REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE DES FRERES MENTOURI CONSTANTINE 1
FACULTE DES SCIENCES DE LA TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE

N° d'ordre : 10/D3C/2024
Série : 02/elect/2024

THESE

Présentée pour obtenir le diplôme de Doctorat 3ème cycle LMD en **Electronique**

Filière :

Génie biomédical

Spécialité :

Biomatériaux et dispositifs pour l'instrumentation biomédicale

Par :

LEHILAHY Mahery

THEME

ANALYSE DES SEQUENCES D'ADN PAR DES TECHNIQUES DE TRAITEMENT NUMERIQUE DU SIGNAL

Année Universitaire

2023-2024

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE DES FRERES MENTOURI CONSTANTINE 1
FACULTE DES SCIENCES DE LA TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE

N° d'ordre : 10/D3C/2024
Série : 02/elect/2024

THESE

Présentée pour obtenir le diplôme de Doctorat en 3ème cycle LMD en **Electronique**

Filière :

Génie biomédical

Spécialité :

Biomatériaux et dispositifs pour l'instrumentation biomédicale

Par :

LEHILAHY Mahery

THEME

ANALYSE DES SEQUENCES D'ADN PAR DES TECHNIQUES DE TRAITEMENT NUMERIQUE DU SIGNAL

Soutenue le : 17-03-2024

Devant le jury composé de :

Président :	**M. Benierbah Said**	Prof.	Université des Frères Mentouri Constantine 1
Rapporteur :	**M. Ferdi Youcef**	Prof.	Ecole Normale Supérieure de Biotechnologie, Taoufik Khaznadar, Constantine
Examinateurs :	**M. Lashab Mohamed**	Prof.	Université Larbi Ben M'hidi, Oum El-bouaghi
	M. Belmeguenai Aissa	Prof.	Université du 20 août 1955, Skikda
	M. Bellal Azzedine	Prof.	Université des Frères Mentouri Constantine 1
	M. Bensouici Tahar	MCA	Université des Frères Mentouri Constantine 1

Année Universitaire

2023-2024

REMERCIEMENTS

Je commence cette page de remerciements en exprimant ma profonde reconnaissance envers Dieu, le Créateur tout-puissant, pour m'avoir doté de la santé, de la persévérance et des opportunités nécessaires à l'accomplissement de ce travail de recherche. Sa grâce est infinie.

Ma plus sincère gratitude se tourne vers Monsieur **Ferdi Youcef**, Professeur à l'École Nationale Supérieure de Biotechnologie Taoufik Khaznadar, Constantine, pour avoir assumé la direction de cette thèse. Votre expertise, vos précieux conseils et votre engagement inébranlable envers la recherche ont été fondamentaux dans la réalisation de cette thèse. Votre influence positive et votre exemplarité professionnelle laisseront une empreinte indélébile dans ma mémoire.

Je tiens à exprimer ma profonde reconnaissance envers Monsieur **Benierbah Said**, Professeur à l'Université des Frères Mentouri Constantine 1, pour l'honneur que vous m'avez accordé en acceptant d'évaluer ce travail et en présidant la soutenance de cette thèse. Mes remerciements les plus sincères et mon profond respect vous sont adressés.

Je souhaite aussi exprimer ma chaleureuse gratitude envers Monsieur **Lashab Mohamed**, Professeur à l'Université Larbi Ben M'hidi, Oum El-Bouaghi, Monsieur **Belmeguenai Aissa**, Professeur à l'Université du 20 août 1955, Skikda, Monsieur **Bellal Azzedine**, Professeur à l'Université des Frères Mentouri Constantine 1, et Monsieur **Bensouici Tahar**, Maître de Conférences de classe A à l'Université des Frères Mentouri Constantine 1. Vos évaluations critiques et sérieuses de ce travail, ainsi que votre acceptation de siéger au sein du jury, sont hautement appréciées. Je vous prie, honorables membres de jury, de bien vouloir accepter mes vifs respects.

Je tiens à exprimer ma profonde reconnaissance envers les responsables du laboratoire Bioengineering de l'Ecole Normale Supérieure de Biotechnologie, Taoufik Khaznadar, Constantine, pour m'avoir accueilli au sein de cette prestigieuse équipe tout au long de l'élaboration de ce travail de recherche.

Aux enseignants du département d'électronique et du parcours « instrumentation biomédicale » de l'Université des Frères Mentouri Constantine 1, j'exprime ma gratitude pour m'avoir ouvert les portes du savoir tout au long de mes années académiques. Je vous prie d'accepter l'expression de ma profonde reconnaissance et de mon plus profond respect.

À mes amis et collègues, en Algérie et au-delà, je vous suis reconnaissant pour votre amitié sincère, votre soutien moral et votre encouragement constants. Une seule page ne saurait suffire pour vous énumérer tous, mais ces mots dans ma langue maternelle résument ma profonde gratitude : MANKASITRAKA SY MANKATELINA.

Enfin, j'adresse ma reconnaissance à ma famille pour le soutien inébranlable et la confiance accordée qui ont été des sources inestimables de motivation. Votre prière, fierté et soutien sont mes ultimes armes et piliers.

DEDICACE

« À ma mère, ma source inépuisable de motivation, cette thèse est dédiée en signe de gratitude pour ton amour inconditionnel, ton soutien indéfectible et la foi que tu as toujours eue en moi.

À mes frères et sœurs, compagnons de vie et de rêves, cette thèse est dédiée en reconnaissance de notre lien inaltérable et de l'inspiration que je puise de vous.

À mon père, qui a traversé les portes vers un monde meilleur et dont la bénédiction continue d'illuminer ma route, cette thèse est dédiée en sa mémoire. »

Votre enfant et frère, LEHILAHY Mahery.

TABLE DES MATIERES

LISTE DES FIGURES

CHAPITRE I

CHAPITRE II

CHAPITRE III

CHAPITRE IV

CHAPITRE V

ANNEXE

LISTE DES TABLEAUX

CHAPITRE I

CHAPITRE II

CHAPITRE III

CHAPITRE V

LISTE DES ABREVIATIONS

ACP : Amplification en Chaîne par Polymérase

ADN : Acide Désoxyribonucléique

ANF : Anti-Notch Filter

ARMA : Autoregressive Moving Average

ARN : Acide Ribonucléique

Arnm : Acide Ribonucléique Messager

Arnpn : Acide Ribonucléique Polynucléaire

Arnt : Acide Ribonucléique de Transfert

AUC : Area Under the Curve

BLAST : Basic Local Alignment Search Tool

CA : Corrélation Approximative

CDS : Coding Sequence

CGR : Chaos Game Representation

COF : Calcul d'Ordre Fractionnaire

CSANF : Conjugate Suppression Anti-Notch Filter

DDJB : DNA Data Bank of Japan

DIT-FFT : Decimation In Time Fast Fourier Transform

DSP : Densité de Spectre de Puissance

EBI : European Bioinformatics Institute

ECG : Electrocardiogramme

EEG : Electroencéphalogramme

EIIP : Electron-Ion Interaction Potential

EMBL : European Molecular Biology Laboratory

EMG : Electromyogramme

EPND : Exon Prediction by Nucleotide Distribution

FARIMA : Fractional Autoregressive Integrated Moving Average

FASTA : Format de Fichier Génomique

FFT : Fast Fourier Transform

FN : Faux Négatif

FP : Faux Positif

GARMA : Generalized Autoregressive Moving Average

GCC : Genetic Code Content

GSP : Genome Signal Processing

HMM : Hidden Markov Model

HMR : Human Mouse Rat

HSANF : Harmonic Suppression Anti-Notch Filter

IMF : Instantaneous Matched Filter

INSDC : International Nucleotide Sequence Database Collaboration

IUPAC : International Union Of Pure And Applied Chemistry

MA : Moving Average

MALDI-TOF-MS : Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry

MATLAB : Matrix Laboratory

MCC : Matthews Correlation Coefficient

MEM : Minimum Entrophy Mapping

MGWT : Modified Gabor Wavelet Transform

MZEF: Michael Zhang's Exon Finder

NA-MEMD-MGWT: Noise Assisted-Multivariate Empirical Mode Decomposition - Modified Gabor Wavelet Transform

NCBI : National Center For Biotechnology Information

NGS : Next-Generation Sequencing

OPG : Optimized Gene Predictor

OSCM : Optimized Spectral Content Measure

PAM : Pulse Amplitude Modulation

PCM : Probabilité Conditionnelle Moyenne

PCR : Polymerase Chain Reaction

QPSK : Quadrature Phase Shift Keying

RIF : Réponse Impulsionnelle Finie

RII : Réponse Impulsionnelle Infinie

ROC : Receiver Operating Characteristic

SAVMD : Sinusoidal-Assisted Variation Mode Decomposition

SB : Signal-Bruit

SCM : Spectral Content Measure

SG : Savitzky-Golay

SH : Suppression des Harmoniques

SMRT : Single-Molecule Real-Time

SOM : Self-Organizing Map

SONF : Statistically Optimal Null Filter

SRM : Spectral Rotation Measure

TDF : Transformée Discrete de Fourier

TEM : Transmission Electron Microscopy

TFD : Transformée de Fourier Discrete

TIFF : Tagged Image File Format

TNS : Traitement Numérique du Signal

UCSC : University of California, Santa Cruz

UTR : Untranslated Region

VM : Variance Minimum

VN : Vrai Négatif

VP : Vrai Positif

XML : Extensible Markup Language

INTRODUCTION GENERALE

Les études en génomique, plus précisément sur l'acide désoxyribonucléique (ADN), n'ont pas commencé dans les années 50 avec la découverte de la structure à double hélice par Watson et Crick. On peut dire qu'elles ont débuté bien avant, avec les travaux de Gregor Mendel (1822-1884), "loi de Mendel", qui énonçait des règles de transmission des caractères chez les espèces diploïdes (comme les humains). Tout cela a montré le vif intérêt que l'on porte sur la compréhension du vivant [1], [2].

Comprendre le vivant en étudiant l'ADN est alors un domaine très large et exigeant. Parmi les défis toujours d'actualité figure l'identification des régions de la séquence d'ADN qui contribuent à la constitution du gène pour coder les protéines chez les eucaryotes (cellules pourvues de noyaux). Ces régions sont communément appelées les exons [2]. Les intérêts pour la bonne identification des exons sont nombreux tels que : la meilleure compréhension du fonctionnement et des rôles des gènes, la structuration de la séquence d'ADN et son annotation, l'anticipation et le suivi des probables anomalies… Ces avantages servent et révolutionnent par la suite d'autres branches scientifiques telles que : la médecine, la génétique, la biologie, l'agronomie, l'archéologie et tant d'autres domaines. En effet, dans le cas de la médecine, l'information génétique, notamment l'identification des exons, est au centre des préoccupations des chercheurs afin de pister les gènes pouvant causer des maladies et anomalies héréditaires [3].

Cependant, les programmes et techniques d'identification des exons utilisés actuellement sont pour la plupart des technologies biologiques complexes en termes de principes de fonctionnement, paramétrage et propriétés biologiques prises en compte. Ces méthodes sont onéreuses, demandant beaucoup d'expérimentations et sont surtout réservées aux spécialistes du domaine concerné [3], [4].

Dans la perspective de vouloir mieux comprendre ce monde du génomique, d'autres domaines scientifiques comme les mathématiques (modèles markoviens, processus stochastiques) et l'informatique (apprentissage machine, bioinformatique) apportent leur contribution. La première et grande contribution est la numérisation de la séquence d'ADN par des techniques de conversion afin de pouvoir les exploiter. Des dizaines de techniques de numérisation sont disponibles. Cependant, le choix de méthodes d'exploitation de ces données numériques reste encore à proposer [5].

Il existe des méthodes qui permettent de faire l'identification des exons mais le nombre des propriétés biologiques prises en compte nécessitent une profonde connaissance et des prérequis dans le domaine.

Depuis près de deux décennies, le domaine de l'électronique, par les techniques de traitement numérique du signal (TNS) est entré dans cette course technologique génomique. Cela a donné naissance à une nouvelle discipline de la génomique qui est l'analyse par les TNS, appelée le « Genomic Signal Processing » (GSP). A la base du GSP sont les deux propriétés observées dans le signal de la séquence d'ADN : la « 3-base periodicity » et la nature « fractale ». Les méthodes GSP développées actuellement exploitent surtout la « 3-base periodicity » qui met en évidence des pics dans les zones exoniques du spectre de la séquence d'ADN. Ces méthodes sont efficaces mais présentent des limites en termes de performance. Ces limites s'observent dans les résultats statistiques et graphiques des tests, et sont souvent causées par la complexité de leurs algorithmes, du choix de la méthode de numérisation, des paramètres des TNS utilisées, de la longueur de la séquence d'ADN et de ses exons [4], [6].

L'objectif de ce travail de thèse est alors de proposer une nouvelle approche en GSP pour le traitement des séquences d'ADN. Le travail est surtout orienté vers l'identification des exons, notamment chez les eucaryotes. L'étude se base sur la propriété spectrale « 3-base periodicity » du signal de l'ADN. Le but est de concevoir un programme dont l'algorithme est centré sur un filtre fractionnaire présentant des bonnes caractéristiques permettant d'extraire efficacement et rapidement les informations pour l'identification des exons.

Les travaux réalisés ont été rassemblés dans ce manuscrit structuré en cinq chapitres :

Le premier chapitre est dédié à la mise en contexte du sujet de thèse en introduisant la génomique, de l'ADN aux méthodes d'identification des exons. Premièrement, nous aborderons l'ADN et ses propriétés fondamentales, puis la notion de synthèse de protéine. Une brève présentation de la génomique moderne incluant les technologies d'acquisition de données génomiques, leur stockage et leur traitement. Enfin, nous parlerons des défis dans l'identification des exons en citant quelques approches majeures existantes.

Le deuxième chapitre est une suite logique du premier, en présentant les notions fondamentales du « Genomic Signal Processing » (GSP) qui nous intéresse, c'est-à-dire le traitement numérique des signaux génomiques (notamment des ADN). Premièrement, nous

présenterons les propriétés de l'ADN qui fondent le GSP. Enfin, nous détaillerons les différentes étapes de l'algorithme standard des méthodes GSP.

Le troisième chapitre présentera les deux grandes classes de méthodes GSP existantes dans la littérature en illustrant par des exemples. Nous en déduirons leurs principaux résultats et limites.

Le quatrième chapitre oriente vers la conception d'une nouvelle approche basée sur le calcul d'ordre fractionnaire (COF) et ses modèles. Premièrement, nous présenterons quelques notions de base sur les fractales, le calcul d'ordre fractionnaire (COF) et quelques modèles fractionnaires ainsi que quelques applications biomédicales. Enfin, nous détaillerons la conception du filtre fractionnaire pour l'identification des exons.

Le dernier chapitre est consacré aux simulations, résultats et discussions sur l'application du filtre fractionnaire pour l'identification des exons. Les résultats seront présentés sous forme graphique et valeurs statistiques afin de les comparer avec d'autres méthodes GSP implémentées et issues de la littérature.

CHAPITRE I - GENERALITES : DE L’ADN A LA GENOMIQUE

I.1. Introduction

Ce premier chapitre présente de manière générale les notions de base en génomique. Cette présentation brève jettera les bases nécessaires pour aborder le thème développé dans ce manuscrit. Nous commencerons par fournir des informations générales sur l'acide désoxyribonucléique (ADN) et présenterons les principes fondamentaux de la biologie moléculaire. Ensuite, nous aborderons la génomique dans sa globalité, en expliquant les concepts clés tels que le séquençage et les bases de données génomiques. Enfin, nous présenterons des notions d'identification des régions codantes (exons) et des approches utilisées dans ce domaine.

I.2. L'ADN en tant que support de l'information génétique

Par définition, l'ADN est la molécule qui contient l'information génétique d'un organisme vivant. Cette information propre définit les caractères de l'organisme vivant (forme, croissance, reproduction, hérédité…) ou en termes biologiques : le phénotype (ensemble des caractères visibles) d'un organisme vivant. L'ADN se trouve essentiellement dans le noyau de la cellule vivante. Cependant, avec les deux types de cellule : les eucaryotes et les procaryotes, sa localisation diffère. Les procaryotes ne contiennent pas de noyaux bien délimités à l'intérieur de la membrane cytoplasmique, ce qui fait que leur ADN est localisé dans une région du cytoplasme appelée nucléoïde. Tandis que pour les eucaryotes, l'ADN est contenu dans un noyau cellulaire bien délimité [7].

Chimiquement, l'ADN est une macromolécule (polymère) composée de monomères appelés nucléotides. Chaque nucléotide est formé de trois unités : une base azotée, un pentose (sucre) qui est le désoxyribose et un ou plusieurs groupes de phosphates. La base azotée peut, selon sa configuration chimique, être une purine ou une pyrimidine [1].

Les purines sont l'adénine (A) et la guanine (G) tandis que les pyrimidines sont la cytosine (C) et la thymine (T). La figure (I.1) montre les différentes configurations chimiques des pyrimidines et purines.

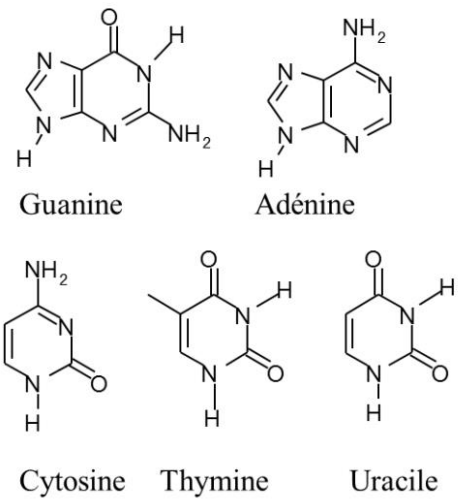


Figure I. 1 : Configuration chimique des pyrimidines et purines. Uracile (base azotée dans les acides ribonucléiques ou ARN) est aussi une pyrimidine [7].

Les groupes "base-sucre" (nucléotides) sont liés entre eux par des liaisons de groupes de phosphates et les bases se lient entre elles par des liaisons d'hydrogène. La notion de "séquence d'ADN" définit la chaîne formée par la succession des nucléotides [2].

Dans la réalité, une séquence d'ADN peut mesurer des mètres et elle est empaquetée dans des structures appelées chromosomes dans le noyau cellulaire, dont la taille est de l'ordre du micromètre. Dans le langage courant, la longueur d'une séquence d'ADN correspond au nombre de nucléotides présents dans la séquence. La figure (I.2) illustre cette opération d'empilement de l'ADN à travers différents états pour former les chromosomes [8].

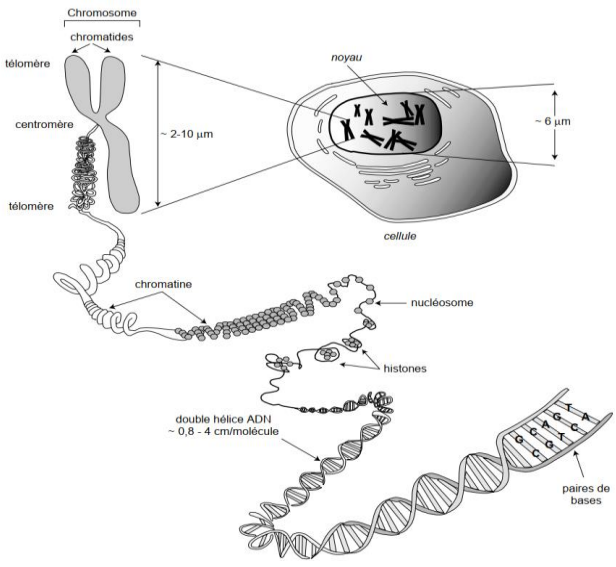


Figure I. 2 : Processus d'empaquetage de l'ADN en chromosome [8].

I.2.1. La structure en double hélice de l'ADN

Le prix Nobel de médecine a été attribué en 1953 à WATSON et CRICK pour leur découverte de la structure en double hélice de l'ADN. En effet, d'une manière très spécifique, la chaîne des nucléotides (base-sucre) forme un brin et ce dernier se lie parallèlement à un autre brin polynucléotide. Cette liaison se fait par des liaisons d'hydrogène de la manière suivante : les adénines (A) se lient uniquement avec les thymines (T) et les cytosines (C) avec les guanines (G). Les deux brins sont donc complémentaires car la séquence de nucléotides de l'un détermine celle de l'autre. En outre, la lecture de la séquence d'ADN peut se faire d'un sens à l'autre selon le brin. Ces deux brins ou chaînes de nucléotides s'enroulent sur un axe commun et donne la structure hélicoïdale ou en double hélice de l'ADN comme illustré dans la figure (I.3) [1], [9].

A l'observation, la configuration la plus courante dans les conditions physiologiques standards est appelée "forme B", là où chaque brin fait un tour complet tous les 3.4 nm sur un axe quasi perpendiculaire et chaque tour complet contient dix nucléotides comme montré dans la figure (I.3). [7], [10].

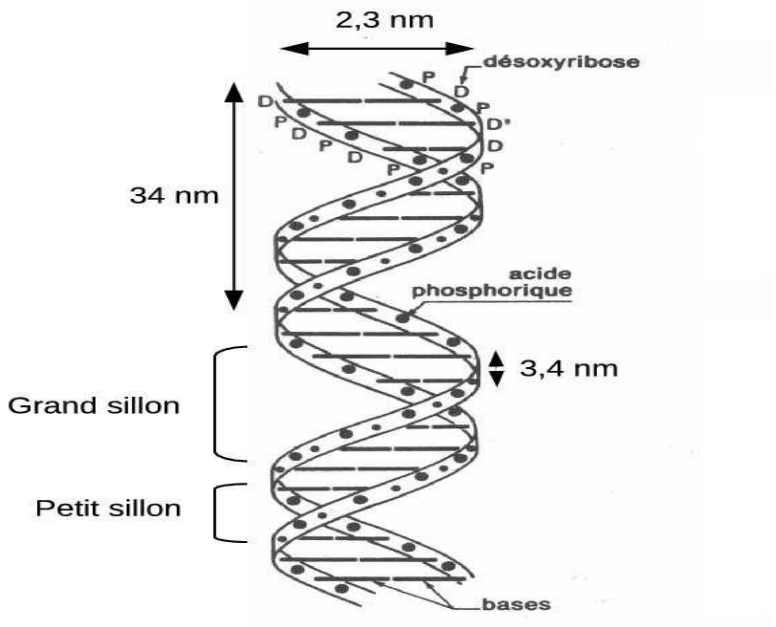


Figure I. 3 : Schéma de la structure en double hélice de l'ADN (forme B) [10].

I.2.2. Les régions de la séquence d'ADN

Auparavant, la séquence d'ADN était subdivisée seulement en deux régions : la région codante des protéines (exons) et la région non codante. Cette dernière fut longtemps décrite

comme ADN poubelle (Junk DNA) par le simple fait qu'il ne code pas directement les protéines [11]. Cependant, avec l'avancée des recherches en biologie moléculaire, une séquence d'ADN peut être subdivisée en plusieurs régions ou zones. Les proportions générales de ces régions dans une séquence d'ADN peuvent varier en fonction du gène spécifique, de l'organisme et du contexte biologique. Dans le génome humain, selon le projet ENCODE, environ 1,5 % des 3,3 milliards de paires de bases (nucléotides) sont dans la région codante (exon), tandis que 80 % de la séquence d'ADN ont des fonctions biochimiques spécifiques pouvant aider à la synthèse des protéines. La figure (I.4) illustre la proportion des régions de la séquence d'ADN dans le génome humain. En général, les régions codantes (exons) représentent une petite partie de l'ensemble de la séquence d'ADN, tandis que les régions non codantes (introns, régions régulatrices, …) peuvent occuper une plus grande partie, surtout chez les eucaryotes [3], [12].

Dans une séquence d'ADN, on peut alors trouver les régions suivantes :

Région promotrice : une partie de la séquence d'ADN située en amont d'un gène. Elle contient des éléments régulateurs, tels que des promoteurs de base et des sites d'activation, qui interagissent avec les facteurs de transcription. En effet, ces derniers se lient à la région promotrice pour initier la transcription du gène. La longueur de la région promotrice peut varier, mais généralement elle est d'environ 100 à 1000 paires de bases (pb) [2], [3].

Région terminatrice : une région qui marque la fin d'un gène et joue un rôle dans la terminaison de la transcription. Elle peut contenir des séquences spécifiques qui signalent à l'ARN polymérase de se détacher de l'ADN et de terminer la transcription. La longueur de la région terminatrice est généralement de quelques dizaines à quelques centaines de paires de bases [2], [3].

Région codante ou exon : une partie de la séquence d'ADN qui contient les codons (triplet, tri-nucléotides), c'est-à-dire les triplets de nucléotides spécifiques qui codent pour les acides aminés. Ce sont les parties codantes de l'ADN qui restent après l'épissage (élimination) des introns. Les codons sont traduits en séquences d'acides aminés lors de la synthèse des protéines. La longueur de la région codante dépend du nombre de codons nécessaires pour coder une protéine spécifique. Elle peut varier de quelques centaines à plusieurs milliers de paires de bases. La taille des exons varie selon les espèces. Cependant, leur proportion dans le génome est assez faible, de l'ordre de 1 % chez les humains (figure I.4) avec une moyenne de 150

nucléotides. Bien que le nombre moyen d'exons soit de 9, il existe des exceptions, comme celui avec 17 106 nucléotides dans le gène humain codant la protéine titine (avec près de 365 exons) [2], [3].

Intron : une région non codante de l'ADN présente seulement dans les gènes eucaryotes. Lors de la transcription, ces régions sont transcrites en ARN messager (ARNm) mais ne sont pas traduites en protéines. Les introns sont ensuite éliminés par un processus appelé épissage, laissant seulement les exons qui sont traduits en protéines. Les introns peuvent varier en longueurs, allant de quelques dizaines à des milliers de paires de bases [2], [3].

Régions régulatrices : des zones qui sont impliquées dans la régulation de l'expression des gènes. Ces régions modulent l'activité des gènes en fonction de signaux internes ou externes. Les régions régulatrices peuvent varier en longueur, allant de quelques dizaines à des milliers de paires de bases [2], [3].

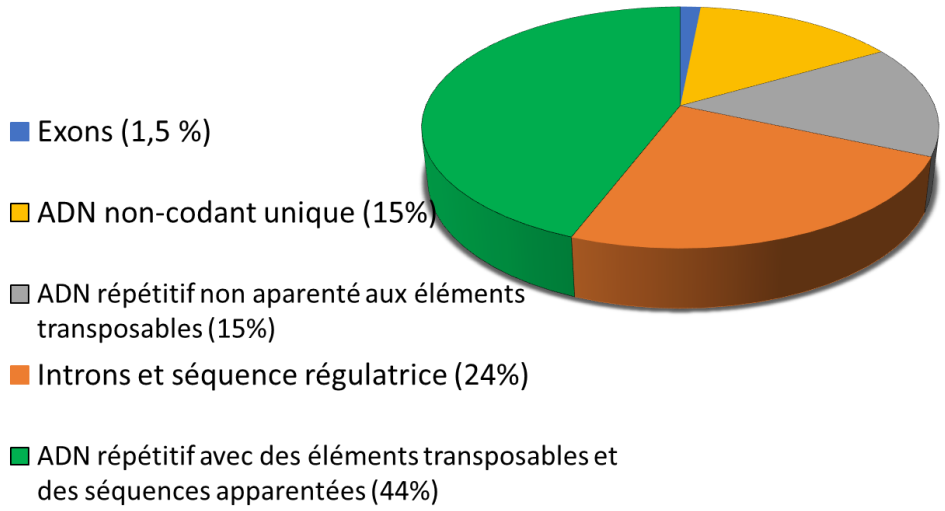


Figure I. 4 : Estimation des types d'ADN dans le génome humain [2], [13].

I.2.3. Notion de gène et de génome

Un gène est une partie de la séquence d'ADN qui va coder une protéine. Les gènes sont responsables de la transmission de l'information génétique d'une génération à l'autre et ils contrôlent l'expression des caractères héréditaires chez les individus. Ils peuvent être activés ou désactivés en fonction des besoins de l'organisme [1], [2]. La manifestation visible du génotype (ensemble des caractères héréditaires) est appelée « phénotype » et cela peut dépendre à la fois du gène mais aussi de l'environnement [39].

Le génome, quant à lui, est l'ensemble de l'ADN d'un organisme, y compris ses gènes et ses séquences non codantes. Le génome contient toutes les informations nécessaires pour le développement, la croissance, la reproduction et la survie de l'organisme. Chez les organismes eucaryotes, le génome est organisé en plusieurs chromosomes, tandis que chez les procaryotes, tels que les bactéries, le génome est plus simple, organisé en un seul chromosome circulaire [3], [4]. Les procaryotes contiennent environ 4000 gènes tandis que chez les eucaryotes, on peut avoir plus de 30 000 gènes. Si on prend l'exemple du génome humain, il comprend approximativement trois milliards de paires de nucléotides et 30 000 gènes environ. Ces gènes ont des longueurs et des rôles différents, seule une partie intervient directement à la synthèse des protéines [38].

I.3. Dogme central en biologie moléculaire

La figure (I.5) illustre la théorie fondamentale en biologie moléculaire, telle qu'elle a été comprise auparavant. Dans cette théorie, c'est l'ADN qui dirige sa propre réplication, sa transcription en ARN et sa traduction en protéine, dans un sens unique. Ce processus de synthèse des protéines sera présenté dans les points suivants. Le terme "dogme central" a été introduit dès l'observation de ce processus. Cependant, avec les nouvelles découvertes, notamment chez les organismes tels que les rétrovirus qui sont capables de faire le processus dans le sens inverse (ARN vers ADN), le dogme redevient théorie à l'échelle de toutes les espèces [9].

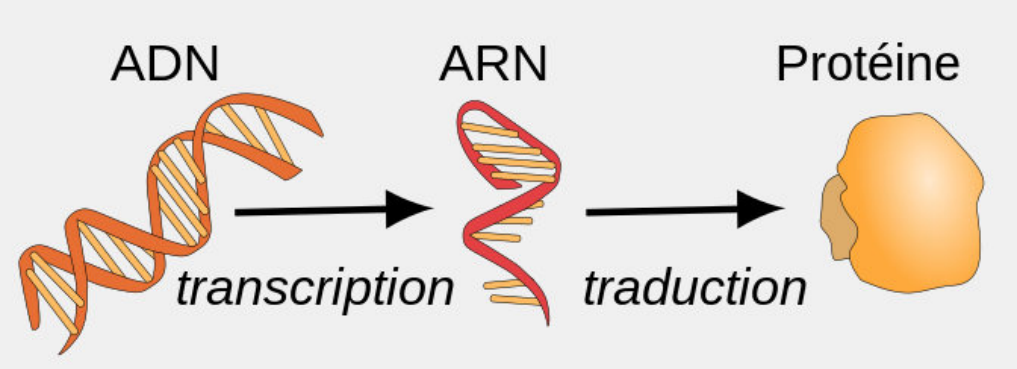


Figure I. 5 : Schéma simplifié du dogme central en biologie moléculaire [9], [14].

I.3.1. Processus de synthèse des protéines

Les protéines sont appelées les blocs constructifs des cellules et de l'organisme. Pour l'être humain et les mammifères, par exemple, ils entrent dans la composition des muscles, de la peau, des ongles, des poils, du sang… Les protéines sont également à la base de nombreuses hormones, d'enzymes et d'anticorps qui sont tous nécessaires à la croissance, la réparation et la défense des tissus du corps humain [14], [15].

Chez les eucaryotes, la synthèse des protéines est la succession de l'opération de transcription, épissage (et excision) et traduction. La figure (I.6) illustre cela.

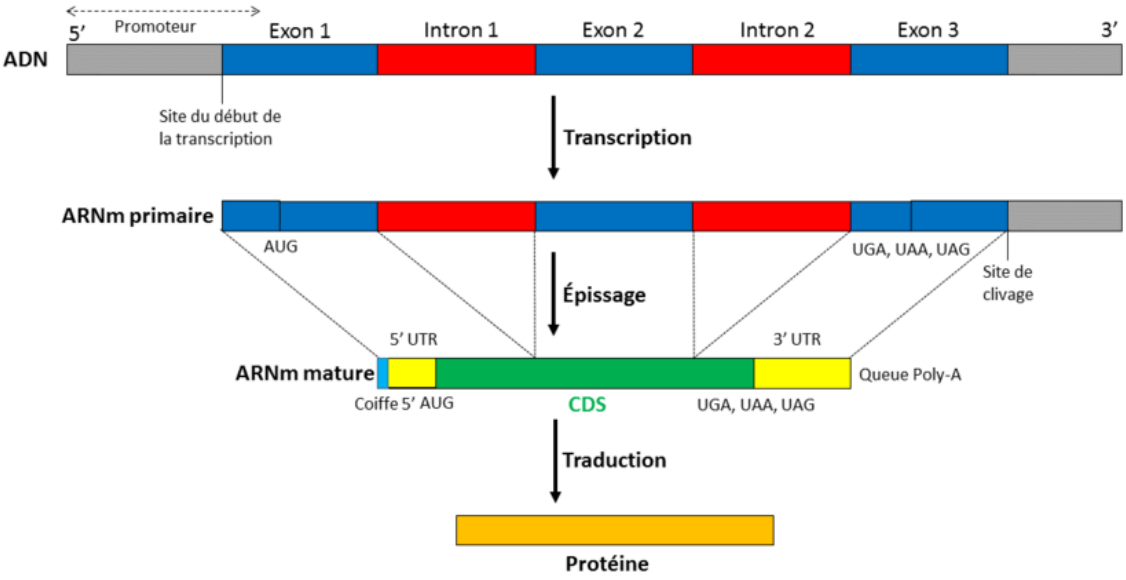


Figure I. 6 : Schéma résumant le processus de synthèse des protéines chez les eucaryotes, de l'ADN à la protéine. CDS (Coding DNA sequence) est la séquence codante proprement dite. UTR (Untranslated Region) est la région non traduite. Queue Poly-A est la polyadénylation, ajout de succession d'Adénine [14], [15].

I.3.2. Transcription

La transcription est le mécanisme de transfert de l'information génétique de l'ADN en acide ribonucléique (ARN), dans le but de synthétiser les protéines. Si la réplication traite la totalité du génome à chaque cycle de la vie cellulaire, la transcription n'est pas fixe et peut se produire à une époque précise de la vie de la cellule en fonction de différents facteurs (environnementaux et/ou du développement) [16].

La transcription se déroule en trois étapes : l'initiation, l'élongation et la terminaison. Cette terminaison n'est pas encore bien comprise chez les procaryotes tandis que chez les eucaryotes elle est mieux observable car elle est bien localisée dans le noyau cellulaire. La transcription fait appel à des activités enzymatiques comme les ARN polymérases : trois pour les eucaryotes, au lieu d'un seul pour les procaryotes. Aussi, elle se produit dans le noyau pour les eucaryotes, mais directement dans le cytoplasme pour les procaryotes. Toutefois, la transcription est assez similaire pour les deux types de cellules [17], [18].

L'ARN ainsi formé est constitué d'une chaîne de bases (nucléotides) comme l'ADN, la seule différence est que les T (thymines) de l'ADN sont remplacés par le U (uracile) dans l'ARN et aussi le pentose de l'ARN est le ribose. Cet ARN s'appelle ARN pré-messager (pré-ARNm) ou l'ARN précurseur (primaire). Il contient à la fois des régions codantes, appelées exons, des régions non codantes, appelées introns, et des régions régulatrices non-traduites (les UTR : untranslated region) [17].

I.3.3. Epissage

L'épissage est un processus essentiel dans la maturation des ARN messagers (ARNm), permettant la conversion de l'ARN pré-messager (pré-ARNm) en ARNm fonctionnel. Lors de l'épissage, il y a élimination des introns (séquences non codantes) et la jonction des exons (séquences codantes) pour générer une séquence d'ARNm mature qui peut être traduite en protéine. Chez les eucaryotes, on parle surtout d'épissage alternatif qui est un processus complexe et dynamique de maturation des ARN permettant la diversification des protéines synthétisables à partir d'un seul gène. Ce type d'épissage est très observé chez les mammifères et c'est le cas pour près de 70% des gènes humains [2].

Comme la transcription, l'épissage classique se déroule aussi suivant plusieurs étapes qui sont : la formation du spliceosome (une machinerie complexe composée de protéines et d'ARN nucléaire appelé ARNpn), la reconnaissance des sites d'épissage, l'élimination des introns par excision (réaction de transestérification), la jonction des exons par ligature pour former l'ARNm mature et enfin l'exportation de l'ARNm mature du noyau vers le cytoplasme, où il va être traduit en protéines [2], [14], [18].

I.3.4. Traduction

Après l'épissage vient l'étape de la traduction qui s'effectue au niveau du cytoplasme de la cellule. C'est l'étape-clé de l'expression génétique où la séquence d'acides aminés de la protéine est déterminée à partir de la séquence nucléotidique de l'ARNm mature. Les ARNm matures sont traduits en protéines par les ribosomes [2], [14].

La traduction se déroule généralement suivant les trois étapes suivantes : initiation, élongation et terminaison. Il existe ainsi des codons spécifiques (start et stop) permettant aux ribosomes de démarrer et d'arrêter la traduction. En effet, la traduction commence par le codon initiation AUG et se termine quand le ribosome repère l'un des codons STOP : UAA, UAG, UGA.

A part les ribosomes, il existe d'autres acteurs de la traduction qui travaillent ensemble tels que : ARN de transfert (ARNt) et les facteurs d'initiation et de terminaison qui sont des protéines impliquées dans l'initiation et la terminaison de la traduction [14].

I.3.5. Code génétique

Structurellement, les protéines sont des macromolécules constituées essentiellement de chaînes d'acides aminés liées par des liaisons peptidiques (courtes chaînes d'acides aminés) et centrées sur un atome de carbone (α-carbone). Dans l'organisme, elles se présentent sous forme d'un complexe en 3D [15].

En effet, chaque acide aminé est codé par des tri-nucléotides (triplet) appelés aussi codons. L'ensemble de ce codage s'appelle le code génétique, résultat d'un long travail qui a valu le prix Nobel en médecine et en physiologie en 1968 à M. Nirenberg, G. Khorana et R. Holley. [19] - [21]

Comme la séquence d'ADN et d'ARN messager sont représentées par la succession de 4 types de nucléotides (base azotée) : A, C, T (U pour ARNm) et G et que le codon est formé de 3 nucléotides, il existe alors 64 types de codons. De plus, le fait que 20 des 22 types d'acides aminés sont directement synthétisés par les codons rend le code génétique universel et redondant car différents codons peuvent synthétiser le même acide aminé [4].

Le tableau (I.1) suivant présente le code génétique standard.

		Deuxième nucléotide								
Premier nucléotide		U		C		A		G		Troisième nucléotide
	U	UUU	phénylalanine	UCU	sérine	UAU	tyrosine	UGU	cystéine	U
		UUC		UCC		UAC		UGC		C
		UUA	leucine	UCA		UAA	**STOP**	UGA	**STOP**	A
		UUG		UCG		UAG		UGG	tryptophane	G
	C	CUU		CCU	protine	CAU	histidine	CGU	arginine	U
		CUC		CCC		CAC		CGC		C
		CUA		CCA		CAA	glutamine	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	isoleucine	ACU	thréonine	AAU	asparagine	AGU	sérine	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	lysine	AGA	arginine	A
		AUG	**Méthionine (START)**	ACG		AAG		AGG		G
	G	GUU	valine	GCU	alanine	GAU	acide aspartique	GGU	glycine	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	acide glutamique	GGA		A
		GUG		GCG		GAG		GGG		G

Tableau I. 1 : Le code génétique qui fait la correspondance entre les codons d'ARN messager et les acides aminés [22].

I.4. La génomique et son importance dans la compréhension des organismes vivants

La génomique est un domaine scientifique axé sur l'étude des génomes d'un organisme. Le terme « génomique » a été introduit dans les années 80 pour désigner l'utilisation de techniques, de méthodes et d'approches pour l'analyse, le séquençage, l'annotation et l'interprétation des génomes. Il vise à comprendre la structure, la fonction, l'évolution et la régulation des génomes, ainsi que leurs rapports aux caractéristiques phénotypiques des organismes. L'étude du génome ouvre de nouvelles perspectives dans de nombreux domaines, y compris la médecine, la préservation de la biodiversité, les sciences de l'environnement et l'agriculture, et même l'anthropologie [23].

La génomique peut aussi être divisée en deux branches complémentaires : fonctionnelle et structurale.

- **La génomique structurale** vise à cartographier et à annoter les différentes régions du génome en utilisant des outils bioinformatiques et des techniques expérimentales pour analyser la séquence d'ADN, prédire la structure des protéines et étudier l'organisation tridimensionnelle du génome [3].
- **La génomique fonctionnelle** quant à elle, a pour but de comprendre la régulation, la traduction et les interactions des protéines avec d'autres molécules cellulaires. Elle utilise des techniques telles que la transcriptomique, la protéomique, et la métabolomique pour étudier l'expression des gènes, les modifications post-traductionnelles des protéines, et les voies métaboliques [3].

La suite de ce travail se focalise surtout sur la génomique structurale dont les informations nécessaires seront présentées dans les points suivants.

I.4.1. Génomique structurale

I.4.1.1. Cartographie : séquençage de l'ADN

Par définition, le séquençage est la méthode permettant de déterminer la succession linéaire exacte des bases azotées A, G, T, C dans une séquence d'ADN donnée. Le terme séquençage existe sous deux appellations : le « reséquençage » qui désigne le séquençage d'ADN suivi de comparaison avec une séquence référentielle, tandis que le « séquençage de novo » désigne l'étude d'une séquence inconnue. Pour le reste de ce travail, le terme séquençage sera utilisé pour généraliser le processus [24].

Depuis la fin des années 70 à nos jours, plusieurs méthodes de séquençage existent. Les défis majeurs sont : le séquençage complet du génome des organismes vivants, amélioration des résultats en réduisant les erreurs, la réduction des coûts, optimisation de la durée et du débit de séquençage…

Les méthodes existantes sont groupées en trois générations :

- **La première génération :** la méthode Sanger (par synthèse enzymatique) et la méthode Maxam-Gilbert (par dégradation chimique). Sanger et Gilbert ont eu le prix Nobel de chimie en 1980 pour leurs méthodes. Le premier organisme séquencé par la méthode de Sanger en 1977 est le virus bactériophage phiX174. La méthode a depuis permis le séquençage de divers organismes, mais c'est surtout cette méthode qui, au bout de dix ans,

avec un coût d'environ trois milliards de dollars, a permis le séquençage du génome humain dans le cadre du projet génome humain (Human Genome Project). Les principaux inconvénients du séquençage de première génération sont le débit et aussi l'utilisation de traceurs radioactifs [24], [25].

- **La deuxième ou nouvelle génération de séquençage (Next Generation of Sequencing - NGS) :** regroupe les méthodes qui ont été développées dans le but d'augmenter le rendement, d'améliorer le rapport qualité-prix, d'optimiser les résultats, mais surtout de démocratiser le séquençage génomique d'où le slogan « une séquence à 1000 USD ». Les NGS sont les plus commercialisées actuellement comme techniques de séquençage. L'utilisation de l'Amplification en Chaine par Polymérase (ACP) (Polymerase Chain Reaction, PCR) comme nouvelle méthode de préparation des séquences à analyser a contribué au progrès du séquençage à haut débit. Parmi les technologies de séquençage NGS on peut citer : le séquençage sur phase solide, la méthode Polony, le pyroséquençage, le séquençage par spectrométrie de masse (MALDI-TOF MS), le séquençage par hybridation, par ligature et le séquençage en temps réel d'une seule molécule. Récemment, des technologies de séquençage miniature utilisant les nanopores (nanotechnologie) ont été développées [7], [26], [27].
- **La troisième génération** regroupe les techniques qui sont en cours de développement, utilisables, mais peu commercialisées. On peut citer l'utilisation des semi-conducteurs pour la miniaturisation, le séquençage optique, le séquençage par microscopie électronique à transmission (TEM) et le "short to long-read shifting" qui est une technique permettant de combiner les méthodes de séquençage de courtes et longues séquences d'ADN [24], [27].

Il est aussi important de présenter le « séquenceur » qui est la machine effectuant le séquençage. Il existe deux grandes familles de séquenceurs ou plateformes de séquençage :

- **Technologie à lecture courte (short-read technology) :** famille de séquenceurs qui utilisent surtout la technologie NGS par synthèse ou par ligature pour l'étude d'une seule molécule ou bien des copies identiques de cette molécule. Illumina, Thermo Fisher Scientific et MGI sont les entreprises qui dominent le marché pour ces types de séquenceurs [27].

- **Technologie à lecture longue (long-read technology) :** dans cette famille de séquenceurs, la lecture de séquences allant de 5000 pb à 30 000 pb voire plus est possible. Ces appareils peuvent ainsi réaliser le séquençage d'une seule et longue molécule. Les sociétés Pacific Biosciences avec sa technologie SMRT (single molecule real-time) est capable de lire plus de 10 000 pb, tandis que la société Oxford Nanopore Technologies a des plateformes capables de lire des molécules allant jusqu'à 1 million de pb. Ces deux sociétés sont actuellement les novateurs pour les « long-read technology » [27].

I.4.2. Bases de données génomiques

Les premiers résultats du séquençage sont généralement optiques (image) mais peuvent aussi être non-optiques, selon la technologie de séquençage utilisée et le type de séquenceur. Si on prend l'exemple du génome humain dans la technologie Illumina, des données en images format TIFF allant jusqu'à 30 To sont d'abord obtenues puis converties par le séquenceur en format binaire (.bcl) d'environ 100 Go (où chaque nucléotide occupe 1 octet) [28]. En effet, la grande quantité de données génomiques issue du séquençage a besoin de plate-forme de stockage et d'exploitation. Avec l'avancée en bioinformatique et de l'internet, des serveurs sont utilisés par des institutions, des laboratoires de recherche, mais aussi des universités pour constituer des bases de données en génomique. Ces bases de données sont consultables en ligne, pour la plupart via les « Genome Browsers », c'est-à-dire des moteurs de recherche spécifiques [29], [30].

Actuellement, les plus importants moteurs de recherches « Genome Browsers » donnant accès aux bases de données génomiques sont :

- « National Center for Biotechnology information (NCBI) » qui est le centre national pour l'information biotechnologique aux Etats-Unis (USA) [31].
- « University of California Santa Cruz (UCSC) Genome Browser » qui est dirigée par l'université de Santa Cruz en Californie (USA) [32].
- « Ensembl Genome Browser » qui est dirigée par l'institut européen de bioinformatique EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute), en Angleterre [33].

Si on prend le cas de « GenBank » de NCBI, cette base de données a atteint un record de 241 015 745 séquences d'ADN pour près de 500 000 espèces archivées en 40 ans (1982 -

décembre 2022), ce qui fait d'elle la plus importante base de données actuellement [34]. Des collaborations existent entre les bases de données et donnent naissance à des grands projets comme le « International Nucleotide Sequence Database Collaboration » INSDC qui est une collaboration entre NCBI, EMBL-EBI et DDJB (Japon) [34], [35].

Ces moteurs de recherches (Genome Browsers) ont des propriétés et des outils plus ou moins communs afin de faciliter l'exploitation des bases de données. En effet, il existe :

- Une interface interactive pour dialoguer avec les utilisateurs.
- Une visualisation des données avec des informations complètes sur les séquences.
- Une option de recherche pour les séquences en utilisant des mots-clés : les noms scientifiques ou les numéros d'accès (ascension number).
- Un programme d'alignement des séquences qui permet d'évaluer, de comparer et calculer des statistiques de pertinence entre différentes données.
  L'outil BLAST (Basic Local Alignment Search Tool) est fréquemment utilisé dans les bases de données génomiques.
- Un catalogue des espèces répertoriées qui est mis à jour quotidiennement.
- Des articles scientifiques liés aux données génomiques.
- La possibilité de télécharger des séquences sous différents formats : FASTA, EMBL (texte), web (XML)…
- La possibilité de chargement de nouvelles séquences par les utilisateurs de la base de données.

Il existe aussi plusieurs types de formats utilisés, des formats avec ou sans annotations (informations sur la séquence). Si on prend le cas du NCBI et du GenBank, le format FASTA est le plus utilisé. C'est un format qui a été développé par William R. Pearson en 1988 et peut stocker chaque nucléotide en un octet. FASTA possède plusieurs extensions comme : «.fna, .ffn, .faa, .frn, .fa » mais l'extension usuelle est «.fasta ». C'est un format dont la première ligne du texte du fichier commence par « > » suivie de la description de la séquence (numéro d'accès, espèce, …) puis les autres lignes sont les nucléotides (pour une séquence d'ADN), comme montré dans la figure (I.7) [36], [37].

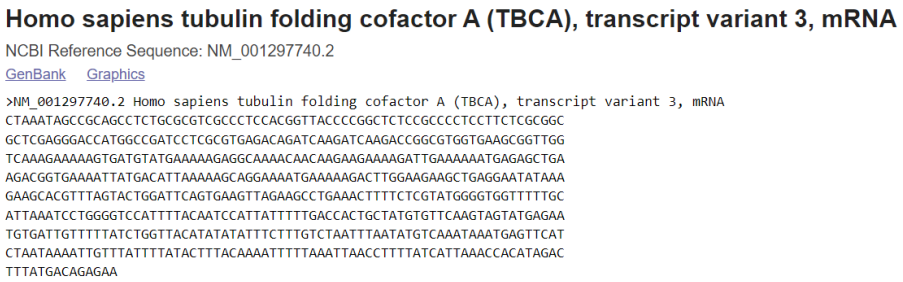
A partir des « Genome Browsers », d'autres bases de données plus réduites ont été développées dans le but de faire des études plus spécifiques. Parmi ces banques de données génomiques, citons : HMR195, BG750, Fly100, Asp67, h170 [38]-[40].

Une fois téléchargé, le format FASTA peut être lu par tout programme de traitement de texte d'un ordinateur ou par des logiciels de simulations comme : 4Peaks, GeoSpiza FinchTV, CubicDesign DNA baser, MATLAB [42].

L'union internationale de la chimie pure et appliquée (International Union of Pure and Applied Chemistry) propose le code IUPAC qui permet de nommer les acides nucléiques. Ce codage est aussi utilisé dans le format FASTA. Le tableau (I.2) montre le code IUPAC pour les nucléotides des séquences d'ADN. [43]

Code acide nucléique dans FASTA	**Signification**	
A	A	Adénine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracile
(i)	i	Iosine (non-standard)
R	A ou G (I)	purine
Y	C, T ou U	Pyrimidines
K	G, T ou U	Bases avec Cétone
M	A ou C	Bases avec groupe amine
S	C ou G	Forte interaction
W	A, T ou U	Faible interaction
B	Pas A (i.e. C, G, T ou U)	**B** vient après A
D	pas C (i.e. A, G, T ou U)	**D** vient après C
H	pas G (i.e., A, C, T ou U)	**H** vient après G
V	ni T ni U (i.e. A, C ou G)	**V** vient après U
N	A C G T U	Acide **N**ucléique
-	Espace de longueur indéterminée	

Tableau I. 2 : Code IUPAC des nucléotides des séquences d'ADN [43].



Homo sapiens tubulin folding cofactor A (TBCA), transcript variant 3, mRNA

NCBI Reference Sequence: NM_001297740.2

GenBank Graphics

```
>NM_001297740.2 Homo sapiens tubulin folding cofactor A (TBCA), transcript variant 3, mRNA
CTAAATAGCCGCAGCCTCTGCGCGTCGCCCTCCACGGTTACCCCGGCTCTCCGCCCCTCCTTCTCGCGGC
GCTCGAGGGACCATGGCCGATCCTCGCGTGAGACAGATCAAGATCAAGACCGGCGTGGTGAAGCGGTTGG
TCAAAGAAAAAGTGATGTATGAAAAAGAGGCAAAACAACAAGAAGAAAAGATTGAAAAAATGAGAGCTGA
AGACGGTGAAAATTATGACATTAAAAAGCAGGAAAATGAAAAAGACTTGGAAGAAGCTGAGGAATATAAA
GAAGCACGTTTAGTACTGGATTCAGTGAAGTTAGAAGCCTGAAACTTTTCTCGTATGGGGTGGTTTTTGC
ATTAAATCCTGGGGTCCATTTTACAATCCATTATTTTTGACCACTGCTATGTGTTCAAGTAGTATGAGAA
TGTGATTGTTTTTATCTGGTTACATATATATTTCTTTGTCTAATTTAATATGTCAAATAAATGAGTTCAT
CTAATAAAATTGTTTATTTTATACTTTACAAAATTTTTAAATTAACCTTTTATCATTAAACCACATAGAC
TTTATGACAGAGAA
```

Figure I. 7 : Exemple de format FASTA pour la séquence NM_001297740.2 qui code le protéine TBCA chez l'être humain [41].

I.4.3. Annotation des génomes

L'annotation des séquences génomiques est la caractérisation des différentes parties du génome, y compris l'identification des gènes, des régions régulatrices, des régions non codantes et d'autres éléments fonctionnels. L'un des objectifs principaux de l'annotation est d'identifier les gènes et de déterminer leurs limites, y compris les exons (régions codantes) et les introns (régions non codantes).

Il existe deux grandes familles de techniques d'annotation : expérimentale (in vivo) et in silico. L'annotation expérimentale utilise des techniques telles que le séquençage direct des ARN, la cartographie des modifications de l'ADN, pour obtenir des informations précises sur les régions du génome. L'annotation in silico, quant à elle, utilise des méthodes informatiques et des algorithmes pour prédire et attribuer des informations fonctionnelles aux séquences génomiques sans recourir à des données expérimentales directes. L'annotation in silico est largement utilisée pour l'identification des exons, la prédiction sur les séquences génomiques (des sites de liaison des facteurs de transcription, la recherche de motifs conservés, etc.) [2], [10].

I.4.3.1. Importance et défis de l'identification précise des exons

L'identification précise des exons chez les eucaryotes et procaryotes revêt une importance capitale pour comprendre les mécanismes moléculaires, annoter les fonctions génétiques et détecter les variantes génétiques. Il a été déjà montré lors de l'explication de la théorie fondamentale de la biologie moléculaire le rôle des exons pour la production de protéines fonctionnelles.

Contrairement aux eucaryotes, les procaryotes ont des gènes relativement simples avec une structure linéaire sans introns. Par conséquent, l'identification des exons chez les procaryotes est moins complexe. Pour les eucaryotes, plusieurs points rendent l'identification plus complexes. Parmi eux sont les faits que :

- Les gènes eucaryotes présentent une structure complexe avec des exons et des introns de nombres et tailles variables. Par exemple, un gène humain a une taille moyenne de 27 Kb (base) alors que le gène de la dystrophine (présent chez l'humain et les mammifères), peut atteindre plus de 2 millions de nucléotides (avec des introns allant jusqu'à 100Kb) [3].
- Certains exons chez les eucaryotes peuvent être très courts (30 à 50 nucléotides), rendant difficile leur détection uniquement sur la base de méthodes statistiques.
- Les codons de démarrage (AUG) et de fin de traduction (codons stop) ne suivent pas toujours les règles standards chez les eucaryotes, ce qui complique leur identification.
- Certains gènes eucaryotes peuvent subir des épissages non conventionnels ou présenter des traductions sans codon AUG, rendant leur identification plus difficile [2].

La suite de ce travail se consacrera beaucoup plus sur les eucaryotes.

I.4.3.2. Méthodes et programmes actuels utilisés pour l'identification des exons

Les approches actuelles d'identification des exons peuvent être groupées en deux catégories :

- Les programmes basés sur les informations de similarité des gènes connus. Dépendants des informations sur les séquences (telle que la transcriptome[1]), il est souvent difficile d'utiliser ces méthodes pour de nouvelles identifications d'exons de gènes inconnus.
- Les programmes (ab initio) qui utilisent des modèles probabilistes tels que les modèles de Markov ainsi que les machines à vecteurs de support. Ces modèles permettent par la suite d'utiliser des capteurs de signal qui exploitent des sites spécifiques de l'ADN (région d'épissage, promotrice et régulatrice) et des capteurs

[1] Transcriptome : ensemble des ARNm, produit de la transcription.

de contenus qui exploitent les caractéristiques des séquences (longueur de la séquence, les exons, les introns).

Les programmes (ab initio) sont les plus utilisés, bien que les deux types de programmes soient souvent combinés. Parmi les programmes existants, on peut citer : Genscan, GeneMark, HMMgene, AUGUSTUS, GeneID, FGENES, MZEF, Morgan, FGENSH, Snap. Parfois, la combinaison de deux programmes permet d'avoir une meilleure identification des exons : GENSCAN/FGENESH , GENESCAN/GENEHMM [44] - [48].

La figure (I.8) illustre l'utilisation du modèle de Markov dans le contexte d'identification des exons. En effet, pour une chaîne de Markov d'ordre k, la probabilité d'apparition d'une base à une position donnée dépend des bases précédentes, et les paramètres du modèle sont définis par les probabilités conditionnelles correspondantes. Ces paramètres du modèle sont différents entre les régions des exons et introns de l'ADN car les probabilités d'apparitions des bases dans ces régions sont différentes [3], [44].

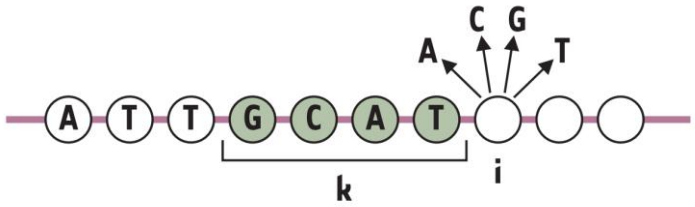


Figure I. 8 : Représentation simplifiée d'une chaine de Markov utilisée dans l'identification des exons [3].

En général, l'identification des gènes chez les organismes procaryotes par ces programmes sont satisfaisantes (avec une probabilité d'erreur inférieure à quelques % pour un gène bactérien). Cependant, la difficulté chez les organismes eucaryotes réside dans la complexité et la diversité structurelle des séquences d'ADN [3].

I.4.3.3. Méthodes alternatives basées sur des nouvelles approches

Les programmes actuels utilisent des éléments tels que les sites d'épissage des introns, les motifs de codons de début et de fin de traduction, les motifs d'acides aminés conservés et les motifs de régions régulatrices, pour prédire les exons dans les séquences d'ADN. Leurs

utilisations pour l'identification des exons montrent leur efficacité. Cependant, ces approches peuvent devenir difficiles et complexes en raison du grand nombre de paramètres à prendre en compte simultanément. La présence et la variation des signaux et motifs dans les génomes peuvent être complexes et difficiles à modéliser avec précision [3].

Une approche alternative pour faciliter l'identification des exons consiste à explorer les techniques de traitement numérique du signal (TNS). Ces techniques permettent d'analyser les séquences d'ADN en utilisant des méthodes plus simples, en se concentrant sur des caractéristiques numériques telles que les profils de fréquence, les motifs de répétition et les schémas de distribution des nucléotides.

L'utilisation de ces techniques de traitement numérique du signal (TNS) offre plusieurs avantages. Elles permettent entre autres de réduire la complexité des programmes d'identification des exons en se concentrant sur des aspects plus quantitatifs et moins dépendants de motifs spécifiques et biologiques. De plus, ces approches peuvent être plus robustes face aux variations génétiques et aux structures génomiques atypiques, offrant ainsi une meilleure précision dans l'identification des exons [6].

I.5. Conclusion

Ce premier chapitre a présenté les notions de base en génomique, allant de l'ADN aux techniques pour l'identification des exons. Nous avons vu que l'identification des exons et de leurs positions dans l'ADN est importante dans la génomique structurale. Les approches actuelles sont constamment mises à jour. D'autres domaines scientifiques s'intéressent depuis un certain temps à la génomique, notamment pour l'identification des exons dans les cellules eucaryotes. Parmi eux sont les techniques de TNS. Ces sujets seront abordés dans les chapitres suivants.

CHAPITRE II - TRAITEMENT NUMERIQUE DU SIGNAL GENOMIQUE

II.1. Introduction

Après le séquençage et l'archivage des données génomiques, plusieurs études peuvent être menées afin de mieux comprendre les informations relatives aux gènes, à commencer par leurs localisations. Pour cela, les scientifiques ont proposé diverses méthodes pour étudier plus efficacement les séquences d'ADN. Une de ces approches est l'utilisation du traitement numérique du signal (TNS) en génomique, également connu sous le nom de « genomic signal processing » (GSP). Ce chapitre présente les bases théoriques du GSP. Nous allons aborder les techniques de numérisation de la séquence d'ADN, les techniques de TNS les plus couramment utilisées. Enfin, les paramètres (critères) d'évaluation de performance des méthodes GSP seront présentés.

II.2. Traitement numérique de signal génomique

II.2.1. Définition

Le GSP est une branche de la génomique qui analyse les séquences d'ADN, ARN et protéiques en utilisant les techniques de TNS. C'est une étude post-séquençage qui consiste le plus souvent à identifier et prédire les zones codantes (les exons) des séquences d'ADN afin d'explorer par la suite les fonctions des gènes et les rôles biologiques des différents éléments de la séquence d'ADN… [49], [50]. Le GSP est apparu dans les années 70 et continue de se développer jusqu'à nos jours en étudiant surtout les gènes des eucaryotes [4].

II.2.2. Les propriétés des séquences d'ADN qui fondent le GSP

L'analyse des séquences d'ADN en utilisant les TNS se base sur des propriétés intrinsèques (pas forcement biologiques) observées dans les séquences d'ADN, surtout des cellules eucaryotes. Parmi ces propriétés : la « 3-base periodicity » et la nature « fractale ».

II.2.2.1. La propriété « 3-base periodicity »

L'analyse spectrale, en utilisant la transformation de Fourier dans l'étude des ADN, a révélé une propriété fondamentale en GSP qui est le "3-base periodicity". Cette propriété s'identifie par la présence d'un grand pic sur le spectre à la fréquence N/3 pour une séquence de longueur N dans les exons tandis que le pic est absent dans les introns [51] - [53].

Les origines de ce pic ont longtemps été débattues mais deviennent de plus en plus claires. En effet, il existe quelques hypothèses qui ont été avancées sur l'origine de la propriété "3-base periodicity".

- ***1ère hypothèse :*** le "3-base periodicity" n'est pas biaisé par le code génétique. Les expériences faites avec des gènes artificiels et naturels codant les mêmes protéines ont tous montrés que les exons se distinguent des introns par les pics présents sur le spectre [51].
- ***2ème hypothèse :*** le "3-base periodicity" ne dépend pas de l'ordre des acides aminés mais de leurs compositions. Cela a été vérifié en changeant aléatoirement les ordres des acides aminés d'un gène et en comparant le spectre de celui-ci avec le spectre du gène original. La distribution des nucléotides (A, G, T, C) dans les codons participe à l'apparition du pic dans les exons. En effet, dans un gène, les nucléotides ont différentes fréquences d'apparition en 1ère, 2ème et 3ème position d'un codon. L'étude faite sur le GroEL a par exemple montré que la fréquence d'apparition de l'adénine (A) sur les 3 positions du codon est quasi-identique alors que les G, T et C semblent apparaître aléatoirement. En faisant par la suite, l'analyse spectrale de la séquence pour chaque base, le spectre du signal avec l'adénine ne montre pas de pic dans la région exonique tandis que les 3 autres spectres (avec, G, T et C) montrent [52]. Cette expérience est illustrée dans la figure (II.1).
- ***3ème hypothèse :*** une corrélation entre la variance de la distribution des 4 nucléotides sur les 3 positions du codon et de la densité spectrale de la séquence à la fréquence f = N/3 a été prouvée. Cela se démontre en créant une matrice dont les éléments correspondent au nombre d'apparitions des nucléotides dans les 3 positions du codon. La matrice est la suivante:

$$P = \begin{pmatrix} P_{A1} & P_{A2} & P_{A3} \\ P_{G1} & P_{G2} & P_{G3} \\ P_{T1} & P_{T2} & P_{T3} \\ P_{C1} & P_{C2} & P_{C3} \end{pmatrix}$$

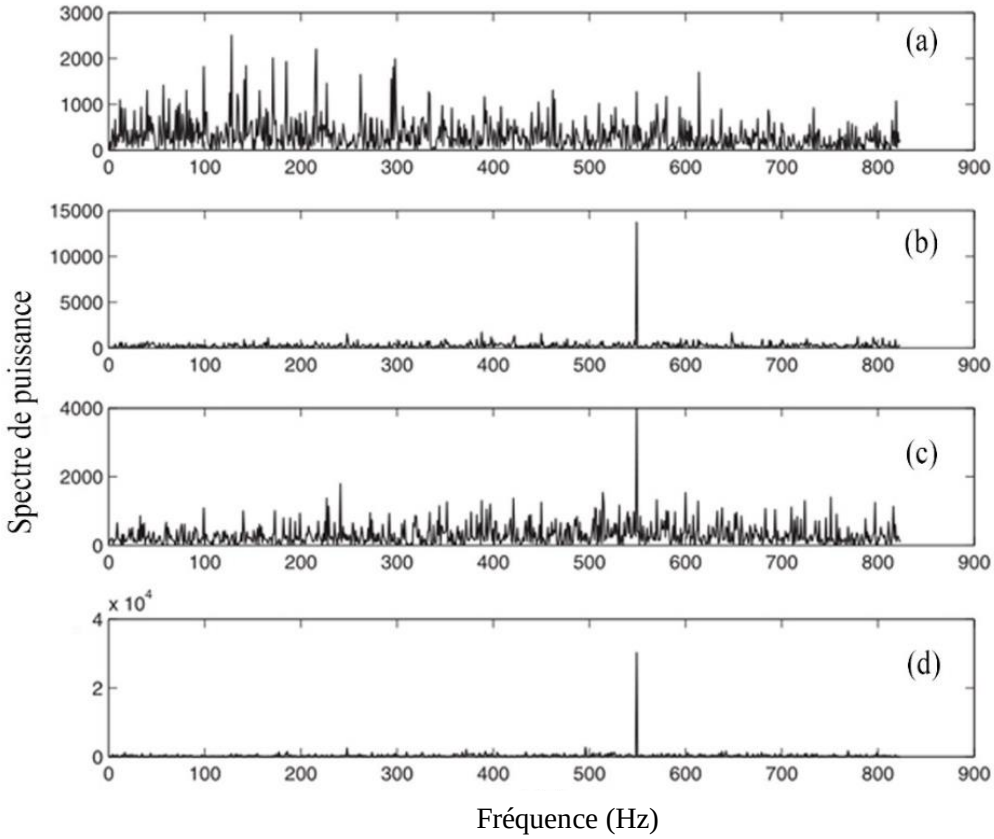


Figure II. 1 : Spectres de puissance de la transformation de Fourier discrète (TFD) de la séquence d'ADN du gène GroEL. (a) Spectre de puissance du TFD de la séquence avec adénine seule. (b) Spectre de puissance du TFD de la séquence avec thymine seule. (c) Spectre de puissance du TFD de la séquence avec cytosine seule. (d) Spectre de puissance du TFD de la séquence avec guanine seule [52].

En utilisant la matrice P, on peut calculer la variance de la distribution de chaque nucléotide par la relation (II.1) suivante :

$$\sigma_x = \sum_{i=1,2,3} \left(P_{xi} - \frac{1}{3} \sum_{j=1,2,3} P_{xj} \right)^2 \qquad \text{(II.1)}$$

Afin de calculer la somme des variances

$$P_{totale} = \sum_{x=A,G,T,C} \sigma_x \qquad \text{(II.2)}$$

De (II.1) et (II.2), on obtient

$$P_{totale} = \sum_{x=A,G,T,C} \sum_{i=1,2,3} \left(P_{xi} - \frac{1}{3} \sum_{j=1,2,3} P_{xj} \right)^2 \qquad \text{(II.3)}$$

et l'équation (II.3) devient

$$P_{totale} = \frac{2}{3} \sum_{x=A,G,T,C} (P_{x1}^2 + P_{x2}^2 + P_{x3}^2 - (P_{x1} * P_{x3} + P_{x1} * P_{x2} + P_{x2} * P_{x3})) \qquad \text{(II.4)}$$

En traçant les fonctions P_{totale} de l'équation (II.4) et le spectre de la puissance à la fréquence f=N/3, $SP_{\left(\frac{N}{3}\right)}$ pour le cas d'un gène donné, une corrélation linéaire peut être observée dans la relation (II.5) :

$$SP_{\left(\frac{N}{3}\right)} = \frac{3}{2} P_{totale} \qquad \text{(II.5)}$$

La figure (II.2) montre cette corrélation décrite dans la relation (II.5). Les séquences d'ADN utilisées dans l'expérience présentent des longueurs variables et les probabilités d'occurrence possibles des nucléotides sur chaque position du codon sont différentes [52].

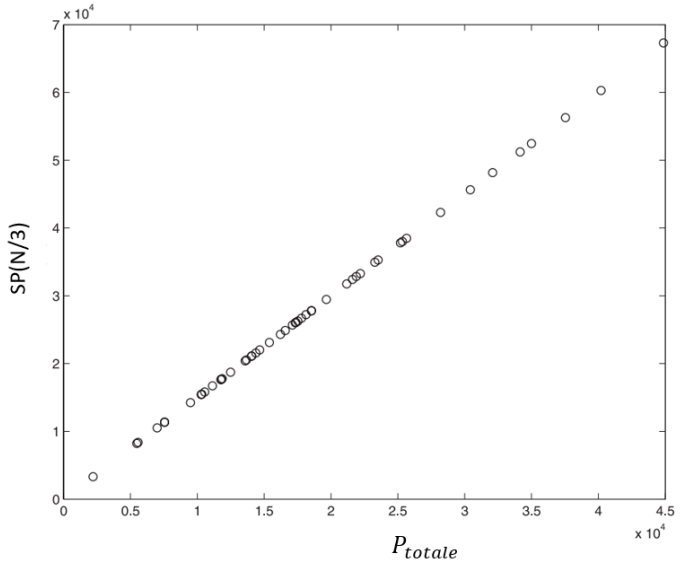


Figure II. 2 : La corrélation entre la répartition des nucléotides aux trois positions du codon et le spectre de puissance à N/3 [52].

- ***4ème hypothèse*** : l'amplitude de la densité spectrale de la séquence d'ADN dépend de sa longueur et de la variance de distribution des nucléotides P_{totale} dans les codons. Cela aussi contribue au rapport signal sur bruit car plus la séquence est longue, plus la variance de distribution des nucléotides grandit et le rapport signal sur bruit aussi [52].
- Il existe d'autres propriétés des exons et introns qui ont été mises en évidence mais qui sont moins exploitées pour l'identification des exons par le GSP. Parmi ces propriétés sont le fait que les introns sont riches en A et T tandis que les exons sont plutôt riches en G et C [54].

II.2.2.2. La propriété fractale sur les séquences d'ADN

Les travaux de Voss dans [55] abordent le sujet des corrélations fractales à longue distance (long-terme) et du bruit 1/f dans les séquences d'ADN. Les corrélations fractales font référence à la structure autosimilaire qui se trouve à différentes échelles dans les séquences d'ADN, tandis que le bruit 1/f se réfère à un type particulier de bruit dont la puissance décroît avec la fréquence selon une relation inverse. Les résultats suggèrent que les séquences d'ADN présentent des corrélations fractales à longue distance, ce qui signifie que des motifs similaires tels que les codons se répètent à différentes échelles. De plus, le bruit 1/f est observé dans ces séquences, ce qui implique une régularité à large échelle. Ces propriétés ont été renforcées par les travaux de Peng et al. [56] sur la propriété fractale des séquences d'ADN chez les eucaryotes plutôt que chez les procaryotes et notamment dans les exons [57]. Ces travaux observent en effet une corrélation entre la longueur de la séquence et ses éléments (purine-pyrimidine). Cela se fait par une linéarité observée dans la représentation log-log de la fonction (fonction de Peng) $F(l) \sim l^{\alpha}$. Avec F étant la fonction de Peng, l la largeur d'une partie de la séquence d'ADN et α l'exposant fractionnaire. D'autres études spécifient que ces propriétés des textes génomiques sont surtout visibles pour les espèces eucaryotes [58].

Les travaux de Jeffrey et Fiser et al. [59], [60] sur l'algorithme appelé "jeu de chaos" (chaos game representation, (CGR)) proposent des représentations en images fractales itératives des textes génomiques. L'objectif était de mieux observer le caractère non-aléatoire, autosimilaire des nucléotides dans la séquence d'ADN (signature génomique).

En effet, sur un plan fermé $X = [0,1]^2$, la fonction CGR est la suivante :

$$X_{n+1} = 1/2(X_n + l_{u_{n+1}}) \qquad \text{(II.6)}$$

Dans la relation (II.6), l_u représente les A, G, T, C avec $u \in \{A, G, T, C\}$ et les coordonnées initiales sont $l_A(0,0), l_G(1,1), l_T(1,0)\ et\ l_C(0,1)$ [61].

Avec le temps, le CGR a été amélioré par l'incrémentation de son ordre et la modification des coordonnées initiales du plan pour plus de précision dans l'observation des motifs similaires qui se répètent dans l'ADN [62], [63]. Par exemple, le plan CGR initial peut devenir $l_A(1,1), l_G(1,-1), l_T(-1,1)\ et\ l_C(-1,-1)$ [62].

La figure (II.3) illustre le CGR du génome mitochondrial humain permettant d'observer des motifs triangulaires similaires qui se répandent dans le plan [62].

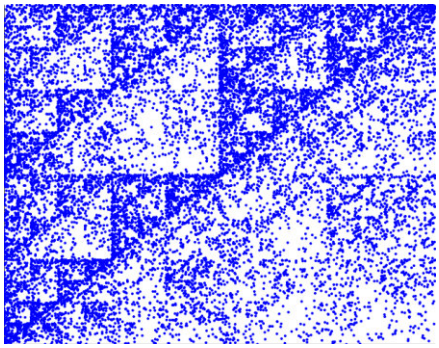


Figure II. 3 : CGR de la séquence complète du génome mitochondrial humain (GenBank - Numéro d'accès : D38116) [62].

II.3. Organigramme standard de GSP pour l'identification des exons

Les méthodes qui existent actuellement dans le GSP suivent les mêmes étapes principales pour l'identification des exons chez les eucaryotes. Après l'acquisition de la séquence d'ADN (via les bases de données), viennent 3 étapes majeures : la numérisation de la séquence, le traitement du signal par les TNS et enfin l'évaluation statistique par des critères (paramètres) [6].

Ces étapes sont illustrées dans la figure (II.4). L'étape concernant l'utilisation des TNS sera détaillée dans le 3ème chapitre de ce manuscrit, tandis que les parties sur la numérisation de la séquence et les paramètres d'évaluation seront présentées dans les points qui suivent.

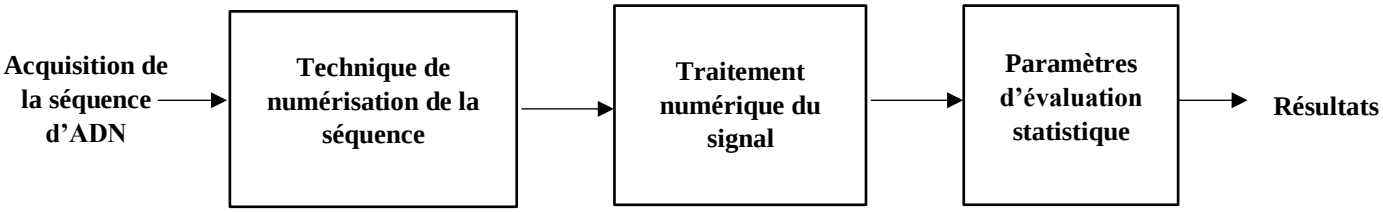


Figure II. 4 : Organigramme standard des méthodes GSP.

II.4. Les méthodes de numérisation de la séquence d'ADN en GSP

La séquence d'ADN doit être numérisée pour être mieux exploitable avec les techniques de traitement numérique de signal (TNS). Cette étape est importante et le choix de la méthode de numérisation peut influencer les résultats à la fin de l'algorithme du GSP [5], [64].

Il existe une vingtaine de méthodes de numérisation de la séquence d'ADN qui peuvent être groupées en deux familles :

- **Numérisation qui n'est pas basée sur les propriétés biologiques (physico-chimiques) des ADN.** Elle consiste en la permutation des caractères A, G, T et C par des valeurs numériques plus ou moins arbitraires. Ces valeurs n'ont pas de significations biologiques mais contribuent à rendre la numérisation optimale pour l'outil de TNS qui va être utilisé après. Ces méthodes peuvent être basées sur des codages binaires, des coordonnées spatiales (cartésiennes, polaires) pour les représentations graphiques en 1D, 2D, 3D et même 4D [64].
- **Numérisation basée sur les propriétés physico-chimiques des ADN.** Elle consiste en la permutation des caractères A, G, T et C par des valeurs numériques qui peuvent avoir des significations biologiques liées à l'ADN. Ces méthodes exploitent des propriétés structurelles des ADN (nucléotides, purine, pyrimidine, groupe amine, groupe cétone, les liaisons covalentes), propriétés moléculaires (masse moléculaire, numéro atomique), propriétés thermodynamiques (énergie, entropie, enthalpie, interaction électron-ion) ... [5].

Le tableau (II.1) suivant regroupe quelques techniques de numérisation de la séquence d'ADN selon leurs principes de base, leurs noms (nous avons gardé leurs appellations originales), la conversion numérique utilisée et leurs propriétés.

Méthodes qui ne sont pas basées sur les propriétés physico-chimiques de l'ADN			
Principe de base	**Nom de la méthode**	**Conversion numérique**	**Propriété de la méthode**
Codage binaire	Voss [51]	$s = [A, G, T, C]\ devient$ $A_n = [1,0,0,0],\ G_n = [0,1,0,0]\ , T_n = [0,0,1,0], C_n = [0,0,0,1]$	Une séquence d'ADN est numérisée en 4 séquences formées de 0 et 1. n étant la position de la base dans la séquence. Cette méthode de numérisation est parmi la plus utilisée dans le GSP.
	Galois field[5]	$\alpha^0 = 1 \leftrightarrow 1 \leftrightarrow C, \alpha^1 = \alpha \leftrightarrow 2 \leftrightarrow T, \alpha^2 = \alpha + 1 \leftrightarrow 3 \leftrightarrow G,$ $0 = 0 \leftrightarrow 0 \leftrightarrow A$	La séquence est transformée dans un codage orthogonal (n, k) selon le polynôme Galois à 4 équations.
Représentation graphique	CGR (Chaos Game Representation) [62], [65]	$A(0,0), G(1,1), C(0,1), T\ (1,0)$	Représentation graphique dans un cercle unité basé sur la propriété fractale de la séquence d'ADN.
	Tetrahedron [5], [66]	$A = k, G = -\frac{2\sqrt{2}}{3}i - \frac{\sqrt{6}}{3}j - \frac{1}{3}k, T = \frac{2\sqrt{2}}{3}i - \frac{1}{3}k,$ $C = -\frac{2\sqrt{2}}{3}i + \frac{\sqrt{6}}{3}j - \frac{1}{3}k$	Une représentation en 3D permettant de visualiser les distances entre les nucléotides et les codons
	Self Organizing Maps (SOM) [67]	$A(0,0,0), G(0,1,0), T(0.289,0.5,0.816), C(0.866,0.5,0)$	La carte auto-adaptative est une représentation avec des coordonnées en 3D. Elle est très utilisée pour les études sur l'ADN avec les réseaux neurones.
	Quaternion [68]	$A = i + j + k, G = -i - j + k, T = i - j - k, C = -i + j - k$	Une représentation sur une échelle en 3D (si j=2) et 4D (si j=3) qui ne fonctionne qu'avec la transformation discrète de Fourier quaternionique.
	Z-curve [69]	$n = A_n + G_n + T_n + C_n$ $\begin{vmatrix} A_n \\ G_n \\ T_n \\ C_n \end{vmatrix} = \frac{n}{4}\begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \end{vmatrix} + \frac{1}{4}\begin{vmatrix} +1+1+1 \\ -1+1-1 \\ +1-1-1 \\ -1-1+1 \end{vmatrix}\begin{vmatrix} x_n \\ y_n \\ z_n \end{vmatrix}$	Conversions des 4 faces du tetrahedron $(A_n, G_n, T_n\ et\ C_n)$ en 3 coordonées indépendantes $\{x_n, y_n, z_n\}$, n représente le nombre des bases. C'est une amélioration du H-curve [70].
Coordonnées cartésiennes	Integer (nombre entier) [71] [5]	$x \leftrightarrow n$ $x \in \{A, G, T, C\}, n \in \{2,3,1,0\}\ ou\ n \in \{1,2,-2,-1\}$	La conversion des A, G, T, C en nombre entier permet d'avoir une seule séquence numérique de l'ADN.

	Real number (nombre réel) [71] [5]	$x \leftrightarrow n$ $x \in \{A, G, T, C\}, n \in \mathbb{R}$	La conversion des A, G, T, C en nombre réel permet d'avoir une seule séquence numérique de l'ADN.
	Nombre complexe [5], [64]	$A = 1 + j, G = -1 - j, T = 1 - j, C = -1 + j$	Il existe d'autres possibilités d'affecter d'autres nombres complexes aux A,G, T, C.
	QPSK (Quadrature Phase Shift Keying) [72]–[74]	$A = 1 + j, G = -1 + j, T = 1 - j, C = -1 - j$	Modulation de phase à quatre états qui permet de représenter les données sur un plans à 2D.
	Pulse Amplitude Modulation (PAM) [72]–[74]	$A = -1.5, G = -0.5, T = 1.5, C = 0.5$	Représentation unidimensionnelle, un peu similaire à la conversion en nombre réel.
	Représentation trigonométrique [75]	$(\sin\theta, -cos\theta) \rightarrow A$; $(cos\theta, -sin\theta) \rightarrow G$; $(sin\theta, cos\theta) \rightarrow T$; $(cos\theta, sin\theta) \rightarrow C$ où $0 < \theta < \frac{\pi}{2}$ et $\theta \neq \frac{\pi}{4}$	Représentation en 2D qui ne cause pas des pertes d'informations de la séquence d'ADN.
Méthodes basées sur des propriétés physico-chimiques de l'ADN			
Propriétés des structures primaires du génome	Représentation dinucléotide [76], [77]	Représentation de 16 dinucléotides (AC, AG, AA, AT, GA, GG, GT, GC, TA, TG, TT, TC, CA, CG, CT, CC) sur la circonférence d'un cercle sur un cercle unité.	Chacun des 16 points de la circonférence se situe à un angle précis permettant de distribuer les dinucléotides de la séquence à numériser.
	DNA Walk [56], [78]	$C\ ou\ T = 1$, $A\ ou\ G = -1$ et le reste = 0	Numérisation basée sur les purines (A et G) et pyrimidines (T et C) et leurs changements cumulés. On obtient une seule séquence numérique.
	Paired numerical [78]	$A\ ou\ T = 1$, $C\ ou\ G = -1$ et le reste = 0	Numérisation basée sur la complémentarité des bases sur les doubles hélices des ADN (A-T et G-C). On peut obtenir 1 ou 2 séquences numériques
	Ring structure (structure en anneaux) [77]	*AG : (0, 1.5), CT : (0, -1.5), CA : (1, 1), TG:(-1, -1), CG : (1, -1), TA : (-1, 1), GA : (1, 0), GT (0.5, -1.25), GC : (-0.5, -1.25), TC : (-1, 0), AC : (-0.5, 1.25), AT : (0.5, 1.25), AA : (0, 1), TT : (0.5, 0), GG: (0, -1), CC : (-0.5, 0)*	Une extension de la représentation dinucléotide en utilisant les propriétés thermodynamiques des ADN : purine, pyrimidine, groupe amine et cétone, liaison covalente, masse moléculaire

	Inter-nucleotide distance encoding (distance inter-nucleotide) [79]	Par exemple : pour une séquence d'ADN « AGTTCTACCAGC », l'encodage pour le 1er A est 6 car il est séparé de 6 nucléotides par rapport au 2ème A et ainsi de suite pour les autres nucléotides. Afin d'avoir une séquence numérique de [9 ; 1 ; 2 ; 3 ; 6 ; 3 ; 1 ; 3 ; 2 ; 1 ; 0].	Utilisation des structures primaires des ADN, de la distances inter-nucléotides du même type. Surtout dans les études de localisation des complexes G-C [80].
	Frequency of occurrence (Fréquence d'apparition) [81]	Pour le génome humain, l'encodage utilisé est le suivant : CG : 0.01, GC : 0.043, CC : 0.047, GT : 0.049, GG : 0.050, AC : 0.054, TC : 0.057, GA : 0.061, TA : 0.067, AG : 0.070, CT : 0.071, TG : 0.074, CA : 0.074, AT : 0.081, AA : 0.097, TT : 0.097.	Considération des fréquences de nucléotides, dinucléotides et des triplets. Ce code est différent selon les espèces à étudier.
	Triplet encoding (encodage des triplet) [82]	$(T_1, \sigma_1), (T_2, \sigma_2), (T_3, \sigma_3), \dots, (T_{64}, \sigma_{64})$, avec T représentant un triplet parmis les 64 codons, et σ est une fonction de répétition des triplets dans la séquence.	Ce codage permet de faire la comparaison entre deux séquences d'ADN en termes de triplet.
	Minimum Entropy Mapping (MEM) : représentation d'entropie minimale [72]	C'est un processus itératif où chaque itération utilise une incrémentation de valeur fixe afin de créer les vecteurs pour A, G, T et C. Puis, l'étape de calcul des valeurs et la numérisation de la séquence suivie d'une analyse spectrale.	Ce codage permet de réduire le bruit et d'extraire les informations utiles en réduisant l'entropie spectrale de la séquence d'ADN.
	Code Walsh d'ordre 4 [83]	$W_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} A \\ G \\ T \\ C \end{pmatrix}$	Une représentation basée sur la structure en double hélices des ADN en considérant A et T complémentaires et G et C aussi.
	Contenu du code génétique (GCC, genetic code content) [84]	Chaque triplet est remplacé par un nombre complexe unique correspondant à l'acide aminé selon le code génétique (tableau I.1). Exemple : Phe (F) = 2.02+189i, Gly (G) = 0.07+60i, His (H) =0.61+152.6i	La partie réelle représente les propriétés hydrophobiques de l'acide aminé tandis que la partie imaginaire est son volume résiduel.

Propriétés biochimiques	Numéro atomique [85]	A = 70, G = 78, T = 66, C = 58	Numérisation en utilisant les numéros atomiques de chaque nucléotide. Une méthode qui était utilisée pour mesurer la dimension fractale entre des espèces (exemple : Humain et Chimpanzé). [86]
	Interaction pseudo potentielle des électrons-ions (EIIP) [87]	A = 0.1260, G = 0.0806, T = 0.1335, C = 0.1340	Utilisation des énergies des électrons libres dans les acides aminés et nucléotides. Une méthode très utilisée dans le GSP comme la représentation de Voss pour l'identification des exons.
	Représentation des masses moléculaires [57], [77], [88]	A = 135.13, G = 151.13, T = 126.1, C = 111.1 ou A = 134, G = 150, T = 125, C = 110	Utilisation des masses moléculaires de chaque nucléotide pour numériser la séquence. La masse moléculaire est en g/mol.
	Propriétés thermodynamiques [89]	TC = 5.6, GA = 5.6, CA = 5.8, TG = 5.8, TA = 6.0, AC = 6.5, GT = 6.5, CT = 7.8, AG = 7.8, AT = 8.6, TT = 9.1, AA = 9.1, CC = 11.0, GG = 11.0, GC = 11.1, CG = 11.9	Utilisation des valeurs d'enthalpie selon les propriétés thermodynamique des interactions entre nucléotides. L'enthalpie étant en kcal/mol.

Tableau II. 1 : Quelques méthodes pour la conversion numérique de la séquence d'ADN.

Les études ont montré que la performance d'une technique GSP est étroitement liée au choix de la méthode de numérisation et aussi de la technique de TNS utilisée [5], [62].

Selon la littérature, la représentation de Voss et l'EIIP sont les méthodes les plus utilisées comme numérisation de la séquence d'ADN dans le but d'identifier les exons [5], [51], [87], [90] - [94].

Le nombre de séquences numériques obtenues par chaque méthode de numérisation n'est pas souvent le même car certaines méthodes donnent une seule séquence numérique à partir de la séquence d'ADN originale comme EIIP, integer (nombre entier), real number (nombre réel), quaternion, QPSK, tandis que d'autres peuvent donner plusieurs séquences numériques à partir de la séquence d'ADN à numériser. C'est le cas de la représentation de Voss qui donne 4 séquences numériques ou le Z-curve qui donne 3 séquences numériques [64].

II.5. Les paramètres d'évaluation des méthodes GSP

Dans plusieurs domaines tels que la biologie, la médecine, … les tests et les modèles sont évalués en utilisant des critères (paramètres) statistiques. De même, les algorithmes GSP utilisés pour identifier les exons font également appel à ces mêmes paramètres statistiques pour distinguer les exons des régions non-codantes. L'évaluation des performances d'une méthode GSP se fait généralement au niveau nucléotidique, en utilisant les valeurs de base tels que : le taux de vrai positif (VP), vrai négatif (VN), faux positif (FP) et faux négatif (FN) [38].

Vrai positif (VP) : désigne les nucléotides correctement identifiés comme codants.

Vrai négatif (VN) : désigne les nucléotides correctement identifiés comme non-codants.

Faux positif (FP) : désigne les nucléotides identifiés comme codants mais qui ne le sont pas.

Faux négatif (FN) : désigne les nucléotides identifiés comme non-codants mais qui ne le sont pas.

Ces 4 valeurs permettront par la suite de calculer les paramètres d'évaluation tels que : la sensibilité (Sn), la spécificité (Sp), l'exactitude (E), la corrélation approximative (CA), la courbe de ROC (receiver operating characteristic) et sa surface AUC (area under the curve) et le F1-mesure. Ces paramètres sont évalués à l'aide d'un seuillage (S). La durée d'exécution du programme peut aussi être pris en compte. Ces paramètres seront définis dans les points suivants, dans le cadre de l'identification des exons [38], [44].

II.5.1. Sensibilité (Sn)

La sensibilité (Sn) mesure la proportion de vrais positifs (VP) parmi tous les exemples positifs (exons présents dans la séquence d'ADN). Elle indique la capacité d'un modèle ou méthode à identifier tous les exons positifs. Elle se calcule par la relation (II.7) suivante.

$$Sn = VP / (VP + FN) \quad \text{(II.7)}$$

La Sn varie de 0 à 1 avec 1 comme meilleure valeur.

II.5.2. Spécificité (Sp)

La spécificité (Sp) mesure la proportion de vrais négatifs (VN) parmi tous les exemples négatifs (régions non-codantes présentes dans la séquence d'ADN). Cela indique la capacité d'un modèle ou méthode à identifier tous les non-codants (ex : introns). Elle se calcule par la relation (II.8) suivante.

$$Sp = VN / (VN + FP) \quad \text{(II.8)}$$

La Sp varie de 0 à 1 avec 1 comme meilleure valeur.

La précision mesure la proportion de vrais positifs (VP) parmi tous les exemples identifiés comme positifs. Elle mesure la capacité du modèle à identifier correctement les exons. Elle se calcule par la relation (II.9) suivante.

$$Précision = VP / (VP + FP) \quad \text{(II.9)}$$

Il convient de noter que, dans la littérature sur l'identification des exons, une même formule est utilisée pour calculer à la fois la précision et la spécificité. Cela est dû au fait que la formule (II.9) ne prend en compte que les vrais positifs et les faux positifs. Les vrais négatifs (régions non-codantes identifiées comme non-codantes) ne sont pas pris en compte car leur nombre est généralement beaucoup plus élevé que celui des faux positifs [38], [44], [50], [93], [95], [96].

La formule (II.9) est alors utilisée pour remplacer la formule (II.8) afin de mesurer la spécificité (Sp).

II.5.3. Exactitude (E)

L'exactitude (E) mesure la proximité des résultats obtenus par rapport à la vraie valeur, ou la valeur de référence. Elle mesure à quel point les résultats sont corrects. Elle se calcule par la relation (II.10) suivante.

$$E = VP + VN \,/\, (VN + VP + FP + FN) \quad \text{(II.10)}$$

La valeur de E varie aussi de 0 à 1. Avec 1 comme la meilleure valeur.

II.5.4. Corrélation approximative (CA)

La corrélation approximative (CA) mesure la corrélation entre les prédictions du modèle et les observations réelles. Cela prend en compte les faux positifs et les faux négatifs. Elle se calcule par la relation (II.12) suivante, à partir de la moyenne de la probabilité conditionnelle (PCM) de la relation (II.11)

$$\text{PCM} = \frac{1}{4}\left[\frac{VP}{VP+FN} + \frac{VP}{VP+FP} + \frac{VN}{VN+FP} + \frac{VN}{VN+FN}\right] \quad \text{(II.11)}$$

$$\text{CA} = (\text{PCM} - 0.5) \times 2 \quad \text{(II.12)}$$

Les valeurs de CA varie entre -1 et 1 tandis que PCM varie entre 0 et 1. 1 étant toujours la meilleure valeur.

II.5.5. F1-mesure

Le F1-mesure (F1) évalue l'efficacité des modèles de classification pour des problèmes à deux classes ou plus. Il est souvent utilisé pour des situations où les données sont déséquilibrées. Le F1-mesure combine la précision (spécificité dans le cas d'identification des exons) et le rappel (sensibilité) en une seule mesure. Elle se calcule par la relation (II.13) en comparant les vraies prédictions (VP) aux erreurs (FN+FP) faites.

$$F1\_mesure = VP/(VP + \frac{1}{2}(FP + FN)) \quad \text{(II.13)}$$

C'est un paramètre alternatif du coefficient de corrélation de Matthews (MCC) mais qui est préférable dans le cadre de l'identification des exons [38]. Le F1-mesure varie entre 0 et 1 avec 1 est le meilleur mesure décrivant une bonne prédiction du modèle.

II.5.6. ROC (Receiver Operating Characteristic) et AUC (Area under the curve)

La courbe ROC (Receiver Operating Characteristic) est un graphique 2D qui représente la performance d'un modèle de classification. Elle est obtenue en traçant le taux de sensibilité (Sn) en fonction du taux de faux positifs (1 - spécificité) à différents seuils de classification. Un modèle qui prédit parfaitement la classe positive aura une courbe ROC qui passe par le coin

supérieur gauche du graphique, tandis qu'un modèle qui prédit au hasard aura une courbe ROC qui suit ou qui est inférieure à la diagonale du graphique [97], [98].

La surface sous la courbe ROC (AUC, pour Area Under the Curve) est une mesure de la performance globale d'un modèle de classification. Elle correspond à la probabilité que le modèle classe un exemple positif choisi au hasard plus haut qu'un exemple négatif choisi au hasard. Une AUC de 1 indique une performance parfaite, tandis qu'une AUC de 0,5 correspond à une performance aléatoire et une AUC<0.5 indique une mauvaise classification [38], [50], [97], [98].

La figure (II.5) illustre les interprétations possibles faites avec un courbe ROC et son AUC.

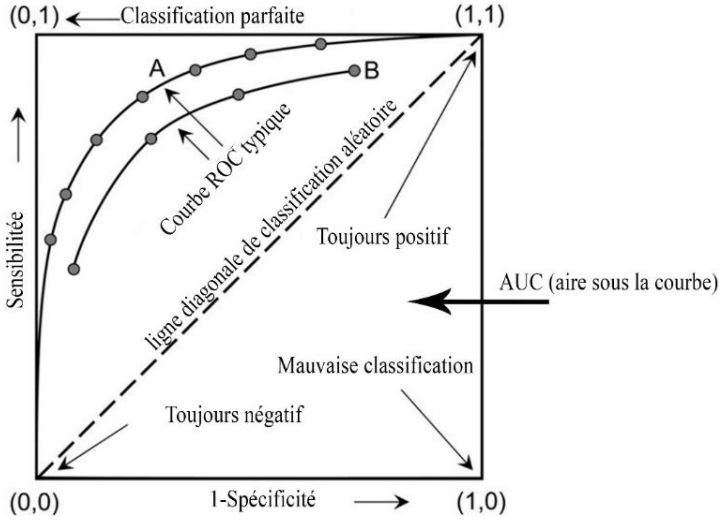


Figure II. 5 : Représentation de la courbe ROC et son AUC, avec les propriétés de classification [99].

II.5.7. Durée d'exécution de l'algorithme

La durée d'exécution de l'algorithme GSP peut être considérée comme un paramètre d'évaluation important, surtout si l'algorithme est destiné à être utilisé dans des applications en temps réel ou pour des analyses de grands volumes de données. Une durée d'exécution courte est préférable [38].

II.5.8. Remarque sur les paramètres d'évaluation

Dans le contexte de l'identification des exons, la région non-codante est généralement beaucoup plus grande que le nombre d'exons, c'est-à-dire les faux positifs (régions non-codantes identifiées à tort comme des exons), qui ont tendance à avoir un impact plus important sur les performances de l'algorithme. Par conséquent, la précision est souvent considérée

comme plus importante que la spécificité dans l'identification des exons, bien que les deux soient importantes. La spécificité est calculée en utilisant la relation (II.9) de la précision dans ce contexte [38], [50], [93], [95], [96].

Dans certains travaux GSP, on utilise aussi l'exactitude moyenne (Em) qui se calcule généralement comme la moyenne de la somme entre la sensibilité (Sn) et la spécificité (Sp) à une valeur de seuil donnée.

Le choix du seuil a un impact sur les performances de l'algorithme et sur les valeurs des autres paramètres d'évaluation. Beaucoup de méthodes GSP ont des difficultés à estimer une valeur de seuil spécifique car souvent cela peut changer en fonction de la séquence d'ADN analysée [6].

II.6. Conclusion

Ce chapitre présente les bases du « Genomic Signal Processing » (GSP) pour l'identification des exons chez les eucaryotes. La propriété spectrale dite « 3-base periodicity » est fondamentale en GSP bien qu'il existe la nature fractale de ces signaux. Nous avons vu que l'algorithme GSP nécessite plusieurs étapes. Bien que les méthodes de numérisation d'ADN soient plus ou moins bien maîtrisées, les techniques de TNS proposent diverses approches qui seront présentées dans le chapitre suivant. L'évaluation des performances des méthodes GSP sont faites généralement au niveau nucléotidique par des paramètres statistiques communément utilisés dans d'autres domaines comme les tests de diagnostics médicaux. La plus grande différence est dans l'estimation de la spécificité d'une méthode GSP qui utilise la formule de précision dans le cadre de l'identification des exons.

CHAPITRE III - TECHNIQUE DE TRAITEMENT NUMERIQUE DU SIGNAL POUR L'IDENTIFICATION DES EXONS

III.1. Introduction

Ce troisième chapitre est une suite logique du précédent. Il présente les méthodes de traitement numérique du signal (TNS) couramment utilisées pour l'identification des exons chez les eucaryotes. Ce chapitre présente alors la deuxième étape principale de l'algorithme du GSP décrite dans la figure (II.4) du chapitre précédent. Premièrement, nous présenterons plusieurs méthodes GSP basées sur les transformations du signal suivies des méthodes GSP basées sur le filtrage numérique.

III.2. Transformée de Fourier et fenêtrage pour l'identification des exons

Dans l'identification des exons des cellules eucaryotes, la transformation de Fourier et les fenêtrages sont les techniques de TNS premièrement introduites. Les points suivants rapportent les méthodes GSP utilisant ces techniques.

III.2.1. Mesure de composant spectral (spectral content measure, SCM)

Les travaux de Fickett [100] et Voss [55] sur la propriété « 3-base periodicity » des exons ont permis par la suite à Tiwari et al. [51] de proposer une méthode appelée « Spectral Content Measure (SCM) ». La SCM identifie les exons en mesurant le rapport signal-bruit (SB) des pics sur le spectre de puissance de la séquence d'ADN à la fréquence f = 1/3. Ce spectre est obtenu en faisant la somme des spectres de chaque signal numérique des bases (nucléotides) à la fréquence f = 1/3.

Les étapes de la SCM se résume comme suit :

Soit une séquence d'ADN x de longueur N qui est numérisée par la représentation de Voss en donnant 4 séquences numériques, une pour chaque base (nucléotides) : $x_A(n), x_G(n), x_T(n) et\ x_G(n)$ où n est la position de chaque base.

La somme des spectres T est :

$$T(k) = \frac{1}{N^2}(|X_A(k)|^2 + |X_G(k)|^2 + |X_T(k)|^2 + |X_C(k)|^2) \qquad \text{(III.1)}$$

avec : $k = 0,1,2,\dots,N-1$

Dans la relation (III.1), $X_A(k), X_G(k), X_T(k)\ et\ X_C(k)$ sont les transformées de Fourier de :$x_A(n), x_G(n), x_T(n) et\ x_G(n)$.

La moyenne de la somme totale des spectres étant aussi calculé par :

$$\bar{T} \equiv \frac{2}{N}\sum_{k=1}^{\frac{N}{2}} T\left(\frac{k}{N}\right) = \frac{1}{N}\left(1 + \frac{1}{N} - \sum_{\alpha} \rho_{\alpha}^{2}\right) \quad \text{(III.2)}$$

Dans la relation (III.2) , $\alpha \in \{A, G, T, C\}$ et ρ_{α} représente la fréquence d'apparition de α.

De (III.1) et (III.2) la SCM calcule le rapport signal-bruit (SB) dans la relation (III.3) :

$$SB = \frac{T\left(\frac{N}{3}\right)}{\bar{T}} \quad \text{(III.3)}$$

Par calcul cumulatif de la distribution de SB, le long de la séquence de différents organismes à analyser, le SCM proposé par Tiwari et al. [51] a fixé une valeur seuil égale à 4 afin de différencier les exons et les introns. En effet, la région de la séquence ayant un rapport signal-bruit SB supérieur ou égal au seuil est identifiée comme exon présentant la propriété « 3-base periodicity », le reste de la séquence sera considérer comme les introns et non-codant.

La SCM utilise une fenêtre de type rectangulaire avec une largeur de $l = 351$ bases (nucléotides) pour parcourir la totalité de la séquence d'ADN jusqu'au point $N - l$. La largeur de la fenêtre rectangulaire était fixée à 351 due au fait que les organismes analysés par Tiwari et al.[51] ont des exons de largeur moyenne de 300 pb. Toutefois, le SCM propose aussi une largeur de fenêtre $l \sim 150$ pour des organismes plus complexes. La performance de SCM dépend à la fois de l et de N (largeur de la fenêtre et de la séquence).

III.2.2. La mesure optimisée du composant spectral (Optimized Spectral Content Measure-OSCM)

Le OSCM est une méthode proposée par Anastassiou [101] comme alternative de la SCM [51]. En utilisant la représentation de Voss [55] pour numériser la séquence d'ADN. L'OSCM calcule le module au carré de la somme (S) à la fréquence $f = \frac{1}{3}$ des transformées de Fourier $X_A, X_G, X_T \; et \; X_C$ des sequences numériques d'ADN $x_A, x_G, x_T \; et \; x_C$ en y ajoutant des paramètres de pondération $a, g, t \; et \; c$ à chaque séquence numérique selon la relation (III.4).

$$|S|^2 = |aX_A(f) + gX_G(f) + tX_T(f) + cX_C(f)|^2 \quad \text{(III.4)}$$

Pour différencier les exons des non-codants, l'OSCM utilise la courbe de $|S|^2$ avec un système de seuillage.

La particularité de OSCM est qu'en ajustant les paramètres $a, g, t\ et\ c$, 3 des 4 séquences numériques obtenues par la représentation de Voss suffisent pour faire l'identification des exons dans la séquence : mettant $c = 0$ par exemple. L'OSCM utilise une fenêtre rectangulaire de largeur l multiple de 3 pour extraire le composant à la fréquence $f = \frac{1}{3}$. Comme la SCM, la performance de l'OSCM dépend à la fois de l et de N (largeur de la fenêtre et de la séquence). Ces deux méthodes de GSP sont toujours considérées comme des références dans l'évaluation de nouvelles approches basées sur la TNS pour l'identification des exons.

III.2.3. Mesure de la rotation spectrale (Spectral Rotation Measure, SRM) »

Cette méthode proposée par Kotlar et Lavner [6] est similaire à l'OSCM. La SRM est basée sur le calcul de la valeur des arguments des coefficients X_b, avec $b \in \{A, G, T, C\}$, à la fréquence $f = \frac{1}{3}$ de chaque séquence numérique x_b en utilisant la transformée de Fourier discrète (TFD) selon la relation (III.5). Le calcul se fait afin de tracer les graphes de distributions de ces arguments.

$$X_b\left(\frac{L}{3}\right) = \sum_{n=0}^{L-1} x_b(n) e^{-j\omega n} \qquad \text{(III.5)}$$

Dans la relation (III.5), L est là largeur de la fenêtre tandis que $\omega = 2\pi f = \frac{2\pi}{3}$.

La SRM se calcule comme suit :

$$|V|^2 = \left|\frac{e^{-j\mu_A}}{\sigma_A} X_A\left(\frac{L}{3}\right) + \frac{e^{-j\mu_G}}{\sigma_G} X_G\left(\frac{L}{3}\right) + \frac{e^{-j\mu_T}}{\sigma_T} X_T\left(\frac{L}{3}\right) + \frac{e^{-j\mu_C}}{\sigma_C} X_C\left(\frac{L}{3}\right)\right|^2 \qquad \text{(III.6)}$$

Dans la relation (III.6), $\mu_A, \mu_G, \mu_T\ et\ \mu_C$ sont respectivement les valeurs moyennes des arguments de $X_A(L/3)$, $X_G(L/3)$, $X_T(L/3)$ et $X_C(L/3)$. Tandis que σ_A, σ_G, σ_T et σ_C sont les déphasages angulaires de $X_A(L/3)$, $X_G(L/3)$, $X_T(L/3)$ et $X_C(L/3)$.

Pour identifier et différencier les exons, la SRM utilise une représentation graphique afin de comparer la longueur des vecteurs crées à partir des $\mu_A, \mu_G, \mu_T\ et\ \mu_C$. En faisant les calculs pour les bases G uniquement, la SRM arrive à mieux détecter les exons de longueurs plus courtes $\sim 120\ pb$.

Comme l'OSCM, la SRM utilise aussi des fenêtres rectangulaires et sa performance dépend surtout de N (longueur de la séquence). Les 16 chromosomes de Saccharomyces cerevisiae (levure du boulanger) ont été testés pour évaluer l'OSCM.

III.2.4. Prédictions des exons par distribution des nucléotides (Exon Prediction by Nucleotide Disctribution, EPND)

Cette méthode a été proposée par Yin et Yau [52], [102] afin d'identifier les exons en utilisant les probabilités de distributions des nucléotides dans les codons. C'est une méthode basée sur la transformée de Fourier discrète (TFD) sans utiliser de fenêtre spécifique. Elle a permis de mieux comprendre l'origine de la propriété « 3-base periodicity » des exons, (la 3ème hypothèse dans le point II.2.2.1). Cependant, l'EPND est limité quand il s'agit des séquences avec beaucoup d'exons et d'introns.

III.2.5. Prédicteur optimisé de gène (Optimized Gene Predictor, OGP)

Le OGP est une méthode introduite par Jiang et al [75]. Basée sur la représentation graphique en 2D via une numérisation des ADN en cordonnées trigonométriques (tableau II.1) afin de mieux conserver les informations de la séquence. La technique utilise une fenêtre coulissante de largeur arbitraire qui s'optimise avec les calculs afin d'avoir une fenêtre proche de la taille de l'exon à analyser. Parmi les avantages de la méthode est qu'elle peut être utilisée pour analyser des séquences inconnues, car le paramètre comme la largeur de la fenêtre l ainsi que la valeur du seuil pour identifier les exons s'ajustent selon la séquence analysée. Cependant, la méthode OGP dépend des choix des valeurs initiales de l et de l'angle θ (de la conversion numérique des exons) qui doivent aléatoirement être ajustées.

III.3. GSP à base de transformée en ondelettes

L'utilisation de la transformée en ondelette (wavelet transform) comme technique de TNS dans l'identification des exons chez les cellules eucaryotes donnent des possibilités d'amélioration des résultats attendus dans le GSP.

Mena-Chalco et al [103] furent les premiers à proposer de telle approche. La méthode s'appelle « Modified Gabor-wavelet transform » (MGWT), une combinaison entre la transformation en ondelette de Gabor (Gabor-wavelet transform : GWT) et la transformée de Fourier à court terme (TFCT) avec une fonction de fenêtrage Gaussienne, aussi appelée transformée de Gabor (TG).

La transformée de Gabor est définie par la relation (III.7) suivante.

$$\psi_{TG}(x,b,a) = e^{-\frac{(x-b)^2}{2}} e^{aj(x-b)} \qquad \text{(III.7)}$$

La transformée en ondelette de Gabor est définie par la relation (III.8).

$$\psi_{GWT}(x,b,a) = e^{-\frac{(x-b)^2}{2}} e^{j\omega_0\left(\frac{x-b}{a}\right)} \quad \text{(III.8)}$$

En modifiant la formule (III.8), Mena-Chalco et al. [103] propose le MGWT qui est définie par la relation (III.9) suivante :

$$\psi_{MGWT}(x,b,a) = e^{-\frac{(x-b)^2}{2a^2}} e^{j\omega_0(x-b)} \quad \text{(III.9)}$$

Cette méthode utilise aussi la représentation de Voss [55] pour numériser les ADN. L'algorithme général de MGWT est le suivant :

1. Numérisation de la séquence d'ADN par la technique de Voss pour obtenir les 4 séquences numériques : $s_A, s_G, s_T \; et \; s_C$.
2. Calcul des transformées MGWT avec la relation (III.10) de chaque séquence numérique à différentes échelles et à fréquence constante ω_0 correspondant à N/3 pour la propriété « 3-base periodicity », N étant la longueur de la séquence.

   $$S_\alpha(b,a) = \int s_\alpha \psi_{MGWT}(x,b,a)dx \quad \text{(III.10)}$$

   Avec
   $\alpha \in \{A,G,T,C\}$, b la position des bases dans la séquence et a est le paramètre d'échelle.

3. Calcul du spectre total T qui est la somme des spectres de chaque séquence numérique par la relation (III.11).
   $$T(b,a) = \sum_\alpha |S_\alpha(b,a)|^2 \quad \text{(III.11)}$$

4. Projection du spectre total T (b, a) sur un repère axé, suivant la relation (III.12).
   $$T_p(b) = \sum_\alpha T(b,a), \quad b = 0,1,\dots,N-1 \quad \text{(III.12)}$$
   Avec N la longueur de la séquence.
5. Opération de seuillage est appliquée à $T_p(b)$ pour identifier les exons avec une valeur de seuil S prédéterminée. Les coefficients de $T_p(b)$ inferieurs à S sont mis à zéro et identifiés comme introns alors que les autres sont des exons.

La performance de la méthode MGWT dépend surtout des fixations du paramètre d'échelle a et de la valeur du seuil S.

III.4. Remarques sur les méthodes GSP basées sur les transformées

En général, la performance des méthodes GSP utilisant les techniques de transformations dépendent de plusieurs et mêmes paramètres : largeur et le type de fenêtre coulissante, longueur de la séquence à analyser, nombre d'exons et d'introns dans la séquence et valeur du seuil pour différencier les exons et les introns. La représentation de Voss [55] est souvent la plus utilisée pour numériser la séquence d'ADN.

Afin d'améliorer les performances des méthodes GSP à base de transformations, les chercheurs ont essayé avec d'autres types de fenêtres et d'autres approches : fenêtre de Barlett [104], ajustement de largeur de fenêtre [105], fenêtre adaptative [106], le DIT-FFT proposé dans [107], des méthodes plus adaptatives comme le SAVMD [50], combinaison entre l'analyse temporelle et fréquentielle [54].

Il existe aussi d'autres approches GSP basées sur la combinaison de TFD, transformée en ondelette et l'utilisation de filtre numérique pour l'identification des exons. L'une de ces méthodes est celle proposée par Abassi et al. [108]. Cette méthode utilise aussi la conversion de Voss [55] pour numériser les ADN, un filtrage numérique à réponse impulsionnelle finie (RIF), un algorithme de corrélation croisée entre la séquence d'ADN numérisé, un train d'impulsion périodique de largeur spécifique (l=270) et l'application de la transformée en ondelette. Cette dernière permettra de réduire le bruit du signal selon les cas. La fixation de seuil optimal utilise la méthode proposée par Kwan et al. [109] qui se base sur les caractéristiques statistiques des valeurs présentes dans le signal de sortie pour un ensemble de séquences d'exons et d'introns utilisés comme données d'entraînement. La moyenne et l'écart-type de ces valeurs sont utilisés pour calculer le seuil S, qui sera ensuite utilisé pour prendre des décisions sur la classification des régions d'exons et d'introns. Dans [110], la transformée de Stockwell (transformée en S) est combinée avec un filtrage numérique afin d'identifier les exons. C'est une approche GSP qui ne dépend pas de l'utilisation de fenêtre spécifique en utilisant la technique de numérisation EIIP.

III.5. Les méthodes GSP basées sur les filtres numériques

Dans cette partie, nous allons présenter plusieurs méthodes GSP basées sur les filtres numériques pour l'identification des exons dans les cellules eucaryotes. Les principaux objectifs de l'utilisation du filtre numérique dans ce domaine sont : une meilleure extraction de la composante « 3-base periodicity » due à la sélectivité du filtre numérique utilisé, non-

dépendance au fenêtrage, une meilleure réduction des bruits du signal et optimisation des temps de calcul, selon les types de filtres utilisés.

III.5.1. Filtre anti-Notch (simple, lattice et en cascade)

Vaidyanathan et al. [111] sont parmi les premiers à proposer une méthode basée sur le filtre de type anti-Notch à réponse impulsionnelle infinie (RII) avec une fréquence centrale centrée sur $\omega_0 = \frac{2\pi}{3}$ et une atténuation de 13 dB. Cette méthode utilise la représentation de Voss en ne gardant que la séquence numérique contenant la base G (Guanine) [112]. La largeur de la bande passante du filtre dépend d'un paramètre R (multiplicateur) comme dans la figure (III.1). La fonction de transfert $H(z)$ est donnée par la relation (III.13) suivante :

$$H(z) = \frac{(1-R^2) - 2Rcos\omega_0 z^{-1}}{1 - 2Rcos\omega_0 z^{-1} + R^2 z^{-2}} \qquad \text{(III.13)}$$

Dans la formule III.13, $cos\omega_0 = \frac{2Rcos\theta}{1+R^2}$

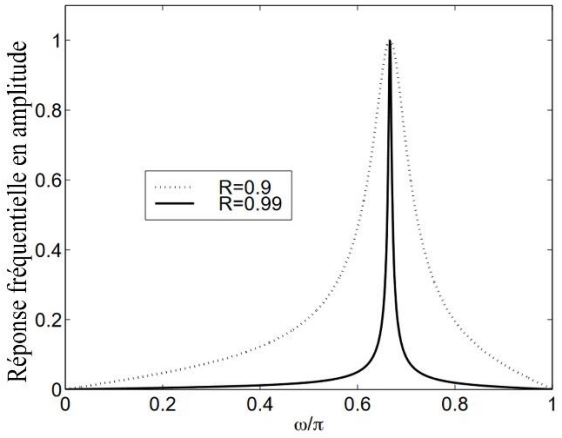


Figure III. 1 : Réponse fréquentielle en amplitude du filtre anti-Notch selon les valeurs du multiplicateur R [111].

L'organigramme de l'algorithme proposé par Vaidyanathan et al. [111] est montré dans la figure (III.2), où $x_G(n)$ est la séquence d'ADN numérisée et $y_G(n)$ est la séquence filtrée, avec G étant la base Guanine et n est la position du nucléotide.

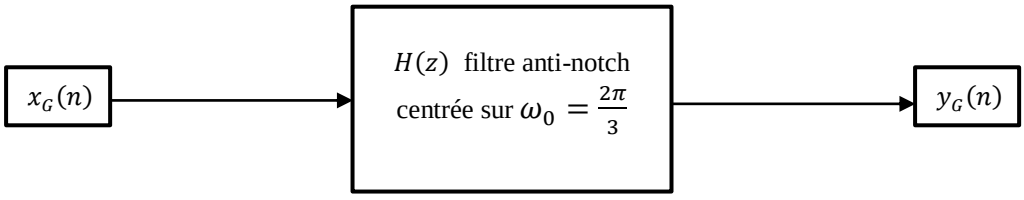


Figure III. 2 : Les étapes de la méthode GSP basée sur le filtre anti-Notch proposée dans [111].

A la sortie du filtre, une somme des modules au carré des signaux est calculée selon la relation (III.14).

$$Y[n] = \sum |y_G(n)|^2 \qquad \text{(III.14)}$$

Puis le tracé de la courbe de Y[n] permet de visualiser les exons et introns suivant un seuillage préétabli.

Vaidyanathan et al. [111] proposent alors deux autres structures à base du filtre anti-Notch, une structure lattice et une structure en cascade.

- **La structure lattice** propose un filtre anti-Notch de fonction de transfert $H(z)$ qui utilise deux coefficients $k_1 = R^2$ et $k_2 = -cos\omega_0$. Avec $\omega_0 = \frac{2\pi}{3}$, cette structure ne dépend que de la valeur de R^2. La figure (III.3) montre les opérations dans cette structure pour déterminer le filtre $H(z) = \frac{Y(z)}{X(z)}$. X(z) et Y(z) sont respectivement les transformées en Z des signaux d'entrée x(n) et de sortie y(n) du filtre H(z).

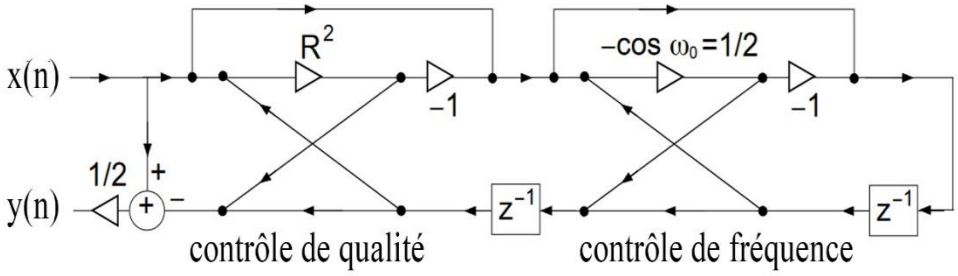


Figure III. 3 : Les étapes du filtrage anti-Notch en structure lattice [111].

- **La structure en cascade**, quant à elle produit un filtre anti-Notch $H(z)$ à partir d'un filtre passe-bas $H_1(z)$ auquel les éléments de retards z^{-1} sont remplacés par z^{-3} et d'un filtre $H_2(z)$ qui à deux zéros à la fréquence $\omega = 0$. Les relations (III.15) et (III.16) montrent la fonction du transfert de ces filtres :

$$H_1(z) = \frac{A_0(z)+A_1(z)}{2} \qquad \text{(III.15)}$$

$$H_2(z) = (1 - z^{-1})^2 \qquad \text{(III.16)}$$

$A_0(z), A_1(z)$ sont des filtres passe-tout de premier et second ordre respectivement.

De (III.15) et (III.16), on obtient le filtre anti-Notch $H(z) = H_1(z^3)\, H_2(z)$.

Vaidyanathan et al. [111] ont testé leur méthode sur le gène F56F11.4 de l'organisme C-elegans (chromosome III). Ils ont mieux détecté les exons de ce gène tout en réduisant les bruits du signal en utilisant la structure en cascade et en prenant la valeur de R = 0.992.

III.5.2. Filtre à taux multiples

Guan et al. [113] proposent l'utilisation d'un filtre RII passe-bande de Tchebychev de type I, de 4ème ordre, stable, de fonction de transfert $H(z)$ selon la relation (III.17). La méthode utilise aussi la numérisation de Voss [55].

$$H(z) = \frac{\sum_{n=0}^{4} b_n z^{-n}}{\sum_{n=0}^{4} a_n z^{-n}} \qquad \text{(III.17)}$$

Les coefficients du filtre sont rapportés dans le tableau (III.1).

n	**a_n**	**b_n**
0	1	1
1	1.98950751345255	0
2	2.98445914522766	-2
3	1.98447099131237	0
4	0.99494334594164	1

Tableau III. 1 : Les valeurs des coefficients du filtre RII de Tchebychev I à taux multiples montrant sa stabilité [113]

Guan et al. [113] testaient leur méthode sur le gène F56F11.4 de l'organisme C-elegans (chromosome III) et ont conclu que leur approche est meilleure que les approches GSP à base de TFD, notamment celle de Tiwari et al. [51], en réduisant les bruit du signal et en montrant les pics correspondants aux exons de ce gène, comme l'illustre la figure (III.4).

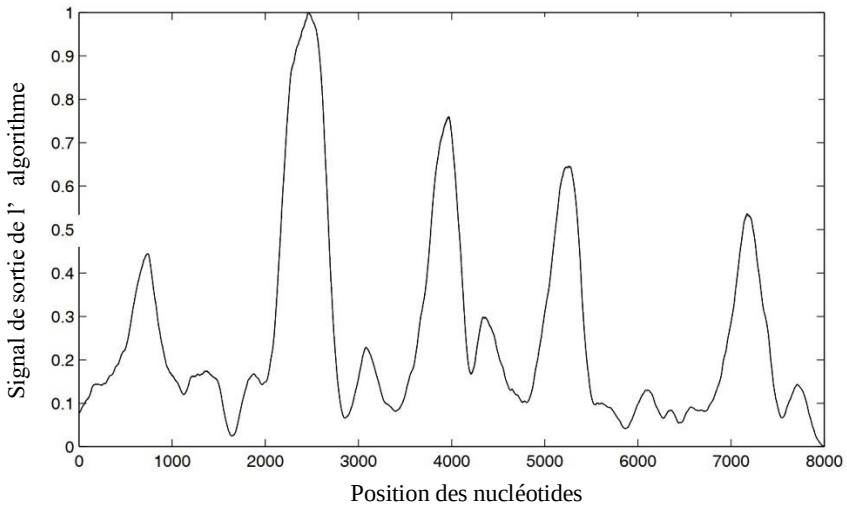


Figure III. 4 : Signal de sortie de la méthode GSP basée sur le filtrage à taux multiple sur le gène F56F11.4 du C.elegans [113].

III.5.3. Filtre optimisé à déphasage nul (Statistically Optimal Null Filter, SONF)

Kakumani et al. [114] proposent une meilleure détection des exons courts et aussi des exons séparés par des introns courts avec une méthode basée sur un filtrage optimisé à déphasage nul. La méthode commence par maximiser le rapport signal sur bruit (S/B) de la sortie à chaque instant en utilisant un filtre adaptatif. Ensuite, la méthode améliore davantage le signal estimé en utilisant un critère d'optimisation des moindres carrés, qui permet de minimiser l'erreur entre la sortie de SONF et l'entrée. Grâce à sa capacité à suivre rapidement les signaux en évolution, SONF permet un traitement plus pratique des signaux de courte durée.

Les étapes de la méthode se résument comme suit :

- Application d'une fenêtre rectangulaire de largeur l=351 en utilisant la numérisation de Voss (4 séquences numériques), x[n] avec n est la position de nucléotide dans la séquence d'ADN.
- Application du filtre SONF à chacune des séquences numériques en utilisant la fonction $\emptyset[n]$ puis utilisation du filtre adaptatif instantané. La sortie de ce filtre adaptatif instantané est noté $v[n]$. Tandis que la sortie du SONF après l'utilisation de la fonction d'échelle $\gamma[n]$ est $y[n]$.
- Evaluation récursive de la sortie du filtre SONF ($y[n]$) par rapport à l'entrée $x[n]$ afin d'obtenir $z[m]$.
- Calcul du module au carré du rapport signal-bruit de chaque séquence filtrée $z[m]$ suivi du calcul de la somme $S(n)$ des 4 valeurs.

- Coulissage de la fenêtre rectangulaire par incrémentation de 1 de sa largeur.
- Traçage de la courbe des $G(n)$ en fonction de n où les pics indiquent la localisation des exons.

Ces étapes sont illustrées dans la figure (III.5) suivante.

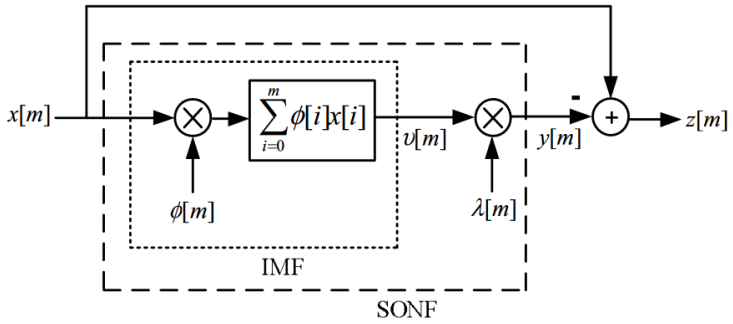


Figure III. 5 : Filtre SONF avec son étape de IMF imbriqué (filtrage adaptatif instantané) [114].

Le SONF a été testé sur différents gènes du Chromosome III de l'organisme C-elegans. Parmi ces gènes sont le **T12B5.13** (numéro d'accès AF100307 avec 1245 bases et 6 exons) et le **C30C11.1** (numéro d'accès L09634 avec 754 bases et 3 exons). Les résultats sont comparés avec ceux des méthodes basées sur la TFD [112] et sont rapportés sur la figure (III.5) pour le cas du gène T12B5.13. De cette figure, on constate que le signal de sortie de la méthode GSP basée sur le filtre SONF présente beaucoup plus de pics au niveau de la région des exons, par rapport à la méthode basée sur la TFD [114].

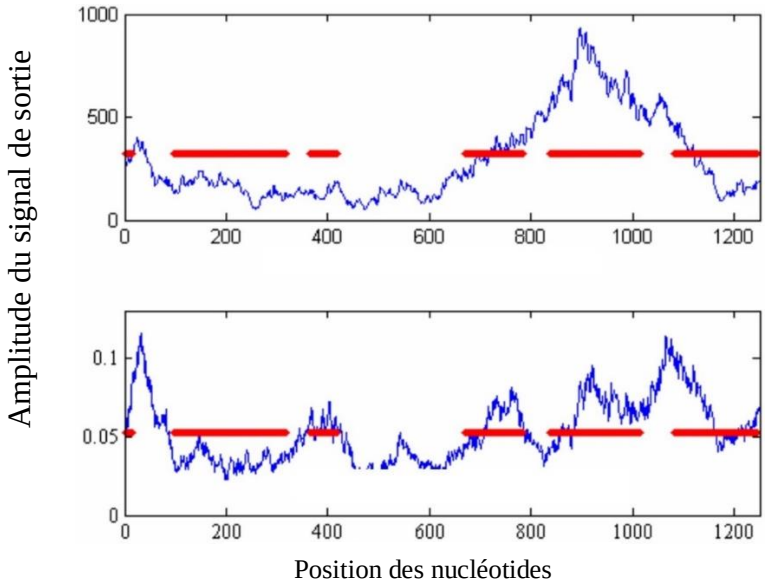


Figure III. 6 : Comparaison de performance entre la méthode basée sur SONF et celle basée sur la TFD sur le gène T12B5.13. En haut, le signal de sortie de la méthode TFD et en bas celui de SONF. Les lignes rouges horizontales représentent les positions réelles des exons selon la base de données NCBI [31].

III.5.4. Filtre à suppression harmonique et filtre à variance minimale adaptatif

Dans leurs travaux, Tomar et al. [115] proposent une méthode basée sur un filtre à suppression des harmoniques (SH) combiné à un autre filtre pour l'estimation de spectre de variance minimum (VM) afin de mieux identifier les exons. Leur méthode utilise une technique de numérisation des ADN dites « simplex » qui donne deux vecteurs à partir de la séquence d'ADN. Une fenêtre de lissage de largeur 351 est aussi utilisée.

Le filtre à suppression des harmoniques (SH) permet ainsi de réduire les harmoniques des fréquences multiples de 2π/3 et de ne laisser passer que les composants de fréquence 2π/3. Puis le signal obtenu est filtré par un filtre passe-bande (de fréquence centrale

$\omega_0 = \frac{2\pi}{3}$) utilisant une technique d'estimation de spectre de variance minimum (VM).

La relation (III.18) donne la fonction de transfert $H(z)$ du filtre SH :

$$H(z) = \frac{1-2R_2 cos\theta_2 z^{-1}+R_2^2 z^{-2}}{1-2R_1 cos\theta_2 z^{-1}+R_1^2 z^{-2}} \prod_{i=1}^{3} \frac{1-2R_1 cos\theta_i z^{-1}+R_1^2 z^{-2}}{1-2R_2 cos\theta_i z^{-1}+R_2^2 z^{-2}} \quad \text{(III.18)}$$

Avec $\theta_i = (0{,}2\pi/6, \pi)$ pour $i = 1{,}2{,}3$, $R_1 = 0.998\ et\ R_2 = 0.898$.

Tomar et al. [115] observent que le filtre SH ne réussit pas à éliminer les composants de fréquence conjuguée : -2π/3 ou 4π/3. C'est pour cela qu'ils proposent la technique d'estimation de spectre de variance minimum.

La relation (III.19) donne la réponse impulsionnelle h du filtre passe-bande (estimation de spectre de variance minimum).

$$h = \frac{R_x^{-1} e}{e^H R_x^{-1} e} \quad \text{(III.19)}$$

Avec e^H : matrice Hermitienne (conjuguée complexe) du vecteur exponentiel e.

R_x : matrice de Toeplitz d'autocorrélation p × p des échantillons dans la fenêtre actuelle.

Le filtre VM est implémenté en structure lattice et permet d'éliminer efficacement les fréquences conjuguées 4π/3.

Tomar et al. [115] testent leur technique sur le gène F56F11.4 du chromosome III du C-elegans en traçant le module au carré du signal de sortie des filtres et en comparant avec la méthode basée sur TFD de Tiwari et al. [51]. Une valeur de seuil $s = 4$ est utilisée pour classer les exons. Ils concluent que leur méthode détecte mieux les exons courts en réduisant efficacement les fréquences harmoniques.

Barman et al. [116] ont aussi proposé une méthode similaire avec un filtre RII anti-Notch en structure lattice, lattice cascadé et suppresseur d'harmonique mais en utilisant la numérisation en nombre complexe de la séquence d'ADN, le « QPSK mapping » (tableau II.1) afin de minimiser la complexité du calcul. Ils concluent que leur méthode est meilleure par rapport à un filtrage avec un seul filtre anti-Nocth en le testant sur le gène F56F11.4a de l'organisme C-elegans. On peut dire que cette méthode est une combinaison de la méthode proposée par Vaidyanathan et al. [111] et Tomar et al. [115].

III.5.5. Méthode basée sur le filtre Tchebychev de type II

Ramachandran et al. [92], proposent quant à eux un algorithme à base de deux filtres RII à déphasage nul : un filtre passe-bande suivi d'un filtre passe-bas. Ils proposent aussi le EIIP (tableau II.2) comme numérisation de la séquence d'ADN. Le filtre de Tchebychev inverse de type II est choisi comme filtre passe-bande avec les spécificités suivantes :

- Atténuation maximum de la bande-passante : 1dB
- Atténuation minimum de la bande coupée : 30 dB
- Bord inférieur et supérieur de la bande passante : 0.664 et 0.670
- Bord inférieur et supérieur de la bande d'arrêt : 0.659 et 0.675
- Ordre du filtre : 6

Pour le filtre passe-bas, c'est encore le filtre de Tchebychev inverse de type II qui est utilisé avec les spécificités suivantes :

- Atténuation maximum de la bande-passante : 1dB
- Atténuation minimum de la bande coupée : 80 dB
- Bord inférieur et supérieur de la bande passante : 0.4 et 0.5
- Ordre du filtre : 14

Les valeurs de fréquences citées ci-dessus sont normalisées à la fréquence de Nyquist .

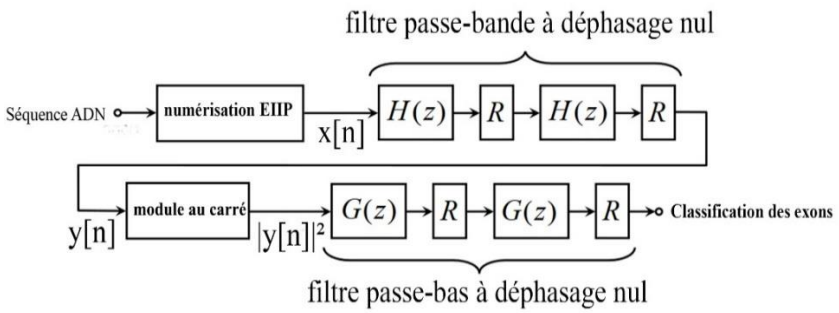


Figure III. 7 : Les étapes de la méthode GSP basée sur le filtre Tchebychev inverse de type II [92].

La figure (III.7) montre l'organigramme de la méthode proposée par Ramachandran et al. [92] pour l'identification des exons. Le filtre passe-bande centré à la fréquence $\omega_0 = \frac{2\pi}{3}$ qui permet d'isoler les composants fréquentiels à f = 1/3 pour la propriété « 3-base periodicity ». Le module au carré permet de mieux mettre en évidence les pics dans le signal et le filtrage passe-bas joue le rôle de démodulateur en amplitude pour réduire d'avantage le bruit. La méthode a été testée sur 5 gènes de la base de données HMR195 [40] qui sont AB009589, AF039307, AF042784, AF009614, et AB003306 (leurs numéros d'accès via Genbank) [117]. La méthode est ensuite comparée avec celle qui est basée sur la TFD [112] et les auteurs concluent que la leur est plus précise et moins complexe à implémenter. Dans [118], Ramachandran et al. ajoutent que le type de filtre peut être autre que Tchebychev à condition qu'il soit très sélectif pour l'opération de filtrage à la fréquence f = 1/3.

Heba et al. [119] reprennent un algorithme similaire à la figure (III.7) mais en choisissant le « 2-bit binary » (tableau II.1) comme technique de numérisation de l'ADN. Cette technique est semblable à la représentation en nombre entier pour la conversion numérique des ADN. Ils évaluent leur méthode avec des gènes de la base de données HMR195 [40] et comparent les résultats avec d'autres méthodes de numérisations, notamment le EIIP et GCC (tableau II.1) et concluent que le choix de la technique de numérisation de l'ADN influence considérablement le reste de la méthode GSP ainsi que le résultat final.

III.5.6. Filtre Savitzky-Golay (SG)

Dans leurs recherches pour l'identification des exons chez les eucaryotes, Singh et al. [93] se concentrent dans la réduction des bruits et suppression des harmoniques indésirables du signal filtré par différents types de filtre anti-Notch. Ils proposent alors l'utilisation d'un filtre de lissage qui est le filtre Savitzky-Golay (SG). Dans cette approche, le filtre anti-Notch dépend du paramètre de contrôle R qui est fixé à 0.992 et de la fréquence centrale $\omega_0 = \frac{2\pi}{3}$ tandis que le filtre SG est contrôlé par la largeur de la fenêtre de coulissage (fixée à l = 351) et l'ordre du filtre est 4. La figure (III.8) montre l'organigramme de la méthode proposée en utilisant la numérisation de Voss, puis un filtrage anti-Notch, une opération de module au carré et le filtre SG prend le relai avant de faire la classification des exons et introns à l'aide d'un seuillage en utilisant des paramètres d'évaluation statistique. Dans la figure (III.8), l'indice $i \in \{A, G, T, C\}$.

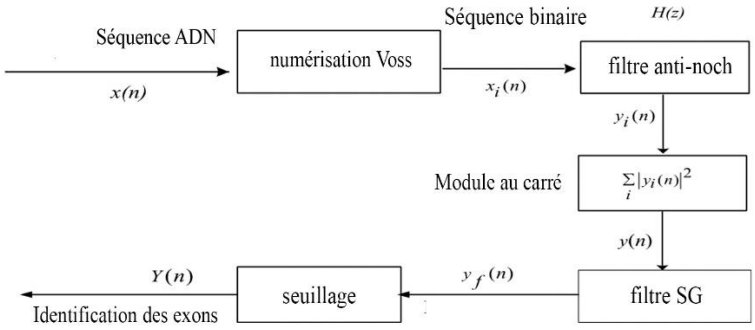


Figure III. 8 : Les étapes de la méthode GSP en utilisant le filtre Savitzky-Golay [93].

Cette méthode fut testée sur cinq gènes différents dont le F56F11.4 du C-elegans et les résultats étaient nettement meilleurs comparés aux méthodes basées sur les filtres anti-Notch [111], sur la transformée en ondelette [108], [120], [121]. Le vrai avantage de cette méthode repose dans les propriétés du filtre Savitzky-Golay [122] : un rapport signal/bruit élevé tout en préservant les caractéristiques importantes du signal tels que les pics et la puissance du signal en amplitude.

III.6. Comparaison générale entre filtres RIF et RII

Kar et al. [91] proposent une étude comparative entre différents filtres RIF et RII. Pour cela, 3 filtres de chaque groupe ont été évalués en suivant l'organigramme de la figure (III.9) suivante :

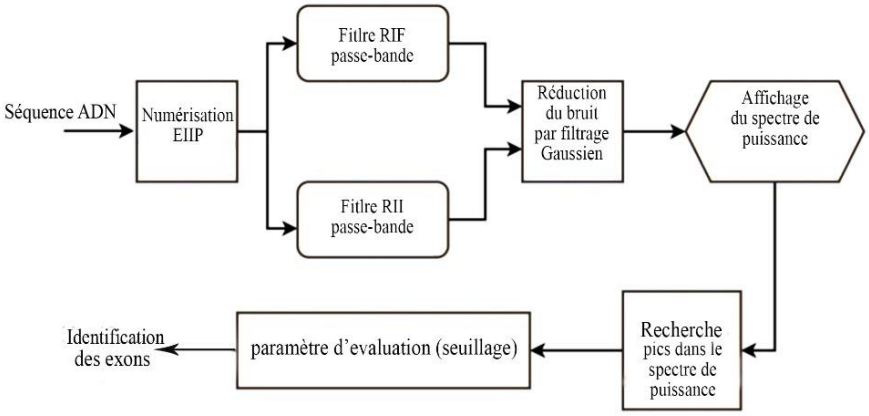


Figure III. 9 : Organigramme du processus d'identification des exons avec filtre RII et RIF [91].

Les caractéristiques des filtres sont :

Pour le RIF :

- Fenêtre Kaiser (filtre de Kaiser) : [0.665,0.667] pour les fréquences normalisées (Nyquist) de la bande passante.
- Filtre Parks-Macllelan : 0.4 dB et 30 dB d'atténuation pour la bande passante et la bande coupée, respectivement.
- Filtre à moindre carré : [0.65,0.68] pour les fréquences normalisées (Nyquist) de coupures de la bande coupée.

Pour le RII :

- Filtre Butterworth : 0.4 dB et 30 dB d'atténuation pour la bande passante et la bande coupée, respectivement.
- Filtre Tchebychev de type II : [0.664,0.672] pour les fréquences normalisées (Nyquist) de la bande passante.
- Filtre elliptique : [0.659,0.678] pour les fréquences normalisées (Nyquist) de la bande coupée.

Kar et al. [91] utilisent la technique EIIP pour la numérisation des séquences d'ADN. L'évaluation de l'étude se faisait généralement avec le gène F.56F11.4a du C-elegans et en comparant les valeurs des paramètres d'évaluation avec une valeur de seuil s=0.2 (2%). Bien que le filtre RIF (Parks-McClellan) donne de bons résultats, la conclusion générale est que les filtres RII sont plus performants que les RIF pour l'identification des exons.

III.7. Conclusion

Les méthodes GSP présentées dans ce chapitre sont considérées comme des méthodes « références » pour l'identification des exons dans les cellules eucaryotes. Elles utilisent des techniques de transformations du signal (Fourier, ondelette) et le filtrage numérique afin d'exploiter la propriété de « 3-base periodicity » qui se manifeste par la présence d'un pic dans la région des exons lors de l'analyse spectrale de la séquence à une fréquence f = N/3 (N étant la longueur de la séquence). Il est intéressant de souligner que les représentations de Voss et d'EIIP sont les plus couramment utilisées par ces méthodes GSP pour convertir la séquence d'ADN en valeurs numériques. Bien que chaque méthode soit évaluée sur différents gènes, le gène F56F11.4 de l'organisme C-elegans est le plus utilisé lorsqu'on applique les paramètres d'évaluation statistique au niveau nucléotidique.

CHAPITRE IV - CONCEPTION D'UN FILTRE FRACTIONNAIRE POUR L'IDENTIFICATION DES EXONS

IV.1. Introduction

Comme présentées dans le chapitre précédent, les méthodes de traitement numérique du signal génomique (GSP) exploitent surtout jusqu'à présent la propriété « 3-base periodicity » et se basent généralement sur les transformations et les filtrages numériques. Les outils fractionnaires ont déjà été utilisés avec succès dans plusieurs domaines, mais aussi dans le traitement de certains signaux biomédicaux [123]. Ce chapitre explore la possibilité de l'utilisation de modèles basés sur le calcul d'ordre fractionnaire dans l'identification des exons chez les eucaryotes. Premièrement, une présentation générale des fractales, du calcul d'ordre fractionnaire et du filtre fractionnaire. Ensuite, nous citerons quelques modèles basés sur le calcul d'ordre fractionnaire. Enfin, nous présenterons la conception du filtre fractionnaire pour l'identification des exons chez les eucaryotes.

IV.2. Définition générale des fractales

« Les fractales sont des objets mathématiques qui ont des propriétés d'autosimilarité (des motifs qui se répètent de manière similaire) à différentes échelles » [124].

En termes de signaux stationnaires, les fractales peuvent être définies comme des signaux qui présentent des propriétés statistiques invariantes dans le temps [125].

Dans le cas des signaux non-stationnaires, les fractales peuvent être utilisées pour représenter la variation à différentes échelles dans le signal. Cela peut être utile pour modéliser des signaux avec des caractéristiques complexes, tels que des signaux biologiques ou des séries temporelles financières [126].

Les fractales ont alors de nombreuses applications dans les domaines scientifiques et technologiques. Elles peuvent être utilisées pour modéliser et caractériser des systèmes dynamiques complexes, et pour extraire des informations utiles à partir de signaux bruités ou de faible qualité. Elles sont utilisées en biologie pour étudier la répartition des structures des plantes, bactéries et feuilles. En géologie, elles permettent d'analyser le relief, les côtes, les cours d'eau et les structures des roches. En médecine, elles sont utilisées pour étudier la structure des organes ainsi que des phénomènes physiologiques comme les battements du cœur. Les fractales sont également utilisées en météorologie pour analyser les nuages, les vortex et les turbulences. Dans le domaine de l'économie et de la finance, elles sont utilisées pour évaluer les risques financiers et tenter de prévoir les crashs boursiers. Les fractales sont également

présentes dans les arts graphiques, la littérature, la musique et le cinéma. Ces exemples montrent l'importance des fractales dans l'étude des structures complexes [127].

IV.2.1. Calcul d'ordre fractionnaire

IV.2.1.1. Définition

Le calcul d'ordre fractionnaire (COF), également appelé calcul fractionnaire, est une branche des mathématiques qui étend les opérations de dérivation et d'intégration à des ordres arbitraires. En mathématiques classiques, ces opérations sont généralisées exclusivement pour des ordres entiers (1, 2, 3, ...). Cependant, en calcul d'ordre fractionnaire, les ordres peuvent être généralisés à des nombres réels, incluant des fractions et même des nombres irrationnels [128].

L'expression ${}_cD_t^{\alpha}x_{\alpha}(t)$ est une représentation de l'opérateur de dérivation d'ordre fractionnaire. Elle indique la dérivée fractionnaire d'une fonction (signal analogique) $x_{\alpha}(t)$ par rapport à la variable t. La valeur α représente l'ordre de la dérivation fractionnaire. Si α est positif (α > 0), cela indique une dérivation d'ordre fractionnaire. Si α est négatif (α < 0), cela indique une intégration d'ordre fractionnaire. Avec c comme borne inférieure et t comme borne supérieure. Ces bornes spécifient l'intervalle sur lequel la dérivation d'ordre fractionnaire est appliquée [123].

La définition de Grünwald-Letnikov est couramment utilisée parmi d'autres en offrant une approche pratique pour calculer les dérivées et intégrales fractionnaires se basant sur des différences finies. Cela permet d'obtenir des approximations numériques des opérations d'ordre fractionnaire, facilitant ainsi l'implémentation et la résolution de problèmes dans le domaine numérique [123].

En prenant α comme nombre réel, c remis à zéro, l'opérateur ${}_cD_t^{\alpha}x_{\alpha}(t)$ devient D^{α}, et selon la définition de Grünwald-Letnikov, $D^{\alpha}x_{\alpha}(t)$ est donné par :

$$D^{\alpha}x_{\alpha}(t) = \lim_{h\to 0} \frac{1}{T_s^{\alpha}} \sum_{i=0}^{\left[\frac{t}{T_s}\right]} \frac{\Gamma(i-\alpha)}{\Gamma(-\alpha)\Gamma(i+1)} x_{\alpha}(t - iT_s) \qquad \text{(IV.1)}$$

Dans la relation (IV.1), "h" et "T_s" sont des paramètres qui contrôlent la discrétisation temporelle de la fonction $x_{\alpha}(t)$ et déterminent la façon dont les échantillons de la fonction sont pris en compte dans le calcul de la dérivée fractionnaire. Cela permet d'obtenir une approximation de la dérivée fractionnaire d'ordre α de la fonction $x_{\alpha}(t)$ en utilisant des

échantillons très proches dans le temps. La fonction gamma $\Gamma(.)$ dont (.) est la partie entière est définie dans la relation (IV.2) suivante :

$$\Gamma(a) = \int_0^{+\infty} e^{-t} t^{a-1} dt \qquad \text{(IV.2)}$$

IV.2.1.2. Application du calcul d'ordre fractionnaire dans le domaine biomédical

Le calcul d'ordre fractionnaire (COF) a apporté des outils comme des modèles et des filtres numériques pour analyser des signaux biomédicaux qui ont des propriétés fractales. Les exemples d'études utilisant des modèles fractionnaires basés sur le COF étaient faits pour l'étude sur la relaxation et la diffusion en résonance magnétique nucléaire, la dynamique du virus de l'hépatite C, les émotions humaines, l'infection du VIH dans les cellules T CD4+, les oscillations posturales et la chimiotaxie bactérienne [129] – [135].

Les filtres fractionnaires ont beaucoup été utilisés pour les signaux des électrocardiogrammes (ECG), les électroencéphalogrammes (EEG) et les signaux EMG (électromyogrammes), en traitement d'image médicale, analyse de la parole, en ophtalmologie (analyse de l'iris), qui peuvent être très complexes et présenter des caractéristiques non linéaires [136] – [150].

Des modèles basés sur les COF sont aussi utilisés pour créer des modèles qui génèrent des signaux fractals ayant les propriétés comme : une fonction de distribution de probabilité à décroissance lente, une fonction d'autocorrélation avec une décroissance asymptotique et une DSP de type 1 / f. Parmi ces modèles, on peut citer le FARIMA (Fractional Autoregressive Integrated Moving Average) et le GARMA (Generalized Autoregressive Moving Average) [151], [152].

IV.3. Conception de filtre basé sur le calcul d'ordre fractionnaire

Du point de vue de la théorie des systèmes, la conception de filtres numériques basés sur le calcul d'ordre fractionnaire consiste à passer un signal d'entrée $x_\alpha(t)$ à travers un filtre ayant une fonction de transfert comme dans la relation (IV.3) suivante.

$$H_\alpha(s) = s^\alpha \qquad \text{(IV.3)}$$

Ce filtre correspond à un dérivateur d'ordre fractionnaire analogique lorsque $\alpha > 0$, et à un intégrateur d'ordre fractionnaire analogique lorsque $\alpha < 0$.

Cette approche décrite dans la relation (IV.3) est utilisée dans beaucoup de travaux [153] – [157]. Afin de pouvoir implémenter le filtre fractionnaire décrit dans la relation (IV.3), un processus de discrétisation est nécessaire pour l'approximation du filtre idéal. Ce processus de discrétisation peut se faire généralement en deux étapes .

- **Premièrement**, une fonction de mappage appropriée entre le plan s et le plan z doit se faire pour obtenir s = M(z). Cela se fait souvent par l'utilisation des règles d'intégrations numériques conventionnelles en utilisant : les lois rectangulaire (Euler), trapézoïdale (Tustin) et de Simpson, ou de combinaison linéaire comme les lois d'Al-Alaoui. Cette dernière ayant montré son efficacité pour le COF [158].
- **Deuxièmement**, une technique de troncature (troncature basée sur la méthode de Muir, troncature de la fonction continue, troncature de la série de puissance…) ou de modélisation du signal déterministe est utilisée pour obtenir la fonction de transfert approximative du filtre numérique suivante [158] :

$$H(z) \approx H_\alpha|_{s=M(z)} \qquad \text{(IV.4)}$$

Cette fonction de transfert (IV.4) approximative de (IV.3) permet de calculer la dérivée et l'intégrale d'ordre α du signal discret d'entrée $x(n) = x_\alpha(nT_s)$, de sorte que les échantillons de sortie y(n) soient aussi proches que possible des échantillons de sortie $y_\alpha(nT) = D^\alpha x_\alpha(t)$ aux instants d'échantillonnage $t_n = nT_s$. La figure (IV.1) montre le modèle de discrétisation de ce processus.

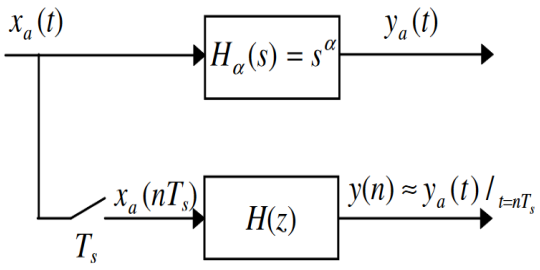


Figure IV. 1 : Schéma synoptique pour l'approximation de la fonction de transfert du filtre idéal [123].

Des exemples d'étude sur les signaux d'électrocardiogramme (ECG) et d'électromyogramme (EMG) [123], [137], [143], utilisent des filtres fractionnaires dérivateurs passe-bas RIF et de lissage RIF qui peuvent être obtenus en combinant les fonctions de transfert approximatives suivantes :

$$s^\alpha \approx H_{\alpha a}(z) = \frac{(1-z)^\alpha}{T_s^\alpha} = \frac{1}{T_s^\alpha}\sum_{i=0}^{\infty} a_i\, z^i \text{ , avec } \alpha < 0 \qquad \text{(IV.5)}$$

et

$$s^{\alpha} \approx H_{\alpha c}(z) = \frac{\left(1-z^{-1}\right)^{\alpha}}{T_s^{\alpha}} = \frac{1}{T_s^{\alpha}} \sum_{i=0}^{\infty} a_i \, z^{-i} \text{ , avec } \alpha < 0 \quad \text{(IV.6)}$$

Les coefficients a_i s'expriment par la relation de récurrence (IV.7)

$$a_0 = 1, a_i = \frac{i-\alpha-1}{i} a_{i-1}, i = 1,2,3, \ldots \quad \text{(IV.7)}$$

Dans les relations (IV.5) et (IV.6), les termes « a » et « c » désignent les propriétés temporelles d'anti-causalité et de causalité.

IV.4. Quelques modèles basés sur le calcul d'ordre fractionnaire

IV.4.1. Le "Fractional Autoregressive Integrated Moving Average" - FARIMA

Le cas général du modèle FARIMA (p, α, q) décrit un modèle fractionnaire autorégressif intégré à moyenne mobile qui peut être utilisé à des signaux biomédicaux avec des dépendances longues et courtes [123], [152]. Avec des valeurs de p et q égales à zéro, un modèle FARIMA (0, α, 0) est défini par l'équation de différence suivante :

$$(1 - B)^{\alpha} x(n) = w\,(n) \quad \text{(IV.8)}$$

Dans la relation (IV.8), x(n) est le signal à générer, w(n) est un processus de bruit blanc gaussien à l'entrée avec une variance σ_{ω}^2 et une moyenne μ_{ω}, et B est l'opérateur de retard défini comme $Bx(n) = x\,(n-1)$.

L'équation (IV.8) peut être écrite, en transformée en Z, comme suit :

$$X\,(z) = H_{\alpha}\,(z)\, W\,(z) \quad \text{(IV.9)}$$

avec

$$H_{\alpha}(z) = (1 - z^{-1})^{-\alpha} = \sum_{n=0}^{\infty} h_{\alpha}(n) z^{-n} \quad \text{(IV.10)}$$

Dans les relations (IV.9) et (IV.10), $H_{\alpha}(z)$ représente la fonction de transfert du filtre numérique idéal (intégrateur d'ordre fractionnaire) issue de la fonction de transfert du filtre analogique $s^{-\alpha}$ par la transformation (mappage plan s au plan z) $s = \frac{1-z^{-1}}{T_S}$ avec $T_S = 1$. Ce filtre numérique idéal d'ordre fractionnaire génère le signal x(n) quand le signal d'entrée est un bruit blanc gaussien w(n). Les coefficients $h_{\alpha}(n)$ se calculent comme dans la relation (IV.7) avec $h = a \; et \; n = i$.

Il est observé dans la relation (IV.8) que : pour 0 < α < 0,5, le signal x(n) généré par le modèle FARIMA présente une dépendance à long-terme. Ce modèle peut alors être utilisé pour générer des processus autosimilaires (fractionnaires) avec un paramètre Hurst H[2] lié au paramètre de différenciation fractionnaire selon la relation $H = \alpha + 0{,}5$ [152].

IV.4.2. Le "Generalized Autoregressive Moving Average" ou GARMA

Une extension du modèle FARIMA est proposée dans [123], [152], en ajoutant un paramètre ν qui permet à la densité spectrale de puissance (DSP) d'être infinie à une fréquence non-nulle, ce qui prend en compte le comportement cyclique d'un signal aléatoire. Cette nouvelle extension est un processus généralisé autorégressif à moyenne mobile, abrégé en GARMA [151].

Le modèle GARMA (p, α, ν, q) d'un processus aléatoire stationnaire à temps discret $x(n)$, n = 0,1,2, 3, est représenté dans la relation (IV.11).

$$A(z^{-1})x(n) = B(z^{-1})(1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha} w(n) \qquad \text{(IV.11)}$$

avec

$A(z^{-1})$: polynôme autorégressif d'ordre p

$B(z^{-1})$: polynôme à moyenne mobile (moving average) d'ordre q

$w(n)$: un bruit blanc Gaussien

α : paramètre qui contrôle la dépendance à court-terme

ν : paramètre qui contrôle la périodicité du processus

$(1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha}$: polynôme fractionnaire aussi appelé polynôme Gegenbauer [159]

z^{-1} : opérateur de retard (décalage)

On remarque que si ν = 1, le polynôme fractionnaire $(1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha}$ devient $(1 - z^{-1})^{-2\alpha}$. Ce dernier est le polynôme fractionnaire commun associé au cas du modèle FARIMA (0,2α,0), qui est caractérisé par une densité spectrale de puissance infinie à une fréquence nulle. Lorsque l'entrée w(n) est un bruit blanc stationnaire, la sortie x(n) est stationnaire si α < 0,5 et | υ | < 1 ou si α < 0,25 et | υ | = 1.

[2] Paramètre de Hurst : utilisé comme mesure de la mémoire à long terme des séries temporelles. Il concerne les autocorrélations des séries chronologiques et le taux auquel elles diminuent à mesure que le décalage entre les paires de valeurs augmente.

La sortie x(n) est inversible si α > -0,5 et | υ | <1 ou α > -0,25 et | υ | = 1.

De la relation (IV.11), on obtient la fonction de transfert d'un filtre fractionnaire en (IV.12).

$$H(z) = (1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha} \frac{B(z^{-1})}{A(z^{-1})} \qquad \text{(IV.12)}$$

Le modèle GARMA (0, α, ν,0) de la relation (IV.11) peut s'écrire comme suit :

$$x(n) = (1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha} w(n) \qquad \text{(IV.13a)}$$

avec comme fonction de transfert : $H_{\alpha,\upsilon}(z) = (1 - 2\upsilon z^{-1} + z^{-2})^{-\alpha}$ (IV.13b)

et l'amplitude de la réponse fréquentielle est donnée par la relation (IV.14) :

$$\left|H_{\alpha,\nu}(f)\right| = \left|e^{-j2\pi f}\left(e^{j2\pi f} + e^{-j2\pi f} - 2\nu\right)\right|^{-\alpha} \qquad \text{(IV.14)}$$

La relation (IV.15) suivante montre la densité spectrale de puissance (DSP) du modèle,

$$S_{xx}(f) = \frac{\sigma_w^2}{[2|\cos(2\pi f) - \nu|]^{2\alpha}}, \; -0.5 \leq f \leq 0.5 \qquad \text{(IV.15)}$$

Avec σ_w^2 la variance du bruit blanc gaussien.

Contrairement au cas du modèle FARIMA, l'amplitude de la DSP du processus GARMA présenté dans la relation (IV.15) est non-bornée et tend vers l'infini à la fréquence de Gegenbauer qui est $f_G = cos^{-1}(\nu)/2\pi$. La figure (IV.2) montre différentes DSP selon la relation (IV.15) avec différentes valeurs de la fréquence de Gegenbauer, c'est-à-dire différentes valeurs de ν.

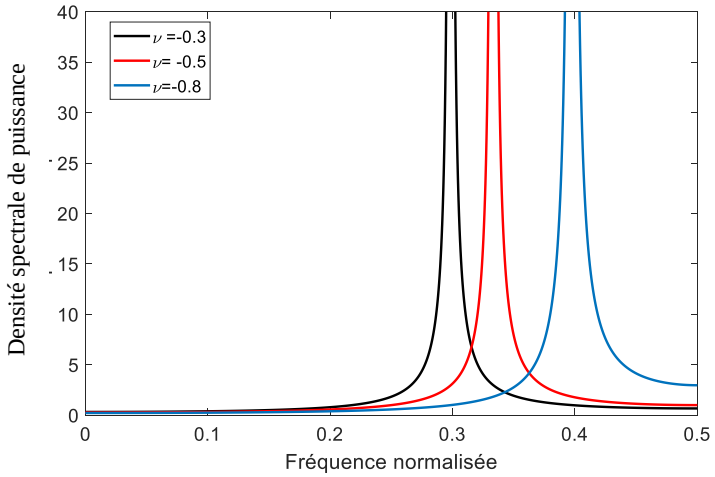


Figure IV. 2 : Densité spectrale de puissance du processus GARMA
avec α = 0.6 pour différentes valeurs de ν = -0.3, -0.5 et -0.8.

La fonction d'autocorrélation r_{xx} est donnée par la relation suivante :

$$r_{xx}(k) = 2\int_0^{0.5} S_{xx}(f)\cos(2\pi kf)\,df,\ k = 0, \pm 1, \ldots \quad \text{(IV.16)}$$

Avec $-0.5 \leq f \leq 0.5$. Cette relation (IV.16) est la transformée de Fourier inverse de la DSP selon le théorème de Wiener-Khintche [3] [151].

Avec $\alpha < 0.5\ et\ |v| < 1$, La relation (IV.16) peut aussi être approximée comme suit :

$$r_{xx}(k) \sim k^{2\alpha-1}\cos(2\pi f_G k)\ \ comme\ k\ \to \infty \quad \text{(IV.17)}$$

Cette autocorrélation décroît selon une loi hyperbolique et présente une oscillation périodique liée à la fonction cosinus [151]. Cela indique que le signal généré par le modèle GARMA (0, α, ν,0) défini par l'équation (IV.13a) est caractérisé par une forte dépendance à long terme avec une composante périodique. Des algorithmes efficaces pour la génération de signaux fractals (avec propriété spectrale $1/f^{\beta}$) peuvent être alors développés en utilisant le modèle GARMA. La figure (IV.3) illustre cette autocorrélation de la séquence générée par le processus GARMA. C'est le cas de l'autocorrélation observée pour dix échantillons de séquence obtenue à partir de dix échantillons du signal d'entrée (bruit blanc) avec une longueur de 5000 (5000 points).

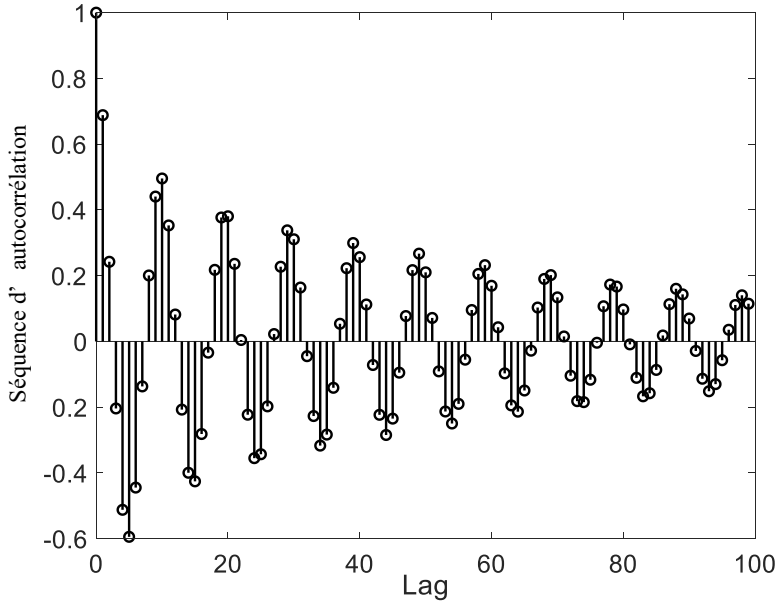


Figure IV. 3 : Echantillon de séquence d'autocorrélation issue du processus GARMA avec α = 0.45, ν = 0.8. Normalisé à 1 à zéro Lag.

[3] Le théorème de Wiener-Khintche affirme que la DSP d'un procédé stochastique stationnaire est analogue à la transformée de Fourier de la fonction d'autocorrélation correspondante.

IV.5. Conception du filtre fractionnaire pour l'identification des exons

IV.5.1. Propriétés fractionnaires des séquences d'ADN

Il a été démontré il y a longtemps que les séquences d'ADN peuvent présenter des propriétés des signaux fractals, autocorrélations à longue distance avec des motifs similaires qui se répètent à différentes échelles et la présence du bruit de type $1/f^{\beta}$. Ces propriétés étaient surtout utilisées pour développer des méthodes de représentations graphiques de la séquence d'ADN. L'autre propriété très importante observée dans la séquence d'ADN est la « 3-base periodicity ». Cette dernière a été depuis longtemps exploitée en utilisant des filtres numériques, des transformations du signal afin d'identifier les exons, comme détaillé dans le 3ème chapitre. En combinant ces propriétés à l'utilisation des modèles basés sur le COF, on peut surement concevoir une nouvelle approche pour l'identification des exons dans les cellules eucaryotes [55] – [58], [62], [63], [65].

IV.5.2. Filtre fractionnaire idéal pour l'identification des exons

On peut concevoir un filtre numérique basé sur les COF en utilisant le modèle GARMA [151]. Un filtre dont les propriétés peuvent extraire les informations utiles dans l'identification des exons chez les eucaryotes.

Dans le cas du modèle GARMA (0, α, ν,0), la fonction de transfert du filtre numérique fractionnaire est donnée par la relation (IV.13b) :

En substituant $z = e^{j2\pi f}$ et $\theta = 2\pi f$, l'amplitude de la réponse fréquentielle est :

$$\left|H_{\alpha,\nu}(\theta)\right| = \left|e^{-j\theta}\left(e^{j\theta} + e^{-j\theta} - 2\nu\right)\right|^{-\alpha} \qquad \text{(IV.18)}$$

Avec la méthode d'approximation trigonométrique d'Euler

$e^{-j\theta} = cos\theta - jsin\theta$ et $e^{j\theta} = cos\theta + jsin\theta$

La relation (IV.18) peut s'écrire comme :

$$\left|H_{\alpha,\nu}(\theta)\right| = \left|e^{-j\theta}(2\cos\theta - 2\nu)\right|^{-\alpha}$$

$$= |(cos\theta - jsin\theta)\,(2cos\theta - 2\upsilon)|^{-\alpha}$$

$$= [2\left|\sqrt{cos^2\theta + sin^2\theta}\right||cos\theta - \upsilon|]^{-\alpha}$$

Avec $|a + ib| = \sqrt{a^2 + b^2}$ et comme $\sqrt{cos^2\theta + sin^2\theta} = 1$

On a alors $\left|H_{\alpha,\nu}(\theta)\right| = [2|cos\theta - \nu|]^{-\alpha}$

Ou bien $\left|H_{\alpha,\nu}(f)\right| = [2|\cos(2\pi f) - \nu|]^{-\alpha}$, $|f| \leq 0.5$ (IV.19)

Pour α>0, l'amplitude de la réponse fréquentielle (IV.19) est non-bornée à la fréquence centrale donnée par :

$$f_G = \frac{cos^{-1}(\nu)}{2\pi}, \ |\nu| < 1 \quad \text{(IV.20)}$$

La fréquence décrite dans la relation (IV.20) est appelée fréquence de Gegenbauer dans le contexte de l'analyse de séries temporelles [159]. A ce niveau, la remarque importante est que valeur de la fréquence centrale du filtre fractionnaire dans la relation (IV.20) ne dépend que de la valeur du paramètre ν.

IV.5.2.1. Fixation des valeurs de ν

La valeur du paramètre ν (contrôlant la périodicité du signal) permet de choisir la valeur de la fréquence centrale du filtre numérique donnée par la relation (IV.20). En se basant sur les études sur l'utilisation des filtres numériques dans l'identification des exons dans le chapitre III (III.5), il est préférable que le filtre fractionnaire soit de type anti-Notch avec une bande passante étroite et une fréquence centrale $f = 1/3$ ou $\omega = 2\pi/3$ pour pouvoir mettre en évidence la propriété « 3-base periodicity » des exons.

Ainsi, si $f = f_G = \frac{cos^{-1}(\nu)}{2\pi} = 1/3$ alors , $\nu = \cos(2\pi f_G)$ avec f_G=1/3

Ce qui donne une valeur de $\nu = -0.5$, respectant les conditions initiales de $|\nu| < 1$ selon la relation (IV.20) pour le modèle GARMA (0, α, ν,0).

IV.5.2.2. Comportement du filtre fractionnaire idéal pour différentes valeurs de α

Le paramètre α (qui contrôle la dépendance à court terme ou à long terme du signal) joue aussi un grand rôle dans les caractéristiques du filtre à concevoir.

Pour différentes valeurs de α > 0 (α=0.1,0.4 et 0.8) avec $\nu = -0.5$, les amplitudes des réponses fréquentielles du filtre sont montrées dans la figure (IV.4) et (IV.5).

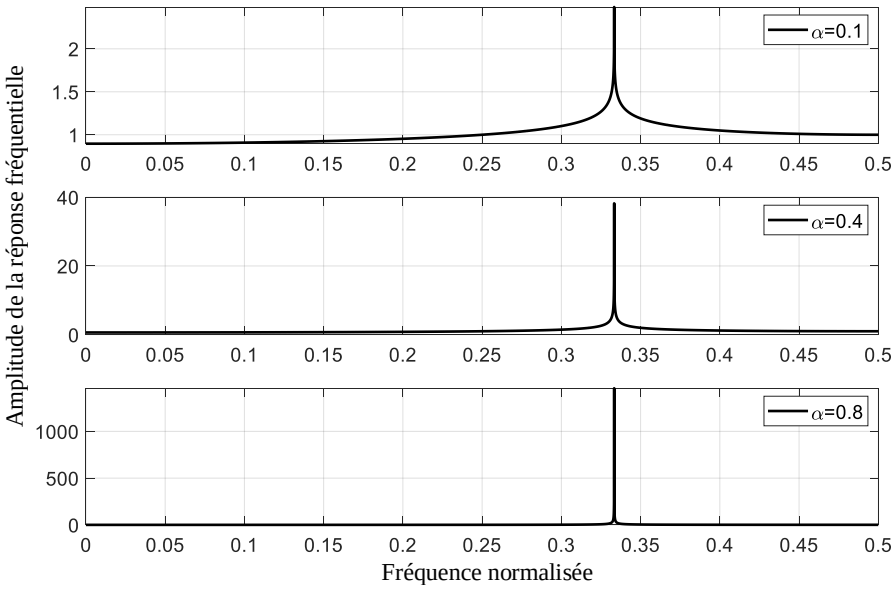


Figure IV. 4 : Réponse fréquentielle du filtre fractionnaire idéal à différentes valeurs de α : 0.1, 0.4, 0.8.

Dans la figure (IV.4), on observe que la valeur de l'amplitude de la réponse fréquentielle varie en fonction de la valeur de α . Plus α grandit, plus l'amplitude grandit.

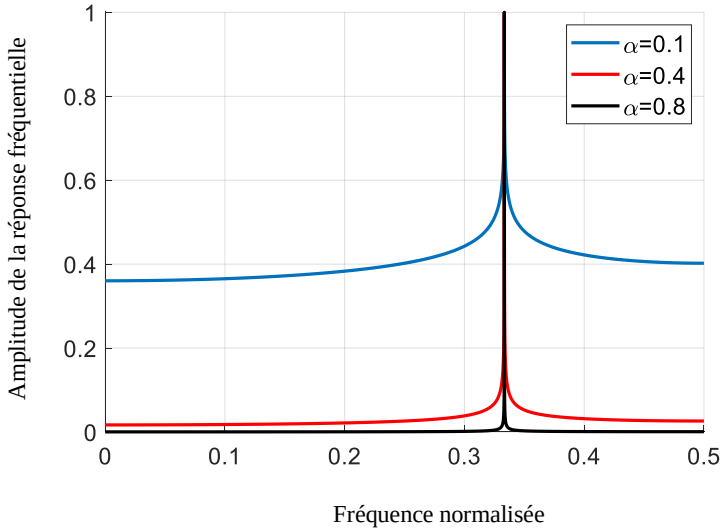


Figure IV. 5 : Amplitude normalisée de la réponse fréquentielle du filtre fractionnaire idéal à différentes valeurs de α : 0.1, 0.4, 0.8.

Pour les mêmes valeurs de α que dans la figure (IV.4) , on observe dans la figure (IV.5) que la largeur de la bande passante du filtre diminue quand α augmente.

Selon les valeurs de α et ν, le filtre fractionnaire idéal obtenu a une grande amplitude de la réponse fréquentielle et une bande passante étroite centrée sur la valeur de la fréquence f_G=1/3. Une autre caractéristique importante de ce filtre est l'absence d'oscillations dans la bande d'arrêt.

IV.5.3. Approximation du filtre fractionnaire idéal

La réponse impulsionnelle du filtre fractionnaire idéal obtenu par le processus GARMA (0, α, ν,0) est donnée par l'expression suivante :

$$h_{\alpha,\nu}(n) = \sum_{k=0}^{\left[\frac{n}{2}\right]} \frac{(-1)^k \Gamma(\alpha+n-k)(2\nu)^{n-2k}}{\Gamma(\alpha)\Gamma(k+1)\Gamma(n-2k+1)} \ , \ n = 0,1,2, ... \qquad \text{(IV.21)}$$

Avec α>0 et |ν|<1 , Γ(.) est la fonction gamma et [n/2] représente la partie entière de n/2 [159].

Dans (IV.21), la séquence $h_{\alpha,\nu}(n)$ avec n=0,1,2…, peut être calculée par l'équation de différence (IV.22) suivante :

$$h_{\alpha,\nu}(n) = 2\nu\left(\frac{\alpha-1}{n}+1\right)h_{\alpha,\nu}(n-1) - \left(2\frac{\alpha-1}{n}+1\right)h_{\alpha,\nu}(n-2) \qquad \text{(IV.22)}$$

avec $h_{\alpha,\nu}(0) = 1$ *et* $h_{\alpha,\nu}(1) = 2\alpha\nu$.

C'est un filtre à réponse impulsionnelle infinie (RII), dont l'expression générale suivante exprime la sortie y(n) :

$$y(n) = \sum_{i=0}^{\infty} h_{\alpha,\nu}(i)x(n-i) \qquad \text{(IV.23)}$$

où x(n) est le signal d'entrée.

Vue la nature du filtre RII, la relation (IV.23) n'est pas directement implémentable tel quel et nécessite une technique de discrétisation par approximation. D'après [123], [158], la technique de troncature conduit à un filtre RIF, tandis que l'approche par des techniques de modélisation de signal donne un filtre RII. De plus, dans le chapitre III (III.6), on a vu que le filtre RII est plus efficace lorsqu'il est appliqué à l'étude de données génomiques pour l'identification des exons [91].

L'estimation des coefficients de la fonction de transfert $H(z) \approx H_\alpha|_{s=M(z)}$ peut alors se faire en utilisant des techniques de modélisation du signal.

Parmi les techniques couramment utilisées sont les méthodes de Padé, Prony, Shanks et Steiglitz-McBride. Les études comparatives faites dans [158], montrent que la méthode d'approximation de Prony est plus performante et moins complexe pour l'estimation des coefficients du numérateur et du dénominateur de la relation (IV.24).

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{i=0}^{M} b_i z^{-i}}{1+\sum_{i=1}^{N} a_i z^{-i}} \quad \text{(IV.24)}$$

Ainsi, afin de calculer les coefficients $a_i \; et \; b_i$ de H(z), la méthode de Prony est utilisée à partir des premiers échantillons L de la réponse impulsionnelle $\mathrm{h}_{\alpha,\nu}(\mathrm{n})$ en (IV.22).

La sortie du filtre fractionnaire approximatif en (IV.23) peut être ensuite calculée par l'équation de différence suivante :

$$y(n) = -\sum_{i=1}^{N} a_i y(n-i) + \sum_{i=0}^{M} b_i x(n-i) \quad \text{(IV.25)}$$

Dans (IV.25), les valeurs M et N doivent être réelles et faibles afin de réduire la complexité des opérations.

IV.5.3.1. Paramétrage du filtre fractionnaire approximatif

Comme pour le filtre fractionnaire idéal, afin de pouvoir utiliser le filtre fractionnaire approximatif dans l'identification des exons, le paramètre ν est fixé à -0.5.

Les comportements du filtre approximatif selon les valeurs de $\alpha > 0$ sont montrés par les figures (IV.6) - (IV.8), en fixant les valeurs $M = N = 3$ et $L = 1000$ comme premiers échantillons de $\mathrm{h}_{\alpha,\nu}(\mathrm{n})$ en (IV.22). Les valeurs de M, N et L ont été estimées de façon empirique par différentes simulations.

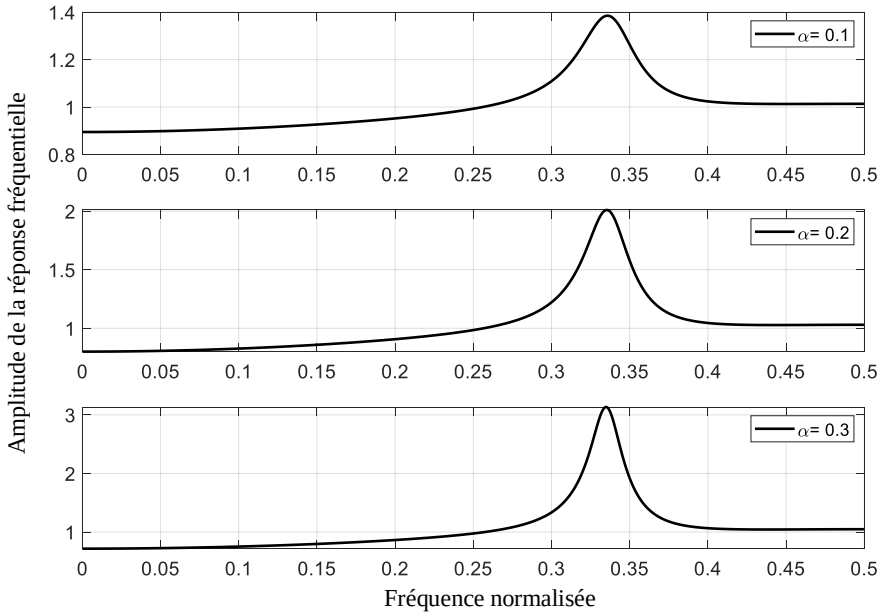


Figure IV. 6 : Réponse fréquentielle du filtre fractionnaire approximatif à différentes valeurs de α : 0.1, 0.2, 0.3.

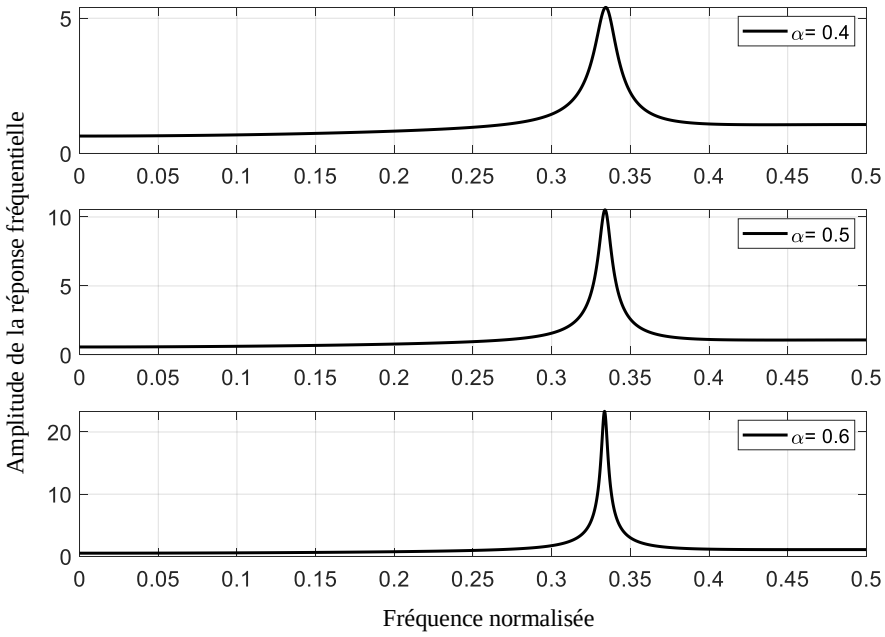


Figure IV. 7 : Réponse fréquentielle du filtre fractionnaire approximatif à différentes valeurs de α : 0.4, 0.5, 0.6.

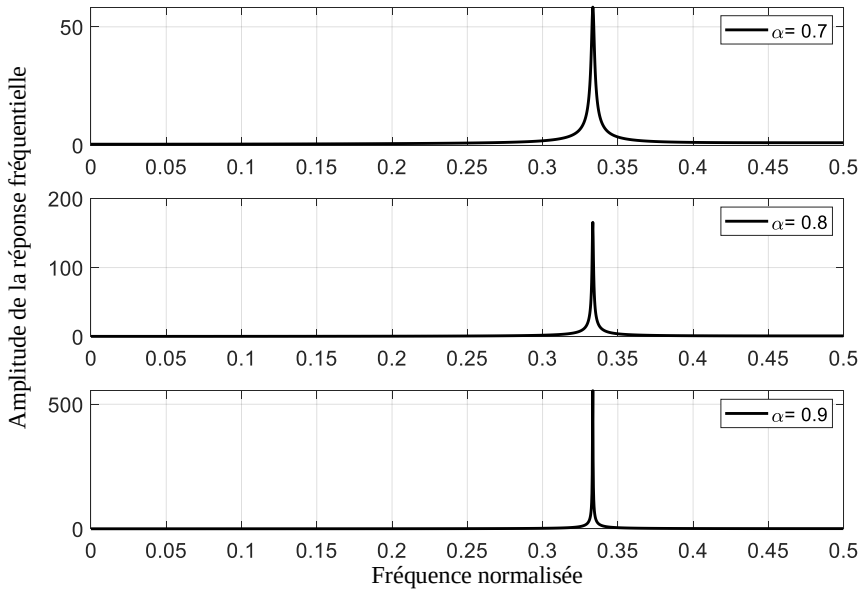


Figure IV. 8 : Réponse fréquentielle du filtre fractionnaire approximatif à différentes valeurs de α : 0.7, 0.8, 0.9.

On observe à partir des figures (IV.6) - (IV.8) que l'amplitude de la réponse fréquentielle du filtre approximatif varie suivant la valeur de α, tandis que la largeur de la bande passante du filtre devient étroite quand α augmente. On observe aussi que le filtre devient de plus en plus sélectif en centrant la fréquence sur f = 1/3 pour $\alpha > 0.5$ avec absence d'oscillations dans la bande d'arrêt. On peut dire alors que le filtre fractionnaire approximatif garde alors les propriétés observées pour le filtre fractionnaire idéal dans les figures (IV.4) et (IV.5).

IV.5.3.2. Stabilité du filtre fractionnaire approximatif

Comme tous les filtres numériques, la stabilité du filtre fractionnaire proposé se vérifie si les racines du dénominateur (les pôles N) de sa fonction de transfert en (IV.24) ont un module strictement inférieur à 1, c'est-à-dire, si les pôles se trouvent dans le cercle unité du plan complexe. Sachant que les modules de ces racines représentent les distances des pôles au point (0,0) dans le plan complexe, ils peuvent se calculer par la relation (IV.26) suivante.

$$|racine| = \sqrt{Re^2 + Im^2} \qquad \text{(IV.26)}$$

où Re est la partie réelle de la racine et Im sa partie imaginaire. Le filtre est alors stable si $|racine| = \sqrt{Re^2 + Im^2} < 1$.

Afin de calculer les coefficients de la fonction de transfert H(z) dans la relation (IV.24), la méthode d'approximation de Prony est utilisable [158].

Ainsi, pour α = 0.5, par approximation de Prony, les coefficients sont $a_0 = 1, a_1 = 1.1589,$ $a_2 = 1.1248,\ a_3 = 0.1685$ et $b_0 = 1, b_1 = 0.6589,\ b_2 = 0.4204,\ b_3 = -0.1013$.

Pour α = 0.6, les coefficients sont $a_0 = 1$, $a_1 = 1.2062$, $a_2 = 1.1881$, $a_3 = 0.2107$ et $b_0 = 1, b_1 = 0.6062,\ b_2 = 0.3444,\ b_3 = -0.1029.$

Pour α = 0.8, ils sont $a_0 = 1, a_1 = 1.2926,\ a_2 = 1.2890,\ a_3 = 0.2933$ et $b_0 = 1, b_1 = 0.4926,\ b_2 = 0.1750$, $b_3 = -0.0734.$

Pour les cas des coefficients calculés précédemment : avec α = 0.6, les modules des pôles sont 0.2163 ,0.8811 et 0.8811 qui sont tous < 1. De même pour α = 0.8, les modules des pôles sont 0.2948, 0.8952 et 0.8952 qui sont tous < 1. Les pôles sont à l'intérieur du cercle unité dans le plan complexe comme montré dans la figure (IV.9) pour le cas de α = 0.5.

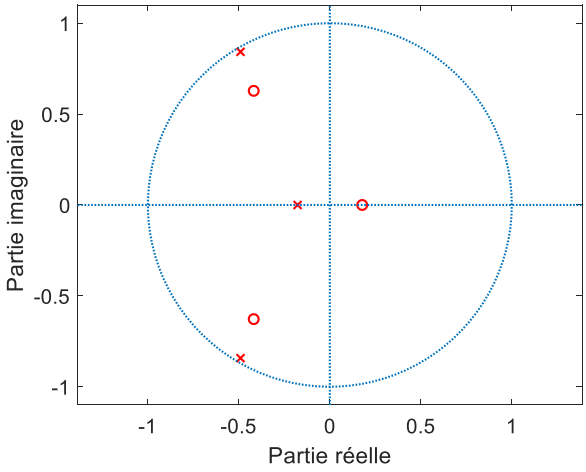


Figure IV. 9 : Distribution des pôles et zéros pour le cas du filtre fractionnaire approximatif issu du processus GARMA avec α = 0.5.

Le filtre fractionnaire résultant de la méthode d'approximation susmentionnée est stable, pour des valeurs de α > 0 (mais surtout α > 0.5) et préservent les propriétés du filtre fractionnaire idéal issu du modèle GARMA qui sont : un fort gain en amplitude de la réponse fréquentielle, bande passante centrée et étroite, pas d'oscillation dans les bandes d'arrêt.

IV.6. Conclusion

Dans ce chapitre, nous avons présenté les notions de bases du COF avec ses outils et modèles. Nous avons vu que ces outils ont déjà fait leurs preuves dans d'autres domaines d'études de signaux complexes tels que les signaux biologiques, signaux financiers. Dans le domaine biomédical, les ECG, EEG et EMG comptent parmi les signaux ayant fait appel aux outils fractionnaires [123]. La conception de certains modèles comme le FARIMA et GARMA permettent de concevoir des filtres numériques dont les caractéristiques sont efficaces pour une nouvelle méthode GSP dans l'identification des exons chez les eucaryotes.

CHAPITRE V - APPLICATION DU FILTRE FRACTIONNAIRE DANS L'IDENTIFICATION DES EXONS : SIMULATIONS, RESULTATS ET DISCUSSION

V.1. Introduction

L'application d'une nouvelle approche en traitement numérique du signal génomique (GSP) basée sur le filtre fractionnaire est présentée dans ce dernier chapitre. Un chapitre consacré aux simulations, résultats et interprétations de cette méthode pour l'identification des exons de la séquence d'ADN chez les cellules eucaryotes. Plusieurs points seront d'abord présentés tels que : les bases des données génomiques utilisées, l'environnement de programmation, l'algorithme et l'organigramme de la méthode proposée. Par la suite, nous présenterons les différents paramétrages du filtre fractionnaire avant de faire l'évaluation de ses performances. La dernière partie du chapitre consiste à présenter et interpréter les résultats en comparant avec ceux des méthodes GSP implémentées et issues de la littérature, présentes dans le chapitre III. Pour le reste de ce chapitre, le terme « méthode proposée » désigne la méthode GSP basée sur l'utilisation du filtre fractionnaire conçu dans le chapitre précédent.

V.2. Bases de données utilisées

Le choix des bases de données pour collecter les séquences d'ADN à tester est une étape primordiale. Parmi les bases de données existantes, nous avons sélectionné deux : NCBI et HMR195 qui seront présentées ultérieurement. Ces séquences sont utilisées dans la majorité des études sur l'identification des exons dans les cellules eucaryotes [50], [54], [91], [93], [95], [107], [108], [111], [113], [119], [121], [160].

V.2.1. HMR195

Dans le premier chapitre, nous avons présenté les bases de données génomiques qui sont disponibles pour le traitement numérique des séquences d'ADN. La plus importante actuellement est le NCBI qui compte 241015745 séquences d'ADN pour 500 000 espèces [34].

Le HMR195 est une base de données issue du NCBI, organisée par Rogic et al. [40], [44]. Cette base de données fut spécialement créée pour évaluer des nouvelles méthodes d'identifications des exons. Elle contient 195 gènes bien localisés provenant de 3 espèces vivantes qui sont : l'homme, le rat et la souris grise, en termes scientifiques : Homo sapiens, Mus musculus et Rattus norvegicus, d'où l'abréviation « HMR ». En ce qui concerne les caractéristiques de l'ensemble des données, le rapport des séquences humaines, souris et rats est de 103 : 82 : 10 c'est-à-dire, 103 séquences humaines, 82 pour la souris et 10 pour le rat. La longueur moyenne des séquences est de 7 096 pb. Le nombre de gènes à exon unique est de 43 et le nombre de

gènes à plusieurs exons est de 152 (2 à 26 exons). Le nombre moyen d'exons par gène est de 4,86. La longueur moyenne des exons est de 208 pb, la longueur moyenne des introns est de 678 pb, et la longueur totale moyenne des gènes est de 4 303 pb.

V.2.2. F56F11.4 du Caenorhabditis elegans (C.elegans)

Depuis la publication de son génome complet en 1998 [24], le Caenorhabditis elegans ou C.elegans (un ver de terre, rond, nématode) est un organisme modèle largement utilisé en génomique et surtout dans les études sur l'identification des exons par les méthodes GSP. Le C.elegans est facile à cultiver en laboratoire et possède un cycle de vie court. Sa transparence permet d'observer facilement les structures internes et les processus biologiques tels que la division cellulaire, la migration cellulaire et la différenciation cellulaire. Le génome du C.elegans contient environ 20 000 gènes, tous identifiés. Dans sa version la plus courante, le gène F56F11.4 du C.elegans contient 8000 pb et 5 exons bien localisés, et constitue le gène de référence en GSP [91].

Le tableau (V.1) résume les informations sur les séquences d'ADN utilisées dans ce travail.

Base de données	Description	Longueur de la séquence (pb)	Information sur les exons	Nombre d'exons
HMR195	Humain :103 séquences Souris :82 séquences Rat :10 séquences	Longueur moyenne des séquences : 7.096 pb	Nombre moyen d'exon par gène : 4.86	43 séquences avec 1 exon 33 séquences avec 2 exons 25 séquences avec 3 exons 21 séquences avec 4 exons 10 séquences avec 5 exons 63 séquences avec [6,26] exons
Caenorhabditis elegans (F56F11.4) NCBI	Numéro d'accès n°AF099922	8000 pb	928…1039 (111 pb) 2528…2857 (329 pb) 4114…4377 (263 pb) 5465…5644 (179 pb) 7255…7605 (350 pb)	5

Tableau V. 1 : Information sur les séquences et base de données utilisées [34], [40].

V.3. Environnement informatique

V.3.1. Logiciel MATLAB

MATLAB (Matrix laboratory), plus précisément la version R2019a, est le logiciel de programmation et de simulation que nous avons utilisé pour nos expériences. L'existence de sous-programmes, fonctions et boîtes à outils (toolbox) préinstallés permettent d'accélérer la programmation de l'algorithme. Parmi les sous-programmes et fonctions que nous avons sollicités, citons : le « Signal processing toolbox 8.2 » et le « Design and simulate streaming

signal processing systems toolbox » pour les traitements numériques des signaux, la génération de signaux, l'analyse spectrale, l'implémentation des filtres et la transformation numérique des signaux, les « Graphics functions » qui regroupent tous les fonctions pour les traçages des courbes, le « Bioinformatics toolbox 4.12 » pour l'importation et la visualisation des données génomiques, le « prony.m » pour la modélisation du signal [42].

V.3.2. Equipement informatique

Le tableau (V.2) suivant présente les caractéristiques de l'équipement informatique que nous avons utilisé pour les expériences :

Fabricant	ASUSTeK COMPUTER INC.
Modèle	GL752VW
Système d'exploitation	Windows 10 Pro 64-bit
Processeur	Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz (8 CPUs), ~2.6GHz
Mémoire vive (Random Access Memory)	8192 MB
Carte graphique intégrée	Intel(R) HD Graphics 530 - 4173 MB
Carte graphique dédiée	NVIDIA GeForce GTX 960M - 6055 MB

Tableau V. 2 : Caractéristiques de l'équipement informatique utilisé.

V.4. Algorithme proposé pour l'identification des exons

La méthode que nous proposons pour l'identification des exons se déroule en 6 étapes dépendantes et séquentielles : de l'acquisition de la séquence d'ADN à l'obtention des résultats. Nous allons détailler chaque étape dans les points suivants. La figure (V.1) montre l'organigramme de l'algorithme proposé. Chaque étape constitue un sous-programme.

V.4.1. 1ère étape : acquisition et lectures des séquences d'ADN

Cette première étape consiste à télécharger la séquence d'ADN depuis sa base de données en ligne. Le téléchargement peut se faire de deux manières :

- Téléchargement d'un fichier format FASTA de la séquence d'ADN puis sauvegarde sur le stockage local de l'ordinateur avant de l'importer dans MATLAB.
- Téléchargement direct de la séquence d'ADN en utilisant la boîte à outils « Bioinformatics toolbox 4.12 ».

Ces deux façons donnent les mêmes résultats.

La séquence d'ADN importée dans MATLAB est une suite de lettre type "…AGTC… " qui sera par la suite lue. La longueur de la séquence peut être déduite à ce niveau. Cette séquence est représentée par le signal x (n) dans la figure (V.1).

V.4.2. 2ème étape : conversion numérique de la séquence d'ADN

Cette étape consiste à convertir la séquence alphabétique des ADN en nombre en utilisant des méthodes de conversion. Le tableau (II.1) dans le chapitre II présente les différentes méthodes de conversions existantes dans la littérature. C'est l'étape de la conversion de x(n) en x(k) dans la figure (V.1) avec $n \in \{A, G, T, C\}, k \in \mathbb{R}$.

V.4.3. 3ème étape : utilisation du filtre fractionnaire

La séquence d'ADN numérisée x(k) est filtrée par le filtre fractionnaire élaboré dans le chapitre (IV) (IV.5). Ce filtre utilise la propriété fractionnaire de l'ADN et permet de mettre en évidence les composants de fréquence f = 1/3 du signal pour la propriété « 3-base periodicity » des zones exoniques.

Afin de minimiser la distorsion et le déphasage du signal de sortie, on propose de filtrer deux fois : une dans le sens direct et l'autre dans le sens inverse, filtrage bidirectionnel (forward-backward) [162] – [164].

Cette 3ème étape est le centre de l'algorithme car elle permet d'extraire les composantes fréquentielles exoniques de la séquence d'ADN analysée. Dans la figure (V.1), le signal x(k) passe le filtre fractionnaire et devient y(k).

V.4.4. 4ème étape : module au carré

Dans cette étape, on effectue une opération de module au carré du signal de sortie du filtre fractionnaire afin de mieux mettre en évidence les pics qui peuvent représenter les zones exoniques et de réduire les autres zones. Dans la figure (V.1), y(k) donne $|y(k)|^2$.

V.4.5. 5ème étape : filtrage passe-bas

Le filtre passe-bas est utilisé pour démoduler en amplitude (en gardant que l'enveloppe du signal) et lisser le signal. Parmi les filtres passe-bas disponibles, le Tchebychev inverse type-II d'ordre 13 a été choisi. Les fréquences de coupures de la bande passante et de la bande d'arrêt sont : 0.5 et 0.6 (valeurs normalisées à la fréquence de Nyquist). L'atténuation maximale et minimale de la bande passante sont 80 dB et 1 dB.

Ces spécifications du filtre passe-bas sont proches de celles utilisées dans [92], [107], [165] pour l'identification des exons. A cette étape, les pics révélés sur le signal de sortie (normalisé) du filtre prédisent la présence et les positions des exons dans la séquence d'ADN analysée.

V.4.6. 6ème étape : résultat et évaluation de la méthode

Cette étape permet d'afficher les résultats graphiques et les valeurs statistiques des paramètres d'évaluation de la performance de la méthode proposée. En utilisant les paramètres d'évaluation statistique comme : la sensibilité (Sn), la spécificité (Sp), l'exactitude (E), la corrélation approximative (CA), F1-mesure, la courbe ROC et son AUC (la surface sous la courbe ROC) ainsi que la durée de la simulation (T). Le signal à la sortie de l'étape 5 est normalisé de 0 à 1 (comme amplitude maximale). Ces paramètres s'obtiennent selon un seuillage qui varie de 0 à 1 avec un pas de 0.01 pour avoir 100 valeurs différentes de chaque paramètre.

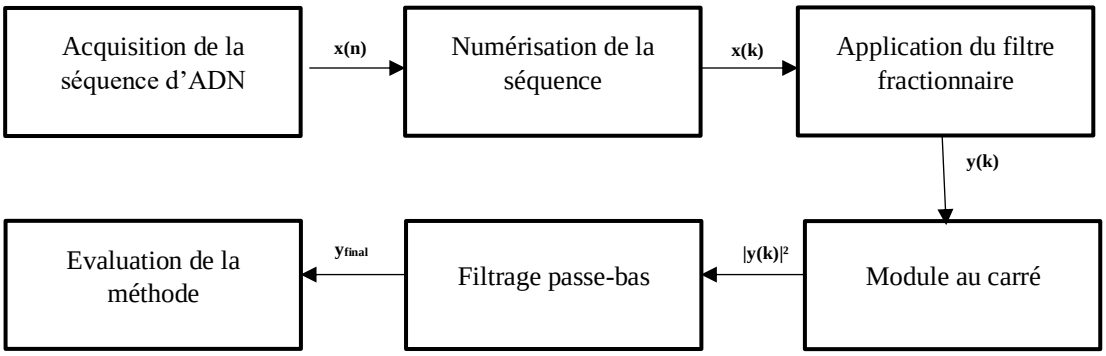


Figure V. 1 : Les étapes de l'algorithme de la méthode proposée.

V.5. Simulations et discussions des résultats

V.5.1. Choix de la valeur du paramètre α du filtre fractionnaire

Dans le chapitre (IV), $\upsilon = -0.5$, l'ordre du filtre $M = N = 3$ et $L = 1000$ ont déjà été fixés, tandis que α n'était pas encore fixé. Les figures (IV.6) - (IV.8) montrent bien l'effet de la valeur de α sur l'allure de la réponse fréquentielle du filtre fractionnaire. On observe aussi que pour $\alpha \geq 0.5$, l'allure de la réponse fréquentielle se stabilise et la largeur de sa bande passante devient plus étroite (sélectivité).

Ainsi, pour le choix de valeur de α, une série de tests a été effectué en variant α de 0.5 à 0.8 avec un pas de 0.001. Le but est de comparer les courbes ROC obtenues, plus précisément les valeurs des AUC de ces courbes pour déduire la valeur de α qui donne la

meilleure performance à l'ensemble de l'algorithme. Rappelons que l'allure du courbe ROC et la valeur de son AUC permettent d'évaluer la performance globale d'une méthode de classification donnée. Nous prenons en compte aussi la valeur de la corrélation approximative (CA) qui désigne la corrélation entre le modèle et la réalité.

Afin de bien justifier cette valeur de α, on utilise les méthodes de numérisation (conversion) de séquence d'ADN les plus connues pour leurs performances dans différents travaux sur l'identification des exons (basés sur le filtrage numérique et les transformées). Les détails de ces méthodes ont déjà été reportés dans le tableau (II.1). La séquence d'ADN de référence est le F56F11.4 du Caenorhabditis elegans, choisie aussi pour sa large utilisation en GSP. D'après les simulations, les valeurs maximales de AUC et CA obtenues sont regroupées dans le tableau (V.3) suivant :

Technique de conversion numérique	AUC_{max}	CA_{max}	α
Voss	0.9316	0.8693	0.567
Nombre complexe	0.9134	0.8806	0.613
EIIP	0.9095	0.8035	0.621
Paired numerical (numérisation par parité)	0.6907	0.6425	0.687
Nombre entier	0.8974	0.8316	0.628
2-bit binaire	0.8499	0.7117	0.606
Numéro atomique	0.8533	0.7737	0.746
PAM (nombre réel)	0.5091	0.4841	0.662
Code Walsh	0.8176	0.7224	0.627
Masse moléculaire	0.8567	0.7713	0.738

Tableau V. 3 : Valeurs des AUC_{max}, CA_{max} α à partir des différentes méthodes de numérisation d'ADN.

D'après le tableau (V.3), pour les dix techniques de numérisation évaluées, la représentation de Voss, le nombre complexe et le EIIP sont les plus performantes avec une valeur de AUC > 0.9 et CA > 0.8.

La figure (V.2) montre la courbe de $AUC = f(\alpha)$ pour les trois techniques de conversion : Voss, EIIP et nombre complexe. On peut observer que pour les valeurs $0.5 < \alpha < 0.8$, les courbes ROC de la représentation de Voss et EIIP ne décroissent pas rapidement comparées à la numérisation utilisant le nombre complexe. En d'autres termes, la performance générale de l'algorithme varie rapidement avec l'utilisation des nombres complexes par rapport à l'utilisation de la représentation de Voss et EIIP, comme méthode de conversion numérique de

la séquence d'ADN. Cependant, il est intéressant de trouver une valeur de α fixe et qui donne des bons résultats même si on change de technique de numérisation (entre Voss et EIIP). Pour cela, on peut procéder à une simple estimation d'une valeur moyenne par la relation suivante :

$$\alpha = \alpha_{Voss} + (\alpha_{EIIP} - \alpha_{Voss})/2 = \ 0.567 + (0.621 - 0.567)/2 = 0.594. \qquad (V.1)$$

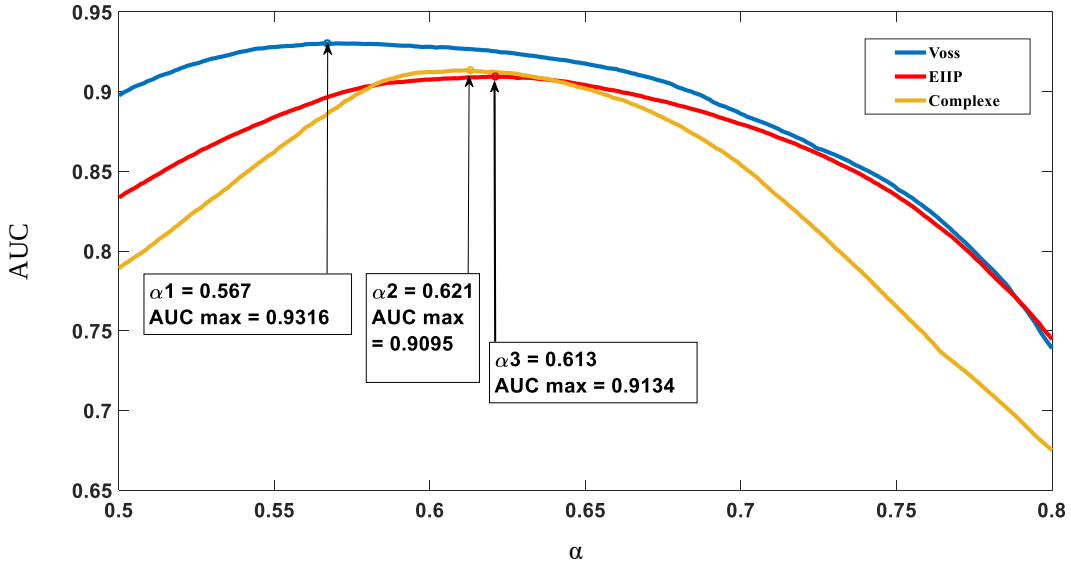


Figure V. 2 : Variation de AUC_{max} en fonction de la valeur de α allant de 0.5 à 0.8.

Pour les trois techniques de numérisation (Voss, EIIP et nombre complexe).

D'après la relation (V.1), les paramètres du filtre fractionnaire proposé pour l'identification des exons sont alors : $\alpha = 0.594,\ \upsilon = -0.5,\ M = N = 3\ et\ L = 1000$. Les coefficients « a » et « b » de la fonction de transfert du filtre approximatif dans la relation (IV.24) en utilisant la méthode de Prony (fonction prony.m dans Matlab) sont : $a_0 = 1, a_1 = 1.2034, a_2 = \ 1.1846, a_3 = 0.2081$ et $b_0 = 1, b_1 = 0.6094\ ,\ b_2 = 0.3492,\ b_3 = -0.1031$ [166]. Les pôles de ce filtre RII sont à l'intérieur du cercle unité, ce qui démontre sa stabilité. La figure (V.3) montre la comparaison entre la réponse fréquentielle du filtre fractionnaire idéal et celui du filtre approximatif en conservant les mêmes caractéristiques tels que : faible ordre, un fort gain en amplitude de la réponse fréquentielle, bande passante étroite, pas d'oscillation dans les bandes d'arrêt.

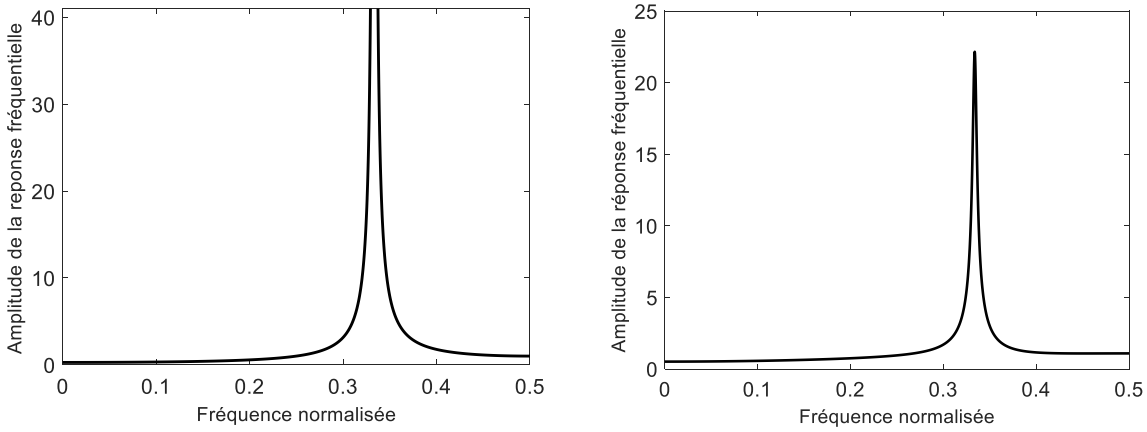


Figure V. 3 : Comparaison entre la réponse fréquentielle du filtre fractionnaire idéal (à gauche) et filtre approximatif (à droite). Avec les paramètres α = 0.594, υ = -0.5, M = N = 3 et L = 1000.

V.5.1.1. Choix de la technique de numérisation d'ADN

En utilisant le filtre fractionnaire avec les paramètres obtenus de point précédent : $\alpha = 0.594,\ \upsilon = -0.5, M = N = 3\ et\ L = 1000$, on a refait les simulations sur les 10 techniques de conversion numérique de l'ADN du tableau (V.3) afin de choisir celles qui seront utilisées par la suite. La séquence du gène F56F11.4 du Caenorhabditis elegans est analysée, et l'évaluation du courbe ROC ainsi que son AUC est montrée dans la figure (V.4).

On observe que parmi les techniques de conversion numérique de séquence d'ADN utilisées dans la figure (V.4a), celle de Voss a comme $AUC = 0.93$, suivie de EIIP et du nombre complexe qui ont comme $AUC = 0.91$. Ce sont les meilleurs résultats de classification. La figure (V.4b) montre les résultats pour les valeurs de AUC entre 0.80 et 0.90 qui correspondent aux techniques de numérisation utilisant le nombre entier, 2-bit binaire et codage Walsh. Ces représentations donnent de bons résultats. Dans la figure (V.4c), on observe des faibles valeurs de AUC avec les techniques de numérisation en utilisant le numéro atomique ($AUC = 0.69$), PAM ($AUC = 0.61$) et nombre réel ($AUC = 0.36$).

De ces résultats, on revient à la déduction faite lors du choix du paramètre α que les représentations de Voss et EIIP donnent de bonnes performances avec le filtre fractionnaire pour l'identification des exons. Ces deux techniques sont aussi les plus utilisées dans la littérature sur les méthodes GSP pour l'identification des exons chez les eucaryotes (basées sur le filtrage numérique et les transformations…) [91], [107], [111], [112], [121], [165] -[167].

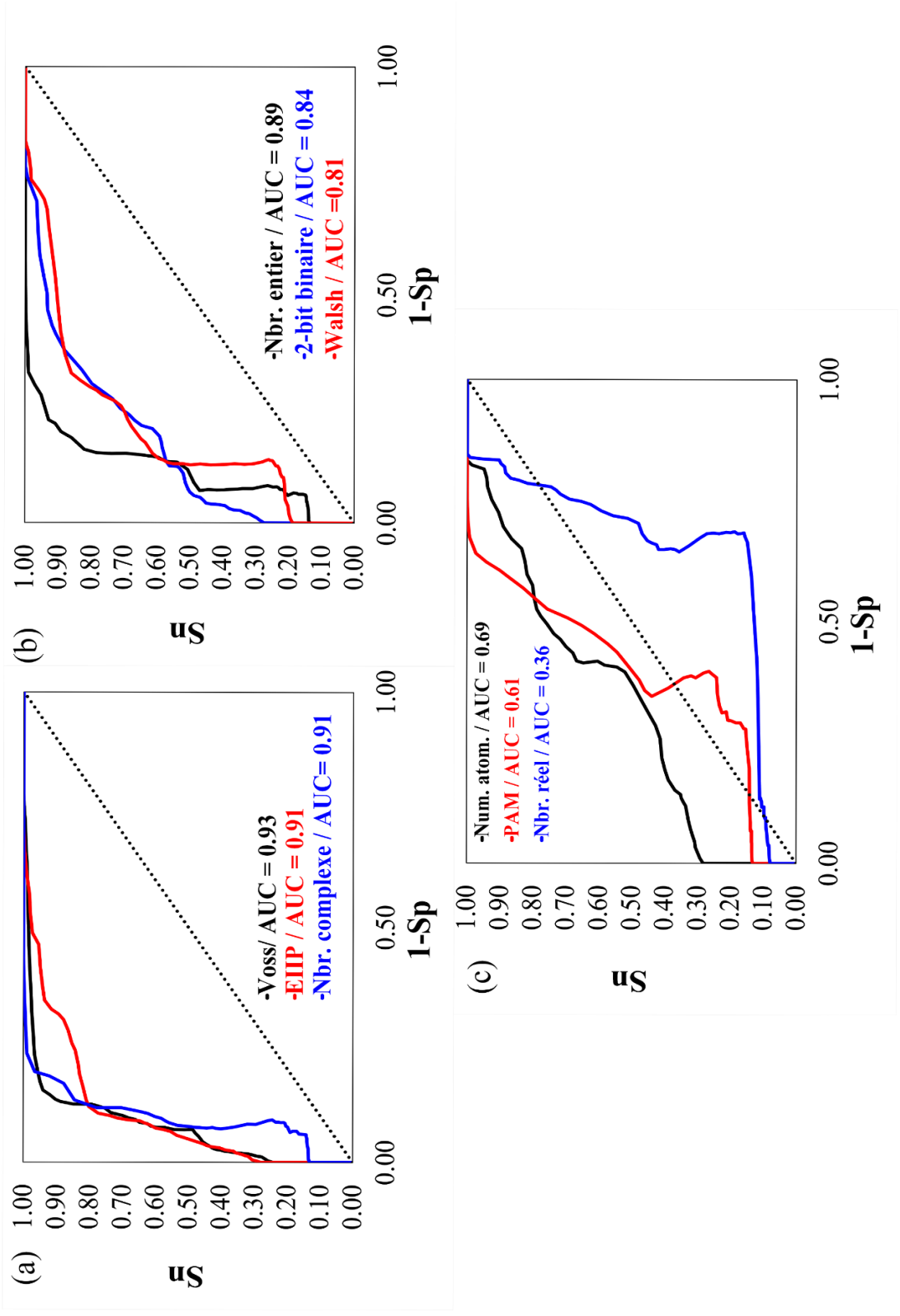


Figure V. 4 : Comparaison des valeurs des AUC et de l'allure des courbes ROC de la méthode proposée pour différentes techniques de numérisation de l'ADN. Sn : sensibilité, Sp : spécificité [166].

V.5.2. Evaluation de la méthode proposée sur le gène F56F11.4 du C.elegans

Dans cette partie, nous allons présenter trois études comparatives en regard de la méthode proposée :

- deux méthodes GSP basées sur les filtres numériques que nous avons implémentés.
- des méthodes GSP basées sur les filtres numériques, dans la littérature.
- des méthodes GSP basées sur les transformées, dans la littérature.

V.5.2.1. Etude comparative avec des méthodes GSP basées sur des filtres numériques

La présente étude compare la méthode proposée avec d'autres méthodes que nous avons simulées. Les deux méthodes GSP parmi les plus utilisées sont : l'utilisation du filtre RII de type « anti-Notch » et le filtre Tchebychev inverse de type II (passe-bande et passe-bas) [92], [111], [119], [165], [168]. Ce sont des méthodes GSP classées comme "références" car leurs algorithmes servent de base pour d'autres méthodes dans le domaine de l'identification des exons dans la séquence d'ADN chez les eucaryotes. Dans la littérature, ces deux méthodes utilisent souvent la représentation de Voss et EIIP comme techniques de conversion numérique des séquences d'ADN. Les organigrammes de leurs algorithmes sont présentés dans la figure (V.5) suivante :

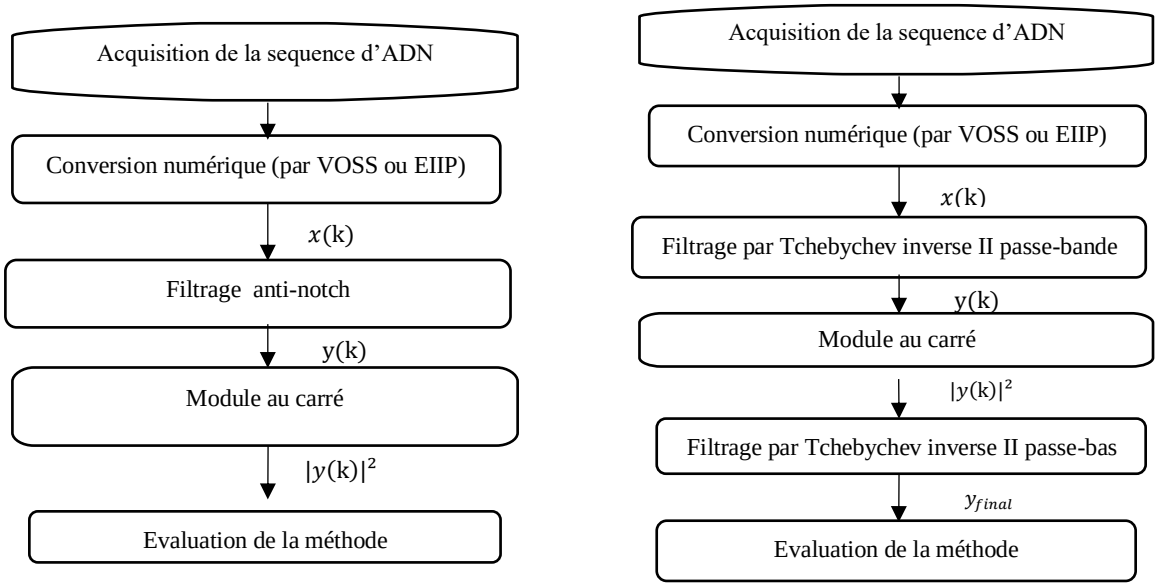


Figure V. 5 : Les étapes des algorithmes d'identification des exons par les méthodes de GSP basées sur le filtrage numérique : anti-Notch (à gauche) et Tchebychev (à droite).

V.5.2.1.a. Evaluation graphique

Les figures (V.6) et (V.7) montrent les signaux de sortie des algorithmes des méthodes GSP comparées. Les méthodes a, b et c sont respectivement les méthodes basées sur le filtre anti-Notch, le filtre Tchebychev inverse II et la méthode proposée (filtre fractionnaire). Les lignes brisées verticales limitent les positions exactes des exons dans la séquence F56F11.4, selon la base de données NCBI.

La première observation est la présence des pics évidents sur les signaux, pour les 3 méthodes (a, b et c). Ces pics présentent les composantes fréquentielles f = 1/3 correspondant à la propriété « 3-base periodicity » pour l'identification des exons. Cependant, il semble y avoir des différences entre les positions de ces zones. En effet, avec la représentation EIIP sur la figure (V.6), la méthode basée sur le filtre anti-Notch (a) et la méthode proposée (c) mettent en évidence des pics dans les zones des exons délimités par NCBI. Quant à la méthode basée sur le filtre Tchebychev inverse II (b), le premier et le troisième pic ne correspondent pas vraiment à la zone des exons selon NCBI.

Avec la numérisation de Voss, dans la figure (V.7), la méthode proposée basée sur le filtre fractionnaire (c), identifie bien les zones des exons par des pics bien visibles dans les régions délimitées par les lignes brisées. Quant aux deux autres méthodes, le filtrage anti-Notch (a), détecte les 5 exons mais le premier pic semble être de très faible amplitude qui peut induire surement des faux positifs (FP) dans les zones introns en cas de valeur de seuil bas. La méthode (b), comme avec la représentation EIIP, présente 2 pics évidents (le 1^{er} et le $3^{ème}$) qui dépassent la zone exacte des exons données par NCBI.

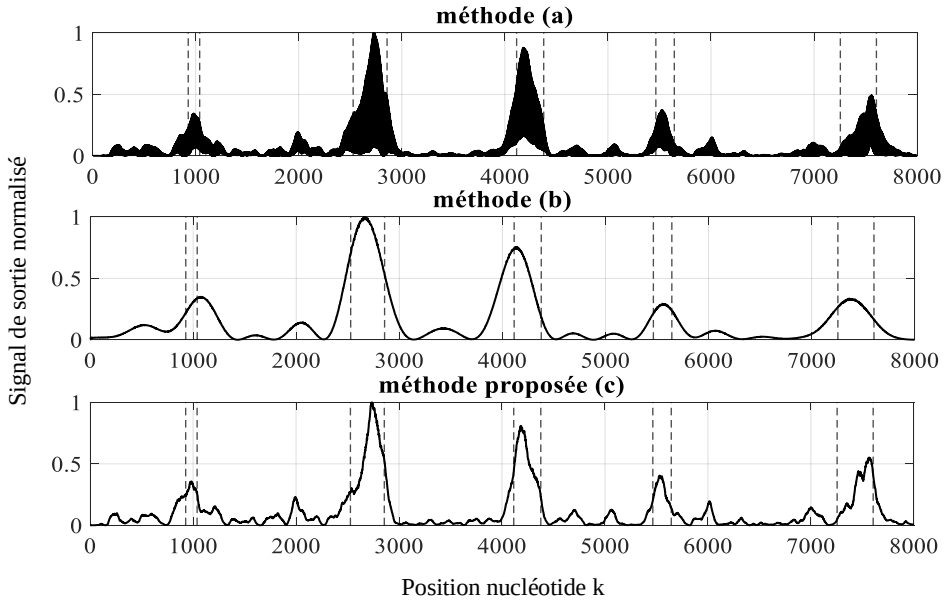


Figure V. 6 : Signal en sortie des méthodes GSP avec la technique de numérisation EIIP. (a) avec le filtre anti-Notch, (b) avec le filtre Tchebychev et la méthode proposée est en (c). Les lignes brisées verticales délimitent les positions exactes des exons selon la base de données du NCBI [166].

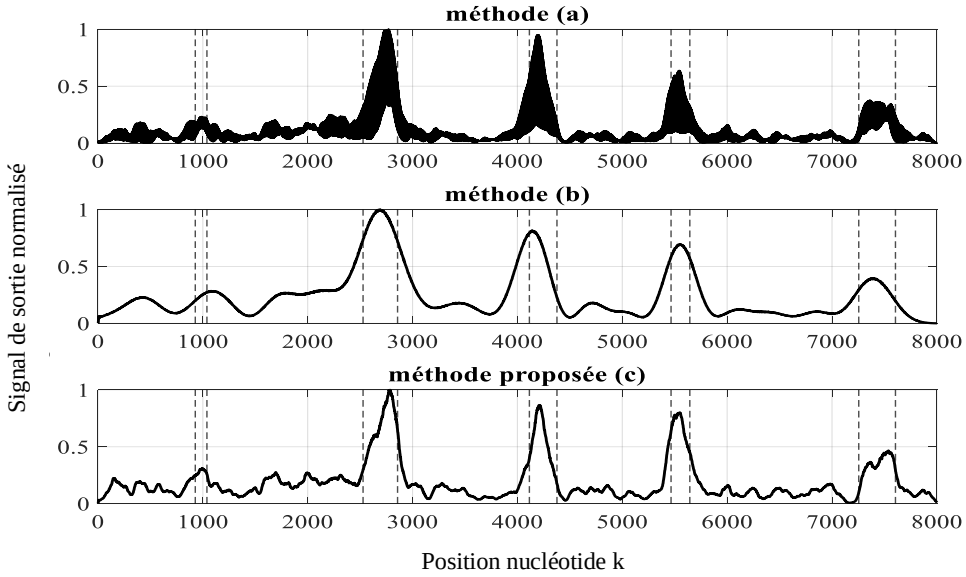


Figure V. 7 : Signal en sortie des méthodes GSP avec la technique de numérisation Voss. (a) avec le filtre anti-Notch, (b) avec le filtre Tchebychev et la méthode proposée est (c). Les lignes brisées verticales délimitent les positions exactes des exons selon la base de données du NCBI [166].

Les figures (V.8b) et (V.8d), montrent les courbes de l'exactitude (E) en fonction des valeurs de seuil (S), pour les trois méthodes GSP susmentionnées. La plage de valeur de seuil choisie pour toutes les expériences étant le même : de 0 à 1 avec un pas de 0.01 pour avoir 100 valeurs (0 % à 100 % du signal).

Premièrement, sur la figure (V.8b) (avec EIIP) et la figure (V.8d) (avec Voss), nous observons que la méthode proposée a la plus grande valeur d'exactitude (E) comparée aux deux autres méthodes. Et cette valeur reste élevée pour un seuil (S) compris dans l'intervalle [0.25 - 0.30] avant de décroître lentement. Pour la méthode basée sur le filtrage anti-Notch, l'intervalle de seuil (S) pour atteindre l'exactitude maximale est : [0.20 - 0.30] tandis que pour la méthode avec le filtre Tchebychev, l'intervalle de seuil est [0.20 - 0.45]. Il est préférable de choisir une méthode dont la valeur de seuil optimal ne fluctue pas sur un grand intervalle pour atteindre une meilleure exactitude comme le cas de la méthode proposée.

Les figures (V.8a) et (V.8c) comparent les courbes du ROC ainsi que les AUC des trois méthodes testées. On observe clairement que dans les deux cas de représentations (EIIP et Voss), la méthode proposée a la plus grande AUC (> 0.9) et tous les points de ses courbes ROC sont au-dessus de la ligne diagonale (Sn = 1-Sp). Vient ensuite la méthode avec le filtre anti-Notch, dont l'allure des courbes ROC est au-dessus de la ligne diagonale mais avec des valeur AUC (< 0.9). En dernier du classement se trouve la méthode basée sur le filtrage Tchebychev inverse II, dont les points de courbes ROC croisent ou sont au-dessous de la ligne diagonale. Cette dernière méthode a par conséquent une mauvaise performance comparée à la méthode proposée basée sur le filtre fractionnaire.

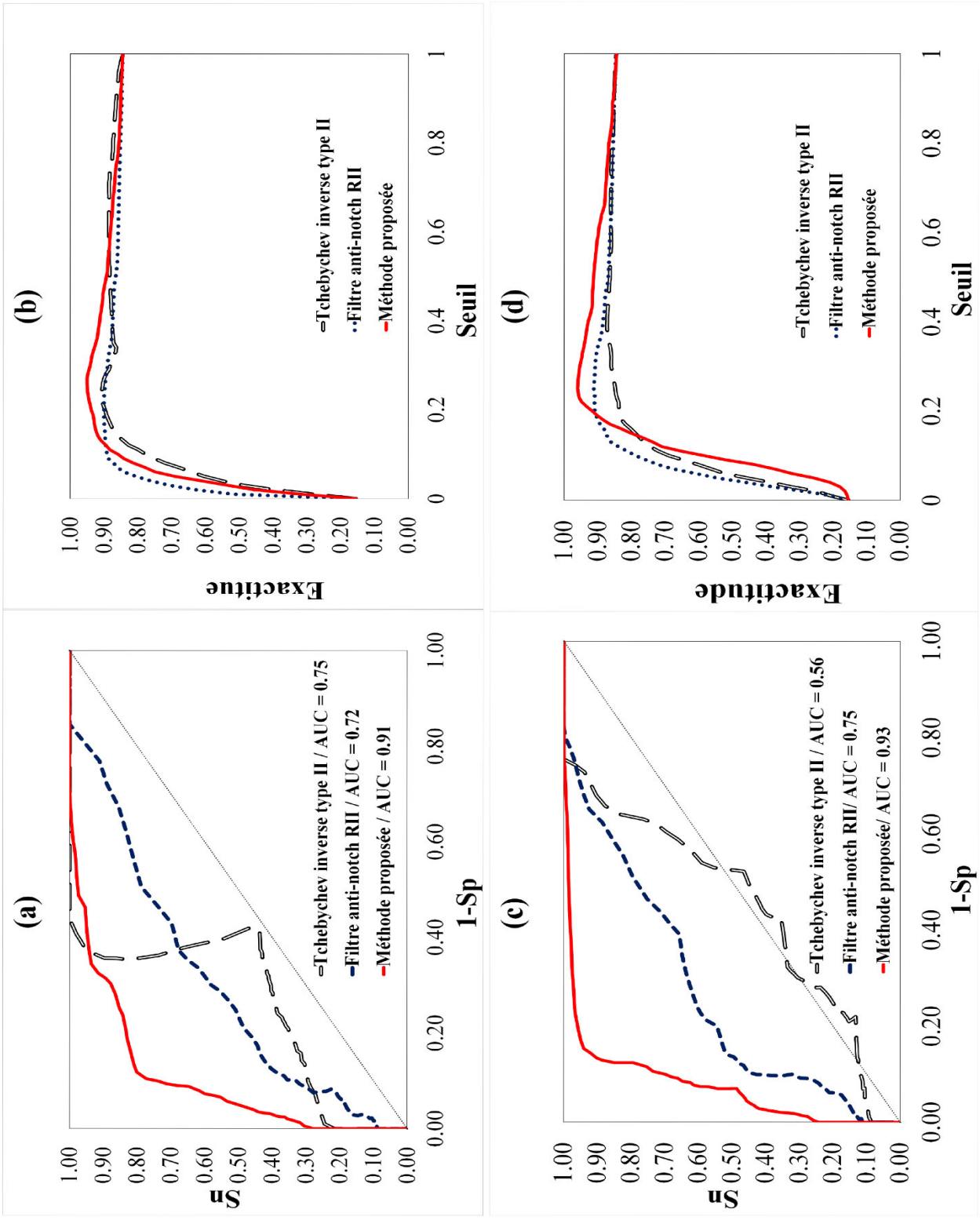


Figure V. 8 : Les courbes d'exactitude (E) en fonction du seuil (S) et de sensibilité en fonction de 1-spécificité (courbe ROC) pour les 3 méthodes GSP comparées. Sn : sensibilité, Sp : spécificité [166].

V.5.2.1.b. Evaluation des paramètres statistiques

Les évaluations graphiques précédentes sont vérifiées par les valeurs des paramètres statistiques en fonction d'un seuillage (S). Ces paramètres sont : AUC, corrélation approximative (CA) et exactitude (E). Sachant que ces paramètres sont calculés à partir de la sensibilité (Sn) et la spécificité (Sp). Ces derniers sont calculés à partir des vrais positifs (VP), vrais négatifs (VN), faux positifs (FP) et faux négatifs (FN). Les définitions de ces paramètres ont été présentées dans le 2ème chapitre (II.5). Les valeurs sont regroupées dans le tableau (V.4).

Séquence du gène F56F11.4 (C. elegans)							
Méthode GSP	**Conversion numérique**	**Sn**	**Sp**	**CA**	**E**	**S (0 à 1)**	**AUC**
Filtre de Tchebychev inverse de type II	EIIP	0.97	0.61	0.73	0.90	0.21	0.75
	VOSS	0.88	0.34	0.44	0.72	0.41	0.56
Filtre anti-Notch RII	EIIP	0.43	0.84	0.59	0.90	0.22	0.72
	VOSS	0.52	0.85	0.64	0.91	0.26	0.75
Méthode proposée	**EIIP**	**0.79**	**0.89**	**0.81**	**0.95**	**0.27**	**0.91**
	VOSS	**0.95**	**0.83**	**0.86**	**0.96**	**0.25**	**0.93**

Tableau V. 4 : Valeurs des paramètres d'évaluation pour les trois méthodes GSP comparées. Sn : sensibilité, Sp : spécificité, CA : corrélation approximative, E : exactitude, S : seuil, AUC : surface sous la courbe ROC [166].

Les valeurs de AUC dans le tableau (V.4) montrent que la méthode proposée est performante avec des $AUC > 0.9$. En effet, avec la représentation de Voss, $AUC = 0.93$ est la valeur la plus élevée, suivie de $AUC = 0.91$ pour EIIP. Les autres méthodes ont toutes des $AUC < 0.9$ qui les classent comme bonnes mais moins performantes que la méthode proposée. La valeur la plus faible est $AUC = 0.56$ et obtenue pour la méthode qui utilise le filtre Tchebychev inverse de type II avec la représentation de Voss. Ces résultats montrent une bonne classification faite avec la méthode proposée en fonction de la sensibilité et la spécificité et de la courbe de ROC.

En termes d'exactitude (E), on observe que la méthode proposée a aussi la plus grande valeur d'exactitude moyenne (E), $E_{Voss} = 0.96$, suivi de $E_{EIIP} = 0.95$. La méthode avec le filtre RII anti-Notch a $E_{Voss} = 0.91$ et $E_{EIIP} = 0.90$, tandis que la méthode avec Tchebychev inverse de type II montre le E le moins élevé : $E_{Voss} = 0.72$. Ces résultats signifient que la méthode proposée est plus proche de la valeur de référence (annotation NCBI), par rapport aux autres méthodes.

En termes de corrélation approximative (CA), les valeurs sont tous > 0.8, pour la méthode proposée avec Voss et EIIP. La valeur de $CA_{Voss} = 0.86$ (pour la méthode proposée) est la plus élevée comparée aux autres méthodes testées (dont les CA sont tous < 0.80). Cela montre une meilleure corrélation entre les prédictions du modèle proposée et les observations réelles (référence NCBI).

On observe aussi que pour la méthode proposée, les valeurs du seuil (S) pour obtenir les valeurs des paramètres susmentionnés sont proches l'une de l'autre pour Voss et EIIP, 0.25 et 0.27, respectivement.

A partir des évaluations graphiques et statistiques, on déduit que la méthode proposée basée sur le filtre fractionnaire est plus performante comparée aux deux autres méthodes avec le filtre anti-Notch et le filtre Tchebychev inverse de type II que nous avons simulées.

On observe aussi que l'utilisation de la technique de numérisation de Voss donne des meilleurs résultats comparés à EIIP. Cette observation est utilisée dans le tableau (V.5) pour comparer les positions exactes des paires de bases délimitant les exons de la séquence de gène F56F11.4 des 3 méthodes GSP. On observe alors, à partir du tableau (V.5) que la méthode proposée identifie les 5 exons et leurs positions sont proches des valeurs référentielles dans NCBI, par rapport aux deux autres méthodes. En effet, graphiquement selon la figure (V.7), on a observé des pics dans les 5 régions exoniques mais analytiquement avec les valeurs de seuillage (S) optimal : la méthode avec le filtre anti-Notch ne délimite pas le 1er exon, tandis que la méthode avec le filtre Tchebychev inverse II n'identifie pas le 1er et le 5ème exon, selon la référence NCBI.

Séquence du gène F56F11.4 (C. elegans)			
Position des exons selon (NCBI)	**Filtre anti-Notch**	**Filtre de Tchebychev inverse II**	**Méthode proposée**
	Voss (S=0.26)	**Voss (S=0.41)**	**Voss (S=0.25)**
928-1039 (111 pb)	Non-délimité	Non-délimité	**938-1039(104 pb)**
2528-2857 (329 pb)	2504-2897 (393 pb)	2465-2939 (474 pb)	**2517-2918(401 pb)**
4114-4377 (263 pb)	4024-4357 (333 pb)	3979-4312 (333 pb)	**4058-4354 (296 pb)**
5465-5644 (179 pb)	5410-5668 (258 pb)	5419-5650(231 pb)	**5413-5665 (252 pb)**
7255-7605 (350 pb)	7285-7598 (313 pb)	Non-délimité	**7298-7609 (311 pb)**

Tableau V. 5 : Position des exons dans la séquence F56F11.4, pb : paire de base, S : seuil (0 à1).

V.5.2.2. Etude comparative avec des méthodes GSP basées sur le filtrage numérique, dans la littérature

Cette étude compare la performance de la méthode proposée avec des méthodes GSP basées sur le filtrage numérique, trouvées dans la littérature. Nous avons pris en considération, les travaux anciens et récents. Parmi les méthodes basées sur le filtrage numérique que nous avons recueillies :

- La méthode utilisant le filtre Savitzky-Golay (Filtre S-G) comme lissage et suppresseur de bruit [93].
- La méthode utilisant un filtrage bidirectionnel [164]. Cette comparaison est intéressante. car dans la 3ème étape de l'organigramme dans la figure (V.1), le filtre fractionnaire est utilisé en mode bidirectionnel.
- Les différentes configurations du filtre "anti-Notch" comme : ANF, CSANF, HSANF, ANF+MA [111], [121]. Nous avons gardé les abréviations en anglais tout en donnant les définitions ci-dessous :

  ANF : filtre anti-Notch (anti-Notch filter).

  CSANF : anti-Notch suppresseur de conjuguées (conjugate suppression anti-Notch filter).

  HSANF : filtre anti-Notch suppresseur d'harmoniques (harmonic suppression anti-Notch filter).

  ANF + MA : filtre anti-Notch à moyenne mobile (moving average anti-Notch filter).

Dans la littérature, ces travaux ont tous été évalués au niveau nucléotidique sur la séquence du gène F56F11.4, avec les paramètres statistiques tels que : système de seuillage (S), la sensibilité (Sn), la spécificité (Sp), l'exactitude moyenne (Em), la valeur de AUC et la durée d'exécution du CPU. Ces travaux utilisent aussi la représentation de VOSS comme technique de conversion numérique. Le tableau (V.6) regroupe les résultats de ces paramètres d'évaluation.

Séquence du gène F56F11.4 (C. elegans)							
Méthodes	**Technique de numérisation**	**Sn**	**Sp**	***Em***	**S (0 à 1)**	**AUC > 0.90**	**Temps CPU (sec)**
Filtre S-G [93]	Voss	0.75	0.59	0.67	0.22	Non	0.049
Filtrage bidirectionnel [164]	Voss	0.76	0.52	0.64	0.13	Non	0.008
CSANF [121]	Voss	0.72	0.53	0.63	0.19	Non	0.011
HSANF [121]	Voss	0.72	0.54	0.63	0.18	Non	0.011
ANF+MA [121]	Voss	0.57	0.66	0.62	0.05	Non	0.005
ANF [111]	Voss	0.60	0.39	0.50	0.11	Non	0.004
Méthode proposée	**Voss**	**0.95**	**0.83**	**0.89**	**0.25**	**Oui**	**0.049**

Tableau V. 6 : Valeurs des paramètres d'évaluation des méthodes GSP basées sur le filtre numérique comparées à la méthode proposée. Sn : sensibilité, Sp : spécificité, Em : exactitude moyenne et S : seuil [166].

On observe à partir des résultats rassemblés dans le tableau (V.6) que les valeurs des paramètres de sensibilité (Sn), spécificité (Sp) et exactitude moyenne (Em) sont supérieures pour la méthode proposée comparée aux autres méthodes citées. En outre, seule la méthode proposée a une valeur de AUC supérieur à 0.9. En termes de durée d'exécution de l'algorithme, pour 100 tours du CPU, la méthode proposée s'exécute en 0.049 secondes, tout comme la méthode utilisant le filtre S-G [93]. Cela est sûrement dû à la dernière étape de filtrage passe-bas pour ces deux méthodes. Contrairement à ANF, CSANF, HSANF, ANF+MA et le filtrage bidirectionnel qui s'exécutent plus rapidement, sans une étape supplémentaire de lissage et filtrage passe-bas à la fin de leurs algorithmes.

Les études comparatives avec des méthodes GSP utilisant les filtres numériques (de la littérature) sur la séquence F56F11.4 montrent que la méthode proposée est plus performante pour l'identification des exons de ce gène, en termes d'évaluation graphique et statistique.

V.5.2.3. Etude comparative avec des méthodes GSP basées sur les transformées, dans la littérature

Nous avons complété l'évaluation de la performance de la méthode proposée en comparant avec des méthodes basées sur les transformées, développées dans la littérature. Ces méthodes prennent en compte les différents types de transformations du signal couramment

utilisées dans le contexte d'identification des exons chez les eucaryotes. Comme l'étude précédente, nous avons pris les travaux anciens et récents tels que :

- La méthode GSP utilisant la transformée de Fourier [101].
- La méthode GSP utilisant la transformée de Stockwell [168].
- La méthode GSP utilisant la transformée en ondelette [95], [103].
- Une méthode GSP plus récente basée sur le traitement adaptatif des signaux et l'analyse spectrale : le « Sinusoidal-assisted variation mode decomposition » (SAVMD) [50].
- Le « noise-assisted multivariate empirical mode decomposition + modified Gabor-wavelet transform » (NA-MEMD-MGWT) [95] qui est une récente version des méthodes GSP basées sur la transformée en ondelette [103].

Selon les auteurs, les deux dernières méthodes susmentionnées sont plus performantes comparées aux méthodes basées sur les transformations de Fourier et en ondelettes tels que le SCM [112] et le MGWT [103].

Dans la littérature, toutes ces méthodes ont été évaluées avec la même séquence du gène F56F11.4 (C. elegans) en utilisant la technique de numérisation de Voss et EIIP (sauf pour le cas du NA-MEMD-MGWT). Les paramètres d'évaluation utilisés sont identiques à la précédente étude (tableau V.6) et les valeurs sont rapportées dans le tableau V.7.

Séquence du gène F56F11.4 (C. elegans)							
Méthodes	**Technique de numérisation**	**Sn**	**Sp**	***Em***	**S (0 à 1)**	**AUC > 0.90**	**Temps CPU (sec)**
Transformée de Fourier [101]	Voss	0.82	0.86	0.85	0.82	Non	0.15
Transformée en ondelette [103]	Voss	0.80	0.82	0.81	0.61	Non	0.35
Transformée de Stokwell [168]	EIIP	0.86	0.87	0.86	0.85	Non	0.28
Méthode proposée	**EIIP**	**0.79**	**0.89**	**0.84**	**0.27**	**Oui**	**0.05**
Méthode proposée	**Voss**	**0.95**	**0.83**	**0.89**	**0.25**	**Oui**	**0.05**

Tableau V. 7 : Valeurs des paramètres d'évaluation des méthodes GSP basées sur les transformées comparées à la méthode proposée. Sn : sensibilité, Sp : spécificité, Em : exactitude moyenne et S : seuil [166].

On observe que toutes les valeurs de Sn, Sp et Em des méthodes GSP dans le tableau (V.7) sont supérieures ou égales à 0.8 sauf pour la méthode proposée avec la représentation EIIP qui a une Sp =0.79. Seule la méthode proposée (avec Voss) a la plus grande valeur de Sn = 0.95. La méthode proposée avec la numérisation de Voss et EIIP ont aussi des valeurs de

AUC supérieures à 0.90. Le temps que met le CPU est plus court pour la méthode proposée comparée aux autres. La valeur de Em de la méthode proposée (avec numérisation de Voss) est la plus élevée. On peut remarquer aussi qu'il faut une valeur de seuil (S) élevée, $S > 0.5$ (50% du signal) à toutes les méthodes utilisant les transformées, contrairement à la méthode proposée qui a une valeur de seuil $S \sim 0.25$.

On peut déduire de ces résultats que la méthode proposée, surtout en utilisant la numérisation de Voss, est plus performante pour la détection des exons du gène F56F11.4.

La comparaison avec la méthode SAVMD [50] est faite en évaluant les valeurs de AUC avec différentes techniques de numérisation de l'ADN. On observe à partir du tableau (V.8) que les valeurs de AUC de la méthode proposée (pour la numérisation de Voss, EIIP et le nombre entier) sont supérieures par rapport à la méthode SAVMD. Cependant, la méthode SAVMD a une valeur de AUC légèrement supérieure avec l'utilisation de la technique de numérisation en nombre complexe.

La comparaison avec la méthode « NA-MEMD-MGWT » se fait selon [95], en utilisant la valeur de la corrélation approximative (CA) (les prédictions du modèle comparées aux observations réelles). La méthode proposée a une valeur de $CA_{Voss} = 0.86$ tandis que la « NA-MEMD-MGWT » a une valeur de $CA_{dinucléotide} = 0.83$. Bien que ces valeurs soient proches, la méthode proposée est légèrement supérieure en utilisant une technique de numérisation (Voss) plus simple que la technique utilisant les dinucléotides par la NA-MEMD-MGWT (tableau II.1).

Séquence du gène F56F11.4 (C. elegans)				
Méthodes GSP	**Technique de numérisation**			
	Nombre entier	**Nombre complexe**	**Voss**	**EIIP**
SAVMD[50]	0.88	0.93	0.90	0.86
Méthode proposée	**0.89**	**0.91**	**0.93**	**0.91**

Tableau V. 8 : Comparaison des valeurs des AUC en fonction de la technique de numérisation des ADN pour la méthode proposée et le SAVMD [50], [166].

V.5.3. Evaluation de la méthode proposée sur la base de données HMR195

Bien que les études sur le gène F56F11.4 soient importantes et que les résultats obtenus avec la méthode proposée sont pertinents, il est intéressant d'élargir les tests.

Cette partie présente alors l'évaluation de la méthode proposée sur la base de données HMR195 que nous avons présentée dans le point (V.2.1) et le tableau (V.1). Cette base de données regroupe 195 gènes de trois espèces différentes ayant des séquences de différentes longueurs avec le nombre d'exons allant de 1 à 26 par séquence.

Trois études seront présentées par la suite :

- Impact et optimisation de la valeur du seuil.
- Impact du nombre d'exons dans la séquence.
- Etude comparative avec d'autres méthodes GSP qui ont été évaluées sur la base de données HMR195.

V.5.3.1. Impact et optimisation de la valeur de seuillage

Cette étude présente l'impact du choix de la valeur de seuillage sur la performance de la méthode proposée. En effet, la variation du seuil peut améliorer ou compromettre les résultats du test. Cela est observé en prenant cas par cas des séquences de différents nombres d'exons de la base de données HMR195. L'évaluation graphique des signaux en sortie de la méthode proposée est rapportée dans la figure (V.9).

Si on prend le cas de la séquence D38752 ayant 1 exon, représentée dans la figure (V.9a), on y observe un pic très prononcé délimité par les lignes brisées rouges (référence de la position exactes des exons selon HMR195). Ce résultat est obtenu en utilisant un seuil de 0.2 (20% de l'amplitude maximale du signal de sortie). En prenant une autre séquence, AF042782 ayant 2 exons, représentée dans la figure (V.9b), on voit aussi 2 pics distincts délimités par les lignes brisées rouges (référence de la position exactes des exons selon HMR195). Ce résultat est obtenu avec un seuil de 0.18. En observant la figure (V.9c), pour la séquence AF049259 contenant 7 exons : avec S = 0.25, 6 sur les 7 pics sont considérés comme vrais positifs, tandis qu'en ajustant le seuil à 0.21 (ligne verticale rouge sur la figure (V.9d), les 7 pics représentants les 7 exons sont pris en compte. La figure (V.10) illustre bien cet impact du choix de la valeur du seuil sur la performance globale de la méthode proposée. En effet, on observe qu'afin de maintenir un taux élevé de l'exactitude de la méthode proposée, la valeur de seuil s'adapte et varie suivant la séquence à analyser.

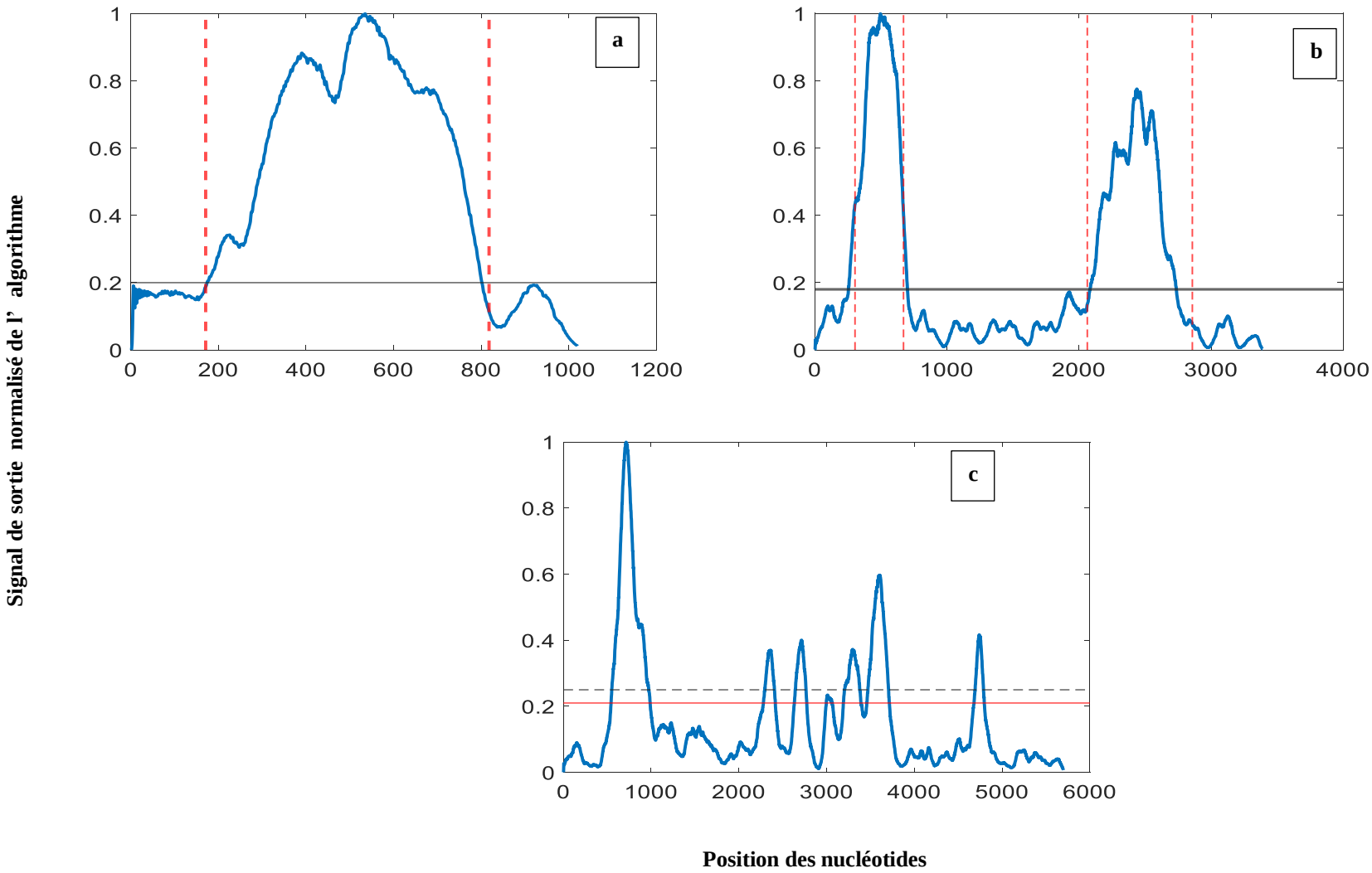


Figure V. 9 : Signaux de sortie de l'algorithme de la méthode proposée pour différentes séquences du HMR195.

a) avec le gène D38752, b) avec le gène AF042782 et c) avec le gène AF049259.

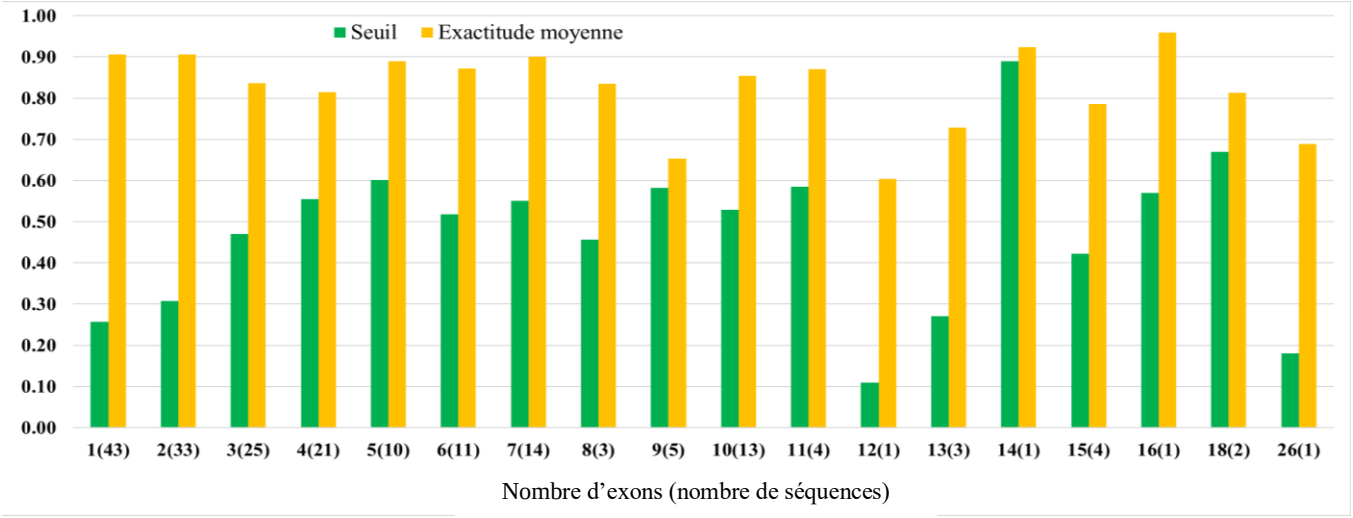


Figure V. 10 : Valeur moyenne de l'exactitude de la méthode proposée suivant la variation du seuillage, pour le cas de la base de données HMR195.

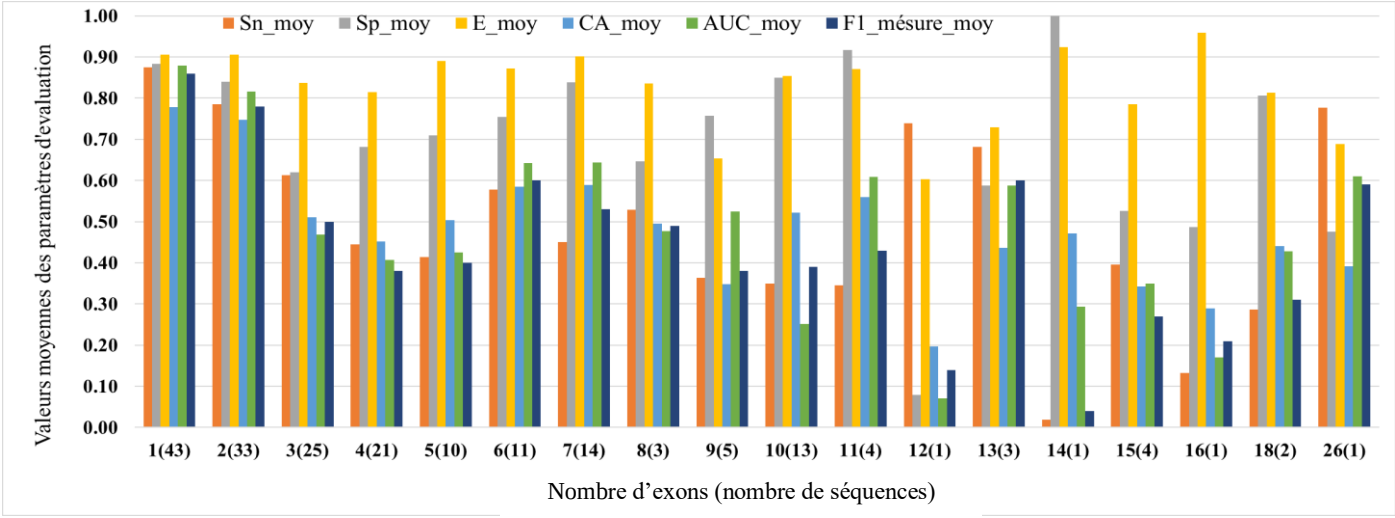


Figure V. 11 : Impact du nombre d'exons sur la performance de la méthode proposée, cas de la base de données HMR195. Sn_moy : sensibilité moyenne, Sp_moy : spécificité moyenne, E_moy : exactitude moyenne, CA_moy : corrélation approximative moyenne, AUC_moy : moyenne des aires sous les ROC.

V.5.3.2. Impact du nombre d'exons sur la performance de la méthode

Cette partie présente une étude sur l'évaluation de la performance de la méthode proposée en fonction du nombre d'exons contenus dans une séquence donnée. En effet, la base de données HMR195 offre cette possibilité, car elle contient des gènes de 1 à 26 exons (par séquence). La figure (V.11) montre que pour les séquences contenant de faibles quantités d'exons, les valeurs des paramètres statistiques sont meilleures, tandis que plus le nombre d'exons augmente, plus ces valeurs diminuent. C'est surtout le cas pour la sensibilité (Sn_{moy}), la corrélation approximative moyenne (CA_{moy}) ainsi que la F1-mesure moyenne.

Base de données HMR195						
Nombre d'exons	**Nombre de séquences**	**Longueur moyenne de séquence**	**E_{moy}**	**CA_{moy}**	**F1-mesure$_{moy}$**	**AUC_{moy}**
1	43	961 pb à 38418 pb	0.91	0.78	0.86	0.88
2	33	1377 pb à 12414 pb	0.91	0.75	0.78	0.82
3	25	923 pb à 17004 pb	0.84	0.51	0.5	0.47
4	21	795 pb à 14357 pb	0.82	0.45	0.38	0.41
5	10	3653 pb à 19730 pb	0.89	0.50	0.4	0.43
6	11	2396 pb à 56500 pb	0.87	0.59	0.6	0.64
7	14		0.90	0.59	0.53	0.64
8	3		0.84	0.49	0.49	0.48
9	5		0.65	0.35	0.38	0.52
10	13		0.85	0.52	0.39	0.25
11	4		0.87	0.56	0.43	0.61
12	1		0.60	0.20	0.14	0.07
13	3		0.73	0.44	0.60	0.59
14	1		0.92	0.47	0.04	0.29
15	4		0.79	0.34	0.27	0.35
16	1		0.96	0.29	0.21	0.17
18	2		0.81	0.44	0.31	0.43
26	1		0.69	0.39	0.59	0.61

Tableau V. 9 : Valeurs des paramètres statistiques pour la base de données HMR195. Sn_{moy} : sensibilité moyenne, Sp_{moy}: spécificité moyenne, CA_{moy} : corrélation approximative moyenne, E_{moy} : exactitude moyenne et AUC : surface sous la courbe ROC.

L'interprétation globale des résultats du tableau (V.9) permet d'observer que les valeurs des paramètres d'évaluation décroissent progressivement quand le nombre d'exons augmente.

Pour les 76 séquences contenant 1 à 2 exons dont la longueur varie entre 961 pb à 38418 pb : les valeurs moyennes de AUC > 0.8, E > 0.9 et CA, F1-mesure sont tous supérieures à 0.7. Pour les séquences avec 3 à 5 exons, les valeurs moyennes des paramètres E > 0.8, CA,

AUC et F1-mesure varient entre 0.45 et 0.5. On peut dire que la performance de la méthode commence à diminuer.

Les séquences ayant des nombres d'exons entre 6 à 26 présentent des valeurs moyennes des paramètres E = 0.81 tandis que CA et F1-mesure sont inférieures à 0.5.

Dans HMR195, 4 séquences uniques ont 12,14, 16 et 26 exons. Ce sont respectivement les séquences : AB002059 (28984 pb), AF039401 (35278 pb), AB013139 (56500 pb) et AF024570 (11494 pb). La méthode proposée est moins performante avec ces séquences. En effet, ces séquences sont parmi les plus longues de HMR195, en plus du grand nombre d'exons. On déduit alors qu'en plus de l'impact de la valeur du seuil et du nombre d'exons dans la séquence, la performance de la méthode proposée dépend aussi de longueur du gène (plutôt que de la longueur de l'exon).

V.5.3.3. Etude comparative sur le HMR195 avec des méthodes GSP dans la littérature

Cette dernière étude consiste à comparer les résultats de la méthode proposée avec ceux de la littérature travaillant sur le HMR195.

Premièrement, les valeurs maximales du paramètre CA sont disponibles pour les méthodes SCM [101] utilisant la transformation de Fourier, la MGWT [103] et NA-MEMD-MGWT [95] utilisant la transformation en ondelette ainsi que la récente méthode SAVMD [50]. En effet, les valeurs maximales de la corrélation approximative (CA) atteintes sur la base de données HMR195 sont respectivement 0.52, 0.61, 0.63 et 0.63 pour SCM [101], MGWT [103], SAVMD [50] et NA-MEMD-MGWT [95], tandis que pour la méthode proposée, la CA_{max} est de 0.98 (cas du gène U44436 à 1 exon). Cela indique une meilleure corrélation globale entre les prédictions de la méthode proposée et les observations réelles.

Une deuxième comparaison est rapportée dans le tableau (V.10) pour les valeurs moyennes des paramètres statistiques : sensibilité (Sn_{moy}), spécificité (Sp_{moy}) et l'exactitude (E_{moy}) des différentes méthodes GSP utilisant les transformées et les filtrages numériques. Ces résultats sont issus de plusieurs travaux utilisant la base de données HMR195 [107], [111], [114], [168], [169]. Contrairement à la méthode proposée, toutes ces méthodes GSP du tableau (V.10) n'arrivaient qu'à tester 100 des 195 séquences de HMR195. Cela explique que la méthode proposée a des valeurs de Sn_{moy} et Sp_{moy} plus basses. Cependant, en termes d'exactitude moyenne (E_{moy}), la méthode proposée se classe en deuxième position avec

E_{moy}=0.86. Cela indique que les résultats de la méthode proposée sont proches des valeurs réelles de référence selon HMR195.

Base de données HMR195				
Méthodes GSP	**Nombre de séquences testées**	**Sn_{moy}**	**Sp_{moy}**	**E_{moy}**
Filtre anti-Notch [111]	100	0.81	0.82	0.83
Filtre SONF [114]	100	0.89	0.86	0.84
Transformée de Stockwell [168]	100	0.88	0.88	0.85
Multirésolution [169]	100	0.89	0.88	0.86
DIT-FFT [107]	100	0.80	0.90	0.88
Méthode proposée	**195**	**0.62**	**0.77**	**0.86**

Tableau V. 10 : Comparaison de la performance de la méthode proposée avec d'autres méthodes GSP, cas de la base de données HM195. Sn_{moy} : sensibilité moyenne, Sp_{moy} : spécificité moyenne, E_{moy}: moyenne de l'exactitude

La dernière étude comparative est faite avec des programmes existant pour l'annotation in silico (ab initio) des exons et les résultats sont groupés dans le tableau (V.11). Ces programmes ont été présentés dans le 1er chapitre de ce manuscrit (I.4.3.2) et détaillés dans [44], [47]. Il est à souligner que ces programmes exploitent simultanément plusieurs propriétés biologiques de l'ADN telles que : la similarité des gènes, la présence de sites spécifiques comme les régions promotrices et régulatrices et les informations sur les zones d'épissage. La méthode proposée quant à elle, se base uniquement sur les propriétés de l'ADN en GSP qui est la « 3-base periodicity ».

Il existe des séquences dépourvues d'exons dans le HMR195, selon les programmes : FGENES, Genie, Genscan, HMMgene, Morgan et MZEF. Ce n'est pas le cas avec GeneMark.hmm et la méthode proposée. Cela influence sûrement les valeurs moyennes de sensibilité (Sn_{moy}), specificité (Sp_{moy}) de la méthode proposée. En effet, cette dernière donne une valeur pour chaque séquence analysée (195 valeurs), tandis que les autres programmes ne considèrent que les valeurs des paramètres pour les séquences qu'elles classent comme pourvues d'exons. On observe alors une valeur plus faible de la moyenne de CA avec une distribution des CA (écart-type) élevée pour la méthode proposée par rapport aux autres programmes.

Bases de données HMR195				
Programmes	**Nombre de séquences testées (séquences dépourvues d'exon)**	**Sn_{moy}**	**Sp_{moy}**	**CA_{moy} (+/- écart-type)**
FGENES	190 (5)	0.86	0.88	0.84 (+/-0.19)
GeneMark.hmm	195 (0)	0.87	0.89	0.84(+/-0.18)
Genie	180 (15)	0.91	0.9	0.89(+/-0.16)
Genscan	192 (3)	0.95	0.9	0.91(+/-0.12)
HMMgene	190 (5)	0.93	0.93	0.91(+/-0.13
Morgan	127 (0)	0.75	0.74	0.7(+/-0.21)
MZEF	119 (8)	0.7	0.73	0.68(+/-0.21)
Méthode proposée	**195 (0)**	**0.62**	**0.77**	**0.61(+/-0.22)**

Tableau V. 11 : Comparaison de la performance de la méthode proposée avec des programmes d'annotation génomiques, cas de la base de données HMR195. Sn_{moy} : sensibilité moyenne, Sp_{moy} : spécificité moyenne, CA_{moy} : corrélation approximative moyenne [44], [47].

V.6. Conclusion

Ce dernier chapitre présente l'application de la nouvelle approche GSP pour d'identification des exons chez les cellules eucaryotes. Cette méthode propose un algorithme centré sur l'utilisation d'un filtre numérique fractionnaire avec un grand gain en amplitude de la réponse fréquentielle, une bande passante étroite et centrée à f = 1/3 pour la propriété « 3-base periodicity » des exons et aussi une absence d'oscillations dans la bande d'arrêt. Après l'étape de calibrage du filtre, vient celui du choix de la technique de numérisation de la séquence d'ADN. Parmi les techniques testées, celle de Voss et EIIP s'avèrent les plus performantes et cette observation confirme leurs utilisations courantes dans l'identification des exons. Plusieurs expériences ont été menées afin d'évaluer analytiquement et graphiquement au niveau nucléotidique, la performance de cette méthode. Les résultats démontrent que sur le gène F56F11.4, la méthode proposée est plus performante par rapport aux autres méthodes GSP implémentées et disponibles dans la littérature. Nous avons ensuite élargi l'évaluation en travaillant sur la base de données HMR195. Cela a permis de comprendre le comportement de l'algorithme proposé, notamment l'impact du choix de la valeur de seuillage afin d'améliorer les résultats. Une autre remarque est que la performance générale de la méthode proposée dépend surtout du nombre d'exons dans la séquence d'ADN, plutôt que de la longueur de la séquence ou de la longueur des exons. Travaillant sur la base de données HMR195, la

comparaison avec des méthodes GSP (dans la littérature) a montré que la méthode proposée présente une bonne performance. La comparaison avec les programmes (non-GSP) d'identification d'exons a révélé que la méthode proposée est légèrement moins performante. Cependant, la méthode proposée est moins complexe et ne prend en compte que les propriétés spectrales de l'ADN, tandis que les programmes (non-GSP) considèrent beaucoup plus les propriétés physico-chimiques de l'ADN.

CONCLUSION GENERALE

Cette thèse est une contribution dans l'analyse des séquences d'ADN par des techniques de traitement numérique du signal (TNS) pour l'identification des exons chez les eucaryotes. Les travaux qui traitent ce sujet sont nombreux et se basent sur les transformations du signal et le filtrage numérique. Le second et le troisième chapitre présentent l'état de l'art de ces méthodes GSP.

L'identification des exons par les TNS suit plusieurs étapes depuis l'acquisition de la séquence au classement des résultats par les paramètres d'évaluation statistique. La numérisation de la séquence est une étape préliminaire importante et différentes techniques sont disponibles selon leur dépendance ou non-dépendance à des propriétés de la séquence d'ADN.

Les méthodes GSP existantes se basent surtout sur la propriété « 3-base periodicity » pour effectuer l'identification des exons. Dans notre travail, nous avons conçu une méthode GSP basée sur le filtre fractionnaire de type anti-Notch. Ce filtre qui est présenté dans le quatrième chapitre de ce manuscrit est conçu à base d'un modèle de type GARMA basé sur le calcul d'ordre fractionnaire.

Ce filtre fractionnaire est facilement ajustable, ne dépendant que de deux paramètres (α et υ) afin d'avoir les caractéristiques optimales pour l'identification des exons tels que : un grand gain en amplitude de la réponse fréquentielle, une fréquence centrale centrée à $f = 1/3$, une absence d'oscillations dans les bandes d'arrêt, un faible ordre (M = N = 3) et une bande passant étroite. L'approximation du filtre idéal est obtenue par la technique de modélisation de Prony. Le filtre « idéal » et le filtre « approximatif » préservent les mêmes propriétés de stabilité en ayant les pôles contenus dans le cercle unité.

L'application de cette nouvelle méthode GSP centrée sur le filtre fractionnaire est présentée dans le dernier chapitre de ce manuscrit. Sur les bases de données génomiques référentielles (NCBI et HMR195), nous avons déduit que la performance la méthode proposée est vérifiée par rapport aux méthodes GSP que nous avons implémentées et trouvées dans la littérature. Parmi les techniques de numérisation disponibles, nous en avons testée une dizaine et conclu que les numérisations de Voss et EIIP restent les plus performantes. Cette performance se présente graphiquement et analytiquement – c'est-à-dire sur le plan statistique – en utilisant les paramètres d'évaluation commune pour les études en GSP.

Pour le cas du gène F56F11.4 du C.elegans, la méthode proposée est généralement plus performante et moins complexe comparée aux méthodes GSP de la littérature. Les études sur

la totalité de la base de données HMR195 ont montré que la méthode proposée est efficace et arrive à analyser toutes les séquences, contrairement aux autres méthodes GSP et aux programmes existants.

En effet, les programmes d'annotation génomiques utilisent plusieurs propriétés biologiques spécifiques de l'ADN afin d'identifier les exons, tandis que la méthode proposée n'utilise que ses propriétés spectrales (non physico-chimiques).

En résumé, la méthode proposée se distingue par sa simplicité d'implémentation bien que son organigramme suive le standard des méthodes GSP. Ce travail de thèse ouvre alors des nouvelles perspectives dans l'utilisation des outils basés sur le calcul d'ordre fractionnaire pour l'identification des exons, notamment chez les eucaryotes. Les perspectives suivantes sont envisageables :

- La conception de nouvelles techniques plus robustes pour la numérisation de l'ADN basée sur le calcul fractionnaire.
- L'approfondissement de l'étude de l'impact du nombre d'exons dans la séquence.
- L'exploitation des propriétés biologiques de la séquence d'ADN.

LISTE DES TRAVAUX SCIENTIFIQUES

La préparation de cette thèse a permis de réaliser les 3 travaux suivants :

- **Publication internationale**

[1] M. Lehilahy et Y. Ferdi, 'Identification of exon locations in DNA sequences using a fractional digital anti-notch filter', *Biomed. Signal Process. Control*, vol. 80, p. 104362, Feb. 2023. Disponible : https://doi.org/10.1016/j.bspc.2022.104362

- **Communication internationale**

[1] M. Lehilahy et Y.Ferdi, 'An Accurate Identification Method of Exons using an Antinotch Fractional Filter', *ICMCS'20,* Abdelmalek Essâadi University, Tetouan, Morocco, October 1-2, 2020.

[2] M. Lehilahy, Y.Ferdi, et Mohamed Reda Lakehal 'A New Fractional Calculus Based filter for exon prediction', *IPPM'1*, Echahid Hamma Lakhdar University, El-Oued, Algeria, February 23-26, 2020.

BIBLIOGRAPHIE

[1] J. D. Watson et F. H. C. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”, *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953.
Disponible : https://doi.org/10.1038/171737a0.

[2] B. Lewin, “GENE VII”. Pearson; First Printing édition (16 dec. 2003). 1056 pages. ISBN-10: 0131439812, ISBN-13: 978-0131439818.

[3] C. Médigue, S. Bocs, L. Labarre, C. Mathé, et D. Vallenet, “L’annotation in silico des séquences génomiques: Bio-informatique (1)” , *médecine/sciences*, vol. 18, no. 2, pp. 237–250, Feb. 2002.
Disponible : https://doi.org/10.1051/medsci/2002182237.

[4] P. P. Vaidyanathan, "Genomics and proteomics: a signal processor's tour”, *in IEEE Circuits and Systems Magazine*, vol. 4, no. 4, pp. 6-29, Fourth Quarter 2004.
Disponible : https://doi.org/10.1109/MCAS.2004.1371584.

[5] N. Yu, Z. Li, et Z. Yu, “Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning”, *Big Data Min. Anal.*, vol. 1, pp. 191–210, Sep. 2018.
Disponible : https://doi.org/10.26599/BDMA.2018.9020018.

[6] S. A. Marhon et S. C. Kremer, “Gene Prediction Based on DNA Spectral Analysis: A Literature Review”, *J. Comput. Biol.*, vol. 18, no. 4, pp. 639–676, Apr. 2011. Disponible : https://doi.org/10.1089/cmb.2010.0184.

[7] Florence Mauger, “Développement de méthodes d’analyse de l’ADN par clivage d’une chimère ARN/ADN et par spectrométrie de masse MALDI-TOF”, *Génétique*. Université Pierre et Marie Curie - Paris VI, 2012.

[8] H. Hayes, “ADN et chromosomes”, *INRAE Prod. Anim.*, vol. 13, no. HS, pp. 13–20, Dec. 2000, Disponible : https://doi.org/10.20870/productions-animales.2000.13.HS.3806.

[9] Pray, L., “Discovery of DNA structure and function: Watson and Crick. ”, *Nature Education,* vol. 1, no. 1, p. 100, 2008.

[10] M. Brut, "Nouvelle approche méthodologique pour la prise en compte de la flexibilité dans les interactions entre molécules biologiques : les Modes Statiques”, thèse de doctorat, Micro et nanotechnologies/Microélectronique, Université Paul Sabatier - Toulouse III, 2009.

[11] S. Ohno, "So much 'junk' DNA in our genome”, *in Evolution of Genetic Systems, Brookhaven Symposium in Biology*, vol. 23, pp. 366-370, 1972.

[12] ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome”, *Nature*, vol. 489, pp. 57–74, 2012. Disponible : https://doi.org/10.1038/nature11247.

[13] L. Ohayon, "Génétique et évolution”, [En ligne]. Disponible : http://ohayon.lucie.free.fr/articles.php?lng=fr&pg=731&tconfig=0. (Consulté le 5 septembre 2023).

[14] K. Macé, E. Giudice, et R. Gillet, "La synthèse des protéines par le ribosome - Un chemin semé d'embûches", *médecine/sciences*, vol. 31, no. 3, pp. 282-290, 2015.
Disponible : https://doi.org/10.1051/medsci/20153103014.

[15] G. A. Petsko et D. Ringe, "Structure et fonction des protéines, 1er édition", DE BOECK SUP, 19 novembre 2008. 190 pages. ISBN-10: 2804158888, ISBN-13: 978-2804158880. (Traduction par D. Charmot et C. Sanlaville).

[16] N. Sonenberg et A. G. Hinnebusch, "Regulation of translation initiation in eukaryotes: mechanisms and biological targets.", *Cell*, vol. 136, no. 4, pp. 731–745, Feb. 2009. Disponible : https://doi.org/10.1016/j.cell.2009.01.042.

[17] N. A. Campbell et J. B. Reece, "Biologie", Éditions du Renouveau Pédagogique Inc., 2004. ISBN: 2-7613-1379-8.

[18] B. M. Alberts, A. Johnson, J. Lewis, et al., "Biologie moléculaire de la cellule", 4ème éd., Paris: Flammarion Médecine-Sciences, 2004. ISBN: 978-2-257-16219-9.

[19] H. Gobind Khorana, "H. Gobind Khorana Nobel Lecture", the nobel prize, [En ligne]. Disponible : https://www.nobelprize.org/prizes/medicine/1968/khorana/biographical/. (Consulté le 6 février 2023).

[20] Robert W. Holley, "Robert W. Holley Nobel Lecture", the nobel prize, [En ligne]. Disponible : https://www.nobelprize.org/prizes/medicine/1968/holley/facts/. (Consulté le 6 février 2023).

[21] Marshall W. Niremberg, "Marshall W. Nirenberg Nobel Lecture", the nobel prize, [En ligne]. Disponible : https://www.nobelprize.org/prizes/medicine/1968/nirenberg/facts/. (Consulté le 6 février 2023).

[22] Communauté Scenari, "Le code génétique", les bases de la biologie moléculaire [En ligne]. Disponible : https://www.supagro.fr/ress-tice/ue1-ue2_auto/Bases_Biologie_Moleculaire/co/_gc_code_genetique.html. (Consulté le 29 mars 2023).

[23] F. Gros, "La génomique et ses applications : promesses et limites", *Comptes Rendus Biologies*, vol. 338, nos. 8-9, pp. 543-546, 2015, ISSN 1631-0691. Disponible : https://doi.org/10.1016/j.crvi.2015.07.003.

[24] J. Lamoril, N. Ameziane, J.-C. Deybach, P. Bouizegarène, et M. Bogard, "Les techniques de séquençage de l'ADN : une révolution en marche. Première partie", *Immuno-Anal. Biol. Spéc.*, vol. 23, no. 5, pp. 260–279, Oct. 2008. Disponible : https://doi.org/10.1016/j.immbio.2008.07.016.

[25] F. S. Collins et L. Fink, "The Human Genome Project.", *Alcohol Health Res. World*, vol. 19, no. 3, pp. 190–195, 1995. Disponible : PMCID: PMC6875757

[26] E. El Fahime et M. M. Ennaji, "Évolution des techniques de séquençage", Les technologies de laboratoire, vol. 2, no.5, 2007. [En ligne].
Disponible : https://revues.imist.ma/index.php/technolab/article/view/334. (Consulté le 24 mars 2023).

[27] T. Korcsmaros, T. Boughtwood, T. Paton, et W. Scott, “Sequencing Buyer’s Guide 2022”. [En ligne]. Disponible : https://frontlinegenomics.com/the-sequencing-buyers-guide-5th-edition/. (Consulté le 10 septembre 2023).

[28] O. Ezratty, "Les technologies de séquençage du génome humain - 5”, [En ligne]. Disponible : https://www.oezratty.net/wordpress/2012/technologies-sequencage-genome-humain-5/. (Consulté le 10 septembre 2023).

[29] Université de Batna 2, "Les banques de données biologiques”, Département d’Écologie et Environnement, Faculté des Sciences de la Nature et de la Vie, Année académique 2021-2022, [En ligne]. Disponible : http://staff.univ-batna2.dz/. (Consulté le 6 juin 2023).

[30] J. Wang, L. Kong, G. Gao, et J. Luo, “A brief introduction to web-based genome browsers”, *Brief. Bioinform.*, vol. 14, no. 2, pp. 131–143, Mar. 2013. Disponible : https://doi.org/10.1093/bib/bbs029.

[31] United States National Library of Medicine, "National Center for Biotechnology Information”, National Library of Medicine, [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov. (Consulté le 06 mai 2023).

[32] Université Santa Cruz “,University of California Santa Cruz Genome Browser”, [En ligne]. Disponible : http://genome.ucsc.edu. (Consulté le 6 mai 2023).

[33] Institut européen de bioinformatique, "EMBL’s European Bioinformatics Institute”, [En ligne]. Disponible : https://www.ebi.ac.uk. (Consulté le 6 mai 2023).

[34] United States National Library of Medicine, "Genbank NCBI statistics”, [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/genbank/statistics/. (Consulté le 6 mai 2023).

[35] INSDC, "International Nucleotide Sequence Database Collaboration - INSDC”, [En ligne]. Disponible : https://www.insdc.org. (Consulté le 6 mai 2023).

[36] W. R. Pearson, “BLAST and FASTA similarity searching for multiple sequence alignment”, *Methods Mol. Biol. Clifton NJ*, vol. 1079, pp. 75–101, 2014. Disponible : https://doi.org/10.1007/978-1-62703-646-7_5.

[37] W. R. Pearson et D. J. Lipman, “Improved tools for biological sequence comparison”, *Proc. Natl. Acad. Sci.* U. S. A., vol. 85, no. 8, pp. 2444–2448, Apr. 1988. Disponible : https://doi.org/10.1073/pnas.85.8.2444.

[38] M. Burset et R. Guigó, “Evaluation of Gene Structure Prediction Programs”, *Genomics*, vol. 34, no. 3, pp. 353–367, Jun. 1996. Disponible : https://doi.org/10.1006/geno.1996.0298.

[39] R. Guigo, “Roderic guigo lab”, [En ligne]. Disponible : https://genome.crg.eu. (Consulté le 10 mai 2023).

[40] S. Rogic, “HMR195 dataset”, [En ligne]. Disponible : https://srogic.wordpress.com/research/. (Consulté le 2 janvier 2023).

[41] D. E. Gordon, J. Hiatt et al., “Homo sapiens tubulin folding cofactor A (TBCA), transcript variant 3, mRNA- NCBI- Reference Sequence: NM_001297740.2”, NCBI, [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/nuccore/NM_001297740.2?report=fasta. (Consulté le 4 juin 2023).

[42] MathWorks, “MATLAB”, [En ligne]. Disponible : https://www.mathworks.com/products/matlab.html. (Consulté le 4 juin 2023).

[43] A. D. Johnson, “An extended IUPAC nomenclature code for polymorphic nucleic acids”, *Bioinforma. Oxf. Engl.*, vol. 26, no. 10, pp. 1386–1389, May 2010. Disponible : https://doi.org/10.1093/bioinformatics/btq098.

[44] S. Rogic, A. K. Mackworth, et F. B. F. Ouellette, “Evaluation of Gene-Finding Programs on Mammalian Sequences”, *Genome Res.*, vol. 11, no. 5, pp. 817–832, May 2001. Disponible : https://doi.org/10.1101/gr.147901.

[45] N. Scalzitti, A. Jeannin-Girardon, P. Collet, O. Poch, et J. D. Thompson, “A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms”, *BMC Genomics*, vol. 21, no. 1, p. 293, Dec. 2020. Disponible : https://doi.org/10.1186/s12864-020-6707-9.

[46] G. Parra, E. Blanco, et R. Guigó, “GeneID in Drosophila”, *Genome Research*, vol. 10, no.4, pp. 511–515, 2000. Disponible : https://doi.org/10.1101/gr.10.4.511.

[47] S. Rogic, B. F. F. Ouellette, et A. K. Mackworth, “Improving gene recognition accuracy by combining predictions from two gene-finding programs”, *Bioinformatics*, vol. 18, no. 8, pp. 1034–1045, Aug. 2002. Disponible : https://doi.org/10.1093/bioinformatics/18.8.1034.

[48] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, et B. Morgenstern, “AUGUSTUS: ab initio prediction of alternative transcripts”, *Nucleic Acids Research*, vol. 34, no. Web Server issue, pp. W435–W439, 2006. Disponible : https://doi.org/10.1093/nar/gkl200.

[49] M. S. Mabrouk, “Advanced Genomic Signal Processing Methods in DNA Mapping Schemes for Gene Prediction Using Digital Filters”, *Am. J. Signal Process.*, vol. 2017, no. 1, pp. 12-24, 2017. Disponible : https://doi.org/10.5923/j.ajsp.20170701.02.

[50] Q. Zheng, T. Chen, W. Zhou, S. A. Marhon, L. Xie, et H. Su, “SAVMD: An adaptive signal processing method for identifying protein coding regions”, *Biomed. Signal Process. Control*, vol. 70, p. 102998, Sep. 2021. Disponible : https://doi.org/10.1016/j.bspc.2021.102998.

[51] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, et R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences”, *Bioinformatics*, vol. 13, no. 3, pp. 263–270, juin 1997. Disponible : https://doi.org/10.1093/bioinformatics/13.3.263.

[52] C. Yin et S. S.-T. Yau, “A Fourier Characteristic of Coding Sequences: Origins and a Non-Fourier Approximation”, *J. Comput. Biol.*, vol. 12, no. 9, pp. 1153–1165, Nov. 2005. Disponible : https://doi.org/10.1089/cmb.2005.12.1153.

[53] A. A. Tsonis, J. B. Elsner, et P. A. Tsonis, “Periodicity in DNA coding sequences: Implications in gene evolution”, J. Theor. Biol., vol. 151, no. 3, pp. 323–331, Aug. 1991.
Disponible : https://doi.org/10.1016/S0022-5193(05)80381-9.

[54] M. Akhtar, J. Epps, et E. Ambikairajah, “Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction”, *IEEE J. Sel. Top. Signal Process*., vol. 2, no. 3, pp. 310–321, Jun. 2008. Disponible : https://doi.org/10.1109/JSTSP.2008.923854.

[55] R. F. Voss, “Evolution of long-range fractal correlations and 1/ f noise in DNA base sequences”, *Phys. Rev. Lett.*, vol. 68, no. 25, pp. 3805–3808, Jun. 1992.
Disponible : https://doi.org/10.1103/PhysRevLett.68.3805.

[56] C.-K. Peng et al., “Long-range correlations in nucleotide sequences”, *Nature*, vol. 356, no. 6365, pp. 168–170, Mar. 1992. Disponible : https://doi.org/10.1038/356168a0.

[57] W. Li et K. Kaneko, “Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence”, *EPL*, vol. 17, no. 7, 1992. Disponible : https://doi.org/10.1209/0295-5075/17/7/014.

[58] A. S. Borovik, A. Yu. Grosberg, et M. D. Frank-Kamenetskii, “Fractality of DNA Texts”, *J. Biomol. Struct. Dyn.*, vol. 12, no. 3, pp. 655–669, Dec. 1994.
Disponible : https://doi.org/10.1080/07391102.1994.10508765.

[59] H. J. Jeffrey, "Chaos game visualization of sequences”, *Computers & Graphics*, vol. 16, no. 1, pp. 25-33, 1992. Disponible : https://doi.org/10.1016/0097-8493(92)90067-6.

[60] A. Fiser, G. E. Tusnády, et I. Simon, “Chaos game representation of protein structures”, *J. Mol. Graph.*, vol. 12, no. 4, pp. 302–304, Dec. 1994. Disponible : https://doi.org/10.1016/0263-7855(94)80109-6.

[61] I. Messaoudi, "Caractérisation des génomes bactériens par la représentation en jeu de Chaos et étude fréquentielle des textures par Transformée de Fourier à deux dimensions”, *Revue électronique Sciences et Technologies de l’Automatique (e-sta)*, 2011. [En ligne]. (Consulté le 10 septembre 2023).

[62] C. Yin, “Encoding and Decoding DNA Sequences by Integer Chaos Game Representation”, *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 26, no. 2, pp. 143–151, 2019. Disponible : https://doi.org/10.1089/cmb.2018.0173.

[63] M. Meraz, R. Carbó, E. Rodriguez, et J. Alvarez-Ramirez, “Fractal correlations in the Covid-19 genome sequence via multivariate rescaled range analysis“, *Chaos Solitons Fractals*, vol. 168, p. 113132, Mar. 2023. Disponible : https://doi.org/10.1016/j.chaos.2023.113132.

[64] H. K. Kwan et S. B. Arniker, “Numerical representation of DNA sequences”, *2009 IEEE International Conference on Electro/Information Technology, Windsor, ON, Canada*, 2009, pp. 307-310.
Disponible : https://doi.org/10.1109/EIT.2009.5189632.

[65] A. Fiser, G. E. Tusnády, et I. Simon, “Chaos game representation of protein structures”, *J. Mol. Graph.*, vol. 12, no. 4, pp. 302–304, Dec. 1994. Disponible : https://doi.org/10.1016/0263-7855(94)80109-6.

[66] F. Castro-Chavez, “A Tetrahedral Representation of the Genetic Code Emphasizing Aspects of Symmetry”, *BIO-Complex*., vol. 2012, no. 2, pp. 1–6, Jun. 2012.
Disponible : https://doi.org/10.5048/BIO-C.2012.2.

[67] T. Kohonen et P. Somervuo, “How to make large self-organizing maps for nonvectorial data”, *Neural Netw. Off. J. Int. Neural Netw. Soc*., vol. 15, no. 8–9, pp. 945–952, Nov. 2002.
Disponible : https://doi.org/10.1016/s0893-6080(02)00069-2.

[68] A. K. Brodzik et O. Peters, “Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences”, *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, Philadelphia, PA, USA, 2005, pp. v/373-v/376 Vol. 5.
Disponible : https://doi.org/10.1109/ICASSP.2005.1416318.

[69] R. Zhang et C. Zhang, “Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences”, *J. Biomol. Struct. Dyn*., vol. 11, pp. 767–82, Mar. 1994.
Disponible : https://doi.org/10.1080/07391102.1994.10508031.

[70] E. Hamori et J. Ruskin, “H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences”, *J. Biol. Chem*., vol. 258, no. 2, pp. 1318–1327, Jan. 1983.
Disponible : https://doi.org/10.1016/S0021-9258(18)33196-X.

[71] P. D. Cristea, “Genetic signal representation and analysis”, *in Proc. SPIE, vol. 4623, Functional Monitoring and Drug-Tissue Interaction*, 5 juin 2002. Disponible : https://doi.org/10.1117/12.491244.

[72] L. Galleani et R. Garello, “The Minimum Entropy Mapping Spectrum of a DNA Sequence”, *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 771-783, février 2010.
Disponible : https://doi.org/10.1109/TIT.2009.2037041.

[73] N. Chakravarthy, A. Spanias, L. D. Iasemidis, et K. Tsakalis, “Autoregressive Modeling and Feature Analysis of DNA Sequences”, *EURASIP J. Adv. Signal Process*., vol. 2004, no. 1, p. 952689, Dec. 2004.
Disponible : https://doi.org/10.1155/S111086570430925X.

[74] G. L. Rosen, “Examining coding structure and redundancy in DNA. How does DNA protect itself from life’s uncertainty ?” , *IEEE Eng. Med. Biol. Mag. Q. Mag. Eng. Med. Biol. Soc*., vol. 25, no. 1, pp. 62–68, Feb. 2006. Disponible : https://doi.org/10.1109/memb.2006.1578665.

[75] X. Jiang, D. Lavenier, et S. S.-T. Yau, “Coding Region Prediction Based on a Universal DNA Sequence Representation Method”, *J. Comput. Biol.*, vol. 15, no. 10, pp. 1237–1256, Dec. 2008.
Disponible : https://doi.org/10.1089/cmb.2008.0041.

[76] Z. Liu, B. Liao, W. Zhu, et G. Huang, “A 2D Graphical Representation of DNA Sequence Based on Dual Nucleotides and Its Application”, *Int. J. Quantum Chem*., vol. 109, pp. 948–958, Jan. 2009.
Disponible : https://doi.org/10.1002/qua.21919.

[77] A. T. M. G. Bari, M. R. Reaz, A. K. M. T. Islam, H.-J. Choi, et B.-S. Jeong, “Effective Encoding for DNA Sequence Visualization Based on Nucleotide’s Ring Structure”, *Evol. Bioinforma*. Online, vol. 9, pp. 251–261, 2013. Disponible : https://doi.org/10.4137/EBO.S12160.

[78] M. Akhtar, J. Epps, et E. Ambikairajah, “On DNA Numerical Representations for Period-3 Based Exon Prediction”, *2007 IEEE International Workshop on Genomic Signal Processing and Statistics, Tuusula, Finland*, Jul. 2007, pp. 1–4. Disponible : https://doi.org/10.1109/GENSIPS.2007.4365821.

[79] A. S. S. Nair et T. Mahalakshmi, “Visualization of genomic data using inter-nucleotide distance signals”, in *Proceedings of IEEE Genomic Signal Processing*, 2005, vol. 408. [En ligne]. (Consulté le 09 septembre 2023).

[80] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, et J. L. Oliver, “CpGcluster: a distance-based algorithm for CpG-island detection”, *BMC Bioinformatics*, vol. 7, p. 446, Oct. 2006. Disponible : https://doi.org/10.1186/1471-2105-7-446.

[81] N. Yu, Z. Yu, F. Gu, et Y. Pan, “Evaluating the Impact of Encoding Schemes on Deep Auto-Encoders for DNA Annotation”, *in Bioinformatics Research and Applications, ISBRA 2017*, vol. 10330, Lecture Notes in Computer Science, Springer, Cham, 2017. Disponible : https://doi.org/10.1007/978-3-319-59575-7_40.

[82] S. Zou, L. Wang, et J. Wang, “A 2D graphical representation of the sequences of DNA based on triplets and its application”, *Eurasip J. Bioinforma. Syst. Biol.*, vol. 2014, no. 1, 2014.
Disponible : https://doi.org/10.1186/1687-4153-2014-1.

[83] R. K. M et N. K. Vaegae, “Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes”, *Biomed. Signal Process. Control*, vol. 58, 2020.
Disponible : https://doi.org/10.1016/j.bspc.2020.101859.

[84] C. Yin et S. S. T. Yau, “Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes”, *in 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Sun Valley, ID, USA, 2008, pp. 223-227.
Disponible : https://doi.org/10.1109/CIBCB.2008.4675783.

[85] T. Holden et al., “ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes”, *Instruments, Methods, and Missions for Astrobiology X*, 2007.
Disponible : https://doi.org/10.1117/12.732283.

[86] F. C. D. L. Torre, J. I. González-Trejo, C. A. Real-Ramírez, et L. F. Hoyos-Reyes, “Fractal dimension algorithms and their application to time series associated with natural phenomena”, *in Journal of Physics: Conference Series*, 2013. Disponible : https://doi.org/10.1088/1742-6596/475/1/012002.

[87] A. S. Nair et S. P. Sreenadhan, “A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)”, *Bioinformation*, vol. 1, no. 6, pp. 197-202, octobre 2006. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891688/. (Consulté le 09 septembre 2023).

[88] H. E. Stanley et al., “Statistical mechanics in biology: how ubiquitous are long-range correlations ?”, *Phys. Stat. Mech. Its Appl.*, vol. 205, no. 1–3, 1994. Disponible : https://doi.org/10.1016/0378-4371(94)90502-9.

[89] K. J. Breslauer, R. Frank, H. Blocker, et L. A. Marky, “Predicting DNA duplex stability from the base sequence”, *Proc. Natl. Acad. Sci.* U. S. A., vol. 83, no. 11, 1986.
Disponible : https://doi.org/10.1073/pnas.83.11.3746.

[90] M. S. Mabrouk, "A Study of the Potential of EIIP Mapping Method in Exon Prediction Using the Frequency Domain Techniques", *Am. J. Biomed. Eng.*, vol. 2, no. 2, pp. 17–22, Aug. 2012. Disponible : https://doi.org/10.5923/j.ajbe.20120202.04.

[91] S. Kar et M. Ganguly, "Study of effectiveness of FIR and IIR filters in Exon identification: A comparative approach", *Mater. Today Proc*., p. S2214785322010392, Mar. 2022. Disponible : https://doi.org/10.1016/j.matpr.2022.02.394.

[92] P. Ramachandran, W.-S. Lu, et A. Antoniou, "Location of exons in DNA sequences using digital filters", *in 2009 IEEE International Symposium on Circuits and Systems (ISCAS)*, Taipei, Taiwan, 2009, pp. 2337-2340. Disponible : https://doi.org/10.1109/ISCAS.2009.5118268.

[93] A. K. Singh and V. K. Srivastava, "Improved filtering approach for identification of protein-coding regions in eukaryotes by background noise reduction using S–G filter", *Netw. Model. Anal. Health Inform. Bioinforma.*, vol. 10, no. 1, p. 19, Dec. 2021. Disponible : https://doi.org/10.1007/s13721-021-00293-8.

[94] A. M. Dessouky, T. E. Taha, M. M. Dessouky, A. A. Eltholth, E. Hassan, et F. E. Abd El-Samie, "Non-parametric spectral estimation techniques for DNA sequence analysis and exon region prediction", *Comput. Electr. Eng.*, vol. 73, pp. 334–348, Jan. 2019. Disponible : https://doi.org/10.1016/j.compeleceng.2018.12.001.

[95] Q. Zheng, T. Chen, W. Zhou, L. Xie, et H. Su, "Gene prediction by the noise-assisted MEMD and wavelet transform for identifying the protein coding regions", *Biocybern. Biomed. Eng.*, vol. 41, no. 1, pp. 196–210, Jan. 2021. Disponible : https://doi.org/10.1016/j.bbe.2020.12.005.

[96] V. Pathak, S. J. Nanda, A. M. Joshi, et S. S. Sahu, "VLSI implementation of anti-notch lattice structure for identification of exon regions in Eukaryotic genes*", IET Comput. Digit. Tech.*, vol. 14, no. 5, pp. 217–229, Sep. 2020. Disponible : https://doi.org/10.1049/iet-cdt.2019.0086.

[97] H. Delacour, A. Servonnet, A. Perrot, J. F. Vigezzi et J. M. Ramirez, "La courbe ROC (receiver operating characteristic): principes et principales applications en biologie clinique", Annales de biologie clinique, vol. 63, no. 2, pp. 145-154, mars 2005. [En ligne].

[98] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognit. Lett*., vol. 27, no. 8, pp. 861–874, Jun. 2006. Disponible : https://doi.org/10.1016/j.patrec.2005.10.010.

[99] P. Ramachandran, Wu-Sheng Lu, et A. Antoniou, "Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA*", IEEE Trans. Biomed. Eng*., vol. 59, no. 6, pp. 1598–1609, Jun. 2012. Disponible : https://doi.org/10.1109/TBME.2012.2190512.

[100] J. W. Fickett and C.-S. Tung, "Assessment of protein coding measures", *Nucleic Acids Res*., vol. 20, no. 24, pp. 6441–6450, 1992. Disponible: https://doi.org/10.1093/nar/20.24.6441.

[101] D. Anastassiou, "Genomic signal processing, Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, vol.16, no. 12, pp. 1073–1081, Dec. 2000. Disponible : https://doi.org/10.1093/bioinformatics/16.12.1073

[102] C. Yin et S. S.-T. Yau, “Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence”, *J. Theor. Biol.*, vol. 247, no. 4, pp. 687–694, Aug. 2007. Disponible : https://doi.org/10.1016/j.jtbi.2007.03.038.

[103] J. P. Mena-Chalco, H. Carrer, Y. Zana, et R. M. Cesar, “Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform”, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 5, no. 2, pp. 198–207, Apr. 2008. Disponible : https://doi.org/10.1109/TCBB.2007.70259.

[104] S. Datta, A. Asif, et H. Wang, “Prediction of protein coding regions in DNA sequences using Fourier spectral characteristics”, *in Proceedings - IEEE Sixth International Symposium on Multimedia Software Engineering, MSE 2004*, 2004. Disponible : https://doi.org/10.1109/mmse.2004.63.

[105] M. K. Choong et H. Yan, “Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves”, *Bioinformation*, vol. 2, no. 7, 2008. Disponible : https://doi.org/10.6026/97320630002273.

[106] D. K. Shakya, R. Saxena, et S. N. Sharma, “An adaptive window length strategy for eukaryotic CDS prediction”, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 5, 2013. Disponible : https://doi.org/10.1109/TCBB.2013.76.

[107] S. Kar, M. Ganguly, et S. Das, “Using dit-fft algorithm for identification of protein coding region in eukaryotic gene”, *Biomed. Eng. Appl. Basis Commun.*, vol. 31, no. 01, p. 1950002, Feb. 2019. Disponible : https://doi.org/10.4015/S1016237219500029.

[108] O. Abbasi, A. Rostami, et G. Karimian, “Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform”, *BMC Bioinformatics*, vol. 12, no. 1, p. 430, Dec. 2011. Disponible : https://doi.org/10.1186/1471-2105-12-430.

[109] J. Y. Y. Kwan, B. Y. M. Kwan, et H. K. Kwan, “Spectral analysis of numerical exon and intron sequences”, *in 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2010*, 2010. Disponible : https://doi.org/10.1109/BIBMW.2010.5703954.

[110] S. S. Sahu et G. Panda, “Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach”, *Genomics Proteomics Bioinformatics*, vol. 9, no. 1–2, pp. 45–55, Apr. 2011. Disponible : https://doi.org/10.1016/S1672-0229(11)60007-7.

[111] P. P. Vaidyanathan et Byung-Jun Yoon, “Digital filters for gene prediction applications”, *in Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.*, Pacific Grove, CA, USA: IEEE, 2002, pp. 306–310 vol.1. Disponible : https://doi.org/10.1109/ACSSC.2002.1197196.

[112] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, et R. Ramaswamy, “Prediction of probable genes by Fourier analysis of genomic sequences”, *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997. Disponible : https://doi.org/10.1093/bioinformatics/13.3.263.

[113] Raymond Guan and J. Tuqan, “Multirate DSP models for gene detection”, *in Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, Pacific Grove, CA, USA: IEEE, 2004, pp. 1641–1645. Disponible : https://doi.org/10.1109/ACSSC.2004.1399436.

[114] R. Kakumani, V. Devabhaktuni, et M. O. Ahmad, “Prediction of protein-coding regions in DNA sequences using a model-based approach”, *in 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, USA: IEEE*, May 2008, pp. 1918–1921.
Disponible : https://doi.org/10.1109/ISCAS.2008.4541818.

[115] V. Tomar, D. Gandhi, et C. Vijaykumar, “Digital signal processing for gene prediction”, *in TENCON 2008 - 2008 IEEE Region 10 Conference, Hyderabad, India: IEEE,* Nov. 2008, pp. 1–5.
Disponible : https://doi.org/10.1109/TENCON.2008.4766648.

[116] S. Barman, S. Biswas, S. Das, et M. Roy, “Performance analysis and simulation of IIR anti-notch filter with various structures for gene prediction application”, *in 2012 5th International Conference on Computers and Devices for Communication (CODEC), Kolkata: IEEE,* Dec. 2012, pp. 1–4.
Disponible : https://doi.org/10.1109/CODEC.2012.6509360.

[117] National Center for Biotechnology Information, "Genbank”, [En ligne].
Disponible : https://www.ncbi.nlm.nih.gov/genbank/. (Consulté le 9 septembre 2023).

[118] P. Ramachandran, Wu-Sheng Lu, et A. Antoniou, “Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA”, *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1598–1609, Jun. 2012.
Disponible : https://doi.org/10.1109/TBME.2012.2190512.

[119] H. Wassfy, M. Elnaby, M. Salem, M. Mabrouk, et A.-A. Zidan, “Advanced DNA Mapping Schemes for Exon Prediction Using Digital Filters”, *Amj. Biomed. Engineering* vol. 2016, pp. 25–31, Apr. 2016.
Disponible : https://doi.org/10.5923/j.ajbe.20160601.04.

[120] T. P. George et T. Thomas, “Discrete wavelet transform de-noising in eukaryotic gene splicing”, *BMC Bioinformatics*, vol. 11, no. S1, p. S50, Jan. 2010.
Disponible : https://doi.org/10.1186/1471-2105-11-S1-S50.

[121] M. K. Hota et V. K. Srivastava, “Identification of protein coding regions using antinotch filters*”, Digit. Signal Process.*, vol. 22, no. 6, pp. 869–877, Dec. 2012.
Disponible : https://doi.org/10.1016/j.dsp.2012.06.005.

[122] A. Savitzky et M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures”, *Anal. Chem.*, vol. 36, no. 8, 1964. Disponible : https://doi.org/10.1021/ac60214a047.

[123] Y. Ferdi, “Some Applications Of Fractional Order Calculus To Design Digital Filters For Biomedical Signal Processing”, *J. Mech. Med. Biol.*, vol. 12, no. 02, p. 1240008, Apr. 2012.
Disponible : https://doi.org/10.1142/S0219519412400088.

[124] E. Renshaw et M. F. Barnsley, “Fractals Everywhere.”, *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 158, no. 2, 1995. Disponible : https://doi.org/10.2307/2983297.

[125] K. Falconer, “Fractal Geometry: Mathematical Foundations and Applications”, John Wiley & Sons, 2004. ISBN-10: 0-470-84861-8 (Hardcover); ISBN-13: 978-0-470-84861-7 (Hardcover); ISBN-10: 0-470-84862-6 (Paperback); ISBN-13: 978-0-470-84862-4 (Paperback).

[126] L. E. Calvet et A. J. Fisher, “Multifractal volatility: Theory, forecasting, and pricing”. 2008. Disponible : https://doi.org/10.1016/B978-0-12-150013-9.X5001-6.

[127] G. Dupuy, Queiros-Condé D., Chaline J. et Dubois , “Le monde des fractales, la nature trans-échelles”, Paris, Ellipses, 476 p. (2ème édition), Cybergeo, 2015. Disponible : https://doi.org/10.4000/cybergeo.27283.

[128] A. Kilbas, H. M. Srivastava, et J. J. Trujillo, “NEW BOOK: 'Theory and Applications of Fractional Differential Equations'”, *Elsevier*, North-Holland Mathematics Studies, vol. 204, Fractional Calculus and Applied Analysis, vol. 9, no. 1, p. 71, 2006.

[129] Y. Ding et H. Ye, “A fractional-order differential equation model of HIV infection of CD4+ T-cells”, *Mathematical and Computer Modelling*, vol. 50, no. 3-4, pp. 386-392, 2009. Disponible : https://doi.org/10.1016/j.mcm.2009.04.019.

[130] A. M. Sabatini, “A statistical mechanical analysis of postural sway using non-Gaussian FARIMA stochastic models”, *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1219-1227, sept. 2000. Disponible : https://doi.org/10.1109/10.867954.

[131] R. Magin, O. Abdullah, D. Baleanu et J. Zhou, “Anomalous diffusion expressed through fractional order differential operators in the Bloch–Torrey equation”, *Journal of Magnetic Resonance*, vol. 190, pp. 255-270, 2008. Disponible : https://doi.org/10.1016/j.jmr.2007.11.007.

[132] L. Song, S. Xu, et J. Yang, “Dynamical models of happiness with fractional order”, *Communications in Nonlinear Science and Numerical Simulation*, vol. 15, no. 3, pp. 616-628, 2010. Disponible : https://doi.org/10.1016/j.cnsns.2009.04.029.

[133] W. M. Ahmad et R. El-Khazali, “Fractional-order dynamical models of love”, *Chaos, Solitons & Fractals*, vol. 33, no. 4, pp. 1367-1375, 2007. Disponible : https://doi.org/10.1016/j.chaos.2006.01.098.

[134] E. Ahmed et H. A. El-Saka, “On fractional order models for Hepatitis C”, *Nonlinear Biomed. Phys.*, vol. 4, no. 1, p. 1, 2010. Disponible : https://doi.org/10.1186/1753-4631-4-1.

[135] A. El-Sayed, S. Rida et A. Arafa, “On the solutions of time-fractional bacterial chemotaxis in a diffusion gradient chamber”, *International Journal of Nonlinear Science*, vol. 7, pp. 1749-3889, Jan. 2009.

[136] J. Lu et M. Xie, “Use fractional calculus in iris localization”, *in 2008 International Conference on Communications, Circuits and Systems*, Xiamen, China, 2008, pp. 946-949. Disponible : https://doi.org/10.1109/ICCCAS.2008.4657925.

[137] Y. Ferdi, J.P. Herbeuval, A. Charef, et B. Boucheham, “R wave detection using fractional digital differentiation”, *ITBM-RBM*, vol. 24, nos 5-6, pp. 273-280, 2003. Disponible : https://doi.org/10.1016/j.rbmret.2003.08.002.

[138] Y. Ferdi, “Impulse invariance-based method for the computation of fractional integral of order 0<α<1”, *Computers & Electrical Engineering*, vol. 35, no. 5, pp. 722-729, 2009. Disponible : https://doi.org/10.1016/j.compeleceng.2009.02.005.

[139] J. Baranowski et P. Piątek, “Fractional Band-Pass Filters: Design, Implementation and Application to EEG Signal Processing”, *J Circuits Syst Comput*, vol. 26, no. 11, p. 1750170, 2017. Disponible : https://doi.org/10.1142/S0218126617501705.

[140] Ferdi, A. Taleb-Ahmed, et M. R. Lakehal, “Efficient Generation of $1/f^{\beta}$ Noise Using Signal Modeling Techniques”, *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1704-1710, juillet 2008. Disponible : https://doi.org/10.1109/TCSI.2008.918173.

[141] M. Benmalek et A. Charef, “Digital fractional order operators for R-wave detection in electrocardiogram signal”, *IET Signal Processing*, vol. 3, no. 5, pp. 381-391, sept. 2009. Disponible : https://doi.org/10.1049/iet-spr.2008.0094.

[142] Y. Ferdi et A. Taleb-Ahmed, “A Procedure for Efficient Generation of 1/f β Noise Sequences”, *in Image and Signal Processing. ICISP 2008*, A. Elmoataz, O. Lezoray, F. Nouboud, et D. Mammass (éds.), vol. 5099, Springer, Berlin, Heidelberg. Disponible : https://doi.org/10.1007/978-3-540-69905-7_56.

[143] A. Goutas, Y. Ferdi, J.-P. Herbeuval, M. Boudraa, et B. Boucheham, “Digital fractional order differentiation-based algorithm for P and T-waves detection and delineation”, *in ITBM-RBM*, vol. 26, no. 2, pp. 127-132, 2005. Disponible : https://doi.org/10.1016/j.rbmret.2004.11.022.

[144] A. Charef et F. Abdelliche, “Fractional wavelet for R-Wave detection in ECG signal”, *Critical Reviews in Biomedical Engineering*, vol. 36, no. 2-3, pp. 79-91, 2008. Disponible :https://doi.org/10.1615/critrevbiomedeng.v36.i2-3.10.

[145] Y. Li, H. Tang, et H. Chen, “Fractional-order derivative spectroscopy for resolving simulated overlapped Lorenztian peaks*”, Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 83-89, 2011. Disponible: https://doi.org/10.1016/j.chemolab.2011.01.013.

[146] A. K. Shukla, R. K. Pandey, et R. B. Pachori, “A fractional filter based efficient algorithm for retinal blood vessel segmentation”, *Biomedical Signal Processing and Control*, vol. 59, p. 101883, 2020. Disponible : https://doi.org/10.1016/j.bspc.2020.101883.

[147] Y. Ferdi, “A Fractional Digital High-Pass Notch Filter”, *in* 2020 *10th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Saint-Etienne, France, 2021, pp. 1-4. Disponible : https://doi.org/10.1109/ISIVC49222.2021.9487545.

[148] K. Assaleh et W. M. Ahmad, “Modeling of speech signals using fractional calculus”, *in 2007 9th International Symposium on Signal Processing and Its Applications*, Sharjah, United Arab Emirates, 2007, pp. 1-4. Disponible : https://doi.org/10.1109/ISSPA.2007.4555563.

[149] J. Wang, Y. Ye, X. Pan, et X. Gao, “Parallel-type fractional zero-phase filtering for ECG signal denoising”, *Biomedical Signal Processing and Control*, vol. 18, pp. 36-41, 2015, ISSN 1746-8094. Disponible : https://doi.org/10.1016/j.bspc.2014.10.012.

[150] B.T. Krishna, “Studies on fractional order differentiators and integrators: A survey”, *Signal Processing*, vol. 91, no. 3, pp. 386-426, 2011, ISSN 0165-1684.
Disponible : https://doi.org/10.1016/j.sigpro.2010.06.022.

[151] Y. Ferdi, “Fast generation of generalized autoregressive moving average processes”, *in 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, Istanbul, Turkey: IEEE, Jun. 2014, pp. 1004–1009. Disponible : https://doi.org/10.1109/ISIE.2014.6864749.

[152] M. R. Lakehal, Y. Ferdi et A. Taleb-Ahmed, “Generation of FARIMA(0,α,0) sequences by recursive filtering: Testing for self-similarity”, *2009 16th IEEE International Conference on Electronics, Circuits and Systems - (ICECS 2009)*, Yasmine Hammamet, Tunisia, 2009, pp. 563-566.
Disponible : https://doi.org/10.1109/ICECS.2009.5410865.

[153] Yan Li, Hu Sheng et YangQuan Chen, “Analytical impulse response of a fractional second order filter and its impulse response invariant discretization”, *Signal Processing*, vol. 91, no. 3, pp. 498-507, 2011, ISSN 0165-1684. Disponible : https://doi.org/10.1016/j.sigpro.2010.01.017.

[154] Y. Ferdi, “Fractional order calculus-based filters for biomedical signal processing”, in *2011 1st Middle East Conference on Biomedical Engineering*, Sharjah, United Arab Emirates, 2011, pp. 73-76.
Disponible : https://doi.org/10.1109/MECBME.2011.5752068.

[155] Y. Ferdi, “Improved lowpass differentiator for physiological signal processing”, in *2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*, Newcastle Upon Tyne, UK, 2010, pp. 747-750.
Disponible : https://doi.org/10.1109/CSNDSP16145.2010.5580319.

[156] G. S. Visweswaran, P. Varshney et M. Gupta, “New Approach to Realize Fractional Power in z-Domain at Low Frequency”, *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 3, pp. 179-183, March 2011. Disponible : https://doi.org/10.1109/TCSII.2011.2110350.

[157] Yan Li, Hu Sheng et Yang Quan Chen, “On distributed order integrator/differentiator”, *Signal Processing*, vol. 91, no. 5, pp. 1079-1084, 2011, ISSN 0165-1684.
Disponible : https://doi.org/10.1016/j.sigpro.2010.10.005.

[158] Y. Ferdi, “Computation of Fractional Order Derivative and Integral via Power Series Expansion and Signal Modelling”, *Nonlinear Dyn.*, vol. 46, no. 1–2, pp. 1–15, Sep. 2006.
Disponible : https://doi.org/10.1007/s11071-005-9000-1.

[159] C.-F. Chung, “Estimating a generalized long memory process”, *J. Econom.*, vol. 73, no. 1, pp. 237–259, Jul. 1996. Disponible : https://doi.org/10.1016/0304-4076(95)01739-9.

[160] J. Meher, P.K. Meher, et G. Dash, “Improved Comb Filter based Approach for Effective Prediction of Protein Coding Regions in DNA Sequences”, *J. Signal and Information Processing*, vol. 2, pp. 88-99, 2011.
Disponible : https://doi.org/10.4236/jsip.2011.22012.

[162] A. K. Roonizi et C. Jutten, “Forward-backward Filtering and Penalized Least-Squares Optimization: A Unified Framework”, *Signal Processing*, vol. 178, p. 107796, 2021. Disponible : https://doi.org/10.1016/j.sigpro.2020.107796.

[163] F. Gustafsson, “Determining the initial states in forward-backward filtering”, *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 988–992, Apr. 1996. Disponible : https://doi.org/10.1109/78.492552.

[164] A. Singh et V. Srivastava, “Bidirectional filtering approach for the improved protein coding region identification in eukaryotes”, *Netw. Model. Anal. Health Inform. Bioinforma.*, vol. 11, Dec. 2022. Disponible : https://doi.org/10.1007/s13721-022-00358-2.

[165] P. Ramachandran, W.-S. Lu, et A. Antoniou, “Improved hot-spot location technique for proteins using a bandpass notch digital filter”, *in 2008 IEEE International Symposium on Circuits and Systems (ISCAS)*, Seattle, WA, USA, 2008, pp. 2673-2676. Disponible : https://doi.org/10.1109/ISCAS.2008.4542007.

[166] M. Lehilahy et Y. Ferdi, “Identification of exon locations in DNA sequences using a fractional digital anti-notch filter”, *Biomed. Signal Process. Control*, vol. 80, p. 104362, Feb. 2023. Disponible : https://doi.org/10.1016/j.bspc.2022.104362.

[167] I. M. El-Badawy, S. Gasser, A. M. Aziz, et M. E. Khedr, “On the use of Pseudo-EIIP mapping scheme for identifying exons locations in DNA sequences”, *in 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, Malaysia: IEEE, Oct. 2015, pp. 244–247. Disponible : https://doi.org/10.1109/ICSIPA.2015.7412197.

[168] S. S. Sahu et G. Panda, “Identification of Protein-Coding Regions in DNA Sequences Using A Time-Frequency Filtering Approach”, *Genomics Proteomics Bioinformatics*, vol. 9, no. 1–2, pp. 45–55, Apr. 2011, Disponible : https://doi.org/10.1016/S1672-0229(11)60007-7.

[169] T. M. Inbamalar et R. Sivakumar, “Improved Algorithm for Analysis of DNA Sequences Using Multiresolution Transformation”, *Sci. World J.*, vol. 2015, pp. 1–9, 2015, Disponible : https://doi.org/10.1155/2015/786497.

ANNEXE

Extrait de code du programme pour l'identification des exons

Ci-dessous est montré l'extrait code du programme MATLAB pour l'identification des exons en suivant les 6 étapes de l'algorithme présenté dans le chapitre V (figure V.1).

C'est le cas de la séquence AF042782 avec 2 exons de la base de données HMR195.

1ère étape : acquisition de donnée FASTA de la séquence :

```
% Chemin d'accès au fichier FASTA
fastaFile = 'D:\...sequenceset\DNASequencesM.fasta';

% Utilisation de fastaread pour lire le fichier
sequences = fastaread(fastaFile);

% Boucle pour parcourir chaque séquence
for i = 1:numel(sequences)
    header = sequences(i).Header;
    sequence = sequences(i).Sequence;

% Séparation de l'en-tête et de la séquence
    disp([header, ' ', sequence]);
end
```

2ème étape : Conversion numérique de la séquence FASTA :

```
% référence NCBI pour les positions des exons
x1=305; x2=672; x3=2063; x4=2858;
sequence='TTGAATTCAATTAAAACATGCTTTTTTGGGGGTAAAAAGAGCAACTTTGA.....';
l= length(sequence);
xA = [];xC = [];xG = [];xT = [];
%%  Application de la numérisation de Voss
for i=1:l
if sequence(i) == 'A' xA(i)=1 ; else xA(i)=0; end
if sequence(i) == 'C' xC(i)=1 ; else xC(i)=0; end
if sequence(i) == 'G' xG(i)=1 ; else xG(i)=0; end
if sequence(i) == 'T' xT(i)=1 ; else xT(i)=0; end
end
```

3ème étape : Application du filtre fractionnaire :

```
ax = 0.594;u = -0.5 ; p = 3; q = p; L = 1000 ; % paramètres du filtre
fractionnaire
```

```
h = Fract_impulse_response(ax,u,L); % sous-programme du filtre
fractionnaire
[b,a] = prony(h,q,p);      % sous-programme de calcul de la méthode de
Prony
nx = 5000; Fs= 1;
[H,f2]=freqz(b,a,nx,'whole',Fs);
Lf=length(f2)/2; f2= f2(1:Lf);
H=H(1:Lf);
yA = filtfilt(b,a,xA);   % filtrage bidirectionel
yG = filtfilt(b,a,xG);
yT = filtfilt(b,a,xT);
yC =filtfilt(b,a,xC);
```

4ème étape : Module au carré :

```
yA= abs(yA).^2; yG= abs(yG).^2;yT= abs(yT).^2;yC= abs(yC).^2; % module au
carré
```

5ème étape : filtrage passe-bas :

```
lp =lowpass_cheby_inv2 ;   % sous-programme de filtre passe-bas de
Tchebychev inverse II

y_finalA = filter(lp,yA);y_finalG = filter(lp,yG);
y_finalT = filter(lp,yT);y_finalC = filter(lp,yC);
```

6ème étape : résultats et évaluation :

- **Evaluation graphique :**

```
S= y_finalA + y_finalG + y_finalT+ y_finalC;
y_final = S/max(S);  % normalisation du signal de sortie
figure(1)
plot(y_final)
xline(x1,'r');xline(x2,'r');xline(x3,'r ');xline(x4,'r');
                        % lignes verticaux de référenciation NCBI
xlabel('Position des nucléotides ')
ylabel('Signal de sortie normalisé')
```

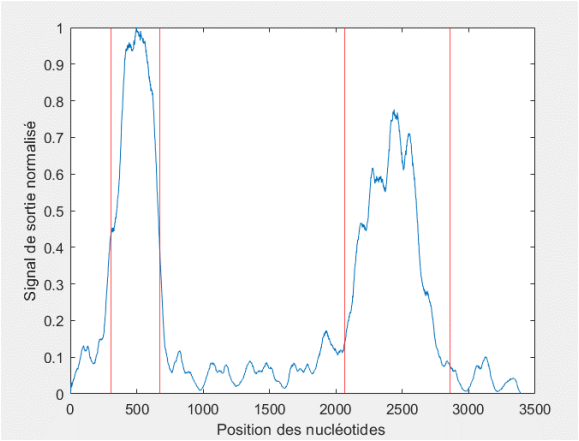


Figure A. 1 : Signal de sortie de l'algorithme. Cas du gène AF042782 avec 2 exons de la base de données HMR195

- **Calcul des paramètres d'évaluation :**

```
%% Initialisation des paramètres d'evaluations

k=0;
Tp1 =0;Tp2 =0;Fn1=0;Fn2 =0;Fp1=0;Fp2=0;Fp3=0;Tn1=0;Tn2=0;Tn3=0;

%%  Seuillage et calcul de VP, VN , FP, FN
for seuil = 0:0.01:1;
for i = x1:x2 ;
    if y_final(i) >= seuil
        Tp1 = Tp1+ 1;
    else
        Fn1 = Fn1 +1;
    end
end

for i = x3:x4 ;
    if y_final(i) >= seuil
        Tp2 = Tp2+ 1;
    else
        Fn2 = Fn2 +1;
    end
end

for i = 1:(x1-1)  ;
    if y_final(i) >= seuil
        Fp1 = Fp1+ 1;
    else
        Tn1 = Tn1 +1;
    end
end

for i = (x2+1) :(x3-1);
    if y_final(i) >= seuil
        Fp2= Fp2+ 1;
    else
        Tn2 = Tn2 +1;
    end
end

for i = (x4+1) :length(y_final)
    if y_final(i) >= seuil
        Fp3= Fp3+ 1;
    else
        Tn3 = Tn3 +1;
    end
end
```

```
k=k+1;

%% Calcul des autres paramètres d'évaluation
Tp(k) = Tp1+Tp2;Fn(k) = Fn1+Fn2;
Fp(k) = Fp1+Fp2+Fp3;Tn(k) = Tn1+Tn2+Tn3;
P(k) = Fp(k) + Tp(k);N(k) = Fn(k) + Tn(k);
Sensitivity(k) = Tp(k)/ ( Tp(k) + Fn(k) );
Specificity(k)  = Tp(k) / (Tp(k) + Fp(k) );
Ac(k) = (Tp(k) + Tn(k)) / (P(k) + N(k));
F1_score(k) = ( 2* Tp(k)) /(( 2*Tp(k)) + Fp(k) + Fn(k));
ACP(k)= (Sensitivity(k)+ Specificity(k)+ (Tn(k) / ( Fp(k) + Tn(k) ))+(Tn(k)
/ ( Fn(k) + Tn(k) )))/4;
AC(k)= ((ACP(k)-0.5)*2);
Tp1 =0; Tp2 =0;
Fn1=0; Fn2 =0;
Fp1=0; Fp2=0; Fp3=0;
Tn1=0; Tn2=0;  Tn3=0;
end
Sn=Sensitivity'
Sp=Specificity'
F1_score= F1_score'
Ac=Ac'
Approx=AC'
%% Affichage du ROC et calcul de AUC
Y = Sensitivity;
X = 1-Specificity;

figure(2)
plot(X,Y,'b--o')
pl = line([0 1],[0 1],'Color','black');
xlabel('1-Spécificité'), ylabel('Sensibilité')
grid ;

XA=[1 X 0];
YA=[1 Y 0];
for k=1:102
AUC(k)= (XA(k)-XA(k+1))*YA(k);
soma=sum(AUC);
end
AUC= soma;
```

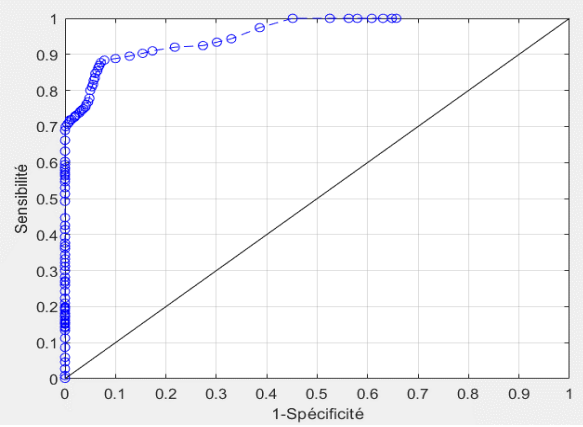


Figure A. 2 : Courbe ROC de la séquence AF042782
avec 2 exons de la base de données HMR195

```
% Trouver la valeur maximale de Approx et son indice correspondant
[maxValue, maxIndex] = max(Approx);

% Récupérer les valeurs correspondantes dans les autres tableaux
valSn = Sn(maxIndex);
valSp = Sp(maxIndex);
valF1_score = F1_score(maxIndex);
valAc = Ac(maxIndex);

% Afficher tous les valeurs paramètres sur une même ligne
fprintf('Max Approx : %.2f, Seuil : %d, Sn : %.2f, Sp : %.2f, Ac : %.2f,
F1_score : %.2f, AUC : %.2f\n ', maxValue, maxIndex, valSn, valSp, valAc,
valF1_score, AUC);
```

Résultats affichés dans la fenêtre « command window » de Matlab : Max Approx : 0.85, Seuil : 19, Sn : 0.88, Sp : 0.93, Ac : 0.93, F1_score : 0.90, AUC : 0.96

RESUME DE LA THESE

ANALYSE DES SEQUENCES D'ADN PAR DES TECHNIQUES

DE TRAITEMENT NUMERIQUES DU SIGNAL

Directeur de thèse : Professeur Ferdi Youcef

L'annotation rapide et efficace des structures génomiques en post-séquençage est un défi scientifique majeur. L'identification des zones codantes de protéines dans les gènes (exons) chez les cellules eucaryotes est alors une étape primordiale dans cette annotation. Les techniques et programmes existants sont souvent complexes en se basant sur les propriétés biologiques spécifiques des séquences d'ADN. L'ère du numérique et aussi l'avancée des techniques de traitements du signal (TNS) permettent d'apporter des nouvelles méthodes alternatives. L'ensemble de ces nouvelles méthodes constitue le « Genomic Signal Processing » (GSP). Les méthodes GSP commencent par la numérisation de la séquence d'ADN et utilisent le TNS, en exploitant la propriété « 3-base periodicity » des exons afin de les identifier. Ces méthodes sont efficaces mais nécessitent de l'amélioration. C'est dans cette perspective que ce travail de thèse est effectué en fournissant une nouvelle approche GSP basée sur le calcul d'ordre fractionnaire. La méthode proposée est un algorithme centré sur le filtrage fractionnaire afin d'identifier les exons chez les eucaryotes. L'évaluation de la méthode se fait en utilisant les critères communs en GSP tels que les bases de données génomiques et les paramètres d'évaluation statistiques. Les études comparatives faites entre les méthodes GSP existantes et la méthode proposée montrent l'efficacité de cette dernière et ouvrent de nouvelles perspectives pour cette étude sur l'identification des exons.

Mots-clés : ADN, identification des exons, eucaryotes, filtre fractionnaire, calcul d'ordre fractionnaire, 3-base periodicity, traitement du signal génomique.

DNA SEQUENCES ANALYSIS USING DIGITAL SIGNAL PROCESSING TECHNIQUES

Thesis Supervisor: Professor Ferdi Youcef

Rapid and efficient annotation of genomic structures in post-sequencing is a significant scientific challenge. Identifying protein-coding regions (exons) in eukaryotic genes is a crucial step in this annotation process. Existing techniques and programs often rely on complex methods based on DNA sequences biological specific properties. The digital era and advancements in digital signal processing techniques (DSP) offer new alternative methods. This collection of new methods is referred to as "Genomic Signal Processing" (GSP). GSP methods start by digitizing the DNA sequence and utilize DSP, exploiting the "3-base periodicity" property of exons for identification. These methods are effective but require improvement. It is in this perspective that this thesis work is carried out by providing a new GSP approach based on fractional order calculation. The proposed method is an algorithm centered on fractional filtering to identify exons in eukaryotes. The method's evaluation is carried out using common GSP criteria such as genomic databases and statistical evaluation parameters. Comparative studies between existing GSP methods and the proposed method demonstrate the effectiveness of the latter and open up new prospects for this study on exon identification.

Keywords: DNA, exon identification, eukaryotes, fractional filter, fractional-order calculus, 3-base periodicity, genomic signal processing.

تحليل سلاسل الحمض النووي باستخدام تقنيات المعالجة الرقمية للإشارات

المشرف: البروفيسور فردي يوسف

تُعتَبَرُ التسمية السريعة والفعّالة للهياكل الجينية في مرحلة ما بعد التسلسل تحديًا علميًا رئيسيًا. يُعتَبَرُ تحديدُ المناطق التي تُشفر البروتين في الجينات (الإكسونات) عند الخلايا اليوكاريوتية خطوةً حاسمةً في هذه التسمية. غالبًا ما تكون التقنيات والبرامج الموجودة معقدة، وذلك بناءً على الخصائص البيولوجية المحددة لتسلسلات الحمض النووي. يُتِيحُ العصر الرقمي إضافةً للتقدم في تقنيات معالجة الإشارات في توفير طُرُقٍ بديلةٍ وجديدةٍ، تُشكِّلُ جميع هذه الأساليب الجديدة ما يُعرَفُ بـ "معالجة الإشارات الجينومية."Genomic Signal Processing (GSP)

تَبْدَأُ طُرُقُ (GSP) بترقيم تسلسل الحمض النووي واستخدام تقنيات معالجة الإشارات، باستغلال خاصية "التواتر الثلاثي "(3-base periodicity) للإكسونات لتحديد هويتها. هذه الأساليب فعالةٌ ولكنها تحتاج إلى التحسين. إنه لهذا الغرض قمنا بتحضير هذه الأطروحة وذلك من أجل تقديم نهج جديد في مجال معالجة الإشارات الجينومية (GSP) يعتمد على حساب النظام الكسري. الطريقة المقترحة هي خوارزمية تركز على التصفية الكسرية لتحديد الإكسونات في الخلايا اليوكاريوتية. يتم تقييم الطريقة باستخدام المعايير الشائعة في معالجة الإشارات الجينومية، مثل قواعد البيانات الجينومية والمعايير الإحصائية للتقييم. تُظهِرُ الدراسات المقارنة التي أُجريت بين الطُرُق الحالية لمعالجة الإشارات الجينومية والطريقة المقترَحة فعاليةَ الأخيرة وتفتح آفاقًا جديدةً لهذه الدراسة في تحديد الإكسونات.

الكلمات المفتاحية: الحمض النووي، تحديد الإكسونات، الكائنات اليوكاريوتية، مرشح كسري، حساب النظام الكسري، التواتر الثلاثي، معالجة الإشارات الجينومية.