Laboratoire des Sciences de l'Image,
de l'Informatique et de la Télédétection

UMR 7005 UDS/CNRS

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur

Thèse

présentée pour l'obtention du grade de

Docteur de l'Université de Strasbourg
Discipline : Informatique

par

Emmanuel Benard

Extension des modèles stochastiques de substitution de nucléotides et son implémentation informatique

Soutenue publiquement le 6 décembre 2011

Membres du jury

Directeur de thèse :	Christian Michel, Professeur, Université de Strasbourg
Rapporteur :	Gilles Bernot, Professeur, Université de Nice Sophia Antipolis
Rapporteur :	Alain Denise, Professeur, Université Paris-Sud 11
Examinatrice :	Dominique Bechmann, Professeure, Université de Strasbourg
Examinateur :	Sylvain Contassot-Vivier, Professeur, Université de Nancy

À ma famille

Remerciements

Mes premiers remerciements vont à mon directeur de thèse, Christian Michel, pour la confiance qu'il m'a accordée durant ces quelques années pendant lesquelles j'ai pu apprécier, entre autres, sa passion pour la recherche, sa rigueur scientifique, sa grande disponibilité et sa patience à toute épreuve.

Je tiens ensuite à remercier Dominique Bechmann, Gilles Bernot, Sylvain Contassot-Vivier et Alain Denise de m'avoir fait l'honneur d'accepter d'être membres du jury, de l'intérêt qu'ils ont porté à mes travaux, ainsi que pour leurs remarques enrichissantes, notamment lors de la séance de discussion.

De façon générale, merci à l'ensemble du personnel du *Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection*, merci à toute l'équipe *Bioinformatique théorique, Fouille de données et Optimisation stochastique*. En particulier aux autres membres de la thématique *bioinformatique théorique*, à savoir et chronologiquement : Gabriel Frey pour ses conseils avisés en début de thèse et Sophie Lèbre pour son investissement au cours des derniers mois, sa patience et ses précieux conseils. Merci à Bruno Belarte, Camille Kurtz, François Petitjean, Aurélie Bertaux, Olivier Khun, Matthieu Hermann, Stéphane Querry et Patrice Roy pour la bonne humeur générale. Quelques remerciements spéciaux à Ogier Maitre et Jonathan Weber, compagnons de rédaction, je pense que nous nous souviendrons longtemps de l'été 2011. Merci également à Maurice Lanselle avec qui les pauses café étaient toujours instructives.

Merci à l'ensemble des personnes du Département Informatique avec lesquelles j'ai pu partager divers enseignements en tant que doctorant-moniteur puis A.T.E.R. et merci à Marie-Claire Hantsch.

Plus personnellement, merci à Guillaume Buwalda, Elsa Dollé, Camille Godron, Jean-Daniel Lomenède, Nicolas Philippe, Mikaël Salson, Élodie Sérurier-Duceau et Nicolas Toublanc pour leur soutien et les moments partagés de détente estivale.

Merci à mes parents, Françoise et Ludovic Bénard, de m'avoir, entre autres, toujours aidé et soutenu. Merci également à ma soeur, Stéphanie Bénard et son compagnon, Damien Lainey, ainsi qu'à ma grand-mère, Louise. Je souhaite adresser mes derniers remerciements personnels à Jackie et Jean Carré, mes premiers instituteurs, qui m'ont fait l'honneur d'assister à ma soutenance malgré les « quelques kilomètres ».

Enfin, merci aux personnes qui auraient dû apparaître dans ces remerciements et que j'aurais éventuellement oubliées.

Table des matières

Introduction **11**
Contexte . . . 11
Objectifs de la thèse . . . 11
Plan . . . 12

1 Éléments biologiques, de théorie des codes et modèles d'évolution **13**
1.1 Rappels de biologie moléculaire . . . 13
1.1.1 Classification du vivant . . . 13
1.1.2 L'ADN . . . 13
1.1.3 L'ARN . . . 14
1.1.4 Protéines . . . 15
1.1.5 Traduction et code génétique . . . 15
1.1.6 Gènes et séquences codantes . . . 16
1.1.7 Épissage . . . 17
1.1.8 Mutations . . . 17
1.1.9 Contenu en G+C et effet CpG . . . 18
1.2 Théorie des codes circulaires dans les gènes . . . 19
1.2.1 Définitions et notations préliminaires . . . 19
1.2.2 Hypothèse initiale des codes sans ponctuation dans les gènes . . . 22
1.2.3 Découverte des codes circulaires dans les gènes . . . 23
1.3 Modèles probabilistes d'évolution de séquences génétiques . . . 24
1.3.1 Propriétés générales des modèles probabilistes de substitution . . . 25
1.3.2 Modèles probabilistes de substitution de nucléotides . . . 26
1.3.3 Modèles d'évolution de motifs avec dépendance au contexte . . . 29
1.3.4 Modèles d'évolution de probabilités d'occurrence de motifs . . . 31

2 Modèles biomathématiques d'évolution **33**
2.1 Équations différentielles des modèles stochastiques d'évolution par substitution . 33
2.1.1 Définitions préalables et notations . . . 34
2.1.2 Dérivation . . . 34
2.1.3 Équation différentielle matricielle . . . 35
2.2 Méthode de résolution analytique . . . 35
2.2.1 Solution de l'équation différentielle . . . 36
2.2.2 Inversion du temps . . . 36
2.2.3 Modèles analytiques basés sur le modèle 3PS . . . 37

2.2.4 Modèle analytique généralisé basé sur le modèle 3PS 38
2.2.5 Généralisation du modèle 3PS à un modèle à quatre paramètres de substitution par site 47
2.3 Méthode de résolution numérique 49
2.3.1 Résolution numérique avec paramètres de substitution fonction du temps 50
2.3.2 Inversion du temps 50
2.4 Modèle numérique inverse avec blocage des substitutions 51
2.4.1 Motivation théorique 51
2.4.2 Principe 51
2.4.3 Algorithme de blocage dynamique des substitutions 52
2.5 Conclusion 54

3 Structures de données et algorithmes d'implantation des formules 55
3.1 Implémentation avec Mathematica et webMathematica 55
3.2 Principales contraintes 56
3.2.1 Taille des formules analytiques 56
3.2.2 Fonctionnement sur un PC standard 56
3.2.3 Contrainte liée à l'unification des modèles analytiques 56
3.3 Description des différentes approches envisagées 56
3.3.1 Approche classique 56
3.3.2 Méthode formelle « pure » 57
3.3.3 Méthode formelle « mixte » 57
3.4 Comparaison des méthodes de calcul « pure » et « mixte » 59
3.4.1 Choix de la méthode utilisée 61
3.5 Conclusion 61

4 Principe et fonctionnalités du logiciel SEGM 63
4.1 Développement d'une première version de SEGM 63
4.2 Seconde version de SEGM 64
4.2.1 Taille des motifs et probabilités d'occurrence initiales 64
4.2.2 Options de base 64
4.2.3 Options conditionnelles 67
4.2.4 Résultats 70
4.3 Principe de fonctionnement du logiciel SEGM 71
4.4 Conclusion 72

5 Applications bioinformatiques 75
5.1 Évolution des sites d'épissage des introns du génome humain 75
5.1.1 Données utilisées 76
5.1.2 Résultats de l'étude évolutive inverse des sites donneurs 76
5.1.3 Résultats de l'étude évolutive inverse des sites accepteurs 83
5.1.4 Conclusion 88
5.2 Évolution inverse du code circulaire X(EUK_PRO) 89
5.2.1 Base de données utilisée 89

5.2.2 Algorithme du modèle d'évolution numérique inverse avec blocage dynamique . . . 91
5.2.3 Résultats . . . 94
5.2.4 Conclusion . . . 101

Conclusion générale et perspectives **103**
Perspectives . . . 103

A Construction de la matrice A_2 avec Kronecker **105**

B Probabilité d'occurrence analytique du dinucléotide AG avec 3PS **107**

C Probabilités d'occurrence analytiques des nucléotides avec 4P **109**

Bibliographie **111**

Liste de publications **117**

Table des figures **119**

Liste des tableaux **121**

Glossaire de biologie moléculaire **123**

Introduction

Contexte

En reprenant l'expérience de Griffith de 1928 sur la transformation de bactéries, Avery, McLeod et McCarthy ont montré en 1944 que l'ADN est le support de l'information génétique et de l'hérédité. Des travaux ultérieurs de Chargaff en 1950 sur la composition de l'ADN et de Hershey et Chase en 1952 sur les phages ont confirmé ce rôle important de l'ADN. Plus tard, l'apparition des techniques de séquençage de l'ADN dans les années 1970 et le développement d'outils informatiques de plus en plus puissants ont permis de séquencer le génome complet de différentes espèces. Par exemple, le séquençage du génome humain lancé par le projet « Human Genome » a été achevé en 2003. À ce jour, selon la base de données GOLD [1], les génomes complets de plus de 1900 espèces [2] ont été séquencés, ce nombre étant en augmentation accélérée avec le développement de nouvelles techniques. Les données obtenues, de plus en plus massives, sont un outil pour l'étude évolutive des espèces au travers de reconstructions d'arbres phylogénétiques ou de calcul de distances d'évolution entre séquences génétiques. Pour cela, une modélisation de l'évolution des séquences génétiques est nécessaire. Des modèles d'évolution de séquences génétiques à base de mutations ont ainsi été développés, le premier ayant été publié par Jukes et Cantor en 1969. Depuis, des modèles plus réalistes mais plus complexes ont été proposés à des fins de reconstruction d'arbres phylogénétiques ou de calcul de distances entre séquences.

Au cours de la première moitié des années 1990, Michel s'est intéressé au développement de modèles d'évolution permettant de déterminer les solutions analytiques des probabilités d'occurrence de motifs génétiques écrits sur un alphabet à deux lettres, puis à quatre lettres dans les années 2000. Ce type de modèles permet d'étudier l'histoire évolutive des probabilités d'occurrence de motifs génétiques dans le sens direct (passé-présent) ou inverse (présent-passé) à partir d'un vecteur de probabilités d'occurrence initiales de motifs et de paramètres de substitution. Bahi et Michel ont par la suite développé des modèles numériques non linéaires faisant intervenir des paramètres de substitution dépendants du temps.

Objectifs de la thèse

Notre thèse s'inscrit dans la continuité des travaux de Michel sur l'extension des modèles analytiques de substitution de nucléotides à des dinucléotides et des trinucléotides ainsi que sur l'étude des codes circulaires découverts par Arquès et Michel dans les gènes d'espèces eucaryotes et procaryotes en 1996, en particulier le code circulaire auto-complémentaire X(EUK_PRO)

1. Genomes OnLine Database v 3.0, `http://genomeonline.org`
2. 1919 espèces au 10 octobre 2011

composé des 20 codons apparaissant préférentiellement dans la phase de lecture des gènes.

Le premier objectif était d'unifier les modèles analytiques étendus aux dinucléotides et trinucléotides proposés par Michel et de développer une application web SEGM calculant les solutions analytiques des probabilités d'occurrence de motifs de taille 1 à 3 (nucléotides, dinucléotides et trinucléotides) et offrant la possiblité de visualiser sous forme de courbes l'évolution de ces probabilités au cours du temps. Cette évolution pouvant être réalisée dans le sens direct ou inverse.

Dans un deuxième temps, nous souhaitions étendre ces modèles de substitution à des motifs de taille quelconque. Toutefois, l'approche classique utilisée pour l'extension des modèles analytiques aux dinucléotides et trinucléotides ne pouvait pas être utilisée pour des motifs de taille supérieure, celle-ci devenant trop coûteuse en ressources (temps de calcul et mémoire utilisée). Cette extension nécessitait donc de définir une nouvelle approche basée sur l'utilisation de sommes et de produits de Kronecker de matrices. Elle a permis d'étendre SEGM à des motifs de taille 4 et 5 (tetranucléotides et pentanucléotides).

Enfin, nous voulions étudier l'histoire évolutive inverse des codes circulaires identifiés dans les gènes eucaryotes et procaryotes en 1996, en particulier celle du code circulaire X(EUK_PRO). Cependant, cette étude évolutive inverse nécessitait le développement d'un nouveau modèle d'évolution numérique pour des raisons que nous exposons dans la suite.

Plan

Dans un souci de simplicité de lecture, nous avons choisi de découper ce manuscrit en cinq chapitres pouvant être lus de façon indépendante. Ainsi, dans un premier chapitre, nous proposons quelques rappels de biologie moléculaire et de théorie des codes puis une brève description de quelques modèles d'évolution de séquences génétiques développés à ce jour. Le deuxième chapitre est consacré à la présentation des modèles mathématiques sous-jacents aux modèles d'évolution analytique et numérique développés au cours de cette thèse. Nous décrivons les différentes stratégies envisagées pour l'implantation du logiciel de recherche SEGM en évoquant les structures de données et algorithmes utilisés dans un troisième chapitre. Nous décrivons ensuite le logiciel SEGM développé durant cette thèse en présentant ses différentes fonctionnalités dans le quatrième chapitre. Un cinquième chapitre propose une application à chacun des deux modèles d'évolution développés durant cette thèse. Nous présentons ainsi une étude évolutive des dinucléotides des sites d'épissage des introns du génome humain avec le modèle analytique de dinucléotides et une étude évolutive inverse des codons du code circulaire identifié par Arquès et Michel en 1996 dans les gènes eucaryotes et procaryotes avec le modèle d'évolution numérique inverse avec blocage dynamique des substitutions. Enfin, nous concluons en abordant les principaux résultats obtenus durant cette thèse ainsi que différentes perspectives envisagées.

Chapitre 1

Éléments biologiques, de théorie des codes et modèles d'évolution

Nous proposons dans ce chapitre quelques rappels de biologie moléculaire et de théorie des codes circulaires nécessaires à la compréhension des différentes notions abordées dans ce manuscrit. Nous introduisons en fin de ce chapitre quelques modèles probabilistes d'évolution de gènes utilisés classiquement en phylogénétique, puis nous décrivons la classe de modèles d'évolution à laquelle nous nous sommes intéressés pour nos travaux de recherche.

1.1 Rappels de biologie moléculaire

1.1.1 Classification du vivant

Les espèces vivantes sont classées en trois domaines : les eucaryotes, les bactéries et les archées. Ces deux derniers domaines sont regroupés sous le nom de procaryotes. Les espèces eucaryotes sont des organismes unicellulaires ou pluricellulaires possédant un noyau cellulaire contrairement aux procaryotes, organismes unicellulaires sans noyau.

1.1.2 L'ADN

Chez toutes les espèces vivantes, l'information génétique est contenue dans les cellules et supportée par une longue chaîne moléculaire constituée de deux brins opposés (bicaténaire) formant une double hélice : l'acide désoxyrybonucléique, appelée ADN. Lorsqu'elle est repliée sur elle-même, une molécule d'ADN se présente sous forme de chromosome. Les deux brins de cette double hélice sont composés d'une suite de nucléotides, eux-mêmes constitués de trois éléments : un groupe phosphate, un sucre et une base azotée (figure 1.1). Il existe quatre variantes de bases azotées : l'Adénine, la Cytosine, la Guanine et la Thymine, respectivement représentées par les lettres A, C, G et T.

Sur chaque brin d'ADN, les nucléotides sont liés entre eux par des liaisons 5'-3' phosphodiester entre le groupe phosphate d'un nucléotide et le sucre d'un autre nucléotide (figure 1.1). Cette structure chimique induit une orientation des brins de l'ADN. La lecture de ces brins s'effectue dans le sens 5' vers 3'.

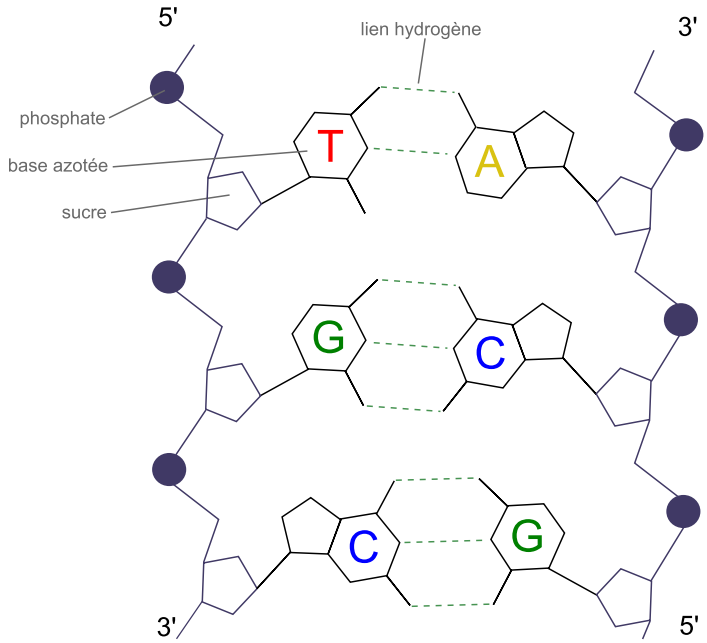


FIGURE 1.1 – Appariement des bases azotées de l'ADN.

Les propriétés chimiques des bases azotées permettent de les regrouper en deux familles : les purines (R), Adénine et Guanine et les pyrimidines (Y), Cytosine et Thymine. Le lien entre les deux brins d'ADN est assuré par la liaison entre les bases azotées des nucléotides. Cet appariement des bases respecte une règle de complémentarité : une purine est appariée à une pyrimidine, et inversement. Cette règle de complémentarité des bases opposées est due aux liaisons hydrogène reliant les bases. En effet, Adénine et Thymine sont liées par une double liaison hydrogène alors que Cytosine et Guanine sont liées par une triple liaison hydrogène, comme l'illustre la figure 1.1.

1.1.3 L'ARN

L'ARN (acide ribonucléique) est une molécule également présente au sein des cellules et possédant de grandes similitudes avec l'ADN. Toutefois, à la différence de l'ADN, l'ARN est généralement monocaténaire, c'est-à-dire composé d'un seul brin, constitué de nucléotides dont les bases azotées sont l'Adénine, la Cytosine, la Guanine et l'Uracile. Cette dernière base remplace la Thymine de l'ADN et possède des propriétés physico-chimiques similaires permettant ainsi à l'Uracile de s'apparier à l'Adénine avec deux liaisons hydrogène lors de l'étape de transcription de l'ADN en ARN.

La transcription est un des trois mécanismes, avec la réplication et la traduction, de la théorie fondamentale de la biologie moléculaire proposée en 1958 par Crick [Cri70]. Cette théorie explique les transferts d'information entre l'ADN, l'ARN et les protéines.

Au cours de la transcription, la séquence d'ARN messager ou ARNm obtenue est identique à celle d'un des brins de l'ADN, dit codant, au remplacement près de la Thymine par l'Uracile. L'ARNm est synthétisé par une enzyme, l'ARN polymérase, par complémentarité du brin transcrit de l'ADN, c'est-à-dire le brin opposé (ou complémentaire) au brin codant. L'ARNm sert de

vecteur de déplacement de l'information génétique vers un complexe moléculaire contenu dans le cytoplasme, le ribosome, responsable de l'étape de traduction de l'ARNm en protéines. Chez les eucaryotes, avant traduction, l'ARNm subit une étape supplémentaire dite d'épissage que nous présenterons dans une sous-section suivante.

D'autres types d'ARN, dits de structure (non codants), peuvent également être synthétisés. Par exemple, des ARN de transfert (ARNt) servent de vecteurs pour l'appariement entre les acides aminés et les codons lors de l'étape de traduction de l'ARNm en protéine au sein du ribosome, des ARN ribosomiques (ARNr) constituent la structure des ribosomes, des ARN nucléaires interviennent dans le processus d'épissage des exons, ou encore des micro ARN (miARN) interviennent dans la régulation de l'expression des gènes.

Des ARN de petite taille, les microARN (miARN), interviennent également dans la régulation de l'expression de gènes.

1.1.4 Protéines

Les protéines sont des macromolécules jouant un rôle essentiel au sein des cellules. Constituées d'une succession de molécules appelées acides aminés, elles sont fabriquées dans le cytoplasme des cellules lors de l'étape de traduction de l'ARNm par un ribosome. Les protéines peuvent avoir plusieurs fonctions, notamment structurale (par exemple pour l'architecture de la cellule), enzymatique (comme l'ARN polymérase), hormonale, ou encore fonction moteur (transport de molécules dans la cellule).

1.1.5 Traduction et code génétique

La synthèse d'une protéine par le ribosome s'effectue par assemblages successifs en chaîne d'acides aminés au fur et à mesure de la lecture d'un ARNm (figure 1.2).

1.1.5.1 Phases de lecture des gènes

La lecture de l'ARNm est réalisée par des triplets de nucléotides (ou trinucléotides) dans le sens 5' vers 3'. Cela implique l'existence de trois phases de lecture $p \in \{0, 1, 2\}$ dans les gènes codants. La phase $p = 0$ est la phase de référence (phase de lecture ouverte) définie par le codon d'initiation ATG marquant le début de la séquence d'ADN à transcrire. La phase $p = 1$ (respectivement $p = 2$) correspond à la phase de lecture décalée d'un (respectivement de deux) nucléotide(s) dans le sens 5'-3'. Les trinucléotides lus dans la phase de lecture $p = 0$ sont appelés codons. Ainsi, si le découpage en codons du début d'une séquence codante en phase de lecture ($p = 0$) donne ATG.TCC.GAT, il devient TGT.CCG.AT en phase de lecture décalée $p = 1$ et GTC.CGA.T en phase de lecture décalée $p = 2$. La phase de lecture est donc déterminante pour la synthèse des protéines. L'assemblage s'arrête lorsque le ribosome détecte un codon STOP (TAA, TAG ou TGA) marquant l'arrêt du processus de traduction.

1.1.5.2 Code génétique

Comme chaque nucléotide peut posséder une des quatre bases azotées, il existe 64 trinucléotides différents (AAA, AAC, ..., TTT). Cependant, les protéines ne sont constituées que de 20 acides aminés différents. Les 64 codons de l'ARNm codent donc pour 20 acides aminés plus les codons STOP marquant l'arrêt de la traduction. Le code génétique est donc **redondant**

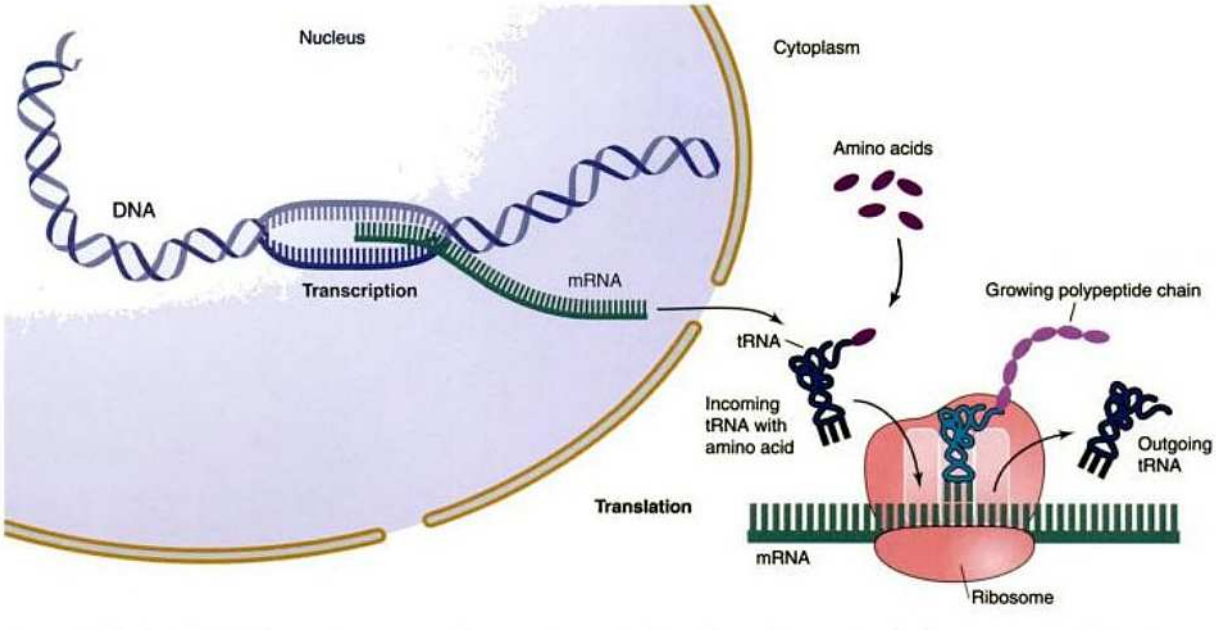


FIGURE 1.2 – Schéma représentant la synthèse des protéines dans le cytoplasme d'une cellule eucaryote. Figure extraite de [Tro07].

ou dégénéré : plusieurs codons codent pour un même acide aminé. Par souci de simplification, nous utiliserons dans la suite de ce manuscrit le même ensemble de bases azotées, soit Adénine, Cytosine, Guanine et Thymine, pour les séquences d'ADN et d'ARN. Le tableau 1.1 donne la correspondance entre les 64 codons et les acides aminés associés lors de l'étape de traduction.

1.1.5.3 Biais d'utilisation du code génétique

Le fait que le code génétique soit redondant n'implique pas que tous les codons codant pour un même acide aminé sont utilisés avec la même fréquence. Parmi eux, certains apparaissent plus fréquemment selon les gènes et les espèces considérés. Cette différence d'utilisation des codons pour le codage des acides aminés est appelé biais d'usage du code. Des études sur le biais d'usage du code dans les gènes de différentes espèces ont été réalisées dès 1980 [GGG+80], révélant notamment un lien entre le biais d'usage du code et l'expressivité des gènes [GGG+81, Ike85, SM94].

1.1.6 Gènes et séquences codantes

Un gène est une séquence d'ADN transcrite en ARN. Les séquences d'ADN transcrites en ARNm et codant pour des protéines sont appelées CDS (Coding Sequences).

Chez les eucaryotes, l'ADN est contenu dans le noyau de la cellule sous forme de plusieurs filaments (chromosomes). Les gènes d'eucaryotes sont morcelés, c'est-à-dire qu'ils sont entrecoupés de séquences codantes (exons) et non codantes (introns).

Dans les cellules procaryotes, l'ADN est présent dans le cytoplasme sous forme d'une unique chaîne circulaire. Cependant, nous distinguons les deux types de gènes procaryotes : les gènes d'archées sont morcelés comme les gènes d'eucaryotes en exons et introns, contrairement aux gènes des bactéries dits continus.

	site 2				
site 1	T	C	A	G	site 3
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA *	TGA *	A
	TTG Leu	TCG Ser	TAG *	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

TABLE 1.1 – Code génétique. ATG est le codon d'initiation standard, TAA, TAG et TGA sont les codons STOP marquant l'arrêt de la traduction. Le code génétique est dit redondant : plusieurs codons peuvent coder le même acide aminé. Cette dégénérescence (ou redondance) se fait essentiellement sur le troisième site des codons.

1.1.7 Épissage

La synthèse des protéines au sein des cellules d'eucaryotes et d'archées nécessite une étape supplémentaire, par rapport aux bactéries, avant la traduction de l'ARNm en protéines. En effet, l'ARNm résultant de la transcription d'un gène contient les mêmes séquences que le brin d'ADN codant (brin non transcrit). Il contient donc des exons et des introns. L'ARNm subit donc un épissage, étape consistant à éliminer les introns de l'ARNm et à fusionner les exons successifs. La molécule d'ARNm obtenue peut alors être utilisée pour la synthèse des protéines. De façon générale, l'ARNm avant épissage est appelé ARN pré-messager. Un phénomène d'épissage alternatif peut se produire : la même molécule d'ARN pré-messager peut engendrer des ARNm différents.

1.1.8 Mutations

Des modifications interviennent dans la molécule d'ADN, impliquant un changement de l'information génétique. On parle alors de mutations. Ces modifications peuvent intervenir à l'échelle d'un ou de quelques nucléotides ou bien à l'échelle des gènes et sont en particulier à l'origine de la diversification des espèces.

1.1.8.1 Mutations ponctuelles

Une mutation est dite ponctuelle lorsqu'elle se réalise à l'échelle des nucléotides. Elle peut intervenir sous trois formes différentes.

Substitution Le remplacement d'un nucléotide par un autre nucléotide dans la molécule d'ADN est une substitution. Ce type de mutation ponctuelle peut avoir des conséquences d'importances différentes selon la région de l'ADN touchée. En effet, elle est sans conséquence si elle intervient dans une région non codante ou non promotrice de l'ADN. En revanche, lorsqu'elle se produit dans une région codante (exon ou gène), elle modifie un codon et éventuellement l'acide aminé codé. Cependant, la redondance du code génétique fait que certaines substitutions de codons n'ont pas de répercussion sur l'acide aminé codé et donc sur la fonction de la protéine codée. Lorsque la substitution dans un codon ne modifie pas l'acide aminé codé elle est dite synonyme. En revanche, lorsqu'une substitution dans une région codante provoque un changement de l'acide aminé codé et donc une possible modification de la fonction de la protéine codée, elle est appelée faux-sens. Si elle a pour conséquence le remplacement d'un acide aminé par un codon stop, elle est dite non-sens.

Mutations ponctuelles décalantes Il arrive qu'un nucléotide (ou une suite de nucléotides, c'est-à-dire un oligonucléotide ou motif génétique) soit inséré dans une séquence d'ADN. Dans ce cas, on parle d'insertion. À l'opposé, on parle de suppression ou de délétion lorsqu'il s'agit de la suppression d'un nucléotide (ou un oligonucléotide) dans la séquence. Dans les régions codantes, ces deux types de mutations ponctuelles sont dites décalantes car elles ont pour conséquence un décalage de la phase de lecture. Par conséquent, elles peuvent avoir des effets plus graves sur la synthèse des protéines (séquence fausse, apparition prématurée d'un codon STOP).

1.1.8.2 Mutations à grande échelle

Des mutations faisant intervenir un grand nombre de nucléotides, voire des gènes peuvent également avoir lieu. Parmi ces mutations à grande échelle, en plus des insertions et des suppressions, nous pouvons distinguer les duplications consistant à copier puis insérer une région entière de la séquence d'ADN, les inversions, provoquant un retournement d'une partie de la séquence d'ADN, les transpositions, correspondant à un déplacement d'une partie de la séquence ainsi que les translocations, échanges de séquences entre chromosomes.

1.1.8.3 Causes des mutations

Au cours de la division cellulaire, le processus de réplication de l'ADN provoque parfois des erreurs dans la recopie des brins d'ADN. Toutefois, des mécanismes de réparation efficaces des erreurs de l'ADN faisant intervenir des enzymes spécifiques permettent de limiter le nombre de mutations conservées. Les mutations peuvent également être causées par des facteurs extérieurs comme par exemple des composés chimiques mutagènes, des radiations, des rayons ultraviolets ou encore des virus.

1.1.9 Contenu en G+C et effet CpG

Le contenu en G+C ou taux de GC est une mesure du contenu en nucléotides G et C d'une séquence génétique, c'est-à-dire la somme des proportions de nucléotides G et C de cette séquence. Ce contenu en G+C est variable selon les régions de l'ADN et selon les génomes.

L'effet CpG est une sous-représentativité du dinucléotide CG présente dans les génomes des vertébrés, le p indiquant la liaison phosphodiester entre les nucléotides. Ce déficit en CG est dû à un processus de méthylation-déamination de la cytosine sur les dinucléotides CG favorisant sa substitution en thymine et ayant pour conséquence une sur-représentativité du dinucléotide TG (CA sur le brin complémentaire). Dans certaines séquences du génome humain, cet effet CpG est moins significatif. Ces régions sont alors appelées ilôts CpG.

Le contenu en G+C et l'effet CpG sont pris en compte par certains modèles d'évolution que nous traitons dans la section 1.3.

1.2 Théorie des codes circulaires dans les gènes

Avant la découverte du code génétique, dans les années 1960, expliquant comment les chaînes d'acides aminés constituant les protéines sont synthétisées à partir de la lecture de suites de nucléotides, Crick et al. ont proposé dès 1957 l'existence d'un code dit « comma-free » ou « sans ponctuation » formé d'un ensemble de 20 trinucléotides codant les 20 acides aminés. Plus récemment en 1996, des travaux statistiques menés par Arquès et Michel sur les gènes ont révélé l'existence d'un autre type de codes dans les gènes, des codes circulaires, possédant des propriétés plus faibles que les codes sans ponctuation. Nous introduisons donc dans cette section les définitions et notations nécessaires à la compréhension des notions de théorie des codes sans ponctuation et des codes circulaires, les codes circulaires identifiés par Arquès et Michel ayant fait l'objet d'une étude évolutive inverse utilisant un modèle d'évolution numérique avec blocage dynamique des substitutions et dont nous présentons les résultats dans le chapitre 5.

1.2.1 Définitions et notations préliminaires

1.2.1.1 Mots, facteurs et ordre lexicographique

Soit $\mathcal{A}$ un alphabet fini. $\mathcal{A}^*$ est l'ensemble des mots de $\mathcal{A}$ contenant le mot vide noté ϵ, élément neutre de $\mathcal{A}^*$, et $\mathcal{A}^+$ est l'ensemble des mots de $\mathcal{A}$ privé du mot vide ϵ. Étant donné un sous-ensemble X de $\mathcal{A}^*$, X^n est l'ensemble des mots sur $\mathcal{A}$ qui sont les produits de n mots de X, c'est-à-dire $X^n = \{x_1x_2\dots x_n | x_i \in X\}$. La cardinalité d'un ensemble X fini est notée $|X|$ et $|u|$ représente la longueur d'un mot u.

Définition 1.1 *Un mot p est un préfixe d'un mot w s'il existe un mot u tel que $w = pu$.*

Définition 1.2 *Un mot s est un suffixe d'un mot w s'il existe un mot u tel que $w = us$.*

Définition 1.3 *Un mot f est un facteur d'un mot w s'il existe deux mots u et v tels que $w = ufv$.*

Définition 1.4 *Un facteur propre (respectivement préfixe propre, suffixe propre) f d'un mot w est un facteur (respectivement préfixe, suffixe) f de w tel que $|f| < |w|$.*

Dans la suite, nous considérons les nucléotides comme des lettres définissant l'alphabet génétique $\mathcal{A}_4 = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. L'alphabet $\mathcal{A}_4^n$ représente alors l'ensemble des 4^n mots (n-nucléotides) de longueur n. Par exemple, l'ensemble des 64 trinucléotides est noté $\mathcal{A}_4^3$. L'alphabet $\mathcal{A}_4$ est

muni de l'ordre total $A < C < G < T$. Par conséquent, $\mathcal{A}_4^+$ est ordonné lexicographiquement : étant donnés deux mots $u, v \in \mathcal{A}_4^+$, u est lexicographiquement plus petit que v, $u < v$, si et seulement si u est un préfixe propre de v ou si il existe deux lettres $x, y \in \mathcal{A}_4$, $x < y$, et trois mots $r, s, t \in \mathcal{A}_4^*$ tels que $u = rxs$ et $v = ryt$.

1.2.1.2 Codes

Définition 1.5 *Un sous-ensemble X de $\mathcal{A}^+$ est un* ***code*** *sur $\mathcal{A}$ si pour tout $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X, n, m \geq 1$, la condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implique $n = m$ et $x_i = x'_i$ pour $i = 1, \ldots, n$.*

Pour tout alphabet de cardinalité $k \geq 1$ et tout mot de longueur $n \geq 1$, $\mathcal{A}_k^n$ est un code. Par exemple, $\mathcal{A}_4^3$ est un code, et plus précisément un code uniforme [BP85]. Par conséquent, tout sous-ensemble de $\mathcal{A}_4^3$ est un code, que nous appelons ici code trinucléotide.

1.2.1.3 Codes sans ponctuation

Définition 1.6 *Un code trinucléotide $X \subset \mathcal{A}_4^3$ est dit* ***code sans ponctuation*** *si pour tout $y \in X$ et $u, v \in \mathcal{A}_4^*$ tels que $uyv = x_1 \cdots x_n$ avec $x_1, \ldots, x_n \in X$, $n \geq 1$, $u, v \in X^*$.*

Différentes catégories de codes sans ponctuation ont été décrites dans [MPP08].

1.2.1.4 Codes circulaires

Définition 1.7 *Un code trinucléotide $X \subset \mathcal{A}_4^3$ est* ***circulaire*** *si pour tout $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$, $n, m \geq 1$, $p \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, les conditions $sx_2 \cdots x_n p = y_1 \cdots y_m$ et $x_1 = ps$ impliquent $n = m$, $p = \epsilon$ et $x_i = y_i$ pour tout $i = 1, \ldots, n$ (figure 1.3).*

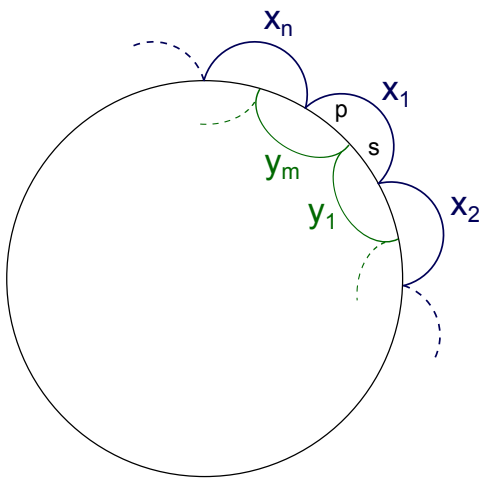


FIGURE 1.3 – Illustration de la définition 1.7 d'un code circulaire.

Selon la définition 1.7, un code circulaire est un ensemble de mots sur un alphabet donné tel que tout mot écrit sur un cercle a au plus une décomposition en mots de ce code circulaire [Las76].

Exemple 1.1 *Soit l'ensemble* $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ *et le mot* $w = ATGGCCCTA$. *Écrit sur un cercle, le mot* w *peut être décomposé de deux façons différentes en mots de* Y *:* ATG, GCC, CTA *et* AAT, GGC, CCT, *comme le montre la figure 1.4.* Y *n'est donc pas un code circulaire. En revanche, l'ensemble* $Y' = \{AAT, ATG, CCT, CTA, GCC, GTC\}$ *obtenu à partir de* Y *en remplaçant* GGC *par* GTC *est un code circulaire car tout mot construit à partir de mots de* Y', *en particulier* w, *n'a qu'une seule décomposition possible en mots de* Y'.

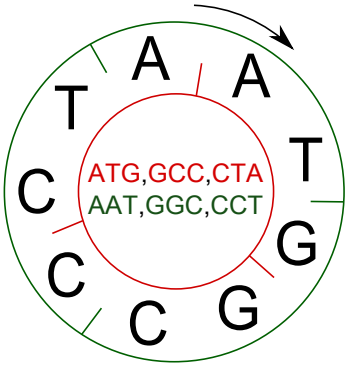


FIGURE 1.4 – Illustration de la double décomposition possible sur un cercle du mot w = ATGGCCCTA en mots de Y = {AAT,ATG,CCT,CTA,GCC,GGC}. L'ensemble Y n'est donc pas un code circulaire.

La phase de construction (phase de lecture) d'un mot w construit par concaténation de mots d'un code circulaire Y peut être retrouvée après lecture d'une suite de quelques nucléotides appelée fenêtre du code circulaire Y et quel que soit l'endroit dans le mot w. La longueur de la fenêtre permettant de retrouver la phase de lecture d'un mot w dépend du code circulaire et correspond à la longueur du plus long mot ambigu, c'est-à-dire pouvant être lu dans au moins deux phases, plus une lettre. Dans le cas des codes trinucléotides circulaires, cette fenêtre est de 13 nucléotides. Le code circulaire X(EUK_PRO) identifié par Arquès et Michel en 1996 dans les gènes eucaryotes et procaryotes possède également une fenêtre de longueur 13 [AM96].

Les codes circulaires ont ainsi des propriétés plus faibles que les codes sans ponctuation. Alors que les 20 trinucléotides d'un code sans ponctuation n'apparaissent que dans la phase de lecture ($p = 0$), les trinucléotides d'un code circulaire peuvent également apparaître dans les deux phases de lecture décalées ($p = 1$ et $p = 2$). Des mots du code circulaire peuvent donc apparaître dans une phase de lecture décalée jusqu'à une longueur de fenêtre de 12 nucléotides (longueur du plus long mot ambigu).

Définition 1.8 *Un code trinucléotide circulaire* $X \subset \mathcal{A}_4^3$ *est* ***maximal*** *si pour tout* $x \in \mathcal{A}_4^3$, $x \notin X$, $X \cup \{x\}$ *n'est pas un code trinucléotide circulaire.*

Définition 1.9 *Un code trinucléotide circulaire contenant exactement* k *éléments est appelé code* k***-trinucléotide circulaire***.

Remarque 1.1 *Un code circulaire 20-trinucléotide est :*

maximal *il ne peut pas être contenu dans un code circulaire trinucléotide de plus grande cardinalité ;*

maximum *aucun code circulaire trinucléotide ne peut contenir plus de 20 mots.*

Définition 1.10 *la* ***fonction de complémentarité*** $\mathcal{C} : \mathcal{A}_4^+ \to \mathcal{A}_4^+$ *est définie par* $\mathcal{C}(A) = T$, $\mathcal{C}(T) = A$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$ *et par* $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ *pour tout* $u, v \in \mathcal{A}_4^+$, *par exemple* $\mathcal{C}(AAC) = GTT$. *Cette fonction de complémentarité sur les mots est étendue aux ensembles de mots : le complémentaire d'un ensemble de trinucléotides est obtenu par application de la fonction de complémentarité* $\mathcal{C}$ *à tous les trinucléotides de l'ensemble.*

Définition 1.11 *la* ***fonction de permutation circulaire*** $\mathcal{P} : \mathcal{A}_4^3 \to \mathcal{A}_4^3$ *associe à chaque trinucléotide* $l_1l_2l_3$ *le permuté* $\mathcal{P}(l_1l_2l_3) = l_2l_3l_1$. *La* $k^{ème}$ *itération de* $\mathcal{P}$ *est notée* $\mathcal{P}^k$. *Cette fonction est également étendue aux ensembles de mots : le permuté d'un ensemble de trinucléotides est obtenu par application de la fonction de permutation* $\mathcal{P}$ *à tous les trinucléotides de l'ensemble.*

Définition 1.12 *Un code circulaire trinucléotide* X_0 *est* ***auto-complémentaire*** *si, pour tout* $x \in X_0$, $\mathcal{C}(x) \in X_0$.

Définition 1.13 *Si* X_0 *est un code circulaire trinucléotide, alors les ensembles de trinucléotides permutés* $\mathcal{P}(X_0)$ *et* $\mathcal{P}^2(X_0)$ *sont notés respectivement* X_1 *et* X_2.

Définition 1.14 *Un code circulaire trinucléotide* X_0 *est* ***auto-complémentaire*** $\mathcal{C}^3$ *si les ensembles* X_0, X_1 *et* X_2 *sont des codes circulaires satisfaisant les propriétés suivantes :*
- $X_0 = \mathcal{C}(X_0)$ *(auto-complémentarité de* X_0*) ;*
- $\mathcal{C}(X_1) = X_2$ *(et* $\mathcal{C}(X_2) = X_1$*).*

1.2.2 Hypothèse initiale des codes sans ponctuation dans les gènes

La découverte du code génétique dans les années 1960 a permis de répondre à deux questions principales sur les mécanismes en jeu dans la synthèse d'une protéine à partir d'une suite de trinucléotides : pourquoi existe-t-il 64 trinucléotides pour coder seulement 20 acides aminés et comment est choisie la phase de lecture ? Pour répondre à ces questions et en partant du principe d'une bijection de 20 trinucléotides sur les 20 acides aminés, Crick et al. ont proposé en 1957 [CGO57] l'existence d'un ensemble de 20 trinucléotides ayant les propriétés de code trinucléotide sans ponctuation. Avec un tel code, tous les trinucléotides codants se trouvent donc dans la même phase de lecture.

Cependant, la construction d'un tel code de 20 trinucléotides est soumise à deux règles principales. Premièrement, dans un code trinucléotide sans ponctuation, les trinucléotides constitués de trois nucléotides identiques sont exclus.

Exemple 1.2 *Si nous prenons par exemple une suite de trois trinucléotides AAA, il y a trois décompositions possibles :*

$$\ldots, AAA, AAA, AAA, \ldots \text{ ou } \ldots A, AAA, AAA, AA \ldots \text{ ou encore } \ldots AA, AAA, AAA, A \ldots,$$

les virgules indiquant les décompositions. Il est donc impossible dans ce cas de retrouver la phase de lecture.

Deuxièmement, un tel code ne peut contenir deux trinucléotides dont l'un est le permuté circulaire de l'autre.

Exemple 1.3 *Considérons le trinucléotide AAC et son permuté circulaire $\mathcal{P}(AAC) = ACA$. Alors la concaténation de trois trinucléotides AAC peut être décomposée de deux manières différentes, empêchant ainsi de retrouver la phase de lecture :*

$$\ldots, AAC, AAC, AAC, \ldots \, ou \ldots A, ACA, ACA, AC \ldots$$

Ainsi, en excluant les quatre trinucléotides AAA,CCC,GGG et TTT et en regroupant les 60 autres trinucléotides en 20 classes de trois trinucléotides constituées chacune des trois trinucléotides permutés circulaires, par exemple AAC, ACA et CAA, un code trinucléotide sans ponctuation ne contient qu'un seul trinucléotide par classe. Un tel code peut donc avoir au maximum 20 trinucléotides, soit autant que le nombre d'acides aminés.

Malgré quelques résultats obtenus par Golomb et al. en 1958 [GGW58, GWD58], cette notion de code sans ponctuation dans les gènes sur l'alphabet génétique $\mathcal{A}_4$ a été rapidement abandonnée après la découverte du codage de la phénylalanine (phe) par le trinucléotide TTT [NM61], un codon exclu par définition d'un tel code sans ponctuation. Une autre raison est aussi la mise en évidence de la présence d'un trinucléotide initiateur (ATG) ou codon start dans les gènes pour déterminer la phase de lecture.

1.2.3 Découverte des codes circulaires dans les gènes

En 1996, Arquès et Michel ont réalisé une étude statistique des occurrences de trinucléotides par phase de lecture dans des génomes d'espèces eucaryotes et procaryotes [AM96]. Ils ont ainsi identifié un ensemble de 20 trinucléotides

$$\begin{aligned} X_0(\text{EUK_PRO}) &= \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,} \\ &\quad \text{GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\} \end{aligned}$$

apparaissant préférentiellement en phase de lecture de référence ($p = 0$) dans les populations de gènes étudiées. Cet ensemble $X(\text{EUK_PRO}) = X_0(\text{EUK_PRO})$ est un code circulaire qui possède également des propriétés biomathématiques fortes.

1.2.3.1 Propriétés biomathématiques du code circulaire X(EUK_PRO)

Le code circulaire identifié par Arquès et Michel en 1996 dans les gènes d'espèces eucaryotes et procaryotes possède des propriétés biomathématiques fortes, parmi celles-ci :

Code maximal $X(\text{EUK_PRO})$ est un code circulaire maximal, c'est-à-dire qui n'est contenu dans aucun autre code circulaire de plus grande cardinalité.

Permutation La permutation circulaire (gauche) $\mathcal{P}$ (définition 1.11) génère $\mathcal{P}(X_0(\text{EUK_PRO})) = X_1(\text{EUK_PRO})$ et $\mathcal{P}(\mathcal{P}(X_0(\text{EUK_PRO}))) = X_2(\text{EUK_PRO})$. Ces trois codes circulaires sont

donc définis par

$$\begin{aligned}
X_0(\text{EUK_PRO}) &= \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,} \\
&\quad \text{GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\} \\
X_1(\text{EUK_PRO}) &= \{\text{AAG,ACA,ACG,ACT,AGC,AGG,ATA,ATG,CCA,CCG,} \\
&\quad \text{GCG,GTG,TAG,TCA,TCC,TCG,TCT,TGC,TTA,TTG}\} \\
X_2(\text{EUK_PRO}) &= \{\text{AGA,AGT,CAA,CAC,CAT,CCT,CGA,CGC,CGG,CGT,} \\
&\quad \text{CTA,CTT,GCA,GCT,GGA,TAA,TAT,TGA,TGG,TGT}\}
\end{aligned}$$

Complémentarité D'après la définition 1.12, X_0(EUK_PRO) est un code circulaire auto-complémentaire. En effet, 10 trinucléotides de X(EUK_PRO) sont complémentaires aux 10 autres. De plus, X_1(EUK_PRO) et X_2(EUK_PRO) sont complémentaires : les 20 trinucléotides de X_1(EUK_PRO) sont complémentaires des 20 trinucléotides de X_2(EUK_PRO).

Code $\mathcal{C}^3$ Bien que la permutation d'un code circulaire Y ne génère pas nécessairement un code circulaire, les codes X_1(EUK_PRO) et X_2(EUK_PRO) obtenus par permutation circulaire de X_0(EUK_PRO) sont des codes circulaires $\mathcal{C}^3$ (définition 1.14). X_0(EUK_PRO) étant le code circulaire identifié en phase de lecture de référence et auto-complémentaire, il est considéré comme le code circulaire $\mathcal{C}^3$ principal et est noté X(EUK_PRO).

Rareté La probabilité d'occurrence du code circulaire $\mathcal{C}^3$ X(EUK_PRO) est de $216/3^{20}$ soit environ 6.10^{-8}, c'est le rapport entre le nombre de codes complémentaires $\mathcal{C}^3$ calculés algorithmiquement (216) et le nombre de codes potentiels ($3^{20} = 3486784401$).

1.3 Modèles probabilistes d'évolution de séquences génétiques

Des modèles probabilistes d'évolution de séquences génétiques à base de substitution de nucléotides ont été développés dès la fin des années 1960 afin d'estimer des distances évolutives entre séquences ou de construire des arbres phylogénétiques. Comme nous l'avons vu dans la section 1.1.8, différents types de mutations interviennent dans l'évolution des séquences génétiques. Cependant, seules les mutations ponctuelles sont généralement prises en compte dans les modèles d'évolution probabilistes et uniquement les substitutions pour une grande majorité. Au cours de cette thèse, nous nous sommes focalisés sur ce dernier type de modèles d'évolution. Comme nous l'avons expliqué précédemment, notre approche des modèles d'évolution n'est pas phylogénétique mais consiste à étudier l'évolution des probabilités d'occurrence de motifs génétiques au cours du temps. Bien que la finalité soit différente, nous utilisons des matrices de substitution de nucléotides similaires aux modèles phylogénétiques. Ainsi, dans cette section, nous proposons dans un premier temps quelques propriétés générales sur les modèles d'évolutions, puis nous présentons les principaux modèles de substitution de nucléotides existants. Nous listons ensuite quelques exemples de modèles de substitution de motifs utilisés pour les approches phylogénétiques. Enfin, nous présentons les principaux modèles de substitution analytiques et numériques développés pour l'étude évolutive des probabilités d'occurrence de motifs génétiques

dont cette thèse poursuit les travaux. Les modèles biomathématiques sous-jacents de ces modèles sont détaillés dans le chapitre 2.

1.3.1 Propriétés générales des modèles probabilistes de substitution

1.3.1.1 Modèle de substitution de type markovien

De façon générale, le processus de substitution intervenant dans les modèles probabilistes de substitution de nucléotides et leurs extensions aux modèles de substitution de motifs est :

markovien : l'état futur d'un nucléotide soumis à substitution ne dépend que de l'état présent de ce nucléotide, il n'est pas nécessaire de connaître l'état passé du nucléotide ;

uniforme : tous les nucléotides de la séquence sont soumis au même processus de substitution ;

homogène : les taux de substitution des nucléotides sont constants durant tout le processus d'évolution ;

indépendant : les nucléotides de la séquence évoluent de façon indépendante ;

stationnaire : le processus de substitution conduit à une distribution constante des états du processus au cours du temps ;

réversible : la probabilité de substitution d'un nucléotide x en un nucléotide y pendant un intervalle de temps Δ est identique à la probabilité de substitution de y en x pendant Δ.

Les modèles de substitution listés dans la suite sont catégorisés suivant leurs propriétés.

1.3.1.2 Notations

Pour tous les modèles de substitution de nucléotides présentés dans cette section, nous notons A^s la matrice $(4,4)$ associée au modèle s et dont les éléments $0 \leq a^s_{i,j} \leq 1$ en ligne i et colonne j, $1 \leq i,j \leq 4$, sont les taux de substitution d'un nucléotide j en un nucléotide i, i et j représentant à la fois l'indice et le nucléotide dans l'ordre lexicographique. Les éléments diagonaux définis par $a^s_{i,i} = -\sum_{j=1,j\neq i}^{4} a_{j,i}$ sont représentés par « - » dans les matrices des taux de substitution instantanée A^s données dans la suite de ce chapitre.

$$A^{\mathrm{s}} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} - & a^s_{1,2} & a^s_{1,3} & a^s_{1,4} \\ a^s_{2,1} & - & a^s_{2,3} & a^s_{2,4} \\ a^s_{3,1} & a^s_{3,2} & - & a^s_{3,4} \\ a^s_{4,1} & a^s_{4,2} & a^s_{4,3} & - \end{pmatrix} \end{array}.$$

1.3.1.3 Distribution stationnaire

Les processus de substitution traités ici sont de type markovien dont la chaîne est irréductible, c'est-à-dire que tout état i est accessible à partir de tout état j, apériodique et possède un nombre fini d'états. Ces modèles de substitution possèdent donc une distribution stationnaire unique $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ vérifiant l'équation $\pi A^s = 0$. Cette distribution stationnaire est un vecteur des fréquences en nucléotides obtenu à l'équilibre du processus d'évolution à base de substitution.

1.3.2 Modèles probabilistes de substitution de nucléotides

De nombreux modèles probabilistes de substitution de nucléotides ont été développés à ce jour. Nous détaillons dans un premier temps les modèles de substitution de nucléotides ayant fait l'objet d'une extension aux dinucléotides et trinucléotides, puis à des motifs de taille quelconque au cours de cette thèse. Nous proposons ensuite un aperçu des principaux modèles généralement utilisés en les regroupant suivant leur distribution stationnaire.

1.3.2.1 Modèles utilisés au cours de cette thèse

Modèle de Jukes et Cantor à un paramètre : JC Le premier modèle d'évolution à base de substitution de nucléotides a été proposé par Jukes et Cantor en 1969 dans [JC69]. Ce modèle possède un paramètre de substitution α identique pour toutes les substitutions entre nucléotides différents. La matrice des taux de substitution associée est donc définie par

$$A^{\mathrm{JC}} = \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}.$$

Modèle de Kimura à deux paramètres : K2P Le modèle JC a été généralisé en 1980 par Kimura avec un modèle faisant intervenir deux paramètres de substitution [Kim80]. L'apparition de ce second paramètre est motivée par la distinction rendue possible entre les transitions α (substitution entre deux purines ou deux pyrimidines : A↔G = C↔T) et les transversions β (substitution entre une purine et une pyrimidine et inversement : A↔T = A↔C = C↔G). Ce modèle permet ainsi de prendre en compte le biais entre le taux de transitions et le taux de transversions présent de façon générale au sein des séquences ADN, comme l'ont montré dès 1977 Vogel et Kopun à partir de l'analyse de protéines apparentées, de variants d'hémoglobine chez l'Homme et de séquences ARN d'espèces apparentées [VK77]. D'autres études ont montré par la suite que ce biais de transitions était plus important dans les séquences d'ADN mitochondrial que dans l'ADN nucléaire ou les chloroplastes [BPWW82, GLG82, CC84, Wak96, YY99]. La matrice des taux de substitution associée au modèle K2P est

$$A^{\mathrm{K2P}} = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \alpha \\ \beta & \alpha & \beta & - \end{pmatrix}.$$

Modèle de Kimura à trois paramètres : K3P Kimura a généralisé le modèle K2P en 1981 dans [Kim81] en incluant un troisième paramètre de substitution autorisant une distinction supplémentaire entre les deux types de transversion possibles. Dans ce modèle défini par la matrice des taux de substitution

$$A^{\mathrm{K3P}} = \begin{pmatrix} - & \gamma & \alpha & \beta \\ \gamma & - & \beta & \alpha \\ \alpha & \beta & - & \gamma \\ \beta & \alpha & \gamma & - \end{pmatrix},$$

les trois paramètres de substitution α, β et γ correspondent respectivement aux transitions A↔G et C↔T, transversions de type I A↔T et C↔G et transversions de type II A↔C et G↔T. Ce modèle peut être ramené au modèle K2P en prenant $\gamma = \beta$ et au modèle JC dans le cas $\alpha = \beta = \gamma$.

Ces trois modèles de substitution de nucléotides ayant par la suite fait l'objet d'une extension aux dinucléotides et trinucléotides [Mic07c, FM06, Mic07a], puis étendus à des motifs de taille quelconque durant cette thèse, nous les renommons dans la suite de ce document pour plus de simplicité. Ainsi, les modèles de substitution de nucléotides JC, K2P et K3P seront respectivement renommés 1P, 2P et 3P et leurs extensions aux motifs 1PS, 2PS et 3PS, explicitant le nombre de paramètres par site de motif.

Ces premiers modèles que nous venons de présenter reposent sur une matrice de substitution symétrique et ont une distribution stationnaire uniforme : $\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$.

D'autres modèles de substitution ont été développés avec des matrices de substitution non nécessairement symétriques. Parmi ceux-ci, certains prennent en compte les fréquences stationnaires en nucléotides dans le processus de substitution.

1.3.2.2 Autres modèles sans fréquence stationnaire

Modèle général SYM Le modèle de substitution de nucléotides le plus général reposant sur une matrice de substitution symétrique et avec une distribution stationnaire uniforme est le modèle SYM détaillé dans [Zha94]. La matrice de substitution de ce modèle possède six paramètres :

$$A^{\mathrm{SYM}} = \begin{pmatrix} - & \gamma_1 & \alpha_1 & \beta_1 \\ \gamma_1 & - & \beta_2 & \alpha_2 \\ \alpha_1 & \beta_2 & - & \gamma_2 \\ \beta_1 & \alpha_2 & \gamma_2 & - \end{pmatrix}.$$

Modèles avec matrices de substitution non symétriques En 1992, Tamura a développé un modèle de substitution de nucléotides tenant compte du biais en contenu G+C et du biais de transitions [Tam92]. Ce modèle repose sur une matrice de substitution à trois paramètres non symétrique étendant celle du modèle de Kimura 2P vu précédemment. Contrairement aux modèles à matrice de substitution symétrique, les taux de substitution d'un nucléotide i en j sont différents des taux de substitution d'un nucléotide j en i :

$$A^{\mathrm{T92}} = \begin{pmatrix} - & (1-\theta)\beta & (1-\theta)\alpha & (1-\theta)\beta \\ \theta\beta & - & \theta\beta & \theta\alpha \\ \theta\alpha & \theta\beta & - & \theta\beta \\ (1-\theta)\beta & (1-\theta)\alpha & (1-\theta)\beta & - \end{pmatrix}.$$

D'autres modèles à base de matrices de substitution non symétriques avec un nombre croissant de paramètres de substitution ont également été développés, comme ceux de Takahata et Kimura en 1981 [TK81] et Blaisdell en 1985 [Bla85] intégrant quatre paramètres, ou encore celui de Gojobori et al. en 1982 à six paramètres de substitution [GLG82].

1.3.2.3 Modèles avec prise en compte des fréquences stationnaires

Modèle F81 En 1981, Felsenstein a proposé une extension du modèle 1P à un paramètre de substitution de Jukes et Cantor en ajoutant les fréquences stationnaires des nucléotides dans la matrice de substitution. Le modèle F81 obtenu est ainsi défini par la matrice

$$A^{\mathrm{F81}} = \begin{pmatrix} - & \alpha\pi_A & \alpha\pi_A & \alpha\pi_A \\ \alpha\pi_C & - & \alpha\pi_C & \alpha\pi_C \\ \alpha\pi_G & \alpha\pi_G & - & \alpha\pi_G \\ \alpha\pi_T & \alpha\pi_T & \alpha\pi_T & - \end{pmatrix}.$$

Modèle HKY De la même façon que le modèle F81 généralise 1P, en 1985 Hasegawa et al. [HKY85] ont publié un modèle intégrant les fréquences stationnaires et effectuant une distinction entre les taux de transition et de transversion. Ce modèle HKY représenté par la matrice de substitution

$$A^{\mathrm{HKY}} = \begin{pmatrix} - & \pi_A\beta & \pi_A\alpha & \pi_A\beta \\ \pi_C\beta & - & \pi_C\beta & \pi_C\alpha \\ \pi_G\alpha & \pi_G\beta & - & \pi_G\beta \\ \pi_T\beta & \pi_T\alpha & \pi_T\beta & - \end{pmatrix}$$

est donc une extension du modèle 2P de Kimura avec prise en compte des fréquences stationnaires.

Modèle TN93 Alors que le modèle de Kimura 3P distingue les deux types de transversion, celui de Tamura et Nei publié en 1993 [TN93] effectue une distinction entre les deux types de transitions. Dans ce modèle, les transitions entre purines (A↔G) ont un taux de substitution différent des transitions entre pyrimidines (C↔T). La matrice de substitution de ce modèle prenant également en compte les fréquences stationnaires est définie par

$$A^{\mathrm{TN93}} = \begin{pmatrix} - & \pi_A & \beta\pi_A & \pi_A \\ \pi_C & - & \pi_C & \alpha\pi_C \\ \beta\pi_G & \pi_G & - & \pi_G \\ \pi_T & \alpha\pi_T & \pi_T & - \end{pmatrix}.$$

Modèles généraux Le modèle réversible le plus général avec fréquences stationnaires est le modèle GTR publié par Tavaré en 1986 [Tav86]. Il repose sur le modèle SYM :

$$A^{\mathrm{GTR}} = \begin{pmatrix} - & \alpha\pi_A & \beta\pi_A & \gamma\pi_A \\ \alpha\pi_C & - & \delta\pi_C & \epsilon\pi_C \\ \beta\pi_G & \delta\pi_G & - & \theta\pi_G \\ \gamma\pi_T & \epsilon\pi_T & \theta\pi_T & - \end{pmatrix}.$$

Enfin, le modèle de substitution de nucléotides le plus général autorise des taux de substitution tous différents. La matrice de substitution de ce modèle appelé « unrestricted » ou G4H

[ROMM90] est donc définie par

$$A^{\mathrm{G4H}} = \begin{pmatrix} - & a & b & c \\ g & - & d & e \\ h & i & - & f \\ j & k & l & - \end{pmatrix}.$$

Lìo et Goldman proposent dans [LG98] une revue des principaux modèles de substitution de nucléotides.

1.3.2.4 Modèles d'évolution non uniformes et non homogènes

Les processus de substitution de nucléotides des modèles présentés jusqu'ici sont uniformes (le même processus de substitution est appliqué sur tous les sites de la séquence) et homogènes (les taux de substitution sont constants durant tout le processus d'évolution). Ces propriétés approchent de façon satisfaisante la réalité biologique. Cependant, les différents types de séquences composant l'ADN n'évoluent pas de façon uniforme. Par exemple, au sein des gènes, les exons et les introns ne sont pas soumis au même processus d'évolution. Il en est de même pour les codons. Nous avons vu que le code génétique est redondant, plusieurs codons peuvent coder pour un même acide aminé. D'après le tableau 1.1, une substitution n'a pas la même conséquence selon le site sur lequel elle intervient. Ainsi, une substitution sur le troisième site d'un codon a plus de chances d'être synonyme (elle ne modifie pas l'acide aminé codé) que sur le premier site.

Pour prendre en compte cette non uniformité du processus de substitution le long de la séquence, Fitch et Markowitz en 1970 [FM70] émettent l'idée d'un modèle d'évolution dont le processus de substitution intervient de façon binaire sur chaque site de la séquence. Dans ce modèle appelé **Covarion** (Concomitantly Variable Codons), chaque site peut être ou non soumis au processus de substitution : dans le premier cas le site peut évoluer, il est dit *On*, dans le second cas, le site n'évolue pas, il est dit *Off*. Tuffley et Steel ont repris l'idée de covarion en 1998 dans un modèle appelé **Rates accross sites** [TS98] dans lequel les sites sont soumis au processus de substitution avec deux vitesses possibles : chaque site peut évoluer durant tout le processus d'évolution de façon lente ou rapide. De son côté, Yang définit dans [Yan93] un modèle plus souple dans lequel chaque site peut évoluer avec une vitesse dépendant d'une loi de distribution Gamma.

Dans [TS98], Tuffley et Steel présentent également une version du modèle covarion non homogène selon le temps d'évolution. Alors que le processus de substitution est constant pour chaque site durant tout le temps d'évolution dans le modèle de Fitch et Markowitz, celui de Tuffley et Steel autorise les sites à alterner entre les états *On* et *Off* tout au long du processus d'évolution. Galtier généralise en 2001 dans [Gal01] le modèle non uniforme de Yang [Yan93] en proposant le modèle **SSRV** (Site Specific Rate Variation) non uniforme et non homogène qui autorise les sites à évoluer avec des vitesses différentes au cours du temps d'évolution.

1.3.3 Modèles d'évolution de motifs avec dépendance au contexte

Des modèles d'évolution ont été également développés pour prendre en compte une dépendance entre sites voisins vis à vis du processus de substitution, en particulier pour les dinucléotides (étude de l'effet CpG) et trinucléotides (étude des codons). Cette dépendance aux sites

voisins entraîne l'existence d'un cône de dépendance : l'état d'un nucléotide au temps $t + T$ dépend de l'état de ses voisins au temps t, eux-mêmes dépendants de l'état de leurs voisins au temps précédent. Nous listons ici quelques uns de ces modèles généralement irréversibles, sans les décrire en détail, ceux-ci n'entrant pas dans la classe des modèles étudiés durant cette thèse.

Modèles d'évolution de dinucléotides Duret et Galtier ont utilisé en 2000 [DG00] un modèle d'évolution de dinucléotides prenant en compte un chevauchement entre dinucléotides successifs pour l'étude de l'effet CpG dans le génome humain. Ce modèle suppose que l'état d'un nucléotide ne dépend que de ses voisins immédiats, entraînant une approximation des fréquences en trinucléotides à partir des fréquences en nucléotides et dinucléotides.

Toujours dans le cadre de l'étude de l'effet CpG, Arndt et al. ont également utilisé en 2003 cette méthode d'approximation des fréquences, appelée approximation de k-cluster, dans [ABH03], pour déterminer les fréquences stationnaires analytiques de nucléotides et dinucléotides générées par un modèle d'évolution reposant sur deux processus de substitution. Un premier processus fait évoluer les nucléotides de façon indépendante et un second processus fait évoluer les paires de nucléotides voisins. Ils proposent également dans [ABH03] une méthode d'estimation des taux de substitution en utilisant une méthode de calcul de maximum de vraisemblance. L'année suivante, Lunter et Hein [LH04] ont introduit un autre modèle de substitution avec dépendance au contexte utilisant l'approximation du k-cluster. Ils décrivent également un algorithme de calcul de vraisemblance de séquences soumises à ce modèle d'évolution leur servant également à estimer les taux de substitution à l'aide d'une méthode d'échantillonnage bayésienne. Plus récemment, Bérard et al. [BGP08] considèrent un modèle de substitution prenant en compte la classe (purine ou pyrimidine) des nucléotides voisins pour permettre de déterminer les fréquences stationnaires analytiques sans utiliser de méthode d'approximation des fréquences.

Modèles d'évolution de codons En 1994, Muse et Gaut [MG94] puis Goldman et Yang [GY94] proposent les premiers modèles d'évolution de codons à base de substitution de nucléotides. Ces modèles reposent sur des matrices de substitution de taille (61,61), les codons stop étant exclus du processus de substitution. Dans le modèle de Muse et Gaut, les taux de substitution entre un codon i et un codon j résultent d'un produit entre la fréquence stationnaire du site du codon où intervient la substitution et un paramètre dépendant de la nature synonyme ou non-synonyme de la substitution entre i et j. Le modèle Goldman et Yang diffère par les paramètres de substitution considérés ainsi que les fréquences stationnaires. En effet, dans ce second modèle, les paramètres de substitution prennent en compte la fréquence stationnaire du codon j, le biais de transitions/transversions, la distance de Grantham [Gra74] séparant les acides aminés codés par les codons i et j, ainsi que la variabilité de la séquence. D'autres modèles de codons ont été développés par la suite comme celui de Pedersen et al. [PWC98] reprenant les paramètres de substitution des modèles de Muse et Gaut et de Goldman en plus d'un paramètre lié à l'effet CpG au sein des codons.

Jensen et Pedersen ont ensuite étendu ces modèles à des pentanucléotides pour prendre en compte la dépendance aux voisins des nucléotides aux extrémités des codons [JP00, PJ01].

1.3.4 Modèles d'évolution de probabilités d'occurrence de motifs

Au cours de cette thèse, nous nous sommes intéressés à une autre classe de modèles d'évolution de motifs. Alors que les modèles que nous avons exposés jusqu'ici avaient pour finalités la reconstruction d'arbres phylogénétiques, la dérivation de distances entre séquences ou encore le calcul de la distribution stationnaire à l'équilibre du processus, les modèles dont cette thèse fait l'objet permettent d'étudier l'évolution des probabilités d'occurrence des motifs au cours du temps d'évolution. Nous distinguons dans la suite les modèles déterminant les solutions analytiques des probabilités d'occurrence des motifs et les modèles numériques non homogènes, c'est-à-dire reposant sur une matrice de substitution non constante au cours du temps d'évolution. Ces deux types de modèles ont pour particularité d'autoriser l'étude de l'évolution des probabilités d'occurrence de motifs dans le sens d'évolution direct (passé-présent) ou inverse (passé-présent), comme nous le montrons, pour le cas des modèles analytiques, dans la section 2.2.2 du chapitre suivant. Une autre spécificité de ces modèles est qu'ils se basent sur des matrices de substitution de motifs dont les paramètres peuvent être différents pour chaque site, ce qui n'était pas le cas pour les modèles cités précédemment.

1.3.4.1 Modèles analytiques

En 1993, Arquès et Michel ont proposé un premier modèle d'évolution déterminant les expressions analytiques des probabilités d'occurrence de codons soumis à un processus de substitution aléatoire, dans le sens d'évolution direct ou inverse, sur l'alphabet à deux lettres $\mathcal{A} = \{R, Y\}$ [AM93]. Ils ont ensuite généralisé ce modèle aux dinucléotides sur l'alphabet génétique $\mathcal{A} = \{A, C, G, T\}$ en 1995 [AM95]. Suite à l'identification du code circulaire X(EUK_PRO) dans les gènes eucaryotes et procaryotes en 1996 (voir section 1.2.3), Arquès, Fallot et Michel ont étendu dans [AFM98] le modèle de substitution de nucléotides de Jukes et Cantor à un paramètre 1P aux trinucléotides pour une étude évolutive de X(EUK_PRO). Ce modèle 1PS a par la suite été généralisé avec deux puis trois paramètres de substitution par site, engendrant les modèles 2PS et 3PS de dinucléotides et de trinucléotides, respectivement dans [Mic07c] et [FM06, Mic07a].

Bien que ces modèles n'aient pas de finalité phylogénétique, les solutions analytiques des probabilités d'occurrence des motifs obtenus avec ceux-ci peuvent être utilisées pour dériver des distances entre séquences, comme le montre par exemple Michel pour les dinucléotides dans [Mic07c].

Parallèlement à ces modèles analytiques reposant sur des matrices de substitution à paramètres constants, des modèles non homogènes ont également été développés.

1.3.4.2 Modèles numériques non homogènes

Bahi et Michel ont introduit en 2004 dans [BM04] un modèle d'évolution de trinucléotides dont la matrice de substitution possède des paramètres dépendants du temps, autorisant ainsi des accélérations ou décélérations du processus de substitution au cours du temps. Sur le principe du modèle de covarion de Tuffley et Steel [TS98] faisant évoluer certains codons suivant leur état *On* ou *Off*, Bahi et Michel proposent dans [BM08] un modèle chaotique de substitution de

trinucléotides basé sur le choix aléatoire d'un sous-ensemble des 64 trinucléotides pouvant être soumis au processus de substitution à chaque T. L'année suivante, ils présentent dans [BM09a] un modèle combinant la propriété de non-homogénéité des paramètres de substitution de [BM04] et l'aspect chaotique de [BM08].

Nous détaillons dans le chapitre 2 la résolution des probabilités d'occurrence des motifs avec les modèles analytiques présentés précédemment ainsi que leur extension à des motifs de taille quelconque et un modèle numérique non homogène développé au cours de cette thèse.

Chapitre 2

Modèles biomathématiques d'évolution

Notre étude de l'extension des modèles analytiques de substitution de nucléotides à des motifs génétiques de taille quelconque réalisée au cours de cette thèse s'inscrit dans la continuité des travaux de Michel sur l'extension de ces modèles aux dinucléotides et trinucléotides exposés dans la section 1.3.4 du chapitre précédent. Ces modèles d'évolution permettent de déterminer et d'étudier l'évolution des probabilités d'occurrence de motifs génétiques au temps t à partir de paramètres de substitution et d'un vecteur de probabilités d'occurrence initiales des motifs. Une propriété importante de ces modèles les rend inversibles : l'étude évolutive peut se faire dans le sens direct (passé-présent) ou inverse (présent-passé). Ils permettent donc d'étudier le passé des séquences génétiques actuelles.

La détermination des probabilités d'occurrence de motifs repose sur la résolution d'une équation différentielle commune aux différents modèles de matrices de substitution utilisés. Après établissement de cette équation, nous nous intéressons à sa résolution par des approches analytiques et numériques. Pour cela, nous proposons dans un premier temps une méthode de résolution classique spécifique à un type de matrice de substitution correspondant au modèle de substitution 3PS de trinucléotides présenté dans la section 1.3.4 du chapitre précédent. Nous présentons ensuite une méthode de résolution plus générale utilisée pour l'extension à des motifs de taille quelconque, valable pour toute une classe de matrices de substitution dont 3PS est un cas particulier. Nous donnons un exemple avec une matrice de substitution à quatre paramètres de substitution. Nous nous intéressons enfin à une méthode de résolution numérique classique qui nous a servi de base au développement d'un modèle d'évolution numérique inverse non homogène de trinucléotides basé sur une matrice de substitution non constante, c'est-à-dire avec des taux de substitution variant au cours du temps.

2.1 Équations différentielles des modèles stochastiques d'évolution par substitution

Le processus de substitution de tous les modèles d'évolution étudiés ici est défini par une équation différentielle déterminant les probabilités d'occurrence au temps $t \geq 0$ des 4^n motifs génétiques de taille n mutant selon des probabilités de substitution constantes. Par exemple, si

$n = 1$, l'équation différentielle portera sur les nucléotides, si $n = 2$, sur les dinucléotides, etc...

2.1.1 Définitions préalables et notations

Soient i et j deux motifs génétiques de taille n, $1 \leq i, j \leq 4^n$. Notons $P_T(j \to i)$ la probabilité de substitution du motif j en le motif i pendant l'intervalle de temps T, $T \geq 0$. La probabilité d'occurrence $P_i(t+T)$ du motif i au temps $t+T$ est égale à la somme des probabilités $P_j(t)$ des 4^n motifs j au temps t multipliées par leurs probabilités de substitution $P_T(j \to i)$ en les motifs i durant T, c'est-à-dire :

$$P_i(t+T) = \underbrace{\sum_j P_j(t) P_T(j \to i)}_{\text{Probabilité d'apparition du motif } i}. \tag{2.1}$$

Ainsi, $P_i(t+T) - P_i(t)$ représente la probabilité que le motif i apparaisse et disparaisse au cours de l'intervalle de temps T.

2.1.2 Dérivation

À partir de l'équation (2.1), la dérivée $P_i'(t) = \frac{\partial P_i(t)}{\partial t}$ de la probabilité d'occurrence du motif i au temps t est

$$\begin{aligned}
P_i'(t) &= \lim_{T \to 0} \left[\frac{P_i(t+T) - P_i(t)}{T} \right] \\
&= \lim_{T \to 0} \left[\frac{\sum_j P_j(t) P_T(j \to i) - P_i(t)}{T} \right] \\
&= \lim_{T \to 0} \left[\frac{\sum_{j \neq i} P_j(t) P_T(j \to i) + P_i(t) P_T(i \to i) - P_i(t)}{T} \right]
\end{aligned}$$

où $P_T(i \to i)$ représente la probabilité que le motif i ne mute pas en un autre motif $j \neq i$ au cours de l'intervalle de temps T. Nous obtenons donc

$$\begin{aligned}
P_i'(t) &= \lim_{T \to 0} \left[\frac{\sum_{j \neq i} P_j(t) P_T(j \to i) - P_i(t)(1 - P_T(i \to i))}{T} \right] \\
&= \lim_{T \to 0} \left[\frac{\sum_{j \neq i} P_j(t) P_T(j \to i) - P_i(t) \sum_{j \neq i} P_T(i \to j)}{T} \right] \\
&= \sum_{j \neq i} P_j(t) \lim_{T \to 0} \left[\frac{P_T(j \to i)}{T} \right] - P_i(t) \sum_{j \neq i} \lim_{T \to 0} \left[\frac{P_T(i \to j)}{T} \right].
\end{aligned}$$

Pour tous motifs i, j, la probabilité de substitution instantanée $P(j \to i)$ du motif j en le motif i est considérée constante au cours du temps. Nous considérons également que T est assez petit pour qu'il n'y ait pas plus d'une substitution par site de motif pendant cette durée. Sous cette hypothèse, nous pouvons écrire

$$P_T(j \to i) = \lim_{T \to 0} P(j \to i) T$$

et par conséquent

$$\lim_{T \to 0} \left(\frac{P_T(j \to i)}{T} \right) = P(j \to i).$$

Finalement, pour tout motif i, la dérivée $P_i'(t)$ vaut

$$\begin{aligned} P_i'(t) &= \sum_{j\neq i} P_j(t)P(j\to i) - P_i(t)\sum_{j\neq i} P(i\to j) \\ &= \sum_{j\neq i} P_j(t)P(j\to i) - P_i(t)(1-P(i\to i)) \\ &= \sum_{j} P_j(t)P(j\to i) - P_i(t). \end{aligned} \tag{2.2}$$

2.1.3 Équation différentielle matricielle

Soit $P_n(t) = [P_i(t)]_{1\leq i\leq 4^n}$ le vecteur colonne composé des probabilités d'occurrence $P_i(t)$ des 4^n motifs de taille n, $1 \leq i \leq 4^n$. L'équation (2.2) nous permet de dériver une équation différentielle matricielle décrivant le processus de substitution pour les motifs génétiques

$$P_n'(t) = M_n \cdot P_n(t) - P_n(t) = (M_n - I_n)\cdot P_n(t) \tag{2.3}$$

où $\cdot$ est le produit matriciel, I_n la matrice identité $(4^n, 4^n)$ et $M_n = [m_{i,j}]_{1\leq i,j\leq 4^n}$ la matrice des probabilités de substitution instantanées dont les éléments $m_{i,j}$ en ligne i et colonne j sont les probabilités de substitution du motif j en le motif i

$$m_{i,j} = P(j\to i).$$

La matrice des probabilités de substitution instantanées M_n est stochastique en colonne. En effet, quel que soit $1 \leq j \leq 4^n$, les éléments de la matrice M_n satisfont l'équation $\sum_{i=1}^{4^n} m_{i,j} = \sum_{i=1}^{4^n} P(j\to i) = 1$. Pour tout $1 \leq j \leq 4^n$, les éléments diagonaux $m_{j,j}$ de M_n satisfont l'équation

$$m_{j,j} = 1 - \sum_{i=1, i\neq j}^{4^n} m_{i,j}.$$

Soit A_n la matrice des taux de substitution instantanée, $A_n = M_n - I_n$. Les colonnes de A_n somment donc à 0. L'équation différentielle matricielle (2.3) devient alors

$$P_n'(t) = A_n \cdot P_n(t). \tag{2.4}$$

Dans la suite de ce chapitre, et pour des raisons de simplicité, les matrices de taux de substitution instantanées A_n seront appelées matrices de substitution.

2.2 Méthodes de résolution analytique pour des matrices de substitution de grande taille

L'équation différentielle (2.4) peut être résolue numériquement quelle que soit la matrice de substitution A_n associée mais aussi analytiquement lorsque cette matrice est diagonalisable et possède certaines propriétés. Les modèles analytiques de Michel pour les dinucléotides [Mic07c] et les trinucléotides [FM06, Mic07b] sont basés sur les modèles de substitution de nucléotides de Kimura à deux et trois paramètres de substitution 2P et 3P [Kim80, Kim81]. La méthode

de diagonalisation classique des matrices de substitution de type 3PS pour les dinucléotides et trinucléotides repose sur leurs structures en blocs. Nous présentons dans un premier temps cette méthode classique. Par ailleurs, une étude approfondie de la construction des matrices de substitution pour des motifs de taille supérieure à 1 nous a permis de développer une méthode de résolution analytique pour toute une classe de matrices de substitution, dont nous verrons un exemple généralisant le modèle 3PS à un modèle à quatre paramètres de substitution par site 4PS. Cette méthode de résolution permet également de générer les solutions analytiques des probabilités d'occurrence pour des motifs de taille quelconque.

2.2.1 Solution de l'équation différentielle

La résolution de l'équation (2.4) est possible par diagonalisation de la matrice A_n. Si A_n est diagonalisable, c'est-à-dire si elle peut-être factorisée tel que $A_n = Q_n \cdot D_n \cdot Q_n^{-1}$, où $D_n(4^n, 4^n)$ est la matrice diagonale des valeurs propres de A_n, $Q_n(4^n, 4^n)$ la matrice des vecteurs propres associés à D_n et $Q_n^{-1}(4^n, 4^n)$ l'inverse de Q_n, alors l'équation (2.4) devient

$$P_n'(t) = Q_n \cdot D_n \cdot Q_n^{-1} \cdot P_n(t). \tag{2.5}$$

Cette équation différentielle (2.5) a pour solution classique [Lan05] :

$$P_n(t) = Q_n \cdot e^{D_n t} \cdot Q_n^{-1} \cdot P_n(0) \tag{2.6}$$

où $e^{D_n t}$ est la matrice diagonale des exponentielles des valeurs propres de la matrice A_n multipliées par le temps d'évolution t, Q_n est la matrice des vecteurs propres associés à A_n et $P_n(0)$ est le vecteur des 4^n probabilités d'occurrence initiales des motifs de taille n au temps $t = 0$.

2.2.2 Inversion du temps

L'équation (2.6) permet donc d'obtenir l'expression analytique de la probabilité d'occurrence de chaque motif i de taille n en fonction d'un vecteur de probabilités d'occurrence initiales de ces motifs, des paramètres de substitution et du temps d'évolution t. Ces probabilités $P_n(t)$ sont donc déterminées au temps t à partir de probabilités initiales (passées) $P_n(0)$, conduisant à une évolution directe (passé-présent). En exprimant $P_n(0)$ en fonction de $P_n(t)$ dans l'équation (2.6), nous obtenons

$$P_n(0) = Q_n \cdot e^{-D_n t} \cdot Q_n^{-1} \cdot P_n(t). \tag{2.7}$$

En renommant $\tilde{P}_n(t) = P_n(0)$ et $\tilde{P}_n(0) = P_n(t)$, l'équation (2.7) peut être réécrite

$$\tilde{P}_n(t) = Q_n \cdot e^{-D_n t} \cdot Q_n^{-1} \cdot \tilde{P}_n(0). \tag{2.8}$$

Ainsi, le remplacement de t par $-t$ dans l'équation (2.6) permet de déterminer les probabilités d'occurrence passées des motifs à partir des probabilités actuelles, ce qui correspond à une évolution inverse (présent-passé). Cette équation a des applications importantes en évolution car ces probabilités initiales associées aux gènes actuels sont connues et peuvent être obtenues à partir des bases de données.

2.2.3 Modèles analytiques basés sur le modèle 3PS

Les modèles d'évolution analytiques développés jusqu'alors pour les dinucléotides et trinucléotides sont basés sur le modèle de Kimura à 3 paramètres de substitution par site nucléotidique 3PS.

2.2.3.1 Construction classique de la matrice de substitution A_n

La matrice de substitution correspondant au modèle 3PS pour des motifs de taille n peut être construite par une formule récursive exprimant la matrice de substitution A_k (4^k,4^k) à partir de matrices de taille inférieure (4^{k-1},4^{k-1}). La matrice A_n est donc calculée récursivement pour k de n à 1 selon la construction suivante :

$$A_k = \begin{pmatrix} A_{k-1} & c_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} \\ c_{n-k+1}I_{k-1} & A_{k-1} & b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} \\ a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} & A_{k-1} & c_{n-k+1}I_{k-1} \\ b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & c_{n-k+1}I_{k-1} & A_{k-1} \end{pmatrix} \tag{2.9}$$

où I_{k-1} est la matrice identité $(4^{k-1}, 4^{k-1})$ avec $I_0 = 1$ et A_{k-1} est la matrice récursive $(4^{k-1}, 4^{k-1})$ avec $A_0 = -\sum_{k=1}^{n} a_k + b_k + c_k$.

Exemple 2.1 *La matrice des taux de substitution A_2 de taille (16,16) associée aux dinucléotides ($n = 2$) est construite récursivement à partir de $k = n = 2$:*

$$A_2 = \begin{pmatrix} A_1 & c_1I_1 & a_1I_1 & b_1I_1 \\ c_1I_1 & A_1 & b_1I_1 & a_1I_1 \\ a_1I_1 & b_1I_1 & A_1 & c_1I_1 \\ b_1I_1 & a_1I_1 & c_1I_1 & A_1 \end{pmatrix} \textit{avec } A_1 = \begin{pmatrix} A_0 & c_2I_0 & a_2I_0 & b_2I_0 \\ c_2I_0 & A_0 & b_2I_0 & a_2I_0 \\ a_2I_0 & b_2I_0 & A_0 & c_2I_0 \\ b_2I_0 & a_2I_0 & c_2I_0 & A_0 \end{pmatrix}.$$

Donc, la matrice de substitution de dinucléotides A_2 vaut :

$$A_2 = \begin{pmatrix}
s & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 \\
c_2 & s & b_2 & a_2 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 \\
a_2 & b_2 & s & c_2 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 \\
b_2 & a_2 & c_2 & s & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 \\
c_1 & 0 & 0 & 0 & s & c_2 & a_2 & b_2 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 \\
0 & c_1 & 0 & 0 & c_2 & s & b_2 & a_2 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 \\
0 & 0 & c_1 & 0 & a_2 & b_2 & s & c_2 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 \\
0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & s & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 \\
a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & s & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 \\
0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & c_2 & s & b_2 & a_2 & 0 & c_1 & 0 & 0 \\
0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & a_2 & b_2 & s & c_2 & 0 & 0 & c_1 & 0 \\
0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & b_2 & a_2 & c_2 & s & 0 & 0 & 0 & c_1 \\
b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & s & c_2 & a_2 & b_2 \\
0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & c_2 & s & b_2 & a_2 \\
0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & a_2 & b_2 & s & c_2 \\
0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & s
\end{pmatrix},$$

avec $s = -(a_1 + b_1 + c_1 + a_2 + b_2 + c_2)$.

2.2.3.2 Méthode de factorisation de Tian et Styan

La matrice de substitution A_n associée aux motifs de taille n possède donc une structure en blocs selon (2.9). Par ailleurs, la matrice de substitution de nucléotides 3P a une structure bicirculante [Kak75]. Cette propriété permet de calculer le déterminant de A_n à partir de sa factorisation en matrices blocs selon la méthode de Tian et Styan [TS01]. Pour cela, nous pouvons réécrire l'équation (2.9) par

$$A_k = \begin{pmatrix} B_0 & B_1 & B_2 & B_3 \\ B_1 & B_0 & B_3 & B_2 \\ B_2 & B_3 & B_0 & B_1 \\ B_3 & B_2 & B_1 & B_0 \end{pmatrix},$$

avec $B_0 = A_{k-1}, B_1 = c_{n-k+1}I_{k-1}, B_2 = a_{n-k+1}I_{k-1}$ et $B_3 = b_{n-k+1}I_{k-1}$. Ceci met en évidence le fait que la matrice A_k a une structure bicirculante par blocs [Kak75]. Nous pouvons donc factoriser A_k par la méthode de factorisation par matrices blocs décrite dans [TS01].

$$A_k = \begin{pmatrix} B_0 & B_1 & B_2 & B_3 \\ B_1 & B_0 & B_3 & B_2 \\ B_2 & B_3 & B_0 & B_1 \\ B_3 & B_2 & B_1 & B_0 \end{pmatrix} = Q_k \cdot \begin{pmatrix} C_0 & & & \\ & C_1 & & \\ & & C_2 & \\ & & & C_3 \end{pmatrix} \cdot Q_k. \tag{2.10}$$

avec C_0, C_1, C_2 et C_3 définis par

$$\begin{aligned} C_0 &= B_0 + B_1 + B_2 + B_3 \\ C_1 &= B_0 + B_1 - B_2 - B_3 \\ C_2 &= B_0 - B_1 + B_2 - B_3 \\ C_3 &= B_0 - B_1 - B_2 + B_3 \end{aligned}$$

et

$$Q_k = \frac{1}{2} \begin{pmatrix} I_{k-1} & I_{k-1} & I_{k-1} & I_{k-1} \\ I_{k-1} & I_{k-1} & -I_{k-1} & -I_{k-1} \\ I_{k-1} & -I_{k-1} & I_{k-1} & -I_{k-1} \\ I_{k-1} & -I_{k-1} & -I_{k-1} & I_{k-1} \end{pmatrix}$$

Les valeurs propres λ_i de la matrice A_k peuvent alors être déduites à partir du calcul des racines du polynôme caractéristique $det(A_k - \lambda I_k)$ de la matrice factorisée $(A_k - \lambda I_k)$ et des règles classiques de calcul de déterminants, en particulier pour deux matrices X et Y, $det(X \cdot Y) = det(X)det(Y)$ et pour une matrice diagonale D, $det(D) = \prod_i d_{ii}$.

Pour $k > 2$ cette factorisation peut être appliquée récursivement sur les matrices C_0, C_1, C_2 et C_3 car elles sont également bicirculantes par blocs par construction.

Remarque 2.1 *La factorisation par blocs de Tian et Styan [TS01] est définie pour des matrices rectangulaires, nous l'appliquons ici dans le cas particulier de matrices carrées.*

2.2.4 Modèle analytique généralisé basé sur le modèle 3PS

La méthode de factorisation par blocs de Tian et Styan vue précédemment permet de déterminer les valeurs propres des matrices de substitution de type 3PS pour des motifs de taille

quelconque à partir du calcul des racines du polynôme caractéristique. Toutefois, le calcul devient très coûteux même pour des petites tailles de motifs. De plus, cette méthode ne permet pas de calculer simplement les vecteurs propres nécessaires à la diagonalisation de la matrice de substitution. Enfin, il est nécessaire de construire la matrice de substitution dans son intégralité pour appliquer cette méthode. Ceci s'avère lourd en temps de calcul pour des tailles de motifs strictement supérieures à 3.

Une analyse approfondie des matrices de substitution de type 3PS pour les nucléotides, dinucléotides et trinucléotides ainsi que des valeurs et vecteurs propres correspondants nous a permis d'identifier des propriétés supplémentaires sur ces matrices de substitution, notamment leur construction possible par des sommes de Kronecker. Cette propriété est à la base d'une méthode généralisée de diagonalisation de matrices de substitution pour des motifs de taille quelconque. Cette méthode, dite de Kronecker, permet ainsi de déterminer simplement les valeurs et vecteurs propres des matrices de substitution de motifs de taille quelconque à partir des valeurs et vecteurs propres des matrices de substitution de nucléotides par site.

Les opérateurs de Kronecker (somme et produit) sont des outils classiques utilisés dans divers champs des mathématiques appliquées comme les processus de Poisson modulés par des chaînes de Markov, par exemple en communication [BS86, SVP03]. Ils sont également utilisés en phylogénie avec les processus de Markov, en particulier pour diagonaliser des matrices de Markov modulées pour l'étude évolutive de taux de substitution au cours du temps (modèle covarion [GJM04, WSSR07, AR09]), ou encore pour déterminer des taux de substitution de dinucléotides ou trinucléotides [MBM+08, LH04, DYH06].

2.2.4.1 Opérations de Kronecker

Dans la suite de ce chapitre, nous noterons $\otimes$ le produit de Kronecker défini pour deux matrices carrées $X(m,m)$ et $Y(n,n)$ par :

$$X \otimes Y = \begin{pmatrix} x_{1,1} \times Y & x_{1,2} \times Y & \dots & x_{1,m} \times Y \\ x_{2,1} \times Y & x_{2,2} \times Y & \dots & x_{2,m} \times Y \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} \times Y & x_{m,2} \times Y & \dots & x_{m,m} \times Y \end{pmatrix}$$

Nous utiliserons également cet opérateur $\otimes$ étendu au produit de k matrices carrées X_i, $1 \leq i \leq k$:

$$\bigotimes_{i=1}^{k} X_i = X_1 \otimes X_2 \otimes \ldots \otimes X_k.$$

Nous noterons $\oplus$ la somme de Kronecker définie pour deux matrices carrées $X(m,m)$ et $Y(n,n)$ par :

$$X \oplus Y = X \otimes I_n + I_m \otimes Y,$$

où I_n et I_m sont les matrices identité de tailles respectives (n,n) et (m,m). De la même façon, nous étendrons cet opérateur à la somme de Kronecker de k matrices carrées X_i, $1 \leq i \leq k$:

$$\bigoplus_{i=1}^{k} X_i = X_1 \oplus X_2 \oplus \ldots \oplus X_k.$$

Remarque 2.2 *Les produit $\otimes$ et somme $\oplus$ de Kronecker ne sont pas commutatifs.*

2.2.4.2 Construction de la matrice de substitution A_n par somme de Kronecker

La construction récursive de la matrice de substitution A_n pour des motifs de taille n (2.9) correspond à une somme de Kronecker des matrices de substitution associées à chaque site k des motifs de taille n. En effet, la matrice A_k des taux de substitution associée au site k peut s'écrire comme la somme de deux matrices telles que

$$A_k = \begin{pmatrix} A_{k-1} & c_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} \\ c_{n-k+1}I_{k-1} & A_{k-1} & b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} \\ a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} & A_{k-1} & c_{n-k+1}I_{k-1} \\ b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & c_{n-k+1}I_{k-1} & A_{k-1} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\ c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\ a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\ b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0 \end{pmatrix} \otimes I_{k-1} + \begin{pmatrix} A_{k-1} & & & \\ & A_{k-1} & & \\ & & A_{k-1} & \\ & & & A_{k-1} \end{pmatrix},$$

où la matrice diagonale composée de sous-matrices $A_{k-1}(4^{k-1}, 4^{k-1})$ est de taille $(4^k, 4^k)$. Ainsi, selon la définition du produit de Kronecker, nous pouvons réécrire l'équation précédente par

$$A_k = \begin{pmatrix} 0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\ c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\ a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\ b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0 \end{pmatrix} \otimes I_{k-1} + I_1 \otimes A_{k-1}, \tag{2.11}$$

avec I_1 la matrice identité $(4, 4)$. L'équation (2.11) correspond alors, par définition, à une somme de Kronecker

$$A_k = \begin{pmatrix} 0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\ c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\ a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\ b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0 \end{pmatrix} \oplus A_{k-1}. \tag{2.12}$$

Soit N_k la matrice des taux de substitution instantanée de nucléotides associée au site $1 \leq k \leq n$ d'un motif de taille n. Cette matrice est donc définie par

$$N_k = \begin{pmatrix} s_k & c_k & a_k & b_k \\ c_k & s_k & b_k & a_k \\ a_k & b_k & s_k & c_k \\ b_k & a_k & c_k & s_k \end{pmatrix}, \tag{2.13}$$

avec $s_k = -a_k - b_k - c_k$.

En remarquant que

$$N_{n-k+1} = \begin{pmatrix} 0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\ c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\ a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\ b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0 \end{pmatrix} + \begin{pmatrix} s_{n-k+1} & 0 & 0 & 0 \\ 0 & s_{n-k+1} & 0 & 0 \\ 0 & 0 & s_{n-k+1} & 0 \\ 0 & 0 & 0 & s_{n-k+1} \end{pmatrix},$$

$$= \begin{pmatrix} 0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\ c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\ a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\ b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0 \end{pmatrix} + s_{n-k+1} \times I_1, \tag{2.14}$$

avec $s_{n-k+1} = -(a_{n-k+1} + b_{n-k+1} + c_{n-k+1})$, l'équation récursive (2.12) peut être réécrite à l'aide d'une construction récursive à base de sommes de Kronecker

$$A_k = (N_{n-k+1} - s_{n-k+1} \times I_1) \oplus A_{k-1},$$

ce qui implique que la matrice de substitution A_n peut être définie par

$$A_n = \bigoplus_{k=1}^{n} (N_k - s_k \times I_1) + A_0 \times I_n,$$

avec $A_0 = -\sum_{k=1}^{n}(a_k + b_k + c_k)$. Nous obtenons donc

$$\begin{aligned} A_n &= \bigoplus_{k=1}^{n} N_k - \bigoplus_{k=1}^{n} (s_k \times I_1) + A_0 \times I_n \\ &= \bigoplus_{k=1}^{n} N_k - \sum_{k=1}^{n} s_k \times I_n + A_0 \times I_n, \end{aligned}$$

Or, comme $A_0 = -\sum_{k=1}^{n}(a_k + b_k + c_k) = -\sum_{k=1}^{n} s_k$, nous pouvons en conclure que

$$A_n = \bigoplus_{k=1}^{n} N_k. \tag{2.15}$$

Pour illustrer la méthode de construction d'une matrice de substitution par sommes de Kronecker, un exemple de construction de matrice de substitution de type 3PS associée aux dinucléotides est proposée dans l'annexe A.

2.2.4.3 Diagonalisation de la matrice de substitution A_n

La matrice de substitution de nucléotides N_k, de type 3P, associée au site k d'un motif de taille n est réelle et symétrique. N_k est donc diagonalisable et quelque soit $1 \leq k \leq n$

$$N_k = R \cdot S_k \cdot R^{-1}$$

avec S_k la matrice spectrale de N_k associée aux nucléotides

$$S_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_k + b_k) & 0 & 0 \\ 0 & 0 & -2(a_k + c_k) & 0 \\ 0 & 0 & 0 & -2(b_k + c_k) \end{pmatrix}$$

et R la matrice des vecteurs propres associés à S_k

$$R = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}. \tag{2.16}$$

Nous pouvons noter que la matrice R n'est pas formelle et qu'elle ne dépend pas des taux de substitution.

Remarque 2.3 *Dans le cas de la matrice de substitution de nucléotides, $n = 1$ et $A_1 = N_1 = Q_1 \cdot D_1 \cdot Q_1^{-1} = R \cdot S_1 \cdot R^{-1}$, ce qui implique que $D_1 = S_1$ et $Q_1 = R$.*

Nous avons vu que la matrice de substitution $A_n(4^n, 4^n)$ pouvait être calculée à partir de la somme de Kronecker des n matrices de substitution de nucléotides $N_k(4,4)$ (2.15). Or, un théorème classique sur les produit et somme de matrices de Kronecker [Lau05] permet de déterminer les matrices spectrale D_n et de vecteurs propres Q_n de A_n à partir des matrices spectrale S_k et de vecteurs propres R respectivement. Il en résulte

$$\begin{cases} D_n = \oplus_{k=1}^n S_k \\ Q_n = \otimes_{k=1}^n R \\ Q_n^{-1} = (\otimes_{k=1}^n R)^{-1} = \otimes_{k=1}^n R^{-1} \end{cases}. \tag{2.17}$$

La matrice de substitution A_n peut ainsi être directement diagonalisée à partir de la somme de Kronecker des n matrices spectrales S_k et du produit de Kronecker des n matrices de vecteurs propres R. L'équation (2.15) peut alors s'écrire

$$A_n = \bigotimes_{k=1}^{n} R \cdot \bigoplus_{k=1}^{n} S_k \cdot \bigotimes_{k=1}^{n} R^{-1}. \tag{2.18}$$

En conclusion, les valeurs et vecteurs propres pour des matrices de substitution de motifs de taille quelconque peuvent être obtenus à partir de sommes et de produits de Kronecker des valeurs et vecteurs propres des matrices de substitution de nucléotides.

2.2.4.4 Solutions analytiques des probabilités d'occurrence des motifs de taille n au temps t

En remplaçant dans l'équation solution (2.6) les matrices Q_n, D_n et Q_n^{-1} par leurs valeurs respectives dans (2.17), nous obtenons la probabilité d'occurrence $P_n(t)$ des motifs de taille n au temps t exprimée en fonction des valeurs et vecteurs propres élémentaires associés aux nucléotides de chaque site k

$$\begin{aligned} P_n(t) &= \bigotimes_{k=1}^{n} R \cdot e^{\bigoplus_{k=1}^{n} S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0) \\ &= \bigotimes_{k=1}^{n} R \cdot \bigotimes_{k=1}^{n} e^{S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0). \end{aligned}$$

Finalement,

$$P_n(t) = \bigotimes_{k=1}^{n} (R \cdot e^{S_k t} \cdot R^{-1}) \cdot P_n(0), \tag{2.19}$$

où $e^{S_k t}$ est la matrice diagonale des exponentielles des valeurs propres des matrices N_k.

Remarque 2.4 *La matrice R des vecteurs propres de la matrice de substitution de nucléotides N_k de type 3P a ses valeurs dans $\{1, -1\}$ et $R \cdot R^T = 4 \times I_1$, avec R^T la transposée de R et I_1 la matrice identité $(4,4)$. R est donc une matrice de Hadamard d'ordre 4 [Had93, Hor07] dont les vecteurs colonnes sont des fonctions de Walsh [Wel70]. Les matrices de vecteurs propres des matrices de substitution A_n peuvent donc également être construites à partir de produits de Kronecker de matrices de Hadamard d'ordre 4 ou récursivement par une méthode de construction de Sylvester [Syl67, Hen05].*

2.2.4.5 Solution analytique pour un motif donné de taille n au temps t

L'équation (2.19) donne les solutions analytiques des 4^n motifs génétiques de taille n sous forme d'un vecteur de probabilités. Pour des tailles de motifs supérieures à 3, la résolution de l'équation nécessite d'importantes ressources en termes de temps de calcul pour un PC standard actuel. Dans certains cas, par exemple lors de l'étude de certains motifs génétiques au sein de génomes, seules les probabilités de quelques motifs peuvent être significatives. Pour éviter le calcul des probabilités des 4^n motifs, nous avons également dérivé une formule calculant la probabilité d'occurrence analytique $P_{i_1}(t)$ au temps t d'un motif i_1 donné de taille n.

Soit $P_{i_1}(t)$ la probabilité d'occurrence du motif d'indice $1 \leq i_1 \leq 4^n$ de taille n au temps t. $P_{i_1}(t)$ peut être directement obtenue à partir du vecteur des probabilités d'occurrence analytiques $P_n(t) = [P_i(t)]_{1 \leq i \leq 4^n}$.

Nous pouvons donc écrire :

$$P_{i_1}(t) = P_n(t)[i_1] = \left(\bigotimes_{k=1}^{n} (R \cdot e^{S_k t} \cdot R^{-1}) \cdot P_n(0) \right) [i_1] \tag{2.20}$$

$$= \left(\bigotimes_{k=1}^{n} (R \cdot e^{S_k t} \cdot R^{-1}) \right) [i_1,] \cdot P_n(0) \tag{2.21}$$

$$P_{i_1}(t) = \left(\bigotimes_{k=1}^{n} R \right) [i_1,] \cdot \bigotimes_{k=1}^{n} e^{S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0) \tag{2.22}$$

Nous voyons donc que le calcul de la probabilité d'occurrence analytique du motif i_1 ne nécessite pas de construire la totalité de la matrice $Q(4^n, 4^n) = (\otimes_{k=1}^{n} R)$, mais seulement la $i_1{}^{\text{ème}}$ ligne : $Q[i_1,] = (\otimes_{k=1}^{n} R)\,[i_1,]$.

De plus, la matrice $Q(4^n, 4^n)$ est construite à partir d'un produit de Kronecker de n matrices $R(4, 4)$. La $i_1{}^{\text{ème}}$ ligne de Q peut donc être déterminée à partir d'un produit de Kronecker de n lignes de R, c'est à dire, une ligne par site $k, 1 \leq k \leq n$, de motif.

Le calcul de cette $i_1{}^{\text{ème}}$ ligne de Q revient donc à faire le produit de Kronecker des lignes de R d'indices correspondant aux indices des nucléotides de chaque site k du motif d'indice i_1. Par exemple, pour calculer la ligne de Q correspondant à la probabilité d'occurrence du motif GCT, de taille $n = 3$ et d'indice $i_1 = 40$, il suffit d'effectuer le produit de Kronecker des 3 lignes de R suivantes, d'indices $\delta(i_1, k)$: la ligne $\delta(i_1, 1) = 3$ correspondant au nucléotide G pour le site $k = 1$, la ligne $\delta(i_1, 2) = 2$ correspondant au nucléotide C pour le site $k = 2$ et la ligne $\delta(i_1, 3) = 4$ correspondant au nucléotide T pour le site $k = 3$. Ainsi, la ligne d'indice $i_1 = 40$ de Q est obtenue par

$$Q\,[40,] = R\,[\delta(i_1, 1),] \otimes R\,[\delta(i_1, 2),] \otimes R\,[\delta(i_1, 3),] = R\,[3,] \otimes R\,[2,] \otimes R\,[4,]\,.$$

Les indices $\delta(i_1, k)$ des nucléotides de chaque site k d'un motif génétique de taille n et d'indice i_1 sont obtenus par la relation suivante

$$\delta(i_1, k) = \lfloor \frac{i_1 - 1}{4^{n-k}} \rfloor\,[4] + 1$$

où $\lfloor \frac{i_1 - 1}{4^{n-k}} \rfloor$ représente la partie entière de $\frac{i_1 - 1}{4^{n-k}}$ et $[\ ]$ est la fonction modulo.

Reprenons l'exemple précédent pour illustrer cette relation. Le motif GCT ($n = 3$) a pour indice $i_1 = 40$. Le calcul des indices $\delta(i_1, k)$ donne donc

$$\begin{aligned}
\delta(i_1,1) &= \lfloor \frac{40-1}{4^{3-1}} \rfloor [4] + 1 = \lfloor \frac{39}{16} \rfloor [4] + 1 = 2\,[4] + 1 = 3 \\
\delta(i_1,2) &= \lfloor \frac{40-1}{4^{3-2}} \rfloor [4] + 1 = \lfloor \frac{39}{4} \rfloor [4] + 1 = 9\,[4] + 1 = 2 \\
\delta(i_1,3) &= \lfloor \frac{40-1}{4^{3-3}} \rfloor [4] + 1 = \lfloor \frac{39}{1} \rfloor [4] + 1 = 3\,[4] + 1 = 4.
\end{aligned}$$

Nous pouvons donc réécrire l'équation (2.22) :

$$P_{i_1}(t) = \bigotimes_{k=1}^{n} R\,[\delta(i_1,k),] \cdot \bigotimes_{k=1}^{n} e^{S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0). \tag{2.23}$$

L'équation (2.23) peut également être simplifiée. En effet, le produit vectoriel

$$\bigotimes_{k=1}^{n} R\,[\delta(i_1,k),] \cdot \bigotimes_{k=1}^{n} e^{S_k t}$$

est effectué entre un vecteur ligne $Q[i_1,]$ et une matrice diagonale e^{St} dont les éléments diagonaux sont les exponentielles des valeurs propres de la matrice de substitution de taille $(4^n, 4^n)$. Cette matrice e^{St} est également obtenue par un produit de Kronecker de n matrices $e^{S_k t}$. Le produit vectoriel $Q[i_1,] \cdot e^{St}$ peut donc être remplacé par un produit scalaire (terme à terme) entre le vecteur ligne $Q[i_1,]$ et un vecteur ligne e^{Lt} composé des éléments diagonaux de e^{St}.

Soit L_k le vecteur ligne associé au site k d'un motif de taille n et composé des éléments diagonaux de S_k. L_k est donc un vecteur ligne des valeurs propres d'une matrice de substitution N_k de nucléotides associée au site k d'un motif. Nous pouvons donc noter $e^{L_k t}$ le vecteur ligne des exponentielles des valeurs propres $\lambda_i, 1 \leq i \leq 4^n$ composé des éléments diagonaux de $e^{S_k t}$. Ainsi, effectuer le produit de Kronecker des n vecteurs ligne $e^{L_k t}$ revient à construire un vecteur ligne composé des éléments diagonaux du produit de Kronecker de n matrices diagonales $e^{S_k t}$.

Nous pouvons donc réécrire l'équation (2.23) :

$$P_{i_1}(t) = \bigotimes_{k=1}^{n} R\,[\delta(i_1,k),] \times \bigotimes_{k=1}^{n} e^{L_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0) \tag{2.24}$$

où $\otimes_{k=1}^{n} e^{L_k t}$ est un vecteur ligne. Cette opération divise par 4^n le nombre d'opérations à effectuer par rapport à l'équation (2.23).

Soient U_{i_1} le vecteur ligne associé au motif d'indice i_1 et résultant du produit scalaire

$$U_{i_1} = \bigotimes_{k=1}^{n} R\,[\delta(i_1,k),] \times \bigotimes_{k=1}^{n} e^{L_k t}$$

et V_{i_1} le vecteur ligne également associé au motif d'indice i_1 et résultant du produit vectoriel

$$V_{i_1} = U_{i_1} \cdot \bigotimes_{k=1}^{n} R^{-1}.$$

Nous pouvons alors déterminer l'expression des éléments i des vecteurs U_{i_1} et V_{i_1} :

$$U_{i_1}\,[i] = \left(\bigotimes_{k=1}^{n} R \right) [i_1, i] \times e^{\lambda_i t},$$

avec $e^{\lambda_i t}$ le $i$ème élément du vecteur ligne $\otimes_{k=1}^{n} R\left[\delta(i_1,k),\right]$ et

$$V_{i_1}[i] = \sum_{j=1}^{4^n} U_{i_1}[j] \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[j,i].$$

Donc,

$$V_{i_1}[i] = \sum_{j=1}^{4^n} \left(\left(\bigotimes_{k=1}^{n} R\right)[i_1,j] \times e^{\lambda_j t}\right) \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[j,i].$$

Ainsi, d'après l'équation (2.24), nous pouvons écrire

$$P_{i_1}(t) = V_{i_1} \cdot P_n(0),$$

et donc

$$\begin{aligned}
P_{i_1}(t) &= \sum_{i_2=1}^{4^n} V_{i_1}[i_2] \times P_{i_2}(0) \\
&= \sum_{i_2=1}^{4^n} \left(P_{i_2}(0) \times \sum_{i_3=1}^{n} \left(U_{i_1}[i_3] \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[i_3,i_2]\right)\right) \\
&= \sum_{i_2=1}^{4^n} \left(P_{i_2}(0) \times \sum_{i_3=1}^{n} \left(\left(\bigotimes_{k=1}^{n} R\right)[i_1,i_3] \times e^{\lambda_{i_3} t} \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[i_3,i_2]\right)\right).
\end{aligned}$$

En exprimant $P_{i_1}(t)$ comme une somme de 4^n exponentielles de valeurs propres $e^{\lambda_{i_2} t}$ associées chacune à une somme de 4^n probabilités d'occurrence initiales nous obtenons :

$$P_{i_1}(t) = \sum_{i_2=1}^{4^n} e^{\lambda_{i_2} t} \times \sum_{i_3=1}^{4^n} \left(\left(\bigotimes_{k=1}^{n} R\right)[i_1,i_2] \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[i_2,i_3] \times P_{i_3}(0)\right)$$

$$P_{i_1}(t) = \sum_{i_2=1}^{4^n} e^{\lambda_{i_2} t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k),\delta(i_2,k)\right]\right) \times \prod_{k=1}^{n} \left(R^{-1}\left[\delta(i_2,k),\delta(i_3,k)\right]\right) \times P_{i_3}(0)\right)$$

$$P_{i_1}(t) = \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k),\delta(i_2,k)\right]\right) \times \prod_{k=1}^{n} \left(R^{-1}\left[\delta(i_2,k),\delta(i_3,k)\right]\right) \times P_{i_3}(0)\right)$$

où $\delta(i_x,k) = \lfloor \frac{i_x-1}{4^{n-k}} \rfloor [4] + 1$.

Nous pouvons donc écrire

$$P_{i_1}(t) = \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k),\delta(i_2,k)\right] \times R^{-1}\left[\delta(i_2,k),\delta(i_3,k)\right]\right) \times P_{i_3}(0)\right). \tag{2.25}$$

Remarque 2.5 *Les matrices des vecteur propres R des matrices N_k étant symétriques, nous avons $R^{-1} = \frac{1}{4}R$. Les coefficients $R[\delta(i_1,k),\delta(i_2,k)] \times R[\delta(i_2,k),\delta(i_3,k)]$ dans l'équation (2.26) sont donc similaires aux éléments des matrices O_k définies dans [LM10] par $O_k[i,j] = R[i,k]R^{-1}[k,j]$.*

D'après la remarque 2.5, l'équation (2.25) peut encore être simplifiée en terme d'opérations à effectuer :

$$P_{i_1}(t) = \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(\frac{1}{4} R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right] \right) \times P_{i_3}(0) \right).$$

Finalement,

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right] \right) \times P_{i_3}(0) \right) \tag{2.26}$$

où $L_k = \{0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)\}$ pour tout site k.

Exemple 2.2 *Nous donnons en annexe B un exemple de calcul de la probabilité d'occurrence analytique obtenue à partir de l'équation (2.26) pour le dinucléotide AG. Ce motif est de taille $n = 2$ et d'indice $i_1 = 3$ parmi $4^n = 16$ dinucléotides.*

2.2.4.6 Comportement des probabilités d'occurrence au cours du temps d'évolution

Comme nous venons de le voir, l'équation (2.26) donne l'expression analytique de la probabilité d'occurrence d'un motif i_1 donné de taille n pour le modèle d'évolution 3PS dans le sens d'évolution direct (passé-présent). Cette expression est composée d'une somme d'exponentielles de termes négatifs ou nuls appartenant à L_k et multipliés par le temps d'évolution t. Ainsi, dans le sens direct, ces exponentielles ont une asymptote horizontale, les probabilités d'occurrence des 4^n motifs tendent vers $\frac{1}{4^n}$.

En revanche, nous avons vu dans la section 2.2.2 qu'il suffit de remplacer le temps d'évolution t par $-t$ dans l'expression analytique des probabilités d'occurrence des motifs pour effectuer une évolution inverse (présent-passé). Par conséquent, dans le sens d'évolution inverse, les termes des exponentielles des expressions analytiques des probabilités d'occurrence des 4^n motifs deviennent positifs ou nuls. Les probabilités d'occurrence analytiques ont donc des asymptotes verticales dans le cas d'une évolution inverse. Ceci pose un problème de validité des probabilités d'occurrence qui nous a amené à développer un modèle d'évolution numérique inverse prenant en compte ce problème et que nous présentons dans la section 2.4.

2.2.4.7 Généralisation de la méthode de Kronecker à d'autres modèles

La méthode de Kronecker est présentée dans la sous-section précédente en utilisant le modèle d'évolution à base de matrices de substitution de type 3P comme les matrices N_k associées aux nucléotides et plus généralement les matrices $\oplus_{k=1}^{n} N_k$ de type 3PS associées aux motifs de taille n. Cette théorie s'applique donc aux cas particuliers à deux paramètres 2P (Kimura 1980) et un paramètre 1P (Jukes et Cantor 1969). Ces matrices ayant une structure bicirculante, leurs vecteurs propres sont constants, comme nous l'avons vu pour la matrice R (2.16) des vecteurs propres de la matrice de substitution de nucléotides N_k. Ces vecteurs propres sont donc indépendants du site k du motif.

Cependant, cette méthode de Kronecker n'est pas intrinsèque à ce type de matrices de substitution 3PS. Elle peut donc être appliquée à d'autres matrices de substitution. En effet, cette

méthode est liée au procédé de construction de la matrice de substitution pour des motifs de taille quelconque et non au nombre de paramètres de substitution. Ainsi, dans le cas le plus général, cette méthode est applicable à des matrices de substitution N'_k à douze paramètres

$$N'_k = \begin{pmatrix} s_{1k} & a_k & b_k & c_k \\ g_k & s_{2k} & d_k & e_k \\ h_k & i_k & s_{3k} & f_k \\ j_k & l_k & m_k & s_{4k} \end{pmatrix}$$

avec $s_{1k} = -g_k - h_k - j_k, s_{2k} = -a_k - i_k - l_k, s_3 = -b_k - d_k - m_k$ et $s_{4k} = -c_k - e_k - f_k$. Cette méthode peut donc s'appliquer en théorie au modèle général G4H présenté dans [Zha94].

La généralisation de la méthode de Kronecker à des modèles de substitution autres que 3PS nécessite donc dans le cas général la prise en compte de la dépendance des valeurs et vecteurs propres des matrices de substitution associées à chaque site k du motif. Dans le cas général, la matrice spectrale S'_k et la matrice des vecteurs propres R'_k sont associées au site k du motif.

Ainsi, à la différence de l'équation 2.19 donnant le vecteur $P_n(t)$ des probabilités d'occurrence des 4^n motifs de taille n avec le modèle 3PS, ce vecteur est dans le cas général

$$P_n(t) = \otimes_{k=1}^{n}(R'_k \cdot e^{S'_k t} \cdot R'^{-1}_k) \cdot P_n(0) \tag{2.27}$$

2.2.5 Généralisation du modèle 3PS à un modèle à quatre paramètres de substitution par site

Les modèles analytiques de substitution pour des motifs génétiques (nucléotides, dinucléotides et trinucléotides) développés jusqu'à présent sont basés sur le modèle 3PS. La méthode de Kronecker de détermination des solutions analytiques des probabilités d'occurrence pour des motifs de taille quelconque que nous venons de présenter nous permet de généraliser ces modèles d'évolution à des modèles d'évolution prenant en compte plus de paramètres de substitution par site de motif. En théorie, cette méthode peut s'appliquer à des modèles prenant en compte jusqu'à douze paramètres de substitution par site de motif. Or, l'obtention de solutions analytiques nécessite au préalable la possibilité de diagonaliser la matrice de substitution, ce qui n'est pas le cas de toutes les matrices de taille $(4, 4)$.

Cependant, il est possible de généraliser le modèle 3PS avec un modèle à quatre paramètres de substitution par site de motif, modèle que nous noterons 4PS et dont la matrice de substitution de nucléotides $N(4P)_k$ associée au site k d'un motif est

$$N(4P)_k = \begin{pmatrix} s_{1k} & c_k & a_k & b_k \\ c_k & s_{2k} & d_k & a_k \\ a_k & d_k & s_{2k} & c_k \\ b_k & a_k & c_k & s_{1k} \end{pmatrix}, \tag{2.28}$$

cette matrice étant également stochastique en colonnes avec $s_{1k} = -a_k - b_k - c_k$ et $s_{2k} = -a_k - c_k - d_k$.

Cette matrice de substitution différencie les transitions et les deux types de transversion comme dans le cas du modèle 3P. De plus, elle effectue une distinction au sein des transversions de type II (A↔T, C↔G) qu'elle différencie en deux sous-types : les transversions de type II faibles

(A↔T) intervenant entre les deux bases azotées A et T, liées par deux liaisons hydrogène, et les transversions de type II fortes (C↔G) intervenant entre les deux bases azotées C et G, liées par trois liaisons hydrogènes. Dans ce modèle b_k est donc le paramètre de substitution correspondant à une transversion de type II faible sur le site k d'un motif et d_k est le paramètre de substitution correspondant à une transversion de type II forte site le site k d'un motif.

2.2.5.1 Solutions analytiques

La matrice de substitution $N(4P)_k$ est réelle, symétrique et diagonalisable. Nous avons donc

$$N(4P)_k = R(4P)_k \cdot S(4P)_k \cdot R(4P)_k{}^{-1} \tag{2.29}$$

avec la matrice spectrale $S(4P)_k$

$$S(4P)_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_k+c_k) & 0 & 0 \\ 0 & 0 & -a_k-b_k-c_k-d_k-r_k & 0 \\ 0 & 0 & 0 & -a_k-b_k-c_k-d_k+r_k \end{pmatrix}, \tag{2.30}$$

$R(4P)_k$ la matrice des vecteurs propres colonnes de $N(4P)_k$

$$R(4P)_k = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & \frac{b_k-d_k-r_k}{a_k-c_k} & \frac{b_k+r_k-d_k}{a_k-c_k} \\ 1 & -1 & -\frac{b_k-d_k-r_k}{a_k-c_k} & -\frac{b_k+r_k-d_k}{a_k-c_k} \\ 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.31}$$

où $r_k = \sqrt{(a_k-c_k)^2+(b_k-d_k)^2}$, et $R(4P)_k{}^{-1}$ la matrice inverse de $R(4P)_k$.

Contrairement aux modèles vus précédemment (1P, 2P et 3P), la matrice des vecteurs propres est formelle et fonction des taux de substitution.

Les solutions analytiques formelles des probabilités d'occurrence des dinucléotides obtenues avec le modèle 4PS étant de taille trop importante pour être présentées dans ce document, nous proposons en annexe C les solutions analytiques formelles des probabilités d'occurrence des nucléotides obtenues avec le modèle de subsitution de nucléotides 4P.

2.2.5.2 Matrices de substitution similaires

La matrice de substitution $N(4P)_k = N(4P_A)_k$ du modèle 4P associée au site k d'un motif présentée précédemment est similaire à deux autres matrices que nous appellerons $N(4P_B)_k$ et $N(4P_C)_k$. Ces matrices sont définies comme suit :

$$N(4P_B)_k = \begin{pmatrix} s_{1k} & c_k & a_k & b_k \\ c_k & s_{2k} & b_k & d_k \\ a_k & b_k & s_{1k} & c_k \\ b_k & d_k & c_k & s_{2k} \end{pmatrix}, s_{1k} = -a_k-b_k-c_k \text{ et } s_{2k} = -b_k-c_k-d_k. \tag{2.32}$$

et

$$N(4P_C)_k = \begin{pmatrix} s_{1k} & c_k & a_k & b_k \\ c_k & s_{1k} & b_k & a_k \\ a_k & b_k & s_{2k} & d_k \\ b_k & a_k & d_k & s_{2k} \end{pmatrix}, s_{1k} = -a_k-b_k-d_k, s_{2k} = -a_k-b_k-d_k. \tag{2.33}$$

Les matrices spectrales et de vecteurs propres colonnes associées aux matrices $N(4P_B)_k$ et $N(4P_C)_k$ sont définies par

$$S(4P_B)_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(b_k+c_k) & 0 & 0 \\ 0 & 0 & -a_k-b_k-c_k-d_k-r_k & 0 \\ 0 & 0 & 0 & -a_k-b_k-c_k-d_k+r_k \end{pmatrix} \tag{2.34}$$

et

$$R(4P_B)_k = \begin{pmatrix} 1 & -1 & -\frac{a_k-d_k+r_k}{b_k-c_k} & -\frac{a_k-d_k-r_k}{b_k-c_k} \\ 1 & 1 & -1 & -1 \\ 1 & -1 & \frac{a_k-d_k+r_k}{b_k-c_k} & \frac{a_k-d_k-r_k}{b_k-c_k} \\ 1 & 1 & 1 & 1 \end{pmatrix}, \tag{2.35}$$

avec $r_k = \sqrt{(b_k-c_k)^2+(a_k-d_k)^2}$, ainsi que

$$S(4P_C)_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_k+b_k) & 0 & 0 \\ 0 & 0 & -a_k-b_k-c_k-d_k-r_k & 0 \\ 0 & 0 & 0 & -a_k-b_k-c_k-d_k+r_k \end{pmatrix} \tag{2.36}$$

et

$$R(4P_C)_k = \begin{pmatrix} 1 & -1 & \frac{c_k-d_k+r_k}{a_k-b_k} & \frac{c_k-d_k-r_k}{a_k-b_k} \\ 1 & -1 & -\frac{c_k-d_k+r_k}{a_k-b_k} & -\frac{c_k-d_k-r_k}{a_k-b_k} \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \tag{2.37}$$

avec $r_k = \sqrt{(a_k-b_k)^2+(c_k-d_k)^2}$.

Remarque 2.6 *La matrice de substitution $N(4P_B)_k$ est la matrice de substitution de nucléotides $TIM1e$ avec des fréquences stationnaires égales [Pos03].*

2.3 Méthode de résolution numérique

Tous les modèles d'évolution ne permettent pas une résolution analytique de l'équation matricielle différentielle (2.3). En effet, cela dépend de la possibilité de diagonaliser la matrice de substitution utilisée et d'obtenir des valeurs et vecteurs propres réels. Ce n'est par exemple pas le cas pour la matrice de substitution de nucléotides à six paramètres de substitution SYM présentée dans la section 1.3.2.2 du chapitre 1. D'autres modèles avec paramètres de substitution non constants, c'est-à-dire dépendants du temps, nécessitent une résolution numérique. Par exemple, les probabilités d'occurrence de trinucléotides calculées avec le modèle non homogène de substitution de trinucléotides de Bahi et Michel [BM04] (chapitre 1 section 1.3.4) ne peuvent pas être déterminées analytiquement.

Nous détaillons donc la résolution numérique de l'équation différentielle (2.3) puis nous proposons un modèle numérique d'évolution inverse de trinucléotides avec blocage dynamique des substitutions reposant sur cette méthode de résolution numérique.

2.3.1 Résolution numérique avec paramètres de substitution fonction du temps

Dans le cas d'une matrice des probabilités de substitution non constantes, l'équation matricielle différentielle associée (2.3) doit être réécrite

$$P_n'(t) = (M_n(t) - I_n) \cdot P_n(t) = A_n(t) \cdot P_n(t), \tag{2.38}$$

où $M_n(t)$ et $A_n(t)$ représentent respectivement la matrice des probabilités de substitution instantanée et la matrice des taux de substitution instantanée au temps t.

Pour résoudre cette équation différentielle, nous supposons qu'il existe des intervalles de temps $t_{h+1} - t_h$, $t_h < t_{h+1}$ $\forall 0 \leq h$, $t_0 < t_1 < \ldots < t$, suffisamment petits pendant lesquels les probabilités d'occurrence des motifs et les taux de substitution sont constants, ce qui implique que $M_n(t)$ et $A_n(t)$ sont constantes pendant ces intervalles de temps. Ainsi, $\forall t \in [t_h, t_{h+1}]$,

$$\begin{aligned} M_n(t) &= M_{n,h}, \\ A_n(t) &= A_{n,h}. \end{aligned}$$

Les matrices $M_{n,h}$ et $A_{n,h}$ changent à chaque pas d'évolution h. Nous pouvons alors réécrire l'équation différentielle (2.38)

$$P_n'(t) = A_{n,h} \cdot P_n(t), \forall t \in [t_h, t_{h+1}]$$

où $A_{n,h} = M_{n,h} - I_n$.

Le vecteur de probabilités d'occurrence au temps t, $t \in [t_h, t_{h+1}]$, de motifs génétiques de taille n est donc

$$P_n(t) = e^{A_{n,h}(t-t_h)} \cdot e^{A_{n,h-1}(t_h - t_{h-1})} \cdot \ldots \cdot e^{A_{n,1}(t_2 - t_1)} \cdot e^{A_{n,0}(t_1 - t_0)} \cdot P_n(0) \tag{2.39}$$

où $e^{A_{n,h}(t-t_h)}$ est l'exponentielle de la matrice $A_{n,h}(t - t_h)$.

Remarque 2.7 *Lorsque l'intervalle de temps $[t_0, t]$ est échantillonné en $h+1$ intervalles égaux $\Delta = t - t_h = t_h - t_{h-1} = \ldots = t_1 - t_0$, l'équation (2.39) peut être réécrite*

$$P_n(t) = e^{A_{n,h}\Delta} \cdot e^{A_{n,h-1}\Delta} \cdot \ldots \cdot e^{A_{n,1}\Delta} \cdot e^{A_{n,0}\Delta} \cdot P_n(0). \tag{2.40}$$

L'algorithme de calcul de l'équation (2.40) nécessite de réaliser les produits matriciels de la « droite »vers la « gauche ». Dans la suite, nous considérons que l'intervalle de temps $[t_0, t]$ est échantillonné en intervalles de temps égaux Δ.

2.3.2 Inversion du temps

En exprimant $P_n(0)$ en fonction de $P_n(t)$ dans l'équation (2.40) nous obtenons

$$P_n(0) = e^{-A_{n,0}\Delta} \cdot e^{-A_{n,1}\Delta} \cdot \ldots \cdot e^{-A_{n,h-1}\Delta} \cdot e^{-A_{n,h}\Delta} \cdot P_n(t).$$

En posant $\tilde{P}_n(t) = P_n(0)$ et $\tilde{P}_n(0) = P_n(t)$, l'équation (2.39) peut être réécrite

$$\tilde{P}_n(t) = e^{-A_{n,0}\Delta} \cdot e^{-A_{n,1}\Delta} \cdot \ldots \cdot e^{-A_{n,h-1}\Delta} \cdot e^{-A_{n,h}\Delta} \cdot \tilde{P}_n(0). \tag{2.41}$$

Le remplacement de t par $-t$ et l'inversion du sens des produits matriciels (de la « gauche » vers la « droite ») dans l'équation (2.39) permet d'obtenir les probabilités d'occurrence passées de motifs à partir des probabilités actuelles.

2.4 Modèle stochastique numérique d'évolution inverse de trinucléotides avec blocage dynamique des taux de substitution

2.4.1 Motivation théorique

Les modèles d'évolution stochastiques analytiques ou numériques de trinucléotides développés à ce jour permettent d'étudier l'évolution des séquences génétiques dans le sens direct (passé-présent) ou inverse (présent-passé).

De façon générale, dans le cas de modèles basés sur 1P, 2P et 3P avec des paramètres de substitution constants, l'étude évolutive directe de motifs de taille n ne pose pas de problème, le vecteur des 4^n probabilités d'occurrence obtenu au temps t est toujours valide puisque les 4^n probabilités sont positives et convergent vers la fréquence stationnaire $\frac{1}{4^n}$ lorsque t tend vers l'infini. Par exemple, dans le cas d'une étude évolutive directe de dinucléotides, $n = 2$, leurs $4^2 = 16$ probabilités d'occurrence convergent vers $\frac{1}{16}$.

En revanche, les probabilités d'occurrence divergent dans le cas d'une étude évolutive inverse : les 4^n probabilités d'occurrence tendent chacune vers l'infini (positif ou négatif) du fait de la positivité des termes exponentiels (equations 2.41 et 2.8), comme nous l'avons expliqué dans la section 2.2.4.6. En conséquence, ces modèles sortent du domaine de validité des probabilités.

Ce problème se pose notamment lors de l'étude évolutive inverse des séquences codant les protéines (gènes). Dans ces séquences, l'absence de codons stop TAA, TAG et TGA se traduit par des probabilités nulles pour ces motifs. Les probabilités de ces codons stop peuvent donc devenir très rapidement négatives en « remontant » le temps suivant les paramètres de substitution choisis, conduisant à la non-validité du modèle stochastique.

2.4.2 Principe

Pour conserver un modèle valide quelque soit le sens d'évolution choisi, nous avons développé un modèle stochastique numérique d'évolution inverse de trinucléotides avec blocage dynamique des taux de substitution.

Dans ce modèle, les paramètres de substitution ne sont pas constants tout au long du processus évolutif, mais sont modifiés en fonction des probabilités d'occurrence des motifs. Si la probabilité d'occurrence d'un motif j devient nulle ou négative, le processus de substitution de ce motif j est bloqué. La probabilité d'occurrence de ce motif reste nulle durant tout le reste du processus d'évolution. Ce blocage empêche donc le motif j d'être substitué en un autre motif $i \neq j$ ainsi que tout autre motif $i \neq j$ d'être substitué en le motif j bloqué. La matrice de substitution dynamique associée à ce modèle est donc symétrique tout au long du processus d'évolution. Par conséquent, ce modèle d'évolution permet de simuler l'utilisation d'une matrice de substitution de taille variable au cours du processus d'évolution en fonction des valeurs des probabilités d'occurrence des motifs. Par exemple, dans le cas d'une évolution inverse de codons, les trinucléotides TAA, TAG et TGA ayant une probabilité nulle, leur probabilité d'occurrence est bloquée à 0 et la matrice de substitution (64,64) est corrigée de façon à ne faire évoluer par la suite que les 61 codons restants, simulant ainsi l'utilisation d'une matrice de substitution de taille (61,61) tant qu'aucun autre codon n'a une probabilité d'occurrence négative ou nulle.

2.4.3 Algorithme de blocage dynamique des substitutions

Nous décrivons ici l'algorithme du processus d'évolution de ce modèle pour des motifs de taille $n = 3$ avec trois paramètres de substitution par site (3PS). Les paramètres de substitution ne sont pas exprimés en fonction du temps t mais sont liés aux probabilités d'occurrence des motifs, elles-mêmes fonction du temps t. La méthode de résolution numérique vue dans la section précédente peut donc être appliquée.

Nous supposons que le temps d'évolution t est échantillonné en durées égales $\Delta = t_{h+1} - t_h, \forall\ 0 \leq h$.

Soit A la matrice initiale des taux de substitution constante de trinucléotides et A_{t_h} la matrice des taux de substitution constante de trinucléotides pendant l'intervalle de temps $\Delta = t_{h+1} - t_h, \forall\ 0 \leq h$. A_{t_0} est alors la matrice des taux de substitution constante entre t_0 et t_1.

Pour chaque intervalle de temps Δ entre $t = t_0$ et $t = t_{max}$, le vecteur $\tilde{P}(t)$ des probabilités d'occurrence des 64 trinucléotides est parcouru pour bloquer les éventuelles probabilités nulles ou négatives, puis normalisé. Il n'est pas nécessaire d'effectuer le contrôle des probabilités d'occurrence devenant supérieures à 1 puisque les 64 probabilités d'occurrence somment à 1. Une probabilité d'occurrence ne peut donc être supérieure à 1 sans qu'aucune autre probabilité ne soit négative.

La matrice des taux de substitution A_{t_h} est obtenue à partir de $A_{t_{h-1}}$ par blocage des substitutions associées aux probabilités bloquées dans $\tilde{P}(t)$. Le vecteur des probabilités d'occurrence à l'étape suivante $\tilde{P}(t+\Delta)$ est ensuite calculé selon l'équation (2.41) en effectuant un produit matriciel de « gauche » à « droite » des matrices exponentielles $e^{-A_{t_h} \times \Delta}$ successives et du vecteur des probabilités d'occurrence initiales. Lorsqu'un blocage a lieu, le processus est « réinitialisé » de telle sorte que le vecteur $\tilde{P}(t)$ soit le nouveau vecteur de probabilités d'occurrence initiales pour le calcul de $\tilde{P}(t+\Delta)$ et pour les étapes suivantes, jusqu'à ce qu'un nouveau blocage se produise. Ce processus est répété jusqu'à ce que $t = t_{max}$.

Le processus d'évolution inverse avec blocage dynamique des substitutions est détaillé dans l'algorithme 1.

Algorithme 1 Algorithme de calcul du vecteur des 64 probabilités d'occurrence de trinucléotides entre $t_0 = 0$ et t_{max} dans le sens d'évolution inverse avec blocage dynamique des taux de substitution.

Entrées : $\tilde{P}(0) = \{\tilde{P}_1(0), \ldots, \tilde{P}_{64}(0)\}$ le vecteur des probabilités d'occurrence initiales des 64 trinucléotides,
A la matrice de substitution initiale,
t_{max} le temps t inverse maximal.

Sorties : $\tilde{P}(t)$ le vecteur des probabilités d'occurrence des 64 trinucléotides au temps d'évolution inverse t.

$A(0) \leftarrow A$
$A_{sauv} \leftarrow$ matrice identité (64,64)
$ensembleProbabilitesBloquees \leftarrow \{\}$
$probaInit \leftarrow \tilde{P}(0)$
$t \leftarrow 0$
tant que $t \leq t_{max}$ **faire**
  $sommeProbabilitesBloquees \leftarrow 0$
  $blocageEnCours \leftarrow$ Faux
  pour tout $1 \leq i \leq 64$ **faire**
    si $\tilde{P}_i(t) \leq 0$ et $i \notin ensembleProbabilitesBloquees$ **alors**
      $blocageEnCours \leftarrow$ Vrai
      $sommeProbabilitesBloquees \leftarrow sommeProbabilitesBloquees + \tilde{P}_i(t)$
      ajouter i à $ensembleProbabilitesBloquees$
    fin si
  fin pour
  $\tilde{P}(t) \leftarrow \tilde{P}(t)/(1 - sommeProbabilitesBloquees)$
  $A(t + \Delta) \leftarrow A(t)$
  pour tout $i \in ensembleProbabilitesBloquees$ **faire**
    mettre à 0 tous les paramètres de substitution des ligne i et colonne i de $A(t + \Delta)$
  fin pour
  pour tout $1 \leq i \leq 64$ **faire**
    $A(t + \Delta)[i, i] \leftarrow -\sum_{j=1, j \neq i}^{64} A(t + \Delta)[j, i]$
  fin pour
  si $blocageEnCours == Vrai$ **alors**
    $probaInit \leftarrow \tilde{P}(t)$
    $A_{sauv} \leftarrow$ matrice identité (64,64)
  fin si
  $A_{sauv} \leftarrow A_{sauv} \cdot e^{-A(t+\Delta)\times\Delta}$
  $\tilde{P}(t + \Delta) \leftarrow A_{sauv} \cdot e^{-A(t+\Delta)\times\Delta} \cdot probaInit$
  $t \leftarrow t + \Delta$
fin tant que
retourner $\tilde{P}(t)$

2.5 Conclusion

L'utilisation d'opérations de Kronecker (sommes et produits de matrices) nous a permis de construire des matrices de substitution et de calculer les probabilités d'occurrence analytiques de motifs génétiques de taille quelconque à partir de modèles de substitution de nucléotides de types 1P, 2P et 3P, précédemment étendus aux dinucléotides et trinucléotides. Cette approche nous a également permis d'étendre d'autres modèles de substitution de nucléotides, comme par exemple le modèle 4P à quatre paramètres de substitution dont 3P est un cas particulier. Cette extension des modèles de nucléotides de types 1P, 2P et 3P à des motifs de taille quelconque nous a servi à étendre l'application web SEGM, que nous présentons dans les deux chapitres suivants, à l'étude évolutive de motifs génétiques de taille 5. Elle a par ailleurs fait l'objet d'une publication en revue internationale à comité de lecture en 2011 [BM11].

D'autre part, le développement d'un modèle d'évolution inverse de trinucléotides avec matrice de substitution non constante bloquant les substitutions des motifs dont la probabilité est nulle est un premier outil pouvant servir à l'étude évolutive inverse des CDS en conservant un vecteur de probabilités d'occurrence des motifs valide. Nous proposons dans le chapitre 5 une application de ce modèle à l'évolution inverse des trinucléotides du code circulaire X(EUK_PRO) présenté dans le chapitre 1.

Chapitre 3

Structures de données et algorithmes d'implantation des formules analytiques des modèles d'évolution des gènes

La version actuelle du logiciel de recherche SEGM (Stochastic Evolution of Genetic Motifs), dont le principe et les fonctionnalités seront décrites dans le chapitre suivant, est une extension d'une première version développée durant la première partie de cette thèse. La première version unifiait les modèles analytiques de substitution de nucléotides 1P, 2P et 3P présentés dans le chapitre 2 étendus aux dinucléotides et trinucléotides [Mic07c, Mic07a] avec les modèles respectifs 1PS, 2PS et 3PS. L'extension théorique des modèles de substitution de nucléotides de type 1P, 2P et 3P à des motifs de taille quelconque présentée dans le chapitre 2 nous a permis développer une seconde version de SEGM, étendue au calcul des probabilités d'occurrence analytiques de motifs de taille 4 et 5 (tetranucléotides et pentanucléotides). Pour réaliser cette extension, l'approche classique pour le calcul des solutions analytiques utilisée dans la première version du SEGM ne pouvait plus être appliquée au calcul des solutions analytiques pour des tetranucléotides et pentanucléotides. Nous en expliquons les raisons après une description de l'approche classique utilisée dans la première version de SEGM. Nous proposons ensuite deux autres méthodes de calcul des solutions analytiques envisagées, dont celle retenue et implantée dans le SEGM. Nous expliquons le choix de la méthode retenue par une étude des temps de calcul et de l'espace mémoire utilisé pour les deux méthodes. Dans toute la suite de ce chapitre nous ne considérons que le modèle de substitution 3PS généralisé aux motifs de différentes tailles, les modèles 1PS et 2PS étant des cas particuliers de 3PS.

3.1 Implémentation avec Mathematica et webMathematica

Le logiciel de recherche SEGM a été programmée en JSP (Java Sever Page) et Mathematica à partir du logiciel de calcul symbolique Mathematica (version 8), de son extension webMathematica (version 3) ainsi que de l'environnement de développement Mathematica Workbench. SEGM est actuellement hébergé sur un serveur Tomcat (version 7).

3.2 Principales contraintes

3.2.1 Taille des formules analytiques

Comme énoncé dans le chapitre 2 présentant les modèles biomathématiques d'évolution développés au cours de cette thèse, les solutions analytiques des probabilités d'occurrence de motifs génétiques de taille n obtenues avec le modèle 3PS contiennent 4^{2n} termes. Une première version de SEGM publiée en 2009 [BM09b] proposait les solutions analytiques pour des nucléotides, dinucléotides et trinucléotides. Les formules analytiques calculées par SEGM étaient alors composées pour les trinucléotides de $4^6 = 4096$ termes. La version actuelle de SEGM propose des solutions analytiques pour des motifs de taille 1 à 5, ce qui nécessite un calcul de formules comportant 4^{10} termes pour les pentanucléotides, soit environ un million de termes. Ce changement d'échelle du nombre de termes nécessaires au calcul des solutions analytiques impose une stratégie de calcul et une implantation des formules efficaces.

3.2.2 Fonctionnement sur un PC standard

La version actuelle du SEGM web server a été développée pour fonctionner sur un PC standard avec processeur Intel Core 2 duo à 3,16 GHz et 8 Go de mémoire vive. Ce PC est également utilisé comme serveur pour d'autres applications web. Les ressources en temps CPU et en mémoire sont donc limitées à celles d'un ordinateur standard actuel grand public.

3.2.3 Contrainte liée à l'unification des modèles analytiques

Le développement de SEGM a pour but de mettre à disposition en ligne une unification des modèles de substitution analytiques de type 1PS, 2PS et 3PS pour des motifs génétiques de taille 1 à 5. Nous utilisons donc la même stratégie de calcul pour toutes les tailles de motifs.

3.3 Description des différentes approches envisagées

3.3.1 Approche classique

La première version de SEGM proposait le calcul des solutions analytiques des probabilités d'occurrence de motifs de taille $n = 1$ à $n = 3$ (nucléotides, dinucléotides et trinucléotides) dans les deux sens d'évolution (direct et inverse).

L'approche classique utilisée consistait à calculer le vecteur des probabilités d'occurrence au temps t des 4^n motifs de taille n. Ce vecteur était obtenu par produits matriciels correspondant à l'équation (2.6) vue au chapitre 2 que nous rappelons ici :

$$P_n(t) = Q_n \cdot e^{D_n t} \cdot Q_n^{-1} \cdot P_n(0)$$

où Q_n, D_n, Q_n^{-1} et $P_n(0)$ sont respectivement la matrice des vecteurs propres, la matrice diagonale des valeurs propres, la matrice inverse des vecteurs propres associées à la matrice des taux substitution instantanée $A_n(4^n, 4^n)$ et le vecteur des probabilités d'occurrence initiales des 4^n motifs de taille n.

Cette méthode nécessitait donc la connaissance au préalable des matrices Q_n, D_n et Q_n^{-1}. Ces matrices étant de taille $(64, 64)$ pour le cas des trinucléotides et donc composées au maximum de 4096 termes « simples », elles étaient directement implantées dans le fichier source jsp. Ces matrices peuvent être calculées à partir de la factorisation par sous-matrices blocs de Tian et Styan exposée dans la sous-section 2.2.3.2 du chapitre 2. La matrice spectrale D_n et la matrice exponentielle $e^{D_n t}$ étant des matrices diagonales, elles étaient générées à la volée à partir d'une liste des 4^n valeurs propres de A_n.

Le vecteur des probabilités d'occurrence initiales $P_n(0)$ des 4^n motifs était enregistré dans un tableau pendant le calcul des solutions analytiques.

L'avantage principal de cette méthode est qu'un seul calcul est nécessaire pour obtenir les solutions analytiques des probabilités d'occurrence des 4^n motifs. La première version de SEGM tirait d'ailleurs partie de cet avantage en affichant les solutions analytiques des 4^n motifs. Cependant, cette méthode s'avère trop coûteuse en temps et en espace pour un passage à l'échelle des pentanucléotides : l'affichage des solutions analytiques des 64 trinucléotides nécessite le traitement de $64^3 = 262144$ termes alors que l'affichage des 1024 pentanucléotides nécessite le traitement de $1024^3 = 1073741824$ termes, soit plus d'un milliard de termes.

3.3.2 Méthode formelle « pure »

Une première méthode de calcul des solutions analytiques envisagée pour la version actuelle de SEGM reposait sur l'équation (2.26) donnant la solution analytique pour un motif donné présentée dans la section 2.2.4.5 du chapitre 2. En effet, les calculs étant effectués par un noyau Mathematica et ce logiciel autorisant le calcul symbolique, il était possible d'implanter l'équation (2.26) pour obtenir la solution analytique pour un motif donné. Dans la suite, nous appellerons cette méthode *pure* car elle peut être entrée sous sa forme close avec uniquement des instructions Mathematica. Pour cette raison, nous ne donnons pas d'algorithme d'implantation de cette méthode. Nous verrons par la suite les avantages et inconvénients de cette méthode.

3.3.3 Méthode formelle « mixte »

La seconde méthode envisagée, appelée *mixte* dans la suite, calcule également la solution analytique pour un motif donné. Contrairement à la méthode *pure*, elle repose sur la création et l'utilisation de fonctions annexes, c'est-à-dire des fonctions qui ne sont pas des instructions Mathematica.

Comme nous l'avons vu dans le chapitre 2, la solution analytique de la probabilité d'occurrence d'un motif i_1 de taille n peut être calculée à partir de l'équation 2.21 que nous rappelons ici :

$$P_{i_1}(t) = \left(\otimes_{k=1}^{n}(R \cdot e^{S_k t} \cdot R^{-1})\right)[i_1,] \cdot P_n(0),$$

où R et R^{-1} sont les matrices et matrices inverses $(4, 4)$ des vecteurs propres associées aux matrices de substitution instantanées N_k associées aux sites k de motifs et S_k les matrices diagonales des valeurs propres associées aux matrices N_k.

Dans la suite, nous appellerons matrice de transition totale la matrice $\otimes_{k=1}^{n}(R \cdot e^{S_k t} \cdot R^{-1})$ de taille $(4^n, 4^n)$ et matrices de transition partielles les matrices de transition $R \cdot e^{S_k t} \cdot R^{-1}$ de

taille $(4, 4)$ associées aux sites k des motifs.

Nous présentons la méthode *mixte* à l'aide de fonctions décrites comme des algorithmes. La fonction principale calculant la solution analytique de la probabilité d'occurrence d'un motif i_1 est donnée dans l'algorithme 2. Elle repose sur le calcul de la ligne $LigneMat(n, i_1)$, donnée dans la fonction 3, cette ligne correspondant au motif i_1 de taille n dans la matrice de transition totale. Le calcul des matrices de transition partielles associées aux sites k est effectué par la fonction $RSRInv(k)$ décrite dans la fonction 4. Les éléments de la ligne de la matrice de transition totale sont déterminés par une fonction $Coeff(k, i, j)$ détaillée dans la fonction 5 calculant l'élément en ligne i et colonne j de la matrice de transition totale en simulant un produit de Kronecker des matrices de transition $RSRInv(k)$.

Algorithme 2 Algorithme de calcul de la solution analytique du motif i_1 de taille n avec la méthode *mixte*.

Entrées : $P(t_0)$, vecteur des probabilités d'occurrence initiales des 4^n motifs.
Sorties : $P_{i_1}(t)$, solution analytique de la probabilité d'occurrence du motif i_1 au temps t.

$P_{i_1}(t) = LigneMat(n, i_1) \cdot P(t_0)$
Factorisation de $P_{i_1}(t)$ par les exponentielles {Fonction `Collect[]` de Mathematica}
Simplification des exponentielles {Fonction `Simplify[]` de Mathematica}
retourner $P_{i_1}(t)$

Fonction 3 $LigneMat(n, i_1)$: calcul de la ligne correspondant au motif i_1 de la matrice de transition totale.

Entrées : i_1, index du motif dont on veut calculer la probabilité d'occurrence analytique,
n, taille des motifs.
Sorties : $LigneMat(n, i_1) = tab$, vecteur ligne des coefficients $Coeff(n, i_1, j)$ associés à chaque probabilité initiale $P_j(0)$.

$tab = [0, \ldots, 0]$ {tableau de 4^n éléments initialisés à 0}
pour tout $1 \leq j \leq 4^n$ **faire**
  $tab[j] = Coeff(n, i_1, j)$
fin pour
retourner tab

Fonction 4 $RSRInv(k)$: calcul de la matrice de transition partielle $RSRInv(k)$ associée au site k.

Entrées : k, numéro du site de motif.
Sorties : $RSRInv(k)$, matrice de transition partielle pour le site k d'un motif.

$RSRInv(k) = R \cdot e^{S_k t} \cdot R^{-1}$
retourner $RSRInv(k)$

Fonction 5 $Coeff(n, i, j)$: calcul de l'élément j de la ligne correspondant au motif i de la matrice de transition totale.

Entrées : n, taille des motifs,
i et j, coordonnées du coefficient à calculer.
Sorties : $res = Coeff(n, i, j)$

$res = 1$ {Coefficient résultant initialisé à 1}
pour tout $1 \leq k \leq n$ **faire**
$res = res \times RSRInv(k)[\delta(i, k), \delta(j, k)]$ {avec $\delta(x, k) = \lfloor \frac{x-1}{4^n - k} \rfloor [4] + 1$}
fin pour
retourner res

3.4 Comparaison des méthodes de calcul « pure » et « mixte »

La méthode *pure* consistant à implanter l'équation (2.26) calcule la solution analytique pour un motif donné en tirant profit de la gestion automatique des structures de données et de l'utilisation de la mémoire par Mathematica. Nous comparons ici le temps de calcul et l'espace mémoire utilisée avec cette dernière avec la méthode algorithmique dite *mixte* basée sur une simulation du produit de Kronecker des matrices de transition partielles à l'aide de fonctions annexes. Les tests ont été effectués sur le PC hébergeant l'application web SEGM. Seule l'étape de calcul est prise en compte pour les tests.

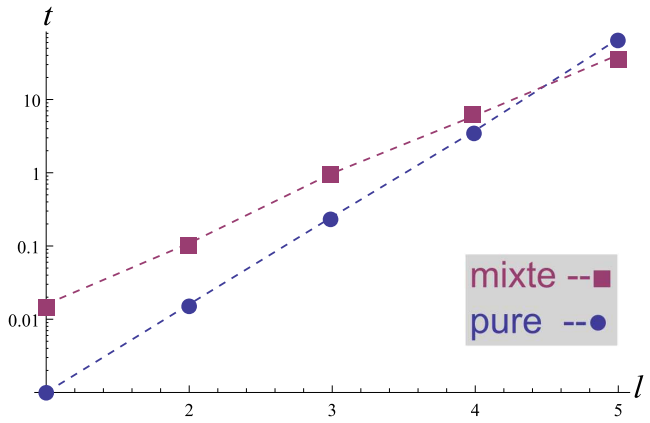
La figure 3.1 donne les temps de calcul de chacune des deux méthodes *pure* et *mixte* pour toutes les tailles de motifs gérées par SEGM.

Pour des motifs de taille 1 à 4, la méthode *pure* est plus rapide que la méthode *mixte*. Le ratio des temps de calcul $\frac{pure}{mixte}$ est de 0.06 pour $l = 1$, et monte à 0.63 pour $l = 4$. Cependant, pour des motifs de taille $l = 5$, le rapport est inversé et la méthode *mixte* est plus rapide avec un ratio $\frac{mixte}{pure}$ de 0.60.

Ce comportement peut être analysé avec les résultats de l'utilisation de la mémoire présentés dans la figure 3.2. Comme attendu, la méthode *pure* s'avère beaucoup plus efficace que la méthode *mixte* en terme d'espace mémoire utilisé, et ce, quelle que soit la taille de motif choisie. L'implantation de l'équation (2.26) telle quelle dans Mathematica sans l'utilisation de fonctions annexes permet de tirer parti de l'utilisation de structures de données optimales par Mathematica.

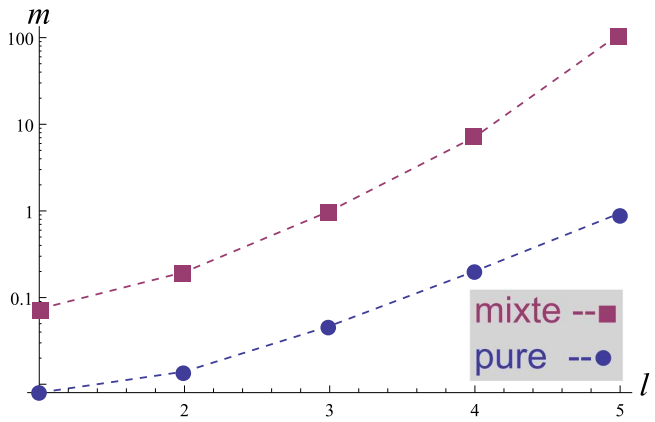
La table 3.1 donne les ratios des temps de calcul d'une solution analytique pour un motif de taille $l - 1$ et pour un motif de taille l, $2 \leq l \leq 5$ avec les deux méthodes *pure* et *mixte*.

Les solutions analytiques des probabilités d'occurrence étant composées de 4^{2n} termes, le passage d'une taille n de motif à $n + 1$ multiplie par $4^2 = 16$ le nombre de termes de chaque solution analytique. Les ratios des temps de calcul obtenus avec la méthode *pure* pour l allant de 2 à 5 suivent cette augmentation du nombre de termes, avec une légère augmentation de $l = 4$ à $l = 5$ (ratio proche de 18). L'espace mémoire utilisé avec la méthode *pure* pour $l = 5$ est de 0.943 Mo contre 0.202 Mo pour $l = 4$. Les ratios des temps de calcul obtenus avec la



	temps (s)	
taille	*mixte*	*pure*
1	0.016	0.001
2	0.109	0.016
3	0.984	0.25
4	6.047	3.781
5	40.703	67.297

FIGURE 3.1 – Courbes de comparaison et temps de calcul de la solution analytique pour un motif donné de taille $l = 1$ à $l = 5$ (en abscisse), avec la méthode *mixte* et la méthode *pure*. L'échelle de temps t (en secondes) en ordonnée sur le graphique est logarithmique. L'abscisse représente la taille des motifs.



	espace (Mo)	
taille	*mixte*	*pure*
1	0.074	0.008
2	0.197	0.014
3	0.994	0.047
4	7.159	0.202
5	111.302	0.943

FIGURE 3.2 – Courbes de comparaison et espace mémoire utilisé pour le calcul d'une solution analytique pour un motif donné de taille $l = 1$ à $l = 5$ (en abscisse), avec la méthode *mixte* et la méthode *pure*. L'échelle de l'espace mémoire utilisé m (en Mo) en ordonnée sur le graphique est logarithmique. L'abscisse représente la taille des motifs.

	ratio $\frac{l}{l-1}$	
taille l	*pure*	*mixte*
2	16.0	6.8
3	15.6	9.0
4	15.1	6.1
5	17.8	6.7

TABLE 3.1 – Ratios des temps de calculs d'une solution analytique pour un motif de taille $l-1$ et pour un motif de taille l avec les méthodes *pure* et *mixte*.

méthode *mixte* pour l allant de 2 à 5 sont compris entre 6.14 et 6.81, sauf dans le cas du passage de $l = 2$ à $l = 3$ où le ratio est de 9.0. Ceci pourrait être expliqué par l'espace mémoire utilisé. En effet, avec la méthode *mixte* pour $l = 3$, 0.994 Mo sont utilisés contre 0.109 Mo pour $l = 2$. Ces résultats suggèrent donc que la gestion de la mémoire par Mathematica diffère lorsque l'utilisation de la mémoire passe de 0.2 Mo à 0.9 Mo environ pour ces calculs. Par ailleurs, le faible ratio obtenu avec la méthode *mixte* et l'utilisation massive de la mémoire relativement à la méthode *pure* pourrait s'expliquer par l'utilisation des fonctions annexes stockant en mémoire certaines données utilisées à plusieurs reprises.

3.4.1 Choix de la méthode utilisée

La méthode *pure* est donc la plus rapide et la plus efficace pour des tailles de motif allant de 1 à 4. En valeur absolue, le gain de temps maximum pour cette méthode est obtenu pour $l = 4$ avec 2,3 secondes. Les gains de temps pour des motifs de taille inférieure sont inférieurs à 1 seconde et sont donc négligeables pour l'application web.

En revanche, le gain de temps de la méthode *mixte* pour des motifs de taille $l = 5$ est, en valeur absolue, d'environ 27 secondes. De plus, l'espace mémoire utilisé dans ce cas étant d'une centaine de Mo, cette méthode permet un gain de temps conséquent par rapport à la méthode *pure* tout en utilisant une partie très limitée de la mémoire disponible. Pour ces raisons, nous avons choisi d'implémenter les modèles de substitution de motifs de taille 1 à 5 dans SEGM avec la méthode *mixte*.

3.5 Conclusion

La méthode classique de calcul des solutions analytiques des probabilités d'occurrence des motifs génétiques de taille 1 à 3 utilisée dans la première version de SEGM que nous avions développée ne pouvait pas être adaptée au calcul des solutions analytiques pour des motifs de taille supérieure. Cependant, la méthode utilisant des opérations de Kronecker (sommes et produits de matrices) vue dans le chapitre 2 nous a permis d'aborder le calcul des probabilités d'occurrence de motifs de taille 4 et 5 avec Mathematica selon deux approches dites *pure* (implantation directe de l'équation (2.26) donnant la solution analytique pour un motif donné) et *mixte* (méthode algorithmique). Bien que la méthode *pure* soit plus rapide et moins coûteuse en mémoire pour le calcul des probabilités d'occurrence de motifs de taille 1 à 4, nous avons implanté la méthode *mixte* dans la version actuelle de SEGM pour le gain de temps de calcul obtenu par rapport à

la méthode *pure* pour les motifs de taille 5.

Chapitre 4

Principe et fonctionnalités du logiciel de recherche SEGM

L'application web SEGM (Stochastic Evolution of Genetic Motifs) disponible à l'adresse

`http://lsiit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/Accueil.jsp`

est un outil mis à la disposition des communautés biologique et bioinformatique pour l'étude évolutive de séquences génétiques à partir de solutions analytiques de probabilités d'occurrence de motifs génétiques de taille 1 à 5. Une première version de SEGM a été mise en ligne en 2008 pour un calcul des solutions analytiques pour des motifs génétiques de taille 1 à 3. Nous en proposons une description rapide dans une première section. La seconde version étend le calcul aux motifs génétiques de taille 4 et 5 avec de nouvelles fonctionnalités. Le principe de fonctionnement et les différentes fonctionnalités de la version actuelle de SEGM sont présentés dans une seconde section.

4.1 Développement d'une première version de SEGM

Suite aux travaux de Michel [Mic07c, Mic07a] et Frey et Michel [FM06] sur la généralisation aux dinucléotides et trinucléotides des modèles stochastiques de substitution de nucléotides de type 1PS, 2PS et 3PS définis au chapitre 1, la première partie de cette thèse a été consacrée au développement d'une application web unifiant ces modèles stochastiques de substitution de nucléotides, dinucléotides et trinucléotides.

Cette application web avait pour objectif de déterminer, pour des motifs de taille $n = 1$ à $n = 3$, les solutions analytiques des probabilités d'occurrence au temps t des 4^n motifs à partir d'un vecteur de probabilités d'occurrence initiales et de paramètres de substitution. Les solutions analytiques pouvaient être calculées dans le sens d'évolution direct (passé-présent) ou inverse (présent-passé).

Dans cette première version, le choix de la taille des motifs à étudier était proposé sur la page d'accueil. L'utilisateur était ensuite dirigé vers la page de SEGM spécifique à la taille de motifs choisie où il devait entrer manuellement le vecteur des probabilités d'occurrence initiales et les paramètres de substitution. Après calcul, les solutions analytiques des 4^n motifs de taille n

étaient affichées, avec les courbes d'évolution quand les conditions étaient vérifiées. Les conditions d'affichage des courbes d'évolution étaient similaires à celles décrites dans la section 4.2.

4.2 Seconde version de SEGM

Le développement de la version actuelle de SEGM étendue aux tétranucléotides et aux pentanucléotides a nécessité une refonte complète de l'application de façon à la rendre plus ergonomique et plus rapide pour effectuer le calcul des solutions analytiques.

4.2.1 Taille des motifs et probabilités d'occurrence initiales

La page d'accueil de la version actuelle du logiciel SEGM propose, à l'aide d'un menu déroulant, le choix entre les cinq tailles n de motifs possibles (figure 4.1).

Après ce choix et contrairement à la première version de SEGM, l'utilisateur est invité à charger un fichier contenant les probabilités d'occurrence initiales des 4^n motifs. Ce changement de stratégie s'explique par le nombre important de probabilités d'occurrence initiales à entrer dans le cas des tétranucléotides (256) et pentanucléotides (1024). La procédure d'entrée des 4^n probabilités d'occurrence initiales est donc automatisée par l'utilisation d'un fichier de type tableur au format xls. Suivant la taille des motifs choisis, un exemple de fichier avec 4^n probabilités d'occurrence initiales est proposé en exemple à partir d'un lien (figure 4.1). L'utilisation d'un fichier de type tableur permet ainsi une gestion automatisée et beaucoup plus rapide des probabilités d'occurrence initiales (figure 4.2). Le vecteur des probabilités d'occurrence initiales doit être trié dans l'ordre lexicographique des motifs.

Le vecteur des 4^n probabilités d'occurrence initiales chargé par l'utilisateur est ensuite vérifié et analysé par le serveur. Le vecteur doit contenir 4^n probabilités initiales positives ou nulles et doit sommer à 1. Si les conditions ne sont pas respectées, des messages d'erreurs explicatifs sont affichés et l'utilisateur est invité à recommencer le processus d'acquisition des données à partir de la page d'accueil. Si les conditions sont respectées, l'utilisateur est dirigé sur la page spécifique au modèle d'évolution pour la taille de motif choisie.

4.2.2 Options de base

Le calcul des solutions analytiques peut être configuré à l'aide de plusieurs options que nous décrivons ici.

4.2.2.1 Sens du temps d'évolution

La première option (figure 4.3) permet de choisir le sens d'évolution des motifs : direct (passé-présent) ou inverse (présent-passé). Par défaut, le sens direct est choisi.

4.2.2.2 Modèle de substitution

La détermination du nombre de paramètres de substitution par site de motif dans l'option 2 (figure 4.3) correspond au choix du modèle de substitution à appliquer aux motifs. Le choix par défaut est de trois paramètres par site de motif, distinguant les transitions et les deux types de transversion pour chaque site, ce qui correspond au modèle 3PS, le plus général du SEGM. Les

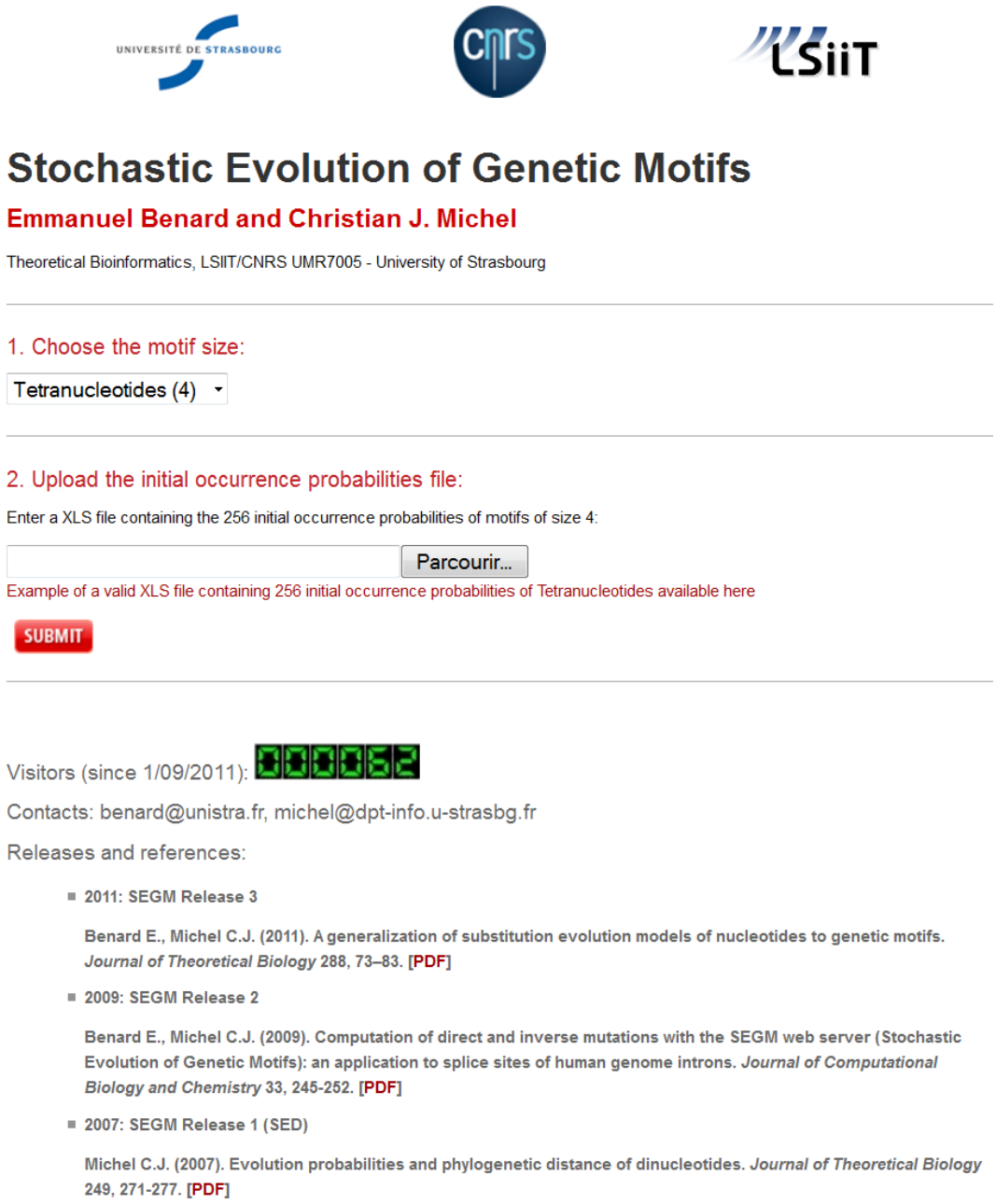
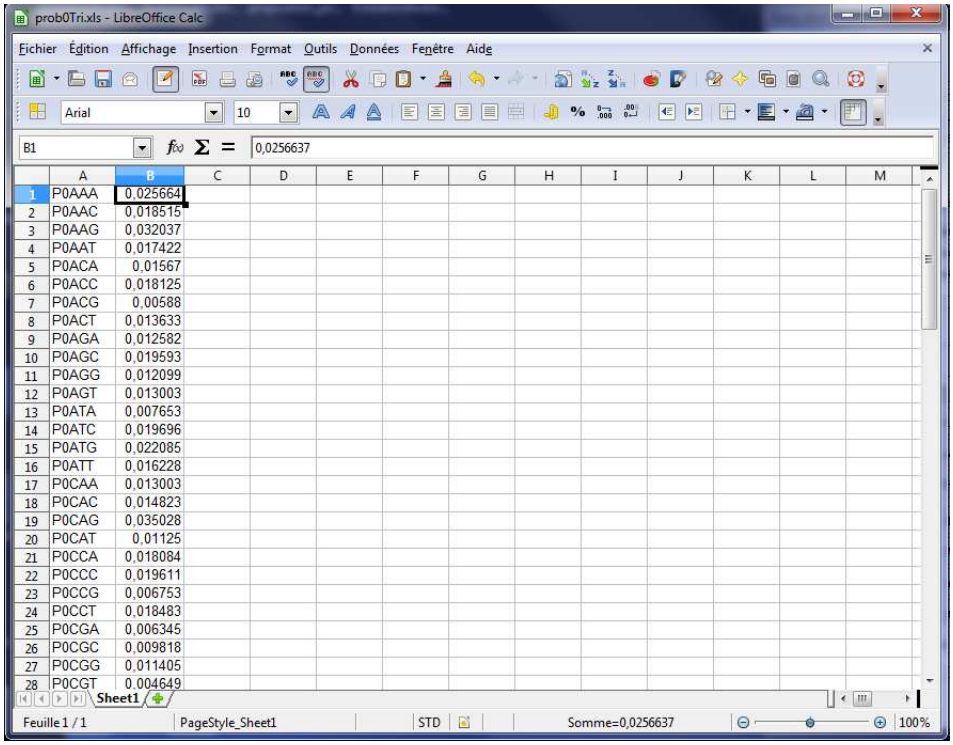


FIGURE 4.1 – Page d'accueil de la version actuelle de SEGM. Le choix de la taille des motifs et le chargement du vecteur des probabilités d'occurrence initiales sont effectués à partir de cette page.

choix correspondant à deux et un paramètre de substitution par site sont associés respectivement aux modèles 2PS et 1PS.



	A	B
1	P0AAA	0,025664
2	P0AAC	0,018515
3	P0AAG	0,032037
4	P0AAT	0,017422
5	P0ACA	0,01567
6	P0ACC	0,018125
7	P0ACG	0,00588
8	P0ACT	0,013633
9	P0AGA	0,012582
10	P0AGC	0,019593
11	P0AGG	0,012099
12	P0AGT	0,013003
13	P0ATA	0,007653
14	P0ATC	0,019696
15	P0ATG	0,022085
16	P0ATT	0,016228
17	P0CAA	0,013003
18	P0CAC	0,014823
19	P0CAG	0,035028
20	P0CAT	0,01125
21	P0CCA	0,018084
22	P0CCC	0,019611
23	P0CCG	0,006753
24	P0CCT	0,018483
25	P0CGA	0,006345
26	P0CGC	0,009818
27	P0CGG	0,011405
28	P0CGT	0,004649

FIGURE 4.2 – Exemple de fichier xls contenant un vecteur de probabilités d'occurrence initiales de trinucléotides. Les probabilités d'occurrence initiales sont triées selon l'ordre lexicographique des trinucléotides.

4.2.2.3 Paramètres de substitution

Les paramètres de susbtitution correspondant au modèle choisi dans l'option 2 peuvent être fixés avec des valeurs numériques dans l'option 3 (figure 4.4). Par défaut, les paramètres sont formels.

Quelque soit le nombre de paramètres de substitution par site choisi, la somme de tous les paramètres doit être inférieure ou égale à 1. Aucun paramètre ne doit être négatif. Une vérification est effectuée sur l'ensemble des paramètres avant le calcul des solutions analytiques. En cas de non respect des conditions sur les paramètres de substitution, des messages d'erreur explicatifs sont affichés et l'utilisateur est invité à corriger ces paramètres.

Remarque 4.1 *Pour une simplicité d'implémentation, la numérotation des sites de motifs de taille n dans SEGM est réalisée à partir de* 0*. Les sites sont donc numérotés de* 0 *à* $n-1$*.*

4.2.2.4 Motifs soumis à évolution

Cette version de SEGM propose de calculer simultanément les probabilités d'occurrence analytique d'au plus quatre motifs parmi les 4^n motifs possibles. Le choix de ces motifs se fait dans l'option 4 (figure 4.4) à l'aide de menus déroulants qui donnent la liste des motifs possibles en fonction de la taille préalablement choisie. Par défaut, un seul motif est choisi. Ce motif est le premier des 4^n motifs dans l'ordre lexicographique, c'est-à-dire A^n, composé de n A consécutifs.

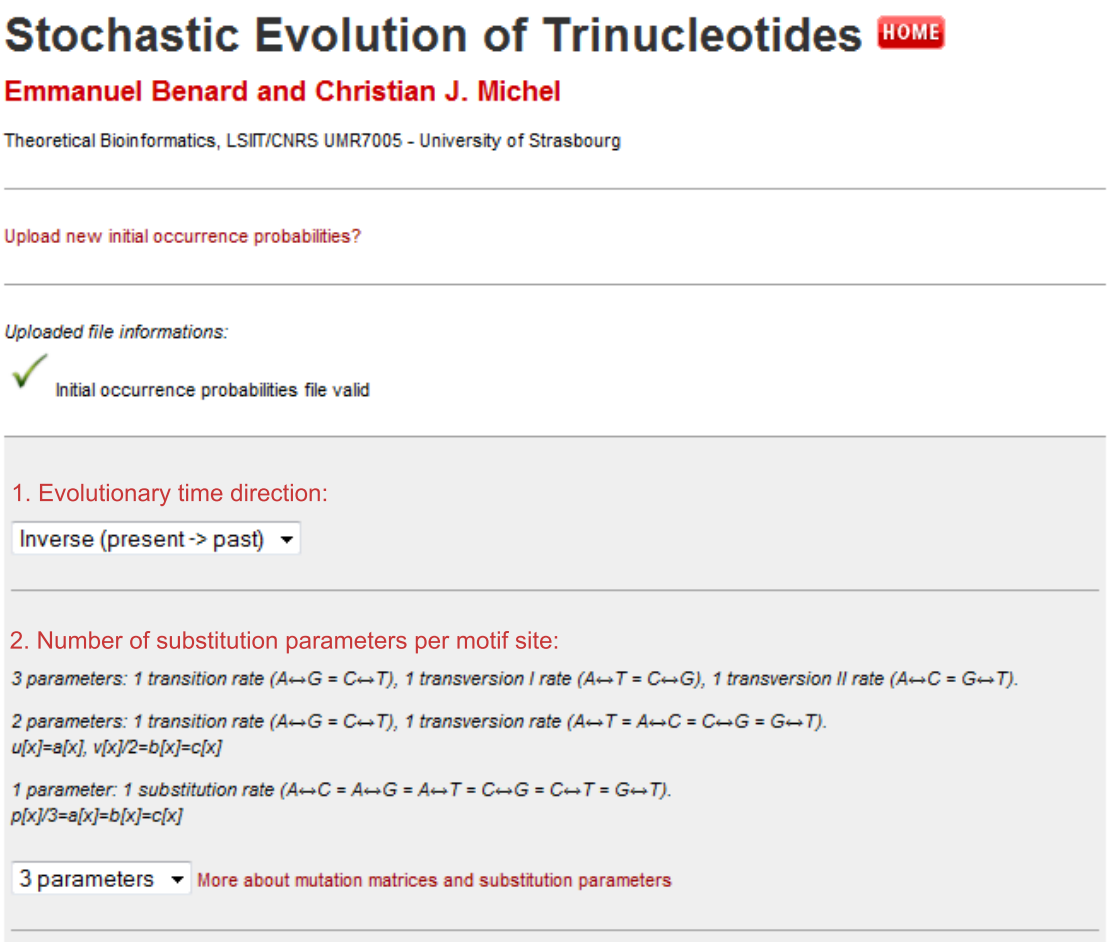


FIGURE 4.3 – Choix du sens d'évolution (option 1) et du nombre de paramètres de substitution par site (option 2).

4.2.2.5 Format de sortie des solutions analytiques

Les solutions analytiques calculées avec la première version du SEGM étaient affichées sur la page web sous forme d'images de façon à les rendre plus lisibles que sous forme textuelle (supprimant la mise en forme). Le format de sortie utilisé par défaut est le format textuel standard, mais difficilement réutilisable dans d'autres applications. Trois autres formats sont également disponibles avec la nouvelle version de SEGM : C, Fortran et TEX (figure 4.4). L'utilisation de ces formats rend donc possible l'intégration et l'utilisation des solutions analytiques dans d'autres programmes utilisateurs.

4.2.3 Options conditionnelles

Lorsque tous les paramètres de substitution sont numériques, d'autres options pour l'affichage de courbes d'évolution et de solutions numériques sont proposées.

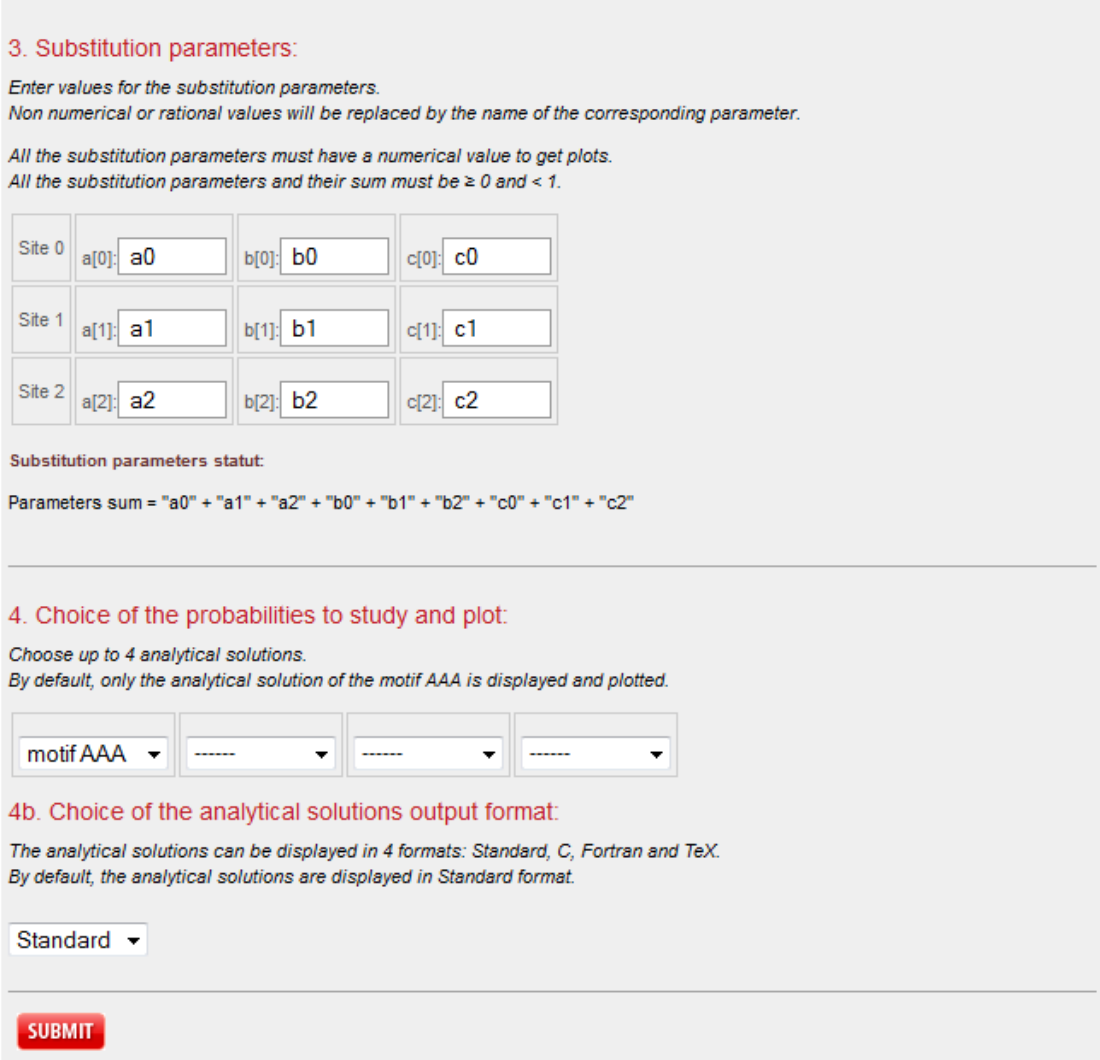


FIGURE 4.4 – Paramètres de substitution pour chaque site (option 3) et choix des motifs à étudier (option 4) et du format de sortie des solutions analytiques (option 4b).

4.2.3.1 Intervalle de temps

L'option 5 (figure 4.5) permet de configurer l'intervalle de temps pour les courbes d'évolution. Quelque soit le sens d'évolution choisi, direct ou inverse, l'intervalle doit être entré en valeurs absolues. Ainsi, pour une évolution inverse, l'utilisateur doit entrer les valeurs $\tilde{tmin} = -tmin$ et $\tilde{tmax} = -tmax$ d'après l'équation (2.8) d'inversion du temps d'évolution. Par exemple, dans le cas d'une évolution inverse entre $tmin = 0$ et $tmax = -3$, l'intervalle est défini par $\tilde{tmin} = 0$ et $\tilde{tmax} = 3$. Lorsque l'intervalle défini par l'utilisateur n'est pas valide, c'est-à-dire $tmin, tmax < 0$ ou $tmin \geq tmax$, l'intervalle par défaut est utilisé ($tmin = 0$ et $tmax = 5$).

4.2.3.2 Valeur de t pour les solutions numériques

Des solutions numériques peuvent être obtenues en fixant t dans l'option 6 (figure 4.5). Comme pour l'intervalle de temps, t doit être positif quelque soit le sens d'évolution choisi. Lorsque l'utilisateur entre une valeur de t non valide, t est laissé formel (valeur par défaut).

3. Substitution parameters:

Enter values for the substitution parameters.
Non numerical or rational values will be replaced by the name of the corresponding parameter.

All the substitution parameters must have a numerical value to get plots.
All the substitution parameters and their sum must be ≥ 0 and < 1.

Site 0	a[0]: 0.2	b[0]: 0.025	c[0]: 0.025
Site 1	a[1]: 0.15	b[1]: 0.25	c[1]: 0.25
Site 2	a[2]: 0.05	b[2]: 0.025	c[2]: 0.025

Substitution parameters statut:

Parameters sum = 1.

4. Choice of the probabilities to study and plot:

Choose up to 4 analytical solutions.
By default, only the analytical solution of the motif AAA is displayed and plotted.

motif AAA | ------ | ------ | ------

4b. Choice of the analytical solutions output format:

The analytical solutions can be displayed in 4 formats: Standard, C, Fortran and TeX.
By default, the analytical solutions are displayed in Standard format.

Standard

5. Time interval for plots (optional):

0 ≤ t min < t max

t min: 0 | t max: 5

6. Time value for numerical solutions (optional):

0 ≤ t

t: t

SUBMIT

FIGURE 4.5 – Options conditionnelles pour l'intervalle de temps des courbes d'évolution (option 5) et valeur du temps d'évolution t pour les solutions numériques (option 6), disponibles lorsque tous les paramètres de substitution (option 3) sont numériques.

4.2.4 Résultats

Les résultats obtenus avec SEGM sont disponibles sous plusieurs formes.

4.2.4.1 Solutions analytiques

Quelque soit la configuration choisie, les solutions analytiques des motifs choisis sont toujours affichées en fonction du temps d'évolution t formel dans la partie résultats (figure 4.6).

Results

Analytical solutions (Standard format):

ProbtATG(t)	$0.015625 + 0.000540842/E^{2*(a0 + b0)*t} - 0.00152927/E^{2*(a1 + b1)*t} - 0.0000528894$ $/E^{2*(a0 + a1 + b0 + b1)*t} + 0.000469227/E^{2*(a0 + a2 + b0 + b2)*t} + 0.00120317/E^{2*(a1 + a2 + b1 + b2)*t} + 0.00125916$ $/E^{2*(a0 + a1 + b1 + c0)*t} - 0.00117735/E^{2*(a1 + b0 + b1 + c0)*t} - 0.000459652/E^{2*(a0 + a2 + b2 + c0)*t} - 0.000696165$ $/E^{2*(a2 + b0 + b2 + c0)*t} + 0.00100502/E^{2*(a0 + a1 + b0 + c1)*t} + 0.000166589/E^{2*(a0 + b0 + b1 + c1)*t} + 0.00231898$ $/E^{2*(a1 + a2 + b2 + c1)*t} - 0.000283983/E^{2*(a2 + b1 + b2 + c1)*t} - 0.000835963/E^{2*(a0 + a1 + c0 + c1)*t} + 0.000572985$ $/E^{2*(a1 + b0 + c0 + c1)*t} + 0.000552281/E^{2*(a0 + b1 + c0 + c1)*t} - 0.00263804/E^{2*(b0 + b1 + c0 + c1)*t} + 0.000495397$ $/E^{2*(a0 + a2 + b0 + c2)*t} + 0.00157556/E^{2*(a1 + a2 + b1 + c2)*t} + 0.00121722/E^{2*(a0 + b0 + b2 + c2)*t} + 0.00013649$ $/E^{2*(a1 + b1 + b2 + c2)*t} - 0.000945947/E^{2*(a0 + a2 + c0 + c2)*t} - 0.000162183/E^{2*(a2 + b0 + c0 + c2)*t} - 0.00164162$ $/E^{2*(a0 + b2 + c0 + c2)*t} + 0.000777826/E^{2*(b0 + b2 + c0 + c2)*t} + 0.00176303/E^{2*(a1 + a2 + c1 + c2)*t} - 0.000420078$ $/E^{2*(a2 + b1 + c1 + c2)*t} + 0.00184492/E^{2*(a1 + b2 + c1 + c2)*t} - 0.00137639$ $/E^{2*(b1 + b2 + c1 + c2)*t} + (0.00036714 - 0.000178117/E^{2*(a0 + a1 + b0 + b1)*t})/E^{2*(a2 + b2)*t} + (0.0026192 - 0.000264094$ $/E^{2*(a1 + a2 + b1 + b2)*t})/E^{2*(b0 + c0)*t} + (-0.00191735 + 0.0000447409/E^{2*(a1 + a2 + b1 + b2)*t})$ $/E^{2*(a0 + c0)*t} + (-0.0000772344 + 0.0000866094/E^{2*(a0 + a2 + b0 + b2)*t} - 0.000258935$ $/E^{2*(a0 + a2 + b2 + c0)*t} + 0.000361008/E^{2*(a2 + b0 + b2 + c0)*t})/E^{2*(b1 + c1)*t} + (0.00230261 + 0.000637553$ $/E^{2*(a0 + a2 + b0 + b2)*t} - 0.00112972/E^{2*(a0 + a2 + b2 + c0)*t} + 0.00052303/E^{2*(a2 + b0 + b2 + c0)*t})$ $/E^{2*(a1 + c1)*t} + (-0.000512409 - 0.000698465/E^{2*(a0 + a1 + b0 + b1)*t} - 0.00038629/E^{2*(a0 + a1 + b1 + c0)*t} - 0.000655334$ $/E^{2*(a1 + b0 + b1 + c0)*t} + 0.00061581/E^{2*(a0 + a1 + b0 + c1)*t} - 0.000362639/E^{2*(a0 + b0 + b1 + c1)*t} - 0.00126494$ $/E^{2*(a0 + a1 + c0 + c1)*t} + 0.000910484/E^{2*(a1 + b0 + c0 + c1)*t} + 0.000596409/E^{2*(a0 + b1 + c0 + c1)*t} - 0.000747581$ $/E^{2*(b0 + b1 + c0 + c1)*t})/E^{2*(b2 + c2)*t} + (0.00224861 + 0.0000702462/E^{2*(a0 + a1 + b0 + b1)*t} - 0.00009901$ $/E^{2*(a0 + a1 + b1 + c0)*t} - 0.000181843/E^{2*(a1 + b0 + b1 + c0)*t} + 0.000928941/E^{2*(a0 + a1 + b0 + c1)*t} + 0.0000760541$ $/E^{2*(a0 + b0 + b1 + c1)*t} - 0.000931705/E^{2*(a0 + a1 + c0 + c1)*t} - 0.000225394/E^{2*(a1 + b0 + c0 + c1)*t} - 0.000123967$ $/E^{2*(a0 + b1 + c0 + c1)*t} + 0.000407291/E^{2*(b0 + b1 + c0 + c1)*t})/E^{2*(a2 + c2)*t}$

FIGURE 4.6 – Exemple de solution analytique obtenue pour le motif ATG avec comme vecteur de probabilités d'occurrence initiales les fréquences d'apparition des codons dans les CDS du génome humain. Les paramètres de substitution sont formels dans cet exemple.

4.2.4.2 Courbes d'évolution

Lorsque tous les paramètres de substitution sont numériques, les courbes d'évolution des motifs choisis sont tracées sur un graphique de comparaison. Un second graphique propose le tracé de la courbe correspondant à la somme des probabilités d'occurrence des motifs choisis (figure 4.7).

4.2.4.3 Solutions numériques

Lorsque tous les paramètres de substitution sont numériques et qu'une valeur numérique a été entrée pour t dans l'option 6, les probabilités des motifs choisis sont données au temps t à la suite des courbes d'évolution (figure 4.8).

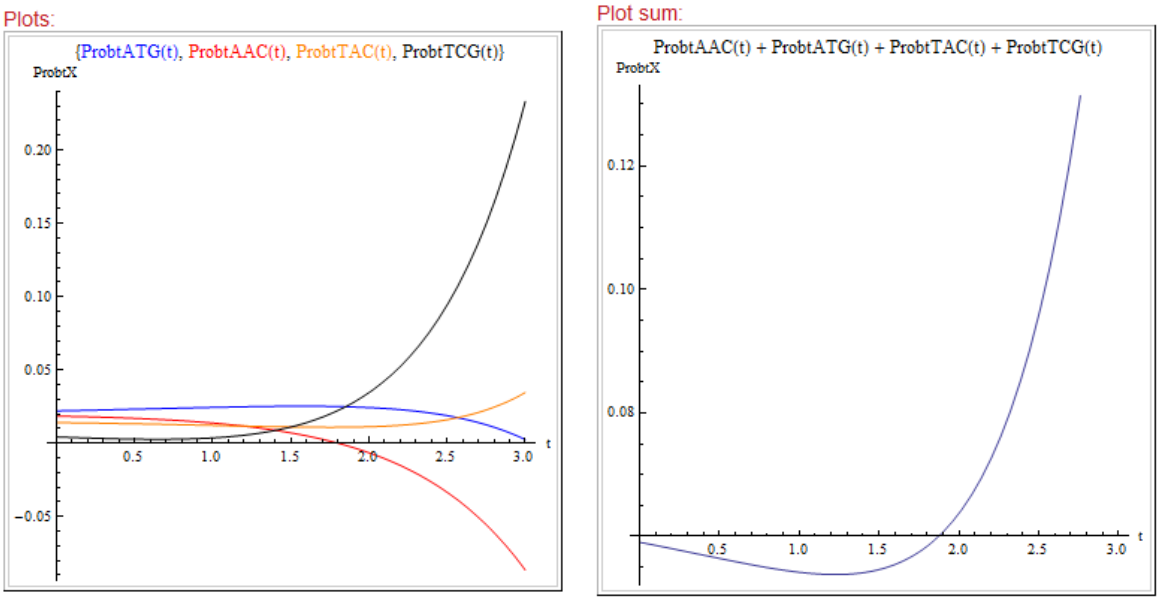


FIGURE 4.7 – Exemple de courbes d'évolution inverse de probabilités d'occurrence des 4 trinucléotides ATG, AAC, TAC et TCG au cours du temps t et courbe d'évolution inverse de la somme de leurs probabilités d'occurrence. L'abscisse et l'ordonnée correspondent respectivement au temps d'évolution t et à la probabilité d'occurrence des motifs.

Numerical solutions:

ProbtATG(t) with t = 1.4	0.0251928
ProbtAAC(t) with t = 1.4	0.0088487
ProbtTAC(t) with t = 1.4	0.011303
ProbtTCG(t) with t = 1.4	0.00877437
Sum: ProbtAAC(t) + ProbtATG(t) + ProbtTAC(t) + ProbtTCG(t) with t = 1.4	0.0541189

FIGURE 4.8 – Exemple de solutions numériques obtenues pour les trinucléotides ATG, AAC, TAC et TCG au temps $t = 1.4$.

4.3 Principe de fonctionnement du logiciel SEGM

Après le choix de la taille $1 \leq n \leq 5$ des motifs à étudier et le chargement du vecteur des probabilités initiales à partir de la page d'accueil de SEGM, l'utilisateur est dirigé vers une page spécifique aux n-nucléotides sur laquelle les différentes options exposées précédemment peuvent être modifiées. Lors du premier chargement de cette page, les valeurs par défaut sont sélectionnées et la solution analytique du premier motif dans l'ordre lexicographique, c'est-à-dire A^n, est calculée. Le nombre de paramètres de substitution et leurs valeurs n'ayant pas encore été définis par l'utilisateur, la solution analytique est donc calculée dans le cas le plus général, c'est-à-dire avec trois paramètres de substitution par site, ces paramètres étant formels (valeurs par défaut). La solution analytique obtenue lors du premier calcul est affichée au format standard.

Une fois le premier calcul effectué, l'utilisateur peut modifier le sens d'évolution, le nombre

de paramètres de substitution par site, les valeurs des paramètres de substitution, les motifs à étudier et le format de sortie des solutions analytiques. Lorsque tous les paramètres de substitution sont numériques, les options conditionnelles pour la gestion des courbes d'évolution et des solutions numériques sont disponibles.

Le même vecteur des probabilités d'occurrence initiales est conservé durant toute la session. Il peut être modifié en ouvrant une nouvelle session par retour à la page d'accueil et chargement d'un nouveau vecteur de probabilités d'occurrence initiales.

Le principe de fonctionnement du logiciel de recherche SEGM est synthétisé dans un diagramme de flux présenté en figure 4.9.

4.4 Conclusion

Nous avons développé une première version du logiciel de recherche SEGM durant la première partie de cette thèse pour unifier le modèle de substitution de nucléotides 3P (modèle de Kimura à trois paramètres de substitution) présenté dans le chapitre 1 et les modèles de dinucléotides et trinucléotides de type 3PS (trois paramètres par site de motif) proposés par Michel et présentés dans le chapitre 2. Ce logiciel est un outil initialement mis à la disposition des communautés biologique et bioinformatique pour étudier l'évolution des probabilités d'occurrence de motifs génétiques de taille 1 à 3 par calcul de solutions analytiques et de courbes d'évolution. Notre extension des modèles de substitution de nucléotides à des motifs de taille quelconque nous a permis d'étendre SEGM au calcul des solutions analytiques pour des motifs de taille 4 et 5.

Par ailleurs, cette extension a nécessité une refonte du logiciel nous permettant d'améliorer son ergonomie, notamment par l'automatisation du chargement des probabilités d'occurrence initiales au moyen d'un fichier de type tableur et l'ajout d'un choix de format de sortie des solutions analytiques de façon à les intégrer dans des programmes utilisateurs.

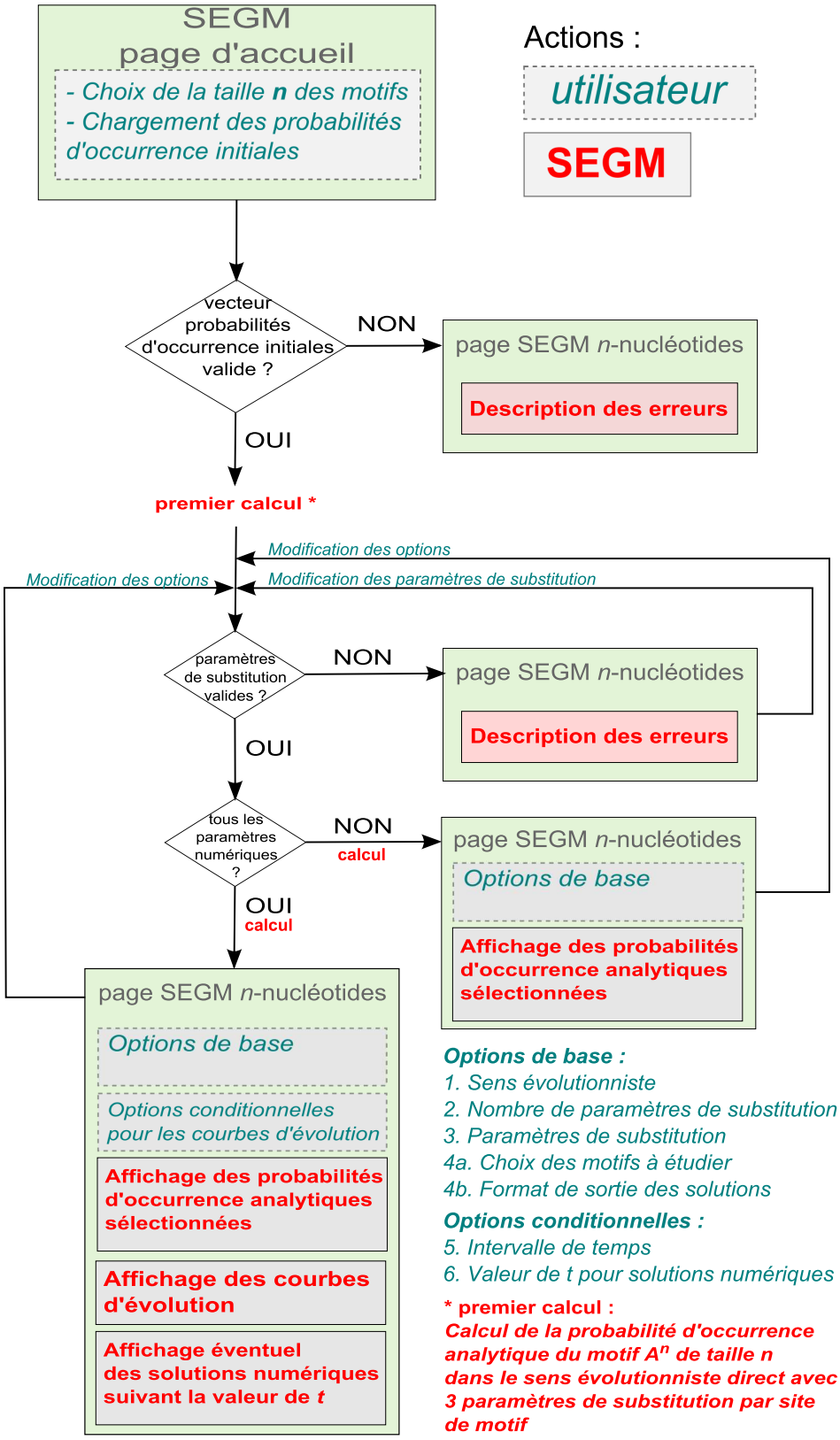


FIGURE 4.9 – Diagramme de flux du logiciel de recherche SEGM.

Chapitre 5

Applications bioinformatiques

Nous exposons dans ce chapitre deux exemples d'application bioinformatique des modèles de substitution de motifs présentés dans le chapitre 2.

La première application est une étude évolutive inverse des dinucléotides composant les sites d'épissage des introns du génome humain. Cette application a été réalisée à partir du modèle de substitution analytique de dinucléotides de type 3PS. Nous montrons notamment dans cet exemple d'application les courbes d'évolution inverse des sites d'épissage obtenues avec SEGM.

La seconde application proposée est une étude évolutive inverse du code circulaire X(EUK_PRO) identifié par Arquès et Michel en 1996 dans les gènes d'espèces eucaryotes et procaryotes [AM96]. Pour cela, nous avons utilisé le modèle d'évolution numérique inverse de trinucléotides avec blocage dynamique des taux de substitution développé au cours de cette thèse.

5.1 Évolution des sites d'épissage des introns du génome humain

Chez les eucaryotes, le processus d'épissage de l'ARNm est rendu possible par l'existence de signaux permettant au complexe moléculaire de détecter les introns à éliminer. Les dinucléotides aux frontières des introns font partie de ces signaux et sont appelés sites d'épissage. Plusieurs études ont révélé que la quasi totalité de ces sites d'épissage étaient composés des dinucléotides GT pour le site donneur en 5' de l'intron et AG pour le site accepteur en 3' de l'intron (figure 5.1). La paire canonique des sites d'épissage donneur-accepteur des introns est donc notée GT-AG. Cependant, dans certains cas, les paires de sites d'épissage sont composées de dinucléotides différents [BSS00, SSH90, Mou00].

En émettant l'hypothèse que les paires non canoniques des sites d'épissage résultent d'un

FIGURE 5.1 – Sites d'épissage canoniques des introns des génomes eucaryotes. Le site canonique donneur (resp. accepteur) à l'extrémité 5' (resp. 3') d'un intron est constitué du dinucléotide GT (resp. AG).

processus aléatoire d'évolution par substitution, nous avons voulu analyser l'évolution inverse des sites d'épissage donneur et accepteur des introns du génome humain avec le modèle analytique de substitution de dinucléotides disponible dans SEGM. Le but de cette analyse évolutive inverse était de retrouver dans le passé des dinucléotides GT pour le site donneur et AG pour le site accepteur avec des probabilités d'occurrence les plus proches de 1.

5.1.1 Données utilisées

Différentes bases de données sur les sites d'épissage existent, comme [BSS01, CGB04, SRH+06]. Pour effectuer cette étude évolutive des sites d'épissage des introns du génome humain, nous avons utilisé les données publiées par Chong et al. en 2004 dans [CGB04] présentant une base de données de sites d'épissage du génome humain : ICE (Information for the Coordinates of Exons). Cette base de données contient environ 92000 paires de sites d'épissage donneur-accepteur extraites de 10803 gènes humains. Chong et al. listent les fréquences d'apparition des 256 paires donneur-accepteur possibles réparties en trois classes : paires de sites d'épissage *pre* des premiers introns des gènes, paires de sites d'épissage *int* internes et paires de sites d'épissage *der* des derniers introns des gènes.

Nous avons effectué une analyse évolutive inverse de chaque site d'épissage donneur et accepteur pour les trois types (premiers, internes et derniers). Les fréquences et probabilités d'occurrence des 16 dinucléotides possibles pour les sites donneur et accepteur des trois types sont présentées respectivement dans le tableau 5.1 et dans le tableau 5.2.

L'objectif de notre étude évolutive inverse est de maximiser les probabilités d'occurrence des dinucléotides GT sur le site donneur et AG sur le site accepteur des trois classes de sites d'épissage tout en conservant bien évidemment des probabilités d'occurrence non négatives pour les autres dinucléotides, ce qui représente une difficulté dans la modélisation et le nombre de solutions possibles. Nous avons ainsi appliqué une fonction Mathematica de recherche de maximum global sur les solutions analytiques des dinucléotides obtenues avec SEGM à partir des probabilités d'occurrence initiales présentées dans les tableaux 5.1 et 5.2. Le modèle d'évolution utilisé pour cette analyse est le modèle 3PS à trois paramètres de substitution par site pour des motifs de taille 2 (dinucléotides).

5.1.2 Résultats de l'étude évolutive inverse des sites donneurs

Le tableau 5.3 contient les ensembles solutions des paramètres de substitution permettant de maximiser la probabilité d'occurrence du dinucléotide GT sur le site donneur et le temps d'évolution inverse t_{max} auquel est obtenue la solution, pour les trois classes de sites d'épissage.

Les probabilités d'occurrence numériques des 16 dinucléotides pour les trois classes de sites donneurs obtenues avec les solutions maximisant GT du tableau 5.3 sont données dans le tableau 5.4.

Din.	Occ. *pre*	$\tilde{P}_{pre}(0)$	Occ. *int*	$\tilde{P}_{int}(0)$	Occ. *der*	$\tilde{P}_{der}(0)$
AA	3	0.000299	22	0.000306	12	0.001127
AC	4	0.000399	6	0.000084	14	0.001315
AG	15	0.001497	21	0.000292	10	0.000940
AT	10	0.000998	97	0.001351	19	0.001785
CA	7	0.000699	10	0.000139	11	0.001034
CC	12	0.001198	15	0.000209	10	0.000940
CG	7	0.000699	2	0.000028	2	0.000188
CT	12	0.001198	10	0.000139	18	0.001691
GA	11	0.001098	25	0.000348	17	0.001597
GC	**126**	**0.012576**	**591**	**0.008230**	**134**	**0.012590**
GG	28	0.002795	31	0.000432	17	0.001597
GT	**9768**	**0.974947**	**70937**	**0.987870**	**10348**	**0.972282**
TA	2	0.000200	13	0.000181	5	0.000470
TC	4	0.000399	3	0.000042	8	0.000752
TG	9	0.000898	20	0.000279	7	0.000658
TT	1	0.000100	5	0.000070	11	0.001034
Total	*10019*	*1.000000*	*71808*	*1.000000*	*10643*	*1.000000*

TABLE 5.1 – Nombre d'occurrences et probabilités d'occurrence correspondantes des 16 dinucléotides sur le site donneur pour les trois classes de sites d'épissage des introns du génome humain extraits de la base de données ICE [CGB04] : sites des premiers introns (*pre*), sites internes (*int*) et sites des derniers introns (*der*). $\tilde{P}_{pre}(0)$, $\tilde{P}_{int}(0)$ et $\tilde{P}_{der}(0)$ sont les vecteurs des probabilités d'occurrence actuelles des 16 dinucléotides pour les trois classes de sites et sont donc les vecteurs des probabilités d'occurrence initiales dans le cadre de l'étude évolutive inverse. En gras : les probabilités d'occurrence des dinucléotides les plus fréquents (GT et GC) sur le site donneur.

Din.	Occ. *pre*	$\tilde{P}_{pre}(0)$	Occ. *int*	$\tilde{P}_{int}(0)$	Occ. *der*	$\tilde{P}_{der}(0)$
AA	5	0.000499	13	0.000181	12	0.001127
AC	16	0.001597	72	0.001003	30	0.002819
AG	**9870**	**0.985128**	**71442**	**0.994903**	**10421**	**0.979141**
AT	13	0.001297	64	0.000891	16	0.001503
CA	22	0.002196	21	0.000292	16	0.001503
CC	10	0.000998	11	0.000153	11	0.001033
CG	9	0.000898	10	0.000139	6	0.000564
CT	7	0.000699	16	0.000223	26	0.002443
GA	2	0.000200	22	0.000307	9	0.000846
GC	7	0.000699	9	0.000125	5	0.000470
GG	20	0.001996	41	0.000571	32	0.003007
GT	2	0.000200	11	0.000153	9	0.000846
TA	7	0.000699	7	0.000098	5	0.000470
TC	5	0.000499	13	0.000181	5	0.000470
TG	14	0.001397	41	0.000571	26	0.002443
TT	10	0.000998	15	0.000209	14	0.001315
Total	*10019*	*1.000000*	*71808*	*1.000000*	*10643*	*1.000000*

TABLE 5.2 – Nombre d'occurrences et probabilités d'occurrence correspondantes des 16 dinucléotides sur le site accepteur pour les trois classes de sites d'épissage des introns du génome humain extraits de la base de données ICE [CGB04] : sites des premiers introns (*pre*), sites internes (*int*) et sites des derniers introns (*der*). $\tilde{P}_{pre}(0)$, $\tilde{P}_{int}(0)$ et $\tilde{P}_{der}(0)$ sont les vecteurs des probabilités d'occurrence actuelles des 16 dinucléotides pour les trois classes de sites et sont donc les vecteurs des probabilités d'occurrence initiales dans le cadre de l'étude évolutive inverse. En gras : les probabilités d'occurrence du dinucléotide le plus fréquent (AG) sur le site accepteur.

Paramètres	a_1	b_1	c_1	a_2	b_2	c_2	t_{max}
Sites *pre*	0.052943	0.063185	0.004859	0.673365	0.056824	0.148824	0.019150
Sites *int*	0.127851	0.013010	0.006527	0.779447	0.032591	0.040574	0.010689
Sites *der*	0.087347	0.082995	0.050481	0.623307	0.077895	0.077975	0.020764

TABLE 5.3 – Ensembles de paramètres de substitution et temps d'évolution inverse t_{max} maximisant la probabilité d'occurrence du dinucléotide GT pour les trois classes de sites d'épissage donneurs *pre*, *int* et *der*. Les paramètres de substitution correspondent au modèle de substitution de dinucléotides 3PS : a_1 et a_2 sont les taux de transition (A↔G, C↔T) sur le premier et le second site, b_1 et b_2 sont les taux de transversion de type I (A↔T, C↔G) et c_1 et c_2 sont les taux de transversion de type II (A↔C, G↔T).

Din.	$\tilde{P}_{pre}(0.019150)$	Δ_{pre}	$\tilde{P}_{int}(0.010689)$	Δ_{int}	$\tilde{P}_{der}(0.020764)$	Δ_{der}
AA	0.000283	-0.000016	0.000306	0.000000	0.001131	0.000004
AC	0.000391	-0.000008	0.000073	-0.000011	0.001313	-0.000002
AG	0.001518	0.000021	0.000292	0.000000	0.000939	-0.000001
AT	0.000000	-0.000998	0.000000	-0.001351	0.000000	-0.001785
CA	0.000698	-0.000001	0.000140	0.000001	0.001047	0.000013
CC	0.001202	0.000004	0.000210	0.000001	0.000933	-0.000007
CG	0.000698	-0.000001	0.000027	-0.000001	0.000172	-0.000016
CT	0.000000	-0.001198	0.000000	-0.000139	0.000000	-0.001691
GA	0.000000	-0.001098	0.000000	-0.000348	0.000000	-0.001597
GC	0.000000	-0.012576	0.000000	-0.008230	0.000000	-0.012590
GG	0.000000	-0.002795	0.000000	-0.000432	0.000000	-0.001597
GT	**0.993709**	**0.018762**	**0.998450**	**0.01058**	**0.992592**	**0.02031**
TA	0.000190	-0.000010	0.000180	-0.000001	0.000464	-0.000006
TC	0.000402	0.000003	0.000041	-0.000001	0.000748	-0.000004
TG	0.000909	0.000011	0.000280	0.000001	0.000661	0.000003
TT	0.000000	-0.000100	0.000000	-0.000070	0.000000	-0.001034

TABLE 5.4 – Probabilités d'occurrence des dinucléotides pour les trois classes de sites d'épissage donneurs obtenues dans le sens d'évolution inverse avec les paramètres de substitution solutions du tableau 5.3. Les colonnes Δ_{pre}, Δ_{int} et Δ_{der} indiquent la différence des probabilités d'occurrence de chaque dinucléotide pour les trois classes de sites d'épissage donneurs : $\Delta_{pre} = \tilde{P}_{pre}(0.019150) - \tilde{P}_{pre}(0)$, $\Delta_{int} = \tilde{P}_{int}(0.010689) - \tilde{P}_{int}(0)$ et $\Delta_{der} = \tilde{P}_{der}(0.020764) - \tilde{P}_{der}(0)$. En gras : probabilités d'occurrence et différence positive du dinucléotide canonique GT pour les trois classes de sites donneurs.

Les courbes d'évolution inverse des probabilités d'occurrence des dinucléotides obtenues avec les ensembles solutions de paramètres de substitution du tableau 5.3 sur le site donneur pour les trois classes de sites d'épissage *pre*, *int* et *der* sont données respectivement dans les figures 5.2, 5.3 et 5.4. Ces courbes d'évolution ont été générées avec SEGM.

Remarque 5.1 *Les solutions analytiques donnant des résultats exacts, contrairement à une modélisation par simulation informatique, les probabilités obtenues sont exactes à la décimale près.*

5.1.2.1 Analyse des résultats

Les probabilités d'occurrence initiales du dinucléotide GT sur le site donneur étant proches de 1 quelle que soit la classe des sites d'épissage, les solutions maximisant les probabilités d'occurrence de GT sont obtenues, comme nous pouvons nous y attendre, pour un temps d'évolution inverse t_{max} très court, comme le montre le tableau des solutions 5.3.

Le modèle de substitution de dinucléotides de type 3PS ne permet pas d'obtenir une probabilité d'occurrence du dinucléotide GT égale à 1 sur les sites donneurs des trois classes de sites d'épissage, mais il permet de s'en approcher très significativement. En effet, d'après le tableau

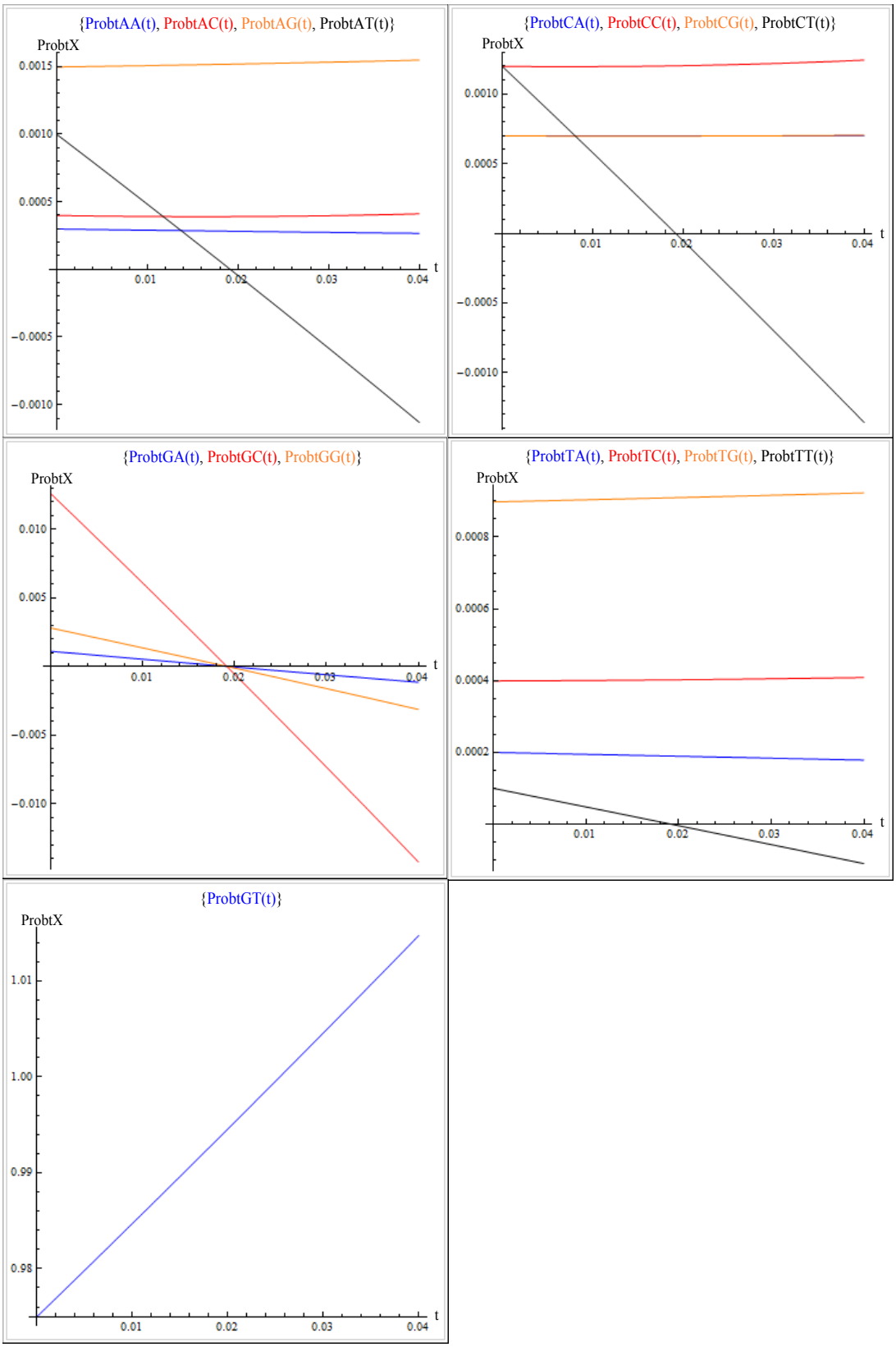


FIGURE 5.2 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.3 pour les sites donneurs des premiers introns (*pre*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.04]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.019150]$ d'après le tableau 5.3. La dernière figure donne l'évolution inverse du dinucléotide GT avec une probabilité maximale. Les courbes ont été générées avec le logiciel de recherche SEGM.

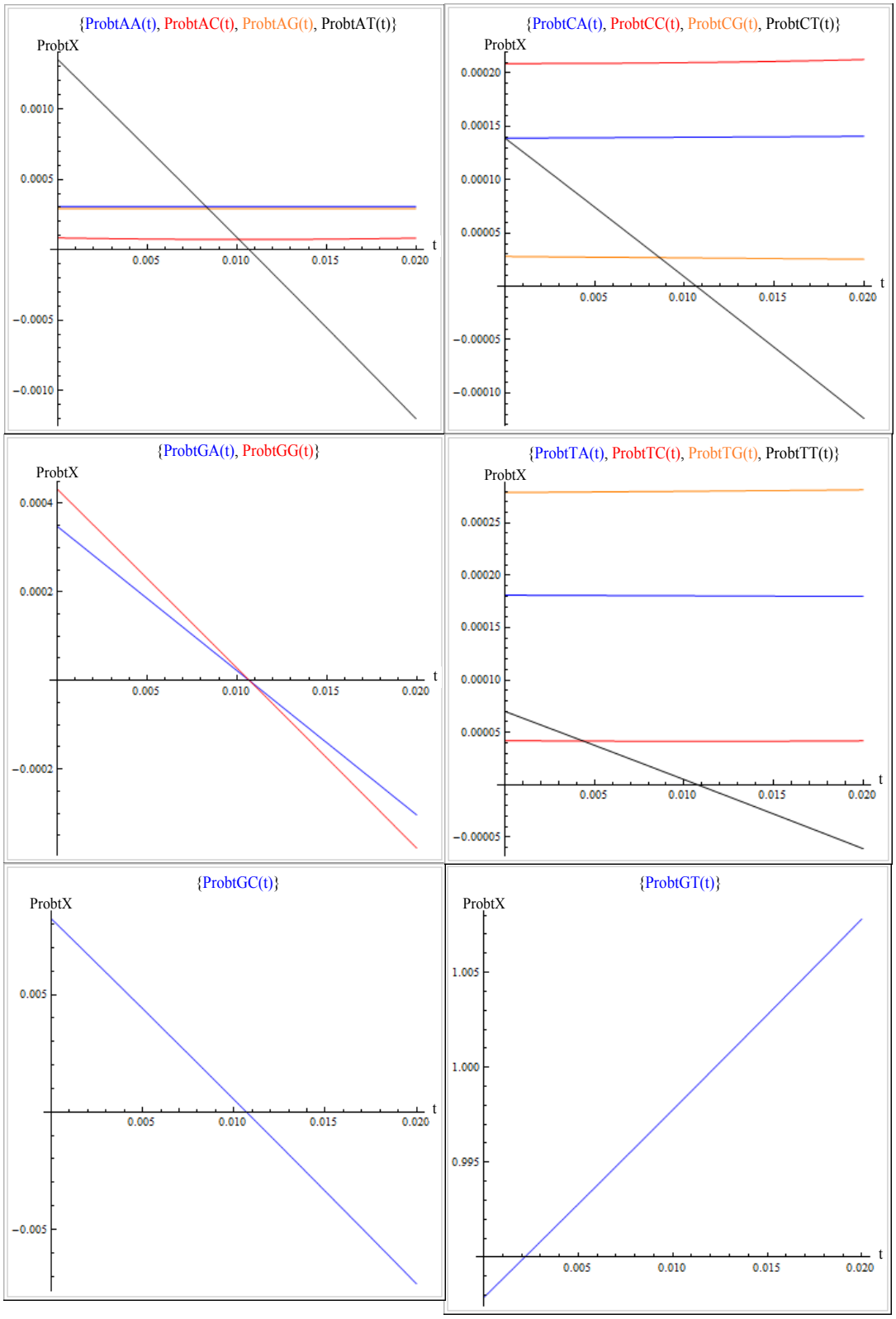


FIGURE 5.3 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.3 pour les sites donneurs internes (*int*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.02]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.010689]$ d'après le tableau 5.3. La dernière figure donne l'évolution inverse du dinucléotide GT avec une probabilité maximale. La courbe du dinucléotide GC est isolée pour des raisons d'échelle. Les courbes ont été générées avec le logiciel de recherche SEGM.

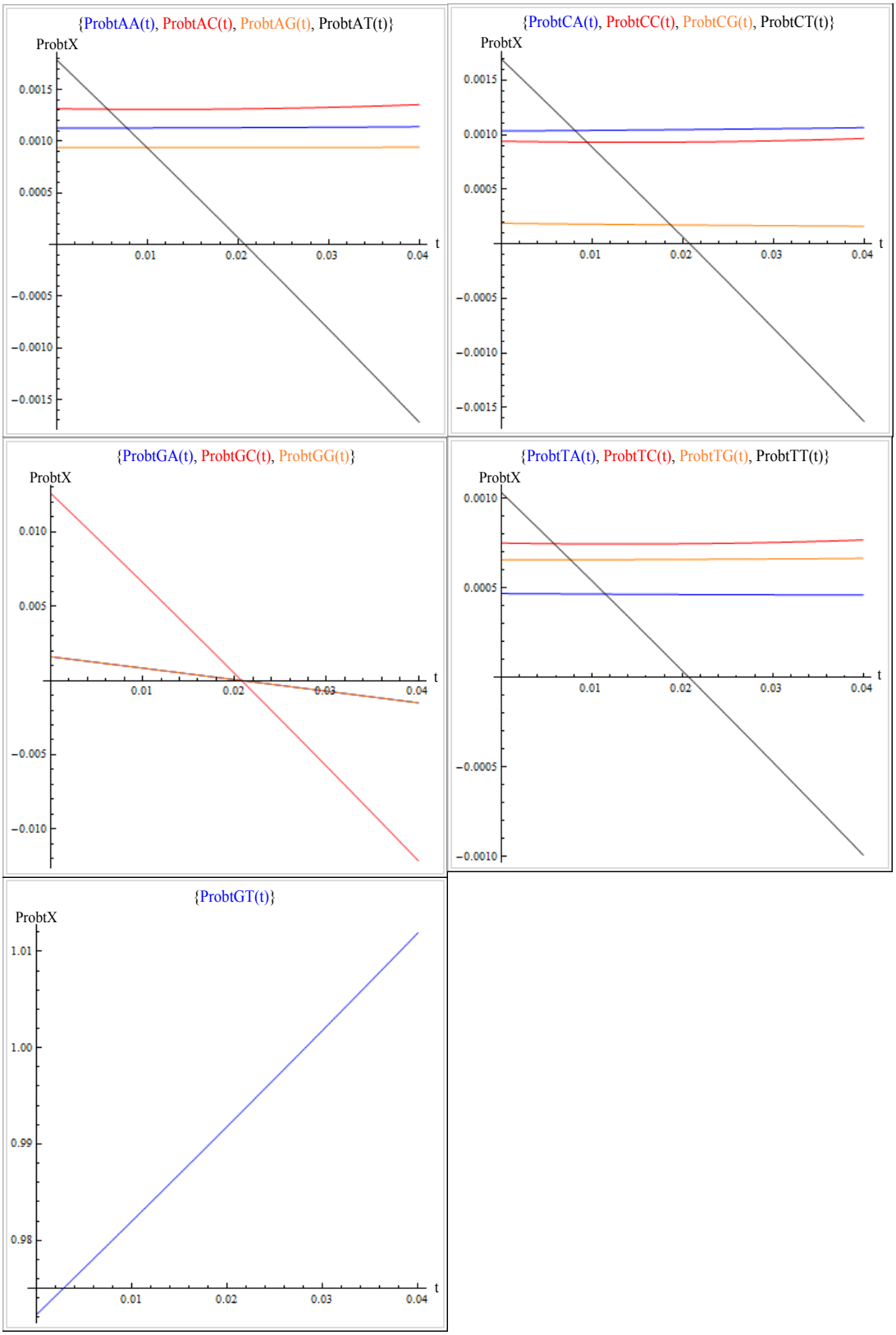


FIGURE 5.4 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.3 pour les sites donneurs des derniers introns (*der*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.04]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.020764]$ d'après le tableau 5.3. La dernière figure donne l'évolution inverse du dinucléotide GT avec une probabilité maximale. Les courbes ont été générées avec le logiciel de recherche SEGM.

5.4, nous pouvons observer que les gains des probabilités d'occurrence de GT sont de 0.018762 pour les sites donneurs des premiers introns (*pre*), de 0.01058 pour les sites donneurs internes (*int*) et de 0.02031 pour les sites donneurs des derniers introns (*der*).

Dans les gènes humains actuels, quelque soit la classe de sites d'épissage donneurs, le dinucléotide le plus fréquent après GT est GC et se détache de tous les autres dinucléotides (tableau 5.1). Les ensembles solutions de paramètres de substitution pour les trois classes de sites donneurs (tableau 5.3) montrent que la substitution la plus fréquente est une transition sur le second site des dinucléotides : 0.67 pour les sites *pre*, 0.78 pour les sites *int* et 0.62 pour les sites *der*. Ce fort taux de transition sur le second site s'explique par la probabilité d'occurrence initiale de GC beaucoup plus importante que les probabilités des autres dinucléotides. En effet, la substitution de GC en GT correspond à une transition a_2 (A↔G et C↔T) sur le second site.

Plus spécifiquement, l'ensemble des paramètres de substitution obtenu pour les sites d'épissage donneurs de la classe *pre* contient un second taux de substitution se détachant des autres. Il s'agit du taux $c_2 = 0.15$, correspondant à une transversion de type II (A↔C et G↔T) sur le second site. Ce taux peut s'expliquer par la probabilité d'occurrence initiale du dinucléotide GG (0.002795, tableau 5.1) qui est supérieure aux probabilités d'occurrence initiales de tous les autres dinucléotides, sauf GT et GC (tableau 5.1).

De façon similaire, le dinucléotide ayant la probabilité d'occurrence initiale la plus importante après celles de GT et GC sur les sites donneurs de la classe *int* est le dinucléotide AT (0.001351, tableau 5.1). Il n'est donc pas étonnant que le taux de substitution $a_1 = 0.13$ correspondant à une transition sur le premier site (A↔G et C↔T) soit le taux de substitution le plus élevé après a_2.

Par ailleurs, pour les trois classes de sites d'épissage donneurs, le taux de substitution c_1, correspondant à une transversion de type II (A↔C et G↔T) sur le premier site, est le plus faible des taux de substitution.

5.1.3 Résultats de l'étude évolutive inverse des sites accepteurs

Les ensembles solutions des paramètres de substitution et du temps d'évolution inverse t maximisant la probabilité d'occurrence du dinucléotide AG sur le site accepteur, pour les trois classes de sites d'épissage, sont listés dans le tableau 5.5.

Les probabilités d'occurrence des 16 dinucléotides pour les trois classes de sites accepteurs obtenues avec les solutions maximisant le dinucléotide AG du tableau 5.5 sont données dans le tableau 5.6.

Les figures 5.5, 5.6 et 5.7 montrent les courbes d'évolution inverse des probabilités d'occurrence des dinucléotides obtenues avec les ensembles solutions de paramètres de substitution du tableau 5.5 sur le site accepteur respectivement pour les trois classes de sites d'épissage *pre*, *int* et der. Ces courbes d'évolution ont été générées avec SEGM à partir des probabilités initiales d'occurrence données dans le tableau 5.2 et des paramètres de substitution des ensembles solutions du tableau 5.5.

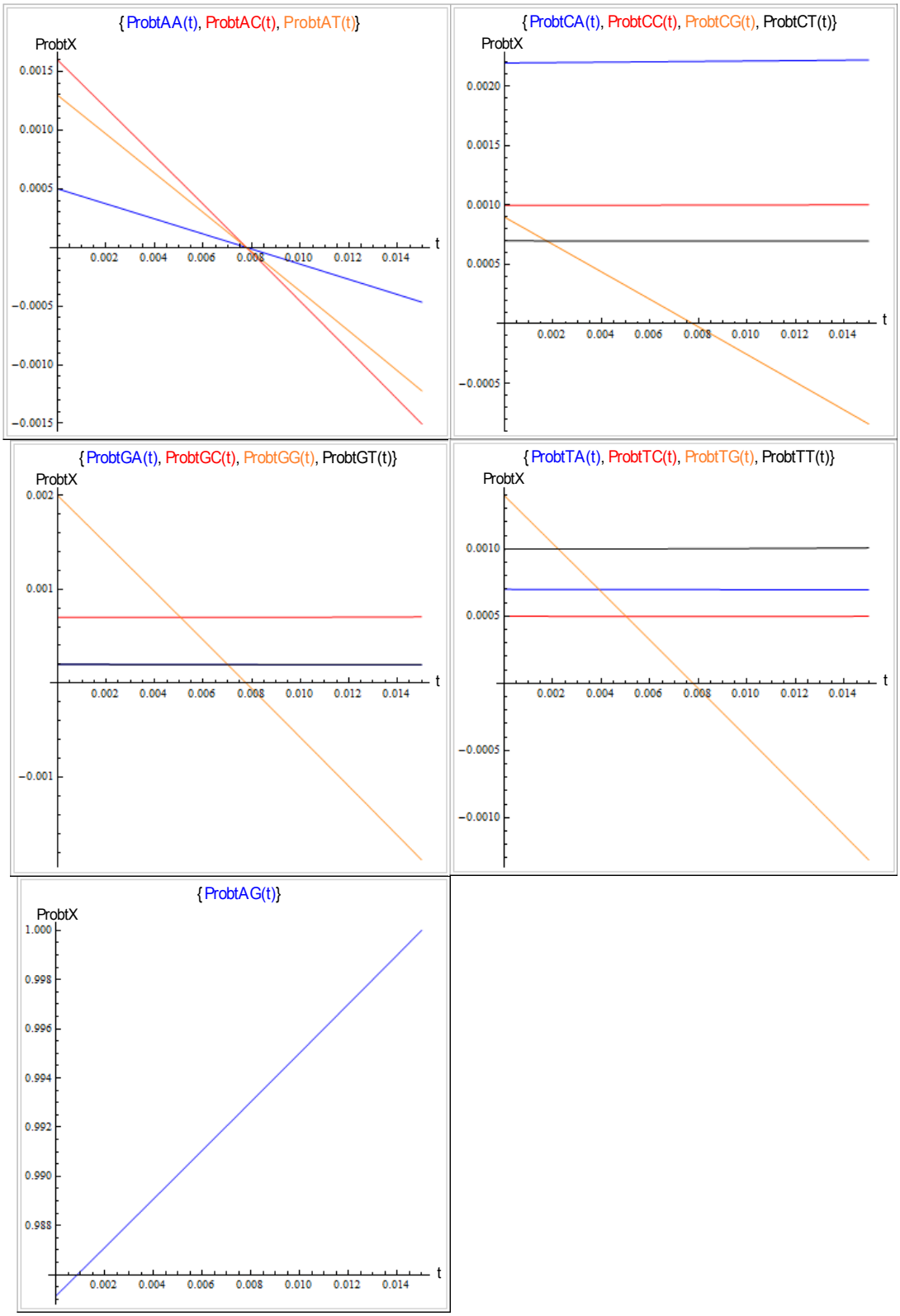


FIGURE 5.5 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.5 pour les sites accepteurs des premiers introns (*pre*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.015]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.007775]$ d'après le tableau 5.5. La dernière figure donne l'évolution inverse du dinucléotide AG avec une probabilité maximale. Les courbes ont été générées avec le logiciel de recherche SEGM.

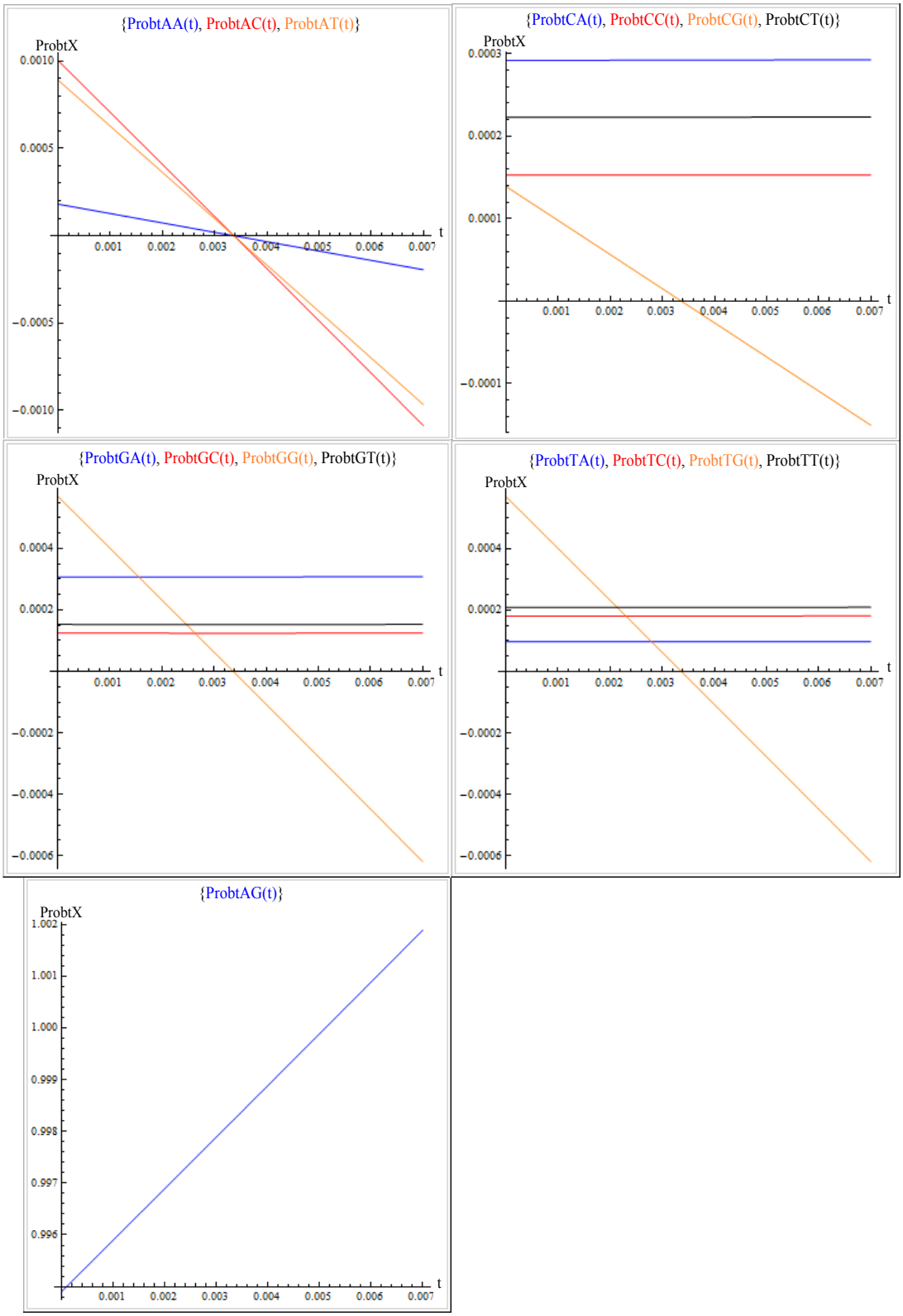


FIGURE 5.6 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.5 pour les sites accepteurs internes (*int*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.007]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.003370]$ d'après le tableau 5.5. La dernière figure donne l'évolution inverse du dinucléotide AG avec une probabilité maximale. Les courbes ont été générées avec le logiciel de recherche SEGM.

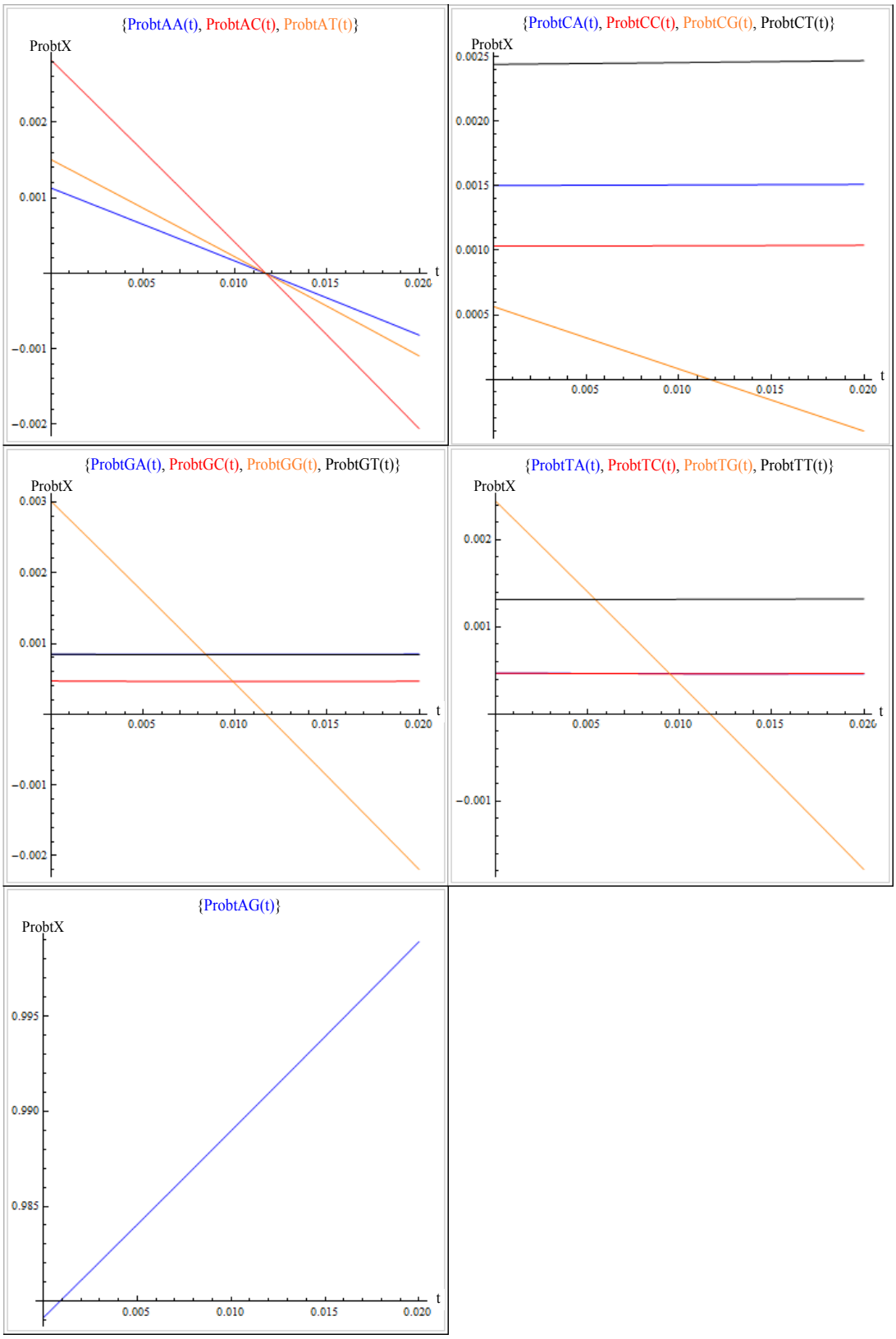


FIGURE 5.7 – Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution du tableau 5.5 pour les sites accepteurs des derniers introns (*der*). L'abscisse donne le temps d'évolution inverse t dans l'intervalle $[0, 0.02]$ et les probabilités d'occurrence des dinucléotides sont en ordonnée. Le modèle est valide pour $t \in [0, 0.011656]$ d'après le tableau 5.5. La dernière figure donne l'évolution inverse du dinucléotide AG avec une probabilité maximale. Les courbes ont été générées avec le logiciel de recherche SEGM.

Paramètres	a_1	b_1	c_1	a_2	b_2	c_2	t_{max}
Sites *pre*	0.260257	0.181874	0.116436	0.064441	0.208015	0.168977	0.007775
Sites *int*	0.170186	0.170166	0.041318	0.053649	0.298965	0.265716	0.003370
Sites *der*	0.263023	0.213572	0.048007	0.098019	0.246564	0.130815	0.011656

TABLE 5.5 – Ensembles de paramètres de substitution et temps d'évolution inverse t_{max} maximisant la probabilité d'occurrence du dinucléotide AG pour les trois classes de sites d'épissage accepteurs *pre*, *int* et *der*. Les paramètres de substitution correspondent au modèle de substitution de dinucléotides 3PS : a_1 et a_2 sont les taux de transition (A↔G, C↔T) sur le premier et le second site, b_1 et b_2 sont les taux de transversion de type I (A↔T, C↔G) et c_1 et c_2 sont les taux de transversion de type II (A↔C, G↔T).

Din.	$\tilde{P}_{pre}(0.007775)$	Δ_{pre}	$\tilde{P}_{int}(0.003370)$	Δ_{int}	$\tilde{P}_{der}(0.011656)$	Δ_{der}
AA	0.000000	-0.000499	0.000000	-0.000181	0.000000	-0.001127
AC	0.000000	-0.001597	0.000000	-0.001003	0.000000	-0.002819
AG	**0.992811**	**0.007683**	**0.998260**	**0.003357**	**0.990607**	**0.011466**
AT	0.000000	-0.001297	0.000000	-0.000891	0.000000	-0.001503
CA	0.002209	0.000013	0.000292	0.000000	0.001507	0.000004
CC	0.000999	0.000001	0.000153	0.000000	0.001036	0.000003
CG	0.000000	-0.000898	0.000000	-0.000139	0.000000	-0.000564
CT	0.000697	-0.000002	0.000223	0.000000	0.002459	0.000016
GA	0.000196	-0.000004	0.000307	0.000000	0.000845	-0.000001
GC	0.000699	0.000000	0.000124	-0.000001	0.000462	-0.000008
GG	0.000000	-0.001996	0.000000	-0.000571	0.000000	-0.003007
GT	0.000196	-0.000004	0.000153	0.000000	0.000841	-0.000005
TA	0.000697	-0.000002	0.000098	0.000000	0.000463	-0.000007
TC	0.000496	-0.000003	0.000181	0.000000	0.000463	-0.000007
TG	0.000000	-0.001397	0.000000	-0.000571	0.000000	-0.002443
TT	0.001001	0.000003	0.000209	0.000000	0.001317	0.000002

TABLE 5.6 – Probabilités d'occurrence des dinucléotides pour les trois classes de sites d'épissage accepteurs obtenues dans le sens d'évolution inverse avec les paramètres de substitution solutions du tableau 5.5. Les colonnes Δ_{pre}, Δ_{int} et Δ_{der} indiquent la différence des probabilités d'occurrence de chaque dinucléotide pour les trois classes de sites d'épissage accepteurs : $\Delta_{pre} = \tilde{P}_{pre}(0.007775) - \tilde{P}_{pre}(0)$, $\Delta_{int} = \tilde{P}_{int}(0.003370) - \tilde{P}_{int}(0)$ et $\Delta_{der} = \tilde{P}_{der}(0.011656) - \tilde{P}_{der}(0)$. En gras : probabilités d'occurrence et différence positive du dinucléotide canonique AG pour les trois classes de sites accepteurs.

5.1.3.1 Analyse des résultats

D'après le tableau 5.6, les variations des probabilités d'occurrence du dinucléotide AG sont de $0.992811-0.985128 = 0.007683$ pour les sites accepteurs des premiers introns (*pre*), de $0.998260-0.994903 = 0.003357$ pour les sites accepteurs internes (*int*) et de $0.990607-0.979141 = 0.011466$ pour les sites accepteurs des derniers introns (*der*). En valeur absolue ces variations sont moins importantes que celles observées pour les sites donneurs (lignes en gras dans les tableaux 5.4 et 5.6).

Même si le modèle d'évolution utilisé ne permet pas de façon analogue au cas GT, de retrouver un dinucléotide AG avec une probabilité d'occurrence égale à 1 sur les sites d'épissage accepteurs (tableau 5.6), le fait que dans les gènes actuels les dinucléotides AG soient plus fréquents sur les sites accepteurs (tableau 5.2) que les dinucléotides GT ne le sont sur les sites donneurs (5.1) implique nécessairement que la variation maximale des probabilités d'occurrence des dinucléotides AG sur les sites accepteurs est inférieure à la variation maximale des probabilités d'occurrence des dinucléotides GT sur les sites donneurs.

Contrairement aux ensembles solutions de paramètres de substitution obtenus pour les sites donneurs dans lesquels le paramètre a_2 se dégage des autres, les ensembles solutions de paramètres de substitution pour les sites accepteurs sont plus homogènes.

Dans le cas des sites accepteurs des premiers introns (*pre*), les deux paramètres de substitution les plus élevés de l'ensemble solution (tableau 5.5) sont $a_1 \simeq 0.26$, correspondant à une transition (A↔G et C↔T) sur le premier site et $b_2 \simeq 0.21$, correspondant à une transversion de type I (A↔T et C↔G) sur le second site. Ces deux paramètres a_1 et b_2 permettent de substituer respectivement les dinucléotides GG et AC en AG, GG et AC étant parmi les plus fréquents (respectivement 0.001996 et 0.002597) dans les gènes humains actuels (tableau 5.2) après AG. L'explication est similaire pour les sites accepteurs des derniers introns (*der*).

Ce raisonnement peut également être appliqué au cas des sites accepteurs internes (*int*) avec les deux paramètres de substitution les plus élevés, b_2 et c_2, correspondant aux deux types de transversion sur le second site et permettant respectivement aux dinucléotides AC et AT d'être substitués en AG. Les dinucléotides AC et AT sont les plus fréquents initialement après AG (tableau 5.5).

Nous pouvons également noter que quels que soient les types de sites d'épissage, les ensembles solutions de paramètres de substitution obtenus ont en commun un très faible taux c_1 de transversion de type II (A↔C et G↔T) sur le premier site. Ce taux c_1 est d'ailleurs le plus faible pour les sites donneurs et accepteurs, sauf dans le cas des sites accepteurs des premiers introns (tableaux 5.3 et 5.5).

5.1.4 Conclusion

Nous nous sommes intéressés à l'évolution inverse des dinucléotides composant les sites d'épissage des introns du génome humain avec le modèle analytique de substitution de dinucléotides implanté dans le logiciel de recherche SEGM. Bien que ces sites d'épissage soient essentiellement composés des mêmes dinucléotides (GT sur le site donneur et AG sur le site accepteur de l'intron), il existe des variations. À partir des données publiées par Chong et al. [CGB04] sur les fréquences d'occurrence des dinucléotides sur les sites d'épissage des introns du génome humain et en émettant l'hypothèse que ces variations sont dues à des substitutions aléatoires au cours de l'évolution, nous avons déterminé quels paramètres de substitution permettent de maximiser les probabilités d'occurrence du dinucléotide GT sur le site donneur et du dinucléotide AG sur le site accepteur au cours d'un processus d'évolution inverse. Les résultats obtenus avec le modèle analytique de substitution de dinucléotides de type 3PS montrent que les taux de substitution

maximisant les probabilités d'occurrence des dinucléotides GT sur le site donneur et AG sur le site accepteur sont cohérents avec le raisonnement scientifique « intuitif » se basant sur les probabilités d'occurrence initiales. L'étude évolutive inverse des sites d'épissage des introns du génome humain a fait l'objet d'une publication en revue internationale à comité de lecture en 2009 [BM09b] avec l'unification des modèles analytiques de substitution de nucléotides de type 3P, dinucléotides et trinucléotides de type 3PS dans le logiciel de recherche SEGM.

5.2 Évolution inverse du code circulaire X(EUK_PRO) avec le modèle d'évolution dynamique par blocage

Depuis sa découverte en 1996, quelques études évolutives ont été menées sur le code ciculaire X(EUK_PRO). Michel a notamment montré en 2007 dans [Mic07a] qu'il était possible d'obtenir une répartition des acides aminés codés proche de celle actuellement codée dans les génomes d'espèces eucaryotes et procaryotes en faisant évoluer les probabilités d'occurrence des trinucléotides du code X(EUK_PRO) dans le sens direct (passé-présent) avec un modèle de substitution analytique de trinucléotides à trois paramètres de substitution par site 3PS.

Le but de notre étude est de tester l'hypothèse d'un code circulaire « primitif » sous-ensemble du code circulaire X(EUK_PRO) identifié en 1996. Nous rappelons ici que tout sous-ensemble d'un code circulaire est un code circulaire. Nous testons cette hypothèse avec le modèle d'évolution numérique inverse de trinucléotides à matrice de substitutions dynamique avec blocage développé au cours de cette thèse.

5.2.1 Base de données utilisée

Nous avons réalisé l'étude évolutive inverse du code ciculaire $X_0 = X$(EUK_PRO) à partir des fréquences d'apparition de codons dans les CDS chromosomiques de plusieurs espèces eucaryotes et procaryotes disponibles dans la base de données NRCUD [1] (Non-Redundant Codon Usage Database). Les fréquences d'apparition des codons proposées dans cette base de données sont elles-même extraites de la base NCBI RefSeq [2]. Les données de NRCUD utilisées pour notre analyse évolutive représentent un total de 76223181 codons issus de séquences chromosomiques de 33 espèces. Les probabilités d'occurrence actuelles des 64 codons calculées à partir de la base de données NRCUD sont listées dans le tableau 5.7. Le processus d'évolution étant réalisé dans le sens d'évolution inverse (présent-passé), le vecteur des probabilités d'occurrence actuelles des 64 codons constituent le vecteur des probabilités d'occurrence initiales $\tilde{P}_i(0)$ des 64 codons pour le modèle d'évolution inverse avec blocage. $\tilde{P}_i(t)$ représente donc la probabilité d'occurrence du codon i au temps d'évolution inverse t.

Remarque 5.2 *Les probabilités d'occurrence des codons stop TAA, TAG et TGA extraites de la base de données NRCUD ne sont pas nulles mais très proches de 0. Nous avons toutefois choisi de conserver les données telles qu'elles sont fournies par NRCUD. Ce choix a une conséquence négligeable puisque dans les résultats obtenus, les probabilités d'occurrence des trois codons stop sont les premières à devenir nulles et ce dans un temps très court.*

1. `http://bioinformatics.forsyth.org/nrcud`, accédée en avril 2011.
2. `http://www.ncbi.nlm.nih.gov/RefSeq`.

Codon i	$\tilde{P}_i(0)$	Codon i	$\tilde{P}_i(0)$
AAA	0.030638	**GAA**	**0.035503**
AAC	**0.021515**	**GAC**	**0.022261**
AAG	0.029151	**GAG**	**0.029154**
AAT	**0.026116**	**GAT**	**0.032760**
ACA	0.013675	GCA	0.016404
ACC	**0.015009**	**GCC**	**0.020644**
ACG	0.011910	GCG	0.018288
ACT	0.014017	GCT	0.019460
AGA	0.011577	GGA	0.016889
AGC	0.012466	**GGC**	**0.019846**
AGG	0.006117	GGG	0.007675
AGT	0.011410	**GGT**	**0.017987**
ATA	0.011288	**GTA**	**0.009444**
ATC	**0.022164**	**GTC**	**0.015520**
ATG	0.023475	GTG	0.019788
ATT	**0.024205**	**GTT**	**0.019033**
CAA	0.020202	TAA	0.001049
CAC	0.010427	**TAC**	**0.014660**
CAG	**0.019749**	TAG	0.000509
CAT	0.012516	TAT	0.016007
CCA	0.014183	TCA	0.013347
CCC	0.009418	TCC	0.012536
CCG	0.011932	TCG	0.012197
CCT	0.011936	TCT	0.016428
CGA	0.007269	TGA	0.000850
CGC	0.012165	TGC	0.007079
CGG	0.006126	TGG	0.011350
CGT	0.009704	TGT	0.007441
CTA	0.008596	TTA	0.013803
CTC	**0.015710**	**TTC**	**0.020533**
CTG	**0.021030**	TTG	0.018185
CTT	0.016712	TTT	0.020963

TABLE 5.7 – Probabilités d'occurrence actuelles des 64 codons issues des séquences chromosomiques d'espèces eucaryotes et procaryotes de la base de données NRCUD. Ces probabilités constituent le vecteur initial $\tilde{P}(0)$ du modèle d'évolution par blocage dynamique. Les probabilités d'occurrence initiales des 20 codons du code circulaire X_0 sont représentées en gras.

i	$\tilde{P}_{X_i}(0)$
0	0.422843
1	0.280374
2	0.228090
3	0.068694

TABLE 5.8 – Probabilités d'occurrence initiales des ensembles de codons X_0, X_1, X_2 et X_3. Elles s'obtiennent en sommant, pour chaque ensemble, les probabilités d'occurrence initiales des codons listées dans le tableau 5.7.

Pour effectuer l'analyse évolutive inverse du code circulaire X_0, nous avons réparti les probabilités d'occurrence initiales $\tilde{P}_i(0)$ des 64 codons en 4 ensembles $X_j, 0 \leq j \leq 3$. Les ensembles X_0, X_1 et X_2 contenant chacun 20 codons correspondent aux codes circulaires identifiés dans les trois phases de lecture des gènes par Arquès et Michel en 1996, et X_3 est l'ensemble composé des 4 triplets AAA, CCC, GGG et TTT. Les ensembles X_0, X_1 et X_2 sont donc identiques aux codes circulaires définis dans la section 1.2.3 du chapitre 1 et rappelés ici :

$$\begin{aligned}
X_0 &= \{\text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,} \\
&\quad \text{GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC}\}, \\
X_1 &= \{\text{AAG,ACA,ACG,ACT,AGC,AGG,ATA,ATG,CCA,CCG,} \\
&\quad \text{GCG,GTG,TAG,TCA,TCC,TCG,TCT,TGC,TTA,TTG}\}, \\
X_2 &= \{\text{AGA,AGT,CAA,CAC,CAT,CCT,CGA,CGC,CGG,CGT,} \\
&\quad \text{CTA,CTT,GCA,GCT,GGA,TAA,TAT,TGA,TGG,TGT}\}.
\end{aligned}$$

5.2.2 Algorithme du modèle d'évolution numérique inverse avec blocage dynamique des substitutions

L'objectif de notre étude étant de tester la possibilité d'obtenir un sous-ensemble « primitif » du code circulaire X_0 avec une probabilité d'occurrence maximale au cours d'un processus d'évolution inverse, nous avons implanté l'algorithme 1 présenté dans le chapitre 2 en ajoutant un balayage des paramètres de substitution. Nous présentons dans cette sous-section l'algorithme d'évolution inverse avec blocage dynamique des substitutions obtenu.

5.2.2.1 Calcul des probabilités d'occurrence dans le sens d'évolution inverse des ensembles X_0, X_1, X_2 et X_3

Les probabilités d'occurrence des ensembles X_0, X_1, X_2 et X_3 au temps d'évolution inverse t sont respectivement notées $\tilde{P}_{X_0}(t)$, $\tilde{P}_{X_1}(t)$, $\tilde{P}_{X_2}(t)$ et $\tilde{P}_{X_3}(t)$. Les probabilités d'occurrence initiales $\tilde{P}_{X_0}(0)$, $\tilde{P}_{X_1}(0)$, $\tilde{P}_{X_2}(0)$ et $\tilde{P}_{X_3}(0)$ sont données dans le tableau 5.8. Elles sont calculées en sommant les probabilités d'occurrence initiales des codons (tableau 5.7) de chaque ensemble X_i, $0 \leq i \leq 3$.

5.2.2.2 Matrice de substitution utilisée

Dans ce modèle numérique, pour chaque ensemble de paramètres de substitution testé, les probabilités d'occurrence inverses des 64 codons doivent être calculées à chaque intervalle de temps $\Delta = t_{h+1} - t_h$. Pour conserver un temps de calcul raisonnable (quelques jours), nous avons choisi d'utiliser une matrice de substitution initiale A (64,64) de type 2PS à six paramètres de substitution, soit deux paramètres par site. Cette matrice est basée sur une matrice de substitution de nucléotides (4,4) de type 2P (modèle de Kimura à deux paramètres défini dans la section 1.3.2 du chapitre 1). Cette matrice A permet donc de distinguer les taux de transition u_k, et de transversion v_k, pour chaque site $1 \leq k \leq 3$ des trinucléotides.

5.2.2.3 Convergence du processus d'évolution inverse

Nous avons vu dans le chapitre 2 qu'avec les modèles d'évolution analytiques classiques, les probabilités d'occurrence de motifs divergent lors d'une étude évolutive inverse, tendent vers les infinis et sortent par conséquent du domaine de validité des probabilités.

Le principe du modèle d'évolution numérique inverse avec matrice de substitution dynamique présenté ici est d'empêcher cette sortie du domaine de validité en bloquant les probabilités d'occurrence des motifs lorsqu'elles deviennent négatives ou nulles. Ce blocage se traduit par une mise à 0 des paramètres de substitution liés aux motifs bloqués dans la matrice de substitution. Cela entraîne donc une modification du processus d'évolution des probabilités d'occurrence des motifs non bloqués liés aux paramètres de substitution mis à 0.

Par exemple, le blocage du codon stop TAG implique de mettre à 0 tous les paramètres de substitution liés à TAG. Ainsi, les 10 paramètres de substitution faisant muter TAG en les motifs AAG, CAG, GAG, TCG, TGG, TTG, TAA, TAC et TAT et inversement sont mis à 0. Le processus d'évolution des motifs précédemment cités est alors modifié.

Par conséquent, le blocage des substitutions pour un motif j donné peut intervenir de deux façons différentes : lorsque la probabilité d'occurrence de ce motif j devient négative ou nulle ou lorsque tous les motifs $i \neq j$ pouvant être substitués en j sont également bloqués. Dans ce dernier cas, le motif j est bloqué avec une probabilité d'occurrence non nulle, il est dit « stable ».

Ce phénomène de double blocage des substitutions engendre une convergence du processus d'évolution inverse, la matrice de substitution $A(t) = M(t) - I$ tendant vers la matrice nulle au fur et à mesure du processus d'évolution.

Dans l'algorithme de calcul 6 utilisé pour notre étude évolutive inverse du code circulaire X_0, la condition d'arrêt est donc la convergence du processus d'évolution, c'est-à-dire lorsque le vecteur de probabilités d'occurrence des motifs ne varie plus. Cette convergence est induite par la nullité de la matrice de substitution $A(t)$.

Algorithme 6 Algorithme de calcul des probabilités d'occurrence inverses des 64 codons avec le modèle d'évolution numérique inverse avec blocage dynamique des substitutions et balayage d'ensembles de paramètres de substitutions.

Entrées : $\tilde{P}(0) = \{\tilde{P}_1(0), \ldots, \tilde{P}_{64}(0)\}$ le vecteur des probabilités d'occurrence initiales des 64 codons.

Sorties : $\tilde{P}(t)$ le vecteur des probabilités d'occurrence des 64 codons au temps d'évolution inverse t.

pour tout $\{u_1, v_1, u_2, v_2, u_3, v_3\}$ tel que $u_1 + v_1 + u_2 + v_2 + u_3 + v_3 == 1$ **faire**
  $A(0) \leftarrow A$ définie avec l'ensemble des paramètres de substitution $\{u_1, v_1, u_2, v_2, u_3, v_3\}$
  $A_{sauv} \leftarrow$ matrice identité (64,64)
  $t \leftarrow 0$
  $continue \leftarrow$ Vrai
  $probaInit \leftarrow \tilde{P}(0)$
  $ensembleProbabilitesBloquees \leftarrow \{\}$
  tant que $continue$ == Vrai **faire**
    $sommeProbabilitesBloquees \leftarrow 0$
    $blocageEnCours \leftarrow$ Faux
    pour tout $1 \leq i \leq 64$ **faire**
      si $\tilde{P}_i(t) \leq 0$ et $i \notin ensembleProbabilitesBloquees$ **alors**
        $blocageEnCours \leftarrow$ Vrai
        $sommeProbabilitesBloquees \leftarrow sommeProbabilitesBloquees + \tilde{P}_i(t)$
        ajouter i à $ensembleProbabilitesBloquees$
      fin si
    fin pour
    $\tilde{P}(t) \leftarrow \tilde{P}(t)/(1 - sommeProbabilitesBloquees)$
    $A(t + \Delta) \leftarrow A(t)$
    pour tout $i \in ensembleProbabilitesBloquees$ **faire**
      mettre à 0 tous les paramètres de substitution des ligne i et colonne i de $A(t + \Delta)$
    fin pour
    pour tout $1 \leq i \leq 64$ **faire**
      $A(t + \Delta)[i, i] \leftarrow -\sum_{j=1, j \neq i}^{64} A(t + \Delta)[j, i]$
    fin pour
    si $blocageEnCours$ == Vrai **alors**
      $probaInit \leftarrow \tilde{P}(t)$
      $A_{sauv} \leftarrow$ matrice identité (64,64)
    fin si
    $A_{sauv} \leftarrow A_{sauv} \cdot e^{-A(t+\Delta)\times\Delta}$
    $\tilde{P}(t + \Delta) \leftarrow A_{sauv} \cdot e^{-A(t+\Delta)\times\Delta} \cdot probaInit$
    $continue \leftarrow \tilde{P}(t + \Delta) \neq \tilde{P}(t)$
    $t \leftarrow t + \Delta$
  fin tant que
  retourner $\tilde{P}(t)$
fin pour

5.2.2.4 Balayage des paramètres de substitution

L'algorithme de calcul 6 applique le processus d'évolution inverse pour tous les ensembles de paramètres $\{u_1, v_1, u_2, v_2, u_3, v_3\}$ valides, c'est-à-dire tel que leur somme est égale à 1, dans l'intervalle $[0, 1]$ avec un pas de 0.05 pour chaque paramètre de substitution. L'intervalle de temps Δ utilisé est 0.005.

5.2.3 Résultats

Nous présentons ici les deux ensembles de paramètres de substitution maximisant la probabilité d'occurrence du code circulaire X_0 dans le sens d'évolution inverse.

Paramètres	u_1	v_1	u_2	v_2	u_3	v_3	t_{max}	$\tilde{P}_{X_0}(t_{max})$
Ensemble E_1	0.05	0.25	0.45	0.00	0.20	0.05	115.895	0.9865
Ensemble E_2	0.05	0.20	0.45	0.00	0.20	0.10	34.550	0.9863

TABLE 5.9 – Ensembles des paramètres de substitution maximisant la probabilité d'occurrence du code circulaire X_0 et temps d'évolution inverse t_{max} auquel est atteinte la convergence du processus d'évolution avec blocage dynamique.

Les deux ensembles de paramètres de substitution E_1 et E_2 listés dans le tableau 5.9 permettent d'obtenir, à la convergence du processus d'évolution inverse, une probabilité d'occurrence de X_0 respectivement de 0.9865 et de 0.9863, donc très proche de 1 contre 0.4228 à l'état initial (tableau 5.8). La probabilité d'occurrence de X_0 n'augmente pas pour toutes les valeurs de paramètres de substitution, elle peut descendre à environ 0.3 avec certains ensembles de paramètres à la convergence du processus d'évolution inverse.

5.2.3.1 Analyse des résultats

Les tableaux 5.10 et 5.11 listent les probabilités d'occurrence des codons stabilisés à la convergence du processus d'évolution avec les ensembles de paramètres de substitution E_1 et E_2. Pour les deux ensembles E_1 et E_2, 9 codons convergent (probabilité d'occurrence stable non nulle). Parmi ces 9 codons, 8 appartiennent au code circulaire X_0 : AAT (E_1 uniquement), ATT, CAG, CTG (E_2 uniquement), GAA, GCC, GGC, TAC et TTC ; 1 appartient à X_2 : TGG. Nous pouvons noter que parmi les codons convergents, le codon TGG de X_2 possède une probabilité d'occurrence inférieure à celles de tous les codons de X_0 pour l'ensemble solutions E_2 (tableau 5.11) et tous sauf le codon TAC pour E_1 (tableau 5.10).

Paramètres de substitution Les deux ensembles de paramètres de substitution E_1 et E_2 maximisant la probabilité d'occurrence de X_0 sont sensiblement identiques (tableau 5.9). Les paramètres de transition u_1, u_2, u_3, sur les trois sites des codons et de transversion v_2 sur le deuxième site sont identiques pour les deux ensembles. Ces ensembles ne diffèrent que de 0.05 pour les paramètres de transversion v_1 et v_3 sur les premier et troisième sites. Les paramètres de transition sont plus élévés que les paramètres de transversion sur les deux derniers sites, ce qui n'est pas le cas pour le premier site. Par ailleurs, ces deux ensembles solutions interdisent les transversions sur le second site.

Ens.	Codon i	$\tilde{P}_i(115.895)$
X_0	AAT	0.041551
	ATT	0.274495
	CAG	0.025132
	GAA	0.335551
	GCC	0.157623
	GGC	0.060315
	TAC	0.012384
	TTC	0.079487
X_2	TGG	0.013461

TABLE 5.10 – Probabilités d'occurrence des 9 codons convergents lors du processus d'évolution inverse avec blocage dynamique. Probabilités obtenues avec l'ensemble de paramètres de subtitution E_1 (tableau 5.9).

Ens.	Codon i	$\tilde{P}_i(34.550)$
X_0	ATT	0.183863
	CAG	0.030161
	CTG	0.102299
	GAA	0.363618
	GCC	0.148850
	GGC	0.061015
	TAC	0.019832
	TTC	0.076706
X_2	TGG	0.013655

TABLE 5.11 – Probabilités d'occurrence des 9 codons convergents lors du processus d'évolution inverse avec blocage dynamique. Probabilités obtenues avec l'ensemble de paramètres de subtitution E_2 (tableau 5.9).

Complémentarité des codons stabilisés Quelque soit l'ensemble E_1 ou E_2, nous observons que parmi les 8 codons convergents de X_0, 6 sont complémentaires. En effet, nous obtenons les 3 paires de codons complémentaires AAT-ATT, GAA-TTC et GCC-GGC avec l'ensemble E_1 et les 3 paires de codons complémentaires CAG-CTG, GAA-TTC et GCC-GGC avec l'ensemble E_2. Si l'on considère l'union des ensembles de codons de X_0 convergents avec E_1 et E_2, seul le codon TAC n'appartient pas à une paire de codons complémentaires. Celui-ci possède également la probabilité d'occurrence la plus faible parmi les probabilités d'occurrence des codons de X_0 convergents, que ce soit avec E_1 (tableau 5.10) ou E_2 (tableau 5.11). Les probabilités d'occurrence des codons complémentaires convergents de X_0 sont toutes supérieures ou égales à 0.025. Concernant le seul codon appartenant à X_2, nous pouvons noter qu'il est le moins fréquent des codons stabilisés avec E_2 (tableau 5.10) et qu'il est légèrement plus fréquent que le codon TAC avec E_1 (tableau 5.10).

Disparition d'ensembles de codons Pour les deux ensembles E_1 et E_2 de paramètres de substitution maximisant la probabilité d'occurrence du code circulaire X_0, les ensembles X_1 et X_3 ont une probabilité d'occurrence nulle à la convergence du processus d'évolution (figures 5.8 et 5.9). La disparition de l'ensemble X_1 entraîne la perte de la propriété $\mathcal{C}^3$ du code circulaire X_0 (définition 1.14 page 22). Par ailleurs, pour les deux ensembles E_1 et E_2, X_2 est représenté dans les deux cas par un seul codon : TGG (tableaux 5.10 et 5.11).

Acides aminés codés Les trinucléotides appartenant au code circulaire X_0 codent pour 12 acides aminés différents. Le tableau 5.12 donne les ensembles de trinucléotides codant chacun des 12 acides aminés. Plusieurs trinucléotides de X_0 codent pour un même acide aminé, donc le codage des acides aminés par X_0 est non bijectif. De façon inattendue, l'ensemble des 9 codons obtenus à l'équilibre du processus inverse avec chacun des ensembles de paramètres de substitution E_1 et E_2 codent pour des acides aminés différents. En effet, à l'équilibre du processus avec E_1, les codons convergents de X_0 AAT, ATT, CAG, GAA, GCC, GGC, TTC et TAC codent respectivement pour les acides aminés Asn, Ile, Gln, Glu, Ala, Gly, Phe et Tyr

Codons de X_0 correspondants	Acide aminé
GCC	***Ala***
AAC, **AAT**	**Asn**
GAC, GAT	Asp
CAG	***Gln***
GAA, GAG	***Glu***
GGC, GGT	***Gly***
ATC, ***ATT***	***Ile***
CTC, *CTG*	*Leu*
TTC	***Phe***
ACC	Thr
TAC	***Tyr***
GTA, GTC, GTT	Val

TABLE 5.12 – Bijection du code circulaire « primitif » avec les acides aminés codés. Les codons convergents de X_0 obtenus avec E_1 et les acides aminés correspondants sont en gras, ceux obtenus avec E_2 sont en italique et l'intersection de E_1 et E_2 est en gras et italique.

(tableau 5.12). Avec l'ensemble de paramètres de substitution E_2, le codon AAT est remplacé par CTG, conduisant à la liste des acides aminés codés suivante : Leu, Ile, Gln, Glu, Ala, Gly, Phe et Tyr.

Courbes d'évolution Les figures 5.8 et 5.9 donnent les courbes d'évolution des probabilités d'occurrence des 4 ensembles de codons X_0, X_1, X_2 et X_3 obtenues respectivement avec les ensembles de paramètres de substitution E_1 et E_2.

Les courbes d'évolution des probabilités d'occurrence des codons de X_0 convergents lors du processus d'évolution inverse sont présentées dans les figures 5.10 et 5.11, respectivement avec les ensembles de paramètres de substitution E_1 et E_2.

Contrairement aux courbes d'évolution obtenues avec les modèles analytiques linéaires (logiciel de recherche SEGM), les courbes générées par le modèle d'évolution inverse avec blocage dynamique des substitutions ont une allure « en dents de scie ».

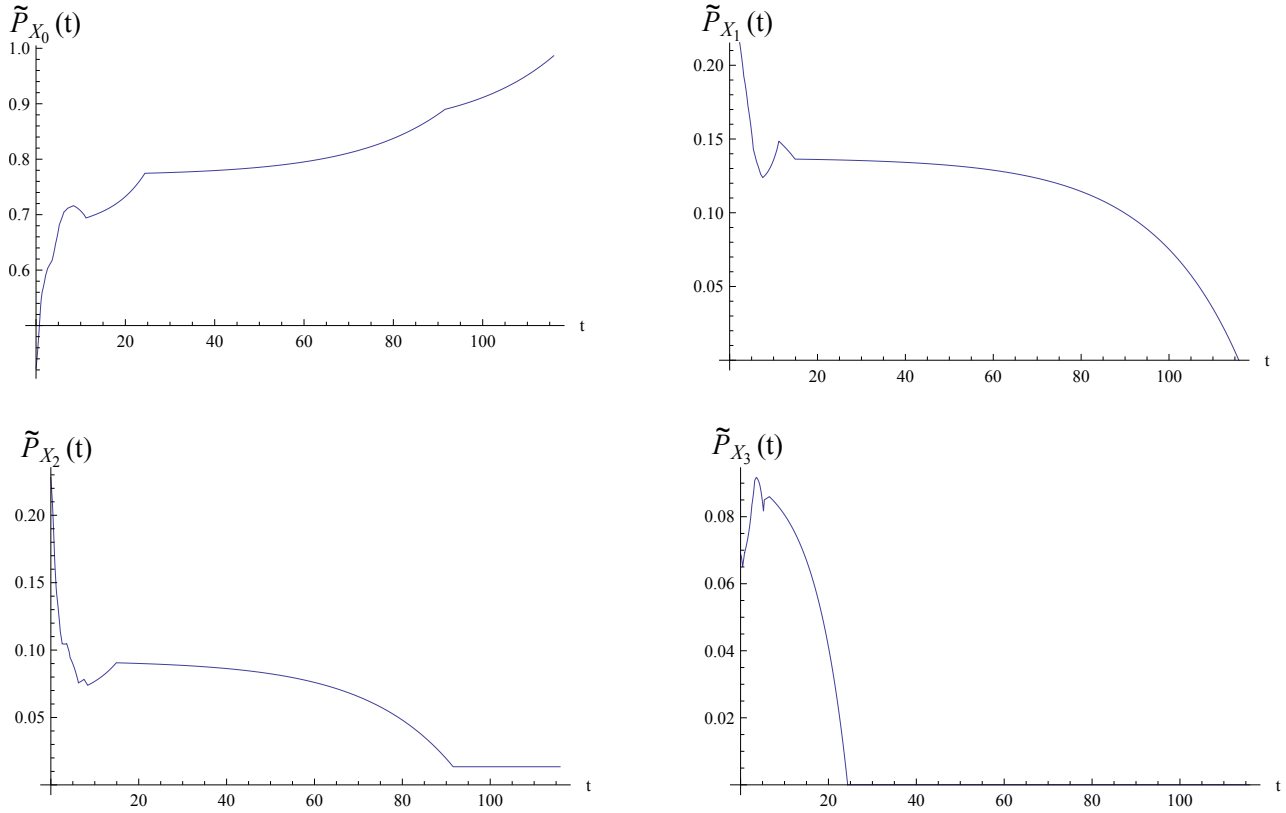


FIGURE 5.8 – Courbes d'évolution des probabilités d'occurrence des 4 ensembles de codons X_0, X_1, X_2 et X_3 obtenues avec l'ensemble des paramètres de substitution E_1 (tableau 5.9) au cours du processus d'évolution inverse avec blocage dynamique des substitutions. L'abscisse représente le temps d'évolution inverse t et les probabilités d'occurrence sont données en ordonnée.

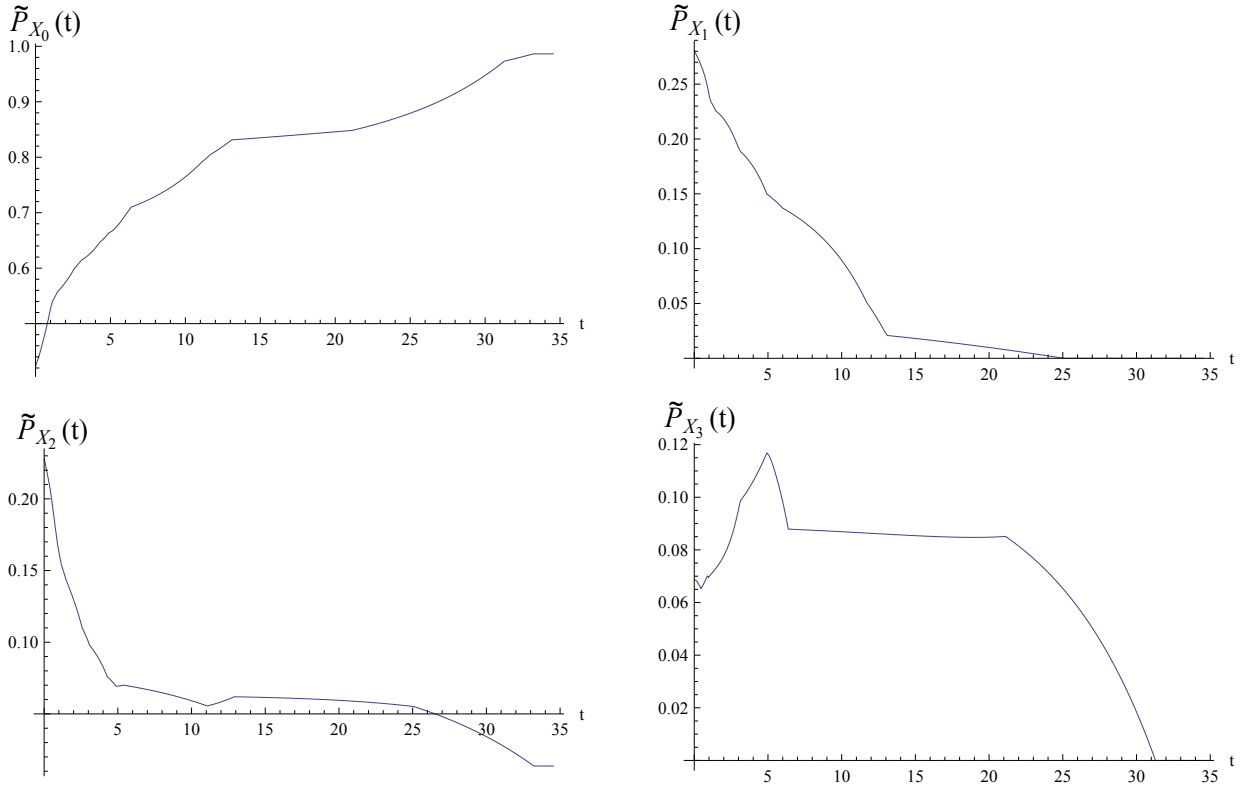


FIGURE 5.9 – Courbes d'évolution des probabilités d'occurrence des 4 ensembles de codons X_0, X_1, X_2 et X_3 obtenues avec l'ensemble des paramètres de substitution E_2 (tableau 5.9) au cours du processus d'évolution inverse avec blocage dynamique des substitutions. L'abscisse représente le temps d'évolution inverse t et les probabilités d'occurrence sont données en ordonnée.

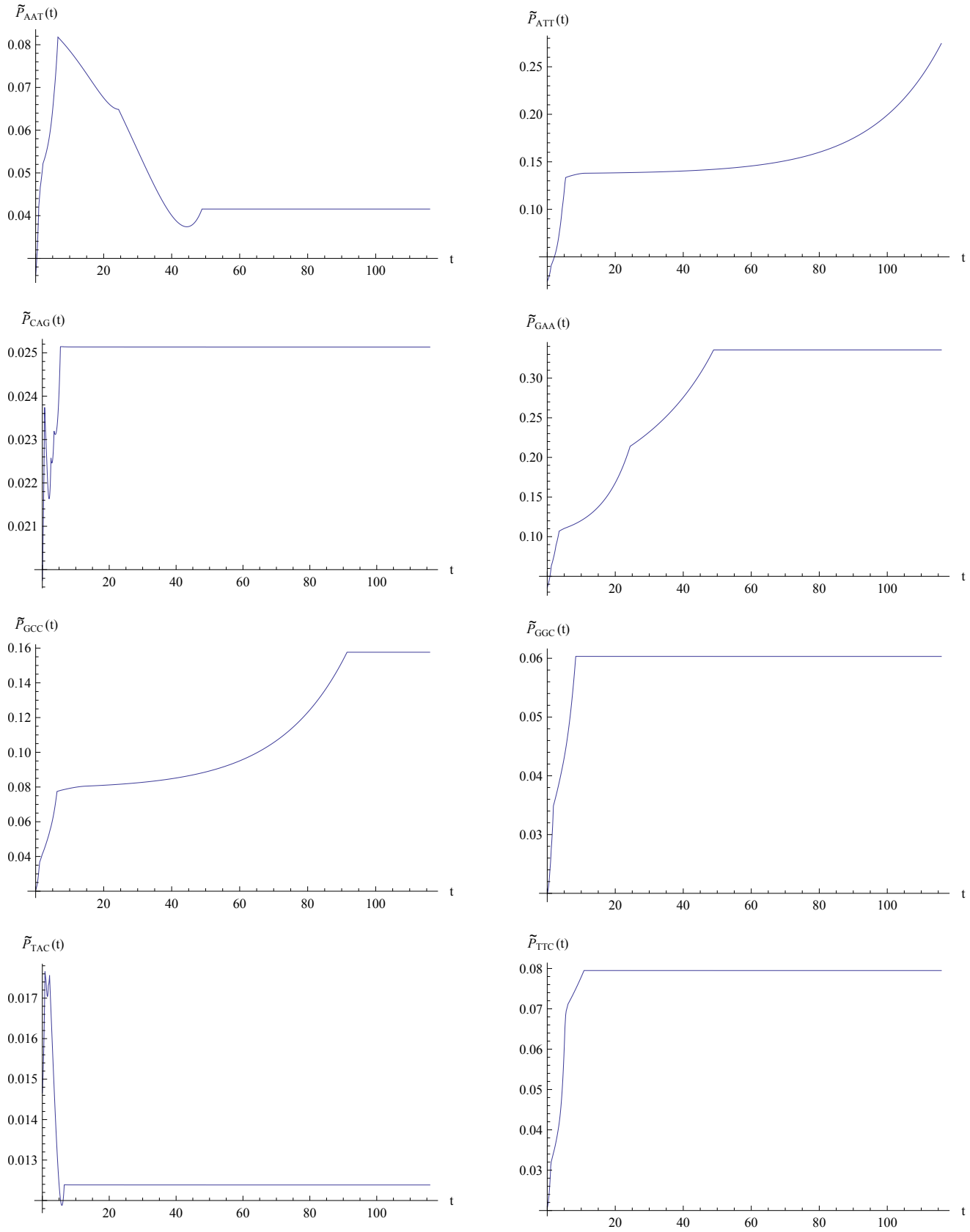


FIGURE 5.10 – Courbes d'évolution des probabilités d'occurrence des 8 codons convergents de X_0 lors du processus d'évolution inverse avec blocage dynamique des substitutions, avec l'ensemble des paramètres de substitution E_1 (tableau 5.9). L'abscisse représente le temps d'évolution inverse t et probabilités d'occurrence sont données en ordonnée.

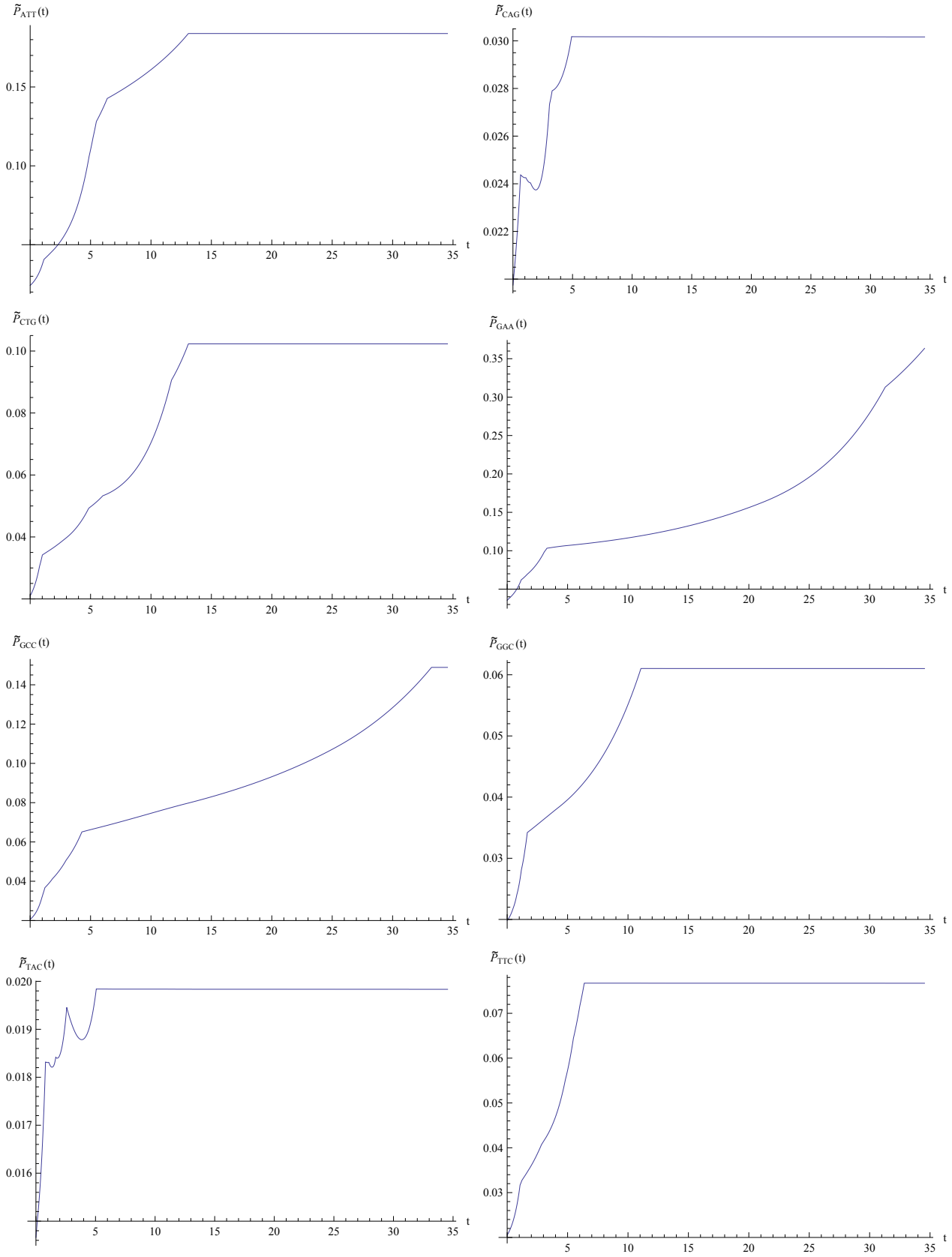


FIGURE 5.11 – Courbes d'évolution des probabilités d'occurrence des 8 codons convergents de X_0 lors du processus d'évolution inverse avec blocage dynamique des substitutions, avec l'ensemble des paramètres de substitution E_2 (tableau 5.9). L'abscisse représente le temps d'évolution inverse t et probabilités d'occurrence sont données en ordonnée.

5.2.4 Conclusion

Nous avons appliqué le modèle numérique d'évolution inverse avec blocage dynamique des substitutions présenté dans le chapitre 2 sur les probabilités d'occurrence des trinucléotides de gènes actuels d'espèces eucaryotes et procaryotes. L'objectif de cette application est d'étudier l'évolution inverse des probabilités d'occurrence actuelles des codons du code circulaire X_0 et de tester l'hypothèse d'un sous-ensemble « primitif » de X_0 avec une probabilité d'occurrence maximale. Les résultats obtenus montrent que deux ensembles de paramètres de substitution permettent de retrouver un sous-ensemble « primitif » de 8 codons de X_0 avec une probabilité proche de 1. Par ailleurs, ce sous-ensemble possède des propriétés inattendues telles que la complémentarité de 6 codons sur 8 ainsi qu'un codage bijectif entre ces 8 codons de X_0 et 8 acides aminés codés. Le modèle numérique d'évolution inverse de trinucléotides avec blocage dynamique des substitutions ainsi que les résultats de l'étude évolutive inverse menée avec ce modèle sur le code circulaire X_0 feront prochainement l'objet d'une soumission pour publication.

Conclusion générale et perspectives

Notre étude des modèles stochastiques analytiques de substitution de nucléotides étendus à des motifs de taille quelconque a fait suite aux travaux de Michel sur l'extension de ces modèles aux dinucléotides et trinucléotides en 2006 et 2007. L'approche classique alors utilisée par Michel pour diagonaliser la matrice de substitution ne peut pas être utilisée pour des motifs de taille supérieure à 3 car elle nécessite des ressources trop importantes (temps de calcul et espace mémoire). Nous avons donc proposé une méthode reposant sur la construction de la matrice de substitution par des sommes de Kronecker. Cette méthode biomathématique permet de déterminer simplement les valeurs et vecteurs propres de la matrice de substitution pour des motifs de taille quelconque à partir des valeurs et vecteurs propres de matrices de substitution de nucléotides. De plus, contrairement à l'approche classique spécifique aux matrices de substitution de type 3PS basées sur le modèle de Kimura à trois paramètres de substitution par site, la méthode généralisée que nous avons proposée est valable pour d'autres types de matrices de substitution, comme la matrice 4P que nous avons donné en exemple au chapitre 2.

D'autre part, nous avons développé un logiciel de recherche SEGM [3] (Stochastic Evolution of Genetic Motifs) qui détermine des solutions analytiques et des courbes d'évolution de probabilités d'occurrence de motifs génétiques de taille 1 à 5 soumis à un processus d'évolution à base de substitutions aléatoires dans les sens d'évolution direct (passé-présent) ou inverse (présent-passé). Les solutions sont calculées à partir d'un vecteur de probabilités d'occurrence initiales et de paramètres de substitution définis par l'utilisateur. La version actuelle de SEGM est une extension d'une précédente version développée au début de cette thèse et qui unifiait les modèles d'évolution analytiques de substitution de nucléotides de Kimura et de dinucléotides et trinucléotides de Michel. Le modèle d'évolution analytique de dinucléotides implanté dans SEGM nous a permis d'étudier l'évolution inverse des sites d'épissage des introns du génome humain.

Par ailleurs, nous nous sommes intéressés à l'étude évolutive inverse des trinucléotides du code circulaire X(EUK_PRO) découvert dans les gènes eucaryotes et procaryotes en 1996 par Arquès et Michel. Pour cela, nous avons développé un modèle d'évolution numérique inverse de trinucléotides avec blocage dynamique des substitutions. Ce modèle permet de contourner l'obstacle de la sortie du domaine de validité des probabilités d'occurrence engendrée par l'absence de codons stop dans les séquences codant les protéines. L'application de ce modèle inverse aux codons de gènes eucaryotes et procaryotes actuels nous a permis de quantifier des ensembles de paramètres de substitution à partir desquels il est possible de retrouver un code circulaire « primitif » sous-ensemble du code circulaire X(EUK_PRO) présent dans les gènes actuels et possédant des propriétés intéressantes.

3. `http://lsiit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html`.

Perspectives

L'extension des modèles stochastiques de substitution de nucléotides à des motifs génétiques de taille quelconque que nous avons proposée est valable pour des modèles dont la matrice de substitution généralise celle du modèle 3PS. Cependant, cette méthode ne peut pas être utilisée pour des modèles dont la matrice de substitution prend en compte les fréquences stationnaires des motifs. En effet, de telles matrices ne peuvent pas être construites à partir d'une unique somme de Kronecker. Le modèle d'évolution de nucléotides HKY présenté dans le chapitre 1 ne peut donc pas être étendu aux motifs de taille quelconque par cette méthode. Nous étudions actuellement la possibilité de réaliser une extension similaire aux motifs génétiques de ce type de modèles de substitution.

De même qu'il est possible de dériver les distances analytiques entre séquences à partir des probabilités d'occurrence analytiques de motifs génétiques calculées avec le modèle de substitution 3PS comme l'a montré Michel dans [Mic07c], nous envisageons de dériver les distances analytiques pour des motifs à partir d'autres modèles de substitution, par exemple des modèles de substitution généralisant 3PS.

La version actuelle de SEGM propose le calcul de solutions analytiques et numériques de probabilités d'occurrence de motifs génétiques de taille 1 à 5 dans les sens d'évolution direct et inverse à partir d'un vecteur de probabilités d'occurrence initiales et de paramètres de substitution. Il permet également de visualiser l'évolution des probabilités d'occurrence des motifs au cours du temps à l'aide de courbes d'évolutions. À notre connaissance, ce logiciel n'a pas d'équivalent à ce jour et nous souhaiterions poursuivre son développement.

Bien que très efficace, l'utilisation d'un chargement automatisé des probabilités d'occurrence initiales au moyen d'un fichier de type tableur oblige actuellement l'utilisateur à entrer des probabilités d'occurrence initiales sous forme de nombres réels et non rationnels. Une des principales améliorations que nous souhaitons apporter à ce logiciel est la possibilité de proposer à l'utilisateur des solutions analytiques totalement formelles, c'est-à-dire avec des probabilités d'occurrence initiales également formelles, mais aussi de proposer des solutions analytiques exactes avec la gestion de probabilités d'occurrence initiales avec des nombres rationnels.

SEGM gère une première opération permettant de sommer les probabilités d'occurrence analytiques et ainsi de « sommer » les courbes d'évolution de ces probabilités d'occurrence. Nous souhaiterions ajouter plus de flexibilité dans l'analyse des courbes d'évolution en incluant d'autres opérations optionnelles telles que la recherche d'intersections, le calcul de dérivées, ou encore le calcul de la probabilité d'occurrence et de la courbe d'évolution correspondant à une expression régulière sur des motifs.

Une extension de SEGM aux motifs de taille supérieure ou égale à 6 permettrait par exemple d'étudier l'évolution de couples de codons (dicodons). Pour cela, nous étudions la possibilité de paralléliser les calculs des solutions analytiques de façon à réduire les temps de calcul en utilisant la compatibilité de Mathematica [4] avec le calcul sur GPU.

À terme, nous souhaiterions développer un logiciel basé sur SEGM capable de générer les probabilités d'occurrence et les courbes d'évolution de motifs génétiques à partir de séquences génétiques (ou de leurs numéros d'ordre dans les principales bases de données génétiques) entrées par l'utilisateur.

4. version 8 actuellement.

Annexe A

Matrice de substitution de dinucléotides A_2 construite avec la méthode de Kronecker

Construction de la matrice de substitution de dinucléotides A_2 $(16, 16)$ à partir de la somme de Kronecker de deux matrices N_1 et N_2 de taille $(4, 4)$ associées aux matrices de substitution de nucléotides respectivement aux sites 1 et 2 du dinucléotide, utilisant l'équation (2.15).

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \oplus \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix}$$

avec $d_1 = -(a_1 + b_1 + c_1)$ et $d_2 = -(a_2 + b_2 + c_2)$. Donc,

$$\begin{aligned} N_1 \oplus N_2 &= \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix} \\ &= \begin{pmatrix} d_1 I_1 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 \end{pmatrix} + \begin{pmatrix} N_2 & 0 & 0 & 0 \\ 0 & N_2 & 0 & 0 \\ 0 & 0 & N_2 & 0 \\ 0 & 0 & 0 & N_2 \end{pmatrix} \end{aligned}$$

où I_1 est la matrice identité $(4, 4)$. Alors,

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 I_1 + N_2 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 + N_2 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 + N_2 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 + N_2 \end{pmatrix}$$

$$= \begin{pmatrix}
d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 \\
c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 \\
a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 \\
b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 \\
c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 \\
0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 \\
0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 \\
0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 \\
a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 \\
0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 \\
0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 \\
0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 \\
b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 \\
0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 \\
0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 \\
0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d
\end{pmatrix} = A_2,$$

avec $d = -(a_1 + b_1 + c_1 + a_2 + b_2 + c_2)$.

Annexe B

Probabilité d'occurrence analytique du dinucléotide AG avec le modèle 3PS

Les indices $1 \leq i \leq 16$ des probabilités d'occurrence $P_i(t)$ et $P_i(0)$ représentent les dinucléotides dans l'ordre lexicographique : $P_1(t) = P_{AA}(t)$, ..., $P_{16}(t) = P_{TT}(t)$.

$$
\begin{aligned}
P_3(t) = \frac{1}{16} \big[& e^0(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) + \\
& P_9(0) + P_{10}(0) + P_{11}(0) + P_{12}(0) + P_{13}(0) + P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) + P_{11}(0) + P_{12}(0) - P_{13}(0) - P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) + P_{13}(0) - P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1)t}(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) - P_{11}(0) - P_{12}(0) - P_{13}(0) - P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& + P_9(0) + P_{10}(0) - P_{11}(0) - P_{12}(0) + P_{13}(0) + P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) - P_{11}(0) + P_{12}(0) + P_{13}(0) - P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_1+b_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) - P_{11}(0) + P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_1+c_1)t}(P_1(0) + P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) - P_7(0) - P_8(0) \\
& - P_9(0) - P_{10}(0) - P_{11}(0) - P_{12}(0) + P_{13}(0) + P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_1+c_1+a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) - P_7(0) - P_8(0) \\
& + P_9(0) + P_{10}(0) - P_{11}(0) - P_{12}(0) - P_{13}(0) - P_{14}(0) + P_{15}(0) + P_{16}(0))
\end{aligned}
$$

$$
\begin{aligned}
&+ e^{-2(a_1+c_1+a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) - P_7(0) + P_8(0) \\
&+ P_9(0) - P_{10}(0) - P_{11}(0) + P_{12}(0) - P_{13}(0) + P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
&+ e^{-2(a_1+c_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) - P_7(0) + P_8(0) \\
&- P_9(0) + P_{10}(0) - P_{11}(0) + P_{12}(0) + P_{13}(0) - P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
&+ e^{-2(b_1+c_1)t}(P_1(0) + P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) - P_7(0) - P_8(0) \\
&+ P_9(0) + P_{10}(0) + P_{11}(0) + P_{12}(0) - P_{13}(0) - P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
&+ e^{-2(b_1+c_1+a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) - P_7(0) - P_8(0) \\
&- P_9(0) - P_{10}(0) + P_{11}(0) + P_{12}(0) + P_{13}(0) + P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
&+ e^{-2(b_1+c_1+a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) - P_7(0) + P_8(0) \\
&- P_9(0) + P_{10}(0) + P_{11}(0) - P_{12}(0) + P_{13}(0) - P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
&+ e^{-2(b_1+c_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) - P_7(0) + P_8(0) \\
&+ P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0))]
\end{aligned}
$$

Annexe C

Probabilités d'occurrence analytiques des nucléotides avec le modèle 4P

Les indices $1 \leq i \leq 4$ des probabilités d'occurrence $P_i(t)$ et $P_i(0)$ représentent les nucléotides dans l'ordre lexicographique : $P_1(t) = P_A(t)$, ..., $P_4(t) = P_T(t)$.

$$\begin{aligned}
P_1(t) = \frac{1}{4}\Big(1 + e^{-2t(a_1+c_1)}\left(P_1(0) - P_2(0) - P_3(0) + P_4(0)\right) \\
+e^{-t(a_1+b_1+c_1+d_1-r_1)}\left(\left(P_1(0) - P_4(0)\right)\left(1 - \frac{b_1-d_1}{r_1}\right) + \left(P_3(0) - P_2(0)\right)\left(\frac{a_1-c_1}{r_1}\right)\right) \\
+e^{-t(a_1+b_1+c_1+d_1+r_1)}\left(\left(P_1(0) - P_4(0)\right)\left(1 + \frac{b_1-d_1}{r_1}\right) - \left(P_3(0) - P_2(0)\right)\left(\frac{a_1-c_1}{r_1}\right)\right)\Big),
\end{aligned}$$

$$\begin{aligned}
P_2(t) = \frac{1}{4}\Big(1 + e^{-2t(a_1+c_1)}\left(-P_1(0) + P_2(0) + P_3(0) - P_4(0)\right) \\
+e^{-t(a_1+b_1+c_1+d_1-r_1)}\left(-\left(P_1(0) - P_4(0)\right)\left(\frac{a_1-c_1}{r_1}\right) - \left(P_3(0) - P_2(0)\right)\left(1 + \frac{b_1-d_1}{r_1}\right)\right) \\
+e^{-t(a_1+b_1+c_1+d_1+r_1)}\left(\left(P_1(0) - P_4(0)\right)\left(\frac{a_1-c_1}{r_1}\right) - \left(P_3(0) - P_2(0)\right)\left(1 - \frac{b_1-d_1}{r_1}\right)\right)\Big),
\end{aligned}$$

$$\begin{aligned}
P_3(t) = \frac{1}{4}\Big(1 + e^{-2t(a_1+c_1)}\left(-P_1(0) + P_2(0) + P_3(0) - P_4(0)\right) \\
+e^{-t(a_1+b_1+c_1+d_1-r_1)}\left(\left(P_1(0) - P_4(0)\right)\left(\frac{a_1-c_1}{r_1}\right) + \left(P_3(0) - P_2(0)\right)\left(1 + \frac{b_1-d_1}{r_1}\right)\right) \\
+e^{-t(a_1+b_1+c_1+d_1+r_1)}\left(-\left(P_1(0) - P_4(0)\right)\left(\frac{a_1-c_1}{r_1}\right) + \left(P_3(0) - P_2(0)\right)\left(1 - \frac{b_1-d_1}{r_1}\right)\right)\Big),
\end{aligned}$$

$$\begin{aligned}
P_4(t) = \frac{1}{4}\Big(1 + e^{-2t(a_1+c_1)}\left(P_1(0) - P_2(0) - P_3(0) + P_4(0)\right) \\
+e^{-t(a_1+b_1+c_1+d_1-r_1)}\left(-\left(P_1(0) - P_4(0)\right)\left(1 - \frac{b_1-d_1}{r_1}\right) - \left(P_3(0) - P_2(0)\right)\left(\frac{a_1-c_1}{r_1}\right)\right) \\
+e^{-t(a_1+b_1+c_1+d_1+r_1)}\left(-\left(P_1(0) - P_4(0)\right)\left(1 + \frac{b_1-d_1}{r_1}\right) + \left(P_3(0) - P_2(0)\right)\left(\frac{a_1-c_1}{r_1}\right)\right)\Big),
\end{aligned}$$

avec $r_1 = \sqrt{(a_1 - c_1)^2 + (b_1 - d_1)^2}$.

Bibliographie

[ABH03] P.F. Arndt, C.B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology*, 10 :313–322, 2003.

[AFM98] D.G. Arquès, J.-P. Fallot, and C.J. Michel. An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5' and 3' regions. *Bull. Math. Bio.*, 60 :163–194, 1998.

[AM93] D.G. Arquès and C.J. Michel. Analytical expression of the purine/pyrimidine codon probability after and before random mutations. *Bull. Math. Bio.*, 55 :1025–1038, 1993.

[AM95] D.G. Arquès and C.J. Michel. Analytical solutions of the dinucleotide probability after and before random mutations. *J. Theor. Biol.*, 175 :533–544, 1995.

[AM96] D.G. Arquès and C.J. Michel. A complementary circular code in the protein coding genes. *J. Theor. Biol.*, 182 :45–58, 1996.

[AR09] E.S. Allman and J.A. Rhodes. The identifiability of covarion models in phylogenetics. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6 :76–88, 2009.

[BGP08] J. Bérard, J.-B. Gouéré, and D. Piau. Solvable models of neighbor-dependent substitution processes. *Mathematical Biosciences*, 211 :56–88, 2008.

[Bla85] B.E. Blaisdell. A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitution at the same site. *J. Mol. Evol.*, 22 :69–81, 1985.

[BM04] J.M. Bahi and C.J. Michel. A stochastic gene evolution model with time dependent mutations. *Bulletin of Mathematical Biology*, 66 :763–778, 2004.

[BM08] J. Bahi and C.J. Michel. A stochastic model of gene evolution with chaotic mutations. *J. Theor. Biol.*, 255 :53–63, 2008.

[BM09a] J. Bahi and C.J. Michel. A stochastic model of gene evolution with time dependent pseudochaotic mutations. *Bull. Math. Bio.*, 71 :681–700, 2009.

[BM09b] E. Benard and C.J. Michel. Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs) : an application to splice sites of human genome introns. *Journal of Computational Biology and Chemistry*, 33 :245–252, 2009.

[BM11] E. Benard and C.J. Michel. A generalization of substitution evolution models of nucleotides to genetic motifs. *J. Theor. Biol.*, 288 :73–83, 2011.

[BP85] J. Berstel and D. Perrin. *Theory of codes.* Academic Press, London, 1985.

[BPWW82] M.N. Brown, E.M. Prager, A. Wang, and A.C. Wilson. Mitochondrial DNA sequences of primates, tempo and mode of evolution. *J. Mol. Evol.*, 18 :225–239, 1982.

[BS86] D.Y. Burman and D.R. Smith. An asymptotic analysis of a queueing system with markov-modulated arrivals. *Oper. Res.*, 34 :105–119, 1986.

[BSS00] M. Burset, I.A. Seledtsov, and V.V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, 28 :4364–4375, 2000.

[BSS01] M. Burset, I.A. Seledtsov, and V.V. Solovyev. Splice DB : database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research*, 29 :255–259, 2001.

[CC84] S.E. Curtis and M.T. Clegg. Molecular evolution of chloroplast DNA sequences. *Mol. Biol. Evol.*, 1 :291–301, 1984.

[CGB04] A. Chong, Z. Guanglan, and V.B. Bajic. Information for the Coordinates of Exons (ICE) : a human splice sites database. *Genomics*, 84 :762–766, 2004.

[CGO57] F.H.C. Crick, J.S. Griffith, and L.E. Orgel. Codes without commas. *Proc. Natl. Acad. Sci.*, 43 :416–421, 1957.

[Cri70] F. Crick. Central dogma of molecular biology. *Nature*, 227 :561–563, 1970.

[DG00] L. Duret and N. Galtier. The covarion between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artificat. *Mol. Biol. Evol.*, 17 :1620–1625, 2000.

[DYH06] J. Darot, C.-H. Yeang, and D. Haussler. Detecting the dependent evolution of biosequences. *Res. Comput. Mol. Biol. Lect. Notes Comput. Sci.*, 3909 :595–609, 2006.

[FM70] W. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4 :579–593, 1970.

[FM06] G. Frey and C. J. Michel. An analytical model of gene evolution with 6 mutation parameters : an application to archeal circular codes. *Comput. Biol. Chem.*, 30 :1–11, 2006.

[Gal01] N. Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18 :866–873, 2001.

[GGG+80] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pavé. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8 :49–62, 1980.

[GGG+81] R. Grantham, C. Gautier, M. Gouy, R. Jacobzone, and R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research*, 9 :43–74, 1981.

[GGW58] S.W. Golomb, B. Gordon, and L.R. Welch. Comma-free codes. *Canad. J. Math.*, 10 :202–209, 1958.

[GJM04] N. Galtier and A. Jean-Marie. Markov-modulated markov chains and the covarion process of molecular evolution. *J. Comput. Biol.*, 11 :727–733, 2004.

[GLG82] T. Gojobori, W.-H. Li, and D. Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, 18 :360–369, 1982.

[Gra74] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185 :862–864, 1974.

[GWD58] S.W. Golomb, L.R. Welch, and M. Delbrück. Construction and properties of comma-free codes. *Biol. Medd. Dan. Vid. Selsk.*, 23, 1958.

[GY94] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11 :725–736, 1994.

[Had93] J. Hadamard. Résolution d'une équation relative aux déterminants. *Bull. Sci. Math. (2)*, Part 1, 17 :240–246, 1893.

[Hen05] M.D. Hendy. *Mathematics of evolution and phylogeny*, chapter Hadamard conjugation : an analytical tool for phylogenetics, pages 143–177. Oxford University Press, 2005.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22 :160–174, 1985.

[Hor07] K.J. Horadam. *Hadamard matrices and their applications.* 2007.

[Ike85] T. Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, 2 :12–34, 1985.

[JC69] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

[JP00] J.L. Jensen and A.-M. K. Pedersen. Probabilistic models of DNA sequence evolution with context dependant rates of substitution. *Adv. in Appl. Probab.*, 32 :499–512, 2000.

[Kak75] Kak. On bicirculant matrices. In Indian National Science Academy, editor, *Proc. Indian National Science Academy*, volume 41 of *A*, 1975.

[Kim80] M. Kimura. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16 :111–120, 1980.

[Kim81] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS*, 78 :454–458, 1981.

[Lan05] K. Lange. Applied probability. Springer, New York, 2005.

[Las76] J.-L. Lassez. Circular codes and synchronization. *Int. J. Comput. Syst. Sciences*, 5 :201–208, 1976.

[Lau05] A.J. Laub. *Matrix analysis for scientists and engineers.* SIAM Publications, Philadelphia, PA, 2005.

[LG98] P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res.*, 8 :1233–1244, 1998.

[LH04] G. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20 :i216–i223, 2004.

[LM10] S. Lèbre and C.J. Michel. A stochastic evolution model for residue insertion-deletion independent from substitution. *Comput. Biol. Chem.*, 34 :259–267, 2010.

[MBM+08] P. Mackiewicz, P. Biecek, D. Mackiewicz, J. Kiraga, K. Baczkowski, M. Sobczynski, and Cebrat S. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. *Lect. Notes Comput. Sci.*, 5103 :100–109, 2008.

[MG94] S.V. Muse and B.S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11 :715–724, 1994.

[Mic07a] C.J. Michel. An analytical model of gene evolution with 9 mutation parameters : an application to the amino acids coded by the common circular code. *Bull. Math. Bio.*, 69 :677–698, 2007.

[Mic07b] C.J. Michel. Codon phylogenetic distance. *Comput. Biol. Chem.*, 31 :36–43, 2007.

[Mic07c] C.J. Michel. Evolution probabilities and phylogenetic distance of dinucleotides. *J. Theor. Biol.*, 249 :271–277, 2007.

[Mou00] S. M. Mount. Genomic sequence, splicing, and gene annotation. *Am. J. Hum. Genet.*, 67 :788–792, 2000.

[MPP08] C.J. Michel, G. Pirillo, and M.A. Pirillo. Varieties of comma-free codes. *Comput. Math. Appl.*, 55 :989–996, 2008.

[NM61] M.W. Nirenberg and J.H. Matthaei. The dependence of cell-free protein synthesis in E. Coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.*, 47 :1588–1602, 1961.

[PJ01] A.-M. K. Pedersen and J.L. Jensen. A dependent-rates model and MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18 :763–776, 2001.

[Pos03] D. Posada. Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Current Protocols in Bioinformatics editorial board Andreas D Baxevanis et a.*, Chapter 6(5) :Unit 6.5, 2003.

[PWC98] A.-M. K. Pedersen, C. Wiuf, and F.B. Christiansen. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.*, 15 :1069–1081, 1998.

[ROMM90] F. Rodríguez, J.L. Oliver, A. Marín, and J.R. Medina. The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, 142 :485–501, 1990.

[SM94] P.M. Sharp and G. Matassi. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.*, 4 :851–860, 1994.

[SRH+06] N. Sheth, X. Roca, M.L. Hastings, T. Roeder, A.R. Krainer, and R. Sachidanandam. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research*, 34 :3955–3967, 2006.

[SSH90] P. Senapathy, M.B. Shapiro, and N.L. Harris. Splice junctions, branch point sites, and exons : sequence statistics, identification, and applications to genome project. *Methods Enzymol.*, 183 :252–278, 1990.

[SVP03] P. Salvador, R. Valadas, and A. Pacheco. Multiscale fitting procedure using markov modulated poisson processes. *Telecommun. Syst.*, 23 :123–148, 2003.

[Syl67] J.J. Sylvester. Thoughts on orthogonal matrices, simultaneous sign-successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers. *Philosophical Magazine*, 34 :461–475, 1867.

[Tam92] K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.*, 35 :356–366, 1992.

[Tav86] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. In Miura RM, editor, *Lectures in mathematics in the life sciences*, volume 17, pages 57–86. Providence (RI) : American Mathematical Society, 1986.

[TK81] N. Takahata and M. Kimura. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, 98 :641–657, 1981.

[TN93] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10 :512–526, 1993.

[Tro07] B.E. Tropp. *Molecular Biology : Genes to Proteins*. Jones and Bartlett Publishers, 2007.

[TS98] C. Tuffley and M. Steel. Modelling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147 :63–91, 1998.

[TS01] Y. Tian and G.P.H. Styan. How to establish universal block-matrix factorizations. *The Electronic Journal of Linear Algebra*, 8 :115–127, 2001.

[VK77] F. Vogel and M. Kopun. Higher frequencies of transitions among point mutations. *J. Mol. Evol.*, 9 :159–180, 1977.

[Wak96] J. Wakeley. The excess of transition among nucleotide substitution : new methods of estimating transition bias underscore its significance. *TREE*, 11 :158–163, 1996.

[Wel70] L.R. Welch. Walsh functions and Hadamard matrices. In *Proc. Walsh Functions Symposium*, 1970.

[WSSR07] H.-C. Wang, M. Spencer, E. Susko, and A.J. Roger. Testing for covarion-like evolution in protein sequences. *Mol. Biol. Evol.*, 24 :294–305, 2007.

[Yan93] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10 :11396–1401, 1993.

[YY99] Z. Yang and A.D. Yoder. Estimation of the transition/transversion rate bias and species sampling. *J. Mol. Evol.*, 48 :274–283, 1999.

[Zha94] A. Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, 39 :315–329, 1994.

Liste de publications

Articles en revues internationales à comité de lecture

E. Benard, C. Michel. *Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs) : an application to splice sites of human genome introns.* Computational Biology and Chemistry, p. 245-252, Vol. 33, 2009

E. Benard, C. Michel. *A generalization of substitution evolution models of nucleotides to genetic motifs.* Journal of Theoretical Biology, p. 73-83, Vol. 288, 2011

Poster dans une conférence nationale

E. Benard, C. Michel. *An extension of SEGM web server to stochastic evolution of tetranucleotides and pentanucleotides.* Session poster. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Montpellier, 2010.

Table des figures

1.1 Appariement des bases azotées de l'ADN. 14
1.2 Synthèse des protéines au sein d'une cellule eucaryote. 16
1.3 Illustration de la définition 1.7 d'un code circulaire. 20
1.4 Double décomposition d'un mot généré par un code non circulaire. 21

3.1 Courbes de comparaison et temps de calcul d'une solution analytique avec les méthodes *mixte* et *pure*. 60
3.2 Courbes de comparaison et espace mémoire utilisé avec les méthodes *mixte* et *pure* pour le calcul d'une solution analytique. 60

4.1 Page d'accueil de SEGM. 65
4.2 Exemple de fichier xls contenant un vecteur de probabilités d'occurrence initiales de trinucléotides. 66
4.3 Choix du sens d'évolution (option 1) et du nombre de paramètres de substitution par site (option 2). 67
4.4 Paramètres de substitution pour chaque site (option 3) et choix des motifs à étudier (option 4) et du format de sortie des solutions analytiques (option 4b). . 68
4.5 Options conditionnelles pour la gestion du temps. 69
4.6 Exemple de solution analytique pour un trinucléotide. 70
4.7 Exemple de courbes d'évolution inverse de probabilités d'occurrence des 4 trinucléotides ATG, AAC, TAC et TCG au cours du temps t et courbe d'évolution inverse de la somme de leurs probabilités d'occurrence. 71
4.8 Exemple de solutions numériques obtenues pour les trinucléotides ATG, AAC, TAC et TCG au temps $t = 1.4$. 71
4.9 Diagramme de flux du logiciel de recherche SEGM. 73

5.1 Sites d'épissage canoniques des introns des génomes eucaryotes. 75
5.2 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites donneurs des premiers introns (*pre*). 80
5.3 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites donneurs internes (*int*). 81
5.4 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites donneurs des derniers introns (*der*). 82

5.5 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites accepteurs des premiers introns (*pre*). 84
5.6 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites accepteurs internes (*int*). 85
5.7 Courbes d'évolution inverse des probabilités d'occurrence des 16 dinucléotides obtenues avec l'ensemble solution des paramètres de substitution pour les sites accepteurs des derniers introns (*der*). 86
5.8 Courbes d'évolution des probabilités d'occurrence des 4 ensembles de codons X_0, X_1, X_2 et X_3 obtenues avec l'ensemble des paramètres de substitution E_1 lors du processus d'évolution inverse avec blocage dynamique des substitutions. 97
5.9 Courbes d'évolution des probabilités d'occurrence des 4 ensembles de codons X_0, X_1, X_2 et X_3 obtenues avec l'ensemble des paramètres de substitution E_2 lors du processus d'évolution inverse avec blocage dynamique des substitutions. 98
5.10 Courbes d'évolution des probabilités d'occurrence des 8 codons convergents de X_0 lors du processus d'évolution inverse avec blocage dynamique. Courbes obtenues avec l'ensemble des paramètres de substitution E_1. 99
5.11 Courbes d'évolution des probabilités d'occurrence des 8 codons convergents de X_0 lors du processus d'évolution inverse avec blocage dynamique. Courbes obtenues avec l'ensemble des paramètres de substitution E_2. 100

Liste des tableaux

1.1 Table du code génétique 17

3.1 Ratio des temps de calculs d'une solution analytique avec les méthodes *pure* et *mixte*. 61

5.1 Nombre d'occurrences et probabilités d'occurrence correspondantes des 16 dinucléotides sur le site donneur pour les trois classes de sites d'épissage des introns du génome humain 77

5.2 Nombre d'occurrences et probabilités d'occurrence correspondantes des 16 dinucléotides sur le site accepteur pour les trois classes de sites d'épissage des introns du génome humain 78

5.3 Ensembles de paramètres de substitution et temps d'évolution inverse t_{max} maximisant la probabilité d'occurrence du dinucléotide GT pour les trois classes de sites d'épissage donneurs *pre*, *int* et *der*. 78

5.4 Probabilités d'occurrence des dinucléotides pour les trois classes de sites d'épissage donneurs obtenues dans le sens d'évolution inverse avec les paramètres de substitution solutions du tableau 5.3. 79

5.5 Ensembles de paramètres de substitution et temps d'évolution inverse t_{max} maximisant la probabilité d'occurrence du dinucléotide AG pour les trois classes de sites d'épissage accepteurs *pre*, *int* et *der*. 87

5.6 Probabilités d'occurrence des dinucléotides pour les trois classes de sites d'épissage accepteurs obtenues dans le sens d'évolution inverse avec les paramètres de substitution solutions du tableau 5.5. 87

5.7 Probabilités d'occurrence actuelles des 64 codons issues des séquences chromosomiques d'espèces eucaryotes et procaryotes de la base de données NRCUD. . . . 90

5.8 Probabilités d'occurrence initiales des ensembles de codons X_0, X_1, X_2 et X_3 calculées à partir des probabilités d'occurrence du tableau 5.7. 91

5.9 Ensembles des paramètres de substitution maximisant la probabilité d'occurrence du code circulaire X_0 et temps d'évolution inverse t_{max} auquel est atteinte la convergence du processus d'évolution avec blocage dynamique. 94

5.10 Probabilités d'occurrence des 9 codons convergents lors du processus d'évolution inverse avec blocage dynamique. Probabilités obtenues avec l'ensemble de paramètres de subtitution E_1 (tableau 5.9). 95

5.11 Probabilités d'occurrence des 9 codons convergents lors du processus d'évolution inverse avec blocage dynamique. Probabilités obtenues avec l'ensemble de paramètres de subtitution E_2 (tableau 5.9). 95
5.12 Bijection du code circulaire « primitif » avec des acides aminés. 96

Glossaire de biologie moléculaire

Accepteur (site) : Extrémité 3' d'un intron, généralement composée du dinucléotide GT (ou GU).

Acide aminé : Élément chimique avec un groupe acide carboxylique (COOH), un groupe amine (NH_2) et une chaîne latérale ou résidu (R) spécifique à chaque acide aminé. Les 20 acides aminés des protéines naturelles sont alanine (Ala, A), cystéine (Cys, C), acide aspartique (Asp, D), acide glutamique (Glu, E), phénylalanine (Phe, F), glycine (Gly, G), histidine (His, H), isoleucine (Ile, I), lysine (Lys, K), leucine (Leu, L), méthionine (Met, M), asparagine (Asn, N), proline (Pro, P), glutamine (Gln, Q), arginine (Arg, R), sérine (Ser, S), thréonine (Thr, T), valine (Val, V), tryptophane (Trp, W) et tyrosine (Tyr, Y).

Acide nucléique : Molécule formée de nucléotides : ADN si le sucre est le désoxyribose et ARN si le sucre est le ribose.

Adénine (A) : Une des 4 bases de l'ADN et l'ARN. Elle appartient à la famille des bases puriques (purines).

ADN (Acide désoxyribonucléique) : Acide nucléique formé de désoxyribonucléotides dont la base peut être l'Adénine, la Cytosine, la Guanine ou la Thymine. L'ADN peut être considéré comme un mot sur l'alphabet des gènes à 4 lettres {A,C,G,T}. Chaque désoxyribonucléotide comporte un acide phosphorique qui relie le carbone 3' du désoxyribose au carbone 5' du désoxyribose du désoxyribonucléotide suivant (liaison phosphodiester), sauf les 2 nucléotides terminaux dont l'un a un carbone 5' libre et l'autre un carbone 3' libre. Ces 2 nucléotides terminaux définissent l'extrémité 5' (phosphate terminal) et l'extrémité 3' (ribose terminal) de l'ADN. Par convention, l'ADN est lu dans le sens 5' vers 3'. L'ADN est généralement bicaténaire avec 2 chaînes orientées en opposition (chaînes antiparallèles : l'extrémité 5' de l'une faisant face à l'extrémité 3' de l'autre et réciproquement) et appariées par complémentarité. Les 2 chaînes de l'ADN sont enroulées l'une autour de l'autre en double hélice telle une échelle en spirale où les bases seraient portées par les barreaux et où les acides phosphoriques et les sucres, par les montants. L'ADN est le support des gènes et de l'hérédité chez tous les eucaryotes, procaryotes et virus à ADN et est transmis aux générations cellulaires futures par duplication. L'ADN peut être ADN codant ou ADN non-codant.

ADN chloroplastique : ADN bicaténaire circulaire en copies multiples dans les chloroplastes.

ADN codant : ADN transcrit soit en ARN messager traduit en protéine soit en ARN de structure non traduit en protéine. L'ADN codant possède une chaîne codante et une chaîne non-codante.

ADN cytoplasmique : ADN bicaténaire circulaire du cytoplasme des cellules des eucaryotes (ADN chloroplastique et ADN mitochondrial).

ADN eucaryotique : ADN bicaténaire essentiellement de l'ADN nucléaire associé à des protéines (histones) pour constituer les différents chromosomes des eucaryotes mais également de l'ADN cytoplasmique non associé à des histones.

ADN mitochondrial : ADN bicaténaire circulaire en copies multiples dans les mitochondries.

ADN non-codant : ADN autre que l'ADN codant, c'est à dire l'ADN non transcrit (ADN d'espacement entre les gènes, ADN répétitif) ou les introns.

ADN nucléaire : ADN bicaténaire linéaire du noyau des cellules des eucaryotes.

ADN procaryotique : ADN bicaténaire circulaire constitutif de l'unique chromosome des procaryotes.

Anticodon : Trinucléotide de l'ARN de transfert s'appariant par complémentarité avec un codon d'un ARN messager lors de la traduction.

Appariement : Formation de liaisons hydrogènes entre des bases complémentaires.

Archée : Voir Procaryotes.

ARN (Acide ribonucléique) : Acide nucléique formé de ribonucléotides dont la base peut être l'Adénine, la Cytosine, la Guanine ou l'Uracile. L'ARN peut être considéré comme un mot sur l'alphabet des gènes transcrits à 4 lettres {A,C,G,U}. Chaque ribonucléotide comporte un acide phosphorique qui relie le carbone 3' du ribose au carbone 5' du ribose du ribonucléotide suivant (liaison phosphodiester), sauf les 2 nucléotides terminaux dont l'un a un carbone 5' libre et l'autre un carbone 3' libre. Ces 2 nucléotides terminaux définissent l'extrémité 5' (phosphate terminal) et l'extrémité 3' (ribose terminal) de l'ARN. Par convention, l'ARN est lu dans le sens 5' vers 3'. L'ARN est généralement monocaténaire et est une transcription par complémentarité de la chaîne non-codante de l'ADN codant, sauf pour les virus à ARN monocaténaire ou bicaténaire qui repésentent un cas particulier de support de gènes. L'ARN est donc identique à la chaîne codante de l'ADN codant. L'ARN peut être ARN messager ou ARN de structure (non codant).

ARN de structure (non codant) : ARN non traduit en protéine (exemples : ARN de transfert, ARN ribosomique, ARN nucléaire, micro ARN).

ARN de transfert (ARNt) : ARN de structure impliqué dans la traduction de l'ARN messager (ARNm) en protéine. Les ARNt existent chez les eucaryotes (noyau, mitochondrie et chloroplaste) et les procaryotes. Chaque ARNt possède un anticodon (dans sa boucle anticodon) complémentaire du codon de l'ARN messager et un trinucléotide CCA (à son extrémité 3') lié à un acide aminé (aminoacyl-ARNt) correspondant à son anticodon.

ARN messager (ARNm) : ARN traduit en protéine. Chez les procaryotes, l'ARNm est une transcription complète de l'ADN pour la traduction de plusieurs protéines (ARNm polycistronique). Chez les eucaryotes, l'ARNm est une transcription incomplète de l'ADN pour la traduction d'une seule protéine (ARNm monocistronique). En effet, l'ARNm est obtenu à partir d'un ARN prémessager (ARNm non-mature, ARNm nucléaire) qui est une transcription complète de l'ADN comportant des exons et des introns. L'épissage des exons avec élimination des introns de l'ARN prémessager forme l'ARNm (ARNm mature, ARNm cytoplasmique). Un même ARN prémessager peut donner différents ARNm (épissage alternatif).

ARN nucléaire : ARN de structure associé à des protéines et impliqué en particulier dans l'épissage des exons.

ARN prémessager : voir ARN messager.

ARN ribosomique (ARNr) : ARN de structure associé à des protéines constituant les ribosomes et impliqué dans la traduction de l'ARN messager en protéine.

Bactérie : Voir Procaryotes.

Base : Élément chimique azoté, aromatique et basique de l'ADN et l'ARN. L'ADN a 2 bases puriques ou purines (R=Adénine ou Guanine) et 2 bases pyrimidiques ou pyrimidines (Y=Cytosine ou Thymine). L'ARN a 2 bases puriques ou purines (R=Adénine et Guanine) et 2 bases pyrimidiques ou pyrimidines (Y=Cytosine et Uracile).

Bicaténaire (double brin) : Organisation moléculaire à 2 molécules.

CDS (Coding Sequence) : partie d'un gène transcrite en ARN messager et codant pour une protéine. Les introns ne sont pas inclus dans les CDS.

Cellule : La plus petite unité structurale et fonctionnelle des espèces. La cellule est limitée par la membrane plasmique et contient le cytoplasme. Chez les procaryotes, la cellule est dépourvue de membrane nucléaire et d'organite membranaire. Chez les eucaryotes, la cellule de plus grande taille possède un noyau entouré d'une membrane nucléaire et des organites cytoplasmiques membranaires (en particulier les mitochondries). Les cellules végétales se distinguent des cellules animales en particulier par la présence de chloroplastes.

Chaîne codante : Aussi appelée chaîne non-sens ou anti-sens. Chaîne de l'ADN codant complémentaire de la chaîne non-codante transcrite. La chaîne codante est donc identique à l'ARN messager chez les procaryotes et à l'ARN prémessager chez les eucaryotes.

Chaîne non-codante : Aussi appelée chaîne sens ou matrice. Chaîne de l'ADN codant transcrite en ARN messager et complémentaire à la chaîne codante.

Chloroplaste : Organite cytoplasmique membranaire des cellules végétales des eucaryotes. L'ADN codant chloroplastique est principalement impliqué dans la photosynthèse.

Chromosome : Élément du noyau cellulaire support du génome des eucaryotes. Le chromosome est composé d'ADNs et de protéines organisés en nucléosomes formant la chromatine. Chaque espèce est caractérisée par un ensemble unique de chromosomes (caryotype). Par extension, le chromosome désigne également le génome des procaryotes et des virus : le chromosome est unique et non associé à des protéines. Les chromosomes se dupliquent pour transmettre le génome dans toutes les générations cellulaires futures. Les gènes des chromosomes sont transcrits soit en ARNs messagers traduits en protéines et soit en ARNs de structure non traduits en protéines.

Code génétique : Règles de correspondance entre les codons et les acides aminés. Comme il y a 4 bases, il y a $4^3 = 64$ codons possibles. Le code génétique est universel (à quelques exceptions près), univoque (un codon code un seul acide aminé) et dégénéré (avec 3 codons de terminaison et 20 acides aminés, plusieurs codons codent un même acide aminé).

Codon : Trinucléotide de l'ADN codant (précisément de la chaîne codante) ou de l'ARN messager spécifiant un acide aminé selon le code génétique. Le codon s'apparie par complémentarité avec l'anticodon de l'ARN de transfert porteur d'un acide aminé lors de la traduction. Le codon ATG codant la méthionine (et moins fréquemment CTG, GTG et TTG chez les procaryotes) est également le codon initiateur de la synthèse d'une protéine. 3 codons de terminaison (codons stop ou codons non-sens) terminent la synthèse d'une protéine : TAA, TAG et TGA.

Cytosine (C) : Une des 4 bases de l'ADN et l'ARN. C'est une base pyrimidique (pyrimidine).

Dinucléotide : Suite (ou séquence) de 2 nucléotides.

Distance génétique : Nombre de mutations neutres qui différencie 2 segments d'ADN homologues chez 2 espèces. En prenant l'hypothèse d'une évolution moléculaire approximativement constante, cette distance est une mesure du temps écoulé depuis la bifurcation évolutive des lignées qui a mené à ces 2 espèces (horloge évolutive).

Donneur (site) : Extrémité 5' d'un intron, généralement composée du dinucléotide AG.

Épissage : Dans des gènes des eucaryotes, élimination de l'intron séparant 2 exons de l'ARN prémessager et formation d'une liaison (épissure) entre l'extrémité terminale 3' d'un exon 5' et l'extrémité initiale 5' de l'exon suivant 3' conduisant à un ARN messager traduit en protéine. Épissage alternatif : formation de différents ARNs messagers à partir d'un même ARN prémessager permettant de produire des protéines apparentées.

Espèce : Ensemble d'êtres vivants partageant des caractères morphologiques et génétiques communs et engendrant une descendance féconde.

Eucaryote : Cellule avec un noyau entouré d'une membrane nucléaire et des organites cytoplasmiques membranaires (par exemple, mitochondries, chloroplastes) capable de se reproduire par mitose. Par extension, l'organisme unicellulaire ou pluricellulaire composé de ce type de cellule.

Exon : ADN codant des eucaryotes alternant avec des introns, transcrit en ARN prémessager, conservé dans l'ARN messager après épissage et traduit en protéine.

Gène : ADN codant transcrit soit en ARN messager traduit en une seule protéine ou en des protéines apparentées (épissage alternatif) (gène codant une ou des protéines), soit en ARN de structure non traduit en protéine (gène codant un ARN de structure). Gène continu : Gène sans intron (chez les procaryotes). Gène discontinu (fragmenté, morcelé, mosaïque) : Gène avec des exons séparés par des introns (chez les eucaryotes).

Génome : Ensemble des nucléotides d'une cellule et par extension, d'un organisme. L'importance d'un génome d'une espèce s'évalue en particulier par le nombre de nucléotides et de gènes. Le génome de l'Homme (noyau) comporterait environ 3 milliards de nucléotides organisés en environ 20000 gènes dans 23 paires de chromosomes.

Guanine (G) : Une des 4 bases de l'ADN et l'ARN. C'est une base purique (purine).

Hexanucléotide : Suite de 6 nucléotides.

Intron : ADN non-codant des eucaryotes alternant avec des exons, transcrit en ARN prémessager (non-mature) puis éliminé après épissage et donc absent de l'ARN messager (mature) et non traduit en protéine. ADN non-codant des eucaryotes situé dans un gène codant un ARN de structure, transcrit en ARN de structure non-mature puis éliminé après maturation de l'ARN et donc absent de l'ARN de structure mature. Il n'existe pas d'intron chez les procaryotes.

Micro ARN (miARN) : ARN de structure (non codant pour une protéine), de petite taille, intervenant dans la régulation de l'expression des gènes.

Mitochondrie : Organite cytoplasmique membranaire de toutes les cellules des eucaryotes. L'ADN codant mitochondrial est principalement impliqué dans la production d'énergie pour la cellule (chaîne respiratoire).

Monocaténaire (simple brin) : Organisation moléculaire à 1 molécule.

Mutation : Changement des nucléotides de l'ADN ou de l'ARN. Mutation ponctuelle (un nucléotide) : substitution, transition, transversion, insertion et délétion. Mutation séquentielle (suite de nucléotides) : Insertion, délétion, répétition (duplication en tandem ou copie adjacente), répétition inversée (copie adjacente en ordre inversé), transposition simple (conservatrice ou déplacement) et transposition réplicative (copie et déplacement). Mutation faux-sens : mutation ponctuelle d'un codon avec changement d'acide aminé. Mutation non-sens : mutation ponctuelle d'un codon spécifiant un acide aminé en un codon de terminaison. Mutation neutre (silencieuse) : mutation ponctuelle d'un codon sans modification de la fonction de la protéine (le codon muté codant soit le même acide aminé soit un acide aminé avec des propriétés physico-chimiques similaires) et ne conduisant pas à de changement phénotypique à l'origine de la sélection naturelle.

Nucléoside : Association d'une base avec un sucre (pentose). Désoxyribonucléoside si le sucre est le désoxyribose et ribonucléoside si le sucre est le ribose.

Nucléotide : Association d'un nucléoside avec un acide phosphorique. Désoxyribonucléotide si le sucre est le désoxyribose et ribonucléotide si le sucre est le ribose.

Pentanucléotide : Suite de 5 nucléotides.

Peptide : Molécule (séquence, chaîne) formée d'acides aminés. Un peptide peut être considéré comme un mot sur l'alphabet des protéines à 20 lettres {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R, S,T,V,W,Y}. Le peptide est déterminé par le code génétique lors de la traduction de l'ARN messager.

Phase de lecture : Une suite de codons dans l'ADN. Une telle suite commençant par un codon d'initiation et se terminant par un codon de terminaison unique (absence de codons de terminaison intermédiaires) est dite phase de lecture ouverte.

Phylogénie : Histoire de l'évolution des espèces à partir d'une espèce commune primitive. Arbre phylogénique (phylogénétique) : Représentation graphique d'une phylogénie sous forme d'un arbre.

Plasmide : ADN bicaténaire circulaire extrachromosomique chez certains procaryotes.

Procaryote (bactérie ou archée) : Organisme unicellulaire ou pluricellulaire (certaines bactéries) filamenteux dépourvu de membrane nucléaire et d'organite membranaire (par exemple, absence de mitochondrie).

Promoteur : Région de l'ADN où s'initie la transcription (boites CAAT et TATA).

Protéine : Molécule (séquence, chaîne) constituée par un ou plusieurs peptides (suite d'acides aminés). Une protéine peut être considéré comme un mot sur l'alphabet des protéines à 20 lettres {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}.

Purine, pyrimidine : Base purique, base pyrimidique, voir Base.

Ribosome : Unité structurale corpusculaire formée d'une grande sous-unité et d'une petite sous-unité constituées d'ARNs ribosomiques associés à des protéines existant chez tous les eucaryotes (noyau, mitochondries et chloroplastes) et les procaryotes et impliquée dans la traduction de l'ARN messager en protéine.

Substitution : Mutation ponctuelle de l'ADN ou de l'ARN avec changement d'une base en une autre base.

Tetranucléotide : Suite de 4 nucléotides.

Thymine (T) : Une des 4 bases de l'ADN. C'est une base pyrimidique (pyrimidine).

Transition : Mutation ponctuelle de l'ADN ou de l'ARN avec changement d'une purine (resp. pyrimidine) en une purine (resp. pyrimidine).

Transversion : Mutation ponctuelle de l'ADN ou de l'ARN avec changement d'une purine (resp. pyrimidine) en une pyrimidine (resp. purine).

Trinucléotide : Suite de 3 nucléotides.

Uracile (U) : Une des 4 bases de l'ARN. C'est une base pyrimidique (pyrimidine).

Virus : Organisme parasite intracellulaire des animaux, végétaux et bactéries incapables de se reproduire hors d'une cellule hôte.

Résumé

Le premier modèle stochastique d'évolution de gènes a été développé par Jukes et Cantor en 1969. Ne prenant en compte qu'un paramètre de substitution par nucléotide, il a ensuite été généralisé, notamment avec les modèles à deux (Kimura 1980) et trois paramètres de substitution (Kimura 1981). Plus tard, ces modèles ont été étendus aux dinucléotides et trinucléotides (Frey et Michel 2006 ; Michel 2007). Ces derniers permettent d'étudier l'évolution des probabilités d'occurrence de motifs génétiques au cours du temps dans les deux sens d'évolution direct (passé-présent) et inverse (présent-passé), à l'aide de solutions analytiques fonctions de paramètres de substitution.

Dans ce travail, nous nous sommes intéressés dans un premier temps à l'extension de ces modèles stochastiques de substitution de nucléotides, dinucléotides et trinucléotides à des motifs génétiques de taille quelconque en proposant une nouvelle approche basée sur les opérateurs de Kronecker. Nous avons également consacré une partie au développement d'un logiciel de recherche SEGM (Stochastic Evolution of Genetic Motifs) permettant le calcul de probabilités d'occurrence analytiques de motifs génétiques de taille 1 à 5 et de courbes d'évolution de ces probabilités. Nous avons réalisé une étude évolutive inverse des sites d'épissage des introns du génome humain en utilisant notamment SEGM. Dans une troisième partie, nous avons développé un modèle stochastique numérique d'évolution de trinucléotides avec blocage dynamique des substitutions, puis nous l'avons appliqué à l'évolution des codons du code circulaire universel X0 découvert dans les gènes d'eucaryotes et de procaryotes en 1996 par Arquès et Michel.
Mots-clés : Modèles stochastiques de substitution, motifs génétiques, probabilités d'occurrence, solutions analytiques, courbes d'évolution, logiciel de recherche, sites d'épissage, codes circulaires.

Abstract

A first stochastic evolution model of genes was developed by Jukes and Cantor in 1969. One substitution type per nucleotide is taken into account with this model. It was then generalised, in particular with two (Kimura 1980) and three substitution parameters per nucleotide (Kimura 1981). An extension of these models to genetic motifs of size 2 (dinucleotides) and 3 (trinucleotides) was proposed by Frey and Michel in 2006 and Michel in 2007. These extended models allow to study evolution of occurrence probabilities of genetic motifs, both in direct (past-present) and inverse (present-past) evolutionary directions, thanks to analytical solutions computed according to substitution parameters.

In this work, we first focused on the extension of these stochastic substitution models for nucleotides, dinucleotides and trinucleotides to genetic motifs of any size, by using a new approach based on Kronecker's operations. A second part was dedicated to the development of a research software SEGM (Stochastic Evolution of Genetic Motifs) allowing the computation of analytical occurrence probabilities of genetic motifs of size 1 to 5 and evolution plots of these motifs. We present an application of these analytical stochastic models to the inverse evolution of splice sites of human genomes introns. In a third part, we developed a numerical stochastic evolution model for trinucleotides with dynamic substitution blockings. We present an application of this model to the inverse evolution of the universal circular code X0 (codon set) discovered in 1996 by Arquès and Michel in eukaryotic and prokaryotic genes.
Keywords : Stochastic models, substitution models, genetic motifs, occurrence probabilities, analytical solutions, evolution plots, research software, splice sites, circular codes.