

MELO DE LIMA Christelle

27 rue professeur Patel 69009 Lyon

Téléphone : 04.78.25.48.40

e-mail : melo@biomserv.univ-lyon1.fr

page personnelle : [http ://melodelima.chez-alice.fr/cv.html](http://melodelima.chez-alice.fr/cv.html)

Française

Née le 1er Juillet 1978

Mariée, un enfant

Docteur de l'Université Claude Bernard Lyon 1 Spécialité Analyse et modélisation des systèmes biologiques

Postes occupés

2005-2007	Attachée Temporaire d'Enseignement et de Recherche (ATER) à l'Université Lyon 2 et Lyon 1. <i>Activités d'enseignement (196 heures équivalent TD) :</i> - à l'Université Lyon 1 - à l'Université Lyon 2 <i>Activité de recherche :</i> - au Laboratoire de Biométrie et Biologie Évolutive UMR CNRS 5558
-----------	--

Formation universitaire

2002-2005	Doctorat de l'Université Claude Bernard Lyon 1 sous la direction de Monsieur Didier Piau (Institut Camille Jordan UMR CNRS 5208) et Monsieur François Rechenmann (INRIA). Titre de la thèse : <i>Développement d'une approche markovienne pour l'analyse de l'organisation des génomes.</i> Financement : bourse MENRT. Thèse soutenue le 28 Novembre 2005.
2001-2002	Diplôme d'Études approfondies d'Analyse et modélisation des systèmes biologiques, Université Claude Bernard Lyon 1. Mention Bien.
2000-2001	Maîtrise de mathématiques, mention ingénierie mathématiques, Université Claude Bernard Lyon 1. Mention Assez Bien.
1999-2000	Licence de mathématiques, Université Claude Bernard Lyon 1. Mention Assez Bien.
1996-1998	Diplôme d'Études Universitaires Générales de mathématiques, Université Claude Bernard Lyon 1.

Activités scientifiques

5 Publications dans des revues internationales à comité de lecture.
3 Publications soumises dans des revues internationales à comité de lecture.
8 Communications dont 2 invités dans des congrès internationaux.
Rapporteur pour la revue internationale *Algorithmica*.
Co-encadrante d'un stage de Master 2 recherche.

Activités d'enseignement

Depuis 09/2002	354 heures équivalent TD réparties entre : - l'Université Claude Bernard en Master (M1) et en Licence (L1, L2 et L3). - l'École Centrale Lyon (équivalent L3). - l'INSA Lyon (équivalent M1). - l'IUT de l'Université Lyon 2 (équivalent L1 et L2).
Disciplines enseignées	Probabilités, Statistique, Analyse, Algèbre, Bioinformatique.

Plan du curriculum vitae

Activité de recherche	page 3
Publications	page 4
Conférences	page 5
Valorisation scientifique	page 7
Travail de thèse	page 9
Activité de recherche postdoctorale	page 15
Activités d'enseignement	page 17

2005-2007 **Activité de recherche postdoctorale** en collaboration avec Christian Gautier et Dominique Mouchiroud, au laboratoire de Biométrie et Biologie Évolutive (UMR CNRS 5558)

Évolution de la structure en isochores chez les poissons.

Réalisations :

- **Analyse comparative** des génomes du chimpanzé (publié dans Lecture Notes in Bioinformatics), de la souris et du poulet.
- **Mise en évidence d'une nouvelle structure dans le génome du Tetraodon, du zebrafish** (soumis à BMC Bioinformatics) **et du fugu** (soumis à Journal of Molecular Evolution)

2002-2005 **Travail de thèse** à l'université Claude Bernard Lyon 1, sous la direction de Didier Piau, Professeur au laboratoire de mathématiques Institut Camille Jordan (UMR CNRS 5208) et de François Rechenmann, Directeur de recherche à l'INRIA.

Réalisations :

- **Développement de modèles probabilistes markoviens pour l'exploration à grande échelle des génomes** à partir de l'utilisation d'états complexes (publié dans Journal of Mathematical Biology).
- Utilisation de méthodes statistiques (analyse multivariée, etc.) et sélection de modèles de Markov cachés pour la discrimination des différentes régions du génome (exons, introns, peptides signaux, etc.) ou de différents génomes entre eux (publié dans International Journal of Foundation of Computer Sciences).
- **Analyse et prédiction des isochores le long du génome humain** (publié dans Gene).

2001-2002 **Stage de DEA** au sein du laboratoire de Biométrie et Biologie Évolutive UMR CNRS 5558, Université Claude Bernard Lyon 1, sous la direction de Christian Gautier.

Réalisation :

- **Développement d'un modèle probabiliste** basé sur les chaînes de Markov cachées pour la localisation des gènes le long du génome humain (publié dans les actes de JOBIM 2004).

Publications dans des revues internationales à comité de lecture

- P5 **Melodelima C.**, Gautier C. and Piau D., A markovian approach for the segmentation of the mouse genome. Soumis à *Journal of Mathematical Biology* in press.
- P4 **Melodelima C.**, Guéguen L., Gautier C. and Piau D., Analysis of the structure of genes using a hidden Markov model. Soumis à *International Journal of Foundations of Computer Science in press*.
- P3 **Melodelima C.** and Gautier C., Segmentation of the chimpanzee genome using a HMM model. *Lecture Notes in Bioinformatics*, 2007 ;4414 :251-262.
- P2 **Melodelima C.**, Guéguen L., Piau D., and Gautier C., A computational prediction of isochores based on hidden Markov models. *Gene*, 2006 ;385 :41-49.
- P1 Gautier C., Navratil V., and **Melodelima C.**, Genomic mapping and molecular processes, *Proceedings of Mathematical and Computational Biology*, 2003 ;2 :22-28.

Publications soumises dans des revues internationales à comité de lecture

- S3 **Melodelima C.** Mouchiroud D., and Gautier C. , Segmentation of fish genomes. Soumis à *BMC Bioinformatics*.
- S2 **Melodelima C.**, Mouchiroud D. and Gautier C., Isochore structure in fish genomes. Soumis à *Molecular Biology and Evolution*.
- S1 **Melodelima C.**, Gautier C. and Piau D., A hidden Markov model to analyse the chicken genome. Soumis à *Mathematical Biosciences*.

Communications invitées dans des congrès internationaux

- Ci2 **Melodelima C.** and Gautier C., Detection and modeling of isochores in mammals using hidden Markov models. Invited *Conference at Physics and Biology*, Wroclaw, Pologne, 2005.
- Ci1 Gautier C., Navratil V., and **Melodelima C.**, Genomic information along the genome : modeling, statistical analysis and evolutionary implications. Invited conference at *IFIP World Computer Congress*, Toulouse, France, 2004 :258-261.

Communications publiées dans des actes de congrès à comité de lecture

Congrès internationaux

- C6 **Melodelima C.** and Gautier C., Segmentation of the chimpanzee genome using a HMM models. *Bioinformatics Research and Development*, Berlin, 2007, *in press*.
- C4 **Melodelima C.**, Guéguen L., Gautier C. and Piau D., Analysis of the structure of genes using hidden Markov model. *The Prague Stringology Conference '06*, Prague, République Tchèque, 2006 :123-136.

Congrès nationaux

- Cn4 **Melodelima C.**, Guéguen L., Piau D. and Gautier C., Prediction of human isochores using a hidden Markov model. *Journées ouvertures à la biologie, l'informatique et les mathématiques*, Lyon, 2005 :427-434.
- Cn2 **Melodelima C.**, Guéguen L., Piau D. and Gautier C., Modelling the length distribution of exons by sums of geometric laws. Analysis of the structure of genes and G+C influence. *Journées ouvertures à la biologie, l'informatique et les mathématiques*, Montréal, 2004 :142-154.

Congrès internationaux

- C5 **Melodelima C.**, Mouchiroud D. and Gautier C., The mosaic genome of fishes *Groupe De Recherche Européen - Arc Rhône Alpin Comparative genomics*, Lyon, 2006.
- C3 **Melodelima C.**, Guéguen L., Piau D., and Gautier C., A computational prediction of isochores based on hidden Markov models, *European Conference on Computational Biology*, Madrid, Espagne, 2005.
- C2 Gautier C., Navratil V. and **Melodelima C.**, Genomic mapping and genome functioning, *Colloque Robert Cedergren*, Montréal, Canada, 2004.
- C1 **Melodelima C.**, Guéguen L., Gautier C. and Piau D., A markovian approach to the analysis of the structure of genes, *European Conference on Computational Biology*, Paris, France, 2003.

Congrès nationaux

- Cn3 Gautier C., Navratil V. and **Melodelima C.**, *Comparative genomic mapping*, Séminaire Agenae, Saint Malo, 2005.
- Cn1 **Melodelima C.**, Piau D. et Gautier C., Analyse de la structure du génome humain par une approche markovienne. 15èmes Rencontres Régionales pour la Recherche, Lyon, 2004.

Valorisation scientifique

Séminaires invités

- Mai 2005 Prédiction et analyse de la structure des génomes par des modèles de Markov cachés. *Séminaire de statistique et modélisation stochastique de Grenoble*. INRIA et Université Joseph Fourier Grenoble 1, invitée par Florence Forbes (chargé de recherche).
- Avril 2005 Prédiction des isochores du génome humain par des modèles de Markov cachés. *Groupe de travail Statistique des séquences biologiques*. Université d'Evry, invitée par Anne-Sophie Tocquet (maître de conférence).
- Juin 2004 Modélisation markovienne de la structure des gènes. Analyse de la structure en isochores. *Groupe de travail Statistique des séquences biologiques*. Université d'Evry, invitée par Sophie Schbath (directeur de recherche).
- Mai 2004 Modélisation markovienne de la structure des gènes : exons de lois non géométriques et influence du taux de G+C. *Laboratoire de probabilités*. Université Claude Bernard Lyon 1, invitée par Jean Bérard (maître de conférence).
- Avril 2004 Modélisation markovienne de la structure des gènes : exons de lois non géométriques et influence du taux de G+C. *Groupe de travail Modélisation de la Biologie*. INSA de Lyon, invitée par Olivier Mazet (maître de conférence).

Groupe de travail

- Juin 2004 Adaptation de l'algorithme de Forward-Backward dans le cas des HMMs à macro-états, *Groupe de travail ADH* animé par Marie-France Sagot (directeur de recherche), Université Claude Bernard Lyon 1.

Expertise scientifique

Rapporteur pour la revue internationale à comité de lecture *Algorithmica* (depuis 2006).

Encadrement

Co-encadrement à 50% du stage de Master 2 recherche d'Alexandra Popa en collaboration avec Christian Gautier "Influence de la recombinaison sur la présence d'isochores chez les poissons".

Connaissances informatiques acquises

Environnements :	Windows, Unix/Linux, Macintosh.
Programmation :	C, C++, Python, Shell, awk, LaTeX, HTML, PostgreSQL.
Statistiques :	Packages R.
Calcul scientifique :	Matlab, Scilab.
Bioinformatiques :	Blast, RepeatMasker, ACNUC, Query, divers logiciels de prédiction de gènes et de motifs.
Centre calculs :	Utilisation des ressources du centre de calculs de l'IN2P3 : Parallélisation des calculs, gestion des tâches sous le système BQS.

Langues

Anglais :	Lu, écrit, parlé.
Allemand :	Notions.

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

Jury :

- Florence Forbes, Chargé de recherche, INRIA, Grenoble
- Christian Gautier, Professeur, Université Claude Bernard Lyon 1
- Didier Piau, Professeur, Université Claude Bernard Lyon 1
- François Rechenmann, Directeur de recherche, INRIA, Grenoble
- Sophie Schbath, Directeur de recherche, INRA, Jouy-en-Josas (rapporteur)
- Claude Thermes, Chargé de recherche, CNRS, Gif-sur-Yvette (rapporteur)

Soutenue le 28 Novembre 2005

Avec les multiples projets de séquençage de génomes, les biologistes ont à leur disposition une abondance de séquences d'ADN d'organismes différents. **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 purement 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 mathématiques et informatiques des séquences d'ADN ont ainsi été des voies de recherche privilégiées 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.

Parmi les éléments importants du génome, il existe le long des chromosomes de mammifères une forte variabilité de la composition en bases G+C. Les régions d'un génome dont la longueur est supérieure à 300 kilobases (kb) et dont la composition en base G+C est homogène sont appelées isochores. 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, la distribution des éléments transposables (fragments d'ADN qui ont la propriété remarquable de se déplacer d'un point à un autre du génome) ou le taux de recombinaison (qui mesure le brassage de l'information génétique). **La localisation et l'analyse des isochores présentent donc un grand intérêt pour la compréhension des génomes.**

L'objectif de ma thèse était de développer des modèles de Markov cachés pour la détection et l'analyse de l'organisation en isochores le long du génome humain. Les méthodes classiques [Isofinder (Oliver et al. 2004), Zcurve method (Zhang et al. 2003)] considèrent uniquement la composition en bases G+C pour localiser les isochores. L'originalité du travail de thèse a été d'intégrer plusieurs propriétés biologiques (comme les compositions en bases, la structure en codon des exons, les distributions des longueurs des exons et des introns, la densité en gènes) qui sont corrélées avec la structuration en isochores afin d'affiner les prédictions du modèle. Pour cela, une approche reposant sur une sélection de modèles HMMs gérant les longueurs des états par la construction d'états complexes a été mise en oeuvre pour permettre la discrimination des différentes régions d'un génome.

Modélisation des séquences d'ADN pour la prédiction d'isochores :

Modélisation des longueurs des différentes régions par un HMM

Dans cette étude, les gènes sont considérés comme composés de deux types de régions : les exons (partie codante) et les introns (partie non codante). Pour modéliser l'enchaînement de ces structures le long d'un gène, il existe principalement deux types d'approches markoviennes : certains algorithmes utilisent les chaînes de Markov cachées (comme HMMgene de Krogh 1997 ou VEIL de Henderson et al. 1997) et d'autres emploient des semi-chaînes de Markov cachées (comme Genscan de Burge et Karlin 1997 ou GenMark.hmm de Borodovsky et Lukashin 1998). Les semi-chaînes de Markov cachées représentent de façon précise la distribution de longueur des gènes mais le nombre de paramètres à estimer est très important. De plus, la complexité des algorithmes utilisés peut être quadratique par rapport à la longueur de la séquence. Ce facteur est particulièrement limitant pour l'analyse des données sur des génomes entiers qui représentent des milliards de paires de bases. Nous nous sommes donc orientés vers les modèles de Markov cachés (HMMs) car le nombre de paramètres à estimer est moindre et car ils permettent la mise en oeuvre d'algorithmes de complexité linéaire par rapport à la longueur de la séquence. Néanmoins, dans un HMM, les régions des séquences sont représentées par des états dont le temps de séjours doit suivre une loi géométrique. Cette contrainte est une limite lors de la modélisation des gènes car la distribution de longueur des exons n'est pas géométrique. Par conséquent, **des convolutions de lois géométriques de paramètres égaux ou différents ont été utilisés** (Figure 1). L'état représentant la région est ainsi remplacé par une juxtaposition de n états qui ont les mêmes probabilités d'émission [P5, Ci2]. Ce regroupement d'états est nommé **macro-état**, (Figure 2) [Cn2].

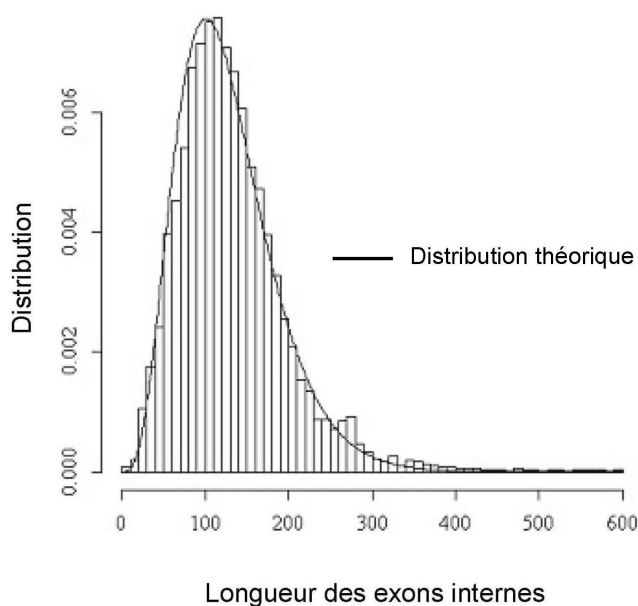


FIG. 1 – : L'histogramme représente la distribution empirique de la longueur des exons internes. La courbe représente la loi théorique obtenue par l'estimation de la distance de Kolmogorov-Smirnov.

Parmi les familles de lois de probabilité, la loi qui se rapproche au mieux de la distribution observée sur un échantillon a été déterminée par l'estimation de la distance de Kolmogorov-Smirnov (Figure 2). L'estimation au maximum de vraisemblance est une méthode souvent proposée dans la littérature mais qui n'a pas été retenue dans cette étude car elle tend à surestimer les distributions des longueurs des grands exons et à sous-estimer de nombreux petits exons autour de la médiane de l'histogramme. **L'estimation à partir de la distance de Kolmogorov-Smirnov permet de modéliser de façon plus précise une majorité d'exons** (Figure 3). Cette méthode a été appliquée aux différents types d'exons et d'introns afin de modéliser leur distribution de longueurs et a permis de prendre en compte, dans les HMMs, les propriétés biologiques suivantes : longueur des exons et des introns

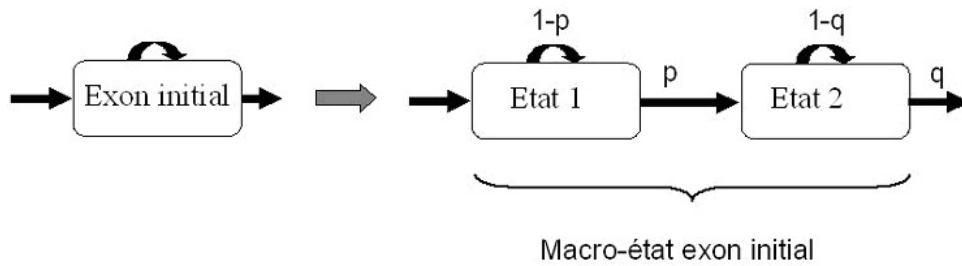


FIG. 2 – : Représentation du macro-état exon initial est composé d’une juxtaposition de deux lois géométrique de paramètres respectifs p et q .

(Melodelima et al. Gene 2006). De plus, la modélisation mise en place a permis une analyse fine des différentes régions constituant un gène (exons initiaux, exons internes, introns ...), en permettant par exemple, l’identifications de nombreux pseudogènes mal annotés dans les banques de données ainsi que l’importance de la modélisation du peptide signal en début de premier exon (Soumis à International Journal of Foundation of Computer Sciences).

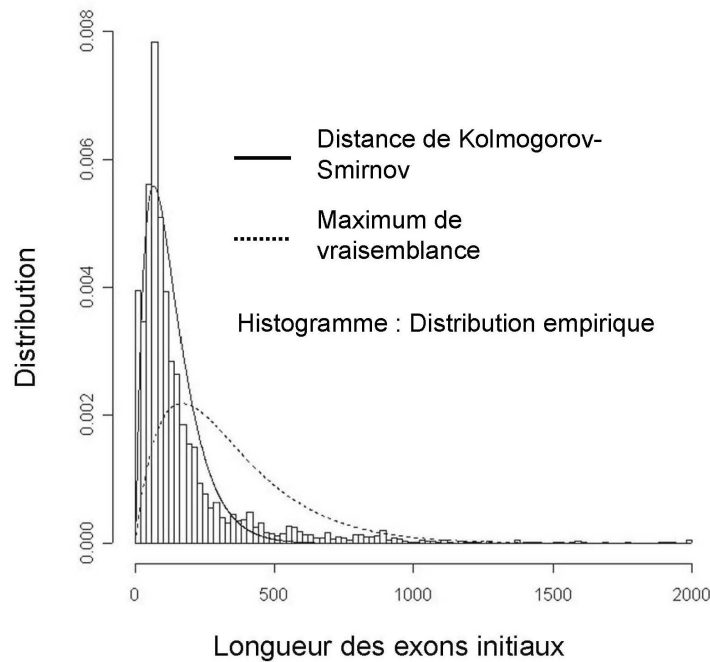


FIG. 3 – : L’histogramme représente la distribution empirique de la longueur des exons initiaux, la courbe pleine caractérise la loi obtenue par l’estimation à partir de la distance de Kolmogorov-Smirnov et la courbe en pointillés décrit la loi obtenue par la méthode du maximum de vraisemblance.

*Algorithme de reconstruction du chemin optimal des **macro-états cachés***

Lorsque les paramètres d’un HMM sont connus et la séquence observée donnée, il es nécessaire de reconstruire le chemin des états associés à cette séquence. Ce chemin représente les successions des différentes régions constituant la séquence. La difficulté réside dans le nombre important de chemins possibles, ce qui rend difficile voire impossible l’évaluation des chemins un à un. Il est donc nécessaire d’utiliser des algorithmes de reconstruction du chemin optimum d’états associés à une séquence observée. Les algorithmes les plus fréquents sont : (i) l’algorithme de Viterbi qui reconstruit la séquence d’états qui a la probabilité maximale de générer la séquence et (ii) l’algorithme de Forward-Backward qui choisit pour chaque base de la séquence, l’état qui a la plus forte probabilité d’être obtenue, connaissant la séquence et indépendamment des états qui le précèdent ou lui succèdent.

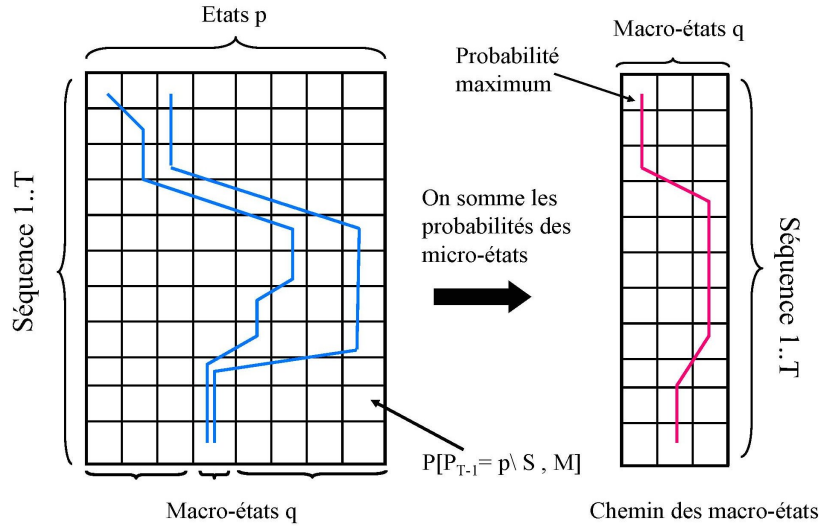


FIG. 4 – Amélioration de l’algorithme de Forward-Backward.

Cependant, l’algorithme de Viterbi néglige le fait que plusieurs trajectoires peuvent être équivalentes. L’algorithme de Forward, quant à lui, somme les 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. **La solution proposée a été de modifier l’algorithme de Forward-backward pour 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 4) [P4].

Prédiction des isochores le long du génome humain

La modélisation des gènes précédemment décrite a donc été utilisée pour caractériser les isochores le long du génome humain en prenant en compte, en plus du taux en G+C, les propriétés biologiques suivantes : longueurs des exons et des introns, densité en gènes, nombre d’exons par gènes et composition en bases qui varient en fonction du G+C. Pour détecter les isochores, trois modèles HMMs (H , L et M) ont été adaptés représentant respectivement les isochores dont le taux en G+C est fort (isochore H), faible (isochore L) et moyen (isochore M). Cette décomposition en trois classes est celle la plus fréquemment proposée dans la littérature (Mouchiroud et al., 1991). Le type d’isochore auquel appartient la séquence correspond au modèle ayant la plus forte probabilité $P(m|S)$, où m appartient à *Model*. La formule de Bayes permet d’obtenir la probabilité *a posteriori* de chaque modèle conditionnellement à la séquence observée S [P2] :

$$P(m|S) = \frac{P(S|m) \times P(m)}{\sum_{m' \in \text{Model}} P(S|m') \times P(m')}.$$

La prédiction des isochores le long du génome humain se décompose ensuite en deux étapes :

1. Les chromosomes sont découpés en fenêtres glissantes de 100 kb avec un chevauchement de 50 kb. Il s’agit de la dimension la plus couramment utilisée dans la littérature (Bernardi 1995, Pavlicek et al. 2002). Elle laisse supposer la présence de gènes à l’intérieur de la plupart des fenêtres.
2. Sur chaque fenêtre, la probabilité $P[m|S]$ est calculée (approche bayésienne) 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 300 kb. Par cette méthode, moins de 5% des fenêtres sont isolées et ces 5% sont réparties aléatoirement. Les probabilités *a priori* de chaque modèle ont été estimées à partir de la fréquence des gènes suivant leur contenu en G+C.

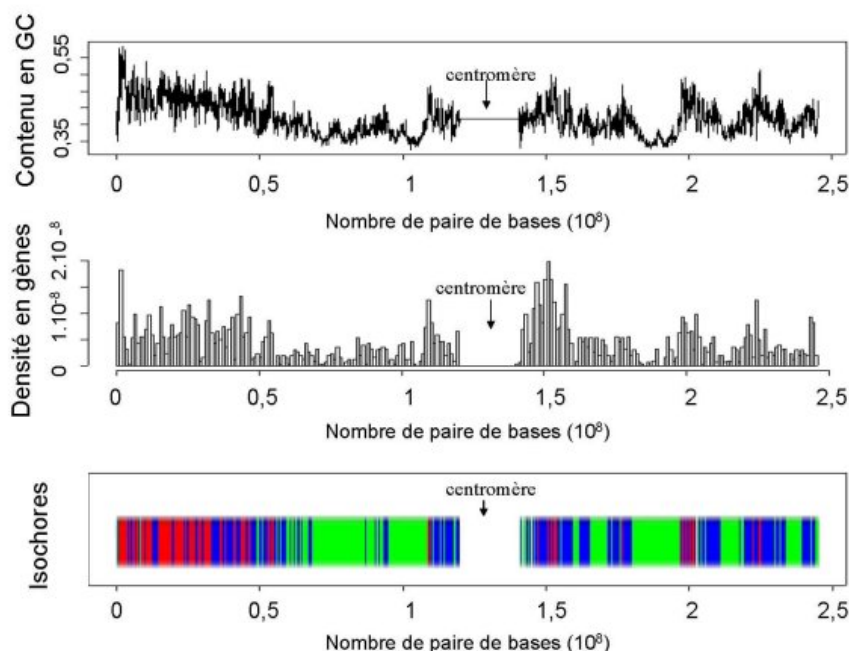


FIG. 5 – La figure du haut représente l'évolution du taux en G+C le long du chromosome 1 humain. La figure du milieu représente la répartition des gènes le long du chromosome 1 humain. La figure du bas représente la répartition des isochores le long du chromosome 1 humain, prédite par notre modèle. En rouge sont représentés les isochores H, en vert les isochores L et en bleu des isochores M.

Impacts biologiques et interprétations

La dernière partie du travail de thèse a été d'interpréter les données obtenues du point de vue biologique. Ainsi, **les modèles développés durant ce travail de thèse produisent une nette segmentation des chromosomes humains en isochores et prennent en compte les diverses propriétés biologiques qui leurs sont liées** (Figure 6). La segmentation obtenue est donc plus précise que les algorithmes classiques, qui se basent uniquement sur le contenu en G+C. De plus, si notre méthode a bien permis de mettre en évidence des ruptures d'homogénéité le long des séquences, elle a également permis d'identifier leurs causes biologiques. Ainsi, une étude complémentaire basée sur une sélection de modèles de Markov cachés a montré que certaines propriétés biologiques interviennent pour classer un gène dans un isochore ou dans un autre, indépendamment du contenu en G+C de la séquence.[Cn4, C4]

Toutefois, **le résultat majeur de cette étude concerne les régions UTRs**. Ce sont des régions non traduites des ARN messagers (ARNm). Les ARNm ont une courte région non traduite dans la direction de l'extrémité 5' du brin, appelée 5'UTR, et une région plus longue non traduite dans la direction de l'extrémité 3' du brin, appelée 3'UTR. La région 3'UTR contient, entre autres, des signaux régulant la stabilité des ARN messagers. Les études conduites ont montré que les régions UTRs possèdent une structure particulière qui dépend du type d'isochore dans lequel elles se situent, c'est-à-dire qu'elles reflètent mieux les caractères de l'isochore que le contenu en G+C des autres régions du gène. 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 serait 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) [P2, C3].

Enfin, même si l'évolution des isochores chez les vertébrés a suscité de nombreux travaux, 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, il a été possible **au cours du travail doctoral de fournir des cartes détaillées de la structure en isochores le long des génomes du chimpanzé, de la souris et du poulet. Une analyse de la compréhension et de l'évolution de l'organisation en isochores de ces génomes, met nettement en évidence une relation entre l'organisation en isochores, la répartition et la structure des**

Références

- Bernardi G (1995) The human genome : organization and evolutionary history. *Ann. Rev. Genet.*, 29 : 445-476.
- Borodovsky M, Lukashin AV (1998) GeneMark.hmm, New solutions for gene finding. *Nucleic Acids Res.*, 26(4) : 1107-1115.
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol Biol.* 268(1) :78-94.
- Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN : a database of homologous vertebrate genes. *Nucleic Acids Res.*, 22(12) :2360-5.
- Henderson J, Salzberg S, Fasman KH (1997) Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, 4 : 127-141.
- 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.
- Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P : IsoFinder (2004) Computational prediction of isochores in genome sequences. *Nucleic Acids Res*, 32(Web Server issue) : W287-292.
- 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. Pavelicek A, Paces P, Clay O, Bernardi G. (2002) A compact view of isochores in the draft human genome sequence. *FEBS Lett.*, 511(1-3), 165-169.
- Zhang CT, Zhang R (2003) An isochore map of the human genome based on the Z curve method. *Gene*, 317(1-2) :127-135.

Évolution de la structure en isochores chez les poissons

Les activités de recherche conduites durant le travail postdoctoral ont représenté une **mobilité thématique vers la biologie**. Les modèles mathématiques et informatiques développés au cours des travaux de thèse ont été utilisés pour conduire des analyses et des interprétations biologiques sur l'évolution des isochores au sein des vertébrés et plus particulièrement chez les poissons actynoptéigiens.

Ainsi, s'il est bien connu que les génomes de mammifères et d'oiseaux possèdent une structure en isochores le long de leur génome, il était également prédit que les poissons, du fait de leur forte homogénéité en base G+C ne possédaient pas *a priori* de structure en isochores. En effet, cette structuration n'avait jusqu'à présent pas pu être observée ni expérimentalement ni par modélisation chez les génomes de poissons.

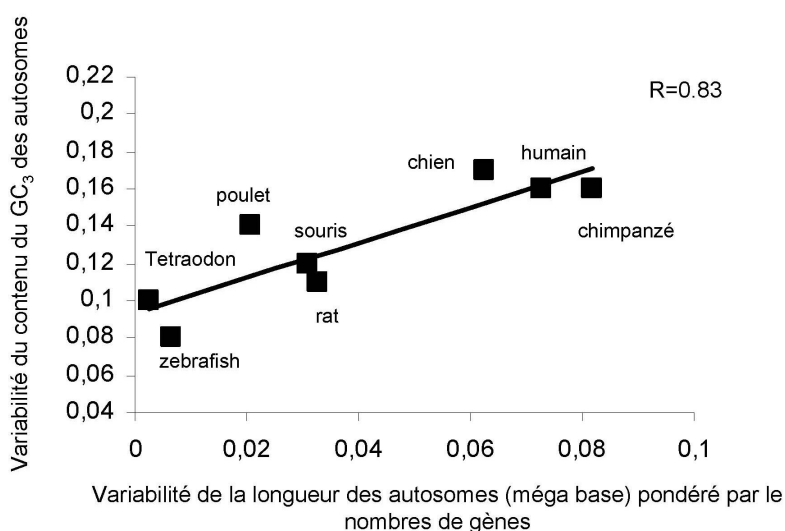


FIG. 6 – Corrélation entre la variabilité du taux en G+C₃ des autosomes et de la variabilité de taille de autosomes pondéré par le nombre de chromosomes. Une forte corrélation est observé avec R=0,83

Toutefois, le travail de recherche post-doctoral a permis d'établir la présence d'isochores chez les poissons et d'expliquer l'évolution des isochores chez les vertébrés.

La méthode de prédiction des isochores développée au cours de ma thèse a été appliquée à l'analyse de l'organisation des génomes de Tetraodon, du fugu et du zebrafish. **Notre étude a clairement mis en évidence l'existence d'une structure en G+C [S1], entièrement nouvelle, le long du génome du Tetraodon.** Ces résultats ont été obtenus par **une approche originale partant de l'hypothèse que les caractéristiques des gènes contenus dans les isochores du génome humain se retrouvent au sein des gènes orthologues des espèces supposées dépourvues d'isochores.** Des gènes orthologues sont des gènes d'espèces différentes qui dérivent d'un même gène ancestral et qui ont divergé à la suite d'un évènement de spéciation. Ils peuvent avoir la même fonction ou non. Cette organisation en G+C est liée aux différences de structures entre les gènes du Tetraodon orthologues aux gènes humains selon leur classe d'isochores. Nous démontrons ainsi qu'il est possible de retrouver, chez une espèce, à partir de gènes orthologues, des traces d'isochores qui ne sont pas

repérables par les autres méthodes utilisées jusqu'à présent [C5, S3]. Une étude similaire réalisée chez le zebrafish a également mis en évidence la présence d'isochores, cette fois principalement pauvre en G+C.

De plus, le travail postdoctoral a permis de montrer l'existence chez les poissons d'une corrélation entre variabilité de la longueur des bras chromosomiques et la variabilité du taux en G+C (Figure 6) [S2]). Une étude complémentaire a été menée pour vérifier si les corrélations précédentes sont également liées à la recombinaison ce qui permet d'étendre les résultats de Meunier et al. 2005, concernant l'homme et les primates, aux génomes de poissons.

Référence

- Melodelima C., Mouchiroud D., and Gautier C. , Segmentation of fish genomes. Soumis à BMC Bioinformatics.
- Melodelima C., Mouchiroud D. and Gautier, Isochore structure in fish genomes. Soumis à Journal of Molecular Evolution.
- Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 21(6) :984-90.

Activités d'enseignement

354 heures équivalent TD :

158 heures de 2002 à 2005 en tant que vacataire.

100 heures en 2005-2006 en tant qu'ATER.

96 heures en 2006-2007 en tant qu'ATER.

- A l'Université Lyon 1 :

Niveau	Filière	Matière	Description	Nombres d'heures	Années
M1	Approches Mathématiques Informatique du Vivant (AMIV)	Bioinformatique	Modèles de Markov	9h de TD 3h de TP ATER	2005-2006
M1	Ecologie Microbienne Evolution	Analyse et modélisation des systèmes	Modèles de Markov	9h de TD 6h de TP ATER	2005-2006
L3	Informatique	Probabilités	Modélisation probabiliste, variables aléatoires, Loi des grands nombres loi Normale estimation et tests	52h de TD Vacataire	2003-2004 2002-2003
L3	Mathématiques Informatique du Vivant (MIV)	Bioinformatique	Bases de Données HTML prédiction de gènes et motifs	9h de TP ATER	2005-2006
L2	Science de la vie	Statistique	Tests d'hypothèses, Test du Chi 2, analyse de variance, regression linéaire, régression multiple	45h de TD 12h de TD 9h de TT ATER et Vacataire	2005-2006 2004-2005 2003-2004
L2	Sciences de la vie	Bioinformatique	Bases de Données HTML prédiction de gènes et motifs	24h de TP ATER	2005-2006
L2	Sciences de la vie	Algèbre	Systèmes de vecteurs calculs matriciel, applications linéaires, changement de base, diagonalisation	9h de TD 9h de TT Vacataire	2003-2004
L1	Mathématiques Appliquées aux Sciences sociales MASS	Statistique	Statistique descriptive unidimensionnelle et bidimensionnelle régression linéaire corrélation de Spearman et de Kendall	9h de TD 40,5h de TD Vacataire	2002-2003

- A l'Ecole Centrale de Lyon :

Probabilités et Statistique (2004/2005) TD 32h, (Equivalent L3)

Modélisation probabiliste, Variables aléatoires, Loi des grands nombres, Théorème central limite, Statistique : estimation et tests.

- A l'INSA de Lyon :

Processus stochastiques (2005/2006) TD 8h, (Equivalent M1)

Modèles de Markov, Modèles de Markov cachés.

- A l'Université Lyon 2 :

Niveau	Filière	Matière	Description	Nombres d'heures	Années
IUT 1ère année équivalent L1	Statistique Traitement Informatique des données (STID)	Statistique	Table de contingence ANOVA Régression linéaire	21h de TD ATER	2005-2006
IUT 1ère année équivalent L1	STID	Probabilité	Variable aléatoire Loi usuelles discrètes discrètes et continues Loi normale	10h de TD ATER	2005-2006
IUT 1ère année équivalent L1	STID	Analyse	Continuité Dérivabilité Primitives	10h de TD ATER	2005-2006
IUT 2ère année équivalent L2	STID	Algèbre	Distance - Norme Diagonalisation et analyse factorielle Projection orthogonale et modèle linéaire	15h de CM 15h de TD ATER	2005-2006
IUT 1ère année équivalent L1	STID	Statistique	Réalisation enquête et traitement données pour l'entreprise Keolis	20h de TD ATER	2005-2006