REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNE MENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE Dr. TAHAR MOULAY – SAIDA

FACULTE DE TECHNOLOGIE

DEPARTEMENT INFORMATIQUE

N° D'ORDRE : .....

THÈSE

Présentée par

KABLI Fatima

Pour l'obtention du diplôme de

DOCTORAT LMD en Informatique

Filière : Informatique

Option : Web et Ingénierie des Connaissances

THEME

Apprentissage Artificiel, Analyse et Fouille de données Complexes

Dirigée par : M.HAMOU Reda Mohamed

Défendu publiquement, en 19/06/2018 :

Devant le jury composé de :

AMINE Abdelmalek	Professeur	Université de Saida	Président
BELALEM Ghalem	Professeur	Université d'Oran 1	Rapporteur
BARIGOU Fatiha	M.C.A	Université d'Oran 1	Rapporteur
TOUMOUH Adil	M.C.A	Université de Sidi Bel Abbès	Rapporteur
HAMOU Reda Mohamed	M.C.A	Université de Saida	Directeur de thèse

Année Universitaire 2017-2018

Thèse préparée au Laboratoire de Gestion des Connaissances et des Données Complexes (GeCoDe)

Université de Saida

Remerciements

Arrivée au terme de la rédaction de cette thèse, je tiens tout d'abord à remercier le grand Dieu le Tout-puissant et Miséricordieux qui m'a donné la force et la patience pour accomplir ce travail.

Je tiens à remercier très sincèrement mon directeur de thèse, monsieur Hamou Reda Mohamed qui a accepté de diriger ce travail. J'ai particulièrement apprécié la qualité de son encadrement, ses conseils, son attention, sa disponibilité et son humanité dans le travail.

Mes vifs remerciements vont également aux membres du jury et examinateurs pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail, professeur Ghalem BELALEM à l'université d'Oran 1. Madame Fatiha Barigou, maître de conférences à l'université d'Oran, Monsieur Toumouh Adil, maître de conférences à l'université Djillali Liabès de Sidi Bel Abbès.

Je tiens également à exprimer ma sincère gratitude au professeur Amine Abdelmalek, pour les conseils, encouragements et suggestions.

Je remercie mes très chers parents, frère et sœurs qui ont toujours été là pour moi.

J'adresse mes plus sincères remerciements à mes professeurs de l'université Tahar Moulay de Saida et l'université d'Oran 1. A tous mes proches et amis, qui m'ont toujours encouragée au cours de la réalisation de ce mémoire.

Kabli Fatima.

Résumé

Apprentissage Artificielle Analyse et Fouille de données Complexes

par KABLI Fatima

Résumé

Le monde numérique connait une quantité énorme d'informations produites quotidiennement, de différents domaines, types et catégories, qui souvent complexes et difficiles à manipuler. Les utilisateurs doivent pouvoir identifier, accéder, évaluer et utiliser efficacement ces différents types des données pour satisfaire leurs besoins. Cependant, les données biologiques ont connu une croissance rapide de jour en jour, collectées in vitro par les biologistes, ce qui a été conduit à la naissance de domaine de bio-informatique pour automatiser le traitement de ce type de données complexes. Le processus ECD, les méthodes d'analyse et de fouille de données ont été largement appliquées pour résoudre de véritables problèmes de bio-informatique, et plusieurs d'autres sont encore ouverts. Dans le but de savoir manipuler les données biologiques moléculaires et de savoir choisir les méthodes appropriées et de proposer de nouvelles approches. Nous avons présentés dans cette thèse les différents champs et problèmes de la bio-informatique et l'intérêt des méthodes d'analyse et de fouille de données biologiques pour résoudre ces problèmes. Notre travail est consacré à la proposition de nouvelles méthodologies basant sur les techniques d'analyse, le processus d'ECD biologique et les méthodes de fouille de données, prendre en compte la complexité des données biologiques moléculaires sous un format primaire (ADN/protéine), dans le but d'améliorer la qualité de réponse aux trois problèmes majeurs de la bio-informatique: L'étude de similarité entre les séquences d'ADN et les séquences de protéine, la classification supervisée des protéines pour la prédiction de leurs fonctions inconnues, la classification non-supervisée des séquences d'ADN pour la prédiction de structure de molécule et le développement de médicaments. L'évaluation de nos approches sur le jeu de données biologiques montre leurs intérêts de traiter la complexité de séquençage des molécules biologiques par les techniques d'analyse et de fouille de données.

Mots clés: Analyse et fouille de données, processus ECD, classification supervisée, classification non-supervisée, règles d'association,Données biologiques moléculaires, ADN, Protéine

Abstract

The digital world knows a tremendous amount of information produced daily, different domains, types and categories, which often are complex and difficult to manipulate. The users need to be able to identify, access, evaluate, and effectively use these different types of data and meet their needs. In order to know how to manipulate the molecular biological data and to know how to choose the appropriate methods and to propose new approaches. In this thesis, we have presented the different fields and problems of bioinformatics and the interest of the methods of analysis and biological data mining to solve these problems. Our work is devoted to propose a new methodologies based on the analysis techniques, biological KDD process and data mining methods, we treated the complexity of molecular biological data in primary format (DNA / protein), with the aim of improving the quality of response to the three major problems of bioinformatics: The study of similarity between DNA sequences and protein, the classification of proteins for the prediction of their unknown functions, the unsupervised classification of DNA sequences for the prediction of molecule structure and the production of

medicaments. the Evaluation of our approaches to the biological dataset shows their interest in treating the complexity of sequencing biological molecules by the data mining and analysis techniques.
Key-words: Analysis and data mining, ECD process, , classification, clustering, association rules, Molecular biological data, DNA, Protein.

ملخص

إن العالم الرقمي يعرف كمية هائلة من المعلومات المنتجة يوميا، من مختلف المجالات وبأنواع وفئات متعددة، والتي غالبا ما تكون معقدة وصعبة المعالجة. يحتاج المستخدمون إلى أن يكونوا قادرين على تحديد هذه الأنواع المختلفة من البيانات والوصول إليها وتقييمها وإستخدامها بفعالية لتلبية إحتياجاتهم، بلمقابل عرفت البيانات البيولوجية نموا سريعا، والتي تجمع في المختبرات من قبل علماء الأحياء، أدى هذا إلى ولادة مجال المعلوماتية الحيوية لمعالجة هذا النوع من البيانات المعقدة. قد تم تطبيق عملية استخراج المعرفة من البيانات وطرق التحليل على نطاق واسع لحل المشاكل المختلفة للمعلوماتية الحيوية والعديد من الطرق الأخرى لا تزال مفتوحة. من أجل معرفة كيفية التعامل مع البيانات البيولوجية الجزيئية ومعرفة كيفية إختيار الطريقة المناسبة وإقتراح مناهج جديدة. لقد قدمنافي هذه الأطروحة مختلف مجالات ومشاكل المعلوماتية الحيوية وطرق التحليل لإستخراج البيانات البيولوجية لحل هذه المشاكل. كرسنا عملنا لإقتراح منهجيات جديدة تقوم على التقنيات التحليلية والخطوات البيولوجية العملية و طرق إستخراج البيانات، أخذنا بعين الإعتبار تعقيد البيانات البيولوجية الجزيئية للبنية الأساسية (الحمض الريبوزي النووي البروتين)، وذلك بهدف تحسين نوعية الإستجابة للمشاكل الرئيسية الثلاث للمعلوماتية الحيوية: دراسة التشابه بين سلاسل الحمض النووي وبين سلاسل البروتين، التصنيف الخاضع للرقابة للبروتينات بهدف التنبؤ بالوظائف الغير معروفة ،التصنيف غير الخاضع للرقابة لتسلسل الحمض النووي للتنبؤ ببنية جزيء (الحمض الريبوزي النووي)وتطوير الأدوية. ويظهر تقييم مناهجنا على مجموعة البيانات البيولوجية أهمية التعامل مع تعقيد تسلسل الجزيئات البيولوجية من خلال إستخراج البيانات وتقنيات التحليل.
الكلمات المفتاحية : تحليل وإستخراج البيانات ، إستخراج المعرفة من البيانات ،التصنيف ، التجمع ، قواعد الترابط ، البيانات البيولوجية الجزيئية ، الحمض الريبوزي النووي ، البروتين.

Table des Matières

Liste des Tableaux

Liste des Figures

Liste des Algorithmes

Dédie humblement ce manuscrit à mes très chers parents, frères et sœurs. . . .

Liste des abréviations

ECD	**E**xtraction de **C**onnaissance à partir de **D**onnées
FD	**F**ouille de **D**onnées
ADN	**A**cide **D**ésoxyribo**N**ucléique
ARN	**A**cide **R**ibo**N**ucléique
IDC	**I**nternational **D**ata **C**orporation
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
IEB	**I**nstitut **E**uropéen de **B**io-informatique
SNP	**S**ingle **N**ucleotide **B**olymorphisms
DDBJ	**D**NA **D**ata **B**ank of **J**apan
MEME	**M**ultiple **EM** for **M**otif **E**licitation
OGH	**O**rganisation du **G**énome **H**umain
MAP	**M**utation **A**cceptée par **P**oint.
BLOSUM	**BLO**cks **SUM**atrix .
SIB	**S**uisse **I**nstitut of **B**ioinformatics.
HGP	**H**uman **G**enome **P**roject.
RF	**R**andom **F**orest.
wwPDB	**W**orld**W**ide **P**rotein **D**ata **B**ank.
UniProt	**U**niversal **P**rotein.
RSL	**R**eprésentation de la **S**équence par **L**ettres.
PIR	**P**rotein **I**nformation **R**esource Sequence Format.
MMC	**M**odèle de **M**arkov **C**aché.
MVS	**M**achines à **V**ecteurs de **S**upport.
IPP	**I**nteractions de **P**rotéines-**P**rotéines.
AC	**A**nalyse de **C**orrélation.
ANVA	**AN**alyse de la **VA**riance.
AG	**A**lgorithmes de la **G**énétiques.
AE	**A**lgorithmes **E**volutionnaire.
ACP	**A**nalyse des **C**omposants **P**rincipales.
SOM	**S**lef **O**rganizing **M**ap.
ACA	**A**lgorithme de **C**luster d'**A**ttributs.
UPGMA	**U**nweighted **P**air **G**roup **M**ethod with **M**ean **A**rithmetic.
CSDAB	**C**lustering **S**patial **B**asé sur la **D**ensité des **A**pplications avec **B**ruit.
RF	**R**andom **F**orest.
IPA	**I**nterface de **P**rogrammation **A**pplicative.
RNA	**R**éseaux de **N**eurones **A**rtificiel.
KNN	**K-N**earest **N**eighbor.
NorDi	Algorithme de **Nor**malisation et la **D**iscrétisation.
FTB	**F**ouille **T**ext **B**iologique.
UICPA	**U**nion **I**nternationale de **C**himie **P**ure et **A**ppliquée.
AC	**A**utomate **C**ellulaire.
P	**P**récision.

R	**R**appel.
F	**F**-mesure.
E	**E**ntropie.

Introduction Générale

Aujourd'hui, les données s'envolent partout autour de nous, la croissance de la quantité de données recueillies et générées a explosé dans les derniers temps grâce à l'automatisation généralisée des diverses activités quotidiennes, des avancées dans les recherches scientifiques et d'ingénierie de haut niveau. La quasi-totalité des données existants sont des données complexes, sont différents l'une de l'autre en terme de structure, format, version, modèle, etc. ces données sont de plusieurs types et catégories, ils ont représenté en différents modes, tableaux, textes, graphiques, images, sont récupérés et bien stockés dans des banques de données publiques et privés. Les données scientifiques produisent dans des domaines divers tels que l'astronomie, l'imagerie médicale, la télédétection, les tests non destructifs, la physique, la science des matériaux et la bio-informatique. Plus particulièrement, les données biologiques qui sont venues de divers champs de recherche tels que la génomique, la protéique, la métabolomique, la phylogénétique et les puces à ADN. Ce type de données a donné naissance aux bases de données biologiques, permettant de trouver des informations à propos de la structure, la fonction, la localisation chromosomique, la similarité de séquence et de structure (génomique et protéique). Les séquences ADN/Protéine/ARN sont considérées comme l'information de base de toute donnée biologique, la source de ces données est le résultat obtenu après les expérimentations in vitro effectuées par les biologistes sur les chromosomes circulaires plus particulièrement sur les génomes des différents organismes et des bactéries. Un génome est exprimé par des milliers de nucléotides (bases) organisé sous la forme d'une séquence (chaine). ARN et Protéine sont conclus à partir de l'ADN suivant un processus de transformation et translation (dogme central de biologie moléculaire), une chaine d'ARN est caractérisée par les même composants de l'ADN, la différence est dans une seule base, cependant la protéine est composée d'un ensemble de petites molécules appelées les acides aminés. La complexité de ces molécules biologiques est cachée dans les composants de base (nucléotides et acide aminé), dans leurs positions, dans la longueur des séquences, dans le changement entre les composants, dans la distribution des composants sur toute la longueur de la séquence, etc. Donc, En raison de la complexité des données et systèmes biologiques, la majorité des ensembles de données biologiques sont assez bruyants et très compliqués. Par conséquent, le développement de méthodes efficaces de traitement est essentiel au succès de l'analyse des données biologiques. L'analyse de ce type de données est une tâche difficile non seulement en raison de sa complexité, mais aussi en raison de l'évolution continue de notre compréhension des mécanismes biologiques. Donc, pour tirer un maximum de connaissances possibles et manipuler les données biologiques de façon générale, plusieurs techniques et outils d'analyse, d'informatique et de mathématique ont été développés. Par conséquent, la science de la bio-informatique est apparue comme domaine multidisciplinaire qui tente de résoudre des problèmes médicaux et biologiques avec des méthodes informatiques. Plusieurs recherches dans ce domaine sont nécessaires car beaucoup de tâches et nombreux problèmes sont encore ouverts. En général, les problèmes liés à l'analyse des données en bio-informatique peuvent être

divisés en trois classes en fonction du type de données biologiques, bio-informatique des séquences, des structures et des réseaux. Le processus d'extraction de connaissance à partir de données (ECD) peut répondre à ces problèmes et aux nouveaux tendances. Le processus d'ECD est un domaine combine des algorithmes, des techniques, de l'apprentissage automatique, de la gestion des connaissances, de l'intelligence artificielle, des techniques mathématiques et statistiques et des bases de données, dans le but de transformer les données en connaissances, c'est-à-dire extraire de nouveaux concepts, connaissances ou relations conceptuelles cachées dans les volumes de données brutes, pour la prise de décision. Le processus d'ECD biologique devrait prendre en compte à la fois les caractéristiques des données biologiques et les exigences générales du processus de découverte de connaissances.
La fouille de données est le cœur du processus ECD, Généralement, les techniques de fouille de données traitent trois problèmes majeurs qui peuvent être observés assez souvent, tel que, la classification supervisée, la classification non-supervisée (clustering) et les règles d'association. Au cours des dernières années, les technologies de fouille de données ont été appliquées à la recherche en bio-informatique avec beaucoup de succès. Il y a beaucoup de problèmes de bio-informatique qui peuvent être considérés comme des problèmes standards de fouille de données afin que les méthodes existantes puissent être appliquées. Plus particulièrement, la fouille de données est probablement l'outil de calcul le plus populaire en biologie moléculaire. Il est considéré comme un véritable défi scientifique pour la compréhension des nouveaux problèmes de génomique et de protéique, dans le but d'aider le biologiste à comprendre les phénomènes biologiques.

Problématique et Objectifs

Les molécules biologiques (ADN/Protéine) peuvent avoir une structure primaire, secondaire, tertiaire et quaternaire, chaque structure a connu une large recherche dans le domaine de biologie et de bio-informatique, dans le but de découvrir des phénomènes biologiques importants. La structure primaire est la structure de base, à partir de laquelle toute autre structure peut être construite et ainsi pour prédire une fonction d'une protéine ou d'un gène particulier peut toujours avoir recours à la structure primaire de base.
Les séquences d'ADN sont représentées à base de nucléotides, sachant que les séquences protéiques sont représentées à base d'acides aminés, donc la complexité de la structure de base de données biologique est cachée dans la complexité de ses petits molécules (structuration, position, longueur, etc.). Ces données sont représentées sous plusieurs formats en chaine de caractères. D'autre part et comme nous l'avons mentionné précédemment les méthodes, les techniques et les algorithmes d'analyse et de fouille de données ont connu une importance considérable dans le domaine de la bio-informatique. Donc dans ce contexte et dans le but de traiter ce type de données biologiques pour extraire la connaissance et l'information importante pour le biologiste, notre problématique est la suivante : « **Comment améliorer le traitement et la manipulation des données biologiques (ADN Protéine)sous un format primaire prend en compte leurs complexités de séquancage, on appliquant les méthodes d'analyse et de fouille de données pour répondre aux besoins biologiques** ».
dans ce contexte, nous avons envisagé deux champs importants :
Le premier est l'analyse des propriétés de petites molécules (nucléotides et acide aminé)

pour proposer de nouvelle représentation aidant dans le développement de nouvelles méthodes pour l'étude de similarité entre les séquences d'ADN et de protéine.
Dans le deuxième champ, nous avons proposé deux méthodologies inspirées du processus ECD biologique avec ses différentes étapes, dans lequel l'étape de modélisation est le cœur de l'ECD.
Avant de passer à la modélisation, nous avons effectué un pré-traitement spécifique et intégré la technique de N-gramme pour l'extraction des descripteurs qui sont pondérés dans la suite. Ce qui nous permet de passer d'un ensemble de données bruits non-structuré à un ensemble de donnée nettoyé, structuré et bien adapté aux méthodes de fouille de données.
Dans l'étape de modélisation :

- Nous avons fait appel à l'algorithme bio-inspiré « automate cellulaire 3D » pour la classification non-supervisée des séquences ADN dans le but de prédire la structure des molécules pour le développement des médicaments.
- Nous avons proposé un nouveau classificateur de protéine, basant sur la construction d'une base de règles, qui a montré les associations existant entre les différents descripteurs (chaine d'acides aminés) dans les protéines, puis faire la classification selon la base des règles pertinentes.

Les méthodologies que nous avons abordées ont traité la complexité des données biologiques sous format primaire, à l'aide des techniques de calcul d'analyse et de fouille de données.
Organisation de thèse :
Le corps de notre thèse s'articule autour de quatre chapitres :
Le premier chapitre intitulé **« Donnée complexes »** comportera deux parties essentielles, une introduction aux données complexes, nous décrirons leurs différents types, catégories et natures. Une deuxième partie s'intéresse aux données biologiques complexes auxquelles nous seront amené à travailler, nous commencerons par une brève vision aux données scientifiques puis nous passerons à représenter les données biologiques de base, nous donnerons les différents types de données (ADN, protéine, ARN), avec les différentes bases de données populaires pour le stockage de ce type de données et les différents formats de représentation. A la fin nous présenterons où se cache la complexité de ces moléculaires biologiques.
Le deuxième chapitre intitulé **« Bio-informatique et Fouille de données biologiques :** nous commencerons par la description des différents champs de la bio-informatique avec les méthodes et algorithmes de base pour traiter et analyser les données biologiques. Par la suite nous nous concentrerons sur les étapes de processus ECD biologique plus particulièrement les techniques d'analyse et de fouille de données, nous présenterons les trois problèmes majeurs (classification non-supervisée, la classification supervisée et l'extraction des règles d'association) des données biologiques qui sont consacré à résoudre les problèmes de bio-informatique comme, l'étude de similarité la classification des gènes et protéine, la prédiction de fonctions et structures etc.
Le troisième chapitre intitulé **« Modélisation des Données Biologiques Complexes » :** Ce chapitre sera consacré à la description détaillée des contributions que nous avons amenées pour réaliser cette thèse et atteindre nos objectifs, nos approches sont regroupées en deux parties principales, la première s'intéresse à la présentation

de notre proposition dans le champ d'étude de similarité des séquences d'ADN et de protéine. Dans la deuxième partie nous présenterons les étapes de processus ECD que nous avons abordé pour résoudre deux problèmes de la bio-informatique, la classification supervisée des protéines dans le but de prédire leurs fonctions et la classification non-supervisée des séquences d'ADN pour la prédiction de structure de molécules aidant au développement des médicaments.
Le quatrième chapitre intitulé **« Expérimentations et résultats » :** suivant la même organisation du troisième chapitre. Nous présenterons les données expérimentales, résultats, comparaisons et discussions par l'application des méthodes proposées sur le jeu des données (d'ADN et protéines), nous nous basons sur des métriques d'évaluations. Le but de ce chapitre est de prouver l'efficacité de nos travaux dans le traitement des données biologiques moléculaires.

Chapitre I

Les Données Complexes

I.1 Introduction

Des avancées rapides dans la collecte de données et la technologie de stockage ont permis l'accumulation d'une grande quantité de données. Les données disponibles sur internet et dans les banques de données, sont souvent complexes ; leur complexité est discutée au niveau de la multiplicité des sources d'information, des formats, des modèles et des versions. L'ensemble de données scientifiques, se produisent dans des domaines divers, l'un des données scientifiques le plus complexe et ayant une certaine importance dans la recherche moderne sont les données biologiques, plus particulièrement les données biologiques moléculaires, telles que les séquences d'ADN, les séquences des protéines et les structures protéiques. Par conséquent, ce chapitre est un état de l'art de différents types et catégories des données complexes, nous discuterons les différents types et catégories des données complexes spécialement les données biologiques moléculaires, on va décrire les deux molécules de base ADN et protéine, qui nous intéressent dans notre travail, les différents formats de représentations et les banques de données biologiques. On terminera avec le but de notre choix de ce type de données pour l'application des méthodes d'analyse et de fouille de données et on va détailler où se trouve la complexité de données biologiques moléculaires.

I.2 Données complexes

Les données du monde réel ne présentent pas tout de la même façon, les données révélées par les sites web se présentent de différents modes : textes, tableaux, images, graphiques. Ainsi, dans d'autres domaines comme le multimédia, l'imagerie médicale, la télédétection, la bio-informatique, les systèmes d'information géographique, etc. Par exemple, si on parle d'un dossier médical, comportent certes des données numériques, qui pouvaient être structuré de façon tabulaire comme les résultats d'analyses biochimiques, mais incorpore également des données textuelles comme les compte rendus d'observations cliniques, des graphiques tels que les tracés d'électro-cardiogramme ou d'électro-encéphalogramme, des images complexes fournies par les radiographies, échographies et scanners.
Dans cette section, nous présenterons quelques natures de données complexes, types et catégories :

I.2.1 Natures de données complexes

I.2.1.1 Données multi-structures

Les données se présentent sous trois différentes façons en termes de structuration, qui sont présentés dans cette section qui sont les données structurées, semi-structurées et non-structurées :

- **Les données structurées:** Les données structurées sont très banales. Elles concernent toutes les données qui peuvent être stockées dans la base de données SQL dans le tableau avec des lignes et des colonnes. Elles ont une clé relationnelle et peuvent être facilement mappées dans des champs préconçus. Aujourd'hui, ces données sont les plus développées et le moyen le plus simple de gérer les informations.
  Mais selon IDC[1]; les données structurées ne représentent que 5 à 10% de tout l'ensemble de données.

- **Les données semi-structurées:**Les données semi-structurées sont des données qui n'ont pas été organisées en un référentiel spécialisé, comme une base de données, mais qui a néanmoins des informations associées, telles que des métadonnées, qui la rendent plus accessible au traitement que les données brutes. Dans le monde numérique, le nombre de documents semi-structurés générés augmente constamment.

- **Les données non-structurées :** Les données non structurées sont des données brutes et non organisées, se réfèrent à des données qui ne correspondent pas parfaitement à la structure traditionnelle de lignes et de colonnes des bases de données relationnelles. Idéalement, toutes informations non-structurées peuvent être facilement transformées en modèles structurés, mais cela prend du temps. Par exemple, un courrier électronique contient un ensemble d'informations tels que le temps d'envoi, le sujet et le contenu du message. Selon une étude d'IDC, les données non structurées représentent plus de 95% de tous les contenus numériques et devraient croître de façon exponentielle.

I.2.1.2 Données multi-sources

Les données multi-sources sont complexes, hétérogènes, dynamiques, distribuées et très grandes. Sont des données issues des simulations informatiques, extrêmement complexes de plusieurs façons, de la géométrie des systèmes de grille à la variété des types de données et de la variété des phénomènes en cours de modélisation. Un aspect particulièrement important de cette complexité est l'hétérogénéité des données, des formats de fichiers aux représentations numériques, des unités et des densités d'échantillonnage, des règles de variété. Ces données sont largement distribuées dans de nombreux systèmes informatiques et de nombreux emplacements géographiques. Une grande partie de ces données sont des observations dynamiques, que ce soit à partir d'instruments satellitaires de détection à distance ou de mesures météorologiques fondées sur le terrain. Et aussi constituent une collection vraiment massive [2].

[1]IDC: International Data Corporation, est disponible dans: http://www.idc.com, consulté le: 2017-08-03

[2]Techopedia est disponible dans: https://www.techopedia.com, consulté le: 2017-08-03.

I.2.1.3 Données temps réel

Les données en temps réel représentent une capture de l'état actuel des données du monde réel. Ce sont des données datées possédant une durée de validité, qui représente la période pendant laquelle elles peuvent être utilisées, permet de gérer l'historique de leur évolution. Ces données peuvent fournir des informations pratiques sur les applications pratiques aux appareils mobiles tels que les téléphones, les ordinateurs portables et les tablettes. Plusieurs systèmes de temps réel ont été proposés pour gérer les données de temps réel, Il s'agit par exemple des applications appliquées pour contrôler des centrales nucléaires, des usines chimiques et des applications de commerce électronique, etc. leurs exploitations par différents applications sont prises en charge par des SGBD temps réel [DMS99].

I.2.1.4 Données multi-modèles

Un modèle de données fait référence aux interrelations logiques et au flux de données entre différents éléments de données impliquées dans le monde de l'information. Il décrit la façon dont les données sont stockées et récupérées, facilite les activités de communication. Les modèles de données permettent de représenter les données requises et le format à utiliser pour les différents processus métier. Il existe trois styles de base de modèle de données qui sont les suivants : modèles de données conceptuelles, modèles de données physiques, modèles de données logiques. [3] Avec deux significations différents, modèle de données théoriques concerne la description formelle et un modèle d'instance qui s'intéresse à l'application d'un modèle théorique.

I.2.2 Catégories de données complexes

Le schéma suivant montre les différentes catégories des données :

[3] Techopedia est disponible dans: https://www.techopedia.com, consulté le: 2017-08-07.

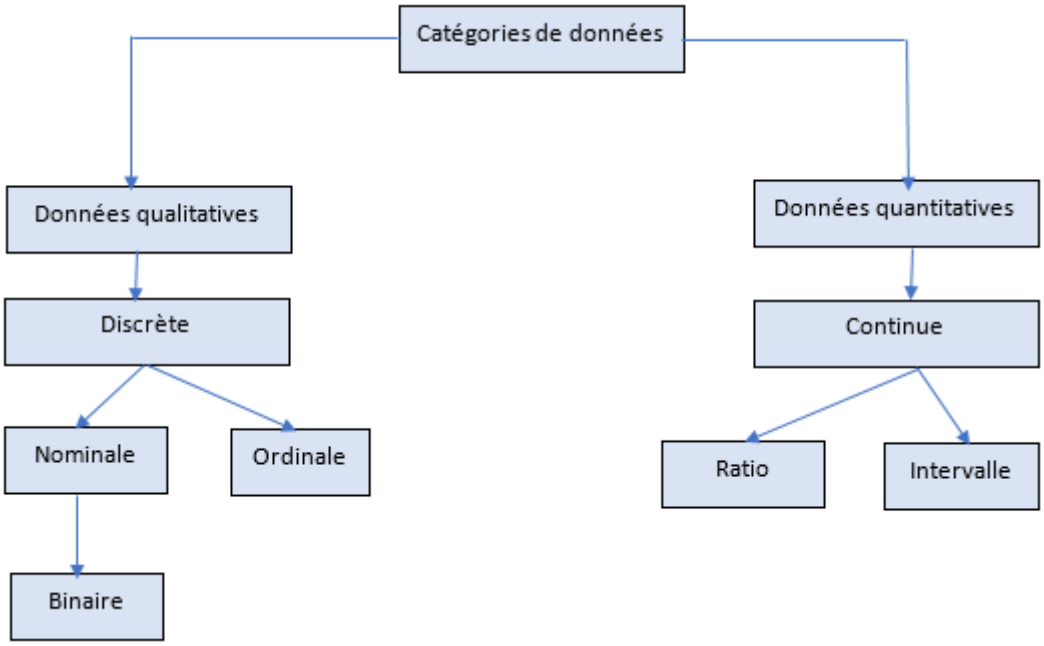

FIGURE I.1: Les catégories des données

I.2.2.1 Données quantitatives

Les données quantitatives sont généralement considérées comme celles qui peuvent être codées numériquement, sont une mesure numérique exprimée non pas au moyen d'une description de langage naturel, mais plutôt en termes de nombres. Cependant, tous les nombres ne sont pas continus et mesurables. Elles sont utiles quand on cherche à décrire le qui, le quoi, le où et le quand Par exemple, le numéro de sécurité sociale est un nombre, mais pas quelque chose que l'on peut ajouter ou soustraire. Selon cette définition, les données qualitatives peuvent être qualifiées de "tout le reste".

- **Les Données discrètes:** Les données discrètes ne peuvent prendre que des valeurs particulières. Il peut y avoir un nombre infini de valeurs. Parfois appelées données thématiques, catégorielles ou discontinues, les données discrètes peuvent être numériques, se divisent en trois catégories: binaire, nominale et ordinale :

- **Les Données Nominales:** Se réfèrent essentiellement à des données catégoriquement discrètes telles que les valeurs des attributs (nom d'une personne, le type de voiture, nom d'un livre). Qui sont facilement à retenir.

- **Les Données Ordinale:** Se réfèrent à des quantités qui ont un ordre naturel, par exemple; le classement des sports préférés, l'ordre de la place des gens dans une ligne, l'ordre des coureurs finissant une course ou plus souvent le choix sur une échelle de notation.

- **Les Données binaires:** Les données binaires sont l'un de type de données représentées ou affichées dans le système à numération binaire. Les données binaires sont la seule catégorie de données pouvant être directement comprises et exécutées par un ordinateur. Il est représenté numériquement par une combinaison de zéro et un.

I.2.2.2 Données qualitatives

Les données qui se rapprochent ou se caractérisent d'une chose ou d'un phénomène, mais ne mesurent pas les attributs, les caractéristiques, les propriétés, etc, Elles sont utiles quand on cherche à expliquer le comment et le pourquoi, se divisent en deux sous catégories :

- **Les Données continues:** Les données continues ne sont pas limitées à des valeurs distinctes, mais peuvent occuper n'importe quelle valeur sur une plage continue. Entre deux valeurs de données continues, il peut y avoir d'autres nombre infini. Les données continues sont toujours essentiellement numériques.
- **Les Données à intervalles:** Les données à intervalles sont comme les données ordinales, sauf qu'on peut dire que les intervalles entre chaque valeur sont également divisés. L'exemple classique d'une échelle d'intervalle est la température Celsius car la différence entre chaque valeur est la même. Par exemple, la différence entre 60 et 50 degrés est un 10 degrés mesurable, tout comme la différence entre 80 et 70 degrés. Le temps est un autre bon exemple d'une échelle d'intervalle dans laquelle les incréments sont connus, cohérents et mesurables. Bien que de nombreux points sur l'échelle soient probablement égaux.
- **Les Données Ratio:** Les données ratio sont des données d'intervalle avec un point zéro naturel. Par exemple, une température est une donnée ratio, tel que 0.0 ne signifie pas « sans chaleur». Les données ratio comme, la taille, le poids, l'activité enzymatique, etc.

I.2.3 Certains types de données complexes

Le schéma ci-dessous représente certain type de données complexes disponibles dans les sites internet et dans des banques de données privées et publics :

FIGURE I.2: Certaine types de données complexes

- **Les images satellitaires:**
  Les images satellitaires sont des images collectées par les satellites d'imagerie de la terre ou d'autres planètes, exploités par les gouvernements et les entreprises à travers le monde. Les sociétés d'imagerie satellitaire vendent des images autorisant à des gouvernements et à des entreprises telles que : Apple Maps et Google Maps.

- **Données scientifiques:**
  Les données scientifiques sont définies comme des informations collectées à l'aide de méthodes spécifiques dans un but spécifique d'étude ou d'analyse. Données recueillies dans une expérience de laboratoire réalisée dans des conditions contrôlées, et reconnues comme la source principale pour la recherche scientifique et sont généralement reconnus par la communauté scientifique.

- **Texte interne de la société :** Texte dans les documents, les journaux, les résultats des enquêtes et les courriels.

- **Photographies et vidéos:** vidéos de sécurité, de surveillance et de trafic.

- **Données radar ou sonar:** profils sismiques véhiculaires, météorologiques et océanographiques.

- **Données des réseaux sociaux:** L'information que les utilisateurs de réseaux sociaux publient, il existe de nombreux types de données sociales, y compris des tweets sur Twitter, des publications sur Facebook, des broches sur Pinterest, etc.

- **Données mobiles:** Messages texte, informations géospatiales.

- **Contenu du site:** à partir de n'importe quel site offrant des contenus non structurés, tels que YouTube, Flickr, Instagram, etc.

I.3 Données biologiques

Comme nous l'avons décrit dans la section précédente, il existe plusieurs types de données complexes, chaque type est différent de l'autre en terme de qualité, nature, version, source et modèle. L'un des types de données complexes est la donnée scientifique, ce type de données se produit dans des domaines divers tels que l'astronomie, l'imagerie médicale, la télédétection, les tests non destructifs, la physique, la science des matériaux et la bio-informatique. Ils peuvent être obtenus à partir de simulations, d'analyse, d'expériences ou d'observations. Les données biologiques sont l'un des données importantes de données scientifique, dans cette section nous décrivons ce type de données, leurs formats de représentation, les banques de données dont lesquelles ces données sont sauvegardées.

I.3.1 Données biologiques moléculaires

Cette section décrit une brève introduction à quelques concepts de base de la biologie moléculaire qui sont pertinentes pour les problèmes du domaine de la bio-informatique. Chaque organisme vivant se compose d'un certain nombre d'organes, Chaque organe se compose d'un certain nombre de tissus, et chaque tissu est une collection de cellules similaires qui se regroupent pour effectuer des fonctions spécialisées. La cellule individuelle est l'unité auto-reproductrice minimale dans toutes les espèces vivantes, Il exécute deux types de fonctions :

- Stockage et transmission de l'information génétique pour maintenir la vie d'une génération à une autre, cette information est stockée sous la forme d'ADN bicaténaire.

- Effectuer les réactions chimiques nécessaires pour maintenir notre vie, à travers des protéines qui sont produites par la transcription des portions d'ADN en molécules étroitement liées appelées ARN, Les ARN guident la synthèse des molécules protéiques, les protéines résultantes sont les principaux catalyseurs pour toutes les réactions chimiques dans la cellule.

Les Trois types de molécules de base l'acide désoxyribonucléique (ADN), l'acide ribonucléique (ARN) et les protéines sont présents dans une cellule : Ci-dessous, nous discuterons de ces trois molécules principales.

I.3.1.1 L'Acide Désoxyribonucléique (ADN)

L'acide désoxyribonucléique (ADN) est le matériel génétique de tous les organismes (avec l'exception de certains virus), il stocke les instructions nécessaires à la cellule pour effectuer les fonctions vitales.
La structure correcte de l'ADN a été déduite par J.D.Watson et F.H.C.Crick En 1953 [WC+53]. Ils ont déduit que l'ADN se compose de deux brins antiparallèles qui sont enroulés l'un autour de l'autre pour former un double hélix. Chaque brin est une chaîne de petites molécules appelées nucléotides. Les types de nucléotides dépendent de type

de bases azotées, qui sont l'adénine (A), la guanine (G), la cytosine (C), la thymine (T).
D'après l'analyse de E.Charga et ses collègues [Sun09], ils sont déduits que la concentration des Thymine et d'adénine est toujours égaux et la concentration de cytosine est toujours égale à la concentration de guanine. Cette observation suggère fortement que A et T ainsi que C et G ont une certaine relation fixe, la figure suivante montre le double hélix d'ADN avec les quatre bases azotiques :

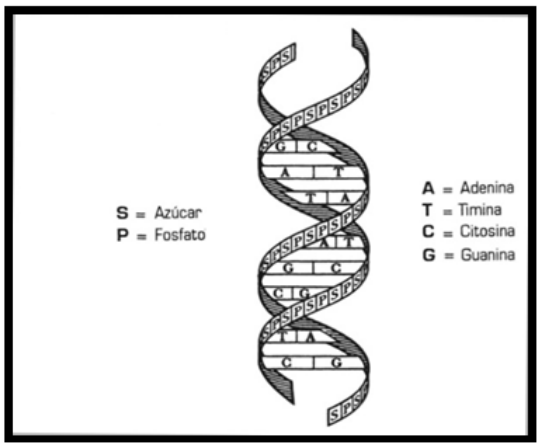

FIGURE I.3: Le double hélix d'ADN

I.3.1.2 L'Acide ribonucléique (ARN)

L'acide ribonucléique (ARN) est l'acide nucléique produit pendant la Transcription (c'est-à-dire obtenir la séquence d'ADN à partir de la séquence d'ARN). Exceptionnellement, l'ARN est utilisé comme matériel génétique au lieu de l'ADN, dans certains organismes, Tels que les virus. L'ARN utilise la base U au lieu de la base T de l'ADN. La base U est chimiquement similaire à la base T. En particulier, U est également complémentaire à A.

I.3.1.3 Protéine

La protéine est une très grande molécule biologique se compose d'une chaîne de molécules plus petites appelées acides aminés. Il existe 20 types d'acides aminés et chacun a des propriétés chimiques et physiques différentes. Les 20 acides aminés obtenus après la traduction de trois nucléotides, est spécifiée par une table de traduction appelée "code génétique". Le code génétique est universel pour tous les organismes. Le tableau ci-dessous montre le code génétique :

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

TABLE I.1: Code génétique

La longueur d'une protéine est dans la plage de 20 à plus de 5000 acides aminés. En moyenne, une protéine contient environ de 350 acides aminés, chacun entre aux est codé par trois bases azotiques (nucléotides) et exprimé par une abréviation (lettre), le tableau ci-dessous représente les 20 acides aminés (nom, abréviation, code) :

Acide glutamique	Glu	E	**Leucine**	Leu	L
Acide aspartique	Asp	D	**Lysine**	Lys	K
Alanine	Ala	A	**Méthionine**	Met	M
Arginine	Arg	R	**Phénylalanine**	Phe	F
Asparagine	Asn	N	**Proline**	Pro	P
Cystéine	Cys	C	**Sérine**	Ser	S
Glutamine	Gln	Q	**Thréonine**	Thr	T
Glycine	Gly	G	**Tryptophane**	Trp	W
Histidine	His	H	**Tyrosine**	Tyr	Y
Isoleucine	Ile	I	**Valine**	Val	V

TABLE I.2: Les abréviations des acides aminés

I.3.1.4 Structures de protéine

Par convention, quatre niveaux d'organisation des protéines peuvent être identifiés : Ce sont les structures primaires, secondaires, tertiaires et quaternaires de la protéine.

- **structure primaire (linéaire):** En biochimie, nous appelons la séquence linéaire d'acides aminés la "structure primaire" d'une protéine, est une chaîne longue avec des liaisons peptidiques. Sachant que chaque acide aminé est identifié en utilisant son abréviation spécifique (voir la table I.2). Cette structure est représentée dans la figure suivante :

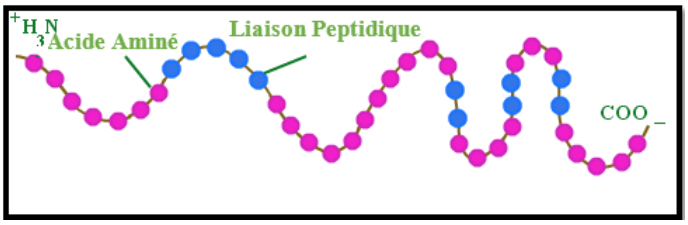

FIGURE I.4: Structure primaire de protéine

- **Structure secondaire:** La structure secondaire de la protéine traite de la conformation de la chaîne peptidique présente dans la molécule de protéine, qui forme un polypeptide après le pliage de la structure primaire, se caractérise par des liaisons hydrogène, il existe deux manières possibles dans lesquelles la chaîne peptidique est organisée, à savoir *alpha* et β. La figure suivante représente la structure secondaire de protéine :

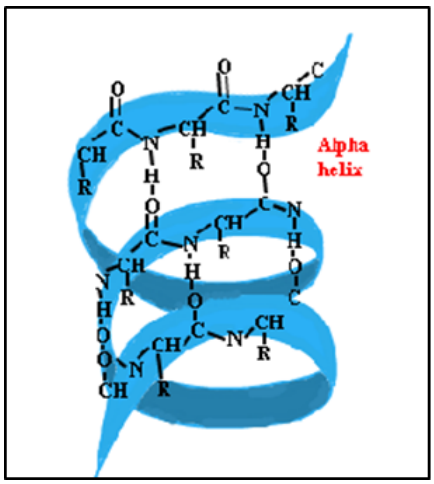

FIGURE I.5: Structure secondaire de protéine

- **Structure tertiaire:** La structure tertiaire est une image tridimensionnelle de protéines qui dépend de la structure secondaire, se caractérise par quatre liaisons différentes, la figure ci-dessous montre la structure tertiaire de protéine :

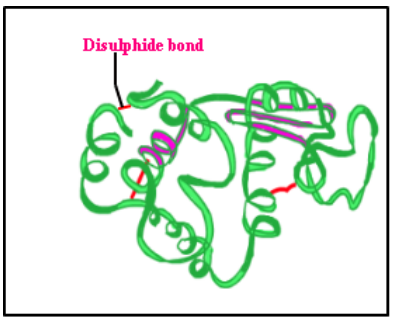

FIGURE I.6: Structure tertiaire de protéine

- **Structure quaternaire :** Par définition, cette structure est l'agencement de plus d'une molécule de protéine dans multi-sous-unités complexes. Les sous-unités de protéines ont interagi entre aux pour construire la structure quaternaire, la figure ci-dessous montre cette structure:

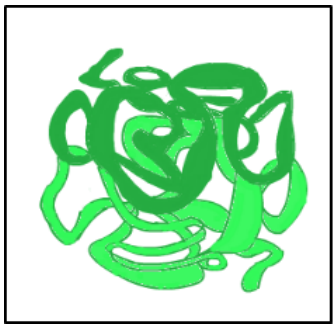

FIGURE I.7: Structure quaternaire de protéine

I.3.2 Dogme central de la biologie moléculaire

Le dogme central a d'abord été énoncé par Francis Crick en 1958 [Cri70] et repris dans un article publié en 1970 [WC58]. Décrit le processus de transfert de l'information de l'ADN à l'ARN puis de l'ARN à une protéine. Il indique que l'information provenant de l'ADN est transférée à l'ARN puis en protéine. En d'autres termes, une fois l'information obtenue dans la protéine, elle ne peut pas retourner à l'acide nucléique. La figure I.8 montre le processus de transfert d'informations génétique qui comprend les deux étapes, transcription et translation :

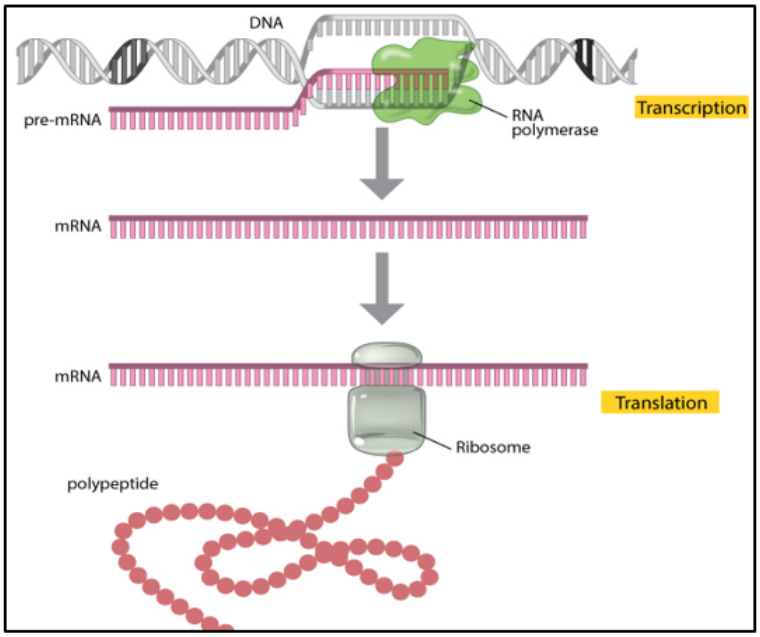

FIGURE I.8: Processus de transcription et traduction

1. **Transcription:** Le processus de transcription de gêne d'ADN à un ARNm à l'aide de l'ARN polymérase est simple, il est comme suit : Tout d'abord; l'ARN

polymérase sépare temporairement deux brins de l'ADN ensuite, il localise le site de départ de la transcription, qui est un marqueur indiquant le début d'un gène. Ensuite, l'ARN polymérase synthétise un ARNm suivant deux règles.

- Les bases A, C et G sont copiées exactement de l'ADN à l'ARNm.
- T est remplacé par U dans l'ARN, nous n'avons que U au lieu de T.

Une fois que l'ARN polymérase atteint le site d'arrêt de transcription (l'extrémité d'un gène), le processus de transcription est arrêté et un ARNm est obtenu.

2. **Translation:** La translation est aussi appelée synthèse des protéines, la synthèse d'une protéine est effectuée à partir d'un ARNm. Le processus de traduction est géré par un moléculaire complexe connu sous le nom de ribosome, à l'aide du ribosome la traduction démarre autour du codon de départ (site de début de traduction), chaque codon contient trois base nucléotides (exemple : ACC, CCT etc.) traduit en un acide aminé (selon la table de code génétique présenté dans la figure 2), une fois le ribosome lit le codon d'arrêt (site d'arrêt de traduction), la traduction s'arrête.

I.3.3 Parties exons et introns dans les gènes

Les segments du gène eucaryote qui sont finalement transcrits en un ARNm sont appelés exons, les segments d'un gène situé entre les exons appelés les introns. Chaque gène eucaryote peut avoir de nombreux introns et chaque intron peut être très long. Par exemple le gène associé à la maladie de la fièvre cystique chez l'homme a 24 introns de longueur totale environ 1 million de bases, alors que la longueur totale de ses exons est seulement 1 kilo de base [Sun09]. Les Introns dans les gènes eucaryotes satisfaire à la règle GT-AG, c'est-à-dire, l'intron se commence avec GT et se termine avec AG.

I.3.4 Bases de données biologiques

Une base de données biologiques est un grand corps de données persistantes organisé ,répertoriant des informations sur les sciences de la vie, collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques.Elles contiennent des informations venant de divers champs de recherche tels que la génomique, la protéomique, la phylogénétique, la métabolomique, et les puces à ADN. Parmi le contenu des bases de données, on trouve des informations à propos de la fonction, de la structure, de la localisation des gènes, les effets cliniques de leurs mutations, ainsi que leurs similarités de séquence et de structure. Les BDD biologiques sont souvent décrites comme des données semi-structurées, et peuvent se présenter sous plusieurs forme (tableaux, xml, ..).Cependant, leur conception, leur développement et maintenance à long terme est un secteur clé de la bioinformatique. Elles permettent aussi aux scientifiques de comprendre et expliquer de nombreux phénomènes et connaissances biologiques ces connaissances facilitent la prise en charge des pathologies, permet la création de nouveaux médicaments et permet la découverte de relations inter-espèces, ect. Historiquement, le processus de collecte des séquences biologiques a commencé avec des séquences d'acides aminés dans les protéines [DSO78]. Puis par la coordination et la normalisation des soumissions par

les institutions engagées dans le maintien de base de données, nombreux travaux de bio-informatique ont concerné sur la technologie des bases de données, ces bases de données incluent des référentiels "publics" de données génétiques et des bases de données "privées" comme celles utilisées par les groupes de recherche ou ceux détenus par des sociétés de biotechnologie.
Il existe des liens denses entre les différentes bases de données biologiques, puisqu'elles contiennent souvent des informations relatives à plusieurs aspects. De plus, il existe un nombre considérable des ressources internet liées à la recherche et à la navigation dans les bases de données et certains algorithmes de traitement et d'inférence de données.
Au cours des dernières années, le nombre de séquences biologiques et des données expérimentales a connu une croissance rapide, ces séquences sont soumises à des banques de données biologiques. Cependant, dans cette partie, nous présenterons les catégories des bases de données biologiques et on fait référence aux types de bases de données contenant des séquences moléculaires biologiques de base.

I.3.4.1 Catégories des bases de données biologiques:

La figure ci-dessous représente les bases de données biologiques classés en quatre catégories différentes:

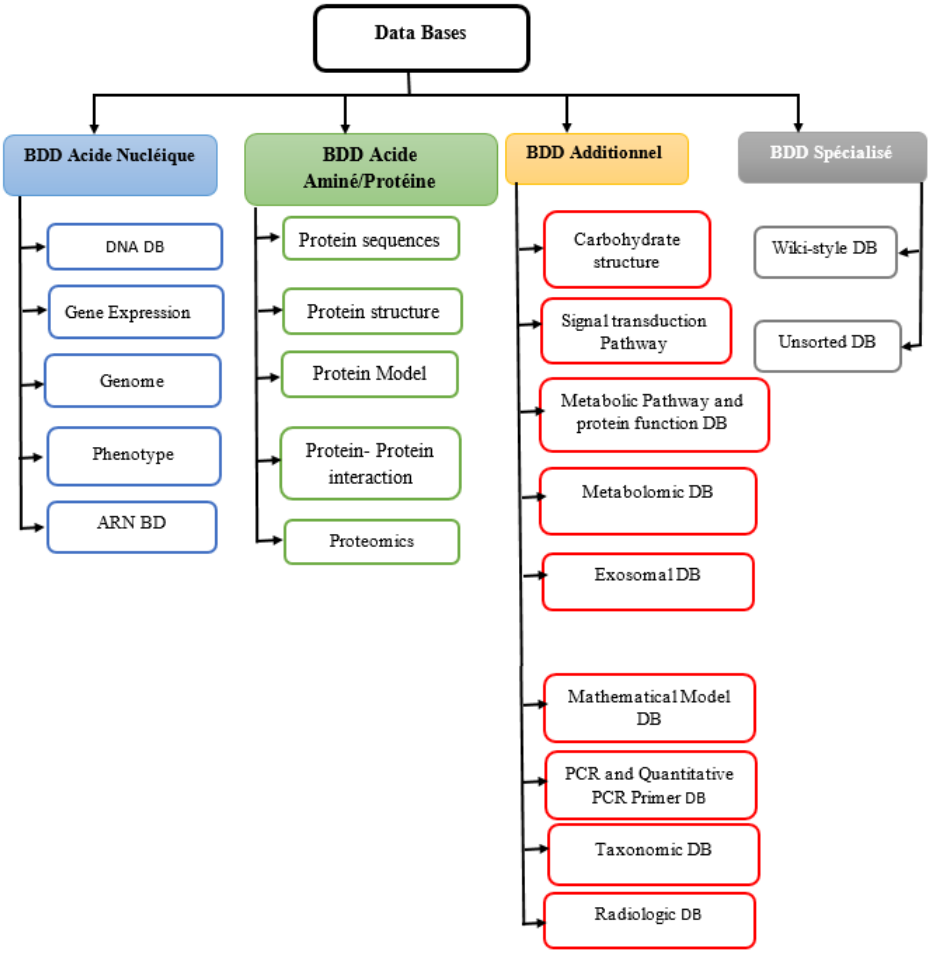

FIGURE I.9: Les catégories des bases de données biologiques

I.3.4.2 Bases de données génomiques (ADN)

GenBank est la base de données génomique la plus connue, maintenues par NCBI[4], Il contient les séquences acides nucléiques et acides aminé de nombreux espèces, son contenu est reflété par deux sources de données, EMBL-EBI [5] et DDBJ [6], qui comportent de nombreuses fonctions liées à la recherche et la navigation des séquences. Ils effectuent des services concernant la soumission de séquences à GenBank, et contiennent également des liens vers divers sites internet de bio-informatiques. De nombreuses bases de données contiennent des informations plus spécialisées sur les séquences génomiques. Comme la base SNP[7], est une archive publique gratuite de variations génétiques au sein et entre différentes espèces développées et hébergées par le Centre national d'information sur la biotechnologie (NCBI) en collaboration avec l'Institut national de recherche sur le génome humain (NHGRI), la base de données du gène promoteur et des séquences régulatrices [8], Les bases de données des motifs hautement conservés de l'ADN MEME [9] ,la base de données pour la normalisation de la nomenclature des gènes (HUGO) [10] et plusieurs d'autres.

I.3.4.3 Bases de données protéiques:

En raison des correspondances entre les séquences d'acides aminés et les séquences codons, les deux types de séquences sont disponibles sur GenBank.
Les bases de données biologiques sont représentées par des séquences d'acides aminés, les aspects fonctionnels, les familles et les domaines de protéines, les structures secondaires et 3D de protéines. La base de données PDB [11] la plus connue, contient des données annotées sur les structures spatiales des protéines et des macromolécules biologiques, il comprend également des données sur leurs séquences, fonctions et maladies connexes.
La banque de donnée de séquences protéiques UniProt [12] est produit par le Consortium UniProt, une collaboration entre l'Institut Européen de Bioinformatique (EBI), l'Institut Suisse de Bioinformatique (SIB) et Ressource d'information sur les protéines (PIR), qui proposent en particulier la recherche de séquences homologues dans la base au moyen d'outils d'alignement de séquences comme FASTA ou BLAST. UniProt est une base de données de séquences protéiques. Son nom dérive de la contraction de "Universal Protein Resource"(base de données universelle de protéines). C'est une base de données ouverte, stable et accessible en ligne, elle est issue de la consolidation de l'ensemble des données produites par la communauté scientifique. UniProt est une base annotée, hiérarchisée où chaque séquence est accompagnée d'un ensemble riche de métadonnées et de liens vers de nombreuses autres bases de données : bibliographiques, phylogénétiques, nucléotidiques. UniProt fournit des informations sur les fonctions de séquences, leurs structures ainsi que des liens vers d'autres bases de données. Il existe également de nombreuses bases de données spécialisées dans des aspects particuliers

[4]NCBI: est disponible dans https://www.ncbi.nlm.nih.gov/, consulté le: 2017-12-21
[5]EMBL-EBI: est disponible dans http://www.ebi.ac.uk/, consulté le 2017-08-08
[6]DDBJ: est disponible dans http://www.ddbj.nig.ac.jp/, consulté le 2017-08-08
[7]SNP: est disponible dans http://snp.cshl.org/, consulté le 2017-08-08.
[8]cisred: est disponible dans http://www.cisred.org/, consulté le 2017-08-08.
[9]MEME: est disponible dans http://meme-suite.org/, consulté le 2017-08-20.
[10]HUGO: est disponible dans https://www.genenames.org/, consulté le 2017-08-03.
[11]PDB: est disponible dans http://www.rcsb.org/pdb/, consulté le 2017-09-01.
[12]UniProt: est disponible dans http://www.uniprot.org/, consulté le 2017-04-03.

des protéines, des fonctions protéiques, les résultats des expériences sur les protéines, telles que les bases de données de protéine de 2D gels [13], des enzymes de restriction [14] et des structures secondaires [15].

I.3.4.4 Bases de données d'ARN

Des informations sur les séquences de ribonucléotides (ARN), les fonctions des molécules d'ARN et leurs structures spatiales, sont disponible dans les bases de données GtRNA[16], PDB[17],Rfam [18], la base de données Rfam , stocke des familles d'ARN non codantes, et aussi contient plusieurs alignements de séquences et modèles de covariance. La base de données GtRNA stocke les séquences génomiques ribonucléotidiques d'ARNt et les structures secondaires. Des données sur les séquences d'acide ribonucléique dans l'ARN peuvent également être trouvées dans GenBank.

I.3.5 Formats de représentation de séquences biologiques

Les séquences (protéine / ADN / ARN) sont représentées dans nombreux formats dans les bases de données, tels que ACE, GDE, PIR, AB1, EMBL, GenBank, CAF, FASTA, FASTAQPHD, Swiss-Prot, Nexus, GFF, texte brut, etc.
[Dav01] montre que la difficulté majeure rencontrée dans l'exécution d'un logiciel d'analyse de séquence est l'utilisation de formats de séquence différents par des programmes différents.
Ces formats sont tous des fichiers ASCII standard, mais ils peuvent différer en présence de certains caractères et mots qui indiquent les informations concernant la séquence et la séquence elle-même. Les formats de séquence les plus couramment utilisés sont discutés ci-dessous :

I.3.5.1 Format de séquence FASTA

FASTA est le plus pratique et le plus répondu, ce format comprend trois parties essentielles :

- Une ligne de commentaire identifiée par un caractère « > » Dans la première colonne suivie du nom et de l'origine de la séquence.
- Présence de la séquence en symboles.
- Un "*" facultatif qui indique la fin de la séquence.
  La présence de "*" peut être essentielle pour lire la séquence correctement par certains programmes d'analyse de séquence.

[13] Abdn:2D Protein data base, Haemophilus, University of Aberdeen, est disponible dans: http://www.abdn.ac.uk, consulté le 2017-04-03.
[14] Rabase: est disponible dans http://rebase.neb.com/rebase/rebase.html, consulté le 2017-09-12.
[15] HSSP: est disponible dans: http://swift.cmbi.kun.nl/swift/hssp/consulté le 2017-09-15.
[16] GtRNA: est disponible dans http://lowelab.ucsc.edu/GtRNAdb/, consulté le 2017-09-12.
[17] PDB: est disponible dans http://www.rcsb.org/pdb/, consulté le 2017-09-01.
[18] Rfam: est disponible dans http://rfam.xfam.org/, consulté le 2017-09-24.

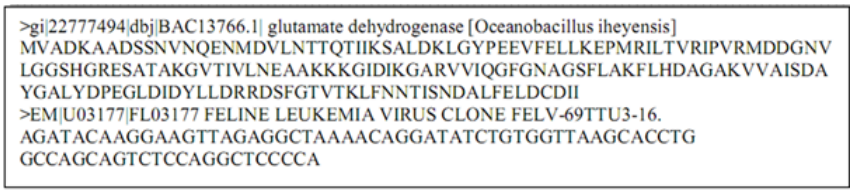

FIGURE I.10: Exemple du format FASTA d'une séquence protéique [Abd10]

I.3.5.2 Format de séquence PIR

A été utilisé par « National Biomedical Research Foundation/ Protein Information Resource (NBRF) » et aussi par d'autres programmes d'analyse de séquence. Ce format caractérise par trois lignes :

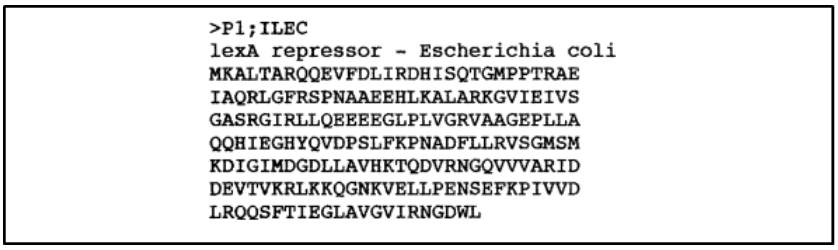

FIGURE I.11: Exemple de format de séquence PIR

La première ligne comprend un caractère initial ">" Suivi d'un code à deux lettres tel que P pour la séquence complète ou F pour le fragment, suivi d'un 1 ou 2 pour indiquer le type de séquence, Puis un point-virgule, puis un nom unique de quatre à six caractères pour l'entrée. Il existe également une deuxième ligne essentielle avec le nom complet de la séquence, suivi par la séquence complète exprimé par des caractères spéciaux selon le types de la séquence (ADN/ ARN/ protéine).

I.3.5.3 Format de séquence GDE

Le format GDE (Genetic Data Environment) est utilisé par un système d'analyse de séquences appelé Genetic Data Environment qui a été conçu par[Smi+94], ce format se compose de plusieurs champs, chacun est encadré par des crochets, et chaque champ comporte des lignes spécifiques, chacune avec une étiquette de nom donnée. Les informations qui suivent chaque balise sont placées entre guillemets ou suivent le nom de la balise par un ou plusieurs espaces.

```
{
name            "Short name for sequence"
longname        "Long (more descriptive) name for sequence"
sequence-ID     "Unique ID number"
creation-date   "mm/dd/yy hh:mm:ss"
direction       [-1|1]
strandedness    [1|2]
type            [DNA|RNA|PROTEIN|TEXT|MASK]
offset          (-999999,999999)
group-ID        (0,999)
creator         "Author's name"
descrip         "Verbose description"
comments        "Lines of comments about a sequence"
sequence        "gctagctagctagctagctcttagctgtagtcgtagctgatgctag
                 ctgatgctagctagctagctagctgatcgatgctagctgatcgtag
                 ctgacggactgatgctagctagctagctagctgtctagtgtcgtag
                 tgcttattgc"
}
```

FIGURE I.12: Exemple de format de séquence GDE [Dav01]

I.3.5.4 Format de séquence EMBL

Ce format permet de décrire une grande quantité d'information, y compris des références bibliographiques, des informations sur la fonction de la séquence, des emplacements d'ARNm et des régions de codage et des positions de mutations importantes. Ces informations sont organisées en champs, chaque séquence est définit par un identifiant. La signification de chacun de ces champs est expliquée par la figure (I.13).

```
ID          identification code for sequence in the database
AC          accession number giving origin of sequence
DT          dates of entry and modification
KW          key cross-reference words for lookup up this entry
OS, OC      source organism
RN, RP, RX, RA, RT, RL  literature reference or source
DR          i.d. in other databases
CC          description of biological function
FH, FT      information about sequence by base position or range of positions
            source range of sequence, source organism
            misc_signal range of sequence, type of function or signal
            mRNA range of sequence, mRNA
            CDS range of sequence, protein coding region
            intron range of sequence, position of intron
            mutation sequence position, change in sequence for mutation
SQ          count of A, C, G, T and other symbols
gaattcgata aatctctggt ttattgtgca gtttatggtt ccaaaatcgc cttttgctgt 60
atatactcac agcataactg tatatacacc caggggcgg aatgaaagcg ttaacggcca 120
.
.
// symbol to indicate end of sequence
```

FIGURE I.13: Un exemple de format de séquence EMBL [Dav01]

I.3.5.5 Format de séquence texte brut

La figure (I.14) décrit la séquence au format texte brut, chaque ligne a un sens bien précis, comme par exemple, un nom, un code, etc.

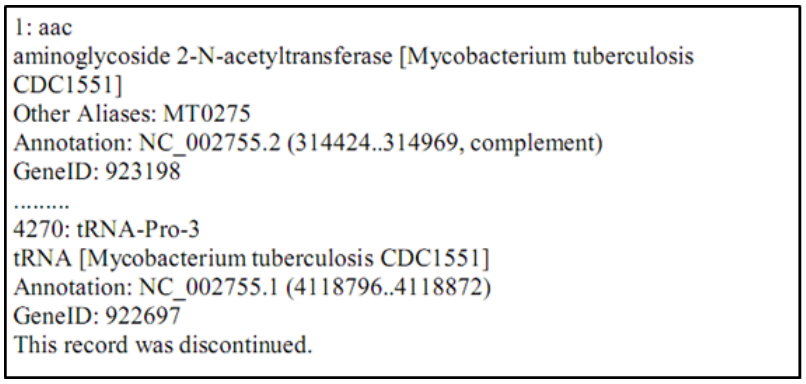

FIGURE I.14: Exemple de format de séquence texte brut [Abd10]

I.3.5.6 Format de séquence SwissProt

Le format d'une entrée dans la base de données de séquences de protéines SwissProt est très similaire au format EMBL, sauf que l'on fournit beaucoup plus d'informations sur les propriétés physiques et biochimiques de la protéine.

I.3.6 Les Types de données biologiques

Cette section décrit également les types de données de base qui peuvent être produits par diverses expériences biologiques par rapport aux trois molécules de base (ADN/protiéne/ARN).

I.3.6.1 Données séquentielles

Les progrès technologiques ont conduit à la collecte de nombreuses quantités de séquences biologiques. En (1954),(prix Nobel en (1972)) est énoncé que, à partir de la structure primaire d'une séquence le repliement d'une protéine dans sa structure fonctionnelle peut se trouver, par conséquent dans sa séquence [Anf73]. Donc la donnée séquentielle est définit comme la structure primaire de la donnée biologique moléculaire, se représente par une chaine de caractères.

I.3.6.2 Données structurales

En bio-informatique, Il existe également toute une gamme de structures, Les données structurales ne sont pas linéaires, permet de déterminer la fonction, permet aussi d'identifier des relations évolutives plus éloignées. Donc la compréhension de la structure aide à la compréhension de fonction.
D'un autre coté, l'exploitation des données structurelles des protéines aider à faciliter la conception des molécules. De plus, l'espace de recherche pour la plupart des problèmes liés aux données structurelles est continus, infinis et nécessitent des algorithmes hautement efficaces et heuristiques.

I.3.6.3 Données d'expressions génétiques

L'expression génétique est le processus par lequel l'information génétique codée dans l'ADN est d'abord convertie en ARN messager et ensuite en un produit fonctionnel (protéine). Par conséquent, les données d'expression génétique sont représentées par une matrice $N*M$. Les N lignes représentent tous les gènes, les M colonnes représentent les échantillons qui peuvent être lié aux (type de tissu, l'âge de l'organisme, conditions environnementales, etc). Les valeurs incluses dans les cellules de la matrice indiquent la variance entre l'expression du gène respectif dans l'échantillon particulier et l'expression du même gène dans un échantillon témoin. Deux outil populaires pour la mesure de l'expression des gènes est le microarray [Aas01]et SAGE [Vel+95].

I.4 Complexité de données biologiques moléculaires

Comme nous venons de le voir tout au long de ce chapitre, les données complexes sont souvent, variées l'une de l'autre en terme de nature, type et qualité. Les données scientifiques, particulièrement les données biologiques moléculaires ont connu une augmentation à des taux explosifs en raison de l'amélioration des technologies existantes et de l'introduction de nouvelles technologies. Le domaine de la bio-informatique a de nombreuses applications dans le monde moderne, y compris les molécules afin d'obtenir de nouvelles connaissances biologiques.
En Biologie computationnelle, les données biologiques sont représentées comme des chaines de caractères d'un ensemble de symboles représentant leurs unités structurales respectives. Dans notre contexte, basées uniquement sur le RSL (Représentation de la séquence par lettres), une séquence d'ADN est représentée sur une chaîne d'alphabet de quatre lettres, et une séquence de protéine est représentée sur une chaîne d'alphabet de vingt lettres. Cependant, il a été reconnu que l'information contenue dans les séquences d'ADN et protéines est extrêmement difficile à comprendre, à reconnaître, à rappeler et à comparer par l'être humain sans des traitements minutieux.
Ces difficultés s'expliquent par une certaine complexité dans l'information génétique (séquence ADN/protéine), donc la question posée est : « Où se trouve la complexité de l'information biologique moléculaire ? » La réponse : Nous résumons la complexité de l'information biologique moléculaire dans les points suivants:

- Une séquence d'ADN/protéine courte peut contenir moins d'informations génétiques, alors qu'une séquence avec plusieurs bases (nucléotidique/acide aminé) peuvent contenir beaucoup plus d'informations génétiques.
- Le changement d'un nucléotide/acide aminé dans la même séquence peut changer la signification de message génétique.
- L'arrangement entre deux séquences peut produire de nombreux résultats différents.
- Il n'existe que peu de motifs d'ADN codants dans les espèces vivants.
- Certaines séquences ne contiennent aucune information génétique appelée (ADN indésirable).
- Un changement dans la position des nucléotides/acides aminés implique un changement correspondant dans le message génétique.

I.5 Conclusion

Dans ce chapitre nous avons discuté les données complexes avec leur nature et catégorie, et nous avons décrit le type des données dont nous avons besoin dans notre contexte, qui sont les données biologiques moléculaires, décrivant en détail les deux types de séquences moléculaires importantes (ADN et Protéine) sous leur différents formats de représentation et les catégories des bases de données biologiques; aussi Nous avons montré où se trouve la complexité de l'information biologique moléculaire.
Dans le chapitre suivant, nous montrerons l'état de l'art de méthodes et techniques d'analyse et de fouille de données pour traiter et extraire l'information appropriée à partir des données biologiques, et aussi nous présenterons les différents problèmes de la biologie liés à l'informatique (bio-informatique), nous nous concentrons sur le traitement de donnée biologique et les problèmes de bio-informatique les plus couramment traités liés aux méthodes d'analyse et de fouille de données.

Chapitre II

Bio-informatique et Fouille de données biologiques

II.1 Introduction

L'ensemble de données biologiques moléculaires complexes disponibles dans les banques de données a connu un accroissement rapide, telles que les séquences d'ADN/ ARN, les séquences de protéines et les structures protéiques. Face à cette énorme quantité de données, le biologiste ne peut pas simplement utiliser les techniques traditionnelles en biologie pour traiter, analyser, gérer et interpréter ces données, donc l'intégration des techniques d'informatique dans le but d'extraire l'information appropriée est nécessaire. Cette science est connue sous le nom bio-informatique, il a été un domaine de recherche actif depuis la fin des années quatre-vingt. Plusieurs recherches dans ce domaine étaient nécessaires et le sont encore, car beaucoup de tâches sont encore en cours. En outre, bien que des progrès considérables ont été réalisés au cours des années, de nombreux problèmes fondamentaux de la bio-informatique sont encore ouverts. L'analyse et la fouille de données ont joué un rôle fondamental dans la compréhension des problèmes émergents en génomique et en protéique.
La fouille de données biologiques est définie comme la découverte d'informations significatives et inconnues cachées dans ce type de données. C'est un domaine émergent depuis le milieu des années quatre-vingt dix. Les techniques de fouille de données traitent trois problèmes majeurs, la classification, le regroupement et l'association. Pour appliquer l'une de techniques on applique le processus d'ECD avec un ensemble de techniques à chaque étape, ces trois types de problèmes peuvent être observés assez souvent. Au cours des dernières années les technologies de fouille de données ont été appliquées à la recherche en bio-informatique avec beaucoup de succès.
Dans le premier chapitre nous avons discuté l'ensemble de données biologiques moléculaires, leurs complexités, formats de représentation et leurs bases de stockage, par conséquent, ce chapitre est un état de l'art des principaux problèmes de la bio-informatique, processus d'ECD et les techniques d'analyse et de fouille de données pour traiter l'information biologique.

II.2 Bio-informatique

II.2.1 Historique

En 1866, Gregor Johann Mendel découvre la génétique, par l'hybridation des Expériences sur les pois a dévoilé certains éléments biologiques appelés gènes, passé de génération en génération. À cette époque, les gens pensaient L'information génétique était véhiculée par certaine "protéine chromosomique" ; Cependant, il n'était pas. Plus tard, en 1869, l'ADN a été découvert. Mais ce n'est qu'en 1944 [Ave+44] que Avery et McCarty a démontré que l'ADN est le principal facteur de l'information génétique.
En 1953, une autre découverte historique a permis de grands progrès en biologie [WC+53] , James Watson et Francis Crick ont déduit la structure tridimensionnelle de l'ADN, qui est une double hélice. En 1956, La première séquence de protéines rapportée était celle de l'insuline bovine constituée de 51 résidus. Plus tard, en 1961, la cartographie de l'ADN au peptide (protéine), nommé sous le nom « code génétique », a été élucidée par Marshall Nirenberg, Par la combinaison de trois nucléotides de la séquence ADN en tant que codon et de cartographier chacun d'entre eux à un acide aminé.
L'utilisation du terme bio-informatique est documentée pour la première fois en 1970 dans une publication de Paulien Hogeweg et Ben Hesper (université d'Utrecht, Pays-Bas), en référence à l'étude des processus d'information dans les systèmes biotiques [Att+11]. À partir des années 70, plusieurs techniques biotechnologiques importantes ont été développées, en 1970 la naissance de l'algorithme Needleman Wunsch [NW70] pour la comparaison de séquences, en 1972 la création de la première molécule d'ADN recombinant par Paul Berg et son groupe, et aussi le lancement de la première banque de données de Protéine par le laboratoire national de Brookhaven en 1973, qu'a été transféré au projet Worlwide Protein Data Bank (wwPDB) [1].
En 1989, la première carte complète du génome de la bactérie Haemophilus in-fluenza a été publiée . L'événement le plus remarquable a été le lancement de Projet du génome humain (HGP) en 1990, Initialement, il était prévu d'être réalisé pendant 15 ans.
La publication de la première ébauche du génome humain en 2000, Un génome humain plus raffiné a également été publié en 2003.
À partir de 2006, la technologie de séquençage de deuxième génération est devenue disponible.
Définition :
La bio-informatique est la science de la gestion, de l'analyse, de l'interprétation et de l'extraction de l'information à partir de séquences et de molécules biologiques. Il s'agit d'un domaine de recherche actif depuis la fin des années quatre-vingt. Combine plusieurs technologies pour traiter les données biologiques, dans le but de découvrir des nouvelles connaissances. [PK07]. La bio-informatique a été occupée par un certain nombre de disciplines connexes. Il s'agit notamment de sciences quantitatives telles que:

- La biologie mathématique et informatique.
- La biométrie et bio-statistique.
- L'informatique.

[1] wwpdb: est disponible dans: http://www.wwpdb.org/, consulté le: 2017-10-24.

- La cybernétique.

Ainsi que les sciences biologiques telles que :

- Évolution moléculaire.
- La génomique et la protéique.
- La génétique.
- La biologie moléculaire et cellulaire.

Par conséquent, les objectifs de la bio-informatique sont :

- Organiser les données de manière à permettre aux chercheurs de créer et d'accéder à l'information biologique.
- Développer des outils facilitant l'analyse et la gestion des données.
- Utiliser les données biologiques pour analyser et interpréter les résultats de manière biologique.

II.2.2 Champs d'application de la bio-informatique

Sans aucun doute, l'objectif de la recherche est de relever les défis scientifiques. Par conséquent, les défis réels en bio-informatique sont de savoir comment résoudre les problèmes scientifiques posés par les biologistes. En raison de la complexité des données biologiques, il existe de nombreuses questions de recherche stimulantes en bio-informatique, mais il est très difficile de fournir une catégorisation complète des problèmes. En général, les problèmes liés à l'analyse des données en bio-informatique peuvent être divisés en trois classes selon le type de données biologiques : (séquences, structures et réseaux).

II.2.2.1 Bio-informatique des séquences

Les séquences d'ADN, d'ARN et de protéines ayant une importance primordiale dans les sciences de la vie, La bio-informatique des séquences, sert à analyser les données issues de l'information génétique contenue dans ces trois types de séquences. De nombreux problèmes de bio-informatique qui se concentrent sur les études de séquences, par l'analyse et la comparaison multiples des séquences, l'identification de séquences à partir de données expérimentales, l'identification des ressemblances entre les séquences, la classification et la régression des séquences, l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur les composants de bases (nucléotides, acides aminés).

1. **Analyse et comparaison des séquences**
   Les séquences de nucléotides et de protéines sont stockées dans des bases de données séquentielles. L'analyse et la comparaison de ces séquences biologiques sont devenues un problème bio-informatique fondamental dans nombreux domaines de la biologie moléculaire moderne. Pour déterminer la fonction d'une nouvelle séquence, une recherche doit être effectuée pour déterminer si une séquence similaire existe déjà dans la base de données. Une séquence de protéine ou ADN peut

avoir des propriétés biologiques similaires à une autre séquence, et cette similitude peut permettre de déduire la fonction de la nouvelle séquence. L'essentiel de problème d'analyse et comparaison de séquence de la bio-informatique est d'examiner comment utiliser efficacement la similarité de séquence pour prédire la fonction d'une séquence nouvellement découverte. Le mot similarité est signifié le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. La recherche dans une base de donnée par l'une des méthodes de similarité permet de quantifié la similarité par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquence. En biologie, deux séquences homologues si elles ont un ancêtre commun, déterminé par un score de similarité. La similarité de deux séquences est généralement déterminée par plusieurs techniques et méthodes d'analyse, d'informatique et de mathématique définit par :

- **Alignement**
  Définition: Un alignement de séquences est un moyen d'arranger des séquences d'ADN, d'ARN ou de protéine pour identifier des régions similaires qui peuvent être une conséquence de relations fonctionnelles, structurales ou évolutives entre les séquences. l'alignement est réalisé par l'insertion des « trous » (symbolisés par des tirets) dans l'une des séquences ou il y a des lettres non identiques (acides aminés ou nucléotides), ce qui est appelé « Mésappariements ». Afin de maximiser le nombre de coïncidences de caractères entre les deux séquences. Certains méthodes représentent l'alignement par une matrice (N, M), N définit les séquences et M les composants des séquences (acides aminés ou nucléotides), les valeurs de la matrice définit par trois scores différents.
  ($match$=+2, $mismatch$=-1, $grap$=-2), signifié (identique, non identique, trou) respectivement. Donc la similarité entre deux séquences est définie par la valeur de score d'alignement, qui est calculé par une fonction de similarité, la plus connue est SUM-OF-PAIR(SP). La distance SP est définie comme:

$$\sum_{1 \leq i < j \leq k} dM(S'_i + S'_j). \tag{II.1}$$

  est égale à la somme du score de distance de chaque paire alignée. On distingue deux types d'alignements de séquences globales et locales, dans l'alignement global toutes les séquences maintiennent une correspondance sur toute leur longueur. D'autre part, dans l'alignement local, seule la partie la plus similaire des séquences est alignée. Il existe deux types d'alignement de séquences, l'alignement par paire et l'alignement multiple.
  L'alignement de séquence par paire, comme suggéré par son nom, est une comparaison de deux séquences biologiques. Les algorithmes les plus connue dans ce type d'alignement est L'algorithme de Needleman-[NW70] et l'algorithme de Smith-Waterman [SW81].
  Alignement Multiple :
  Alignement multiple est utilisé pour identifier des régions de similarité qui peuvent indiquer des relations fonctionnelles, structurelles ou évolutives entre plusieurs séquences biologiques. Nous considérons un ensemble de k séquences ADN / protéine $S = S1, S2, ..., Sk$. Un alignement multiple

M de S est un ensemble de k séquences de même longueur S'1, S'2, ..., S'K qui sont obtenues en insérant des espaces. L'exemple suivant montre l'alignement multiple de 4 séquences d'ADN.
$S1 = ACG\_\_GAGA.$
$S2 = \_CGTTGACA.$
$S3 = AC\_T\_GA\_A.$
$S4 = CCGTTCAC\_.$
Par conséquent, il faut se décider quels alignements sont plus susceptibles d'avoir eu lieu parce que les séquences sont effectivement liées, ou tout simplement par hasard.

- **Schéma de notation :** Lors de la comparaison des séquences de protéines, on cherche des preuves que les séquences ont divergé d'un ancêtre commun par un processus de mutation/sélection. Trois types de base de mutations: les substitutions (lorsque le résidu est modifiée à une autre), les insertions et les suppressions (lorsque le résidu est ajouté ou supprimé de l'une des deux séquences). Les insertions et les suppressions sont appelées lacunes (gaps).
  Le score total de l'alignement est la somme de chacune des mutations. Certaines substitutions d'acides aminés sont plus susceptibles de se produire que d'autres basée sur les propriétés chimiques. Selon [Tri09], la sérine et la thréonine ont un groupe hydroxyle réactif, qui forme facilement des liaisons hydrogène avec une variété de substrats polaires. Par conséquent, les substituts efficaces de la serine ou de la thréonine devraient se produire assez fréquemment. Ces probabilités de substitution sont habituellement représentées sous la forme d'une table, appelée matrice de substitution. Chaque matrice est représentée par vingt lignes et vingt colonnes (pour les vingt acides aminés standards) ; la valeur dans une cellule donnée représente la probabilité d'une substitution d'un acide aminé par un autre.
  Les matrices PAM : Les matrices PAM ont été développées par [DSO78], et dérivées de 1 572 mutations observées dans 71 familles de protéines étroitement liées. Les matrices PAM sont normalisées de telle sorte que la matrice PAM 1 ait une mutation pour 100 acides aminés et soit appropriée pour des séquences de notation très similaires, Les matrices PAM pour comparer des séquences de similarité inférieure sont calculées à partir de la multiplication répétée de la matrice PAM1 par elle-même. PAM2 équivaut à deux substitutions par cent acides aminés, elle est définie par: $PAM2 = PAM1^2$. PAM30 et PAM70 sont couramment utilisés dans la pratique.

- **Les matrices BLOSUM :** BLOSUM est un autre type de matrice de substitution pour l'étude de similarité, ont été développé par [HH92], les matrices BLOSUM sont basées sur les alignements observés, sans considérer les protéines étroitement liées comme les matrices PAM, Plusieurs matrices BLOSUM ont été construites, en utilisant différents degrés de conservation des protéines : le pourcentage de conservation utilisé a été ajouté au nom. Par exemple, BLOSUM80 correspond à la matrice construite avec des séquences identiques à plus de 80%, les substitutions de tous les acides

aminés ont été calculées à l'aide de l'équation suivante :

$$S_{ij} = \frac{1}{\lambda} log \frac{(p_{ij})}{(q_i q_j)} \tag{II.2}$$

Où, p_{ij} est la probabilité que deux acides aminés i et j se remplacent dans une séquence homologue, et q_i, q_j sont les probabilités de trouver les acides aminés i et j.

Algorithme de Smith-Waterman : L'algorithme de Smith Waterman [SW81] a été inventé par Temple F.Smith et Michael S.Waterman, définissant un alignement optimal correspond au meilleur valeur de score de différents correspondance entre les acides aminés ou nucléotides des deux séquences, le score est calculé par l'utilisation de matrice de substitution, en construisant une matrice D, indexée par $i_1[1 :: n]$ et j_2 $[1 :: m]$, un indice représentant chacune des deux séquences de longueur n et m respectivement.

$$D(i,j) \ = Max \begin{cases} 0 \\ D(i-1, j-1) + s(x_i, y_j) \\ D(i-1, j) - d \\ D(i, j-1) \end{cases} \tag{II.3}$$

Où s (xi; yj) est le score pour la substitution du résidu i de x par le résidu j de y donné par la matrice de substitution de notation, d le coût linéaire des écarts. L'option '0' correspond au début d'un nouvel alignement, si le meilleur alignement devient négatif, il vaut mieux commencer un nouveau au lieu d'étendre l'ancien. Les conditions initiales pour i = 0 ou j = 0 sont définies par D (i; j) = 0. Compte tenu de D, l'alignement peut alors être facilement obtenu en récupérant le chemin dans la matrice qu'il fallait suivre pour calculer F (n, m). Ce processus s'appelle « back tracking ».

L'algorithme BLAST : BLAST [Alt+90] est l'outil le plus largement utilisé pour la recherche d'homologie dans les bases de données ADN/protéines. Cherche des régions de similarité locale entre une séquence requête et les séquences de base de données. Il a été utilisé par de nombreux biologistes pour découvrir des relations fonctionnelles et évolutives entre les séquences et identifier les membres des familles de gènes, il existe quatre types de BLAST définit dans le tableau suivant :

	ADN	Protéine
ADN	BLASTN	TBLASTN
Protéine	BLASTX	BLASTP

Table II.1: Les types de l'algorithme BLAST

Principe de l'algorithme BLAST :

Algorithme 1 L'Algorithme BLAST

1: La décomposition de la requête en plusieurs parties (K-Uplets) et construit un dictionnaire de mots.
2: Fixer une valeur de seuil de similarité S par l'utilisateur.
3: Construit un dictionnaire pour chaque k-uplet de la séquence requête, en calculant le score d'alignement par une matrice de similarité (BLOSUM62, PAM etc.), les k-uplets qui ont une valeur de score supérieur à la valeur de seuil S sont sélectionnés.
4: BLAST vérifie si les k-uplets du dictionnaire présentent dans la BDD, ça signifié qu'il existe une région homologue entre la séquence requête et la séquence de BDD, cette homologie est présentée par un score minimum égale au seuil S.
5: Pour chaque séquence homologue à la séquence requête, BLAST étends l'alignement dans les deux directions de la région homologue et recalculer le score de similarité si il est supérieur au seuil S, l'alignement est conservé pour l'analyse finale.
6: BLAST vérifie quels sont les alignements biologiquement pertinents, par une analyse de la distribution des scores d'alignement entre la séquence requêtes et les séquences de la BDD, cette analyse est faite par un calcul de l'espérance mathématique E-value, si E-value est inclut dans l'intervalle (10-10 à 10-200), ça signifié que l'alignement est biologiquement significatif.
7: Les séquences ayant des zones de similitude avec la séquence requête respectant la valeur E-value, sont sélectionné comme séquences biologiquement similaires à la séquence requête.

2. **L'algorithme FASTA :** FastA [LP85], est l'un des plus anciens logiciels connus dans le domaine de la bio-informatique. De même que son équivalent pour les protéines FastP [PL88], le programme FASTA ne considère que les séquences présentant une région de forte similitude avec la séquence recherchée. Il applique ensuite localement à chacune de ces meilleures zones de ressemblance un algorithme d'alignement optimal. La codification numérique des séquences, c'est-à-dire la décomposition de la séquence en courts motifs (nommés uplets) transcodés en entiers, confère à l'algorithme l'essentiel de sa rapidité, selon [2] FASTA se compose de quatres étapes différentes:
   Etape 1: Lorsqu'une séquence est comparée à une base de données, la première étape est effectuée pour chaque séquence présente dans cette base de données.

   (a) Les régions les plus denses en identités entre les deux séquences sont recherchées. Ces régions sont appelés points chauds ou "hot spots".

   (b) C'est le paramètre "ktup" qui détermine le nombre minimum de résidus consécutifs identiques. Généralement : ktup = 2 pour les protéines - ktup = 6 pour l'ADN.

   (c) Recherche des meilleures diagonales : plusieurs "hot spots" dans une même région génère des diagonales de similarité sans insertion ni délétions. Ces diagonales sont les régions ayant le plus de similarité. Elles sont représentées par un graphique de points ou "dotplot".

[2] Cours: Algorithmes et programmes de comparaison de séquences: est disponible dans: http://biochimej.univ-angers.fr/page2/BIOINFORMATIQUE, consulté le: 2017-10-02.

Etape 2:

(a) Les dix meilleures diagonales sont réévaluées à l'aide d'une matrice de substitution et les extrémités de ces diagonales sont coupées afin de conserver les régions ayant les plus hauts scores seulement. Cette recherche de similitude est faite sans insertions ni délétions.

(b) Le score le plus élevé obtenu est appelé le score "init1". Il est attribué à la région ayant le plus fort score parmi les 10 analysées.

Etape 3:

(a) Les diagonales trouvées à l'étape 1 dont le score dépasse un certain seuil ("cutoff"), sont reliées entre elles pour étendre la meilleure similarité.

(b) Ces nouvelles régions contiennent des insertions et/ou des délétions.

(c) Le score des nouvelles régions est calculé en combinant le score des diagonales reliées diminué d'un score de pénalité de jonction des diagonales.

(d) Le score le plus élevé obtenu à cette étape s'appelle le score "initn".

(e) Cette étape permet d'éliminer les segments peu probables parmi ceux définis à l'étape précédente.

Etape 4:

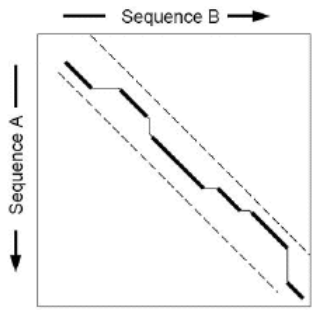

FIGURE II.1: Recherche des meilleures diagonales entre deux séquences A et B par FASTA

(a) La région initiale qui a généré le score"init1" est de nouveau évaluée avec un algorithme de programmation dynamique sur une fenêtre de résidus dont la largeur est déterminée par le paramètre "ktup". Le nouveau score est "opt".

(b) Les séquences de la base de données sont classées selon leurs scores "initn" ou "opt".

(c) Les séquences sont alignées avec la séquence cible à l'aide de l'algorithme de Smith et Waterman : le score final est le score Smith rt Waterman.

La sortie de FASTA se décompose en trois parties :

- colonne 1 : échelle de valeurs
- colonne 2 : nombre de séquences dans la banque donnant un "z-score" = valeur.
- colonne 3 : nombre de séquences dans la banque donnant une "E-value" = valeur.
- "init1" = "initn" = "opt" : 100% de similarité.
- "initn" > "init1" : plusieurs régions de similarité reliées par des gaps.
- "initn" > "opt" : pas de similarité.

3. **Similarité par représentation graphique :** Récemment, la représentation graphique est bien considérée, qui peut offrir une inspection visuelle des données et fournir un moyen simple pour faciliter l'analyse de similarité et la comparaison des séquences biologiques. En raison de son excellente maniabilité, actuellement, plusieurs méthodes basées sur la représentation graphique ont été largement appliquées dans les domaines pertinents de la bio-informatique. Jusqu'à présent, plusieurs méthodes de représentation graphique ont été proposées, Pour donner une caractérisation numérique aux séquences biologiques sur la base de différents espaces à multiples dimensions:

   (a) [Ran+03] Nouvelle représentation graphique 2D des séquences d'ADN et leur caractérisation numérique.

   (b) [GRB01]Nouvelle représentation graphique 2-D des séquences d'ADN de faible dégénérescence.

   (c) [Liu+06]Nouvelle représentation graphique 2D des séquences d'ADN et son application PNN-curve.

   (d) [QQ07] Nouvelle représentation graphique 2D de la séquence d'ADN à base de double nucléotides. aussi [QF07], [YSW09], ont proposé des méthodes de représentation graphique 3D des séquences d'ADN. D'autres représentations 4D et 5D ont été développées dans d'autres travaux [TLZ15; CD05; Lia+07] pour résoudre le problème de la dégénérescence par la représentation graphique.

4. **L'identification des séquences à partir des données expérimentales:** C'est un problème important de bio-informatique pour déterminer les séquences biologiques à partir des données générées par des expériences dans les laboratoires humides. En raison des différences significatives entre les technologies et les méthodes utilisées pour générer des données expérimentales, le problème d'identification de séquence présente plusieurs variantes qui sont totalement différentes du point de vue de calcul. Pour mieux comprendre, nous discutons dans la suite sur les problèmes d'identification des séquences d'ADN et des séquences de protéines.

Identification de séquence d'ADN à partir de données de séquençage : Pour identifier des séquences d'ADN, la technologie dite de séquençage d'ADN est largement utilisée. Dans le séquençage de l'ADN, plusieurs copies de la séquence d'ADN originale sont coupées en millions de fragments d'une manière différente, de sorte qu'un fragment d'une copie peut se superposer des fragments d'un autre. Pour un ensemble de fragments donné, l'assemblage de séquence consiste à aligner et à fusionner des fragments pour reconstituer la séquence d'ADN d'origine.
Identification de séquences de protéines à partir de données de spectrométrie de masse : Dans l'identification de séquences de protéines, une protéine est d'abord digérée en peptides par des protéases telles que la trypsine. Ensuite, le spectromètre de masse en tandem brise les peptides en fragments encore plus petits et enregistre la masse de chaque fragment dans un spectre de masse, Le problème d'identification de la séquence peptidique consiste à dériver la séquence d'un peptide à partir de son spectre de masse. Les séquences peptidiques identifiées sont en outre assemblées pour déduire des séquences de protéine dans la procédure dite d'inférence de protéine.

5. **Classification des séquences biologiques:** Dans la biologie moléculaire, certaine corrélation entre la séquence primaire de l'ADN et les propriétés fonctionnelles et structurales des protéines, ça pose un problème de prédiction de propriétés des séquences et des relations qui ne sont pas encore comprises, dans ce cas, les méthodes de prévision d'informatique et d'apprentissage machine doivent être appliquées pour faire progresser notre compréhension. Le classement des séquences biologiques est un problème important et stimulant en biologie computationnelle. Du point de vue biologique, il aide à identifier des régions de séquences intéressantes et des domaines protéiques qui sont liés à une fonction biologique particulière, nombreuses approches de classification efficaces ont été appliquées pour classifier les données biologiques, comme les modèles génératifs (profil HMMs [Kro+94], [Bal+94]), des modèles discriminatifs (les SVMs [JDH00], [Esk+03], [LN03]), Et des modèles basés sur des graphes [TSS05].

II.2.2.2 Bio-informatique structurale

L'un des plus importantes branches de la biologie moléculaire est la biologie structurale, qui concerne principalement la « structure tertiaire » des macromolécules biologiques. Les macromolécules réalisent la plupart des fonctions cellulaires sur la base de structures tertiaires, la structure tertiaire d'une macromolécule est sa structure tridimensionnelle, par conséquent, de nombreux chercheurs en bio-informatique se concentrent sur l'étude des structures tertiaires des macromolécules telles que l'ARN et les protéines. Les problèmes de la bio-informatique liés à la structure tertiaire des macromolécules peuvent être divisés en trois catégories : (1) Analyse multiples de structures, (2) prédiction de structure, et (3) prédiction basée sur la structure.

1. **Analyse de structure multiple** Nombreuses applications dans les sciences de la vie concentrent sur la comparaison de structures multiples et la découverte de modèles communs à partir d'un ensemble de structures (ARN/ Protéine). L'analyse de structure multiple de protéines peut être utilisée pour caractériser des familles de protéines fonctionnellement ou structurellement liées, et aussi pour révéler les relations entre les séquences, les structures et les fonctions des

protéines. Deux problèmes major de la bio-informatique dans cette catégorie, sont : l'alignement structurel et la découverte de motifs structuraux.
Alignement structurel : Tenter d'établir l'homologie entre deux ou plusieurs structures basées sur leur conformation tridimensionnelle, cette procédure est généralement appliquée aux structures tertiaires protéiques, qui peut transférer des informations sur une protéine bien connue à des protéines inconnues qui peuvent être alignées structurellement.
Découverte de motifs structuraux : Un motif structurel est un ensemble récurrent de résidus spatialement proches en trois dimensions, mais pas nécessairement adjacents dans la séquence. De tels motifs sont utiles pour révéler des relations évolutives et fonctionnelles intéressantes entre les protéines lorsque la similarité des protéines est très faible [Zen15].

2. **Prédiction de structure** En raison de la difficulté d'obtenir la structure tertiaire de chaque ARN et protéine par des expériences de laboratoire humides, Pour la prédiction de la structure des protéines, un grand nombre d'outils de prédiction ont été développés au cours des 20 dernières années, ces outils adoptent différents principes tels que la modélisation par homologie [Vya+12], le threading de protéines [LRW95] et les méthodes ab initio [YSB03].

3. **Prédiction basée sur la structure :** Les fonctions des protéines dans la cellule sont déterminées par la structure tertiaire, cette structure est une information importante dans plusieurs applications de bio-informatique qui peut être unifiée sous le nom « prédiction Basée sur la structure » comme la prédiction des interactions protéines-protéines (IPP), la prédiction des fonctions protéiques et des cibles médicamenteuses.

II.2.2.3 Bio-informatique des réseaux

Les systèmes biologiques complexes sont généralement présentés et analysés comme des réseaux, où les sommets représentent des unités biologiques et les arcs représentent les interactions entre les unités, les différents types de réseaux biologiques comprennent, les réseaux de co-expression de gènes, les réseaux métaboliques, les réseaux de signalisation, les réseaux PPI ,...etc Selon [Zen15], le grand nombre de problèmes de bio-informatique liés au réseau peuvent être classé en trois catégories : (1) analyse de réseau, (2) inférence de réseau, et (3) prédiction assistée par réseau.

1. **Analyse de réseau:** Pour mieux comprendre l'organisation et la structure des grands réseaux biologiques, leurs structures et propriétés topologiques, leurs propriétés dynamiques et les relations fonctionnalité-topologie on devrait être passé par l'analyse des réseaux. L'analyse de réseau devient la méthodologie clé pour l'étude des systèmes biologiques complexes. Il existe de nombreuses tâches d'analyse de réseau biologique, comme le problème de comparaison de réseau est défini comme un processus de comparaison entre deux ou plusieurs réseaux biologiques provenant de différentes espèces, conditions ou types d'interactions. Les résultats de la comparaison peuvent indiquer quelles interactions protéiques sont susceptibles d'avoir des fonctions équivalentes à travers les espèces. De plus, il est également possible de révéler l'évolution sous-jacente des protéines, des réseaux et même de l'espèce entière. Généralement, trois types de méthodes de

calcul sont disponibles pour la comparaison de réseau: (1) alignement de réseau, (2) intégration de réseau, et (3) interrogation de réseau.

2. **Réseau d'inférence:** Il est très difficile de déterminer la structure du réseau par la validation expérimentale de toutes les paires d'interactions entre les unités biologiques. Une approche plus pratique est d'inférer la structure du réseau à partir de la preuve indirecte cachée dans les données expérimentales biologiques. Le sujet de l'inférence de réseau biologique est d'un grand intérêt a été largement étudié, il existe de nombreuses méthodes pour déduire différents types de réseaux biologiques tels que les réseaux de régulation des gènes et les réseaux PPI, et aussi plusieurs méthodes de calcul ont été développées, Pour dériver la structure de réseau sous-jacente entre les protéines.

3. **Prédiction assistée par réseau:** Les réseaux à grande échelle d'interactions moléculaires disponibles à l'intérieur de la cellule permettent d'étudier de nombreux problèmes bio-informatiques dans le contexte d'un réseau. Les applications typiques comprennent la prédiction des fonctions protéiques, prédiction des gènes pathogènes et drogue-cibles. L'idée de base de cette prédiction assistée par réseau est d'utiliser l'information de corrélation entre les entités biologiques dans le réseau pour améliorer la précision de la prédiction.

II.2.2.4 Bio-informatique fonctionnelle

Une grande partie de la recherche bio-informatique est impliquée dans la collecte, le stockage et la récupération de grandes quantités d'informations diverses. Les biologistes souhaitent non seulement effectuer ces tâches, mais également acquérir de nouvelles connaissances déduisant des interactions fonctionnelles entre des données biologiques aussi diverses. Donc, l'application des approches informatiques pour atteindre ces objectifs de la biologie fonctionnelle est appelé la bio-informatique fonctionnelle. L'aspect fonctionnel est signifié par plusieurs champs :

1. **Fonctions d'une protéine :**
   Le terme fonction d'une protéine peut avoir de grandes implications, peut être décrire des données liées à plusieurs niveaux, et peut avoir différentes interprétations dans différents contextes biologiques, cela dépend toujours du contexte en ce qui concerne le tissu, l'organe et le taxon. Chaque type de protéine consiste en une séquence précise d'acides aminés qui lui permet de se replier en une forme particulière pour effectuer ses tâches. La fonction de la protéine n'est pas un terme bien défini; la fonction est un phénomène complexe associé à de nombreux niveaux qui se chevauchent mutuellement : (biochimique, cellulaire, médiation par l'organisme, développemental et physiologique). Ces niveaux de chevauchement sont entrelacés de manière complexe. D'un autre côté, les protéines ne sont pas des morceaux rigides de matériau. Ils peuvent avoir des pièces mobiles conçues avec précision dont les actions mécaniques sont couplées à des événements chimiques [Alb+02].
   Cependant, Les méthodes de prédiction de la fonction protéique sont des techniques que les chercheurs en bio-informatique utilisent pour attribuer des rôles biologiques ou biochimiques aux protéines, plusieurs aspects de la bio-informatique s'inscrit dans la prédiction des fonctions protéique :

- Comparaison de séquences de protéines : Les protéines de séquence similaire sont généralement homologues [Ree+87] et ont donc une fonction similaire. Par conséquent, les protéines dans un génome nouvellement séquencé sont annotées de façon routinière en utilisant les séquences de protéines similaires dans d'autres génomes.

- Le développement de bases de données de domaines protéiques, permet de trouver des domaines connus dans une séquence de requête et les prédictions de fonction d'une manière plus réaliste dans les domaines protéiques.

- Certains serveurs de prédiction de fonctions tels que RaptorX [3], peuvent également prédire le modèle 3D d'une séquence, puis utiliser une méthode basée sur la structure pour prédire les fonctions basées sur le modèle 3D.

2. **Les peptides signaux :** Les peptides signaux courts dirigent certaines protéines vers un endroit particulier tel que les mitochondries, et il existe divers outils pour la prédiction de ces signaux dans une séquence protéique [MHA00]. Par exemple, signalP qui a été mis à jour plusieurs fois en tant que méthodes, est amélioré [Pet+11] dans la suite.

3. **Fonction des motifs structurels :** Les sites actifs sont des portions localisées dans les structures protéiques, responsables pour exprimer des fonctions. Les propriétés structurelles et physiques de site actif sont responsables de l'expression de fonction. La similarité des propriétés structurales et physiques des sites actifs donnera lieu à des similarités de fonctions dans différentes protéines. Ces structures pourraient être des cavités dans la structure des protéines qui peuvent servir à la liaison avec d'autres protéines[Ree+87].
   Dans de nombreux cas, au lieu de la structure protéique complète, la structure 3D d'un motif particulier représentant un site actif ou un site de liaison peut être ciblée. [SW10; Wan+13; GJ16; GMJ16] La méthode SALSA (Structuralally Aligned Local Sites of Activity) [Wan+13], utilisent les propriétés chimiques calculées des acides aminés individuels pour identifier les sites locaux biochimiquement actifs. Des bases de données telles que Catalytic Site Atlas [PBT04] ont été développées qui peuvent être recherchées en utilisant de nouvelles séquences protéiques pour prédire des sites fonctionnels spécifiques.

4. **Développement de médicaments :** L'un des principaux aspects dans la science de la vie et l'étude des données biologiques est le développement de médicaments, selon [Len02], la bio-informatique joue un rôle principal dans ce domaine. Cependant, Le séquençage complet des génomes de l'homme et d'autres espèces représente une puissante incitation à développer de nouveaux médicaments. Les projets de séquençage et d'analyse des génomes ont considérablement accru nos connaissances sur les protéines codées par le génome humain. Cette nouvelle source de connaissances pourrait formidablement accélérer les stades précoces du processus de développement de médicaments ou même permettre de développer des médicaments personnalisés pour les différents patients. Le développement d'un médicament est lié aux séquences moléculaires de base (protéine), beaucoup de ces protéines pouvant être un candidat prometteur pour le développement

[3] RaptorX: http://raptorx.uchicago.edu/, consulté le: 2017-04-03

d'une thérapie pour les personnes dont la maladie est causée par les anomalies dans les gènes. Il y a plusieurs manières d'identifier des protéines qui puissent servir de cibles dans des programmes de développement de médicaments. Une manière consiste à chercher des changements dans le spectre d'expression des protéines, leur localisation, ou leurs modifications post-traductionnelles dans les organismes atteints par des maladies. Une autre manière de procéder consiste à conduire des recherches sur les tissus ou les types cellulaires dans lesquels des gènes particuliers sont exprimés. L'analyse du génome humain devrait augmenter le nombre de cibles de médicaments. Plusieurs techniques et méthodes d'informatique, de statistique et de mathématique peuvent conduire au développement d'un médicament, comme la prédiction de structure 3D, comparaison de séquences, remplacer ou de compléter les tests des médicaments chez les animaux, par des programmes informatique, etc.

5. **Génomique fonctionnelle :** Le but de la génomique fonctionnelle est de comprendre la fonction d'un plus grand nombre de gènes ou de protéines, éventuellement tous les composants d'un génome, de générer et de synthétiser les connaissances génomiques et protéiques dans une compréhension des propriétés dynamiques d'un organisme. Cela fournirait une image plus complète que les études de gènes uniques dans l'objectif de comprendre la relation entre le génome d'un organisme et son phénotype. Le terme génomique fonctionnel est souvent utilisé pour désigner les nombreuses approches et techniques pour étudier les gènes et les protéines d'un organisme, la génomique fonctionnelle peut également inclure des études de la variation génétique naturelle au cours du temps (comme le développement d'un organisme) ou de l'espace (comme les régions de son corps) et aussi des perturbations fonctionnelles telles que des mutations [Gor08].

II.3 Processus ECD

Les progrès des technologies de l'information ont facilité le stockage et la distribution des données au cours des deux dernières décennies. Des quantités énormes de données ont été accumulées à un rythme très rapide. Cependant, ces données ne sont parfois pas significatives dans leurs formats initiaux mais ce que les utilisateurs veulent, c'est la connaissance cachée dans ces données. Ces connaissances peuvent être considérées comme des caractéristiques des données, beaucoup plus précieux que les données originales. Pour cela, un nouveau domaine technologique a vu le jour au milieu des années 90 pour la découverte de connaissances et d'informations à partir de données. C'est ce qu'on appelle (ECD) l'extraction des connaissances à partir des données (knowledge Discovery in databases en anglais) ou simplement la fouille de données (Data Mining) [CHY96; Fay+96]. La découverte de connaissances dans les bases de données (ECD) est un processus général de découverte de modèles cachés dans les données pour une meilleure prise de décision. Le processus ECD est défini en cinq étapes principales : préparation des données et définition de problème, pré-traitement des données, fouille de données, évaluation, interprétation et mise en œuvre. Les informations découvertes doivent être :

1. **Nouveau :** Le bon sens ou les faits connus ne sont pas ce qui est recherché.

2. **Correct :** La sélection ou la représentation inappropriée des données entraînera des résultats incorrects. Les informations extraites doivent être soigneusement vérifiées par les experts du domaine.

3. **Significatif :** L'information minée doit signifier quelque chose et peut être facilement comprise.

4. **Applicable :** L'information minée devrait pouvoir être utilisée dans un certain domaine.

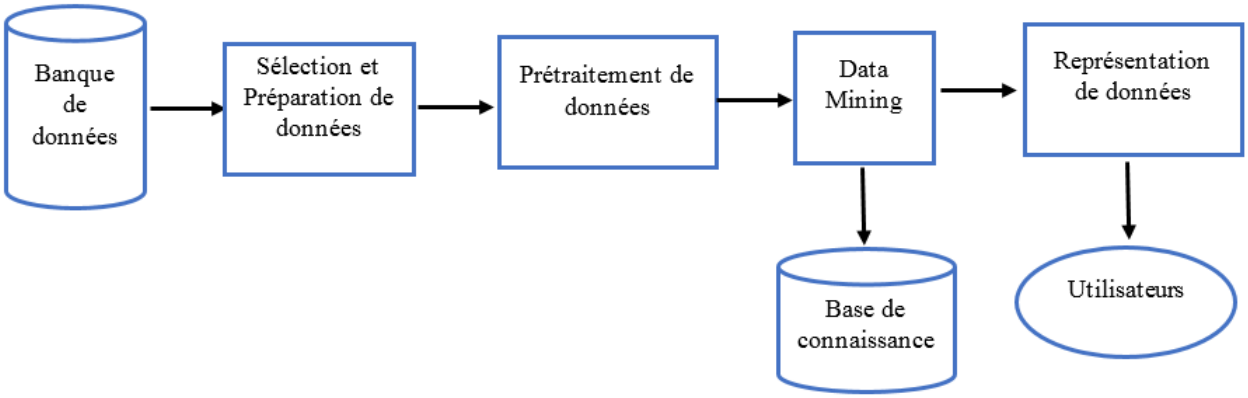

FIGURE II.2: Le processus ECD

II.3.1 Définition de problèmes

Le but de cette étape est de définir à la fois l'objectif de l'étude et le problème (objectifs et attentes) et de définir les connaissances de domaine qui peuvent être nécessaires. Ceci est effectué de manière itérative ; l'utilisateur entreprend des étapes ultérieures plusieurs fois avant de parvenir à une définition de problème satisfaisante. On distingue deux types d'objectifs : la vérification et la découverte [Fay+96], le but de la vérification est de limité à la vérification de l'hypothèse de l'utilisateur nécessite une spécification préalable des hypothèses par le décideur ; tandis que les découvertes prédisent de manière autonome et expliquent de nouvelles connaissances. Les processus de découverte de connaissances biologiques devraient prendre en compte à la fois les caractéristiques des données biologiques et les exigences générales du processus de découverte de connaissances (ECD), dans le but de réaliser les objectifs préalablement fixés et résoudre le problème posé.

II.3.2 Préparation de données

Avant toute application, les données nécessaires à l'analyse sont déterminées d'abord, ces données peuvent alors être collectées à partir de données déjà existantes telles que des fichiers, des bases de données, des entrepôts de données ou des datamarts [Pyl99], si l'ensemble de données construit de cette manière devient très important, une forme représentative réduite de celui-ci peut être obtenue par une procédure d'échantillonnage. Lorsque les données sont sélectionnées, se sont placés dans une

représentation standard, dans un format de table, où les instances et les points sont placés dans des lignes et des colonnes respectivement.

II.3.3 Pré-traitement de données

Les données dans les applications de fouille de données sont généralement, incomplètes et incohérentes. Les redondances peuvent également se produire en raison de l'intégration de données provenant de diverses sources. Le but principal de l'étape de pré-traitement des données est de gérer ces types de données pour améliorer leur qualité. En outre, la transformation et la réduction des données peuvent aider à améliorer la précision et l'efficacité des techniques de FD. Les taches de pré-traitement des données de base peuvent être organisées dans les étapes suivantes :

- **Nettoyage des données**
  Cette étape implique des techniques pour remplir les valeurs manquantes, lissant le bruit, manipulation de valeurs aberrantes, détection et suppression des données redondantes, les valeurs aberrantes dans les ensembles de données sont généralement supprimées. En outre, les données incomplètes sont filtrées par les techniques de filtrage pour une analyse multidimensionnelle sur des bases de données volumineuses.

- **Transformation des données**
  La transformation des données à une représentation appropriée est nécessaire. Les données peuvent être codées en une représentation vectorielle. De plus, les données peuvent être transformées d'un espace de grande dimension à une dimension inférieure pour trouver des caractéristiques plus importantes et réduire l'effort à la phase de fouille de données.

- **Réduction des données**
  Le but de cette étape est la réduction de la taille des données, plusieurs méthodes ont été réalisés quelques une utilisé pour construire le cube de données, pour stocker le résumé multidimensionnel des données. Quelques d'autres pour éliminer les attributs inutiles, Ils existent plusieurs méthodes de: Analyse de corrélation (AC), Analyse de la variance (ANVA), machine a vecteur de support (MVS), les algorithmes génétiques(GA), analyse des composants principales (ACP), etc.

- **Discrétisation** De nombreux algorithmes dans la FD et la grande majorité des algorithmes d'extraction des règles d'associations et leurs dérivations fonctionnent uniquement avec des attributs catégoriels. Il est donc nécessaire d'utiliser des techniques de discrétisation des données afin de réduire le nombre de valeurs pour un attribut continu donné en divisant la plage de l'attribut en intervalles. En d'autres termes, le but des techniques de discrétisation est de convertir les attributs numériques en attributs nominaux. Il existe trois axes par lesquels les méthodes de discrétisation peuvent être classées: global/local, supervisé/non supervisé et statique/dynamique[DKS95]. La discrétisation est le processus de conversion de variables de valeur continue en valeurs discrètes où un nombre limité d'étiquettes sont utilisées pour représenter les variables d'origine. Les valeurs discrètes peuvent avoir un nombre limité d'intervalles dans un spectre continu, alors que les valeurs continues peuvent être infiniment nombreuses. Il existe un certain

nombre d'algorithmes pour la discrétisation des données comme, l'algorithme de raisonnement booléen, From file with cuts, Equal frequency binning, Naive algorithm, etc.

II.3.4 Fouille de données (FD)

La fouille de données signifie « recherche d'information utile dan un grand ensemble de données », on la traduit en anglais par l'expression «Data Mining ». Il existe plusieurs définitions générales de FD, nous en avons sélectionné quelques unes :

- *Han et Kamber:[HPK11]* l'extraction des informations intéressantes non triviales, implicites, préalablement inconnues et potentiellement utiles, à partir de grandes BDDs .
- *Frawley et Piateskishapiro:[FPSM92]* l'extraction d'informations originales, auparavant inconnues potentiellement utiles à partir de données.

la FD est donc définie comme l'ensemble d'algorithmes et techniques d'apprentissage automatiques permettent d'obtenir de connaissances exploitables, à partir d'une base de données. Il s'agit un processus de sélection, d'exploitation, de modification et de modélisation de grande BDDs afin de découvrir de relations implicites et des régularités entre les données. D'autre part ce qui confondent un concept de FD et celui de ECD, et les considèrent comme synonymes. Alternativement, Duval voit FD comme simplement un élément essentiel intervenant dans le processus de découverte de connaissances dans des BDDs. Les techniques de FD peuvent être divisées en deux grandes catégories, descriptives et prédictives; y compris la description, le regroupement, l'association, la classification, la prédiction, l'analyse de similarité. Le choix de méthodes utilisées lors de l'exécution des tâches dépend essentiellement du type de données extraites. Beaucoup de concepts sont utiles pour le même but de miner l'information cachée des données. Parmi eux, les algorithmes, les bases de données, les statistiques, l'apprentissage automatique et la récupération d'information.

II.3.5 Evaluation et interprétation

Les informations extraites doivent être interprétées par des experts humains. Les résultats interprétés sont ensuite évalués par leur nouveauté, leur exactitude, leur compréhensibilité et leur utilité. Seules les informations transitant par ce processus de filtrage peuvent être utilisées dans des applications réelles. L'évaluation des méthodes de FM pour parvenir à une décision finale, c'est-à-dire la sélection du meilleur modèle, nécessite une comparaison des résultats obtenus à partir de diverses méthodes FD. Plusieurs critères sont utilisés à cet effet, y compris des mesures pour évaluer leur performance et leur exactitude ainsi que l'évaluation du temps et des ressources nécessaires.

II.4 Fouille de données biologiques

Après avoir concentré sur l'accumulation de données dans les banques de données biologiques, nous concentrons dans cette partie sur l'analyse de ces données. L'analyse de l'énorme quantité de données est une tâche difficile, non seulement en raison de sa

complexité mais aussi en raison de l'évolution continue de notre compréhension des mécanismes biologiques. Les approches classiques de l'analyse des données biologiques ne sont plus efficaces et ne produisent qu'une quantité très limitée d'informations, par rapport aux nombreux mécanismes biologiques complexes en étude actuels. Donc la nécessité d'utiliser des outils informatiques et de développer de nouvelles approches pour l'analyse des données biologiques aider à comprendre les corrélations qui existent entre, d'une part, les structures et les fonctions des séquences biologiques et, d'autre part, les mécanismes génétiques et biochimiques, donc les méthodes de fouille des données biologiques peuvent répondre à ces nouvelles tendances. La fouille de données est le noyau du processus ECD [Fay+96], dans le but de découvrir des informations à partir de données, utilise des technologies provenant de différents domaines de l'informatique et des sciences de l'information. Les trois principales sont les bases de données, l'apprentissage automatique et les statistiques. La technologie de base de données gère les données pour une sélection pratique. La technologie d'apprentissage automatique apprend des informations ou des modèles à partir de données de manière automatique, et les statistiques trouvent les caractéristiques ou les paramètres statistiques des données. Data mining est probablement l'outil de calcul le plus populaire en biologie moléculaire. Vise à extraire des motifs, des fonctionnalités, des regroupements/classifications de séquences biologiques avec différentes techniques et algorithmes. De nombreux problèmes de bio-informatique peuvent être classé en tant que problèmes de fouille de données standard, en raison de la complexité du processus et des méthodes de fouille, les gens ne peuvent pas facilement utiliser les techniques de data mining pour résoudre leurs problèmes de bio-informatique. Cependant, certains problèmes de bio-informatique ne peuvent pas être modélisés par les méthodes de fouille existantes, ce qui rend nécessaire de développer de nouvelles techniques et solutions de fouille de données. En outre, il reste très difficile de fournir des estimations de rendement de certains algorithmes de bio-informatique. Ce problème viendra plus sérieux quand il n'y a pas de jeux de données de référence, ce qui rend difficile d'évaluer avec précision la performance des algorithmes. Par conséquent, les recherches devraient être consacrées au développement d'algorithmes de validation efficaces pour l'évaluation des résultats de fouille de données dans les applications de bio-informatique. Dans les sections II.4.1, II.4.2 et II.4.3, nous présenterons les trois problèmes de bases de fouille de données biologiques la classification supervisé, classification non-supervisé et les règles d'association.

II.4.1 Classification Non-supervisée des données biologiques

Le taux de croissance des bases de données biologiques est également exponentiel. Il est nécessaire d'explorer et d'analyser de telles données massives pour déduire des informations inhérentes. La classification non-supervisé a été reconnu comme l'une des techniques de fouille de données les plus couramment utilisées, est un moyen prometteur pour l'analyse et l'extraction des connaissances cachées dans les données biologiques. Au cours des dernières années, plusieurs travaux de recherche ont été menés sur l'analyse de clusters et un grand nombre d'algorithmes ont été mis au point, particulièrement pour les données biologiques. Les données biologiques contiennent une grande quantité de connaissances qui peuvent être inconnues mais utiles, la fonction biologique n'est pas tout jour déterminée par un seul gène ou protéine mais aussi par les relations complexes entre eux. Donc lorsque nous rencontrons de nouvelles données biologiques, nous essayons toujours de rechercher les fonctionnalités qui peuvent les décrire, les

comparant avec des données que nous connaissions déjà. Le but de clustering des séquences génomiques est de regrouper les données en clusters selon une mesure de similarité [DD06], les gènes qui partagent des propriétés semblables inclut dans le même cluster peuvent impliquer des relations dans les voies fonctionnelles. Ainsi, le clustering pourrait fournir un moyen de comprendre la fonction des gènes pour lesquels l'information n'a pas été disponible précédemment [JTZ04a]. De plus, le clustering peut être utilisé comme une étape de pré-traitement avant une étape de sélection de caractéristiques ou un algorithme de classification, dans le but de limiter l'analyse à une catégorie spécifique et éviter la redondance.

II.4.1.1 Défis de clustering

Le regroupement consiste à grouper des données similaires en un ensemble fini de clusters séparés. Cette tâche est également appelée segmentation, le nombre de clusters et les catégories ne sont pas connus à l'avance. Les principaux défis concernant l'application du clustering aux séquences biologiques sont : (1) la définition de la distance appropriée entre les objets, (2) le choix de l'algorithme de clustering, et (3) l'évaluation des résultats finaux. En particulier, l'évaluation des résultats du clustering est une tâche non triviale.

- **La définition de la distance appropriée entre les objets** La similarité entre les objets est définie par le calcul de la distance entre eux, le choix de la fonction de distance joue un rôle important dans l'analyse de cluster. Supposons que le jeu de données X contienne n points, c'est-à-dire, $X = x1, x2, ..., xn$. La distance entre deux points de données xi et xj est notée D (xi, xj). Toute fonction de distance D doit satisfaire les propriétés suivantes :

  1. D $(xi, xj) \geq 0$ pour tout xi, xj $\subseteq X$ et D (xi, xj) $= 0$ seulement si $i = j$.
  2. D $(xi, xj) = D(xj, xi)$ pour tout xj $\subseteq X$.
  3. D $(xi, xj) \leq D(xi, xl) + D(xl, xj)$ pour tout xi, xj, xl $\subseteq X$.

  Plusieurs mesures de distance (Euclidean, Manhattan, Chebyshev, etc.), peuvent être calculées, selon le problème étudié. Cependant, ces mesures de distance ne sont pas toujours adéquates pour détecter les corrélations entre les objets [Wan+02]. D'autres systèmes largement utilisés pour déterminer la similarité entre les gènes utilisent les coefficients de corrélation de Pearson et Spearman [HPK11], qui mesurent la similarité entre les formes de deux motifs de gêne. Cependant, ils ne sont pas robustes en ce qui concerne les valeurs aberrantes. La corrélation cosinus s'est avérée plus robuste sur les valeurs aberrantes car elle calcule le cosinus de l'angle entre les vecteurs de valeur de gêne.

- **Choix de l'algorithme de clustering:** Plusieurs algorithmes ont été conçus précédemment pour le clustering des données d'expressions génétique dans le but de résoudre plusieurs problèmes biologiques, certains de ces algorithmes se sont avérés bons dans le domaine biologique, en obtenant de bons résultats et en accélérant le processus d'analyse. Ainsi qu'ils divisent les échantillons de données en différents groupes. (1) les échantillons de données dans le même cluster sont similaires et (2) les échantillons de données de différents clusters sont

dissemblables. Il existe de nombreux algorithmes de clustering dans la littérature, qui peuvent être classées en différentes catégories. Parmi les algorithmes de clustering existants, les méthodes les plus largement utilisées sont les méthodes de partitionnement [Her+99; DK03], les méthodes hiérarchiques [Eis+98], selef organizing map (SOM) [Tam+99], algorithmes classiques ou adaptés aux données sur l'expression génique [FM07; JTZ04b; DD06] et de nouveaux algorithmes, qui traitent spécifiquement des données d'expression génétique, ont été récemment proposés [GL08]. Nous discutons dans la suite les algorithmes les plus connus et les plus utilisés :

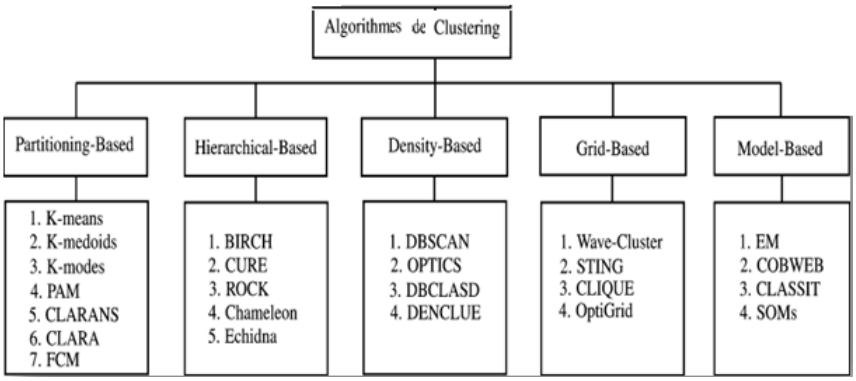

FIGURE II.3: Les Catégories des algorithmes de clustering

Nous discutons dans cette section des algorithmes les plus connus et les plus utilisés :

II.4.1.2 Méthodes de partitionnement

Cette famille d'algorithmes de clustering fonctionne de la même manière que k-means [Mac+67]. K-means est l'un des algorithmes de clustering le plus simple et le plus rapide, Il prend le nombre de cluster (k) comme entrée, puis diviser de façon aléatoire les objets en k cluster. En-suite, il calcule itérativement le centre de chaque cluster et déplace chaque objet au cluster le plus proche. Cette procédure est répétée jusqu'à ce qu'aucun autre objet ne soit déplacé vers un autre cluster. Malgré sa simplicité, le k-means présente certains inconvénients majeurs, tels que la sensibilité aux valeurs aberrantes, et la sensibilité à la valeur de K, parce que le nombre de cluster doit être connu à l'avance donc la pertinence des résultats finaux liée à cette valeur de K. L'algorithme suivant présente les étapes de K-means.

Algorithme 2 L'Algorithme K-moyennes

1: Entrée:
2: Ensemble de données N X1,.....XN
3: Paramètre K: nombre de clusters
4: Sortie
5: une partition de K groupes C1, C2 ,....Ck
6: Début
7: Initialisation aléatoire des centres Ck;
8: Répéter
9: Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche: $X_i \epsilon Ck\, si \forall |X_i - \mu k| = min|X_i - \mu k|$ Avec μ_k le centre de la classe K;
10: Représentation: Calculer les centres associe à la nouvelle partition:
11: $mu_k = \frac{1}{N} \sum x \in ckXi$
12: Jusqu'à convergence de l'algorithme vers une partition stable;
13: Fin.

Plusieurs algorithmes de clustering ont été proposés pour surmonter les inconvénients de k-means. L'algorithme « genetic weighted k-mean » [Wu08] est une hybridation d'un algorithme génétique et d'un algorithme weighted k-mean, les auteurs montrent qu'il obtient de meilleurs résultats que les k-means classique, en termes de qualité de cluster et de sensibilité de clustering aux partitions initiales. L'algorithme c-moyens floues (fuzzyc means) [DK03], un autre type d'algorithme de clustering appliqué aux données génomique, relié chaque gène à tous les clusters via un vecteur d'indices réels, les valeurs de ce vecteur variées entre 0 et 1, pour un gène donné, un indice proche de 1 indique une forte association au cluster et aussi si l'indice proche de 0 indique l'absence d'association forte au cluster correspondant. Le vecteur des indices définit ainsi l'appartenance d'un gène par rapport aux différents clusters.
L'algorithme de cluster d'attributs (ACA) [Au+05], adopte l'idée de k-means pour regrouper les gènes, en remplaçant la mesure de distance par la mesure de redondance d'interdépendance entre les attributs et la notion de moyenne par le concept de mode.

II.4.1.3 Algorithmes Hiérarchiques

Dans la classification hiérarchique, les clusters sont générés dans une hiérarchie, où chaque niveau de la hiérarchie fournit un regroupement particulier des données, allant d'un cluster unique (où tous les points sont placés dans le même cluster) à n clusters (où chaque point comprend un cluster). Le regroupement hiérarchique peut être agglomératif ou divisif pour former le dendrogramme hiérarchique. Les algorithmes agglomératifs (approche ascendante) considèrent initialement chaque objet de données comme un cluster individuel, et à chaque étape, fusionnent les clusters les plus proches jusqu'à ce que tous les clusters soient fusionnés à un seul. Les algorithmes divisifs (approche descendante) commencent par un cluster contenant tous les objets de données, à chaque étape divise un cluster jusqu'à ce que seuls les clusters qui contient un objet individuel restent. Par exemple, l'algorithme agglomératif UPGMA (Unweighted Pair Group Method with Arithmetic Mean)[Eis+98] et l'algorithme divisif deterministic annealing [Alo+99], pour le clustering des génes.

II.4.1.4 Clustering basé sur la densité

Dans les approches de classification basées sur la densité, les clusters sont considérés comme des régions dans l'espace de données dans lequel les points sont densément situés et sont séparés par des régions de densité de point faible (bruit). Ces régions peuvent avoir des formes arbitraires et les points à l'intérieur d'une région peuvent être distribués de manière arbitraire. Les algorithmes de ce type de classification sont caractérisés par :

1. La densité à la place de mesure de distance.
2. Si la distance entre deux points est inférieure à une valeur fixée, donc les points sont voisins.
3. Si le nombre de voisins d'un point dépasse un certain seuil, ce point est dense.

Algorithme DBSCAN : (clustering spatial basé sur la densité des applications avec bruit) [Est+96] Est une technique de clustering basée sur la densité, utilise deux paramètres importants :

1. La distance ε (rayon du voisinage d'un point).
2. Le nombre minimum de points requis pour former un cluster MinPts.

Trois types de points différents : un point central c'est le point qui a un nombre minimal de voisin (MinPts). Un point de frontières : c'est un point ou le nombre de ses voisins est inférieur à MinPts, et un point aberrante, c'est un point ni centrale ni frontière. **Voisinage :** Pour détecter les voisins d'un point P il faut vérifier les deux valeurs ε et MinPts, tel que x est un voisin de y si la distance entre eux est inférieurs à ε, et y doit avoir un nombre suffisants de points pour former un cluster respectant la valeur MinPts. Le pseudo code suivant présente les étapes de l'algorithme BDSCAN :

Algorithme 3 L'Algorithme BDSCAN

Entrée:
Les données D
La distance ε
nombre minimum de points MinPts.
Début
Pour chaque point P non visité des données D
Marquer P comme BRUIT
$VisitéPtsVoisins = epsilonVoisinage(D, P, \varepsilon)$
Si (tailleDe(PtsVoisins) $<$ MinPts)
Marquer P comme BRUIT
Sinon
C++
etendreCluster(D, P, PtsVoisins, C, eps, MinPts)
Ajouter P au cluster C
Fin Si
Pour chaque point P' de PtsVoisins
Si P' n'a pas été visité marquer P' comme visité
PtsVoisins' = epsilonVoisinage(D, P', ε)
Si tailleDe(PtsVoisins') $\geq$ MinPts
PtsVoisins = PtsVoisins $\cup$ PtsVoisins'
Fin Si
Si P' n'est membre d'aucun cluster
Ajouter P' au cluster C
Fin Si
Fin Pour
Fin Pour
epsilonVoisinage(D, P, ε)
Retourner tous les points de D qui sont à une distance inférieure à epsilon de P.
FIN

II.4.1.5 Algorithmes évolutionnaires

Toutes les instances de base d'algorithmes évolutifs partagent un certain nombre de propriétés communes, qui sont mentionnées ici pour caractériser le prototype d'un algorithme évolutif général : Les algorithmes évolutifs utilisent le processus d'apprentissage collectif d'une population d'individus. Habituellement, chaque individu représente (ou encode) un point de recherche dans l'espace de solutions potentielles à un problème donné.

1. Une AE utilise des mécanismes inspirés de l'évolution biologique, tels que la reproduction, la mutation, la recombinaison et la sélection.

2. Les solutions candidates au problème d'optimisation jouent le rôle des individus dans une population, et la fonction de conditionnement physique détermine la qualité des solutions. L'évolution de la population se fait ensuite après l'application répétée des opérateurs ci-dessus.

Algorithmes génétiques pour le clustering: Algorithmes génétiques (AG) [Dav91; Gol+89; Mic13], présenté par John Holland, sont des méthodes efficaces pour résoudre

de nombreux problèmes de recherche et d'optimisation. Les AG sont basées sur le principe de la génétique naturelle et la théorie évolutive des gènes. Les Algorithmes génétiques, sont populaires pour le clusering [BM02a; MB03; BM02b], pour utiliser les AG dans le but de clustering, il faut d'abord choisir un mode d'encodage approprié pour représenter une solution de clustering possible en tant que chromosome. Parmi les différentes approches de codage, deux sont largement utilisés : le codage par points et le codage centralisé.

Codage pa Lr points :

Dans les techniques de codage par points [Lu+04a; Lu+04b], la longueur d'un chromosome est identique au nombre de points est dessinée de 1, ..., K, K est le nombre de clusters, si le gène i est considéré comme valeur de j, alors le ième point est attribué au jème cluster. Les techniques de codage basées sur des points sont simples, mais souffrent de longues longueurs chromosomiques et donc des taux de convergence lents. Une population est donc un ensemble de chromosomes. Chaque chromosome code un point de l'espace de recherche. L'efficacité de l'algorithme génétique va donc dépendre du choix du codage d'un chromosome.

Codage centralisé:

Dans le codage centralisé [BM02a; MB03], les centres de grappe sont encodés dans les chromosomes. Par conséquent, chaque chromosome est de longueur $(K * d)$, où d est la dimension des données. Ici aussi, K peut être varié, ce qui entraîne des chromosomes à longueur variable. L'avantage de l'encodage basé sur le centre est que la longueur du chromosome n'est pas très grande, par conséquent, elle a généralement un taux de convergence plus rapide que les techniques de codage par points. Les algorithmes génétiques ont plusieurs applications dans différents domaines de fouille de données, tels que la télédétection [BMM07; MM09], la fouille de Web [OMV02; Pic+02], la fouille de données multimédia [BTT05; Chi+00], la fouille de texte [DGP05; AAMA04] et la fouille de données biologiques [Ban07; TA08]. La force des AG est continuellement explorée dans ces domaines. Différentes tâches de bio-informatique telles que l'analyse des séquences génétiques, la cartographie des gènes, le fragment d'acide désoxyribonucléique (ADN), la recherche de gènes, l'analyse des microarrays, l'analyse du réseau de régulation des gènes, la construction d'arbres phylogénétiques, la prédiction de la structure et l'analyse de l'ADN, de l'acide ribonucléique (ARN) et de la protéine , et l'ancrage moléculaire avec la conception de ligand ont été effectués en utilisant les AG.

II.4.2 Classification supervisée des données biologiques

La classification décide de la classe d'un échantillon de données non classé. Il doit y avoir au moins deux classes prédéfinies. Donc pour classifier un échantillon il faut définir ses attributs et le modèle de classification pour décider la classe à laquelle l'échantillon de données il appartient. L'ensemble de données devisé en deux parties, les données d'apprentissage pour construire le modèle de classification et les données de test sont utilisées pour évaluer la performance de ce modèle. L'une des principales tâches de la bio-informatique est la classification des données biologiques. Avec l'augmentation rapide de la taille des banques de données biologiques, il est essentiel d'utiliser des programmes informatiques pour automatiser le processus de classification. La classification des échantillons biologiques est une technique importante de fouille de don-nées, Surtout dans le contexte de prédiction de la catégorie des séquences protéique. Une question critique dans la classification des données biologique est le nombre limité

d'échantillons qui sont disponibles, il est donc difficile d'évaluer la signification statistique des résultats. De plus, le nombre élevé de données biologique pourrait introduire du bruit en ce qui concerne le modèle de classification, différents algorithmes de classification ont été proposés pour le traitement des données de biologique. Les algorithmes de classification les plus utilisés exploités dans l'analyse des données biologique appartiennent à quatre catégories : Random forest (RF), classificateurs bayésiens, réseaux de neurones et machines à vecteurs de support (SVM).

II.4.2.1 Random forest (RF) :

RF est largement utilisé dans les applications de la bio-informatique parce que les modèles de classification RF ont une haute précision de prédiction et peuvent fournir des informations supplémentaires concerne la fonctionnalité des gènes et protéines. RF est une évolution de l'arbre de décision exploité pour l'analyse des données biologique, un arbre de décision dérive de l'algorithme simple de division et de conquête (divide-and-conquer). Dans sa structure arborescente, les feuilles représentent des classes et les branches représentent des conjonctions de caractéristiques qui mènent à ces classes. A chaque nœud de l'arbre, l'attribut qui divise le plus efficacement les échantillons en différentes classes est choisi. ID3 [Qui86] Et C4.5[Qui93] sont les algorithmes les plus courants des arbres de décision. RF [Bre01] est un ensemble d'arbres de décision individuels, où chaque arbre dans la forêt construit avec un sous-ensemble aléatoire d'échantillons de données donc pour prédire la classe d'un nouvel échantillon de donnée, chaque arbre de décision dans la forêt lance un vote non pondéré pour l'échantillon, le vote majoritaire détermine la classe de l'échantillon. [DUDA06] ont montré la bonne performance de RF pour les données génomiques bruyantes et multi-classes. Ce type de modèle de classification est populaires, ont plusieurs applications en bio-informatique[EZ13] telles que l'annotation de séquence, la structure de protéines et la prédiction de fonction et les interactions protéine-protéine.

II.4.2.2 Machines à vecteurs de support (SVM)

Machines à vecteur de support, est un algorithme d'apprentissage, vise à rechercher l'hyperplan qui sépare le mieux les classes de données (une classe est définie comme la classe positive et une autre classe est appelée la classe négative)

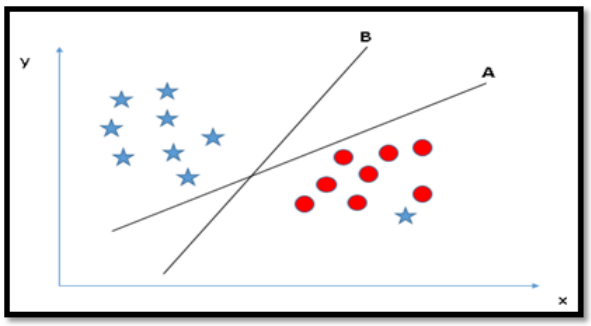

FIGURE II.4: Concepte de base de SVM

Le but de SVM est de trouver une fonction de classification qui soit capable de faire la distinction entre les échantillons des deux classes dans les données d'entraînement.

Géométriquement, si les échantillons de deux classes peuvent être séparés linéairement, la fonction de classification f (x) est linéaire correspond à un hyperplan de séparation passant par le milieu des deux classes, une fois cette fonction générée, un nouvel échantillon de données Xn peut être classé en testant simplement le signe de la fonction. Autrement dit, le nouvel échantillon de données sera affecté à la classe positive si f (x)> 0. Sinon, cet échantillon sera affecté à la classe négative. Les SVM ont été appliquées à un certain nombre de tâches qui impliquent, la classification des gènes de levure en catégories fonctionnelles par [Bro+00], décrivent l'application des SVM à la reconnaissance des tissus du cancer du côlon.) [MCM00] l'application des SVM au problème de la reconnaissance des ARN fonctionnels dans l'ADN génomique,[Seg+03] utilisent la SVM pour développer un schéma de classification basé sur le génome pour le sarcome à cellules claires, et plusieurs d'autre tâches[Nob+04].

II.4.2.3 Réseau de Neurones Artificiels (RNA)

Un réseau de neurones artificiels est un modèle mathématique basé sur des réseaux de neurones biologiques(ANN en anglais), il se compose d'un groupe interconnecté de neurones (unités) artificiels qui sont organisées en couches. La couche d'entrée se compose simplement des données d'origine, et les nœuds de couche de sortie représentent les classes. Généralement, seules les unités appartient à deux couches consécutives sont connectées. Une unité reçoit des informations provenant de plusieurs unités appartenant à la couche précédente, ces unités sont reliées par des liens et chaque lien a un poids. Les avantages des réseaux neurones incluent leur tolérance élevée aux données bruyantes. Un modèle simple d'un réseau de neurone est montré par figure II.5.

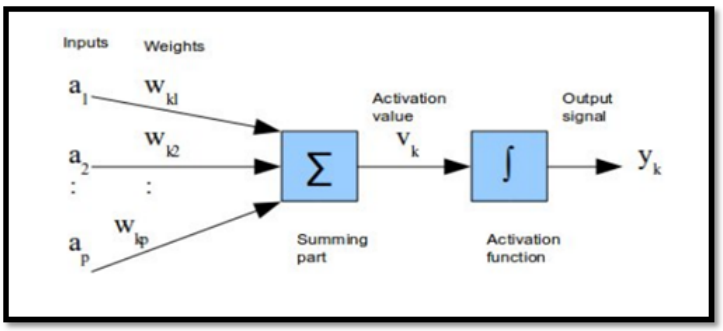

Figure II.5: Simple modèle de réseau de neurone

Les ANNS ont été utilisés dans nombreuses cas pour résoudre des problèmes bio-informatique et se sont révélés très efficaces. Dans [LLB09], une revue des avantages et des inconvénients des réseaux neuronaux dans le contexte de l'analyse des information biologique représenté sous format microarray. [BFL05] ont présenté un système de classement protéique multi-classe basé sur des réseaux neuronaux, des ANNS ont également été appliqués pour prédire la structure tertiaire des protéines, telle que, [DS12] ont proposé une technique basé sur RNA pour la prédiction de la structure secondaire de protéine, et aussi[Ple+08] ont développé une méthode basée sur le réseau neuronal pour la détection de peptides signal dans les protéines.

II.4.2.4 Classificateurs bayésiens

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des hypothèses. il appartenant à la famille des classifieurs linéaires. En termes simples, un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. c'est à dire:

$$P(X \setminus C) = \sum_{i=1}^{n} P(X_i \setminus C)) \tag{II.4}$$

Où $X = (X1, ..., Xn)$ et C est une classe. Cette probabilité est considérée comme la probabilité a posteriori de la classe compte tenu des données, est habituellement calculé en utilisant le théorème de Bayes. L'estimation de cette distribution de probabilités à partir d'un ensemble de données d'apprentissage est un problème difficile, car elle peut nécessiter un très grand ensemble de données pour explorer de manière significative toutes les combinaisons possibles. Le classificateur Naïve Bayes est connu pour être une méthode robuste, qui montre une bonne performance en termes de la précision de classification. [VD14] a fourni la technique pour la classification des séquences de protéines de maladies en utilisant la méthode bayésienne naïve.

II.4.2.5 Autre Algorithmes de classification

D'autres algorithmes de classification ont été proposé, l'approche K-NN pour la classification a été populaire dans la Biologie [DK02; DS12]. Étant donné une base de données de séquences pré-classées, donc pour classer une nouvelle séquence en cherchant les k séquences similaire dans la base de données. La classe qui contient la majorité de ces k séquences est la classe attribuée à la nouvelle séquence non classée. Les approches basées sur le HMM [Rab89] pour la classification des séquences de protéines ont démontré leur efficacité pour la détection de motifs de résidus conservés dans un ensemble de séquences de protéines [Edd+95; HK96]. Et plusieurs d'autres applications des méthodes de classification de fouille de données appliquées sur les données biologiques [EZ13].

II.4.3 Extraction des Règles d'association à partir des données biologiques

L'extraction des règles d'association est une autre tâche importante de fouille de données consiste à découvrir des relations intéressantes entre les attributs des grandes bases de données. L'objectif de cette analyse est de fournir au décideur des connaissances précieuses sur un certain domaine modélisé par une base de données de transactions. Les règles d'association ont été d'abord introduites pour l'analyse du panier de marché pour représenter les ensembles d'articles susceptibles d'être achetés ensemble. Chaque transaction contient un seul prédicat ("achat") avec plusieurs occurrences. Généralement, le nombre de règles d'association croît de façon rapide avec le nombre de transactions et d'attributs, il devient très difficile de les lire et de les interpréter. Les règles induites peuvent être classées en trois catégories :

1. **Règles utiles :** qu'un expert humain peut comprendre et utiliser.

2. **Règles triviales :** les règles qui sont valides mais jamais utilisés.

3. **Règles faibles :** Les règles qui ne sont pas acceptables par l'expert et ne sont pas compréhensibles.

II.4.3.1 Définition et principe

Dans sa version la plus commune, une règle d'association est caractérisé comme suit $A = a1, a2, ..., am$ ensemble d'items, et $T = i1, i2, ..., in$ un ensemble de transactions, Chaque transaction est associé à un sous-ensemble de A, une règle d'association est définie par : X $\rightarrow$ Y, dans laquelle X, Y $\subseteq$ A et X$\cap$Y = $\oslash$. Il existe de nombreuses mesures qui ont été proposées pour évaluer une règle intéressante, le support et la confiance reflètent respectivement l'utilité et la certitude des règles découvertes. Ils sont définis comme suit : Le support : Soit X $\rightarrow$ Y une règle d'association. Le support de X $\rightarrow$ Y, est la fraction des transactions dans la base de données qui contient X $\cup$ Y. $Sup(X \rightarrow Y) = \frac{N_{xy}}{N}$ Où N est le nombre de tuple dans la base de données relationnelle et X_{XY} le nombre de tuples qui contiennent l'ensemble des éléments X $\cup$ Y. La confiance : Soit X $\rightarrow$ Y une règle d'association. La confiance de X $\rightarrow$Y est la faction des transactions dans la base de données qui contient X $\cup$ Y sur le nombre de transactions qui ne contiennent que X.
$Conf(X \rightarrow Y) = \frac{Sup(X \cup Y)}{Sup(X)} = \frac{N_{xy}}{N_z}$ Où X_{XY} est le nombre de tuples qui contiennent l'ensemble des éléments X $\rightarrow$Y et N_X le nombre de tuples qui contiennent X. Le problème d'extraction des règles d'association est de générer toutes les règles d'association qui ont un support et une confiance supérieurs au seuil minimal de support (minsup) et au seuil minimal de confiance (minconf), spécifié par l'utilisateur. Par exemple, la règle AB $\rightarrow$C, le support = 20%, et la confiance = 60%indique que lorsque A et B se produisent, C se produit également dans 60%des cas, et que les trois événements se produisent en même temps dans 20% de tous les cas. L'utilisateur fixe un seuil minimal de support et un seuil minimal de confiance, selon ces seuils la sélection des items fréquents et la génération des règles d'association sont effectués. L'analyse d'association est également applicable à plusieurs domaines, y compris la bio-informatique, le diagnostic médical, le web mining et l'analyse de données scientifiques etc. Un domaine d'application pratique dans la biologie est une donnée d'expression génétique, qui comprend un grand nombre d'attributs (gènes), et les associations entre différents gènes sont souvent particulièrement importantes. Dans l'exemple de [EZ13] suivant, nous présentons l'extraction des associations entres les données génétique :

Transaction ID	Gene 1	Gene 2	Gene 3	Gene 4
T1	0	1	0	0
T2	0	0	0	1
T3	1	1	0	0
T4	0	0	1	0
T5	1	1	1	0

TABLE II.2: Exemple de matrice d'expression génétique binaire

Dans cet exemple, la base de données se compose de quatre gènes différents
$I = gene1, gene2, gene3, gene4$.
Il existe cinq transactions. Un exemple d'une règle pour cette base de données pourrait

être gene1, gene2 ⇒ gene3. Le support de cette règle est 1/5 (= 20%), car une seule transaction (T5) sur cinq contient gene1, gene2 et gene3. La confiance de la règle est 1/2(= 50%), car le support de l'itemset gene1, gene2 est 2/5 et le support de itemset gene1, gene2, gene3 est 1/5. Le problème d'exploitation des Règles d'association peut être décomposé de deux sous-tâches principales : sélection des éléments fréquents : Dans le but de trouver tous les ensembles d'éléments qui satisfont un seuil minimal de support spécifié par l'utilisateur (minsup). Ces objets sont appelés articles fréquents. Génération de règles : Dans le but d'extraire des objets fréquents toutes les règles qui satisfont un seuil minimal de confiance (minconf)spécifié par l'utilisateur.

II.4.3.2 Algorithmes d'extraction des règles d'association

- **L'algorithme Apriori**
  Apriori [AS+94] sert à construire des itemsets candidats fréquents et vérifier ensuite lesquels d'entre eux sont effectivement fréquents, Pour la génération de l'ensemble de k-itemsets candidats fréquents, l'ensemble des (k - 1) itemsets candidats fréquents est exploité, le processus d'extraction des items fréquents présente en deux étapes :

  1. L'ensemble des itemset candidats Ci est généré.
  2. Calculer les fréquences des itemsets générés.
  3. Garder les itemset avec la valeur de support supérieur au seuil minimale de support (minsup).
  4. Garder les règles avec une valeur de confiance supérieure au seuil minimale de confiance (minConf).
  5. **L'algorithme FP-Growth**

Algorithme 4 L'Algorithme Apriori

Début
Entrée
Base de données D
Seuil minimum de support min_s
Seuil minimum de confiance min_c
Sortie
Ensemble de règles d'association R
$F1 = 1 - Itemsets fréquents$
Pour$(k = 2; Fk - 1 \neq \oslash; k++) faire$
$Ck = Apriori_g en(Fk - 1)$
Pour(chaque transaction t de D)faire
Pour (chaque candidat c de C_K) faire
Si (c est inclus dans t) alors
c.support++ ;
Finsi
Finpour
Finpour
Fk =c $\in (C_K / \ c.support \geqslant min_s$
Finpour
$F = U_K F_K$
$R = Generer_R egles(F)$
FIN

FP-GROWTH a été proposé par Han et al [HPY00] est appliqué sans la génération de candidats. Le FP-Growth adopte une stratégie de division-et-conquête pour découvrir tous les items fréquents, basant sur la structure de données arborescentes. FP-Growth comporte les étapes suivantes :

Algorithme 5 L'Algorithme FP-Growth

Début
La construction de l'arbre FP-Tree
Calculer la valeur de support de chaque transaction
Calculer la fréquence et la priorité de chaque item
Ordonner les items selon la priorité, on obtient la base des transactions
Créer l'arbre FP-Tree.
Créer la racine de l'arbre, l'étiqueter comme "NULL", et pour chaque transaction on créer un sous-arbre jusqu'à la dernière transaction
L'extraction des itemset fréquents à partir de l'arbre FP-Tree
L'ensemble des items fréquents sont extraites directement à partir du FP-Tree en utilisant la méthode de partage et conquête (divide-and-conque)
Fin

- **L'Algorithme de GenMiner**
  D'autre type d'algorithme d'extraction de règles d'association récent basant sur le traitement particulier des données génomique, le plus connue est GenMiner [MPP08], GenMiner est une implémentation de la découverte de règles d'association

dédiée à l'analyse des données génomiques. Il permet l'analyse des ensembles de données biologiques de multiples sources, représentées comme deux valeurs discrètes, telles que des annotations de gènes, et des valeurs continues. Met en œuvre l'algorithme NorDi pour la normalisation et la discrétisation des valeurs discrètes, L'algorithme de discrétisation normale (NorDi) [MPP07] a été développé pour améliorer la discrétisation des mesures d'expression génétique en items, cette phase est essentielle pour extraire les règles d'association pertinentes. Cet algorithme détecte d'abord les valeurs aberrantes, à l'aide de la méthode des valeurs aberrantes de Grubbs et du test de normalité de Jarque-Bera, puis supprimer ces valeurs. GenMiner Profite les avantages de l'algorithme Close pour générer efficacement les règles d'association non redondantes avec un faible support et une haute confiance. Close est une approche basée sur des itemsets fermés fréquents [Pas+05] utilisant une sémantique basée sur l'opérateur de fermeture de connexion de Galois pour extraire une RA minimale non redondante sans perte d'information.

II.4.3.3 Application des RAs dans la bio-informatique

En bio-informatique, les ARs ont été utilisé pour analyser un ensemble assez variable de données biologiques, les données d'expression génétique, les séquences biologiques, les données structurales biologiques, les réseaux d'interactions protéiques etc. Dans [Par+09], L'analyse d'association été utilisé pour prédire les fonctions protéiques à partir d'un réseau d'interactions protéiques. [ML98] ont proposé un algorithme de classification de séquences biologiques pour prédire la structure secondaire, Autre approche de prédiction de la structure secondaire a été proposée par Birzele et Kramer [BK06]. Basé sur l'extraction des motifs fréquents et la machine à vecteurs de support (SVM).

II.4.4 Fouille de textes biologiques (FTB)

II.4.4.1 Définition 1

Texte mining définie comme un processus à forte intensité de connaissances dans lequel un utilisateur interagit avec une collection de documents par l'utilisation d'une suite d'outils d'analyse. L'objectif principal de fouille de texte est de récupérer des connaissances cachées dans les textes et de les présenter sous une forme concise et simple aux utilisateurs. Deux directions principales de texte mining peuvent être définies :

- **Extraction d'information:** Selon [AZ12] L'extraction d'informations à partir du texte est une tâche importante dans la fouille de texte. Le but général de l'extraction d'information est de découvrir des informations structurées à partir de textes non structurés ou semi-structurés. A été largement étudié dans diverses communautés de recherche, y compris le traitement du langage naturel, la recherche d'information et le Web mining.

- **Recherche d'information:** La recherche d'information consiste à trouver des documents de nature non-structurée (habituellement du texte) qui satisfont un besoin d'information à partir d'une grande collection de données par l'application d'un ensemble de méthodes, procédures et opérations.

II.4.4.2 Définition 2

La Fouille de texte en biologie moléculaire, définie comme l'extraction automatique d'informations caché dans les gènes, les protéines et leurs relations fonctionnelles à partir de documents textuelle [KV05] Les chercheurs ont besoin d'exploiter l'énorme volume d'informations biologiques, ainsi que la disponibilité d'outils de récupération de données et de fouille de données performants et efficaces, a donné la naissance à un nouveau domaine de recherche et d'application appelée fouille de texte bio-informatique (FTB). D'autres termes pour FTB sont fouille de texte biologique (logique) et fouille de texte biomédicale. Plusieurs études ont catégorisé les tâches de FTB de différents points de vue. Cohen et Hersh [CH05] fournissent une catégorisation de haut niveau identifiant les tâches principales comme suit :

- **Reconnaissance de l'entité nommée :** L'objectif est de trouver et classer les éléments atomiques dans le texte dans des catégories prédéfinies.
- **Classification textuelle :** L'objectif est de déterminer si un document a des caractéristiques particulières, habituellement il inclut certains types d'informations.
- **Extraction de Synonyme et d'abréviation :** Cette tâche traite du problème que de nombreuses entités biologiques ont plusieurs noms, donc dans la littérature biomédicale il y a beaucoup de synonymes et d'abréviations.
- **Extraction de relations :** Le but d'extractions de relations est de trouver un type spécifique de relation entre une paire d'entités de types donnés.
- **Génération d'hypothèses :** L'objectif est de trouver des relations qui ne sont pas présentes dans le texte mais qui sont déduites par d'autres relations explicites.

II.4.4.3 Notions de base

- **Corpus :** Un corpus en biologie moléculaire est un ensemble de documents textuels utilisé dans le développement et l'évaluation d'un problème donnés. Un corpus permet d'obtenir des résultats de performance, qui pourraient être utilisés pour comparer des solutions distinctes au même problème. Il existe plusieurs corpus accessibles au public qui peuvent être utilisés, tels que GENETAG [Tan+05], GENIA [Kim+03], PennBioIE [Coh05] et BioNLP-Corpora [Joh+07].
- **Tokenization :** Il est nécessaire de diviser les textes en langage naturel en unités significatives, appelées tokens. Un token est un groupe de caractères qui est catégorisé selon un ensemble de règles. Le processus de fragmentation d'un texte en ses tokens constitutifs est connu sous le nom de tokenisation. C'est l'une des tâches les plus importantes dans la fouille de texte. Ainsi, plusieurs solutions de tokenization ont été développées pour plusieurs domaines et langues. Par exemple, Open NLP [4] a des modèles pour les documents biomédicaux dans plusieurs langues, et SPECIALIST NLP [BMS00] supporte également le texte biomédical.

[4]Open NLP est disponible dans: http://opennlp.sourceforge.net, consulté le: 2017-08-24

- **Normalisation des mots :** Dans la plupart des cas, les variantes morphologiques des mots ont des interprétations sémantiques semblables, qui peuvent être considérées comme équivalentes. Pour cette raison, plusieurs solutions sont utilisées pour regrouper les différentes formes ambigu d'un mot.7645 Afin qu'ils puissent être analysés comme un seul élément. Pendant ce processus, les termes des textes en langage naturel sont représentés par des mots principaux plutôt que par les mots originaux. Deux approches distinctes de ce problème sont utilisées, stemming et lemmatisation. Le stemming est basé sur le principe d'association de préfixes et de suffixes à la racine d'un mot. Le stemming est utilisé par les moteurs de recherche afin d'élargir les recherches et de proposer des résultats plus complets. D'autre part, la lemmatisation est une méthode plus robuste, car elle trouve le terme racine du mot variant.

- **Stopwords :** L'une des techniques les plus utilisées est la suppression de mots qui sont déjà connus pour être non informatif aux processus de reconnaissance et de normalisation, et pour Réduire le dictionnaire et les tailles de corpus. [Int13]

- **Evaluation :** Après le développement du système, il est nécessaire de calculer des mesures qui fournissent l'effet précis et global de l'application. Pour obtenir des mesures, les prédictions doivent être classées comme suit:
  Vrai Positive (VP): Le nombre de séquences attribués à une catégorie convenablement (Séquences attribués à leurs vraies catégories.
  Vrai Négative (FP): Le nombre de séquences attribués à une catégorie inconvenablement (Séquences attribués à des mauvaises catégories)
  Faux Positive (FN): Le nombre de séquences inconvenablement non attribués (Qui auraient dû être attribués à une catégorie mais qui ne l'ont pas été).
  Faux Négative (VN): Le nombre de séquences non attribués à une catégorie convenablement (Qui n'ont pas à être attribués à une catégorie, et ne l'ont pas été)
  Après avoir obtenu le nombre de prédictions correctes et erronées, trois mesures sont couramment utilisées pour refléter le comportement du système : La précision, le rappel et la F-mesure. Ces mesures peuvent prendre des valeurs entre 0 (faible performance) et 1 (meilleure performance). La précision mesure la capacité d'un système à ne présenter que des éléments pertinents, le rappel mesure la capacité d'un système à présenter tous les éléments pertinents, et la F-mesure est la moyenne harmonique de la précision et du rappel.

II.4.4.4 Applications de FTB

Allant du monde des données biologiques, la fouille de texte fournit le moyen de calcul pour naviguer dans le vaste monde de la connaissance biologique, actuellement encapsulé par la littérature scientifique non structurée et toujours croissante. Tandis que les méthodes de fouille de texte se sont concentrées en grande partie sur la non-structuration des textes dans la littérature scientifique. La fouille de texte peut être incorporée avec des approches d'apprentissage par machine pour obtenir de meilleures performances à l'extraction de données biologiques textuelles. Plusieurs thématiques de bio-informatique ont été traitées en tant qu'un problème de fouille de texte. La prédiction de la fonction [GH04; HR00], l'annotation du génome [WKG00; Huy+98], et même la prédiction de la structure [MBJ00] :

- Deux groupes [BLM04; CRA00] ont récemment tenté d'améliorer la recherche d'homologie, en appliquant une analyse de texte.
- Le programme PSI-BLAST est une version itérative de programme original BLAST, BLAST pouvait trouver que des alignements locaux sans lacunes, comme le suggère son nom, PSI-BLAST réduit la sensibilité des recherches d'homologie, de sorte qu'une similitude éloignée avec une séquence de requête peut être détectée. [MKS00] Étaient l'un des premiers groupes à incorporer l'exploration de texte avec PSI-BLAST pour des recherches d'homologie séquentielles améliorées.
- Détection de corrélation significative entre la composition d'acides aminés d'une protéine et sa localisation sous-cellulaire [Ced+97; NO82]. La localisation sous-cellulaire d'une protéine peut être prédite avec une précision raisonnable à partir de sa composition en acides aminés [RH98; NK92].
- [SKS01] A tenté d'améliorer encore le problème de localisation sous-cellulaire de protéine en intégrant l'exploration de texte. Ils ont expérimenté la prédiction de la localisation sous-cellulaire pour les protéines de levure, en fonction de la séquence et de l'information de la littérature sur les protéines.
- Dans certaines applications, La base de données MEDLINE (Medical Literature Analysis and Retrieval System Ondine) est analysée pour les résumés de la littérature qui mentionnent le nom de la protéine ou un synonyme. Après avoir appliqué l'enlèvement nécessaire des stop words, la suppression et l'élimination des termes dérivés qui se sont produits dans quelques documents
- Et plusieurs d'autres méthodes de fouille de texte intégrants pour le traitement de problèmes de bio-informatique [AM06].

II.5 Conclusion

Sans aucun doute, la fouille de données est la pierre angulaire de la bio-informatique moderne. Au cours des dernières années, les techniques de fouille de données ont été appliquées avec succès pour résoudre de nombreux problèmes critiques dans les sciences de la vie. Nous avons présenté un état de l'art de plusieurs techniques pour illustrer comment modéliser un véritable problème de bio-informatique en tant que problème de fouille données, en employant des algorithmes existants ou en développant de nouveaux algorithmes. Notez qu'il est impossible de couvrir tous les problèmes de bio-informatique dans un seul chapitre. Ce chapitre est présenté les travaux les plus connus et qui nous intéresse dans nos contributions. Ce chapitre a été devisé en deux importantes parties : la définition, l'historique et les problèmes de bio-informatique, et le processus ECD avec les méthodes de fouilles de données biologiques. Dans notre travail, nous avons choisi trois problèmes majeurs de la bio-informatique, et nous décrirons dans le chapitre suivant nos méthodologies dans le but de traiter l'information biologique complexe par les méthodes d'analyse et de fouille de données pour résoudre ces trois problèmes.

Chapitre III

Modélisation des Données Biologiques Complexes

III.1 Introduction

Des quantités énormes de données de différents types et natures ont été accumulées à un rythme très rapide, les développements massifs de la biologie moléculaire complexe ADN/PROTEIN/ARN, ont connu une croissance rapide du volume. L'analyse de ce type de données est une tâche difficile en raison de sa complexité et de ses multiples facteurs corrélés, et aussi en raison de l'évolution continue de notre compréhension des mécanismes biologiques, qui nécessite d'élaborer des approches de haute performance et d'apporter de nouvelles idées adaptés à ce type de données. Plusieurs techniques de bio-informatique ont été proposées pour résoudre plusieurs problèmes de la biologie moléculaire. L'ECD biologique et plus précisément, les techniques d'analyse et de fouille de données ont vu une large applicabilité dans le domaine de la recherche en biologie et en bio-informatique, aidant à accélérer et à approfondir la recherche dans la biologie moléculaire moderne.
Dans les deux premiers chapitres, nous avons présenté un état de l'art à propos de la complexité de l'information biologique et sa diversité, et les principaux problèmes de la bio-informatique et techniques d'analyse et de fouille de données biologiques. Cependant, ce chapitre représente la partie conceptuelle de nos travaux qui visent à améliorer la qualité de réponse aux deux problèmes de la bio-informatique traitant la donnée biologique dans son format primaire : d'un part l'étude de comparaison et similarité et d'autre part l'extraction de la connaissance biologique par des méthodologies inspirées du processus ECD biologique.
Ce chapitre, se compose de trois parties importantes ; la première représente le type de représentation abordée, la deuxième traite le problème de comparaison des séquences ADN et protéine et la troisième partie est consacrée aux deux problèmes importants de bio-informatique (1) la prédiction des structures moléculaires pour le développement des médicaments et (2) la classification supervisée des séquences protéiques, et à la fin on conclurons avec une conclusion.

III.2 Représentation de l'information biologique

Les données biologiques stockées dans les banques de données biologiques, sont différentes en termes de nature et de la façon de représentation. L'objectif de traitement de données biologiques est le facteur essentiel de choisir le type de données traitées,

dans notre thèse et comme nous l'avons noté dans le premier chapitre, nous avons choisi les données sous un format primaire de base dans le but de prédire la fonction des données biologiques(ADN/Protién), prend en considération leurs complexités. Une séquence d'ADN à la représentation d'une chaîne d'alphabet de quatre lettres N = A, C, G, T (quatre bases : adénine, cytosine, 'guanine et thymine), les séquences de protéines à la représentation des chaînes sur l'alphabet à 20 lettres des symboles d'acides aminés: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y.
Cependant, Quel que soit le problème biologique lié aux (ADN/Protéine), leur représentation primaire est considérée comme un paramètre important pour résoudre ce problème, car cette représentation(structure primaire) est le base de toute autre structure. D'un autre côté, notre choix à été effectué sur ce type de représentation dû au fait que c'est la seule structure qui largement garantit le sens de l'information biologique(moins perte d'information).
le traitement et l'analyse de ce type de donnée est compté sur beaucoup de techniques et outils. Dans la suite nous présenterons la conception de nos travaux de thèse.

III.3 Etude de similarité entre les séquences biologiques(ADN/Protéine)

La comparaison des séquences est un instrument fondamental pour comprendre les origines et les fonctions des séquences biologiques. Dans les organismes vivants, la diversité génétique est obtenue par des modifications de séquences préexistantes plutôt que par une création de novo. Plusieurs outils et techniques d'étude de similarité ont été proposés, permettant de calculer le degré de similarité entre les séquences de structure inconnue appelée « séquence de requête», et les séquences dont la structure et la fonction connues préalablement «séquences de référence», Une telle recherche peut entraîner un grand nombre de correspondances entre la séquence de requête et les séquences de référence. Sachant que la similarité de la séquence requête avec la séquence de référence ne signifié pas toujours la relation et l'homologie de leur fonction biologique, mais elle est considérée comme une preuve forte et constituant la base pour analyser et prédire la signification des séquences inconnues. Pour cette raison, un effort de recherche est en cours dans le développement de nouvelles approches automatisées pour étudier la similarité et la dissimilarité des séquences biologiques (ADN/ARN/Protéine).
Comme nous l'avons mentionné dans le deuxième chapitre, grande diversité d'approches consacrées à la comparaison des séquences biologiques de bases (similarité/ dissimilarité) ont été proposées :

- Les approches basées sur la notion d'alignement de séquence (alignement local et alignement global).
- Les approches classiques de programmation dynamique pour aligner de manière optimale deux séquences.
- Les approches heuristiques, moins précises, mais plus rapides et donc mieux adaptées aux comparaisons dans de grandes bases de données.

L'application de ces différentes approches en terme de séquencage, fréquence et alignement des composants de base, donne des résultats de similarité différente entre les séquences testées (selon l'efficacité de chacune), donc les séquences qui partagent

habituellement des similitudes importantes montrent que les gènes partagent des fonctions similaires ou un ancêtre commun au niveau de la séquence primaire (nucléotide ou acide aminé), permettant ainsi la compréhension de l'évolution des espéces.
comme nous l'avons mentionné dans la partie(I.4) du premier chapitre, que la complexité de la séquence moléculaire(ADN ou protiéne) est défini sous plusieurs champs, donc l'analyse et la compréhension de cette complexité aide a comprendre les relations qui ont existé entre ces séquences et la détection ou la prédiction des fonctions biologiques des espèces et leurs homogénéités.
Dans ce contexte, nous avons proposé deux méthodes traitant le problème de comparaison de séquences dont l'objectif est de détecter les homologies des séquences d'ADN et protéines cachées, en se basant sur les propriétés biologiques (chimique et physique) de la structure de base (nucléotidique ou acide aminée), qui fournissent des informations supplémentaires et contribuent à améliorer la recherche de similarité. Une séquence d'ADN ou protéine peut avoir des propriétés biologiques similaires à une autre séquence, même si le séquençage est différent. par conséquent; notre contribution combine trois points importants (1) la transformations des séquences moléculaires en terme de propriétés chimique et physique (2) le calcule de la fréquences d'apparition des composants de base (3) le calcule de la position moyenne des composants de base.

III.3.1 Analyse de similarité des séquences d'ADN

Il est devenu un problème difficile d'obtenir l'information directement de la séquence primaire d'ADN, de trouver un moyen efficace pour analyser un fragment de séquence correspondant aux fonctions génétiques, de comparer une séquence avec d'autres séquences etc. Donc les aspects de l'analyse statistique, mathématique et informatique doivent être combinés pour savoir manipuler ce type de données et gérer les ambiguïtés d'une base nucléotidique donnée dans une séquence ADN avec une position quelconque.

III.3.1.1 Transformation

Pour cela ; Nous avons proposé une approche de comparaison de séquence basée sur la transformation de la chaîne symbolique en d'autres chaînes décrivant les propriétés chimiques des bases nucléotidiques.
Pour définir les propriétés chimiques des nucléotides nous avons fait appel à la notation UIPAC [1] (L'Union internationale de chimie pure et appliquée) [CB85], c'est l'autorité reconnue pour le développement de règles à adopter pour la nomenclature, les symboles et la terminologie des éléments chimiques et de leurs dérivés, L'IUPAC dispose également d'un système de codage pour identifier les bases nucléotidiques et les acides aminés. Ces codes peuvent comporter un code à une lettre ou un code à trois lettres, facilitent et réduisent les séquences d'acides aminés qui constituent des protéines. Les bases nucléotidiques sont constituées de purines (adénine et guanine) et de pyrimidines (cytosine et thymine ou uracile). Ces bases nucléotidiques constituent l'ADN et l'ARN. Ces codes de base de nucléotide rendent le génome d'un organisme beaucoup plus petit et plus facile à lire et aussi sert à gérer l'ambiguïté à une position donnée dans une séquence nucléotidique, le tableau suivant présente le code UIPAC des nucléotides.

[1] UIPAC: est disponible dans https://iupac.org/, consulté le 2017-06-03.

Code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil

Code	Base
R	A or G (Purines)
Y	C or T (Pyrimidines)
S	G or C
W	A or T
K	G or T
M	A or C

Code	Base
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	Any base

TABLE III.1: Codes de nucléotide IUPAC

Le deuxième sous tableau d'IUPAC, montre que les quatre bases d'ADN (A, C, G, T) peuvent constituer trois classes selon leur propriété chimique, la classification RY, la classification MK et la classification WS :

- Structures chimiques des bases (RY) : Le groupe purine $R = A, G$ et le groupe pyrimidine $Y = C, T$;
- Groupes fonctionnels des bases (MK) : Groupe aMino $M = A, C$ et groupe keto $K = G, T$;
- La force des liens H entre les bases (WS) : Le groupe H de faible liaison $W = A, T$ et le groupe H de forte liaison $S = C, G$.

Alors, selon ces trois classifications la transformation d'une séquence ADN de base nucléotidique a trois séquences symboliques est effectuée, la séquence primaire $X = S1S2S3.\ldots\ldots Sn.$ avec la longueur n, est présenté par ΦRY, ΦMK,ΦWS
$\Phi RY(X) = \Phi RY\,(S1)\,\Phi RY\,(S2).....\Phi RY(Sn)$
$\Phi MK(X) = \Phi MK\,(S1)\,\Phi MK\,(S2).....\Phi MK(Sn)$
$\Phi WS(X) = \Phi WS\,(S1)\,\Phi WS\,(S2).....\Phi WS(Sn)$

$$\Phi_{RY}\ (S_i) = \begin{cases} R\ if(S_i) \in R \\ Y\ if(S_i) \in Y \end{cases} \quad i = 1, 2....., n \tag{III.1}$$

$$\Phi_{MK}\ (S_i) = \begin{cases} M\ if(S_i) \in M \\ K\ if(S_i) \in K \end{cases} \quad i = 1, 2....., n \tag{III.2}$$

$$\Phi_{WS}\ (S_i) = \begin{cases} W\ if(S_i) \in W \\ S\ if(S_i) \in S \end{cases} \quad i = 1, 2....., n \tag{III.3}$$

Donc chaque séquence d'ADN utilisée pour l'étude de comparaison est représentée par les trois séquences symboliques selon les trois formules ci-dessus. Pour un fragment de longueur 14 de la séquence de bêta globine de l'humain $BH = \acute{}ATGGTGCACCTGAC\check{}$, les trois séquences symboliques qui peuvent représenter les propriétés chimiques de la séquence BH sont :
$\Phi RY(BH) = RYRRYRYRYYYRRY.$
$\Phi MK(BH) = MKKKKKMMMMMKKMM.$
$\Phi WS(BH) = WWSSWSSWSSWSWS.$

III.3.1.2 Calcul de fréquence et position

Pour chaque séquence obtenue, nous nous concentrons sur l'information du groupe de mutations, pour les trois séquences symboliques, il y a douze mutations possibles:

$R \to R$, $R \to Y$, $Y \to R$, $Y \to Y$, $M \to M$, $M \to K$, $K \to M$, $K \to K$, $W \to W$, $S \to W$, $W \to S$, $S \to S$.
Il est devenu difficile de calculer la similitude des séquences d'ADN sans invoquer certaines statistiques et mathématiques et à travers ces mutations, nous avons combiné deux notions importantes, la fréquence et la position des composants. Nous avons calculé la fréquence de chaque information de mutation définie par la formule de [SH12] suivante :

$$f_{UV} = \frac{nomre\,de\,mot\,UV}{n-1} \tag{III.4}$$

UV: la mutation pour les trois classifications. Pour la classification RY, les fréquences notées f_{ry}, f_{yr} , f_{rr}, f_{yy}, les fréquences de la classification MK noté par : f_{mk}, f_{km}, f_{mm} f_{kk} et pour la classification WS les fréquences indiquées par: f_{ws} , f_{sw}, f_{ss}, f_{ww}.
Par conséquent, si nous calculons la similarité de deux séquences en terme de la fréquence de ses composants, et si les deux séquences ont la même fréquence de composants mais dans deux directions de séquençage différents, Nous les obtenons identiques mais la position de leurs composants est complètement différente, qui peuvent prouver réellement qu'il n'y a pas de relation biologique entre eux. Donc, l'information insuffisante dans un vecteur de représentation de séquence basée sur la fréquence de composants est une raison importante qui provoque de faibles résultats de similarité, Pour cette raison et pour améliorer la qualité d'étude de similarité entre les séquences d'ADN et traiter la complexité des séquences biologiques (ADN), qui a été considéré comme un défi majeur dans le domaine de comparaison de séquences en bioinformatique. Nous avons introduit la notion de position des bases nucléotidiques dans une séquence ADN ; Nous nous basons sur les mutations présentées ci-dessus pour calculer leurs positions moyennes, Nous avons proposé la formule suivante :

$$P_{UV} = \frac{(\sum_{i=0}^{k}(Position\,UV))}{k*(n-1)} \tag{III.5}$$

K: Le nombre de mots UV.
n: Le nombre des composants de la séquence d'ADN.

III.3.1.3 Analyse de similarité

Chaque séquence représentée par deux vecteurs de douze composants, un vecteur représente la fréquence des douze mutations et un autre représente la position moyenne de ces derniers. La similarité entre deux séquences d'ADN est calculée par la distance euclidienne entre les vecteurs de fréquence (SF) et entre les vecteurs de position moyenne (SP), telle qu'une plus petite valeur de distance euclidienne signifiant une grande similarité entre deux séquences d'ADN. Cependant, nous proposons la formule suivante pour calculer la distance moyenne entre deux séquences d'ADN :

$$S = \frac{SF_{(S1,S2)} + SP_{(S1,S2)}}{2} \tag{III.6}$$

SF définit la distance euclidienne entre deux séquences en terme de la fréquence des composants (vecteurs de fréquence à 12 composants), SP définit la distance euclidienne entre deux séquences en terme de la position des composants (vecteurs de position moyenne à 12 composants). Donc, la similarité (S)entre deux séquences d'ADN est définie par la valeur moyenne de la similarité de fréquence des mutations et la

position moyenne des mutations. Pour l'exemple de fragment de la séquence de béta globine **HB**, à partir des trois séquences, nous obtenons deux vecteurs de douze composants, le premier représente la fréquence des mutations et le second représente la position moyenne. Pour le mot RR, sa fréquence est calculée selon la formule (III.4) : $F(RR) = 2/((14-1)) = 0.15$ et sa position moyenne est calculé par : $P(RR) = (((3+12)/2))/((14-1)) = 0.57$, le tableau suivant montre une matrice de fréquence et position de groupe de mutations :

Mutation / F-P	RR	RY	YR	YY	MM	MK	KM	KK	WW	WS	SW	SS
Fréquence	0.153	0.38	0.30	0.15	0.30	0.15	0.15	0.38	0.23	0.30	0.38	0.07
Position	0.50	0.40	0.40	0.65	0.63	0.34	0.61	0.30	0.0	0.50	0.55	0.38

TABLE III.2: Exemple de fréquence et position des mutations

L'évaluation de cette approche d'étude de similarité entre les séquences d'ADN est effectuée sur une base de séquences codantes du premier exon du gène de bêta globine pour onze espèces, les résultats obtenus avec une discussion seront montrés dans le chapitre "Expérimentations et Résultats".

III.3.2 Analyse de similarité des séquences protéiques

Lorsque nous discutons de la fonction d'un gène, nous parlons automatiquement de la fonction de ses protéines, une protéine est une suite de symboles ou de caractère appelée acide aminé. La complexité du séquençage de ces acides aminés est cachée dans la façon dont ces acides aminés sont structurés, dont la répétition des unités sur toute la longueur de la séquence et aussi dont la répartition des acides aminés sur la base de propriétés physico-chimiques. Donc l'analyse des séquences de protéine doit effectuer du point de vue de leur complexité structurelle pour découvrir la signification des messages qui porte.
D'un autre côté, l'étude de similarité d'une séquence protéique inconnue(requête) avec une autre protéine d'une fonction connue(référence) est effectuée pour inférer la fonction d'une nouvelle protéine pour plusieurs raisons. Tout d'abord, lorsque la ressemblance entre deux protéines est supérieure à 40%, il est probable que les deux protéines ont une fonction biologique similaire, et une fonction générale similaire avec une ressemblance partagée aussi faible que 25% [WKG00]. Plus important encore, lorsque deux séquences de protéines sont très similaires, partagent également des sous-unités structurelles et des domaines protéiques. Ces domaines protéiques sont souvent impliqués dans des fonctions biologiques spécifiques. Dans ce cas, la similarité des séquences est un très bon prédicteur de la similarité fonctionnelle.

III.3.2.1 Transformation

Donc, pour ces raisons nous avons proposé une technique d'étude de similarité entre les séquences protéiques en terme des propriétés physico-chimiques des acides aminés. Il est difficile d'obtenir l'information de la séquence primaire de protéine directement ; Dans notre approche, nous nous basons sur les propriétés physico-chimiques des acides

aminés les plus importantes. La charge physique de l'acide aminé est une propriété physique importante qui divise les vingt acides aminés en trois catégories, les protéines qui doivent se lier à des molécules positivement chargées ont des surfaces riches en acides aminés chargés négativement comme le glutamate et l'aspartate, tandis que les protéines se sont liées à des molécules chargées négativement ont des surfaces riches en chaînes chargées positivement comme la lysine et l'arginine. D'autres acides aminés sans charge sont inclus dans la catégorie du neutre [Urr04]. Les trois catégories sont montrées dans le tableau suivant :

Acide Aminé (Abréviation)	**Charge**	**Abréviation de classe**
Arginine (R),Lysine(K), Histidine(H),	+	**Z**
Aspartate (D), Glutamate (E)	-	**M**
Glutamine(Q), Threonine(T), Serine(S), Alanine(A), Cysteine(C), Methionine(M), Valine(V), Asparagine(N), Glycine(G), Isoleucine(I), Leucine(L), Phenylalanine(F), Tryptophan(W), Proline(P), Trysosine(Y).	**N**	**N**

TABLE III.3: Classement des acides aminés selon la charge physique

Les acides aminés peuvent être placés dans d'autre catégorie selon leurs importances (acides aminés essentiels, semi-essentiels et non essentiels). Les acides aminés essentiels ou indispensables sont les acides importants pour le corps. En d'autres termes, ne peuvent pas être synthétisés de novo par l'organisme ou bien qui ont synthétisés à une vitesse faible ou insuffisante, et sont donc besoin d'un apport externe. Les acides aminés non essentiels sont les acides qui peuvent être produits à partir d'autres acides aminés ou se produisent par l'organisme lui-même, L'arginine et l'histidine forment le groupe des acides aminés dits semi-essentiels qui ont consommé dans l'alimentation dans certaines circonstances. Le tableau III.4 montre les trois catégories différentes :

Classe	**l'acide Aminé**	**Abréviation de la classe**
Essentiel	K, V, L, I, M, F, W, T	L
Semi-essentiel	R, H.	I
Non-essentiel	A, N, D, C, Q, E, G, P, S, Y.	F

TABLE III.4: Classement des acides aminés selon leur importance

D'autre classification importante des acides aminés basée sur le radical R, qui est considéré comme la partie variable des acides aminés, déterminant les caractéristiques structurelles des protéines, selon les propriétés chimiques de R les acides aminés divisés en quatre catégories montrés dans le tableau suivant :

Classe	l'acide Aminé	Abréviation de la classe
Non-Polar	G, A, C, V, L, I, M, F, P, W.	P
Polar	N, Q, S, T, Y.	G
Acidic Polar	D, E.	E
Basic Polar	R, H, K	S

TABLE III.5: Classification des acides aminés par polarité

Les acides aminés peuvent être classés en six groupes principaux, en fonction de leur structure et des caractéristiques chimiques générales de groupe R. Le tableau 3.6 montre les six classes différentes.

Classe	l'acide Aminé	Abréviation de la classe
Aliphatic	G, A, V, L, I.	A
Hydroxyl	C, S, T, M	H
Cyclic	P	C
Aromatic	F, Y, W,	R
Basic	H, K, R	B
Acidic and their Amide	D, E, N, Q	T

TABLE III.6: Classification des acides aminés Basé sur les caractéristiques chimiques des groupes R

Selon les propriétés physico-chimiques détaillées ci-dessus, nous avons présenté chaque protéine de séquence primaire $X = S1...S2...S3...Sn$. Avec la longueur n, par les quatre classifications (charge, polarité, essentielle et groupe R) CR, PY, ES et RG, par ϕCR, ϕPR, ϕES et ϕRG.
$\Phi CR(X)= \Phi CR(S1)\ \Phi CR(S2)..... \ \Phi CR(Sn)$.
$\Phi PR(X)= \Phi PR(S1)\ \Phi PR(S2)..... \ \Phi PR(Sn)$.
$\Phi ES(X)= \Phi ES(S1)\ \Phi ES(S2)..... \ \Phi ES(Sn)$.
$\Phi RG(X)= \Phi RG(S1)\ \Phi RG(S2)..... \ \Phi RG(Sn)$.
Exemple: Pour la séquence de protéine "MVHLTPEEKSA", chaque acide aminé présenté par sa propre catégorie (abréviation de classe) selon les quatre classifications. Sa représentation :
$\Phi CR(X) = NNZZNNNMNNZ$.
$\Phi PR(X) = HABAHCTTBHB$.
$\Phi ES(X) = PPSPGPGGSGS$.
$\Phi RG(X) = LLILLFFFLFI$.

III.3.2.2 Analyse de fréquence

Les classes d'abréviations spécifiées dans les tableaux ci-dessus considérées comme un nouveau composant pour représenter la séquence de protéine selon les quatre classifications, regroupées en vecteur de seize composants présentant la fréquence des nouveaux

composants acides aminés avec les mêmes propriétés calculées par la formule (III.4):
UV est le nombre d'acides aminés appartenant à la même catégorie.
n est la longueur de la séquence protéique.
par exemple la fréquence de la catégorie G désignée par f_G, T désignée par f_T , la même chose pour toutes les catégories du quatre classifications.

III.3.2.3 Analyse de position

Comme mentionné ci-dessus dans la partie d'étude de similarité entre les séquences d'ADN, la fréquence d'apparition des composants de base (nucléotides) est une information insuffisante.
Pour les chaines protéiques, la présence d'acides aminés avec la même fréquence tout au long de la séquence ne suffit pas à déterminer la similarité ou la différence entre deux chaînes protéiques, car une différence au niveau du site d'apparition d'acides aminés peut conduire à une différence radicale dans la fonction protéique.
Cependant, La détermination de similarité entre deux chaînes de protéines en terme de position de leurs composants est très importante pour l'étude de la similarité entre aux.
A partir de l'étape de transformation d'une séquence protéique en quatre séquences symboliques respectant les propriétés physico-chimiques des acides aminés, nous avons passé à une étape de représentation vectorielle selon deux critères importants, (la fréquence et la position) de ses nouveaux composants.
Dans cette étape chaque séquence de protéine est représentée par un vecteur de size dimensions représente la position moyenne des composants, nous appliquons la formule (III.5):
K : est le nombre de mots UV.
n : est le nombre des acides aminés de la séquence protéique.

III.3.2.4 Etude de similarité

Chaque séquence protéique est représentée par deux vecteurs de size dimensions, un vecteur représente la fréquence d'apparition des nouveaux composants et un autre représente leurs positions moyennes. La similarité est calculée comme celle de l'ADN, nous avons fait appel à la distance euclidienne entre les vecteurs de fréquence et entre les vecteurs de position moyenne, telle qu'une grande similarité entre deux séquences protéique est définis par une valeur minimale de distance euclidienne. Cependant, nous avons appliqué la même formule pour l'étude de similarité entre les séquences d'ADN (III.6)
Telle que ; SF définit la distance euclidienne entre deux séquences protéiques en terme de fréquence de leurs composants (vecteurs de fréquence à 16 composants), SP Définit la distance euclidienne entre deux séquences protéiques en terme de position moyenne de leurs composants. Donc, la similarité entre deux séquences protéiques est définie par la valeur moyenne de la similarité de fréquence d'apparition et la position moyenne.
Dans le but d'évaluer cette technique et examiner son efficacité ; nous avons effectué des expérimentations sur une base de séquences de protéine de bêta globine de treize espèces différentes, les résultats obtenus avec une discussion sont montrés dans le chapitre "Expérimentation et Résultats".

III.4 Processus d'ECD Biologique

Comme nous l'avons vu dans les deux précédents chapitres, deux points très importants : d'un côté, les données biologiques (ADN/ ARN/Protéine) sont complexes, et l'analyse de ce type de données sous un format de base est une tâche difficile non seulement en raison de sa complexité, pour tirer un maximum de connaissance cachées. D'un autre côté, L'ECD biologique (inclus les différentes méthodes de fouille de données) est encore un « art », pratiqué avec succès par des groupes de recherche en bio-informatique qui s'occupent de résoudre des problèmes de la génomique et protéique. Pour cela, nous présentons le processus d'ECD particulièrement sur les données biologiques (ADN et protéine) sous un format primaire, dans le but de résoudre deux problèmes de bio-informatique majors : le premier est la prédiction de la structure des molécules basant sur l'apprentissage non-supervisé des séquences d'ADN et le deuxième est la classification des protéines inconnues se basant sur les règles d'associations. Dans son contexte biologique ; le processus d'ECD est particulièrement standard, nous présentons dans la suite nos deux méthodologies inspirées de l'ECD qui est représenté par cinq étapes : La définition du problème, la collection des données, le pré-traitement des données, la modélisation et la validation. Pour chaque étape, différentes techniques de traitements spécifiques d'une étape à une autre peuvent être appliquées selon le type et la nature de la donnée traitée.

III.4.1 Définition du problème

Le but de cette étape est de définir à la fois l'objectif de l'étude et le problème (objectifs et attentes) et de définir les connaissances de domaine qui peuvent être nécessaires. Ceci est effectué de manière itérative ; L'utilisateur entreprend des étapes ultérieures plusieurs fois avant de parvenir à une définition de problème Convaincante .
Le processus de découverte et d'extraction de connaissances biologiques devrait prendre en compte à la fois les caractéristiques des données biologiques et les exigences générales du processus. Sachant que, les techniques de fouille de données (FD) sont considérées comme le cœur battant de l'ECD, il se compose de trois problèmes majeurs : classification, clustering et association.
Dans ce cadre, nos méthodologies ont été inspirées du processus ECD et basées sur les techniques de fouille de données pour résoudre deux problèmes biologiques connus dans la bio-informatique, les deux méthodologies adoptées sont communes dans quatre étapes de l'ECD, la différence est dans l'étape de modélisation, tels que ; le clustering pour résoudre le premier problème et la classification et l'association pour le deuxième problème :

III.4.1.1 Prédiction de la structure des molécules pour le développement des médicaments

De plus en plus, les développeurs de médicaments recherchent des molécules importantes et en particulier des protéines comme option thérapeutique.
L'un des importants aspects de techniques de bio-informatiques est de faciliter l'emploi des nouvelles et puissantes technologies dans le processus de découverte de médicaments. Ce processus est effectué à partir de l'identification de gènes jusqu'à la modélisation de protéines, la connaissance du génome (sa structuration, son séquençage et sa localisation) a donné aux chercheurs la possibilité de prédire multiples protéines, donc

un gène constituant en quelque sorte la recette de confection d'une protéine. Mais le problème général est encore loin d'être résolu.
Le volume total d'informations biologique disponible et les techniques de bio-informatique différents ont donné la naissance d'un nouveau paradigme de prédiction de structure pour la découverte des nouveaux médicaments. Nous abordons dans notre travail l'implémentation de méthode de fouille de données pour faire face à ce paradigme.
Le problème posé est : comment découvrir la structure des nouveaux médicaments (molécules) à partir des gènes représentés par sa structure de base (primaire) en utilisant la technique de fouille de donnée (clustering) ? Plus de détail dans la partie modélisation (III.4.4).

III.4.1.2 Classification de protéine inconnue pour la prédiction de leur fonction

La classification des objets biologiques est l'une des tâches fondamentales et traditionnelles des sciences de la vie ; la catégorisation des gènes et des protéines elle-même est devenue un sujet de recherche important, tels que les gènes / protéines connus sont classés dans des catégories empiriques déterminées a priori, qui reflètent nos connaissances actuelles sur les fonctions cellulaires et biochimique. Les principales difficultés de la classification des protéines découlent du fait que les données sont larges, bruyantes, hétérogènes et redondantes ; que les classes elles-mêmes sont très différentes en termes de leurs caractéristiques ; et aussi du point de vu complexité des données, les chaînes de protéines se caractérisent par la longueur et le nombre de composants (limité à vingt composants de base) et la complexité de séquençage et les caractéristiques des acides aminés. Pour ces raisons, il existe un besoin constant de méthodes efficaces pour répondre à ces exigences. Dans ce contexte, nous avons proposé une méthodologie adapté à cette portée, qui sert à classifier les protéines inconnues, nous avons touché le problème de complexité de séquençage et complexité des composants (acide aminés), par l'extraction des relations entre ces composants de base basant sur les règles d'associations pertinentes ; selon lesquelles nous proposons une technique de classification des protéines; plus de détails dans la partie modélisation (III.4.4).

III.4.2 Collection des données

Dans le web, il existe un nombre considérable de bases de données biologiques, sont de différents type et nature, comme nous l'avons détaillé dans le chapitre "Données Complexes".
Donc cette étape de processus ECD, exige de collecter un ensemble de données à partir de différents sites de base de données biologiques qui représentent les données d'apprentissage sur lesquelles nous appliquons nos méthodologies, ce qui concerne notre système nous nous sommes basés sur trois aspects différents pour la collection des données d'apprentissage :

1. **Collection direct :** Nous avons téléchargé des séquences biologiques sous forme de fichier de séquences sur le web.

2. **Collection par l'interface IPA :** Ce type de collection est effectué à partir des interfaces (IPA) dans les sites de base de données biologiques.

3. **Collection par programme :** Nous avons développé un programme java qui sert à connecter les sites de base de données biologiques et télécharger les séquences (ADN/protéine) selon les critères de l'utilisateur (URL, type de la donnée biologique, famille de la séquence, identifiant de l'espèce).

III.4.3 Pré-traitement des données biologiques

Pour un traitement ultérieur, les données collectées dans l'étape précédente, sont représentées sous un format non structuré, peuvent être incomplètes et non interprétables par la machine, nécessitent un nettoyage et une normalisation, dans le but de représenter les données (ADN/Protéine) sous un format structuré et cohérent, nous sommes appuyés sur les techniques de pré-traitement détaillé dans le schéma suivant:

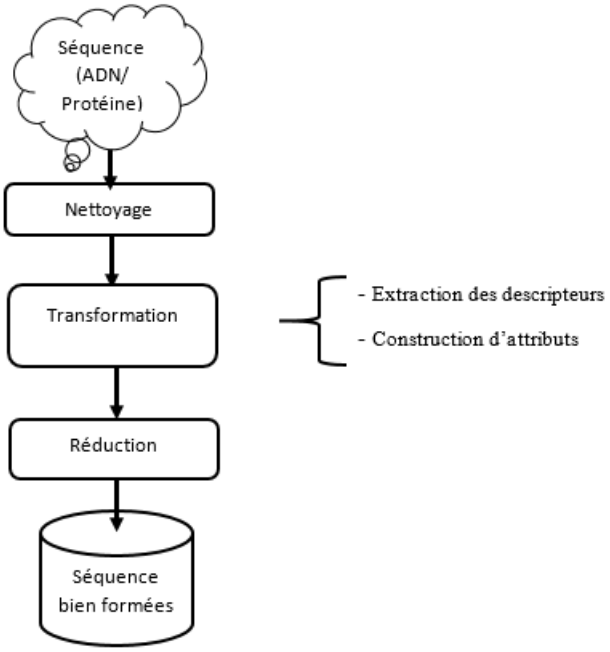

FIGURE III.1: Les étapes de pré-traitement des séquences biologiques

III.4.4 Nettoyage

Pour le nettoyage des textes bruts, l'utilisateur élimine tous les caractères indésirables (caractére, mot clé, chiffre ect...), donc si on parle d'une séquence biologique (ADN/Protéine) nous parlons des composants de bases (nucléotides/acides aminées), la question posée quels sont les éléments qui doivent être nettoyés dans une séquence de base ADN ou protéine? Donc, comme nous l'avons vu dans le premier chapitre, il existe plusieurs formats de représentation des séquences biologiques disponibles dans l'internet, nous avons choisi la plus simple et la plus connu (FASTA). La partie de représentation de séquence de ce format peut comporter des caractères et composants inhabituels qui ont rendu la séquence incompréhensible ou signifient d'autre fonction ou une structure spécifique, dans ce contexte, nous avons éliminé les séquences qui comportent ces caractères et composants : Éliminer les séquences inconnues et les séquences redondantes, dans ces deux étapes l'intervention d'un expert du domaine est

obligatoire.
Les codes nucléotides inhabituels :
N : Représente n'importe quelle base
R : Représente le base A ou G (purine)
Y : Représente la base C ou T (Pyrimidine)
- : ne présente aucune base.
Codes acides aminés inhabituels :
B : Représente Gln ou Glu
Z : Représente Asn ou Asp
X : Représente n'importe quel résidu
- : aucun résidu correspondant.

III.4.5 Transformation

L'étape de transformation permet de passer d'une séquence ADN ou protéine brute à une liste de termes.

1. **Extraction des descripteurs** Cette étape consiste à extraire des attributs à partir des séquences (ADN), en appliquant la technique N-Gramme la plus connue en fouille de texte [Mil+06], l'application de cette technique se fait par le déplacement plusieurs fois de n cases sur la séquence ADN/protéine, le déplacement est effectué par un seul caractère jusqu'à la fin de la séquence. Dans chaque déplacement, on prend une photo. Toutes ces photos représentent tous les n-gramme qu'on puisse générer. Par exemple, pour générer tous les 3-Grammes dans le segment de la séquence d'ADN - « CCAGCTGCAT » - on obtient CCA, CAG, AGC, GCT, CTG, TGC, GCA, CAT.
   Avantage de N-gramme pour les séquences biologiques :
   La séquence biologique caractérisée par sa longueur et la limite de ses composants, nous savons bien qu'une séquence ADN/protéine est définie comme une chaine de caractères et ne pas représenter comme une phrase, donc son traitement est spécial, ce n'est pas comme les textes bruts, la représentation Sac a Mot, sac de phrase, la représentation conceptuelle, représentation par racination, ne marche plus avec une séquence biologique (chaine de caractères non séparée). Donc la seule représentation qui convient avec les séquences biologiques de base est la représentation N-Gramme.

2. **Codage : Attributs-valeurs** Le principe de cette étape est de coder chaque composant obtenu après l'application de la technique n-gramme pour générer un vecteur pour chaque séquence, donc une base de N séquences est représentée par une matrice de N ligne et M colonnes (N représente les séquences ADN pour l'apprentissage et M représente les descripteurs T_{ij} obtenues par la technique de n-gramme).

	C_1	C_2	C_3	C_M
V_1	T_{11}	T_{12}	T_{13}	T_{1M}
V_2	T_{21}	T_{22}	T_{11}	T_{2M}
V_3	T_{31}	T_{32}	T_{33}	T_{3M}
V_N	T_{N1}	T_{N2}	T_{N3}	T_{NM}

Matrice de N séquences et M composants

Dans notre système nous avons utilisés les pondérations suivantes :

- **Pondération TF-IDF:** Cette pondération définie par la fréquence d'apparition du composant (gramme) par le poids de ce mot dans l'ensemble de séquences est appelé en anglais (frequency*inversed documents frequency (TF*IDF)):

$$TF - IDF(i) = TF(i) * \log(\frac{N}{(N_i)}) \tag{III.7}$$

  TF(i) : Le nombre de Séquence comprenant le mot i.
  N : Le nombre de séquences dans l'ensemble de données.

- **Pondération binaire :** dans cette pondération, le composant est pondéré par 1 si il existe dans la séquence et par 0 si il n'existe pas.

- **Pondération Fréquentielle :** Cette pondération représente à quelle fréquence un composant se produit dans une séquence (ADN/Protéine).
  Meilleur valeur de N-gramme : Après l'application de la technique de n-gramme pour l'extraction des descripteurs, il suffit de déterminer la meilleure valeur de n, dans ce contexte on ne peut pas choisir la valeur de n de façon aléatoire mais on doit réaliser une certain expérimentation et varier cette valeur. En ce qui concerne notre système nous avons choisi la technique qui se base sur la fréquence des n-grammes [155] ; dans le but d'identifier les descripteurs de faibles fréquences pour les éliminer à la fin, consiste à calculer le pourcentage de présence des descripteurs dans l'ensemble de données, et le comparer avec un taux (soit x%) tel que X est une constante variée entre 5 et 25, donc les descripteurs avec une meilleure valeur de x sont conservées à la fin.

III.4.6 Réduction

La phase précédente s'intéresse à l'extraction de descripteurs, en raison de la forte dimensionnalité des séquences biologiques ADN/Protéine, on obtient un nombre considérable des attributs (composants), certains ne sont pas absolument utilisables et certains d'autres sont redondants, ça sert automatiquement à dégrader la performance des techniques de fouille de données dans l'étape de modélisation. Donc on ne peut pas garder un terme et éliminer l'autre, Cependant nous avons appliqué une méthode pertinente de sélection d'attributs, qui nous a permis de passe d'un ensemble de N séquences avec M attributs (descripteurs) à un sous ensemble de N séquence avec M' attributs tel que M'$\leq$ M.
Algorithme Hybride de Sélection d'attributs:
cette approche [MK14] fusionne la précision des algorithmes enveloppants et la rapidité des algorithmes filtrants se compose de deux phases différentes:
La première étape (filtrant) consiste à choisir le meilleur sous ensemble d'attribut selon le critère fréquence d'attribut dans un fichier, le choix de la meilleure valeur de n-gramme détaillé précédemment permet aussi d'ignorer les attributs qui ont un pourcentage de fréquence inférieur à une valeur X.
La sélection selon la fréquence des attributs ne tient pas en compte la redondance des attributs, pour cela, une matrice de corrélation attribut-attribut est construite, sachant

que une valeur de corrélation élevée signifie l'existence de redondance, donc pour garder l'attribut le plus pertinent, il faut calculer deux indices importants α et β associés à chaque attribut, sont exprimés par les formules suivantes :

$$\alpha = \frac{nombre\,des\,chaines\,de\,F_i\,dans\,lequels\,X\,apparait}{Nombre\,totale\,des\,chaines\,de\,F_i} * 100 \qquad \text{(III.8)}$$

$$\beta = \frac{nombre\,des\,chaines\,de\,\cup\,F_i\,dans\,lequels\,X\,apparait}{Nombre\,totale\,des\,chaines\,de\,\cup\,F_i} * 100 \qquad \text{(III.9)}$$

Les attributs sont triés dans l'ordre décroissant selon la valeur absolu $\mid \alpha - \beta \mid$ dans une liste Ψk.
La deuxième étape (enveloppante) : permet de construire le sous ensemble finale d'attribut pertinents sans redondance Φk, par le calcule de taux d'erreur Emin de classification (effectué par l'algorithme SVM). La sélection est commencée par l'attribut X de la valeur $\mid \alpha - \beta \mid$ la plus élevé, l'attribut X est inséré dans l'ensemble Φk et supprimé de l'ensemble Ψk selon la valeur Emin, Ce processus est répété jusqu'un Ψk soit vide. Donc l'ensemble obtenu après la réduction est l'ensemble Φk. Le schéma suivant représente les étapes de l'algorithme hybride de sélection :

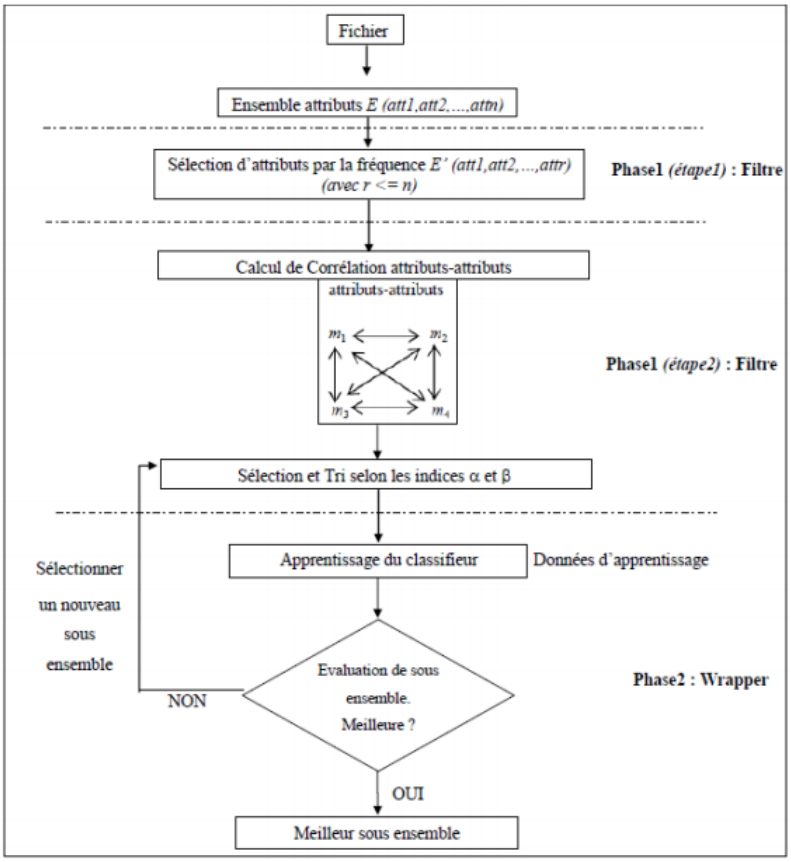

FIGURE III.2: Processus de l'algorithme hybride de sélection [MK14]

III.4.7 Modélisation des données biologiques

Une fois que les données sont traitées dans les étapes (pré-traitement et transformation), il est nécessaire de les interpréter en fonction de l'objectif de recherche déterminé

dans l'étape de définition de problème, donc dans cette étape nous présentons nos deux méthodologies :

III.4.7.1 Regroupement des séquences ADN par l'automate cellulaire 3D

Plusieurs algorithmes et techniques de classification non supervisée (regroupement) ont été développés et appliqués sur les données de différentes natures, comme les algorithmes hiérarchiques, les algorithmes de partitionnement, les algorithmes de bio-inspiré etc. Vue de l'efficacité des algorithmes de bio-inspiré et son application, nous avons appliqué le regroupement des séquences ADN par l'algorithme bio-inspiré de l'automate cellulaire 3D [Ham+12] dans le but de traiter le problème de prédiction des structures de molécule. La conception du système décrit dans la figure suivante :

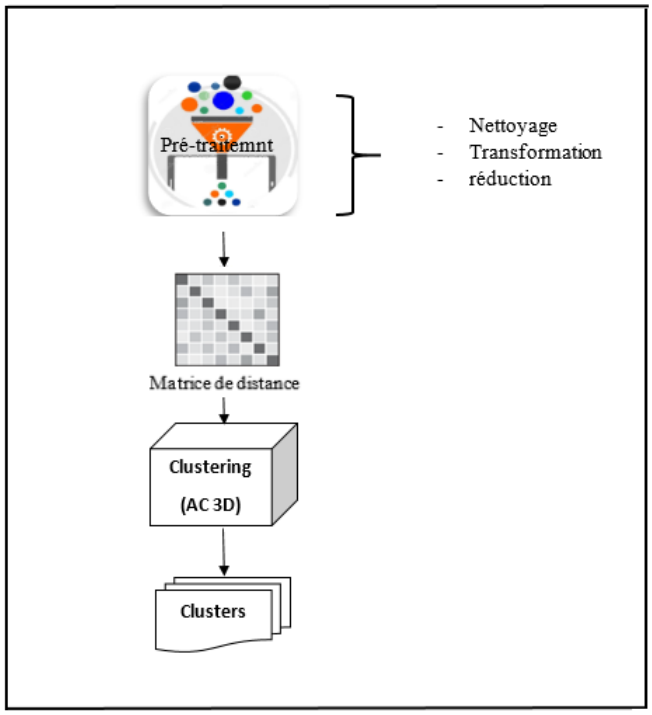

FIGURE III.3: Clustering des séquences d'ADN par AC3D

1. **Matrice de similarité**
   Une matrice de similarité (N * N) représente la similitude entre un nombre de points de données. Dans notre étude, les points de données représentent les vecteurs de fréquence (les séquences d'ADN), chaque élément de matrice représente la similarité entre deux vecteurs i et j où la similarité $(i, j) = D(V_i, V_j)$, N est le nombre de séquences ADN (donnée d'apprentissage) et D est la distance entre i et j. Dans notre système nous avons choisi trois métriques de distance de similarité sont : la distance euclidienne, la distance de Minkowsky 4 et la distance cosinus.
   La distance euclidienne : dans un espace n-dimensionnel, la distance euclidienne entre les vecteurs Xi et Xj est définie sous la forme suivante :

$$D(T_i, T_j) = \sqrt{\sum_k ((X_k(T_i)) - (X_k)(T_j))^2} \tag{III.10}$$

La distance Minkowsky 4 : dans un espace n-dimensionnel, la distance minkowsky entre les vecteurs Xi et Xj est définie sous la forme suivante :

$$D(T_i, T_j) = \sqrt[4]{\sum_k ((X_k(T_i)) - (X_k)(T_j))^4} \tag{III.11}$$

La distance du cosinus : dans un espace n-dimensionnel, la distance du cosinus entre les vecteurs Xi et Xj est défini sous la forme suivante :

$$\cos(x, y) = \frac{T_i.T_j}{\parallel T_i \parallel . \parallel T_j \parallel} \tag{III.12}$$

Dans lesquelles la diagonale de la matrice de distance (euclidienne et minkosky) est égale à 0, donc les distances qui convergent vers zéro signifie une forte similarité entre les séquences ADN. Cependant, la diagonale de la matrice de distance (cosinus) est égale à 1, donc les distances qui convergent vers 1 signifie une forte similarité entre les séquences ADN.

2. **La source d'inspiration de l'automate cellulaire 3D (AC3D)**
L'inspiration d'AC fournie par les tissus cellulaires biologiques pour définir les éléments d'un système cellulaire abstrait, en abstraction d'un tissu cellulaire, la collecte de cellules devient un espace cellulaire discret. L'état interne complexe d'une cellule biologique est réduit à une variable d'état numérique ou symbolique qui prend ses valeurs dans un ensemble d'étapes raisonnablement simple. Les règles complexes et les interactions qui régissent la dynamique temporelle sont abstraites par une fonction automatique ou une règle qui montrent que la variable a doit être datée dans le temps, en tenant compte des interactions d'une cellule avec ses voisins, à partir d'une configuration initiale donnée de l'espace cellulaire.

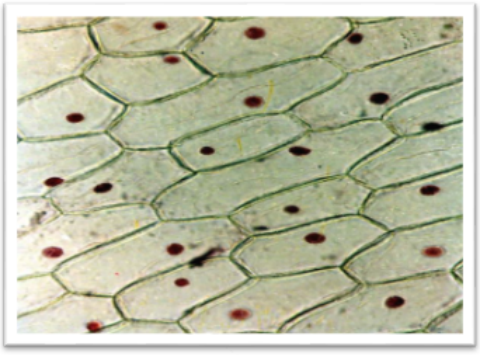

FIGURE III.4: Tissu cellulaire biologique[Ham+12]

En termes plus précis et formels, un système cellulaire abstrait se compose les éléments suivants :
Espace cellulaire : La collection de cellules dans le système s'appelle "l'espace cellulaire". En général, c'est un réseau de cellules de D dimension. En ce qui concerne les systèmes cellulaires à un niveau abstrait, le réseau est généralement considéré comme fini.
Variable de temps : La dynamique du système cellulaire se déroule selon un axe temporel qui peut être discret ou continu.

Ensemble d'état : L'état d'une cellule représente l'information spécifiant l'état actuel de la cellule, c'est la seule façon qui peut influencer l'avenir du système cellulaire.
Voisinage : Le voisinage d'une cellule est l'ensemble des cellules (y compris la cellule Elle-même) dont l'état peut influencer directement l'état futur de la cellule.
Fonction de transition : C'est la fonction qui indique que l'état d'une cellule se déploie dans le temps. Cela dépend uniquement de l'état de cellule par rapport aux (voisinage, position de la cellule et au temps).

3. **Description de (AC3D)** Nous avons appliqué les automates cellulaires 3D [Ham+12] ; des résultats satisfaisants ont été obtenus par l'application de cette méthode sur la base de données textuelle et offrent une meilleure représentation de l'espace. Dans notre travail nous avons appliqués AC 3D pour le regroupement des séquences ADN.
Représentation
D'un point de vue formelle, l'automate cellulaire 3D est défini par le quadruplet (U, V, E, F) où : U = (U_1,U_2,U_3....,U_n) : représente un ensemble de cellules de la grille 3D.
V = (V_1,V_2,V_3....,V_n) : représente les voisins des cellules.
V_i = (U_j,U_k,....U_m: représente les voisins de la cellule i.
E (mort, vivant, isolé et actif) : représente tous les états possibles de la cellule.
F représente toutes les fonctions de transition locale, déterministe ou probabiliste, qui permet d'évoluer à chaque étape temporelle l'état des cellules en fonction de l'état de la cellule elle-même et les états de ses voisins.
L'automate cellulaire 3D est défini comme un réseau de cellules de l'espace 3D appartient à la famille (k, r) tel que k est le nombre de cas possibles de la cellule et r est l'environnement de la cellule. Le AC a quatre cas possibles ($k = 4$) définis dans l'ensemble U, l'état mort est représenté par la valeur 0, l'état vivant est représenté par la valeur -1, l'état isolé est représenté par la valeur -2, et finalement l'état actif est représenté le nombre de séquences d'ADN d'apprentissage.

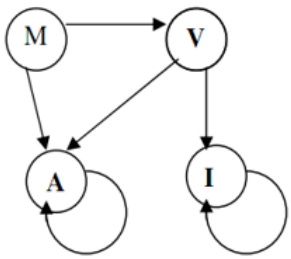

FIGURE III.5: Schématisation de l'automate [Ham+12]

Voisinage
Le voisinage dans l'automate cellulaire 3D est basé sur le voisinage de Moore, consiste de 9 cellules (8 cellules voisines + la cellule elle-même), avec la dimension 3D on obtient 27 cellules, Comme la montre la figure suivante :

FIGURE III.6: Voisinage de Moore 3D [Ham+12]

Procédures

Pour le regroupement des séquences ADN en clusters selon leur similarité, à chaque itération la cellule change son état selon trois règles de transition, la procédure de AC 3D est définie par:

Calcule de la taille de l'automate (grille 3D) : Selon le nombre N des séquences ADN de l'apprentissage, le calcul de la taille est effectué par : X*X*X, la valeur de X est définie par la partie entière de la racine de N plus 1 divisé par le coefficient 3 qui signifié l'espace 3 dimension de l'automate pour représenter les classes.

$$X = \frac{entier\sqrt{N} + 1}{3} \tag{III.13}$$

Donc, la taille de la grille avec 500 instances devient 8*8*8.

Règles : pour le regroupement des séquences ADN en clusters selon leur similarité, à chaque itération la cellule change son état selon les trois règles de transition décrites dans l'algorithme suivant:

Algorithme 6 L'algorithme de l'automate cellulaire 3D

Entrée:
Matrice de similarité D (S_i, S_j).
Initialiser les cellules de l'automate à l'état Morte (état=0).
Sortie : Les clusters.
Début
Pour chaque cellule C_{ij} de l'automate faire
  Si (cellule est morte) Alors
    $C_{ij} \leftarrow$ donnée
    Voisinage de la cellule C_{ij} devient vivant.
  Fin si
  Si la cellule C_{ij} est vivante Alors
    Vérifier le voisinage
  Fin si
  Sinon
  Si le voisinage contient au moins une cellule active Alors
    La cellule $C_{ij} \leftarrow$ données similaires
    Le voisinage de la cellule C_{ij} devient vivant
  Sinon
  Voisinage de C_{ij} devient isolé
  Fin si
  Si la cellule est isolée Alors
    inchangée (reste isolée).
  Fin si
Fin Pour
FIN

III.4.7.2 Classification des protéines par les règles d'association

Le but que nous avons fixé est de proposer un classificateur efficace de protéines en traitant la complexité de l'information protéique sous sa structure primaire de base, dans le but d'obtenir une base de règles optimisées. Pour la représentation des séquences de protéines on appliquant la technique de n-gramme et une bonne stratégie de filtrage des règles d'association. La classification des protéines est une activité importante pour le biologiste dans le but de répondre aux besoins biologiques. Pour cette raison, nous présentons un cadre global inspiré par le processus d'ECD biologique pour classer les protéines, basant sur la base des règles d'association pertinentes.

1. **Architecture générale du système:** Notre système de classification des protéines se compose de cinq étapes importantes décrites dans le schéma suivant :

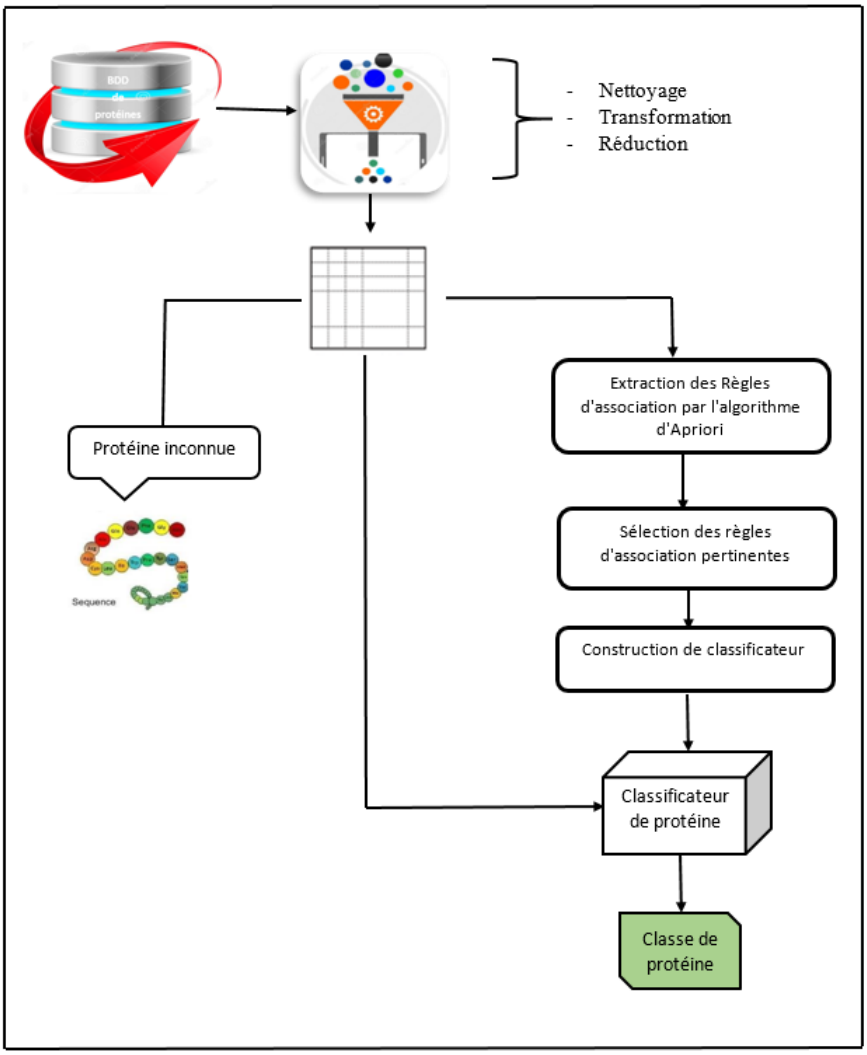

FIGURE III.7: Conception de base de notre système de classification CSP

2. **Extraction des règles d'association:** Pour construire un modèle de classification des protéines, nous nous sommes concentrés sur la génération de règles d'association, l'extraction de ces règles définies comme la recherche de relations entre les éléments dans un ensemble de données. L'exemple le plus connu est l'analyse du panier de marché; tous les achats effectués sont enregistrés dans une base de données, chaque achat est défini comme une transaction et chaque produit défini comme élément.
   Pour extraire les règles d'association entre les séquences de protéines, nous avons défini une nouvelle représentation de la séquence protéique (transaction, éléments) à adapter pour extraire les règles, nous nous sommes concentrés sur les étapes suivantes :
   Transactions et items
   Les séquences de protéines sont représentées dans de nombreux formats différents dans bases de données, ces formats sont tous des fichiers ASCII standard, mais ils peuvent différer en présence de certains caractères et mots qui indiquent où se

trouvent les différents types d'informations et la séquence elle-même. Le plus pratique est FASTA, qui consiste à identifier une partie de la séquence protéique et l'autre partie représente la séquence en symboles (20 acides aminés). Dans notre classificateur, chaque identifiant définit la transaction (achats dans le panier du marché) et chaque descripteur (retenu par la technique N-gram) est défini comme un attribut (élément).
L'extraction des règles d'association nécessite une table de données. C'est une table (Séquence / valeur) où chaque ligne représente une séquence protéique (transaction) et chaque colonne représente un attribut (gramme). L'intersection d'une ligne et d'une colonne représente le poids du descripteur.
Nous obtenons une matrice M de i lignes (protéines) et j + 1 colonnes, de sorte que j représente l'ensemble d'éléments plus l'étiquette de la classe prédéfinie.
Génération des règles par Algorithme Apriori
La génération de règles est une opération consiste à transformer un ensemble d'éléments en règles. [GB02; Zak+97; HGN00] ont montré que l'application des algorithmes Apriori, AIS et FP-growth sur la même base de données avec les mêmes seuils de confiance et de support donne les mêmes résultats globaux (les mêmes règles d'association). C'est parce que les algorithmes d'extraction des règles d'association sont tous plus ou moins proches l'un de l'autre. Selon une étude comparative de [Abd10], l'algorithme d'Apriori est généralement jugé plus ou moins efficace par rapport aux autres algorithmes d'extraction des règles d'association. Par conséquent, dans notre étude, nous avons appliqué l'algorithme Apriori sur l'ensemble des éléments obtenus à l'étape précédente.
L'application de l'algorithme Apriori retenue des règles d'association écrites sous le format X $\rightarrow Y$; indique que les protéines qui contiennent la sous-séquence d'acides aminés de l'ensemble X ont tendance à contenir la sous-séquence d'acides aminés de l'ensemble Y. sachant que le cas où le seuil de support et de confiance de la règle sont supérieurs aux seuils fixés par l'utilisateur.

3. **Sélection des règles d'association significatives**
Il est bien connu que, les règles d'association extraites par l'application d'un algorithme d'extraction sont diversifiées et d'énorme quantité, ce qui nous a amenés au problème original de fouille de données. Donc ; Comment extraire les règles significatives (RS) pour la manipulation et la prise de décision dans l'étape de classification ? Par conséquent ; Nous avons suivi une stratégie efficace pour sélectionner les règles significatives à notre problème parmi celles qui existent. Selon le problème abordé dans notre cas et les différentes méthodes de sélection des règles significatives [HF99; CCF06; Han+96; Bas+02], nous avons gardé les règles d'association avec les caractéristiques suivantes :
Les règles écrites sous la forme $X \rightarrow Classe$ De sorte que la partie droite (conséquence) définit les éléments de l'attribut classe (dans le but de sélectionner les règles adaptables avec le problème de classification, ce qui nous a permis de prédire la classe d'une protéine inconnue).
Une règle d'association $R_a : A \rightarrow C$ est plus générale qu'une autre règle d'association $R_b : B \rightarrow C$ si $A \subset B$. Donc, R_a est plus significative que R_b.

4. **Modèle de classification:** Dans les sections précédentes, nous avons discuté la génération des règles d'association et la sélection de celles qui sont significatives. Cette section a porté sur la construction du modèle de classification selon les

règles d'association significatives; Notre classificateur prédit la classe d'une nouvelle protéine non classifiée à partir des classes prédéfinies. Avant de classer la nouvelle protéine P, nous devons la représenter, nous avons appliqué la technique du n-gramme et calculé la fréquence de chaque composant, donc la protéine est transformée en un vecteur de fréquence $(T_1, T_2, ...T_n)$.
Nous avons déterminé tout d'abord les règles pertinentes parmi les règles significatives obtenues. selon notre type de données, une règle est dite pertinente si elle contient dans sa partie gauche les composants de la protéine P non classée. Nous nous sommes basés sur la formule suivante pour définir les règles d'association pertinentes :
Si (X $\rightarrow$ *classe*) et (A $\subseteq X$) alors (A $\rightarrow Classe$).
(X $\rightarrow Classe$) : la règle significative.
X : ensemble d'éléments du côté droit de la règle.
A : ensemble d'éléments dans la protéine non classifiée.
Cette formule signifie que, l'existence des éléments de A dans l'ensemble des éléments X signifiant qu'il y a une forte probabilité que la classe prédéfinie dans le côté droit de la règle soit définie comme la classe de la protéine P inconnue et c'est suffisante pour sélectionner la règle (X $\rightarrow Classe$) en tant que règle pertinente. Ce processus est répété sur toutes les règles significatives, nous finissons cette étape avec un ensemble de règles pertinentes RR, de sorte que $RR \subseteq RS$ et le nombre de règles dans RR est inférieur ou égal à celui de RS.
Basant sur l'ensemble RR ; Nous avons calculé le pourcentage des règles pertinentes PR par rapport aux règles significatives de chaque classe prédéfinie (donnée d'apprentissage), selon la formule suivante :

$$PRR = \frac{nombres\ des\ regles\ pertinents\ Ci}{nombre\ des\ régles\ sinificatives\ Ci} * 100 \qquad \text{(III.14)}$$

La classe avec la valeur maximale de PRR est définie comme la classe de la nouvelle protéine (P) non classifiée . L'algorithme suivant décrive les étapes de notre classificateur (CSP):

Algorithme 7 L'Algorithme de classification supervisée de protéine (CSP)

Entrée:
  Régles significatives : RS R1,R2, R3,. . . . . ..Rn.
  Vectorisation de la séquence protéique P inconnue (attribut, valeur)
  Initialiser la fréquence des régles pertinants dans chaque classe Ci a 0 : FRC j= 0
Sortie :
  La classe de C_P
Début
Pour chaque composant t_i de P faire
    Pour chaque règle R_j de C_i faire
    Si($t_i \subseteq R_j$.gauche)
    $FR_{ci} \leftarrow FR_{ci} + +$
    Fin si
  Fin Pour
$PRR = \frac{nombres\ des\ regles\ pertinents\ Ci}{nombre\ des\ régles\ sinificatives\ Ci} * 100$
C_P = Maximum (PRR)
  Fin Pour
FIN

III.5 Évaluation et interprétation

Cette étape est importante dans nos méthodologies, sert à évaluer la qualité des méthodes de fouille de données, et interpréter les résultats obtenus, cette tache dépend de la technique de fouille de données dans le processus de l'ECD (classification supervisée, non-supervisée et règle d'association). les métriques de validation que nous avons choisies sont les suivantes :

III.5.1 Rappel (R)

Le rappel est une métrique d'évaluation permet de mesurer la capacité du système à fournir toutes les solutions pertinentes; pour la classification supervisée des protéines, le rappel est définie comme le nombre de protéines pertinentes récupérées en fonction du nombre de protéines pertinentes existantes.

$$R = \frac{nombre\,de\,protein\,correctement\,assigné\,\text{à}\,\,la\,\,classe\,\,i}{nombre\,de\,protein\,appartient\,\text{à}laclasse\,i} \tag{III.15}$$

Et pour la classification non-supervisée des séquences ADN est défini par la formule suivante:

$$R(i,k) = \frac{N_{ik}}{N_{ci}} \tag{III.16}$$

N : nombre de séquences ADN sélectionnées pour l'expérimentation.
I : Le nombre de clusters prédéfinis par un expert (biologiste).
K : Le nombre de clusters construits par le système de classification.
N_{ik} :Le nombre de séquences ADN prédéfinies appartient au C_i et le système les classées dans le cluster C_k.

N_{Ci} : Le nombre de séquences ADN préalablement classifiées par un expert qu'ils appartient à la classe C.

III.5.2 Précision (P)

Le Précision est une métrique d'évaluation permet de mesurer la capacité de système pour refuser les solutions non pertinentes; pour la classification supervisée des protéines, la précision est définie par le nombre de protéines pertinentes trouvées par rapport au nombre total de protéines proposées :

$$P = \frac{nombre\,de\,protein\,correctement\,assigné\,\text{à}\,la\,classe\,i}{nombre\,de\,protein\,assigné\,\text{à}\,la\,classe\,i} \tag{III.17}$$

Et pour la classification non-supervisée des séquences ADN la précision est définie par la formule suivante :

$$P(i,k) = \frac{N_{ik}}{N_k} \tag{III.18}$$

N_k :Le nombre de séquences ADN classés par le système dans la classe N_k.

III.5.3 F-mesure (F)

Est une mesure populaire, combine l'intérêt des mesures de précision et de rappel, qui mesure la capacité de système de donner toutes les solutions pertinentes et à refuser d'autres ; pour la classification supervisée des protéines, f-mesure est définie par:

$$F = 2 * \frac{Rappel\ *\ Précision}{Rappel\ +\ Précision} \tag{III.19}$$

Et pour la classification non-supervisée des séquences ADN est F-mesure est définie par la formule suivante :

$$F = \sum_i \frac{N_{ci}}{N} \max[k] \frac{2 * rappel(i,k) * précision(i,k)}{(rappel(i,k) + précision(i,k))} \tag{III.20}$$

III.5.4 Entropie(E)

L'entropie permet de mesurer la perte d'information de système pour la classification non-supervisée est définie par :

$$E = \sum_i^k \frac{N_k}{N} \times (-\sum précission(i,k) \times \log précission(i,k)) \tag{III.21}$$

III.6 Conclusion

Le traitement et l'analyse de la donnée complexe disponible dans les banques de données et dans le web pour l'extraction de l'information appropriée est un problème crucial dans les domaines de recherche d'information. L'un des données complexes qui connait une large gamme de recherche est la donnée biologique de base, sa meilleure façon de

traitement et d'analyse aidant à comprendre plusieurs phénomènes biologiques importants. Dans ce chapitre nous avons décrit la partie conceptuelle de nos travaux, se compose de deux objectifs principaux, le premier est l'analyse des propriétés physiques et chimiques des composants de base des molécules biologiques avec certaine notion mathématique pour la comparaison des séquences génomiques et protéiques et l'étude de similarité existant entre eux. En revanche, nous avons détaillé le système conceptuel de deux problèmes de bio-informatique: (1) la classification non-supervisée de séquences ADN par un algorithme de bio-inspiré (l'automate cellulaire 3D) pour la prédiction des structures de molécules et le développement des médicaments, (2) la classification supervisée des protéines par la construction d'un nouveau classificateur basé dans son contexte sur les règles d'association entre les composants de protéines. les deux problèmes sont basés sur l'ECD biologique, qui se compose de cinq étapes: La définition du problème, la collection des données biologiques, le pré-traitement des données biologiques, la modélisation avec les techniques de fouille de données (classification, regroupement, règles d'association), et la validation.

Chapitre IV

Expérimentations et Résultats

IV.1 Introduction

Le chapitre précédent s'inscrit dans le cadre conceptuel de nos travaux de thèse. Cependant, ce chapitre est consacré à représenter la partie expérimentale pour évaluer les modèles proposés. D'abord, dans le but de pouvoir évaluer les deux méthodes proposées pour l'étude de similarité des séquences d'ADN et protéiques, nous avons utilisé le jeu de donnée mise dans la plupart des travaux de comparaison de séquences biologiques qui présentent les séquences ADN et protéines de béta-globine. Pour la prédiction des structures moléculaires et le développement des médicaments, nous avons appliqué la méthodologie expliquée dans le chapitre précédent, sur une base de données de séquences ADN primaire de la banque de donnée GenBank et discuté à propos des résultats obtenus. Ce qui concerne le problème de classification supervisée des protéines, les expérimentations ont été effectuées sur cinq familles de protéines extraites de la banque de données Uniprot, les résultats obtenus sont comparés avec cinq classificateurs de la plateforme WEKA[1].
La validation des résultats est basée sur les métriques d'évaluation connues dans le domaine de fouille de données : Rappel, précision, F-mesure et Entropie. Les résultats obtenus ont satisfait notre objectif et ouvre une porte vers des nouvelles méthodologies d'analyse et de fouille de données pour le traitement des données biologiques complexes.

IV.2 Etude de similarité des séquences biologiques

Afin de présenter les étapes abordées pour l'étude de similarité entre les séquences biologiques dans le chapitre précédent, Cette partie s'intéresse à l'évaluation de leurs efficacités sur un jeu de données de séquences d'ADN et protéine :

IV.2.1 Similarité des séquences d'ADN

IV.2.1.1 Jeu de données (Beta-globine)

Pour évaluer notre méthode d'étude de similarité entre les séquences biologiques (ADN), Nous avons utilisé dans l'expérimentation un ensemble de séquences d'ADN dérivées de l'ensemble de données obtenu par la banque de données NCBI, comprend le premier exon de gènes de la béta globine de 11 espèces différentes représentées dans le tableau IV.1.

[1]WEKA: est disponible dans: http://weka.wikispaces.com/, consulté le: 2017-08-13.

Espèces	Séquences d'ADN
Humain	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Chèvre	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGG TCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGG GGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Maki	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGG GGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Souris	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGG GCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Lapin	ATGGTGCATCTGTCCAGTGACGAGAAGTCTGCGGTCACTGCCCTGTGG GGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGG GGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorille	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzé	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTT GGTATCAAGG

TABLE IV.1: Le premier exon de gène bêta globine pour 11 espèces

IV.2.1.2 Étude de similarité

Pour appliquer notre approche d'étude de similarité entre les séquences ADN bien détaillée dans le chapitre de modélisation, nous montrons dans cette section les expérimentations sur la base de séquences ADN de bêta globine pour l'évaluation de notre approche, nous avons procédé la comparaison par paire de tous les 11 espèces (deux à deux). Cette section se compose de trois étapes différentes procédés comme suit :

1. **Calcul de Fréquences des mutations** Pour chaque séquence de la base, nous avons calculé la fréquence des douze composants (groupe de mutation) pour les 11 espèces, le tableau suivant représente les résultats obtenus :

	f_{RR}	f_{RY}	f_{YR}	f_{YY}	f_{MM}	f_{MK}	f_{KM}	f_{KK}	f_{WW}	f_{WS}	f_{SW}	f_{SS}
Humain	0.3297	0.2308	0.2308	0.2088	0.1978	0.1978	0.1868	0.4176	0.1209	0.2967	0.2857	0.2967
Gallus	0.3407	0.2308	0.2308	0.1978	0.2418	0.2308	0.2198	0.3077	0.0989	0.2747	0.2637	0.3626
Maki	0.3626	0.2198	0.2198	0.1978	0.1319	0.2418	0.2308	0.3956	0.1429	0.3187	0.3077	0.2308
Lapin	0.3483	0.2472	0.2360	0.1685	0.1798	0.1910	0.1910	0.4382	0.1011	0.3146	0.3034	0.2809
Rat	0.3297	0.2418	0.2418	0.1868	0.1978	0.2198	0.2088	0.3736	0.1758	0.2747	0.2637	0.2857
Bovine	0.3882	0.2118	0.2118	0.1882	0.1765	0.2118	0.2000	0.4118	0.1294	0.2824	0.2706	0.3176
Opossum	0.3077	0.2308	0.2308	0.2308	0.2308	0.2198	0.2088	0.3407	0.1319	0.3407	0.3297	0.1978
Gorille	0.3478	0.2283	0.2283	0.1957	0.1957	0.1957	0.1848	0.4239	0.0978	0.3043	0.2935	0.3043
Souris	0.3118	0.2258	0.2258	0.2366	0.1935	0.2043	0.1935	0.4086	0.1183	0.3118	0.3011	0.2688
Chèvre	0.3882	0.2118	0.2118	0.1882	0.1647	0.2353	0.2235	0.3765	0.1059	0.2941	0.2824	0.3176
Chimpanzé	0.3462	0.2308	0.2308	0.1923	0.1923	0.1923	0.1827	0.4327	0.1250	0.2981	0.2885	0.2885

TABLE IV.2: Fréquences des mutations de 11 espèces

2. **Calcule de la position moyenne des mutations** Comme deuxième étape est le calcul de position moyenne des mutations pour chaque séquence de la base, pour les douze composants (mutations) de 11 espèces, le tableau suivant représente les résultats obtenus :

	P_{RR}	P_{RY}	P_{YR}	P_{YY}	P_{MM}	P_{MK}	P_{KM}	P_{KK}	P_{WW}	P_{WS}	P_{SW}	P_{SS}
Humain	0.5502	0.4746	0.4956	0.4274	0.4621	0.4512	0.4551	0.5480	0.5325	0.4554	0.4573	0.5539
Gallus	0.5115	0.4558	0.4762	0.5317	0.5490	0.4668	0.4670	0.4922	0.5336	0.4259	0.4286	0.5837
Maki	0.5608	0.4632	0.4841	0.4194	0.5220	0.4590	0.4636	0.5250	0.4632	0.4505	0.4505	0.6332
Lapin	0.5814	0.4479	0.4414	0.4569	0.4789	0.4243	0.4613	0.5457	0.4969	0.4539	0.4557	0.5807
Rat	0.5326	0.4500	0.4695	0.5171	0.4847	0.4879	0.4922	0.5048	0.4457	0.4585	0.4592	0.5917
Bovine	0.5387	0.4654	0.4876	0.4419	0.5192	0.4163	0.4187	0.5600	0.4920	0.4716	0.4747	0.5316
Opossum	0.5165	0.4731	0.4950	0.4861	0.4987	0.4599	0.4610	0.5346	0.3974	0.4747	0.4751	0.6258
Gorille	0.5635	0.4695	0.4896	0.4070	0.4571	0.4463	0.4501	0.5535	0.4855	0.4596	0.4622	0.5637
Souris	0.5877	0.4424	0.4644	0.4506	0.5269	0.4493	0.4528	0.5218	0.4399	0.4520	0.4531	0.6146
Chèvre	0.5258	0.4654	0.4876	0.4684	0.5277	0.4382	0.4409	0.5460	0.5033	0.4701	0.4735	0.5316
Chimpanzé	0.5502	0.4780	0.4956	0.4163	0.4606	0.4514	0.4555	0.5468	0.5851	0.4587	0.4603	0.5288

TABLE IV.3: Position moyenne des mutations pour 11 espèces

IV.2.1.3 Résultats de similarité

A base des deux tableaux précédents, pour chaque paire de séquence nous avons calculé la distance euclidienne entre les vecteurs de douze dimensions pour la fréquence et la position moyenne. Donc ce calcul permet de détecter la similarité d'un coté en terme de nombre de composants de séquences et d'autre côté de la position moyenne de ces composants. Selon notre stratégie bien détaillée dans le chapitre de conception, la

similarité entre les 11 espèces est calculée par la similarité moyenne de fréquence et position moyenne, tel que une grande similitude entre deux séquences d'ADN définies par une valeur minimale de distance euclidienne. Les résultats expérimentaux montrés dans le tableau suivant :

Espèces	Humain	Gallus	Maki	Lapin	Rat	Bovine	Opossum	Gorille	Souris	Chèvre	Chimpanzé
Humain	0.0	0.1563	0.1260	0.0791	0.1209	0.0877	0.1624	0.045	0.0980	0.099	0.0451
Gallus		0.0	0.1847	0.1705	0.1316	0.1578	0.1902	0.1718	0.1644	0.1293	0.1771
Maki			0.0	0.1136	0.1270	0.1295	0.1278	0.1186	0.0873	0.1229	0.1484
Lapin				0.0	0.1253	0.1013	0.1601	0.0672	0.0917	0.1112	0.0909
Rat					0.0	0.1338	0.1242	0.1318	0.1110	0.1266	0.1485
Bovine						0.0	0.1772	0.0842	0.1242	0.0524	0.1001
Opossum							0.0	0.1583	0.1134	0.1677	0.1957
Gorille								0.0	0.0962	0.1017	0.0711
Souris									0.0	0.1287	0.1326
Chèvre										0.0	0.1150
Chimpanzé											0.0

TABLE IV.4: Matrice de similarité pour les 11 séquences de gène de bêta globine

Comme nous l'avons remarqué dans le tableau IV.4, les résultats obtenus montrent:
Les cases en jaune montrent l'existence d'une grande similarité entre la séquence humain avec le gorille, l'humain avec chimpanzé et le chimpanzé avec le gorille qui sont exprimés par une valeur minimale de distance euclidienne moyenne, selon plusieurs travaux de recherches biologique et génomique, le gorille est beaucoup plus proche de l'humain et chimpanzé et l'humain partage 98% de son matériel génétique avec le chimpanzé, et chacun de chimpanzé et gorille appartient à la famille des hominidés.
Les cases en vert montre la similarité entre la séquence de souris avec le maki et la souris avec le rat, cette similarité est exprimée par des valeurs minimales de distance euclidienne. de point de vie biologique, les souris et le rat sont de la même famille des mammifères Muridae.
La case en gris , montre la similarité entre la séquence de bêta globine de chèvre et de la vache, sachant que ces deux espèces sont de la même famille de mammifères bovidés.
L'opossum et le Gallus sont loin des autres espèces, car l'opossum est l'espèce le plus éloigné des mammifères restants et le Gallus est le seul animal non mammalien parmi tous les autres animaux de l'ensemble de données. néanmoins, les neuf espèces restantes sont des mammifères.
On globalité, Certaine similarité existe entre toutes les 11 espèces, selon notre propre analyse d'ADN et les résultats expérimentaux sur la base de bêta globine, la relation entre eux est présentée dans le dendrogramme suivant :

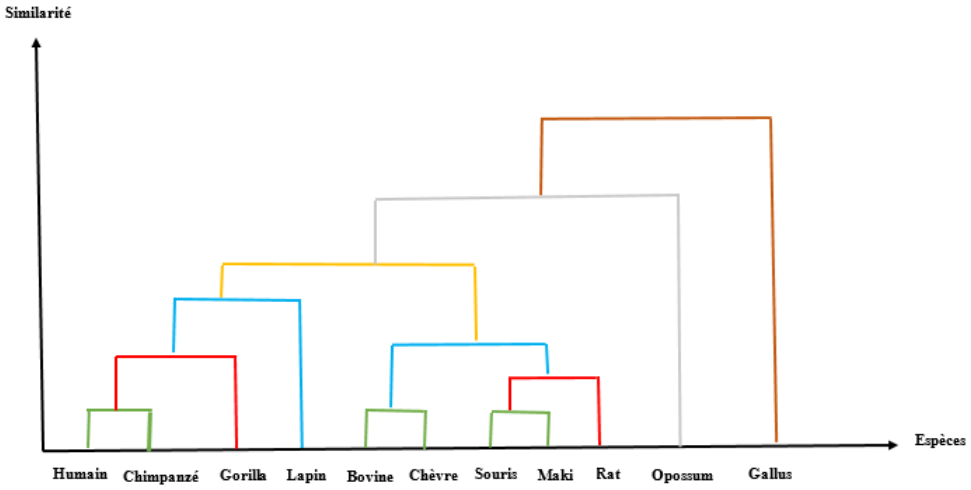

FIGURE IV.1: Le dendrogramme de relation entre les 11 espèces.

En conclusion, les résultats expérimentaux ne sont pas un accident mais confirment l'intuition de l'approche proposée et montrent leur efficacité, coïncident avec le sens de l'évolution de 11 espèces,de sorte que notre analyse de composant de séquences ADN de base se basant sur les propriétés chimiques, fréquences et position moyenne pour l'étude de similarité est performant.

IV.2.2 Similarité des séquences protéiques

IV.2.2.1 Jeu de données (protéine de Bêta globine)

Nous avons effectués un ensemble d'expériences sur une base de protéine de 13 espèces extraites de la banque de données NCBI définies comme une série de bases de données sur les biotechnologies et la bio-médecine, et une ressource importante pour les outils et services de bio-informatique. Les séquences de protéine présentent dans le tableau IV.5 :

Espèces	Séquences de Protéine
Gorille	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTP DAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPEN FKLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Gallus	MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSP TAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVDPENFR LLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
Humain	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTP DAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPEN FRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Chèvre	MLSAEEKASVLSLFAKVNVEEVGGEALGRLLVVYPWTQRFFEHFGDLSSADAI LGNPKVKAHGKKVLDTFSEGLKQLDDLKGAFASLSELHCDKLHVDPENFRLL GNVLVVVLARRFGGEFTPELQANFQKVVTGVANALAHRYH
Maki	MTLLSAEENAHVTSLWGKVDVEKVGGEALGRLLVVYPWTQRFFESFGDLSSP SAVMGNPKVKAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCDKLHVDPQN FTLLGNVLVVVLAEHFGNAFSPAVQAAFQKVVAGVANALAHKYH
Souris	MVHLTDAEKAAVSGLWGKVNADEVGGEALGRLLVVYPWTQRYFDSFGDLSS ASAIMGNAKVKAHGKKVITAFNDGLNHLDSLKGTFASLSELHCDKLHVDPEN FRLLGNMIVIVLGHHLGKDFTPAAQAAFQKVVAGVAAALAHKYH
Lapin	MVHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSA HAVMSNPKVKAHGKKVLAAFSEGLNHLDNLKGTFAKLSELHCDKLHVDPEN FRLLGNVLVVVLSHHFGKEFTPQVQAAYQKVVAGVANALAHKYH
Rat	MVHLTDAEKAAVNGLWGKVNPDDVGGEALGRLLVVYPWTQRYFDSFGDLS SASAIMGNPKVKAHGKKVINAFNDGLKHLDNLKGTFAHLSELHCDKLHVDPE NFRLLGNMIVIVLGHHLGKEFSPCAQAAFQKVVAGVASALAHKYH
Bovine	MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTAD AVMNNPKVKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDKLHVDPENF KLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANALAHRYH
Chimpanzé	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTP DAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPEN FRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Salmo Salar	MVDWTDAERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPA AIMGNPAVAKHGKTVMHGLDRAVQNLDDIKNAYTALSVMHSEKLHVDPDNF RLLADCITVCVAAKLGPTVFSADIQEAFQKFLAVVVSALGRQYH
Âne	MGLTAHDRQLINSTWGKLCAKTIGQEALGRLLWTYPWTQRYFSSFGNLN SADAVFHNEAVAAHGEKVVTSIGEAIKHMDDIKGYYAQLSKYHSETLHV DPLNFKRFGGCLSIALARHFHEEYTPELHAAYEHLFDAIADALGKGYH
Grenouille	MVQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSN PGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPE NFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH

Table IV.5: Séquences de protéines de bêta globine pour 13 espèces

IV.2.2.2 Analyse de fréquence et position

Chaque séquence de protéine a été transformée en quatre séquences symboliques correspond aux propriétés physiquo-chimiques des acides aminés, donc chaque séquence est représentée par deux vecteurs de 16 dimensions, le premier tableau montre la fréquence de chaque composant pour les 13 espèces de la base de protéine et le deuxième montre la position moyenne de chacun des composants :

	f_P	f_G	f_E	f_S	f_Z	f_M	f_N	f_L	f_I	f_F	f_A	f_H	f_C	f_R	f_B	f_T
Gorille	0.574	0.162	0.101	0.155	0.156	0.102	0.748	0.456	0.075	0.469	0.435	0.109	0.048	0.088	0.163	0.163
Gallus	0.565	0.190	0.088	0.156	0.158	0.089	0.760	0.466	0.089	0.452	0.418	0.130	0.034	0.096	0.158	0.171
Human	0.578	0.163	0.102	0.156	0.158	0.103	0.747	0.452	0.082	0.473	0.438	0.110	0.048	0.089	0.158	0.164
Chèvre	0.559	0.166	0.124	0.152	0.153	0.125	0.729	0.444	0.083	0.479	0.444	0.097	0.028	0.090	0.153	0.194
Maki	0.571	0.197	0.088	0.143	0.144	0.089	0.774	0.445	0.075	0.486	0.445	0.123	0.034	0.089	0.144	0.171
Souris	0.578	0.170	0.095	0.156	0.158	0.096	0.753	0.432	0.082	0.493	0.473	0.116	0.021	0.082	0.158	0.158
Lapin	0.544	0.190	0.095	0.170	0.171	0.096	0.740	0.445	0.089	0.473	0.432	0.116	0.027	0.089	0.171	0.171
Rat	0.571	0.163	0.095	0.170	0.171	0.096	0.740	0.425	0.089	0.493	0.445	0.103	0.034	0.082	0.171	0.171
Bovine	0.566	0.159	0.117	0.159	0.160	0.118	0.729	0.479	0.069	0.458	0.431	0.104	0.028	0.097	0.160	0.188
Chimpanzé	0.578	0.163	0.102	0.156	0.158	0.103	0.747	0.452	0.082	0.473	0.438	0.110	0.048	0.089	0.158	0.164
Salmo Salar	0.574	0.203	0.095	0.128	0.129	0.095	0.782	0.442	0.082	0.483	0.435	0.156	0.041	0.075	0.129	0.170
Âne	0.565	0.156	0.116	0.163	0.164	0.116	0.726	0.432	0.089	0.486	0.445	0.082	0.034	0.089	0.164	0.192
Grenouille	0.500	0.226	0.110	0.164	0.166	0.110	0.731	0.379	0.110	0.517	0.386	0.138	0.021	0.124	0.166	0.172

TABLE IV.6: Fréquences des composants de séquences pour 13 espèces

	p_P	p_G	p_E	p_S	p_Z	p_M	p_N	p_L	p_I	p_F	p_A	p_H	p_C	p_R	p_B	p_T
Gorille	0.493	0.499	0.397	0.551	0.555	0.400	0.502	0.475	0.571	0.506	0.506	0.392	0.485	0.518	0.573	0.477
Gallus	0.482	0.468	0.498	0.584	0.588	0.501	0.482	0.484	0.616	0.493	0.494	0.429	0.505	0.485	0.588	0.496
Humain	0.496	0.503	0.400	0.555	0.559	0.402	0.501	0.474	0.586	0.509	0.510	0.394	0.488	0.521	0.559	0.480
Chèvre	0.491	0.517	0.429	0.550	0.553	0.432	0.501	0.486	0.644	0.488	0.497	0.400	0.538	0.514	0.553	0.503
Maki	0.510	0.489	0.377	0.529	0.533	0.379	0.508	0.479	0.555	0.510	0.523	0.362	0.505	0.521	0.533	0.498
Souris	0.504	0.472	0.402	0.555	0.559	0.405	0.500	0.494	0.586	0.491	0.509	0.418	0.594	0.497	0.559	0.464
Lapin	0.495	0.501	0.403	0.548	0.552	0.406	0.500	0.488	0.568	0.499	0.509	0.362	0.545	0.521	0.552	0.502
Rat	0.498	0.484	0.402	0.555	0.559	0.405	0.499	0.490	0.587	0.493	0.504	0.481	0.463	0.497	0.559	0.450
Bovine	0.493	0.515	0.427	0.542	0.545	0.430	0.501	0.480	0.644	0.499	0.508	0.384	0.538	0.484	0.545	0.510
Chimpanzé	0.496	0.503	0.400	0.555	0.559	0.402	0.501	0.474	0.586	0.509	0.510	0.394	0.488	0.521	0.559	0.480
Salmo Salar	0.492	0.510	0.456	0.526	0.530	0.459	0.500	0.493	0.515	0.504	0.503	0.461	0.442	0.534	0.530	0.505
Âne	0.486	0.531	0.396	0.572	0.576	0.399	0.499	0.492	0.616	0.486	0.499	0.444	0.505	0.521	0.576	0.451
Grenouille	0.493	0.423	0.594	0.544	0.548	0.598	0.474	0.457	0.567	0.517	0.507	0.379	0.591	0.538	0.548	0.498

TABLE IV.7: Positions des composants de séquences de bêta globine pour 13 espèces

IV.2.2.3 Résultat de similarité

Dans cette section, nous présentons les résultats de similarité que nous avons obtenus par le calcul de distance euclidienne entre les vecteurs de 16 dimensions. Les deux tableaux ci-dessous montrent les résultats de similarité par paires en termes de fréquence

et de position des acides aminés selon les quatre catégories à base du tableau de fréquence et de position précédents:

	Gorille	Gallus	Humain	Chévre	Maki	Souris	Rabbit	Rat	Bovine	Chimpanzé	Salmon salar	Âne	Grenouille
Gorille	0.0	0.173	0.023	0.110	0.078	0.123	0.076	0.103	0.116	0.023	0.146	0.092	0.313
Gallus		0.0	0.168	0.140	0.226	0.179	0.183	0.171	0.150	0.168	0.183	0.170	0.205
Humain			0.0	0.096	0.079	0.120	0.074	0.103	0.103	0.000	0.147	0.085	0.309
Chèvre				0.0	0.140	0.112	0.096	0.146	0.041	0.096	0.182	0.102	0.275
Maki					0.0	0.135	0.071	0.155	0.133	0.079	0.172	0.150	0.332
Souris						0.0	0.095	0.146	0.118	0.120	0.207	0.122	0.122
Lapin							0.0	0.158	0.096	0.074	0.175	0.127	0.290
Rat								0.0	0.1581	0.103	0.140	0.089	0.334
Bovine									0.0	0.103	0.192	0.125	0.279
Chimpanzé										0.0	0.147	0.085	0.309
Salmon Salar											0.0	0.178	0.284
Âne												0.0	0.335
Grenouille													0.0

TABLE IV.8: Résultat de similarité en termes de fréquence entre les 13 espèces

	Gorille	Gallus	Humain	Chévre	Maki	Souris	Rabbit	Rat	Bovine	Chimpanzé	Salmon salar	Âne	Grenouille
Gorille	0.0	0.054	0.012	0.060	0.062	0.060	0.056	0.053	0.053	0.012	0.091	0.063	0.158
Gallus		0.0	0.056	0.088	0.058	0.085	0.053	0.082	0.073	0.056	0.079	0.098	0.147
Humain			0.0	0.057	0.060	0.055	0.056	0.048	0.056	0.000	0.089	0.059	0.155
Chèvre				0.0	0.085	0.075	0.068	0.064	0.050	0.057	0.111	0.033	0.145
Maki					0.0	0.058	0.069	0.075	0.094	0.060	0.048	0.096	0.147
Souris						0.0	0.069	0.046	0.092	0.055	0.090	0.073	0.073
Lapin							0.0	0.051	0.072	0.056	0.101	0.070	0.120
Rat								0.0	0.082	0.048	0.110	0.047	0.139
Bovine									0.0	0.056	0.120	0.065	0.168
Chimpanzé										0.0	0.089	0.059	0.155
Salmon Salar											0.0	0.128	0.156
Âne												0.0	0.147
Grenouille													0.0

TABLE IV.9: Résultat de similarité en termes de position entre les 13 espèces

A partir des résultats obtenus dans les deux tableaux précédents et selon la stratégie que nous avons appliquée, la similarité entre les 13 espèces est calculée par la similarité moyenne de fréquence et position moyenne, Les résultats expérimentaux montrés dans le tableau suivant :

	Gorille	Gallus	Humain	Chévre	Lemur	Souris	Lapin	Rat	Bovine	Chimp -anzze	Salmo salar	Ane	Grenou ille
Gorille		0.113	0.017	0.085	0.070	0.091	0.066	0.078	0.084	0.017	0.118	0.078	0.235
Gallus			0.112	0.114	0.142	0.132	0.118	0.127	0.111	0.112	0.131	0.134	0.176
Humain				0.077		0.087	0.065	0.075	0.079	0.0	0.118	0.072	0.232
Chévre					0.112	0.093	0.082	0.105	0.045	0.077	0.147	0.068	0.210
Lemur						0.096	0.070	0.115	0.114	0.069	0.110	0.123	0.239
Souris							0.082	0.096	0.105	0.087	0.149	0.098	0.098
Lapin								0.105	0.084	0.065	0.138	0.099	0.205
Rat									0.120	0.075	0.125	0.068	0.236
Bovine										0.079	0.156	0.095	0.223
Chimpan zé											0.118	0.072	0.232
Salmo salar												0.153	0.220
Ane													0.241
Grenouill e													

TABLE IV.10: Résultat de similarité en termes de fréquence et position entre les 13 espèces

Comme nous l'avons observé dans le tableau IV.10:

- Il existe une grande similarité entre la séquence protéique du bêta globine de l'humain, gorille et chimpanzé.
- Une similarité remarquable entre le bêta globine de bovine et de la chèvre qui sont de la même famille de mammifères bovidé qui ont aussi certain similarité avec la séquence de l'âne.
- une similitude remarquable aussi entre la la séquence de bêta globine de la souris et du rat et la souris avec le lémur, tel que la souris et le rat appartiennent à la même famille de mammifères Muridae.
- Chacun de la grenouille, Salmo Salar, opossum et Gallus sont loin des autres espèces, parce que l'opossum est l'espèce le plus éloigné des mammifères et le Gallus, Salmon Salar et grenouille sont des animaux de famille mammifères parmi tous les autres animaux de l'ensemble de données.

Selon les valeurs de la distance euclidienne, une certaine similitude entre les 13 espèces, parce qu'elles appartiennent toutes à la famille des animaux. Le résultat obtenu, ce n'est pas un accident, mais montre la relation au sens de l'évolution entre les 13 espèces. Le dendrogramme suivant décrit la relation entre ces espèces, selon notre propre stratégie:

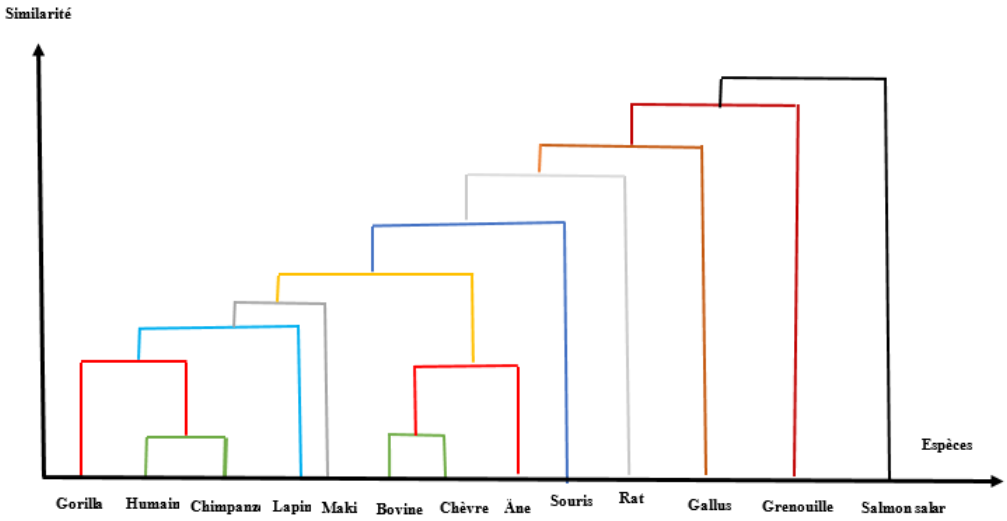

FIGURE IV.2: Le dendrogramme de relation entre les 13 espèces

En conclusion, Les propriétés physico-chimiques des acides aminés sont des informations très importantes, notre proposition est l'exploitation de ces propriétés pour analyser la similitude des séquences de protéines. Nous avons présenté chaque séquence de protéine par quatre séquences symboliques suite aux quatre principales classifications biologiques des acides aminés, regroupées en deux vecteurs de fréquence et de position. Les distances euclidiennes entre les vecteurs (paire à paire) permettant de déduire la similarité entre les séquences de protéines de 13 espèces différentes, les résultats de l'évaluation coïncide avec le sens de l'évolution de ces espèces, ce qui montre l'importance et l'efficacité de notre stratégie d'étude de similarité entre les séquences protéique.

IV.3 Expérimentation pour le regroupement des séquences d'ADN par AC3D

IV.3.1 Ensemble de données

Nous avons utilisé dans notre expérimentation un ensemble de données de biologie moléculaire (séquences d'ADN) obtenu à partir de HS3D [2] est un ensemble de données d'Homo Sapiens (Exon, Intron et Splice), extraites de la banque de donnée GenBank Rel.123.

IV.3.2 Identification de meilleure valeur de N-gram

Comme nous l'avons vu dans le chapitre précédent, avant d'appliquer l'automate cellulaire 3D pour le regroupement des séquences ADN, nous devons passer par les étapes de pré-traitement d'ECD biologique, cette section décrit les résultats obtenus après l'application de la technique de n-gram sur les données pré-traitées.
Pour identifier la meilleure taille de n-grammes, nous avons mené plusieurs expériences

[2]HS3D: ensemble de données de sites d'épilation d'Homo Sapiens est disponible dans: http://www.sci.unisannio. It/docenti/rampone/, consulté le: 2015-08-12

sur la valeur de n entre 2 et 8, afin de ne pas promouvoir une valeur de n sur l'autre ; Cependant, nous avons appliqué une phase de sélection de variable en fonction de la fréquence des n-grammes. Dans le but d'éliminer les variables qui ont un pourcentage de présence inférieur ou égal à un taux x%, x varié entre 5 et 25. Le tableau IV.11 montre les résultats obtenus :

N	Distincte N-grammes	Filtre <= 05%	Filtre <= 15%	Filtre <= 25%
2	16	16	16	16
3	64	64	54	50
4	256	202	105	2
5	1024	735	56	1
6	4096	2365	36	0
7	16384	2653	21	0
8	65536	36521	12	0

TABLE IV.11: Filtrage de n-grammes

On peut constater que, n-gram de taille cinq, six, sept et huit sont nombreux pour le taux 5% mais faibles pour le taux 15%, et trop faibles pour le taux 25%. Le nombre de n-gram de taille 4 pas trop dispersé pour 5% et 15% mais trop faible aussi pour 25%, alors les valeurs de 4 jusqu'à 8 sont trop dispersés par rapport au nombre total de n-gram. Cependant, la valeur 2 et 3 de N montre une faible dispersion de 5% à 25%. Ces résultats montrent que les valeurs 2 et 3 de N sont les meilleures. Donc ces deux valeurs de N sont sélectionnées pour le regroupement par l'automate cellulaire 3D.

IV.3.3 Résultats de l'apprentissage non-supervisée par AC 3D

Comme il est indiqué dans la section précédente, la valeur de n choisi est n=2 et n=3, se basant sur ces deux valeurs de N cette section se devise en deux parties :

1. **Pondération :** Les données d'apprentissage sont représentées par une matrice de fréquence (séquence * descripteur), chaque ligne est un vecteur de séquence et chaque colonne est un descripteur, l'élément à la ligne i et la colonne j est la fréquence du descripteur. les fréquences ont été calculé pour 2 et 3-gram.

2. **Matrice de similarité et clustering :** Pour la classification non-supervisée, le calcul de distance est considéré comme un facteur très important, La collection des séquences sera vue comme étant une matrice de similarité (séquence, séquence), l'élément à la ligne i et la colonne j est la valeur de distance entre deux séquences d'ADN. Pour le calcul de distance entre les vecteurs de séquences nous avons utilisé trois métriques importantes (Euclidienne, Cosine, Minkowsky 4), Les expérimentations de clusteing par l'AC 3D sont effectuées sur la matrice de distance (séquence, séquence), sachant que la taille de n-gram variée entre 2 et 3 avec les trois métriques de distances et un seuil de similarité entre 0.01 et

0.35. Cependant, Les deux tableaux suivants (IV.12, IV.13) montrent le nombre de classe obtenu et les résultats de F-measure et d'entropie en termes de temps.

Distance / seuil	Euclidian				Cosine				Minkowsky 4			
	Classe	Temps	F	E	Classe	Temps	F	E	Classe	Temps	F	E
0.01	13	567	0.58	0.3	2	180	0.73	0.09	8	398	0.60	0.23
0.05	7	347	0.56	0.36	3	141	0.71	0.08	10	463	0.56	0.31
0.10	4	224	0.51	0.31	2	129	0.61	0.21	5	264	0.59	0.27
0.15	3	181	0.57	0.32	2	137	0.69	0.13	4	139	0.54	0.32
0.20	4	222	0.49	0.4	2	129	0.71	0.11	4	226	0.68	0.14
0.25	3	191	0.43	0.41	3	129	0.81	0.05	3	139	0.51	0.32
0.30	2	144	0.54	0.32	3	141	0.75	0.08	3	146	0.49	0.46
0.35	2	186	0.53	0.41	2	180	0.73	0.09	2	141	0.48	0.49

TABLE IV.12: Résultats de clustering par automate cellulaire 3D avec 2-grammes

Distance / Seuil	Euclidian				Cosine				Minkowsky 4			
	Classe	Temps	F	E	Classe	Temps	F	E	Classe	Temps	F	E
0.01	40	2396	0.51	0.32	2	200	0.71	0.10	23	1410	0.58	0.31
0.05	28	1687	0.59	0.25	3	141	0.73	0.08	12	786	0.59	0.29
0.10	14	905	0.63	0.11	2	204	0.70	0.12	6	442	0.70	0.12
0.15	4	329	0.61	0.21	3	281	0.84	0.04	7	483	0.61	0.23
0.20	4	435	0.69	0.12	3	273	0.71	0.13	4	323	0.58	0.46
0.25	6	454	0.68	0.15	4	328	0.64	0.21	5	395	0.69	0.13
0.30	3	261	0.56	0.3	6	386	0.55	0.42	2	146	0.68	0.18
0.35	3	274	0.58	0.27	4	315	0.46	0.50	70	4027	0.58	0.37

TABLE IV.13: Résultats de clustering par automate cellulaire 3D avec 3-grammes

En termes de qualité de clustering par AC 3D, on remarque que :

- La distance euclidienne et Minkowsky 4 ont donné une bonne classification non supervisée, les résultats sont déterminés par les cases jaunes et vertes, montrent que les valeurs de f-mesure et d'entropie avec 3-gram comparé au 2-gram.
- Le résultat obtenu avec la distance cosinus est déterminé par les cases bleu, montre que AC 3D fonctionne mieux en termes de f-mesure et entropie, qu'ils élèvent à 0,84, 0.05 respectivement pour 3-gram.
- Lorsque le nombre de grammes augmente, le temps d'exécution aussi augmente.
- En ce qui concerne le nombre de classes, il varie entre 3 et 4 classes pour les meilleures valeurs de f-mesure et d'entropie.
  Par conséquent, à la fin de cette interprétation, nous avons conclu que :
- Les 3-grammes offrent un bon temps de calcul et des performances de clustering. Ce résultat confirme les résultats décrits dans le domaine de la catégorisation du texte [Seb02].
- La mesure de distance cosinus donne des meilleurs résultats par rapport aux mesures de distances Euclidienne et Minkowsky 4 pour le clustering par d'automates cellulaires 3D.
- La mesure de distance Minkowsky 4 est meilleur que la distance Euclidienne pour cet ensemble de données, et peut être non pour d'autre ensemble selon le type de jeu de données.
- En terme d'entropie, les résultats sont performants ce qui a permet d'obtenir moins de perte d'information.
- Pour la base d'apprentissage de séquences ADN, le nombre de classes (clusters) obtenu est trois (correspondant aux meilleures valeurs de f-mesure et d'entropie).

IV.3.4 Analyse de Clusters

Selon les résultats de la classification non supervisée par AC 3D, nous n'avons conclu que le nombre de groupes (clusters) est trois, se divise les séquences ADN d'apprentissage en trois groupes différents. Cependant, cette section s'intéresse à analyser les clusters obtenus et découvrir la façon selon laquelle les séquences ADN sont distribuées autour des clusters. La distribution est proche de celle de référence, mais il y a certaine déférence montrée par la valeur de f-mesure obtenue. Pour cette raison, et pour répondre à notre problème de prédiction des structures moléculaire et développer les médicaments, et parce que la prédiction est effectuée à base de protéine, nous avons transformé les séquences ADN de chaque cluster en séquences protéique s'accordent au code génétique standard, tel que trois nucléotides successifs représentant un acide aminé et une chaîne d'acides aminés représente une protéine.
Pour chaque chaine protéique obtenue, nous avons analysé les composants de base de protéine (acide aminé), tel que leurs type, fonction et structuration est une clé de base permet de distinguer une protéine de l'autre, et aussi l'apparition d'un acide aminé dans une séquence et pas dans autre signifié certaine différence biologique importante. Nous avons calculé la fréquence d'apparition de chaque acide aminé dans la séquence protéique prend en compte le cluster où se trouve la séquence. Dans ce cas, nous

avons appelé la mesure TF-IDF connue en fouille de texte qui peut répondre à ce besoin parce qu'il prend en considération l'importance du terme dans un document et dans l'ensemble de données, cependant, dans notre cas l'acide aminé définis comme un mot et la séquences définis comme le document où se trouve le mot et le cluster définis comme un ensemble de données. La figure suivante montre la distribution des 20 acides aminés dans chaque cluster selon la valeur de TF-IDF :

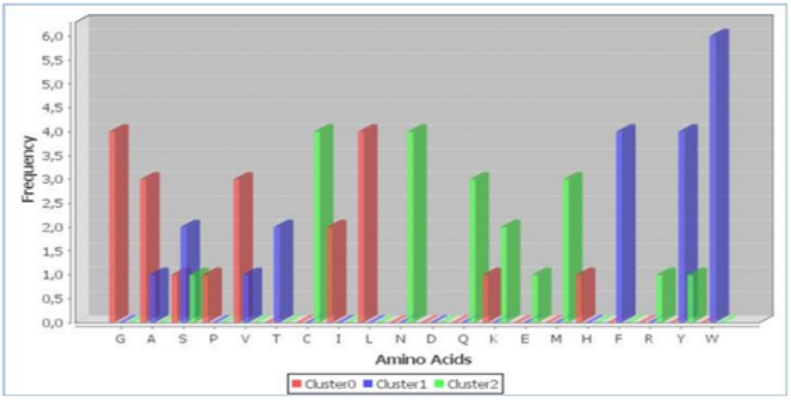

FIGURE IV.3: TF-IDF des acides aminés pour les trois clusters

Les séquences dans les mêmes clusters ont certaines structures similaires ; Certains résultats de recherche montrent les séquences qui ont une certaine similitude dans leurs structures ont généralement une certaine similitude dans leurs fonctions. L'analyse de fréquence des acides aminés dans les clusters montre :

1. **Cluster 1 :** Nous remarquons que il existe une grande proportion des acides aminés (G, A, V, L, I) dans le groupe 1, par rapport à une faible fréquence de (S, P, K, H), cependant, la fréquence des acides aminés restants converge vers 0, on peut dire que ne plus existes par rapport à la fréquence des autres acides aminés.

2. **Cluster 2 :** En ce qui concerne ce groupe, les acides aminés (F, W, Y, S, T) sont les plus fréquents, (V et A) aient une faible fréquence, mais le reste n'existe pas.

3. **Cluster 3 :** le troisième groupe contient les séquences similaires qui ont Les acides aminés N, Q, C, les plus fréquents, M et K avec une grande fréquence, et R, E avec une basse fréquence.
   L'analyse de ces Clusters montre que les séquences de plus de similarité dans leurs structures (composition) sont regroupées dans le même groupe et après la transformation en chaînes d'acides aminés, on constate que ces séquences ont une certaine similitude dans la fréquence des acides aminés impliqués dans leur construction.

IV.3.5 Interprétation

Avant d'expliquer ce qu'on a conclu après l'analyse des clusters obtenus, il s'agit tout d'abord d'expliquer une notion purement biologique, qui placer nos dans le bon chemin pour comprendre l'importance des résultats. Pour synthétiser une molécule(médicaments),

le biologiste identifié la cible de la maladie (fragment ADN) qui ne fonctionne pas correctement et causé l'apparition de la maladie.
Donc, le biologiste doit produire des antibiotiques ou médicaments contre cette maladie, se basant dont son structure de base sur la protéine, si la maladie est génétique, des modifications sont effectuées au fragment ADN, le gène modifié est transféré dans une bactérie d'expérience biologique s'appelle E.Coli. Le Transfert est effectué par la recombinaison de l'ADN isolé avec ADN de la bactérie, qui produit la protéine ou l'hormone souhaitée, et toutes les colonies qui descendent de celle-ci. Les molécules obtenues sont comparées aux molécules produites réellement par la cellule humaine.
Donc pour synthétiser des protéines qui seront utilisées dans le programme de développement du produit pharmaceutique, A quel point nos résultats de clustering peuvent aider à prédire la structure des gènes pour le développement de médicament ?

1. Le biologiste localise les gènes responsables de la maladie.

2. Le biologiste définit la fonction de médicament pour inhiber cette maladie.

3. La détermination de la fonction de médicament peut aider à déterminer les acides aminés qui peuvent être impliqués pour synthétiser une protéine souhaitée (molécule, médicament), donc le biologiste (expert du domaine) propose l'ensemble des acides aminés qui peuvent construire la protéine pour attaquer la maladie.

4. Selon les résultats de distribution des acides aminés dans les trois clusters, un des clusters est sélectionné selon les acides aminés proposés par le biologiste.

5. Le cluster sélectionné permet de déterminer les gènes appropriés (séquences d'ADN) qui ont certaine similarité dans le nombre et la qualité des composants de base.

6. Le cluster localise l'ensemble des gènes qui ont des composants et de structure proche de gêne utilisé pour la production de protéine d'intérêt thérapeutique (médicament).

7. Dans le même cluster sélectionné, on peut définir les gènes les moins concentrés (faible concentration d'acides aminés) avec la même fonction de médicaments.
   La figure suivante montre le processus de prédiction de molécule pour la production de protéine inscrit dans le développement des médicaments.

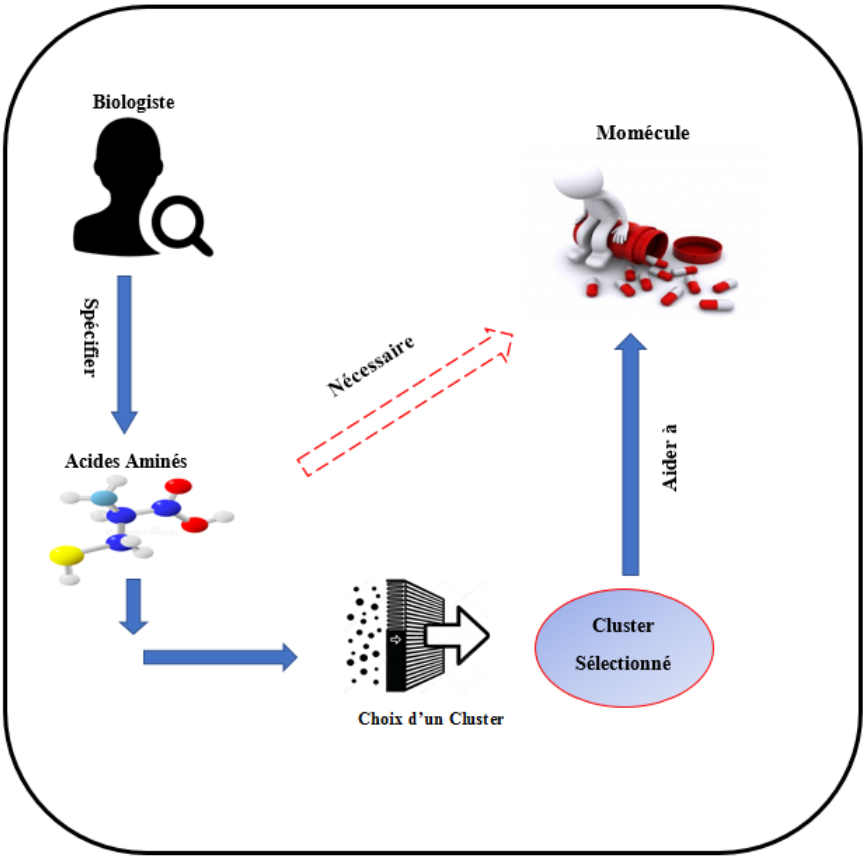

FIGURE IV.4: Sélection de cluster aidant à production de médicament et leurs variations

IV.4 Expérimentation pour la classification de protéine par les règles d'association

IV.4.1 Base de Données protéiques

pour la classification de protéine, nos expériences ont été effectuées sur 5 familles de protéines (collagène alpha-1, bêta-globine, somatotropine, antigène tumoral cellulaire P53, Aldehyde déshydrogénase) extraites de la banque de données UniProt, chaque famille contient un ensemble séquences de protéines de différentes espèces, tel que chaque protéine écrite en forme FASTA. La figure suivante montre la séquence de protéine bêta globine dans la base Uni-prot.

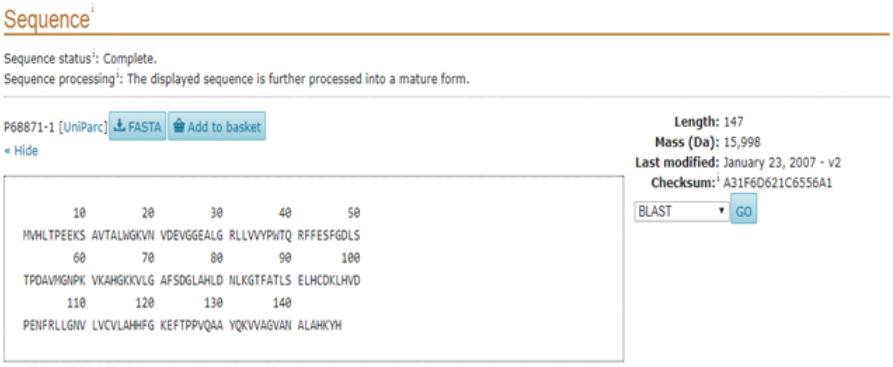

FIGURE IV.5: La séquence de bêta globine de l'humain sous format FASTA dans la base UniProt

Cette section s'intéresse à montrer les expérimentations de notre système de classification sur la base de protéines de 5 familles de protéine, nous l'avons divisée en deux parties : La base d'apprentissage : Contient 60% de la totalité de séquences de la base de données, est utilisé pour appliquer le modèle de classification supervisé de protéine. La base de validation (base de teste) : Contient l'ensemble de données qui reste de la base d'apprentissage (40% de la base), est utilisé pour valider le modèle de classification supervisée avec trois métriques de validation détaillés dans le chapitre précédent : Rappel, Précision et F-mesures.

IV.4.2 Filtrage de n-gramme

Les deux parties de la base ont suivi les étapes décrites dans la conception du système (III.4.7.2), la première étape s'inscrit dans le pré-traitement de séquences protéiques, la deuxième est la transformation de la séquence de protéine à un ensemble de descripteurs par la technique de n-gramme, pour choisir la meilleure valeur de n-gramme, nous avons effectué des expérimentations sur la base d'apprentissage. Selon notre travail de classification non-supervisée par AC 3D [KHA16] et le travail publié par [MER06], les fréquences du n-gramme de taille quatre, cinq, six, sept et huit sont trop dispersées et nombreuses, les meilleures valeurs de N sélectionnées sont deux et trois. Pour cela, nous avons effectué des expériences avec 2 et 3-grammes ; le tableau IV.14 montre le nombre de descripteurs produits pour les deux valeurs de N.

Filter% / N	Distincts N- grams	Filtre <=05%	Filtre <=15%	Filtre <=25%
2	448	424	383	342
3	6195	6170	6129	6008

TABLE IV.14: Filtrage du n-gramme(CSP)

Selon le tableau ci-dessus, on constate que 2 et 3 grammes ne sont pas trop dispersés de 5% à 25%, mais le nombre de descripteurs pour 3-grammes est très grand que 2-grammes, qui équivalant à 448. Dans notre système, les descripteurs sont définis comme des items pour l'extraction des règles d'association. Nous savons bien que, si le nombre d'attributs est important, nous obtenons un grand nombre de règles d'association. Donc nous revenons au problème d'origine de fouille données ? " Comment extraire les informations significatives à partir d'un grand ensemble de données », nous concluons que, la meilleure taille de n-gram est 2, ce qui offre une faible dispersion et un nombre raisonnable d'attributs qui facilitent l'extraction des règles d'association.

IV.4.3 Extraction des règles d'association

Pour l'extraction des règles d'association, et comme nous avons détaillé dans la partie conception, chaque séquence protéique est défini comme une transaction (identifiant de la protéine dans le format FASTA) et les descripteurs définis l'ensemble des items (itemset).Nous avons appliqué l'algorithme Apriori sur la base d'apprentissage (60%) de Uni-Prot, avec la valeur de confiance 0,9 et la valeur de support 0,1.
Pour le filtrage des règles d'association obtenu et la sélection celles qui sont significatives pour notre étude.
Nous avons appliqué la stratégie de sélection des règles significatives (bien détaillé au stade de modélisation), le tableau ci-dessous décrit le nombre total de règles obtenues par l'algorithme Apriori avec le nombre et le pourcentage de règles significatives.

Règles / Classes	Nombre totale des règles d'associations	Nombre des règles significatives	pourcentage (%)
Beta Globine	240000	7049	2.10%
Collagen alpha-1	40000	2071	5.17%
Somatotropin	80000	8963	11.20%
Cellular tumor antigen p53	128744	7938	6.16%
Aldehyde dehydrogenase	7850	773	9.84%

TABLE IV.15: Les Règles significatives pour les cinq classes de protéines

Afin d'analyser le tableau ci-dessus nous avons remarqué que :

- Il existe un grand nombre de règles, jusqu'à un maximum de 240000 règles pour la bêta-globine et jusqu'à un minimum de 7850 règles pour l'Aldéhyde déshydrogénase.
- Lorsque le nombre de règles totales augmente, le nombre de règles significatives augmente aussi.

- Le nombre de règles significatives représente 2,1% à 11,2% du nombre total de règles. Donc, nous avons conclu que : La stratégie de filtrage des règles d'association adopté dans notre système est importante, sert à réduire de manière efficace le nombre des règles d'association, qui facilite la manipulation d'un grande nombre de règles d'association ; qui sont appelés dans la suite par notre système pour la classification supervisée de protéines.

IV.4.4 Mesure des performances de notre système de classification supervisée(CSP)

Afin de représenter la base de protéines en matrice de données (Instances / Attributs) avec des poids de fréquence. Nous avons extrait les règles d'association et obtenu en totalité une base de 26794 règles significatives. les séquences de la base de teste seront vectorisés suivant les étapes "pré-traitement, extraction de descripteurs, codage: (attribut-valeur) détaillés dans le chapitre précédent. Sur la base des règles, nous avons appliqué notre modèle de classification (CSP)[KHA18].
Pour mieux valider l'efficacité de (CSP), nous avons comparé son rappel, sa précision et f-mesure avec d'autres modèles de classification supervisée de même type (classificateur basé sur les règles d'association). Nous avons mis en œuvre les méthodes de classification de la plate-forme WEKA suivants :

- **PART:**
  Selon [PJB14], PART est un algorithme efficace, produit un ensemble de règles «listes de décision» et compare les nouvelles données avec chaque règle de la liste, en fonction du résultat obtenu, à chaque itération PART crée un arbre de décision partielle (C4.5) et choisir la meilleure feuille comme règle pour la classification.

- **One-R:**
  One-R [BD06] est un algorithme simple de classification basé sur les règles d'association, peut déduire des règles de classement généralement simples mais précises d'un ensemble d'instances. Sert à créer une règle pour chaque attribut de donnée d'apprentissage, puis choisir la règle avec le taux d'erreur minimum, One-R est également capable de gérer les attributs numériques et les valeurs manquantes.

- **JRip** (Java Repeated Incremental Pruning) [Raj+11] :
  JRip est l'un des algorithmes basiques et les plus populaires. A chaque étape générée un ensemble de règles pour chaque classe de l'ensemble de données, à l'aide de l'erreur réduite incrémentielle (RIPPER), les classes sont examinées en taille croissante. L'état d'arrêt basé sur la longueur de description minimale (MDL). Une fois qu'un ensemble de règles est produit la classification est faite.

- **Table de décision** [Koh95]
  Table de décision c'est un nombre ordonné de règles (si-alors), défini comme une méthode précise pour la prédiction numérique à partir d'arbres de décision, plus compactes et plus compréhensibles que les arbres de décision.

- **Règle conjoncturelle (Conjunctive Rule) :**
  Cette classe de classificateur manipule une seule règle conjoncturelle. Une règle consiste en ensemble des antécédents et une conséquence (valeur de classe/étiquette). Pour la classification d'une nouvelle instance, en calculant le gain d'information de chaque antécédent en utilisant l'erreur réduite d'élagage (REP). La moyenne pondérée du taux de précision des données d'élagage est utilisée pour la classification. [PJB14].
  Les résultats de comparaison et d'évaluation sont montrés dans le tableau ci-dessous :

Classificateur / Mesure	Notre classificateur (CSP)	Conjunctive Rule	One-R	J-Rip	Decision Table	PART
Rappel	0.92	0.651	0.735	0.882	0.824	0.842
Précession	0.901	0.473	0721	0.879	0.85	0.843
F-mesure	0.941	0.544	0.696	0.878	0.809	0.842

TABLE IV.16: Rappel, précision, f-mesure pour mesurer la performance de modèle de classification CSP

IV.4.5 Évaluation

Comme nous l'avons noté dans le tableau ci-dessus, nous avons procédé à la comparaison de notre système avec les cinq classificateurs weka sur la base de données protéique (UniProt), la base d'entrée est la matrice de fréquence (Instance/Attributs) que nous avons construits précédemment. Selon les résultats obtenus nous évaluions l'efficacité de notre classificateur en terme de :

IV.4.5.1 Rappel

La valeur du rappel égal à 0,92 est considérée comme une bonne valeur comparativement aux autres classificateurs varie entre 0.651 pour le classificateur de règle connective et un maximum rappel 0.882 pour le classificateur JRip, selon le rappel de notre systeme ça signifie que la quasi-totalité des solutions sont pertinentes.

IV.4.5.2 Précision

Les résultats obtenus révèlent que notre système a un meilleur score de précision égale à 0.941, avec un score de précision très faible pour la méthode "Règle conjonctive", et un score moyen pour chaque un de (table de décision, JRip, PART et One-R), ces résultats de précision signifient que notre classificateur est très précis à refuser les solutions non pertinentes comparé avec les autres systèmes.

IV.4.5.3 F-mesure

Pour la mesure qui combine la précision et le rappel (f-mesure), nous avons obtenu un meilleur score de 0,92, ce qui prouve que notre classificateur a une bonne capacité à donner les solutions pertinentes et à refuser d'autres.
Par conséquent, le résultat obtenu prouve la performance d'un classificateur basé sur la technique du n-gramme et les règles d'association significatives pour la classification des séquences complexes de protéines sous la structure primaire de base.

IV.5 Conclusion

Ce chapitre a été inscrit dans le cadre pratique de nos travaux de thèse, présente les résultats des expérimentations de nos approches effectuées sur l'ensemble des données biologiques (ADN/protéine). Afin d'évaluer leurs importances et efficacités par des métriques d'évaluation et des systèmes de comparaison, nos travaux de thèse ont présenté de bons résultats et efficacités. Pour l'étude de similarité nos méthodes ont présenté une efficacité apparue par les résultats de similarité entre les séquences ADN de différentes espèces qui coïncident avec la relation au sens de l'évolution existe réellement entre ces espèces. D'un autre côté, dans l'objectif de procéder les techniques de fouille de données afin d'extraire des connaissances biologiques. Les résultats d'évaluation montrent l'importance de transformer les séquences biologiques en ensembles de descripteurs par la technique de n-gramme avant d'appliquer les techniques de fouille. L'analyse de résultat de clustering par l'automate cellulaire 3D des séquences ADN, nous a donnés une nouvelle vision pour la prédiction des structures des protéines impliquées dans le développement des médicaments. L'extraction des règles d'association entre les composants de base de protéine et leur implémentation pour construire un classificateur de protéine donne de bons résultats de classification supervisée par apport des autres classificateurs de même type.

Conclusion Générale

Dans cette thèse, une nouvelle problématique a été posée, s'inscrit dans le domaine d'analyse et de fouille de données complexes, plus précisément les données biologiques moléculaires, dans l'objectif de traiter la complexité des données biologiques d'ADN et de protéines par les techniques d'analyse et de fouille de données.
Nous avons d'abord commencé par un état de l'art qui couvre les différents types, catégories et natures de données complexes (les données multi-structures, multi-modèles, multi-versions, les données scientifiques, les données biologiques, etc.), nous avons bien détaillé les données biologiques complexes qui nous a intéressées le plus dans nos travaux. Ensuite nous avons décrit les problèmes de bio-informatiques et le processus d'ECD biologiques avec les méthodes et techniques d'analyse et fouille de données liées à résoudre ces problèmes. Ce qui a conduit à résoudre la problématique posée.
Pour montrer nos contributions, nous avons décrit dans les deux derniers chapitres l'étape de modélisation et d'expirimention, constituant deux parties importantes.

1. **Etude de similarité entre les séquences d'ADN et de protéine :**Cette partie a été consacrée à la définition de nos méthodes d'étude de similarité entre les séquences d'ADN et de protéines, nous avons traité la complexité des composants de base de deux type de séquences, par la proposition de nouveaux représentations basant sur les propriétés chimiques et physiques de la structure de base (nucléotidique ou acide aminée). Permet de fournir des informations supplémentaires et contribuent à améliorer la recherche de similarité.

2. **L'élaboration de processus d'ECD biologique et méthodes de fouille :** Cette partie a été liée à l'élaboration des étapes de processus ECD pour le traitement de deux problèmes importants de bio-informatique, prend en priorité la complexité des séquences biologiques :

    - Prédiction de séquençage de molécules pour le développement de médicaments à partir des séquences d'ADN.
    - Classification supervisé des séquences protéiques.

Les deux problèmes ont suivis les mêmes étapes de processus ECD, définition de problème, collection de données, le pré-traitement des données, la modélisation et la validation. Pour chaque étape, nous avons implémentés les techniques de traitements spécifiques pour passer d'une étape à l'autre.
Pour le premier problème, nous avons décrit dans l'étape de modélisation l'algorithme de clustering AC 3D sur les séquences ADN, pour le deuxième problème, nous avons détaillé notre classificateur de protéine se basant sur la base de règles d'association entre les composants de base (acides aminé)des séquences protéiques .
L'évaluation de nos travaux d'étude de similarité d'ADN et de protéine sur les bases de

bêta globine, montrent un bon résultat de similarité coïncide avec la relation au sens de l'évolution entre un ensemble des espèces.
La classification non-supervisée des séquences d'ADN par l'automate cellulaire 3D, a montré un bon résultat en terme de rappel, précision et f-mesure et d'entropie. Cependant, l'analyse des clusters obtenus nous a donnés une nouvelle vision pour la prédiction de la structure des molécules et le développement de médicaments.
La classification supervisée des séquences de protéine par notre classificateur est basée sur la technique de n-gramme et les règles d'associations pertinentes, a montré un bon résultat en termes de rappel, précision, et f-mesure par apport des autres algorithmes de classification de la plateforme WEKA.
Finalement, nous sommes arrivés à la conclusion suivante : Le traitement et l'analyse des données complexes pour l'extraction de l'information appropriée aux utilisateurs doit avoir une analyse préalable des propriétés et natures de la données complexes, avec un bon choix d'algorithmes et techniques de traitement adaptables avec le type de données manipulés. Ce travail est une porte vers d'autre recherche et travaux future montrés dans la section suivante.

Perspectives

Travaux Future :
Les principaux travaux qu'on a planifiés à faire au futur sont :

- Proposition d'un nouveau algorithme d'extraction d'association adapté avec les données biologiques (ADN/ protéine).

- L'extraction de règles d'association des séquences d'ADN pour détecter les parties codantes et non codantes.

- Nous prévoyons également d'appliquer l'algorithme GenMiner et de le rendre adaptable aux séquences protéiques de base, puis nous appliquerons notre classificateur sur la base de règles obtenus.

- Nous avons prévu de développer un système en ligne pour la comparaison de séquence(ADN /Protéine) pair et multiple basant sur les propriétés physiques et chimiques.

- Nous prévoyons aussi à développer un système d'analyse des anomalies de séquences d'ADN de l'être humain causant de la maladie génétique "cancer du sein", pour la prédiction de différents séquences qui peuvent développer par le corps du patient pour attaquer l'effet de médicaments contre la maladie, nous appliquerons les méthodes de fouille de données.

Publications de L'auteur

Conférences nationales

- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, Structure Molecules Prediction from DNA Sequences Based On Clustering By 3d Cellular Automata Approach And N-Grams Technique , 1st National Conference on Embedded and Distributed Systems (EDiS'2015), Oran, Algérie, 15 – 16 Novembre 2015.

Conférences internationales

- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, Similarity Analysis Of DNA Sequences Based On The Chemical Properties Of Nucleotide Bases, Frequency And Position Of Group Mutations, Sixth International Conference on Computer Science and Information Technology (CCSIT'2016), Zurich, Suisse le 2-3 janvier 2016.
- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, DNA Sequences Analysis Based On The Chemical Properties Frequency And Position Of Nucleotide Bases, First International Conference on Business Intelligence and Applications (ICBIA - 2016),Blida, Algérie, 01 – 03 Mars 2016.
- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, Protein sequences similarity analysis based on physico-chemical properties, the Maltepe University, International Student Congress, Istanbul, Turkey, 28-29 Avril 2016.
- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, New Classification System for Protein Sequences, international Conference on Embedded and Distributed Systems (EDiS'2017), Oran, Algeria, 17 – 18 December 2017.

Journaux

- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, DNA Data Clustering by Combination Of 3d Cellular Automata And N-Grams For Structure Molecule Prediction, 2016, International Journal of Bioinformatics Research and Applications (IJIBRA), Vol.12, No.4, pp.299 – 311,2016.
- Fatima Kabli, Reda Mohamed Hamou, Abdelmalek Amine, 2017, Protein classification using n-gram technique and association rules, International Journal of Software Innovation (IJSI)volume 6 issue 2, articles 6, pp. 77-89,2018.

Chapitres de Livre

- Fatima Kabli, Complex Biological Data Mining And Knowledge Discovery, Chapter 16: Handbook of Research on Biomimicry in Information Retrieval and Knowledge Management, pp.303–320, 2018

Bibliographie

[AAMA04] John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken. "Combining information extraction with genetic algorithms for text mining". In: *IEEE Intelligent Systems* 19.3 (2004), pp. 22–30 (cit. on p. 48).

[Aas01] Kjersti Aas. "Microarray data mining: A survey". In: *NR Note, SAMBA, Norwegian Computing Center* (2001) (cit. on p. 23).

[Abd10] MANSOUL Abdelhak. "Fouille de Données Biologiques: Etude Comparative et Experimentation". In: (2010) (cit. on pp. 20, 22, 80).

[Alb+02] B Alberts et al. *Molecular biology of the cell 4th edition: International student edition*. 2002 (cit. on p. 36).

[Alo+99] Uri Alon et al. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". In: *Proceedings of the National Academy of Sciences* 96.12 (1999), pp. 6745–6750 (cit. on p. 45).

[Alt+90] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410 (cit. on p. 30).

[AM06] Sophia Ananiadou and John McNaught. *Text mining for biology and biomedicine*. Artech House London, 2006 (cit. on p. 58).

[Anf73] Christian B Anfinsen. "Principles that govern the folding of protein chains". In: *Science* 181.4096 (1973), pp. 223–230 (cit. on p. 22).

[AS+94] Rakesh Agrawal, Ramakrishnan Srikant, et al. "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994, pp. 487–499 (cit. on p. 53).

[Att+11] TK Attwood et al. "Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective". In: *Bioinformatics-trends and methodologies*. InTech, 2011 (cit. on p. 26).

[Au+05] Wai-Ho Au et al. "Attribute clustering for grouping, selection, and classification of gene expression data". In: *IEEE/ACM transactions on computational biology and bioinformatics* 2.2 (2005), pp. 83–101 (cit. on p. 45).

[Ave+44] Oswald T' Avery et al. "SYMPOSIUM FEBRUARY 2, 1979". In: *The Journal of Experimental Medicine* 79.2 (1944), pp. 137–158 (cit. on p. 26).

[AZ12] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012 (cit. on p. 55).

[Bal+94] Pierre Baldi et al. "Hidden Markov models of biological primary sequence information." In: *Proceedings of the National Academy of Sciences* 91.3 (1994), pp. 1059–1063 (cit. on p. 34).

[Ban07] Sanghamitra Bandyopadhyay. *Analysis of biological data: a soft computing approach*. Vol. 3. World Scientific, 2007 (cit. on p. 48).

[Bas+02] Yves Bastide et al. “Pascal: un algorithme d’extraction des motifs fréquents”. In: *Techniques et Sciences Informatiquess* 21.1 (2002), pp. 65–95 (cit. on p. 80).

[BD06] Gaya Buddhinath and Damien Derry. “A simple enhancement to one rule classification”. In: *Department of Computer Science & Software Engineering. University of Melbourne, Australia* (2006) (cit. on p. 103).

[BFL05] Konstantinos Blekas, Dimitrios I Fotiadis, and Aristidis Likas. “Motif-based protein sequence classification using neural networks”. In: *Journal of Computational Biology* 12.1 (2005), pp. 64–82 (cit. on p. 50).

[BK06] Fabian Birzele and Stefan Kramer. “A new representation for protein secondary structure prediction based on frequent patterns”. In: *Bioinformatics* 22.21 (2006), pp. 2628–2634 (cit. on p. 55).

[BLM04] Daniel G Brown, Ming Li, and Bin Ma. “Homology search methods”. In: *The Practical Bioinformatician*. World Scientific, 2004, pp. 217–244 (cit. on p. 58).

[BM02a] Sanghamitra Bandyopadhyay and Ujjwal Maulik. “An evolutionary technique based on K-means algorithm for optimal clustering in RN”. In: *Information Sciences* 146.1 (2002), pp. 221–237 (cit. on p. 48).

[BM02b] Sanghamitra Bandyopadhyay and Ujjwal Maulik. “Genetic clustering for automatic evolution of clusters and application to image classification”. In: *Pattern recognition* 35.6 (2002), pp. 1197–1208 (cit. on p. 48).

[BMM07] Sanghamitra Bandyopadhyay, Ujjwal Maulik, and Anirban Mukhopadhyay. “Multiobjective genetic clustering for pixel classification in remote sensing imagery”. In: *IEEE transactions on Geoscience and Remote Sensing* 45.5 (2007), pp. 1506–1511 (cit. on p. 48).

[BMS00] Allen C Browne, Alexa T McCray, and Suresh Srinivasan. “The specialist lexicon”. In: *National Library of Medicine Technical Reports* (2000), pp. 18–21 (cit. on p. 56).

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 49).

[Bro+00] Michael PS Brown et al. “Knowledge-based analysis of microarray gene expression data by using support vector machines”. In: *Proceedings of the National Academy of Sciences* 97.1 (2000), pp. 262–267 (cit. on p. 50).

[BTT05] Renato Bueno, Agma JM Traina, and Caetano Traina. “Accelerating approximate similarity queries using genetic algorithms”. In: *Proceedings of the 2005 ACM symposium on Applied computing*. ACM. 2005, pp. 617–622 (cit. on p. 48).

[CB85] Athel Cornish-Bowden. “IUPAC-IUB symbols for nucleotide nomenclature”. In: *Nucleic Acids Res* 13.3021 (1985), p. 30 (cit. on p. 61).

[CCF06] Martine Cadot, Pascal Cuxac, and Claire François. “Aide à l’interprétation des règles d’association composées.” In: *EGC 2006*. 2006, pp. 31–37 (cit. on p. 80).

[CD05] Rui Chi and Kequan Ding. “Novel 4D numerical representation of DNA sequences”. In: *Chemical Physics Letters* 407.1 (2005), pp. 63–67 (cit. on p. 33).

[Ced+97] Juan Cedano et al. “Relation between amino acid composition and cellular location of proteins”. In: *Journal of molecular biology* 266.3 (1997), pp. 594–600 (cit. on p. 58).

[CH05] Aaron M Cohen and William R Hersh. “A survey of current work in biomedical text mining”. In: *Briefings in bioinformatics* 6.1 (2005), pp. 57–71 (cit. on p. 56).

[Chi+00] Patrick Chiu et al. “A genetic algorithm for video segmentation and summarization”. In: *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*. Vol. 3. IEEE. 2000, pp. 1329–1332 (cit. on p. 48).

[CHY96] Ming-Syan Chen, Jiawei Han, and Philip S. Yu. “Data mining: an overview from a database perspective”. In: *IEEE Transactions on Knowledge and data Engineering* 8.6 (1996), pp. 866–883 (cit. on p. 38).

[Coh05] AM Cohen. “Linking biological literature, ontologies and databases: mining biological semantics”. In: *Association for Computational Linguistics*. 2005, pp. 17–24 (cit. on p. 56).

[CRA00] Jeffrey T Chang, Soumya Raychaudhuri, and Russ B Altman. “Including biological literature improves homology search”. In: *Biocomputing 2001*. World Scientific, 2000, pp. 374–383 (cit. on p. 58).

[Cri70] Francis Crick. “Central dogma of molecular biology”. In: *Nature* 227.5258 (1970), pp. 561–563 (cit. on p. 15).

[Dav01] W Mount David. “Bioinformatics: sequence and genome analysis”. In: *Google Scholar* (2001) (cit. on pp. 19, 21).

[Dav91] Lawrence Davis. “Handbook of genetic algorithms”. In: (1991) (cit. on p. 47).

[DD06] Susmita Datta and Somnath Datta. “Evaluation of clustering algorithms for gene expression data”. In: *BMC bioinformatics* 7.4 (2006), S17 (cit. on pp. 43, 44).

[DGP05] G Desjardins, R Godin, and R Proulx. “A genetic algorithm for text mining”. In: *WIT Transactions on Information and Communication Technologies* 35 (2005) (cit. on p. 48).

[DK02] Mukund Deshpande and George Karypis. “Evaluation of techniques for classifying biological sequences”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2002, pp. 417–431 (cit. on p. 51).

[DK03] Doulaye Dembélé and Philippe Kastner. “Fuzzy C-means method for clustering microarray data”. In: *Bioinformatics* 19.8 (2003), pp. 973–980 (cit. on pp. 44, 45).

[DKS95] James Dougherty, Ron Kohavi, and Mehran Sahami. “Supervised and unsupervised discretization of continuous features”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 194–202 (cit. on p. 40).

[DMS99] Claude Duvallet, Zoubir Mammeri, and Bruno Sadeg. “Les SGBD temps réel”. In: *Technique et science informatiques* 18.5 (1999), pp. 479–517 (cit. on p. 7).

[DS12] Arundhati Deka and Kandarpa Kr Sarma. “Artificial neural network aided protein structure prediction”. In: *Int. J. Comput. Appl* 48.18 (2012), pp. 33–37 (cit. on pp. 50, 51).

[DSO78] MO Dayhoff, RM Schwartz, and BC Orcutt. “22 A Model of Evolutionary Change in Proteins”. In: *Atlas of protein sequence and structure*. Vol. 5. National Biomedical Research Foundation Silver Spring, MD, 1978, pp. 345–352 (cit. on pp. 16, 29).

[DUDA06] Ramón Díaz-Uriarte and Sara Alvarez De Andres. “Gene selection and classification of microarray data using random forest”. In: *BMC bioinformatics* 7.1 (2006), p. 3 (cit. on p. 49).

[Edd+95] Sean R Eddy et al. “Multiple alignment using hidden Markov models.” In: *Ismb*. Vol. 3. 1995, pp. 114–120 (cit. on p. 51).

[Eis+98] Michael B Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868 (cit. on pp. 44, 45).

[Esk+03] Eleazar Eskin et al. “Mismatch string kernels for SVM protein classification”. In: *Advances in neural information processing systems*. 2003, pp. 1441–1448 (cit. on p. 34).

[Est+96] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231 (cit. on p. 46).

[EZ13] Mourad Elloumi and Albert Y Zomaya. *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*. Vol. 23. John Wiley & Sons, 2013 (cit. on pp. 49, 51, 52).

[Fay+96] Usama M Fayyad et al. *Advances in knowledge discovery and data mining*. Vol. 21. AAAI press Menlo Park, 1996 (cit. on pp. 38, 39, 42).

[FM07] Limin Fu and Enzo Medico. “FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data”. In: *BMC bioinformatics* 8.1 (2007), p. 3 (cit. on p. 44).

[FPSM92] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. “Knowledge discovery in databases: An overview”. In: *AI magazine* 13.3 (1992), p. 57 (cit. on p. 41).

[GB02] Bart Goethals and Jan Van den Bussche. “Relational association rules: getting W armer”. In: *Pattern Detection and Discovery* (2002), pp. 145–159 (cit. on p. 80).

[GH04] Toni Gabaldón and Martijn A Huynen. “Prediction of protein function and pathways in the genome era”. In: *Cellular and Molecular Life Sciences CMLS* 61.7-8 (2004), pp. 930–944 (cit. on p. 57).

[GJ16] Leonardo D Garma and André H Juffer. “Comparison of non-sequential sets of protein residues”. In: *Computational biology and chemistry* 61 (2016), pp. 23–38 (cit. on p. 37).

[GL08] Jiajun Gu and Jun S Liu. “Bayesian biclustering of gene expression data”. In: *BMC genomics* 9.1 (2008), S4 (cit. on p. 44).

[GMJ16] Leonardo D Garma, Milagros Medina, and André H Juffer. “Structure-based classification of FAD binding sites: A comparative study of structural alignment tools”. In: *Proteins: Structure, Function, and Bioinformatics* 84.11 (2016), pp. 1728–1747 (cit. on p. 37).

[Gol+89] David E Goldberg et al. *Genetic algorithms in search, optimization, and machine learning*. 1989 (cit. on p. 47).

[Gor08] Gavin J Gordon. *Bioinformatics in cancer and cancer therapy*. Springer Science & Business Media, 2008 (cit. on p. 38).

[GRB01] Xiaofeng Guo, Milan Randic, and Subhash C Basak. “A novel 2-D graphical representation of DNA sequences of low degeneracy”. In: *Chemical Physics Letters* 350.1 (2001), pp. 106–112 (cit. on p. 33).

[Ham+12] Reda Mohamed Hamou et al. “Visualization and clustering by 3D cellular automata: Application to unstructured data”. In: *arXiv preprint arXiv:1211.5766* (2012) (cit. on pp. 74–77).

[Han+96] Jiawei Han et al. “DMQL: A data mining query language for relational databases”. In: *Proc. 1996 SiGMOD*. Vol. 96. 1996, pp. 27–34 (cit. on p. 80).

[Her+99] Ralf Herwig et al. “Large-scale clustering of cDNA-fingerprinting data”. In: *Genome research* 9.11 (1999), pp. 1093–1105 (cit. on p. 44).

[HF99] Jiawei Han and Yongjian Fu. “Mining multiple-level association rules in large databases”. In: *IEEE Transactions on knowledge and data engineering* 11.5 (1999), pp. 798–805 (cit. on p. 80).

[HGN00] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. “Algorithms for association rule mining—a general survey and comparison”. In: *ACM sigkdd explorations newsletter* 2.1 (2000), pp. 58–64 (cit. on p. 80).

[HH92] Steven Henikoff and Jorja G Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919 (cit. on p. 29).

[HK96] Richard Hughey and Anders Krogh. “Hidden Markov models for sequence analysis: extension and analysis of the basic method”. In: *Bioinformatics* 12.2 (1996), pp. 95–107 (cit. on p. 51).

[HPK11] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011 (cit. on pp. 41, 43).

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. “Mining frequent patterns without candidate generation”. In: *ACM sigmod record*. Vol. 29. 2. ACM. 2000, pp. 1–12 (cit. on p. 54).

[HR00] Sridhar S Hannenhalli and Robert B Russell. “Analysis and prediction of functional sub-types from protein sequence alignments1”. In: *Journal of molecular biology* 303.1 (2000), pp. 61–76 (cit. on p. 57).

[Huy+98] Martijn Huynen et al. “Homology-based fold predictions for Mycoplasma genitalium proteins1”. In: *Journal of molecular biology* 280.3 (1998), pp. 323–326 (cit. on p. 57).

[Int13] Phenotype-Genotype Integrator. "Bethesda: National Center for Biotechnology Information". In: *Back to cited text* 14 (2013) (cit. on p. 57).

[JDH00] Tommi Jaakkola, Mark Diekhans, and David Haussler. "A discriminative framework for detecting remote protein homologies". In: *Journal of computational biology* 7.1-2 (2000), pp. 95–114 (cit. on p. 34).

[Joh+07] Helen L Johnson et al. "Corpus refactoring: a feasibility study". In: *Journal of Biomedical Discovery and Collaboration* 2.1 (2007), p. 4 (cit. on p. 56).

[JTZ04a] Daxin Jiang, Chun Tang, and Aidong Zhang. "Cluster analysis for gene expression data: a survey". In: *IEEE Transactions on knowledge and data engineering* 16.11 (2004), pp. 1370–1386 (cit. on p. 43).

[JTZ04b] Daxin Jiang, Chun Tang, and Aidong Zhang. "Cluster analysis for gene expression data: a survey". In: *IEEE Transactions on knowledge and data engineering* 16.11 (2004), pp. 1370–1386 (cit. on p. 44).

[KHA16] Fatima Kabli, Reda Mohamed Hamou, and Abdelmalek Amine. "DNA data clustering by combination of 3D cellular automata and n-grams for structure molecule prediction". In: *International Journal of Bioinformatics Research and Applications* 12.4 (2016), pp. 299–311 (cit. on p. 101).

[KHA18] Fatima Kabli, Reda Mohamed Hamou, and Abdelmalek Amine. "Protein Classification Using N-gram Technique and Association Rules". In: *International Journal of Software Innovation (IJSI)* 6.2 (2018), pp. 77–89 (cit. on p. 103).

[Kim+03] J-D Kim et al. "GENIA corpus—a semantically annotated corpus for biotextmining". In: *Bioinformatics* 19.suppl_1 (2003), pp. i180–i182 (cit. on p. 56).

[Koh95] Ron Kohavi. "The power of decision tables". In: *European conference on machine learning*. Springer. 1995, pp. 174–189 (cit. on p. 103).

[Kro+94] Anders Krogh et al. "Hidden Markov models in computational biology: Applications to protein modeling". In: *Journal of molecular biology* 235.5 (1994), pp. 1501–1531 (cit. on p. 34).

[KV05] Martin Krallinger and Alfonso Valencia. "Text-mining and information-retrieval services for molecular biology". In: *Genome biology* 6.7 (2005), p. 224 (cit. on p. 56).

[Len02] Thomas Lengauer. "Bioinformatics-From Genomes to Drugs Volume II: Applications". In: *METHODS AND PRINCIPLES IN MEDICINAL CHEMISTRY* 14.2 (2002) (cit. on p. 37).

[Lia+07] Bo Liao et al. "On the similarity of DNA primary sequences based on 5-D representation". In: *Journal of Mathematical Chemistry* 42.1 (2007), pp. 47–57 (cit. on p. 33).

[Liu+06] Xiao Qing Liu et al. "PNN-curve: A new 2D graphical representation of DNA sequences and its application". In: *Journal of theoretical biology* 243.4 (2006), pp. 555–561 (cit. on p. 33).

[LLB09] Lee J Lancashire, Christophe Lemetre, and Graham R Ball. "An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies". In: *Briefings in bioinformatics* 10.3 (2009), pp. 315–329 (cit. on p. 50).

[LN03] Li Liao and William Stafford Noble. "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships". In: *Journal of computational biology* 10.6 (2003), pp. 857–868 (cit. on p. 34).

[LP85] David J Lipman and William R Pearson. "Rapid and sensitive protein similarity searches". In: *Science* 227.4693 (1985), pp. 1435–1441 (cit. on p. 31).

[LRW95] Christian M-R Lemer, Marianne J Rooman, and Shoshana J Wodak. "Protein structure prediction by threading methods: evaluation of current techniques". In: *Proteins: Structure, Function, and Bioinformatics* 23.3 (1995), pp. 337–355 (cit. on p. 35).

[Lu+04a] Yi Lu et al. "FGKA: A fast genetic k-means clustering algorithm". In: *Proceedings of the 2004 ACM symposium on Applied computing*. ACM. 2004, pp. 622–623 (cit. on p. 48).

[Lu+04b] Yi Lu et al. "Incremental genetic K-means algorithm and its application in gene expression data analysis". In: *BMC bioinformatics* 5.1 (2004), p. 172 (cit. on p. 48).

[Mac+67] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297 (cit. on p. 44).

[MB03] Ujjwal Maulik and Sanghamitra Bandyopadhyay. "Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification". In: *IEEE Transactions on geoscience and remote sensing* 41.5 (2003), pp. 1075–1081 (cit. on p. 48).

[MBJ00] Liam J McGuffin, Kevin Bryson, and David T Jones. "The PSIPRED protein structure prediction server". In: *Bioinformatics* 16.4 (2000), pp. 404–405 (cit. on p. 57).

[MCM00] EJ Moler, ML Chow, and IS Mian. "Analysis of molecular profile data using generative and discriminative methods". In: *Physiological Genomics* 4.2 (2000), pp. 109–126 (cit. on p. 50).

[MER06] F Mhamdi, M Elloumi, and R Rakotomalala. "Extraction et Sélection des n-grammes pour le Classement des Protéines". In: *Atelier EGC 2006, Lille ENIC, Villeneuved'Ascq* (2006) (cit. on p. 101).

[MHA00] Kerstin ML Menne, Henning Hermjakob, and Rolf Apweiler. "A comparison of signal sequence prediction methods using a test set of signal peptides". In: *Bioinformatics* 16.8 (2000), pp. 741–742 (cit. on p. 37).

[Mic13] Zbigniew Michalewicz. *Genetic algorithms+ data structures= evolution programs*. Springer Science & Business Media, 2013 (cit. on p. 47).

[Mil+06] Ethan Millar et al. “Performance and scalability of a large-scale n-gram based information retrieval system”. In: *Journal of digital information* 1.5 (2006) (cit. on p. 71).

[MK14] Faouzi Mhamdi and Mehdi Kchouk. “Algorithme Hybride de Sélection d’attributs pour le Classement des protéines”. In: *EGC 2014* (2014) (cit. on pp. 72, 73).

[MKS00] Robert M Maccallum, Lawrence A Kelley, and Michael JE Sternberg. “SAWTED: structure assignment with text description—enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons”. In: *Bioinformatics* 16.2 (2000), pp. 125–129 (cit. on p. 58).

[ML98] Bing Liu Wynne Hsu Yiming Ma and Bing Liu. “Integrating classification and association rule mining”. In: *Proceedings of the fourth international conference on knowledge discovery and data mining*. 1998 (cit. on p. 55).

[MM09] Anirban Mukhopadhyay and Ujjwal Maulik. “Unsupervised pixel classification in satellite imagery using multiobjective fuzzy clustering combined with SVM classifier”. In: *IEEE transactions on geoscience and remote sensing* 47.4 (2009), pp. 1132–1138 (cit. on p. 48).

[MPP07] Ricardo Martinez, Claude Pasquier, and Nicolas Pasquier. “GenMiner: mining informative association rules from genomic data”. In: *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*. IEEE. 2007, pp. 15–22 (cit. on p. 55).

[MPP08] Ricardo Martinez, Nicolas Pasquier, and Claude Pasquier. “GenMiner: mining non-redundant association rules from integrated gene expression data and annotations”. In: *Bioinformatics* 24.22 (2008), pp. 2643–2644 (cit. on p. 54).

[NHB06] Ashesh Nandy, Marissa Harle, and Subhash C Basak. “Mathematical descriptors of DNA sequences: development and applications”. In: *Arkivoc* 9.2006 (2006), pp. 211–238.

[NK92] Kenta Nakai and Minoru Kanehisa. “A knowledge base for predicting protein localization sites in eukaryotic cells”. In: *Genomics* 14.4 (1992), pp. 897–911 (cit. on p. 58).

[NO82] Ken Nishikawa and Tatsuo Ooi. “Correlation of the amino acid composition of a protein to its structural and biological characters”. In: *The Journal of Biochemistry* 91.5 (1982), pp. 1821–1824 (cit. on p. 58).

[Nob+04] William Stafford Noble et al. “Support vector machine applications in computational biology”. In: *Kernel methods in computational biology* (2004), pp. 71–92 (cit. on p. 50).

[NW70] Saul B Needleman and Christian D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of molecular biology* 48.3 (1970), pp. 443–453 (cit. on pp. 26, 28).

[OMV02] Antoine Oliver, Nicolas Monmarché, and Gilles Venturini. “Interactive Design of Web Sites with a Genetic Algorithm.” In: *ICWI*. 2002, pp. 355–362 (cit. on p. 48).

[Par+09] Sung Hee Park et al. “Prediction of protein-protein interaction types using association rule based classification”. In: *BMC bioinformatics* 10.1 (2009), p. 36 (cit. on p. 55).

[Pas+05] Nicolas Pasquier et al. “Generating a condensed representation for association rules”. In: *Journal of Intelligent Information Systems* 24.1 (2005), pp. 29–60 (cit. on p. 55).

[PBT04] Craig T Porter, Gail J Bartlett, and Janet M Thornton. “The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data”. In: *Nucleic acids research* 32.suppl_1 (2004), pp. D129–D133 (cit. on p. 37).

[Pet+11] Thomas Nordahl Petersen et al. “SignalP 4.0: discriminating signal peptides from transmembrane regions”. In: *Nature methods* 8.10 (2011), p. 785 (cit. on p. 37).

[Pic+02] Fabien Picarougne et al. “Web searching considered as a genetic optimization problem”. In: *Local search two day workshop, London, UK*. 2002, pp. 16–17 (cit. on p. 48).

[PJB14] Vaishali S Parsania, NN Jani, and Navneet H Bhalodiya. “Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis”. In: *International Journal of Darshan Institute on Engineering Research & Emerging Technologies* 3.1 (2014) (cit. on pp. 103, 104).

[PK07] Andrzej Polanski and Marek Kimmel. “Sequence Alignment”. In: *Bioinformatics* (2007), pp. 155–185 (cit. on p. 26).

[PL88] William R Pearson and David J Lipman. “Improved tools for biological sequence comparison”. In: *Proceedings of the National Academy of Sciences* 85.8 (1988), pp. 2444–2448 (cit. on p. 31).

[Ple+08] Dariusz Plewczynski et al. “Prediction of signal peptides in protein sequences by neural networks”. In: *Acta Biochim Pol* 55.2 (2008), pp. 261–267 (cit. on p. 50).

[Pyl99] Dorian Pyle. *Data preparation for data mining*. Vol. 1. morgan kaufmann, 1999 (cit. on p. 39).

[QF07] Zhao-Hui Qi and Tong-Rang Fan. “PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization”. In: *Chemical Physics Letters* 442.4 (2007), pp. 434–440 (cit. on p. 33).

[QQ07] Zhaohui Qi and Xiaoqin Qi. “Novel 2D graphical representation of DNA sequence based on dual nucleotides”. In: *Chemical Physics Letters* 440.1 (2007), pp. 139–144 (cit. on p. 33).

[Qui86] J. Ross Quinlan. “Induction of decision trees”. In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 49).

[Qui93] J Ross Quinlan. “C4. 5: Programming for machine learning”. In: *Morgan Kauffmann* 38 (1993), p. 48 (cit. on p. 49).

[Rab89] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286 (cit. on p. 51).

[Raj+11] Anil Rajput et al. “J48 and JRIP rules for e-governance data”. In: *International Journal of Computer Science and Security (IJCSS)* 5.2 (2011), p. 201 (cit. on p. 103).

[Ran+03] Milan Randić et al. “Novel 2-D graphical representation of DNA sequences and their numerical characterization”. In: *Chemical Physics Letters* 368.1 (2003), pp. 1–6 (cit. on p. 33).

[Ree+87] Gerald R Reeck et al. ““Homology” in proteins and nucleic acids: a terminology muddle and a way out of it”. In: *Cell* 50.5 (1987), p. 667 (cit. on p. 37).

[RH98] Astrid Reinhardt and Tim Hubbard. “Using neural networks for prediction of the subcellular location of proteins”. In: *Nucleic acids research* 26.9 (1998), pp. 2230–2236 (cit. on p. 58).

[Seb02] Fabrizio Sebastiani. “Machine learning in automated text categorization”. In: *ACM computing surveys (CSUR)* 34.1 (2002), pp. 1–47 (cit. on p. 97).

[Seg+03] Neil H Segal et al. “Classification and subtype prediction of adult soft tissue sarcoma by functional genomics”. In: *The American journal of pathology* 163.2 (2003), pp. 691–700 (cit. on p. 50).

[SH12] Long Shi and Hailan Huang. “DNA sequences analysis based on classifications of nucleotide bases”. In: *Affective Computing and Intelligent Interaction*. Springer, 2012, pp. 379–384 (cit. on p. 63).

[SKS01] BJ Stapley, Lawrence A Kelley, and Michael JE Sternberg. “Predicting the sub-cellular location of proteins from text using support vector machines”. In: *Biocomputing 2002*. World Scientific, 2001, pp. 374–385 (cit. on p. 58).

[Smi+94] Steven W Smith et al. “The genetic data environment an expandable GUI for multiple sequence analysis”. In: *Bioinformatics* 10.6 (1994), pp. 671–675 (cit. on p. 20).

[Sun09] Wing-Kin Sung. *Algorithms in bioinformatics: A practical introduction*. CRC Press, 2009 (cit. on pp. 12, 16).

[SW10] Roy D Sleator and Paul Walsh. “An overview of in silico protein function prediction”. In: *Archives of microbiology* 192.3 (2010), pp. 151–155 (cit. on p. 37).

[SW81] Temple F Smith and Michael S Waterman. “Identification of common molecular subsequences”. In: *Journal of molecular biology* 147.1 (1981), pp. 195–197 (cit. on pp. 28, 30).

[TA08] Christopher M Taylor and Arvin Agah. “Data mining and genetic algorithms: Finding hidden meaning in biological and biomedical data”. In: *Computational Intelligence in Biomedicine and Bioinformatics*. Springer, 2008, pp. 49–68 (cit. on p. 48).

[Tam+99] Pablo Tamayo et al. “Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation”. In: *Proceedings of the National Academy of Sciences* 96.6 (1999), pp. 2907–2912 (cit. on p. 44).

[Tan+05] Lorraine Tanabe et al. “GENETAG: a tagged corpus for gene/protein named entity recognition”. In: *BMC bioinformatics* 6.1 (2005), S3 (cit. on p. 56).

[TLZ15] Chengjie Tan, Shanshan Li, and Ping Zhu. “4D Graphical representation research of DNA sequences”. In: *International Journal of Biomathematics* 8.01 (2015), p. 1550004 (cit. on p. 33).

[Tri09] Thomas Triplet. *Classification, clustering and data-mining of biological data*. The University of Nebraska-Lincoln, 2009 (cit. on p. 29).

[TSS05] Koji Tsuda, Hyunjung Shin, and Bernhard Schölkopf. “Fast protein classification with multiple networks”. In: *Bioinformatics* 21.suppl_2 (2005), pp. ii59–ii65 (cit. on p. 34).

[Urr04] Dan W Urry. “The change in Gibbs free energy for hydrophobic association: Derivation and evaluation by means of inverse temperature transitions”. In: *Chemical physics letters* 399.1 (2004), pp. 177–183 (cit. on p. 65).

[VD14] S Vijayarani and S Deepa. “Naïve Bayes Classification for Predicting Diseases in Haemoglobin Protein Sequences”. In: *J. of Computational Intelligence and Informatics* 3.4 (2014) (cit. on p. 51).

[Vel+95] Victor E Velculescu et al. “Serial analysis of gene expression”. In: *Science* 270.5235 (1995), p. 484 (cit. on p. 23).

[Vya+12] VK Vyas et al. “Homology modeling a fast tool for drug discovery: current perspectives”. In: *Indian journal of pharmaceutical sciences* 74.1 (2012), p. 1 (cit. on p. 35).

[Wan+02] Haixun Wang et al. “Clustering by pattern similarity in large data sets”. In: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. ACM. 2002, pp. 394–405 (cit. on p. 43).

[Wan+13] Zhouxi Wang et al. “Protein function annotation with structurally aligned local sites of activity (SALSAs)”. In: *BMC bioinformatics*. Vol. 14. 3. BioMed Central. 2013, S13 (cit. on p. 37).

[WC+53] James D Watson, Francis HC Crick, et al. “Molecular structure of nucleic acids”. In: *Nature* 171.4356 (1953), pp. 737–738 (cit. on pp. 11, 26).

[WC58] JD Watson and FH Crick. “On protein synthesis”. In: *The Symposia of the Society for Experimental Biology*. Vol. 12. 1958, pp. 138–163 (cit. on p. 15).

[WKG00] Cyrus A Wilson, Julia Kreychman, and Mark Gerstein. “Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores”. In: *Journal of molecular biology* 297.1 (2000), pp. 233–249 (cit. on pp. 57, 64).

[Wu08] Fang-xiang Wu. “Genetic weighted k-means algorithm for clustering large-scale gene expression data”. In: *BMC bioinformatics* 9.6 (2008), S12 (cit. on p. 45).

[YSB03] Xin Yuan, Yu Shao, and Christopher Bystroff. “Ab initio protein structure prediction using pathway models”. In: *Comparative and functional genomics* 4.4 (2003), pp. 397–401 (cit. on p. 35).

[YSW09] Jia-Feng Yu, Xiao Sun, and Ji-Hua Wang. “TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications”. In: *Journal of theoretical biology* 261.3 (2009), pp. 459–468 (cit. on p. 33).

[Zak+97] Mohammed Javeed Zaki et al. “New Algorithms for Fast Discovery of Association Rules.” In: *KDD*. Vol. 97. 1997, pp. 283–286 (cit. on p. 80).

[Zen15] He Zengyou. *Data Mining for Bioinformatics Applications*. Woodhead Publishing, 2015 (cit. on p. 35).