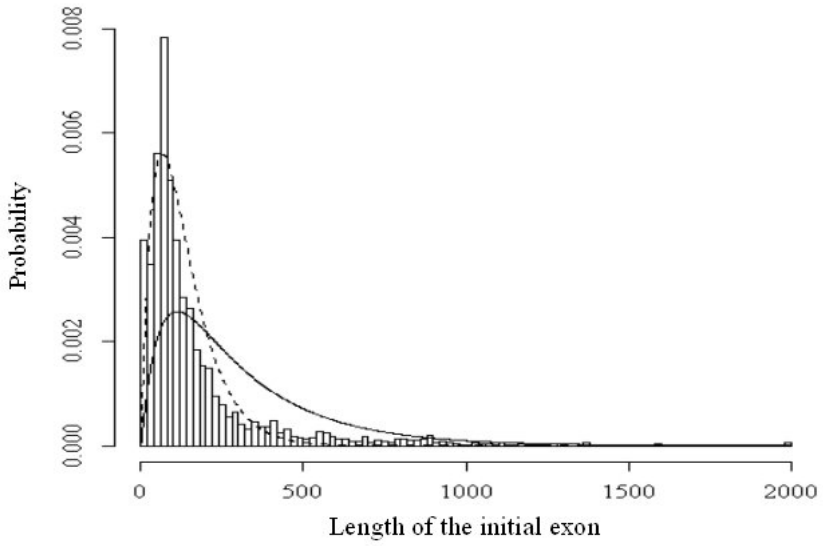


Figure 1: *Empirical distribution of the length of the initial coding exon. The histogram represents the empirical distribution of the length of the initial exons in a multi-exons gene. The dotted line describes the theoretical distribution, obtained by the Kolmogorov-Smirrnov distance. The full line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

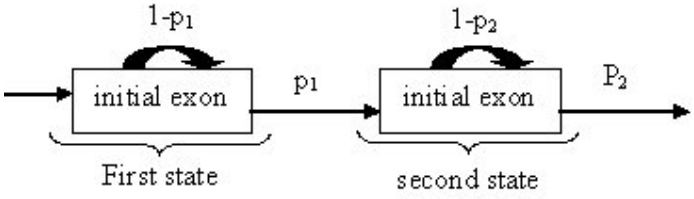


Figure 2: *The initial exon length distribution is modelled by a sum of two geometrical laws of parameters p_1 and p_2, this region is represented by two states, with the duration state: p_1 and p_2. The probabilities of emission of these two states are the same.*

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.

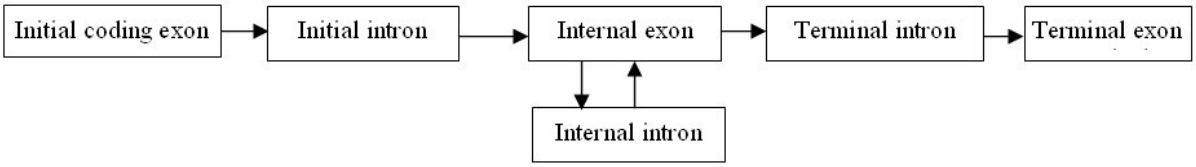


Figure 3: *A representation of our definition of gene.*

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:

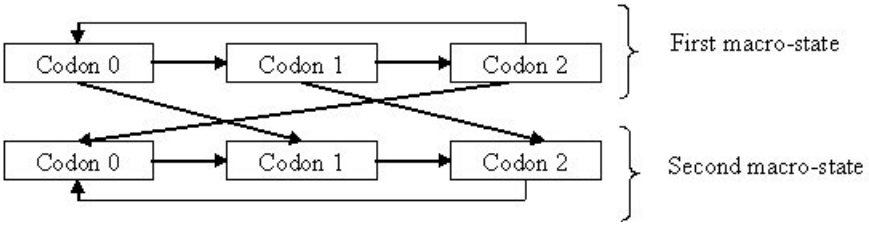


Figure 4: *This figure represents the initial exon HMM that is separated into two macros-states to take into account the length distribution of initial exons (sum of two geometrical laws). Each macro-state is split into three states that represent the position of the nucleotide in the codon, the two 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.

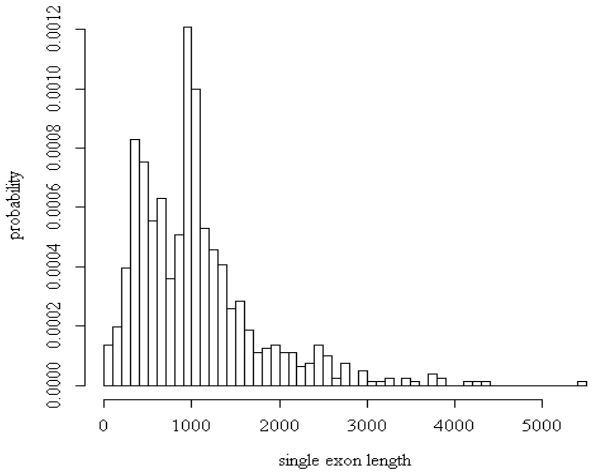


Figure 5: *Empirical distribution of the length of single exons (that is, of intronless genes).*

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

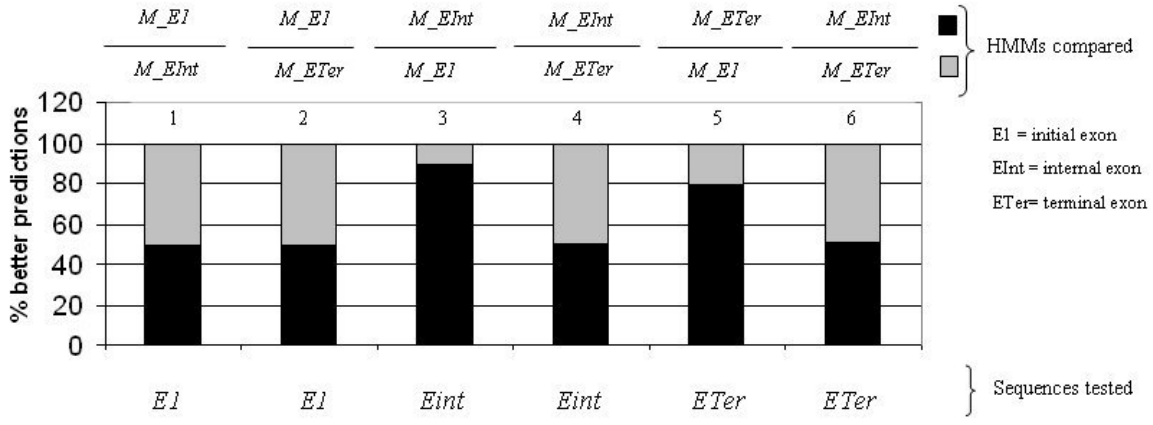


Figure 6: *Models learned on different sequences (initial, internal and terminal exons) were pairwise compared on the same sequences to determine the best predictions.*
For instance, in histogram 1, the likelihood of each first exon is computed under models learned on M_E1 and M_Eint. The black bar represents 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).

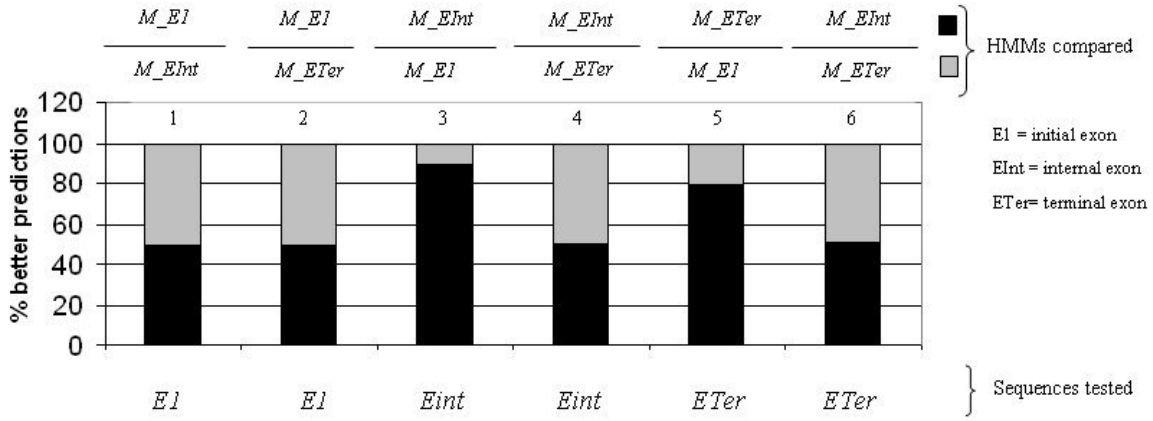


Figure 7: *The models learned on different sequences (initial, internal and terminal exons) were pairwise compared on the initial exons to determine the best predictions.*
The model M_IE$_{80}$ predicts better initial exons than all different models tested.

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.

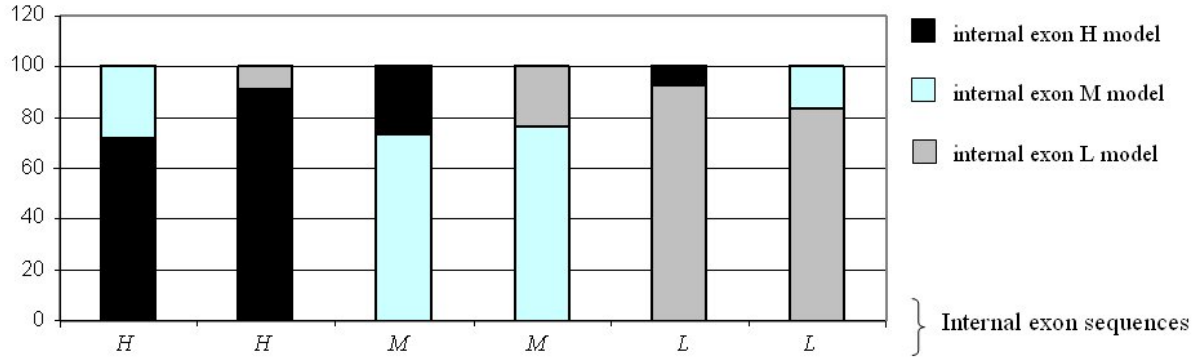


Figure 8: *The models learned on different sequences (internal exons of classes H, M and L) were pairwise compared on the same sequences to determine the best predictions.*

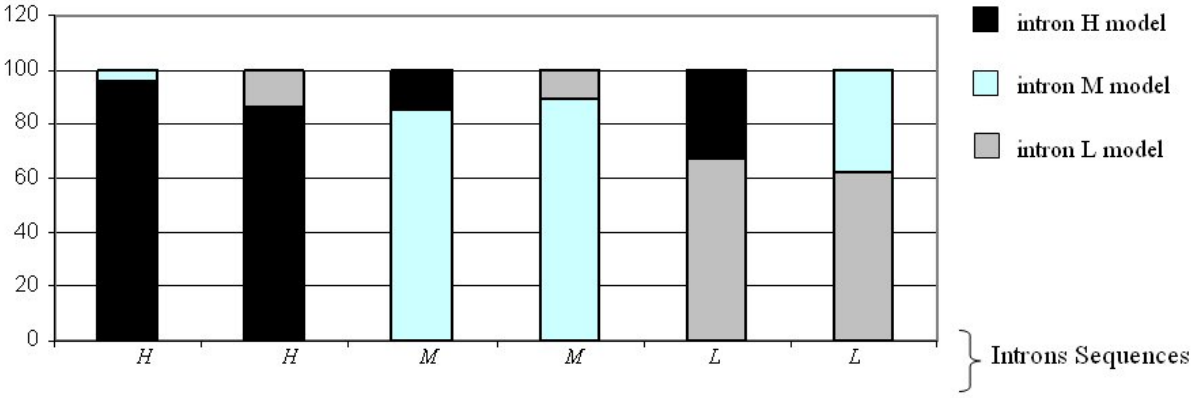


Figure 9: *The models learned on different sequences (introns of classes H, M and L) were pairwise compared on the same sequences to determine the best predictors.*

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.

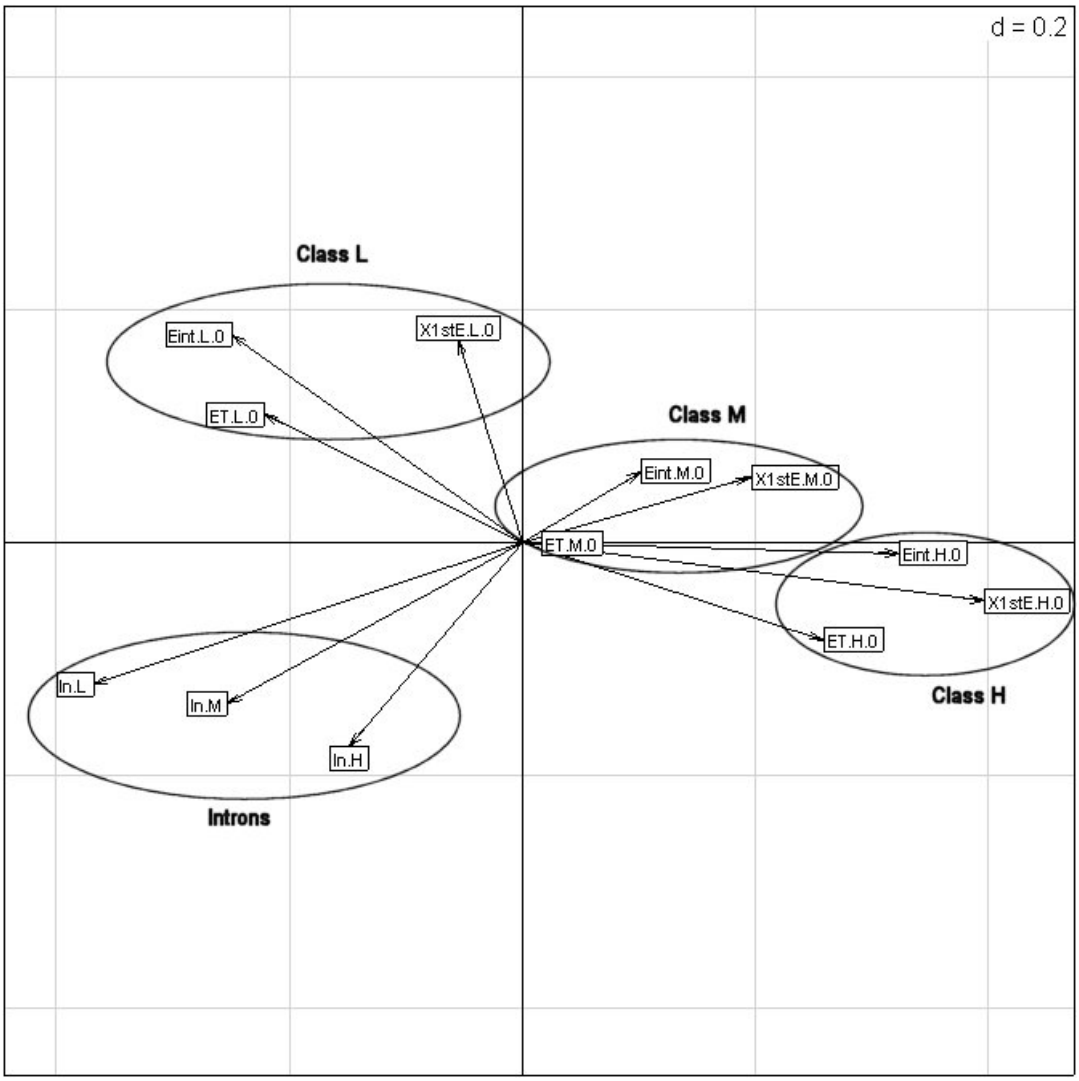


Figure 10: *Correspondence analysis of the emission probabilities of the different state models in reading frame 0.*
EInt.H.0=internal exon model of class H and reading frame 0
EInt.M.0=internal exon model of class M and reading frame 0
EInt.L.0=internal exon model of class L and reading frame 0
ETer.H.0=terminal 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 $p_1 < p_2 < p_3$:

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

5.8.3 Article 3

Genomic mapping and molecular process.

Gautier C., Navratil V. et Melodelima C.

Proceedings of Mathematical and Computational Biology, 2003 ; vol 2 : 22-28.

Genomic mapping and molecular processes

Christian Gautier, Vincent Navratil, Christelle Melo de Lima
LBBE, Universit Lyon I, 43 bd 11 novembre,69622 Villeurbanne Cedex, FRANCE

Abstract

This paper describe both management and analysis of genomic mapping data. A UML representation of both vertebrates genome maps and evolutionary relationships between gene is presented. Statistical analysis has focused on isochore organisation, subtitution rate and skew. Natural selection versus mutational bias is discussed.

Keywords:Genomic mapping, evolution

1 Introduction

Gene density, gene structure as well as genomic sequence statistical properties vary along genomes and define regions that are more or less homogeneous. These regions interact with three levels of biological constraints: i) genetic information including genes, regulatory elements, ...; ii) the management process of this information (transcription or replication units, recombination process, ...); iii) the spatial organisation of the genomic DNA with the different packaging level of chromatin (nucleosomes and higher order organisation). An important step in understanding the functioning of genomes is to associate statistical properties of sequences to each of these biological constraints. However this analysis must take into account evolutionary processes that have generated and that maintain these associations. Genomic patterns so results from a combination of several levels of constraints, natural selection and mutational bias. Inferring processes from patterns is a very complex task, this paper tries to show that taking into account spatial organisation could be of great help in genomic sequence analysis. This strategy needs developping new methodological tools both to manage and to analyse data. In this paper we will mainly focus on data base management in wich we are embedded from several years [1]. However some reference to statistical developments will be made.

Prokaryotic and eukaryotic genomes have quite different behaviors relatively to spatial patterns. We will present separatly these two groups of organisms and will focuse for eucaryotes on vertebrates.

2 Procaryotic genome patterns

Complete procaryotic genome has been the first homogeneous unit describe [2] and its discussion take always an important role in the debate between mutation bias versus selection in genome patterning. Sueka, as soon as in xxxx, determined experimentally the G+C content of bacterial genomes. It appears that those genomes show a very large range of G+C content. This raises the debate between two hypothesis: i) G+C content is linked to the fitness of the organism and its level results from a natural selection process; ii) the G+C content range inside bacteria results from a varability of the mutational bias. So the G+C content variability takes place inside the neutralist vs selectionist debate. Due to the fact that G.C link is stronger than A.T link, it has been postulated that optimal growth temperature (T_{opt})determine selection pressure acting on G+C content. Correlation studies between T_{opt} and G+C content have ruled out this hypothesis. Sueoka propose then his hypothesis of neutral modification of the replication apparatus implying variation of mutational bias. More recent studies [3] have precised relationships between temperature and G+C content. If no relation exists between genomic G+C content and T_{opt}, T_{opt} is correlated to the G+C content of genome regions coding for helix part of ribosomal RNA. This shows that probably two evo-

lutionary processes act on G+C content: mutation bias patterns the whole genome and results in the large range of bacteria genomic content, natural selection maintains a high G+C content in the region where RNA molecule structure must be conserved for high T_{opt}.

Superposition of genome patterning due to mutational bias and natural selection process is also examplifies by several structures linked to replication process. Full discussion, bibliography and continously updated analysis of complete bacterial genomes is provided by J. Lobry on the web site: http://pbil.univ-lyon1.fr/software/Oriloc/ . These data are generated by specific functions inside the R package project *SeqinR*. We will just summarize here the two main structures. It could be demonstrated that under the hypothesis of similar mutational process on each strand the equilibrium is characterized by equal amount of A and T on one side and of C and G on the other side [4]. Transforming bacterial genomes in a walk on a line define by a +1 step for C (resp. A) and -1 step for G (resp. T) put into evidence strong tendencies that delimitates for most genome two regions corresponding to the replicons. See for example Borrelia burgdorferi NC001318 or Escherichia coli CFT073 NC004431 from the Oriloc site. This approach provides an efficient estimation of the localisation of replication origin and terminus [5]. Biological interpretation of the pattern relies on the dissymetry of replication: replication works continuouly during lagging strand replication but discontinuouly when leading strand is processed (and so lagging strand is generated). The fact that this process implies that leading strand remains in a single strand state longer than the lagging strand has been particularly related to the dissymetry of mutations. This regional statistical pattern of bacterial genomes is one of the best examples where a spatial statistical analysis of genomes has led to put into evidence a biological mechanism. However another strong dissymetry can be shown between lagging and leading strands. The same type of walk along the genome with +1 when a gene is on the leading strand and -1 if it is coded on the lagging one shows also a very clear tendency (see Oriloc site). Genes are more often coded on the leading strand than on the lagging strand. This asymmetry in gene direction is considered as resulting from natural selection acting to avoid head-on collisions between replication and transcription [6].

3 Vertebrates genomes

3.1 available data

Since human genome sequencing mouse and rat genome sequence have been determined and dog genome will be known before the end of 2004. Available data on vertebrate genome is so increasing at a very great exponential rate, probably greater than the rate at which data base can be efficiently updated and sequences annotated. Moreover biological information necessary to annotate sequence is not accumulated at the same rate and if genomic sequence itself is quite well known it is far to be the same for genetic elements (genes, promoters, ...). So presently many genes are only determined through mathematical estimation and comparative analysis.

Genomic sequence is not the only means to position genetic elements on the genome and so *sequence maps* must be compared to other maps. "Cheaper" maps (particularly radiation hybrid maps "RH maps") allow comparaisons between model species for which sequence is known to species of interest (farm animals for example) for which RH map are built.

Genetic analysis of families allow to build genetic map, allowing both to position genes and to determine genetic distances and local recombination rate. It is important to note that genetic distances between markers is not a linear function of the physical distance measured by the number of nucleotides between the markers. We will see that the varaition of recombination rate along the genome could have important statistical consequences on sequences. Building genetic maps is a complex task, particularly for vetebrate having large generation timespan. A more efficient approach is provided by radiation hybrid mapping which provided distances well correlated to physical distances (RH mapping).

When two vertebrate genome maps are compared, conserved segments appears in which genes have similar orders. We will not discuss here of a precise definition of *conserved segments* keeping

as criteria a "good" order similarity. More generally the set of genes belonging to a chromosome in one species and which orthologs belong to the same chromosome of another species constitute a *synteny*. Fig. 1 provides an example of conserved segment between human chromosome 12 and mouse chromosome 6.

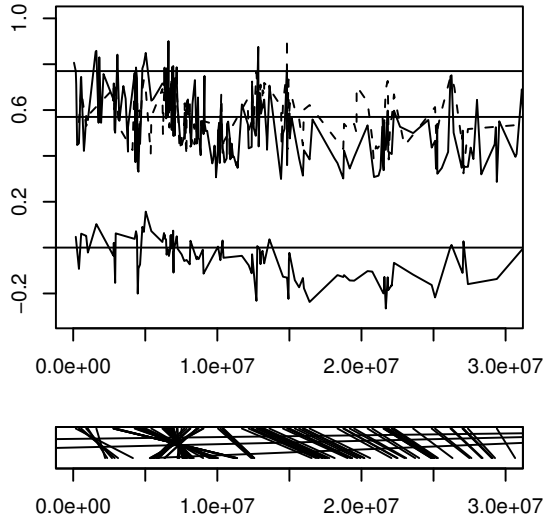


Figure 1: Synteny between human chromosome 12 and mouse chromosome 6

To relate genome map and biological processes, position of many biological informations must be compared. These informations could be statistical ones, as C+G content in third codon position of genes, but also could refers to gene expression (tissue in which the gene is expressed through EST or SAGE data) or evolutionnary process. In this last case the concept of orthology and paralogy is important. Due to duplications inside genomes sequences with great similarity could refer to different genes. More precisely orthologous genes are defined as genes in two different species such that the path linking them in the phylogenetic tree does not go through a duplication. Association of genome maps of two species is based upon association of orthologous genes. So it is of great importance to have a precise strategy to estimate orthologous pairs. Two strategies are used in litterature one is the double reciprocal best fit using blast, the second one use the analysis of the gene family phylogeny to directly verify the preceeding condition. Here we use the second approach, this can be largely automated by use of Hovergen data base [7] and the retrieval software which is able to select trees having specific patterns. The substitution process parameters (particularly K_S and K_A) can then be computed on orthologous pairs and mapped on one of the two genomes.

3.2 GemCore

GeM is a project that implies a collaboration between our laboratory and a computer scientist one (Helix, INRIA). The aim is to build a knowledge base devoted to comparative genomic mapping. In a first step [1] an UML modelling has been made as well as au GUI interface dedicated to graphical representation and request on human and mouse genome. To improve the efficiency of the software a translation in a relational date base has recently be made by V. Navratil using Postgres. This data base presently include human, mouse and rat genome data and is designed to be extended to take into account data from farm animals (pig, cow, chicken) as well as dog (for medical pupose). GeM is able to manage simultaneously all types of mapping. Moreover this new implementation include new data as:

- a direct link to sequence through the ACNUC sequence management system [8,9]
- statistical data (G+C content in each codon position, Ks between human and mouse)
- expression data (EST)
- polymorphism

Some parts of the UML modelling of GemCore is presented in the figure 2. The recursive definition of maps (due to the reflexive relation *mapsOn*) allows the inclusion of a smaller map inside a larger one, for example a local RH map inside a chromosome map. Sequence can be access through the ACNUC software (see http://pbil.univ-lyon1.fr/databases/acnuc/acnuc_gestion.html for implementation details). GenomicElement (for instance proteic genes) belongs to the set of their orthologous conterparts in the species present in

the base (OrthoGroup). These sets is then grouped inside HomolFam to represent all homologous relationships.

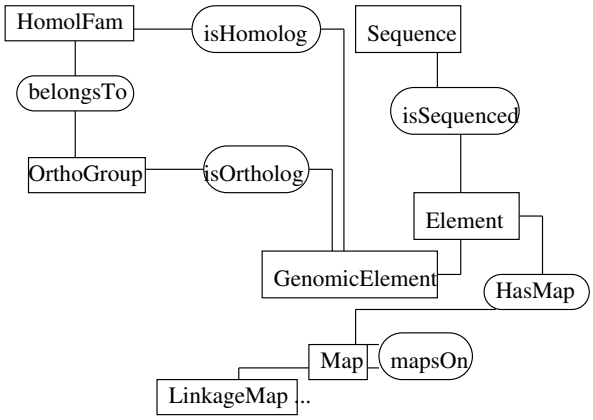


Figure 2: A part of the UML scheme of GemCore

A connection between GemCore and the R statistical software is under development using the Rdbi package. Fig. 1 results from these developments.

3.3 some regional patterns in vertebrates genomes

The great increase of available data on vertebrate genomes allows the study of the regional patterns of many genome characteristics (some related review can be founded in [10,11]. The regional pattern of nucleotide frequencies has been studied since 1976 [12] in the framework of isochore organisation. It will be presented in the next section. Substitution process is clearly involved in the generation and maintenance of such structure. The next section will summarize some results on its spatial patterns. Prokaryotic genes are organised in operon in wich gene have similar expression patterns. Works have been done to search for "operon like" eukaryotic stucture that is a clustering of genes having similar expression pattern. A short section will give some recent results from the litterature on this point. At last implications on genomic sequences of the physical structure of DNA (particularly its different folding level) has been studied since the begining of the genomic era. New data on human genome as well as the use of new methodologies have allowed new results that will be preesented in the last section.

3.3.1 isochores

The main pattern of vertebrate genome is isochore organisation which separate genome in a mosaic of regions whith a more or less homogeneous G+C content. Isochore have been discover by G. Bernardi using gradient density centrigugation [12]. Availability of genomic sequences have then allowed a much more precise description of this genome organisation. The main feature is that all type of genomic sequences is involved in isochore organisation. This results in a strong correlation between c+G content in exons, intons, intergenic regions (see for example the figure 2 of [13]). The best criteria to discriminate between isochore classes is $C+G_{III}$. In this paper we will consider that heavy (H) isochores are defined by $C+G_{III} > 77\%$ and light isochores are defined by $C+G_{III} < 57\%$. Other regions are defined as medium isochores (M). Properties of isochores are summarized here:

- density of genes is greater in H than L
- genes are larger in L than in H, this is particularly true for introns
- Lines ares more frequent in L isochores
- Sines (Alu, B1) are more frequent in H isochores

Isochores have been described mainly in mammals and birds. Figure 1 summarized some statistic features of isochores on a conserved segment between the human chromosome 12 and mouse chromosome 6. Positions of orthologous gene are link by segments in the botton of the graph. It can be clearly seen a large inversion near the telomeric end of human chromosome. The upper part of the graph is related to isochore organisation. The top solid line shows the G+C content in codons position III of human gene, the dash line is the coterpart for mouse genome. The two lines have been synchronised using interpolation between orthologous genes. The bottom line shows Δ = (human G+C_{III} - mouse G+C_{III}) for orthologous pairs. The right part show a large L isochore with negative Δ. This is characteristic of the *minor shift* : G+C content has a smaller variance in murids genome than in other mammals. In L isochores murids genome is

richer in G+C than other mammals and reciprocally murids genome has a lower G+C content than other mammals in H isochore.

Curiously the taxons where isochores was present were those in which homeothermy is known and this fact has taken an important role in proposals for the mechanisms imbedded in appearence and maintenance of isochores. Bernardi (see review in [14]) have argued for a natural selection process implying an adaptation to temperature. However the recent finding of isochores organisation in reptiles seems to ruled out the "homeothermy hypothesis" [15]. Presently no life trait or environmental factor can be related to a possible selective pressure for isochore organisation. Mutational bias provided a neutral alternative to natural selection. Two type of mutational bias have been proposed either a variation along the genome of mutational bias [16] or bias gene conversion [17]. Availability of polymorphism data is qickly increasing and seems to rules out mutational bias giving strong argument for bias gene conversion [18].

3.3.2 substitution process

It is clear that a natural selection pressure exists on nucleotides that determine the coded proteins. So it is necessary to eliminate this natural selection effect when studying relationships between substitution process and regional organisation of genomes. That may be done in taking into account only non coding region (like pseudogenes for instance) or inside coding region in using only silent substitutions. In this last case an index (K_S) of substitution has been defined to estimate the total number of substitution that have take place during evolution from the common ancester to the two considered species. Apart from some trivial processes K_S can be a property of the substitution process only if this process is stationary. A complete discussion of K_S is not in the scope of this paper and we just presented here some results:

- Matassi et al have studied the regional organisation of K_S between human and mouse [19]. This work predate the human genome sequencing and mouse genetic map have been used to localise genes. Analysis is based upon nonparametic correlation coefficient and simulations. The result was the existence of a regional organisation of K_S and its independance from the isochore organisation.
- Duret et al [13] have shown that G+C content is not at equilibrium and that the isochores organisation of mammalian genomes is vanishing.

3.3.3 expression pattern

Several studies proposes that genes having some similar expression pattern show a significant tendency to be linked in genome. So housekeeping genes in human [21], essentials genes in yeast [22] or similar expression breadth m Vanishing GC-Rich Isochores in Mammalian Genomes. ouse genes [23] made cluster in the genome. This is an important pattern suggesting that selection could apply on rearrangement to generate a suitable expression pattern. However it must be quoted that the large number of genes now available lead to statistically significant figures corresponding to very small feature. Correlation of less than 0.1 are often quoted for instance. That raised a methodological question that need to be more precicely examined.

4 conclusion

Many biological processes work on genome with a spatial component. Transcription, replication are clear examples. Their functionning imply constraints that are superimposed one on the other and that interact with genetic information. Taking into account the spatial component of observed pattern help to interpret them, procaryotic genomes give clear example of this approach.

Eucaryotic genomes undergo supplementary processes, particularly those liked to recombination. Variation of the recombination rate, existence of bias conversion raise promising hypothesis in the understanding of isochore organisation. It also imply to reinforced the link with population genetic, for example through the relationships between recombination and selection efficiency. In this context the tremendous increase of polymorphism data that will be generated particularly by EST data, can open very exciting new approaches.

In this context the availability of formal modelling of all this very different type of data and of the knowledge in the field of genomic mapping will be a crucial step to efficiently manage the tremendous data accumulation. It is the aim of the project GeM.

Moreover works must be done to develop mathematical tools capable of separate patterns having different scale. A good example is given by recent work on the analysis of sequence patterns linked to folding of eucaryotic DNA, particularly in relation with nucleosome. Wavelet analysis can focuse on some structure by choice of the used filter [24]. Markov modeling seems also to be a very promising tool to compare large regions uncluding different types of sequence (exon, intron, transposable elements, ...).

References

[1]Bronner G, Spataro B, Gautier C, Rechenmann F. (2000) GeMCore, a Knowledge Base Dedicated to Mapping Mammalian Genomes LNCS . 2066:12-23

[2] Sueoka N, Marmur J, Doty P: Heterogeneity in deoxyriboneucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. Nature 1959, 183:1427-1431.

[3]Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997 Jun;44(6):632-6.

[4] Lobry, JR, Lobry C (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol., 16, 719-723

[5] Picardeau M, Lobry JR, Hinnebusch BJ: Physical mapping of an origin of bidirectional replication at the center of the Borrelia burgdoferi linear chromosome. Mol Microbiol 1999, 32:437-445.

[6] Brewer BJ: When polymerases collide: replication and the transciptional organization of the E. coli chromosome.(1988) Cell, 53:679-686.

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

[8] Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. (1985) Comput Appl Biosci.1:167-172

[9] Perriere G, Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot J, Charavay C, Gouy M, Duret L, Deleage G. Integrated databanks access and sequence/structure analysis services at the PBIL.(2003) Nucleic Acids Res.31:3393-3399

[10] Gautier C (2000) Compositional bias in DNA. Current Opinion in Genetics and Development, 10, 656-661

[11] Duret L. Evolution of synonymous codon usage in metazoans. (2002) Curr Opin Genet Dev. 12:640-649

[12] Macaya, G., Thiery, J.P., Bernardi, G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level J. Mol. Biol. 108, 237254.

[13] Duret L., Semon M., Piganeau G., Mouchiroud D., Galtier N. (2002) Vanishing GC-Rich Isochores in Mammalian Genomes. Genetics, 162, 1837-1847

[14] Bernardi G. (2000) Isochores and the evolutionary genomics of vertebrates Gene, 241, 3-17

[15] Hughes, S, Zelus D, Mouchiroud D (1999) Warm-blooded isochore structure in Nile crocodile and turtle. Mol Biol Evol 16,1521-1527

[16] Wolfe KH, Sharp PM, Li WH (1989) Nature, 337, 283-285

[17] Eyre Walker A (1993) Proc R Soc Lond B 252, 237-243

[18] Smith NG, Eyre Walker A (2001) Mol Biol Evol 18,982-996

[19] Matassi G., Sharp P.M., Gautier C. (1999) Chromosomal location effects on gene sequence evolution in mammals. Current Biology , 9, 786-790

[20]Piganeau G., Mouchiroud D., Duret L., Gautier C. (2002) Expected relationship between the silent substitution rate and the GC content: Implication for the evolution of isochores. J. Mol. Evol., 54, 129-133

[21] Lercher M.J.,Urrutia M.O., Hurst L.D. (2002) Nature Genetics, 31, 180-183

[22]Pl1 C, Hurst L.D.(2003) Evidence for coevolution of gene order and recombination rate. nature genetics, 33,392-395

[23] Hurst L.D., Williams E.J.B. (2000) Gene, 261, 107-114

[24] Audit B., Vaillant C., Arneodo A., d'Aubenton-Carafa Y., Thermes C. (2002) J. Mol. Biol., 316, 903-918

Table des figures

Liste des tableaux

RESUME en français

Le séquençage à grande échelle a permis d'accéder aux génomes complets de nombreux organismes. Les modèles de Markov cachés sont une des méthodes probabilistes les plus utilisées pour l'analyse des séquences. L'objectif de ces travaux est de participer à l'analyse de l'organisation et à la compréhension de l'évolution des génomes. Une méthode de prédiction des isochores adaptée au génome humain a été développée. L'originalité de cette approche consiste à identifier des ruptures d'homogénéité des séquences mais surtout leurs causes biologiques, comme l'influence des régions UTRs lors du classement des gènes dans un isochore. Cette méthode a ensuite été appliquée au génome du tetraodon, pour lequel l'existence d'une structure en mosaïque nouvelle le long de son génome a été mise en évidence.

TITRE en anglais

Detecting and modeling isochores in genomes with hidden Markov models

RESUME en anglais

Large scale sequencing has given access to many complete genomes. Hidden Markov models are one of the most used statistical methods for the analysis of genomics sequences. The objective of this work is to contribute to the analysis and comprehension of the evolution of genomes. A method for predicting isochores adapted to the human genome was developed. The originality of this approach consists in identifying breaks of homogeneity in the sequences. It also consists in identifying their biological impact, such as the UTRs influence may have on the classification of genes in an isochore. This method was then applied to the genome of the tetraodon, when we discovered the existence of a new mosaic structure.

DISCIPLINE

Biologie

MOTS-CLES

HMM, isochore, genomes organisation

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

UMR5558 - Laboratoire de Biométrie et Biologie Evolutive
Batiment Gregor Mendel - Université Claude Bernard Lyon1
43, bv du 11 novembre 1918
69622 Villeurbanne