Prédiction markovienne in silico des régions constantes et variables des lentivirus

Aurélia Quillon

N° d'ordre : 234-2006 Année 2006

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
6 décembre 2006

par
Aurélia Boissin
Quillon

TITRE :

Prédiction markovienne *in silico* des régions constantes et variables des lentivirus

Directeurs de thèse : Dr Caroline Leroux et Pr Didier Piau

Jury :	Elisabeth Gassiat	Rapporteur
	Manolo Gouy	Président
	Caroline Leroux	Directrice
	François Mallet	Examinateur
	Didier Piau	Directeur
	Bruno Torrésani	Rapporteur

N° d'ordre : 234-2006 Année 2006

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
6 décembre 2006

par
Aurélia Boissin
Quillon

TITRE :

Prédiction markovienne *in silico* des régions constantes et variables des lentivirus

Directeurs de thèse : Dr Caroline Leroux et Pr Didier Piau

Jury :	Elisabeth Gassiat	Rapporteur
	Manolo Gouy	Président
	Caroline Leroux	Directrice
	François Mallet	Examinateur
	Didier Piau	Directeur
	Bruno Torrésani	Rapporteur

UNIVERSITE CLAUDE BERNARD - LYON I

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J.F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur R. GARRONE
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur G. ANNAT
Secrétaire Général	M. G. GAY

SECTEUR SANTE

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Odontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : M. le Professeur J.P. SCHARFF
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J.C. AUGROS

A mes filles,
mes plus belles découvertes.

Merci...

Plus on partage, plus on possède...
Léonard Nimoy

J'ai vécu cette thèse comme une grande aventure scientifique et humaine. J'en sors aujourd'hui grandie et plus riche qu'avant. Je tiens à remercier ici toutes les personnes qui ont rendu cette aventure possible.

J'ai effectué ma thèse au sein de l'UMR 754 "Rétrovirus et Pathologie Comparée". Je remercie le Professeur Jean-François Mornex pour m'avoir accueillie dans son unité de recherche.

Je remercie mes directeurs de thèse Caroline Leroux et Didier Piau pour m'avoir fait confiance et avoir accepté de me confier cette thèse. Un immense merci à Caroline qui a relevé le défi d'encadrer une doctorante en mathématiques. Merci pour votre patience, votre attention, vos encouragements et pour m'avoir permis de découvrir l'univers étrange de la biologie. Je remercie également très sincèrement Didier de m'avoir guidée dans l'univers non moins étrange des mathématiques. Nos longues discussions m'ont été précieuses. Merci à vous deux. Sans vous, cette thèse n'existerait pas.

Elisabeth Gassiat et Bruno Torrésani ont accepté de rapporter sur cette thèse. Je tiens à leur exprimer toute ma reconnaissance pour l'intérêt qu'ils ont manifesté pour mon travail, pour leur lecture attentive de mon manuscrit et pour les remarques pertinentes qu'ils ont formulé. Je remercie également Manolo Gouy et François Mallet de m'avoir fait l'honneur de faire partie du jury.

J'adresse ma sincère gratitude aux membres du comité de pilotage pour les conseils et les remarques qu'ils m'ont adressé au cours de la thèse et qui m'ont permis d'avancer toujours un peu plus loin sans trop m'égarer.

J'ai une pensée pour les membres de l'Institut Camille Jordan que j'ai côtoyés pendant ces quelques années et qui, par leur ouverture aux biomathématiques, ont su m'apporter des remarques précieuses. Merci tout particulièrement à Anne, Jean, Gabriela, Frédérique, Jean-Baptiste... Une pensée toute spéciale pour Pierre qui m'a patiemment écouté lui exposer mes problèmes farfelus. Sa gentillesse, sa disponibilité et ses nombreuses suggestions ont grandement contribué au chapitre sur les motifs caractéristiques.

Cette thèse n'aurait sans doute pas été aussi agréable sans les formidables compagnons d'aventure qui m'ont accompagnée. Un immense merci aux membres de l'UMR 754 pour les passionnantes discussions que nous avons eu et notamment à Tim qui a eu la gentillesse de partager un peu de son immense sagesse et de son enthousiasme débordant avec moi. Merci à toutes celles qui ont su apporter la joie et la gaieté à l'UMR 754. Merci notamment à Momo, Kathy, Angela, Sylvie... pour tous ces moments de détente si importants. Je veux aussi remercier tout spécialement les membres, passés et présents, de l'équipe 3 qui ont partagé mon quotidien et même parfois mon bureau : Monia, Geneviève, Fabienne, Christophe, Joëlle, Florent, Séverine, José, Emilie, Barbara, Christine, Nicolas, Chiara... Enfin, merci à tous ceux que je n'ai pas cités ici mais sans qui cette aventure n'aurait pas été tout à fait la même.

Pour terminer ces remerciements, je voudrais dire à mes parents, à ma famille, combien leur soutien et leur amour m'ont été précieux pendant les bons et les mauvais jours de cette thèse. Merci à mes parents de m'avoir donné envie de toujours tout comprendre un peu plus, de m'avoir permis de faire ces études et d'être aujourd'hui docteur. Merci de m'avoir écoutée patiemment vous raconter mes étranges recherches et merci pour la fierté que j'ai vu dans vos yeux. J'ai une petite pensée particulière pour mon papa qui connaîtra bientôt le plaisir d'écrire ces lignes. Merci à mes filles Laura et Eloïse qui, par leur innocence et leurs sourires, m'ont permis de trouver l'énergie nécessaire pour continuer à avancer quand rien ne fonctionnait. Enfin, merci à mon mari

Olivier pour m'avoir supportée quand je faisais des bonds partout parce que j'avais trouvé LE résultat, pour m'avoir portée quand je n'y croyais plus, et pour tout ce que les mots ne peuvent pas exprimer. Merci...

I have yet to see any problem, however complicated, which, when you looked at it in the right way, did not become still more complicated.

Poul Anderson

Table des matières

Introduction

Les rétrovirus sont des virus enveloppés à ARN qui peuvent infecter un grand nombre de vertébrés comme les oiseaux, les moutons ou les hommes. Ils constituent la famille des *Retroviridae*. La particularité des rétrovirus est la transformation de leur génome ARN en ADN avant son intégration dans le génome de la cellule hôte. Cette transformation, appelée rétrotranscription, dépend d'une enzyme spécifique : la transcriptase inverse ou RT (Reverse Transcriptase). Le genre lentivirus est un des genres de la famille des *Retroviridae*. Les lentivirus sont responsables de maladies chroniques d'évolution lente chez les hommes et les animaux. Parmi le genre lentivirus, HIV (Human Immunodeficiency Virus) infecte les hommes, EIAV (Equine Infectious Anemia Virus) infecte les équidés, SIV (Simian Immunodeficiency Virus) infecte les singes, SRLV (Small Ruminant LentiVirus) infecte les petits ruminants (caprins et ovins), BIV (Bovine Immunodeficiency Virus) infecte les bovins et FIV (Feline Immunodeficiency Virus) infecte les félins.

Les lentivirus ont la particularité d'accumuler de nombreux changements génétiques au cours de leur réplication. La très grande instabilité génétique des lentivirus est répartie de façon hétérogène le long de leur génome. Le gène *env*, et plus particulièrement la partie qui code la glycoprotéine de surface (SU), est la région du génome qui accumule le plus grand nombre de mutations. Ces mutations sont en partie dues au manque de fidélité réplicative de la RT durant la rétrotranscription de l'ARN viral en ADN, à l'absence d'activité correctrice d'erreurs de la RT, au taux de réplication virale élevé et aux phénomènes de recombinaison qui surviennent dans les cellules coinfectées. Tous ces éléments ont pour conséquence le fait que chaque copie de génome

lentiviral est potentiellement différente de la copie de génome à partir de laquelle elle a été générée.

De façon intéressante, les mutations de la région du gène *env* codant la glycoprotéine de surface des lentivirus n'apparaissent pas de façon uniforme mais sont localisées dans des zones spécifiques, appelées régions variables (V). Ces régions variables sont séparées par des régions dites constantes (C) qui ne présentent pas, ou présentent peu, de variabilité génétique. En moyenne, selon le lentivirus considéré, entre 10% et 35% des acides aminés de la SU varient d'un isolat à l'autre et plus de 70% des acides aminés variables sont situés dans les régions variables. Des régions constantes et variables ont pu être définies chez tous les lentivirus (Burns *et al.*, 1993, Leroux *et al.*, 1997b, Modrow *et al.*, 1987, Valas *et al.*, 2000, Zheng *et al.*, 1997a, Suarez and Whetstone, 1995, Pancino *et al.*, 1993).

L'accumulation de mutations dans les régions variables peut être la conséquence d'un taux de mutations localement plus élevé ou de mutations survenant au même taux tout le long de la SU mais subissant des mécanismes de sélection permettant l'élimination de la plupart des variants dans les régions constantes ou bien d'une combinaison de ces deux phénomènes. Quoi qu'il en soit, la grande plasticité des génomes lentiviraux leur permet d'échapper efficacement au système immunitaire tout en gardant leur identité. La plupart des acides aminés codés par les régions variables se situent à l'extérieur de la glycoprotéine de surface, alors que les acides aminés codés par les régions constantes sont sur la partie interne de la SU. Cependant, la réplication virale fait intervenir des molécules uni-dimensionnelles et se déroule à un moment où l'information relative à la structure tri-dimensionnelle de la particule virale ne semble pas disponible. Ainsi, l'hypothèse selon laquelle les taux de mutations des régions constantes et variables seraient différents implique que cette différence soit due à des signaux spécifiques codés par la séquence nucléotidique du virus. Il est possible que ces éventuels signaux correspondent à des interactions entre la séquence virale et la RT.

Le but de mon projet de recherche doctoral a été de déterminer s'il existe des signatures spécifiques des régions constantes et des régions variables qui pourraient expliquer l'accumulation de mutations dans les régions variables.

Des modèles mathématiques, basés sur les séquences des lentivirus et capables de différencier et de caractériser ces deux types de régions, ont été développés. La méthodologie employée est basée sur les modèles de Markov cachés. Ces modèles ont souvent été utilisés pour étudier des séquences d'ADN, en particulier pour analyser statistiquement les différences de composition le long des séquences. Ce sont des outils destinés à décrire l'hétérogénéité des séquences. Ils permettent de décomposer une séquence en une succession de régions homogènes sans connaître *a priori* la taille, la position ni la composition de ces régions. Dans les modèles de Markov cachés, chaque type de région est caractérisé par sa composition statistique en mots de nucléotides ou en mots d'acides aminés et la succession des régions homogènes est modélisée par une chaîne de Markov inobservable, appelée la chaîne cachée. Les modèles de Markov cachés ont été initialement introduits dans le contexte de la reconnaissance du langage (Rabiner, 1989) mais sont maintenant des outils majeurs de l'analyse des génomes (Churchill, 1989, Krogh, 1994, Krogh *et al.*, 1994, 2001, Nicolas *et al.*, 2002, Peshin and Gelfand, 1999).

Des modèles de Markov cachés, basés sur des séquences du lentivirus EIAV, ont été développés. Ces modèles permettent de prédire convenablement les régions constantes et les régions variables de ce lentivirus. Ces résultats suggèrent qu'il existe une différence de composition en mots de nucléotides et d'acides aminés entre les régions C et V d'EIAV qui pourrait expliquer l'accumulation de mutations dans les régions variables. Les résultats obtenus pour le lentivirus EIAV ont ensuite été étendus aux lentivirus HIV, SIV et SRLV. En nous basant sur des séquences des SU de HIV, de SIV et de SRLV, nous avons développé des modèles de Markov cachés capables de prédire avec une grande précision les régions constantes et variables de chacun de ces lentivirus. Nous avons également mis au point un modèle commun à tous les lentivirus qui, en se basant uniquement sur des séquences de HIV, d'EIAV, de SIV et de SRLV, est capable de correctement prédire les régions C et V de l'ensemble de ces lentivirus mais également des lentivirus BIV et FIV. Ceci indique que les régions constantes, comme les régions variables, de tous ces lentivirus présentent des points communs. Des signatures caractéristiques de chaque type de région, c'est-à-dire des mots de nucléo-

tides ou d'acides aminés spécifiques des régions constantes et des régions variables, des différents lentivirus ont pu être extraites à partir des modèles de Markov cachés définis. L'étude des signatures spécifiques des régions C et V des lentivirus permettra de mieux comprendre les mécanismes à l'origine de l'accumulation de mutations dans les régions variables.

Chapitre 1

Le contexte biologique

Sommaire

Les lentivirus sont responsables de maladies chroniques et/ou dégénératives d'évolution lente qui peuvent infecter les hommes ou les animaux. Afin de pouvoir combattre ces virus, il est nécessaire de mieux les comprendre. L'analyse de l'information contenue dans les génomes des lentivirus participe à une meilleure connaissance des infections dont ils sont la cause.

Bien que la génétique en tant qu'étude de l'hérédité remonte au XVII^{e} siècle, la mise en évidence de l'acide désoxyribonucléique, ou ADN, comme support de l'information génétique ne date que de 1944. La structure en double hélice de l'ADN a été élucidée en 1953 par Watson et Crick (Watson and Crick, 1953). Ce n'est que dans les années 60 avec la découverte de l'acide ribonucléique, ou ARN, par Jacob et Monod (Jacob and Monod, 1961) et du code génétique par Nirenberg et Matthaei (Nirenberg *et al.*, 1963), puis dans les années 70 avec la découverte de procédés permettant le clonage de séquences d'ADN que la génétique moderne a connu son véritable essor. A l'heure actuelle, les avancées de la génétique ont permis notamment le séquençage complet de près de 300 génomes dont celui de l'homme en 2000.

Après avoir décrit brièvement les différents éléments support de l'information génétique, nous présenterons les lentivirus. Nous décrirons en détail un lentivirus en particulier : le virus EIAV (Equine Infectious Anemia Virus) responsable de l'anémie infectieuse équine.

1.1 Notions de génétique

1.1.1 La cellule

La cellule est l'unité élémentaire de tout être vivant. Le nombre de cellules peut être très variable. Il existe des organismes unicellulaires comme les bactéries ou les levures et des organismes pluricellulaires, constitués de milliers de milliards de cellules, tel l'homme composé de 10^{14} cellules.

Les cellules peuvent être classées selon leur organisation en deux empires (Cavalier-Smith, 2004) : les cellules *procaryotes* et les cellules *eucaryotes*. Les cellules procaryotes sont relativement simples et ne contiennent qu'un seul compartiment dans lequel se trouvent tous les constituants de la cellule dont la molécule d'ADN. La taille des cellules procaryotes est de l'ordre de 1 à 10 μm. Les cellules eucaryotes sont en général plus grandes que les cellules procaryotes et ont une taille de l'ordre de 10 à 100 μm. Elles sont organisées en plusieurs compartiments : la membrane qui isole du milieu, le cytoplasme qui contient les organites et le noyau qui contient l'ADN.

Tous les organismes pluricellulaires sont constitués de cellules eucaryotes. Les organismes unicellulaires peuvent être quant à eux procaryotes comme les bactéries ou eucaryotes comme les levures.

1.1.2 Les acides nucléiques

L'acide désoxyribonucléique

L'acide désoxyribonucléique, ou ADN, est une longue molécule que l'on retrouve dans la plupart des organismes vivants. C'est le support de l'information génétique. La molécule d'ADN possède une structure en forme de double hélice découverte en 1953 par James Dewey Watson et Francis Crick (Watson and Crick, 1953). Elle est formée de deux brins parallèles, chacun étant une chaîne linéaire et orientée de nucléotides. Les nucléotides, ou bases, sont au nombre de quatre : adénine (A), cytosine (C), guanine (G) et thymine (T). Ils se succèdent pour former l'ADN. Une séquence d'ADN peut ainsi être vue comme un enchaînement de lettres prises dans un alphabet à 4 lettres $\mathcal{A} = \{A, C, G, T\}$. Pour des raisons chimiques, la molécule d'ADN est orientée, l'attachement des nucléotides ne se faisant que dans un sens : de 5' vers 3'.

Les nucléotides sont complémentaires et peuvent être associés deux à deux. La règle d'appariement entre les nucléotides est la suivante :

- l'adénine ne peut se lier qu'avec la thymine ;
- la cytosine ne peut se lier qu'avec la guanine.

L'existence de ces liens fait qu'un brin d'ADN s'associe avec le brin complémentaire inverse où chaque nucléotide est lié au nucléotide complémentaire (Figure 1.1). Les deux brins de la molécule d'ADN s'enroulent ensuite et présentent une conformation spatiale en double hélice.

L'acide ribonucléique

L'acide ribonucléique, ou ARN, est une macromolécule linéaire assez similaire à l'ADN. Cependant, contrairement à l'ADN, l'ARN ne possède qu'un seul brin et ne s'enroule pas en double hélice. Comme l'ADN, l'ARN est une succession de nucléotides où la thymine (T) est remplacée par l'uracile

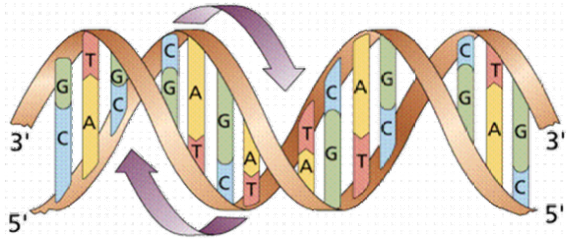

FIG. 1.1 – *Structure schématique de l'ADN*

(U). L'ARN est le messager de l'information génétique codée dans l'ADN. Il joue un rôle majeur dans les opérations de synthèse des protéines à partir de l'information codée dans l'ADN.

Il existe différents types de molécules d'ARN qui assurent chacun une fonction particulière dans la messagerie de l'information génétique. Les principaux types d'ARN sont :

L'ARN messager, ou ARNm : il se forme au contact de l'ADN et son rôle consiste à transcrire une séquence d'ADN puis à transporter l'information génétique recueillie du noyau vers le cytoplasme. L'ARN messager va ensuite se placer sur une unité d'assemblage des protéines, le ribosome, où il sera traduit pour élaborer une séquence d'acides aminés composant les protéines.

L'ARN de transfert, ou ARNt : les ARNt sont des molécules qui se placent sur les sites du ribosome où va être lu l'ARN messager. Ils ont un rôle fondamental lors de la synthèse des protéines : celui de "traduire" les codons (triplet de lettres) de l'ARNm en acides aminés. Les ARNt sont de courtes chaînes d'ARN capables de reconnaître un codon et qui portent à leur extrémité l'acide aminé associé. Il existe 61 sortes d'ARNt, une pour chaque codon qui code un acide aminé.

L'ARN ribosomal, ou ARNr : il représente 80 % de l'ARN total d'une cellule. Associé à des protéines, il forme le ribosome qui constitue la tête de lecture de l'information génétique transcrite par l'ARN messager. C'est dans le ribosome que sont enchaînées les séquences d'acides aminés qui constituent les molécules de protéines.

L'ARNm et l'ARNt interviennent dans la traduction de l'information génétique en protéines : l'ARNm transporte l'information portée par l'ADN et l'ARNt utilise cette information pour fabriquer les protéines.

1.1.3 Les protéines

Les protéines ont été découvertes en 1838 par le chimiste hollandais Gerhard Mulder. Elles sont composées d'un enchaînement d'acides aminés déterminé en fonction de l'information génétique présente dans les gènes. Leur synthèse se fait en deux étapes : la transcription, où l'ADN codant le gène associé à la protéine est transcrit en ARN messager, et la traduction, où l'ARN messager est traduit en protéine en fonction du code génétique. L'assemblage d'une protéine se fait acide aminé par acide aminé.

Les protéines sont indispensables à la vie, assurant des fonctions très diverses dans l'organisme :

- transport : elles transportent d'autres molécules comme l'hémoglobine qui sert à transporter l'oxygène des poumons aux organes ;
- hormonales : elles transmettent des messages à travers l'organisme, comme l'insuline ou l'adrénaline ;
- structure : elles donnent une forme aux cellules lorsqu'elles forment le cytosquelette et leur permettent de se mouvoir lorsqu'elles forment un flagelle ;
- enzymatiques : elles sont indispensables à la réalisation de la plupart des réactions biochimiques dans la cellule ;
- défense contre les micro-organismes , tels les anticorps qui détectent et neutralisent les agents pathogènes ;
- etc.

La fonction des protéines leur est conférée par leur structure tertiaire, c'est-à-dire la manière dont les acides aminés sont agencés les uns par rapport aux autres dans l'espace et, dans le cas de protéines formés de plusieurs chaînes polypeptidiques, leur structure quaternaire, c'est-à-dire la manière dont les différentes chaînes s'associent.

1.1.4 Les mécanismes liés à l'information génétique

Comme nous l'avons vu, la protéine est la résultante de l'information génétique contenue dans l'ADN. La synthèse des protéines à partir de l'ADN implique plusieurs mécanismes complexes : la réplication de l'ADN, la transcription et la traduction que nous allons décrire dans cette section.

1.1.4.1 La réplication de l'ADN

La réplication de l'information génétique résulte d'un dédoublement de l'ADN qui s'effectue lors de la division cellulaire. Elle permet de donner naissance à deux molécules filles identiques. Durant la réplication, les deux brins de la molécule d'ADN s'écartent. Un nouveau brin d'ADN est ensuite synthétisé face à chacun des deux brins grâce à l'incorporation, selon la règle de complémentarité des bases, des nucléotides qui sont dispersés dans le noyau. Chaque nouvelle molécule est identique à la molécule d'ADN initiale. Deux nouvelles molécules d'ADN sont ainsi construites, composées chacune d'un brin de l'ancienne molécule et d'un brin nouvellement formé. La figure 1.2 illustre le mécanisme de réplication de l'ADN.

Malgré la très grande fidélité de la réplication de l'ADN, il arrive que des erreurs apparaissent. Afin de permettre la conservation de l'information génétique, il existe des mécanismes de réparation de l'ADN. Un dommage sur l'un des brins d'ADN peut en général être réparé grâce à l'information portée par le brin complémentaire. Après détection et élimination du brin d'ADN défectueux par un système enzymatique complexe, la cellule synthétise, à l'aide d'une enzyme spécifique, l'ADN polymérase, un nouveau brin d'ADN en se servant comme matrice du brin d'ADN restant. Certaines erreurs ne sont cependant pas réparées et conduisent à des changements permanents de la séquence d'ADN, appelés mutations. Il existe plusieurs types de mutations :

- La substitution correspondant au remplacement d'un nucléotide par un autre ;
- L'insertion correspondant à l'ajout d'un ou plusieurs nucléotides ;
- La délétion correspondant à la suppression d'un ou plusieurs nucléotides ;

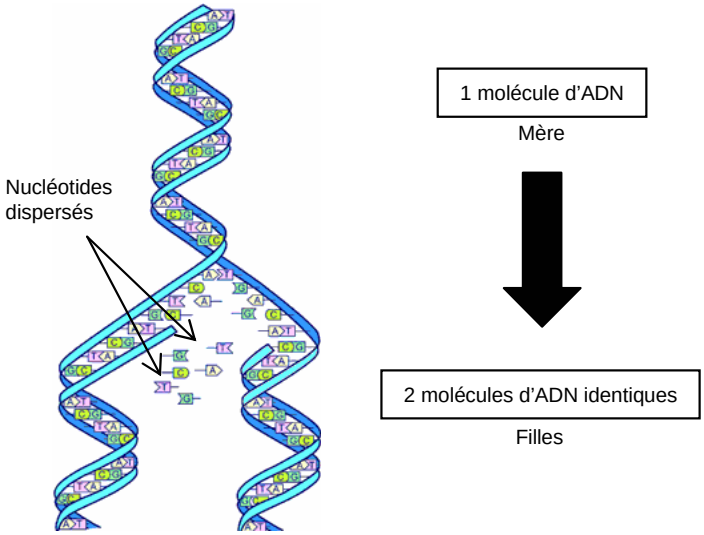

FIG. 1.2 – *Principe de la réplication de l'ADN (d'après "Acide désoxyribonucléique." Wikipédia, l'encyclopédie libre. http ://fr.wikipedia.org/wiki/ADN)*

Les mutations contribuent à la variabilité de l'information génétique et aux maladies.

1.1.4.2 La transcription

La transcription est la copie d'une molécule d'ADN en une molécule d'ARN. Le principe de la transcription est assez similaire à celui de la réplication de l'ADN. Comme lors de la réplication, les deux brins de la molécule d'ADN se séparent et un brin d'ARN est synthétisé selon le principe de complémentarité des bases en incorporant de l'uracile en place de la thymine. Chez les eucaryotes, la transcription de l'ADN en ARN a lieu dans le noyau. Les gènes des organismes eucaryotes sont constitués d'une alternance de régions codantes, appelées exons, et de régions non codantes, appelées introns. La molécule d'ARN directement synthétisée lors de la transcription à partir du modèle ADN reste dans le noyau et est traitée par un complexe enzymatique. Ce mécanisme s'appelle l'épissage : les introns sont excisés, les exons restant se reliant ensuite entre eux. L'ARN produit passe alors dans le

cytoplasme et devient ARN messager mature.

Il faut bien noter qu'un gène peut être associé à plusieurs protéines. L'épissage alternatif permet de créer plusieurs types de molécules d'ARN messagers différents, et donc plusieurs types de protéines différentes, à partir d'un même gène. Lors de l'épissage alternatif, seuls certains exons sont conservés pour former des ARN messagers matures qui seront traduits en protéines. Selon les exons conservés, un même gène peut ainsi coder plusieurs protéines (Figure 1.3).

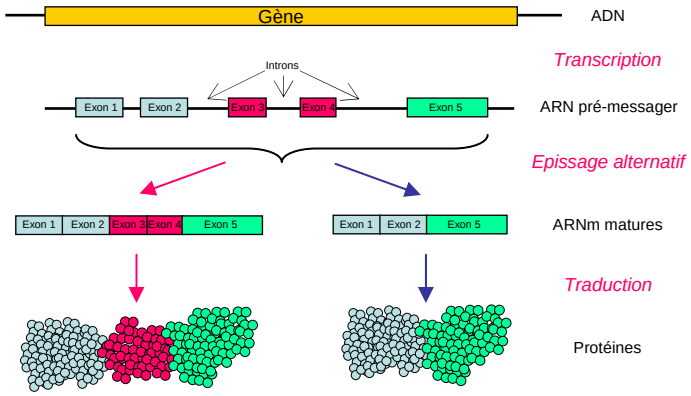

FIG. 1.3 – *Principe de l'épissage alternatif d'un gène*

1.1.4.3 La traduction

La traduction est l'interprétation des codons de l'ARNm en acides aminés. Elle utilise le code génétique standard (Figure 1.4) qui est un système de correspondances permettant à l'ARN, écrit dans un alphabet à 4 lettres (les nucléotides), d'être traduit en une séquence protéique, écrite dans un alphabet à 20 lettres (les acides aminés). Il permet de faire correspondre à chacun des 64 triplets de nucléotides un acide aminé ou un signal d'arrêt de la traduction (codon stop). En outre, le codon ATG (Met), appelé codon-initiateur, permet de commencer la traduction. Comme il n'existe que 20 acides aminés différents, certains acides aminés sont codés par plusieurs codons. On dit que le code génétique est dégénéré. La leucine (L), par exemple,

est codée par 6 codons différents : CTA, CTC, CTG, CTT, TTA et TTG.

Première Position	Seconde Position T	Seconde Position C	Seconde Position A	Seconde Position G	Troisième Position
T	TTT Phe [F] TTC Phe [F] TTA Leu [L] TTG Leu [L]	TCT Ser [S] TCC Ser [S] TCA Ser [S] TCG Ser [S]	TAT Tyr [Y] TAC Tyr [Y] TAA *Ter* [stop] TAG *Ter* [stop]	TGT Cys [C] TGC Cys [C] TGA *Ter* [stop] TGG Trp [W]	T C A G
C	CTT Leu [L] CTC Leu [L] CTA Leu [L] CTG Leu [L]	CCT Pro [P] CCC Pro [P] CCA Pro [P] CCG Pro [P]	CAT His [H] CAC His [H] CAA Gln [Q] CAG Gln [Q]	CGT Arg [R] CGC Arg [R] CGA Arg [R] CGG Arg [R]	T C A G
A	ATT Ile [I] ATC Ile [I] ATA Ile [I] ATG Met [M]	ACT Thr [T] ACC Thr [T] ACA Thr [T] ACG Thr [T]	AAT Asn [N] AAC Asn [N] AAA Lys [K] AAG Lys [K]	AGT Ser [S] AGC Ser [S] AGA Arg [R] AGG Arg [R]	T C A G
G	GTT Val [V] GTC Val [V] GTA Val [V] GTG Val [V]	GCT Ala [A] GCC Ala [A] GCA Ala [A] GCG Ala [A]	GAT Asp [D] GAC Asp [D] GAA Glu [E] GAG Glu [E]	GGT Gly [G] GGC Gly [G] GGA Gly [G] GGG Gly [G]	T C A G

FIG. 1.4 – *Code génétique standard.* La première colonne indique le premier nucléotide, la première rangée indique le deuxième et la dernière colonne indique le dernier nucléotide. Dans chaque case sont indiqués la totalité du codon, les trois premières lettres du nom de l'acide aminé correspondant et, entre crochets, le code à une lettre de cet acide aminé dans un alphabet à 20 lettres. D'après "The genetic code" http ://psyche.uthct.edu/shaun/SBlack/geneticd.html.

Après la transcription de l'ADN en ARN dans le noyau de la cellule, l'ARN migre dans le cytoplasme où s'effectue la traduction de l'ARN en protéine (Figure 1.5).

Les protéines peuvent subir des modifications post-traductionnelles telles que l'ajout de sucres. On parle alors de glycoprotéines, que l'on retrouve dans les protéines codées par le gène *env* des rétrovirus.

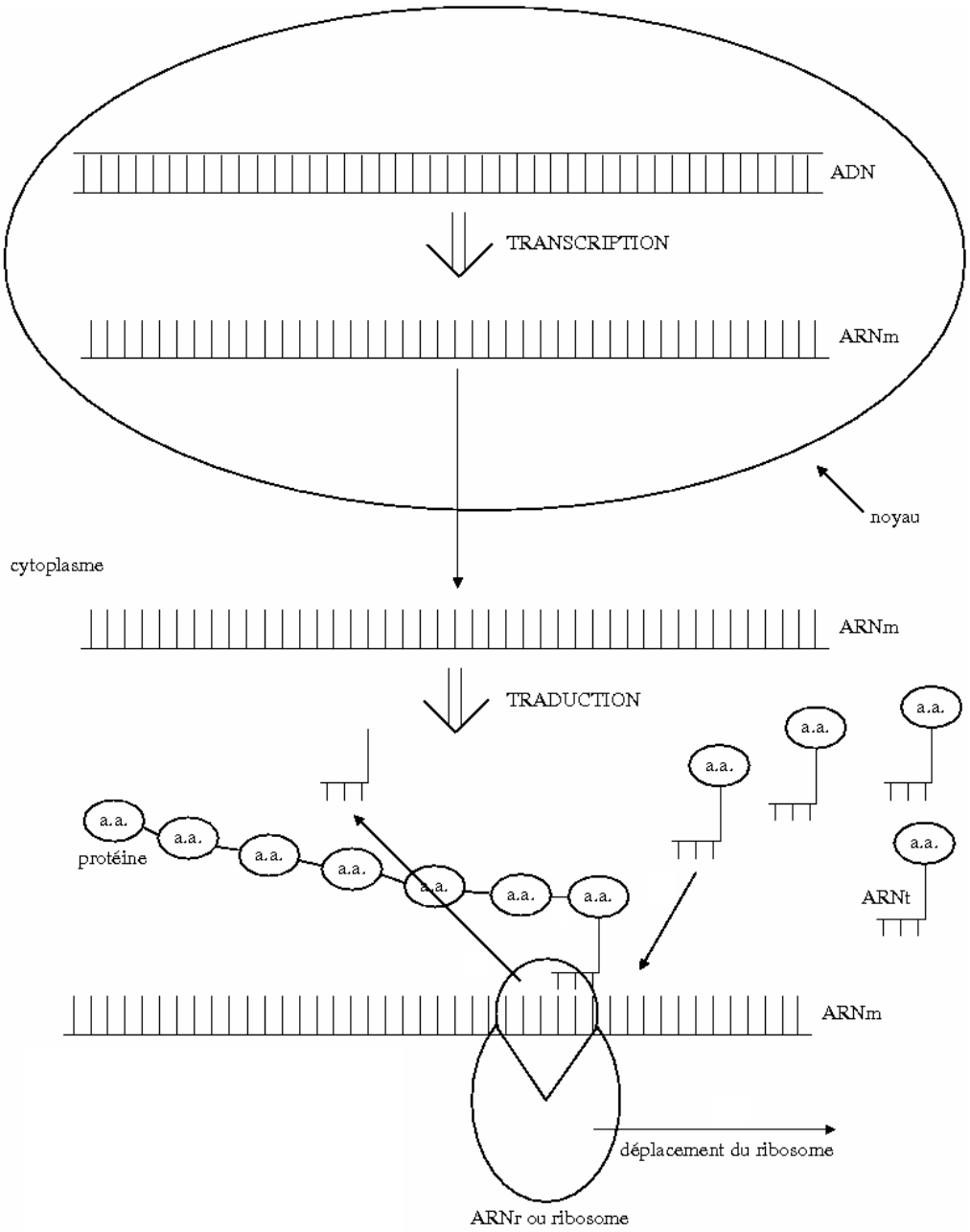

FIG. 1.5 – *Principe de la synthèse des protéines (Bionet 10 :12, 1 May 2005).* Pendant la transcription, l'information relative à la structure de la protéine passe de l'ADN à l'ARNm. L'ARNm passe ensuite dans le cytoplasme. Lors de la traduction, l'information provenant de l'ARNm détermine l'ordre d'assemblage dans le ribosome des acides aminés (a. a.) portés par les ARNt.

1.2 Les rétrovirus

Les rétrovirus sont des virus enveloppés à ARN. Ils constituent la famille des *Retroviridae*. Leur génome est constitué de deux copies d'ARN simple brin. La particularité des rétrovirus est la transformation de leur génome ARN simple brin en ADN double brin avant son intégration dans le génome de la cellule hôte. Le processus permettant la transformation de l'ARN en ADN est appelé rétrotranscription. Il fait intervenir une protéine spécifique codée par les rétrovirus : la RT (Reverse Transcriptase), une ADN polymérase ARN dépendante.

Les rétrovirus peuvent être distingués en deux catégories selon la complexité de leur génome (Coffin, 1992) : les virus simples, qui ne comportent que les gènes *gag, pol, pro* et *env* indispensables à leur multiplication, et les virus complexes, qui possèdent, en plus de ces quatre gènes, des gènes dits régulateurs ou accessoires. Le gène *gag (group-specific antigen)* code les protéines constitutives du core de la particule virale. Le gène *pol (polymérase)* code diverses activités enzymatiques essentielles à la réplication virale dont la RT et l'intégrase. Le gène *pro (protéase)* code la protéase virale. Le gène *env (enveloppe)* code les protéines constitutives de l'enveloppe du virus : la glycoprotéine de surface (SU) et la glycoprotéine transmembranaire (TM). Les gènes accessoires codent les protéines de régulation de l'expression des gènes viraux (Coffin *et al.*, 1997, Levy, 2006).

La famille des *Retroviridae* regroupe des virus divisés en sept genres, regroupés en deux sous-familles, selon des homologies de séquence nucléotidique et de structure génomique dans la région du gène *pol* (ICTV 2005) :

- La sous-famille des *Orthoretrovirinae* est composée de six genres : Alpharétrovirus, Bêtarétrovirus, Gammarétrovirus, Deltarétrovirus, Epsilonrétrovirus et Lentivirus. Le genre Lentivirus est constitué des rétrovirus complexes EIAV (Equine Infectious Anemia Virus), HIV (Human Immunodeficiency Virus), SIV (Simian Immunodeficiency Virus), FIV (Feline Immunodeficiency Virus), BIV (Bovine Immunodeficiency Virus) et SRLV (Small Ruminant LentiVirus).
- La sous-famille des *Spumaretrovirinae* comprend le genre Spumavirus

constitué de rétrovirus complexes non pathogènes comme HFV (Human Foamy Virus).

1.3 Les lentivirus

1.3.1 Les infections lentivirales

Les lentivirus sont des rétrovirus qui infectent aussi bien l'homme, avec le virus HIV, que les animaux, avec par exemple le virus EIAV qui infecte les équidés (chevaux, ânes, mulets) ou le virus SIV qui infecte les singes (Tableau 1.1). Ils sont responsables de maladies chroniques et/ou dégénératives aboutissant à la mort après ont une période de latence allant de quelques mois à plusieurs années. L'infection par les lentivirus entraîne une inflammation chronique et/ou dégénérative de certains organes comme le poumon ou la glande mammaire. Certains lentivirus comme HIV sont également responsables d'un déficit immunitaire. Tous les lentivirus persistent tout au long de la vie de l'individu infecté malgré l'existence d'une réponse immunitaire spécifique. La persistance des lentivirus est due à leur capacité à s'intégrer dans l'ADN de l'hôte et à échapper au système immunitaire. La possibilité des lentivirus d'échapper au système immunitaire provient essentiellement de leur très grande variabilité génomique et antigénique et de leur capacité d'infecter les cellules du système immunitaire.

1.3.2 La structure de la particule virale

Les lentivirus sont des particules sphériques dont le diamètre peut varier de 80 à 100 nm (Figure 1.6). Le génome viral, constitué de deux molécules d'ARN, et les protéines enzymatiques comme la RT et l'intégrase sont contenus dans la *capside* composée des protéines codées par le gène *gag*. La membrane interne, ou *matrice*, entoure la capside. L'enveloppe virale est formée d'une bicouche lipidique qui contient des protéines d'origine cellulaire et dans laquelle sont insérées les glycoprotéines de surface et transmembranaires virales.

TAB. 1.1 – Manifestations cliniques principales des infections lentivirales.

Hôte	Virus	Pathologies induites
Ongulés		
Bovidés	BIV (Bovine Immunodeficiency Virus)	Lymphadénopathies
Petits ruminants	SRLV (Small Ruminant Lentivirus)	Pneumonie interstitielle diffuse, Encéphalopathie, Arthrite Amaigrissement, Mammite
Equidés	EIAV (Equine Infectious Anemia Virus)	Anorexie, Hémorragies, Encéphalite, Fièvre, Amaigrissement, Anémie, Thrombopénie, Oedème, Glomérulonéphrite
Primates		
Singe	SIV (Simian Immunodeficiency Virus)	Déficit immunitaire, Syndrome neurologique, Arthrite
Homme	HIV (Human Immunodeficiency Virus)	Déficit immunitaire, Lymphadénopathies, Syndrome neurologique, Pneumonie interstitielle diffuse
Carnivores		
Félidés	FIV (Feline Immunodeficiency Virus)	Déficit immunitaire, Lymphadénopathies, Syndrome neurologique, Amaigrissement

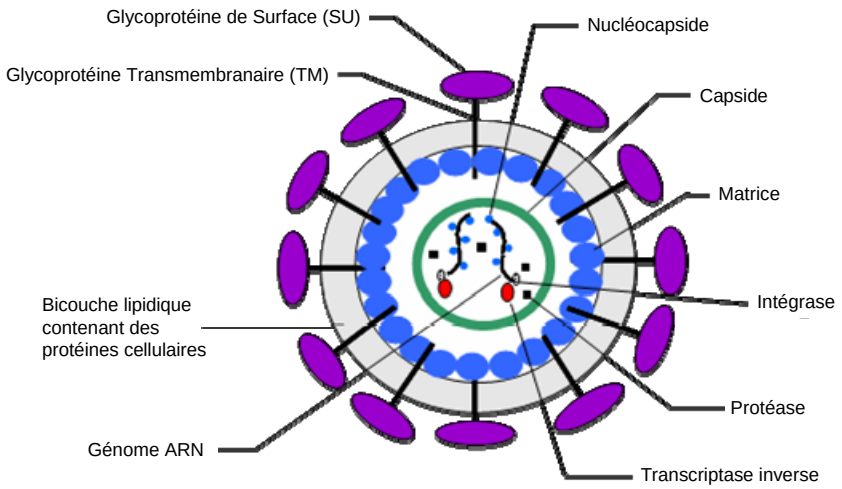

FIG. 1.6 – *Structure schématique de la particule lentivirale. D'après (Coffin et al., 1997).*

1.3.3 Le cycle lentiviral

Le cycle de réplication des lentivirus (Figure 1.7) commence par une interaction entre les glycoprotéines de surface de l'enveloppe virale et une protéine membranaire située à la surface de la cellule cible servant de récepteur. Cette interaction permet la fusion des membranes cellulaires et virales et l'entrée dans la cellule cible de la capside. La capside est ensuite dégradée, permettant la libération dans le cytoplasme de la cellule de l'ARN rétroviral. L'ARN viral est rétrotranscrit grâce à la RT présente dans la capside virale. La rétrotranscription conduit à la synthèse d'un brin d'ADN complémentaire qui est ensuite dupliqué pour obtenir de l'ADN viral double brin circulaire. L'ADN est transporté vers le noyau de la cellule hôte accompagné de l'intégrase virale et s'intègre dans l'ADN de cette cellule sous la forme de provirus. La machinerie cellulaire permet ensuite la transcription de l'ADN proviral en des ARN viraux messagers qui sont alors exportés dans le cytoplasme afin d'y être traduits en protéines virales. Les protéines constitutives de la capside sont assemblées, permettant l'encapsidation de copies de l'ARN génomique. Les virions formés sont enfin libérés. Les protéines de l'enveloppe virale sont incorporées au niveau de la membrane cellulaire lors du bourgeonnement. Les particules virales libres peuvent alors infecter de nouvelles cellules cibles et recommencer un cycle de réplication.

1.3.4 La variabilité génétique lentivirale

Les rétrovirus font partie des génomes qui évoluent le plus rapidement. Comme tous les virus à ARN, les rétrovirus ont un taux de mutation très élevé. Alors que les virus à ADN connaissent entre 10^{-7} et 10^{-11} mutations par nucléotide et par cycle (Drake, 1991), les rétrovirus subissent de 10^{-6} à 10^{-4} mutations par nucléotide et par cycle (Pathak and Temin, 1990). Les rétrovirus peuvent échapper au système immunitaire de leur hôte ou s'adapter et résister aux traitements. La grande variabilité génétique des rétrovirus représente ainsi un obstacle majeur à la mise au point de vaccins.

Parmi les rétrovirus, les lentivirus sont caractérisés par une variabilité génétique importante. Cette section est consacrée aux mécanismes qui en-

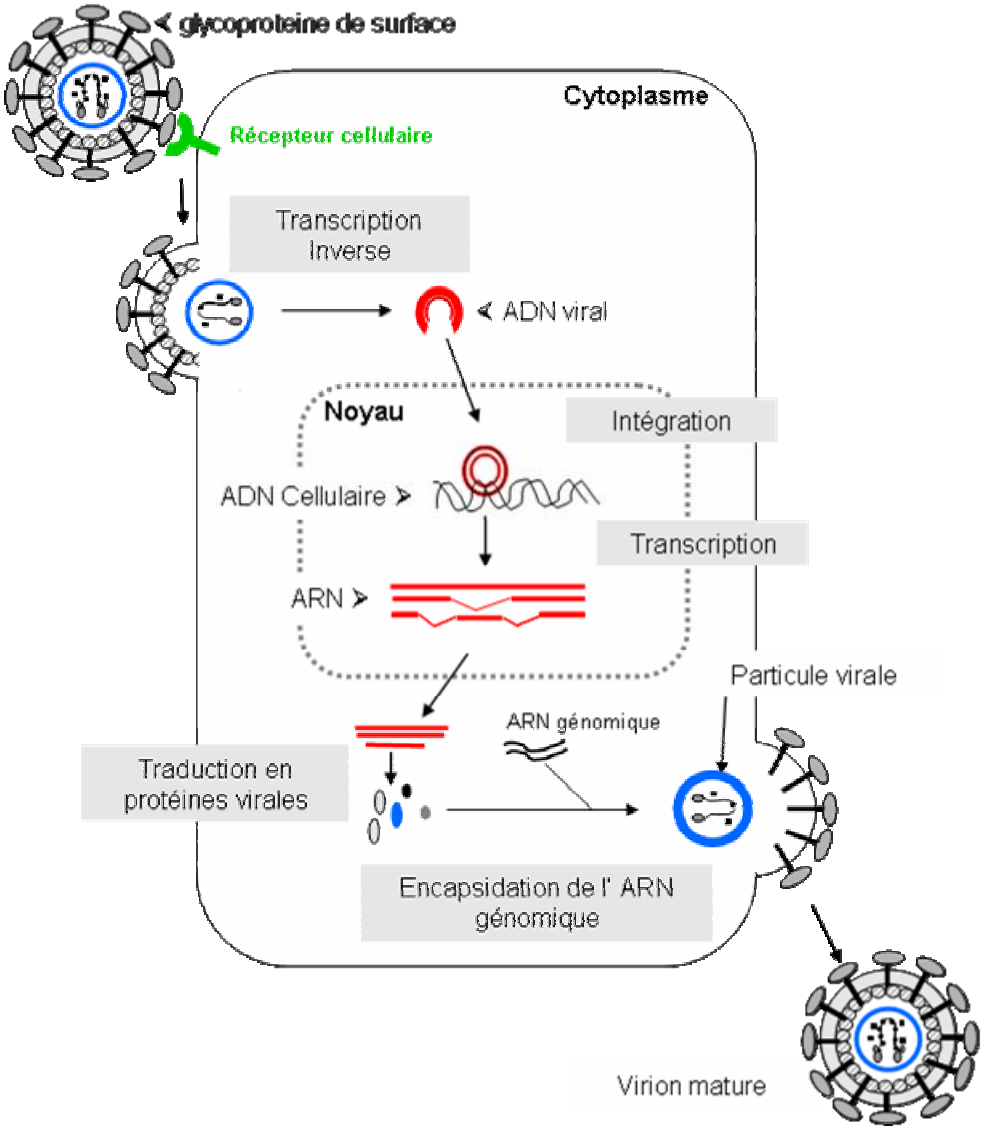

FIG. 1.7 – *Schéma du cycle lentiviral dans la cellule cible (Leroux et al., 2005)*

gendrent cette variabilité.

1.3.4.1 Les sources de la variablité génétique

Au cours de leur vie, les rétrovirus subissent trois systèmes de réplication différents : la génération d'ADN double brin à partir de l'ARN simple brin par la RT, la réplication de l'ADN viral intégré au génome de l'hôte par l'ADN polymérase cellulaire et la transcription de l'ADN viral en ARN par l'ARN polymérase II. Bien qu'il soit possible que des mutations soient introduites dans le génome viral lors de n'importe laquelle de ces étapes de réplication, la grande variabilité génétique des rétrovirus est généralement attribuée au manque de fidélité de la RT lors de la rétrotranscription. De nombreuses estimations des taux d'erreur des RT rétrovirales ont été conduites *in vitro*. Toutes ces études suggèrent que le taux d'erreur de la RT est de l'ordre de 10^{-4} à 10^{-5} erreurs par base et par cycle (Coffin *et al.*, 1997, Roberts *et al.*, 1988, Preston *et al.*, 1988, Takeuchi *et al.*, 1988, Weber and Grosse, 1989, Preston, 1997). Les génomes lentiviraux étant constitués en moyenne de 10 000 bases, ce taux d'erreur est compatible avec l'observation de 0,1 à 1 erreur par génome et par cycle de rétrotranscription (Wain-Hobson, 1989). *In vivo*, la fidélité de la RT peut être influencée par de nombreux facteurs comme le pH, la concentration en nucléotides ou la composition des séquences (Eckert and Kunkel, 1990, Bakhanashvili and Hizi, 1993, Bebenek and Kunkel, 1993) . Cependant, on admet que chaque copie de génome rétroviral diffère de son parent par au moins une base, ce qui conduit à la coexistence, au sein d'un même hôte, de génomes viraux génétiquement apparentés mais distincts, appelés quasiespèces (Wain-Hobson, 1993). Les quasiespèces ont des propriétés biologiques différentes et leur coexistence permet à la population virale dans son ensemble de survivre à la réponse immunitaire de l'hôte.

Les erreurs générées lors de la rétrotranscription ne sont pas corrigées et persistent dans le génome. En effet, contrairement aux polymérases cellulaires, la RT ne possède pas de mécanisme de correction des erreurs (Battula and Loeb, 1976). La transcriptase inverse est ainsi 10 à 200 fois moins fidèle que les ADN polymérases qui ont une activité de réparation (Bebenek and

Kunkel, 1993). De plus, la rétrotranscription ayant lieu dans le cytoplasme de la cellule hôte, les erreurs commises par la RT ne peuvent pas être corrigées par la machinerie de réparation cellulaire de l'ADN qui est localisée dans le noyau de la cellule hôte.

Les phénomènes de recombinaison entre les génomes viraux peuvent augmenter la diversité des populations virales. En général, les deux brins d'ARN qui constituent le génome des lentivirus sont identiques. Le brin d'ADN synthétisé est le même lorsque la RT utilise un seul brin d'ARN ou lorsqu'elle utilise des portions des deux brins lors de la rétrotranscription. Cependant, sous certaines conditions, il peut arriver que deux brins génétiquement différents soient présents simultanément dans une cellule coinfectée. Dans ce cas, la RT peut utiliser des portions de chaque brin d'ARN pour générer un ADN recombinant. Des recombinaisons *in vivo* ont été décrites pour HIV-1 et HIV-2 (Clavel *et al.*, 1989, Gao *et al.*, 1994, Robertson *et al.*, 1995b,a). Le taux minimum de recombinaison pour HIV-1 a été estimé à 2,8 évènements de recombinaison par génome et par cycle (Zhuang *et al.*, 2002).

Si les erreurs de la RT et les phénomènes de recombinaison sont la base de la diversité génétique des lentivirus, cette diversité semble aussi due au grand nombre de virions présents, au taux de réplication élevé (10^{12} virions par jour pour HIV-1) (Coffin, 1995) et à la présence d'une pression sélective.

1.3.4.2 Hétérogénéité de la distribution des mutations dans le génome

Les différents gènes des lentivirus n'évoluent pas à la même vitesse. La très grande plasticité des génomes des lentivirus est particulièrement évidente dans le gène *env*, et notamment dans la région SU codant la glycoprotéine de surface. Cependant, les mutations ne se répartissent pas de manière homogène le long du gène *env*. La région du gène *env* qui code la glycoprotéine de surface est constituée d'une succession de régions dites constantes (C) qui ne présentent pas, ou présentent peu, de variabilité génétique et de régions dites variables (V) qui présentent de nombreuses mutations. Des études sur la variabilité génétique des lentivirus ont conduit à la définition de régions

constantes et variables chez tous les lentivirus (Figure 1.8) (Burns *et al.*, 1993, Leroux *et al.*, 1997b, Modrow *et al.*, 1987, Valas *et al.*, 2000, Zheng *et al.*, 1997b, Suarez and Whetstone, 1995, Pancino *et al.*, 1993).

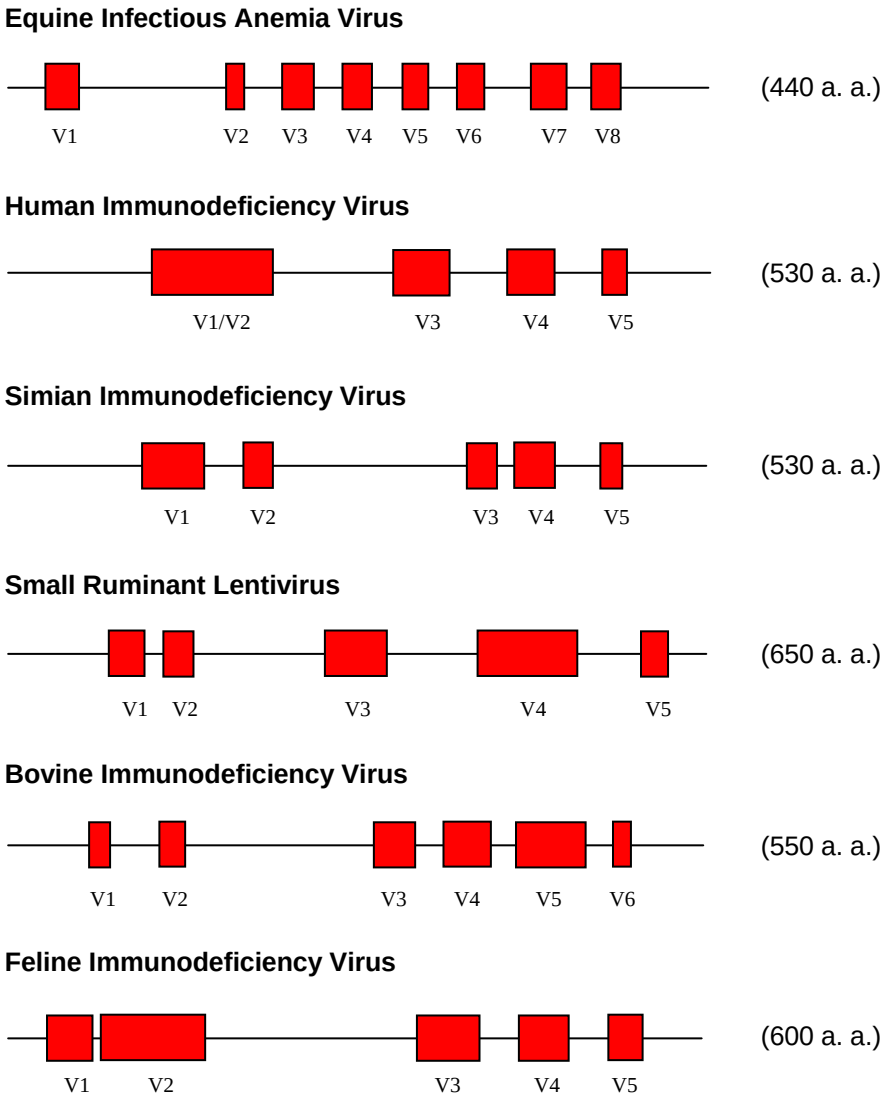

FIG. 1.8 – *Organisation schématique des glycoprotéines de surface des lentivirus*. La position des régions variables V (■) et des régions constantes (—) est représentée. La longueur moyenne en acides aminés (a.a.) de chaque glycoprotéine de surface est indiquée à côté des graphiques.

1.4 Le lentivirus EIAV

Le lentivirus EIAV (Equine Infectious Anemia Virus) est un membre de la famille des *Retroviridae*, appartenant au genre *Lentivirus*. Il infecte naturellement les équidés et est responsable de l'anémie infectieuse équine. Cette maladie a été décrite pour la première fois en France par Lignée en 1843 (Lignée, 1843). Le virus EIAV est transmis mécaniquement par des insectes hématophages, dont les Tabanidés (Foil *et al.*, 1987), ou par du matériel à usage vétérinaire mal stérilisé. Alors que la plupart des lentivirus induisent une maladie dégénérative d'évolution lente conduisant à la mort en quelques mois ou quelques années, le virus EIAV se distingue des autres lentivirus par une évolution d'une maladie chronique vers un stade asymptomatique. L'anémie infectieuse équine peut se manifester par une altération chronique de l'état général accompagnée de fièvres récurrentes. Sur le plan clinique, on décrit classiquement trois phases (Figure 1.9). La phase aiguë de l'infection se caractérise par une forte augmentation de la température associée à une baisse concomitante du nombre de plaquettes sanguines, appelée thrombocytopénie. Ce premier accès fébrile survient dans les 10-15 jours suivant l'infection. En général, l'animal survit à cette première phase et entre alors dans la phase chronique de l'infection. La phase chronique est caractérisée par des épisodes d'hyperthermie récurrents, toujours associés à une thrombocytopénie. Elle peut durer de 6 à 12 mois. Au cours de la phase asymptomatique de la maladie, qui peut durer plusieurs décennies, les équidés ne présentent plus aucune manifestation clinique de l'infection mais demeurent infectés à vie et capables de transmettre le virus. L'infection expérimentale de poneys a permis de montrer que le nombre et la fréquence des épisodes fébriles pouvaient varier selon les animaux, même infectés avec la même souche virale (Leroux *et al.*, 1997b, Hammond *et al.*, 2000, Leroux *et al.*, 2001, Cook *et al.*, 2003).

EIAV possède le génome le plus petit et le plus simple parmi les lentivirus. En plus des gènes *gag, pol* et *env* qui codent les protéines de structure et les activités enzymatiques, le génome d'EIAV contient trois autres gènes (*tat, rev* et *S2*) qui codent des protéines régulatrices. Comme les autres lentivirus, EIAV présente une grande instabilité génétique qui conduit à la coexis-

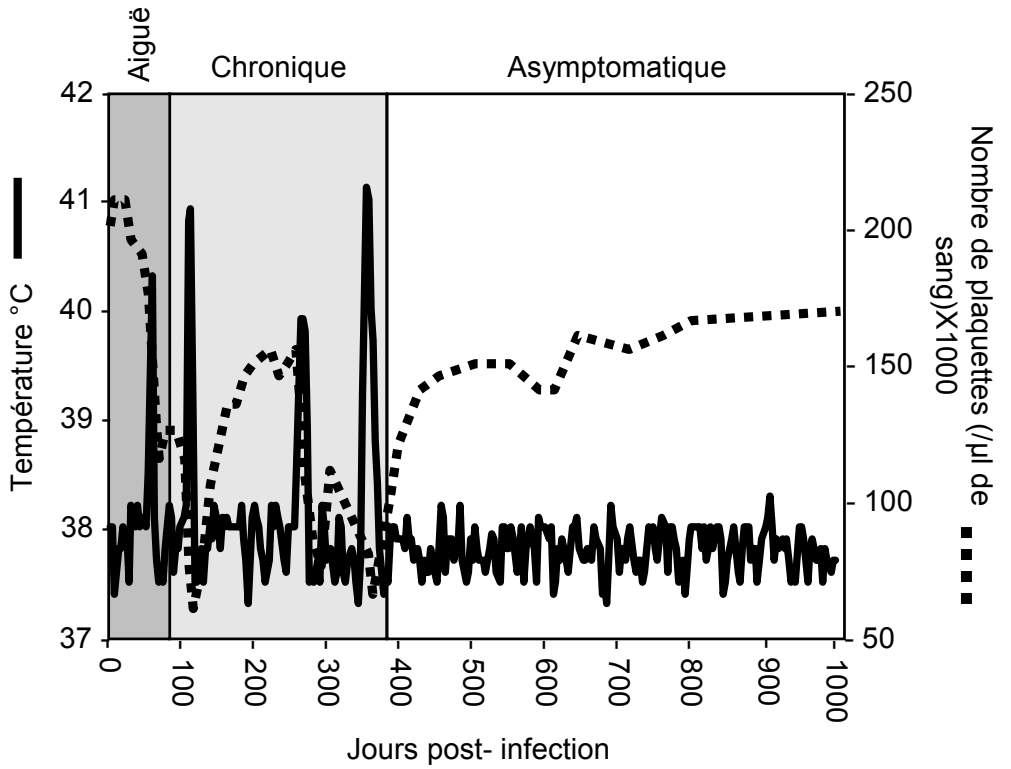

FIG. 1.9 – *Profil clinique de poneys infectés expérimentalement par l'EIAV.* Les épisodes fébriles sont définis par une température rectale supérieure à 39°C associée à une diminution du nombre de plaquettes en dessous de 100 000/μl de sang. D'après Leroux *et al.* (2004).

tence de quasiespèces. Chaque épisode fébrile au cours de l'anémie infectieuse équine est associé à l'émergence d'une nouvelle quasiespèce majoritaire (Leroux *et al.*, 1997a,b).

La variabilité génétique d'EIAV est répartie de façon hétérogène le long de son génome. Le gène *env*, qui code la glycoprotéine de surface et la glycoprotéine transmembranaire, constitue la région du génome qui accumule le plus de mutations. Des études portant sur l'évolution du virus ont permis de définir, sur la base d'alignements de séquences déduites en acides aminés de la région du gène *env* qui code la glycoprotéine de surface, huit régions variables V1 à V8, séparées par neuf régions constantes C1 à C9 (Leroux *et al.*, 1997b). Les régions variables ont été définies comme des ensembles d'acides aminés présentant plus de 30% de divergence par rapport à la séquence consensus (Figure 1.10).

1.5 Le matériel génétique disponible

Ce travail a porté sur la modélisation probabiliste des régions constantes et variables des lentivirus. Dans cette section, nous décrivons, pour les différents lentivirus, les ensembles de séquences nucléotidiques de la région du gène *env* codant la glycoprotéine de surface ayant été utilisées pour entraîner les modèles puis pour les tester. La table 1.2 indique les numéros d'accès de ces séquences dans la base de données GenBank.

EIAV : Nous avons utilisé 187 séquences de la région du gène *env* codant la SU d'EIAV (Craigo *et al.*, 2002, Leroux *et al.*, 2001, 1997b, Zheng *et al.*, 1997a,b, 2000).

Ensemble d'entraînement : 94 séquences.

Ensemble de test : 93 séquences.

Nous avons considéré les 8 régions variables V1 à V8 et les 9 régions constantes C1 à C9 définies par Leroux *et al.* (1997b).

HIV : Nous avons utilisé 155 séquences de la région du gène *env* codant la SU du virus HIV-1. Cet ensemble de séquences est composé de la séquence de référence HIV-1 HXB2 et de séquences réprésentatives des sous-types suivants : A (21 séquences), B (27 séquences), C (26 séquences), D (18 séquences), E (19 séquences), F (3 séquences), G (21 séquences), H (2 séquences), et 17 séquences de formes recombinantes.

Ensemble d'entraînement : 78 séquences.

Ensemble de test : 77 séquences.

Nous avons considéré les régions variables V1 à V5 définies par Modrow *et al.* (1987). Cependant, les régions V1 et V2 n'étant séparées que par une petite région constante composée uniquement de quelques acides aminés, elles ont été considérées comme une unique région variable V1/V2.

SIV : Nous avons utilisé 61 séquences de la région du gène *env* codant la SU de SIV.

Ensemble d'entraînement : 46 séquences.

Ensemble de test : 15 séquences.

Nous avons considéré les régions variables V1 à V5 définies par Burns

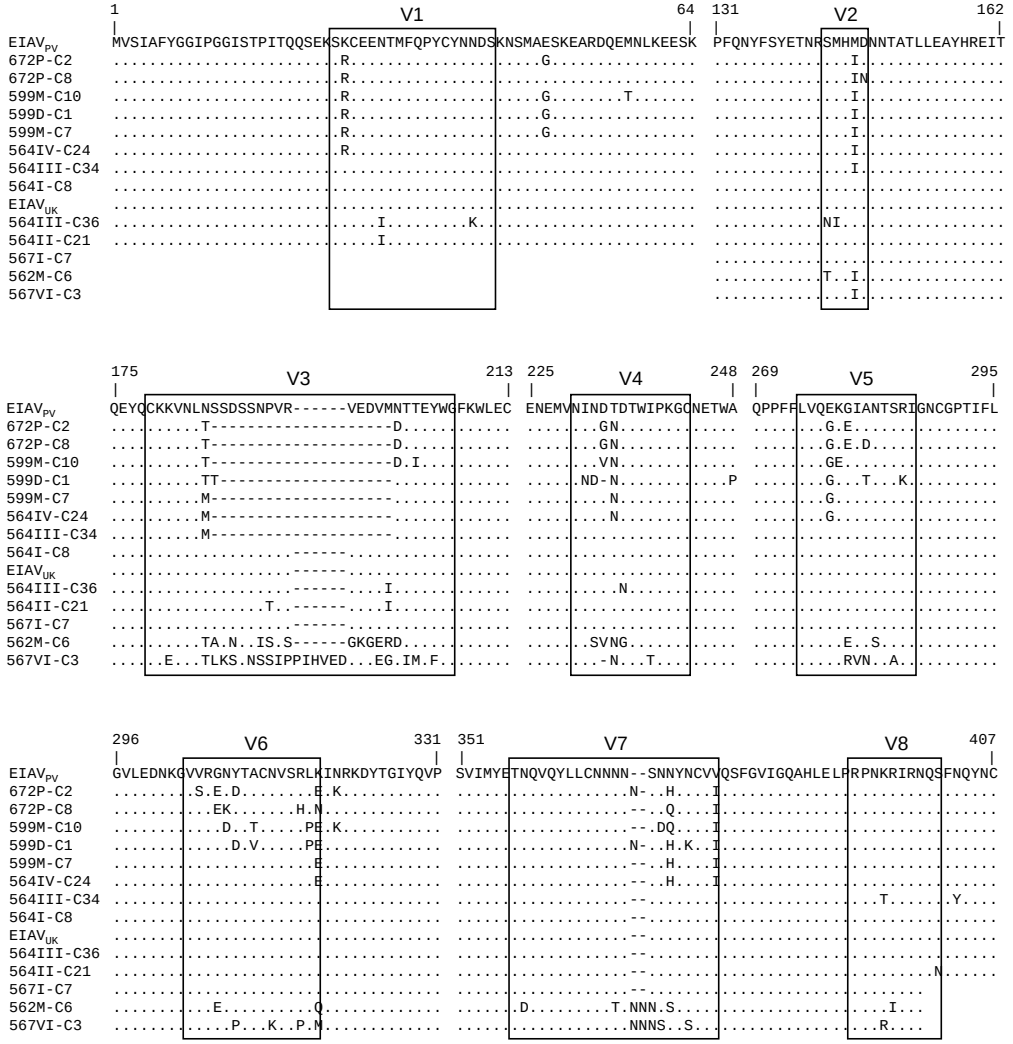

FIG. 1.10 – *Comparaison de séquences déduites en acides aminés de la région SU du gène env d'EIAV.* Les séquences déduites en acides aminés ont été comparées à la séquence du virus EIAV_{PV}. La position des acides aminés est indiquée au dessus des séquences. Seuls les acides aminés différents de ceux d'EIAV_{PV} sont montrés. Les points indiquent des résidus identiques à ceux de la séquence d'EIAV_{PV} ; les tirets indiquent des délétions. Les régions variables sont encadrées.

et al. (1993).

SRLV : Nous avons utilisé 68 séquences de la région du gène *env* codant la SU des SRLV.

Ensemble d'entraînement : 51 séquences.

Ensemble de test : 17 séquences.

Nous avons considéré les régions variables V1 à V5 définies par Valas *et al.* (2000).

BIV : Nous avons utilisé 13 séquences de la région du gène *env* codant la SU de BIV.

Nous avons comparé les régions prédites par nos modèles aux régions variables V1 à V6 définies par Suarez and Whetstone (1995).

FIV : Nous avons utilisé 16 séquences de la région du gène *env* codant la SU de FIV.

Nous avons comparé les régions prédites par nos modèles aux régions variables V1 à V5 définies par Pancino *et al.* (1993).

TAB. 1.2 – Numéros d'accès dans GenBank des séquences utilisées dans cette étude

EIAV	AF005104 à AF005151 (sauf AF005113, AF005136 et AF005145 à AF005148) ; AF016316 ; AF298666 à AF298762 (sauf AF298752 et AF298691 à AF298694) ; AF429316 à AF429353
HIV	K03455 ; AB032740, AB03274 ; AF133821 ; AF190127, AF190128 ; AF197340 ; AF209205, AF209208 ; AF219261, AF219272 ; AF322202 à AF322214 ; AF411964, AF411965 ; AF413978, AF413979 ; AF413987 ; AF443113 à AF443115 ; AF457079 à AF457090 (sauf AF457082 à AF457084, AF457086 et AF457089) ; AF460972, AF460974 ; AF484478, AF484493 ; AF484507 à AF484519 (sauf AF484508, AF484510, AF484512 et AF484517) ; AF529572, AF529573 ; AF530576 ; AF544007, AF544008 ; AJ417424 à AJ417431 ; AY037268 à AY037270 ; AY037280 à AY037283 ; AY158533 à AY158535 ; AY173957, AY173958 ; AY217545 ; AY228556, AY228557 ; AY253305 à AY253322 (sauf AY253307, AY253309, AY253315 à AY253316 et AY253319) ; AY255823 à AY255827 ; AY322184 à AY322191 (sauf AY322186 et AY322188) ; AY357571 à AY357576 (sauf AY357574) ; AY358069 à AY358073 (sauf AY358070) ; AY371155 à AY371163 (sauf AY371158 à AY371162) ; AY423908 à AY423928 ; AY494965 à AY494974 (sauf AY494967 à AY494968, AY494970 et AY494972) ; AY505010, AY505011 ; AY535509 à AY535513 ; AY563169 ; AY818641 à AY818643
SIV	AF075269 ; AF103818 ; AF131870 ; AF188114 à AF188116 ; AF328295 ; AF334679 ; AF382828, AF382829 ; AF447763 ; AY033233 ; AY159321, AY159322 ; AY169968 ; AY221508 à AY221513 ; AY290709 à AY290716 ; AY523865 à AY523867 ; AY587015 ; AY588946 ; AY599198 à AY599201 ; AY611488 ; L20008, L20009 ; L20098, L20099 ; L40990 ; M29975 ; M33262 ; M58410 ; M66437 ; M83293 ; U04005 ; U10897 à U10898 ; U25712 à U25715 ; U25744, U25745 ; U58991 ; U72748
SRLV	A15114 ; AF015180 ; AF156858 à AF156877 ; AF338227 ; AF474005 à AF474007 ; AF479638 ; AJ400718 à AJ400721 ; AY039765 à AY039784 ; L06906 ; M31646 ; M33677 ; M34193 ; M60609, M60610 ; M60855 ; S51392 ; S55323 ; U35795 à U35804 (sauf U35797, U35802 et U35803) ; U51910
BIV	L43126 à L43132 ; M32690 ; NC_001413 ; L04972 ; U80989 à U80991
FIV	M25381 ; M36968 ; L00608 ; M59418 ; X57001 à X57002 ; M73964 à M73965 ; X60725 ; L06725 ; X69494 à X69502 (sauf X69495, X69500 et X69501)

Chapitre 2

Les modèles markoviens de l'ADN

Sommaire

Ces dernières années, les grands projets de séquençage des génomes de différents organismes et le développement du transcriptome avec l'essor des puces à ADN, ont produit, et continuent à produire, une quantité gigantesque d'information. Afin d'aider les biologistes à analyser cette masse d'information, divers modèles mathématiques ont été introduits. Ces modèles permettent d'extraire une partie de l'information contenue dans les séquences

génomiques. Des modèles mathématiques ont ainsi été utilisés, par exemple, pour identifier des gènes, déterminer la structure de protéines ou étudier l'hétérogénéité des séquences d'ADN (Borodovsky and McIninch, 1993, Krogh *et al.*, 2001, Churchill, 1989). L'information contenue dans les séquences génomiques est extrêmement complexe. Un bon modèle a pour but de révéler les caractéristiques importantes de la séquence mais pas d'expliquer la nature dans ses moindres détails. Différents modèles probabilistes ont été introduits pour modéliser les séquences d'ADN, comme les modèles de Markov ou les modèles de Markov cachés.

Une séquence peut être vue comme une succession de lettres prises dans un alphabet $\mathcal{A}$ à quatre lettres qui représentent les quatre nucléotides adénine (A), cytosine (C), guanine (G) et thymine (T)

$$\mathcal{A} = \{A, C, G, T\}$$

lorsqu'il s'agit d'une séquence nucléotidique ou dans un alphabet à vingt lettres qui représentent les vingt acides aminés

$$\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

lorsqu'il s'agit d'une séquence déduite en acides aminés. Une séquence de longueur n sera ainsi représentée par la suite

$$x_{1:n} = x_1, x_2, \ldots, x_n$$

où pour chaque position i de la séquence, x_i est la réalisation d'une variable aléatoire X_i à valeurs dans $\mathcal{A}$ qui décrit le nucléotide ou l'acide aminé présent à cette position.

Ce chapitre décrit les aspects mathématiques des modèles de Markov et des modèles de Markov cachés utilisés pour analyser les séquences d'ADN.

2.1 Les modèles indépendants M0

2.1.1 Définition

Les modèles les plus simples pour décrire les séquences d'ADN sont les modèles de Markov d'ordre 0, aussi appelé modèles indépendants. Ces modèles permettent de rendre compte de la composition en nucléotides ou en acides aminés qui est l'une des principales caractéristiques d'une séquence génomique. Dans les modèles indépendants M0, on suppose que les variables aléatoires X_i sont i.i.d. (indépendantes et identiquement distribuées).
Le fait que les variables aléatoires X_i soient identiquement distribuées signifie qu'elles suivent toutes la même loi μ, déterminée par les quantités :

$$\mu = (\mu(a))_{a \in \mathcal{A}}$$

où $\mu(a) = P(X_i = a)$ pour tout $a \in \mathcal{A}$.

De cette façon, la probabilité de trouver un nucléotide ou un acide aminé donné dans une séquence ne dépend pas de la position i dans la séquence et est la même tout au long de cette séquence.

En plus d'être identiquement distribuées, les modèles indépendants supposent que les variables aléatoires X_i sont distribuées de manière indépendante. Cela signifie par exemple que

$$P(X_i = a, X_j = b) = P(X_i = a)P(X_j = b) = \mu(a)\mu(b)$$

pour toutes positions i et j et pour tous $a, b \in \mathcal{A}$.

Un modèle indépendant est entièrement défini par :

- le nombre M de caractères différents dans l'alphabet $\mathcal{A}$. Le nombre de caractères différents sera 4 pour une séquence nucléotidique et 20 pour une séquence déduite en acides aminés,
- la loi μ.

2.1.2 Estimation des paramètres

Lorsque l'on dispose d'une séquence d'ADN $x_{1:n} = (x_i)_{1 \leq i \leq n}$ de longueur $n$que l'on souhaite modéliser à l'aide d'un modèle indépendant, il faut déterminer les paramètres de la loi μ qui reflètent le mieux cette séquence. Notons $\theta = \{\mu\}$ les paramètres à estimer d'un modèle indépendant. Le problème consiste à déterminer les valeurs de θ qui maximisent la vraisemblance de la séquence étudiée $x_{1:n}$; c'est-à-dire à choisir l'estimateur du maximum de vraisemblance θ^* tel que

$$\theta^* = \text{argmax}_\theta \{P(X_{1:n} = x_{1:n}|\theta)\},$$

où la fonction argmax prend en entrée une variable et une expression arithmétique et retourne la valeur de la variable pour laquelle l'expression est maximale.

La vraisemblance d'une séquence correspond à la probabilité d'observer cette séquence sous le modèle choisi. Sous un modèle indépendant, elle est donnée par

$$P(X_{1:n} = x_{1:n}|\theta) = \prod_{i=1}^{n} P(X_i = x_i|\theta).$$

On en déduit alors

$$P(X_{1:n} = x_{1:n}|\theta) = \prod_{a \in \mathcal{A}} \mu(a)^{N_n(a)}$$

où $N_n(a)$ est le nombre d'occurrences de la lettre $a \in \mathcal{A}$ dans la séquence étudiée. Dans le cas du modèle indépendant M0, l'estimateur du maximum de vraisemblance est

$$\hat{\mu}(a) = \frac{N_n(a)}{n}, \ a \in \mathcal{A}.$$

En d'autres termes, la probabilité d'apparition de chaque lettre est estimée par sa fréquence d'apparition observée.

2.1.3 Les limites du modèle M0

Intéressons-nous aux fréquences des codons. Selon l'indépendance des variables aléatoires sous le modèle M0, on a

$$P(X_i = a_1, X_{i+1} = a_2, X_{i+2} = a_3) = P(X_i = a_1)P(X_{i+1} = a_2)P(X_{i+2} = a_3)$$

pour tout $a_{1:3} \in \mathcal{A}^3$. Ainsi, par exemple, la fréquence du codon ACG est donnée par

$$P(X_{i:i+2} = ACG) = P(X_i = A)P(X_{i+1} = C)P(X_{i+2} = G)$$

D'après Krogh *et al.*, les fréquences d'apparitions observées dans le génome d'*Escherichia Coli* pour les quatres nucléotides A, C, G et T sont respectivement 23,66%, 23,30%, 27,89% et 23,15%. A partir de ces fréquences, nous pouvons calculer les fréquences attendues des codons sous un modèle indépendant M0. Les fréquences observées dans le génome d'*Escherichia coli* des codons sont cependant très différentes des fréquences calculées sous un modèle M0 (Table 2.1). Les modèles indépendants M0 ne semblent ainsi pas très pertinents pour modéliser les séquences d'ADN.

2.2 Les modèles de Markov

Si l'on souhaite modéliser avec une bonne précision les séquences d'ADN, il n'est pas raisonnable de supposer les positions successives X_i indépendantes. C'est ce que montrent les différences entre les fréquences d'apparitions des codons observées chez *Escherichia coli* et les fréquences prédites avec un modèle qui suppose les X_i indépendantes. Afin d'obtenir un modèle qui prenne mieux en compte les caractéristiques des séquences d'ADN, il faut donc introduire une dépendance entre les positions de la séquence.

2.2.1 Les modèles de Markov d'ordre 1

Les modèles les plus simples qui permettent de tenir compte de la dépendance entre les variables aléatoires X_i sont les modèles de Markov d'ordre 1, notés M1.

2.2.1.1 Définition

Les modèles M1 supposent que la valeur de X_i est influencée par la valeur de X_{i-1} ; c'est-à-dire que chaque lettre dépend de la lettre présente à la

Co-don	Freq Obs	Freq Calc	Co-don	Freq Obs	Freq Calc	Co-don	Freq Obs	Freq Calc	Co-don	Freq Obs	Freq Calc
AAA	3,5	1,3	CAA	1,3	1,4	GAA	4,3	1,6	TAA	*	*
AAC	2,4	1,4	CAC	1,1	1,5	GAC	2,2	1,7	TAC	1,4	1,4
AAG	1,1	1,6	CAG	3,0	1,7	GAG	1,8	1,8	TAG	*	*
AAT	1,4	1,3	CAT	1,2	1,4	GAT	3,2	1,5	TAT	1,5	1,3
ACA	0,5	1,4	CCA	0,8	1,5	GCA	2,0	1,7	TCA	0,6	1,4
ACC	2,5	1,5	CCC	0,4	1,6	GCC	2,5	1,8	TCC	0,9	1,5
ACG	1,4	1,7	CCG	2,6	1,8	GCG	3,6	2,0	TCG	0,8	1,6
ACT	0,9	1,4	CCT	0,6	1,5	GCT	1,6	1,6	TCT	0,9	1,4
AGA	0,1	1,6	CGA	0,3	1,7	GGA	0,6	1,8	TGA	*	*
AGC	1,6	1,7	CGC	2,4	1,8	GGC	3,2	2,0	TGC	0,7	1,6
AGG	0,1	1,8	CGG	0,4	2,0	GGG	1,0	2,2	TGG	1,4	1,8
AGT	0,7	1,5	CGT	2,5	1,6	GGT	2,8	1,8	TGT	0,5	1,5
ATA	0,3	1,3	CTA	0,3	1,4	GTA	1,1	1,5	TTA	1,1	1,3
ATC	2,7	1,4	CTC	1,0	1,5	GTC	1,5	1,6	TTC	1,8	1,4
ATG	2,5	1,5	CTG	2,5	1,5	GTG	2,7	1,8	TTG	1,2	1,5
ATT	2,8	1,3	CTT	0,9	1,4	GTT	1,9	1,5	TTT	1,9	1,2

TAB. 2.1 – Comparaison des fréquences d'apparitions (en pourcent) des codons observées dans le génome d'*E. coli* et calculées à partir d'un modèle M0. "Freq Obs" donne la fréquence d'apparition observée et "Freq Calc" la fréquence d'apparition calculée. "*" signifie qu'il s'agit d'un codon stop. D'après (Krogh, 1994)

position précédente. Dans les modèles de Markov M1, les séquences d'ADN sont modélisées par des chaînes de Markov d'ordre 1.

Définition 2.2.1 *On appelle* ***chaîne de Markov*** *d'ordre 1, une suite de variables aléatoires* (X_i) *à valeurs dans un ensemble* $\mathcal{A}$ *telle que, pour toute suite d'observations* $x_{1:n} \in \mathcal{A}^n$ *et pour tout* i*,*

$$P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) = P(X_i = x_i | X_{i-1} = x_{i-1}). \quad (2.1)$$

De plus, la chaîne de Markov est dite ***homogène*** *si, pour tout* i *et tous* $a, b \in \mathcal{A}$*,*

$$P(X_i = b | X_{i-1} = a) = p(a, b).$$

La propriété (2.1) s'appelle la propriété de Markov. Elle s'énonce souvent de la façon suivante : « Le futur ne dépend du passé qu'à travers le présent ». En d'autres termes, la probabilité de lire un nucléotide ou un acide aminé à la position i ne dépend pas de tous les nucléotides ou acides aminés précédents mais seulement du nucléotide ou de l'acide aminé présent à la position $i-1$.

Par la suite, nous ne considérerons que des chaînes de Markov homogènes ; c'est-à-dire dont les probabilités de transitions entre les nucléotides ou les acides aminés ne dépendent pas de la position le long de la séquence.

L'alphabet $\mathcal{A}$ est appelé **espace des états** et les nucléotides ou les acides aminés qui composent l'ensemble $\mathcal{A}$ sont les **états** de la chaîne de Markov. Les probabilités $p(a, b)$ sont appelées **probabilités de transition**. Elles vérifient $p(a, b) \geq 0$ pour tous $a, b \in \mathcal{A}$. Ainsi, $p(a, b)$ est la probabilité d'avoir le nucléotide ou l'acide aminé b à une position donnée sachant que le nucléotide ou l'acide aminé précédent est a. La matrice

$$P = (p(a, b))_{(a,b) \in \mathcal{A} \times \mathcal{A}}$$

est appelée **matrice de transition** et vérifie

$$\sum_{b \in \mathcal{A}} p(a, b) = 1.$$

Pour totalement définir un modèle de Markov, il faut aussi se donner la loi initiale ν qui indique les probabilités

$$\nu(a) = P(X_1 = a)$$

pour tout $a \in \mathcal{A}$. Comme ν est une loi, on a $\nu(a) \geq 0$ pour tout $a \in \mathcal{A}$ et

$$\sum_{a \in \mathcal{A}} \nu(a) = 1.$$

Une chaîne de Markov homogène d'ordre 1 peut être représentée graphiquement par l'ensemble de ses états reliés entre eux par des flèches qui indiquent les transitions entre les états (Figure 2.1).

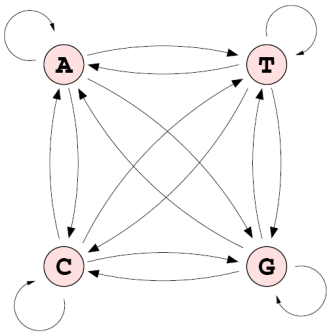

FIG. 2.1 – *Graphe associé à une chaîne de Markov à valeurs dans l'alphabet* $\mathcal{A} = \{A, C, G, T\}$. *La flèche partant de l'état* A *vers l'état* C *est associée à la probabilité de transition* $p(A, C)$.

Un modèle de Markov d'ordre 1 est entièrement défini par :

- le nombre M de caractères différents dans l'alphabet $\mathcal{A}$,
- la matrice de transition P,
- la loi initiale ν qui définit les probabilités d'apparition de chaque caractère de $\mathcal{A}$ à la première position de la séquence.

2.2.1.2 Estimation des paramètres

Les paramètres à estimer d'un modèle de Markov M1 sont $\theta = \{\nu, P\}$.

Par la suite, nous supposerons les chaînes de Markov étudiées irréductibles, c'est-à-dire que tout état est atteignable à partir de tout autre état par des probabilités de transition strictement positives. Pour tous a et b appartenant à $\mathcal{A}$, il existe donc q tel que

$$p^q(a,b) > 0.$$

Les chaînes étudiées seront également supposées apériodiques, ce qui signifie que, pour tous $a, b \in \mathcal{A}$,

$$\text{pgcd}\{q, p^q(a,b) > 0\} = 1.$$

L'espace des états $\mathcal{A}$ étant fini, les chaînes étudiées admettent une unique mesure stationnaire π qui vérifie

$$\pi = \pi P.$$

La distribution initiale ν sera alors choisie comme étant l'unique distribution stationnaire π. L'ensemble des paramètres à étudier est donc réduit à $\theta = \{P\}$.

Afin d'ajuster au mieux un modèle de Markov M1 aux séquences d'ADN que l'on souhaite étudier, il faut déterminer l'estimateur du maximum de vraisemblance de θ. Considérons une séquence $x_{1:n}$. La vraisemblance de cette séquence selon le modèle est alors

$$P(X_{1:n} = x_{1:n}) = \nu(x_1) \prod_{i=2}^{n} p(x_{i-1}, x_i) = \nu(x_1) \prod_{a,b \in \mathcal{A}} p(a,b)^{N_n(ab)}$$

où $N_n(ab)$ représente le nombre d'occurrences du dinucléotide $ab \in \mathcal{A}^2$ dans une séquence de longueur n.

Les estimateurs du maximum de vraisemblance des paramètres d'un modèle de Markov M1 sont

$$\hat{p}(a,b) = \frac{N_n(ab)}{N_{n-1}(a)} \quad \text{pour tous } a, b \in \mathcal{A}.$$

La mesure stationnaire correspondante est donnée par

$$\hat{\pi}(a) = \frac{N_n(a)}{n} \quad \text{pour tout } a \in \mathcal{A}.$$

2.2.2 Les modèles de Markov d'ordre m

Dans le paragraphe précédent, nous avons défini les modèles de Markov d'ordre 1. Cette définition peut être étendue à des modèles de Markov d'ordre m, notés Mm. Les modèles Mm supposent que la valeur de X_i est influencée par la valeur de $X_{i-1}, \ldots, X_{i-m}$; c'est-à-dire que chaque lettre dépend des m lettres présentes aux positions précédentes. Les probabilités de transition d'un modèle d'ordre m sont données par

$$P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_{i-m} = x_{i-m}).$$

Mathématiquement, un modèle de Markov d'ordre m est un cas particulier de modèle de Markov d'ordre 1. En effet $(X_n)_n$ suit un modèle d'ordre m si et seulement si $(X_{n:n+m-1})_n$ suit un modèle d'ordre 1. Les résultats valables pour un modèle d'ordre 1 restent donc valables pour un modèle d'ordre m. En particulier, il est possible d'estimer les paramètres d'un modèle de Markov d'ordre m par maximum de vraisemblance en posant

$$\hat{p}(w, b) = \frac{N_n(wb)}{N_{n-1}(w)} \quad \text{pour tout } w \in \mathcal{A}^m \text{ et tout } b \in \mathcal{A}$$

où $N_n(w)$ et $N_n(wb)$ représentent respectivement le nombre d'occurrences du mot de longueur m, $w \in \mathcal{A}^m$, et du mot de longueur $m+1$, $wb \in \mathcal{A}^{m+1}$, dans une séquence de longueur n.

2.2.3 Les limites des modèles de Markov

Intéressons nous aux dinucléotides CG, notés CpG pour ne pas confondre avec la paire de nucléotides C-G situés en vis à vis sur les deux brins de la molécule d'ADN. Dans le génome des vertébrés, la méthylation de l'ADN est spécifique des doublets CpG. Le niveau important de méthylation et le taux élevé de mutations des cytosines méthylées sont supposés être à l'origine de la sous-représentation des CpG dans les génomes des vertébrés. Cependant, pour des raisons biologiques, la méthylation est inhibée dans certaines régions des génomes. Dans ces régions, la proportion de CpG est plus importante que dans le reste des génomes. De telles régions sont appelées îlots CpG (Bird, 1986, Gardiner-Garden and Frommer, 1987, Larsen *et al.*, 1992).

Contrairement aux modèles indépendants qui ne tiennent pas compte de l'ordre des lettres, les modèles de Markov d'ordre 1 permettent de prendre en compte la composition en dinucléotides des séquences d'ADN. Ils fournissent donc un bon outil pour étudier la proportion de CpG le long des génomes des vertébrés. Cependant, dans les modèles de Markov, la probabilité de voir apparaître G sachant que l'on vient de lire C ne dépend pas de l'endroit où l'on se trouve dans la séquence. Dans un modèle de Markov, la séquence est supposée statistiquement homogène, c'est-à-dire que la probabilité d'apparition du dinucléotide CpG est la même tout le long de la séquence. Or, ce n'est pas le cas dans les génomes des vertébrés.

En plus des îlots CpG, les séquences d'ADN présentent de nombreux autres exemples d'hétérogénéité. Par exemple, les séquences sont composées de gènes et de régions intergéniques, d'introns et d'exons, de régions constantes et de régions variables...

Les modèles de Markov ne permettent pas de prendre en compte l'hétérogénéité des séquences et ne sont donc pas totalement adaptés à la modélisation des séquences d'ADN.

2.3 Les modèles de Markov cachés M1-Mm

Afin de modéliser au mieux les séquences d'ADN, il faut introduire un nouveau type de modèles qui permette de traduire l'hétérogénéité de ces séquences. Les modèles de Markov cachés sont une approche statistique pour intégrer cette hétérogénéité.

2.3.1 Définition

Le principe des modèles de Markov cachés est de considérer une séquence hétérogène comme une succession de régions statistiquement homogènes (Rabiner, 1989, McDonald and Zucchini, 1997). Chaque type de région est alors décrit par un modèle de Markov Mm, $m \geq 0$, spécifique. La succession des régions homogènes est elle-même décrite par une chaîne de Markov, $(S_i)_{1 \leq i \leq n}$, qui n'est pas directement observable. A chaque nucléotide X_i de la séquence

va ainsi correspondre un régime caché S_i qui indique le modèle Mm selon lequel le nucléotide X_i a été généré. Les états S_i de la chaîne de Markov décrivant la succession des régions homogènes sont appelés les **états cachés** du modèle et l'ensemble des états cachés est noté $\mathcal{S}$. Les nucléotides X_i sont appelés les **observations** et l'ensemble des observations est noté $\mathcal{A}$.

Un modèle de Markov caché est donc caractérisé par l'imbrication de deux processus : un processus caché $S = (S_1, \ldots, S_n)$ et un processus observable $X = (X_1, \ldots, X_n)$. Dans le cas d'une séquence d'ADN, le processus caché correspond à l'alternance des régions homogènes le long de la séquence et le processus observable correspond à la séquence d'ADN observée.

Dans tous les modèles considérés ici, la chaîne des états cachés est une chaîne de Markov du premier ordre homogène (M1), ce qui signifie que la probabilité d'être dans un certain état à une position donnée de la séquence ne dépend que de l'état précédent :

$$T(k,l) = P(S_{i+1} = l | S_i = k)$$

pour tous $k, l \in \mathcal{S}$. La matrice T ainsi définie est appelée **matrice de transition** et vérifie

$$\sum_{l \in \mathcal{S}} T(k,l) = 1.$$

Modèles M1-M0 Dans les modèles M1-M0, qui sont les modèles de Markov cachés décrits dans la littérature (Rabiner, 1989, Durbin *et al.*, 1998), les observations X_i apparaissent, conditionnellement à l'état caché correspondant S_i, indépendamment les unes des autres. Dans ce type de modèle, la probabilité d'émettre une lettre $a \in \mathcal{A}$ dépend seulement de l'état $k \in \mathcal{S}$ associé (Figure 2.2).

La **matrice d'émission** est alors définie, pour tout i, par

$$E(k,a) = P(X_i = a | S_i = k).$$

Cette matrice vérifie la propriété

$$\sum_{a \in \mathcal{A}} E(k,a) = 1.$$

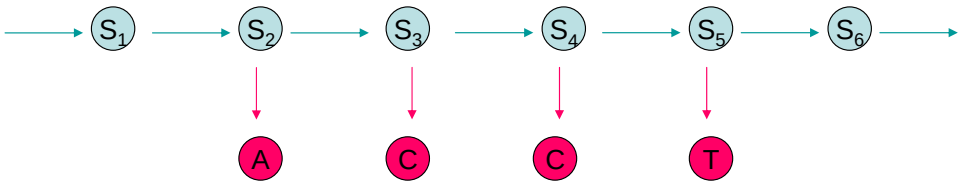

FIG. 2.2 – *Graphe associé à un modèle de Markov caché d'ordre 0.*

Modèles M1-Mm Ces modèles, introduits par Churchill en 1989 (Churchill, 1989), permettent de prendre en compte la composition en $m+1$-nucléotides de la séquence (Muri, 1997). Les modèles M1-Mm supposent que, conditionnellement à l'état caché $k \in \mathcal{S}$, la probabilité d'émettre une lettre $a_{m+1} \in \mathcal{A}$ dépend des m lettres précédentes $a_{1:m} \in \mathcal{A}^m$ (Figure 2.3).

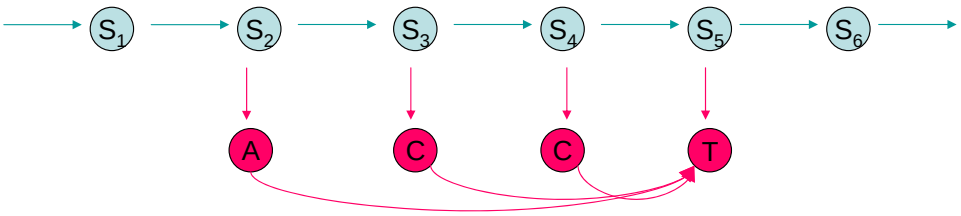

FIG. 2.3 – *Graphe associé à un modèle de Markov caché d'ordre m.*

Les termes de la matrice d'émission $E(k, a_{1:m}, a_{m+1})$ sont donnés, pour tout i, par

$$E(k, a_{1:m}, a_{m+1}) = P(X_{i+1} = a_{m+1} | X_{i-m+1:i} = a_{1:m}, S_{i+1} = k).$$

Cette matrice vérifie la propriété

$$\sum_{a_{m+1} \in \mathcal{A}} E(k, a_{1:m}, a_{m+1}) = 1.$$

Un modèle de Markov caché est entièrement défini par :

- le nombre $N = \text{card}(\mathcal{S})$ d'états cachés. Ici, le nombre d'états cachés correspond au nombre de types de régions homogènes dans une séquence d'ADN,
- le nombre $M = \text{card}(\mathcal{A})$ de caractères différents dans l'alphabet $\mathcal{A}$,
- la matrice de transition T,
- la matrice d'émission E,
- les distributions initiales ν_T et ν_E qui permettent de décrire le début de la séquence des observations.

2.3.2 Estimation des paramètres

Etant donnée une séquence d'ADN $x_{1:n}$, le problème est d'estimer les paramètres du modèle de Markov caché qui caractérise le mieux cette séquence. Notons θ l'ensemble des paramètres d'un modèle de Markov caché à estimer, c'est-à-dire la matrice de transition, la matrice d'émission et les distributions initiales. Le problème consiste à déterminer les valeurs de $\theta = \{T, E\}$ qui maximiseront la probabilité de la séquence $x_{1:n}$ selon le modèle de Markov caché correspondant.

Dans l'estimation des paramètres d'un modèle de Markov caché, nous devons distinguer deux situations : l'estimation lorsque la séquence des états cachés est connue et l'estimation lorsque la séquence des états cachés est inconnue.

2.3.2.1 Estimation lorsque la séquence des états cachés est connue

Lorsque la séquence des états cachés $s_1, \ldots, s_n$ correspondant aux observations $x_1, \ldots, x_n$ est connue, il est possible d'estimer séparément les probabilités de transition entre états et les probabilités d'émission des nucléotides ou des acides aminés par maximum de vraisemblance. La vraisemblance d'une séquence observée $x_{1:n}$ associée à la séquence des états cachés $s_{1:n}$ est donnée par :

$$P_\theta(x_{1:n}, s_{1:n}) = \nu_T(s_1)\nu_E(x_{1:m}) \prod_{i=2}^{n} T(s_{i-1}, s_i) \prod_{i=1}^{n-m} E(S_i, x_{i:m+i-1}, x_{m+i})$$

Les estimateurs du maximum de vraisemblance des paramètres d'un modèle M1-Mm sont alors donnés par les fréquences de transition et d'émission observées sur la séquence :

$$\hat{T}(k,l) = \frac{t(k,l)}{\sum_{s\in\mathcal{S}} t(k,s)}$$

et

$$\hat{E}(k, a_{1:m}, a_{m+1}) = \frac{e(k, a_{1:m}, a_{m+1})}{\sum_{a\in\mathcal{A}} e(k, a_{1:m}, a)}.$$

où $t(k,l)$ représente le nombre de transitions observées de l'état caché $k \in \mathcal{S}$ vers l'état caché $l \in \mathcal{S}$ et $e(k, a_{1:m}, a_{m+1})$ représente le nombre d'occurrences observées du mot $a_{1:m}a_{m+1}$ alors que l'état correspondant à l'observation a_{m+1} est k.

2.3.2.2 Estimation lorsque la séquence des états cachés est inconnue

Lorsque la séquence des états cachés $s_1, \ldots, s_n$ correspondant aux observations $x_1, \ldots, x_n$ n'est pas connue, les estimateurs du maximum de vraisemblance de $\theta = \{T, E\}$ ne peuvent pas être calculés directement. Dans ce cas, il existe différents algorithmes qui permettent d'estimer les paramètres des modèles de Markov cachés. L'algorithme le plus connu est l'algorithme *expectation-maximisation* (EM) qui est un algorithme itératif permettant de maximiser localement la vraisemblance de données incomplètes (Dempster *et al.*, 1977). Dans le contexte particulier des modèles de Markov cachés, cet algorithme est aussi appelé algorithme de *Baum-Welch* (Baum *et al.*, 1970). A partir d'un ensemble initial de paramètres θ_0, cet algorithme consiste en l'alternance de deux phases : une phase d'estimation (E) de la séquence des états cachés sous la valeur courante du paramètre θ et une phase de maximisation (M) de la vraisemblance (Shamir, 2001, Ephraim, 2002). Chaque itération de l'algorithme de Baum-Welch permet ainsi de définir un nouvel ensemble de paramètres θ' qui augmente la vraisemblance :

$$P_{\theta'}(x_{1:n}) \geq P_{\theta}(x_{1:n}).$$

Une itération de l'algorithme de Baum-welch pour une observation $X = x_{1:n}$ peut être décrite par :

Phase E : Calcul de la probabilité de deux états successifs k et l sous la valeur courante $\theta_j = \{T_j, E_j\}$ de θ pour tous $k \in \mathcal{S}$ et $l \in \mathcal{S}$.

Pour cela, définissons les variables forward $f_k(i)$ et backward $b_k(i)$ pour toute séquence $X = x_{1:n}$ et tout état $k \in \mathcal{S}$:

$$f_k(i) = P(X_{1:i} = x_{1:i}, S_i = k),$$

et

$$b_k(i) = P(X_{i+1:n} = x_{i+1:n} | S_i = k, X_{i-m+1:i} = x_{i-m+1:i}).$$

Ces variables se calculent facilement grâce à la récurrence avant-arrière. Le calcul des $f_k(i)$ se fait par la récurrence avant :

$$f_l(i+1) = E(l, x_{i-m+1:i}, x_{i+1}) \cdot \sum_{k \in \mathcal{S}} f_k(i) T(k, l).$$

Le calcul des $b_k(i)$ se fait par la récurrence arrière :

$$b_k(i) = \sum_{l \in \mathcal{S}} E(l, x_{i-m+1:i}, x_{i+1}) b_l(i+1) T(k, l).$$

Le calcul de la probabilité de deux états successifs k et l sous la valeur courante de θ est alors effectué de la façon suivante :

$$P(S_i = k, S_{i+1} = l | X, \theta_j) = \frac{f_k(i) \cdot T_j(k, l) \cdot E_j(l, x_{i-m+1:i}, x_{i+1}) \cdot b_l(i+1)}{P(X)},$$

avec

$$P(X) = \sum_{k \in \mathcal{S}} f_k(n) = \sum_{k \in \mathcal{S}} f_k(i) b_k(i).$$

Phase M : Calcul de la nouvelle valeur θ_{j+1} de θ qui augmente la vraisemblance des observations.

Introduisons :

$$t(k, l) = \frac{1}{P(X)} \cdot \sum_{i=1}^{n-1} f_k(i) \cdot T_j(k, l) \cdot E_j(l, x_{i-m+1:i}, x_{i+1}) \cdot b_l(i+1)$$

et

$$e(k, a_{1:m}, a_{m+1}) = \frac{1}{P(X)} \cdot \sum_{i=m+1}^{n} f_k(i) b_k(i) \cdot 1_{\{X_{i-m:i}=a_{1:m+1}\}}.$$

Alors :

$$T_{j+1}(k, l) = \frac{t(k, l)}{\sum_{s \in \mathcal{S}} t(k, s)}$$

et

$$E_{j+1}(k, a_{1:m}, a_{m+1}) = \frac{e(k, a_{1:m}, a_{m+1})}{\sum_{a \in \mathcal{A}} e(k, a_{1:m}, a)}.$$

Partant d'une initialisation aléatoire θ_0 de θ, les phases E et M sont exécutées alternativement jusqu'à ce que la différence entre les vraisemblances obtenues lors de deux itérations successives soit inférieure à un certain seuil ε fixé arbitrairement. L'algorithme de Baum-Welch permet de générer une suite $\theta_0, \ldots, \theta_j, \ldots$ de paramètres qui augmente à chaque itération la vraisemblance. La suite de paramètres θ_j converge vers un maximum local. Cependant, si le point de départ de l'algorithme θ_0 se situe dans un voisinage du maximum global θ^*, alors l'algorithme de Baum-Welch permet de converger vers l'estimateur du maximum de vraisemblance (EMV) θ^* (Wu, 1983, Muri, 1997).

2.3.3 La reconstruction de la séquence des états cachés

Lorsque l'on étudie une séquence d'ADN hétérogène avec un modèle de Markov caché, l'un des principaux problèmes que l'on se pose après l'estimation des paramètres du modèle est la reconstruction de la séquence des états cachés $s_1, \ldots, s_n$ à partir de la séquence observée $x_1, \ldots, x_n$. On cherche ainsi à « décoder »une séquence $x_{1:n}$ afin d'identifier des régions statistiquement homogènes, chaque région pouvant se répartir sur plusieurs plages de la séquence. Supposons par exemple que l'on souhaite étudier une séquence $x_{1:16}$ de longueur 16 composée de deux types de régions notés 1 et 2. La figure 2.4 représente la séquence d'ADN étudiée ainsi que la séquence des états cachés correspondante.

Pour identifier les différentes régions de cette séquence, nous allons utiliser un modèle de Markov caché avec deux états cachés 1 et 2. On a alors $S =$

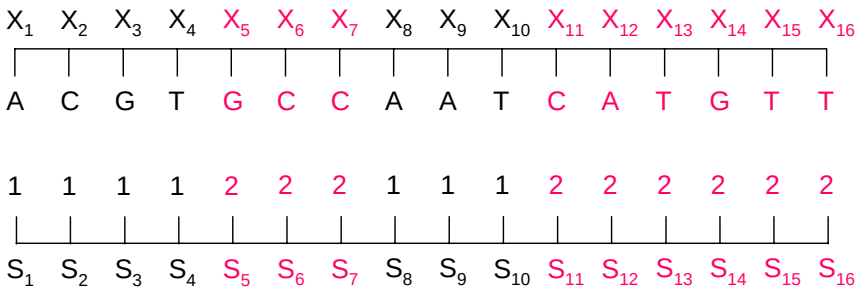

FIG. 2.4 – *Modèle de Markov caché à 2 états : identification de deux régions homogènes d'une séquence d'ADN*

$\{1, 2\}$. Chaque nucléotide A, C, G ou T de la séquence appartient donc soit à l'état caché 1, soit à l'état caché 2. La reconstruction de la séquence des états cachés nous permet ainsi de délimiter les deux types de régions homogènes.

Pour résoudre le problème de la reconstruction de la séquence des états cachés, les deux principaux algorithmes utilisés sont l'algorithme de Viterbi (Viterbi, 1967) et l'algorithme forward-backward (Rabiner, 1989, Durbin *et al.*, 1998). Ces deux algorithmes proposent une approche différente du problème. L'algorithme de Viterbi permet de reconstruire la séquence S des états cachés qui est la plus probable pour une séquence X alors que l'algorithme forward-backward permet de déterminer pour chaque position de la séquence l'état caché s_i qui est individuellement le plus probable.

2.3.3.1 Algorithme de Viterbi

Etant donnée une séquence observée $X = x_{1:n}$, il est possible de déterminer la séquence des états cachés $S^* = s^*_{1:n}$ correspondant à la séquence X la plus probable grâce à un algorithme de programmation dynamique connu

sous le nom d'algorithme de Viterbi.

$$S^* = \text{argmax}_{(S=s_{1:n})} P(S_{1:n} = s_{1:n} | X_{1:n} = x_{1:n})$$
$$S^* = \text{argmax}_{(S=s_{1:n})} P(X_{1:n} = x_{1:n}, S_{1:n} = s_{1:n})$$
$$S^* = \text{argmax}_{(S=s_{1:n})} P(X, S)$$

Cet algorithme est une adaptation au problème de la reconstruction de la séquence des états cachés par Viterbi en 1967 d'un algorithme plus général définit par Bellman en 1957. (Bellman, 1957)
Soit X une séquence de longueur n. Le principe de l'algorithme de Viterbi est de déterminer de manière récursive la probabilité de la chaîne S la plus probable pour X se terminant dans l'état k pour tout état $k \in \mathcal{S}$.
Notons

$$v_k(i) = \max_{S|S_i=k} P(x_1, \ldots, x_i, S)$$

la probabilité de la chaîne la plus probable permettant de générer le préfixe $x_{1:i}$ qui se termine dans l'état k. Il est possible de calculer les variables $v_k(i)$ grâce à la formule de récurrence sur i suivante :

$$\begin{aligned} v_l(i+1) &= \max_{S|S_{i+1}=l} P(x_1, \ldots, x_{i+1}, S) \\ &= P(X_{i+1} = x_{i+1} | S_{i+1} = l) \cdot \max_{k \in \mathcal{S}} [v_k(i) \cdot P(S_{i+1} = l | S_i = k)] \\ &= E(l, S_{i+1}) \cdot \max_{k \in \mathcal{S}} [v_k(i) \cdot T(k, l)] \end{aligned}$$

Le meilleur état final est mémorisé à chaque itération de l'algorithme de Viterbi dans un pointeur

$$\Pi_i(l) = \text{argmax}_{k \in \mathcal{S}} [v_k(i-1) \cdot T(k, l)]).$$

La chaîne la plus probable est alors obtenue en conservant la succession des états qui a permis d'obtenir S^* tel que

$$P(X, S^*) = \max_{k \in \mathcal{S}} v_k(n).$$

L'algorithme de Viterbi s'écrit alors :

1. Initialisation

$$v_k(1) = \nu_E(k)E(k, x_1) \qquad k \in \mathcal{S}$$

2. Récurrence

   Pour i allant de 2 à n :

$$v_l(i) = E(l, S_i) \cdot \max_{k \in \mathcal{S}} [v_k(i-1) \cdot T(k, l)] \qquad k \in \mathcal{S}$$
$$\Pi_i(l) = \text{argmax}_{k \in \mathcal{S}} [v_k(i-1) \cdot T(k, l)] \qquad k \in \mathcal{S}$$

3. Fin

$$P(X, S^*) = \max_{k \in \mathcal{S}} v_k(n)$$
$$s_n^* = \text{argmax}_{k \in \mathcal{S}} v_k(n)$$

4. Rétro-propagation

   Cette étape permet de retrouver la meilleure chaîne $S^* = s_{1:n}^*$ à partir des valeurs stockées dans le pointeur Π.

   Pour i allant de $n-1$ à 1 :

$$s_i^* = \Pi_{i+1}(s_{i+1}^*)$$

En pratique, l'exécution de l'algorithme de Viterbi conduit à multiplier entre elles de très petites probabilités. Pour éviter les erreurs de calculs, il est alors préférable de travailler plutôt avec $\log(v_l(i))$, ce qui limite les problèmes d'approximation que l'on peut rencontrer avec un ordinateur.

2.3.3.2 Algorithme forward-backward

Etant donnée une séquence observée $X = x_{1:n}$, une autre façon de reconstruire la séquence des états cachés $S^{**} = s_{1:n}^{**}$ correspondant à la séquence X est de déterminer, non pas la séquence la plus probable, mais, pour chaque position i de de la séquence, l'état caché s_i^{**} qui est individuellement le plus

probable. L'avantage de l'algorithme forward-backward est donc de maximiser le nombre d'états cachés individuellement corrects. Cependant, en déterminant séparément pour chaque position i l'état caché le plus probable, cet algorithme ne permet pas de prendre en compte les interactions entre états voisins. La séquence des états cachés obtenue à l'aide de l'algorithme forward-backward peut contenir des états voisins dont la probabilité de transition du premier vers le second est très faible, voire nulle.
L'algorithme forward-backward (Baum *et al.*, 1970) calcule pour chaque position i de la séquence, la probabilité de chaque état $k \in \mathcal{S}$, conditionnellement à la séquence observée X :

$$p_i(k) = P(S_i = k|X) \qquad k \in \mathcal{S}.$$

La quantité $p_i(k)$ peut être facilement déterminée à l'aide des variables forward $f_k(i)$ et backward $b_k(i)$ définies de la façon suivante :

$$\begin{aligned} f_k(i) &= P(X_{1:i} = x_{1:i}, S_i = k) \qquad k \in \mathcal{S}, \\ b_k(i) &= P(X_{i+1:n} = x_{i+1:n}|S_i = k, X_{i-m+1:i} = x_{i-m+1:i}) \qquad k \in \mathcal{S}. \end{aligned}$$

On a alors :

$$p_i(k) = \frac{f_k(i) \cdot b_k(i)}{P(X)}.$$

On n'a pas besoin de calculer $P(X)$ puisque, par exemple,

$$P(X) = \sum_{k \in \mathcal{S}} f_k(n) = \sum_{k \in \mathcal{S}} f_k(i) b_k(i).$$

L'algorithme forward-backward calcule les valeurs des variables $f_k(i)$ et $b_k(i)$ grâce à une récurrence "avant-arrière".

La récurrence avant : Elle permet de calculer les valeurs de la variable forward $f_k(i)$:

1. Initialisation

$$f_k(1) = \nu_E(k) E(k, x_1) \qquad k \in \mathcal{S}$$

2. Récurrence avant
   Pour i allant de 1 à $n-1$:

$$f_l(i+1) = E(l, x_{i-m+1:i}, x_{i+1}) \cdot \sum_{k \in \mathcal{S}} f_k(i) T(k,l).$$

La récurrence arrière : Elle permet de calculer les valeurs de la variable backward $b_k(i)$:

1. Initialisation

$$b_k(n) = 1 \qquad k \in \mathcal{S}$$

2. Récurrence arrière
   Pour i allant de $n-1$ à 1 :

$$b_k(i) = \sum_{l \in \mathcal{S}} E(l, x_{i-m+1:i}, x_{i+1}) b_l(i+1) T(k,l).$$

A partir des valeurs de $p_i(k)$ calculées grâce à l'algorithme forward-backard, il est possible de déterminer, pour chaque position i de la séquence, l'état caché s_i^{**} le plus probable :

$$s_i^{**} = \text{argmax}_{k \in \mathcal{S}} \left[p_i(k) \right] \qquad \text{pour } i = 1, \ldots, n.$$

2.3.4 Applications des modèles de Markov cachés à l'analyse des séquences génomiques

Les modèles de Markov cachés, ou HMM (Hidden Markov Models), ont été très largement utilisés pour l'analyse des séquences génomiques depuis les travaux de Churchill en 1989.

Un des domaines dans lesquel les modèles de Markov cachés ont été utilisés avec succès est l'alignement multiple de séquences. Krogh *et al.* et Baldi *et al.* ont développé en 1994 des modèles de Markov cachés pour le problème de l'alignement multiple de séquences de familles de protéines (Krogh *et al.*, 1994, Baldi *et al.*, 1994). Les modèles mis au point identifient les caractéristiques statistiques importantes de familles de protéines comme les globines,

les immunoglobines ou les kinases et peuvent ainsi être utilisés pour la classification des protéines ou l'alignement multiple des séquences.

D'autres modèles de Markov cachés ont été développés afin de permettre la prédiction des structures secondaires des protéines (Asai *et al.*, 1993), de la topologie des protéines de membranes (Krogh *et al.*, 2001) ou l'identification de sites importants sur le plan fonctionnel ou structurel sur des séquences de protéines de la famille des globines ou des phycobiliprotéines (Bickel *et al.*, 2002).

Un des domaines dans lequel les modèles de Markov cachés ont été le plus utilisés est la prédiction des gènes. Plusieurs logiciels basés sur des modèles de Markov cachés ont été développés comme VEIL (Henderson *et al.*, 1997) ou GENSCAN (Burge and Karlin, 1997). En 1994, Krogh *et al.* a developpé le logiciel ECOPARSE qui permet la détection des gènes dans le génome de la bactérie *Escherichia coli*. Ce logiciel a ensuite été généralisé pour donner HMMgene qui a été appliqué à l'homme (Krogh, 1997) et à la drosophile (Krogh, 2000). Le logiciel le plus populaire est GeneMark (Borodovsky and McIninch, 1993) ou son amélioration GeneMark.hmm (Lukashin and Borodovsky, 1998). Audic et Claverie ont également développé en 1998 une méthode de prédiction des régions codant les protéines dans les génomes microbiens (Audic and Claverie, 1998).

L'autre principale utilisation des modèles de Markov cachés est l'analyse de l'hétérogénéité des séquences. En 1989, Churchill a ouvert la voie de l'étude de l'hétérogénéité des séquences génomiques à l'aide des modèles de Markov cachés en les considérant comme une succession de segments homogènes qui peuvent être modélisés par les états cachés d'un modèle markovien. Il a ainsi utilisé les modèles de Markov cachés pour segmenter les ADN mitochondriaux de levures, de l'homme ou de la souris ainsi que le génome du bactériophage lambda (Churchill, 1989). Il a aussi étudié les variations de la distribution des dinucléotides CpG le long de la séquence de l'α-globine humaine (Churchill, 1992). Les modèles de Markov cachés ont également été appliqués à l'étude des fréquences des di-, tri-, et tétranucléotides (Burge *et al.*, 1992), à l'identification de régions homogènes dans les génomes du bactériophage lambda, du virus HIV-1, de la bactérie *Bacillus subtilis* (Muri,

1997, Nicolas *et al.*, 2002), de la levure *Saccharomyces cerevisiae* (Peshin and Gelfand, 1999) et de l'intron 7 du gène α-fétoprotéine du chimpanzé et de l'homme (Boys *et al.*, 2000) (Boys *et al.* 2000). Ils sont également très performants dans la recherche des éléments transposables (Andrieu *et al.*, 2004).

Chapitre 3

Prédiction des régions constantes et variables du gène *env* d'EIAV

Sommaire

La région du gène *env* d'EIAV qui code la glycoprotéine de surface est composée d'une succession de régions variables (V), présentant de nombreuses mutations, et de régions constantes (C), présentant peu de mutations. Afin de déterminer s'il existe des signatures spécifiques des régions C et V codées par les séquences de la SU du lentivirus EIAV, des modèles de Markov cachés ont été développés.

Cette section présente les différents modèles de Markov cachés que nous avons développés afin de prédire la succession des régions constantes et variables le long de la région du gène *env* d'EIAV codant la SU.

3.1 Choix de l'ordre et du nombre d'états cachés des modèles

Des modèles de Markov cachés de différents ordres m ont été utilisés pour tenter de différencier les régions C et V de la SU d'EIAV. A l'heure actuelle, il n'existe pas de procédure statistique bien définie permettant de déterminer l'ordre idéal d'un modèle de Markov caché. Selon l'ordre m choisi, un modèle de Markov caché M1-Mm ne prendra pas en compte la même information. Les régions homogènes mises en évidence ne seront *a priori* pas les mêmes. Ainsi, une région homogène identifiée par un modèle M1-M0 est caractérisée par les fréquences d'apparition des différents nucléotides alors qu'avec un modèle M1-M1, ce sont les fréquences d'apparition des dinucléotides qui caractérisent les régions. Un modèle ayant un ordre trop petit ne permettra pas de rendre compte des propriétés biologiques liées aux oligonucléotides. D'un autre côté, la mise au point d'un modèle d'ordre élevé nécessite l'estimation d'un grand nombre de paramètres et augmente ainsi les risques liés au surentraînement des modèles. Afin de déterminer l'ordre du modèle de Markov caché le plus adapté au problème de la segmentation des régions constantes et variables d'EIAV, nous avons considéré plusieurs modèles en augmentant progressivement l'ordre des modèles.

Le nombre d'états cachés utilisés dans la modélisation correspond au nombre de types de régions homogènes que l'on cherche à identifier dans les séquences étudiées. Des modèles ayant $N = 2$ états cachés ont été développés pour tenter de différencier deux types de régions homogènes, correspondants aux régions C et V. Des modèles à $N = 9$ d'états cachés ont été également définis pour leur permettre d'utiliser un état caché pour décrire chaque région variable V1 à V8 et un autre état caché pour décrire l'ensemble des régions constantes. La quantité d'informations différentes disponible sur les régions

constantes étant limitée par la redondance des séquences dans ces régions, il ne semble pas possible de tenter de définir un état caché caractéristique de chacune des 9 régions constantes sans surentraîner les modèles. Soulignons que rien n'impose aux modèles stochastiques, tels que nous les avons définis, d'utiliser les N états à sa disposition pour décrire le découpage de la SU d'EIAV en régions C et V. Les modèles pourraient, par exemple, regrouper certaines régions C et V en un même état. De même, rien ne dit *a priori* que différentes portions d'une même région, C ou V, ne soient pas décrites par des états différents. Il semble en tout cas évident que notre démarche ne peut aboutir que si les séquences étudiées sont fortement structurées quant à leurs propriétés statistiques.

3.2 Modèles prédictifs des régions C et V de la SU d'EIAV

Dans le but de différencier les régions constantes et variables de la SU d'EIAV, différents modèles de Markov cachés ont été développés. Pour cela, 187 séquences de la région du gène *env* d'EIAV codant la glycoprotéine de surface ont été utilisées. Parmi les séquences disponibles, 94 séquences, sélectionnées de façon aléatoire, ont constitué l'ensemble d'apprentissage des modèles. Cet ensemble a permis d'estimer les différents paramètres des modèles de Markov cachés. Les modèles définis ont ensuite été testés sur les 93 séquences qui n'avaient pas été utilisées lors de l'élaboration des modèles et qui formaient donc l'ensemble de test.

3.2.1 Estimation par l'algorithme de Baum-Welch

Les paramètres de modèles de Markov cachés de différents ordres m ont été estimés à l'aide de l'algorithme de Baum-Welch. Cet algorithme permet de déterminer les paramètres des modèles de Markov cachés sans connaissances préalables concernant la longueur, la position ou la composition des différentes régions homogènes à identifier. Des modèles avec $N = 2$ et $N = 9$

états cachés ont été définis sur les séquences nucléotidiques de la SU d'EIAV. Des modèles d'ordre 0, 1 ou 2 ont également été entraînés avec l'algorithme de Baum-Welch sur des séquences déduites en acides aminés. Cependant, aucun des modèles de Markov cachés dont les paramètres ont été estimés avec l'algorithme de Baum-Welch ne permet d'identifier les régions constantes et variables de la SU d'EIAV. Quels que soient l'ordre et le nombre d'états cachés du modèle considéré, la séquence des états cachés prédite oscille et passe continuellement d'un état caché à l'autre (Figure 3.1). Il est important de noter que, pour chaque séquence de l'échantillon de test et pour chaque position, un des états cachés apparaît comme nettement plus probable, avec une probabilité de l'ordre de $0,9$. En d'autres termes, les oscillations entre les états cachés ne proviennent pas d'une trop grande similarité entre les régions identifiées mais plutôt de l'incapacité du modèle à identifier des plages homogènes de longueur significative.

Ainsi, cette méthode ne permet pas d'identifier des régions homogènes qui correspondent aux régions constantes et variables de la SU d'EIAV.

3.2.2 Estimation par l'algorithme de Baum-Welch avec matrice d'émission fixée

Les modèles de Markov cachés dont les paramètres ont été estimés avec l'algorithme de Baum-Welch ne permettant pas de différencier les régions C et V de la SU d'EIAV, nous avons donc défini une variante de l'algorithme de Baum-Welch permettant d'estimer les paramètres de modèles de Markov cachés. Ce nouvel algorithme, basé sur l'algorithme de Baum-Welch et la méthode du maximum de vraisemblance lorsque la séquence des états cachés est connue, permet d'introduire de l'information sur la composition des différentes régions. Nous allons décrire en détails le fonctionnement de cet algorithme original, mis au point spécifiquement pour cette problématique.

Le nouvel algorithme diffère de l'algorithme classique de Baum-Welch dans l'estimation de la matrice d'émission E. Alors que l'algorithme de Baum-Welch permet de converger vers l'estimateur du maximum de vraisemblance de la matrice E après une phase d'estimation de la position des

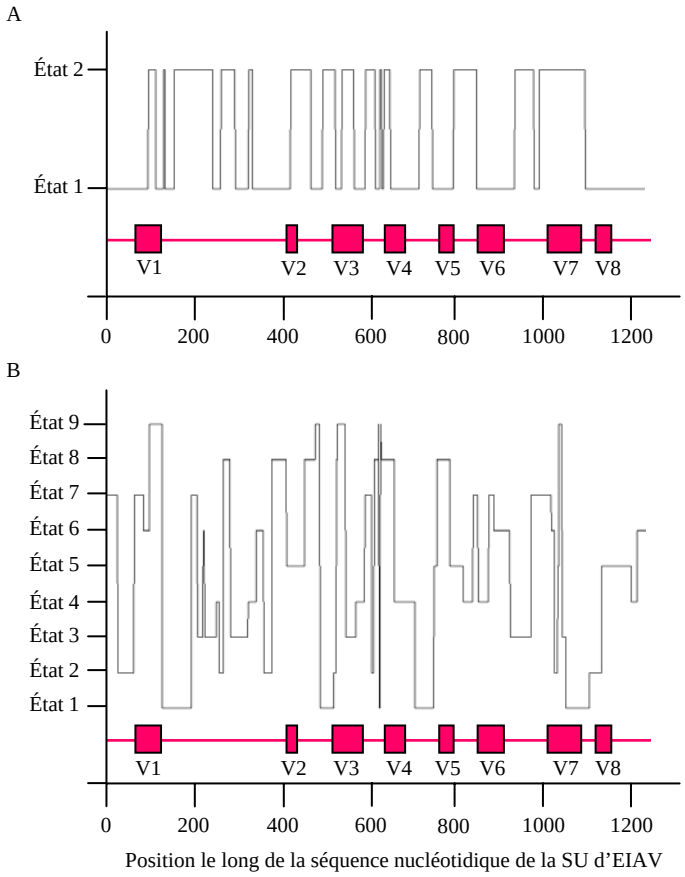

FIG. 3.1 – *Régions prédites par un modèle de Markov caché dont les paramètres ont été estimés avec l'algorithme de Baum-Welch entraîné sur les séquences nucléotidiques de la SU d'EIAV.* Le graphique présente les régions prédites par le modèle mathématique (—). L'organisation schématique de la SU d'EIAV, avec la position des 8 régions variables V1 à V8 (■) et des 9 régions constantes (—), est représentée en dessous du graphique. (A) Modèle de Markov caché d'ordre 5 avec 2 états cachés. (B) Modèle de Markov caché d'ordre 5 avec 9 états cachés.

différentes régions, le nouvel algorithme calcule directement l'estimateur du maximum de vraisemblance de la matrice E par comptages des fréquences des mots de nucléotides ou d'acides aminés sur chacune des régions telles qu'elles ont pu être définies préalablement grâce à un alignement de séquences. Pour cela, les séquences d'entraînement sont alignées à l'aide du programme ClustalX utilisant les paramètres par défaut. Les différentes régions constantes et variables sont ensuite identifiées par comparaison avec les régions constantes et variables précédemment définies (Leroux *et al.*, 1997b). La matrice d'émission E peut être vue comme un assemblage des matrices d'émission des observations pour chaque état. Le nouvel algorithme consiste alors à estimer séparément les matrices d'émission de chaque type de région. Les matrices d'émission des différents états sont ensuite assemblées pour former la matrice d'émission E. Pour chaque type de région, la matrice d'émission correspondante est estimée de façon classique en comptant le nombre de passages d'un ensemble de m lettres à la $m+1$-ième puis en divisant par le nombre total de passages par cet ensemble de m lettres, m représentant l'ordre du modèle de Markov caché. Un pseudo-compte de valeur $+1$ est également ajouté au nombre de passages afin d'éviter que certains paramètres de la matrice d'émission E ne soient nuls et de limiter les risques de surentraînement. Une fois la matrice d'émission E estimée, le nouvel algorithme estime les probabilités de transition entre états de la matrice T à l'aide de l'algorithme de Baum-Welch, mais en gardant toutes les probabilités d'émission à leur valeur calculée. La phase M de l'algorithme de Baum-Welch est modifiée afin d'omettre la maximisation des probabilités d'émission. Enfin, la phase E de l'algorithme de Baum-Welch et la phase M' de maximisation des probabilités de transition du nouvel algorithme sont exécutées alternativement comme dans l'algorithme de Baum-Welch classique.

Ce nouvel algorithme étant une modification de l'algorithme de Baum-Welch qui permet de fixer la valeur de la matrice E, nous l'avons appelé algorithme de Baum-Welch avec matrice d'émission fixée. Concrètement, l'algorithme de Baum-Welch avec matrice d'émission fixée produit une suite de modèles $(T_n, E_n)_{n\geq 0}$ de la façon suivante :

Phase Initialisation : La matrice des probabilités de transition entre états T_0 est initialisée de manière aléatoire. La matrice d'émission E_0 est définie par comptages directs sur les séquences d'entraînement :

$$E_0(k, a_{1:m}, a_{m+1}) = \frac{e(k, a_{1:m}, a_{m+1})}{\sum_{a \in \mathcal{A}} e(k, a_{1:m}, a)},$$

où $e(k, a_{1:m}, a_{m+1})$ représente le nombre d'occurrences observées du mot $a_{1:m}a_{m+1}$ alors que l'état correspondant à l'observation a_{m+1} est k.

Phase Estimation (E) : Calcul de la probabilité $P_{k,\ell}$ de deux états successifs k et ℓ sous la valeur courante (T_n, E_n) pour tous $k \in \mathcal{S}$ et $\ell \in \mathcal{S}$ et pour la séquence $X = x_{1:n}$.

$$P_{k,\ell} = P(S_i = k, S_{i+1} = \ell | X, (T_n, E_n)).$$

Cette probabilité peut être calculée de la même façon que dans l'algorithme de Baum-Welch à l'aide des variables forward $f_k(i)$ et backward $b_k(i)$ définies par :

$$f_k(i) = P(x_1, ..., x_i, S_i = k),$$

et

$$b_k(i) = P(x_{i+1}, ..., x_n \,|\, S_i = k, x_{i-m+1}, ..., x_i).$$

On a :

$$P_{k,\ell} = \frac{f_k(i) \cdot T_n(k, \ell) \cdot E_n(\ell, x_{i-m+1:i}, x_{i+1}) \cdot b_\ell(i+1)}{\sum_{k \in \mathcal{S}} f_k(n)},$$

Phase Maximisation (M') : Calcul de (T_{n+1}, E_{n+1}) à partir des formules

$$T_{n+1}(k, \ell) = \frac{t(k, \ell)}{\sum_{s \in \mathcal{S}} t(k, s)},$$

où

$$t(k, \ell) = \sum_{i=1}^{n-1} f_k(i) \cdot T_n(k, \ell) \cdot E_n(\ell, x_{i-m+1:i}, x_{i+1}) \cdot b_\ell(i+1),$$

et

$$E_{n+1} = E_0.$$

Phase Fin : Les phases E et M' sont exécutées alternativement jusqu'à convergence de l'algorithme.

L'algorithme de Baum-Welch avec matrice d'émission fixée converge vers les estimateurs du maximum de vraisemblance $(\widetilde{T}, E)$ sachant la matrice d'émission E. Le modèle $(\widetilde{T}, E)$ fournit une vraisemblance plus faible que le modèle $(\widehat{T}, \widehat{E})$ obtenu grâce à l'algorithme de Baum-Welch classique.

A l'aide de l'algorithme de Baum-Welch avec matrice d'émission fixée que nous venons de définir, nous avons estimé les paramètres de modèles à $N = 2$ états cachés de différents ordres m. Un modèle de Markov caché d'ordre 2 dont les paramètres ont été estimés avec le nouvel algorithme fournit une délimitation claire de certaines régions variables de la SU d'EIAV pour toutes les séquences nucléotidiques de l'ensemble de test. Cependant, toutes les régions variables ne sont pas identifiées par un modèle M1-M2 (Figure 3.2A). En considérant des modèles d'ordre supérieur ($m \geq 3$), nous améliorons encore la prédiction des régions C et V. Les régions C et V de la plupart des séquences de l'échantillon de test sont prédites correctement avec des modèles d'ordre 3 ou 4 sur les observations avec, en général, moins d'une dizaine de nucléotides prédits dans une mauvaise région sur les 1350 nucléotides qui composent en moyenne les séquences de test. Pour un modèle M1-M3, seuls les nucléotides des régions variables V2 et V5 sont mal reconnus et ils sont en général prédits comme appartenant à une région constante. Pour $m = 5$, les régions prédites par le modèle de Markov caché correspondent presque parfaitement aux régions définies à partir d'un alignement de séquences pour toutes les séquences de l'échantillon de test (Figure 3.2B).

Afin de différencier les régions variables V1 à V8 et les régions constantes, des modèles de Markov cachés à $N = 9$ états ont été définis. Pour cela, nous avons entraîné un état caché avec chacune des huit régions variables et un état caché avec l'ensemble des régions constantes. Un modèle de Markov caché d'ordre $m = 5$ dont les paramètres ont été estimés avec l'algorithme de Baum-Welch avec matrice d'émission fixée entraîné sur des séquences nucléotidiques fournit une délimitation précise des régions C et V. De plus, chaque région variable présente un signal distinct de celui des autres régions variables

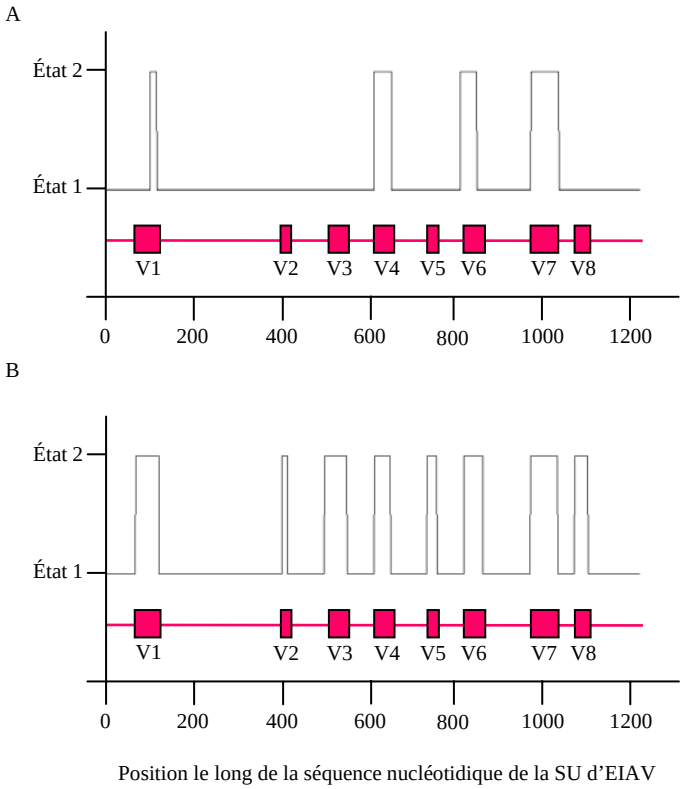

FIG. 3.2 – *Régions prédites par un modèle de Markov caché à 2 états dont les paramètres ont été estimés avec l'algorithme de Baum-Welch à matrice d'émission fixée entraîné sur les séquences nucléotidiques de la SU d'EIAV.* Le graphique présente les régions prédites par le modèle mathématique (—). L'organisation schématique de la SU d'EIAV, avec la position des 8 régions variables V1 à V8 (■) et des 9 régions constantes (—), est représentée en dessous du graphique. (A) Modèle de Markov caché d'ordre 2 avec 2 états cachés. (B) Modèle de Markov caché d'ordre 5 avec 2 états cachés.

(Figure 3.3).

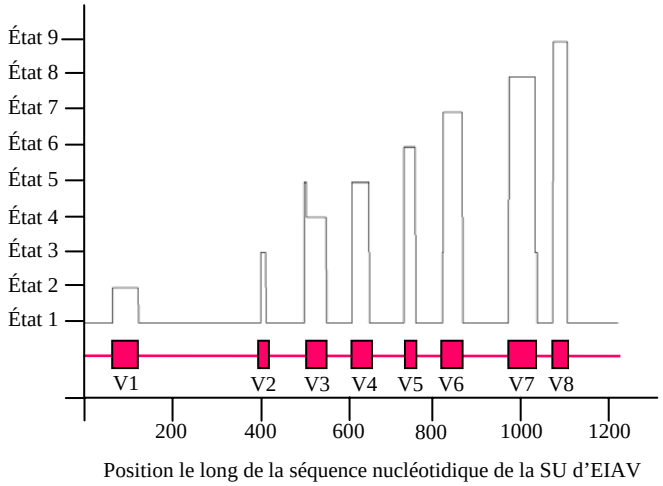

FIG. 3.3 – *Régions prédites par un modèle de Markov caché d'ordre 5 à 9 états dont les paramètres ont été estimés avec l'algorithme de Baum-Welch à matrice d'émission fixée entraîné sur les séquences nucléotidiques de la SU d'EIAV.* Le graphique présente les régions prédites par le modèle mathématique (—). L'organisation schématique de la SU d'EIAV, avec la position des 8 régions variables V1 à V8 (■) et des 9 régions constantes (—), est représentée en dessous du graphique.

Des modèles de Markov cachés à 2 ou à 9 états cachés ont égalemnt été entraînés avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences déduites en acides aminés de la SU d'EIAV. Un modèle de Markov caché d'ordre 1 à 2 états est capable de localiser les régions C et V avec une grande précision (Figure 3.4A). Un modèle d'ordre 1 avec 9 états cachés qui permet une bonne prédiction des régions constantes et de chacune des régions variables de la SU d'EIAV a aussi été défini (Figure 3.4B).

3.2.3 Estimation par maximum de vraisemblance

Après avoir aligné les séquences d'entraînement et déterminé l'emplacement des régions variables, nous avons calculé les estimateurs du maximum

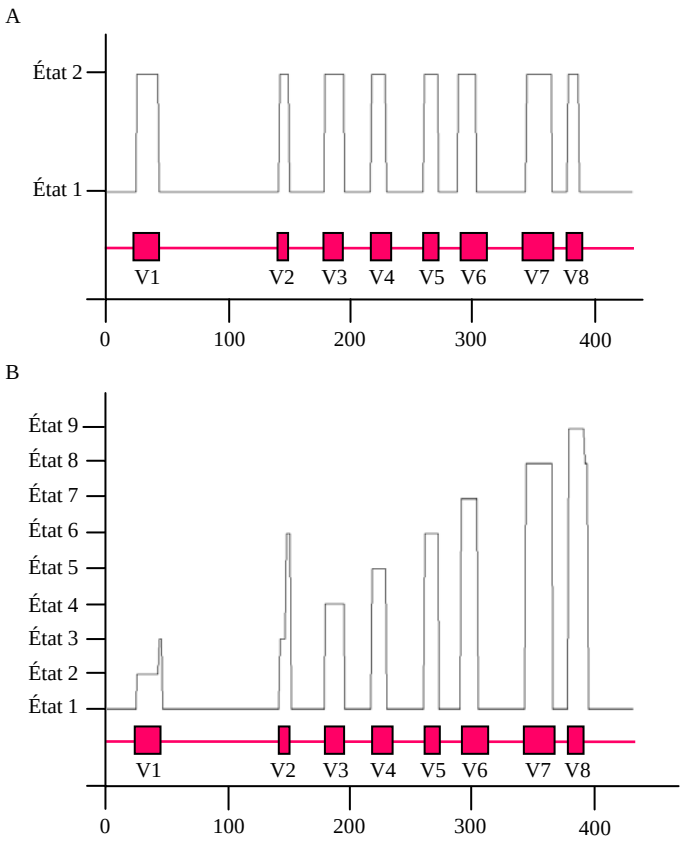

FIG. 3.4 – *Régions prédites par des modèles de Markov cachés d'ordre 1 dont les paramètres ont été estimés avec l'algorithme de Baum-Welch à matrice d'émission fixée entraîné sur les séquences déduites en acides aminés de la SU d'EIAV.* Le graphique présente les régions prédites par le modèle mathématique (—). L'organisation schématique de la SU d'EIAV, avec la position des 8 régions variables V1 à V8 (■) et des 9 régions constantes (—), est représentée en dessous du graphique. (A) Modèle de Markov caché avec 2 états, entraîné sur les régions constantes et variables. (B) Modèle de Markov caché avec 9 états, entraîné sur les régions 8 régions variables et la réunion des régions constantes.

de vraisemblance des paramètres de modèles de Markov cachés de plusieurs ordres. Les probabilités d'émission de chaque état ont été estimées à l'aide de comptages directs des fréquences des mots de nucléotides ou d'acides aminés de longueur $m + 1$, où m est l'ordre du modèle, sur chaque type de région. Les probabilités de transition entre états ont également été estimées à partir des fréquences de transition comptées sur les séquences d'entraînement. Que ce soit pour les modèles à $N = 2$ ou $N = 9$ états cachés, les résultats sont similaires à ceux obtenus en estimant les paramètres des modèles à l'aide de l'algorithme de Baum-Welch avec matrice d'émission fixée.

3.3 Conclusions

Nous avons montré qu'il était possible d'identifier les régions constantes et variables de la SU d'EIAV à l'aide de modèles de Markov cachés, aussi bien sur des séquences nucléotidiques que sur des séquences déduites en acides aminés. Les séquences des états cachés reconstruites à partir de modèles basés sur l'algorithme de Baum-Welch avec matrice d'émission fixée n'oscillent pas entre les différents états comme dans les modèles basés sur l'algorithme de Baum-Welch classique. L'information concernant les fréquences des mots de nucléotides ou d'acides aminés introduite dans les modèles à l'aide de l'algorithme de Baum-Welch avec matrice d'émission fixée permet une meilleure modélisation des compositions statistiques des différents types de régions. La matrice de transition obtenue sans information préalable sur la longueur et l'ordre des régions permet d'identifier de longues régions statistiquement homogènes qui peuvent ensuite être comparées aux régions C et V précédemment définies. Les régions prédites par les modèles correspondent presque parfaitement aux C et V définies par alignements. La région constante globale utilisée dans nos modèles pour modéliser l'ensemble des régions constantes C1 à C9 ne correspond à aucune région réelle mais représente une moyenne de l'ensemble des ces régions. Cependant, nos résultats montrent que les régions constantes ont des compositions statistiques en nucléotides ou en acides aminés suffisamment proches pour être reconnues par un seul état caché. De la

même façon, les régions variables partagent une certaine homogénéité dans leur composition et peuvent être reconnues ensemble dans un modèle à 2 états cachés. Chaque région variable conserve également une composition spécifique qui peut être reconnue par un état distinct dans un modèle à 9 états cachés.

3.4 Etude de la robustesse des modèles sur EIAV

L'algorithme de Baum-Welch avec matrice d'émission fixée a permis de mettre au point des modèles de Markov cachés, d'ordre 5 sur les séquences nucléotidiques et d'ordre 1 sur les séquences déduites en acides aminés, capables de prédire presque parfaitement la succession des régions C et V de la SU d'EIAV. Afin de tester la robustesse des modèles définis et de vérifier que ces modèles n'avaient pas subi de biais lors de leur élaboration, nous avons soumis les modèles développés à divers tests.

3.4.1 Test du surentraînement des modèles

Un biais possible des modèles de Markov que nous avons développés est un éventuel surentraînement. Celui-ci apparaît lorsque le nombre de paramètres estimés pour définir le modèle est trop important par rapport à la quantité de données dont on dispose pour entraîner le modèle. Ce problème concerne donc principalement les modèles d'ordre élevé ayant un grand nombre d'états cachés comme le modèle M1-M5 à 9 états cachés sur les séquences nucléotidiques qui est beaucoup plus riche en paramètres que le modèle M1-M1 à 2 états cachés sur les séquences déduites en acides aminés. En effet, la matrice d'émission E du modèle M1-M5 sur les séquences nucléotidiques comprend 36 864 termes alors que la matrice d'émission E du modèle M1-M1 sur les séquences déduites en acides aminés n'en comprend que 800.

Lorsque le nombre de données est insuffisant par rapport au nombre de paramètres à estimer, le modèle va avoir tendance à "sur-apprendre" les données dont il dispose. Au niveau de la matrice d'émission E d'un modèle de Markov caché, cela se traduit par une sur-représentation des émissions obser-

vées sur les séquences d'apprentissage. Ainsi, les émissions bien représentées dans les séquences d'apprentissage vont apparaître avec une forte probabilité, au détriment des émissions sous-représentées dans l'échantillon d'apprentissage. La matrice d'émission E va être une matrice "déséquilibrée" avec certains coefficients très élevés (proches de 1) et d'autres plus faibles (proches de 0). Le modèle va alors simplement apprendre à reconnaître des motifs particuliers de longueur $m+1$, où m est l'ordre du modèle, sur chaque type de région. Nous avons ainsi vérifié que les modèles développés n'étaient pas trop spécifiques des données d'entraînement et qu'ils demeuraient capables d'identifier les régions constantes et variables de nouvelles séquences de test.

Afin de tester si les modèles n'avaient pas été surentraînés, nous avons vérifié que les matrices d'émission E des modèles développés était bien équilibrées. En étudiant les termes des matrices E, il n'apparaît pas de grandes différences entre les coefficients de chaque matrice qui semblent tous être du même ordre. Il est possible que les différences entre les coefficients soient également diminuées par l'introduction d'un pseudo-compte lors de l'estimation de la matrice E de chaque modèle.

Nous avons également vérifié que les modèles n'identifiaient pas les régions C et V en reconnaissant des motifs spécifiques de chaque type de région mais avaient bien réussi à caractériser chaque région par sa composition en mots de nucléotides ou d'acides aminés. Pour cela, nous avons entraîné des modèles de Markov cachés en utilisant des séquences d'apprentissage ayant le moins possible de motifs en commun avec les séquences de test. Les séquences d'apprentissage et les séquences de test n'ayant alors quasiment aucun motif commun, surtout dans les régions variables, il était impossible au modèle d'apprendre à reconnaître des motifs sur les séquences d'apprentissage puis de retrouver les mêmes motifs sur les séquences de test pour identifier les régions C et V. Afin d'obtenir des séquences d'apprentissage aussi différentes que possible des séquences de test, nous avons entraîné les modèles sur des séquences du virus qui étaient présentes au début de la maladie induite chez les chevaux par EIAV et nous les avons testés sur des séquences issues de prélèvements réalisés au cours d'épisodes fébriles survenant plusieurs mois après le début de l'infection (Leroux *et al.*, 1997b). A cause de la variabilité

du virus tout au long de l'évolution de la maladie, les séquences d'entraînement et de test présentaient $7,8\%$ ($\pm 1,3$) de divergence au niveau des acides aminés. Les séquences d'entraînement et de test présentaient en particulier $43,8\%$ ($\pm 20,2$) de divergence dans la troisième région variable V3 (Figure 3.5A).

Les divergences importantes entre les séquences d'apprentissage et les séquences de test ne perturbent pas les modèles de Markov cachés développés qui parviennent à prédire correctement les régions constantes et les régions variables, notamment la région V3, de séquences n'ayant aucun motif commun avec les séquences de l'échantillon d'apprentissage. Les modèles mis au point ne se contentent donc pas de différencier les types de régions en identifiant uniquement des motifs caractéristiques de chaque région et ne semblent pas souffrir d'un problème de surentraînement.

3.4.2 Influence de l'ordre et de la position des régions variables

Un autre biais possible des modèles de Markov cachés que nous avons définis, notamment les modèles à 9 états cachés, est que ces modèles aient appris que toutes les séquences sont composées de 8 régions variables qui apparaissent toujours dans le même ordre et quasiment à la même position le long des séquences. Ces modèles identifieraient alors les 8 régions variables en se référant à leur ordre et leur position et non à leur composition en nucléotides ou en acides aminés.

Afin de tester ce biais, des séquences assemblées artificiellement à partir des séquences disponibles ont été créées. Dans ces séquences artificielles, le nombre, l'ordre, la taille et la position des régions variables étaient différents de ceux des séquences réelles. Nous avons, par exemple, introduit, en plus des 8 régions variables habituelles, une copie de 15 acides aminés de la région variable V7 dans la région constante C2 séparant V1 et V2 (Figure 3.5B). Les séquences ainsi créées étaient constituées de 9 régions variables. De plus, l'ordre et la position des régions variables étaient alors inédits puisque une nouvelle région V7 succédait à la région V1.

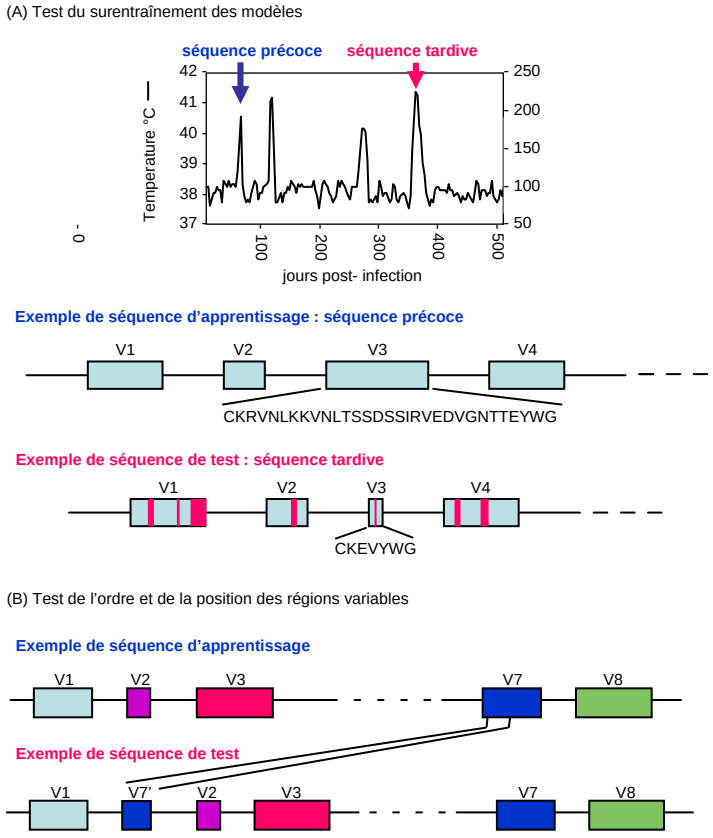

FIG. 3.5 – *Construction des séquences d'apprentissage et des séquences de test permettant de tester la robustesse des modèles.* (A) Les séquences d'apprentissage sont issues de prélèvements précoces et les séquences de test sont issues de prélèvement tardifs. La structure schématique des séquences d'apprentissage et de test avec la position des régions variables est représentée. Les mutations survenues dans les régions variables sont modélisées en pointillés. (B) La structure schématique des séquences d'apprentissage et de test avec la position des régions variables est représentée.

Les modèles entraînés avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur les séquences originales ont été testés sur les séquences assemblées artificiellement. Ils permettent de prédire parfaitement l'enchaînement des nouvelles régions C et V le long des séquences artificielles et notamment la région variable additionnelle (Figure 3.6).

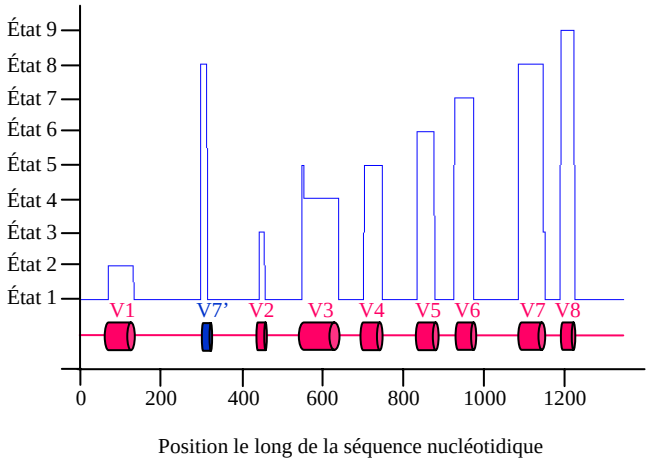

FIG. 3.6 – *Régions prédites par un modèle de Markov caché sur une séquence assemblée artificiellement.* Le graphique représente les régions prédites par un modèle de Markov caché d'ordre 5 entraîné avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences nucléotidiques et testé sur une séquence assemblée artificiellement dans laquelle 15 acides aminés de la région variable V7 ont été insérés dans la région constante C2.

Les modèles de Markov mis au point dans les sections précédentes ne s'appuient donc ni sur l'ordre, ni sur la longueur, ni sur la position des régions variables pour arriver à les différencier des régions constantes. Au contraire, ils doivent se baser sur des différences statistiques entre les compositions en mots de nucléotides ou d'acides aminés de longueur $m+1$, où m est l'ordre du modèle, pour prédire la succession des régions C et V le long de la SU d'EIAV.

Chapitre 4

Etude de la séparation des régions constantes et variables du gène *env* d'EIAV

Sommaire

Nous avons développé des modèles de Markov cachés capables de différencier les régions C et V du lentivirus EIAV. Ces modèles s'appuient sur la composition en mots de nucléotides ou d'acides aminés de chaque type de région. La capacité des modèles développés à distinguer, avec une grande précision, les régions C et V de la SU d'EIAV suggère donc que chaque type

de région possède une composition en mots de nucléotides et d'acides aminés qui lui est propre.

Dans ce chapitre, nous allons analyser plus en détail les différences entre les compositions en acides aminés des régions C et V de la région du gène *env* d'EIAV qui code la glycoprotéine de surface. Nous nous intéresserons à la nature des acides aminés présents dans les régions constantes et variables. Nous effectuerons ensuite une analyse en composantes principales afin de décrire les ressemblances et dissemblances entre les compositions en mots d'acides aminés des régions C et V d'EIAV. Enfin, nous définirons une distance afin de quantifier les différences entre les régions constantes et variables de la SU d'EIAV.

4.1 Analyse descriptive de quelques propriétés chimiques

Des modèles de Markov cachés d'ordre 1 permettent de prédire avec une grande précision l'alternance des régions C et V le long de séquences déduites en acides aminés de la SU d'EIAV. Ce résultat suggère une différence importante dans la composition en mots d'acides aminés de longueur 2 entre les ces deux types de régions. Dans cette section, nous allons déterminer si les acides aminés qui composent les régions constantes et les régions variables ont des propriétés chimiques différentes.

Les acides aminés sont des molécules organiques. Ils ont tous en commun un squelette carboné et deux groupes fonctionnels : une amine et un acide carboxylique. La différence entre les vingt acides aminés se situe au niveau de leur radical. On dénombre vingt radicaux différents. Classiquement, on regroupe les acides aminés par famille, en fonction des propriétés chimiques de leur radical. Par exemple, les acides aminés peuvent être basiques ou acides, hydrophobes ou hydrophiles (Figure 4.1).

Nous avons calculé les fréquences d'apparition des acides aminés acides, basiques, hydrophiles et hydrophes dans les 9 régions constantes et les 8 régions variables de la SU d'EIAV. Pour chacune de ces quatres propriétés,

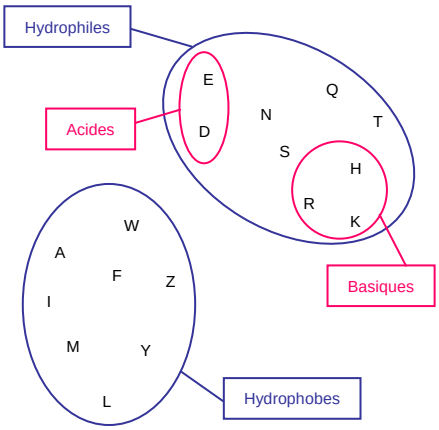

FIG. 4.1 – *Diagramme de Wenn de quelques propiétés chimiques des acides aminés.*

nous avons comparé les fréquences d'apparition des acides aminés dans les régions constantes et dans les régions variables. Des tests de Mann-Whitney ont également été réalisés afin de déterminer si les acides aminés apparaissaient plus fréquemment dans un certain type de région en fontion de leurs propriétés chimiques (Figure 4.2).

Les comparaisons effectuées montrent que les acides aminés qui composent les régions constantes et ceux qui composent les régions variables de la SU d'EIAV ne diffèrent pas significativement par leurs propriétés acido-basiques. En revanche, les acides aminés hydrophiles apparaissent significativement plus fréquemment dans les régions variables que dans les régions constantes. Inversement, les acides aminés hydrophobes sont plus fréquents dans les régions constantes.

Les acides aminés dont les radicaux sont hydrophobes ont peu d'affinités avec les molécules d'eau entourant la protéine. La chaîne a tendance à se replier de façon à les regrouper au centre de la molécule, sans contact direct avec le milieu aqueux. Inversement, les acides aminés hydrophiles ont tendance à se disposer à la périphérie de façon à être en contact avec l'eau. La différence de composition en acides aminés hydrophobes et hydrophiles entre les régions constantes et variables peut donc être mise en rapport avec la structure tridimensionnelle de la glycoprotéine de surface. Les régions constantes vont

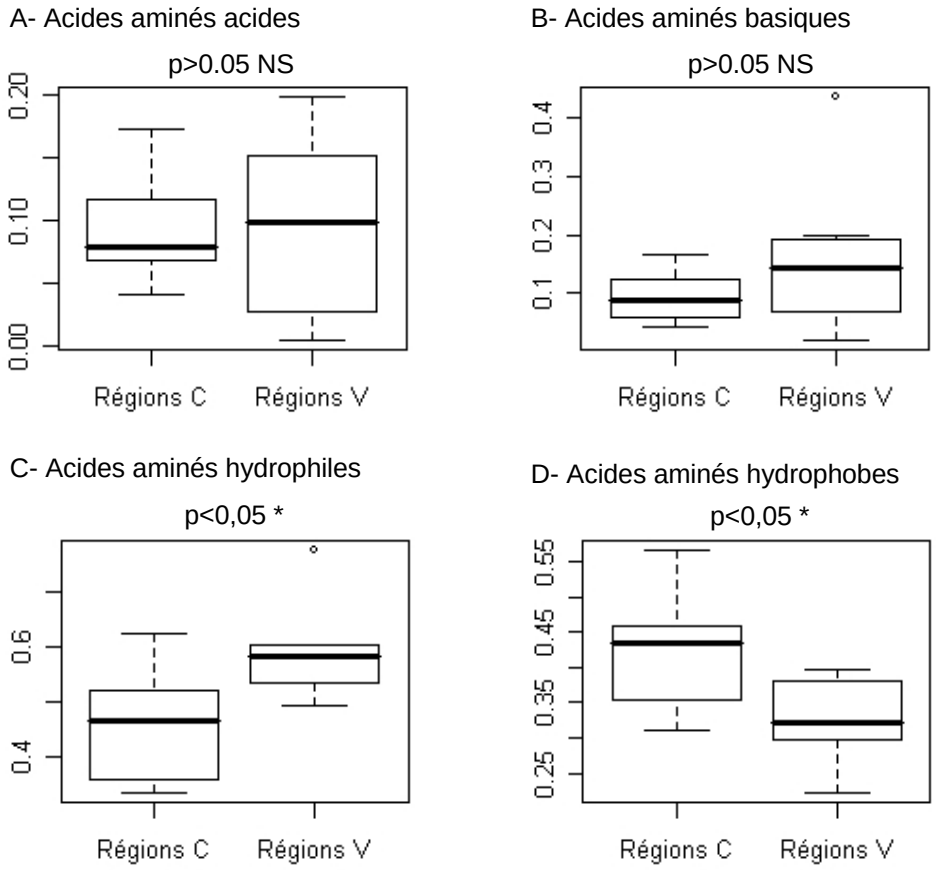

FIG. 4.2 – *Fréquences d'apparition des acides aminés dans les régions constantes et variables de la SU d'EIAV en fonction de leurs propriétés chimiques.* Les fréquences d'apparition dans les régions C et V des acides aminés acides (A), basiques (B), hydrophiles (C) et hydrophes (D) sont représentées sous formes de boxplots. Les boxplots, ou boîtes à moustaches, sont constitués d'une boîte délimitée en bas par le premier quartile Q1 et en haut par le troisième quartile Q3. La médiane, ou deuxième quartile, est représentée par un trait horizontal à l'intérieur de la boîte. L'extrémité de la moustache inférieure correspond à la valeur minimum qui est supérieure à Q1-1,5(Q3-Q1) et celle de la moustache supérieure à la valeur maximum qui est inférieure à Q3+1,5(Q3-Q1). Les valeurs extrêmes sont indiquées par des ronds. Les p-valeurs des tests de Mann-Whitney sont indiquées au dessus des graphiques. NS signifie que les fréquences d'apparition ne sont pas significativement différentes entre les régions C et V.

avoir tendance à se retrouver au centre de la structure tertiaire alors que les régions variables de l'enveloppe seront exposées à la surface de la particule virale.

4.2 Analyse en Composantes Principales

Dans les modèles de Markov cachés capables de prédire la succession des régions C et V des séquences déduites en acides aminés la SU d'EIAV, chaque type de région est représenté par un modèle de Markov d'ordre 1. Ces modèles sont basés sur les fréquences d'apparition des $20 * 20 = 400$ mots de deux acides aminés dans chaque type de région. Cela signifie que chaque type de région peut être décrit par 400 variables quantitatives représentant les fréquences des mots d'acides aminés de longueur 2.

L'ACP (Analyse en Composantes Principales) est une méthode statistique classique pour la description synthétique de tableaux de données dans lesquels des individus sont décrits par des variables quantitatives multiples. Cette méthode permet une réduction de l'information. Les variables quantitatives qui décrivent les individus étudiés sont regroupées au sein de facteurs synthétiques appelés composantes principales. Les individus sont ensuite positionnés par rapport à ces composantes principales issues d'une combinaison linéaire des variables descriptives. Le positionnement des individus par rapport aux composantes principales nouvellement définies peut mettre en évidence des groupes d'individus ainsi que les variables qui ont amené à la création de ces groupes.

Une ACP normée a été réalisée afin de tenter de séparer les 17 régions de la SU d'EIAV en deux groupes, un groupe correspondant aux 9 régions constantes, l'autre aux 8 régions variables, en fonction de la fréquence des mots de deux acides aminés dans ces régions. Les régions C et V de la SU d'EIAV ont été projetées sur le plan décrit par les deux premières composantes principales (Figure 4.3).

Il apparaît que l'ACP ne permet pas de séparer les régions de la SU d'EIAV en deux groupes, que ces groupes correspondent aux régions constantes

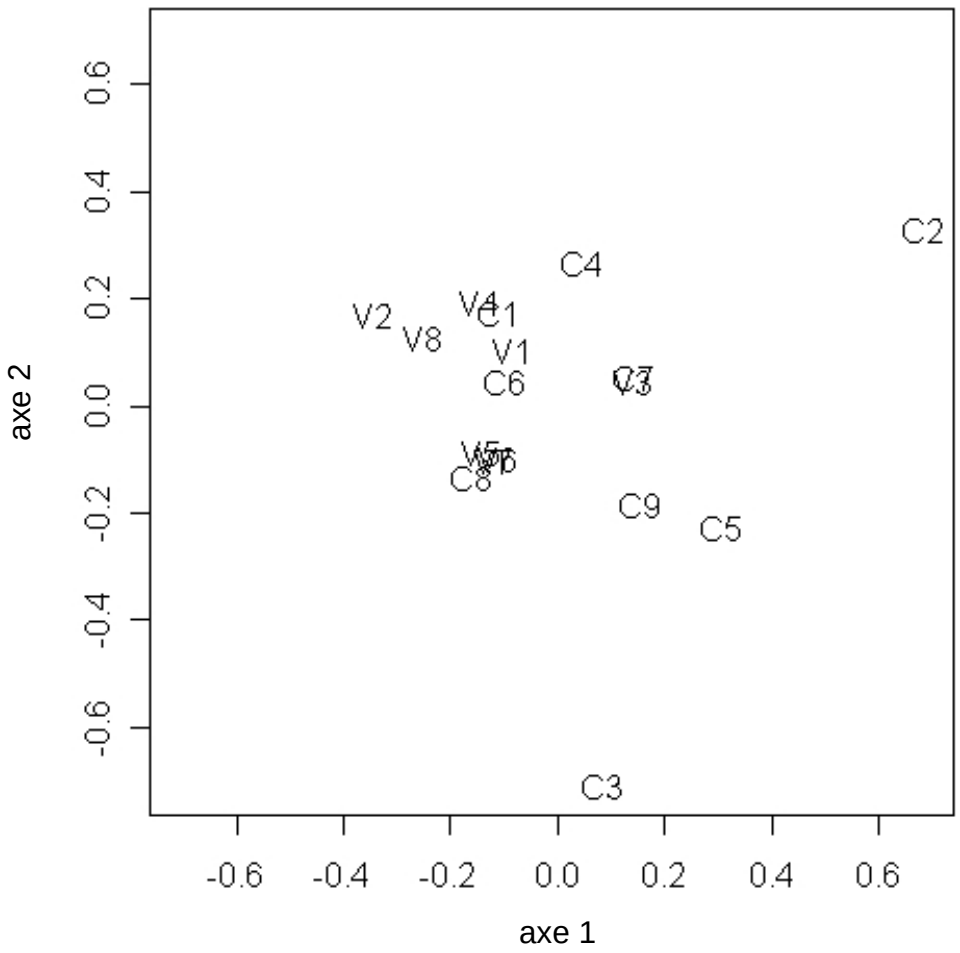

FIG. 4.3 – *Analyse en composantes principales des régions constantes et variables de la SU d'EIAV.* Une ACP a été effectuée sur les régions C et V de la SU d'EIAV en utilisant les fréquences des mots de deux acides aminés comme variables. Le graphique représente la projection des 9 régions constantes C1 à C9 et des 8 régions variables V1 à V8 sur le plan défini par les deux premières composantes principales. L'axe 1 et l'axe 2 sont les axes principaux associés aux deux composantes principales qui permettent de décrire le mieux les régions C et V de la SU d'EIAV. Ces deux axes permettent d'expliquer 23,3% de la variance.

d'une part et variables d'autre part ou non. En se basant sur l'ACP, toutes les régions semblent avoir des compositions en mots de deux acides aminés très proches.

Alors que les modèles de Markov cachés basés sur les fréquences des mots de deux acides aminés permettent de différencier les régions C et V de la SU d'EIAV, l'analyse en composantes principales basée sur ces mêmes fréquences ne parvient pas à séparer ces deux types de régions. Les modèles de Markov cachés sont donc plus adaptés au problème de la séparation des régions constantes et variables de la SU d'EIAV et révèlent une différence plus subtile entre le groupe des régions constantes et celui des régions variables. De simples combinaisons linéaires des fréquences des mots de deux acides aminés ne sont pas suffisantes pour discriminer ces deux groupes de régions.

4.3 Quantification de la séparation des régions constantes et variables d'EIAV

Les modèles de Markov développés ont permis de différencier l'ensemble des régions constantes de l'ensemble des régions variables mais également de distinguer chacune des 8 régions variables. Ce résultat indique que les régions constantes et variables possèdent des compositions statistiques en mots de nucléotides et d'acides aminés distinctes. Les 8 régions variables de la SU d'EIAV sont également statistiquement différentes, chacune pouvant être modélisée par une chaîne de Markov cachée qui lui est spécifique.

Une méthode pour quantifier les différences entre les chaines de Markov qui représentent les régions C et V d'EIAV est de considérer l'entropie relative, aussi appelée divergence de Kullback-Leibler, entre ces modèles (Billingsley, 1961b, Miller, 1955, Johnson and Sinanovic, 2001, Victor, 2000, Paninski, 2003, Pritchard and Scott, 2004).

4.3.1 Utilisation d'une métrique

Notons P et Q les matrices de transition de deux chaînes de Markov et π la mesure stationnaire associée à la matrice P. Soit $x = x_{1:n}$ une séquence de longueur n émise selon la matrice d'émission P. D'après la loi des grands nombres usuelle pour les chaînes de Markov, le nombre d'indices k entre 1 et $n-1$ tels que $x_k = i$ et $x_{k+1} = j$ est presque sûrement asymptotiquement, quand n tend vers l'infini, de la forme $n\pi(i)P(i,j) + o(n)$, où $o(n)$ désigne un terme T asymptotiquement négligeable, c'est-à-dire que la limite de T/n vaut 0. La vraisemblance de la suite d'observations x selon la matrice de transition P vaut donc

$$\ell_P(x) = \prod_{(i,j)} P(i,j)^{n\pi(i)P(i,j)} \cdot \varepsilon_P(x),$$

où le terme $\varepsilon_P(x)$ est exponentiellement négligeable, c'est-à-dire que la limite de $(1/n)\log(\varepsilon_P(x))$ vaut 0, ce que l'on notera $\varepsilon_P(x) = \exp(o(n))$.

De même, la probabilité d'émettre l'observation x selon la matrice de transition Q est donnée par

$$\ell_Q(x) = \prod_{(i,j)} Q(i,j)^{n\pi(i)P(i,j)} \cdot \varepsilon_Q(x),$$

où $\varepsilon_Q(x) = \exp(o(n))$.

On a donc

$$\ell_Q(x) = \ell_P(x) \cdot \exp(-nH(P|Q) + o(n)),$$

où $H(P|Q)$ désigne l'entropie relative des deux chaînes de Markov P et Q et vaut

$$H(P|Q) = \sum_{i,j} \pi(i)P(i,j) \log \frac{P(i,j)}{Q(i,j)}.$$

En général, $H(P|Q)$ et $H(Q|P)$ ne coïncident pas. Nous utilisons une forme symétrisée de l'entropie relative définie par

$$\delta(P,Q) = H(P|Q) + H(Q|P).$$

Ainsi, $\delta(P,Q) = \delta(Q,P)$. Comme il se doit, la vraisemblance de x sous le modèle Q est asymptotiquement négligeable devant sa vraisemblance sous

le modèle P puisque x est engendrée par P. On a donc $H(P|Q) \geq 0$ et $H(P|Q) > 0$ dès que $P \neq Q$. Il en résulte que $\delta(P, Q) = 0$ si et seulement si $P = Q$. Cependant, il est important de noter que δ n'est pas une distance au sens mathématique du terme puisqu'elle ne vérifie pas l'inégalité triangulaire : il existe des matrices P, Q et R telles que $\delta(P, R) > \delta(P, Q) + \delta(Q, R)$.

4.3.2 Etude de la distance entre les régions constantes et variables d'EIAV

L'entropie relative symétrisée δ a été calculée entre les chaînes de Markov qui modélisent les 9 régions constantes C1 à C9 et les 8 régions variables V1 à V8 de la SU d'EIAV (Tableau 4.1).

TAB. 4.1 – *Entropie relative symétrisée entre les régions constantes et variables de la SU d'EIAV.* Le tableau indiquent les valeurs des entropie relatives $\delta(R_i, R_j)$ où R_i représente la ligne du tableau et R_j la colonne. La valeur $4,81$ indiquée en rouge dans le tableau représente ainsi la valeur de l'entropie relative $\delta(C5, V1)$.

δ	C1	C2	C3	C4	C5	C6	C7	C8	C9	C	V	V1	V2	V3	V4	V5	V6	V7	V8
C1	0	4.40	5.08	4.63	4.65	5.08	4.26	4.70	3.52	2.55	3.70	3.62	2.80	3.67	3.77	3.37	3.70	4.00	3.25
C2		0	6.25	5.77	5.93	6.38	5.90	4.65	5.37	2.58	4.60	4.46	3.47	4.53	4.72	4.96	4.94	4.72	3.78
C3			0	6.00	6.62	5.69	5.77	5.35	5.35	2.85	4.86	4.78	4.07	4.51	4.71	4.87	5.18	4.76	4.73
C4				0	6.36	5.71	5.22	5.06	5.09	2.65	4.95	4.70	3.48	4.85	5.14	5.04	4.13	5.26	4.28
C5					0	6.86	6.08	6.17	5.00	2.97	4.88	4.81	3.64	4.96	5.35	4.95	5.41	4.71	4.54
C6						0	6.21	4.70	5.25	3.18	4.70	4.60	3.40	4.75	4.74	4.75	4.67	5.20	3.90
C7							0	4.92	4.85	2.37	4.71	5.15	3.93	5.16	4.99	4.93	4.63	4.47	4.88
C8								0	4.34	2.66	4.68	3.77	3.20	4.25	4.11	4.80	4.04	4.50	3.36
C9									0	2.48	3.84	3.92	3.22	4.16	4.32	4.13	4.84	4.00	3.32
C										0	3.35	3.64	3.19	3.70	3.78	3.62	3.99	2.96	3.01
V											0	2.25	2.07	1.91	2.11	2.09	2.21	1.87	2.36
V1												0	2.41	3.79	3.26	3.07	3.59	3.64	3.57
V2													0	2.79	2.73	2.71	2.59	2.77	2.34
V3														0	3.63	3.77	3.60	4.54	3.41
V4															0	3.57	3.66	4.09	4.03
V5																0	4.08	3.77	3.69
V6																	0	4.25	3.52
V7																		0	3.83
V8																			0

Il apparaît que, quelle que soit la région constante considérée, la matrice d'émission qui modélise cette région Ci est toujours plus proche de la matrice d'émission C qui modélise l'ensemble des régions constantes que de n'importe

quelle matrice modélisant les régions variables Vj. En d'autres termes, on a, pour tous i et j,

$$\delta(Ci, C) < \delta(Ci, Vj).$$

De façon similaire, les différentes régions variables Vi sont toujours plus proches de la région globale V que de n'importe quelle région constante Cj et on a, pour tous i et j,

$$\delta(Vi, V) < \delta(Vi, Cj).$$

Afin de visualiser les ressemblances et dissemblances entre les régions constantes et variables de la SU d'EIAV, les entropies relatives symétrisées, utilisées pour étudier les distances entre ces régions, ont été représentées sous la forme d'un dendogramme. Bien que l'entropie relative symétrisée δ ne soit pas une vraie métrique, il est possible de représenter graphiquement les distances entre les différentes régions par un arbre non enraciné, calculé à l'aide du programme Kitsch (Phylip 3.5c) en utilisant la matrice des distances précédemment définie (Figure 4.4).

Le dendogramme montre une séparation entre deux groupes, le premier groupe étant constitué des régions constantes C1 à C9 et de la région globale C, le second des régions variables V1 à V8 et de la région globale V. La séparation entre les régions constantes et variables de la SU d'EIAV est valable quel que soit l'ordre dans lequel les régions sont entrées dans le programme Kitsch.

4.3.3 Utilisation d'un test statistique basé sur l'entropie relative

Afin de quantifier davantage la séparation entre les régions C et V de la SU d'EIAV, nous avons mis au point un test statistique asymptotique qui permet de différencier, ou non, les matrices de transition empiriques de deux régions différentes. Ce test repose sur la convergence en loi de l'entropie relative vers une loi du χ^2. Après avoir démontré cette convergence en loi, nous l'utiliserons pour construire un test statistique que nous appliquerons

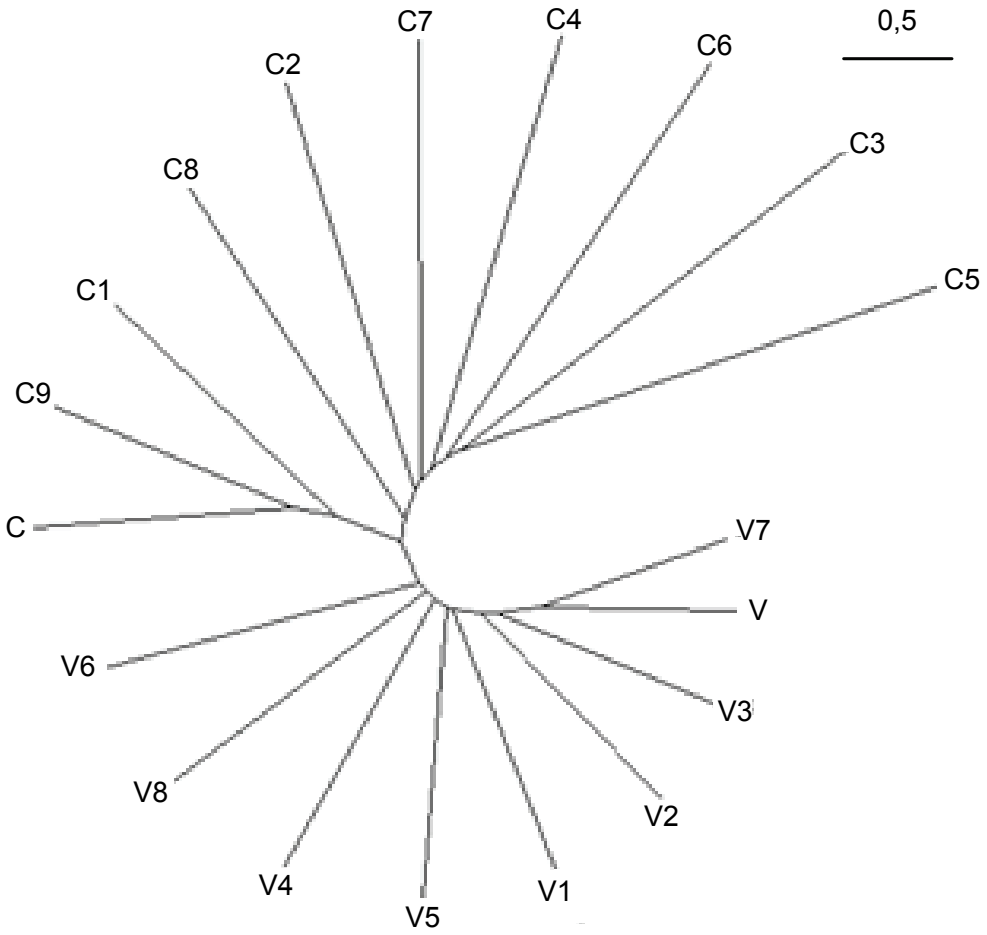

FIG. 4.4 – *Représentation graphique des distances entre les régions constantes et variables d'EIAV.* Une matrice des distances entre les régions C et V de la SU d'EIAV a été calculée en utilisant la forme symétrisée de l'entropie relative δ entre les chaînes de Markov modélisant ces régions. Un dendogramme a été évalué avec le programme Kitsch (Phylip 3.5c) utilisant les paramètres définis par défaut puis représenté à l'aide du programme Unrooted (http ://pbil.univ-lyon1.fr/software/unrooted.html).

aux matrices de transition empiriques estimées sur les différentes régions de la SU d'EIAV.

4.3.3.1 Démonstration d'une convergence en loi

Considérons une chaîne de Markov irréductible sur un nombre fini k d'états, de matrice de transition q et de distribution stationnaire p, avec p_x strictement positif pour tout x. On note ℓ le nombre d'arêtes utilisées par la chaîne, c'est-à-dire le nombre de couple d'états (x, y) tels que $q(x, y) > 0$. On a donc $k \leq \ell \leq k^2$. Lorsque $\ell = k$, la chaîne de Markov se déplace de façon déterministe sur un cercle discret orienté. Ce cas peut être exclu si on le souhaite. En revanche, dès que $\ell \geq k + 1$, plusieurs trajectoires sont possibles et la chaîne est réellement aléatoire. Enfin, $\ell = k^2$ signifie que toutes les transitions sont autorisées et que la chaîne se déplace sur le graphe complet augmenté de toutes les boucles. La dimension $D(q)$ de la chaîne est donnée par

$$D(q) := \ell - k.$$

Cet entier correspond au nombre de paramètres libres parmi les coefficients non nuls de la matrice q, c'est-à-dire à la dimension du simplexe formé par les matrices de transition subordonnées à q, au sens où tous les coefficients qui correspondent à des coefficients nuls de q doivent être nuls également.

L'estimateur du maximum de vraisemblance $\widehat{q}$ de q obtenu par comptages le long d'une trajectoire de longueur n est un estimateur consistant de q lorsque n tend vers l'infini. La relation

$$\widehat{q}(x, y) = q(x, y) + z_{xy}/\sqrt{n} + o(1/\sqrt{n}),$$

permet de définir un vecteur $(z_{xy})_{xy}$ gaussien centré, indexé par les arêtes (x, y), dont la matrice de covariance est une fonction explicite de p et q.

Considérons l'entropie relative de la mesure empirique, fournie par la trajectoire observée, par rapport à la mesure théorique, fournie par p et q. Cette entropie aléatoire est définie par

$$H(\widehat{q}|q) := \sum_{(x,y)} \widehat{p}_x \widehat{q}(x, y) \log(\widehat{q}(x, y)/q(x, y)),$$

où la somme sur (x, y) comprend ℓ termes et $\widehat{p}$ représente la mesure stationnaire de $\widehat{q}$. On peut également considérer

$$H(q|\widehat{q}) := \sum_{(x,y)} p_x q(x,y) \log(q(x,y)/\widehat{q}(x,y)).$$

En utilisant le développement de Taylor au deuxième ordre de la fonction logarithme, on constate que les deux quantités $H(\widehat{q}, q)$ et $H(q, \widehat{q})$ sont telles que, lorsque n devient grand,

$$H = h/(2n) + o(1/n), \quad h := \sum_{(x,y)} z_{xy}^2 p_x / q(x,y).$$

Nous allons montrer que l'entropie relative réduite h suit une loi du χ^2.

On rappelle que si l'on considère une séquence i.i.d. de longueur n, de loi p sur k états, alors la fréquence empirique $\widehat{p}$ est telle que $2nH(\widehat{p}, p)$ converge en loi vers une loi du χ^2 avec $k-1$ degrés de liberté.

Soit N_x, respectivement N_{xy}, le nombre de passages par le sommet x, respectivement l'arête (x, y) jusqu'au temps n. Considérons

$$\xi_{xy} := (N_{xy} - q(x,y)N_x)/\sqrt{N_x}.$$

D'après (Billingsley, 1961a), la matrice $(\xi_{xy})_{xy}$ converge en loi, lorsque n tend vers l'infini, vers une matrice gaussienne centrée $(g_{xy})_{xy}$. Cette matrice est telle que les lignes $(g_{xy})_y$ sont indépendantes, c'est-à-dire que $\mathbb{E}(g_{xy}g_{zt}) = 0$ pour tous $x \neq z$ et tous y et t. Enfin, pour tous x, y et z,

$$\mathbb{E}(g_{xy}g_{xz}) = -q(x,y)q(x,z) \quad (y \neq z), \qquad \mathbb{E}(g_{xy}^2) = q(x,y)(1 - q(x,y)).$$

Comme N_x/n converge presque sûrement vers p_x, la normalisation $1/\sqrt{N_x}$ peut être remplacée par $1/\sqrt{np_x}$. Ceci permet d'obtenir la convergence en loi

$$np_x(\widehat{q}(x,y) - q(x,y))^2 \to g_{xy}^2.$$

Ainsi, les vecteurs $(g_{xy})_y$ sont indépendants. De plus, pour chaque x fixé, $(g_{xy})_y$ admet les mêmes covariances que la loi gaussienne limite obtenue pour une suite i.i.d. de loi $q(x, \cdot)$. Par ailleurs, la variable aléatoire

$$H_x := \sum_y q(x,y) \log(q(x,y)/\widehat{q}(x,y))$$

correspond à l'observation d'un processus i.i.d. pendant un temps qui correspond au nombre de visites de x avant n, soit $np_x + o(n)$.

On en déduit que $2(np_x)H_x$ converge en loi vers une loi du χ^2 avec $D_x(q)$ degrés de libertés, où $D_x(q)+1$ représente le nombre de sommets y tels que $q(x,y) > 0$. Par indépendance des limites en loi des $2np_xH_x$, leur somme $2nH$ converge en loi vers un χ^2 avec $D(q)$ degrés de libertés, où $D(q)$ est la somme sur x des $D_x(q)$.

En conclusion, l'entropie relative réduite h suit une loi du χ^2 avec $D(q)$ degrés de liberté.

4.3.3.2 Définition d'un test statistique

A partir du résultat précédent, nous avons défini un test d'égalité du χ^2 qui permet de décider si les matrices de transition empiriques $\widehat{q}_1$ et $\widehat{q}_2$ de deux chaînes de Markov sont ou non des estimateurs de la même matrice de transition théorique q.

Loi asymptotique

Supposons que l'on dispose de deux observations indépendantes de longueur n de la même chaîne de Markov de matrice de transition q. Les deux observations fournissent deux estimateurs $\widehat{q}_1$ et $\widehat{q}_2$ de q qui vérifient les relations

$$\widehat{q}_i(x,y) = q(x,y) + z_{xy}^{(i)}/\sqrt{n} + o(1/\sqrt{n}), \quad i = 1, 2,$$

où les vecteurs $(z_{xy}^{(1)})_{xy}$ et $(z_{xy}^{(2)})_{xy}$ sont indépendants et suivent la même loi que le vecteur $(z_{xy})_{xy}$ défini précédemment.

L'entropie relative réduite entre les deux observations vaut asymptotiquement

$$h(\widehat{q}_1, \widehat{q}_2) := \sum_{(x,y)} (z_{xy}^{(1)} - z_{xy}^{(2)})^2 \alpha_{xy},$$

où α_{xy} peut valoir indifféremment $p_x^{(1)}/q_1(x,y)$ ou $p_x^{(2)}/q_2(x,y)$ ou $p_x/q(x,y)$. Si l'on utilise $\alpha_{xy} = p_x/q(x,y)$, alors $h(\widehat{q}_1, \widehat{q}_2)$ suit exactement la loi de $2h(\widehat{q}, q)$. Cette relation reste asymptotiquement vraie pour les autres choix de α_{xy}.

Pour déterminer si $\widehat{q}_1$ et$\widehat{q}_2$ correspondent à la même chaîne de Markov q ou non, il est possible de réaliser un test d'égalité du χ^2 utilisant le fait que $nH(\widehat{q}_1, \widehat{q}_2)$ est asymptotiquement χ^2 avec $D(q)$ degrés de liberté. En particulier,

$$\mathbb{E}(H(\widehat{q}_1, \widehat{q}_2)) \sim D(q)/n.$$

Dans le cas plus général où $\widehat{q}_1$ et $\widehat{q}_2$ sont basés sur des séquences indépendantes de longueurs respectives n_1 et n_2, il est possible d'obtenir le même type de résultat. Dans ce cas, $\ell H(\widehat{q}_1, \widehat{q}_2)$ est asymptotiquement χ^2 avec $D(q)$ degrés de liberté, où ℓ représente la moyenne harmonique de n_1 et n_2 définie par la relation

$$\frac{2}{\ell} = \frac{1}{n_1} + \frac{1}{n_2}.$$

Le même type de résultat est également valable lorsque l'on utilise l'entropie relative symétrisée δ. Dans ce cas, la loi de $\frac{1}{2}\ell\delta(\widehat{q}_1, \widehat{q}_2)$ est asymptotiqment χ^2 avec $D(q)$ degrés de liberté et, en particulier,

$$\mathbb{E}(\delta(\widehat{q}_1, \widehat{q}_2)) \sim 2D(q)/\ell.$$

p-valeurs approchées

Afin de décider si deux matrices de transition empiriques correspondent à la même chaîne de Markov, il suffit donc de réaliser un test d'égalité en utilisant la convergence en loi de l'entropie relative vers un χ^2. Cela conduit à déterminer la p-valeur d'un évènement $\{X \geq t\}$ où X suit une loi du χ^2 à d degrés de liberté. Lorsque le calcul simple des p-valeurs d'un χ^2 de grande dimension n'est pas possible, une majoration efficace des p-valeurs peut être obtenue grâce aux bornes exponentielles de Cramér. On obtient ainsi que, pour tout $t \geq d$, la probabilité qu'une loi du χ^2 à d degrés de liberté soit supérieure à t est majorée par

$$e^{-t/2}(te/d)^{d/2}.$$

Intéressons-nous, par exemple, à une loi du χ^2 à $d = 380$ degrés de liberté. Cette approximation nous indique que la p-valeur pour $t = 460$, qui vaut $0,3\%$, est inférieure à $2,45\%$, et que la p-valeur pour $t = 480$, qui vaut $0,03\%$, est inférieure à $0,04\%$.

4.3.3.3 Application aux régions constantes et variables d'EIAV

Le test statistique asymptotique défini dans la section précédente a été appliqué aux régions C et V de la SU d'EIAV. Les valeurs de l'entropie relative symétrisée δ précédemment calculées ont été utilisées afin de déterminer si les matrices de transition empiriques obtenues à partir des séquences déduites en acides aminés de deux régions différentes pouvaient refléter la même chaîne de Markov q. Tous les coefficients des matrices de transitions empiriques étant strictement positifs grâce aux pseudo-comptes introduits, la chaîne q est de dimension $D(q) = 400 - 20 = 380$.

Les tests d'égalité du χ^2 entre les chaînes de Markov modélisant les 9 régions constantes et les 8 régions variables ont fourni des p-valeurs très proches de 0. La plus grande p-valeur a été obtenue pour les régions variables V1 et V2 et est égale à $4 \cdot 10^{-17}$. Les p-valeurs étant très petites, nous pouvons conclure que les chaînes de Markov définies pour modéliser les régions C et V de la SU d'EIAV ne reflétaient pas la même composition statistique en mots de deux acides aminés. Ainsi, chacune des 9 régions constantes et des 8 régions variables possède une signature qui lui est spécifique.

4.3.4 Conclusions

La définition d'une distance entre les chaînes de Markov modélisant les 9 régions constantes et les 8 régions variables de la SU d'EIAV permet d'analyser la séparation entre ces régions. La représentation graphique sous la forme d'un dendogramme de la matrice des distances entre les différentes régions montre une séparation entre le groupe des régions constantes et le groupe des régions variables. De plus, l'analyse des distances entre les régions confirme que la région globale C introduite dans les modèles que nous avons développés permet de modéliser correctement l'ensemble des régions constantes. Toutes les régions constantes étant plus proches de la région C que de n'importe quelle région variable, cette région peut être utilisée par les modèles développés pour différencier les régions constantes des régions variables. De la même façon, la région globale V offre une bonne modélisation de l'ensemble des régions variables.

Le test statistique que nous avons défini montre que les matrices de transition qui modélisent les 17 régions de la SU d'EIAV reflètent des compositions en acides aminés différentes. Ceci confirme la différence entre les régions C et V mais également le fait que chaque région variable possède une signature qui lui est propre et peut donc être modélisée par un état caché spécifique.

Chapitre 5

Prédiction des régions constantes et variables du gène *env* des lentivirus

Sommaire

Dans les chapitres précédents, nous avons défini des modèles de Markov à deux états cachés capables de distinguer les régions C et V sur les séquences nucléotidiques et déduites en acides aminés de la SU d'EIAV. Ceci suggère que les séquences de la région du gène *env* d'EIAV codant la glycoprotéine de surface sont fortement structurées quant à leur composition en mots de nucléotides ou d'acides aminés.

Ce chapitre est consacré à l'extension de ce résultat à tous les autres lentivirus. Des modèles à 2 états cachés ont été développés afin de différencier les régions C et V des lentivirus HIV, SIV, SRLV, BIV et FIV.

5.1 Modèles prédictifs des régions C et V de HIV

5.1.1 Utilisation des modèles prédictifs des régions C et V d'EIAV

Les modèles entraînés sur les séquences de la région du gène *env* d'EIAV codant la glycoprotéine de surface ne permettent pas de différencier les régions C et V de la SU de HIV. Nous avons tenté d'identifier les régions C et V le long de séquences nucléotidiques de HIV à l'aide de modèles de différents ordres entraînés sur des séquences nucléotidiques de la SU d'EIAV. Aucun des modèles testés ne permet de prédire la succession des régions C et V de la SU de HIV (Figure 5.1A). De même, un modèle de Markov caché d'ordre 1 entraîné sur des séquences déduites en acides aminés de la SU d'EIAV ne permet pas de prédire la succession des régions constantes et variables de la SU de HIV (Figure 5.1B).

5.1.2 Définition de modèles spécifiques prédictifs des régions C et V de HIV

Des modèles de Markov cachés à 2 états spécifiques ont été développés afin d'identifier les régions constantes et variables sur les séquences nucléotidiques et déduites en acides aminés de la SU de HIV. Pour entraîner et tester ces modèles, nous avons utilisé un panel de 155 séquences représentatives des différents sous-types de HIV-1. Parmi les séquences disponibles, 78 séquences, sélectionnées de façon aléatoire, ont constitué l'ensemble d'entraînement des modèles. Les modèles définis ont ensuite été testés sur les 77 séquences qui n'ont pas été utilisées lors de l'élaboration des modèles et qui forment l'en-

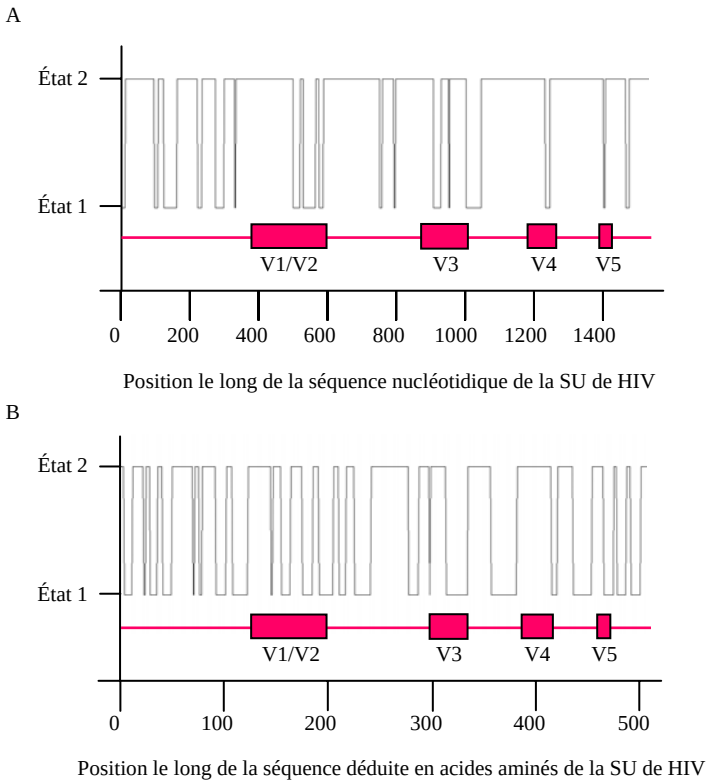

FIG. 5.1 – *Régions prédites sur une séquence de la SU de HIV par un modèle de Markov caché entraîné sur des séquences de la SU d'EIAV.* Les graphiques représentent les régions prédites sur la séquence HIV-1 HXB2 par des modèles de Markov cachés entraînés sur des séquences d'EIAV (—). L'organisation schématique de la SU de HIV, avec la position des régions variables V1 à V5 (■) et des régions constantes (—), est représentée en dessous des graphiques. (A) Modèle de Markov caché d'ordre 5 avec deux états entraîné sur des séquences nucléotidiques de la SU d'EIAV. (B) Modèle de Markov caché d'ordre 1 avec deux états entraîné sur des séquences déduites en acides aminés de la SU d'EIAV.

semble de test. La séquence HXB2, utilisée classiquement comme séquence de référence en virologie, a toujours été intégrée à l'ensemble de test afin que tous les modèles développés puissent être testés sur la même séquence.

L'algorithme de Baum-Welch ne permettant pas de définir des modèles capables de différencier les régions C et V, l'algorithme de Baum-Welch avec matrice d'émission fixée a été utilisé pour estimer les paramètres de modèles de Markov à 2 états cachés.

Des modèles de Markov cachés de différents ordres ont été testés afin de prédire la succession des régions constantes et variables sur les séquences nucléotidiques de la SU de HIV. A partir de l'ordre 2, certaines régions variables sont correctement identifiées sur la plupart des séquences de l'ensemble de test. Un modèle d'ordre 5 permet de localiser avec une grande précision la totalité des régions C et V de la SU de HIV sur toutes les séquences de l'ensemble de test (Figure 5.2).

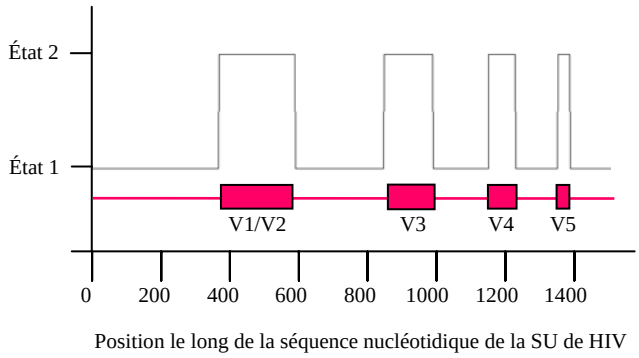

FIG. 5.2 – *Régions prédites par un modèle de Markov caché d'ordre 5 avec deux états entraîné avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences nucléotidiques de la SU de HIV.* Le graphique représente les régions prédites sur la séquence de test HIV-1 HXB2 par un modèle de Markov caché entraîné sur des séquences de HIV (—). L'organisation schématique de la SU de HIV, avec la position des régions variables V1 à V5 (■) et des régions constantes (—), est représentée en dessous du graphique.

L'algorithme de Baum-Welch avec matrice d'émission fixée a été utilisé

pour estimer les paramètres de modèles de Markov cachés à deux états cachés à partir de séquences déduites en acides aminés de la SU de HIV. Un modèle d'ordre 1 permet de prédire avec une grande précision la position des régions C et V sur toutes les séquences de l'ensemble de test (Figure 5.3).

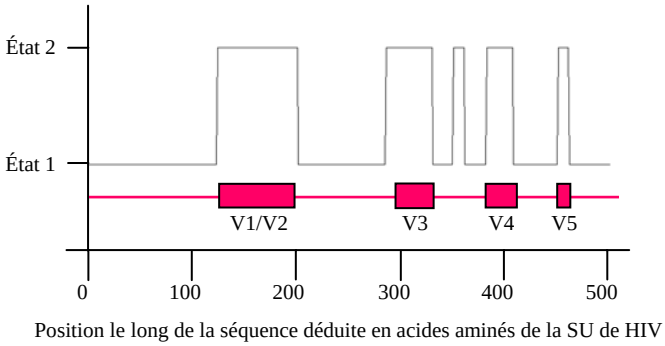

FIG. 5.3 – *Régions prédites par un modèle de Markov caché d'ordre 1 avec deux états entraîné avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences déduites en acides aminés de la SU de HIV.* Le graphique représente les régions prédites sur la séquence de test HIV-1 HXB2 par un modèle de Markov caché entraîné sur des séquences de HIV (—). L'organisation schématique de la SU de HIV, avec la position des régions variables V1 à V5 (■) et des régions constantes (—), est représentée en dessous du graphique.

Les estimateurs du maximum de vraisemblance des paramètres d'un modèle de Markov caché à 2 états d'ordre 5 ont été calculés à l'aide de comptages directs sur les séquences nucléotidiques de l'ensemble d'entraînement. Le modèle ainsi défini permet une bonne prédiction de la succession des régions C et V le long des séquences de la SU de HIV. Un modèle de Markov caché d'ordre 1 à deux états cachés dont les paramètres ont été estimés par maximum de vraisemblance à l'aide de comptages directs sur les séquences d'entrainement déduites en acides aminés fournit également une très bonne prédiction des régions C et V de la SU de HIV.

5.2 Modèles prédictifs des régions C et V de SIV et de SRLV

Les modèles basés sur les séquences de la SU d'EIAV ou de la SU de HIV ne permettent pas de correctement prédire les régions C et V de la SU des lentivirus simiens SIV ou des petits ruminants SRLV. Afin d'identifier ces régions C et V, nous avons développé des modèles de Markov cachés à 2 états spécifiques de ces deux lentivirus.

Pour entraîner et tester des modèles spécifiques de SIV, nous avons utilisé 61 séquences de la région du gène *env* codant la glycoprotéine de surface de SIV. Les trois quarts des séquences, soit 45 séquences, sélectionnées de façon aléatoire, ont servi à estimer les paramètres des modèles de Markov cachés et constituent l'ensemble d'entraînement. L'ensemble de test est formé des 16 séquences n'ayant pas été utilisées lors de l'élaboration des modèles.

De la même façon, les trois quarts des séquences disponibles de la SU de SRLV, soit 51 séquences, sélectionnées de façon aléatoire, ont été utilisées pour entraîner des modèles de Markov cachés spécifiques de SRLV. Ces modèles ont été ensuite été testés sur les 17 séquences de la SU de SRLV qui n'avaient pas servi lors de l'entraînement des modèles.

Des modèles de différents ordres ont été entraînés sur des séquences nucléotidiques de la SU de SIV ou de la SU de SRLV. A partir de l'ordre 2, certaines régions variables des SU de SIV et de SRLV sont correctement identifiées sur la plupart des séquences de l'ensemble de test par le modèle entraîné sur des séquences du lentivirus correspondant. Des modèles de Markov cachés d'ordre 5 permettent de localiser avec une grande précision la totalité des régions C et V de la SU de SIV et de la SU de SRLV (Figure 5.4).

Des modèles d'ordre 1 avec 2 états cachés ont également été entraînés sur des séquences déduites en acides aminés de la SU de SIV et de la SU de SRLV. Ils permettent de prédire avec une très grande précision la position des régions C et V sur toutes les séquences des ensembles de test de SIV et SRLV (Figure 5.5).

Les estimateurs du maximum de vraisemblance des paramètres de mo-

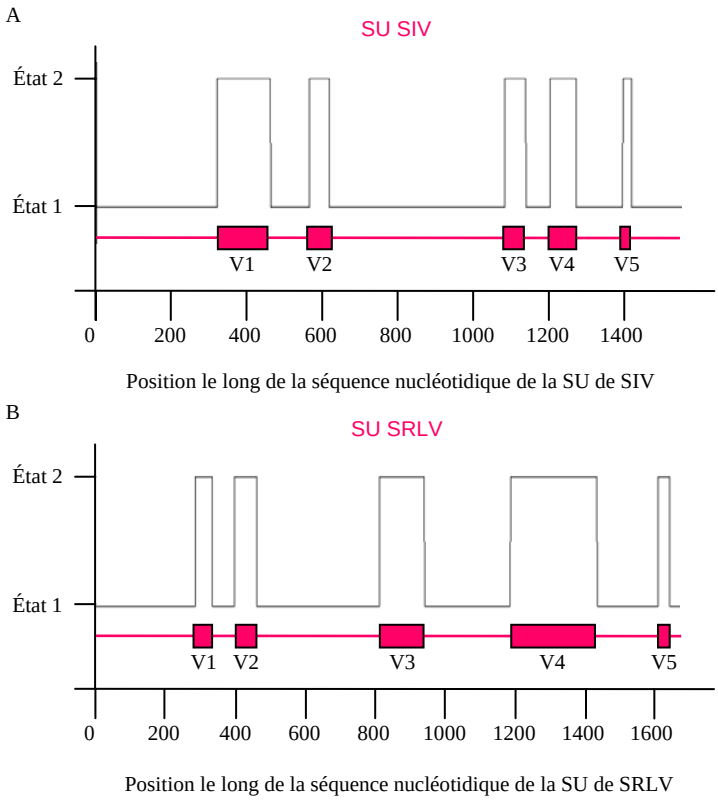

FIG. 5.4 – *Régions prédites par des modèles de Markov cachés entraînés avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences nucléotidiques des SU de SIV et de SRLV.* Les graphiques représentent les régions prédites par des modèles de Markov cachés (—). L'organisation schématique des SU de SIV et de SRLV, avec la position des régions variables V1 à V5 (■) et des régions constantes (—), est représentée en dessous des graphiques. (A) Modèle de Markov caché d'ordre 5 avec deux états entraîné sur des séquences nucléotidiques de la SU de SIV. (B) Modèle de Markov caché d'ordre 5 avec deux états entraîné sur des séquences nucléotidiques de la SU de SRLV.

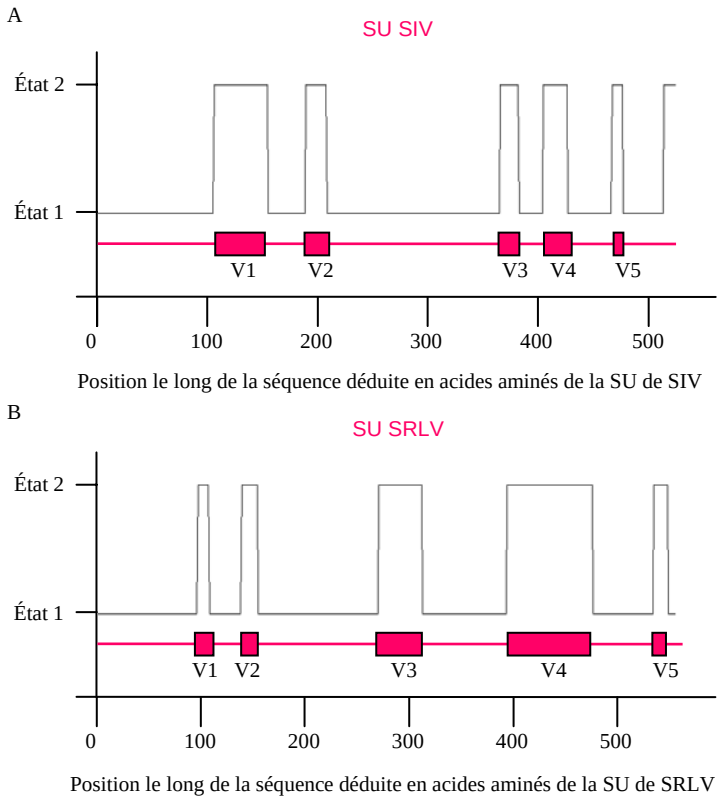

FIG. 5.5 – *Régions prédites par des modèles de Markov cachés entraînés avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences déduites en acides aminés des SU de SIV et de SRLV.* Les graphiques représentent les régions prédites par des modèles de Markov cachés (—). L'organisation schématique des SU de SIV et de SRLV, avec la position des régions variables V1 à V5 (■) et des régions constantes (—), est représentée en dessous des graphiques. (A) Modèle de Markov caché d'ordre 1 avec deux états entraîné sur des séquences déduites en acides aminés de la SU de SIV. (B) Modèle de Markov caché d'ordre 1 avec deux états entraîné sur des séquences déduites en acides aminés de la SU de SRLV.

dèles de Markov cachés à 2 états ont été déterminés par comptages directs sur les séquences d'entraînement des fréquences d'émission et de transition. Les résultats sont très similaires à ceux obtenus avec des modèles dont les paramètres ont été estimés avec l'algorithme de Baum-Welch avec matrice d'émission fixée. Des modèles d'ordre 5 permettent une bonne prédiction de la succession des régions C et V le long des séquences nucléotidiques de la SU de SIV et de la SU de SRLV. De même, des modèles de Markov cachés d'ordre 1 fournissent une très bonne prédiction des régions C et V sur les séquences déduites en acides aminés des SU de SIV et de SRLV.

5.3 Modèles combinés prédictifs des régions C et V des lentivirus

Des modèles de Markov cachés spécifiques des lentivirus EIAV, HIV, SIV et SRLV ont été développés. Chaque modèle permet de prédire avec une grande précision les régions C et V de la SU du lentivirus correspondant aux séquences ayant servi à entraîner le modèle. Cependant, aucun des modèles développés ne permet de prédire les régions constantes et variables des autres lentivirus.

Afin de définir des modèles de Markov cachés capables de prédire convenablement les régions C et V des SU de tous les lentivirus, nous avons entraîné des modèles de Markov cachés à 2 états sur une combinaison de séquences d'entraînement de plusieurs lentivirus. Un ensemble d'entraînement a été constitué à partir de 94 séquences de la SU d'EIAV, 78 séquences de la SU de HIV, 46 séquences de la SU de SIV et 51 séquences de la SU de SRLV sélectionnées de façon aléatoire. Les modèles définis ont été testés sur 93 séquences de la SU d'EIAV, 77 séquences de la SU de HIV, 15 séquences de la SU de SIV et 17 séquences de la SU de SRLV. Les modèles combinés, basés sur des séquences d'EIAV, de HIV, de SIV et de SRLV, ont également été utilisés pour tenter de prédire les régions C et V de la SU de deux lentivirus n'ayant pas servi à définir les modèles : BIV et FIV. Treize séquences de la SU de BIV et 16 séquences de la SU de FIV ont été intégrées à l'ensemble

de test.

Un modèle de Markov caché d'ordre 5 avec 2 états cachés (Figure 5.6) a été entraîné à l'aide de l'algorithme de Baum-Welch avec matrice d'émission fixée sur les séquences nucléotidiques d'EIAV, de HIV, de SIV et de SRLV. De même, un modèle de Markov caché d'ordre 1 avec 2 états cachés (Figure 5.7) a été entraîné à l'aide de l'algorithme de Baum-Welch avec matrice d'émission fixée sur les séquences déduites en acides aminés de la SU de ces quatre virus. Bien que le nombre, la longueur et la position des régions variables diffèrent d'un lentivirus à l'autre, les modèles obtenus sont capables de prédire avec exactitude la position des régions C et V sur les séquences de test d'EIAV, de HIV, de SIV et de SRLV. De façon surprenante, ces modèles sont également capables de différencier les régions V1 et V2 de la SU de HIV alors que ces régions avaient été définies comme une unique région variable V1/V2 dans l'ensemble d'entraînement des modèles. Enfin, les modèles combinés, entraînés uniquement sur des séquences d'EIAV, de HIV, de SIV et de SRLV, permettent de prédire certaines régions variables des lentivirus BIV et FIV. Les régions variables de BIV et de FIV prédites convenablement ne sont pas les mêmes d'une séquence de test à l'autre. Néanmoins, selon les séquences de test, toutes les régions variables de BIV et de FIV peuvent être prédites par les modèles combinés.

5.4 Conclusions

Des modèles de Markov cachés permettant d'identifier les régions C et V des lentivirus EIAV, HIV, SIV et SRLV ont pu être développés. Les séquences des glycoprotéines de surface de ces quatre virus sont donc très structurées quant à leur composition en mots de nucléotides et d'acides aminés. Pour chaque lentivirus, les régions constantes et les régions variables peuvent être caractérisées par des signatures qui leur sont propres. Cependant, les modèles basés sur les séquences de la SU d'un lentivirus ne permettent pas d'identifier les régions C et V des autres lentivirus. Les signatures des régions C et V obtenues sont donc spécifiques de chaque lentivirus. Néanmoins, en en-

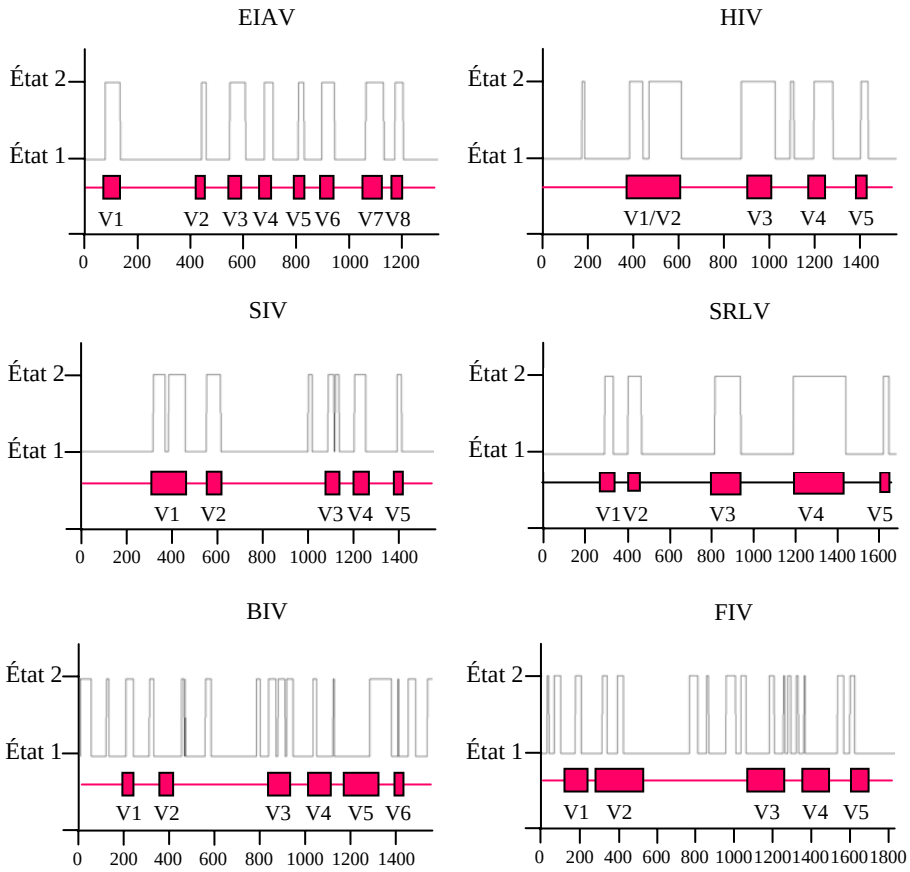

FIG. 5.6 – *Régions prédites par un modèle de Markov caché combiné sur des séquences nucléotidiques de la SU d'EIAV, de HIV, de SIV, de SRLV, de BIV et de FIV.* Les graphiques représentent les régions prédites par un modèle de Markov cachés d'ordre 5 entraîné avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences de la SU d'EIAV, de HIV, de SIV et de SRLV (—). L'organisation schématique des SU des lentivirus, avec la position des régions variables (■) et des régions constantes (—), est représentée en dessous du graphique.

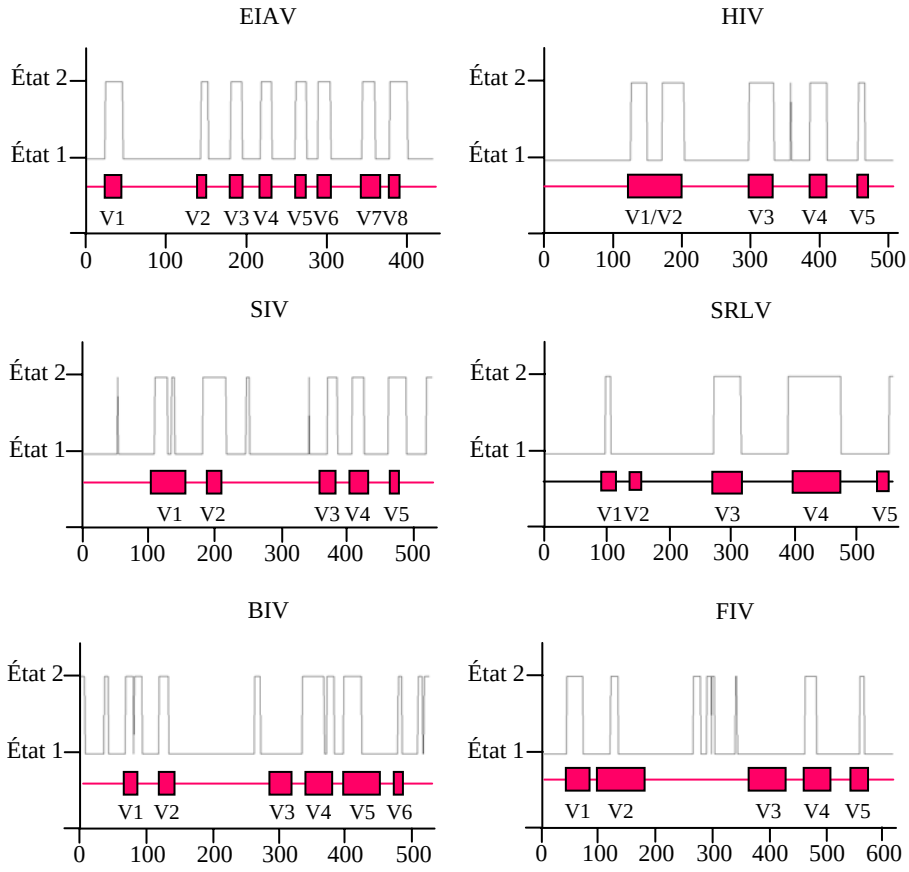

FIG. 5.7 – *Régions prédites par un modèle de Markov caché combiné sur des séquences déduites en acides aminés de la SU d'EIAV, de HIV, de SIV, de SRLV, de BIV et de FIV.* Les graphiques représentent les régions prédites par un modèle de Markov cachés d'ordre 1 entraîné avec l'algorithme de Baum-Welch avec matrice d'émission fixée sur des séquences de la SU d'EIAV, de HIV, de SIV et de SRLV (—). L'organisation schématique des SU des lentivirus, avec la position des régions variables (■) et des régions constantes (—), est représentée en dessous du graphique.

traînant des modèles de Markov cachés sur une combinaison de séquences des différents lentivirus, il est possible de prédire avec une grande précision les régions C et V de tous les lentivirus utilisés pour élaborer les modèles. Les régions C et V de lentivirus n'ayant pas été utilisés pour entraîner les modèles tels que BIV et FIV peuvent également être prédites. Les régions C et V de tous les lentivirus présentent donc des caractéristiques suffisamment proches pour être reconnues par un même modèle de Markov caché.

Chapitre 6

Motifs caractéristiques des régions constantes et variables des lentivirus

Sommaire

Pour parvenir à distinguer convenablement les régions constantes et variables des lentivirus, les modèles de Markov cachés à deux états que nous avons définis ont identifié une composition en mots de nucléotides ou en mots d'acides aminés spécifique de chaque type de région, suggérant que certains mots de nucléotides ou mots d'acides aminés, que nous appellerons motifs, apparaissent préférentiellement dans les régions constantes ou dans les régions variables.

Pour différencier les régions constantes des régions variables, les modèles développés s'appuient sur les fréquences d'apparition des différents motifs dans ces régions. Par la suite, nous considèrerons qu'un motif est caractéristique des régions constantes, ou des régions variables, lorsqu'il apparaît fréquemment dans les régions constantes, ou dans les régions variables, et peu fréquemment dans l'autre type de région.

Une méthode permettant d'extraire les motifs qui sont caractéristiques de chaque type de région à partir des modèles de Markov cachés précédemment définis a été mise en place, puis appliquée aux modèles de Markov cachés capables d'identifier les régions C et V des différents lentivirus.

6.1 Méthode d'extraction de motifs caractéristiques

Afin de différencier les régions C et V des lentivirus, les modèles de Markov cachés développés utilisent une matrice d'émission spécifique de chaque type de région. Les matrices d'émission qui modélisent les compositions en mots de nucléotides ou d'acides aminés des régions C et V des lentivirus utilisent pour cela de très nombreux paramètres. Un modèle de Markov caché d'ordre 5 entraîné sur des séquences nucléotidiques décrit, par exemple, chaque type de région à l'aide de 4096 paramètres.

Le but de l'extraction de motifs caractéristiques des régions C et V des lentivirus est de parvenir à identifier les différentes régions à l'aide d'un nombre restreint de motifs. Pour cela, l'extraction consiste à définir une liste MC et une liste MV contenant un nombre limité de mots de nucléotides ou d'acides aminés caractéristiques respectivement des régions C et des régions V. Ces deux listes peuvent être définies par

$$\begin{aligned} MC &= \{w/ \quad N_n(C,w) \geq s_c \cdot N_n(w)\}, \\ MV &= \{w/ \quad N_n(V,w) \geq s_v \cdot N_n(w)\}, \end{aligned}$$

où $N_n(C,w)$ et $N_n(V,w)$ représentent respectivement le nombre d'occurrences du motif w dans les régions constantes et dans les régions variables d'une séquence de longueur n. $N_n(w)$ représente le nombre d'occurrences de w dans cette même séquence. Les variables s_c et s_v représentent les seuils de fréquence à partir desquels un motif w peut être considéré comme caractéristique des régions C et V.

L'extraction de motifs caractéristiques revient donc à définir le couple (s_c^*, s_v^*) qui permet d'obtenir les ensembles de motifs MC^* et MV^* qui dé-

crivent le mieux les régions C et V des lentivirus avec moins de paramètres que les modèles de Markov cachés définis.

Soit X une séquence et W_X l'ensemble des motifs qui composent cette séquence. On considère deux groupes de séquences : les séquences de l'ensemble $SeqC$ dont tous les états cachés sont de type constant et les séquences de l'ensemble $SeqV$ dont tous les états cachés sont de type variable. Une façon de définir les ensembles MC^* et MV^* est de maximiser la probabilité

$$P = P(\text{motifC}|X \in SeqC) + P(\text{motifV}|X \in SeqV),$$

où motifC représente l'événement

$$\text{"card}\{w/\ w \in W_X \cap MC\} > \text{card}\{w/\ w \in W_X \cap MV\}\text{"}$$

et motifV représente l'événement

$$\text{"card}\{w/\ w \in W_X \cap MV\} > \text{card}\{w/\ w \in W_X \cap MC\}\text{"}.$$

En d'autres termes, on définit les ensembles MC^* et MV^* tels que les séquences constantes soient composées d'une majorité de motifs caractéristiques des régions constantes et les séquences variables d'une majorité de motifs caractéristiques des régions variables.

La probabilité P est estimée par une simulation numérique de type Monte Carlo. Pour cela, on génère un grand nombre de séquences de longueur L selon le modèle de Markov caché dont on souhaite extraire des motifs. Seules les séquences dont tous les états cachés appartiennent au même type sont conservées. La probabilité P est estimée pour tous les couples (s_c, s_v) possibles. Le couple (s_c^*, s_v^*) qui maximise P permet de définir les ensembles MC^* et MV^* qui caractérisent le mieux les régions C et V.

6.2 Application aux régions constantes et variables des lentivirus

Des motifs caractéristiques des régions constantes et variables des différents lentivirus ont été extraits à partir des modèles définis dans les chapitres

précédents. La méthode d'extraction définie dans la section précédente a été utilisée en estimant la probabilité P sur 1500 séquences de longueur 20.

Les modèles entraînés sur les séquences de la région du gène *env* codant la glycoprotéine de surface des lentivirus EIAV, HIV, SIV ou SRLV ont permis de définir des motifs caractéristiques des régions constantes et variables de ces lentivirus. Des mots de 2 acides aminés ont été extraits des modèles d'ordre 1 entraînés sur les séquences déduites en acides aminés. De la même façon, des mots de longueur $m+1$ ont été extraits des modèles d'ordre m entraînés sur les séquences nucléotidiques (Table 6.1).

Les modèles combinés capables de reconnaître la plupart des régions constantes et variables de tous les lentivirus ont permis de définir des motifs caractéristiques des régions C et V de tous les lentivirus, et non plus spécifiques des régions C et V de chaque lentivirus. A partir des 400 paramètres utilisés pour modéliser chaque type de régions dans le modèle combiné d'ordre 1 entraîné sur les séquences déduites en acides aminés, nous avons extrait 216 motifs caractéristiques des régions constantes des lentivirus et 85 motifs caractéristiques des régions variables (Table 6.2).

De même, 2080 motifs caractéristiques des régions constantes et 1007 motifs caractéristiques des régions variables ont été définis à partir du modèle combiné d'ordre 5 entraîné sur les séquences nucléotidiques des SU d'EIAV, HIV, SIV et SRLV (Table 6.3) qui utilise 4096 paramètres pour caractériser chaque type de région.

Les motifs caractéristiques que nous avons extraits permettent de caractériser les régions C et V des lentivirus avec beaucoup moins de paramètres que les modèles de Markov cachés définis dans les chapitres prcédents. La méthode d'extraction mise au point permet de ne conserver qu'environ la moitié des paramètres des modèles de Markov cachés pour caractériser les régions variables et le quart pour caractériser les régions constantes. Le nombre de paramètres étant plus restreint, il est alors possible d'envisager d'analyser en détail les signatures caractéristiques des régions C et V des lentivirus.

Lentivirus	Modèle	(s_c^*, s_v^*)	P^*	card(MC^*)	card(MV^*)
EIAV	M1-M1aa	(65,45)	1	163	130
	M1-M3n	(80,35)	0,88	108	86
	M1-M4n	(85,25)	0,93	396	361
	M1-M5n	(95,10)	0,83	795	879
HIV	M1-M1aa	(85,50)	0,99	155	109
	M1-M3n	(60,45)	0,89	184	51
	M1-M4n	(75,40)	0,81	437	383
	M1-M5n	(100,5)	0,95	663	2722
SIV	M1-M1aa	(85,30)	0,97	207	109
	M1-M3n	(75,35)	0,85	170	23
	M1-M4n	(85,35)	0,73	452	180
	M1-M5n	(85,30)	0,67	1973	1047
SRLV	M1-M1aa	(70,45)	0,99	241	106
	M1-M3n	(75,40)	0,78	112	76
	M1-M4n	(70,55)	0,89	548	177
	M1-M5n	(80,45)	0,72	1690	989

TAB. 6.1 – *Extraction de motifs caractéristiques des régions constantes et variables des lentivirus EIAV, HIV, SIV et SRLV.* Les chiffres du tableau indiquent, pour chaque lentivirus, et pour chaque modèle entraîné sur les séquences de la SU de ce lentivirus, le couple des seuils (s_c^*, s_v^*) qui permet d'obtenir la probabilité P^* maximale ainsi que les cardinaux des ensembles MC^* et MV^* contenant les motifs caractéristiques des régions constantes et variables. Un modèle M1-M$m$aa représente un modèle de Markov caché d'ordre m entraîné sur des séquences déduites en acides aminés. Un modèle M1-Mmn représente un modèle de Markov caché d'ordre m entraîné sur des séquences nucléotidiques.

MC	AE, AG, AI, AK, AM, AP, AQ, AR, AV, AW, AY, CA, CC, CD, CG, CH, CL, CP, CQ, CR, CW, DA, DF, DG, DH, DM, DP, DQ, DY, EA, EC, EF, EH, EI, EL, EM, EP, EQ, EV, EW, FA, FC, FD, FE, FF, FG, FH, FI, FK, FL, FM, FP, FQ, FS, FT, FV, FW, GA, GE, GF, GG, GI, GL, GM, GV, GW, GY, HD, HE, HF, HG, HN, HP, HQ, HR, HS, HT, HW, HY, IE, IF, IH, II, IK, IL, IQ, IS, IT, IV, IW, IY, KA, KC, KD, KF, KL, KP, KQ, KT, KY, LA, LE, LG, LH, LK, LL, LM, LQ, LR, LW, LY, MA, MC, MI, ML, MM, MP, MQ, MR, MS, MT, MV, MW, MY, NF, NM, NR, PA, PC, PD, PE, PF, PI, PL, PP, PQ, PT, PV, PW, QA, QC, QE, QF, QH, QI, QL, QM, QP, QQ, QR, QS, QT, QW, RA, RC, RE, RF, RH, RM, RR, RS, RV, SC, SE, SF, SL, SP, SV, SW, SY, TH, TI, TP, TQ, TV, TY, VA, VF, VG, VH, VI, VK, VL, VM, VP, VS, VT, VW, VY, WA, WD, WE, WH, WK, WL, WM, WN, WP, WQ, WR, WS, WV, WW, WY, YD, YE, YF, YG, YH, YI, YK, YM, YP, YQ, YV, YY.
MV	AD, AL, AN, CE, CF, CI, CK, CS, CY, DD, DI, DK, DR, DT, DV, EY, FR, GK, GN, GT, HH, HI, HV, IC, ID, IN, KG, KI, KK, KM, LD, LF, MD, MG, MH, MN, ND, NH, NK, NL, NN, NQ, NS, NT, NY, PH, PK, PN, PR, QG, QN, RI, RK, RL, RN, RP, RQ, RT, RW, SH, SM, SN, SR, SS, ST, TA, TD, TK, TL, TM, TN, TR, TS, TT, VC, VN, VR, WG, WI, WT, YA, YL, YN, YR, YW.

TAB. 6.2 – *Mots de deux acides aminés caractéristiques des régions constantes et variables des lentivirus.* Le modèle combiné M1-M1, entraîné sur les séquences déduites en acides aminés des SU d'EIAV, HIV, SIV et SRLV, a permis d'extraire respectivement 216 et 85 motifs caractéristiques des régions constantes et des régions variables des lentivirus parmi les 400 mots de deux acides aminés possibles.

MC	AAAACC, AAAAGC, AAAATT, AAACAG, AAACAT, AAACCC, AATCCG, AATCCT, AATCGA, AATCGG, AATCTA, AATCTC, AATCTG, AATGCC, AATGCG, CGGTTT, CGTAAA, CGTACC, CGTACG, CGTACT, CGTAGA, CGTAGC, CGTATT, CGTCAG, CGTCAT, CGTCCA, CGTCTG, GCCTTT, GCGACA, GCGACC, GCGACG, GCGATA, GCGATC, GCGATT, GCGCAT, GCGCCC, GCGCCT, GCGCTC, GCGCTG, GCGCTT, GCGGAA, GCGGCA, GCGGCT, GCGGGT, GCGGTA, TAATTT, TACACA, TACACC, TACACG, TACACT, TACAGT, TACATA, TACATG, TACATT, TACCCA, TACCCC, TACCCG, TTGTCC, TTGTCG, TTGTGA, TTGTGC, TTGTGG, TTGTTG, TTTAAA, TTTAAC, TTTAAG.
MV	AACTAT, AACTCT, AACTGA, AACTGT, AACTTT, AAGACG, ACTTAA, ACTTGA, ACTTGC, ACTTTA, ACTTTC, ACTTTT, AGAGCC, AGATGA, AGCACA, CCGCGT, CCGGCA, CCGGGT, CCGTCA, CCTAAC, CCTAAG, CCTACA, CCTAGA, CCTCAT, CTAGGT, CTAGTG, CTAGTT, CTGTCG, CTGTTC, CTGTTT, GATATA, GATCTT, GATGAC, GATGCG, GATGTA, GATTGA, GATTGT, GCAACA, GCAAGG, GGCGTT, GGCTAA, GGCTCC, GGCTCT, GGGACG, GGGATT, GGGCGA, GGGCGC, GGGTAG, GTCGGC, GTCGGG, GTGAAC, GTGAGC, GTGAGG, GTGAGT, TGAGCA, TGAGCG, TGAGGA, TGAGTA, TGAGTC, TGATAC, TTAGGA, TTAGTC, TTATAA, TTATTG, TTCAAC, TTCAGA,

TAB. 6.3 – *Exemples de mots de six nucléotides caractéristiques des régions constantes et variables des lentivirus.* Le modèle combiné M1-M5, entraîné sur les séquences nucléotidiques des SUs d'EIAV, HIV, SIV et SRLV, a permis d'extraire respectivement 2080 et 1007 motifs caractéristiques des régions constantes et des régions variables des lentivirus parmi les 4096 mots de six nucléotides possibles.

Conclusions et perspectives

L'objectif de mon travail de recherche était de rechercher la présence de signaux caractéristiques qui pourraient expliquer l'accumulation de mutations dans des zones spécifiques des génomes lentiviraux. Nous avons choisi d'étudier le découpage en régions constantes et régions variables de la SU des lentivirus à l'aide de modèles de Markov cachés.

De tels modèles, basés uniquement sur les séquences nucléotidiques ou déduites en acides aminés de la SU d'EIAV, ont été développés afin de différencier les régions C et V de ce lentivirus. Ces modèles ont été entraînés avec l'algorithme de Baum-Welch, qui est classiquement utilisé dans la littérature pour estimer les paramètres de modèles de Markov cachés. Aucun des modèles entraînés avec l'algorithme de Baum-Welch n'a permis de prédire correctement la segmentation de la SU d'EIAV en régions constantes et variables. L'échec de l'utilisation de l'algorithme de Baum-Welch pour l'apprentissage des paramètres des modèles nous a conduit à définir une variante de cet algorithme plus adaptée à notre problème : l'algorithme de Baum-Welch avec matrice d'émission fixée. L'utilisation de cet algorithme a permis de définir des modèles de Markov à 2 états cachés robustes, capables de différencier avec une grande précision les régions C et V de la SU d'EIAV. Des modèles ayant 9 états cachés, capables de prédire l'ensemble des régions constantes d'une part et chacune des 8 régions variables d'EIAV d'autre part, ont également été définis.

Il est important de souligner que les modèles que nous avons définis sont capables d'identifier les régions variables de séquences du gène *env* d'EIAV sans les comparer à des séquences connues. Nos modèles ne sont pas basés

sur des alignements ou sur des calculs de divergences entre séquences. Par ailleurs, la très grande variabilité de la longueur des régions, ainsi que les différents tests que nous avons fait subir à nos modèles, ont permis de prouver que ces modèles n'utilisaient ni l'ordre, ni la position relative des régions, ni leurs longueurs pour prédire la succession des régions C et V le long de la SU d'EIAV. Au contraire, les modèles de Markov cachés développés doivent s'appuyer sur des différences statistiques entre les compositions des régions C et V en mots de nucléotides ou d'acides aminés de longueur $m+1$, où m représente l'ordre du modèle, afin de distinguer ces deux types de régions.

Les modèles que nous avons mis au point montrent que toutes les régions constantes peuvent être convenablement modélisées par un unique état caché. Ceci qui prouve qu'elles possèdent des propriétés statistiques suffisamment similaires. Les régions variables V1 à V8 peuvent être modélisées soit par un seul état général, comme dans les modèles à 2 états cachés, soit par 8 états spécifiques de chacune de ces régions, comme dans les modèles à 9 états cachés. Les régions variables partagent donc, de la même façon que les régions constantes, suffisamment de propriétés statistiques communes pour être modélisées par un état unique. Dans le même temps, chaque région variable présente un profil statistique qui lui est propre et qui permet de la différencier des autres régions variables.

Les différences entre les régions constantes et variables d'EIAV ont pu être confirmées par un test statistique. Ce test montre que les 17 matrices de transition qui modélisent les 9 régions constantes et les 8 régions variables d'EIAV, basées sur les compositions statistiques en mots de nucléotides ou d'acides aminés de ces régions, sont bien toutes différentes. La définition d'une distance entre les matrices de transition des régions C et V a également montré que si toutes les régions sont différentes deux à deux, il existe néanmoins deux groupes principaux de régions : le groupe des régions constantes et celui des régions variables.

Il est intéressant de noter que les modèles entraînés uniquement sur des séquences de la SU d'EIAV ne parviennent pas à identifier les régions C et V des autres lentivirus. La composition statistique en acides aminés et en nucléotides est donc différente d'un lentivirus à l'autre. Des modèles de

Markov cachés à 2 états cachés et spécifiques de chaque lentivirus ont été définis. Ces modèles permettent de prédire avec une grande exactitude les différentes régions du lentivirus pour lequel ils ont été définis. Ainsi, quel que soit le lentivirus considéré, les régions constantes et variables de ce lentivirus présentent une différence de composition en acides aminés et en nucléotides.

Même si les lentivirus présentent des compositions génétiques distinctes, les régions C et V des lentivirus EIAV, HIV, SIV et SRLV possèdent des propriétés statistiques suffisamment similaires pour pouvoir être reconnues par un même modèle. Le modèle commun, entraîné sur une combinaison de séquences des SU de ces quatre lentivirus (93 séquences de la SU d'EIAV, 77 séquences de la SU de HIV, 15 séquences de la SU de SIV et 17 séquences de la SU de SRLV) prédit presque parfaitement les régions C et V d'EIAV et de HIV. Les régions C et V de SIV et SRLV sont également prédites avec une grande précision. Le taux d'erreurs commises par le modèle commun lors de la prédiction est corrélé au nombre de séquences présentes dans l'ensemble d'apprentissage. Ainsi, les régions C et V des lentivirus EIAV et HIV sont mieux prédites que celles des lentivirus SIV et SRLV car l'ensemble d'apprentissage du modèle commun comporte plus de séquences des SU d'EIAV et de HIV.

Le modèle commun permet de prédire convenablement la plupart des régions variables des lentivirus BIV et FIV, bien que les séquences de ces virus n'aient pas été utilisées pour entraîner le modèle. Ceci suggère que les compositions statistiques en mots de nucléotides et en mots d'acides aminés de tous les lentivirus, bien que toutes différentes, présentent des propriétés communes. Nous pouvons supposer que si le nombre de séquences disponibles de la SU de BIV et de FIV était suffisant pour les intégrer à l'ensemble d'apprentissage du modèle commun, nous pourrions améliorer grandement la prédiction des régions C et V de ces virus.

De façon surprenante, le modèle commun permet de différencier les régions variables V1 et V2 de HIV alors que ces deux régions avaient été considérées comme une unique région variable V1/V2 lors de l'entraînement du modèle. L'entraînement du modèle sur un ensemble combiné de séquences de plusieurs lentivirus lui permet d'apprendre les propriétés générales des régions

C et V de tous les lentivirus. Le modèle est moins dépendant des séquences d'apprentissage. Ainsi, bien que la séparation entre les deux premières régions variables de HIV n'apparaisse pas dans les séquences d'apprentissage, le modèle commun parvient à identifier des propriétés caractéristiques des régions constantes des lentivirus et à prédire correctement la région constante qui sépare les régions V1 et V2 de HIV.

Nous avons mis en évidence que les acides aminés qui composent les régions C et V présentent des propriétés chimiques différentes. Les régions constantes sont riches en acides aminés hydrophobes alors que les régions variables sont riches en acides aminés hydrophiles. Ces résultats sont à rapprocher de la structure tertiaire de l'enveloppe du virus EIAV. La structure tertiaire d'une protéine, qui décrit la manière dont celle-ci se replie dans l'espace, dépend de sa séquence, c'est-à-dire de la succession linéaire des acides aminés la constituant. La séquence d'une protéine comporte une certaine proportion d'acides aminés hydrophobes et hydrophiles. Leurs interactions avec les molécules d'eau conditionnent, entre autres paramètres, la manière dont la chaîne polypeptidique se replie. Les acides aminés hydrophobes auront tendance à éviter l'eau. Inversement les acides aminés hydrophiles vont chercher à rester à proximité du milieu aqueux. Ainsi, dans le cas de la protéine de l'enveloppe, il se forme un coeur hydrophobe au centre de la structure tertiaire, tandis que les acides aminés hydrophiles sont exposés en surface. Les modifications des acides aminés exposés à la surface de la particule virale induisent des modifications de leur reconnaissance par certains composants de la réponse immunitaire. Il a ainsi été établi dans le cas de l'infection par EIAV que les variations de la SU modulent la sensibilité à la réponse dite humorale médiée par les anticorps (Leroux *et al.*, 2001, Howe *et al.*, 2002).

Les régions variables, riches en acides aminés hydrophiles, vont se présenter sur la surface extérieure de la glycoprotéine de surface de l'enveloppe et ne pourront pas être reconnues par les anticorps du système immunitaire qui identifient la particule virale. La grande variabilité génétique des régions variables permet ainsi aux lentivirus d'échapper à une partie du système immunitaire.

Les propriétés statistiques permettant de différencier les régions C et V

sont codées par les séquences nucléotidiques des lentivirus. Des modèles de Markov cachés d'ordre 5 prédisent avec une grande précision l'alternance des régions C et V sur les séquences nucléotidiques de tous les lentivirus. Ces modèles sont basés sur les fréquences de mots de 6 nucléotides correspondant à 1 ou 2 codons complets. Cette longueur est compatible avec le nombre de nucléotides situés dans le voisinage de la paume de la RT lors de la rétrotranscription (Beard *et al.*, 1998, Bebenek *et al.*, 1997, Kohlstaedt *et al.*, 1992). Ceci suggère que les mutations pourraient résulter d'une interaction entre les chaînes d'ARN et la protéine RT. Un mécanisme de mutation, dépendant de la séquence du lentivirus et de la chaîne interne de la RT, pourrait modifier la vitesse et/ou la précision de la rétrotranscription. Certaines portions de séquences favoriseraient l'apparition de mutations en permettant un passage plus rapide de la RT. Inversement, d'autres portions de la chaîne d'ARN pourraient présenter des séquences spécifiques ralentissant la RT et lui imposant ainsi une plus grande fidélité réplicative.

Les modèles de Markov cachés ont montré que les régions C et V des lentivirus pouvaient être identifiées par des signaux statistiques spécifiques. Des motifs caractéristiques des régions C et V des lentivirus ont pu être extraits des modèles développés. La comparaison des motifs spécifiques aux lentivirus EIAV, HIV, SIV et SRLV permettra de dégager les points communs et les particularités de ces lentivirus. Les motifs caractéristiques des régions variables des lentivirus pourront également être comparés aux contextes de mutations connus. Des études ont montré par exemple que, dans les génomes rétroviraux, l'hypermutation G→A survenait principalement dans des contextes dinucléotidiques spécifiques comme GpG ou GpA (Vartanian *et al.*, 1994, Wain-Hobson *et al.*, 1995, Vartanian *et al.*, 2002). L'analyse des signaux mis en évidence pourrait permettre une meilleure connaissance des mécanismes conduisant à la répartition hétérogène des mutations dans la région du gène *env* codant la glycoprotéine de surface des lentivirus.

L'enrichissement du modèle commun capable de prédire correctement les régions C et V de n'importe quelle séquence lentivirale avec des séquences d'entraînement représentatives de tous les lentivirus pourra conduire au développement d'un logiciel. Ce logiciel permettra d'identifier les régions C et

V des lentivirus sur de nouvelles séquences sans avoir besoin de les comparer par alignement aux séquences existantes. Ce logiciel sera particulièrement utile pour des séquences très divergentes, et donc difficiles à aligner avec des séquences connues, ou pour identifier les régions C et V d'un nouveau lentivirus qui apparaîtrait.

Notre étude a également conduit à la définition d'une nouvelle méthode d'estimation des paramètres de modèles de Markov cachés : l'algorithme de Baum-Welch avec matrice d'émission fixée. Cet algorithme peut être utilisé dans les problèmes d'identification de régions homogènes lorsque l'on souhaite introduire, dans les modèles de Markov cachés développés, des informations sur la composition statistique de chaque type de région, sans vouloir, ou sans pouvoir, apporter d'informations sur l'ordre et la longueur des régions.

Jusqu'à présent, les modèles de Markov cachés avaient été appliqués à la biologie pour trouver des différences de structures dans de grandes séquences de plusieurs milliers de bases ou dans des génomes complets. L'analyse réalisée ici se situe au niveau d'une portion d'un gène pour rechercher une différence beaucoup plus fine. Nous pouvons envisager d'étendre la méthode développée à l'étude de la variabilité d'autres portions de génome, ou d'autres virus, ou à la recherche de caractéristiques très précises dans les gènes qui participeront à la compréhension du fonctionnement de certains organismes.

Article : *In silico* segmentations of lentivirus envelope sequences

soumis à *BMC Bioinformatics*

In silico segmentations of lentivirus envelope sequences

Aurélia Boissin-Quillon[1], Didier Piau[2] and Caroline Leroux[1]

[1] UMR754 INRA-ENVL-UCBL"Rétrovirus et Pathologie Comparée", IFR 128 BioSciences Lyon-Gerland, Université Claude Bernard Lyon 1, 69007 Lyon, France
[2] Institut Fourier UMR 5582 CNRS-UJF, Université Joseph Fourier (Grenoble 1), 100 rue des Maths, BP 74, 38402 Saint Martin d'Hères, France

Email: Aurélia Boissin-Quillon - aurelia.quillon@univ-lyon1.fr; Didier Piau - Didier.Piau@ujf-grenoble.fr; Caroline Leroux*- caroline.leroux@univ-lyon1.fr;

*Corresponding author

Abstract

Background: The gene encoding the envelope of lentiviruses exhibits a considerable plasticity, particularly the region which encodes the surface (SU) glycoprotein. Interestingly, mutations do not appear uniformly along the sequence of SU, but they are clustered in restricted areas, called variable (V) regions, which are interspersed with relatively more stable regions, called constant (C) regions. We look for specific signatures of C/V regions, using hidden Markov models constructed with SU sequences of the equine, human, small ruminant and simian lentiviruses.

Results: Our models yield clear and accurate delimitations of the C/V regions, when the test set and the training set were made up of sequences of the same lentivirus, but also when they were made up of sequences of different lentiviruses. Interestingly, the models predicted the different regions of lentiviruses such as the bovine and feline lentiviruses, not used in the training set. Models based on composite training sets produce accurate segmentations of sequences of all these lentiviruses.

Conclusions: Our results suggest that each C/V region has a specific statistical oligonucleotide composition, and that the C (respectively V) regions of one of these lentiviruses are statistically more similar to the C (respectively V) regions of the other lentiviruses, than to the V (respectively C) regions of the same lentivirus.

Background

Retroviruses are RNA viruses infecting vertebrates and many non vertebrates. Virus particles are spherical and surrounded by an envelope. Their viral replication is dependent of the RT (Reverse Transcriptase), a viral RNA-dependent DNA-polymerase. The lentivirus genus is part of the retrovirus family. Lentiviruses infect animals and humans and cause slowly progressing diseases. Among the lentivirus genus, HIV-1 and HIV-2 (Human Immunodeficiency Virus type 1 and 2) infect humans, EIAV (Equine Infectious Anemia Virus) infects equids, SRLV (Small Ruminant LentiVirus) infects goats and sheep, SIV (Simian Immunodeficiency Virus) infects non primate monkeys, BIV (Bovine Immunodeficiency Virus) infects bovines and FIV (Feline Immunodeficiency Virus) infects felines.

The considerable plasticity of the genome of lentiviruses is quite obvious in the *env* gene, encoding the envelope, particularly in the region encoding the surface (SU) glycoprotein forming spikes. Causes of this plasticity are, among other factors, the low fidelity of the viral reverse transcriptase (RT) during the retrotranscription of the viral RNA genome into DNA, the lack of proofreading activity of the RT, the high level of virus replication, and some recombination events in co-infected cells [1–4].

Interestingly, SU mutations do not appear uniformly along the *env* gene, but are clustered in restricted and specific areas defined as variable (V) regions flanked by constant (C) regions. On average, and depending on the lentivirus considered, from 10 % to 35 % of the amino-acids in SU vary between isolates, and more than 70 % of these variable amino-acids are located in V regions. Such C/V segmentations hold for all the lentiviruses [5–11].

It is unclear whether the accumulation of mutations in V regions is mainly due to locally high intrinsic mutation rates, or if mutations occur at similar rates at every SU sites with subsequent selection mechanisms eliminating most variants from the C regions. In any case, the plasticity of these genomes allows them to escape immune control very efficiently, while keeping their identity. Most of amino acids encoded by the V regions are on the outside of SU, while the amino acids encoded by the C regions are in the internal part. In this respect, one should note that the replication acts on one-dimensional molecules, at a moment when most of the information about their three-dimensional conformation seems unavailable. In other words, if the intrinsic mutation rates are indeed different in C regions and in V regions, this might be due to some specific signals encoded by the nucleotide (linear) viral sequence itself, possibly corresponding to interactions with the RT. To test this hypothesis, we developed a mathematical model based on lentivirus sequences, as simple and robust as possible, able to localize and to characterize their C/V segmentation of the SU region. Our approach was based on HMMs (hidden Markov models). These

models are tailored to describe heterogeneous sequences, since they basically break down a given sequence into a succession of locally homogeneous subsequences. HMMs were initially introduced in the context of speech recognition [12] and they are now major tools of the analysis of genomic and proteomic sequences [13–19]. In sequence analysis, each of the subsequences called a region, is described by the value of a Markov chain, called the hidden state, taken from a finite collection of values. Each state is characterized by its own statistical composition in nucleotides or in amino-acids. The succession of states itself is ruled by a master Markov chain, called the hidden chain.

Our main findings are as follows. Using SU sequences of EIAV, HIV, SRLV or SIV to train the HMMs, we obtained clear and accurate delimitations of the C and V regions of these lentiviruses. This suggests that the statistical composition of the C regions is markedly different from the statistical composition of the V regions. Additionally, we developped combined models, based on EIAV, HIV, SIV and SRLV sequences. These were able to predict simultaneously the C and V regions of every lentivirus in the collection above. Our combined models also predicted the C/V segmentation of other lentiviruses which were not used in the training sets: BIV and FIV. This indicates that the C and V regions are statistically distinct and that the V regions of all the lentiviruses share common statistical signatures.

Results

C/V segmentations of EIAV

We first tried to differentiate the C and V regions of the EIAV SU, using HMMs with $N = 2$ hidden states, for different orders m. The parameters of the models were estimated on training sets of 94 nucleotide sequences, by the EM algorithm. We used various training sets, dividing at random our complete set of sequences. Then, none of the various HMMs is able to identify the C and V regions of the EIAV SU. We obtained hidden states sequences which oscillated and repeatedly jumped from one hidden state to the other (data not shown). Hence, this method was not reliable to identify homogeneous regions corresponding to the C and V regions of the EIAV SU.

By contrast, fixed EM, as described in section Methods, yielded a clear delimitation of the known C and V regions on the whole test set, for a HMM of order $m = 2$. HMMs of higher orders $m \geq 3$ gave even more accurate predictions. For $m = 5$, the fit of the predicted C and V regions with the segmentations deduced from alignments was almost perfect (Figure 1A).

To differentiate the variable regions V1 to V8 and the C regions, we then used HMMs with $N = 9$ states. Thus, we trained one hidden state with each variable region and one hidden state with the constant regions

as a whole, and we estimated the parameters of a HMM of order $m = 5$ by the fixed EM algorithm. This yields a precise delimitation of the C and V regions, each V region showing a distinct signal (Figure 1B). Estimating the parameters of the models with the direct counting methods gave similar results.
Finally, HMMs with $N = 2$ or $N = 9$ hidden states, able to locate the C and V regions on deduced amino-acid sequences, were trained by the fixed EM algorithm and the direct counting method. We obtained accurate predictions of the C and V regions on the test sequences, with every training method, using a HMM of order $m = 1$ (Figures 2A and 2B).
The reconstructed sequences of the hidden states did not oscillate between the different hidden states as in the models based on the EM algorithm. The transition matrix obtained without prior information on the length of the regions allowed to identify long homogeneous regions and to compare them to the C and V regions previously defined.
At this point, we developped models with an unique C region. This C region do not fit a real region but represent an average of all the constant regions. There is no guarantee *a priori* that the constant regions are grouped together and can be modeled by an unique state. However, the C region introduced in our models allowed to predict all the constant regions with an amazing accuracy.

Tests of the models of EIAV C/V regions

Since our models were able to predict the C and V regions on both deduced amino-acid and nucleotide sequences of EIAV SU, we put them under trial in several directions. First, we checked that the models were not overfitted, keeping in mind that pseudo-counts were introduced to limit the overfitting problem. We checked whether the models were not overly specific of the training data, and whether it was possible to make them encompass new data tests. To perform such tests, the models were trained using sequences sharing a minimal amount of motifs with the test sequences. For example, we trained the models on virus sequences, which were present at the beginning of the disease induced in horses by EIAV, and we tested them on virus sequences at later stages of the disease [6]. Because of the variations due to viral replication during the time course of the EIAV infection, the training and test sequences displayed 7.8 % ($\pm$1.3) of divergence at the amino-acid level. In particular, the test and training sequences displayed 43.8 % ($\pm$20.2) of divergence in the third V region (V3). Despite this important level of divergence between the training and test sequences, the models correctly predicted the C and V regions, notably V3.
To check that the models were not simply following the order and positions of the V regions along the sequence, we also assembled artificial sequences with a greater number of V regions than in the real ones.

For instance, we inserted a copy of 15 amino-acids, taken from V7, into C2. The models which were trained with the fixed EM algorithm on the original sequences, managed to predict perfectly the additional V region of these modified sequences (Figure 2C).

Combined C/V models

Models based on EIAV sequences were unable to predict C and V regions of HIV, SIV or SRLV SU sequences (Figure 3). Hence, we developed a new specific model for each lentivirus. We trained models of order $m = 1$ on deduced amino-acid sequences and models of order $m = 5$ on nucleotide sequences, on 78 HIV sequences, 45 SIV sequences and 51 SRLV sequences respectively, using either the fixed EM algorithm or the direct counting methods. These new models, specific to each lentivirus, were indeed able to predict the C and V regions of test sequences of the corresponding virus, but failed to predict the C and V regions of the other lentiviruses. On the contrary, a combined HMM of order $m = 1$ with $N = 2$ hidden states, trained on a composite training set of EIAV, HIV, SIV and SRLV deduced amino-acid sequences, was powerful enough to localize accurately the V regions of test sequences of EIAV (V1 to V8), HIV (V1 to V4), and SIV or SRLV (V1 to V5). Rather to our surprise, the model also discriminated V1 and V2 of HIV, although these two regions were given as a unique region V1/V2 in the training set (Figure 4). The C and V regions of EIAV, HIV, SIV and SRLV were also predicted with great accuracy by HMMs of order $m = 5$ with $N = 2$ hidden states, trained on the corresponding nucleotide sequences. Finally, the combined models, trained on EIAV, HIV, SIV and SRLV sequences, were able to predict C and V regions of two lentiviruses which were not used to train them, namely BIV and FIV (Figure 4).

Separation of the EIAV C/V regions

The models developed in our study allow us to differentiate the C and V regions of EIAV and to distinguish each of the 8 variable regions. This indicates that the C and V regions have distinct statistical composition and that the 8 variable regions are statistically distinct too. A classical method to quantify the differences between the Markov chains which represent the C and V regions of EIAV, is to consider the relative entropy, also named Kullback-Leibler divergence, between these models, see [20–25]. The relative entropy of two Markov chains is given by

$$H(P|Q) = \sum_{i,j} \pi(i) P(i,j) \log \frac{P(i,j)}{Q(i,j)},$$

where P and Q are the transition matrix of the two Markov chains and π the invariant distribution associated to P. We used a symmetrized form of the relative entropy, defined as

$$\delta(P,Q) = H(P|Q) + H(Q|P).$$

The computation of the symmetrized relative entropy between the Markov chains modeling the 9 constant regions and the 8 variable regions of EIAV (see Table 2) indicates that the different C (respectively V) regions are closer to the global C (respectively V) region than to any of the V (respectively C) regions. Furthermore, the V regions are closer to each other than to any of the C regions.

To quantify this overall feeling, we first used the symmetrized relative entropy δ to study the distances between the C and V regions, representing them by a dendogram. Note that δ is not a true metric because it does not satisfy the triangle inequality. However, one can visualize the distances between the different regions by an unrooted tree, computed by the program Kitch (Phylip 3.5c) using the distance matrix previously estimated (Figure 5). The dendogram shows a distinct separation between a first group, made of the C regions, and a second group, made of the V regions. This confirms the fact that the C and V regions of EIAV differ in their statistical composition.

To further quantify this separation between the C and V regions, we built an asymptotic statistical test for the empirical transition matrices of two different regions, based on the following considerations. Assuming in general that $\widehat{q}_1$ and $\widehat{q}_2$ are empirical transition matrices of the same Markov chain with theoretical transition matrix q, based on two independent sequences of length L of the Markov chain, one can show that $L\,H(\widehat{q}_1, \widehat{q}_2)$ is asymptotically χ^2-distributed with $D(q)$ degrees of freedom, where $D(q)$ denotes the "dimension" of the Markov chain, that is, $D(q)$ is the number of nonzero coefficients in q minus the number of states (see the Appendix for more details). When every transition has positive probability and q has size M, $D(q) = M^2 - M$. In particular,

$$E(H(\widehat{q}_1, \widehat{q}_2)) \sim D(q)/L.$$

In the still more general case when $\widehat{q}_1$ and $\widehat{q}_2$ are based on independent sequences of unequal lengthes L_1 and L_2 respectively, a similar result holds, namely that $\ell\,H(\widehat{q}_1, \widehat{q}_2)$ is asymptotically χ^2-distributed with $D(q)$ degrees of freedom, where ℓ denotes the harmonic mean of L_1 and L_2, defined by the relation

$$\frac{2}{\ell} = \frac{1}{L_1} + \frac{1}{L_2}.$$

Using the symmetrized entropy δ, one sees that the distribution of $\frac{1}{2}\ell\delta(\widehat{q}_1, \widehat{q}_2)$ is asymptotically χ^2 with

$D(q)$ degrees of freedom, and in particular,

$$E(\delta(\widehat{q}_1, \widehat{q}_2)) \sim 2D(q)/\ell.$$

Using this result, one can perform χ^2 tests of equality between the C and V regions of EIAV. This yields p-values very close to zero. The biggest p-value is obtained for the two variable regions V1 and V2 and is $4 \cdot 10^{-17}$. Since the p-values are so small, one can conclude that the Markov chains previously defined to model the C and V regions of EIAV do not reflect the same statistical composition in words of amino acids. Hence, each of the 9 constant regions and the 8 variable regions has a specific signature.

Discussion

We report that HMMs are able to predict the C/V segmentations of various lentiviruses, based only on their deduced amino-acid sequences or their nucleotide sequences, with an amazing accuracy and a great robustness.

We would like to stress the fact that our algorithms identify the V regions without any comparaison by alignment with known sequences. The models developped in this study are not based on computations of divergences between sequences. Furthermore, the lengths of the regions exhibit a great variability, and the numbers of regions themselves may be, and indeed are sometimes, different from one sequence to another. These, and the various tests presented in section Results, prove that the models do not rely on the relative positions of the regions, nor on their lengths, to identify C/V segmentations of the sequences. On the contrary, they have to rely only on some statistical differences between the compositions in words of nucleotides or amino-acids of length $1 + m$, where m is the order of the model.

More detailed consequences of the performances of the models are as follows. First, all the C regions can be suitably modeled by a unique state. This proves that they have similar statistical properties. The V regions can be modeled either by one state or by several states. This suggests that V regions share common properties, when compared to C regions, and, at the same time, that each V region has its own statistical profile.

To highlight similarities and differences between data, a classical statistical method is based on Principal Components Analysis (PCA). Knowing that first order HMMs were able to differentiate between the C and V regions of EIAV and used only frequencies of words of two amino-acids, we performed a PCA of the 9 constant regions and the 8 variable regions of EIAV, using the frequencies of $20 \times 20 = 400$ words of two amino-acids as variables . Figure 6 shows a projection of the C and V regions of EIAV onto the plane

defined by the two first principal axes. One sees that, contrary to our method based on HMMs, PCA does not allow to separate the EIAV regions into two groups, whether these groups correspond to the C regions and the V regions or not. With PCA, all the regions seem to have nearly the same statistical composition in words of two amino-acids, although it is not the case. Thus our method, based on HMMs, is able to reveal rather subtle differences between the group of V regions and the group of C regions.

It may be of interest to note that a model, trained on EIAV sequences only, failed to identify the C and V regions of other lentiviruses, and conversely. Hence, the genetic compositions of the *env* genes of these different lentiviruses are distinct. However, the C and V regions of EIAV, HIV, SIV and SRLV do share some properties which are similar enough, so as to be recognized by a unique HMM, trained on a combined pool of EIAV, HIV, SIV and SRLV SU sequences. This combined model also predicts the C/V segmentation of BIV and FIV, whose sequences were not used to train the model. This supports the conclusion that the statistical compositions in words of nucleotides or amino-acids of the envelope genes of all these lentiviruses share some common features.

Models of order $m = 5$ of nucleotide sequences, based on the frequencies of words of length 6, predict with an amazing accuracy the C/V segmentations. These words correspond to one or two complete codons. This length is also compatible with the number of nucleotides that are in the neighborhood of the palm of RT during the retrotranscription [26–28]. This suggests that some mechanism of inaccurate nucleotide substitution, possibly due to sequence-specific variations and in interaction with the side chains of the RT protein, might modify the speed and/or the precision of the passage along the portion of the RNA chain which the RT copies.

Conclusions

The constant and variable regions of the lentiviruses EIAV, HIV, SLRV, SIV, BIV, and FIV can be identified by rather crude mathematical models based on HMMs. We attempt at present to extract the nature of the statistical signals which allow to distinguish between these regions. In this spirit, it has been reported that the retroviral `G` $\rightarrow$ `A` hypermutation occurs mainly in specific dinucleotide contexts, like `GpG` and `GpA` [29, 30]. Hence, one of our objectives now is to compare to known contexts of mutation the nucleotide words which are, as the present study shows, statistically characteristic of the variable regions of these lentiviruses.

Methods

Biological data

This section describes the sets of SU nucleotide sequences, used to train and to test the models (Table 1).

- EIAV: 187 sequences [6, 9, 31–34].

  Training set: 94 sequences. Test set: 93 sequences.

  According to the regions described in [6], we considered 8 variable regions (V1 to V8) and 9 constant regions (C1 to C9).

- HIV: 155 HIV-1 sequences. The panel is composed of the HIV-1 HXB2 sequence and representative sequences from the following subtypes: A (21 sequences), B (27 sequences), C (26 sequences), D (18 sequences), E (19 sequences), F (3 sequences), G (21 sequences), H (2 sequences), and 17 sequences of recombinant forms.

  Training set: 78 sequences. Test set: 77 sequences.

  Variable regions V1 to V5 are as defined in [7]. However, V1 and V2 are considered as a unique variable region V1/V2, since these variable regions are separated by a small constant region composed of only a few nucleotides.

- SIV: 61 sequences. Training set: 45 sequences. Test set: 15 sequences.

  Variable regions V1 to V5 are as defined in [5].

- SRLV: 68 sequences. Training set: 51 sequences. Test set: 17 sequences.

  Variable regions V1 to V5 are as defined in [8].

- BIV: 13 sequences. Test set: 13 sequences.

  We compared the predicted regions with the variable regions V1 to V6 previously defined in [10].

- FIV: 16 sequences. Test set: 16 sequences.

  We compared the predicted regions with the variable regions V1 to V5 previously defined in [11].

Hidden Markov models

We recalled in the introduction that HMMs involve pairs of random processes, called respectively the hidden process and the observed process. In our context, the hidden process $(S_i)_{1 \leq i \leq L}$ describes the succession of homogeneous regions along a sequence of length L. For every i, S_i belongs to a given finite set of size N and is called the hidden state at position i. The observed process $(X_i)_{1 \leq i \leq L}$ describes the

nucleotide sequence or the deduced amino-acid sequence. For every i, X_i belongs to a given finite alphabet $\mathcal{X}$ of size M and is called the observation at position i. For instance, $M := 4$ and $\mathcal{X} := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ for nucleotide sequences, and $M := 20$ for deduced amino-acid sequences.

We use HMMs of type M1Mm, hence the hidden process is a first order Markov chain. That is, the value S_i at position i depends probabilistically on the value S_{i-1} at position $i-1$. The transition matrix T is defined as

$$T(s' \,|\, s) := P(S_i = s' \,|\, S_{i-1} = s),$$

for every states s and s', that is, $T(s' \,|\, s)$ denotes the probability that $S_i = s'$, conditionally on the fact that $S_{i-1} = s$. In turn, conditionally on the state process, the observed process is an inhomogeneous Markov chain of order m whose transition probabilities at position i depend only on S_i. The emission matrix B is defined as

$$B(x \,|\, s, x_1, ..., x_m) := P(X_i = x \,|\, X_{i-1} = x_1, \ldots, X_{i-m} = x_m, S_i = s),$$

for every state s and every observations x, x_1, …, x_m. In words, the state at a given position depends on the state at the previous position, and the observation at a given position depends on the state at the same position and on the m previous observations.

Hence, the full model is specified by the pair of matrices (T, B) and by some initial distributions.

Parameter estimation

The estimation of the best model (T, B) for a given training set of sequences is usually based on maximum likelihood methods. Assume first that the segmentation of the observed sequences is available, that is, that one knows the state sequences. Then, the parameters of the model can be estimated with direct counting methods. For every states s and s', one sets

$$\widehat{T}(s' \,|\, s) := \frac{N(ss')}{N(s)},$$

where $N(s)$, respectively $N(ss')$, denotes the number of times the letter s, respectively the word ss', appears in the state sequence, that is,

$$N(s) := \sum_{i=1}^{L} \mathbf{1}\{S_i = s\}, \qquad N(ss') := \sum_{i=1}^{L-1} \mathbf{1}\{S_i = s, S_{i+1} = s'\}.$$

Likewise, for every observations x, x_1, …, x_m, and every state s, one sets

$$\widehat{B}(x \,|\, s, x_1, ..., x_m) := \frac{N(x_m \cdots x_1 x \,|\, s)}{N(x_m \cdots x_1 * \,|\, s)},$$

with

$$N(x_m \cdots x_1 * \,|\, s) := \sum_{z \in \mathcal{X}} N(x_m \cdots x_1 z \,|\, s).$$

Here $N(x_m \cdots x_1 z \,|\, s)$ is the number of times the word $x_m \cdots x_1 z$ appears in the observation sequence while the state is s at the position of the observation x, that is,

$$N(x_m \cdots x_1 z \,|\, s) := \sum_{i=m+1}^{L} \mathbf{1}\{X_{i-m} = x_{i-m}, X_{i-1} = x_{i-1}, X_i = z, S_i = s\}.$$

As is well known, maximum likelihood estimators are sensitive to overfitting. To avoid such problems, we added constant pseudo-counts n_0 to every $N(s)$, $N(ss')$ and $N(x_m \cdots x_1 z \,|\, s)$, equal to $n_0 := 1$.

Reconstruction algorithms

When the segmentation of the training sequences is not available, the maximum likelihood estimators $(\widehat{T}, \widehat{B})$ of (T, B) cannot be directly computed. But there exists several algorithms which estimate iteratively the parameters of the models with no foreknowledge of either the observation process or the state process. The most classical one is the expectation-maximization (EM) algorithm, introduced by [35]. In the context of hidden Markov chains, this algorithm is known as the Baum-Welch algorithm, see [13, 36] for a detailed description of the algorithm, and [12]. To compute maximum likelihood estimates of the parameters, this algorithm alternates E-steps and M-steps until convergence. During each E-step, the algorithm estimates the missing data (the hidden states sequence), computing the most likely state sequence with respect to the current value of the parameters, obtained through the preceding M-step. During each M-step, the algorithm maximizes the transition and emission probabilities, using the state sequence computed during the preceding E-step. There is no guarantee that the EM algorithm should produce a sequence $(T_n, B_n)_{n \geq 0}$ of models which converges to $(\widehat{T}, \widehat{B})$. Indeed, starting from an unspecified initial point (T_0, B_0), the algorithm can get stuck in one of many local maxima of the likelihood. But there exists a neighborhood of $(\widehat{T}, \widehat{B})$, such that, for any (T_0, B_0) in this neighborhood, $(T_n, B_n)_{n \geq 0}$ indeed converges to $(\widehat{T}, \widehat{B})$ ([37, 38]). To introduce some information about the composition of the different regions, we also define and use a new algorithm based on the EM algorithm and on direct counting methods. The details of this new algorithm are as follows. The emission matrix B, corresponding to the transition probabilities between observations for each state, is defined by counting on training sequences. Then one estimates iteratively the state transition probabilities of the T matrix with the EM algorithm, keeping every emission probabilities at their calculated value. The M-step of the EM algorithm is modified, to omit the usual maximization of the emission probabilities. Then, the E-step and the maximization of the transition probabilities are performed

as in the classical EM algorithm. We call this new algorithm fixed EM algorithm (which stands for EM algorithm with fixed emission probabilities). In details, the fixed EM algorithm produces a sequence $(T_n, B_n)_{n\geq 0}$ of models as follows.

Step Initiation: The transition probabilities T_0 are initialized using random values. The emission matrix B_0 is defined by counting on training sequences as follows:

$$B_0(x \,|\, s, x_1, ..., x_m) := \frac{N(x_m \cdots x_1 x \,|\, s)}{N(x_m \cdots x_1 * \,|\, s)},$$

with the same notations than in the section "Parameter estimation".

Step Estimation (E): Computation of the probability $P_{k,\ell}$ of every successives states k and ℓ in $\mathcal{S}$, under the current value (T_n, B_n).

$$P_{k,\ell} = P(S_i = k, S_{i+1} = \ell \,|\, x_1, ..., x_L, (T_n, B_n)).$$

This probability can be computed using the forward and backward variables $f_k(i)$ and $b_k(i)$ defined by:

$$f_k(i) = P(x_1, ..., x_i, S_i = k),$$

and

$$b_k(i) = P(x_{i+1}, ..., x_n \,|\, S_i = k, x_{i-m+1}, ..., x_i).$$

We have:

$$P_{k,\ell} = \frac{f_k(i) \cdot T_n(k,\ell) \cdot B_n(\ell, x_{i-m+1:i}, x_{i+1}) \cdot b_\ell(i+1)}{\sum_{k\in\mathcal{S}} f_k(n)},$$

Step Maximization (M): Computation of (T_{n+1}, B_{n+1}), through the formulas

$$T_{n+1}(k,\ell) = \frac{t(k,\ell)}{\sum_{s\in\mathcal{S}} t(k,s)},$$

where

$$t(k,\ell) = \sum_{i=1}^{n-1} f_k(i) \cdot T_n(k,\ell) \cdot B_n(\ell, x_{i-m+1:i}, x_{i+1}) \cdot b_\ell(i+1),$$

and

$$B_{n+1} = B_0.$$

Step End: The steps E and M are executed alternatively until convergence.

The fixed EM algorithm converges to the maximum likelihood estimators $(\widetilde{T}, \widetilde{B})$ conditioned by the emission matrix B. The model $(\widetilde{T}, \widetilde{B})$ yields a lower likelihood than the model $(\widehat{T}, \widehat{B})$ obtained with the EM algorithm. Experimentally, on EIAV sequences, we observe that the convergence of fixed EM occurs 10 times faster than the convergence of the EM algorithm. We defined the fixed EM algorithm to introduce some information about the number N of types of regions and the statistical composition in words of nucleotides or amino-acids of these regions. On the contrary, we introduced no information about the order or the position of the regions along the sequence.

In both EM and fixed EM algorithms, to reconstruct the hidden states sequence and to identify the predicted C and V regions, one determines the sequence of the most probable hidden states, that is, one computes at each position i of the sequence the likelihood of the different hidden states $(S_i = s)$ conditionally on the observed sequence and one selects the state with the highest likelihood. The likelihood of the hidden states for each position is computed using the classical forward-backward algorithm, described by [39].

Acknowledgements

ABQ is a recipient of an INRA fellowship.

References

1. Coffin JM: **Genetic variation in AIDS viruses**. *Cell* 1986, **46**:1–4.

2. Preston BD: **Reverse transcriptase fidelity and HIV-1 variation**. *Science* 1997, **275**(5297):228–229.

3. Preston BD, Poiesz BJ, Loeb LA: **Fidelity of HIV-1 reverse transcriptase**. *Science* 1988, **242**(4882):1168–1171.

4. Roberts JD, Bebenek K, Kunkel TA: **The accuracy of reverse transcriptase from HIV-1**. *Science* 1988, **242**(4882):1171–1173.

5. Burns DP, Collignon C, Desrosiers RC: **Simian immunodeficiency virus mutants resistant to serum neutralization arise during persistent infection of rhesus monkeys.** *Journal of Virology* 1993, **67**(7):4104–4113.

6. Leroux C, Issel CI, Montelaro RC: **Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease episode in an experimentally infected pony**. *Journal of Virology* 1997, **71**(12):9627–9639.

7. Modrow S, Hahn BH, Shaw GM, Gallo RC, Wong-Staal F, Wolf H: **Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions**. *Journal of Virology* 1987, **61**(2):570–578.

8. Valas S, Benoit C, Baudry C, Perrin G, Mamoun RZ: **Variability and immunogenicity of caprine arthritis-encephalitis virus surface glycoprotein**. *Journal of virology* 2000, **74**(13):6178–6185.

9. Zheng YH, Sentsui H, Nakaya T, Kono Y, Ikuta K: **In vivo dynamics of Equine Infectious Anemia Viruses emerging during febrile episodes : Insertions/duplications at the principal neutralizing domain**. *Journal of Virology* 1997, **71**(7):5031–5039.

10. Suarez DL, Whetstone CA: **Identification of hypervariable and conserved regions in the surface envelope gene in the bovine lentivirus**. *Virology* 1995, **212**(2):728–733.

11. Pancino G, Fossati I, Chappey C, Castelot S, Hurtrel B, Moraillon A, Klatzmann D, Sonigo P: **Structure and variations of feline immunodeficiency virus envelope glycoproteins**. *Virology* 1993, **192**(2):659–662.

12. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proceedings of the IEEE* 1989, **77**(2):257–286.

13. Churchill GA: **Stochastic models for heterogeneous DNA sequences**. *Bulletin of Mathematical Biology* 1989, **51**:79–94.

14. Krogh A: **A hidden Markov model that finds genes in *E. coli* DNA**. *Nucleic Acids Research* 1994, **22**(22):4768–4778.

15. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D: **Hidden Markov models in computational biology: Applications to protein modeling**. *Journal of Molecular Biology* 1994, **235**:1501–1531.

16. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes**. *Journal of Molecular Biology* 2001, **305**(3):567–580.

17. Nicolas P, Bize L, Muri F, Hoebeke M, Rodolphe F, Ehrlich SD, Prum B, Bessières P: **Mining *Bacilius subtilis* chromosome heterogeneities using hidden Markov models**. *Nucleic Acids Research* 2002, **30**(6):1418–1426.

18. Peshin L, Gelfand MS: **Segmentation of yeast DNA using hidden Markov models**. *Bioinformatics* 1999, **15**(12):980–986.

19. Salzberg SL, Delcher AL, Kasif S, White O: **Microbial gene identification using interpolated Markov models**. *Nucleid Acids Research* 1998, **26**(2):544–548.

20. Billingsley P: **Statistical methods in Markov chains**. *The Annals of Mathematical Statistics* 1961, **32**:12–40.

21. Miller GA: **Note on the bias of information estimates**. In *Information Theory in Psychology: Problems and Methods*. Edited by Quastler H, Glencoe, Illinois: The Free Press 1955:95–100.

22. Johnson D, Sinanovic S: **Symmetrizing the Kullback-Leibler distance**[citeseer.ist.psu.edu/johnson01symmetrizing.html].

23. Victor JD: **Asymptotic Bias in Information Estimates and the Exponential (Bell) Polynomials**. *Neural Computation* 2000, **12**:2797–2804.

24. Paninski L: **Estimation of entropy and mutual information**. *Neural Computation* 2003, **15**:1191–1253.

25. Pritchard G, Scott DJ: **The eigenvalues of the empirical transition matrix of a Markov chain**. *Journal of Applied Probability* 2004, **41A**:347–360.

26. Beard WA, Bebenek K, Darden TA, Li L, Prasad R, Kunkel TA, Wilson SH: **Vertical-scanning mutagenesis of a critical tryptophan in the minor groove binding track of HIV-1 reverse transcriptase. Molecular nature of polymerase-nucleic acid interactions**. *Journal of Biological Chemistry* 1998, **273**(46):30435–30442.

27. Bebenek K, Beard WA, Darden TA, Li L, Prasad R, Luton BA, Gorenstein DG, Wilson SH, Kunkel TA: **A minor groove binding track in reverse transcriptase**. *Nature structural Biology* 1997, **4**(3):194–197.

28. Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA: **Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor**. *Science* 1992, **256**(5065):1783–1790.

29. Vartanian JP, Meyerhans A, Sala M, Wain-Hobson S: **G→A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription**. *Proceedings of the National Academy of Sciences* 1994, **91**(8):3092–3096.

30. Wain-Hobson S, Sonigo P, Guyader M, Gazit A, Henry M: **Erratic G→A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus**. *Virology* 1995, **209**(2):297–303.

31. Craigo JK, Leroux C, Howe L, Steckbeck JD, Cook SJ, Issel RC Charles Iand Montelaro: **Transient immune suppression of inapparent carriers infected with a principal neutralizing domain-deficient equine infectious anaemia virus induces neutralizing antibodies and lowers steady-state virus replication**. *Journal of General Virology* 2002, **83**:1353–1359.

32. Leroux C, Craigo JK, Issel CI, Montelaro RC: **Equine Infectious Anemia Virus genomic evolution in progressor and nonprogressor ponies**. *Journal of Virology* 2001, **75**(10):4570–4583.

33. Zheng YH, Sentsui H, Kono Y, Ikuta K: **Mutations occurring during serial passage of Japanese equine infectious anemia virus in primary horse macrophages**. *Virus Research* 2000, **68**:93–98.

34. Zheng YH, Nakaya T, Sentsui H, Kameoka M, Kishi M, Hagiwara K, Takahashi H, Kono Y, Ikuta K: **Insertions, duplications and substitutions in restricted gp90 regions of equine infectious anaemia virus during febrile episodes in an experimentally infected horse**. *Journal of General Virology* 1997, **78**:807–820.

35. Dempster A, Laird N, Rubin D: **Maximum likelihood from incomplete data via the EM algorithm**. *Journal of the Royal Statistical Society* 1997, **39**:1–38.

36. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis*. Cambridge University Press 1998.

37. Wu C: **On the convergence properties of the EM algorithm**. *The Annals of Statistics* 1983, **11**:95–103.

38. Muri F: **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*, University Paris V 1997.

39. Baum LE, Petrie T, Soules G, Weiss N: **A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains**. *The Annals of Mathematical Statistics* 1970, **41**:164–171.

40. **Unrooted** [http://pbil.univ-lyon1.fr/software/unrooted.html].

Figures

Figure 1 - Regions predicted by the hidden Markov models on the nucleotide sequence of EIAV SU

The graphic displays the regions predicted by our mathematical model (—). The schematic organization of EIAV SU, with the position of the 8 variable regions V1 to V8 (hatched boxes) and of the 9 constant regions (—), as defined by classical amino-acid multiple alignments, is represented under the graphic. (A) HMM of order 5 with 2 hidden states, trained on the variable (V) and constant (C) regions. (B) HMM of order 5 with 9 hidden states, trained on the 8 V regions and on the reunion of the C regions.

Figure 2 - Regions predicted by the hidden Markov models on the deduced amino-acid sequence of EIAV SU

The graphic displays the regions predicted by our mathematical model (—). The schematic organization of EIAV SU with the position of the 8 variable regions V1 to V8 (hatched boxes) and of the 9 constant regions (—), as defined by classical amino-acid multiple alignments, is represented under the graphic. (A) First order HMM with 2 hidden states, trained on the V and C regions. (B) First order HMM with 9 hidden states, trained on the 8 V regions and on the reunion of the C regions. (C) First order HMM with 9 hidden states tested on an artificial sequence, where 15 amino-acids of the V7 sequence are inserted into the constant region C2 located between V1 and V2.

Figure 3 - Regions predicted on HIV sequences by hidden Markov models trained on EIAV sequences

The graphic displays the regions predicted on the HIV-1 HXB2 sequence by our mathematical models (—). The schematic organization of HIV SU with the position of the variable regions (hatched boxes) and of the constant regions (—), as defined by classical amino-acid alignments, is represented under the graphic. (A)

HMM of order 5 with 2 hidden states trained on nucleotide sequences of EIAV SU. (B) First order HMM with 2 hidden states trained on deduced amino-acid sequences of EIAV SU.

Figure 4 - Regions predicted by the combined hidden Markov model trained on EIAV, HIV, SIV, and SRLV SU

The graphs display the regions predicted by the first order combined HMM, on sequences of EIAV, HIV, SIV, SRLV, BIV and FIV SU. The schematic organization of SU with the position of the variable regions (hatched boxes) and of the constant regions (—), as defined by classical amino-acid alignments, is represented under the graphics.

Figure 5 - Graphic representation of the distances between the C and V regions of EIAV

A distance matrix between the C and V regions is computed with the symmetrised form of the relative entropy. A dendogram is evaluated with the Kitsch (Phylip 3.5c) program with the default parameters and drawn with the Unrooted program [40].

Figure 6 - Principal Components Analysis of the C/V regions of EIAV

Plot of the two first axes of the principal components analysis of the composition in words of two amino-acids of the constant C1 to C9 and variable V1 to V8 regions of EIAV.

Tables

Table 1 - GenBank accession numbers of the sequences used in this study.

Table 2 - Symmetrized relative entropy between the C/V regions of EIAV.

δ	C1	C2	C3	C4	C5	C6	C7	C8	C9	C	V	V1	V2	V3	V4	V5	V6	V7	V8
C1	0	4.40	5.08	4.63	4.65	5.08	4.26	4.70	3.52	2.55	3.70	3.62	2.80	3.67	3.77	3.37	3.70	4.00	3.25
C2		0	6.25	5.77	5.93	6.38	5.90	4.65	5.37	2.58	4.60	4.46	3.47	4.53	4.72	4.96	4.94	4.72	3.78
C3			0	6.00	6.62	5.69	5.77	5.35	5.35	2.85	4.86	4.78	4.07	4.51	4.71	4.87	5.18	4.76	4.73
C4				0	6.36	5.71	5.22	5.06	5.09	2.65	4.95	4.70	3.48	4.85	5.14	5.04	4.13	5.26	4.28
C5					0	6.86	6.08	6.17	5.00	2.97	4.88	4.81	3.64	4.96	5.35	4.95	5.41	4.71	4.54
C6						0	6.21	4.70	5.25	3.18	4.70	4.60	3.40	4.75	4.74	4.75	4.67	5.20	3.90
C7							0	4.92	4.85	2.37	4.71	5.15	3.93	5.16	4.99	4.93	4.63	4.47	4.88
C8								0	4.34	2.66	4.68	3.77	3.20	4.25	4.11	4.80	4.04	4.50	3.36
C9									0	2.48	3.84	3.92	3.22	4.16	4.32	4.13	4.84	4.00	3.32
C										0	3.35	3.64	3.19	3.70	3.78	3.62	3.99	2.96	3.01
V											0	2.25	2.07	1.91	2.11	2.09	2.21	1.87	2.36
V1												0	2.41	3.79	3.26	3.07	3.59	3.64	3.57
V2													0	2.79	2.73	2.71	2.59	2.77	2.34
V3														0	3.63	3.77	3.60	4.54	3.41
V4															0	3.57	3.66	4.09	4.03
V5																0	4.08	3.77	3.69
V6																	0	4.25	3.52
V7																		0	3.83
V8																			0

Additional Files

Appendix: On the discrimination of Markov chains through their empirical transition matrices

EIAV	AF005104 to AF005151 (except AF005113, AF005136 and AF005145 to AF005148); AF016316; AF298666 to AF298762 (except AF298752 and AF298691 to AF298694); AF429316 to AF429353
HIV	K03455; AB032740, AB03274; AF133821; AF190127, AF190128; AF197340; AF209205, AF209208; AF219261, AF219272; AF322202 to AF322214; AF411964, AF411965; AF413978, AF413979; AF413987; AF443113 to AF443115; AF457079 to AF457090 (except AF457082 to AF457084, AF457086 and AF457089); AF460972, AF460974; AF484478, AF484493; AF484507 à AF484519 (except AF484508, AF484510, AF484512 and AF484517); AF529572, AF529573; AF530576; AF544007, AF544008; AJ417424 to AJ417431; AY037268 to AY037270; AY037280 to AY037283; AY158533 to AY158535; AY173957, AY173958; AY217545; AY228556, AY228557; AY253305 to AY253322 (except AY253307, AY253309, AY253315 to AY253316 and AY253319); AY255823 to AY255827; AY322184 to AY322191 (except AY322186 and AY322188); AY357571 to AY357576 (except AY357574); AY358069 to AY358073 (except AY358070); AY371155 to AY371163 (except AY371158 to AY371162); AY423908 to AY423928; AY494965 to AY494974 (except AY494967 to AY494968, AY494970 and AY494972); AY505010, AY505011; AY535509 to AY535513; AY563169; AY818641 to AY818643
SIV	AF075269; AF103818; AF131870; AF188114 to AF188116; AF328295; AF334679; AF382828, AF382829; AF447763; AY033233; AY159321, AY159322; AY169968; AY221508 to AY221513; AY290709 to AY290716; AY523865 to AY523867; AY587015; AY588946; AY599198 to AY599201; AY611488; L20008, L20009; L20098, L20099; L40990; M29975; M33262; M58410; M66437; M83293; U04005; U10897 to U10898; U25712 to U25715; U25744, U25745; U58991; U72748
SRLV	A15114; AF015180; AF156858 to AF156877; AF338227; AF474005 to AF474007; AF479638; AJ400718 to AJ400721; AY039765 to AY039784; L06906; M31646; M33677; M34193; M60609, M60610; M60855; S51392; S55323; U35795 to U35804 (except U35797, U35802 and U35803); U51910
BIV	L43126 to L43132; M32690; NC_001413; L04972; U80989 to U80991
FIV	M25381; M36968; L00608; M59418; X57001 to X57002; M73964 to M73965; X60725; L06725; X69494 to X69502 (except X69495, X69500 and X69501)

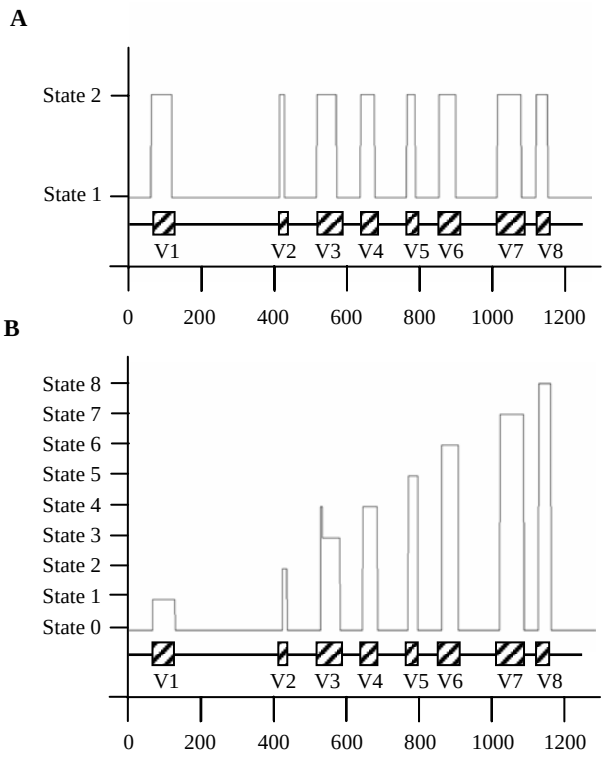
A
State 2
State 1
V1
V2
V3
V4
V5
V6
V7
V8
0
200
400
600
800
1000
1200
B
State 8
State 7
State 6
State 5
State 4
State 3
State 2
State 1
State 0
V1
V2
V3
V4
V5
V6
V7
V8
0
200
400
600
800
1000
1200
Position on the nucleotide sequence of EIAV SU

Figure 1

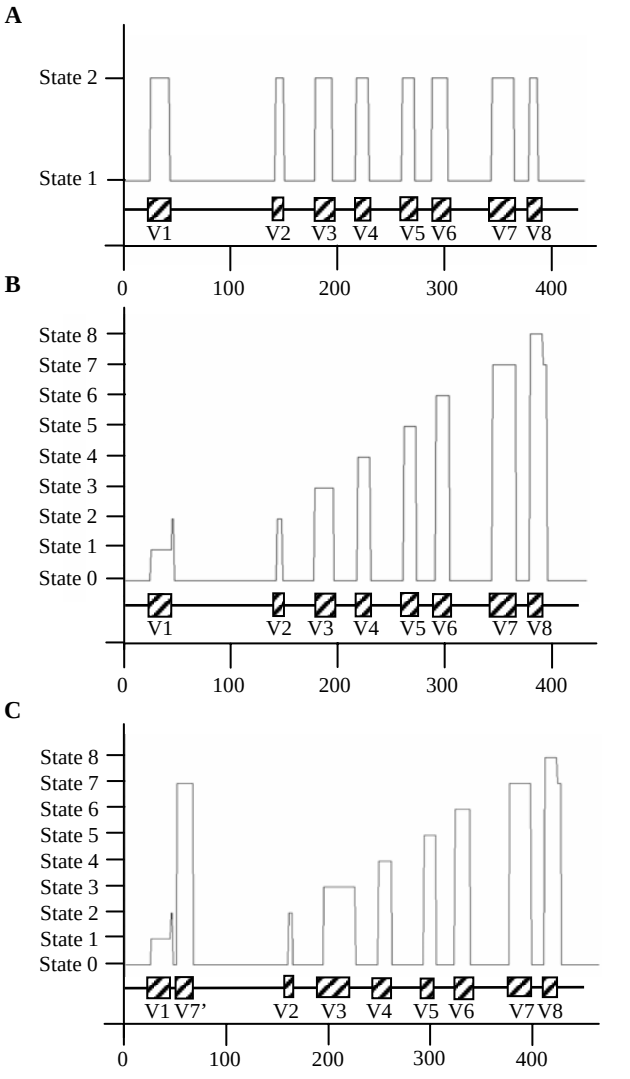
A
State 2
State 1
V1
V2
V3
V4
V5
V6
V7
V8
0
100
200
300
400
B
State 8
State 7
State 6
State 5
State 4
State 3
State 2
State 1
State 0
V1
V2
V3
V4
V5
V6
V7
V8
0
100
200
300
400
C
State 8
State 7
State 6
State 5
State 4
State 3
State 2
State 1
State 0
V1
V7'
V2
V3
V4
V5
V6
V7
V8
0
100
200
300
400
Position on the deduced amino-acid sequence of EIAV SU

Figure 2

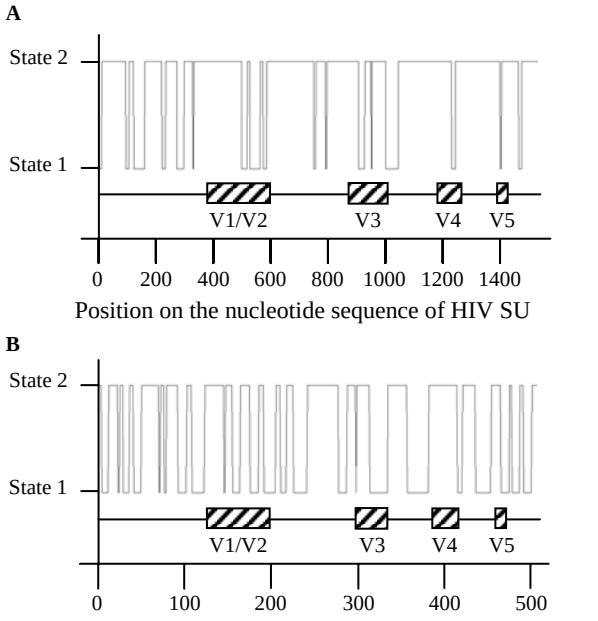
A
State 2
State 1
V1/V2
V3
V4
V5
0
200
400
600
800
1000
1200
1400
Position on the nucleotide sequence of HIV SU
B
State 2
State 1
V1/V2
V3
V4
V5
0
100
200
300
400
500
Position on the deduced amino-acid sequence of HIV SU

Figure 3

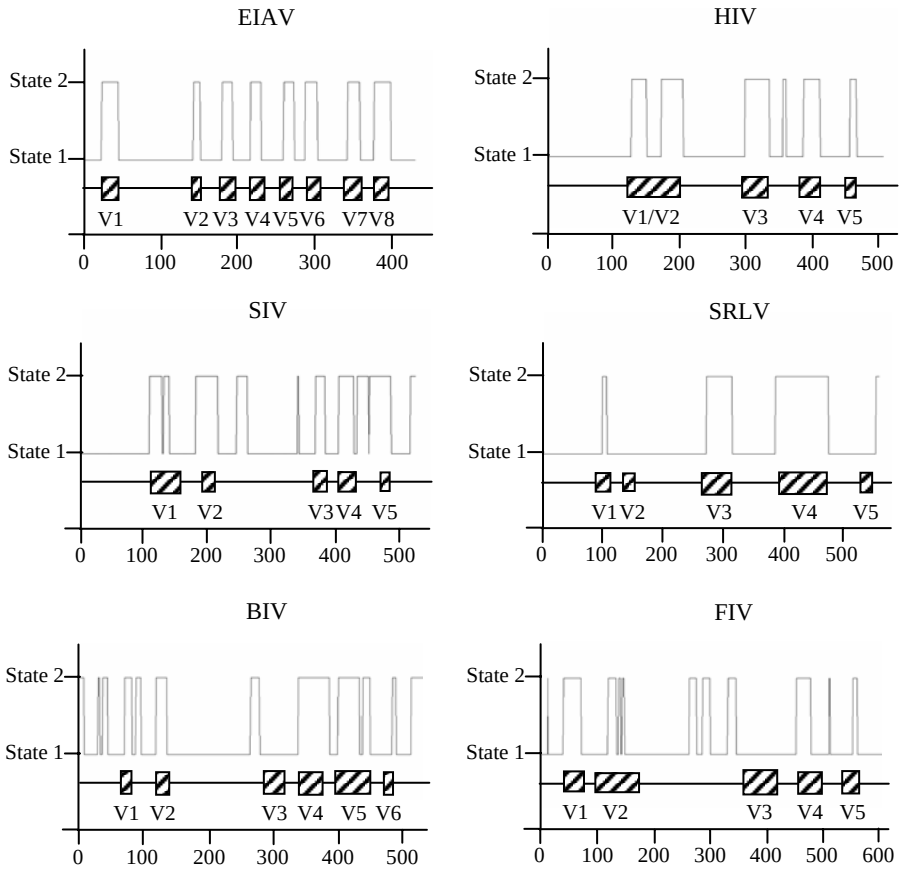
EIAV
State 2
State 1
V1
V2 V3 V4 V5 V6
V7 V8
0
100
200
300
400
HIV
State 2
State 1
V1/V2
V3
V4
V5
0
100
200
300
400
500
SIV
State 2
State 1
V1
V2
V3 V4
V5
0
100
200
300
400
500
SRLV
State 2
State 1
V1 V2
V3
V4
V5
0
100
200
300
400
500
BIV
State 2
State 1
V1
V2
V3
V4
V5
V6
0
100
200
300
400
500
FIV
State 2
State 1
V1
V2
V3
V4
V5
0
100
200
300
400
500
600
Position on the deduced amino-acid sequences

Figure 4

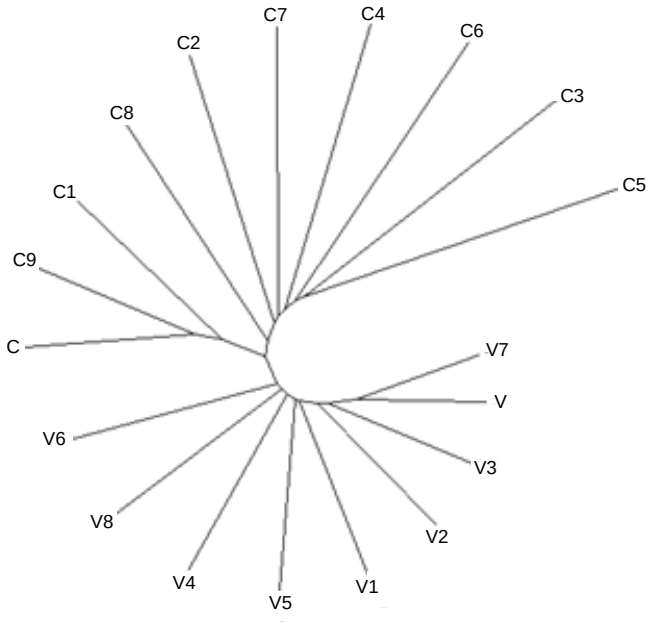

Figure 5

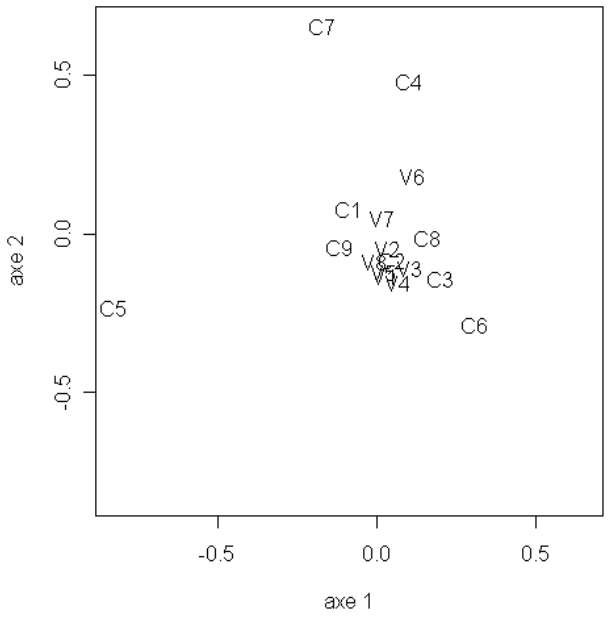

Figure 6

Appendix: On the discrimination of Markov chains through their empirical transition matrices

Aurélia Boissin-Quillon[1], Didier Piau[2] and Caroline Leroux[1]

[1] UMR754 INRA-ENVL-UCBL"Rétrovirus et Pathologie Comparée", IFR 128 BioSciences Lyon-Gerland, Université Claude Bernard Lyon 1, Domaine de Gerland, 69007 Lyon, France
[2] Institut Fourier UMR 5582 CNRS-UJF, Université Joseph Fourier (Grenoble 1), 100 rue des Maths, BP 74, 38402 Saint Martin d'Hères, France

Email: Aurélia Boissin-Quillon - aurelia.quillon@univ-lyon1.fr; Didier Piau - Didier.Piau@ujf-grenoble.fr; Caroline Leroux*- caroline.leroux@univ-lyon1.fr;

*Corresponding author

Introduction

We consider irreducible Markov chains on a finite number k of states, with transition matrix q and stationary distribution p. The number ℓ of edges used by the chain is the number of couples of states (x, y) such that $q(x, y) > 0$, hence $k \leq \ell \leq k^2$. When $\ell = k$, the chain moves deterministically on an oriented discrete circle, hence one can exclude this case if necessary. On the contrary, as soon as $\ell \geq k + 1$, several trajectories are possible and the chain is truly random. Finally, $\ell = k^2$ means that all the transitions are allowed, hence the chain moves on the complete graph with loops. The dimension $D(q)$ of the chain is

$$D(q) := \ell - k,$$

This the number of free parameters among the nonzero coefficients of q, that is, the dimension of the simplex formed by the transition matrices subordinated to q, in the sense that the coefficients corresponding to coefficients of q equal to 0 have to be equal to 0 too.
The maximum likelihood estimator $\widehat{q}$ of q uses countings along a trajectory of length n and is a consistent estimator of q when n goes to infinity. The relation

$$\widehat{q}(x, y) = q(x, y) + z_{xy}/\sqrt{n} + o(1/\sqrt{n}),$$

defines a Gaussian centered vector $(z_{xy})_{xy}$, indexed by the edges (x, y), and whose covariance matrix is an explicit function of p and q.

We consider the relative entropy of the empirical measure, given by the observed trajectory, with respect to the theoretical measure, given by p and q. This random entropy is defined by

$$H(\widehat{q},q) := \sum_{(x,y)} \widehat{p}_x\, \widehat{q}(x,y)\, \log(\widehat{q}(x,y)/q(x,y)),$$

where the sum indexed by (x,y) has ℓ terms and $\widehat{p}$ denotes the stationary distribution of $\widehat{q}$. One can also consider the entropy

$$H(q,\widehat{q}) := \sum_{(x,y)} p_x\, q(x,y)\, \log(q(x,y)/\widehat{q}(x,y)).$$

Using second-order Taylor series approximations of the logarithm function, one sees that both $H(\widehat{q},q)$ and $H(q,\widehat{q})$ are such that, when n becomes large,

$$H = h/(2n) + o(1/n), \quad h := \sum_{(x,y)} z_{xy}^2\, p_x/q(x,y).$$

In this appendix we show that the reduced relative entropy h follows a quite simple χ^2 distribution and we draw some statistical consequences from this result.

Convergence in distribution

Let N_x, respectively N_{xy}, denote the number of times the vertice x, respectively the edge (x,y), is visited up to time n. Consider

$$\xi_{xy} := (N_{xy} - q(x,y)N_x)/\sqrt{N_x}.$$

According to [P. Billingsley (1960). Statistical Inference in Markov Chain. The Stanford meetings of the Institute of Mathematical Statistics. Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago, Ill. 1961], the matrices $(\xi_{xy})_{xy}$ converge in distribution, when n goes to infinity, to a Gaussian centered matrix $(g_{xy})_{xy}$ distributed as follows. The vectors $(g_{xy})_y$ are independent for different states x, hence the covariance of g_{xy} and g_{zt} is 0 for every $x \neq z$ and every y and t. Finally, for every x, y and z,

$$E(g_{xy}g_{xz}) = -q(x,y)q(x,z) \quad (y \neq z), \qquad E(g_{xy}^2) = q(x,y)(1-q(x,y)).$$

Since N_x/n converges almost surely to p_x, one can replace the factor $1/\sqrt{N_x}$ by $1/\sqrt{np_x}$. This remark yields the following convergence in distribution:

$$np_x(\widehat{q}(x,y) - q(x,y))^2 \to g_{xy}^2.$$

In addition, we recall that, if one observes an i.i.d. sequence with theoretical distribution p on k states, then the empirical distribution $\widehat{p}$ is such that $2nH(\widehat{p}, p)$ converges in distribution to a χ^2 distribution with $k-1$ degrees of freedom.

The vectors $(g_{xy})_{xy}$ are independent. Furthermore, for each fixed x, $(g_{xy})_y$ admits the same covariances that the limit gaussian distribution obtained for an i.i.d. sequence of distribution $q(x,\cdot)$. In addition, the random variable

$$H_x := \sum_y q(x,y)\log(q(x,y)/\widehat{q}(x,y))$$

corresponds to the observation of this i.i.d. process during a time which corresponds to the number of visits of x before n, that is, a random number of visits which is $np_x + o(n)$. Hence, $2(np_x)H_x$ converges in distribution to the χ^2 distribution with $D_x(q)$ degrees of freedom, where $D_x(q)+1$ equal the number of y such that $q(x,y) > 0$. By independance of the limits in distribution of the $2np_xH_x$, their sum $2nH$ converges in distribution to the χ^2 distribution with $D(q)$ degrees of freedom, where $D(q)$ is the sum indexed by x of the $D_x(q)$.

In conclusion, h follows the χ^2 distribution with $D(q)$ degrees of freedom.

Statistical applications

Assume that one has two independent sequences of observations of the same Markov chain with transition matrix q. This yields two estimators $\widehat{q}_1$ and $\widehat{q}_2$ of q, based respectively on the countings $N^{(1)}$ and $N^{(2)}$. We proved the relations

$$\widehat{q}_i(x,y) = q(x,y) + z_{xy}^{(i)}/\sqrt{n} + o(1/\sqrt{n}), \quad i = 1, 2,$$

where the two families $(z_{xy}^{(1)})_{xy}$ and $(z_{xy}^{(2)})_{xy}$ are independent and follow the distribution of $(z_{xy})_{xy}$ described in the previous section. The reduced relative entropy between the two sequences of observations is asymptotically equal to

$$h(\widehat{q}_1, \widehat{q}_2) := \sum_{(x,y)} (z_{xy}^{(1)} - z_{xy}^{(2)})^2 \alpha_{xy},$$

where α_{xy} can be indifferently $p_x^{(1)}/q_1(x,y)$ or $p_x^{(2)}/q_2(x,y)$ or $p_x/q(x,y)$. If one uses $\alpha_{xy} = p_x/q(x,y)$, then $h(\widehat{q}_1, \widehat{q}_2)$ follows exactly the distribution of $2h(\widehat{q}, q)$, and the same result holds asymptotically for the other choices of $\alpha_{x,y}$. Hence, to determine whether $\widehat{q}_1$ and $\widehat{q}_2$ correspond to the same Markov chain or not, one can use the fact that, if they do, $nH(\widehat{q}_1, \widehat{q}_2)$ is asymptotically χ^2 with $D(q)$ degrees of freedom. In particular,

$$E(H(\widehat{q}_1, \widehat{q}_2)) \sim D(q)/n.$$

If one wishes to work instead with the symmetrized form of the relative entropy, one can use

$$\zeta := n(H(\widehat{q}_1, \widehat{q}_2) + H(\widehat{q}_2, \widehat{q}_1)) = \sum_{(x,y)} (N_{xy}^{(1)} - N_{xy}^{(2)}) \log \left(\frac{N_{xy}^{(1)} N_x^{(2)}}{N_x^{(1)} N_{xy}^{(2)}} \right).$$

Since ζ is asymptotically twice a χ^2 with $D(q)$ degrees of freedom, one can compute the p-value of the event $\{\zeta \geq t\}$ for every $t \geq 2D(q)$.

To compute upper bounds of the p-values of χ^2 distributions of large dimension d, one can use exponential Cramer bounds. This yields that, for every $t \geq d$, the probability that a χ^2 distribution with d degrees of freedom is greater than t is at most

$$e^{-t/2}(te/d)^{d/2}.$$

This approximation yields for instance that, if $d = 400 - 20 = 380$, the p-value for $t = 460$ is less than 2.47 % and the p-value for $t = 480$ is less than 0.04 %, to be compared to the true p-values 0.3% and 0.03% respectively.

Bibliographie

Andrieu, O., Fiston, A. S., Anxolabéhère, D., and Quesneville, H. 2004. Detection of transposable elements by their composition bias. *BMC Bioinformatics*, 5 :94.

Asai, K., Hayamizu, S., and Handa, K. 1993. Prediction of protein secondary structure by the hidden Markov model. *CABIOS*, 9(2) :141–146.

Audic, S. and Claverie, J.-M. 1998. Self-identification of protein-coding regions in microbial genomes. *Proceedings of the National Academy of Sciences*, 95 :10026–10031.

Bakhanashvili, M. and Hizi, A. 1993. The fidelity of the reverse transcriptases of human immunodeficiency viruses and murine leukemia virus, exhibited by the mispair extension frequencies, is sequence dependent and enzyme related. *FEBS letters*, 319 :201–205.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. 1994. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences*, 91 :1059–1063.

Battula, N. and Loeb, L. A. 1976. On the fidelity of DNA replication. Lack of exodeoxyribonuclease activity and errorcorrecting function in avian myeloblastosis virus DNA polymerase. *Journal of Biological Chemistry*, 251(4) :982–986.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41 :164–171.

Beard, W. A., Bebenek, K., Darden, T. A., Li, L., Prasad, R., Kunkel, T. A., and Wilson, S. H. 1998. Vertical-scanning mutagenesis of a critical tryp-

tophan in the minor groove binding track of HIV-1 reverse transcriptase. molecular nature of polymerase-nucleic acid interactions. *Journal of Biological Chemistry*, 273(46) :30435–30442.

Bebenek, K., Beard, W. A., Darden, T. A., Li, L., Prasad, R., Luton, B. A., Gorenstein, D. G., Wilson, S. H., and Kunkel, T. A. 1997. A minor groove binding track in reverse transcriptase. *Nature structural Biology*, 4(3) :194–197.

Bebenek, K. and Kunkel, T. A. 1993. The fidelity of retroviral reverse transcriptase. In Skalka, A. M. and Goff, S. P., editors, *Reverse Transcriptase*, pages 85–102. Cold Spring Harbor Laboratory Press.

Bellman, R. 1957. *Dynamic Programming*. Princeton University Press.

Bickel, P. J., Kechris, K. J., Spector, P. C., Wedemayer, G. J., and Glazer, A. N. 2002. Finding important sites in protein sequences. *Proceedings of the National Academy of Sciences*, 99 :14764–14771.

Billingsley, P. 1961a. *Statistical inference for Markov processes*, volume 2 of *Statistical Research Monographs*. The University of Chicago Press.

Billingsley, P. 1961b. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, 32 :12–40.

Bird, A. P. 1986. CpG rich islands and the function of DNA methylation. *Nature*, 321 :209–213.

Borodovsky, M. and McIninch, J. 1993. GENMARK : Parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2) :123–133.

Boys, R. J., Henderson, D. A., and Wilkinson, D. J. 2000. Detecting homogeneous segments in DNA sequences using hidden Markov models. *Journal of the Royal Statistical Society, Series C*, 49(2) :269–285.

Burge, C., Campbell, A. M., and Karlin, S. 1992. Over and under–representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, 89 :1358–1362.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268 :78–94.

Burns, D. P., Collignon, C., and Desrosiers, R. C. 1993. Simian immunodeficiency virus mutants resistant to serum neutralization arise during persistent infection of rhesus monkeys. *Journal of Virology*, 67(7) :4104–4113.

Cavalier-Smith, T. 2004. Only six kingdoms of life. *Proceedings of the Royal Society B*, 271 :1251–1262.

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

Churchill, G. A. 1992. Hidden Markov chains and the analysis of genome structure. *Computer and Chemistry*, 16(2) :107–115.

Clavel, F., Hoggan, M. D., Willey, R. L., Strebel, K., Martin, M. A., and Repaske, R. 1989. Genetic recombination of human immunodeficiency virus. *Journal of Virology*, 63(3) :1455–1459.

Coffin, J. M. 1992. Structure and classification of retroviruses. In Levy, J. A., editor, *The retroviridae*, pages 19–49. Plenum Press.

Coffin, J. M. 1995. HIV population dynamics *in vivo* : Implications for genetic variation, pathogenesis and therapy. *Science*, 267 :483–489.

Coffin, J. M., Hughes, S. H., and Varmus, H. E. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press.

Cook, R. F., Cook, S. J., Berger, S. L., Leroux, C., Ghabrial, N. N., Gantz, M., Bolin, P. S., Mousel, M. R., Montelaro, R. C., and Issel, C. J. 2003. Enhancement of equine infectious anemia virus virulence by identification and removal of suboptimal nucleotides. *Virology*, 313(2) :588–603.

Craigo, J. K., Leroux, C., Howe, L., Steckbeck, J. D., Cook, S. J., and Issel, Charles I.and Montelaro, R. C. 2002. Transient immune suppression of inapparent carriers infected with a principal neutralizing domain-deficient equine infectious anaemia virus induces neutralizing antibodies and lowers steady-state virus replication. *Journal of General Virology*, 83 :1353–1359.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 :1–38.

Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences*, 88 :7160–7164.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*. Cambridge University Press.

Eckert, K. A. and Kunkel, T. A. 1990. High fidelity DNA synthesis by *Thermus aquaticus* DNA polymerase. *Nucleic Acids Research*, 18 :3739–3743.

Ephraim, Y. 2002. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6) :1518–1569.

Foil, L. D., Adams, W. V., McManus, J. M., and Issel, C. J. 1987. Bloodmeal residues on mouthparts of Tabanus fuscicostatus (Diptera : Tabanidae) and the potential for mechanical transmission of pathogens. *Journal of Medical Entomology*, 24(6) :613–616.

Gao, F., Yue, L., Robertson, D. L., Hill, S. C., Hui, H., Biggar, R. J., Neequaye, A. E., Whelan, T. M., Ho, D. D., Shaw, G. M., Sharp, P. M., and Hahn, B. H. 1994. Genetic diversity of human immunodeficiency virus type 2 : evidence for distinct sequence subtypes with differences in virus biology. *Journal of Virology*, 68(11) :7433–7447.

Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2) :261–282.

Hammond, S. A., Li, F., McKeon, Sr, B. M., Cook, S. J., Issel, C. J., and Montelaro, R. C. 2000. Immune responses and viral replication in long-term inapparent carrier ponies inoculated with equine infectious anemia virus. *Journal of Virology*, 74(13) :5968–5981.

Henderson, J., Salzberg, S., and Fasman, K. H. 1997. Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, 4(2) :127–141.

Howe, L., Leroux, C., Issel, C. J., and Montelaro, R. C. 2002. Equine infectious anemia virus envelope evolution in vivo during persistent infection progressively increases resistance to in vitro serum antibody neutralization as a dominant phenotype. *Journal of Virology*, 76(21) :10588–10597.

Jacob, F. and Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3 :318–356.

Johnson, D. and Sinanovic, S. 2001. Symmetrizing the Kullback-Leibler distance. Disponible à l'adresse :
http ://citeseer.ist.psu.edu/johnson01symmetrizing.html.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, 256(5065) :1783–1790.

Krogh, A. 1994. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Research*, 22 :4768–4778.

Krogh, A. 1997. Two methods for improving performance of a HMM and their application for gene finding. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 179–186.

Krogh, A. 2000. Using database matches with HMMgene for automated gene detection in Drosophila. *Genome Research*, 10(4) :523–528.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology : Applications to protein modeling. *Journal of Molecular Biology*, 235 :1501–1531.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a Hidden Markov Model : application to complete genomes. *Journal of Molecular Biology*, 305 :567–580.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. 1992. CpG islands as gene markers in the human genome. *Genomics*, 13(4) :1095–1107.

Leroux, C., Cadoré, J.-L., and Montelaro, R. C. 2004. Equine Infectious Anemia Virus (EIAV) : what has HIV's country cousin got to tell us ? *Veterinary Research*, 35 :485–512.

Leroux, C., Chastang, J., Greenland, T., and Mornex, J.-F. 1997a. Genomic heterogeneity of small ruminant lentiviruses : existence of heterogeneous populations in sheep and of the same lentiviral genotypes in sheep and goats. *Archives of Virology*, 142(6) :1125–1137.

Leroux, C., Craigo, J. K., Issel, C. I., and Montelaro, R. C. 2001. Equine Infectious Anemia Virus genomic evolution in progressor and nonprogressor ponies. *Journal of Virology*, 75(10) :4570–4583.

Leroux, C., Issel, C. J., and Montelaro, R. C. 1997b. Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease episode in an experimentally infected pony. *Journal of Virology*, 71(12) :9627–9639.

Leroux, C., Montelaro, R. C., Sublime, E., and Cadoré, J. L. 2005. EIAV (equine infectious anemia virus) : mieux comprendre la phatogenèse des infections lentivirales. *Virologie*, 9 :1–12.

Levy, J. A. 2006. HIV pathogenesis : knowledge gained after two decades of research. *Advances in Dental Research*, 19(1) :10–16.

Lignée, M. 1843. Mémoire et observations sur une maladie de sang, connue sous le nom d'anhémie hydrohémie, cachexie acquise du cheval. *Recueil de Médecine Vétérinaire*, 20 :30–44.

Lukashin, A. V. and Borodovsky, M. 1998. GeneMark.hmm : new solutions for gene finding. *Nucleic Acids Research*, 26(4) :1107–1115.

McDonald, I. L. and Zucchini, W. 1997. *Hidden Markov and other models for discrete-valued times series.* Chapman and Hall.

Miller, G. A. 1955. Note on the bias of information estimates. In Quastler, H., editor, *Information Theory in Psychology : Problems and Methods*, pages 95–100. The Free Press, Glencoe, Illinois.

Modrow, S., Hahn, B. H., Shaw, G. M., Gallo, R. C., Wong-Staal, F., and Wolf, H. 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates : prediction of antigenic epitopes in conserved and variable regions. *Journal of Virology*, 61(2) :570–578.

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 Paris V.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B., and Bessières, P. 2002. Mining *Bacilius subtilis* chromosome heterogeneities using hidden Markov models. *Nucleic Acids Research*, 30(6) :1418–1426.

Nirenberg, M. W., Matthaei, J. H., Jones, O. W., Martin, R. G., and Barondes, S. H. 1963. Approximation of genetic code via cell-free protein synthesis directed by template RNA. *Federation Proceedings*, 22 :55–61.

Pancino, G., Fossati, I., Chappey, C., Castelot, S., Hurtrel, B., Moraillon, A., Klatzmann, D., and Sonigo, P. 1993. Structure and variations of feline immunodeficiency virus envelope glycoproteins. *Virology*, 192(2) :659–662.

Paninski, L. 2003. Estimation of entropy and mutual information. *Neural Computation*, 15 :1191–1253.

Pathak, V. K. and Temin, H. M. 1990. Broad spectrum of *in vivo* forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle : Substitutions, frameshifts, and hypermutations. *Proceedings of the National Academy of Sciences*, 87 :6019–6023.

Peshin, L. and Gelfand, M. S. 1999. Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, 15(12) :980–986.

Preston, B. D. 1997. Reverse transcriptase fidelity and HIV-1 variation. *Science*, 275(5297) :228–229.

Preston, B. D., Poiesz, B. J., and Loeb, L. A. 1988. Fidelity of HIV-1 reverse transcriptase. *Science*, 242(4882) :1168–1171.

Pritchard, G. and Scott, D. J. 2004. The eigenvalues of the empirical transition matrix of a Markov chain. *Journal of Applied Probability*, 41A :347–360.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. 1988. The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882) :1171–1173.

Robertson, D. L., Hahn, B. H., and Sharp, P. M. 1995a. Recombination in aids viruses. *Journal of molecular evolution*, 40(3) :249–259.

Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H. 1995b. Recombination in HIV-1. *Nature*, 374(6518) :124–126.

Shamir, R. 2001. Algorithms for molecular biology. Lecture notes. Disponible à l'adresse :
http ://www.math.tau.ac.il/∼rshamir/algmb/01/scribe01/lec05.ps.gz.

Suarez, D. L. and Whetstone, C. A. 1995. Identification of hypervariable and conserved regions in the surface envelope gene in the bovine lentivirus. *Virology*, 212(2) :728–733.

Takeuchi, Y., Nagumo, T., and Hoshino, H. 1988. Low fidelity of cell-free DNA synthesis by reverse transcriptase of human immunodeficiency virus. *Journal of Virology*, 62 :3900–3902.

Valas, S., Benoit, C., Baudry, C., Perrin, G., and Mamoun, R. Z. 2000. Variability and immunogenicity of caprine arthritis-encephalitis virus surface glycoprotein. *Journal of virology*, 74(13) :6178–6185.

Vartanian, J. P., Henri, M., and Wain-Hobson, S. 2002. Sustained G→A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. *Journal of General Virology*, 83 :801–805.

Vartanian, J. P., Meyerhans, A., Sala, M., and Wain-Hobson, S. 1994. G→A hypermutation of the human immunodeficiency virus type 1 genome : evidence for dCTP pool imbalance during reverse transcription. *Proceedings of the National Academy of Sciences*, 91(8) :3092–3096.

Victor, J. D. 2000. Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Computation*, 12 :2797–2804.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13 :260–269.

Wain-Hobson, S. 1989. HIV genome variability *in vivo*. *AIDS*, 3 :S13–S18.

Wain-Hobson, S. 1993. The fastest genome evolution ever described : HIV variation in situ. *Current opinion in genetics and development*, 3(6) :878–883.

Wain-Hobson, S., Sonigo, P., Guyader, M., Gazit, A., and Henry, M. 1995. Erratic G→A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus. *Virology*, 209(2) :297–303.

Watson, J. D. and Crick, F. H. C. 1953. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171 :737–738.

Weber, J. and Grosse, F. 1989. Fidelity of human immunodeficiency virus type 1 reverse transcriptase in copying natural DNA. *Nucleic Acids Research*, 17 :1379–1393.

Wu, C. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1) :95–103.

Zheng, Y.-H., Nakaya, T., Sentsui, H., Kameoka, M., Kishi, M., Hagiwara, K., Takahashi, H., Kono, Y., and Ikuta, K. 1997a. Insertions, duplications and substitutions in restricted gp90 regions of equine infectious anaemia virus during febrile episodes in an experimentally infected horse. *Journal of General Virology*, 78 :807–820.

Zheng, Y. H., Sentsui, H., Kono, Y., and Ikuta, K. 2000. Mutations occurring during serial passage of japanese equine infectious anemia virus in primary horse macrophages. *Virus Research*, 68(1) :93–98.

Zheng, Y.-H., Sentsui, H., Nakaya, T., Kono, Y., and Ikuta, K. 1997b. In vivo dynamics of Equine Infectious Anemia Viruses emerging during febrile episodes : Insertions/duplications at the principal neutralizing domain. *Journal of Virology*, 71(7) :5031–5039.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. 2002. Human immunodeficiency virus type 1 recombination : Rate, fidelity, and putative hot spots. *Journal of Virology*, 76(22) :11273–11282.

RESUME

Les lentivirus présentent une évolution rapide du gène *env*, notamment dans la région codant la glycoprotéine de surface (SU). Un fait remarquable est que les mutations de la SU sont localisées dans des zones spécifiques, appelées régions variables (V), séparées par des régions dites constantes (C). Afin de déterminer s'il existe des signatures spécifiques des régions C et V, nous avons développé des modèles de Markov cachés, ou HMM (Hidden Markov Models), basés sur la composition en oligonucléotides de chaque type de région, capables de différencier les régions C et V des lentivirus. Nous avons entraîné des modèles de Markov cachés sur des séquences des SU des lentivirus équins, humains, simiens et des petits ruminants. Nous avons obtenu une délimitation claire des régions C et V de tous ces lentivirus ainsi que des lentivirus bovins et félins qui n'avaient pas été utilisés pour définir les modèles. Nos résultats suggèrent que les régions C et V des lentivirus ont des compositions statistiques en mots de nucléotides et d'acides aminés différentes. Des signatures caractéristiques des régions C et V ont été extraites à partir des modèles définis.

TITLE

In silico markovian prediction of variable and constant regions of lentivirus genomes

ABSTRACT

Lentiviruses exhibit a considerable plasticity of the *env* gene, particularly in the region which encodes the surface (SU) glycoprotein. Interestingly, the SU mutations are clustered in restricted areas, called variable (V) regions and interspaced by more stable regions, called constant regions (C). To determine if the C and V regions are characterized by specific signatures, we developed HMM (Hidden Markov Models), based on the statistical oligonuclotide composition of each type of region, aimed at differentiating the C and V regions of the lentiviruses. We trained hidden Markov models on SU sequences of the equine, human, simian and small ruminant lentiviruses. We obtained a clear delimitation of the C and V regions of all these lentiviruses and of bovine and feline lentiviruses which were not used to define the models. Our results suggest that the C and V regions of the lentiviruses have statistical differences in their composition in words of nucleotides or amino-acids. Specific signatures of the C and V regions have been extracted from our models.

MOTS-CLES

Modèles de Markov cachés, séquences, lentivirus, modélisation, mutations.
Hidden Markov Models, HMM, sequences, lentivirus, modelling, mutations.

INTITULE ET ADRESSE DU LABORATOIRE

UMR 754 INRA-ENVL-UCBL «Rétrovirus et Pathologie Comparée»
50 avenue Tony Garnier
69366 Lyon cedex 07