REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE FERHAT ABBAS SETIF-1

FACULTE DE TECHNOLOGIE

THESE

Présentée au Département d'électronique

Pour l'obtention du diplôme de

DOCTORAT EN SCIENCES

Filière : Electronique **Option : Electronique**

Par

MIHI Assia

THEME

Détection d'événements par les méthodes intelligentes dans les séquences biomoléculaires

Soutenue le 27/04/2019 devant le Jury:

Ferhat Hamida Abdelhak	**Professeur**	**Univ. F. Abbas Sétif 1**	**Président**
Boucenna Noureddine	**M.C.A.**	**Univ. F. Abbas Sétif 1**	**Directeur de thèse**
Chemali Hamimi	**Professeur**	**Univ. F. Abbas Sétif 1**	**Examinateur**
Abdessemed Foudil	**Professeur**	**Univ. M. Ben Boulaid Batna 2**	**Examinateur**
Boumehraz Mohamed	**Professeur**	**Univ. M. Khider Biskra**	**Examinateur**
Aidel Salih	**Professeur**	**Univ. B. El Ibrahimi B.B.A.**	**Examinateur**
Benmahammed Khier	**Professeur**	**Univ. F. Abbas Sétif 1**	**Invité**

Remerciements

Je tiens à remercier avant tous, le bon Dieu qui m'a donné la force et la patience tout au long de la préparation de cette thèse.

Je tiens à remercier en premier lieu Docteur Boucenna Noureddine, mon directeur de thèse, pour m'avoir encadré durant ces années. Je le remercie pour ses qualités humaines et son soutien qui m'ont permis de mener à bien ce travail.

Merci aussi au Professeur Benmahammed Khier pour m'avoir lancé dans le domaine de la bioinformatique et d'avoir accepté de faire partie des membres de mon jury. Je le remercie de m'avoir aidé durant ces années.

J'adresse ma plus vive gratitude aux membres de mon jury. Monsieur Ferhat Hamida Abdelhak, Professeur à l'université de Sétif, Monsieur Chemali Hamimi, Professeur à l'université de Sétif, Monsieur Abdessemed Foudil, Professeur à l'université de Batna, Monsieur Boumehraz Mohamed, Professeur à l'université de Biskra et Monsieur Aidel Salih, Professeur à l'université de Bordj Bou Arréridj; pour me faire l'honneur de prendre part à mon jury, je les remercie tout particulièrement pour l'attention et le temps qu'ils y ont consacré pour lire ce manuscrit.

Je remercie également l'ensemble du personnel du département d'électronique de l'université de Sétif, particulièrement, Fadhila et Hekmet.

Enfin, un grand merci à ma famille pour son soutien, en particulier à mon frère Toufik et ma sœur sana.

Dédicace.

A mes parents.

A mes sœurs Ifriquia, Europa, Leila et Sana.

A mon frère Toufik.

A toute ma famille.

Table des matières

Symboles et abréviations

Ensembles

$\mathbb{R}$ l'ensemble des réels.

$\mathbb{Z}$ l'ensemble des entiers.

$\mathbb{N}^*$ l'ensemble des naturels non nuls.

Fonctions

$\psi_{a,b}$ une ondelette de facteur d'échelle a et de paramètre de translation b.

ψ ondelette mère.

Espaces fonctionnels

$L^2(\mathbb{R})$ l'espace des fonctions continues d'une variable réelle et de carré intégrable.

Opérateurs et symboles

$\langle .,. \rangle$ le produit scalaire.

$\|.\|$ la norme.

$\bigcup$ union des espaces.

$\bigcap$ intersection des espaces.

$\oplus$ somme directe (orthogonale).

$\perp$ sous-espaces orthogonaux.

$\tilde{h}$ le filtre symétrique de h.

$\hat{\psi}$ la transformée de Fourier de ψ.

$\lim .$ la limite.

Abréviations

ANFIS Adaptative Neuro Fuzzy Inference System.

SIF Systèmes d'Inférence Floue.

ADN Acide DésoxyriboNucléique.

RNA Réseau de Neurones Artificiels.

NCBI National Center for Biotechnology Information.

Liste des figures

Chapitre 1 : La bioinformatique

Chapitre 2 : Analyse fréquentielle des signaux ADN

Chapitre 3 : La prédiction de gènes

Chapitre 4 : Résultats et discussion

Liste des tableaux

Chapitre 1 : La bioinformatique

Chapitre 4 : Résultats et discussion

Introduction générale

La biologie est définie comme l'étude de vivant depuis l'étude des interactions entre espèce et populations jusqu'aux études des fonctions des tissus et des cellules au sein d'un organisme. Dans le cadre de leurs travaux, les biologistes ont à collecter et à interpréter un grand nombre de données. Un ensemble de technologies expérimentales complexes permet de collecter des données plus vite qu'elles ne peuvent être interprétées. Avec l'importante quantité de séquence d'ADN, comment parvenir à comprendre quelles portions de celles-ci contrôlent les divers mécanismes chimiques de la vie ?

On dispose maintenant de l'information génétique exhaustive sur un nombre croissant d'organismes vivants et il est aujourd'hui possible d'aborder de manière globale un certain nombre de problèmes complexes dont on n'avait jusqu'à présent qu'une connaissance fragmentaire : voies métaboliques, interaction de la cellule avec l'extérieur, mécanismes globaux de régulation et contrôle.

L'accélération du séquençage, permise en particulier par la robotisation et la parallélisation des méthodes d'analyse, nécessite un soutien de plus en plus important de l'outil informatique. Ensuite l'informatique est un outil incontournable pour extraire et analyser l'information contenue dans ces gigabases (1 Gbase = 10^9 nucléotides) de séquence. Le volume de données à traiter est considérable. Il est clairement impossible de caractériser expérimentalement tous les gènes contenus dans ces séquences [1].

La bioinformatique, discipline récente, traite des différents aspects de ce nouveau champ de la connaissance et s'appuie bien sûr à la fois sur les concepts de la biologie et de l'informatique, et sur des outils issus de la chimie et de la physique.

La bioinformatique est un domaine de recherche qui propose et développe des modèles, des méthodes et des outils afin d'analyser l'information disponible (donnée de séquences, etc.), et produire de nouvelles connaissances pour mieux comprendre la biologie et tenter de répondre aux questions évoquées ci-dessus.

C'est un domaine de recherche multidisciplinaire qui s'appuie largement sur des développements pointus en statistiques et en reconnaissance de motifs. Les scientifiques travaillant dans ce domaine ont des origines thématiques variées parmi lesquelles les mathématiques, l'informatique,… Cependant, la biologie, productrice des données en bioinformatiques, est une science à la fois très pointue et très générale. La bioinformatique est donc un domaine dans lequel l'interprétation est source de pièges pour tous ceux qui

recherchent des motifs et font des prédictions, sans avoir une complète compréhension de l'origine de données, ni de ce qu'elles signifient. En mettant à la disposition des chercheurs en biologie des bases de données, des interfaces utilisateur et des outils informatiques et statistiques, la bioinformatique permet de comparer des séquences d'ADN et de gérer des résultats qui sont potentiellement significatifs. Les outils bioinformatiques peuvent donner également à l'utilisateur l'occasion d'interpréter de manière excessive des données et de fournir un sens à ce qui n'en pas réellement [2].

L'un des premiers aspects de la bioinformatique des séquences est celui qui traite du séquençage, c'est-a-dire connaitre l'enchainement complet des bases nucléotidiques qui constituent un génome. En effet, d'une part, les suites de nucléotides sont perçues comme des mots appartenant au langage génétique dont il faut décider s'ils correspondent ou non a des gènes. Pour déterminer si une séquence est codante, on peut utiliser des outils informatiques de prédiction capables d'identifier un gène selon plusieurs critères comme la composition en bases. Divers logiciels ont été développés pour reconnaitre les gènes sur la base d'informations compilées manuellement et regroupées en bases de données, sources de vérifications et comparaisons. Ces logiciels sont fondés sur l'utilisation de méthodes informatiques et mathématiques diverses comme les réseaux de neurones, l'analyse discriminante ou encore les méthodes probabilistes utilisant des chaines de Markov.

En plus de ces techniques de prédiction de gènes, les approches basées sur l'informatique souple (soft computing) ont gagné en popularité ces derniers temps. Les outils de prédiction de gènes basés sur des techniques de calcul souple ont de bonnes performances. Ces techniques, en particulier les réseaux neuronaux, apparait comme un outil puissant dans la prédiction de gènes. En effet, les techniques hybrides donnent des résultats prometteurs, mais ils sont appliqués de manière très limitée [3].

La capacité d'apprentissage et d'adaptation des réseaux neuronaux artificiels (RNA) et la représentation des connaissances à travers des logiques floues lorsqu'elles sont réunies ensemble, on peut exploiter les avantages de chacun d'eux [4].

Cependant, la théorie des ensembles flous joue un rôle important dans le traitement des connaissances imprécises lors de la prise de décisions en bioinformatique et en biologie computationnelle [5,6]. Par conséquent, les ensembles flous ont attiré l'attention de plus en plus et l'intérêt pour la reconnaissance de formes, l'analyse de données, etc. La logique floue a été largement utilisée pour fournir de la flexibilité aux algorithmes classiques. La gestion de l'incertitude est un point clé de cette approche. L'incertitude joue un rôle majeur en biomédecine et en sciences biomédicales car la plupart des recherches effectuées dans ce

domaine sont expérimentales et sont affectées par les erreurs associées aux mesures. Les approches floues ont été explorées par la bioinformatique dans une certaine mesure [5].

Le système neuro-flou est basé sur un système flou entraîné par l'algorithme d'apprentissage de RNAs. Il combine la capacité de raisonnement inexacte de la théorie des ensembles flous et l'auto-adaptation et la capacité d'organisation de RNA pour obtenir une structure computationnelle plus puissante. L'approche neuro-flou, en d'autres termes, combine la capacité des ensembles flous à modéliser des systèmes vaguement définis, et la capacité d'apprentissage d'un RNA. Le système neuro-flou résultant peut être entraîné en utilisant les données disponibles pour régler les fonctions d'appartenance. C'est une méthode largement préférée dans les études d'exploration de données dans le but de la prédiction [7]. Les systèmes neuro-flous sont généralement représentés comme des RNAs multicouches spéciaux, tels que les systèmes d'inférence neuro-flous adaptatifs ANFIS (Adaptative Neuro Fuzzy Inference System).

Dans cette thèse, nous nous sommes penchés particulièrement sur le problème de la prédiction de gène. Pour ce faire, une approche basée sur un system prédictif neuro-flou adaptatif (ANFIS) a été proposée pour la prédiction de séquences d'ADN. De ce point de vue, les techniques d'analyse du signal tels que la transformée de Fourier, les ondelettes et les bancs de filtres ont été utilisés. L'application de tels outils d'analyse spectrale à des signaux issus de codages nucléotidiques des séquences ADN apparaît utile pour mettre en évidence des séquences particulières dans la structure nucléosomale de l'ADN et plus généralement l'extraction d'informations structurelles.

Parmi les techniques permettant d'analyser le signal d'ADN dans le domaine des fréquences, la plus connue est la transformée de Fourier. La transformée de Fourier est l'un des outils centraux du traitement du signal. Récemment, la transformée en ondelettes a été introduite avec un certain succès pour l'étude des séquences d'ADN. Déjà utilisée dans de nombreux domaines des sciences fondamentales et appliquées, la transformée en ondelettes apparaît très bien appropriée pour l'étude des propriétés fractales de certains processus.

Afin de réaliser des traitements locaux sur des données (ou signaux) contenant des éléments informatifs dans certaines bandes de fréquence, nous nous intéressons dans cette thèse à l'étude de bancs de filtres. Ces outils permettant de représenter efficacement certaines informations structurées dans les données et qui peuvent s'appliquer dans de nombreuses situations telles que l'extraction d'informations.

Nous poursuivrons ici cette démarche par l'analyse de codages structurels pour la mise au point d'une méthodologie de prédiction des gènes. En effet, Les séquences nucléotidiques

extraites d'une base de données sont riches de redondances et de biais statistiques. L'étude de ces biais est extrêmement informative, car elle renseigne sur les phénomènes qui en sont responsables, elle permet de mieux comprendre la façon dont la cellule vivante exploite son information génétique. Une fois que les mécanismes impliqués sont caractérisés, l'analyse et la recherche systématique de ces biais dans de nouvelles séquences deviennent alors des outils précieux pour effectuer des prédictions sur leurs propriétés [1]. Ainsi, la base de la plupart des méthodes de reconnaissance de régions codantes est les biais de position et de composition imposés à la séquence d'ADN dans les régions codantes par le code génétique et par la distribution des acides aminés dans les protéines [8].

Cette thèse s'organise en quatre chapitres :

Le premier chapitre présente une introduction à la bioinformatique et ces différents aspects. Le chapitre regroupe aussi les notions de biologie moléculaire nécessaires à notre étude. Nous débutons par une introduction générale aux séquences biologiques, en partant de l'ADN, jusqu'à la construction d'une protéine.

Le deuxième chapitre formalise la notion de codage nucléotidique et présente les concepts mathématiques nécessaires à l'étude de signaux issus de codages nucléotidiques. Le chapitre introduit la transformée de Fourier, la transformée en ondelettes et les bancs de filtres ; et présente la manière dont ceux-ci peuvent être utilisées pour effectuer une étude de type spectral d'un signal.

Par la suite, le chapitre 3 se concentre sur un phénomène bien précis, celui de prédiction de gène. Il s'agit d'un domaine abondamment étudié en bioinformatique et pour lequel des algorithmes performants ont été développés au moyen d'un formalisme très complet. Ceux-ci seront abordés dans ce chapitre où nous verrons les différentes techniques provenant du domaine de l'apprentissage machine (réseaux de neurones artificiels et modèles de Markov caché) qui permettront de combler les lacunes des techniques d'alignement dans le problème qui nous intéresse. Les réseaux de neurones avec leur capacité à apprendre présentent de meilleures performances lorsqu'ils sont associés à la logique floue. Notre choix se porte ainsi sur des réseaux neuro-flous adaptatifs (ANFIS). Dans ce chapitre nous présentons un algorithme évolutif pour apprendre les paramètres d'un tel réseau.

Le quatrième chapitre de ce manuscrit présente l'essentiel des résultats obtenus lors de ce travail de thèse concernant l'étude des séquences ADN par l'intermédiaire de codages nucléotidiques. Les concepts théoriques et les méthodologies développées dans les trois premiers chapitres vont nous permettre de réaliser les simulations numériques.

L'interprétation de ces résultats et la mise au point de la méthodologie de prédiction proposée sont effectuées dans ce chapitre.

En conclusion, nous résumons les principaux résultats de notre travail et nous abordons enfin une discussion sur les perspectives de travail qui découlent de cette thèse.

Chapitre 1

La bioinformatique

1.1. Introduction :

L'informatique, qu'elle soit considérée comme une science ou une technologie, tient une place croissante dans le développement des recherches en biologie. Il suffit pour s'en convaincre de considérer le grand nombre de revues, de conférences et plus généralement de publications à la frontière de l'informatique et de la biologie depuis une quinzaine d'années. Au delà de ce simple constat, peut-être est-il intéressant de comprendre et d'expliquer pourquoi. Pour ce faire, il convient de se plonger dans l'histoire de ces deux disciplines scientifiques et d'en trouver les points de liaison caractéristiques.

1.2. Qu'est ce que la bioinformatique ?

La bioinformatique, discipline en évolution permanente, est l'application d'outils et de techniques informatiques et mathématiques à la gestion et à l'analyse des données biologiques. Le terme bioinformatique est relativement récent et, tel qu'il est défini ici, il empiète sur d'autres termes comme biologie computationnelle, biologie *in silico* (grâce au silicium des microprocesseurs) ou d'autres expressions de ce genre [9].

1.3. Biologie et informatique :

L'un des aspects les plus stimulants, lorsque l'on travaille en informatique et en biologie, est de constater à quel point ces disciplines sont riches en nouvelles techniques et en nouveaux résultats.

La biologie est une science déjà ancienne relativement à l'informatique. La génétique par exemple, qui occupe aujourd'hui une place centrale dans les sciences de la vie, est née il y a un siècle grâce aux premières études des lois de l'hérédité par le moine Gregor Mendel. La découverte fondamentale de la structure de l'acide désoxyribonucléique (ADN) et la première mise en évidence de la structure d'une protéine datent elles des années 1950. Comme dans de nombreux domaines scientifiques, de nouveaux axes de recherche en biologie, reposent aujourd'hui sur des techniques et des concepts plus récents. La dernière décennie a vu le lancement et l'aboutissement du Projet Génome Humain, qui aidera à déterminer la position et la nature de tous les gènes et bien plus encore. Nous sommes actuellement dans l'âge d'or

de la recherche en biologie, un moment important du point de vue médical, scientifique et philosophique de l'histoire humaine [9].

En comparaison, l'informatique est une science relativement récente. Depuis le premier ordinateur l'informatique, tout comme la biologie, n'a cessé d'évoluer depuis. Tout, de nos jours, depuis nos communications jusqu'à notre agriculture en passant par le monde de la finance, est intimement lié aux ordinateurs et à leur programmation. L'ordinateur est devenu la principale métaphore pour expliquer un grand nombre de choses. De nombreuses problématiques en biologie de la cellule, de l'organisme ou des populations s'appuient sur des méthodes informatiques pour proposer et tester des hypothèses.

Réciproquement, de remarquables découvertes en biologie ont trouvé un écho en informatique, les programmes capables d'évoluer dits génétiques ou les réseaux neuronaux en sont des exemples. L'échange d'idées et de concepts entre la biologie et l'informatique est, en soi, une incitation à la découverte.

1.4. La révolution biologique :

La révolution biologique a été initiée par quatre événements majeurs : la découverte de l'ADN en tant que support de l'information génétique (1944) et de sa structure en double hélice (1953), ainsi que la mise en place du dogme central de la biologie moléculaire (1958) et le déchiffrage du code génétique (1962). Ces événements constituent les fondements de la génomique. Le tableau 1.1 regroupe les événements majeurs qui sont intervenus en biologie (incolore), en informatique (en vert) et en bioinformatique (en bleu) depuis 1944 [10].

Tableau 1.1. Chronologie non exhaustive des événements marquants survenus en biologie, en informatique et en bioinformatique.

Année	Auteurs	Evénement
1944	Avery	Démonstration de l'ADN en tant que support de l'information génétique
1946		ENIAC, premier ordinateur totalement électrique et Turing-complet
1947	Bell Labs	Invention du transistor
1951		UNIVAC, premier ordinateur commercialisé
1953	Sanger, Thompson	Détermination de la séquence des chaines A et B de l'insuline
	Watson, Crick	Modèle de la structure en double hélice de l'ADN
1956	IBM	Premiers disques durs commercialisés (5 Mo)
1958	Crick	Enonciation du dogme central de la biologie moléculaire
	Texas instruments	Réalisation du premier circuit intégré

1962	Matthaei	Déchiffrage du code génétique
1965	Dayhoff	Premier atlas de la séquence et de la structure des protéines
1967	Fitch	Construction d'arbres polygénétiques
1969	US DoD	ARPANET, premier réseau d'informatique permanent
	Bell Labs	Naissance du future Unix
1970	Smith, Wilcox	Première enzyme de restriction spécifique isolée
	Needleman, Wunsch	Algorithme d'alignement global optimal entre deux séquences
1971	Intel	Premier microprocesseur (Intel 4004)
1973		Annonce de la Protein Data Bank (PDK)
1974	Chou, Fasman	Algorithme de prédiction des structures secondaires des protéines
1977	Sanger	Méthode de séquençage par synthèse enzymatique de l'ADN
	Sanger, Air	Premier génome complet séquencé : phage ϕX174
	Maxam, Gilbert	Méthode de séquençage chimique de l'ADN
	Staden	Suite d'analyse de séquences d'ADN Staden
1980	EMBL	Première banque de séquences nucléiques
1981	Anderson	Séquence du génome mitochondrial humain
	Smith, Waterman	Algorithme d'alignement local optimal entre deux séquences
	IBM	Commercialisation du premier PC (IBM PC 5150)
1982	GenBank	Banque Américaine de séquences nucléiques
		Naissance du réseau des réseaux : Internet
	Commodore	Commercialisation du 1er ordinateur : Commodore 64
1983	Mullis	Invention de la réaction en chaine de la polymérase (PCR)
1985	Lipman, Pearson	FASTA, premier programme de recherche de séquences par similarité
	Gouy	ACNUC, programme d'interrogation des banques de séquences
	Philips, Sony	Invention du CD-Rom (Compact Disc Read Only Memory) (650 Mo)
1986	Bairoch	SWISS-PROT, Banque de séquences protéiques
	DDBJ	Banque japonaise de séquences nucléiques
1987	Applied Biosystems	Premier séquenceur automatique (ABI 370)
	Burke	Création du vecteur de clonage Yeast Artificial chromosome (YAC)

	Kulesh	Apparition de la technologie des puces à ADN
1988		Lancement du projet international de séquençage du génome humain
	Higgins, Sharp	CLUSTAL, programme d'alignement multiple
1989	O'Connor	Création du vecteur de clonage Bacterial Artificial Chromosome (BAC)
	CERN	Invention du World Wide Web et du langage HTML
1990	Altschul	BLAST, programme de recherche de séquences par similarité
1991	Adams	Création et utilisation à grande échelle du séquençage partiel d'ADNc (EST)
	Roberts	GRAIL, programme de localisation de gènes
1993	Etzold, Argos	SRS, programme d'interrogation des banques de séquences
	NCSA	Premier navigateur Web : Mosaic
1994	Thompson	CLUSTALW
1995	Fleischman	Premier organisme vivant séquencé : *Haemophilus influenzae* (1.8 Mb)
	Fraser	Plus petit organisme vivant séquencé : *Mycoplasma genitalium* (580 Kb)
	Jong, Brenner	Création de la librairie biologique open source BioPerl
	DVD Forum	Spécifications du DVD (Digital Versatile Disc) (8.5 Go)
1996	Walsh, Barrell	Premier organisme eucaryote séquencé : *Saccharomyces cerevisiae* (12.1 Mp)
	Affymetrix	Commercialisation de la première puce à ADN
1997	Blattner	Génome complet d'*Escherichia coli* (4.7 Mb)
	Altschul	Gapped BLAST et PSI- BLAST
	Burge, Karlin	GenScan, prédiction de la structure complète des gènes dans l'ADN génomique humain
1998		Premier organisme pluricellulaire séquencé : *Caenorhabditis elegans* (97 Mp)
2000	Dennis, Surridge	Génome d'*Arabidopsis thaliana* (100 Mp)
	Adams	Génome de *Drosophila melanogaster* (180 Mb)
	AMD	Premier microprocesseur x86 atteignant 1 GHz
2001	Lander	Publication préliminaire du génome humain par le *Human Genome Project* (2.9 Gb)
	Venter	Publication préliminaire du génome humain par *Celera Genomics* (2.9 Gb)
	Ensembl	Navigateur de génome Ensembl
	NCBI	Navigateur de génome du NCBI

2002	Waterson	Séquence préliminaire du génome de la souris (2.5 Gb)
	UCSC	Navigateur de génome de l'UCSC
2004	IHGSC	Finalisation du génome humain (3.2 Gp)
	ENCODE PC	ENCODE, projet d'identification de tous les éléments fonctionnels du génome humain
	Blu-ray Disc Association	Spécifications du Blu-ray Disc (50 Go)
2005	Roche, 454	Séquenceur automatique haut-débit de 2ème génération par pyroséquençage : GS20
	AMD	Premier processeur x86 double coeurs
2006	Folding@Home	Utilisation de la puissance des processeurs de flux des GPGPU (nouvelles cartes graphiques) des PC et du processeur Cell BE des PlayStation3 pour accélérer jusqu'à 30x les calculs de dynamique moléculaire
2007	Illumina, Solexa	Séquenceur automatique haut-débit de 2ème génération par synthèse microfluidique : Genome Analyser
	Applied Biosystems	Séquençage automatique haut-débit de 2ème génération par ligation : système SOLID
	Hitachi	Premier disque dur atteignant 1To de capacité
	HVD Forum	Spécifications du HVD (Holographic Versatile Disc) (200 Go)
2008	Helicos	Séquenceur automatique haut-débit de 2ème génération par synthèse sans pré-amplification
2009…		Prés de 1000 projets de séquençage de génomes eucaryotes en cours…

La bioinformatique, dont Margaret O. Dayhoff parmi les pionniers grâce à son atlas de la séquence et de la structure des protéines (1965), va rapidement s'avérer indispensable, notamment dans la gestion des données, aboutissant naturellement aux premières banques de données biologiques dont la PDB (2000).

C'est à partir de 1977, avec l'apparition de deux nouvelles approches de séquençage de l'ADN que la bioinformatique prend réellement son envol. La production de séquences par ces méthodes est l'occasion de créer les nouvelles banques de données EMBL (2009) et GenBank (2009) afin de répertorier ces séquences nucléiques, et de développer de nouveaux algorithmes permettant de traiter les données biologiques. Ces derniers ont abouti aux outils majeurs de la bioinformatique que sont FASTA (1988), CLUSTALW (1994) et BLAST (1997).

L'outil de recherche de similitude de séquences, Basic Local Alignement Search Tool, ou BLAST, est sans aucun doute celui qui est le plus utilisé [11]. Le programme BLAST sert à la recherche, dans de grandes bases de données de séquences moléculaires, des séquences qui

présentent des régions de similitude avec la séquence entrée, fournie par l'expérimentateur. Les algorithmes d'alignement local tentent de trouver dans des paires de séquences des régions isolées qui ont un haut degré de similitude. BLAST cherche alors des régions de la séquence cible ayant des similitudes avec la séquence soumise par l'utilisateur.

Avec l'apparition des séquenceurs automatiques et de nouveaux outils de biologie moléculaire, la production des séquences s'accélère et l'on voit apparaître des projets de séquençage de génomes complets qui aboutissent vers la fin du siècle.

Après dix années d'efforts, l'arrivée des premières séquences préliminaires du génome Humain (2001) marque la fin de l'ère génomique et l'entrée dans l'ère post-génomique. Cependant, il a fallu attendre jusqu'en 2004 pour obtenir de la part de l'International Human Genome Sequencing Consortium une version que l'on peut considérer comme finalisée (2004).

Actuellement, grâce aux techniques de séquençage haut-débit, les projets de séquençage se sont multipliés de sorte que la communauté scientifique a accès à 1059 génomes complets et publiés [10].

1.5. Séquençage de l'ADN :

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné. La séquence d'ADN contient l'information nécessaire aux êtres vivants pour survivre et se reproduire. Déterminer cette séquence est donc utile aussi bien pour les recherches visant à savoir comment vivent les organismes que pour des sujets appliqués. En médecine, elle peut être utilisée pour identifier, diagnostiquer et potentiellement trouver des traitements à des maladies génétiques et à la virologie. En biologie, l'étude des séquences d'ADN est devenue un outil important pour la classification des espèces.

1.5.1. Séquençage haut débit :

On désigne par séquençage haut débit (HTS pour *high-throughput sequencing*) aussi appelé NGS pour *next-generation sequencing* un ensemble de méthodes apparues à partir de 2005 produisant des millions de séquences en un *run* et à faibles coût. Le pyroséquençage appartient à ces nouvelles techniques. Elles se caractérisent par l'utilisation d'approches massivement parallèles, permettant de séquencer des centaines de milliers de fragments simultanément. Elles s'affranchissent des étapes de clonage et de constitution de banques génomiques. Elles permettent de séquencer à partir de molécules uniques d'ADN.

1.5.2. Séquenceur d'ADN :

Un séquenceur de gènes est un appareil capable d'automatiser l'opération de séquençage de l'ADN. Un séquenceur sert à déterminer l'ordre des bases nucléiques d'un échantillon d'ADN et à le présenter, après traitement, sous forme d'une suite de lettres, appelée read ou *lecture*, représentant des nucléotides. Les grands projets de séquençage, tels ceux de déchiffrage de génomes entiers, ne sont concevables que s'il existe des appareils permettant d'augmenter la productivité des agents humains. On peut considérer certains séquenceurs comme des appareils optiques, vu qu'ils analysent les signaux lumineux émis par des fluorochromes fixés aux nucléotides.

Le principe de la réaction de séquençage utilisée dans les séquenceurs est dérivé de celui de la méthode de Sanger. Il se fonde toujours sur l'utilisation de di-désoxy-nucléotides (dd-NTP), mais peaufiné par l'utilisation de marqueurs fluorescents à la place de marqueurs radioactifs.

Les instruments les plus modernes de séquençage automatique de l'ADN sont capables de lire jusqu'à 384 échantillons marqués à la fluorescence d'un coup (*run*) et réaliser jusqu'à 24 *runs* en une journée. Ces instruments n'effectuent que la séparation des brins et la lecture des pics ; les réactions de séquençage, la purification et la resuspension dans un tampon approprié doivent se faire séparément, le plus souvent à la main.

Figure 1.1. Plusieurs séquenceurs d'ADN (Source : http://fr.wikipedia.org/).

1.6. Alignement de séquences :

Comparer des séquences par alignement permet de déterminer leur degré de similarité. Aligner deux séquences, c'est rechercher le maximum d'appariements entre les lettres qui les composent (nucléotides) [12].

Les scénarios :

- **Alignement global :** Alignement de deux séquences sur la totalité de leur longueur.
- **Alignement local :** Alignement de deux séquences sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similitude. Cette propriété en fait un outil idéal, rapide et efficace, de recherche dans les bases des données en comparant une séquence inconnue avec les séquences de la banque.
- **Alignement multiple :** alignement portant sur plusieurs séquences à la fois et dans leur intégralité. Il permet de mettre en évidence les relations entre séquences en comparant les séquences 2 à 2.

1.7. Les banques de données biologiques :

L'origine des banques de données biologiques remonte à l'utilisation des premiers ordinateurs par des cristallographes ou des biochimistes. Parmi ceux-ci, Margaret Dayhoff, biochimiste américaine, fut la première à voir l'intérêt de rassembler toutes les données sur les séquences des protéines afin d'étudier leurs relations évolutives et de les classer en familles. Elle publia le premier atlas de protéines contenant la séquence et la structure de 65 d'entre elles (*Atlas of Protein Sequence and Structure*) en 1965. Cet atlas fut périodiquement mis à jour et diffusé sur papier jusqu'en 1978. Distribué sur support magnétique à partir de 1978, il est désormais disponible en ligne depuis 1981, via Internet (Margaret Dayhoff). Cet atlas, de plus en plus volumineux, est devenu, en 1984, la banque de données P.I.R. (*Protein Information Resource*) de la National Biomedical Research Foundation (N.B.R.F.), la première concernant les protéines et qui reste une référence pour leur analyse. Elle contenait, en 2004, quelque 283 000 séquences protéiques qui totalisaient 96 millions d'acides aminés.
Parallèlement et de manière concertée, deux banques de données de séquences nucléiques ont pris leur essor de chaque côté de l'Atlantique, en 1982. Aux États-Unis, la GenBank a pris corps au L.A.N.L. (Los Alamos Nuclear Laboratory) avec Doug Brutlag et Temple Smith ; depuis 1987, elle est gérée et distribuée par le N.C.B.I. (National Center for Biotechnology Information). La banque nucléique européenne, quant à elle, a pris le nom du laboratoire au sein duquel elle a été développée à Heidelberg (Allemagne) : E.M.B.L. (European Molecular Biology Laboratory). Depuis 1997, une antenne spéciale pour l'informatique a été créée à Cambridge : l'E.B.I. (European Bioinformatics Institute), pour poursuivre le développement de la banque E.M.B.L [13].
Étant donné le travail considérable que représente le maintien de ces deux banques, les organismes ont décidé de joindre leurs efforts.

1.8. Organisation des séquences biomoléculaires :

1.8.1. Structure des gènes :

L'unité la plus petite de la vie est la cellule. Les organismes peuvent être unicellulaires, comme les bactéries, de nombreux champignons et algues, ou pluricellulaires, comme les végétaux et les animaux. Toutes les caractéristiques d'un organisme sont déterminées par la structure et la fonction des cellules qui le composent.

Le jeu complet d'ADN d'un organisme est appelé un génome. Un génome est composé de longues molécules d'ADN, qui sont à leur tour les principaux composants des chromosomes. Chaque chromosome contient une molécule d'ADN portant de nombreux gènes. Les génomes de la plupart des organismes procaryotes sont constitués d'un seul chromosome, tandis que les génomes des Eucaryotes comportent plusieurs chromosomes. Chez les Eucaryotes, la plupart des chromosomes sont situés dans le noyau, mais les mitochondries et les chloroplastes contiennent chacun un type unique de chromosome [14].

1.8.2. Eléments de construction de l'ADN :

L'ADN comporte trois types de composants chimiques : du phosphate, un sucre appelé désoxyribose et quatre bases azotées- l'adénine (A), la guanine (G), la thymine (T) et la cytosine (C). Deux de ces bases, l'adénine et la guanine possèdent une structure à deux cycles, caractéristique d'une substance chimique appelée purine. Les deux autres bases, la cytosine et la thymine, ont une structure à un seul cycle, d'un type appelé pyrimidine.

Les composants chimiques de l'ADN sont organisés en groupes appelés nucléotides, composés chacun d'un groupement phosphate, d'une molécule de désoxyribose et de l'une des quatre bases. La figure 1.2 présente la structure des quatre nucléotides de l'ADN.

L'ADN est une double hélice :

L'ADN est composé de deux chaînes (brins) appariées de nucléotides, enroulées en une double hélice. Les deux brins nucléotidiques sont maintenus ensemble par de faibles associations entre les bases de chaque brin formant une structure semblable à un escalier en colimaçon (Figure 1.3). Le squelette de chaque brin est un polymère sucré répété de phosphate et de désoxyribose. Les liaisons sucre-phosphate du squelette sont appelées liaisons phosphodiester.

Les carbones des groupements sucrés sont numérotés de 1' à 5'. Une partie de la liaison phosphodiester est établie entre le phosphate et le carbone 5' du désoxyribose, tandis que l'autre se trouve entre le phosphate et le carbone 3' du de désoxyribose. On dit donc de chaque squelette sucre- phosphate qu'il a une polarité 5'-3'.

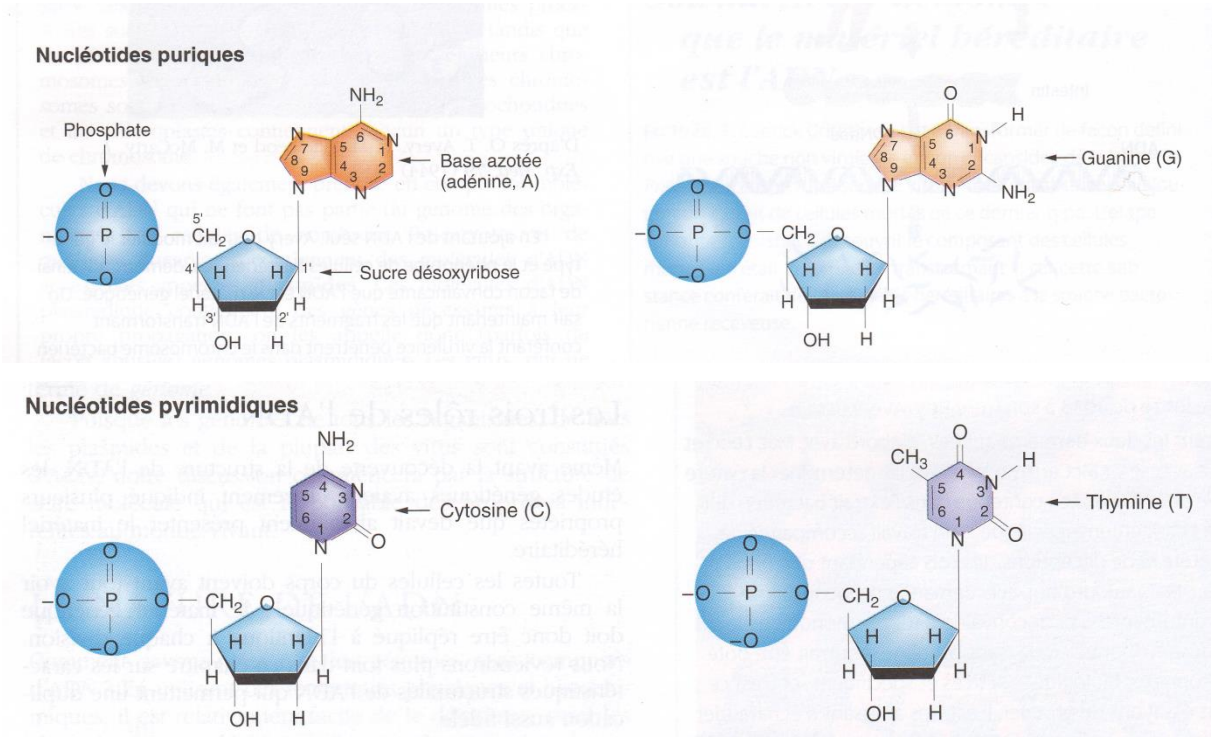
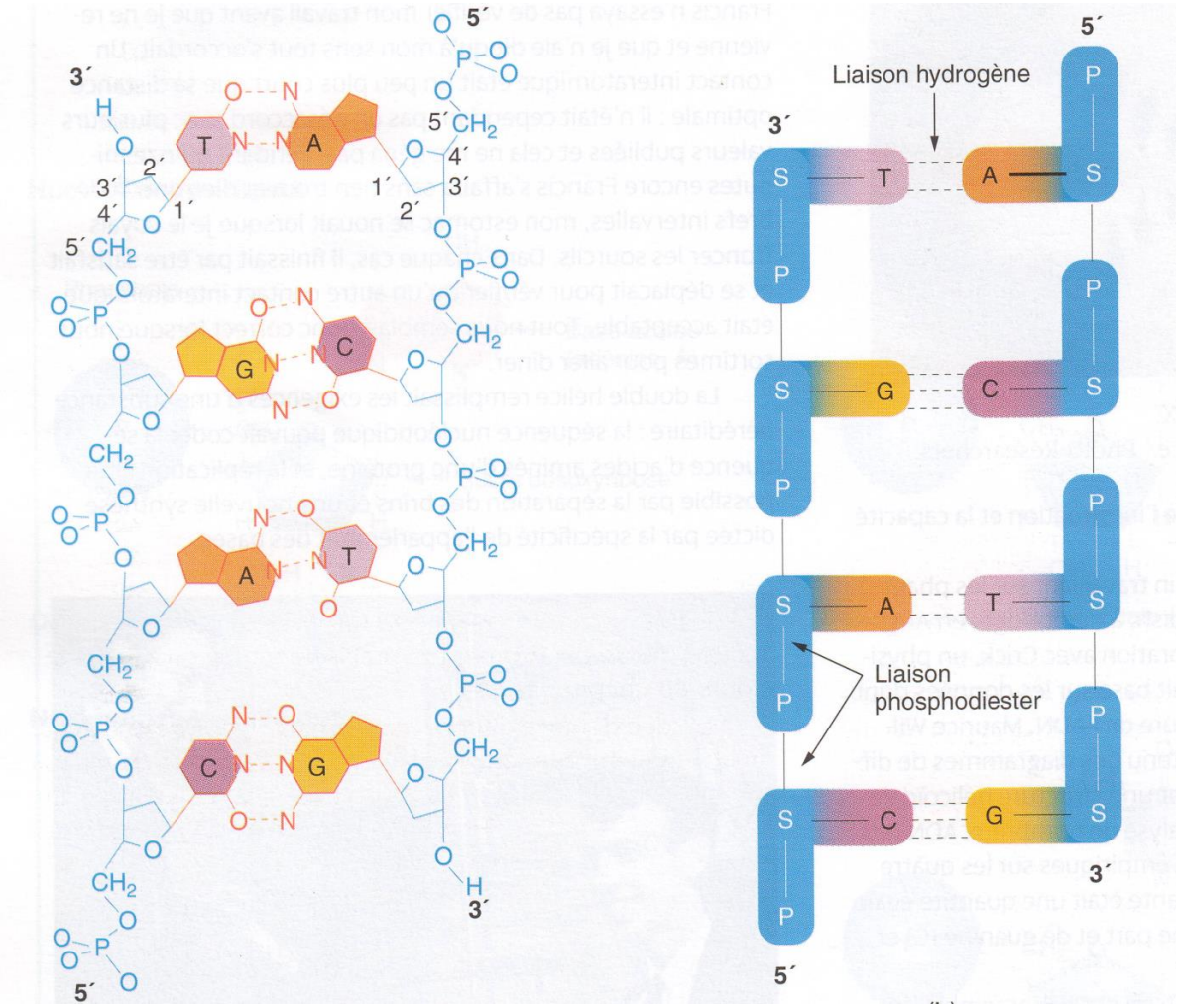


Figure 1.2. Structure chimique des quatre nucléotides.

Dans la molécule d'ADN double brin, les deux squelettes sont en sens opposé ou antiparallèles, comme le montre la figure 1.3. L'un des brins est orienté de 5' en 3' (5'→3'), de gauche à droite. L'autre brin est orienté de 5' en 3' de droite à gauche, autrement dit de 3' en 5' de gauche à droite (3'→ 5').

Les bases de l'ADN interagissent selon une règle simple : il existe seulement deux paires de bases : A .T et G . C. Les bases de ces deux paires sont dites complémentaires l'une de l'autre. Les bases de ces deux paires sont dites complémentaires l'une de l'autre. Ceci signifie que sur n'importe quelle 'marche' de l'escalier d'ADN en colimaçon, les seules associations de bases qui puissent exister entre les deux brins sans déformer fondamentalement la molécule d'ADN double brin sont A .T et G . C.

L'association de A avec T et de G avec C se fait par l'intermédiaire de liaisons hydrogène. La paire G . C possède trois liaisons hydrogène, alors que la paire A .T n'en contient que deux.



(a) (b)

Figure 1.3. L'organisation des composants de l'ADN.

(a) Un schéma précis des constituants chimiques montrant le squelette sucre-phosphate en bleu et les liaisons hydrogène entre les bases au centre de la molécule.

(b) Une version simplifiée du même segment, soulignant l'arrangement antiparallèle des nucléotides, qui sont représentés sous la forme de structures en L avec des « orteils » de phosphate en 5' et des « talons » de sucre en 3'.

La structure de l'ADN reflète sa fonction :

Comment la structure de l'ADN remplit-elle les exigences d'une molécule héréditaire? En premier lieu, la duplication. Chaque brin sert de matrice (guide d'alignement) pour la synthèse de son brin complémentaire. Si par exemple, un brin a la séquence AAGGCTGA (lue dans le sens 5' vers 3'), alors on sait automatiquement que son brin complémentaire aura la séquence TTCCGACT (dans le sens 3' vers 5'). Les deux brins d'ADN se séparent et chacun sert de matrice pour la construction d'un nouveau brin complémentaire. La deuxième exigence

concernant l'ADN est d'avoir un contenu informationnel. Ce besoin informationnel est rempli par sa séquence nucléotidique, qui ressemble à un langage écrit.

1.8.3. Principales régions d'un gène :

La taille et la fonction des gènes sont diverses. Cependant, on peut définir pour la plupart des gènes certaines caractéristiques topographiques.

Le gène doit non seulement pouvoir être transcrit en un ARN fonctionnel mais que cet ARN doit être également être fabriqué au bon endroit et au bon moment lors du développement de l'organisme. C'est seulement dans ce cas qu'un gène est parfaitement fonctionnel. Pour que cela soit possible, à l'une des extrémités du gène se trouve une région régulatrice, un segment d'ADN constitué d'une séquence nucléotidique spécifique qui lui permet de recevoir et de répondre à des signaux provenant d'autres parties du génome ou de l'environnement. En dernier lieu, ces signaux d'activation sont convertis en protéines régulatrices qui se fixent à la région régulatrice du gène et initient la transcription dans la région adjacente, qui code l'ARN. A l'autre extrémité du gène se trouve une région qui contient les signaux pour terminer le transcrit.

Il existe deux types de gènes, ceux-ci qui codent des protéines (la majorité) et ceux qui codent des ARN fonctionnels.

De nombreux gènes Eucaryotes contiennent de mystérieux segments d'ADN appelés introns, intercalés dans la région transcrite du gène. Les introns ne contiennent aucune information de séquence concernant les produits fonctionnels des gènes, tels les protéines. Ils sont transcrits en même temps que les régions codantes (appelées exons), mais ils sont ensuite excisées du transcrit initial.

Les Eucaryotes sont les organismes dont les cellules possèdent un noyau délimité par une membrane. Les animaux, les plantes et les champignons sont tous des Eucaryotes, à l'intérieur du noyau se trouvent les chromosomes.

Le génome de la plupart des Procaryotes est contenu dans le seul chromosome. Dans presque tous les cas, ce chromosome est une double hélice d'ADN circulaire fermée. Il existe quelques exceptions, telles que la bactérie *Borrelia burgdorfei*, chez laquelle le chromosome est une double hélice unique d'ADN linéaire.

Les gènes bactériens sont assez proches les uns des autres, avec assez peu d'espace intergénique et les introns sont extrêmement rares.

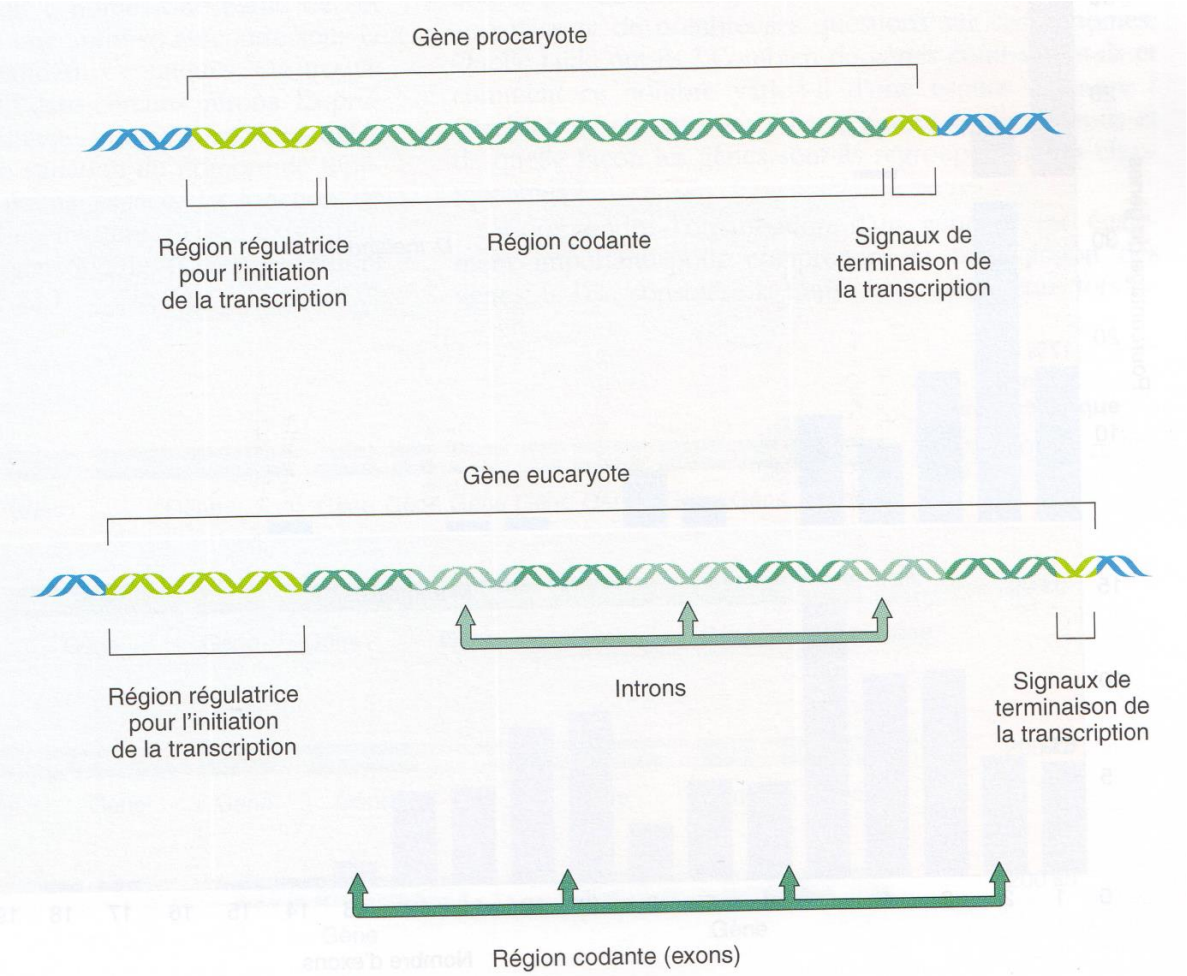


Figure 1.4. La structure générale d'un gène chez les Procaryotes et les Eucaryotes.

1.8.4. Gènes et les ARN :

L'ARN est fabriqué au cours d'un processus qui copie la séquence nucléotidique de l'ADN, la synthèse de l'ARN est appelée transcription. L'ADN est considéré comme le manuel d'instructions nécessaires pour produire tous les ARN dont la cellule a besoin.

Bien que l'ARN et l'ADN soient tous deux acides nucléiques, l'ARN est différent sur plusieurs points importants :

1. L'ARN est une chaîne nucléotidique simple brin et non une double hélice.
2. Le sucre contenu dans les nucléotides de l'ARN est le ribose et non le désoxyribose.
3. Les nucléotides de l'ARN portent les bases adénine, guanine et cytosine, mais la base pyrimidique uracile (abrégée en U) remplace la thymine. Toutefois, l'uracile forme des liaisons hydrogène avec l'adénine, exactement comme la thymine.

1.8.5. Classes d'ARN :

Les ARN peuvent être regroupés en deux grandes classes. Certains ARN sont des intermédiaires dans le processus de décodage des gènes en chaînes polypeptidiques. Nous appellerons ces ARN, des ARN 'informationnels'. Dans l'autre classe, c'est l'ARN lui-même qui est le produit final fonctionnel.

Les ARN informationnels :

Pour la plupart des gènes, l'ARN est seulement un intermédiaire dans la synthèse du produit fonctionnel final, qui est une protéine. L'ARN informationnel de cette majorité de gènes est toujours un ARN messager (ARNm).

La séquence des nucléotides dans l'ARNm est convertie en une séquence d'acides aminés appartenant à une chaîne polypeptidique, par un processus appelé traduction.

Les ARN fonctionnels :

Il est important de souligner que l'action de ces ARN fonctionnels se situe exclusivement au niveau des ARN ; ils ne sont jamais traduits en polypeptides.

Les principales classes d'ARN fonctionnels interviennent dans différentes étapes du traitement informationnel de l'ADN en protéine. Deux de ces classes d'ARN fonctionnels se rencontrent chez tous les organismes :

- Les molécules d'ARN de transfert (ARNt) agissent comme des transporteurs qui apportent des acides aminés à proximité de l'ARNm au cours du processus de traduction (synthèse protéique).
- Les ARN ribosomaux (ARNr) sont des composants des ribosomes. Ceux-ci sont de gros ensembles macromoléculaires qui servent de guides pour coordonner l'assemblage de la chaîne d'acides aminés d'une protéine.

Il existe deux types de gènes, ceux-ci qui codent des protéines (la majorité) et ceux qui codent des ARN fonctionnels.

1.8.6. La traduction utilise un brin d'ADN comme matrice :

Les deux brins d'ADN de la double hélice se séparent localement et l'un des brins séparés sert de matrice (de guide) pour la synthèse d'ARN. Pour chaque gène, un seul brin est utilisé, et il s'agit toujours du même brin pour ce gène (figure 1.5).

Ensuite, des nucléotides libres qui ont été synthétisés dans la cellule s'alignent sur la matrice d'ADN. Les ribonucléotides libres A s'alignent faces aux T dans l'ADN, les G face aux C, les C face aux G et les U face aux A. Ce processus est catalysé par l'enzyme ARN polymérase,

qui se fixe et se déplace le long de l'ADN en ajoutant des ribonucléotides à la chaîne d'ARN en cours d'élongation.

L'ARN s'allonge toujours dans le sens 5'→3', comme le montre la figure 1.6, le fait que l'ARN soit synthétisé dans le sens 5'→ 3' signifie le brin matrice doit être orienté dans le sens 3'→ 5'.

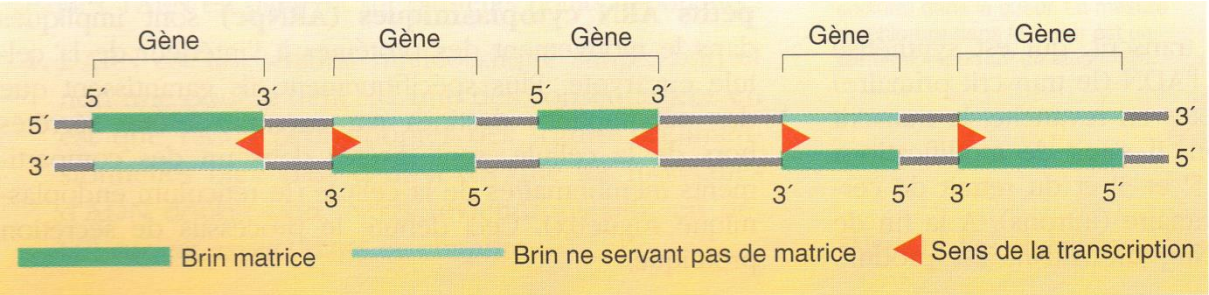


Figure 1.5. Brins d'ADN utilisés comme matrices pour la transcription.

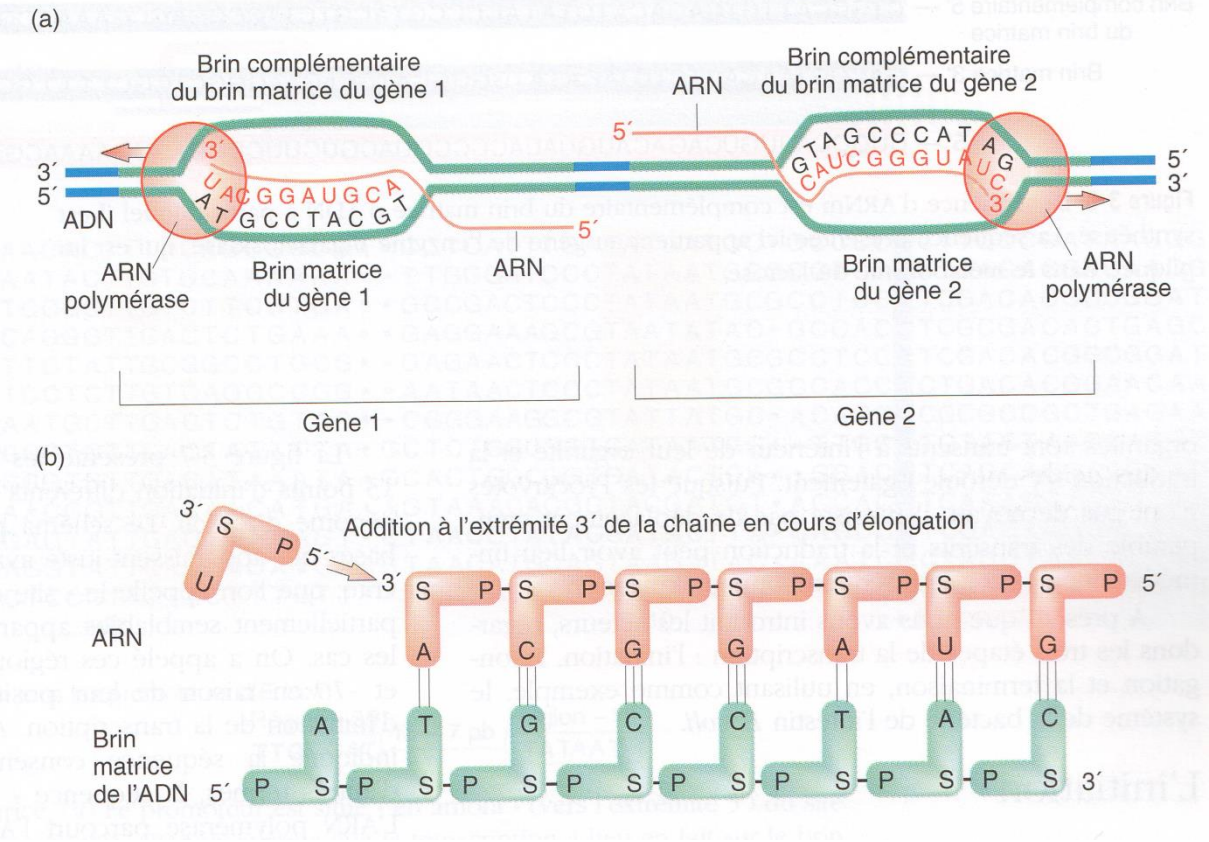


Figure 1.6. Transcription de deux gènes.

(a) L'ARN polymérase se déplace à partir de l'extrémité 3' du brin matrice. Certains gènes sont transcrits à partir d'un brin de la double hélice tandis que d'autres gènes utilisent d'autre brin comme matrice.

(b) Un uracile est ajouté à l'extrémité 3' du transcrit pour le gène 1.

L'initiation et l'élongation :

Une séquence d'ADN à laquelle se fixe une ARN polymérase pour initier la transcription est appelée un promoteur. Un promoteur fait partie de la région régulatrice adjacente à la région codante d'un gène. Rappelons nous, puisqu'un transcrit d'ARN est synthétisé dans le sens 5'→3', on examine par convention le gène également dans le sens 5'→3' (l'orientation du brin complémentaire du brin matrice). En utilisant ce principe, le promoteur se trouve au début du gène.

L'ARN polymérase parcourt l'ADN à la recherche d'une séquence promotrice, se fixe à l'ADN à cet endroit, le déroule et commence la synthèse d'une molécule d'ARN au niveau du site d'initiation de la transcription (figure 1.7).

La terminaison :

Lorsque l'ARN polymérase reconnaît des séquences nucléotidiques spécifiques dans l'ADN qui agissent comme des signaux de terminaison de la synthèse de la chaîne, le brin d'ARN et la polymérase sont libérés de la matrice d'ADN.

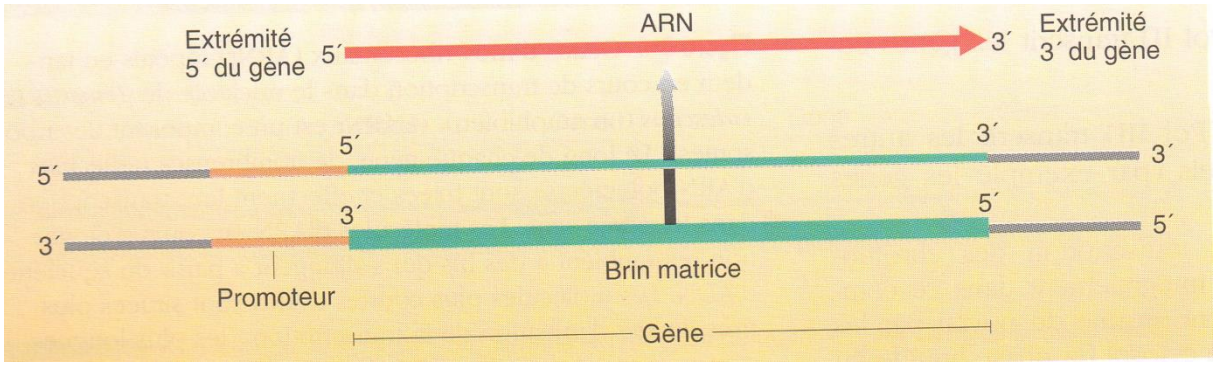


Figure 1.7. Convention utilisée pour désigner les extrémités 5' et 3' d'un gène.

1.8.7. La structure d'une protéine :

Une protéine est une chaîne d'acides aminés que l'on désigne parfois par le terme polypeptides.

La traduction :

La séquence d'acides aminés spécifique d'un polypeptide est déterminée par la séquence nucléotidique du gène qui la code. La séquence de nucléotides de l'ADN d'un gène est transcrite en une séquence équivalente d'ARNm. Un groupe de ribosomes parcourt l'ARNm, chacun débutant à l'extrémité 5' et avançant sur toute la longueur de l'ARNm pour gagner l'extrémité 3'. Au fur et à mesure qu'un ribosome avance, il "lit" la séquence nucléotidique de l'ARNm trois nucléotides à la fois. Chaque groupe de trois est appelé un triplet ou codon et

correspond à un acide aminé spécifique. Puisqu'il y a 4 nucléotides différents dans l'ARNm, il y a 4×4×4 = 64 codons différents possibles. Ces codons et les acides aminés aux quels ils correspondent sont présentés dans la figure 1.8 et le tableau 1.2.

Le tableau 1.2 montre que le nombre de codons pour un même acide aminé varie, allant de un (tryptophane, UGG) à six (Sérine, UCU, UCC, UCA, UCG, AGU ou AGC), on ne sait pas exactement pourquoi.

Certains acides aminés peuvent être apportés au ribosome par plusieurs types d'ARNt portant des anticodons différents, tandis que d'autres acides aminés sont apportés au ribosome par un seul ARNt.

Deuxième lettre

Première lettre	U	C	A	G	Troisième lettre
U	*UUU, UUC* : *Phe* *UUA, UUG* : *Leu*	*UCU, UCC, UCA, UCG* : *Ser*	*UAU, UAC* : *Tyr* *UAA* : *Stop* *UAG* : *Stop*	*UGU, UGC* : *Cys* *UGA* : *Stop* *UGG* : *Trp*	U C A G
C	*CUU, CUC, CUA, CUG* : *Leu*	*CCU, CCC, CCA, CCG* : *Pro*	*CAU, CAC* : *His* *CAA, CAG* : *Gln*	*CGU, CGC, CGA, CGG* : *Arg*	U C A G
A	*AUU, AUC, AUA* : *Ile* *AUG* : *Met*	*ACU, ACC, ACA, ACG* : *Thr*	*AAU, AAC* : *Asn* *AAA, AAG* : *Lys*	*AGU, AGC* : *Ser* *AGA, AGG* : *Arg*	A C A G
G	*GUU, GUC, GUA, GUG* : *Val*	*GCU, GCC, GCA, GCG* : *Ala*	*GAU, GAC* : *Asp* *GAA, GAG* : *Glu*	*GGU, GGC, GGA, GGG* : *Gly*	A C A G

Figure 1.8. Code génétique.

Tableau 1.2. Les 20 acides aminés habituels chez les organismes vivants.

Acide aspartique	Asp	Leucine	Leu
Acide glutamique	Glu	Lysine	Lys
Alanine	Ala	Méthionine	Met
Arginine	Arg	Phénylalanine	Phe
Asparagine	Asn	Proline	Pro
Cystéine	Cys	Sérine	Ser
Glutamine	Gln	Thréonine	Thr
Glycine	Gly	Tryptophane	Trp
Histidine	His	Tyrosine	Tyr
Isoleucine	Ile	Valine	Val

Certaines espèces d'ARNt peuvent apporter leur acide aminé spécifique en réponse à plusieurs codons.

L'initiation :

Chez les Eucaryotes, AUG est généralement le premier codon d'un polypeptide. Chez *E. coli*, AUG, GUG, et à de rares occasions UUG, servent de codons d'initiations. Comment les codons d'initiation corrects sont ils sélectionnés au milieu de tous les codons AUG et GUG qui se trouvent dans une molécule d'ARNm ? Chez les bactéries, les codons d'initiation sont précédés de séquences appelées séquences de ***Shine-Dalgarno*** qui s'apparient avec l'extrémité 3' de l'ARNr 16S du ribosome à côté du codon d'initiation. Il y a une région de séparation courte et variable entre la séquence de *Shine- Dalgarno* et le codon d'initiation.

L'élongation :

Plusieurs protéines appelées facteurs d'élongation, qui guident la fixation et le déplacement des ARNt et du ribosome participent à l'élongation.

La terminaison :

Certains codons ne spécifient aucun acide aminé. Ces codons, UAG, UGA et UAA sont appelés codons stop ou codon de terminaison. On peut les considérer comme les signes de ponctuation qui terminent le message codé dans l'ARNm. Les codons stop sont souvent appelés codons non-sens car ils ne désignent aucun acide aminé. C'est trompeur, car la ponctuation constitue une partie importante du sens de tout message. Il est intéressant de constater que les trois codons stop ne sont pas reconnus par ARNt mais par des facteurs protéiques appelés facteurs de libération.

1.9. Conclusion :

L'objectif de ce chapitre est dans un premier temps de donner un aperçu des grands thèmes qui constituent la bioinformatique contemporaine, en précisant notamment les techniques informatiques qui s'en trouvent au centre. Ce chapitre présente aussi des bases biologiques, où des généralités relatives aux acides nucléiques et aux protéines sont données.
La combinaison des deux aspects informatique et biologie permet d'accroitre significativement la compréhension des mécanismes biologiques, tout en utilisant les problématiques inhérentes a la biologie afin de développer les méthodes qui leur sont propres.

Chapitre 2

Analyse fréquentielle des signaux ADN

2.1. Introduction :

Depuis le début des années 90, l'intérêt des mathématiciens, physiciens et informaticiens pour l'analyse statistique des séquences d'ADN n'a pas cessé de croître. En effet, les immenses progrès de la biologie moléculaire et les grands projets de séquençage ont révélé l'extraordinaire complexité des génomes. Afin de mieux comprendre l'organisation et l'évolution des génomes, il est apparu nécessaire d'introduire de nouveaux concepts et de nouvelles techniques d'analyse du signal.

Pour caractériser quantitativement les propriétés statistiques, les lois de distribution des nucléotides et les possibles corrélations, il faut associer à la séquence d'ADN, et donc au texte d'alphabet {A, C, T, G} pour les quatre bases adenine, cytosine, thymine et guanine, un signal digital.

2.2. Construction de signaux ADN par codage :

Une séquence ADN contient des informations de différentes natures (structurelles, comme les propriétés de courbure locale de la double hélice, ou fonctionnelles, comme la localisation des gènes), cette abondance d'information pouvant a priori rendre difficile l'analyse d'une propriété particulière. En utilisant un signal unidimensionnel, le but est de s'affranchir, dans la limite du possible, des informations pour se focaliser sur un signal plus spécifique [15].

Grâce au codage, une séquence nucléotidique peut être associée à un signal, le codage permet d'associer un signal à un brin d'ADN en considérant une séquence de nucléotides comme un mot construit sur un alphabet nucléotidique. On associe un signal à un tel mot via une application définie sur l'alphabet et à une valeur numérique.

Les deux brins d'une macro-molécule d'ADN étant complémentaires, l'étude des éléments constitutifs d'une séquence ADN peut se ramener à l'étude des mots construits sur un alphabet de quatre lettres {A, C, G, T}, représentant les quatre bases nucléotidiques constitutives d'un des deux brins de l'ADN.

Pour caractériser quantitativement les propriétés statistiques, les lois de distribution des nucléotides et les possibles corrélations, il faut associer à la séquence de longueur *N*, et donc au texte d'alphabet {A, C, T, G} représentant les quatre bases nucléotidiques constitutives d'un des deux brins de l'ADN, un signal digital *x(n), n=0,1,…,N-1*. La construction d'un signal à partir d'une séquence de nucléotides se résume par la donnée d'un codage. Il en existe de nombreux et le choix du codage dépend de la propriété que l'on souhaite mettre en évidence. Nous donnerons quelques exemples couramment rencontrés dans la littérature:

- **Codages du type purine-pyrimidine :**

Les codages les plus simples sont certainement les codages binaires qui séparent les nucléotides en deux familles, en attribuant à chacune de ces familles une valeur différente. Le codage défini explicitement par les relations suivantes,

$$\tau(A)=1, \tau(G)=1,\ \tau(C)=-1,\ \tau(T)=-1 \tag{2.1}$$

permet de faire la distinction entre les purines A et G et les pyrimidines C et T. Il est naturellement appelé codage purine-pyrimidine [16].

- **Codages mono-nucléotidiques :**

Un codage mono-nucléotidique est un codage permettant d'analyser la distribution de la position d'une des bases A, C, G ou T dans un mot, en associant trois des bases à la même valeur. Il existe donc principalement quatre codages mono-nucléotidiques différents. Si l'on souhaite, par exemple, étudier la répartition en adénine d'une séquence ADN, il suffit de définir le codage de la manière suivante,

$$\tau(A)=1\ ,\ \tau(G)=-1/3\ ,\ \tau(C)=-1/3\ ,\ \tau(T)=-1/3 \tag{2.2}$$

où la valeur −1/3 est choisie telle que, si les concentrations en A, C, G et T sont égales, le bruit associé possède une moyenne nulle.

Si les concentrations en nucléotides s'écartent notablement de l'équipartition, il y a lieu de modifier le codage comme suit. Supposons que les concentrations en A, C, G et T sont respectivement c_A, c_C, c_G et c_T , et posons $c = c_C + c_G + c_T$. Il suffit de redéfinir le codage de la manière suivante, τ (A) = 1 et τ (*l*) = −*cl*/*c*, pour *l* appartenant à {C,G,T}.

- **Codage pourcentage en GC :**

La concentration en bases C et G est une des caractéristiques du génome les plus étudiées.

Par exemple, la densité en gènes semble être corrélée à la concentration en GC, les régions riches en gènes étant des régions à haute concentration en GC.

Pour représenter la concentration en bases C et G le long d'un brin d'ADN, on utilise le codage suivant, appelé codage pourcentage en GC,

$$\tau(C)=1,\ \tau(G)=1,\ \tau(A)=0,\ \tau(T)=0 \tag{2.3}$$

Dans cette étude, nous nous intéresserons plus particulièrement au codage proposé par D. Anastassiou [17]. Nous montrerons dans ce qui suit que ce codage semble refléter plus particulièrement des propriétés spécifiques des séquences étudiées telles que la configuration spatiale de la double hélice et la distribution de la composition nucléotidique. Ceci justifiera l'étude privilégiée que nous ferons de ce codage dans la suite du travail par rapport aux autres codages.

Le codage est défini comme suit :

Dans une séquence d'ADN, des nombres *a*, *t*, *c*, *g* sont attribués aux caractères A, T, C, G, respectivement. Un bon choix des nombres *a*, *t*, *c* et *g* peut fournir des propriétés potentiellement utiles aux séquences numériques *x(n)*.

Par exemple, si nous choisissons des paires complexes conjugués $t = a^*$ et $g = c^*$, alors le brin d'ADN complémentaire est représenté par:

$$\tilde{x}(n)=x^*(-n+N-1),\ n=0,1,\ldots,N-1 \tag{2.4}$$

Dans ce cas, tous les codes produiront des séquences numériques conjugués symétriques ont des propriétés mathématiques intéressantes, y compris une phase linéaire. Un tel code (le plus simple parmi tous ceux possibles) est le suivant:

$$a=1+j,\ t=1-j,\ c=-1-j,\ g=-1+j \tag{2.5}$$

2.3. Etude fréquentielle des signaux ADN :

Afin d'explorer le contenu fréquentiel des signaux issus de codages des séquences ADN du génome, nous avons procédé à une analyse par la transformée de Fourier, la transformée en ondelettes et les bancs de filtres. Ces outils de traitement du signal permettent de représenter efficacement certaines informations structurées dans les signaux.

Certaines applications intéressantes des techniques de traitement du signal pour la prédiction de la structure de l'ADN, la détection, l'extraction de caractéristiques et la classification des gènes ont déjà fait la preuve de leur utilité [18,19].

Dans cette partie, nous allons présenter les principes des méthodes que nous avons utilisées.

2.3.1. Les ondelettes :

Le traitement du signal a pour principal objet la description des signaux liés au monde réel dans un but de traitement, d'identification, de compression, de compréhension ou de

transmission. Dans ce contexte, les transformations linéaires ont toujours joué un très grand rôle, et parmi ces dernières, la plus célèbre et la plus anciennement étudiée est la transformation de Fourier (1822). Cette transformation permet, comme chacun sait, d'explorer la composition fréquentielle du signal et par ses propriétés de lui appliquer facilement des opérateurs de filtrage. Lors de cette transformation le signal est décomposé sur un ensemble de signaux de «base» qui sont les cosinus et sinus ou l'exponentielle imaginaire, mais, très tôt dans l'histoire du traitement du signal, il est apparu que la décomposition obtenue n'était pas toujours la plus satisfaisante et la première transformation en ondelettes (le nom n'est pas encore utilisé) est proposée par Haar en 1910 ; il serait plus judicieux de parler alors de «paléo-ondelette». La transformée en ondelettes est un outil qui découpe les données, les fonctions ou les opérateurs en composantes fréquentielles suivant une résolution adaptée à l'échelle. Les précurseurs conscients de cette technique ont été des mathématiciens (Calderon 1964), des physiciens (Aslaken et Klauder en 1968, Paul en 1985), et surtout des ingénieurs (ou des chercheurs en sciences pour l'ingénieur) comme Esteban et Galand (1977), Smith et Barnwell (1986), Vetterli (1986), nous pourrions parler dans leur cas de «pré-ondelette». Mais le premier à avoir utilisé la méthode et le premier à avoir proposé le nom d'ondelettes fut Jean Morlet (1983). Le problème traité par Morlet était celui de l'analyse de données issues de sondages sismiques effectués pour des recherches géologiques ; ces données faites de nombreux transitoires sont particulièrement adaptées à une technique d'analyse conservant la notion de localisation de l'événement tout en fournissant une information sur son contenu fréquentiel ce qui est tout l'intérêt de ce type de transformation. Les résultats obtenus par Morlet et formalisés par le physicien Alex Grossmann ont rapidement éveillé l'attention de nombreux chercheurs et bientôt des bases mathématiques solides ont été mises en place faisant apparaître la notion de base orthogonale (Y.Meyer 1985), d'analyse multirésolution (S. Mallat 1989) et d'ondelettes à support compact (I. Daubechies 1988). Les ondelettes modernes étaient nées [20].

- **Pourquoi les ondelettes ?**

La plupart des signaux du monde réel ne sont pas stationnaires, et c'est justement dans l'évolution de leurs caractéristiques (statistiques, fréquentielles, temporelles, spatiales) que réside l'essentiel de l'information qu'ils contiennent. Or l'analyse de Fourier (2.6) propose une approche globale du signal, les intégrations sont faites de moins l'infini à plus l'infini, et toute notion de localisation temporelle (ou spatiale pour des images) disparaît dans l'espace de Fourier ; il faut donc trouver un compromis, une transformation qui renseigne sur le

contenu fréquentiel tout en préservant la localisation afin d'obtenir une représentation espace/échelle du signal.

• **Transformée de Fourier :**

$$T^{fourier} f(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j\omega t} dt \tag{2.6}$$

• **Transformée en ondelettes :**

$$T^{ond} f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \psi(\frac{t-b}{a}) dt \tag{2.7}$$

Dans cette expression, *a* est le facteur d'échelle et *b* le paramètre de translation. La variable *a* joue le rôle de l'inverse de la fréquence : plus *a* est petit moins l'ondelette (la fonction analysante) est étendue temporellement, donc plus la fréquence centrale de son spectre est élevée. On peut également interpréter cette expression comme une projection du signal sur une famille de fonctions analysantes $\psi_{a,b}$ construite à partir d'une fonction "mère" ψ conformément à l'équation suivante :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi(\frac{t-b}{a}) \tag{2.8}$$

On notera que la norme est conservée lors du changement de facteur d'échelle :

$$\begin{aligned} \left\| \psi_{a,b} \right\|^2 &= \int_{-\infty}^{+\infty} \frac{1}{a} \left| \psi(\frac{t-b}{a}) \right|^2 dt \\ &= \frac{1}{a} \int_{-\infty}^{+\infty} \left| \psi(x) \right|^2 a dx \\ &= \left\| \psi \right\|^2 \end{aligned} \tag{2.9}$$

On pourra noter :

$$T^{ond} f(a,b) = \left\langle f, \psi_{a,b} \right\rangle \tag{2.10}$$

La résolution spatio-temporelle est calculée de la même manière que précédemment : Si la «largeur» temporelle de ψ (l'écart type) est prise comme unité : σ = 1 alors on peut calculer la «largeur» de $\psi_{a,0}$ avec l'équation suivante :

$$\begin{aligned} \sigma_t^2 &= \int t^2 \left| \psi_{a,0}(t) \right|^2 dt \\ &= \int t^2 \frac{1}{a} \left| \psi(\frac{t}{a}) \right|^2 dt \end{aligned}$$

$$= \int a^2 x^2 \frac{1}{a} |\psi(x)|^2 \, a dx$$

ce qui donne : $\sigma_t = a$.

On peut de même calculer l'occupation fréquentielle de l'ondelette en calculant l'écart type pour la transformée de Fourier $\psi_{a,0}$ de $\psi_{a,0}$; en prenant comme unité l'écart type de la transformée de Fourier de l'ondelette mère ψ :

$$\begin{aligned} \sigma_\omega^2 &= \int \omega^2 \left| \psi_{a,0}(\omega) \right|^2 d\omega \\ &= \int \omega^2 \frac{1}{a} \left| a\psi(a\omega) \right|^2 d\omega \\ &= \int \frac{\xi^2}{a^2} \frac{1}{a} \left| a\psi(\xi) \right|^2 \frac{d\xi}{a} \end{aligned} \qquad (2.11)$$

On trouve $\sigma_\omega = \frac{1}{a}$. De sorte que le pavé élémentaire dans l'espace temps-fréquence est de surface constante tandis que la résolution temporelle est proportionnelle à a et que la résolution fréquentielle est inversement proportionnelle à a comme on le voit sur la figure 2.1.

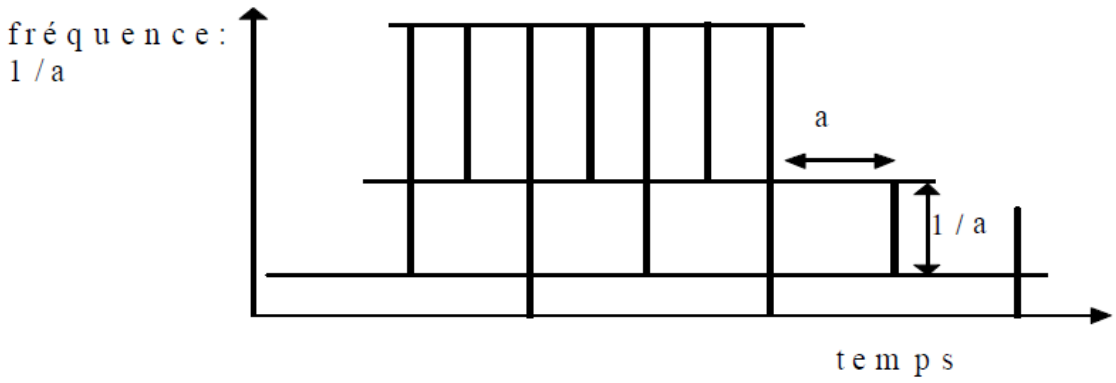


Figure 2.1. Pavage temps-fréquence pour la transformée en ondelette discrète.

Les premières ondelettes utilisées (en dehors de l'ondelette de Haar que nous étudierons plus loin) ont été l'ondelette de Morlet, une gaussienne modulée par une exponentielle complexe, et le chapeau mexicain, en réalité la dérivée seconde d'une gaussienne.

- Ondelette de Morlet :

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{-i\omega_0 x} \qquad (2.12)$$

- Chapeau mexicain :

$$\psi(x) = \frac{2}{\sqrt{3}} \pi^{-\frac{1}{4}} (1 - x^2) e^{-\frac{x^2}{2}} \tag{2.13}$$

La figure 2.2 présente le chapeau mexicain pour deux valeurs du facteur d'échelle : a = 1 pour la courbe la plus localisée et a = 2 pour la courbe la plus étendue (la figure 2.5 présente la réponse fréquentielle pour a = 1). La figure 2.3 présente la partie réelle de l'ondelette de Morlet pour deux valeurs du facteur d'échelle, on pourra comparer avec la figure 2.4 où on constate que la fenêtre d'analyse reste constante lors du changement d'échelle (fréquence).

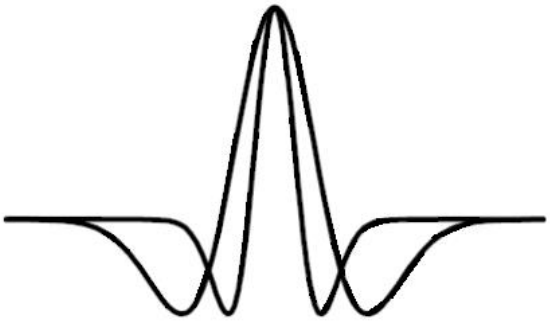


Figure 2.2. Chapeau mexicain.

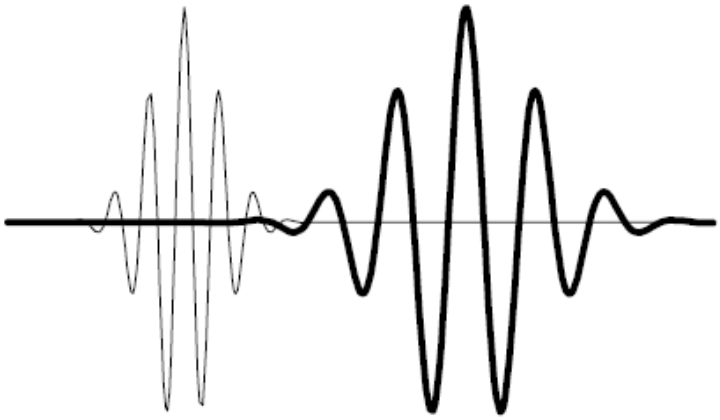


Figure 2.3. Ondelette de Morlet (partie réelle).

Les représentations fréquentielles des ondelettes de Morlet (ω_o= 5), figure 2.4, illustrent que la largeur spectrale de l'ondelette varie en fonction du facteur d'échelle inversement à la largeur spatiale.

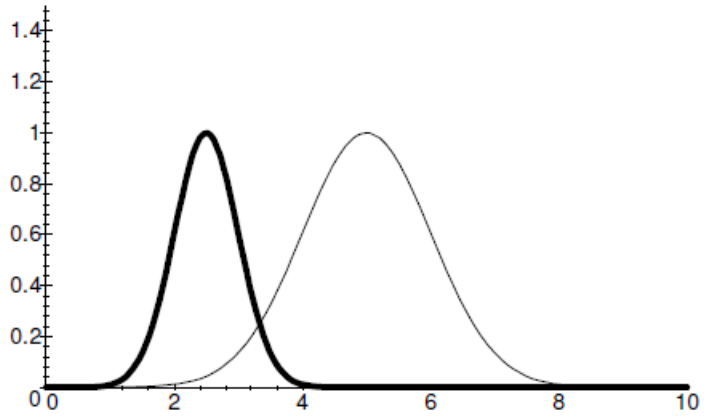


Figure 2.4. Ondelette de Morlet : $\psi(\omega)$.

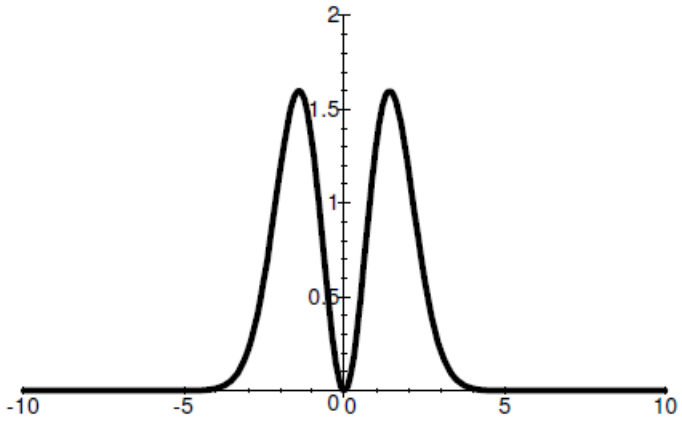


Figure 2.5. Chapeau mexicain : $\psi(\omega)$.

Le facteur d'échelle a et le pas de translation b sont des réels et la transformation en ondelettes est continue et donc redondante. Le plan temps fréquence est sur-analysé. Il est donc évident qu'une discrétisation de la transformée doit être envisagée si on souhaite obtenir une transformation non redondante. Le pavage temps-fréquence obtenu par la transformation en ondelettes (figure 2.1) suggère une méthode de discrétisation exponentielle pour les échelles et pour le temps. Dans l'expression $\psi(\frac{t-b}{a})$ le pas de translation à l'échelle a est b/a.

On posera donc :

$$a = a_0^m \text{ et } b = nb_0a_0^m \text{ avec } a_0,\ b_0,\ n,\ m \in \mathbb{Z}$$

d'où l'expression de la transformée en ondelettes discrète (2.14) donnée ci-après.

- **Transformée en ondelettes discrète :**

$$T^{ond} f(m.n) = a_0^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t)\psi(a_0^{-m}t - nb_0)dt \tag{2.14}$$

Si on choisit $a_0 = 2$ et $b_0 = 1$, on parle alors de transformée dyadique.

2.3.2. Analyse multirésolution :

2.3.2.1. Espaces d'approximation :

Nous nous plaçons dans l'espace $L^2(\mathbb{R})$ des fonctions continues d'une variable réelle et de carré intégrable. Une analyse à la résolution *j* de la fonction *f* sera obtenue par action d'un opérateur linéaire A_j sur *f*, tel que :

$$A_j f \in V_j \tag{2.15}$$

V_j étant un sous espace de L^2, A_j sera un projecteur.

On construira une analyse multirésolution à l'aide de sous-espaces V_j emboîtés les uns dans les autres, tels que le passage de l'un à l'autre soit le résultat d'un changement d'échelle (zoom).

Par exemple, dans le cas dyadique on aura :

$$f(x) \in V_j \Leftrightarrow f(\frac{x}{2}) \in V_{j+1} \tag{2.16}$$

ce qui correspond à une dilatation d'un facteur 2. L'espace V_{j+1} contient des signaux plus "grossiers" que l'espace V_j et il est clair que :

$$V_{j+1} \subset V_j \tag{2.17}$$

L'axiomatique correspondante peut s'exprimer comme suit :

Soit un ensemble de sous espaces de L^2 tels que :

$$..... \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset ..... \subset V_{j+1} \subset V_j \subset$$

$$\overline{\bigcup_{j \in z} V_j} = L^2(\mathbb{R}) \tag{2.18}$$

$$\bigcap_{j \in z} V_j = \{0\} \tag{2.19}$$

$$\forall j \in \mathbb{Z} \quad \text{si} \quad f(x) \in V_j \Leftrightarrow f(2^{-1}x) \in V_{j+1} \quad (\ ou\ f(2^j x) \in V_0) \tag{2.20}$$

$$\forall k \in \mathbb{Z} \quad \text{si} \quad f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0 \qquad \text{(invariance par translation)} \tag{2.21}$$

Cet ensemble définit une analyse multirésolution de $L^2(\mathbb{R})$.

Remarque 1 : La propriété 2.18 assure la convergence de l'analyse et peut aussi s'écrire :

$$\lim_{j\to-\infty} V_j = L^2(\mathbb{R}) \tag{2.22}$$

Dans ces conditions, on peut montrer qu'il existe une fonction dite fonction d'échelle qui par dilatation et translation engendre une base orthonormée de V_j . Cette fonction sera notée :

$$\varphi(x) \in L^2(\mathbb{R}) \tag{2.23}$$

et les fonctions de bases sont construites suivant la relation :

$$\varphi_{j,n}(x) = 2^{\frac{-j}{2}}\varphi(2^{-j}x - n) \text{ avec } n \in \mathbb{Z} \tag{2.24}$$

Il suffit d'ailleurs que $\varphi(.,-n)$ soit une base de V_o.

La base sera orthonormée si :

$$\int_{-\infty}^{+\infty} \varphi(x)\varphi^*(x+n)dx = \delta(n) \quad \forall n \in \mathbb{Z} \tag{2.25}$$

Rappelons que le produit scalaire est défini par :

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(x)g^*(x)dx \tag{2.26}$$

(pour des fonctions réelles ou complexes)

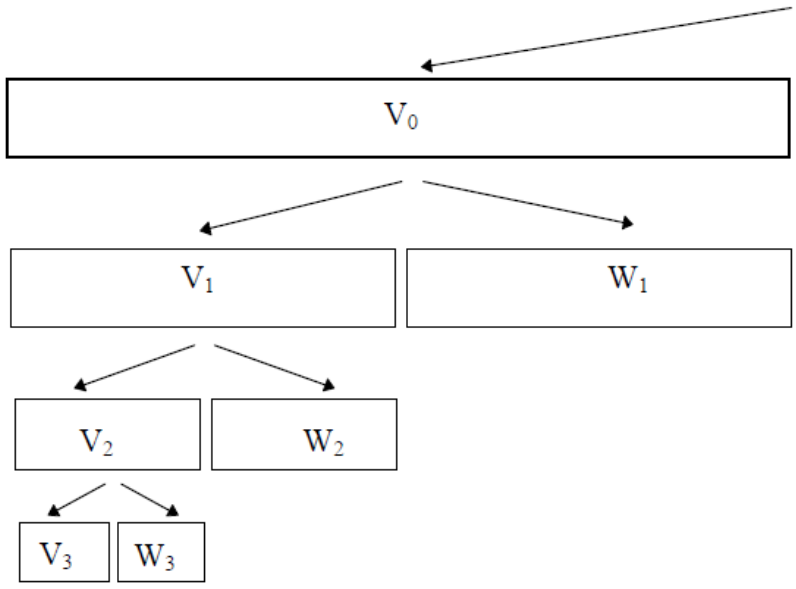


Figure 2.6. Schéma de l'analyse multirésolution.

La relation d'orthogonalité entre les fonctions de base pour une échelle donnée 2.25 pourra donc s'écrire :

$$\left\langle \varphi_{j,n}, \varphi_{j,k} \right\rangle = \delta(n-k) \quad \forall n,k,j \in \mathbb{Z} \tag{2.27}$$

L'action du projecteur sur *f* fournira sa décomposition sur la base des fonctions d'échelle et les coefficients de cette décomposition constituent l'approximation à l'échelle *j* de *f*.

$$A_j f = \sum_n \left\langle f, \varphi_{j,n} \right\rangle \varphi_{j,n} \tag{2.28}$$

On pose :

$$a_n^j = \left\langle f, \varphi_{j,n} \right\rangle \tag{2.29}$$

l'approximation à la résolution *j* de la fonction *f* sera définie par la suite discrète des nombres (réels ou complexes) a_n^j.

Une suite numérique formée par échantillonnage d'un signal continu réel pourra être considérée comme une approximation à une résolution donnée du signal continu.

2.3.2.2 Espaces des détails

L'espace des détails vient compléter l'analyse. On peut définir pour chaque V_j son complément orthogonal W_j dans V_{j-1} tel que :

$$V_{j-1} = V_j \oplus W_j$$

$$L^2(\mathbb{R}) = \oplus_{j \in \mathbb{Z}} W_j$$

Comme W_{j-1} est orthogonal à V_{j-1}, alors W_{j-1} sera orthogonal à W_j ; cette propriété s'écrit :

$$\forall j,k \neq j \quad \text{alors} \quad W_j \perp W_k \tag{2.30}$$

Les sous-espaces W_j ne forment pas une famille d'espaces emboîtés, mais les propriétés d'échelle et d'invariance par translation sont conservées.

Dans ces conditions, on peut montrer qu'il existe une fonction appelée ondelette qui par dilatations et translations engendre une base orthonormée des W_j et donc de L^2.

Cette fonction est notée :

$$\psi(x) \in L^2(\mathbb{R}) \tag{2.31}$$

et les fonctions de base sont construites suivant la relation :

$$\psi_{j,n}(x) = 2^{-\frac{1}{2}}\psi\left(2^{-j}x - n\right) \text{ avec } \quad n \in \mathbb{Z} \tag{2.32}$$

L'orthonormalité de la base d'ondelettes s'écrit :

$$\left\langle \psi_{j,n}, \psi_{i,k} \right\rangle = \delta(j-i)\delta(n-k) \quad \forall j,i,n,k \in \mathbb{Z} \tag{2.33}$$

L'approximation à l'échelle immédiatement plus fine pourra donc être reconstruite en utilisant les détails du signal fournis par sa projection sur la base de W_j suivant la relation suivante :

$$A_{j-1}f = A_j f + \sum_n \left\langle f, \psi_{j,n} \right\rangle \psi_{j,n} \tag{2.34}$$

On notera D_j le projecteur sur W_j et le signal de détail sera décrit par la suite numérique :

$$d_n^j = \left\langle f, \psi_{j,n} \right\rangle \tag{2.35}$$

donc :

$$D_j f = \sum_n \left\langle f, \psi_{j,n} \right\rangle \psi_{j,n} \tag{2.36}$$

et la formule de reconstruction s'écrit :

$$A_{j-1}f = A_j f + D_j f \tag{2.37}$$

Le signal de détail est constitué d'une suite numérique dont les éléments sont aussi les coefficients de la transformée en ondelettes.

Le schéma de la décomposition est représenté symboliquement sur la figure 2.6. Dans laquelle la largeur des rectangles symbolisant les sous-espaces est proportionnelle à la densité de l'échantillonnage réalisé par la projection du signal dans le sous-espace considéré.

Exemple : En exemple, on peut présenter l'analyse de Haar. Les sous espaces V_j sont définis par :

$$V_j = \left\{ f \in L^2(\mathbb{R}) \text{ telles que } \forall k \in \mathbb{Z} \text{ on a } f_{\left[2^j k, 2^j(k+1)\right[} = \text{ constante} \right\} \tag{2.38}$$

Le sous espace V_j est l'ensemble des fonctions constantes sur les intervalles de largeur 2^j. Les fonctions de base sont construites à partir de la fonction d'échelle $\varphi(x)$ égale à 1 de 0 à 1 et nulle partout ailleurs :

$$\varphi(x)=\begin{cases}0 & si \quad x<0\\ 1 & si \quad 0\le x<1\\ 0 & si \quad 1\le x\end{cases} \tag{2.39}$$

La fonction ondelette est définie par :

$$\psi(x)=\begin{cases}0 & si \quad x<0\\ 1 & si \quad 0\le x<1/2\\ -1 & si \quad 1/2\le x<1\\ 0 & si \quad 1\le x\end{cases} \tag{2.40}$$

Les figures 2.7 et 2.8 présentent les représentations graphiques de ces fonctions.

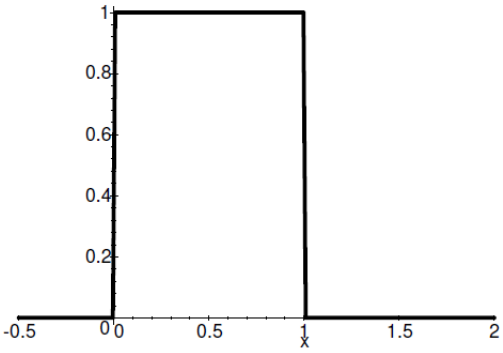


Figure 2.7. Fonction d'échelle de l'analyse de Haar.

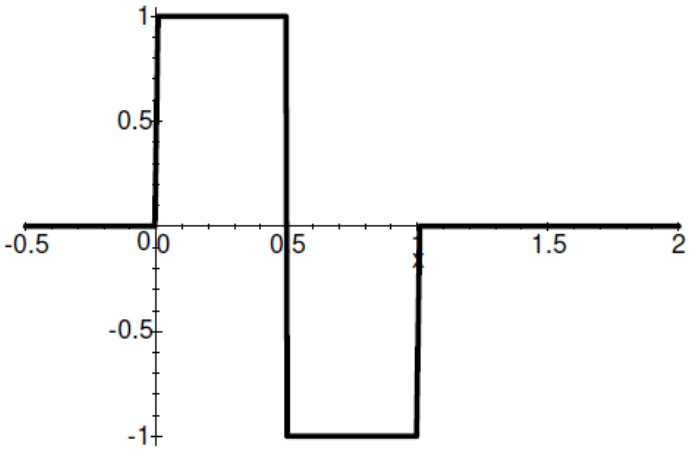


Figure 2.8. Ondelette mère de l'analyse de Haar.

Nous prenons une fonction quelconque pour illustrer la décomposition aux échelles *j = 0* et *j = 1*. Cette fonction est présentée sur le graphe de la figure 2.9.

Ses projections sur le sous espace V_o et sur le sous espace V_1 sont présentées dans la figure 2.10.

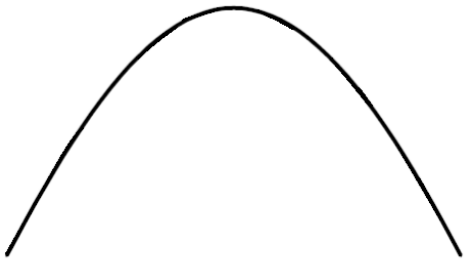


Figure 2.9. Fonction exemple.

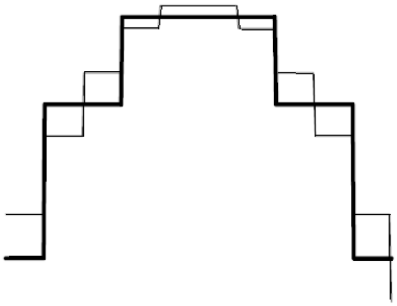


Figure 2.10. Signaux d'approximation $A_0 f$ (en trait fin) et $A_1 f$.

La projection sur W_1, donc le signal de détail, est donnée sur la figure 2.11.

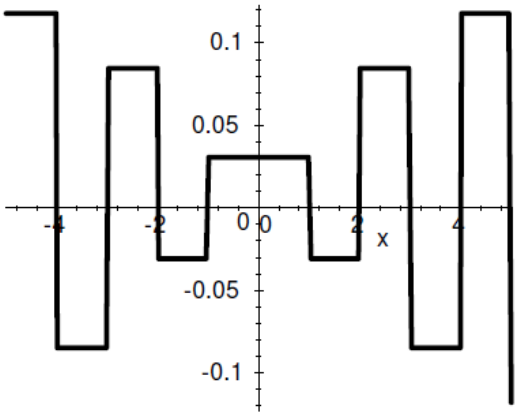


Figure 2.11. Signal de détail $D_1 f$.

On vérifie bien que :

$$A_0 f = A_1 f + D_1 f$$

Conformément à l'équation 2.37.

2.3.2.3. Algorithme d'analyse :

- **Projection sur les fonctions d'échelle :**

Le point clef est fourni par la décomposition de $a_n^j = \langle f, \varphi_{j,n} \rangle$ en fonction de $a_n^{j-1} = \langle f, \varphi_{j-1,n} \rangle$.

Par construction $\varphi(x)$ est une fonction de V_0 ; comme $V_0 \subset V_{-1}$ on peut décomposer $\varphi(x)$ sur la base de V_{-1}. Et donc $\exists h[n]$ suite numérique avec $n \in \mathbb{Z}$ telle que :

$$\varphi(x) = \sum_n h[n] \varphi_{-1,n}(x) \tag{2.41}$$

avec, conformément à 2.24, $\varphi_{-1,n}(x) = 2^{\frac{1}{2}} \varphi(2x-n)$, soit :

$$h[n] = \langle \varphi, \varphi_{-1,n} \rangle \tag{2.42}$$

La suite numérique $h[n]$ sera considérée comme étant la réponse impulsionnelle d'un filtre numérique.

La construction de cette suite peut être menée à partir de la donnée de $\varphi(x)$. On pourra donc définir une analyse multirésolution indifféremment en partant de la fonction d'échelle ou du filtre numérique associé.

Montrons que la décomposition est la même pour des échelles quelconques.

On a :

$$\varphi(x) = \sum_n h[n] 2^{1/2} \varphi(2x-n) \tag{2.43}$$

donc :

$$\varphi_{j,n}(x) = 2^{-j/2} \sum_k h[k] 2^{1/2} \varphi\left(2\left(2^{-j}x - n\right) - k\right) \tag{2.44}$$

ce qui en regroupant les indices et les exposants conduit à l'équation :

$$\varphi_{j,n} = \sum_k h[k] \varphi_{j-1,k+2n} \tag{2.45}$$

Donc on peut calculer les coefficients $a_n^j = \langle f, \varphi_{j,n} \rangle$ de l'approximation à la résolution j :

$$a_n^j = \sum_k h[k] \langle f, \varphi_{j-1,k+2n} \rangle \tag{2.46}$$

si on pose $l = k + 2n$, cette expression s'écrit :

$$a_n^j = \sum_l h[l-2n]\langle f, \varphi_{j-1,l}\rangle \tag{2.47}$$

et si on note :

$$\tilde{h}[n] = h[-n] \tag{2.48}$$

la séquence retournée ou le filtre symétrique de *h*, on obtient :

$$a_n^j = \sum_l \tilde{h}[2n-l]\langle f, \varphi_{j-1,l}\rangle \tag{2.49}$$

et on aura finalement l'équation récursive suivante :

$$a_n^j = \sum_l \tilde{h}[2n-l]a_l^{j-1} \tag{2.50}$$

Si on considère a_n^j comme une séquence numérique indexée par *n*, le calcul précédent peut être interprété comme un produit de convolution entre $\tilde{h}$ et a^{j-1} évalué pour un indice sur deux ; ou encore comme le filtrage de la séquence a^{j-1} par le filtre de réponse impulsionnelle $\tilde{h}$ suivi par un sous-échantillonnage de rapport 2.

- **Projection sur les fonctions ondelettes :**

Un schéma analogue est bâti à partir de la décomposition de l'ondelette de W_0 sur la base de V_{-1} :

$$\psi = \sum_n g[n]\varphi_{-1,n} \tag{2.51}$$

ou de façon plus détaillée :

$$\psi(x) = \sum_n g[n]\sqrt{2}\varphi(2x-n) \tag{2.52}$$

ce qui conduit à l'équation de construction de $g[n]$ suivante :

$$g[n] = \langle \psi, \varphi_{-1,n}\rangle \tag{2.53}$$

$g[n]$ sera également considérée comme la réponse impulsionnelle d'un filtre numérique ; nous verrons que ce filtre est lié au filtre $h[n]$ et qu'il peut être construit à partir de ce dernier.
Un calcul analogue en tous points au précédent permet d'écrire les coefficients de détail :

$$d_n^j = \langle f, \psi_{j,n}\rangle \tag{2.54}$$

$$d_n^j = \sum_k g[k] \left\langle f, \varphi_{j-1,k+2n} \right\rangle \tag{2.55}$$

On introduit également le filtre symétrique dont la réponse impulsionnelle correspond à la séquence $g[n]$ retournée :

$$\tilde{g}[n] = g[-n] \tag{2.56}$$

La décomposition en ondelettes à l'échelle j pourra donc s'écrire :

$$d_n^j = \sum_l \tilde{g}[2n-l] \left\langle f, \varphi_{j-1,l} \right\rangle \tag{2.57}$$

ou encore :

$$d_n^j = \sum_l \tilde{g}[2n-l] a_l^{j-1} \tag{2.58}$$

Cette relation sera interprétée de la même manière que précédemment [21-24].
La figure 2.12 résume l'algorithme récursif d'analyse multirésolution de Mallat.

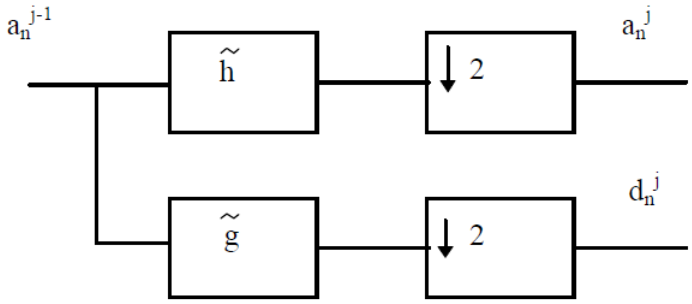


Figure 2.12. Algorithme d'analyse de Mallat.

2.3.3. Bancs de filtres :

2.3.3.1. Traitement en sous bandes :

La notion d'information contenue dans un signal est liée à sa représentation spectrale. Pour un signal quelconque ses caractéristiques, autres que sa forme temporelle, proviennent en grande partie de la répartition (de la forme) de sa densité spectrale. L'intérêt est d'avoir une idée de sa répartition de la quantité d'information dans chaque sous bande.
Les bancs de filtres sont un outil du traitement de signal qui permet, entre autre, d'obtenir une représentation particulière de l'information contenue dans un signal. Contrairement à la transformée de Fourier qui permet d'obtenir une représentation dans le domaine spectral, la représentation donnée par les bancs de filtres reste du domaine temporel. L'idée de base est

d'obtenir une série de signaux représentatifs d'une bande de fréquence du signal original. En simplifiant et en supposant que le filtre idéal existe, le signal est filtré par un ensemble de *M* filtres passe-bandes à supports disjoints (voir figure 2.13), ce qui permet d'obtenir *M* signaux correspondant chacun à une partie du spectre du signal original.

2.3.3.2. Découpage du spectre :

Le découpage du spectre en *M* sous-bandes peut s'effectuer de manière simple à partir de la représentation par transformée de Fourier, ici nous nous intéresserons au découpage à l'aide de filtres, et plus particulièrement de filtres à réponse impulsionnelle finie (RIF) auxquelles on peut donner des caractéristiques de phases intéressantes comme la phase linéaire, la minimalité de phase, etc ...

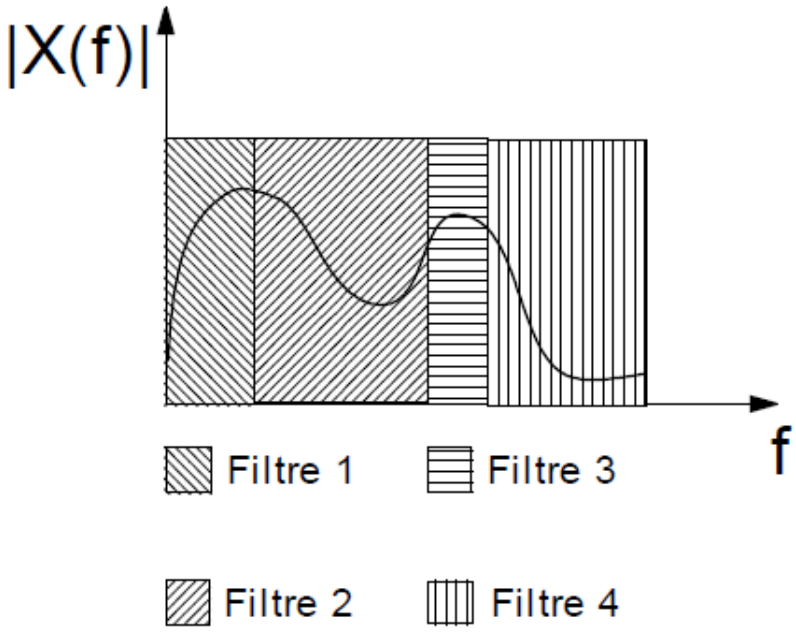


Figure 2.13. Découpage en 4 bandes.

Dans une telle représentation, à un échantillon de l'entrée correspondent 4 échantillons en sortie. Il y a redondance de l'information car chaque x_n^k n'occupe qu'une partie du spectre (figure 2.14). Les x_n^k sont des signaux à bande étroite, on doit donc pouvoir les échantillonner à une fréquence plus basse que la fréquence d'échantillonnage initiale sans perte d'information. Une simplification dans le traitement consiste à définir des largeurs de bandes identiques pour chaque filtre, ainsi pour un traitement en *M* voies définir la largeur de bande à $1/2M$ permet de ramener la fréquence d'échantillonnage dans chaque sous bande à $1/M$ (figure 2.15). Le sous échantillonnage provoque des repliements de spectre, mais sans recouvrement si les filtres sont idéaux.

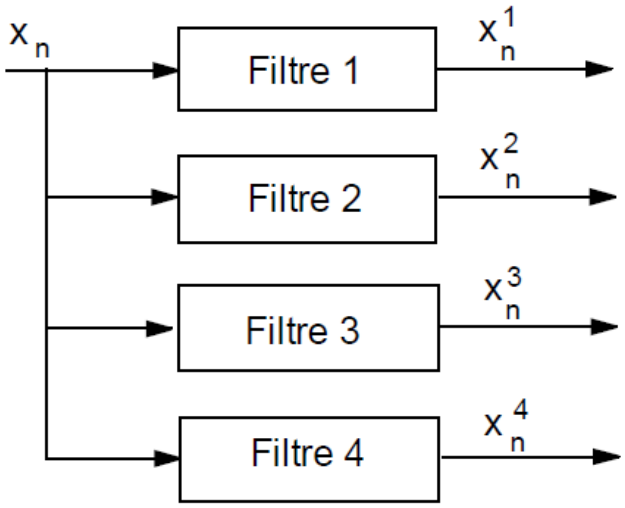


Figure 2.14. Banc de filtre à 4 voies.

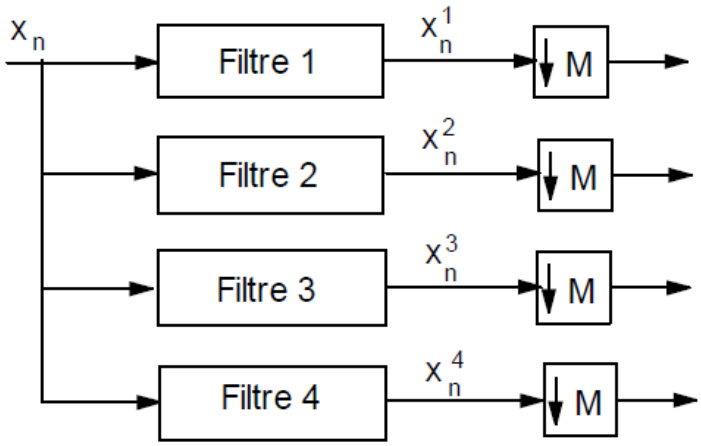


Figure 2.15. Sous échantillonnage.

Un tel banc de filtres sera dit uniforme (largeurs de bandes identiques) à décimation critique (facteur de sous échantillonnage égal au nombre de bandes).

Sachant que dans la pratique les filtres ne sont pas idéaux, il y aura repliement avec recouvrement au cours du sous échantillonnage. Cependant on dispose ici de *M* voies, donc *M* informations de repliement.

2.3.3.3. Caractéristiques des filtres :

Dans la plupart des applications, il s'agit de filtres RIF à phase linéaire. La réponse impulsionnelle est supposée être symétrique [25].

L'intérêt principal des filtres RIF réside dans la possibilité d'obtenir des filtres à phase linéaire. Nous allons voir ce que ceci signifie en terme de fonction de transfert.

- **Filtre RIF :**

Un filtre RIF possède une fonction de transfert polynomiale de la forme :

$$H(z)=\sum_{n=0}^{n-1}h_n z^{-n} \tag{2.59}$$

La transformée de Fourier correspondante est :

$$H\left(e^{j\omega}\right)=\sum_{n=0}^{N-1}h_n e^{-jn\omega}=\pm\left|H\left(e^{j\omega}\right)\right|e^{j\varphi(\omega)} \tag{2.60}$$

Où $\varphi(\omega)$ est fonction continue de ω, le signe $\pm$ signifie que la phase du filtre peut présenter des sauts de π ($e^{j\pi}=-1$). Les coefficients h_n étant réels,

on a les relations :

$$\left.\begin{array}{c}\left|H\left(e^{j\omega}\right)\right|=\left|H\left(e^{-j\omega}\right)\right| \\ \varphi(\omega)=-\varphi(-\omega)\end{array}\right\}\ 0\le\omega<\pi \tag{2.61}$$

Si nous voulons imposer la contrainte de phase linéaire, cela peut s'exprimer de deux façons :

$$\begin{aligned}&\text{Cas 1 : } \varphi(\omega)=-\alpha\omega \qquad -\pi\le\omega\le\pi \\ &\text{Cas 2 : } \varphi(\omega)=\beta-\alpha\omega \qquad -\pi\le\omega\le\pi\end{aligned} \tag{2.62}$$

où α est le temps de propagation de groupe en nombre d'échantillons.

- **Cas 1 :**

Le gain complexe s'exprime sous la forme:

$$H\left(e^{j\omega}\right)=\sum_{n=0}^{N-1}h_n e^{-jn\omega}=\pm\left|H\left(e^{j\omega}\right)\right|e^{-j\alpha\omega} \tag{2.63}$$

Soit par identification des parties réelles et imaginaires,

$$\begin{aligned}\pm\left|H\left(e^{j\omega}\right)\right|\cos(\alpha\omega)&=\sum_{n=0}^{N-1}h_n\cos(n\omega) \\ \pm\left|H\left(e^{j\omega}\right)\right|\sin(\alpha\omega)&=\sum_{n=1}^{N-1}h_n\sin(n\omega)\end{aligned} \tag{2.64}$$

Les h_n sont obtenus en effectuant le rapport des deux égalités de (2.39):

$$tg(\alpha\omega) = \frac{\sum_{n=1}^{N-1} h_n \sin(n\omega)}{h_0 + \sum_{n=1}^{N-1} \cos(n\omega)} \tag{2.65}$$

Deux cas sont à envisager:

1. $\alpha = 0$ qui se traduit par $h_n = 0$ pour $n \neq 0$ et h_0 arbitraire. C'est un simple gain sans intérêt.

2. $\alpha \neq 0$ dans ce cas (2.65) se met sous la forme :

$$\sum_{n=0}^{N-1} h_n \cos(n\omega) \sin(\alpha\omega) = \sum_{n=0}^{N-1} h_n \sin(n\omega) \cos(\alpha\omega)$$

qui peut se condenser en : $\sum_{n=0}^{N-1} h_n \sin\left((\alpha - n)\omega\right) = 0$.

Si l'équation admet une solution, elle est unique.

On peut vérifier assez facilement qu'une solution est donnée par:

$$\alpha = \frac{N-1}{2}$$

$$h_n = h_{N-1-n} \qquad 0 \leq n \leq N-1 \tag{2.66}$$

Les conséquences en sont:

1. à une valeur de *N*, correspond une valeur de α ,

2. pour cette valeur de α , la réponse présente une symétrie particulière.

- si *N* est impair, α est entier,
- si *N* est pair, α est non entier.

Exemples de réponses impulsionnelles:

- pour *N* = 11 (figure 2.16), on a α = 5 centre de symétrie = h_5

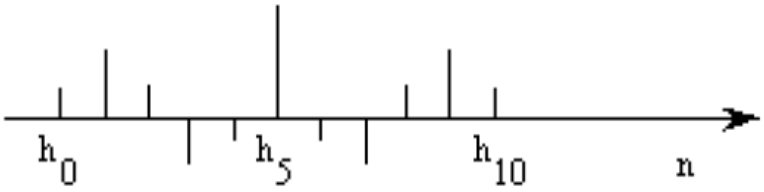


Figure 2.16. Réponse symétrique (*N* impair).

- pour *N* = 10 (figure 2.17), on a α = 4.5 centre de symétrie entre deux échantillons.

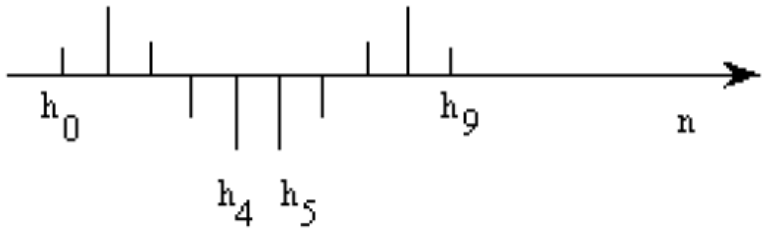


Figure 2.17. Réponse symétrique (*N* pair).

- **Cas 2 :**

Pour cette contrainte: $\varphi(\omega) = \beta - \alpha\omega$

Le raisonnement précédent conduit à la solution:

$$\alpha = \frac{N-1}{2}$$

$$\beta = \pm\frac{\pi}{2}$$

$$h_n = -h_{N-1-n} \qquad 0 \le n \le N-1 \tag{2.67}$$

Pour ce type de filtres la réponse impulsionnelle est antisymétrique (figures 2.18 et 2.19). Cette antisymétrie impose $h_{(N-1)/2} = 0$ pour N impair.

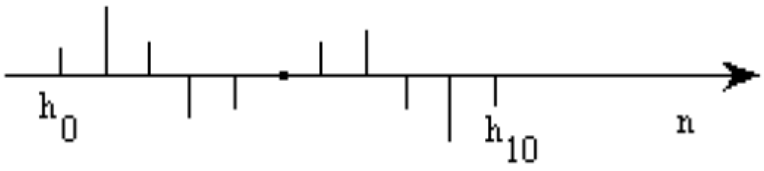


Figure 2.18. Réponse antisymétrique (*N* impair).

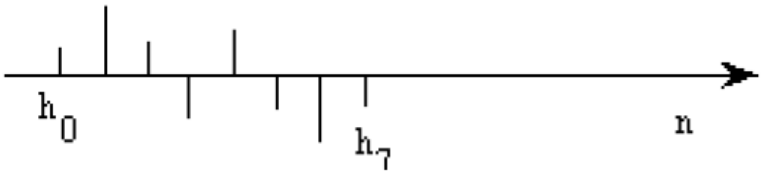


Figure 2.19. Réponse antisymétrique (*N* pair).

2.3.3.4. Comportement fréquentiel :

Ces filtres permettent, par des manipulations simples, de définir les réponses en fréquence et d'en déduire l'utilisation possible des filtres (passe-bas, passe-haut, etc...). La forme générale de la réponse en fréquence est donnée par :

$$H\left(e^{j\omega}\right)=\sum_{n=0}^{N-1}h_n e^{-jn\omega}=\pm\left|H\left(e^{j\omega}\right)\right|e^{j(\beta-\alpha\omega)} \tag{2.68}$$

- **Réponse symétrique (N impair) :**

Les h_n vérifient $h_n=h_{N-1-n}$, et pour cette forme $\alpha=(N-1)/2$, soit:

$$H\left(e^{j\omega}\right)=e^{-j\omega\frac{(N-1)}{2}}\sum_{n=0}^{\frac{N-1}{2}}a_n\cos\left(n\omega\right) \tag{2.69}$$

Où : $\pm\left|H\left(e^{j\omega}\right)\right|=\sum_{n=0}^{\frac{N-1}{2}}a_n\cos\left(\omega\right)$ et $a_0=h_{N-1}/2$ et $a_n=2h_{\frac{N-1}{2}-n}$ pour $(n=1,...,(N-1)/2)$.

2.3.3.5. Méthodes de synthèse des filtres RIF :

Les méthodes de synthèse que nous allons considérer permettent de satisfaire des contraintes de réponses en amplitudes, sachant que par raison de symétrie de la réponse impulsionnelle, la phase résultante sera linéaire en fréquence.

On définit un gabarit de réponse en amplitude de la façon suivante (figure 2.20):

- $\Delta f=f_2-f_1$ bande de transition,
- $R_c=\dfrac{f_1+f_2}{\Delta f}$ raideur de coupure,
- δ_1 : ondulation tolérée en bande passante,
- δ_2 : ondulation tolérée en bande affaiblie.

Tout filtre dont la réponse en amplitude se situe dans le gabarit défini satisfait les contraintes spécifiées. Pour des raisons de facilité de réalisation et de réduction en coût de calcul, le meilleur filtre sera celui d'ordre minimum.

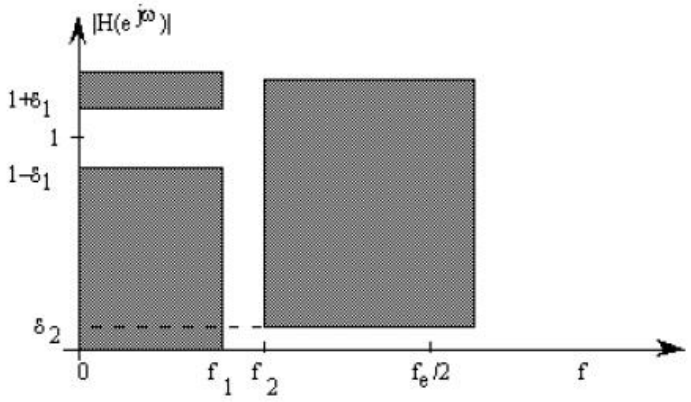


Figure 2.20. Définition du gabarit.

- **Méthode de la fenêtre :**

La réponse en fréquence d'un filtre numérique (figure 2.21) est périodique de période 1. Il est alors facile de l'exprimer sous forme de série de Fourier, série dont les coefficients h_n sont ceux de la réponse impulsionnelle:

$$H\left(e^{j\omega}\right)=\sum_{n=-\infty}^{+\infty}h_n e^{-j\omega n} \tag{2.70}$$

Avec:

$$h_n=\frac{1}{2\pi}\int_{-\pi}^{+\pi}H\left(e^{j\omega}\right)e^{j\omega n}d\omega=\int_{-1/2}^{1/2}H\left(e^{j2\pi f}\right)e^{j2\pi nf}df \tag{2.71}$$

Dans (2.71) on prendra $H\left(e^{j2\pi f}\right)=\left|H\left(e^{j2\pi f}\right)\right|$; c'est à dire une phase nulle. Avec le gabarit idéal de la figure 2.21:

$$h_n=\int_{-f_c}^{f_c}e^{j2\pi nf}df \tag{2.72}$$

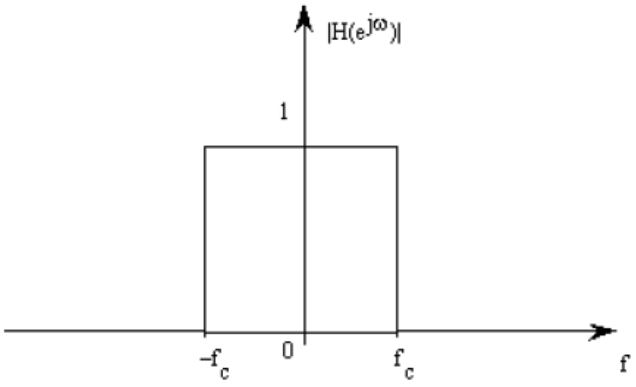


Figure 2.21. Gabarit idéal.

Les h_n obtenus sont définis pour $n \in \mathbb{Z}$ (figure 2.22), ce qui n'est pas utilisable de manière directe en pratique.

Pour obtenir un filtre de longueur ou d'ordre fini on tronquera le nombre des h_n utilisés :

$$h_n = 0 \qquad |n| > \frac{N-1}{2}$$

On obtient:

$$H(z) = h_0 + \sum_{n=1}^{\frac{N-1}{2}} \left(h_{-n} z^n + h_n z^{-n} \right)$$

La causalité est obtenue en multipliant $H(z)$ par le terme $z^{-\frac{N-1}{2}}$ ce qui introduit un retard de $\frac{N-1}{2}$, donc une phase linéaire, mais ne change pas la réponse en amplitude:

$$H_{causale}(z) = z^{-\frac{N-1}{2}} H(z) \tag{2.73}$$

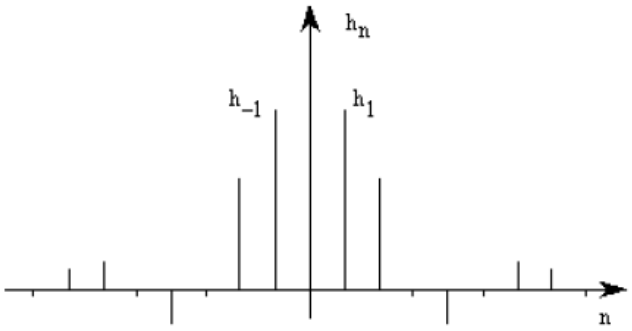


Figure 2.22. Réponse impulsionnelle.

- **Effet de la troncature :**

Tronquer le nombre de coefficients h_n revient à multiplier par une fenêtre rectangulaire (figure 2.23) définie par:

$$\begin{cases} w_n = 1 & |n| \le \dfrac{N-1}{2} \\ w_n = 0 & \textit{ailleurs} \end{cases}$$

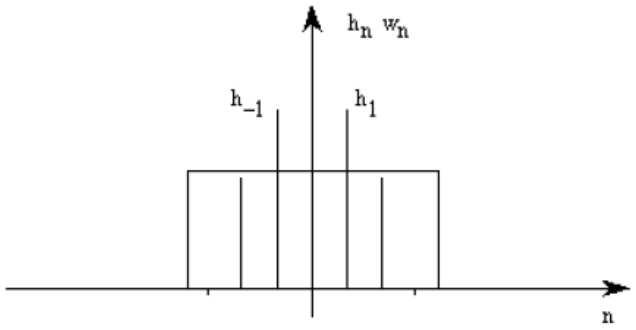


Figure 2.23. Troncature temporelle.

Ce qui traduit la convolution de $H\left(e^{j\omega}\right)$ avec $W_R\left(e^{j\omega}\right)$. Or pour un filtre passe-bas $H\left(e^{j\omega}\right)$ vaut : $H\left(e^{j\omega}\right)=\begin{cases}1 & 0\le\omega\le\omega_c\\ 0 & \omega_c\le\omega\le\pi\end{cases}$ où ω_c définit la pulsation de coupure du filtre.

Et $W_R\left(e^{j\omega}\right)$ qui est la transformée de Fourier d'une fonction rectangulaire est un rapport de sinus. La convolution introduit donc sur $H\left(e^{j\omega}\right)$ des ondulations qui sont dues à la forme de $W_R\left(e^{j\omega}\right)$.

$$W_R\left(e^{j2\pi f}\right)=\sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} e^{-j2\pi nf}=\frac{\sin\left(N\pi f\right)}{\sin\left(\pi f\right)} \tag{2.74}$$

La figure (2.24) traduit l'effet de la fenêtre rectangulaire sur un filtre passe-bas, c'est une représentation du phénomène de Gibbs caractéristique de la convergence en moyenne quadratique des séries de Fourier de fonctions discontinues.

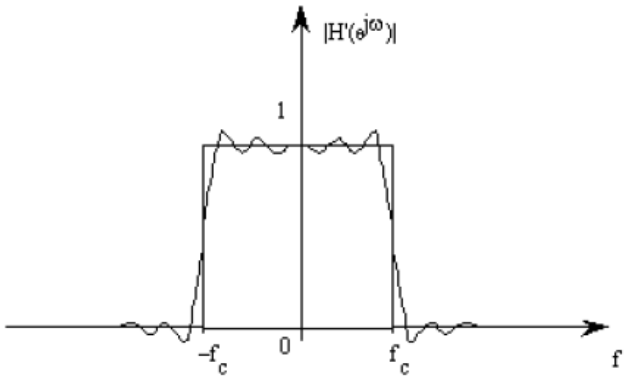


Figure 2.24. Phénomène de Gibbs.

- **Définition de fenêtres :**

Les fenêtres utilisées pour la pondération des coefficients d'un filtre RIF doivent conserver la symétrie de la réponse impulsionnelle, autrement dit doivent vérifier $w_0 = w_{N-1}$ dans leur version causale. On donne l'expression temporelle de quelques fenêtres couramment utilisées dans la synthèse de filtres.

Fenêtre de Hanning : $w_n = 0.5 - 0.5\cos(\frac{2\pi n}{N})$ $\quad |n| \le \frac{N-1}{2}$

Fenêtre de Hamming : $w_n = 0.54 - 0.46\cos(\frac{2\pi n}{N})$ $\quad |n| \le \frac{N-1}{2}$

Fenêtre de Kaiser : $w_n = \frac{\frac{1}{2} I_0\left[\alpha\sqrt{1-(\frac{2n}{N})^2}\right]}{I_0(\alpha)}$ $\quad |n| \le \frac{N}{2}$

Le paramètre α de la fenêtre de Kaiser contrôle l'atténuation du filtre passe-bas. $I_0(x)$ est la fonction de Bessel modifiée et les conceptions pratiques utilisent environ 20 termes de :

$$I_0(x) = 1 + \sum_{k=1}^{\infty}\left[\frac{(0.5x)^2 k}{k!}\right]^2$$

2.3.3.6. Conversion de fréquence :

La conversion de fréquence, au sens changement de la fréquence d'échantillonnage, a de nombreuses applications.

- **Réduction d'un facteur M :**

Soit un signal *x(n)* correspondant à un échantillonnage à la fréquence 1, et supposons que son spectre occupe toute la bande $\left(\left|X\left(e^{j2\pi f}\right)\right|\right) \neq 0, f \in \left[-\frac{1}{2}, \frac{1}{2}\right]$.

La réduction de fréquence d'un facteur *M* faisant passer la bande d'échantillonnage de $\left[-\frac{1}{2}, \frac{1}{2}\right]$ à $\left[-\frac{1}{2M}, \frac{1}{2M}\right]$, il est nécessaire de filtrer avant de sous échantillonner pour éviter les phénomènes de repliements dus à la nouvelle périodisation du spectre (figure 2.26).

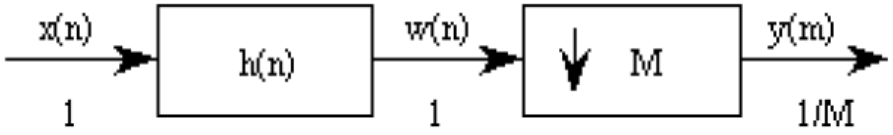


Figure 2.25. Principe de réduction de fréquence.

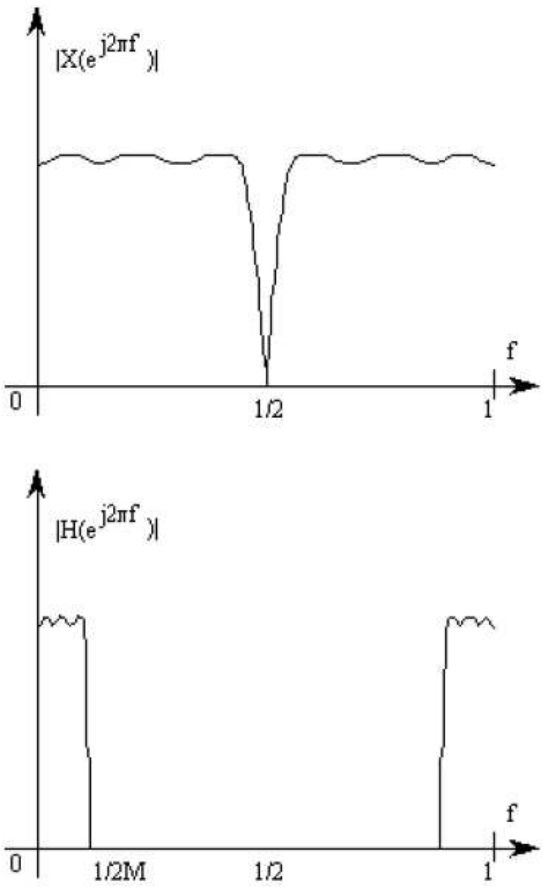


Figure 2.26. Filtrage antirepliement.

Le filtre est défini par:

$$\left|H\left(e^{j2\pi f}\right)\right| = \begin{cases} 1 \ \ pour & |f| < \dfrac{1}{2M} \\ 0 \ \ ailleurs & \end{cases} \tag{2.75}$$

A la sortie du filtre on obtient la séquence

$$w(n) = \sum_{k=-\infty}^{+\infty} h(k)x(n-k)$$

et la séquence $y(m)$ s'en déduit par: $y(m) = w(mM)$.

La relation entrée/sortie s'écrit:

$$y(m) = \sum_{k=-\infty}^{+\infty} h(k)\,x(mM-k)$$

Le spectre de $y(m)$ est donné par la figure (2.27).

L'ensemble filtrage/sous échantillonnage n'est pas invariant dans le temps, puisque pour une séquence $x(n)$ donnée il y a M façons de calculer la séquence $y(m)$. En d'autres termes si on sous échantillonne le signal $x(n-p)$ alors la sortie n'est pas $y(m-\frac{p}{M})$ sauf si $p = qM$, $q \in \mathbb{Z}$.

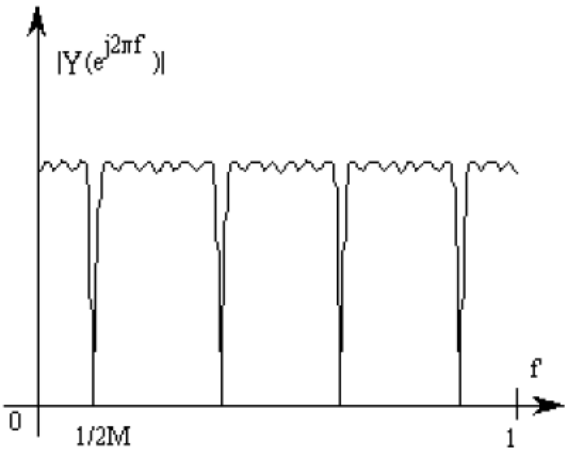


Figure 2.27. Spectre du signal après sous-échantillonnage.

Etablissons les relations entre les transformées en *z* desquelles on déduira les relations entre les transformées de Fourier.

Soit la séquence *v(n)* telle que:

$$v(n)=\begin{cases} w(n) & pour\ n=0,\pm M,\pm 2M,... \\ 0 & ailleurs \end{cases} \tag{2.76}$$

Cela se traduit par:

$$v(n)=w(n)\sum_{m=-\infty}^{+\infty}\delta(n-mM) \tag{2.77}$$

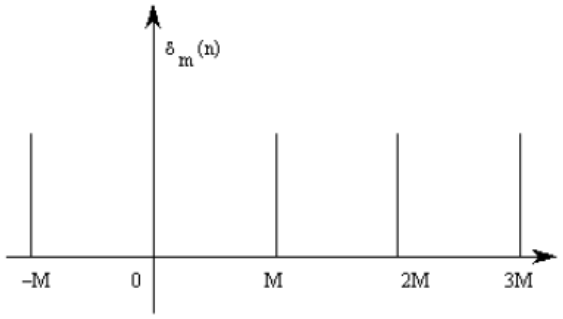


Figure 2.28. Train d'impulsions équivalent pour le sous-échantillonnage.

Où $\sum_{m=-\infty}^{+\infty}\delta(n-mM)$ (figure 2.28) est un train d'impulsions unités périodiques qui peut encore s'exprimer sous la forme:

$$\sum_{m=-\infty}^{+\infty}\delta\left(n-mM\right)=\frac{1}{M}\sum_{r=0}^{M-1}e^{j2\pi\frac{r}{M}n} \tag{2.78}$$

On obtient par substitution $v\left(n\right)=w\left(n\right)\frac{1}{M}\sum_{r=0}^{M-1}e^{j2\pi\frac{r}{M}n}$ et comme $y\left(m\right)=v\left(mM\right)=w\left(mM\right)$, il s'ensuit:

$$Y\left(z\right)=\sum_{m=-\infty}^{+\infty}v\left(mM\right)z^{-m}=\sum_{m=-\infty}^{+\infty}v\left(m\right)z^{-\frac{m}{M}} \tag{2.79}$$

Remplaçant *v(m)* par son expression en fonction de *w(m)* :

$$Y\left(z\right)=\frac{1}{M}\sum_{r=0}^{M-1}W\left(e^{-j2\pi\frac{r}{M}}z^{\frac{1}{M}}\right) \tag{2.80}$$

Compte tenu de ce que $W\left(z\right)=H\left(z\right)X\left(z\right)$, l'expression de $Y(z)$ s'écrit finalement:

$$Y\left(z\right)=\frac{1}{M}\sum_{r=0}^{M-1}H\left(e^{-j2\pi\frac{r}{M}}z^{\frac{1}{M}}\right)X\left(e^{-j2\pi\frac{r}{M}}z^{\frac{1}{M}}\right) \tag{2.81}$$

Soit sur le cercle unité $z=e^{j2\pi f}$, $f\in\left[-\frac{1}{2},\frac{1}{2}\right]$:

$$Y\left(e^{j2\pi f}\right)=\frac{1}{M}\sum_{r=0}^{M-1}H\left(e^{j2\pi\frac{\left(f-r\right)}{M}}\right)X\left(e^{j2\pi\frac{\left(f-r\right)}{M}}\right) \tag{2.82}$$

Dans le cas d'un filtre idéal, $H\left(e^{j2\pi\frac{\left(f-r\right)}{M}}\right)=0$ pour $r\neq 0$, d'où:

$$Y\left(e^{j2\pi f}\right)=\frac{1}{M}X\left(e^{j2\pi\frac{f}{M}}\right)$$

Dans la pratique, le filtre passe-bas doit faire en sorte que les composantes en fréquence pour $\left|f\right|>\frac{1}{2M}$ soient négligeables. Ce qui signifie que tous les termes tels que $r\neq 0$ sont presque nuls, d'où: $Y\left(e^{j2\pi f}\right)\approx\frac{1}{M}X\left(e^{j2\pi\frac{f}{M}}\right)$.

Si le spectre de la séquence $x(n)$ était effectivement limité à la bande $\left[-\frac{1}{2M},\frac{1}{2M}\right]$ alors on aurait exactement: $Y(z)=\frac{1}{M}X\left(z^{\frac{1}{M}}\right)$.

2.3.3.7. Banc de filtres à M voies :

La figure (2.29) donne une réalisation de banc de filtres. Dans la mesure où chaque sous-bande est traitée à une fréquence *M* fois plus petite que la fréquence d'échantillonnage d'entrée et qu'il y a *M* sous-bandes, il y a conservation de la quantité d'informations.

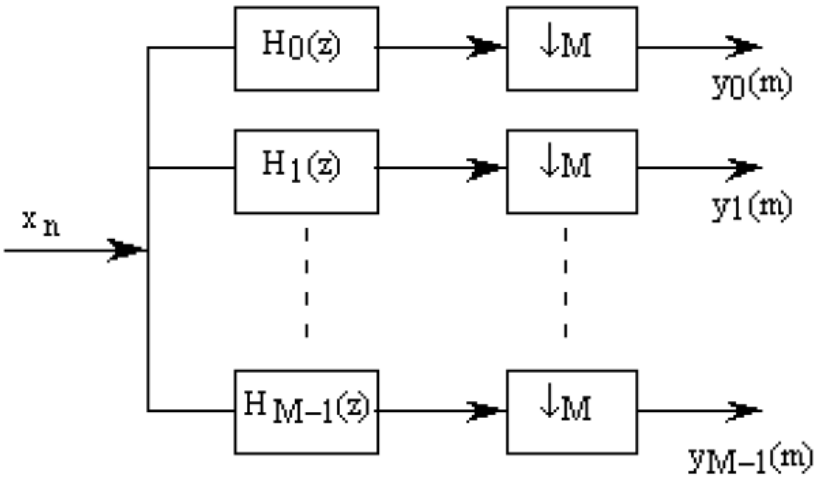


Figure 2.29. Principe du traitement en sous-bandes.

Dans la structure de la figure 2.29 toutes les voies sont échantillonnées de manière synchrone, c'est à dire que tous les *M* échantillons d'entrée, on calcule les $y_r(m)$. Dans la mesure où l'on travaille à la fréquence $\frac{f_e}{M}$ on peut traiter les $y_r(m)$ de manière séquentielle au rythme de f_e. Lorsque $n = mM$ on a affaire à un échantillon dans la bande $M-1$ pour $y_{M-1}(m)$, l'instant suivant $n=mM+1$ c'est pour $y_{M-2}(m)$ et ainsi de suite jusqu'à l'instant $n=mM+M-1$ où l'on calcule $y_0(m)$. A l'échantillon suivant on recommence.

Soit alors :

$$y_r(m)=\sum_{n=-\infty}^{+\infty}x_n h_r(mM-n)$$

2.4. Conclusion :

Dans ce chapitre nous avons donné une présentation courante des outils mathématiques utilisés. Dans ce cas sont la décomposition du signal à l'aide de la transformée de Fourier, de la théorie des ondelettes ainsi qu'une révision des équations de base des bancs de filtres.

La transformée de Fourier fait partie des outils fondamentaux dans le traitement du signal. L'analyse de la transformée en ondelettes représente un intérêt particulier car elle décompose les signaux à une seule dimension dans un plan bidimensionnel. Cette transformée fournit des informations du signal dans le temps et simultanément à travers les domaines par une série d'opérations de convolution entre le signal en cours d'analyse et d'intégration en ondelettes de la fréquence du signal, en tenant compte de divers facteurs d'échelle.

Les bancs de filtres forment une classe d'outils fondamentaux en traitement du signal. Dans beaucoup d'applications il peut être intéressant de séparer le signal d'entrée en plusieurs composantes en sous-bandes. Ceci permet en effet de situer la ou les bandes de fréquence où l'on peut trouver l'information.

Chapitre 3

La prédiction de gènes

3.1. Introduction :

Les séquences nucléiques (ADN ou ARN) comme les séquences protéiques, comportent de l'information intrinsèque qui peut être l'objet de nombreuses analyses.

Une des principales tâches de la bioinformatique, est de créer des logiciels de gestion et d'analyse de l'information, qui permettent d'annoter de façon efficace et complète chaque portion de génomes (fonction, structure des protéines potentiellement produites, taux d'expression d'un gène aux différentes étapes de la vie d'un organisme…).

La principale information contenue dans le génome, c'est les gènes eux-mêmes. En présence d'une information brute comme la séquence d'un génome, le biologiste cherchera en première analyse à identifier les différents gènes qu'elle contient, pour étudier les protéines à partir de ceux-ci.

La suite de ce chapitre aborde la description des différents algorithmes performants développés au moyen d'un formalisme très complet pour analyser le génome.

3.2. Recherche de gènes :

Lorsqu'une grande portion d'ADN a été cartographiée et séquencée, l'objectif suivant est de donner un sens à cette séquence brute, de dégager sa ou ses fonctions. Dans ce but la recherche de gènes puis la caractérisation de ceux-ci est entreprise.

La recherche exhaustive de tous les gènes contenus dans un grand génome est en général une tâche complexe. Chez les eucaryotes supérieurs, celle-ci est compliquée par la « dilution » de l'information pertinente dans les séquences non codantes : ADN répété, régions intergéniques. La présence d'introns de grande longueur séparant des petits exons rend assez délicat l'assemblage exact des zones codantes [1].

Le problème de l'identification des gènes dans les séquences d'ADN génomique par des méthodes computationnelles a attiré beaucoup d'attention de la part des chercheurs. Un certain nombre d'approches ont été développées qui intègrent plusieurs types d'informations incluant des capteurs de signal d'épissage (le processus de recherche de signaux limites, c'est-à-dire les codons d'initiation /de terminaison et les sites d'épissage, est appelé détection de signal 'signal sensing en anglais'), des propriétés de composition d'ADN codant et non codant et

dans certains cas une recherche d'homologie dans les bases de données pour prédire des structures entières du gène dans les séquences génomiques. Les méthodes de comparaison de séquences deux à deux sont au cœur des analyses bioinformatiques. Elles permettent notamment d'interroger des bases de données à partir d'une séquence requête. De ce fait, la comparaison de séquences est souvent la première analyse vers l'attribution d'une fonction potentielle à une séquence. En propageant ainsi une information d'un génome à un autre, ces méthodes permettent d'identifier les caractéristiques de familles de protéine, de construire des modèles par homologie… [2].

Les programmes de détection de gènes dans une séquence d'ADN génomique non annotée utilisent des approches variées. Ceux qui donnent les meilleurs résultats sont ceux qui combinent les approches basées sur la reconnaissance de texture et les approches basées sur la reconnaissance de motifs. Les méthodes de recherche de texture s'appuient sur le fait que la distribution des nucléotides dans une région codante est différente de celle d'une région non codante, c'est une recherche par le contenu (content sensor), qui analyse les propriétés statistiques de la distribution des nucléotides ou des acides aminés dans la séquence. Par contre, la reconnaissance des motifs se ramène à la recherche sur une séquence d'ADN certaines séquences nucléotidiques ou d'un site correspondant à une fonction biologique. Un motif peut être considéré comme une séquence d'acides aminés qui définit une sous-structure dans une protéine pouvant être reliée à une fonction ou un groupement structural. Par exemple, les programmes de la famille GRAIL (Gene Recognition and Analysis Internet Link) (1994) développés par l'Oak Ridge National Laboratories utilisent les réseaux neuronaux pour combiner différentes données statistiques. Il localise les gènes dans une séquence d'ADN anonyme en reconnaissant les caractéristiques liées aux régions codant les protéines et aux limites des régions codantes, puis combine les caractéristiques reconnues en utilisant un système de réseau neuronal [8]. Les réseaux neuronaux également appelés modèles connexionnistes sont des modèles statistiques utilisés en reconnaissance de forme et en classification originalement développés dans les années 40 comme modèles mathématiques de mémoire. Un réseau de neurones est construit de nœuds physiques et de connexions. L'idée qui se cache derrière les réseaux neuronaux est que, en travaillant tous ensemble, ces petits éléments peuvent accomplir des taches complexes. Un réseau de neurones est composé d'un ensemble de nœuds qui sont connectés sur la base d'une topologie fixée, où chaque nœud possède une entrée et une sortie vers d'autres nœuds. Les réseaux neuronaux sont utilisés de façon intensive dans certains problèmes de bioinformatique : par exemple, les programmes de prédiction et recherche de gènes [2].

Parmi les programmes couramment utilisés pour la prédiction de gènes, on peut citer GENSCAN (1997) qui combine l'information statistique sur le contenu en s'appuyant sur un modèle probabiliste de structure de gènes. C'est un outil de recherche des gènes dans les génomes très élaboré basé sur des analyses statistiques pour construire des processus de Markov caché [26]. En effet, en soumettant une séquence d'ADN à des différents tests statistiques des fréquences relatives des nucléotides sont obtenues. Les fréquences obtenues sont utilisées pour calculer des probabilités équivalentes. Pour fabriquer une séquence qui reproduise ces statistiques, une approche consiste à procéder itérativement. En utilisant un générateur de nombres aléatoires et en utilisant ces probabilités, il est possible d'engendrer progressivement une pseudo-séquence reproduisant les biais observés dans les séquences considérées. Cette méthode de fabrication aléatoire où l'état i+1 est produit à partir de l'état i en utilisant une table de probabilité est bien connue dans le domaine des mathématiques appliquées, elle s'appelle une chaine ou processus de Markov. Leur portée dans le domaine de l'analyse des séquences biologiques est cependant est beaucoup plus vaste. En raffinant ce type de modèle et en lui faisant prendre en compte les fréquences dans les régions non codantes, les codons de démarrage et de terminaison, on peut en faire un outil de recherche automatique des gènes dans les génomes [2].

Il existe un très grand nombre de programme de détection de gènes, les programmes très connus : Genemark (1993), Glimmer (1998), FGENEH (1994), GeneID (1992), Genie (1996), GeneParser (1995), SORFIND (1992), GenLang (1994), Xpound (1994).

Les systèmes de recherche de gènes, tels que le système GeneParser, sont construits en utilisant la programmation dynamique et les réseaux de neurones. Les méthodes de programmation dynamique décrivent une classe générique d'algorithmes utilisés comme une méthode générale d'optimisation. Les algorithmes de programmation dynamique résolvent des problèmes d'optimisation, problèmes possédant un grand nombre de solutions possibles, parmi lesquelles une ou un petit nombre de meilleures solutions à extraire. Un algorithme de programmation dynamique trouve la meilleure solution en réduisant le problème original en un ensemble de sous-problèmes plus petits et plus faciles à résoudre. La programmation dynamique commence donc par résoudre des sous-problèmes puis stocke chaque solution intermédiaire dans un tableau avec un score et finalement choisit la séquence de solutions qui correspond au score maximal [2].

GeneLang, similaire à GeneParser, est un système de reconnaissance de formes syntaxiques, qui utilise les outils et les techniques linguistiques computationnelles pour trouver des gènes et d'autres caractéristiques d'ordre élevé dans les données de séquences biologiques. Par

exemple, la théorie du langage formel utilise un ensemble de règles, appelé grammaire, pour définir les ensembles de chaînes valides sur un alphabet donné. La motivation pour construire une grammaire de gène est la disponibilité des analyseurs qui peuvent reconnaître si la chaîne d'entrée satisfait aux règles spécifiées par une grammaire de gène. Ainsi, dans GeneLang, les motifs sur l'alphabet d'ADN sont décrits en utilisant un ensemble de règles et un analyseur à usage général, implémenté dans le langage de programmation logique Prolog. Ainsi, le système traite les composants d'un gène tels que les sites donneurs et accepteurs, les introns et les exons, les codons de départ et d'arrêt, etc. comme des mots qui sont formés sur l'alphabet d'ADN à quatre caractères. Le gène suivant dans le niveau de cette hiérarchie syntaxique, c'est-à-dire le gène, est une phrase valide formée par ces mots. Les gènes d'une séquence d'ADN peuvent être reconnus d'une manière analogue à la reconnaissance de phrases grammaticalement correctes en langue anglaise [27].

Cependant, Xpound a été développé pour détecter des régions codantes dans les séquences d'ADN par le modèle probabiliste de Markov. Les résultats sont présentés en termes de probabilité pour chaque nucléotide dans une séquence à coder. Tandis que le programme SORFIND vise à la prédiction des exons à travers une analyse des cadres de lecture ouverts (ORF pour Open Reading Frame) épissables [28].

Plusieurs programmes ont été conçus utilisant différents types de modèles de Markov afin de capturer les différences de composition entre les régions codantes et les régions non codantes de l'ADN. Ces programmes de prédiction de gène dans une séquence d'ADN génomique ont atteint un état de maturité relative avec le développement de systèmes basés sur des modèles de Markov cachés (par exemple GENSCAN et Genie) et des modèles apparentés (par exemple GeneID, FGENESH et l'algorithme largement utilisé GENEMARK) [26]. La plupart de ces programmes ont été mis à la disposition du public via des serveurs Internet.

Burset & Guigo (1996) dans [29] offre une introduction à la découverte de gènes par ordinateur et met en évidence certaines des forces et des faiblesses des méthodes actuellement disponibles. La question de la précision prédictive de ces méthodes a été abordée par une comparaison exhaustive des méthodes disponibles en utilisant un grand ensemble de séquences de gènes de vertébrés construits par Burset et Guigo. Cet ensemble de données de 570 gènes provenant de nombreux organismes différents a été utilisé dans [29] pour comparer l'efficacité de nombreux différents programmes de détection de gènes. Les auteurs concluent que la précision prédictive de tous ces programmes reste plutôt faible, moins de 50% des exons étant identifiés exactement par la plupart des programmes. Ainsi, le développement de nouvelles méthodes (et / ou l'amélioration des méthodes existantes) continue d'être important.

Plus récemment, une analyse comparative des techniques de calcul logiciel pour les programmes de prédiction de gènes est réalisée (2013) [3]. Au cours de cette analyse, les programmes sont testés sur des séquences issues de l'ensemble de données HMR195. La performance de RBFN (radial basis function network) Combining (2007) est comparée à celle de deux programmes de prédiction de gènes populaires (AUGUSTUS et FGENESH). La méthode de combinaison RBFN combine les prédictions de trois outils de recherche de gènes populaires (GENSCAN, HMMgene et Glimmer). Un réseau de neurones est utilisé pour combiner les paramètres de précision de ces outils. En utilisant un algorithme génétique, les paramètres pondérés équitables de la RBFN sont calculés. Enfin, l'évaluation intégrative des outils est réalisée à l'aide d'un réseau de neurones entraînés. Les résultats de cette combinaison montrent que la méthode proposée est efficace pour combiner des programmes de recherche de gènes et atteint une plus grande précision au niveau de l'exon que comme un outil de prédiction de gène seul [3].

Glimmer est un système de recherche de gènes largement utilisé dans la communauté génomique procaryote. En utilisant une technique de calcul appelée modèles de Markov interpolés (IMM), Glimmer peut trouver environ 98% de tous les gènes dans le génome bactérien. Cependant que, le système de modèle de Markov caché HMMgene est le système de recherche de gène qui s'est également prouvé efficace pour la découverte de gènes humains. L'approche d'AUGUSTUS est la prédiction de gène avec un modèle de Markov caché et un nouveau sous-modèle d'intron. Chaque programme de détection de gènes a ses forces et ses faiblesses, et certains systèmes semblent mieux fonctionner sur des organismes spécifiques [30]. Ces programmes sont souvent optimisés pour le travail sur des séquences appartenant à un même organisme (*E. coli*, levure, homme, drosophile, *arabidopsis thaliana*..) ou bien appartenant à une même catégorie d'organismes (eucaryotes, vertébré, procaryotes..) que les séquences tests à partir desquelles les programmes ont été développés [2].

Les programmes sont évalués dans la référence [3] sur toutes les séquences de l'homme, de la souris et du rat à partir de l'ensemble de données HMR195. Les résultats indiquent que les outils de prédiction de gènes combinés à l'aide de techniques de calcul souple sont comparables aux prédicteurs de gènes populaires. Les auteurs ont rapporté dans [3] que les outils de prédiction de gènes basés sur des techniques de calcul souple ont de bonnes performances. Les techniques informatiques douces, en particulier les réseaux neuronaux, semblent être un outil puissant dans la prédiction génétique. Les techniques hybrides donnent des résultats prometteurs, mais ils sont appliqués de manière très limitée.

3.3. Techniques de calcul souples pour la prédiction des gènes :

Le calcul souple est l'approche moderne de la construction d'un système de calcul intelligent. Le calcul souple est le mélange de méthodologies qui fournissent des capacités de traitement flexible de l'information pour traiter les problèmes du monde réel. De nos jours, les techniques de calcul souple sont identifiées comme des alternatives intéressantes aux méthodes de calcul intensif (hard computing methods). Les méthodes de calcul intensif traditionnelles sont souvent peu pratiques pour les problèmes du monde réel. Ils ont toujours besoin d'un modèle systématique précis et ont souvent besoin de beaucoup de temps de calcul. Contrairement aux méthodes de calcul intensif, les méthodes de calcul souple permettent de résoudre les problèmes d'imprécision, d'incertitude, d'apprentissage et d'approximation pour parvenir à la flexibilité, à la robustesse, aux solutions peu coûteuses et à la prise de décision humaine.

Certaines propriétés des techniques de calcul souple les rendent appropriées pour les tâches de séquençage. Ces techniques peuvent être facilement adaptées aux conditions changeantes. Ils sont capables de traiter des ensembles de données très volumineuses avec des données manquantes et bruitées et peuvent être utilisés pour extraire des relations cachées à partir de ces données. Une propriété unique de calcul souple est qu'elle est profondément impliquée dans l'apprentissage à partir de données expérimentales, ce qui la rend apte à la prédiction de gènes. Au cours de la prédiction des gènes, des motifs spécifiques dans la séquence d'ADN sont reconnus. Des techniques de calcul souple ont été largement utilisées dans les problèmes de reconnaissance de motifs. Le calcul souple comprend plusieurs techniques, les plus importantes étant les réseaux neuronaux, les algorithmes génétiques et la logique floue. L'importance des techniques de calcul souple réside dans le fait qu'elles sont complémentaires et non compétitives. Dans de nombreux cas, un problème peut être résolu en utilisant un réseau de neurones, logique floue et algorithme génétique en combinaison plutôt qu'une seule technique [3].

La section suivante décrit l'application de ces techniques de calcul logiciel dans le domaine de la prédiction de gènes

3.3.1. Systèmes hybrides :

Un système hybride intègre deux ou plusieurs techniques pour résoudre un problème. L'exemple le plus courant comprend un réseau de neurones combiné à une logique floue. La logique floue est une technique relativement nouvelle basée sur une logique à valeurs multiples qui permet de définir plusieurs valeurs entre des valeurs conventionnelles telles que 0 et 1. Elle fournit une méthode pour traiter l'imprécision et l'incertitude. L'idée principale

derrière la logique floue est d'approximer la prise de décision humaine en utilisant des termes de langage naturel au lieu de termes quantitatifs. L'un des plus grands avantages de la logique floue est qu'elle simplifie les systèmes complexes. Le système hybride est connu sous le nom de systèmes neuro-flous. Les systèmes neuro-flous ont été utilisés récemment dans la prédiction de gènes [3].

3.3.2. Systèmes neuro-flous :

Une nouvelle approche pour prédire les sites d'épissage basée sur un système d'inférence floue basé sur un réseau adaptatif (ANFIS) est discutée dans la référence [31]. Ici, les données de séquence sont divisées en trois ensembles de données en utilisant cinq stratégies de prétraitement: coder les nucléotides, extraire les propriétés statistiques, ignorer les caractéristiques faiblement corrélées, normaliser les modèles et réduire les caractéristiques redondantes. Enfin, le réseau est formé en utilisant différents algorithmes d'apprentissage. L'ANFIS surpasse les algorithmes de classification bien connus. Une tentative récente d'utilisation du réseau neuro-flou pour prédire des sites d'épissage est présentée dans la référence [32]. Le réseau neuro-flou est également basé sur l'ANFIS. Contrairement à l'approche précédente, des exemples de séquences de sites d'épissage vrai et faux sont utilisés ici pour définir des règles floues. La plus grande contribution de cette méthode est d'atteindre une précision de prédiction élevée en utilisant des réseaux neuro-flous plus petits.

3.4. Orientations du travail :

L'intérêt de ce travail de thèse est de construire un système prédictif prenant appui sur l'intégration des réseaux de neurones et des systèmes d'inférence floue. L'utilisation conjointe des réseaux de neurones et de la logique floue, permet de tirer les avantages des deux méthodes : les capacités d'apprentissage de la première et la lisibilité et la souplesse de la seconde.

L'approche développée dans cette thèse décrit un système de localisation de gène. Ce système localise des gènes dans une séquence d'ADN des organismes eucaryotes et procaryotes en reconnaissant des caractéristiques liées à des régions codantes et leurs limites.

Un certain nombre de mesures de codage ont été proposées sur la base de la fréquence du nucléotide d'un segment de séquence de longueur fixe. Le réseau neuro-flou est entraîné en se basant sur des données empiriques. Ensuite ces mesures sont introduites au réseau pour l'évaluation finale des séquences d'ADN candidates.

Par ailleurs, une méthode plus empirique est proposée pour résoudre ce type de problème de prédiction. Des algorithmes d'analyse spectrale ont été proposés, basés sur la transformée de Fourier, la transformée en ondelettes et banc de filtres, pour identifier des régions codant pour

les protéines dans des séquences d'ADN. L'analyse des séquences d'ADN exploitent les observations empiriques sur le spectre des séquences d'ADN.

Une étape essentielle dans l'annotation des génomes est la différenciation des régions codant pour les protéines à partir des régions non codantes, cela est souvent appelé épissage. Les algorithmes d'épissage basés sur l'analyse spectrale reposent sur des différences observées empiriquement entre les spectres des régions codantes et non codantes.

Le système neuro-flou utilisé comme une approche de prédiction utilise une connaissance préalable d'au moins quelques régions de codage pour réaliser la tâche de prédiction. Par suite, l'algorithme peut prédire des différentes limites des différentes régions codantes pour une séquence d'ADN.

3.5. Intégration de la logique floue dans les réseaux de neurones :

3.5.1. Limitation des réseaux de neurones :

Une limitation de l'utilisation des réseaux de neurones réside dans leur conception car il est difficile de trouver, de manière immédiate, à partir des données du problème à traiter, la meilleure architecture neuronale conduisant aux performances désirées. Cette détermination reste essentiellement empirique et soumise à la réalisation d'essais successifs qui pénalisent le délai de conception du système de traitement de données envisagé.

Les stratégies neuro-floues, consistant à intégrer certains aspects relevant du concept de logique floue, contribuent dans une certaine mesure à résoudre ces problèmes [33].

3.5.2. Les systèmes d'inférence floue :

La logique floue à été introduite, en 1965 par L.A. Zadeh, pour généraliser la logique binaire ou logique booléenne qui présente certains inconvénients. Dans le cadre de la logique booléenne, il n'est pas possible d'exprimer des affirmations nuancées : les faits sont soit vrais soit faux, ce qui ne représente pas la conception réelle que possède l'être humain par rapport à son environnement. Un autre défaut de la logique binaire est constitué par le fait qu'un système de traitement de l'information basé sur cette logique ne permet pas de prendre des décisions dans le cas où des informations manquent ou sont incomplètes, contrairement à l'être humain qui est capable d'effectuer des raisonnements par défaut.

C'est ainsi qu'en logique floue, les variables binaires de la logique booléenne sont remplacées par des variables linguistiques décrivant, par des mots de la langue usuelle, une certaine situation. Les variables linguistiques sont caractérisées par des degrés d'appartenance à des ensembles flous.

Les Systèmes d'Inférence Floue (SIF) exploitent les caractéristiques de la logique floue afin de permettre la construction de systèmes de traitement de l'information simples et performants. Un élément essentiel intervenant dans la conception des SIF consiste à écrire des règles d'inférence mettant en relation des grandeurs d'entrée avec des grandeurs de sortie, de la forme : « SI *situation* ALORS *décision* » où situation et décision sont des propositions du type « (*x* est *A*) *ET* (*y* est *B*) *OU* (*z* est *C*)… », signifiant que *x*, *y* et *z* sont des variables appartenant aux sous-ensembles flous *A*, *B* et *C*. Un ensemble de règles de cette forme constitue la base de règles du SIF, exploitée par le moteur d'inférence.

Les fonctions d'appartenance aux sous-ensembles flous, associées à chaque variable, forment la base de connaissances du SIF.

Ainsi, la conception d'un SIF consiste à déterminer judicieusement, en se basant sur des connaissances empiriques relatives au problème à résoudre, les sous-ensembles flous définissant les classes de situations et les classes de décisions ainsi que la base de règles les faisant correspondre.

Les variables d'entrée sont des grandeurs numériques ; à ce titre, elles ne peuvent donc pas être directement exploitées par le moteur d'inférence qui ne traite que des grandeurs linguistiques représentées par des ensembles flous. L'opération, consistant à transformer des grandeurs numériques en grandeurs linguistiques, est la fuzzification. Elle consiste simplement à attribuer à chaque variable numérique un degré d'appartenance à un ensemble flou, généralement appartenant à l'intervalle [0,1].

Après avoir parcouru la base de règles, le moteur d'inférence fournit un résultat pour chaque règle. Il est alors nécessaire de rassembler ces conclusions partielles afin d'obtenir une conclusion globale sous la forme d'un unique ensemble flou, décrivant la décision finale obtenue. Cette étape est connue sous le nom d'agrégation des résultats partiels.

Elle est généralement mise en œuvre en prenant le maximum des différents résultats partiels. Finalement, la sortie du SIF se présente sous la forme dune grandeur numérique. Celle-ci est obtenue à partir de l'ensemble flou décrivant la conclusion fournie par le moteur d'inférence. Cette opération est appelée défuzzification. Elle peut être menée à bien en calculant, par exemple, l'abscisse du centre de gravité de la fonction d'appartenance de la conclusion obtenue précédemment.

Il apparaît que l'avantage principal de l'utilisation d'un SIF, pour l'analyse et le traitement de données, consiste en la possibilité d'intégrer des observations effectuées sur les données à traiter. Ainsi, la conception d'un SIF exploite la connaissance et l'expertise humaine qui sont généralement difficiles à mettre en application lors de la conception de

systèmes de traitement classiques, celle-ci utilisant des méthodes analytiques basées sur l'exploitation de données numériques.

3.5.3. Les structures neuro-floues :

Il apparait que les systèmes d'inférence floue et neuronaux présentent des caractéristiques complémentaires qu'il est intéressant de regrouper dans une même structure, associant les avantages des deux techniques. C'est ce qui fait l'objet des approches neuro-floues.

Deux principales orientations coexistent pour l'obtention de structures neuro-floues. La première consiste à améliorer les réseaux de neurones du point de vue de l'interprétation de leur action tandis que la deuxième vise à conférer aux systèmes d'inférence floue une capacité d'apprentissage semblable à celle des réseaux de neurones.

Ces deux approches exploitent la complémentarité des propriétés des réseaux de neurones et de la logique floue, ainsi qu'une certaine équivalence formelle existant entre ces deux concepts. Leur objectif commun est d'obtenir des systèmes de traitement de l'information délivrant, avec un temps de calcul suffisamment faible, des décisions pouvant être justifiées en un langage naturel à partie des données traitées et robustes en présence d'imprécisions et d'incertitudes, imitant ainsi le comportement d'un être humain.

3.5.3.1. Réseaux de neurones flous :

L'approche consistant à intégrer des concepts appartenant à la logique floue dans les réseaux de neurones se situe à différents niveaux d'abstraction. En effet, une première possibilité consiste à généraliser les neurones « conventionnels » qui prennent une décision après avoir effectué une somme pondérée de leurs entrées, en éléments réalisant des opérations telles que l'évaluation d'une règle d'inférence floue. La nouvelle structure ainsi obtenue est représentée sur la figure 3.1 où x_1, x_2,..., x_n sont des entrées, u_e la décision et par exemple $\mu_{A_i}(x_1)$ désigne le degré d'appartenance de x_1 au sous-ensemble flou A_i.

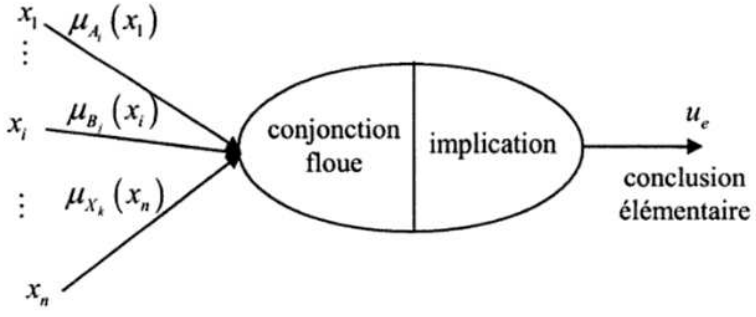


Figure 3.1. Neurone « flou ».

Les poids synaptiques de ces neurones généralisés, qui dépendent dans ce cas des entrées, réalisent la fuzzification de celles-ci, tandis que la fonction de décision correspond à l'opération d'implication floue. Les neurones flous construits de cette manière représentent alors une règle d'inférence et réalisent le processus d'évaluation de cette règle.

De tels neurones flous étant définis, il est possible de les assembler en couches puis en réseaux et de soumettre à un apprentissage les poids synaptiques et les fonctions de décision. Les différentes conclusions partielles de chaque couche sont alors traitées par la couche suivante, jusqu'à l'obtention d'une décision finale, donnée par la dernière couche.

Un réseau de neurones flous réalisant la mise en œuvre d'un système d'inférence floue de type Sugeno, à deux entrées et une sortie, muni d'une base de règles comportant cinq règles, est donné par la figure 3.2.

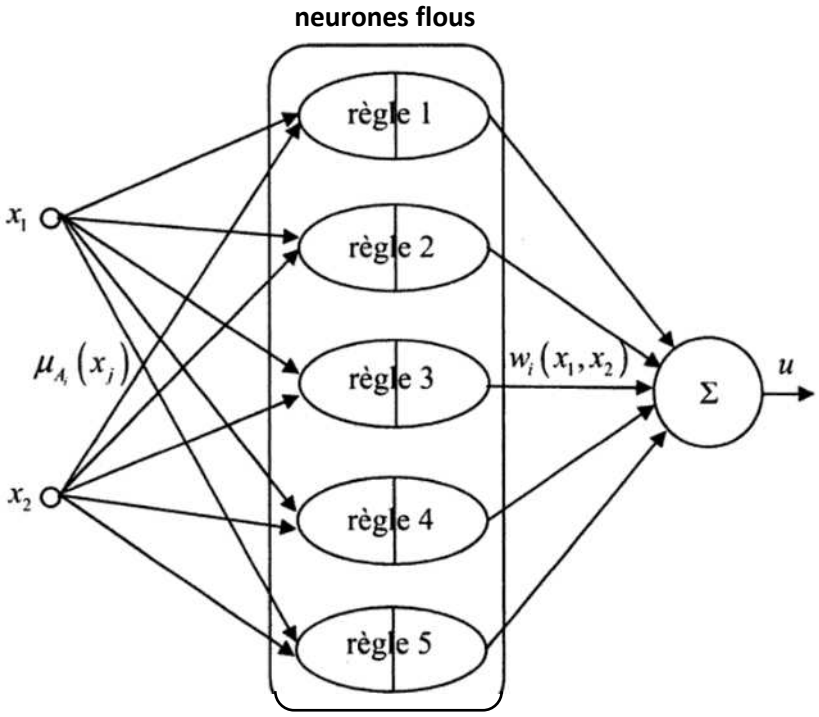


Figure 3.2. Réseau de neurones flous.

L'utilisation de neurones flous conduit à une structure de réseau à une seule couche cachée puisque chaque neurone réalise à lui seul les opérations de calcul des prémisses et d'implication d'une règle.

Cette approche de fuzzification des réseaux de neurones permet ainsi de donner à la sortie de chaque neurone une interprétation, puisqu'il s'agit alors du résultat d'une règle d'inférence floue de type Sugeno. De plus, les règles représentées par une couche de neurones étant évaluées simultanément, la complexité de la production d'une décision s'en trouve fortement

diminuée, puisque la charge de calcul se trouve distribuée sur plusieurs opérateurs fonctionnant en parallèle.

3.5.3.2. Systèmes d'inférence floue à base de neurones standards :

L'approche de construction de structures neuro-floues, basée sur l'introduction de nouveaux types de neurones présentant des caractéristiques floues, nécessite de modifier la nature des neurones mis en œuvre, et donc les stratégies d'apprentissage utilisées. Une autre approche, possible, permet d'intégrer une interprétation floue dans les réseaux de neurones utilisés dans le cadre de la commande de processus, tout en utilisant des neurones standards, c'est-à-dire prenant une décision à partir d'un potentiel synaptique constitué d'une somme pondérée de leurs entrées. Cette stratégie permet d'utiliser des structures neuronales et des algorithmes d'apprentissage déjà existants et éprouvés.

Une raison, justifiant la mise en œuvre d'une telle approche réside dans l'exploitation de la capacité de calcul d'un réseau de neurones qui est alors assimilé à un calculateur massivement parallèle permettant l'élaboration d'une décision avec un temps de calcul très réduit. Cette solution consiste à construire les différentes parties d'un système d'inférence floue, indépendamment les unes des autres, au moyen de réseaux de neurones qui sont, ensuite, assemblés pour fournir une structure neuronale globale dans laquelle peuvent être identifiées les différentes opérations réalisées dans un système floue. Dans ce cas, l'interprétation donnée aux réseaux de neurones ne se situe pas au niveau des neurones élémentaires, mais dans des groupes de neurones constituant des sous-réseaux. Cette méthode est illustrée par la figure 3.3 qui montre cette spécialisation des sous-réseaux de neurones, dans le cas de l'implémentation neuronale d'un SIF de type Sugeno.

Dans cette figure, les réseaux de neurones RN_1 à RN_p sont entraînés à calculer les degrés d'appartenance de chaque grandeur d'entrée et à évaluer les prémisses de p règles implantées sous la forme de connexions. Quant au réseau RN, il subit un apprentissage en vue de calculer les fonctions numériques apparaissant en conclusion des règles. De même, les opérateurs produits, notés Π, intervenant dans les inférences, peuvent être des multiplieurs conventionnels ou bien des réseaux de neurones entraînés pour émuler ces opérateurs.

3.5.4. Systèmes d'inférence neuro-flous adaptatifs :

Une architecture particulière de réseaux neuro-flous a été développée par Jang et Sun pour l'identification paramétrique. Il s'agit de l'algorithme ANFIS (Adaptive Neuro-Fuzzy Inference System) qui permet l'identification de paramètres en utilisant une règle d'apprentissage hybride combinant l'algorithme de rétropropagation du gradient et la méthode

des moindres carrés. L'algorithme ANFIS est basé sur une représentation graphique d'un SIF de type Sugeno, munie d'une capacité d'apprentissage de type neuronal.

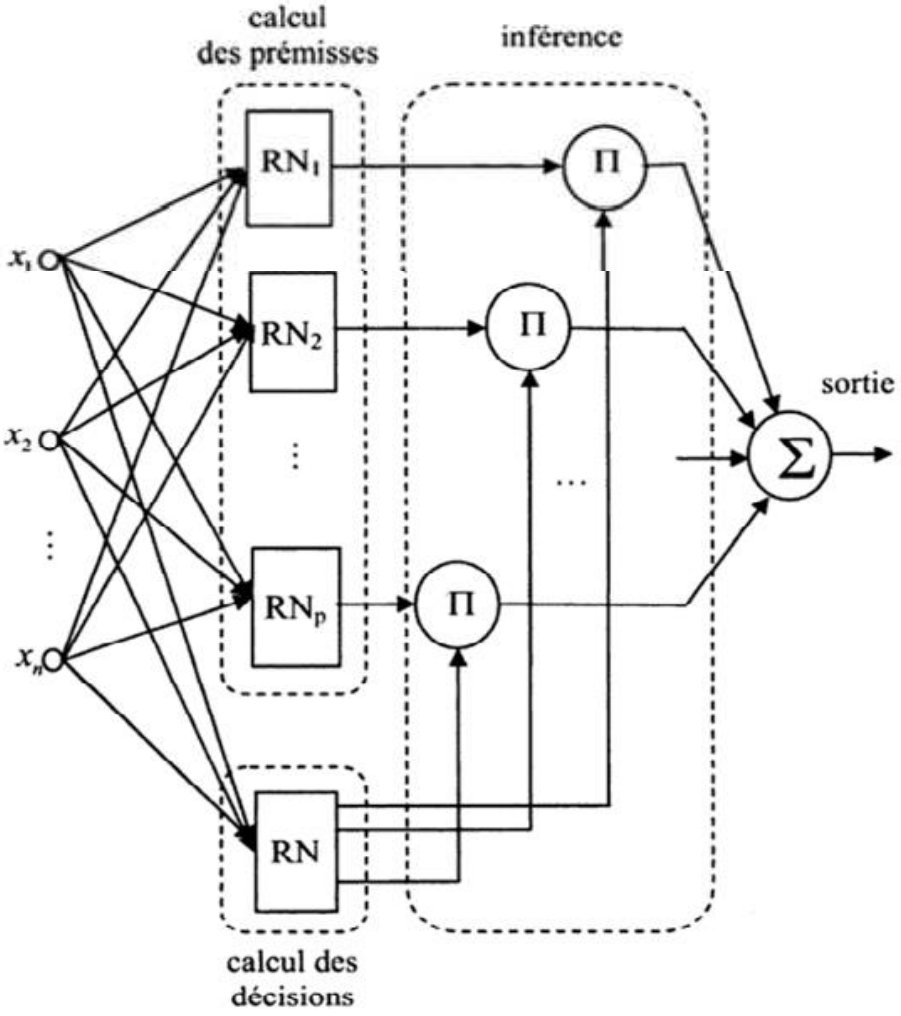


Figure 3.3. Réseau neuro-flou à neurones standards.

A titre d'illustration, considérons un SIF de type Sugeno, muni de deux règles dont les conclusions sont des fonctions linéaires des entrées x et y :

SI x est A_1 ET y est B_1, ALORS $f_1 = p_1 x + q_1 y + r_1$

SI x est A_2 ET y est B_2, ALORS $f_2 = p_2 x + q_2 y + r_2$

L'algorithme ANFIS permet de construire un réseau implémentant ces règles selon l'architecture suivante (figure 3.4).

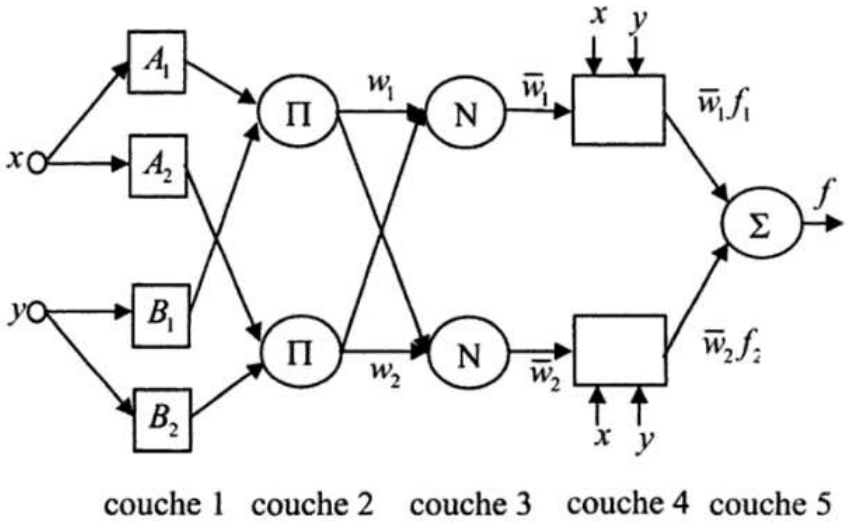


Figure 3.4. Exemple d'architecture d'un réseau neuro-flou.

Dans cette structure :

- La couche 1 calcule les degrés d'appartenance, compris entre 0 et 1, de chaque variable d'entrée,
- La couche 2 calcule les prémisses de chaque règle comme étant le produit Π des degrés d'appartenance des variables intervenant dans les prémisses de chaque règle :

$$w_i = \mu_{A_i}(x)\mu_{B_i}(y), i = 1, 2$$

- La couche 3, constituée d'opérateurs de normalisation notés N, normalise les résultats de la couche :

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1, 2$$

- la couche 4 évalue la conclusion de chaque règle :

$$\overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i)$$

- la couche 5 fournit, dans ce cas, le résultat final :

$$f = \sum_i w_i f_i \Big/ \sum_i w_i = \sum_i \overline{w_i} f_i$$

La règle d'apprentissage de cette structure est basée sur une méthode d'apprentissage de type descente de gradient, analogue à celle utilisée dans le cas des réseaux de neurones non bouclés.

Les fonctions d'appartenance peuvent être définies par des triangles ou des trapèzes, ou, si on désire faciliter la détermination des classes par des techniques d'apprentissage, par des fonctions gaussiennes, chacune définie par son centre c_i et sa variance σ_i^2.

3.6. Conclusion :

Dans ce chapitre nous avons abordé la description des outils et algorithmes de détection de gènes couramment utilisés en bioinformatique. Certains programmes de détection de gènes combinent plusieurs approches et intègrent des informations supplémentaires pour valider les prédictions, comme la comparaison de séquences des gènes prédits avec des gènes connus et répertoriés dans les banques.

Les techniques de calcul souple peuvent bien fonctionner pour la prédiction de gènes en raison de leur capacité à gérer l'incertitude dans les séquences de données. En revanche, les techniques hybrides peuvent atteindre un niveau significatif de précision dans l'identification des gènes.

Les systèmes neuro-flous sont des systèmes flous formés par un algorithme d'apprentissage inspiré de la théorie des réseaux de neurones. La performance d'un réseau neuro-flou dépasse celles d'autres méthodes en termes de décision.

Chapitre 4

Résultats et discussion

4.1. Introduction :

Plusieurs caractéristiques et mesures ont été proposées pour identifier des gènes dans des régions d'ADN non caractérisé. Cette tâche est difficile et fait actuellement l'objet de nombreux efforts de recherche. Les caractéristiques de reconnaissance, pour localiser les gènes dans la séquence d'ADN, sont généralement liées aux régions codant pour la protéine et aux frontières des régions codantes. La base de la plupart des méthodes de reconnaissance de région codante est les biais de position imposés à la séquence d'ADN dans les régions codantes par le code génétique et par la distribution des acides aminés dans les protéines [34-41].

Les algorithmes existants pour identifier des régions codant pour des protéines dans des séquences d'ADN exploitent plusieurs observations empiriques attribuées à des séquences d'ADN, telles que l'observation empirique selon laquelle le spectre des régions codant les protéines se manifeste. Et il existe une variété de techniques de calcul basées sur les qualités statistiques des exons dans le génome. Sur la base des informations structurelles fournies par les séquences d'ADN, les recherches ont montré que les régions codantes se comportent statistiquement de manière aléatoire par rapport aux régions non codantes. De nombreuses études ont été menées pour extraire les segments caractéristiques, révéler certaines structures cachées, distinguer les régions codantes des régions non codantes dans les séquences d'ADN, et explorer la similarité structurelle entre les séquences d'ADN [18].

Dans cette étude, une approche basée sur ANFIS a été présentée pour la prédiction de séquences d'ADN. Le but de cette étude est de tester certaines caractéristiques proposées extraites de la séquence de nucléotides et de présenter ces caractéristiques dérivées au système neuro-flou comme entrée. Afin de tester leur capacité à détecter les motifs. En se concentrant sur l'extraction des caractéristiques des séquences d'ADN, deux ensembles de caractéristiques ont été utilisés. La prédiction a été effectuée par extraction de caractéristiques à l'aide de la transformée en ondelettes et le second ensemble de caractéristiques a été extrait de la fréquence de transition des nucléotides.

4.2. Signaux ADN obtenus par codage :

Une séquence d'ADN peut être considérée comme un texte de quatre lettres (A, C, G et T) où A, C, G et T se réfèrent respectivement aux bases adénine, cytosine, guanine et thymine. L'analyse de l'ADN nécessite d'abord de convertir le texte d'ADN en une séquence numérique. Cela peut être fait sur la base de la distinction des quatre bases, en utilisant une représentation complexe des bases :

$$a = 1 + j\,,\ t = 1 - j\,,\ c = -1 - j\,,\ g = -1 + j\ .$$

Afin de la mise au point d'une méthodologie de prédiction des gènes dans des séquences d'ADN des organismes eucaryotes et procaryotes. Des séquences d'*Escherichia coli* et de *Caenorhabditis elegans* ont été utilisées avec une structure génétique déterminée. Les données sont extraites de la base de données NCBI (National Center for Biotechnology Information).

Nous illustrons ici un exemple de signal obtenu avec le codage appliqué sur une séquence d'*E. coli*. Les résultats sont l'amplitude et la phase non enveloppée (unwrapped phase) des coefficients du signal complexe (figure 4.1). Comme on peut le constater, la phase présente des comportements oscillatoires irréguliers. Ces comportements reflètent la structure en nucléotides de cette séquence.

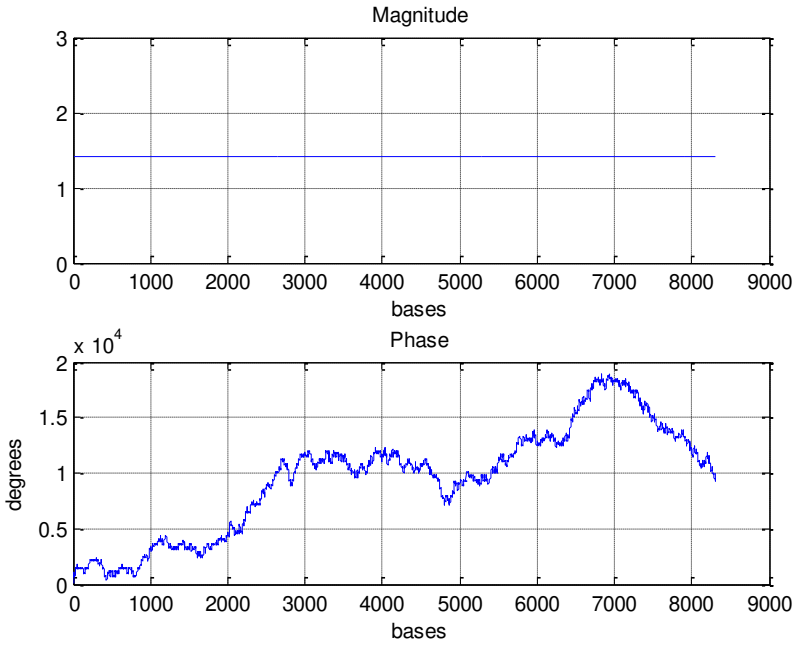


Figure 4.1. Le signal associé à une séquence exemplaire d'*E. coli* calculé avec le codage complexe utilisé.

La phase non enveloppée nous permet de corriger les angles de phase du radian dans un vecteur de phases en ajoutant des multiples de ± 2π lorsque les sauts absolus entre les éléments consécutifs de ce vecteur sont supérieurs ou égaux à la tolérance de saut, en général de π radians. Alors que, la phase enveloppée (wrapped phase) signifie que tous les points de phase sont limités à la plage - π ≤ phase < + π. Lorsque la phase réelle est en dehors de cette plage, la valeur de la phase est augmentée ou diminuée d'un multiple de 2π pour que la valeur de la phase soit comprise entre +/- π.

4.3. Analyse et caractérisation de signaux ADN par la transformée de Fourier :

L'analyse des signaux, obtenus avec le codage envisagé, par la transformée de Fourier va nous permettre de mettre en évidence l'existence d'une composante fréquentielle qui a une interprétation de nature structurelle.

Les figures 4.2 à 4.4 montrent le résultat obtenu à partir de l'analyse de Fourier de trois séquences du génome du *C. elegans*. Une séquence codante, une séquence non codante et une séquence contenant une région codante et une région non codante.

Le spectre de l'ADN codant pour les protéines a un pic à la fréquence k = N/3, où N est la longueur de la séquence. Cette fréquence correspond à une période de trois échantillons de la longueur de chaque codon. Par exemple, dans la figure 4.2, nous avons tracé le spectre pour une région codante de longueur N = 273 à l'intérieur du génome de *C. elegans*, démontrant un pic à la fréquence k = 91. Les différentes oscillations observées correspondent à la distribution de nucléotides, telles que la périodicité de 3 paires de bases (pb) dans les séquences codantes qui reflète la structure en codons de ces régions. Par contre, les spectres obtenus pour le segment de l'ADN non codant et le segment contenant une région codante et une région non codante ne présentent pas le comportement de période 3. Le comportement de ces spectres est qualitativement le même comme le montre les figures 4.3 et 4.4.

On ce qui concerne la phase, les phases obtenues présentent un caractère oscillant. La phase correspondant à la région codante oscille autour d'une tendance de pente négative. Tandis que, la phase des deux autres séquences présente le même type d'oscillations mais autour d'une tendance de pente positive.

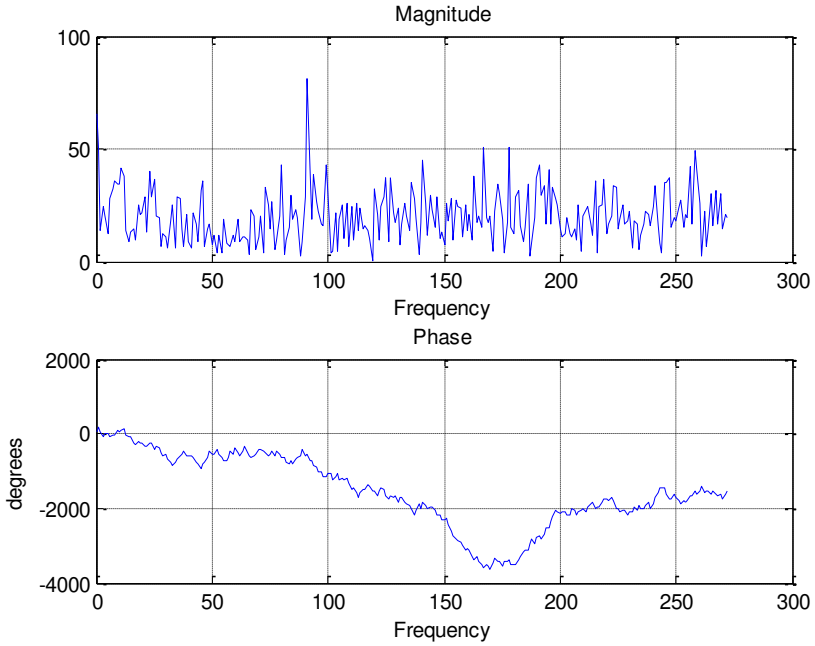


Figure 4.2. Le spectre d'une région codante d'ADN.

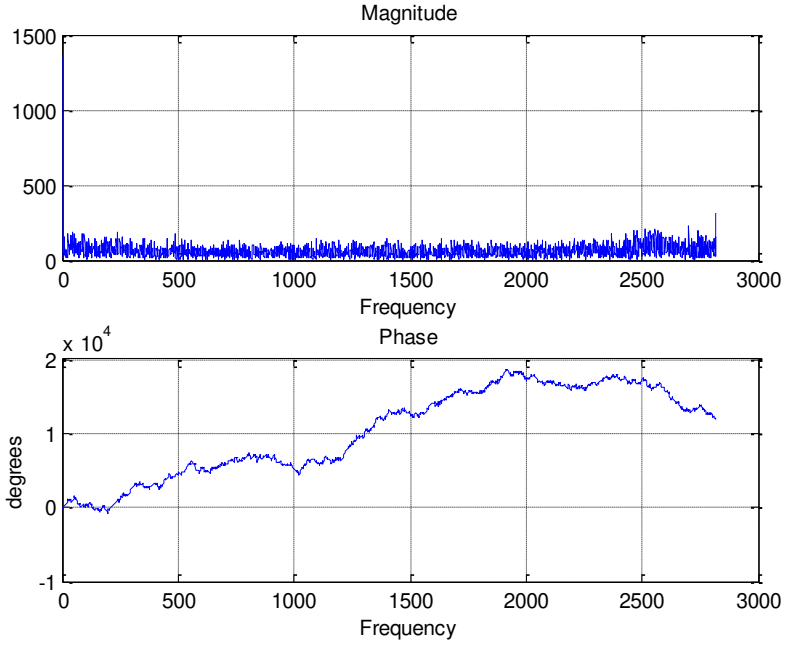


Figure 4.3. Le spectre d'une région non codante d'ADN.

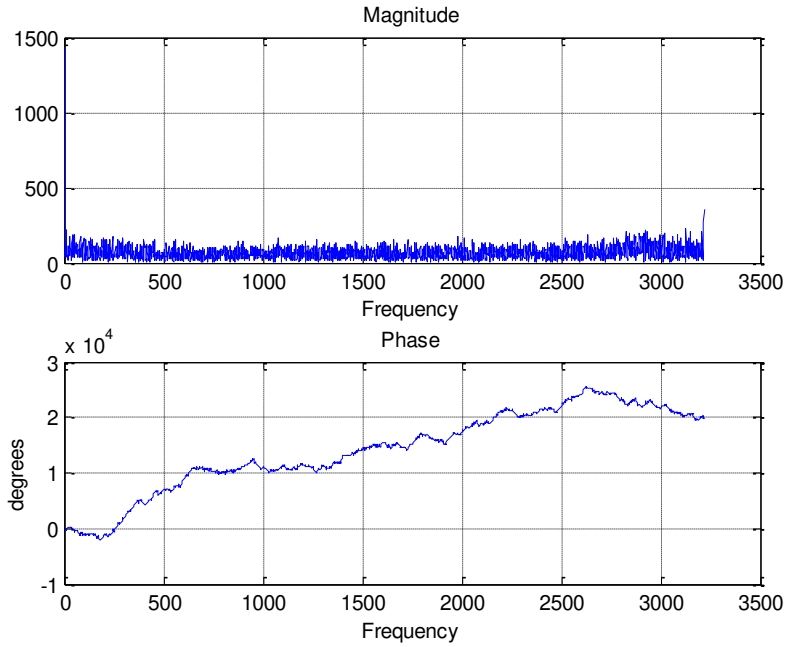


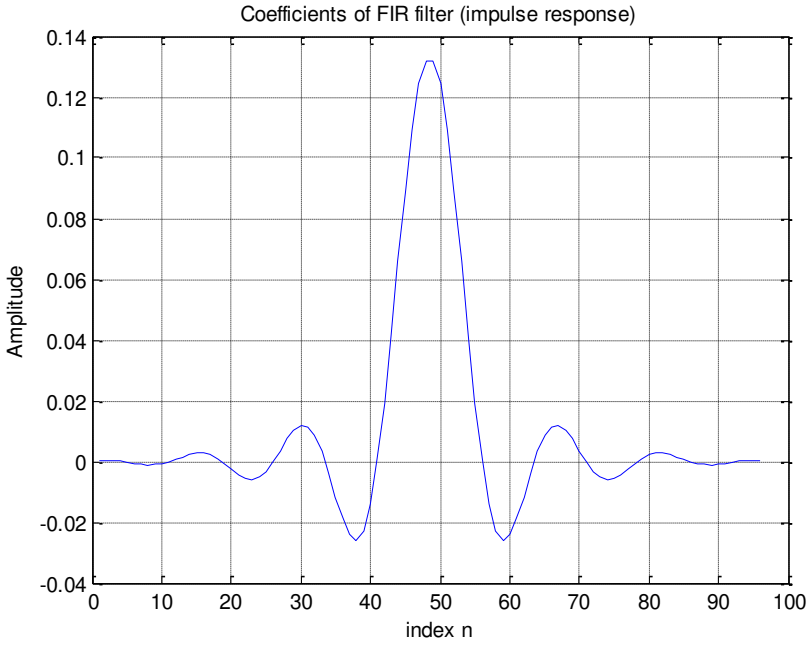
Figure 4.4. Le spectre d'une séquence d'ADN contenant une région non codante et une région codante.

4.4. Analyse et caractérisation de signaux ADN par banc de filtres :

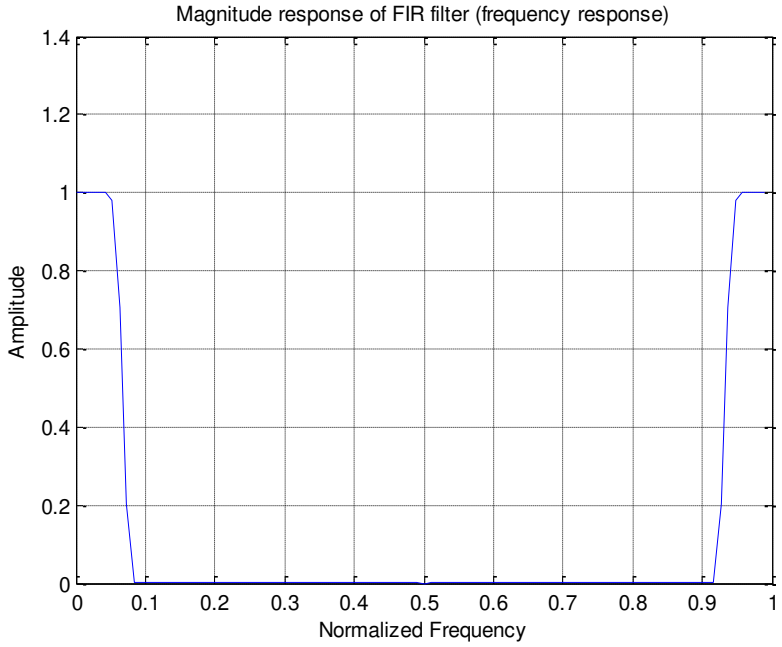
Une analyse par bancs de filtres a été effectuée sur des séquences d'ADN en examinant les caractéristiques spécifiques de la composition nucléotidique dans la séquence d'ADN. Pour cet effet, une séquence d'ADN, avec 8310 nucléotides, du génome d'*E. coli* est utilisé comme séquence d'étude dont le signal complexe associé est présenté dans la figure 4.1. Cette séquence est composée de 7 gènes. De plus, deux séquences de deux gènes extraites du même génome d'*E. coli* sont utilisées. L'un des deux gènes est de longueur 624 nucléotides et l'autre de longueur 2430 nucléotides.

Un banc de filtres de 8 sous bandes avec une décimation d'un facteur 6 est utilisé, la fenêtre de troncature est de type Kaiser. Notons qu'il est intéressant de pouvoir travailler avec une redondance faible pour les bancs de filtres. Pour cet exemple nous avons travaillé avec le banc de filtres avec les paramètres : le banc de filtres est composé de $M = 8$ filtres, le facteur de décimation $D = 6$, ce qui fait donc une redondance totale pour le banc de filtres est $k = M / D = 4/3 \approx 1.33$. Dans ce cas, le banc de filtres est dit un banc de filtres suréchantillonné. L'ordre des filtres, i.e. la longueur des filtres, est 97. Plus précisément, les filtres seront supposés de longueur $2k'L$ avec $k' \in \mathbb{N}^*$, où L est le plus petit multiple commun de M et D et en prenant en compte la symétrie et la taille impaire de la réponse impulsionnelle. La pulsation de

coupure du filtre passe bas est $\omega_c = \frac{\pi}{M}$. La figure 4.5 présente la réponse impulsionnelle et fréquentielle du filtre passe bas.



(a)



(b)

Figure 4.5. (a) Réponse impulsionnelle et (b) réponse fréquentielle du filtre passe bas.

Du point de vue de l'implémentation de la méthode de filtrage, on peut remarquer que l'analyse étant essentiellement locale (aux effets de recouvrement près), il est possible de séparer un signal à analyser en plusieurs sous-signaux de moindre taille et se recouvrant partiellement.

Les signaux obtenus à l'aide d'un tel banc de filtres sont illustrés dans les figures 4.6 à 4.8. On peut constater que, les signaux de phase après filtrage sont beaucoup plus lisses et continues qu'auparavant, i.e. par rapport à la phase du signal présenté dans la figure 4.1. Les signaux de phase ainsi obtenus sont caractérisés par la présence de tendances (auxquelles se superposent des fluctuations) de pente positive et d'autres de pente négative.

A propos des signaux d'amplitude, on s'aperçoit que ces signaux oscillent autour d'une valeur finie positive de 0.05.

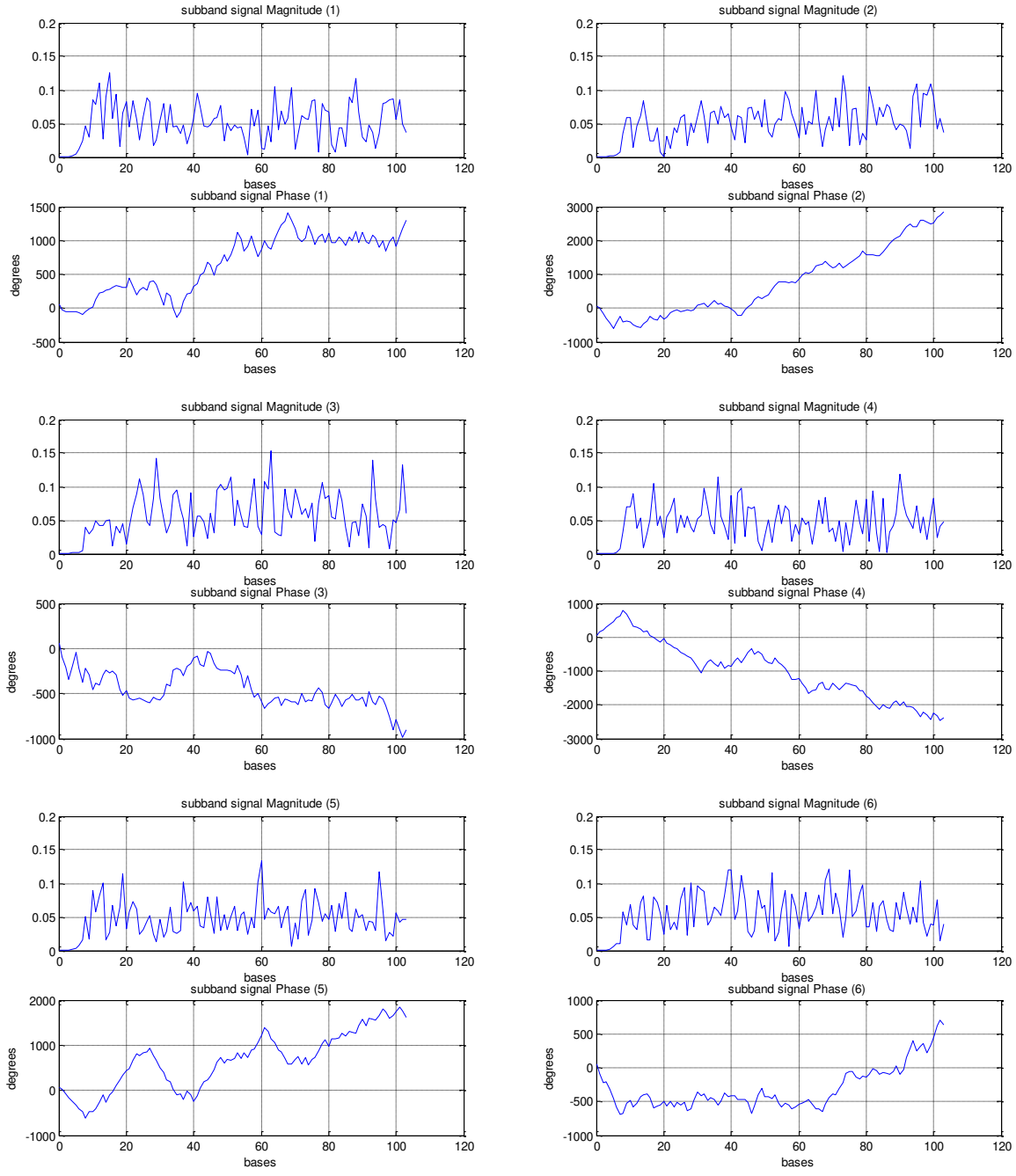


Figure 4.6. Les signaux obtenus après filtrage du signal associé à la séquence du gène de longueur 624 nucléotides.

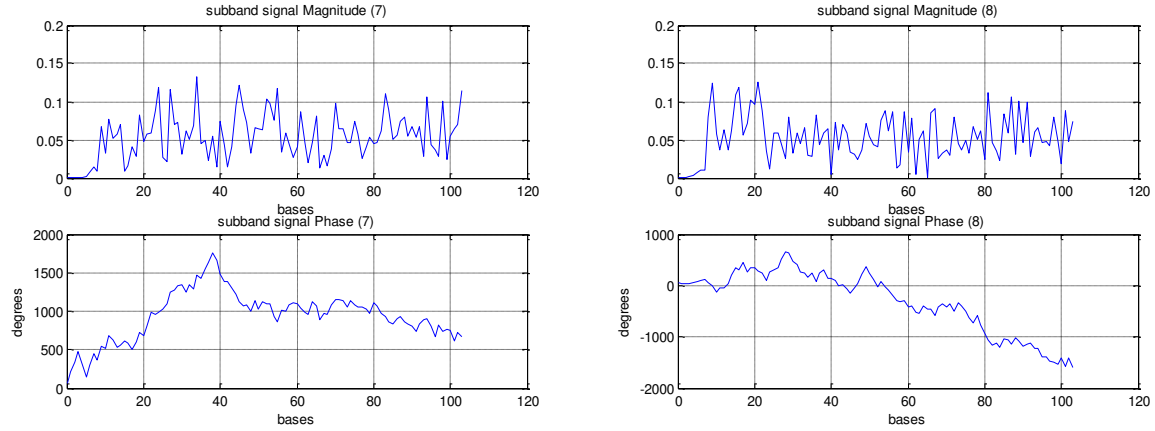


Figure 4.6. (La suite)

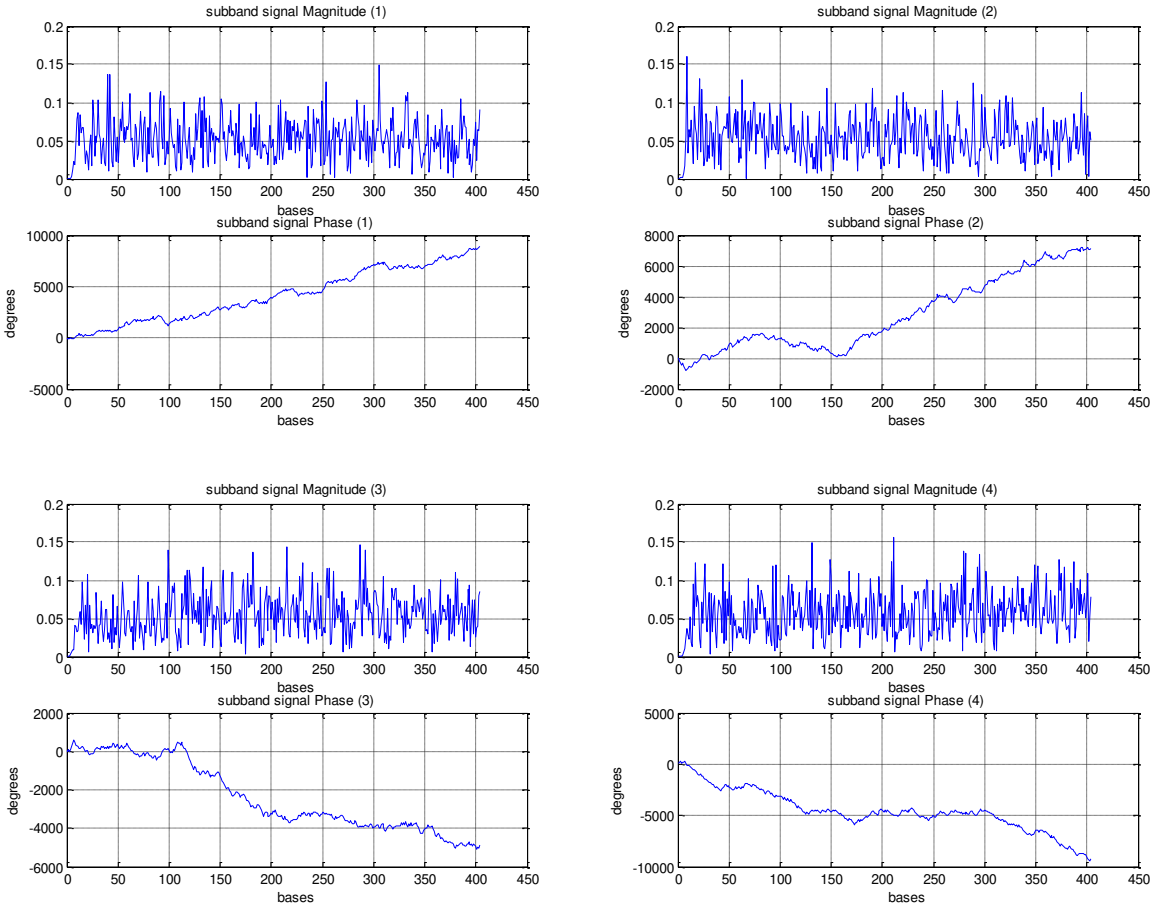


Figure 4.7. Les signaux obtenus après filtrage du signal associé à la séquence du gène de longueur 2430 nucléotides.

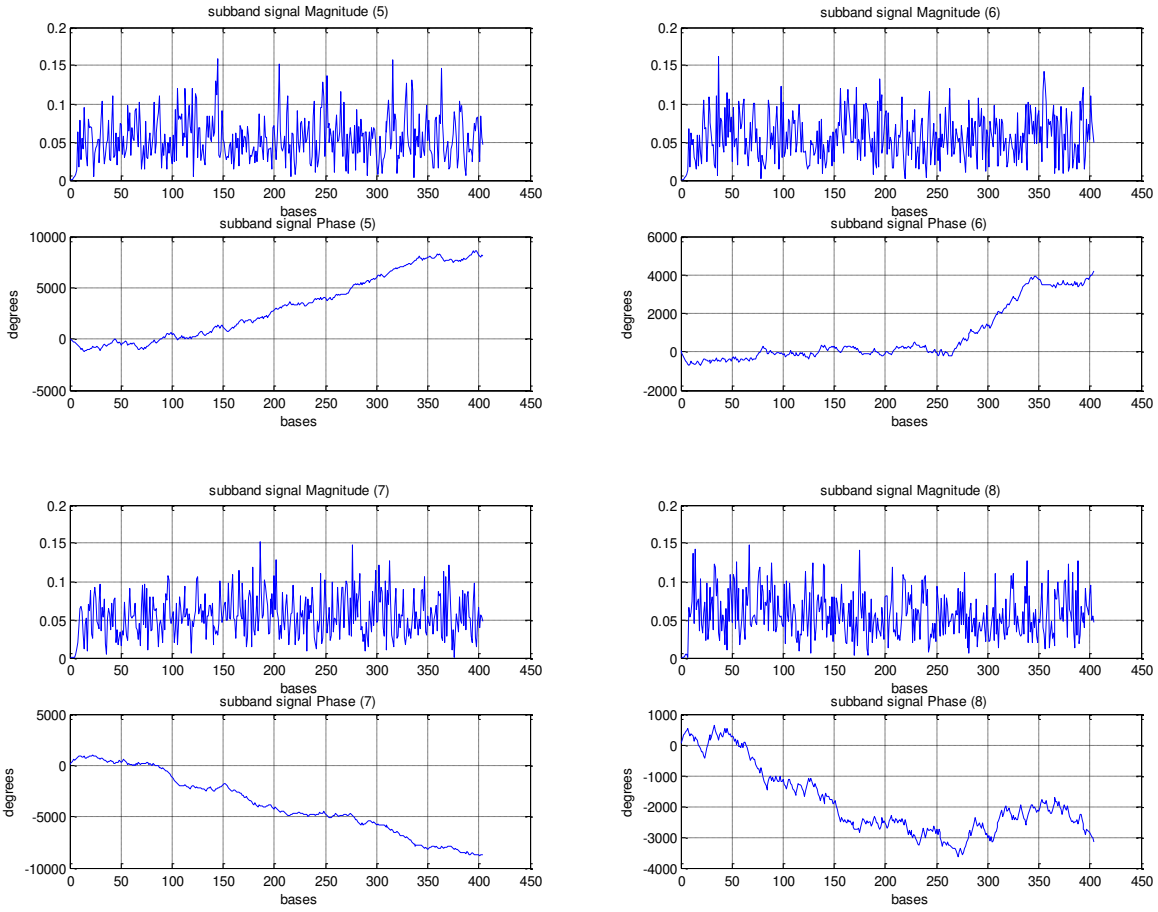


Figure 4.7. (La suite)

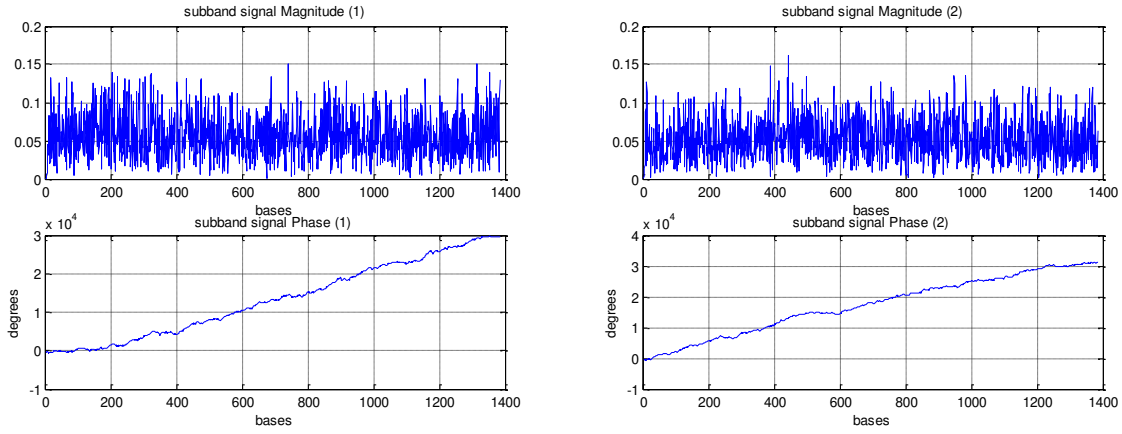


Figure 4.8. Les signaux obtenus après filtrage du signal associé à la séquence composé de 7 gènes.

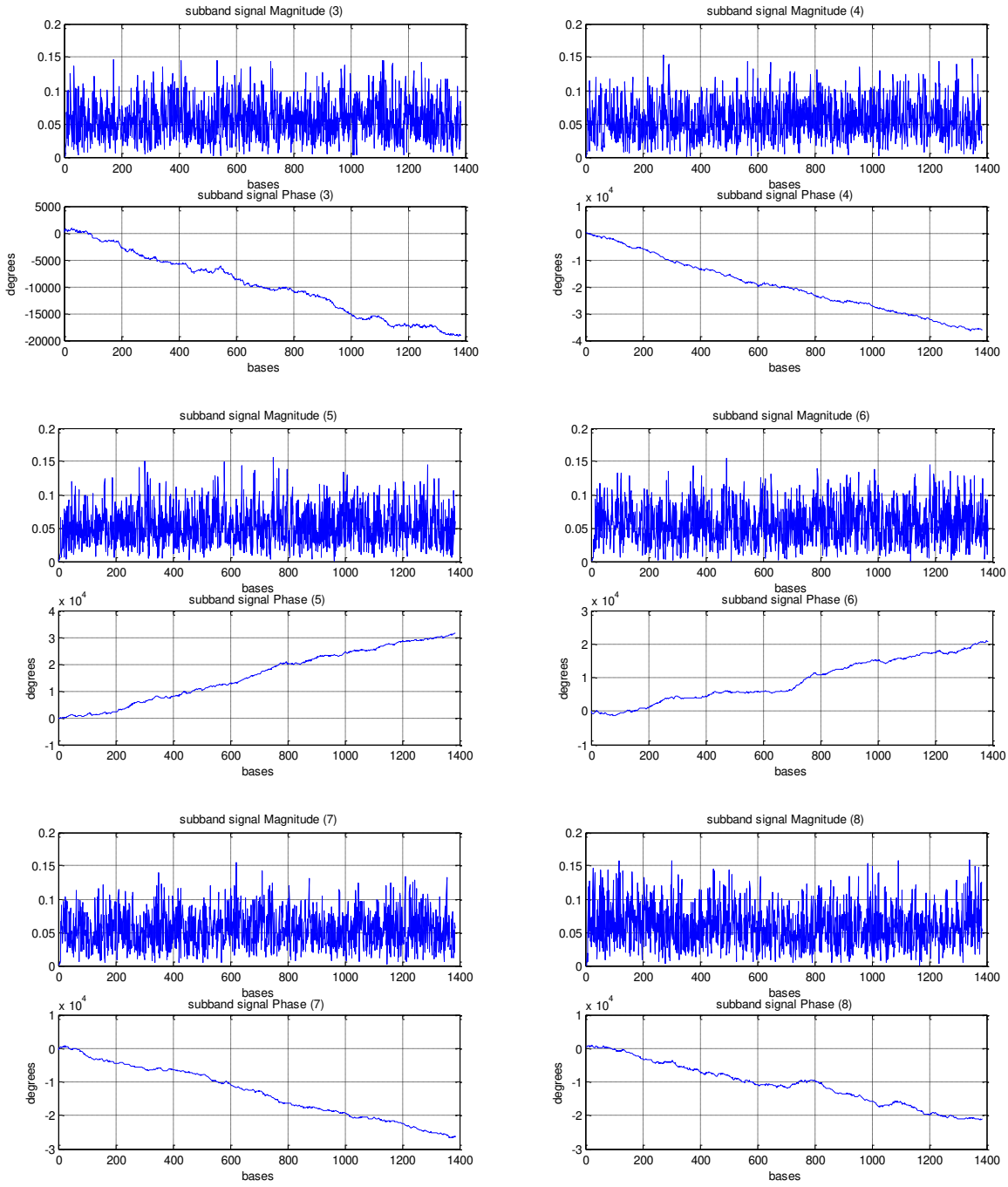


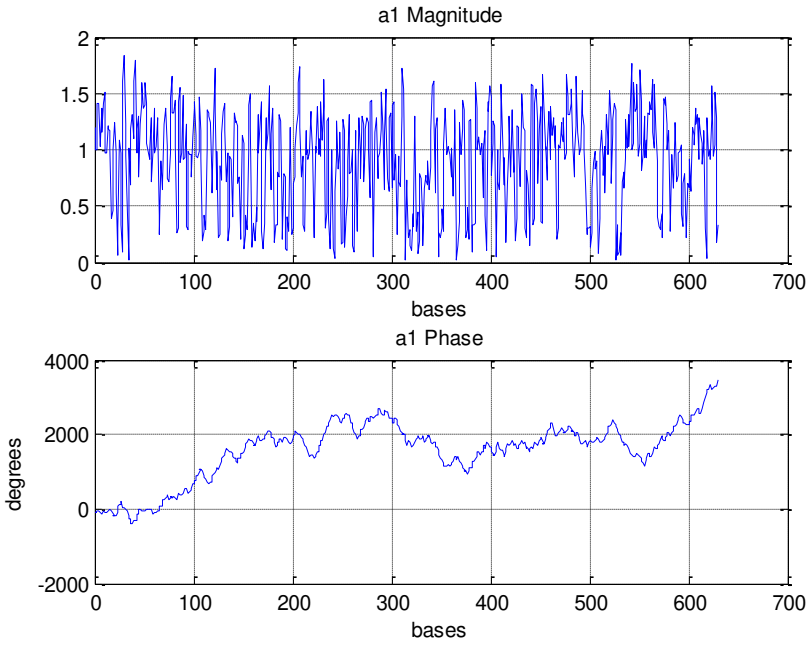
Figure 4.8. (La suite)

4.5. Analyse et caractérisation de signaux ADN par la transformée en ondelettes :

L'application d'ondelettes est utile dans l'analyse de séquences d'ADN pour la prédiction de structure de séquence, la comparaison de séquences et sa classification. L'analyse en ondelettes fournit une description utile de la structure inhérente aux séquences d'ADN. En tant que méthode de traitement de signal multirésolution, la transformée en ondelettes peut présenter une meilleure résolution de fréquence dans les basses fréquences et une meilleure résolution temporelle dans les hautes fréquences. La transformée en ondelettes

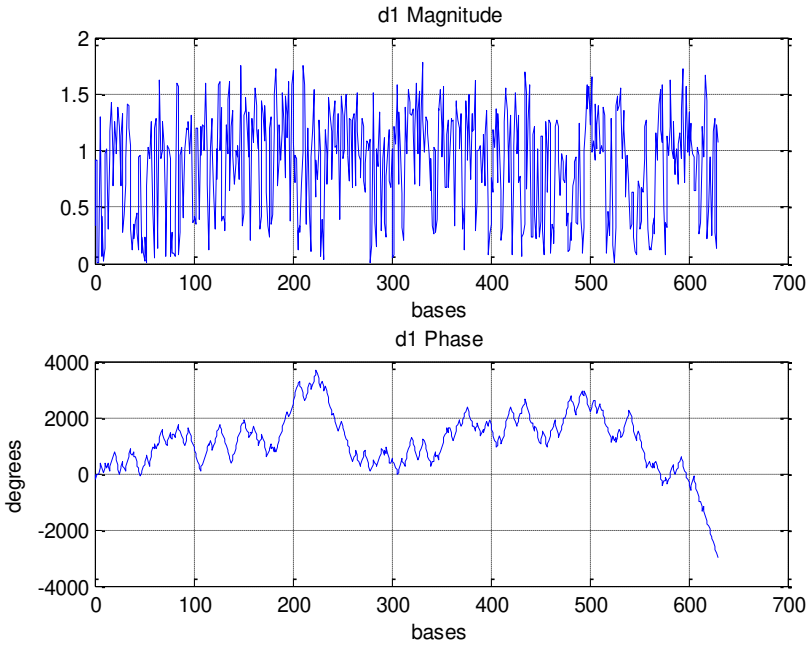
décompose un signal en plusieurs groupes de coefficients. Différents vecteurs de coefficients contiennent des informations sur les caractéristiques de la séquence à différentes échelles. Les coefficients aux échelles grossières capturent les caractéristiques globales et brutes du signal tandis que les coefficients à l'échelle fine contiennent des détails [18,42-48].

Cette étape applique la transformée en ondelettes à la séquence. Les coefficients d'ondelette d'approximation et de détail au premier niveau ont été calculés, comme le montre la figure 4.9. Les résultats sont l'amplitude et la phase non enveloppée des coefficients d'ondelettes.



(a)

Figure 4.9. Transformée en ondelettes d'une séquence d'*E. coli*.
Amplitude et phase de (a) approximation (b) détail.



(b)

Figure 4.9. (La suite)

Comme on peut le constater, les graphes de phase ainsi obtenu sont très irréguliers faits de fluctuations de faible amplitude autour des tendances. Ces tendances sont induites par le biais dans la composition des nucléotides de la chaîne d'ADN.

Les deux échelles correspondent à l'observation des fluctuations de la séquence d'ADN. L'un filtre la composante basse fréquence (les tendances basses fréquences où semblent osciller lentement) (figure 4.9 (a)) et l'autre révèle les fluctuations locales (haute fréquence) de la fréquence des transitions entre les nucléotides sur la séquence d'ADN (figure 4.9 (b)). Par conséquent, les coefficients d'ondelettes de détail et d'approximation calculés du signal d'ADN ont été utilisés comme vecteurs représentant le signal.

4.6. Sélection de caractéristiques:

4.6.1. Premier ensemble de caractéristiques:

Pour chaque segment d'ADN, (2 × la longueur du segment) phases et amplitudes des coefficients d'ondelettes ont été obtenues. Afin de réduire la dimensionnalité des vecteurs caractéristiques, les caractéristiques suivantes ont été utilisées pour représenter la distribution temps-fréquence du signal ADN:

a. Amplitude :

1. La moyenne des amplitudes des coefficients d'ondelette dans chaque sous-bande.
2. L'écart-type des amplitudes des coefficients d'ondelette dans chaque sous-bande.

3. Le nombre de pics des amplitudes des coefficients d'ondelettes qui ont une valeur de crête supérieure à la valeur moyenne dans chaque sous-bande.

4. Le nombre de pics des amplitudes des coefficients d'ondelette qui ont une valeur de crête inférieure à la valeur moyenne dans chaque sous-bande.

5. Le nombre de fonds des amplitudes des coefficients d'ondelette qui ont une valeur supérieure à la valeur moyenne dans chaque sous-bande.

6. Le nombre de fonds des amplitudes des coefficients d'ondelettes qui ont une valeur inférieure à la valeur moyenne dans chaque sous-bande.

b. Phase :

1. Le nombre de pics des phases de coefficients d'ondelettes dans chaque sous-bande.

2. Le nombre de fonds des phases de coefficients d'ondelette dans chaque sous-bande.

3. Le nombre de passages par zéro où les positions de croisement se situent avant le milieu de la longueur de la fenêtre dans chaque sous-bande.

4. Le nombre de passages par zéro où les positions de croisement se situent après le milieu de la longueur de la fenêtre dans chaque sous-bande.

Par conséquent, la taille du vecteur d'entrée est 2×6+2×4=20, ce qui est une dimensionnalité réduite des vecteurs caractéristiques. Des tests sont effectués avec différents types d'ondelettes et l'ondelette de Daubechies d'ordre 4 (db4) est sélectionnée qui a donné une efficacité de prédiction maximale.

Localiser la position d'un gène consiste à déterminer les positions de ses régions codantes. Afin de localiser les gènes et de déterminer les positions de ses régions codantes, les caractéristiques extraites ont été utilisées pour discriminer les limites des gènes et des régions codantes. Le tableau 4.1 présente les caractéristiques extraites calculées à l'intérieur de cinq fenêtres exemplaires autour de diverses régions d'intérêt dans une séquence d'ADN, qui sont les limites des gènes et des régions codantes, d'une séquence exemplaire d'*E. coli*. Et ailleurs dans la séquence d'ADN, les caractéristiques sont extraites d'un exemple de fenêtre aléatoirement choisi dans une région codante. Avant d'exécuter les simulations, nous normalisons linéairement toutes les données dans l'intervalle [0, 1]. D'après le tableau 4.1, on peut constater que la plupart des caractéristiques extraites des cinq régions spéciales de la séquence d'ADN sont différentes les unes des autres. Intuitivement, ils peuvent servir de paramètres utiles pour discriminer et identifier les gènes et les séquences codantes dans les séquences d'ADN et peuvent influencer significativement la performance globale de la prédiction. Quelques cas de similarités ont été trouvés d'une ou deux caractéristiques entre deux régions pour les caractéristiques d'amplitude et les caractéristiques de la phase

d'approximation, et de deux caractéristiques entre trois régions pour les caractéristiques de la phase de détail. Ces similitudes peuvent être dues aux corrélations à longue portée dans les séquences d'ADN. Les corrélations à longue portée observées dans diverses études ont été soulevées par plusieurs auteurs. A. Arneodo et al. dans [15] ont rapporté que l'existence de corrélations à longue portée dans les séquences d'ADN a nécessité beaucoup d'efforts de la part des chercheurs pour adapter leurs techniques afin de maîtriser la présence. Les auteurs de l'article ont confirmé la présence de corrélations à longue portée dans les séquences d'ADN par une analyse par ondelettes et ont mentionné que les corrélations à longue distance observées dans les séquences de gènes pourraient résulter de l'alternance d'introns et d'exons. Une autre interprétation possible de l'existence de corrélations à longue portée dans les séquences d'ADN est l'existence de segments fortement répétés. C'est une caractéristique des génomes eucaryotes supérieurs qui contiennent un grand nombre de copies de séquences de nucléotides non codantes. Ces conclusions sont susceptibles d'être pertinentes pour les génomes eucaryotes et procaryotes.

Tableau 4.1. Les caractéristiques extraites du gène exemplaire d'*E. coli*.

Caractéristiques	Début du gene		Fin du gene		Début de la région codante		Fin de la région codante		ailleurs	
	Approx.	Detail	Approx.	Detail	Approx.	Detail	Approx.	Detail	Approx.	Detail
Amplitude										
1.	1.00260	0.81969	1.01970	0.76715	1.03370	0.73147	1.0054	0.81524	0.86763	0.93553
2.	0.35675	0.43066	0.40428	0.45234	0.42143	0.47906	0.38229	0.42430	0.41904	0.44777
3.	14	15	16	12	15	12	17	14	13	17
4.	1	1	4	3	1	4	3	3	4	2
5.	2	5	3	4	3	3	2	3	2	7
6.	12	12	16	12	12	14	17	15	14	13
Phase										
1.	11	11	9	6	12	9	9	5	6	6
2.	11	11	0	6	11	10	9	5	6	7
3.	2	1	0	1	3	1	3	2	0	1
4.	0	0	0	0	3	0	0	3	0	3

4.6.2. Deuxième ensemble de caractéristiques:

On sait que la séquence d'ADN présente différents types de motifs en fonction de leurs fonctions biologiques. Les recherches ont montré que les régions codantes se comportent statistiquement de manière aléatoire par rapport aux régions non codantes. Cette différence statistique peut être exploitée pour identifier la région d'intérêt (par exemple la région codante, ...) dans une nouvelle séquence d'ADN qui n'a pas encore été annotée. Dans cette thèse, nous considérons la fréquence des transitions de base à base à l'intérieur d'une région. Sur la base de la séquence de caractères ADN, un ensemble de caractéristiques est généré. Ceux-ci sont montrés dans les tableaux 4.2 à 4.6. Chaque ligne de la table contient le nombre de transitions d'une base spécifique à chacune des quatre bases normalisées par la longueur de la fenêtre [52]. Dans les tableaux suivants, un exemple calculé à partir de la séquence d'*E. coli* est présenté. Comme on peut le voir à partir des tableaux, les fréquences de transition ont été calculées pour cinq régions différentes de la séquence d'un gène exemplaire, autour des limites de gènes et de régions codantes et ailleurs dans la séquence, afin de refléter les différentes propriétés.

Tableau 4.2. Fréquences de transition à l'intérieur de la région autour du début du gène.

	A	C	G	T
A	0.114750	0.016393	0.081967	0.114750
C	0.032787	0	0.032787	0.032787
G	0.098361	0.049180	0.049180	0.049180
T	0.081967	0.032787	0.065574	0.131150

Tableau 4.3. Fréquences de transition à l'intérieur de la région autour de la fin du gène.

	A	C	G	T
A	0.098361	0.032787	0.016393	0.114750
C	0.049180	0.098361	0.032787	0.065574
G	0.032787	0.081967	0.032787	0.032787
T	0.081967	0.049180	0.098361	0.065574

Tableau 4.4. Fréquences de transition à l'intérieur de la région autour du début de la région codante.

	A	C	G	T
A	0.114750	0.016393	0.065574	0.098361
C	0.032787	0	0.032787	0.032787
G	0.065574	0.032787	0.049180	0.065574
T	0.081967	0.049180	0.065574	0.180330

Tableau 4.5. Fréquences de transition à l'intérieur de la région autour de la fin de la région codante.

	A	C	G	T
A	0.098361	0.049180	0.049180	0.098361
C	0.081967	0.032787	0.032787	0.049180
G	0.032787	0.065574	0.065574	0.049180
T	0.065574	0.065574	0.065574	0.081967

Tableau 4.6. Fréquences de transition à l'intérieur d'une région ailleurs dans la séquence.

	A	C	G	T
A	0.114750	0.032787	0.049180	0.114750
C	0.114750	0.016393	0.032787	0.049180
G	0.032787	0.098361	0.049180	0.032787
T	0.049180	0.065574	0.081967	0.049180

Sur la base de l'information structurelle donnée par la séquence d'ADN, comme indiqué dans les tableaux, nous pouvons constater que le plus grand nombre de mesures basées sur la fréquence du nucléotide pour les différents types de séquences d'ADN, limites de gènes et de régions codantes, ont des distributions différentes de l'occurrence des nucléotides. Chacune de ces mesures peut avoir un pouvoir discriminant. Quelques cas de similarités sont trouvés dans une mesure entre deux régions et seulement dans un cas il y a trois valeurs de mesures similaires entre deux régions. A l'exception, du contenu en dinucléotide CG est statistiquement similaire pour les cinq types de régions mentionnées ci-dessus. Cependant, ce n'est pas le cas pour les autres fréquences de transition où elles présentent une différence significative.

4.7. Système d'inférence neuro-flou adaptatif:

Dans cette étude, nous utilisons des réseaux de systèmes d'inférence floue (ANFIS) basés sur un réseau adaptatif pour construire les modèles neuro-flous individuels dans l'ensemble neuro-flou, qui consiste en quatre modèles individuels ANFIS. ANFIS est un système flou de type Sugeno dans une structure de réseau à cinq couches. Des fonctions d'appartenance à une cloche généralisées sont utilisées sur chaque entrée. Le nombre de fonctions d'appartenance pour chaque entrée est de deux. L'algorithme de rétropropagation est utilisé pour l'apprentissage des fonctions d'appartenance, tandis que l'algorithme des moindres carrés moyens (least mean squares algorithm LSE) détermine les coefficients des paramètres linéaires dans la partie conséquente des règles. Chaque modèle ANFIS a été implémenté en utilisant le progiciel MATLAB (MATLAB version 8.6 avec la boîte à outils de logique floue).

4.8. Modèle d'ensemble neuro-flou:

Sur la base des paramètres proposés pour caractériser le comportement du signal génomique, quatre ANFIS ont été utilisés pour les 20 caractéristiques du premier ensemble et pour les 16 caractéristiques du second ensemble. Pour le premier ensemble, les quatre ANFIS: ANFIS 1, ANFIS 2, ANFIS 3 et ANFIS 4 se réfèrent aux entrées caractérisant l'amplitude de détail, la phase de détail, l'amplitude d'approximation et la phase d'approximation, respectivement. Pour le second ensemble, chaque ligne du tableau de la fréquence des transitions est l'entrée de l'un des quatre ANFIS. Un test est également effectué par l'entrée à quatre dimensions des colonnes de la table des transitions. Les quatre sorties uniques des quatre ANFIS sont combinées pour avoir une seule sortie. La stratégie de combinaison de sortie de notre modèle d'ensemble neuro-flou est une simple somme des valeurs de sortie obtenues à partir de chaque réseau neuro-flou individuel. La sortie unique obtenue sera considérée comme la sortie de l'ensemble, ce qui indique la prédiction. Dans cette méthode de combinaison simple, les réseaux neuro-flous individuels ont le même poids et effectuent la même tâche. Une prédiction d'une instance est effectuée en fonction de la prédiction obtenue à partir de chaque ANFIS.

4.8.1. Apprentissage d'ensemble :

Dans cette approche, une fenêtre rectangulaire de 61 nucléotides glisse sur la séquence d'ADN est utilisée et le contenu de la fenêtre est transformé en entrée pour le système neuro-flou. L'activation d'une seule sortie indique si le nucléotide situé au centre de la fenêtre appartient au premier nucléotide d'un codon dans les limites des régions codantes ou du gène. La sélection de la longueur de fenêtre appropriée est une question très importante, car le choix

de la longueur de la fenêtre peut avoir un effet significatif sur les résultats de la prédiction. Dans la présente étude, la longueur de la fenêtre a été choisie de 61 nucléotides. Ceci est dû au fait que nous avons trouvé que de meilleures précisions de prédiction résultent d'une petite fenêtre car elles fournissent des informations plus concentrées au système. Une grande fenêtre entraîne une mauvaise résolution pour détecter les limites de régions codantes et de gènes. À mesure que la taille de la fenêtre augmente, la sortie du système a tendance à avoir des pics avec des valeurs plus petites, ce qui réduit les performances de prédiction par rapport au système entraîné avec une fenêtre plus petite.

Le système neuro-flou est entraîné sur des exemples de régions codantes et non codantes; la sortie montre des pics d'amplitude 1, 2, 3 et 4 qui indiquent respectivement le début du gène, la fin du gène, le début et la fin de la région codante. La sortie 0 est référencée ailleurs dans la séquence.

Des séquences d'*E. coli* et de *C. elegans* ont été utilisées avec une structure génétique déterminée. Les caractéristiques sont générées à chaque fois que la fenêtre est décalée; l'ensemble complet de données d'apprentissage est obtenu en faisant glisser la fenêtre sur la séquence. La fenêtre est décalée d'un nucléotide à la fois. En faisant glisser la fenêtre le long d'une séquence d'ADN, le système est capable de générer des signaux continus qui montrent la prédiction sur la longueur de la séquence.

La précision de la prédiction des différentes régions est testée sur un ensemble de séquences de test. Les sections suivantes présentent un résumé des résultats obtenus en utilisant cette méthodologie.

4.8.2. Apprentissage et tests :

Dans cette expérience, les ensembles de données (séquences d'*E. coli* et de *C. elegans*) ont été divisés en deux ensembles de données distincts: les données d'apprentissage et l'ensemble de données de test, pour l'apprentissage et l'évaluation de nos modèles. L'ensemble de données d'apprentissage a été utilisé pour optimiser les modèles en ajustant les paramètres de la fonction d'appartenance pour mieux correspondre aux données, tandis que l'ensemble de données de test a été utilisé pour contrôler le sur-ajustement (overfitting) des modèles. Où l'ensemble de test est utilisé pour déterminer quand l'entraînement doit être terminé pour éviter le sur-ajustement. Le sur-ajustement est un problème courant dans la construction du modèle ANFIS, qui se produit lorsque le modèle ANFIS est surentraîné sur les données.

Dans cette étude, nous avons employé une méthode appelée la méthode d'arrêt-précoce (early-stopping method), c'est une stratégie populaire pour éviter le sur-ajustement lors de

l'apprentissage d'un modèle ANFIS. Où le nombre d'itérations d'apprentissage (époques) est déterminé tant que les performances sur les données d'évaluation s'améliorent jusqu'à ce que l'effet de sur-ajustement apparaisse et la performance commence à se dégrader. On peut le distinguer par l'inspection directe de l'erreur de test, lorsque l'erreur quadratique moyenne (MSE) devient plus grande avec l'augmentation du nombre d'itérations d'entraînement ceci indique le début d'un sur-ajustement.

Le critère de validation des modèles est l'erreur quadratique moyenne (MSE). Ce critère permet de juger de la qualité des modèles, en donnant une estimation globale et numérique de l'écart entre les résultats prédits et les données observées. Le modèle est bien optimisé si la valeur du MSE est proche de zéro.

Pour *E. coli*, l'ensemble d'apprentissage a été extrait de trente gènes avec une structure de gène déterminée expérimentalement. La longueur moyenne d'un gène est de 704 nucléotides. La performance de généralisation des réseaux est validée en utilisant des données extraites de cent cinquante autres gènes. Les ensembles d'apprentissage et de tests des quatre ANFIS, fenêtrés par une fenêtre rectangulaire composée de 61 nucléotides, ont été formés par des vecteurs de 20 dimensions (dimension des vecteurs caractéristiques extraits) pour le premier ensemble de caractéristiques et de 16 dimensions pour le second ensemble. Chaque ANFIS est entraîné en utilisant 200 séquences d'ADN. Les quatre ANFIS sont entraînés pour 100 époques d'entraînement pour le premier ensemble de caractéristiques et 600 époques d'entraînement pour le deuxième ensemble. Le pas d'adaptation des paramètres a une valeur initiale de 0,01. À la fin des époques d'apprentissage, la courbe de convergence de l'erreur de réseau (erreur quadratique moyenne) de chaque ANFIS a été calculée. Les valeurs de convergence finales des quatre ANFIS sont comprises entre 0,0238 et 0,0192 pour le premier ensemble de caractéristiques, tandis que les valeurs de convergence finales sont entre 0,0232 et 0,0213 pour le deuxième ensemble de caractéristiques fourni au système par lignes et entre 0,0229 et 0,0213 lorsque cet ensemble de caractéristiques fourni au système par des colonnes. Chaque ANFIS est ensuite testé; le nombre total d'exemples de test est de 200 séquences d'ADN constituées de 112 régions codantes.

Pour *C. elegans*, chaque ANFIS est entraîné en utilisant 200 séquences d'ADN choisies aléatoirement de trois gènes. La longueur d'un gène est comprise entre 2748 et 49262 nucléotides. La longueur d'un exon varie entre quelques centaines et plusieurs centaines de nucléotides avec une moyenne de 239 nucléotides. Les modèles individuels d'ANFIS sont entraînés en utilisant des sous-ensembles de gènes. Un grand nombre de données entraîne un temps de calcul important. Pour cette raison, le nombre possible de données sélectionnées doit

être limité. Par conséquent, différents sous-ensembles de données et un nombre minimal et différent de gènes sont nécessaires. Pour cet effet, une méthodologie a été suivi le principe est de faire positionné la fenêtre d'extraction de caractéristiques sur les segments critiques dans la séquence d'ADN tels que les limites des gènes et des exons. Chaque ANFIS est entraîné pour 100 époques d'entraînement pour le premier ensemble de caractéristiques et 600 époques d'entraînement pour le deuxième ensemble. Dans les deux cas, le pas d'adaptation des paramètres a une valeur initiale de 0,01. À la fin des époques d'entraînement, les valeurs de convergence d'erreur quadratique moyenne finale des quatre ANFIS sont comprises entre 0,0206 et 0,0194 pour le premier ensemble de caractéristiques. Alors que pour le second ensemble de caractéristiques, fourni au système par lignes et par colonnes, les valeurs de convergence finales sont comprises entre 0,0205 et 0,0200 et entre 0,0205 et 0,0203, respectivement. Sur un ensemble de test contenant 202 séquences d'ADN constituées de 112 exons chaque ANFIS est ensuite testé.

Afin d'évaluer la performance prédictive des modèles ANFIS, des paramètres statistiques tels que l'indice d'accord (IA), le carré du coefficient de corrélation (R^2), l'erreur quadratique moyenne de normalisation (NMSE) et le biais fractionnaire (FB) ont été utilisés. Ces critères ont été évalués en utilisant les équations suivantes:

$$IA = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left(|O_i - O_m| + |P_i - P_m|\right)^2} \tag{4.1}$$

$$R^2 = \frac{\left(\sum_{i=1}^{N}(O_i - O_m)(P_i - P_m)\right)^2}{\sum_{i=1}^{N}(O_i - O_m)^2(P_i - P_m)^2} \tag{4.2}$$

$$NMSE = \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(P_i)^2} \tag{4.3}$$

$$FB = \frac{P_m - O_m}{0.5(P_m + O_m)} \tag{4.4}$$

Où O_i est la valeur observée et P_i est la valeur prédite, O_m et P_m sont les valeurs moyennes des valeurs observées et prédites, respectivement. Et N est le nombre de données d'entraînement.

IA varie entre 0 et 1; un maximum d'IA représente un accord parfait entre les valeurs observées et prédites. L'accord parfait est atteint lorsque R^2 est égal à 1. FB représente une mesure de l'accord entre la valeur moyenne des valeurs prédites et celle des valeurs observées. L'accord parfait entre les valeurs observées et prédites est indiqué lorsque FB = NMSE = 0.
Les valeurs de IA, R^2, NMSE et FB obtenues sont fournies dans les tableaux 4.7 et 4.8 pour les séquences d'ADN d'*E. coli* et de *C. elegans*, respectivement. Il apparaît que les valeurs des paramètres de performances statistiques des quatre modèles ANFIS présentent une légère différence lors de l'entraînement en utilisant les différents ensembles de caractéristiques.

Tableau 4.7. Résumé des mesures statistiques des quatre modèles ANFIS entraînés sur différents ensembles de caractéristiques extraits de séquences d'ADN d'*E. coli*.

Ensembles de caractéristiques	Modèle ANFIS	Mesures statistiques			
		IA	R^2	NMSE	FB
Premier ensemble	1	0.7626	1.7441	0.1241	9.2361×10^{-6}
	2	0.6512	1.3096	0.2086	-1.6194×10^{-5}
	3	0.8414	1.5100	0.0898	3.2399×10^{-6}
	4	0.5686	1.1865	0.3290	9.1632×10^{-6}
Deuxième ensemble (lignes)	1	0.7020	1.1894	0.1585	-5.9430×10^{-7}
	2	0.6632	1.3800	0.1890	-5.5228×10^{-7}
	3	0.4986	1.7432	0.5990	-3.0529×10^{-7}
	4	0.7758	1.3436	0.1136	-1.8334×10^{-6}
Deuxième ensemble (colonnes)	1	0.5876	1.7482	0.2939	7.8592×10^{-8}
	2	0.5903	1.2879	0.2803	-2.1880×10^{-6}
	3	0.5553	1.7861	0.3786	1.7372×10^{-6}
	4	0.6319	1.0155	0.2235	-5.6125×10^{-7}

Tableau 4.8. Résumé des mesures statistiques des quatre modèles d'ANFIS entraînés sur différents ensembles de caractéristiques extraits de séquences d'ADN de *C. elegans*.

Ensembles de caractéristiques	Modèle d'ANFIS	Mesures statistiques			
		IA	R^2	NMSE	FB
Premier ensemble	1	0.6918	1.7995	0.1742	1.7231×10^{-6}
	2	0.4637	1.6809	0.9641	$- 2.1846 \times 10^{-6}$
	3	0.8805	1.7536	0.0765	$- 5.7807 \times 10^{-6}$
	4	0.4725	1.7174	0.8497	5.9567×10^{-6}
Deuxième ensemble (lignes)	1	0.5004	1.7699	0.6041	4.7563×10^{-7}
	2	0.7918	1.1098	0.1056	1.3820×10^{-6}
	3	0.7377	1.2098	0.1340	$- 1.1198 \times 10^{-6}$
	4	0.5029	1.6749	0.5829	$- 1.2633 \times 10^{-7}$
Deuxième ensemble (colonnes)	1	0.5531	1.3835	0.3840	2.7299×10^{-6}
	2	0.5443	1.7072	0.4013	$- 2.3332 \times 10^{-7}$
	3	0.8412	1.4645	0.0853	$- 3.2417 \times 10^{-7}$
	4	0.5725	1.6587	0.3172	$- 6.4802 \times 10^{-7}$

Comme indiqué dans les tableaux, les valeurs d'IA étaient entre 0,4986 et 0,8414, les valeurs de R^2 étaient entre 1,1865 et 1,7861, les valeurs de NMSE étaient entre 0,0898 et 0,5990, et les valeurs absolues des valeurs de FB étaient entre $7{,}8592 \times 10^{-8}$ et $1{,}6194 \times 10^{-5}$ pour les séquences d'ADN d'*E. coli*. Cependant, pour les séquences d'ADN de *C. elegans*, les valeurs d'IA étaient comprises entre 0,4637 et 0,8805, les valeurs de R^2 étaient comprises entre 1,1098 et 1,7995, les valeurs de NMSE étaient entre 0,0765 et 0,9641 et les valeurs absolues des valeurs de FB étaient entre $2{,}33332 \times 10^{-7}$ et $5{,}9567 \times 10^{-6}$. Du fait que, l'ajustement entre les valeurs de sortie du modèle et les valeurs observées a montré un bon accord, il a été conclu que le système neuro-flou pouvait être utilisé avec succès pour la prédiction.

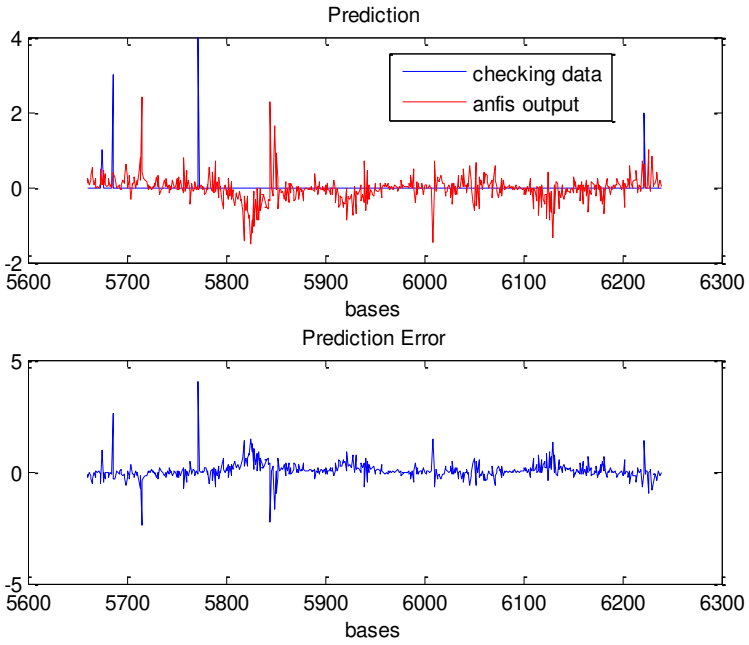
4.9. Résultats des tests et discussion:

4.9.1. *E. Coli*:

4.9.1.1. Premier ensemble de caractéristiques:

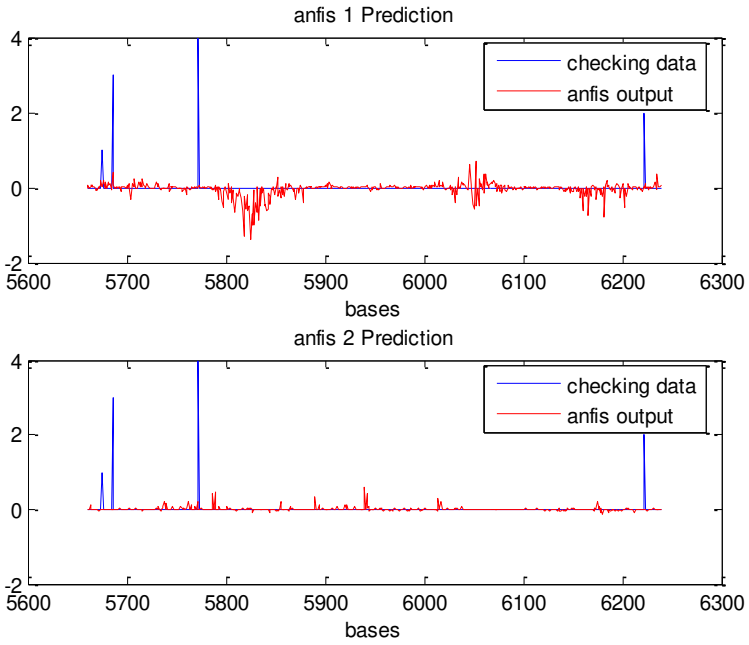
Nous illustrons ici les résultats obtenus avec le codage appliqué sur un exemple d'une séquence d'*E. coli*. La séquence considérée contient un seul gène. La prédiction finale de la structure du gène est montrée par la figure 4.10. Où la prédiction, l'erreur de prédiction,

montrée dans le deuxième tracé, et les quatre valeurs estimées d'ANFIS sont présentées. Les lignes verticales représentées sur la figure sont les sites prédits, qui indiquent une bordure décalée de la région codante interne. Pour la position du gène, le système génère une prédiction correcte, cependant, les valeurs de sortie estimées sont très petites. Les résultats présentés montrent que la représentation d'entrée caractérisant l'amplitude d'approximation utilisée par ANFIS 3 a un effet significatif sur la capacité du système à prédire les limites des régions.

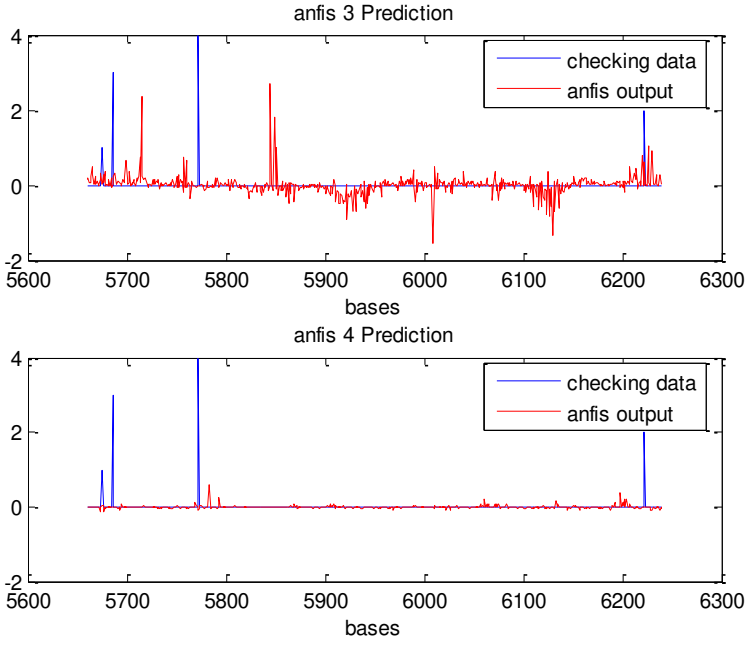


(a)

Figure 4.10. Résultats de prédiction d'une séquence d'*E. coli* en utilisant le système neuro-flou (Premier ensemble de caractéristiques).
(a) Prédiction et erreur de prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



(b)



(c)

Figure 4.10. (La suite)

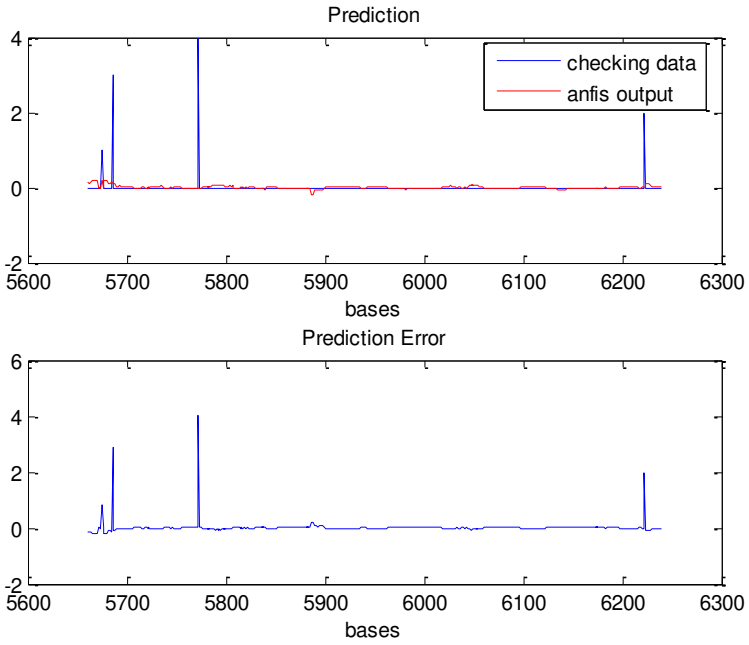
4.9.1.2. Deuxième ensemble de caractéristiques:

Afin de montrer la différence de performance entre les deux représentations d'entrée, le premier et le deuxième ensemble de caractéristiques, les résultats du test du système neuro-flou en utilisant une représentation d'entrée constituée des valeurs générées par la fréquence des transitions d'une base spécifique à chacune des quatre bases sont présenté dans les figures

suivantes. Le test de prédiction a été réalisé sur la même séquence d'*E. coli* utilisée dans le test précédent, i.e. le test avec le premier ensemble de caractéristiques.

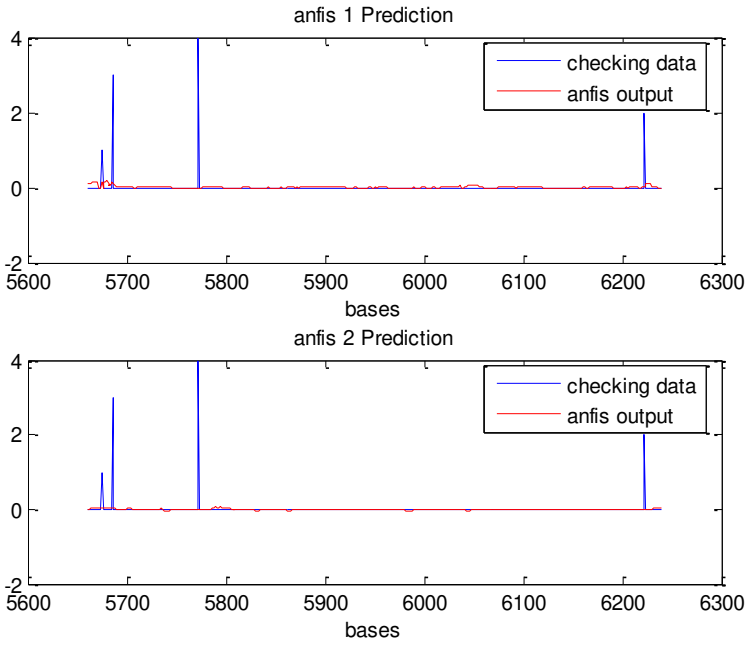
- **Deuxième ensemble de caractéristiques fourni au système par lignes:**

Comme le montre la figure 4.11, avec le deuxième ensemble de caractéristiques, le système neuro-flou n'est pas capable de générer la prédiction correcte. Les fréquences des transitions de nucléotides séquentielles, fournies au système par lignes, qui se produisent dans la fenêtre d'entrée ne caractérisent pas la structure du gène.

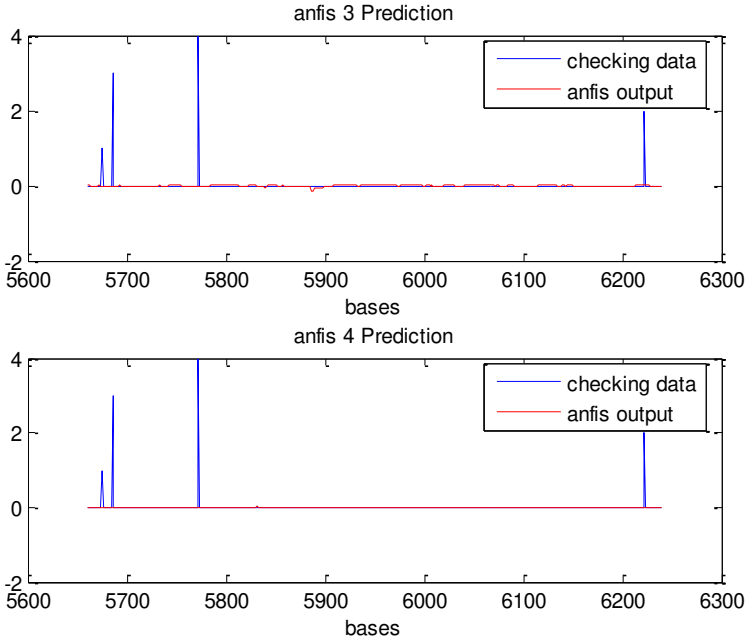


(a)

Figure 4.11. Résultats de prédiction d'une séquence d'*E. coli* en utilisant le système neuro-flou (Deuxième ensemble de caractéristiques fourni au système par lignes).
(a) Prédiction et erreur de prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



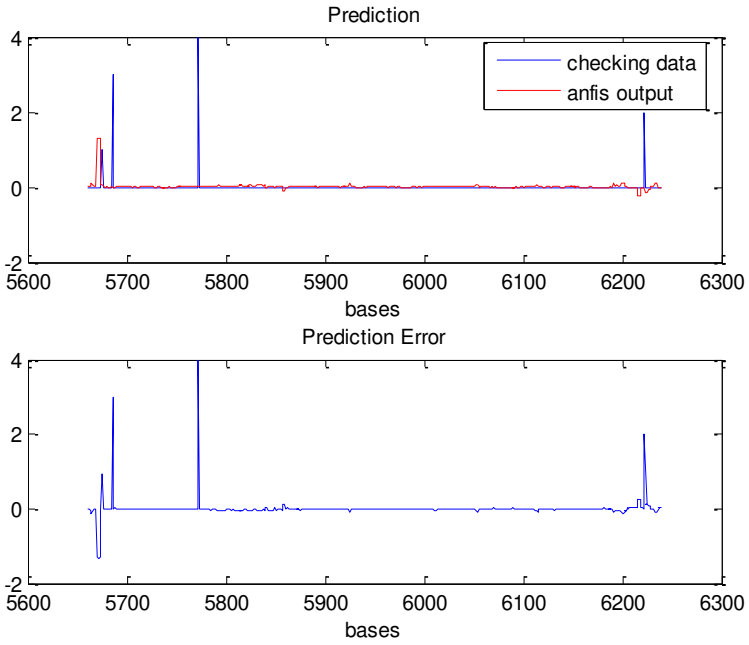
(b)



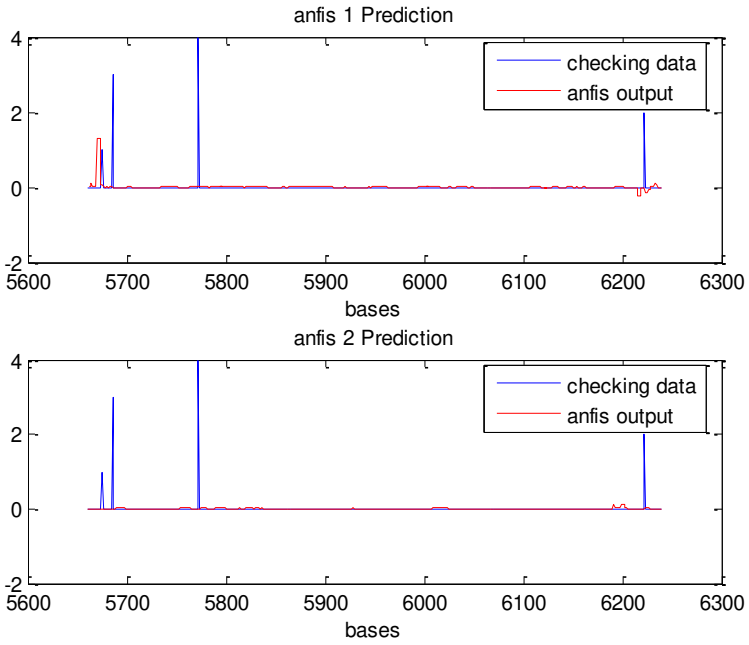
(c)

Figure 4.11. (La suite)

- **Deuxième ensemble de caractéristiques fourni au système par colonnes:**

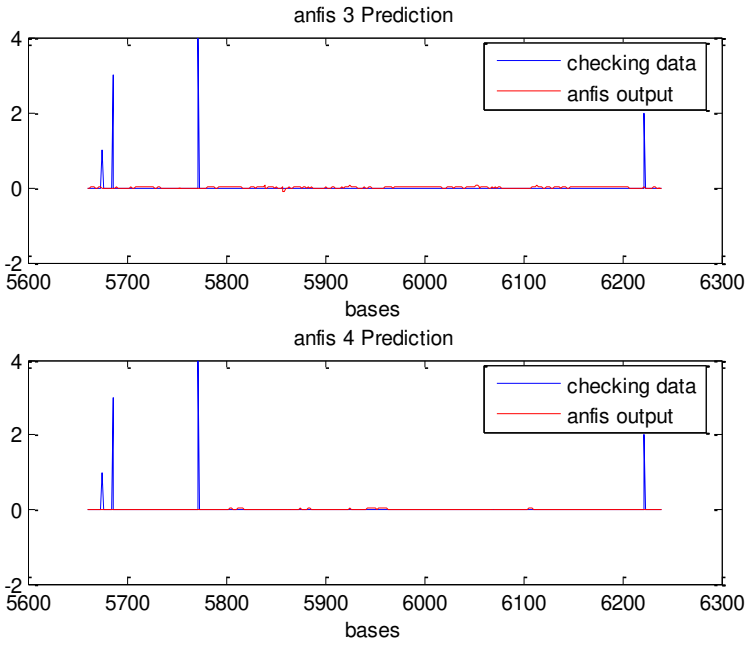


(a)



(b)

Figure 4.12. Résultats de prédiction d'une séquence d'*E. coli* en utilisant le système neuro-flou (Deuxième ensemble de caractéristiques fourni au système par colonnes).
(a) Prédiction et erreur de prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



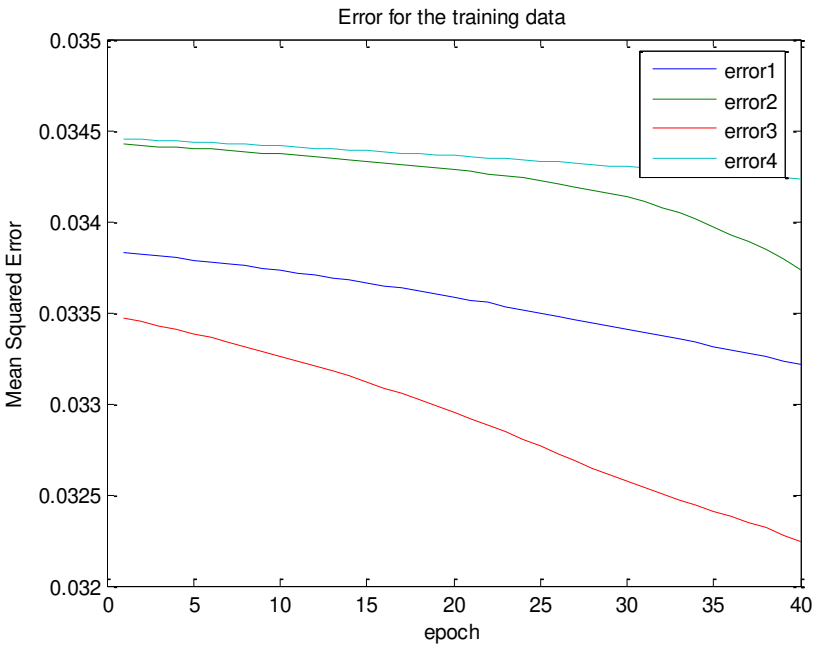
(c)

Figure 4.12. (La suite)

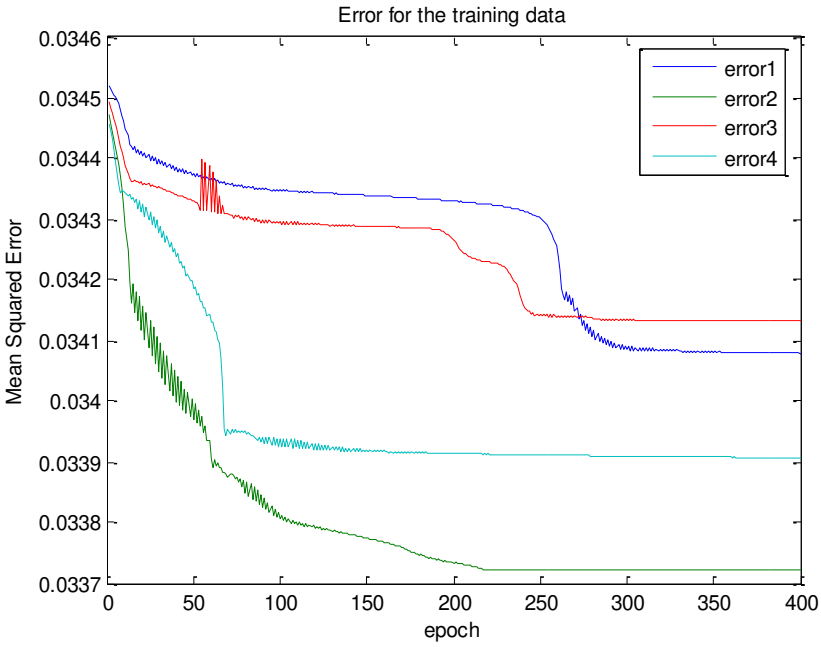
Comme le montre la figure 4.12, à l'exception de la prédiction correcte du début du gène, la représentation d'entrée n'affecte pas significativement la précision de prédiction du système neuro-flou. La sortie d'ANFIS 1 présente la prédiction correcte. Par conséquent, la première colonne, présentée par la sortie d'ANFIS 1, du tableau donne les informations utiles pour la prédiction du début du gène. La première colonne de la table, contient les probabilités que chacune des quatre bases sera suivie par la base A à l'intérieur de la fenêtre, caractérise le début du gène.

Les résultats indiquent que, pour *E. coli*, le système neuro-flou avec les caractéristiques d'ondelettes produit une meilleure prédiction que le système neuro-flou entraîné en utilisant les caractéristiques de la fréquence de transitions. En outre, il convient de souligner que le système neuro-flou avec la représentation d'entrée le deuxième ensemble de caractéristiques a moins de temps d'apprentissage par rapport au système neuro-flou avec la représentation d'entrée le premier ensemble de caractéristiques où l'apprentissage du système neuro-flou est assez lent.

• Erreur pour les données d'entraînement:

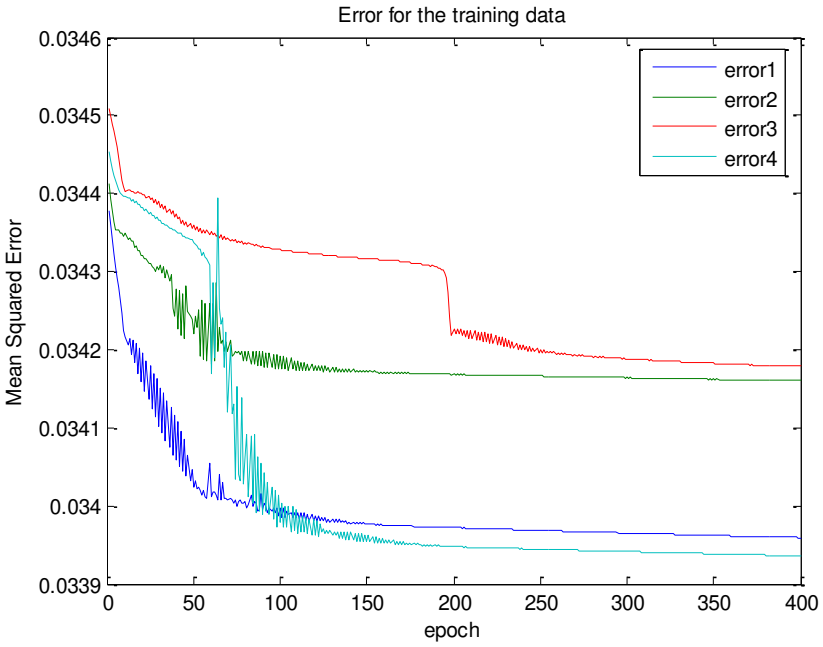


(a)



(b)

Figure 4.13. Les courbes de convergence d'erreur de réseau de chaque ANFIS (*E. coli*).
(a) Premier ensemble (b) Deuxième ensemble (lignes) (c) Deuxième ensemble (colonnes)



(c)

Figure 4.13. (La suite)

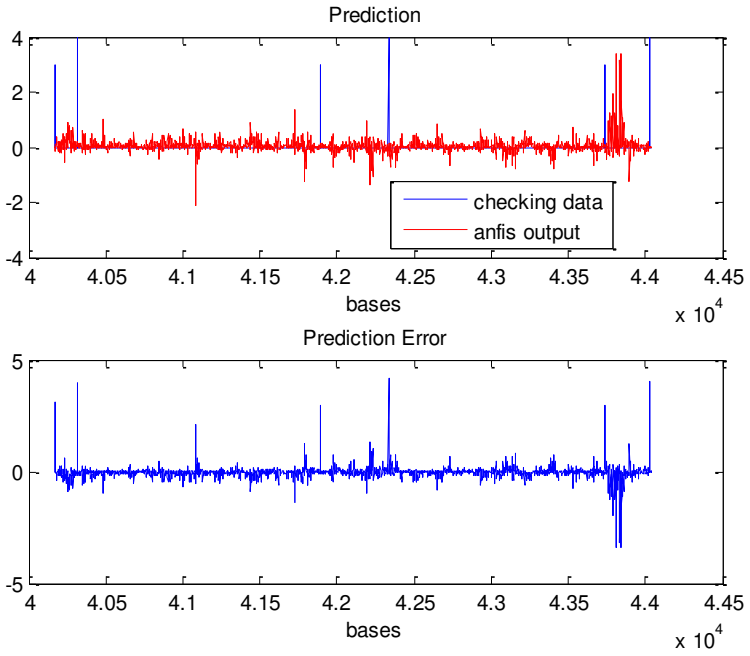
La courbe de convergence de l'erreur de réseau (erreur quadratique moyenne) de chaque ANFIS a été calculée comme le montre la figure 4.13. Comme illustré, les 40 premières époques (figure 4.13 (a)) et 400 époques (figure 4.13 (b) et 4.13 (c)) pour le premier ensemble de caractéristiques et le deuxième ensemble de caractéristiques, respectivement, sont présentés. A partir de ces courbes on constate que les courbes de convergence pour le premier ensemble de caractéristiques semblent régulières tandis que les mesures d'erreur pour le second ensemble de caractéristiques subissent des combinaisons consécutives d'une augmentation suivie d'une diminution. A partir du comportement irrégulier des courbes d'erreur observées, ces paramètres ne permettent pas de discriminer sans ambiguïté entre les différentes régions d'une séquence d'ADN. Ainsi, comprendre l'origine de la corrélation observée entre la discrimination et le comportement régulier de la courbe d'erreur.

Dans la présente étude, toutes les caractéristiques utilisées dans les descriptions ont différents niveaux de saillance (saliency). La saillance des caractéristiques fournit un moyen de choisir les caractéristiques qui conviennent le mieux à la prédiction. Sur la base de l'analyse de la courbe de convergence d'erreur quadratique moyenne de chaque ANFIS, l'ANFIS qui présente l'erreur minimale ses paramètres d'entrée ont un impact sur la performance de la prédiction et pourraient fournir des meilleures prédictions.

4.9.2. *C. elegans*:

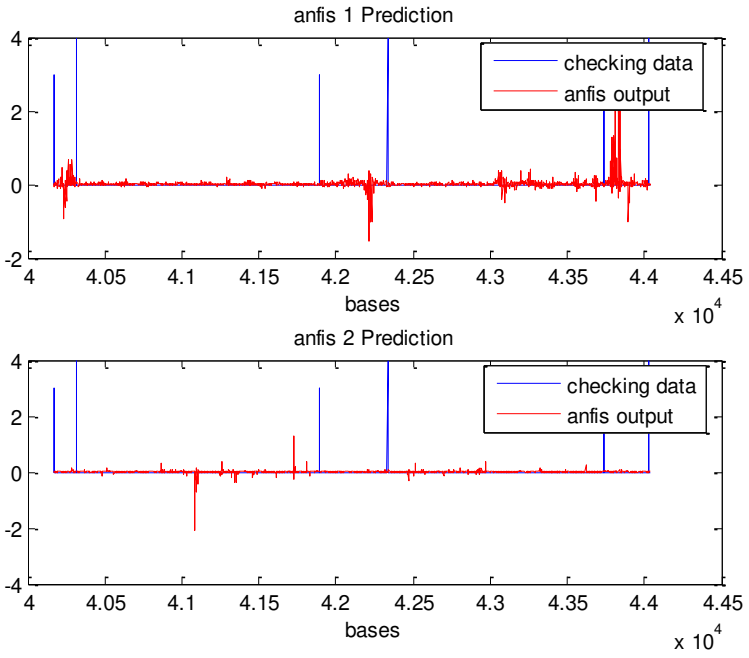
4.9.2.1. Premier ensemble de caractéristiques:

La reconnaissance de la position d'exons successifs d'une séquence de *C. elegans*, contenant trois exons, obtenue en utilisant le premier ensemble de caractéristiques est représentée dans la figure 4.14. La figure montre que le début du troisième exon de la séquence est correctement prédit. Selon la figure, le début du troisième exon dans la séquence est indiqué par un signal relativement élevé dans les sorties du ANFIS 1 et du ANFIS 3. Les sorties du ANFIS 2 et du ANFIS 4 présentent des signaux relativement faibles. Les résultats présentés montrent que la représentation d'entrée caractérisant d'amplitude de détail et d'amplitude d'approximation utilisée par ANFIS 1 et ANFIS 3 a un effet significatif sur la capacité du système à prédire les limites des régions.

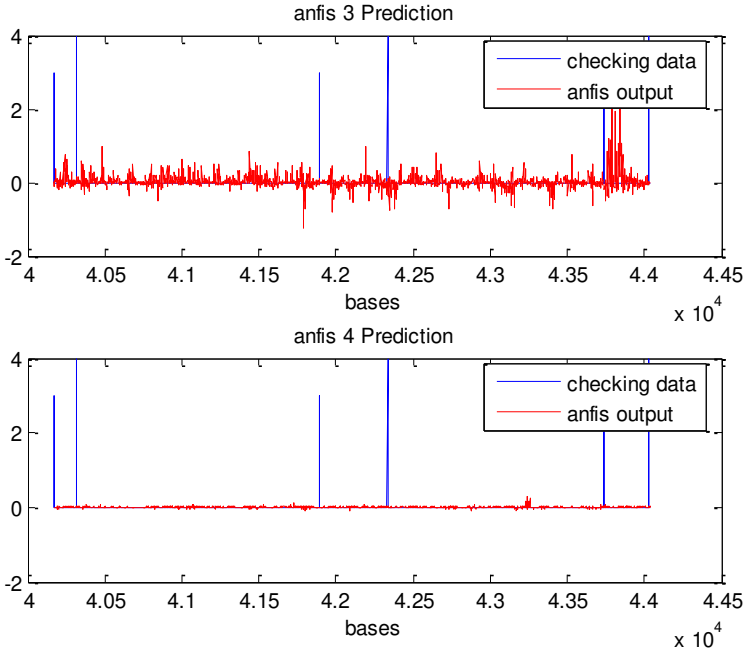


(a)

Figure 4.14. Résultats de prédiction d'une séquence de *C. elegans* en utilisant le système neuro-flou (Premier ensemble de caractéristiques).
(a) Prédiction et erreur de prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



(b)



(c)

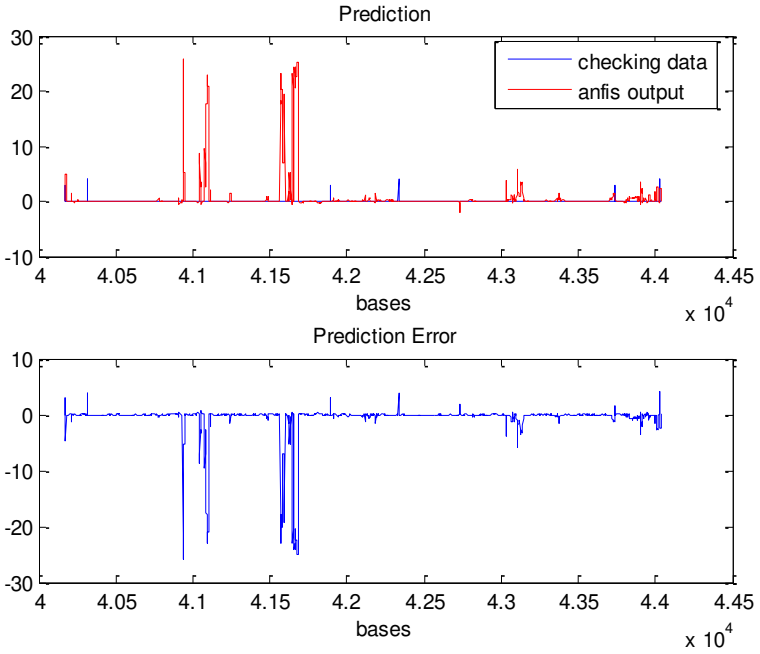
Figure 4.14. (La suite)

4.9.2.2. Deuxième ensemble de caractéristiques:

Pour la même séquence, les résultats de la prédiction du système avec le second ensemble de caractéristiques sont montrés dans les figures 4.15 et 4.16.

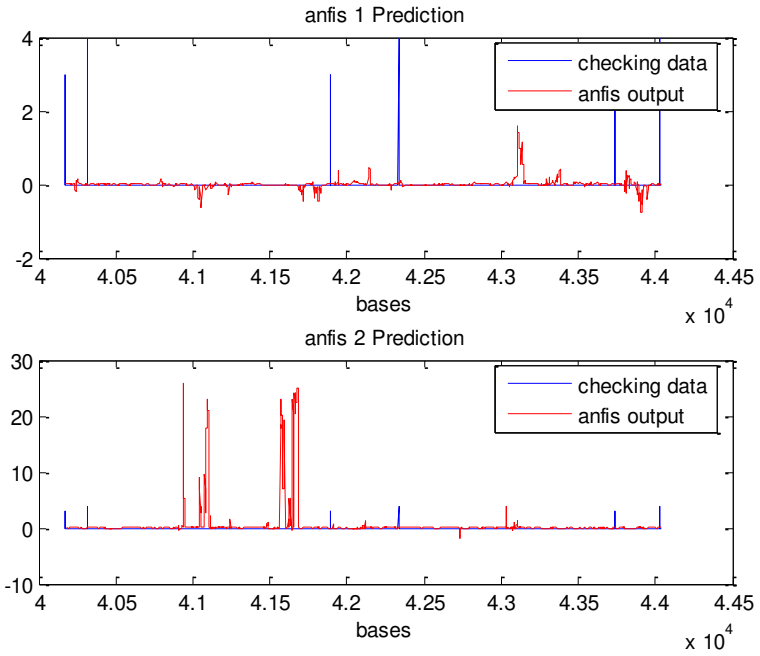
• Deuxième ensemble de caractéristiques fourni au système par lignes:

Dans ce cas, une prédiction correcte du début du premier exon et de la fin du troisième exon, indiquée par la sortie du ANFIS 3, et le début du troisième exon, indiqué par la sortie du ANFIS 4, sont obtenus (figure 4.15). Par conséquent, les statistiques que les bases G et T trouvées suivies par n'importe quelle base permet de prédire des exons dans la séquence d'ADN donnée avec une précision raisonnable.

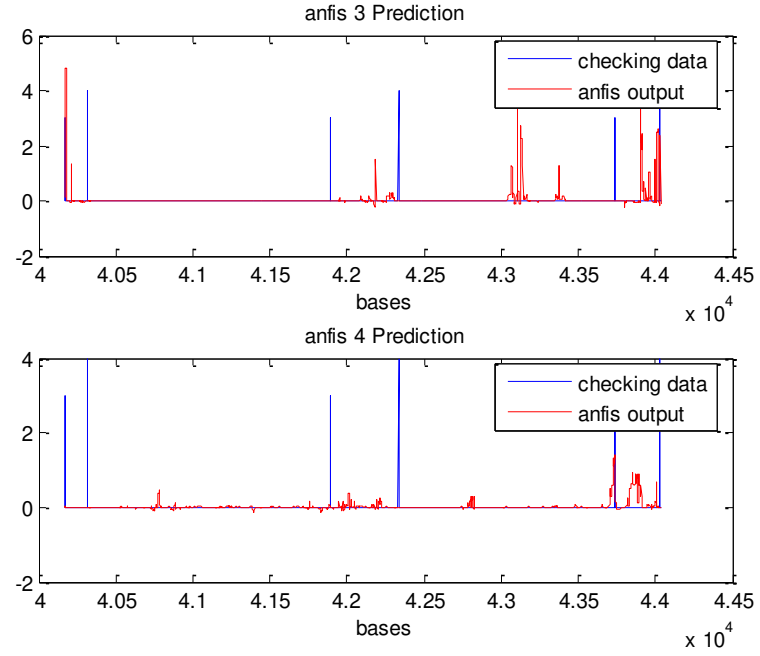


(a)

Figure 4.15. Résultats de prédiction d'une séquence de *C. elegans* en utilisant du système neuro-flou (Deuxième ensemble de caractéristiques fourni au système par lignes).
(a) Prédiction et erreur de Prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



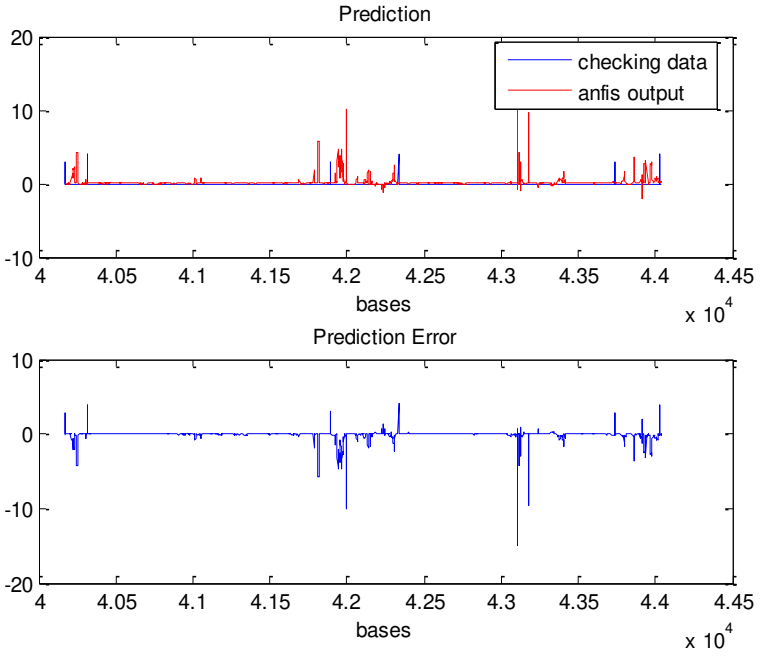
(b)



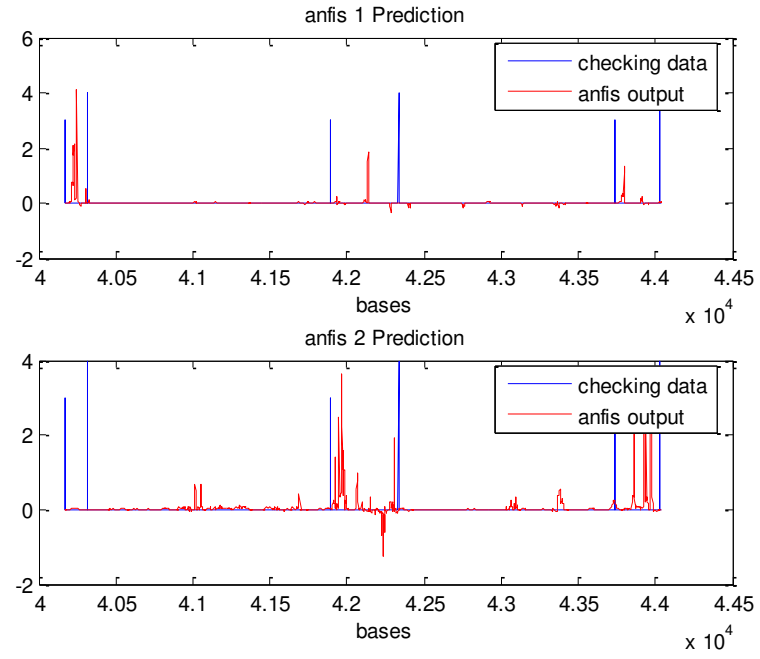
(c)

Figure 4.15. (La suite)

• Deuxième ensemble de caractéristiques fourni au système par colonnes:

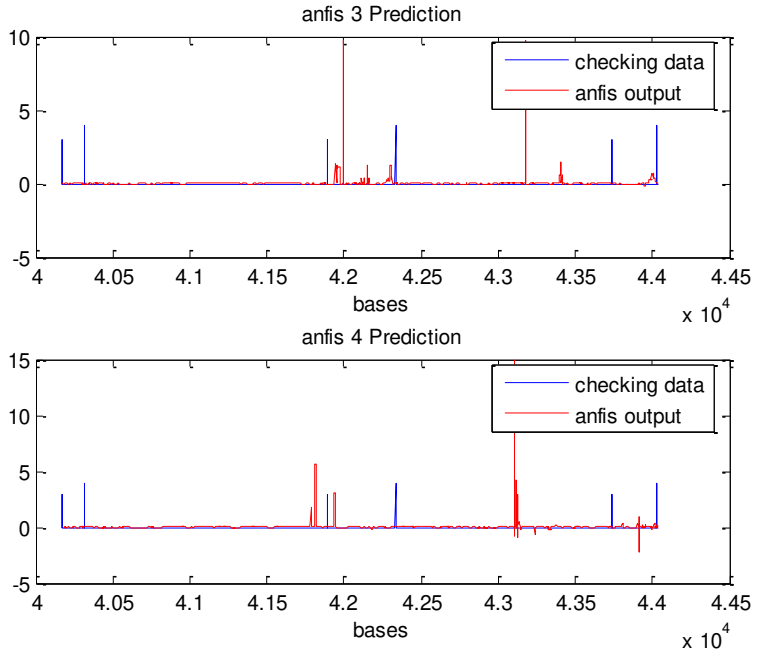


(a)



(b)

Figure 4.16. Résultats de prédiction d'une séquence de *C. elegans* en utilisant le système neuro-flou (Deuxième ensemble de caractéristiques fourni au système par des colonnes). (a) Prédiction et erreur de prédiction (b) Prédictions ANFIS 1 et ANFIS 2 (c) Prédictions ANFIS 3 et ANFIS 4



(c)

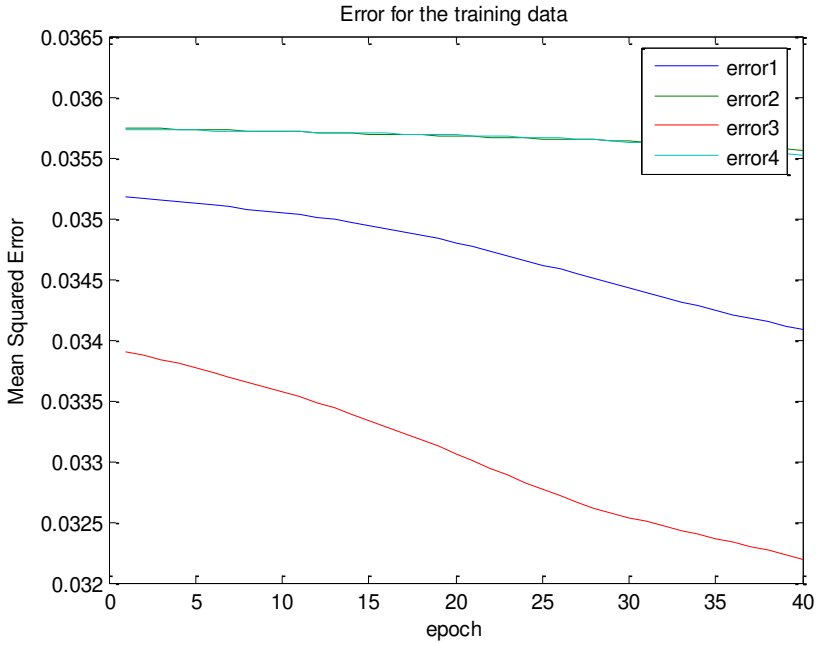
Figure 4.16. (La suite)

L'information individuelle à la sortie de chaque unité ANFIS est présentée dans la figure 4.16. La présence des exons dans la séquence est indiquée par des pics relativement élevés dans les sorties du ANFIS 1 et du ANFIS 2. Les résultats montrent que la détection des exons dans la séquence de test est améliorée en fournissant les paramètres caractéristiques par colonnes au système. Ainsi, les statistiques calculées dans les colonnes 1 et 2 pour les caractères particuliers A et C, les probabilités que chacune des quatre bases seront suivies par les bases A et C, sont de bons candidats pour la prédiction des exons dans une séquence d'ADN.

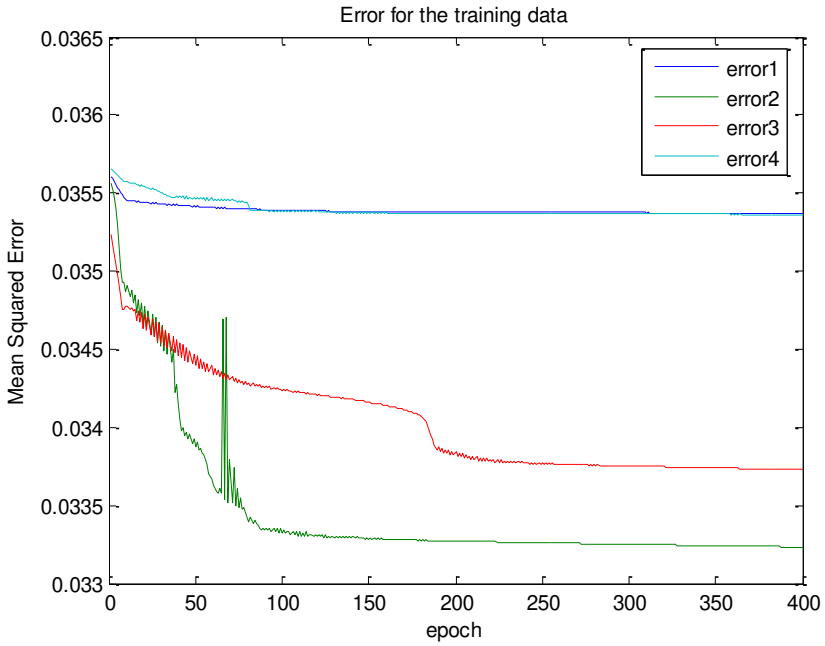
Nous avons trouvé que, pour *C. elegans*, le système neuro-flou avec le second ensemble de caractéristiques donnait de meilleurs résultats que le système avec le premier ensemble de caractéristiques, conduisant à une amélioration significative des résultats. Les mesures de fréquence calculées dans les colonnes 1 et 2, lignes 3 et 4 rendent la prédiction finale d'exon plus fiable. Ces mesures présentent un certain pouvoir discriminant. Cela souligne la nécessité d'inclure les compositions GA, GC, TA et TC dans le cadre des informations à fournir au système neuro-flou.

• Erreur pour les données d'entraînement:

L'évolution de l'erreur d'entraînement, pour les différentes données, durant les époques d'apprentissage est représentée dans la figure suivante.

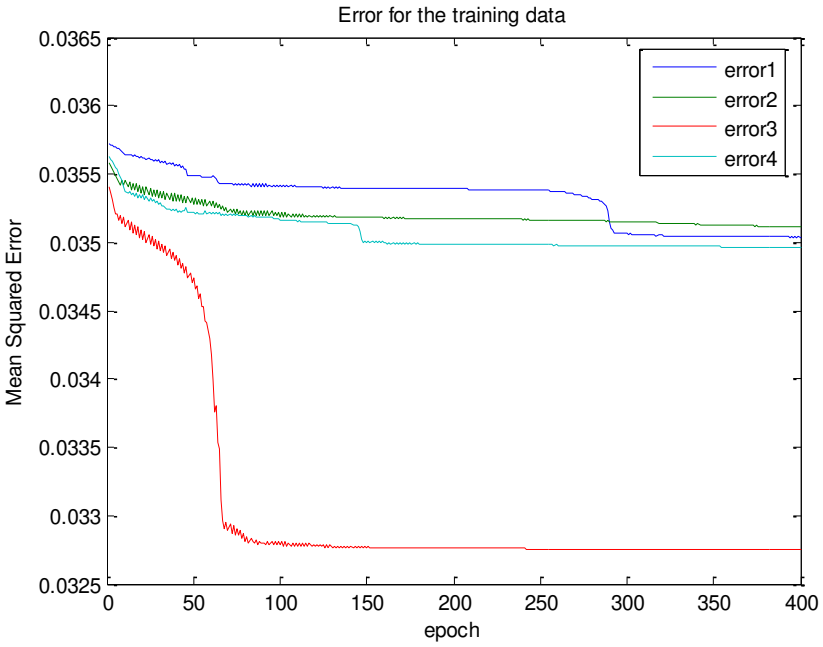


(a)



(b)

Figure 4.17. Les courbes de convergence d'erreur de réseau de chaque ANFIS (*C. elegans*).
(a) Premier ensemble (b) Deuxième ensemble (lignes) (c) Deuxième ensemble (colonnes)



(c)

Figure 4.17. (La suite)

La progression de l'erreur (erreur quadratique moyenne) de chaque ANFIS par rapport aux époques a été examinée. Comme le montrent les figures 4.17 (b) et 4.17 (c), on se rend compte que les fluctuations sont moins prononcées que celles obtenues pour la séquence d'*E. coli*, comme signature de la performance et la précision de la prédiction.

4.9.3. Fonctions d'appartenance :

Les fonctions d'appartenance finales des paramètres d'entrée (après entraînement) par rapport aux fonctions d'appartenance initiales (avant entraînement) ont été examinées. Certaines fonctions d'appartenance initiales du premier et du second ensemble de caractéristiques sont présentées ci-dessous. La figure 4.18 et la figure 4.20 montrent les fonctions d'appartenance initiales des caractéristiques d'amplitude d'approximation utilisées par ANFIS 3 comme entrées et la première colonne du tableau de la fréquence de transitions les entrées de l'ANFIS 1 (second ensemble de caractéristiques fourni au système par colonnes), respectivement. La fonction d'appartenance de chaque paramètre d'entrée a été divisée en deux régions. Nous pouvons voir que les fonctions d'appartenance initiales sont équidistantes avec suffisamment de chevauchement dans la plage d'entrée.

Les fonctions d'appartenance des entrées de l'ANFIS 3 après entraînement, comme indiqué ci-dessous dans la figure 4.19, ne changent pas radicalement. Les résultats d'apprentissage pour les entrées d'ANFIS 1, après les périodes d'apprentissage, sont montrés dans la figure 4.21. On constate que ces fonctions d'appartenance comparées à celles d'avant l'entraînement

présentent des changements considérables. Nous voyons comment les fonctions d'appartenance finales tentent de capturer les caractéristiques locales de l'ensemble de données d'apprentissage.

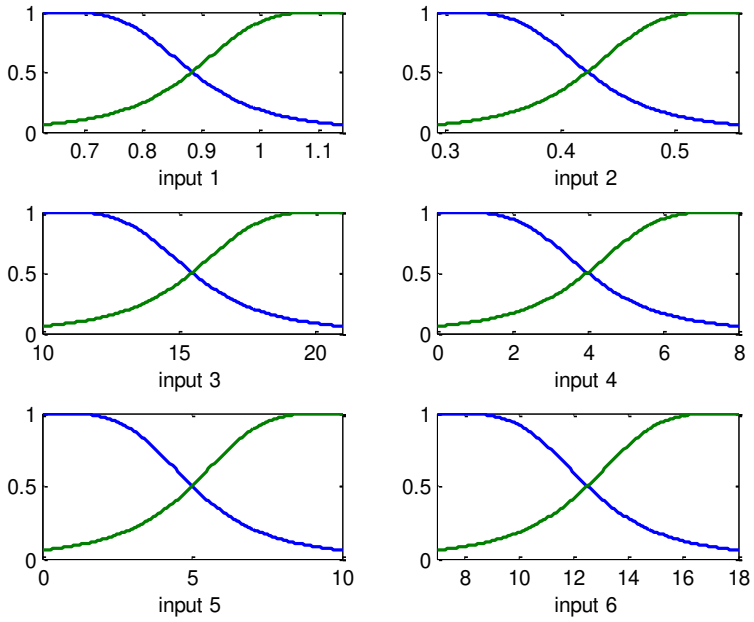


Figure 4.18. Fonctions d'appartenance initiales des caractéristiques de l'amplitude d'approximation (entrées de l'ANFIS 3).

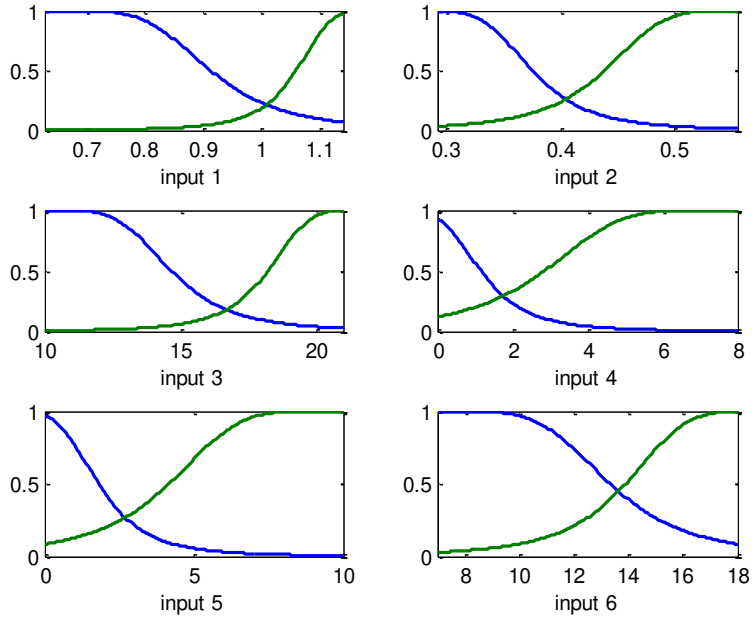


Figure 4.19. Fonctions d'appartenance finales des caractéristiques de l'amplitude d'approximation (entrées de l'ANFIS 3).

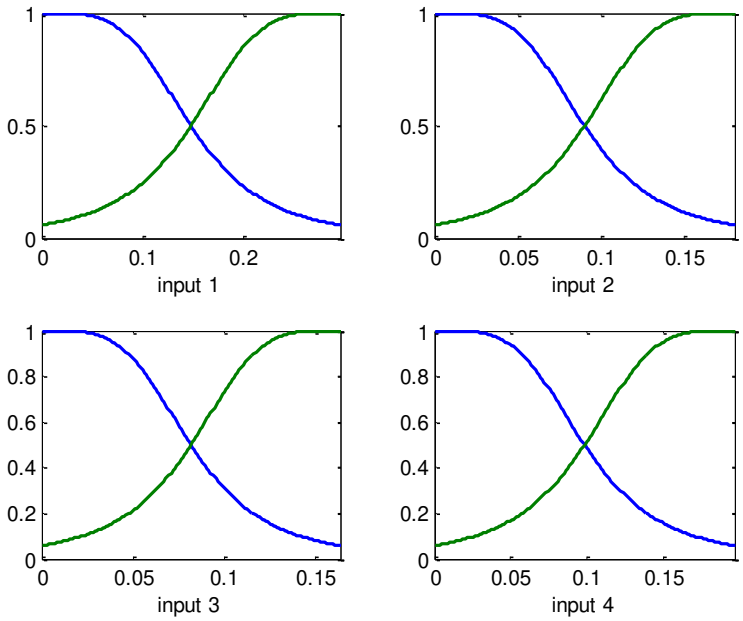


Figure 4.20. Fonctions d'appartenance initiales de la première colonne du tableau de la fréquence des transitions (entrées de l'ANFIS 1).

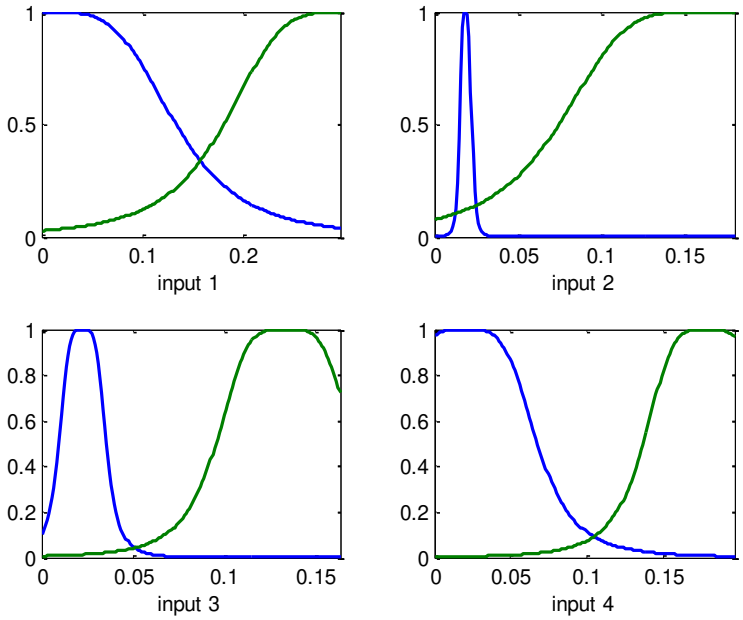


Figure 4.21. Fonctions d'appartenance finales de la première colonne du tableau de la fréquence des transitions (entrées de l'ANFIS 1).

Sur la base de l'examen des fonctions d'appartenance de chaque entrée, on peut mentionner qu'en général, les fonctions d'appartenance des entrées du premier ensemble de

caractéristiques ne changent pas radicalement tandis que celles des entrées du deuxième ensemble présentent des changements considérables.

4.9.4. Performances de prédiction des régions de codage:

Une évaluation de la performance du système neuro-flou avec les paramètres d'entrée proposés pour les ensembles de données des séquences d'*E. coli* et de *C. elegans* apparaît dans les tableaux 4.9 et 4.10, respectivement. Les statistiques de précision sont indiquées pour les régions codantes prédites par le système neuro-flou, telles que testées sur l'ensemble de données. Pour chaque ensemble de caractéristiques, le pourcentage de régions codantes prédites correctement, c'est-à-dire, prédites exactement (les deux extrémités correctes) est indiqué dans la colonne 3. Les pourcentages de régions codantes prédites partiellement (une extrémité correcte), non prédites (non chevauchées par une région codante prédite) ou fausses (ne chevauchant pas une région codante annotée) sont énumérés dans les colonnes 4, 5 et 6, respectivement.

Tableau 4.9. Précision du système neuro-flou pour différents ensembles de caractéristiques testés sur des séquences d'ADN d'*E. coli*.

Ensembles de caractéristiques	Nombre de regions codantes	Regions codantes prédites			
		Exactement (%)	Partiellement(%)	non prédites (%)	fausses (%)
Premier ensemble	112	50	26	24	25
Deuxième ensemble (lignes)	112	27	48	25	0
Deuxième ensemble (colonnes)	112	0	25	75	0

Tableau 4.10. Précision du système neuro-flou pour différents ensembles de caractéristiques testés sur des séquences d'ADN de *C. elegans*.

Ensembles de caractéristiques	Nombre de regions codantes	Régions codantes prédites			
		Exactement (%)	Partiellement(%)	non prédites (%)	fausses (%)
Premier ensemble	112	28	36	36	28
Deuxième ensemble (lignes)	112	50	43	7	57
Deuxième ensemble (colonnes)	112	29	71	0	43

Les mesures standards d'évaluation de la précision des prédictions par exon sont montrées dans le tableau 4.11 telles que mesurées pour les deux ensembles de données de tests des séquences d'ADN d'*E. coli* et *C. elegans*. La sensibilité et la spécificité sont les deux mesures les plus utilisées pour évaluer la précision de la prédiction. Habituellement, la sensibilité (Sn) et la spécificité (Sp) sont définies comme:

Sn = nombre d'exons corrects / nombre d'exons réels (sensibilité)

Sp = nombre d'exons corrects / nombre d'exons prédits (spécificité)

Le tableau donne également (troisième colonne) leur moyenne (Sn + Sp) / 2, exons non prédits (Missing Exons : ME) et faux exons (Wrong Exons : WE). Qui sont définis par:

ME = Exons non prédits / nombre d'exons réels (Exons non prédits)

WE = Faux exons / nombre d'exons prédits (Faux Exons)

Tableau 4.11. Précision évaluée en termes de paramètres de sensibilité et de spécificité.

Ensembles de caractéristiques	*E. coli*					*C. elegans*				
	Sn	Sp	Avg.	ME	WE	Sn	Sp	Avg.	ME	WE
Premier ensemble	0.50	0.49	0.49	0.24	0.25	0.28	0.30	0.29	0.36	0.30
Deuxième ensemble (lignes)	0.27	0.36	0.31	0.25	0	0.50	0.33	0.41	0.07	0.38
Deuxième ensemble (colonnes)	0	0	0	0.75	0	0.29	0.20	0.24	0	0.30

En entraînant les réseaux neuro-flous sur des données empiriques, le système neuro-flou semble, sur la base des résultats de simulation rapportée dans le tableau 4.11, avoir capturé certaines des caractéristiques de la relation entre ces quantités et la présence de régions codantes.

La comparaison des données de précision, le système neuro-flou pour différents ensembles de caractéristiques testés sur des séquences d'ADN d'*E. coli*, montre que le système neuro-flou pour le premier ensemble de caractéristiques est significativement plus précis que le système neuro-flou pour les autres ensembles de caractéristiques. Par contre, la précision du système neuro-flou pour différents ensembles de caractéristiques testés sur des séquences d'ADN de *C. elegans*, présentées dans le tableau 4.11, montre que la représentation d'entrée le second

ensemble, fourni au système par lignes, utilisée par le réseau a un effet significatif sur la capacité du réseau à apprendre.

Près de 36% des exons des séquences d'ADN de *C. elegans* sont des exons courts. Les exons courts sont facilement non prédits, en particulier ceux qui sont bordés par de longs introns, en raison d'un manque de caractéristiques de discrimination présentes dans ces segments. Avec la représentation du second ensemble (lignes), 40% de ces exons ont été correctement prédits, 40% ont été partiellement prédits et 20% ont été non prédits.

Les mesures de précision des programmes existants conçus pour prédire la structure des gènes sont présentées dans le tableau 4.12. Les résultats pour tous les programmes sauf GENSCAN et RBFN Combining sont du tableau 1 de la référence [29]; Les résultats de GENSCAN et RBFN Combining sont respectivement de [53] et [3].

Tableau 4.12. Performance des programmes pour les ensembles de test.

		Précision par nucléotide				Précision par exon				
Programme	Sequences	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
GENSCAN	570 (8)	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
FGENEH	569 (22)	0.77	0.88	0.78	0.80	0.61	0.64	0.64	0.15	0.12
GeneID	570 (2)	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
Genie	570 (0)	0.76	0.77	0.72	n/a	0.55	0.48	0.51	0.17	0.33
GenLang	570 (30)	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	562 (0)	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL II	570 (23)	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	561 (0)	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	570 (28)	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	478 (1)	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	478 (1)	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09
RBFN com.	-	0.92	0.94	0.91	-	0.79	0.80	0.80	0.11	0.06

Remarque. Les mesures standards de la précision de prédiction par nucléotide et par exon (décrites précédemment) sont données. Au niveau des nucléotides, la sensibilité (Sn), la spécificité (Sp), la corrélation approximative (AC) et la valeur du coefficient de corrélation (CC) sont données à des fins de comparaison.

Résumons les meilleures performances (précision) de prédiction des régions codantes dans les séquences d'ADN d'*E. coli* et *C. elegans* en utilisant le système neuro-flou pour différents ensembles de caractéristiques, au niveau de l'exon. Pour *E. coli*, le système neuro-flou proposé avec le premier ensemble de caractéristiques offre les meilleures performances avec la sensibilité (Sn 0,50), la spécificité (Sp 0,49) et la précision moyenne de 0,49, ainsi que les exons non prédits (ME 0,244) et faux exons (WE 0,25). Pour *C. elegans*, le système neuro-flou avec les deuxièmes caractéristiques (lignes) utilisées en entrée fournit les meilleures

performances avec une sensibilité (Sn 0,50), spécificité (Sp 0,33) et précision moyenne de 0,41, ainsi que des exons non prédits (ME 0,07) et faux exons (WE 0,38).

La comparaison des données de précision montre que le système neuro-flou avec la représentation d'entrée proposée est significativement plus précis au niveau de l'exon par toutes les mesures de précision que de nombreux programmes existants (voir le tableau 4.12). La capacité prédictive de notre système est aussi bonne que celle d'autres systèmes de recherche de gènes. En particulier, les performances du système neuro-flou sont comparables à celles de Genie, GenLang et GeneParser3.

Le système neuro-flou avec la représentation d'entrée utilisée a capturé certaines caractéristiques de discrimination présentes dans les segments d'ADN. Les résultats obtenus montrent que ce système est capable de prédire correctement une grande proportion de l'ensemble de tests. En raison de la nature difficile du problème de la prédiction de gène, le grand nombre de programmes de prédiction de gènes a été développé sur la base de techniques de calcul complexes, par exemple HMM et programmation dynamique. Le système neuro-flou avec les caractéristiques proposées, il fournit une solution simple pour traiter les données. De plus, le système neuro-flou est flexible pour inclure de nouvelles caractéristiques ou informations dans un système de reconnaissance des régions codantes à partir desquelles des informations de plus sur les régions codantes sont obtenues. Les tentatives se concentrent actuellement sur l'intégration de caractéristiques supplémentaires dans notre modèle afin d'améliorer la discrimination et la capacité prédictive des régions codantes.

Cette étude a donné lieu à la publication [54] où les résultats obtenus ont été publiés.

4.10. Conclusion :

Le réseau neuro-flou décrit dans les sections précédentes de ce chapitre est un système prédictif de la structure de gène. La sensibilité et la spécificité de ce système de reconnaissance des gènes montrent que la capacité prédictive de notre système est aussi bonne que celle d'autres systèmes de recherche de gènes. Par entraînement du réseau neuro-flou sur des données empiriques, ce dernier semble avoir capturé une partie de la partie la plus essentielle de la relation entre ces quantités et la présence de régions codantes.

Conclusion générale

Un certain nombre de systèmes de prédiction de gènes ont été développés. Ces systèmes utilisent une variété de techniques de calcul sophistiquées, y compris le réseau de neurones, la programmation dynamique, les arbres de décision, le raisonnement probabiliste et les chaînes de Markov cachées. Lorsque les méthodes ci-dessus sont comparées avec une approche basée sur l'informatique souple. L'approche basée sur l'informatique souple est plus robuste et tolérante aux données bruitées et incomplètes. La capacité d'apprentissage et d'adaptation des réseaux neuronaux artificiels et la représentation des connaissances à travers des logiques floues réunies permettent d'exploiter les avantages de chacun d'entre eux.

L'hybridation neuro-floue est un système flou augmenté de réseaux de neurones pour améliorer certaines de ses caractéristiques comme la flexibilité, la vitesse et l'adaptabilité. La logique floue vise à modéliser les modes imprécis de raisonnement et les processus de pensée qui jouent un rôle essentiel dans la remarquable capacité humaine à prendre des décisions rationnelles dans un environnement d'incertitude et d'imprécision. La plupart des systèmes biologiques se comportent de manière floue, l'interaction et l'activité de différents gènes atteignent des niveaux différents.

Des implémentations réussies d'ANFIS dans l'ingénierie des biosystèmes ont été rapportées dans la littérature. En raison de la complexité des biosystèmes, l'accent a été mis sur la mise en œuvre de techniques d'intelligence artificielle qui offrent l'avantage de modéliser des relations non linéaires avec un nombre réduit de paramètres d'entrée, contrairement aux modèles analytiques qui sont trop complexes pour être pratiques et difficiles à implémenter. Ils permettent une prédiction de la dynamique du système en fournissant suffisamment de données lors du processus d'apprentissage.

Ce travail de thèse présente une application d'ANFIS pour la prédiction de régions codantes dans les séquences d'ADN. Quatre ANFIS ont été utilisés pour la prédiction où les prédictions des quatre ANFIS ont été combinées. Deux mesures ont été proposées sur la base des coefficients d'ondelettes et de la fréquence des transitions à l'intérieur d'une fenêtre de longueur fixe comme entrées.

Une analyse des signaux obtenus à partir de séquences d'ADN par les techniques d'analyse du signal tels que la transformée de Fourier, les ondelettes et les bancs de filtres a été effectuée pour mettre en évidence des caractéristiques au sein de ces signaux. Quelques

observations nous a conduit à proposer l'utilisation de la transformation en ondelettes comme outil d'analyse des séquences d'ADN.

Quelques conclusions concernant la pertinence des caractéristiques sur la prédiction des séquences d'ADN ont été obtenues. Dans le cas procaryote, c'est-à-dire *E. coli*, nous avons trouvé que les caractéristiques liées aux coefficients d'ondelettes donnaient de meilleurs résultats que les mesures de la fréquence de transitions. D'autre part, le second ensemble de mesures fournit une indication utile des limites des exons dans les séquences d'ADN du nématode *C. elegans*.

L'originalité de ce travail réside dans les caractéristiques proposées pour la prédiction, sur la base des coefficients d'ondelettes et de la fréquence des transitions, de même que l'architecture du réseau neuro-flou proposée, qui a donné lieu à la publication suivante : « *Prediction of DNA sequences using adaptive neuro- fuzzy inference system* ». Ce travail s'inscrit dans le contexte et l'objectif initial de l'étude : « *Détection d'évènements par les méthodes intelligentes dans les séquences biomoléculaires* ».

La prédiction des gènes est un problème en général de nature difficile. C'est un problème ouvert et important en bioinformatique. Dans cette thèse, nous avons développée un système prédictif neuro-flou. La perspective liée à ce travail est actuellement en phase d'amélioration : Les tentatives se concentrent actuellement sur l'intégration de caractéristiques additionnelles dans notre modèle afin d'améliorer sa capacité prédictive. En utilisant un réseau neuro-flou comme moyen de base pour combiner des informations provenant de différentes sources, nous avons obtenu un cadre flexible pour inclure de nouvelles informations dans notre système de reconnaissance de gène alors qu'une compréhension plus profonde et donc plus d'informations sur les gènes sont acquises.

Bibliographie

[1] F. Dardel et F. Képès, *Bioinformatique Génomique et post-génomique*, Les éditions de l'école polytechnique, Paris, France, 2006.

[2] C. Gibas et P. Jambeck, *Introduction à la bioinformatique : Concepts fondamentaux et outils logiciels*, éditions O'REILLY, Paris, France, 2001.

[3] N. Goel, S. Singh and T. C. Aseri, A comparative analysis of soft computing techniques for gene prediction, Elsevier, Analytical Biochemistry, Vol. 438, N° 1, pp. 14–21, 2013.

[4] K. Raza and R. Parveen, Soft computing approach for modelling genetic regulatory networks', in: N. Meghanathan, D. Nagamalai and N. Chaki (Eds.): Advances in computing & information technology, Springer-Verlag Berlin Heidelberg, 2013, pp.1-11.

[5] C.L. Chuang, C.M. Chen and J. A. Jiang, Infer Genetic/ Transcriptional Regulatory Networks by Recognition of Microarray Gene Expression Patterns Using Adaptive Neuro Fuzzy Inference Systems, in: Y. Jin and L. Wang (Eds): Fuzzy Systems in Bioinformatics and Computational Biology, Springer-Verlag Berlin Hendelberg, 2009, pp.217-233.

[6] Z. Wang and V. Palade, Fuzzy gene mining: a Fuzzy-based framework for cancer microarray data analysis, in: Y. Q. Zhang and J. C. Rajapake (Eds.): Machine learning in bioinformatics, John Wiley & Sons, Inc, Hoboken, New Jersey, 2009, pp.111-133.

[7] I. Batmaz and G. Koksal, Overview of knowledge discovery in databases process and data mining for surveillance technologies and EWS, in: Bioinformatics: concepts, methodologies, tools and applications, Volume 1, Information Resources Management Association, USA, 2013, pp.42-71.

[8] Y. Xu, R.J. Mural, J.R. Einstein, M.B. Shah and E.C. Uberbacher, GRAIL: A Multi-Agent Neural Network System for gene Identification, Proceeding of the IEEE, Vol. 84, N° 10, pp.1544-1551, 1996.

[9] J.D. Tistall, *Introduction à perl pour la bioinformatique*, traduction de Laurent Mouchard et Guénola Ricard, Editions O'REILLY, Paris, France, 2002.

[10] N. Garniere, Développement d'une suite logicielle pour l'analyse et l'annotation intégrative automatiques de transcrits et de protéines. Application aux banques d'ADNc de l'annélide polychète *Alvinella pompejana*, Thèse de doctorat, Université de Strasbourg, Spécialité : Bioinformatique, 2009.

[11] G. Gibson et S. V. Muse, *Précis de génomique*, 1re édition, Editions De Boeck Université, Bruxelles, Belgique, 2004.

[12] G. Coutouly, E. Klein, E. barbieri et M. Kriat, *Travaux dirigés de biochimie, biologie moléculaire et bioinformatique*, Biosciences et techniques, collection dirigée par J. Figarella et A. Calas, 3ème édition, Edition DOIN, France, 2006.

[13] D. Tagu et J.L. Risler, *Bio-informatique : Principes d'utilisation des outils*, Collection Savoir faire, Editions Quae, France, 2010.

[14] A.J.F. Griffiths, W.M. Gelbart, J.H. Miller et R.C. Lewontin, *Analyse génétique moderne*, Editions De Boeck Université, Bruxelles, Belgique, 2001.

[15] A. Arneodo, Y. D'Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Mury and C. Thermes, Wavelet based fractal analysis of DNA sequences, Elsevier, Physica D, Vol. 96, N° 1, pp.291-320, 1996.

[16] S. Nicolay, Analyse de séquences ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles, Thèse de doctorat, Université de Liège, 2006.

[17] D. Anastassiou, Genomic signal processing, IEEE signal processing magazine, Vol. 18, N° 4, pp.8-20, 2001.

[18] X.Y. Zhang, F. Chen, Y.T. Zhang, S.C. Agner, M. Akay, Z.H. Lu, M.M.Y. Waye, and S. K.W. Tsui, Signal Processing Techniques in Genomic Engineering, Proceeding of the IEEE, Vol. 90, N° 12, pp.1822-1833, 2002.

[19] E.R. Dougherty, I. Shmulevich and M.L. Bittner, Genomic Signal Processing: The Salient Issues, EURASIP Journal of Applied Signal Processing, Hindawi Publishing Corporation, N°1, pp. 146-153, 2004.

[20] F. Truchetet, *Ondelettes pour le signal numérique*, Editions Hermes, Paris, France, 1998.

[21] T.K. Sarkar and C. Su, A Tutorial on Wavelets from an Electrical Engineering Perspective, Part 2: The continuous Case, IEEE Antennas and Propagation Magazine, Vol. 40, N° 6, 1998.

[22] I. Daubechies, The Wavelet Transform, Time-Frequency Localization and Signal Analysis, IEEE Transaction on Information Theory, Vol. 36, N° 36, 1990.

[23] T.K. Sarkar, C. Su, R. Adve, M. Salazar-Palma, I. Garcia-Castillo and R.R. Boix, A Tutorial on Wavelets from an Electrical Engineering Perspective, Part 1: Discrete Wavelet Techniques, IEEE Antennas and Propagation Magazine, Vol. 40, N° 5, pp. 49-70, 1998.

[24] S.G. Mallat, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 11, N° 7, pp. 674-693, 1989.

[25] P.P. Vaidyanathan, Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial, Proceedings of the IEEE, Vol. 78, N° 1, pp. 56-93, 1990.

[26] C. Burg and S. Karlin, Finding gene in genomic DNA, Current Opinion in structural Biology, Vol. 8, N° 3, pp.346-354, 1998.

[27] G.B. Singh, *Fundamentals of Bioinformatics and Computational Biology: Methods and exercises in MATLAB*, Springer, USA, 2015.

[28] M.J. Bishop, *Guide to Human Genome Computing*, ACADEMIC PRESS (An imprint of Elsevier Science), Second Edition, UK, 2003.

[29] M. Burset and R. Guigo, Evaluation of Gene Structure Prediction Programs, GENOMICS, by Academic Press, Inc, USA, Vol. 34, N° 0298, pp. 353-367, 1996.

[30] S.L. Salzberg, Gene discovery in DNA sequences, IEEE Intelligent Systems and their Applications, Vol. 14, N° 6, pp. 44-48, 1999.

[31] E. Al-Dauod, Identifying DNA splice sites using pattern statistical properties and fuzzy neural networks, EXCLI Journal, Vol. 8, pp. 195-202, 2009.

[32] F. Moghimi, M.T.M. Shalmani, A.K. Sedigh and M. Kia, Two new methods for DNA splice site prediction based on neuro fuzzy network and clustering, Neural Computing & Applications, Vol. 23, N° 1, pp. 407-414, 2013.

[33] P. Borne, M. Benrejeb et J. Haggège, *Les réseaux de neurones : Présentation et Applications, Méthodes et pratiques de l'ingénieur*, Volume 15 : Automatique, Collection dirigée par Pierre Borne, Editions TECHNIP, Paris, France, 2007.

[34] A.E. Hassanien, M.G. Milanova, T.G. Smolinski and A. Abraham , Computational Intelligence in Solving Bioinformatics Problems: Reviews, Perspectives and Challenges, Studies in computational intelligence, Computational intelligence in Bioinformatics, Springer, Berlin, Heidelberg, Vol. 151, pp. 3-47, 2008.

[35] A. Nikolova, V. Mladenov and G. Tsenov, Performance Comparison of Techniques for DNA Sequence Prediction using Neural Networks, 4th International Symposium on Communications, Control and Signal Processing (ISCCSP), 3-5 March 2010, Limassol, Cyprud, Publisher: IEEE, pp. 1-5, 2010.

[36] S.B. Arniker, H. K. Kwan, N.F. Law and D.P.K. Lun, DNA numerical representation and neural network based human promoter prediction system, India Conference (INDICON), 16-18 December 2011, Hyderabad, India, 2011 Annual IEEE, pp. 1-4, 2011.

[37] I. Tabus and J. Astola, Gene feature selection, in: E. R. Dougherty, I. Shmulevich, J. Chen and Z. Jane Wang (Eds): Genomic Signal Processing and Statistics, EURASIP Book Series on Signal Processing and Communications, Hindawi Publishing Corporation, New York, USA, 2005, pp. 67-92.

[38] L. Jourdan, C. Dhaenens, and E.G. Talbi, Evolutionary feature selection for Bioinformatics, in: G.B. Fogel, D.W. Corne, and Y. Pan (Eds.): Computational Intelligence in Bioinformatics, IEEE Press Series on Computational Intelligence, Published by John Wiley & Sons, Inc., 2008, Hoboken, New Jersey, 2008, pp. 117-140.

[39] J. Malone, Soft computing in Bioinformatics: Genomic and Proteomic applications, in: B. Prasad (Eds.): Soft Computing in Industry, STUD FUZZ 226, Springer-Verlag Berlin Heidelberg, 2008, pp. 135-150.

[40] Z.R. Yang, *Machine Learning Approaches to Bioinformatics*, Science, Engineering and Biology Informatics, Vol. 4, World Scientific Publishing, Singapore, 2010.

[41] G. Schaefer, T. Nakashima, and Y. Yokota, Fuzzy classification for gene expression data analysis, in: A. Kelemen, A. Abraham and Y. Chen (Eds.): Computational

Intelligence in Bioinformatics, Studies in Computational Intelligence 94, Springer-Verlag Berlin Heidelberg, 2008, pp. 209-218.

[42] P. Lio, Wavelets in Bioinformatics and Computational Biology: State of Art and Perspectives, Bioinformatics, Vol. 19, N° 1, pp. 2-9, 2003.

[43] A. Bucur, J. van Leeuwen, N. Dimitrova and C. Mittal, Frequency Sorting Method for Spectral Analysis of DNA Sequences, International Conference on Bioinformatics and Biomedicine, 2008 BIBM '08. IEEE, Philadelphia, USA, pp. 43-50, 2008.

[44] M. Roy and S. Barman Mandal, Spectral analysis of coding and non-coding regions of a DNA sequence by parametric method, India Conference (INDICON), Kolkata, India, 2010 Annual IEEE, pp. 1-4, 2010.

[45] K. Deergha Rao and M.N.S. Swamy, Analysis of Genomics and Proteomics Using DSP Techniques, IEEE Transactions on Circuits and systems I, Vol. 55, N° 1, pp. 370 – 378, 2008

[46] S. Mitra and T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*, John Wiley & Sons, Inc., New Jersey, USA, 2003.

[47] J. Zhao, X. W. Yang, J. P. Li and Y.Y. Tang, DNA Sequences Classification Based on Wavelet Packet Analysis, in: Y.Y. Tang, V. Wicherhauser, P.C. Yuen and C.H. Li (Eds): Wavelet Analysis and its applications, Second International Conference, WAA 2001, Hong Kong, China, December 2001, Springer Proceeding 2001, pp. 424-429.

[48] A. Krishnan and K.B. Li, Protein string algorithms: Protein sequence analysis using wavelet transforms, in: S. Bandyopadhay, U. Maulik, and J. T. Wang (Eds.): Analysis of Biological Data, a Soft Computing Approach, Science, Engineering and Biology Informatics, Vol. 3, World Scientific Publishing, Singapore, 2007, pp. 109-132.

[49] I. Messaoudi, A.E. Oueslati and Z. Lachiri, Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA, EURASIP Journal on Bioinformatics and Systems Biology, Vol. 2014, N° 1, pp.1-13, 2014.

[50] Z. Heidari, D.R. Roe, R. Galindo-Murillo, J.B. Ghasemi and T.E. Cheatham III, Using Wavelet Analysis to Assist in Identification of Significant Events in Molecular Dynamics Simulations, Journal of chemical information and modeling, Vol. 56, N° 7, pp. 1282-1291, 2016.

[51] H.H. Huang and S.B. Girimurugan, Discrete Wavelet Packet Transform Based Discriminant Analysis for Whole Genome Sequences, Statistical applications in genetics and molecular biology, Vol. 18, N° 2, pp.15-34, 2019.

[52] B.J. Yoon and P.P. Vaidyanthan, Identification of CPG islands using a bank of IIR lowpass filters, 2004 IEEE 11th Digital signal processing Workshop, Taos Ski Valley, USA, pp. 315-319, 2004.

[53] C. Burg and S. Karlin, Prediction of Complete Gene Structures in Human Genomic DNA, Journal of Molecular Biology, Vol. 268, pp. 78-94, 1997.

[54] A. Mihi, N. Boucenna and K. Benmahammed, Prediction of DNA sequences using adaptative neuro-fuzzy inference system, International Journal of Biomathematics, World Scientific Publishing Company, Singapore, Vol. 11, N° 4, pp. 1850047- (1-38), 2018.

Résumé : La prédiction précise et la détection des régions d'ADN ou de leurs structures sous-jacentes sont des difficultés persistantes pour les chercheurs. L'extraction de caractéristiques et la classification fonctionnelle de séquences génomiques constituent un domaine de recherche intéressant. De nombreuses techniques de calcul ont déjà été appliquées, y compris le réseau neuronal artificiel, le modèle non linéaire, le spectrogramme et les techniques statistiques.
Dans cette thèse, certaines caractéristiques sont extraites des coefficients d'ondelettes et un second ensemble de caractéristiques est extrait de la fréquence de transition des nucléotides. Ces deux ensembles de caractéristiques sont examinés. Le but était d'étudier les capacités de ces paramètres pour prédire les segments critiques dans la séquence d'ADN. Le système neuro-flou a été utilisé pour la prédiction. Les performances du système neuro-flou ont été évaluées en termes de performance d'entraînement et de précision de prédiction. Deux séquences génomiques des organismes: procaryotes et eucaryotes ont été utilisées, à titre d'exemple, des séquences d'*Escherichia coli* et de *Caenorhabditis elegans* ont été sélectionnées.

Mots-clés: Séquence d'ADN; Système d'inférence neuro-floue adaptative (ANFIS); Logique floue; Transformée en ondelettes; Signal génomique.

Abstract: Accurate prediction and detection of the DNA regions or their underlying structural patterns are constant difficulties for researchers. Feature extraction and functional classification of genomic sequences is an interesting area of research. Many computational techniques have already been applied including the artificial neural network, nonlinear model, spectrogram and statistical techniques.
In this thesis, some features are extracted from the wavelet coefficients and second set of features are extracted from the frequency of transition of nucleotides. These two features sets are examined. The purpose was to investigate the abilities of these parameters to predict critical segments in the DNA sequence. The neuro-fuzzy system was used for prediction. The performance of the neuro-fuzzy system was evaluated in terms of training performance and prediction accuracies. Two genomic sequences of the organisms: prokaryotic and eukaryotic were used, as an example *Escherichia coli* and *Caenorhabditis elegans* sequences were selected.

Keywords: DNA sequence; Adaptive neuro-fuzzy inference system (ANFIS); Fuzzy logic; Wavelet transform; Genomic signal.

ملخص: يعتبر التنبؤ الدقيق والكشف عن مناطق الحمض النووي أو الأنماط الهيكلية الكامنة لها صعوبات مستمرة للباحثين. استخراج ميزة وتصنيف وظيفي لتسلسلات الجينوم هو مجال مثير للاهتمام للبحث. وقد تم بالفعل تطبيق العديد من التقنيات الحاسوبية بما في ذلك الشبكة العصبية الاصطناعية ، والنموذج اللاخطي ، والطيفي ، والتقنيات الإحصائية.
في هذه الأطروحة، يتم استخراج بعض الميزات من معامل المويجات ويتم استخراج المجموعة الثانية من الميزات من وتيرة انتقال النيوكليوتيدات. يتم فحص هاتين المجموعتين من الميزات. كان الغرض من ذلك هو دراسة قدرات هذه المعلمات للتنبؤ بالقطع الحرجة في تسلسل DNA. تم استخدام النظام العصبي غامض للتنبؤ. تم تقييم أداء النظام الغامض العصبي من حيث أداء التدريب ودقة التنبؤ. تم استخدام اثنين من تسلسلات الجينوم من الكائنات الحية : بدائية النواة وحقيقية النواة ، على سبيل المثال تم اختيار تسلسلات *Escherichia coli* و *Caenorhabditis elegans*.

كلمات البحث: تسلسل الحمض النووي. نظام استدلال عصبي غامض متكيف (ANFIS) ؛ المنطق الغامض ؛ تحول موجي؛ إشارة الجينوم.