REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **système d'information optimisation et de décision (SIOD)**

Prédiction d'interaction protéine-protéine

Par :
NESSIGHA AHLAM

Année universitaire : 2020/2021

Dédicace

Toutes les lettres ne sauraient trouver les mots qu'il faut... Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect...Aussi, c'est tout simplement que. Je dédie ce modeste travail à:

À celui qui m'a donné tout ce qu'il avait pour que je puisse réaliser ses espoirs envers lui, à celui qui me poussait en avant pour atteindre le désiré, à celui qui possédait l'humanité avec tout pouvoir, à celui qui veillait sur mon éducation avec d'énormes sacrifices traduits dans sa sanctification de la science, à ma première école de la vie,

Mon père, cher à mon cœur, que Dieu préserve sa vie.

À celle qui lui a donné le plus grand plaisir de toute la tendresse et de la tendresse, à celle qui a été patiente avec tout, qui m'a nourri avec le droit aux soins et a été mon soutien dans l'adversité, et sa prétention à moi pour le succès, elle m'a suivi pas à pas dans mon travail, à qui j'étais soulagée chaque fois que je me souvenais de son sourire sur mon visage.

Source de tendresse, ma mère est l'ange le plus cher dans le cœur et l'œil. Que Dieu la récompense pour moi.

A ma chère sœur NASSIMA, qui a toujours été de mon côté, je te souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité.

A mon frère AYMEN, qui a toujours été de mon côté, et m'avez aide, je vous souhaite un avenir plein de réussit dans ton vie professionnelle.

À mes proches, ma sœur LOUBNA et mes petits frères, AMIR et ISMAIL, je leur souhaite une vie heureuse et réussie.

À mes sœurs qui n'ont pas donné naissance à ma mère MAJDA et HIND, ma chère tante HOUDA, et à mes chers amis et sœurs, RAYAN et NESRINE.

Une spéciale dédicace à tous ceux qui m'ont aimé le bien et le succès et qui m'ont soutenu tout au long des années passé et qui était la raison de mon succès. Et à tous ce qui ont enseigné moi au long de ma vie scolaire.

Remerciement

Je tiens premièrement à prosterner remerciant Allah le tout puissant de m'avoir donné le courage et la patience pour terminer ce travail.

Je veux adresser les grands remerciements les plus sincères à mon cher encadreur **Dr Belounnar Saliha** *pour m'avoir honoré par son encadrement, ses conseils précieux, sa patience et ses nobles valeurs humaines.*

Je tiens également à remercier messieurs les membres de jury pour l'honneur qu'ils m'ont fait en acceptant de siéger à notre soutenance.

Finalement, je tiens à exprimer ma profonde gratitude à ma famille qui ma soutenue et encouragée tout au long de mes études.

RÉSUMÉ

Les interactions protéines jouent un rôle majeur dans divers processus et fonctions biologiques des cellules vivantes.Par conséquent, les étudier et les identifier correctement est essentiel pour comprendre les versets moléculaires dans la cellule.

La biologie offre de nombreuses méthodes expérimentales pour détecter les interactions protéiques, mais malgré leur développement, elles sont coûteuses et prennent du temps.

Ce travail propose un système pour étudier les interactions des paires des protéines à partir de leurs chaînes peptidiques pour créer des modèles de prédiction à l'aide de méthodes d'apprentissage automatique. Ces modèles sont utilisés pour prédire les interactions des paires protéine-protéine.

Le système est conçu, implémenté et validé des données résultant de la prédiction des interactions protéine-protéine.

Mots clés: Protéine, Acides aminés, Chaînes peptidiques, Interaction protéine-protéine, Apprentissage automatique, Classification (SVM).

ملخص

تلعب تفاعلات البروتينات دورا رئيسيا في العمليات و الوظائف البيولوجية المختلفة في الخلايا الحية. لذلك فإن دراستها و تحديدها بشكل صحيح أمر بالغ الأهمية لفهم الآيات الجزيئية داخل الخلية.

يقدم علم الأحياء طرق تجريبية عديدة من أجل الكشف عن تفاعلات البروتينات و لكن بالرغم من تطورها إلا أنها مكلفة وتستغرق وقتا.

يقترح هذا العمل نظاما لدراسة تفاعلات ازواج البروتين. انطلاقا من سلاسلها البيبتيدية لإنشاء نماذج للتنبؤ باستخدام أساليب التعلم الآلي. تستخدم هذه النماذج للتنبؤ بتفاعلات زوج البروتين-بروتين.

تم تصميم النظام و تنفيذه و التحقق من صحة البيانات الناتجة عن التنبؤ بتفاعلات البروتين-بروتين.

الكلمات المفتاحية : البروتين ، احماض امينة ، سلاسل بيبتيدية ، تفاعل البروتين- بروتين ، التعليم الآلي ، التصنيف (SVM).

ABSTRACT

Proteins interactions play a major role in the different biological processes and functions. So studying and correctly identifying them is crucial to understanding the molecular function inside the cell.

Biology offers several experimental methods for detecting protein interactions, but despite their development, they are both time-consuming and expensive.

This work proposes a system for studying protein-pairs interactions from their peptide chains to create predictive models using machine learning methods. These models are used to predict protein-protein interactions.

The system is designed, implemented and validated data generated from prediction of protein-protein interactions.

Key words: Protein, amino acids, peptide chains, protein-protein interaction, machine learning, classification (SVM).

Table des matières

Table de matières I

Liste des figures V

Liste des tables VII

Introduction générale 1

1 PROTEINE ET SON INTERACTION 3

1.1 Introduction 4
1.2 Définition de la Bioinformatique 4
1.3 Application de la Bioinformatique 4
1.4 Terminologies biologiques 5
1.4.1 ADN 5
1.4.2 ARN et ARNm 6
1.4.3 La reproduction 7
1.4.4 Les acides aminés 8
1.4.5 Codon 9
1.4.6 La Traduction 9
1.5 La protéine 11
1.5.1 Définition de la protéine 11
1.5.2 La structure de la protéine 11
1.5.2.1 Structure primaire 11
1.5.2.2 Structure secondaire 12
1.5.2.3 Structure tertiaire 12
1.5.2.4 Structure quaternaire 12
1.5.3 Interaction protéine 14
1.5.4 Type interaction 14
1.5.4.1 Homo-oligomère ou hétéro-oligomère 14
1.5.4.2 Obligatoire ou non-obligatoire 14
1.5.4.3 Permanente ou transitoire 14
1.5.5 Mécanismes de régulation des protéines 15

1.5.5.1 La localisation 15
1.5.5.2 La concentration locale 15
1.5.5.3 L'environnement physico-chimique local 15
1.5.6 Réseau d'interaction protéine-protéine 15
1.5.7 Les bases de données d'interactions protéine-protéine 17
1.5.8 Prédiction d'interactions 17
1.5.9 Méthodes de prédiction d'interaction 18
1.5.9.1 Méthodes de conservation du contexte génomique 18
1.5.9.2 Méthodes de co-évolution 18
1.5.9.3 Méthodes basées sur la structure 19
1.5.9.4 Méthodes basées sur les domaines 19
1.5.9.5 Méthodes d'apprentissage 19
1.6 Conclusion 20

2 METHODE D'APPRENTISSAGE 21

2.1 Introduction 22
2.2 L'apprentissage 22
2.2.1 L'apprentissage supervisé 22
2.2.2 L'apprentissage non supervisé 24
2.2.3 L'apprentissage semi-supervisé 24
2.3 Le Support Victor Machin (SVM) 25
2.3.1 Principe de la technique SVM 25
2.3.1.1 L'hyperplan 25
2.3.1.2 Marge 26
2.3.1.3 Maximiser la marge 26
2.3.1.4 Vecteurs supports 27
2.3.2 SVM à marge dure 28
2.3.3 SVM à marge souple 29
2.3.4 SVM a Kernel 30
2.3.5 SVM multi classes 30
2.3.5.1 Les méthodes de décomposition 31
2.3.5.1.1 Une-contre-reste (1vsR) 31

2.3.5.1.2 Une-contre-une (1vs1) 32
2.3.6 Avantage du SVM 33
2.4 Evaluation de l'apprentissage et sélection de modèle 34
2.4.1 La sélection de modèle 34
2.4.2 L'évaluation d'un modèle 34
2.5 Méthodes d'apprentissage 34
2.6 Conclusion 35

3 CONCEPTION DU SYSTEME 36

3.1 Introduction 37
3.2 Méthodologie suivie 37
3.3 Conception globale du système 38
3.4 Conception détaillé 38
3.4.1 Module préparation des données 39
3.4.1.1 La conversion 40
3.4.1.2 Extraction des paires de protéines 40
3.4.1.3 Réduction des données 40
3.4.1.4 Validation croisée 40
3.4.2 Le module prédiction 43
3.4.2.1 Création du modèle 43
3.4.2.2 Teste du modèle 43
3.5 Conclusion 45

4 IMPLEMENTATION ET RESULTAT 46

4.1 Introduction 47
4.2 Environnement et outils de programmation 47
4.2.1 Langages de programmation 47
4.2.1.1 MATLAB 47
4.2.2 Outils utilisés 48
4.2.2.1 Libsvm 48
4.2.2.1.1 Paramètres SVM pris en charge par la librairie 48
4.2.2.2 La matrice SM_BLOSUM62 50
4.3 Système de prédiction d'interaction protéine-protéine 51

4.3.1 Base de données utilisé 51
4.3.2 Préparation des données 52
4.3.3 Prédiction 53
4.4 Interface de l'application de la prédiction 53
4.5 Expérimentations et résultats 56
4.5.1 Première expérimentation 56
4.5.2 Deuxième expérimentation 57
4.5.3 Troisième expérimentation 57
4.5.4 Quatrième expérimentation 58
4.5.5 Cinquième expérimentation 58
4.5.6 Après l'exécution 60
4.6 Discussion des résultats 61
4.7 Conclusion 62

5 Conclusion générale 63

6 Bibliographie 65

Liste des figures

1.1 Structure du nucléotide **(Raven et al., 2014)**……..…5

1.2 La structure de l‟ADN **(Raven et al., 2014)** ……6

1.3 La reproduction…….……8

1.4 La structure d’acide amine. ……..…...……8

1.5 La structure des acides aminés liés ……...……..……..9

1.6 La traduction. ……..……10

1.7 Les 4 structures des protéines. **[5]**……13

1.8 Visualisation d’un réseau d’interaction protéine-protéine chez la levure du boulanger. **[Jeong et al., 2001]**……...……..16

2.1 Les deux types de classifications…….……...……..…23

2.2 Hyperplan à séparation optimale……...…26

2.3 Exemple d’un hyperplan optimal **[20]** ……...……27

2.4 L’hyperplan H optimal, vecteurs supports et marge maximale……...……28

2.5 Données d'apprentissage avec une marge maximale……...……29

2.6 Marge souple……...……...……29

2.7 Points non linéairement séparable et leur projection vers un autre espace d'une dimension supérieur…….……30

2.8 Résolution des cas d’indécision dans la méthode 1vsR……...……31

2.9 Les surfaces rouges représentent le « reject decision »……32

2.10 Classification multi class par paire……...……33

3.1 Conception global……38

3.2 Conception détaillé…….……..……39

3.3 Conception du module Préparation des donnes. ……42

3.4 Conception du module prédiction……...……..…44

4.1 Matlab……..……47

4.2 Matrice SM_BLOSOM62……50

4.3 Les quatre groupes qui composent notre base de données……...…51

4.4 Le contenu de chaque groupe dans la base de données…….…..52

4.5 L'interface principale du système prédiction d'interaction protéine-protéine……..…54

4.6 L'interface de traitement pour le système de prédiction d'interaction protéine-protéine54

4.7 L'interface Help……...……55

4.8 Précision et Rappel de notre méthode…….……59

4.9 L'interface de l'application après la fin du processus de traitement et l'apparition des résultats……60

4.10 La classification réelle et la classification prédictive de base de test2……..61

Liste des tables

1.1 Description de quelques bases de données d'interactions. **[26]** ……..….17

4.1 Matrice de confusion de la première expérimentation sur les données de test……....56

4.2 Matrice de confusion de la deuxième expérimentation sur les données de test……57

4.3 Matrice de confusion de la troisième expérimentation sur les données de test……...57

4.4 Matrice de confusion de la quatrième expérimentation sur les données de test……..58

4.5 Matrice de confusion de la cinquième expérimentation sur les données de test……...58

Introduction générale

Depuis le début de la biologie. L'étude de la fonction des organismes biologiques (organes, cellules, protéines, gènes) est une préoccupation constante. Et la découverte croissante de ces organismes biologiques a incité la biologie à se spécialiser dans différentes branches (physiologie, biologie cellulaire, biologie moléculaire, génétique, etc.). Chaque découverte élargit le champ du questionnement, permettant l'émergence de nouvelles disciplines, qui il s'agit de développer de nouvelles méthodes d'analyse de la fonction des organismes biologiques.

L'un de ces organismes est la protéine, qui remplit rarement sa fonction seule. Elle interagit avec d'autres protéines pour accomplir sa fonction biologique. Par conséquent, les interactions protéine-protéine (IPP) jouent un rôle majeur dans les processus et les fonctions biologiques des cellules vivantes. Y compris les cycles métaboliques, la transcription et la réplication de l'ADN et les cascades de signalisation… C'est pour ca, l'identification et la caractérisation correcte des interactions protéiques est essentielle pour comprendre les mécanismes moléculaires à l'intérieur de la cellule.

Au cours des dernières décennies, de nombreuses techniques expérimentales innovantes ont été développées pour la détection des IPP. Une énorme quantité de différents types de données PPI ont été collectées. Cependant, les méthodes expérimentales sont coûteuses et prennent du temps, donc les paires PPI actuelles obtenues par des méthodes expérimentales ne couvrent qu'une petite partie des réseaux PPI complets. De plus, les méthodes expérimentales à grande échelle souffrent généralement de taux élevés de prédictions faussement négatives et faussement positives. Ainsi, il est d'une grande importance pratique de développer des méthodes de calcul fiables pour faciliter la détermination de l'IPP.

Un certain nombre de techniques de calcul ont été proposées pour fournir des informations complémentaires ou des preuves à l'appui des méthodes expérimentales, et elle a rencontré un grand succès dans son domaine. Ces méthodes s'appuient sur des informations sur les protéines pour prédire leurs interactions. Mais elle trouve des problèmes au manque des informations.

Alors dans notre travaille, on étudiera les interactions protéine-protéine à partir de leurs chaînes protéiques uniquement, en utilisant les méthodes d'apprentissage qui a réussi à prédire les IPP.

Effectivement, nous avons proposé un système basé sur la méthode de classification SVM (Machine à Vecteurs Supports) en utilisant les chaines protéiques pour créé des modèles de prédiction , puis prédire les interactions protéine-protéine. Et nous avons fourni une interface à ce système pour le rendre facile à utiliser et à bien comprendre.

Ce mémoire comporte quatre chapitres organisés comme suit : Dans le premier chapitre, nous présentons le domaine de la bioinformatique et le contexte biologique relatif à cette thèse. Aussi nous introduisons les concepts de protéine et d'interaction protéine-protéine avec les méthodes de prédiction d'interaction protéine-protéine.

Le deuxième chapitre, présente les principaux concepts de l'apprentissage.son différentes problématiques. Et nous présentons plus particulièrement la méthode de classification SVM que nous utilisons dans notre projet ou nous détaillons son principe et ses avantages.

Le troisième chapitre est consacré à la conception de notre système, nous avons utilisé des schémas illustratifs globale et détaillée. Et dans le dernier chapitre, nous avons expliqué les étapes d'implémentation nécessaires pour réaliser le système conçu, et aussi les expérimentations et la discussion des résultats obtenus.

Le mémoire est terminé par une conclusion générale issue de notre travail contenant les perspectives envisagées.

CHAPITRE 1 :
PROTEINE ET SON INTERACTION

CHAPITRE 1:PROTEINE ET SON INTERACTION

1.1 Introduction

L'étude des interactions protéine-protéine (IPP) peut être très importante pour comprendre les fonctions cellulaires biologiques. Cependant, la découverte des IPP dans les laboratoires est longue et coûteuse. Pour cette raison, il y a eu beaucoup d'efforts récents pour développer des techniques de prédiction informatique des interactions protéine-protéine, car cela peut compléter les procédures de laboratoire et fournir un moyen peu coûteux de prédire le groupe d'interactions le plus probable sur toute la gamme de protéines.

L'objectif de ce chapitre est de présenter le contexte biologique relatif à cette thèse. Nous introduisons les notions de protéine et d'interaction protéine-protéine avec les méthodes de prédiction d'interaction protéine-protéine.

.

1.2 Définition de la Bioinformatique

La bioinformatique est une discipline émergente qui s'appuie sur les forces de l'informatique, les mathématiques et la technologie de l'information qui déterminerait l'analyse de l'information génétique. La bioinformatique tire partie des synergies entre sciences computationnelles et biologiques. **[1]** donc on peut dire aussi que la bioinformatique est l'application d'informatique sur la biologie.**[2]**

1.3 Application de la Bioinformatique

La Bioinformatique n'est pas limité dans ses applications et en général peut être appliqué à toute recherche informatique pour résoudre des problèmes biologiques. Les applications courantes de la bioinformatique sont énumérées ci-dessous : **[1]**

- La bioinformatique des séquences, qui traite de l'analyse de données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines qu'il code. Cette branche s'intéresse en particulier à l'identification des ressemblances entre les séquences, à l'identification des gènes ou de régions biologiquement pertinentes dans l'ADN ou dans les protéines, en se basant sur l'enchaînement ou séquence de leurs composants élémentaires (nucléotides, acides aminés).

- La bioinformatique structurale, qui traite de la reconstruction, de la prédiction ou de l'analyse de la structure 3D ou du repliement des macromolécules biologiques (protéines, acides nucléiques), au moyen d'outils informatiques.
- La bioinformatique des réseaux, qui s'intéresse aux interactions entre gènes, protéines, cellules, organismes, en essayant d'analyser et de modéliser les comportements collectifs d'ensembles de briques élémentaires du Vivant. Cette partie de la bioinformatique se nourrit en particulier des données issues de technologies d'analyse à haut débit comme la protéomique ou la transcriptomique pour analyser des flux génétiques ou métaboliques.
- La bioinformatique statistique et la bioinformatique des populations. **[1]**

1.4 Terminologies biologiques

1.4.1 ADN

L'ADN est une molécule très longue, composée d'une succession de nucléotides accrochés les uns aux autres par des liaisons phosphodiester. Il existe quatre nucléotides différents : A(adénosine), C(cytidine), G(guanosine), et T(thymidine), dont l'ordre d'enchaînement est très précis et correspond à l'information génétique. **[30]**

Un nucléotide d'ADN (Figure 1) a 3 composants :

- Un sucre (désoxyribose).
- Un composant d'acide phosphorique (phosphate).
- Une base d'azote (un des quatre types : Adénine ou Adénosine (A), Guanine (G), Cytosine (C) et Thymine (T)).

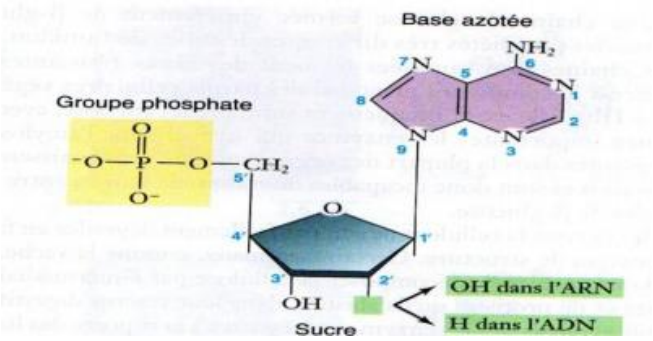


Figure 1.1 :- Structure du nucléotide **(Raven et al., 2014).**

L'ADN est composé de deux brins se faisant face, et formant une double hélice. Ceci est rendu possible par la complémentarité des nucléotides qui peuvent interagir par des liaisons hydrogènes. Il y a deux liaisons hydrogènes entre **A** et **T** et trois entre **C** et **G**, ce qui conduit aux interactions possibles : A-T et T-A, d'une part et G-C et C-G, d'autre part. Les brins d'ADN sont orientés dans le sens 5' vers 3' (et ceci en raison de notations liées à la géométrie du désoxyribose).

Les deux brins d'une double hélice sont complémentaires et antiparallèles, c'est-à-dire assemblés tête bêche (l'extrémité 5' de l'un est en contact avec l'extrémité 3' de l'autre et inversement).

Nous pouvons également dire que l'ADN est la molécule portant l'information génétique dans le noyau**.** **[4]**

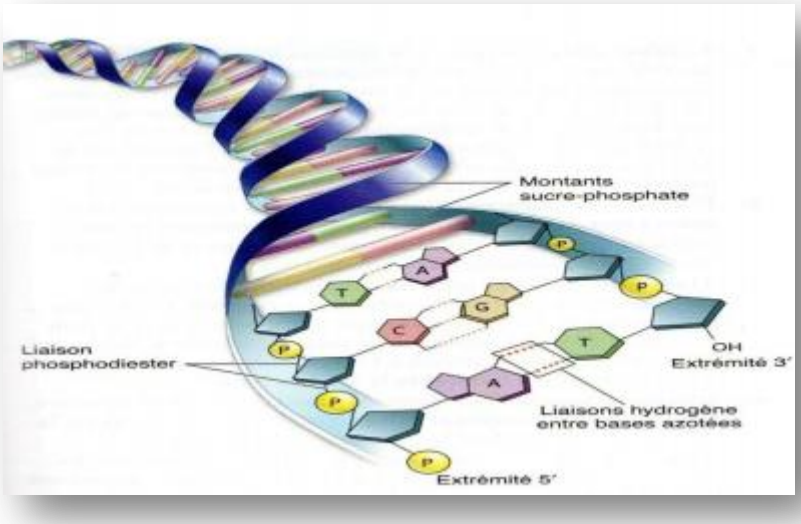


Figure 1.2 :- La structure de l'ADN **(Raven et al., 2014)** .

1.4.2 ARN et ARNm

ARN : L'acide ribonucléique (ARN) est l'acide nucléique obtenu à partir d'un des deux brins d'ADN. Produit au cours du processus de transcription (c'est-à-dire de l'ADN à l'ARN).

ARNM : est une molécule qui porte une séquence nucléotidique qui forme le code génétique sur lequel les protéines sont fabriquées**.** **[29]**

1.4.3 La reproduction

Est la première étape de la fabrication des protéines [4] .Sont réalisées à l'intérieur du noyau dans les cellules qui ont un noyau et ceux qui n'ont pas un noyau dans le cytoplasme et au cours de laquelle le traitement biomoléculaire de la molécule ARNm.

Éléments principaux du processus de reproduction :

- ADN.
- Enzyme ARN Polymères.
- Quatre types de nucléotides sont inclus dans la synthèse d'ARN.
- L'énergie d'ATP.

La reproduction passe par les étapes suivantes :

- **L'étape de démarrage** : L'enzyme ARN polymérase est associée au début du gène, élimine la circonférence et ouvre la série ADN après avoir cassé les liaisons hydrogène entre les paires de bases nucléotidiques. L'enzyme commence par lire la séquence des bases sur l'une des liaisons ADN de la chaîne clonée et reliant les noyaux correspondants pour former une chaîne d'ARN Les nucléotides de l'acide nucléique sont comparés aux nucléotides en fonction de l'intégration des bases azotées.

- **L'étape d'allongement** : L'enzyme ARN polymérase se déplace le long du gène pour lire les informations sur la molécule d'ADN et lie les noyaux d'ARN en fonction de leur séquence dans la chaîne clonée de l'ADN, conduisant à l'allongement de la molécule d'ARN.

- **L'étape de fin** : L'enzyme atteint la fin du gène et l'élongation allongée de l'ARN est séparée et l'enzyme se décompose et la chaîne d'ADN se ferme.

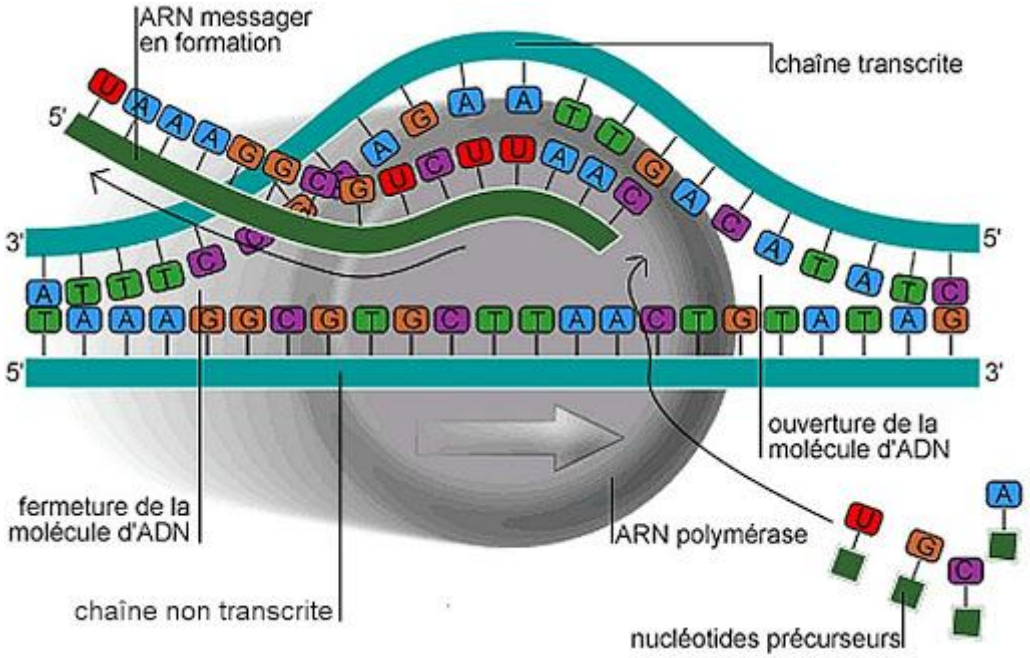


Figure 1.3:- la reproduction.

1.4.4 Les acides aminés

Il y'a 20 acide aminée, chaque acide aminée a La formule générale suivante :

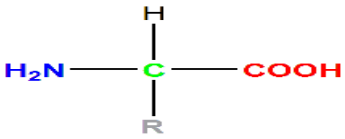


Figure 1.4 :- La structure d'acide amine.

Les acides aminés se distinguent donc par la nature de leur radical **R** plus communément appelé chaine latérale. Ces dernières se distinguent par leur dimension, leur forme, leur charge, leur capacité de contracter des liaisons hydrogènes et leur réactivité chimique. [3]

Pour construire une protéine, il est nécessaire de se munir d'un mécanisme permettant de lier les acides aminés entre eux : la liaison peptidique. Son principe consiste à lier le groupe α-carboxyle d'un acide aminé à la fonction α-amide d'un autre acide aminé par une liaison amide. [3]

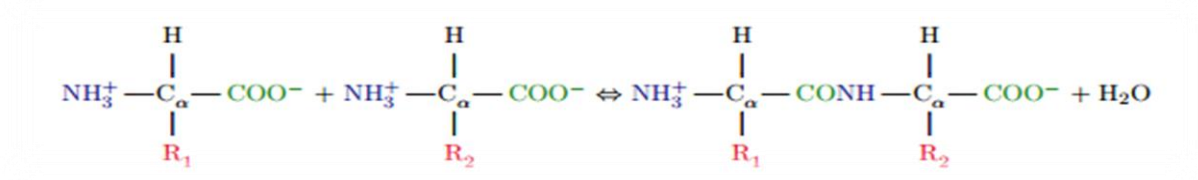


Figure 1.5 :- La structure des acides aminés liés.

1.4.5 Codon

Est une séquence de trois nucléotides sur un acide ribonucléiques messager(ARNm) spécifiant l'un des 22 acides aminés proteinogénes dont la succession sur l'ARN messager détermine la structure primaire de la protéine a synthétiser .Il existe quatre bases nucléiques A ,T,C,G de sorte qu'il existe 4^3 =64 codons différents, on distingue le codon de démarrage AUG et le codon d'arrêt UAG ,UGA,UAA.

1.4.6 La Traduction

La traduction est un processus permettant la synthèse d'une chaîne polypeptidique (protéine) à partir d'un brin d'ARN messager (ARNm) et assure l'expression des gènes portés par l'ADN, et constitue la deuxième grande étape de ce processus après la transcription (qui consiste en la conversion de l'ADN en ARNm).

Acteurs de la traduction : La traduction nécessite un grand nombre d'acteurs dont les plus importants sont :

- L'ARNm.
- Le ribosome .
- L'ARN de transfert (ARNt).
- L'acide aminé.

La traduction se passe par les étapes suivantes :

- **L'étape de démarrage :** L'ARNm est placé sous la micro-unité et ARNt et l'acide aminé méthionine est placé dans la position du ribosome P. L'ARNt est connu comme le marqueur "AUG" aqueux dans l'ARNm à travers l'antioxydant. Ils sont connectés sous le grand corps pour former le complexe de départ. Le deuxième ARNt de l'acide

aminé est situé sur le site A du ribosome selon le second brin de l'ARNm La liaison peptidique entre le premier acide.

- **L'étape d'allongement :** Le ribosome se déplace par un marqueur sur l'ARN m .Conduisant à la séparation de l'ARNt de l'acide aminé et du site P. L'emplacement du second ARNt portant le dipeptide change du site A au site P et le site A devient vide pour recevoir un nouvel ARNt portant un troisième acide aminé. Ainsi, les mêmes étapes sont répétées et la chaîne peptidique est absorbée par la quantité d'acide aminé.

- **L'étape de fin :** Le ribosome est l'un des arrêts sur l'ARNm .La chaîne peptidique résultante se décompose et le dernier ARNt est séparé et séparé sous les unités de ribosome. Le premier acide aminé de la chaîne peptidique est également implanté.

Les ribosomes peuvent répéter le cycle et former une autre série de peptides.

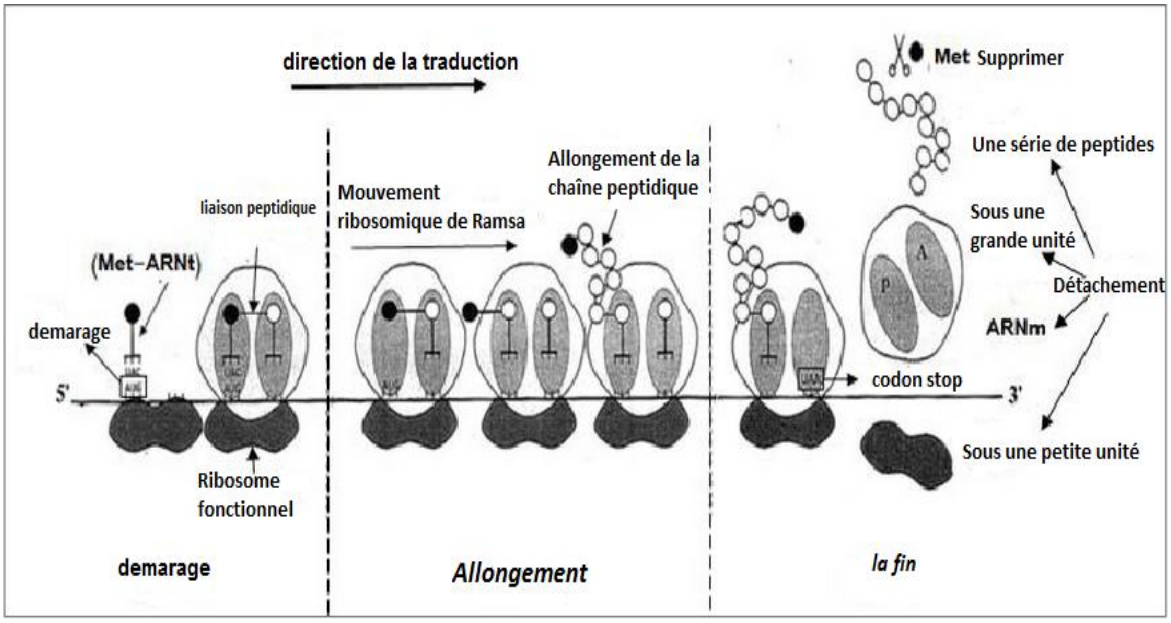


Figure 1.6 :- La traduction.

1.5 La protéine

1.5.1 Définition de la protéine

Une protéine est une collection de chaînes polypeptidiques non ramifiées d'acides aminés liés entre eux par des liaisons peptidiques.

Ces protéines jouent un rôle essentiel dans la cohésion des structures morphologiques et dans le fonctionnement cellulaire. On citera pour mémoire, quelques grands groupes de protéines :

- Les enzymes (catalyseurs biologiques, responsables de la plupart des réactions chimiques de la cellule).
- Les anticorps (responsables de la défense des organismes supérieurs, ils forment, dans le sang, des complexes avec les corps étrangers).
- Les protéines de stockage.
- Les protéines de transport.
- Les hormones (certaines hormones sont de nature protéiques).
- Les histones (liées à l'ADN, elles participent au contrôle de l'expression génétiques).
- Les protéines de structure et de soutien. **[28]**

1.5.2 La structure de la protéine

Elles se composent d'un enchaînement linéaire d'acides aminés liés par des liaisons peptidiques. Cet enchaînement possède une organisation tridimensionnelle (ou repliement) qui lui est propre. De la séquence au repliement, il existe quatre niveaux de structuration de la protéine.

1.5.2.1 Structure primaire

Dans cette structure, la protéine prend la forme d'une chaîne constituée d'un groupe d'acides aminés liés par des liaisons peptidiques aux enzymes ribosomales.

1.5.2.2 Structure secondaire

La structure secondaire décrit le repliement local de la chaîne principale d'une protéine et il existe deux principales catégories de structures secondaires selon l'échafaudage de liaisons hydrogène, et donc selon le repliement des liaisons peptidiques : les hélices α, les feuillets ß.

Cette structure maintient sa cohésion au moyen de liaisons hydrogène qui se forment entre les deux groupes CO et NH.

1.5.2.3 Structure tertiaire

La série de peptides contenant des structures secondaires implique des régions interstitielles appelées zones d'inflexion, Constitue une sous-unité. Cette structure maintient sa cohérence par :

- Liaisons hydrogène entre les fonctions chimiques des racines R.
- Les associations entre les groupes négatifs et positifs dans les racines R.
- Attirer des racines avides d'eau.
- Les ponts de soufre produits entre deux racines de deux acides aminés Cys. **[28]**

1.5.2.4 Structure quaternaire

Deux ou plusieurs séries ont une structure triangulaire et chaque série est appelée une unité. Sous les unités, lier ensemble avec des liaisons faibles telles que des liaisons hydrogène. ou plus rarement des liaisons covalentes (ponts disulfures).Et sa c'est dans le cas de protéines formées de plusieurs sous unités. **[28]**

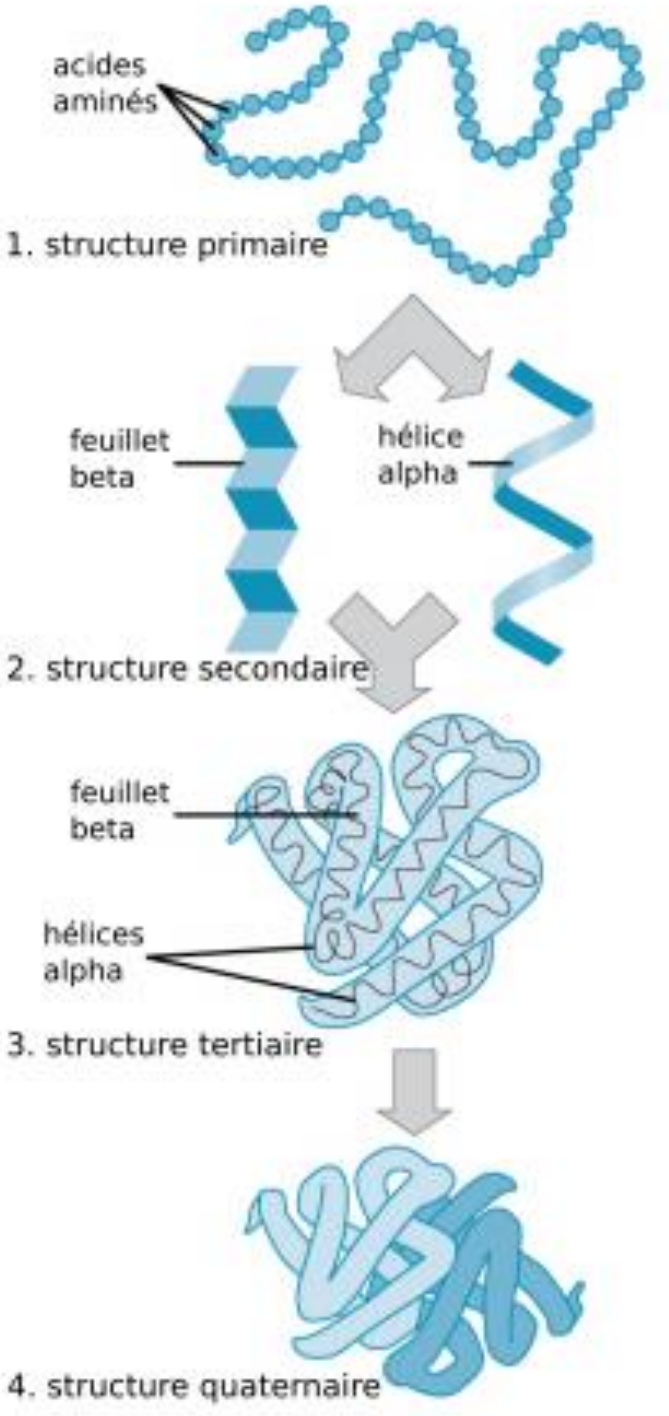


Figure 1.7 :- Les 4 structures des protéines. [5]

1.5.3 Interaction protéine

La plupart des protéines assurent leurs fonctions biologiques en interagissant avec une ou plusieurs autres protéines. Elles peuvent former de larges complexes protéiques tels que le protéasome, qui est un assemblage d'environ 50 sous-unités protéiques agissant ensemble pour dégrader d'autres protéines, et jouant un rôle primordial dans l'homéostasie. Ces interactions sont très diverses, selon leurs composition, leurs affinités ou leur nature permanente ou transitoire. **[24]**

1.5.4 Type interaction

1.5.4.1 Homo-oligomère ou hétéro-oligomère.

Les interactions peuvent exister entre protéines identiques ou différentes, l'interaction peut avoir lieu sur une même surface pour les deux monomères (isologue), ou sur deux surfaces différentes (hétérologue), auquel cas la formation d'agrégats est possible. **[5]**

1.5.4.2 Obligatoire ou non-obligatoire

Les complexes formés peuvent être obligatoires ou non. Une interaction est obligatoire si les monomères impliqués n'ont pas de structure stable in vivo en l'absence de cette interaction. Dans ce cas, il est fréquent que la fonction des protéines impliquées soit dépendante de cette interaction. La plupart des complexes hétéro-oligomèriques impliquent des interactions non-obligatoires, ce qui signifie que les protéines sont stables en l'absence d'interactions. **[5]**

1.5.4.3 Permanente ou transitoire

On peut aussi distinguer les interactions selon leur dynamique. Les interactions permanentes sont très stables, et les protéines impliquées ne sont présentes que sous leur forme complexée. Les interactions transitoires sont beaucoup plus dynamiques, les partenaires s'associent et se dissocient rapidement in vivo. Les interactions transitoires peuvent être faibles, c'est à dire dépendantes de la concentration de chacun des partenaires dans le milieu. L'interaction est contrôlée pas un équilibre oligomérique dynamique en solution, dans

laquelle les interactions se font et se défont continuellement. Les interactions transitoires peuvent aussi être fortes lorsqu'elles sont contrôlées par un mécanisme moléculaire. L'équilibre oligomérique dépend alors d'une phosphorylation ou d'une déphosphorylation, ou de la présence d'un autre partenaire tel que la guanosine triphosphate. Bien souvent, les interactions obligatoires sont aussi permanentes. [5]

1.5.5 Mécanismes de régulation des protéines

Les interactions sont régulées par différents mécanismes. On identifie 3 types de contrôles :

1.5.5.1 La localisation

L'association de deux protéines dépend d'une rencontre des surfaces d'interaction, et requiert une co-localisation dans l'espace et le temps, c'est à dire une co-expression ou co-localisation dans un compartiment. Dans le cas où les protéines sont dans des compartiments différents, des transports dirigés entre les différentes localisations sont nécessaires. [5]

1.5.5.2 La concentration locale

Ce paramètre est contrôlé par divers mécanismes, tels que : l'expression des gènes, les niveaux de sécrétion, la dégradation des protéines, le stockage temporaire, l'environnement moléculaire, et la diffusion ou viscosité du milieu. [5]

1.5.5.3 L'environnement physico-chimique local

Les affinités mutuelles des composants d'un complexe peuvent être modifiées par la présence d'une molécule effectrice (e.g. ATP, Ca2+) ou par un changement des conditions physiologiques. [5]

1.5.6 Réseau d'interaction protéine-protéine

L'ensemble des interactions protéine-protéine ayant lieu dans un organisme, un organe

ou un type cellulaire donné est appelé interactome. Un interactome est souvent représenté par un réseau, dans lequel les nœuds correspondent à des protéines, et où un arc entre deux nœuds signifie l'existence d'une interaction entre les deux protéines correspondantes. Ce réseau, appelé réseau d'interaction protéine-protéine, est non dirigé. Cependant, cette représentation ne tient pas compte de la dynamique et de la localisation des interactions protéine-protéine dans la cellule. **[27]**

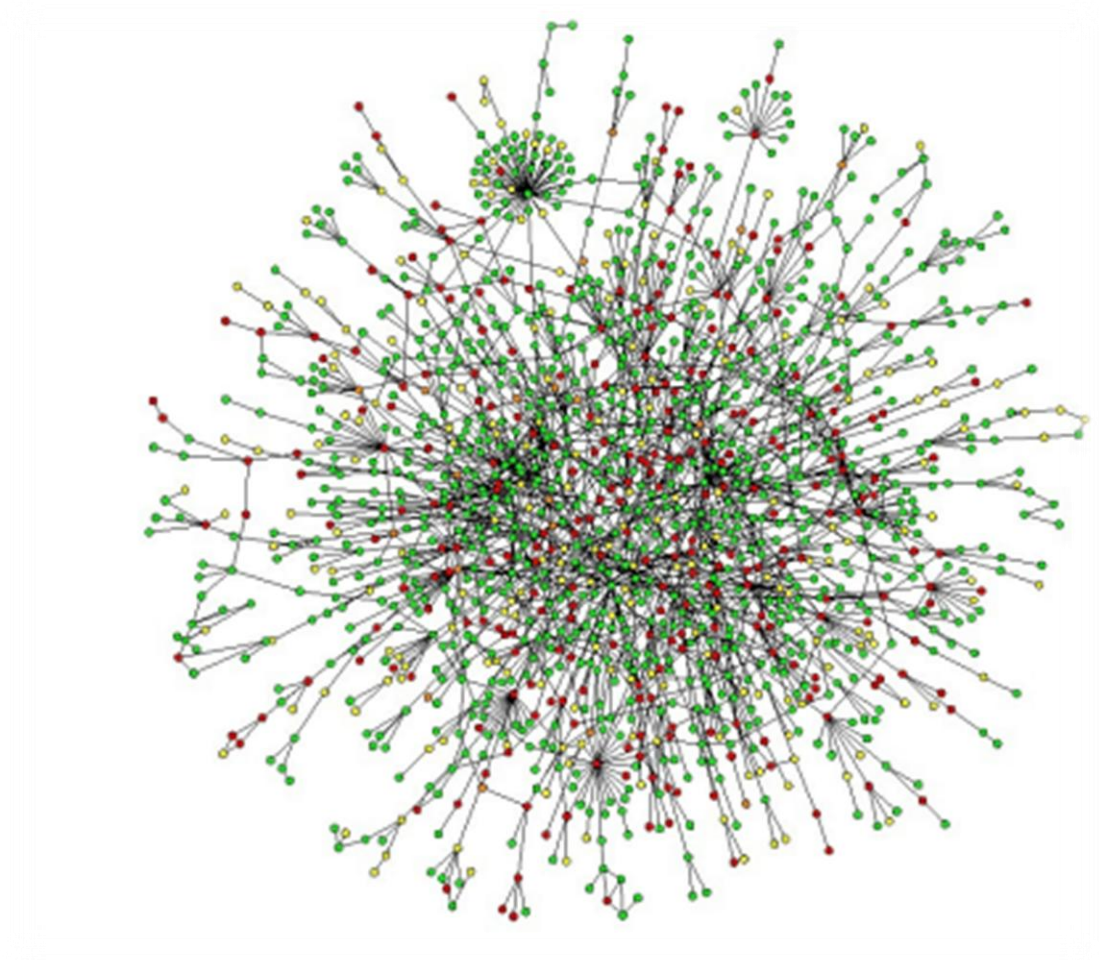


Figure 1.8 :- Visualisation d'un réseau d'interaction protéine-protéine chez la levure du boulanger. **[Jeong et al., 2001].**

1.5.7 Les bases de données d'interactions protéine-protéine

Les données d'interactions entre protéines sont publiquement disponibles dans différentes bases de données. Ces bases des données diffèrent par le type d'organisme couvert et les politiques d'acquisition des données. D'une part, les données d'interactions peuvent être succinctes ou détaillées, et d'autre part elles peuvent être récupérées de façon automatique ou entrées manuellement par une personne qui extrait ces informations de la littérature et des différents cribles effectués. Les principales d'entre elles sont:

Nom	Description	couverture organismes	Type de molécules
IntAct	Détaillée	Large	Tout
MINT	Détaillée	Large	Protéines
DIP	Détaillée	Large	Protéines
MatrixDB	Détaillée	Limité	protéines de la matrice extracellulaire
InnateDB	Détaillée	Limité	Protéines
MIPS	Détaillée	Mammifères	Protéines
MPACT	Détaillée	Levure	Protéines
BioGRID	Succincte	Limité	Protéines
HPRD	Succincte	Humain	Protéines
MPIDB	Succincte	Microbes	Protéines

TABLE 1.1 :– Description de quelques bases de données d'interactions. **[26]**

1.5.8 Prédiction d'interactions

La prédiction des interactions permet d'augmenter le nombre d'interactions disponibles, notamment concernant les protéines peu ou pas étudiées. Les approches prédictives développées permettent de réduire l'espace des interactions à tester expérimentalement en proposant des interactions probables. De plus, ces méthodes prédictives permettent de proposer des interactions entre les protéines d'organismes peu étudiés, ou pour des systèmes inter-espèce dans lesquels très peu de données expérimentales d'interactions sont disponibles.

La prédiction d'interactions a aussi l'avantage d'être applicable à grande échelle (génome et protéome) à peu de frais. Les méthodes de prédictions sont basées sur les séquences des protéines, et les caractéristiques structurales et génomiques liées aux interactions et aux relations fonctionnelles. [5]

1.5.9 Méthodes de prédiction d'interaction

1.5.9.1 Méthodes de conservation du contexte génomique

L'analyse comparative des génomes, et en particulier de la conservation des contextes génomiques à travers les espèces, a permis de mettre en évidence des liens fonctionnels entre des gènes ou entre les protéines que ces derniers codent. Ces interactions fonctionnelles ne sont pas nécessairement des interactions physiques**.** [6] Différentes méthodes ont été comparées par Huynen et al. [Huynen et al, 2000].et parmi ces méthodes :

Transfert par interologues

Cette méthode est basée sur le fait que deux protéines liées fonctionnellement ont tendance à co-évoluer. Si l'on considère deux protéines A et B en interaction dans un organisme donné, leurs orthologues A' et B' dans un autre organisme ont une forte probabilité d'interagir [7]**.** Cette méthode a permis de prédire des interactions chez l'homme, C .elegans et D. melanogaster à partir de données d'interaction de levure**.** [8]

1.5.9.2 Méthodes de co-évolution

Les protéines qui interagissent physiquement évoluent en général de manière coordonnée, conservant ainsi les contacts entre elles [Pazos et al., 1997]. Ainsi, les méthodes basées sur ce principe sont susceptibles de prédire des interactions pas seulement fonctionnelles mais vraiment physiques.et parmi ces méthodes :

Méthode des profils phylogénétiques

Cette méthode n'utilise pas les interactions protéine-protéine déjà connues. Ce sont les relations d'homologie entre les protéines dans différents organismes qui sont utilisées**.**

Chaque protéine est représentée par un vecteur booléen dans lequel l'absence ou la présence d'un orthologue dans différents organismes est indiqué. Les protéines ayant des vecteurs similaires, donc des profils phylogénétiques proches sont identifiées et ont une forte probabilité de participer à un même complexe ou une même voie de signalisation, et donc d'interagir physiquement. **[9]**

1.5.9.3 Méthodes basées sur la structure

D'autres méthodes utilisent des informations plus proches de la structure tridimensionnelle de la protéine, et cherchent d'abord à identifier les sites d'interaction. Ceci peut se faire par la reconnaissance de motifs de résidus [Kini et Evans, 1996], ou en utilisant les propriétés de la topologie de l'interface, de la surface accessible au solvant ou de l'hydrophobicité [Jones et Thornton, 1997]. **[6]**

1.5.9.4 Méthodes basées sur les domaines

Les domaines sont considérés comme les entités élémentaires de construction des protéines. Ce sont des unités structurales et/ou fonctionnelles qui sont conservées au cours de l'évolution. Chaque domaine contribue à la structure globale de la protéine et à ses différentes fonctions. L'hypothèse que les protéines interagissent entre elles par l'intermédiaire de leurs domaines est largement acceptée. L'idée est alors d'inférer des informations sur les interactions domaine-domaine en se basant sur les interactions protéine-protéine, puis de prédire des interactions protéine-protéine à partir des ces interactions domaine-domaine inférées**.** Mais ces méthodes sont évidemment très dépendantes de l'état de nos connaissances, car une grande partie de ces règles demeurent inconnues. C'est pourquoi on considère que les interactomes issues de ces approches sont encore peu fiables**.** **[25]**

1.5.9.5 Méthodes d'apprentissage

Différentes méthodes de classification avec apprentissage ont été utilisées pour prédire des interactions entre des protéines, ou entre des domaines. L'idée est d'apprendre les caractéristiques des paires de protéines en interaction, afin de prédire pour deux protéines quelconques si elles interagissent ou non, ou quelle est la probabilité qu'elles interagissent.

1.6 Conclusion

Nous avons introduit dans ce chapitre les terminologies biologique et le contexte des interactions protéine-protéine. Et nous avons également parcouru ici un ensemble de méthodes de prédiction d'interaction .Dans le chapitre suivant, nous nous intéressons aux méthodes d'apprentissage qui peuvent être utilisées pour la prédiction d'interactions protéine-protéine.

CHAPITRE 2 :

METHODE D'APPRENTISSAGE

CHAPITRE 2 : METHODE D'APPRENTISSAGE

2.1 Introduction

L'objectif de ce chapitre consiste à présenter les principaux concepts de l'apprentissage statistique utilisés à cette thèse. Nous commençons par introduire les différentes problématiques d'apprentissage Nous présentons plus particulièrement la méthode utiliser les Séparateurs à Vaste Marge. Enfin, nous abordons le problème de l'évaluation de l'apprentissage et celui de la sélection de modèle.

2.2 L'apprentissage

Cette notion englobe toute méthode permettant de construire un modèle de la réalité a partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement le modèle. Il existe deux tendances principales en apprentissage, celle issue de l'intelligence artificielle et qualifiée de symbolique, et celle issue des statistiques et qualifiée de numérique**.** **[13]** Ces méthodes sont utilisées dans de nombreuses applications comme la reconnaissance de formes, le diagnostic médical, la bioinformatique, les interfaces cerveau-machine…etc.

2.2.1 L'apprentissage supervisé

L'apprentissage supervisé désigne un type d'algorithme d'apprentissage automatique qui utilise un ensemble de données connues (appelé ensemble de données d'apprentissage) pour réaliser des prévisions. L'ensemble de données d'apprentissage contient des données d'entrée et des valeurs de réponses. L'objectif de l'algorithme d'apprentissage supervisé consiste à créer un modèle à partir de ces éléments capable de réaliser des prévisions sur les valeurs de réponse pour un nouvel ensemble de données. Un ensemble de données de test est souvent utilisé pour valider le modèle. Les ensembles de données d'apprentissage plus volumineux produisent généralement des modèles à la puissance prédictive plus élevée qui peuvent facilement s'appliquer aux nouveaux ensembles de données. **[12]**

CHAPITRE 2 : METHODE D'APPRENTISSAGE

L'apprentissage supervisé inclut deux catégories d'algorithmes :

- **Classification**: pour les valeurs de réponse catégoriques, où les données peuvent être divisées en « classes » spécifiques [12]. Par exemple, pour prédire si un mail est SPAM ou non, la valeur de réponse peut prendre deux valeurs possibles :

  Y € {SPAM, NON SPAM}.

Quand l'ensemble des valeurs possibles d'une classification dépasse deux éléments, on parle de classification multi-classes.

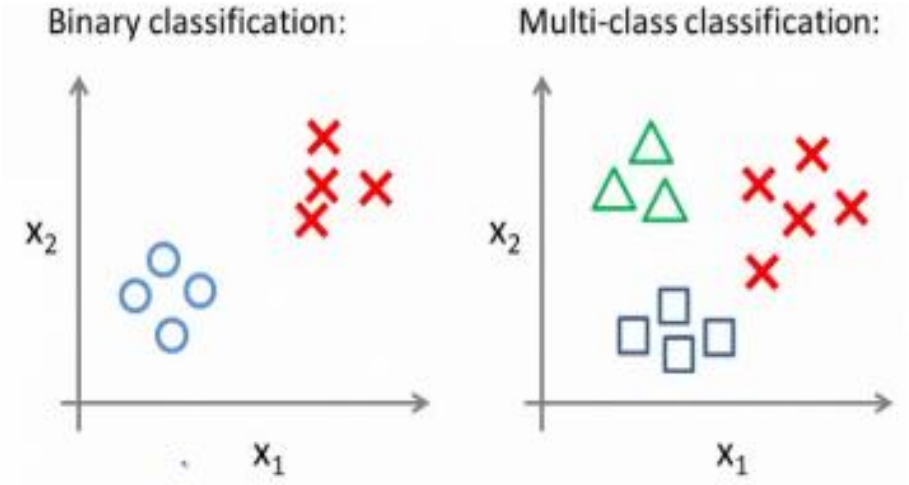


Figure 2.1 :- les deux types de classifications.

Parmi les algorithmes de classification en trouve :

- Machines à vecteurs de support.
- Réseaux de neurones.
- Classificateur bayésien naïf.
- Arbres de décision.
- Analyse discriminante.
- Plus proches voisins (KNN).

Chacun de ses algorithmes a ses propres propriétés mathématiques et statistiques. En fonction des données d'entrainement et nos features (caractéristiques), on optera pour l'un ou l'autre de ces algorithmes. Toutefois, la finalité est la même.

- **Régression**: pour les valeurs de réponse continue **[12].** Pour illustrer cette notion, on peut penser à un algorithme qui prend en entrée des caractéristiques d'un véhicule, et tentera de prédire le prix du véhicule.

Parmi les algorithmes de régression on trouve :

- Régression linéaire.
- Régression non linéaire.
- Modèles linéaires généralisés.
- Arbres de décision.
- Réseaux de neurones.

2.2.2 L'apprentissage non supervisé

L'apprentissage non supervisé représente une autre problématique d'apprentissage statistique, dans laquelle les étiquettes des données en entrée ne sont pas connues durant le processus d'apprentissage. L'objectif consiste alors à identifier une structure sous-jacente parmi les données. Un des problèmes étudiés dans ce cadre consiste à utiliser les attributs des données pour les regrouper entre elles de façon à ce que les données similaires se retrouvent dans le même groupe, et les données dissimilaires dans des groupes différents. Un des algorithmes les plus connus pour résoudre ce problème de partitionnement des données est celui des *k* moyennes. Un autre problème concerne la réduction de la dimension dans le cas de données de très grandes dimensions. **[10]**

2.2.3 L'apprentissage semi-supervisé

L'apprentissage semi-supervisé est une problématique d'apprentissage statistique qui utilise lors de l'apprentissage à la fois les données étiquetées, généralement en petite quantité car coûteuses à obtenir, et un grand nombre de données non étiquetées. L'apprentissage semi-supervisé se situe à la frontière entre l'apprentissage non supervisé et l'apprentissage supervisé. Il peut être vu comme de l'apprentissage non supervisé avec des contraintes ou

de l'apprentissage supervisé avec des informations supplémentaires sur la distribution des exemples. Il apparaît que les données non étiquetées, lorsqu'elles sont utilisées en conjonction avec une petite quantité de données étiquetées, peuvent améliorer sensiblement la précision de l'apprentissage**.** **[10]**

2.3 Le Support Victor Machin (SVM)

SVM est une méthode de classification binaire par apprentissage supervisé, elle fut introduite par Vapnik en 1995. Cette méthode est donc une alternative récente pour la classification. Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié. Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonction dites noyau (kernel) qui permettent une séparation optimale des données. **[14]**

2.3.1 Principe de la technique SVM

Le principe de base des SVM consiste de ramener le problème de la discrimination à celui, linéaire, de la recherche d'un hyperplan optimal. Deux idées ou astuces permettent d'atteindre cet objectif :

- La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes.
- Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau (kernel) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire de plus grande dimension.

2.3.1.1 L'hyperplan

L'objectif est de produire un classificateur qui fonctionnera bien sur des exemples

invisibles, c'est-à-dire qu'il se généralise bien. Il existe plusieurs classificateurs linéaires possibles qui peuvent séparer les données, mais il n'y en a qu'une qui maximise la marge (maximise la distance entre elle et le point de données le plus proche de chaque classe). Ce classificateur linéaire est appelé hyperplan à séparation optimale. **[15]**

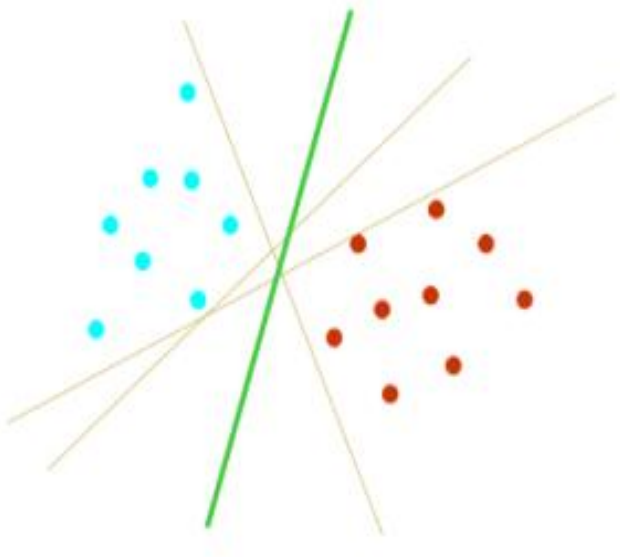


Figure 2.2 :- Hyperplan à séparation optimale.

2.3.1.2 Marge

Il existe une infinité d'hyperplans capable de séparer parfaitement les deux classes d'exemples. Le principe des SVM est de choisir celui qui va maximiser la distance minimale entre l'hyperplan et les exemples d'apprentissage (la distance entre l'hyperplan et les vecteurs supports), cette distance est appelée "la marge". **[16]**

2.3.1.3 Maximiser la marge

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. **[20]** Dans le schéma qui suit, la partie droite nous montre

qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé.

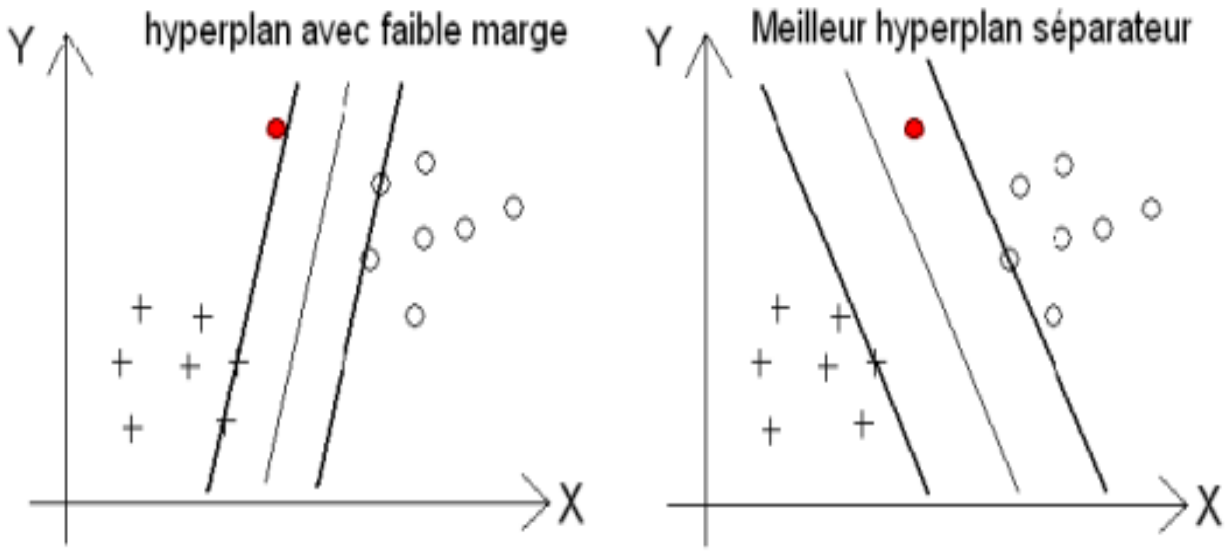


Figure 2.3:- Exemple d'un hyperplan optimal **[20]**

En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal.

2.3.1.4 Vecteurs supports

Pour une tâche de détermination de l'hyperplan séparable des SVM est d'utiliser seulement les points les plus proches (les points de la frontière entre les deux classes des données) parmi l'ensemble total d'apprentissage, ces points sont appelés "vecteurs supports".

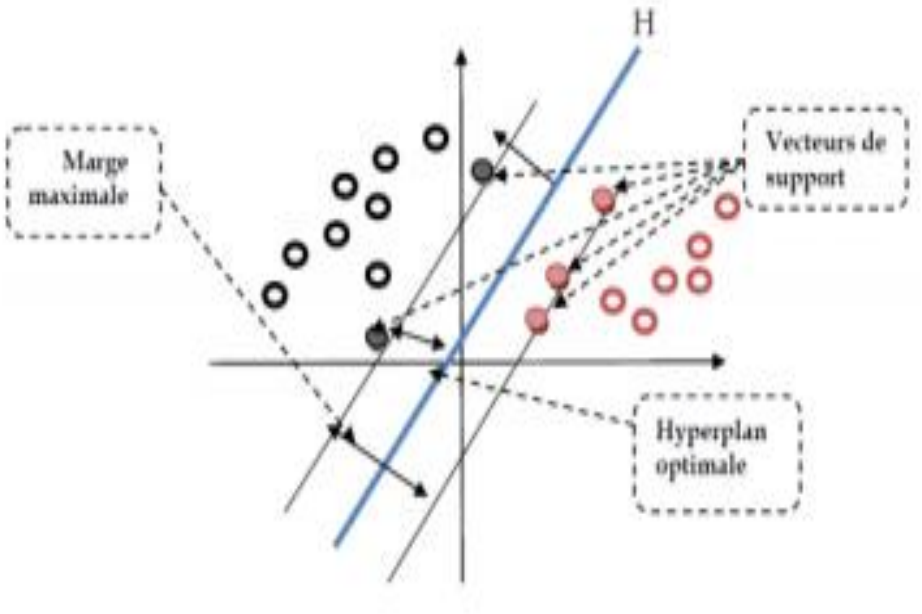


Figure 2.4 :- L'hyperplan H optimal, vecteurs supports et marge maximale.

2.3.2 SVM à marge dure

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe (au milieu) des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le (plus sûr). En eet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. On appelle cette distance (marge) entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de séparateurs à vaste marge (SVM a marge dure). **[14]**

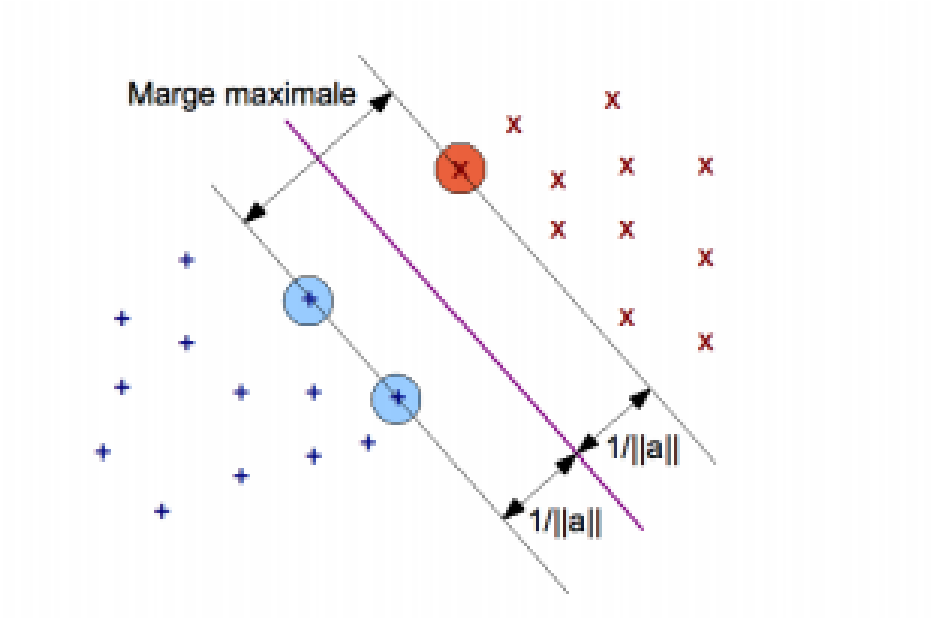


Figure 2.5 :- Données d'apprentissage avec une marge maximale.

2.3.3 SVM à marge souple

La marge souple est considérée comme une relaxation de la marge dure justifiée par la présence des exemples mal classifiés appartenant a la marge (dite erreur de marge) conduisant a une impossibilité de classification avec un séparateur linéaire. **[17]**

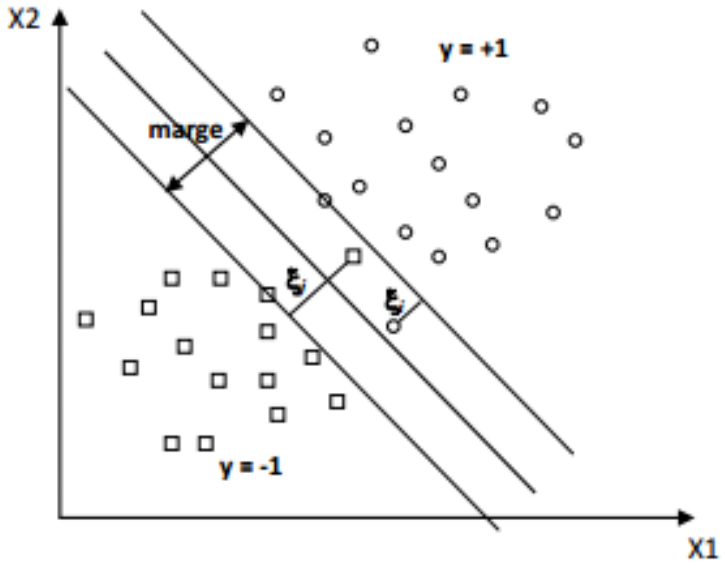


Figure 2.6 :- Marge souple.

2.3.4 SVM a Kernel

Les limites de l'approche a marge souple s'expose avec les données non linéairement séparable a tout point de l'espace, la motivation derrière l'utilisation des fonction Kernel est la possibilité de projecter les valeurs des données vers un autre espace d'une dimension supérieur ou la séparation linéaire est possible, ce qui mène a utiliser ces approches pour classifier.

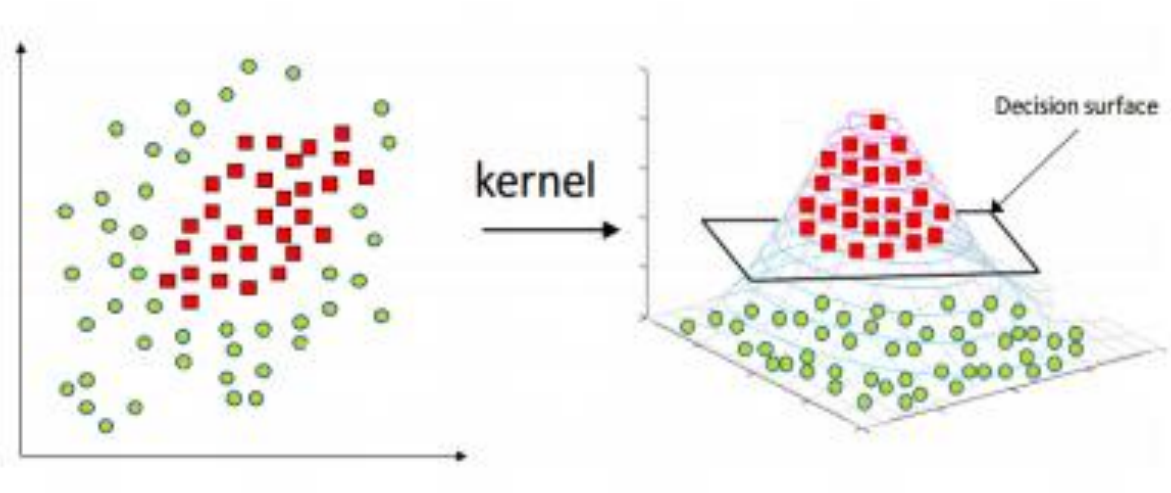


Figure 2.7 :- Points non linéairement séparable et leur projection vers un autre espace d'une dimension supérieur.

On général le choix d'un Kernel correspond a :

- Choisir une mesure de similarité pour les données.
- Choisir une représentation linéaire aux données.
- Choisir un espace de fonction pour l'apprentissage. **[18]**

2.3.5 SVM multi classes

À l'origine, Les machines à vecteur support sont binaires. Cela dépend de trouver le meilleur hyperplan. Mais dans le monde réel, la plupart des problèmes sont multiclasse. Dans de tels cas, on ne cherche pas à affecter un nouvel exemple à l'une de deux classes mais à l'une parmi plusieurs, c-à-d que la décision n'est plus binaire et un seul hyperplan ne suffit plus.

Les méthodes des machines à vecteur support multi classe, réduisent le problème multi classe à une composition de plusieurs hyperplans biclasses permettant de tracer les frontières de décision entre les différentes classes. Ces méthodes décomposent l'ensemble d'exemples en plusieurs sous ensembles représentant chacun un problème de classification binaire. Pour chaque problème un hyperplan de séparation est déterminé par la méthode SVM binaire. On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple**.** **[19]**

2.3.5.1 Les méthodes de décomposition

Les méthodes de décomposition permettent d'aborder un problème de discrimination à catégories multiples comme une combinaison de problèmes de calcul de dichotomies. On trouve plusieurs méthodes de décomposition et parmi ces méthodes on a :

2.3.5.1.1 Une-contre-reste (1vsR)

L'approche "un contre tous" est la plus simple et la plus ancienne des méthodes de décomposition. Elle consiste à utiliser un classifieur binaire (à valeurs réelles) par catégorie. chaque classifieur sépare la classe **i** de toutes les autres classes. **[21]**

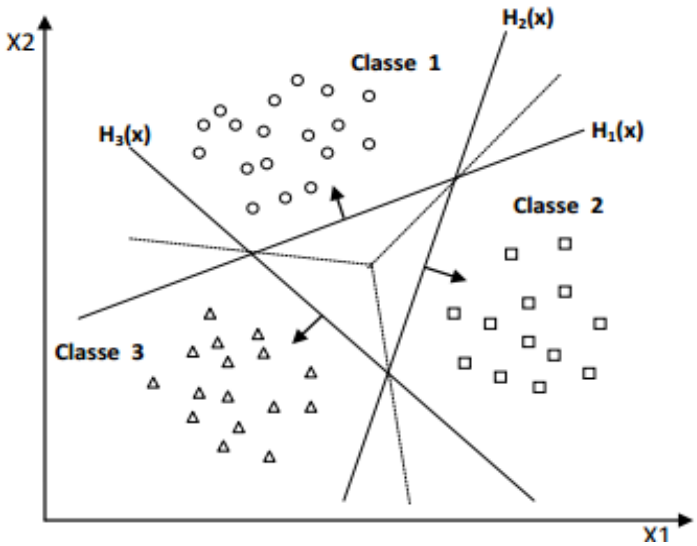


Figure 2.8 :– Résolution des cas d'indécision dans la méthode 1vsR.

La méthode 1vsR peut être utilisée pour découvrir même les cas de rejet où un exemple n'appartient à aucune des K classes**.** [22] Soulent les classificateur on classifier le nouvel exemple a quel classe il appartenant. Le problème est que l'exemple peut être vérifié pour plus d'une classe, ce qui produit des régions d'ambiguïté. Cet exemple sera rejeté est ne sera affecté a aucune classe, cet écartement de l'exemple est appelé Reject decision.

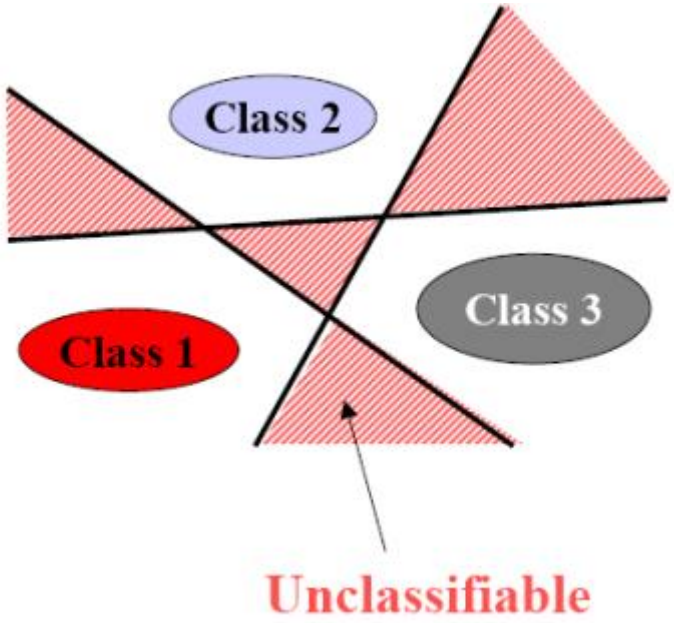


Figure 2.9 :- Les surfaces rouges représente le « reject decision ».

2.3.5.1.2 Une-contre-une (1vs1)

Cette méthode, appelée aussi "pairwise", Ordinairement attribuée à Knerr et ses co-auteurs**.** [23] Elle consiste à utiliser un classifieur pour chaque paire de classes. Au lieu d'apprendre *K* fonctions de décisions, la méthode 1vs1 discrimine chaque classe de chaque autre classe. L'affectation d'un nouvel exemple se fait par liste de vote. On teste un exemple par le calcul de sa fonction de décision pour chaque hyperplan. Pour chaque test, on vote pour la classe à laquelle appartient l'exemple (classe gagnante). Un nouvel exemple est affecté à la classe la plus votée.

Malheureusement, la fonction peut être vérifiée pour plusieurs classes, ce qui produit

des zones d'indécisions. La méthode de vote affecte dans ce cas, un exemple aléatoirement à l'une des classes les plus votées. [19]

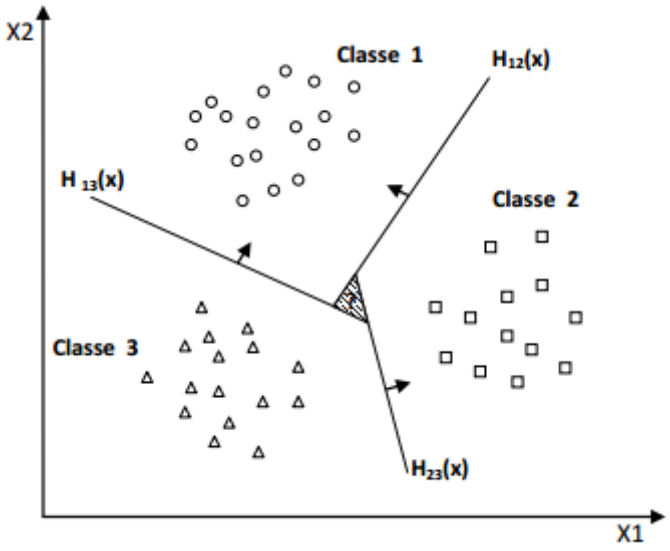


Figure 2.10 : - Classification multi class par paire.

2.3.6 Avantage du SVM

En le comparant aux autres méthodes d'apprentissage tel que les réseaux de neurones et les arbres de décision, les SVMs sont en avance dans plusieurs points :

- Les SVMs permettent de traiter plusieurs problèmes de datamining : classification, Régression, clustering, détection des outliers...etc.

- Les SVMs permettent de traiter les données numériques et symboliques, ce qui le favorisées dans plusieurs applications complexes tel que le textmining, la Reconnaissance des images, la reconnaissance vocale, les séquences biologiques...etc.

- Les capacités de généralisation et la simplicité d'entrainement des SVMs sont bien au- delà des autres méthodes.

- Les SVMs sont très efficaces sur les données à nombre élevé d'attributs, même avec Peu d'exemples. Elles n'imposent aucune limite sur le nombre d'attributs sauf les limites imposées par le hardware. **[19]**

2.4 Evaluation de l'apprentissage et sélection de modèle

La performance en généralisation d'une méthode d'apprentissage se rapporte à sa capacité de prédiction sur des données de test indépendantes. La problématique de l'évaluation de cette performance apparaît dans deux types de problèmes [Hastie et al., 2008]:

2.4.1 La sélection de modèle

Une méthode d'apprentissage est caractérisée par différents hyper paramètres. La sélection de modèle revient à choisir les valeurs de ces hyper paramètres. Plusieurs critères peuvent être utilisés pour faire ce choix, comme la stabilité ou une mesure de performance.

2.4.2 L'évaluation d'un modèle

Cette évaluation consiste à estimer l'erreur de prédiction d'un modèle sur de nouvelles données, une fois celui-ci choisi.

Dans le cas où l'on dispose de beaucoup d'exemples, l'approche la plus simple pour les deux problèmes consiste à diviser ces exemples en trois ensembles : un ensemble d'apprentissage, un ensemble de validation et un ensemble de test. L'ensemble de validation est utilisé pour estimer l'erreur de prédiction pour la sélection de modèle et l'ensemble de test est utilisé pour évaluer l'erreur de généralisation du modèle choisi. **[10]**

2.5 Méthodes d'apprentissage

Plusieurs approches supervisées ont été utilisées pour l'inférence de réseaux biologiques [Qi & Noble, 2011], dont : la régression logistique [Lin et al., 2004], les méthodes à noyaux [Yamanishi et al., 2004; Ben-Hur & Noble, 2005], les arbres de décision [Zhang et al., 2004] et les forêts aléatoires [Qi et al., 2005]. Ces classifieurs ont

été comparés pour la prédiction de PPI chez la levure [Qi et al., 2006], et les méthodes qui ont obtenu les meilleures performances sont les forêts aléatoires et les séparateurs à vaste marge.

Dans le cas des réseaux sociaux, Al Hasan et al. [2006] ont également comparé plusieurs algorithmes de classification supervisée pour l'inférence d'un réseau de co-citations entre publications scientifiques à partir d'informations sur la topologie et sur les caractéristiques des nœuds. L'approche utilisant les SVM obtient les meilleurs résultats. Par ailleurs, Lichtenwalter et al. [2010] ont proposé deux méthodes d'ensemble, basées sur le bagging et les forêts aléatoires, pour la prédiction de liens dans le cadre supervisé**. [10]**

2.6 Conclusion

Dans ce chapitre, nous avons essayé de présenter de manière simple et complète le concept de système d'apprentissage introduit par Vladimir Vapnik, les « Support Vector Machine » Nous avons donné une vision générale des SVM. Cette méthode de classification est basée sur la recherche d'un hyperplan qui permet de séparer au mieux des ensembles de données. Nous avons exposé les cas linéairement séparable et les cas non linéairement séparables qui nécessitent l'utilisation de fonction noyau (kernel) pour changer d'espace. Cette méthode est applicable pour des tâches de classification à deux classes, mais il existe des extensions pour la classification multi classe.

CHAPITRE 3 :

CONCEPTION DU SYSTEME

3.1 Introduction

Selon les méthodes expérimentales et les laboratoires, la prédiction des interactions protéiques est coûteuse, prend du temps et ajouter des résultats inexacts a cause de nombreux facteurs tels que le milieu réactionnel et le manque d'équipement et d'outils expérimentaux. Par conséquent, des méthodes techniques ont été proposées pour remplissez et remplacez ces problèmes et ont obtenu un grand succès par rapport aux méthodes expérimentales.

Le but de notre sujet est de concevoir un système qui permet la prédiction des interactions protéine-protéine en utilisant seulement leurs Chaînes des acides aminés, nous allons dans ce chapitre présenter les différentes étapes de conception d'un tel système. Nous avons suivis la méthode descendante qui consiste en le découpage du système en module fonctionnel globales puis le découpage de chaque modules récursivement en sous modules jusqu'à arriver à des modules élémentaire.

3.2 Méthodologie suivie

Pour atteindre notre programme, nous avons d'abord préparé la base de données sur laquelle nous allons effectuer les traitements à partir de bases de données contenant des chaînes protéiques.

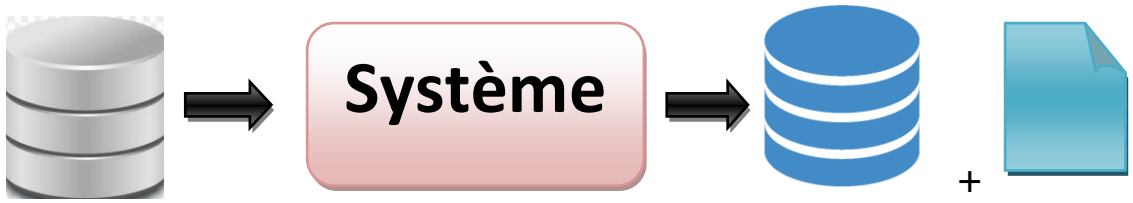
Puisque chaque protéine exprimée dans la base de données avec une série d'acides aminés, nous l'avons d'abord convertie en une forme qui nous aide dans le calcul en utilisent une matrice de substitution SM_BLOSUM62. Après avoir converti toutes les protéines de notre base de données, nous formerons la base des paires des protéines à partir faire une concaténation entre chaque deux protéine. Mais cela rendra la base de données très volumineuse, ce qui prendra beaucoup de temps pour la traiter, et pour résoudre ce problème, nous avons suivi une méthode de réduction des données qui minimise la taille de la base.

Ensuite, après avoir obtenu la base de données d'une taille optimale des paires de protéines ,en utilise la méthode de « k-fold cross validation (validation croisée) » qui divise a chaque fois notre base de données en deux base (base d'entrainement , et base de teste) . On répète l'opération 5 fois en sélectionnant des autres échantillons d'entrainement et de test. Puis et à l'aide de notre méthode d'apprentissage qui basée sur les machines à vecteurs supports, Nous avons créé des modèles de prédiction du PPI avec les bases

d'entrainement que nous avons utilisé pour classer les paires dans les bases du testes en paires de protéines positives (interagissent entre elles) et les paires de protéines négatives (n'interagissent pas entre elles). Puis nous l'avons comparé à la classification réelle des bases de données, ce qui nous a donné un taux d'erreur de 0.02 % .ce qui indique l'efficacité de notre méthode.

3.3 Conception globale du système

La conception globale de notre système est illustrée dans le schéma suivant :



Base de données Des chaines protéines

Bases des paires protéines classifiées

Modèle de prédiction

Figure 3.1 :- conception globale.

Ce système a comme entrée une base de données contenant des chaînes de protéines sous forme des chaînes d'acides aminés. Et a comme sortie le résultat de la prédiction qui est les paires des protéines Interagir (positive) et non interagir (négative) dans les bases du test avec des modèles utilisant dans la prédiction d'interaction protéines.

3.4 Conception détaillé

Notre système est composé de deux modules :

Module préparation des données :

Responsable de recevoir la base de données saisie et de la convertir en format matricielle pour faciliter son étude puis de sélectionner les paires de protéines qui seront utilisées dans l'étude.

Module de la prédiction

C'est la partition responsable d'avoir des modèles puis de classer les éléments des bases de test en deux classes (positive et négative) selon ces modèles.

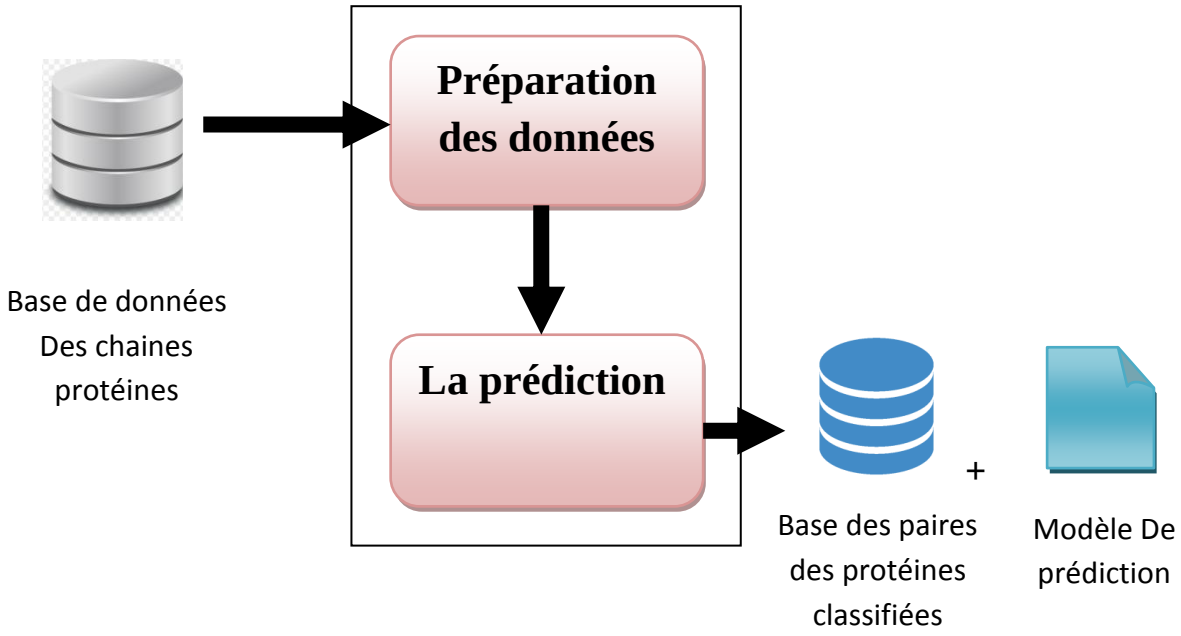


Figure 3.2 :- Conception détaillé.

3.4.1 Module préparation des données

L'entré de ce module est une base de données des chaines protéines.la base est composée par 4 sous bases (deux exprimant des protéines positives Connu pour leur interaction et deux négatives connus pour ne pas interagir). Ensuite nous passons par une suite d'opérations dans les quatre étapes suivantes :

3.4.1.1 La conversion

Pour prédire le PPI par séquences, l'un des principaux défis de calcul est de trouver un moyen approprié de décrire complètement les informations importantes du PPI. Pour résoudre ce problème, Nous représentons la séquence protéique comme une matrice de représentation matricielle de substitution (SMR). La séquence protéique de longueur L donnée peut être représentée comme une matrice L × 20, basée sur une matrice de substitution. Nous utilisons la matrice SM_BLOSUM62 comme matrice de substitution. Où chaque fois que nous remplaçons un acide aminé de la séquence protéique par la ligne de longueur 20 correspondante de la matrice jusqu'à ce que nous substituions tous les acides de la chaîne et nous les gardons tous dans la même matrice .Ensuite, Nous mettons chacun sur une ligne dans une nouvelle matrice correspondant. Nous faisons cela afin de transformer les protéines des quatre bases.

3.4.1.2 Extraction des paires de protéines

Après avoir obtenu quatre matrices chiffrées de l'étape précédente, nous faisons une concaténation entre chacune des deux matrices (positives ensemble et négatives ensemble). On obtient deux matrices avec des précédents 1 pour chaque ligne de la matrice des paires positives, et 2 pour chaque ligne de la matrice des paires négatives. Ensuit, Nous avons ajouté les deux matrices dans une seule matrice en les plaçant l'une sous l'autre.

3.4.1.3 Réduction des données

Le but de cette étape est d'obtient une représentation réduite de l'ensemble de données qui est beaucoup plus petit en volume mais qui produit les mêmes (presque les mêmes) résultats analytiques. On utilisons le pattern de reconnaissance 2D_LDA (Pattern Recogntion) qui vise à trouver la transformation optimale (projection) telle que la structure de classe de l'espace original de haute dimension soit préservée dans l'espace de basse dimension. Après avoir obtenu une base de données idéale, nous avons précédé les paires positives qui étaient précédées du numéro 1 avec le numéro 1 et les paires négatives qui étaient précédées du numéro 2 avec le numéro -1.

3.4.1.4 Validation croisée

Une technique pour évaluer comment les résultats de l'analyse statistique se généralisent à un ensemble de données indépendant. . Il est principalement utilisé dans les contextes où

l'objectif est la prédiction, et l'on souhaite estimer la précision avec laquelle le modèle prédictif fonctionne dans la pratique.

Lors d'un cycle de validation croisée, nous divisons notre base de donne de données en sous-ensembles complémentaires, effectuons l'analyse sur un sous-ensemble (appelé ensemble d'apprentissage ou d'entrainement contenant 80% de la base de données totale), et validons l'analyse sur l'autre sous-ensemble (appelé ensemble de validation ou ensemble de test contient les 20% autres de la base de données totale). 5 tours de validation mutuelle sont effectués en utilisant 5 sections différentes où à chaque fois que nous changeons le contenu des deux bases en changeant la partie sélectionnée pour la base de test. Cela nous permet d'utiliser toutes nos données de base de données pour l'entrainement ainsi que pour les tests.

Ensuit, pour chaque base de données résultante, nous séparons le contenu de la première colonne de chaque ligne et le plaçons dans une nouvelle base de données. Autrement dit, chaque base de données sera divisée en deux parties, une section contenant des caractéristiques (Features) et une section contenant les classes correspondante(Label).

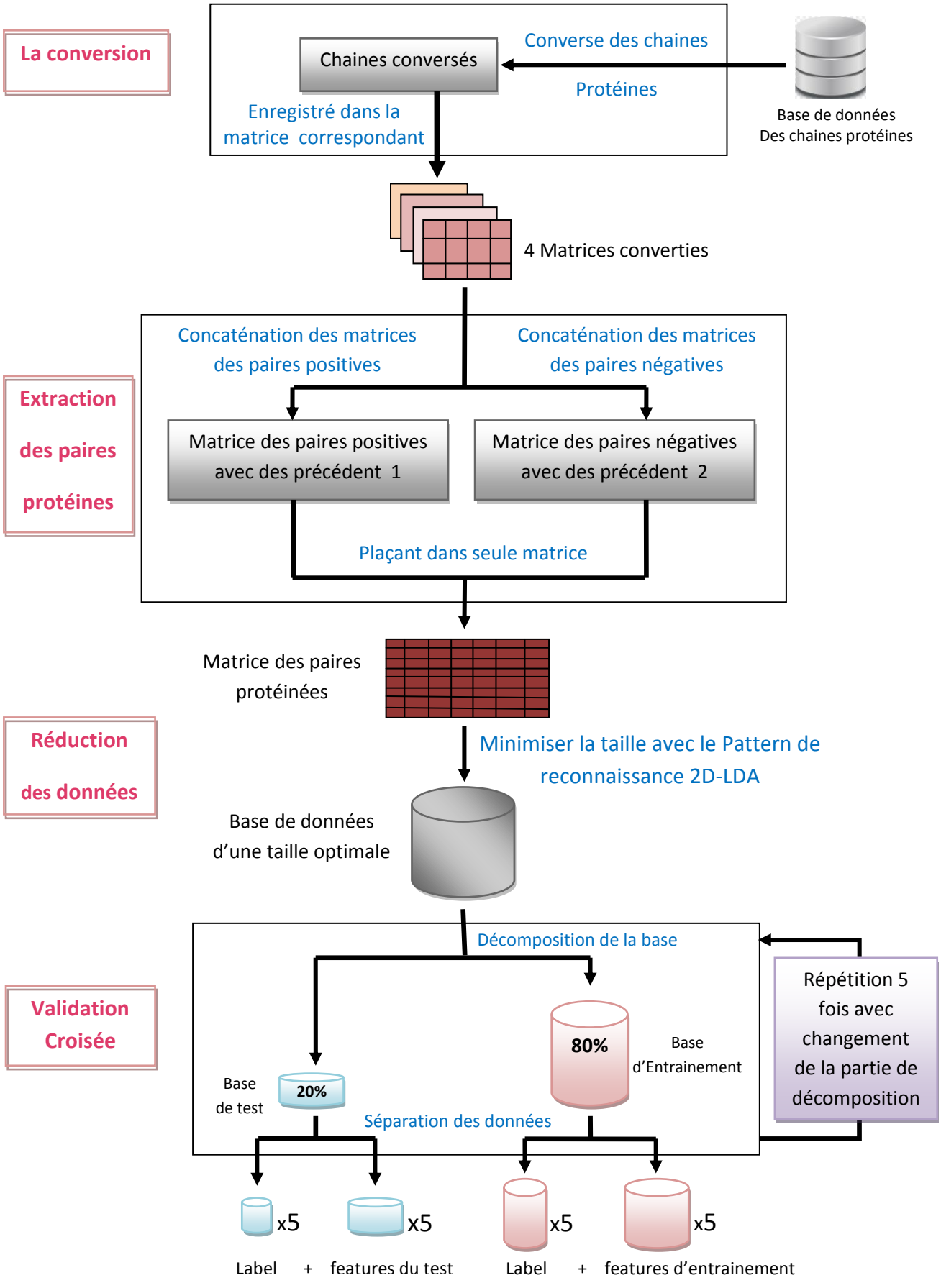


Figure 3. 3 :- Conception du module Préparation des données.

3.4.2 Le module prédiction

Lorsque l'on utilise des méthodes d'apprentissage, on utilise généralement deux groupes de données principaux : le groupe d'entraînement et le groupe de test. Pour cela ce module a come entrée un groupe du base d'entrainement et un groupe des base de test qui nous l'avons obtenu de le module précédente. Ensuit on passe par les étapes suivant :

3.4.2.1 Création du modèle

Depuis que nous avons cinq bases de données attachées à cinq bases de données de leur classification. Dans cette partie, nous avons créé cinq modèles de prédiction à l'aide de SVM où dans chaque modèle nous changeons la base utilisée.

3.4.2.2 Teste du modèle

Cette étape utilisée pour évaluer les performances de généralisation du modèle. La qualité de ce modèle est alors jugée à sa capacité à réduire l'erreur de test.

Le modèle qui maximise la précision, Rappel (Recall) et taux de reconnaissance (Accurecy) sur les données de test peut être utilisé dans la prédiction. Pour cela, nous avons distribué les bases de données restantes afin d'étudier chaque modèle, où à chaque fois en utilisant SVM et l'un des cinq modèles, nous classons les données d'une base, puis comparons les résultats avec la classification réelle et calculons son Précision, Rappel (Recall) et son taux de reconnaissance (Accurecy) et à la fin nous avons calculé le moyen des taux de reconnaissances obtenus et le taux d'erreur des modèles pour voir l'efficacité de notre méthode.

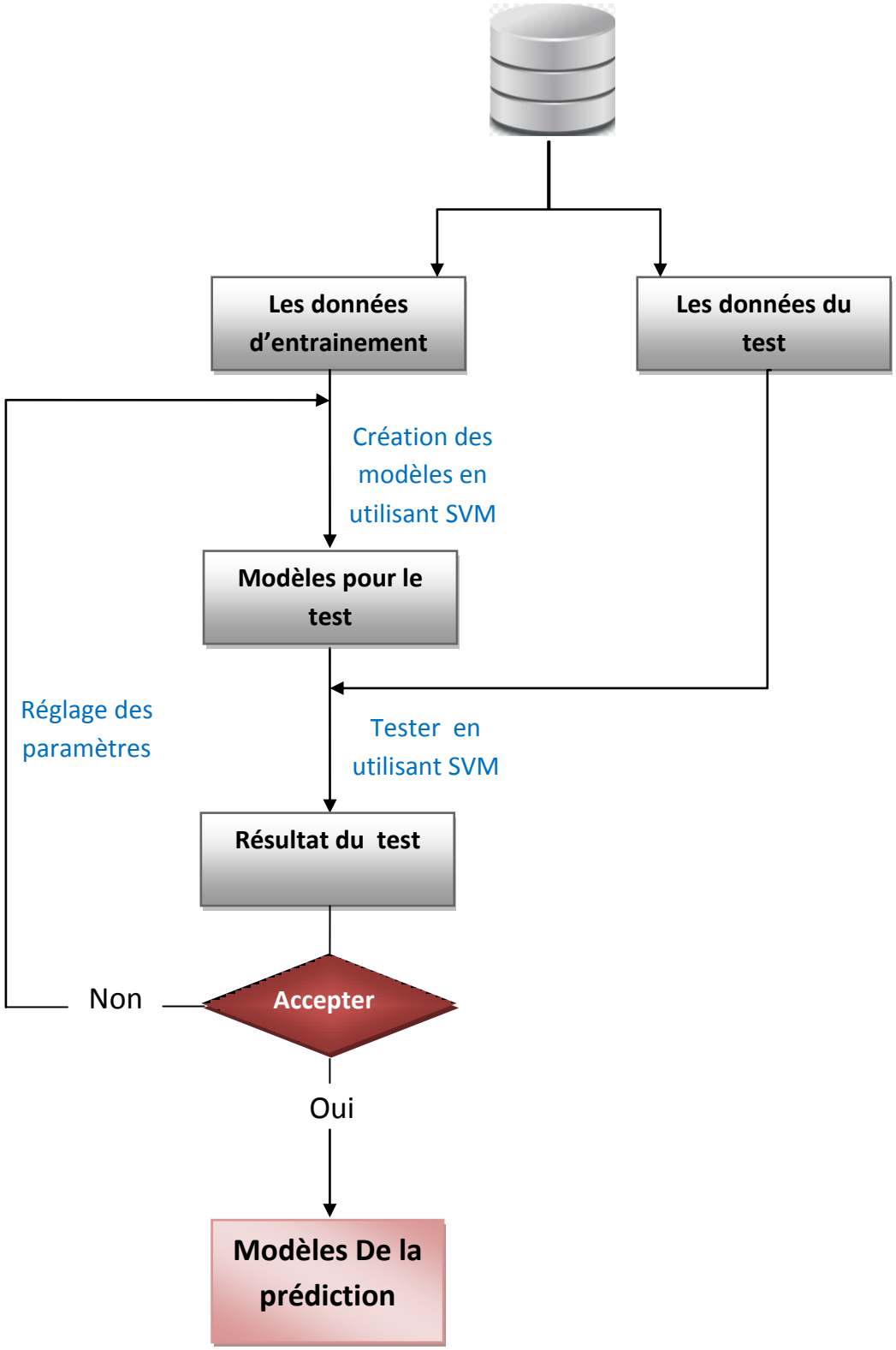


Figure 3. 4 :- Conception du module prédiction.

3.5 Conclusion

Dans ce chapitre, nous avons présenté notre méthode proposée. Ensuite nous avons présenté la conception de notre système, la conception globale des deux modules (préparation des données et Prédiction). Enfin nous avons détaillé la conception de chaque module avec les schémas correspondent.

Dans le chapitre suivant, nous allons présenter la mise en œuvre de notre conception. En outre, nous allons présenter les différents détails de l'implémentation tels que les outils et langages de programmation que nous avons utilisés. Nous présentons également les résultats obtenus.

CHAPITRE 4 :

IMPLEMANTATION ET RESULTAT

4.1 Introduction

Après avoir détaillé dans le chapitre précédent l'approche de mise en œuvre pour les parties de notre système, Nous allons dans ce qui suit détaillé l'implémentation des différents modules de conception pour réaliser le système conçu.

Étant donné que la conception détaillée prend en compte la plate-forme de réalisation, Nous commençons dans la première partie en définissant l'environnement de programmation et les outils utilisés dans le développement. Puis nous avons présentés l'interface de notre système. Ensuite nous expliquerons toutes les expérimentations que nous avons appliquées sur la méthode proposée et les résultats obtenus.

4.2 Environnement et outils de programmation

4.2.1 Langages de programmation

4.2.1.1 MATLAB

Figure 4.1 : Matlab.

MATLAB (« matrice laboratoire ») est un logiciel commercial de calcul interactif, et un langage de programmation de quatrième génération émulé par un environnement de développement du même nom. Il est utilisé à des fins de calcul numérique. Développé par la société The MathWorks, la première version est met en 1984, MATLAB permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, de créer des interfaces utilisateurs.

MATLAB vous aide à faire passer vos idées au-delà du bureau. Vous pouvez exécuter vos analyses sur des ensembles de données plus volumineux et évoluer vers des clusters et des nuages. Le code MATLAB peut être intégré à d'autres langages, comme le C, C++, Java, et Fortran.et on peut l'utiliser en Linux, Unix, Mac OS, Windows. Ce qui vous permet de déployer des algorithmes et des applications au sein de systèmes Web, d'entreprise et de production. **[12]**

4.2.2 Outils utilisés

4.2.2.1 Libsvm

LIBSVM est une bibliothèque développée par Chih-Chung Chang et ChihJen Lin, deux chercheurs du département informatique de l'université national de Taiwan, depuis l'année 2000. Cette bibliothèque est développée pour implémenter dans la résolution des problèmes liée a la classification avec l'utilisation des machine a vecteur support, les machine a vecteur que ce soit la classification mono-class, binaire, multi-classes et même pour la régression. Elle est disponible en langage C++ et en JAVA, et compatible avec diverses plateformes logicielles (Python, R, MATLAB, Perl, Ruby, Weka, Common LISP, CLISP, Haskell, LabVIEW, interfaces PHP, C# .NET, extensions CUDA).

Une utilisation typique de LIBSVM implique deux étapes : d'abord, la formation d'un ensemble de données pour obtenir un modèle et la seconde, est utilisant le modèle pour prédire l'information d'un ensemble de données de test. Pour SVC et SVR, LIBSVM peut également produire des estimations de probabilité**.** **[31]**

4.2.2.1.1 Paramètres SVM pris en charge par la librairie [32]

An d'entamer le processus d'apprentissage avec la LIBSVM, certains paramètres sont à renseigner selon qu'on souhaite faire une classification, une régression ou autre, le choix des bons paramètres est déterminant pour obtenir des résultats satisfaisants. Ci-dessous la liste des paramètres LIBSVM les plus utilisés. À noter que dans le cas où ces paramètres ne sont pas renseignés, la librairie utilise les valeurs par défaut.

-s svm type : C'est le type de l'algorithme SVM à utiliser, peut être l'une des fonctions :

C, SVC, NU SVC, ONE CLASS, EPSILON SVR, NU SVR.

-t kernel type : C'est le type de la fonction noyau à utiliser, peut être défini à : LINEAR, POLY,RBF ou SIGMOID.

$$\text{Linaire} = u_0 * v$$

$$\text{polynomiale} = (\text{gamma} * u_0 * v + \text{coef0})^{\text{degré}}$$

$$\text{Radiale} = \exp(-\text{gamma} * |u - v|_2)$$

$$\text{sigmoïde} = \tanh(\text{gamma} * u_0 * v + \text{coef0})$$

Tels que **u** représente la transposé du vecteur contenant les valeurs des attributs de l'ensemble d'apprentissage, et **v** le vecteur des labels (étiquettes). **Le gamma**, **degré** et **coef0** sont des paramètres (rentrés par l'utilisateur).

Paramètre des fonctions noyau

-d degré **:** Paramètre degré de la fonction noyau, par défaut 3.

-g gamma **:** Paramètre gamma de la fonction noyau, par défaut 1.

-r coef0 **:** Paramètre coef0 de la fonction noyau, par défaut 0.

Paramètres dépendants du type SVM choisi

-c cost **:** C'est le paramètre C (coût), qui représente la pénalité de l'erreur, à renseigner lors de L'utilisation du type SVM C-SVC, epsilonSVR et nu-SVR, par défaut le coût est égal à 1.

-wi weight **:** pour changer le paramètre C à (weight*C), s'il n'est pas renseigné weight est égale à 1 sa valeur par défaut, et par conséquent neutre.

-n nu **:** Paramètre nu du type nu-SVC, One-class-SVM et nu-SVR, par défaut 0.5.

-p epsilon **:** Paramètre epsilon de la fonction de perte (Loss Function) pour le type epsilon-SVR, par défaut égal à 0.1.

4.2.2.2 La matrice SM_BLOSUM62 [33]

Les matrices de similarité ou matrices de substitution sont des matrices utilisées en bioinformatique pour réaliser des alignements de séquences biologiques reliées évolutivement. Elles permettent de donner un score de similarité ou de ressemblance entre deux acides aminés. Ces matrices, M, sont des matrices 20 x 20 (pour les 20 acides aminés protéinogènes standards) qui recensent l'ensemble des scores M(a,b) obtenus lorsqu'on substitue l'acide aminé **a** à l'acide **b** dans un alignement. Plus le score M (a,b) est élevé, plus la similarité entre les deux acides aminés **a** et **b** est importante.

La matrice SM_BLOSUM62 que nous avons utilisé est basée sur le contenu en information des substitutions. Elle est calculée à partir des fréquences de substitution d'acides aminés dans des blocs de séquence conservés, sans insertion, présentant au moins 62 % de conservation de séquence. Les acides aminés sont indiqués par leur code à une lettre. Les coefficients de la matrice sont exprimés en demi-bits d'information :

- Une valeur nulle indique une substitution neutre.
- Un score positif correspond à une substitution surreprésentée et donc probablement favorable.
- Un score négatif correspond une substitution sous-représentée et donc probablement défavorable.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	6	-2	-2	-3	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	2	0	-4	-3	0
R	-2	8	-1	-2	-5	1	0	-3	0	-4	-3	3	-2	-4	-3	-1	-2	-4	-3	-4
N	-2	-1	8	2	-4	0	0	-1	1	-5	-5	0	-3	-4	-3	1	0	-6	-3	-4
D	-3	-2	2	9	-5	0	2	-2	-2	-5	-5	-1	-5	-5	-2	0	-2	-6	-5	-5
C	-1	-5	-4	-5	-13	-4	-5	-4	-4	-2	-2	-5	-2	-4	-4	-1	-1	-3	-4	-1
Q	-1	1	0	0	-4	8	3	-3	1	-4	-3	2	-1	-5	-2	0	-1	-3	-2	-3
E	-1	0	0	2	-5	3	7	-3	0	-5	-4	1	-3	-5	-2	0	-1	-4	-3	-4
G	0	-3	-1	-2	-4	-3	-3	8	-3	-6	-5	-2	-4	-5	-3	0	-2	-4	-5	-5
H	-2	0	1	-2	-4	1	0	-3	11	-5	-4	-1	-2	-2	-3	-1	-3	-4	3	-5
I	-2	-4	-5	-5	-2	-4	-5	-6	-5	6	2	-4	2	0	-4	-4	-1	-4	-2	4
L	-2	-3	-5	-5	-2	-3	-4	-5	-4	2	6	-4	3	1	-4	-4	-2	-2	-2	1
K	-1	3	0	-1	-5	2	1	-2	-1	-4	-4	7	-2	-5	-2	0	-1	-4	-3	-3
M	-1	-2	-3	-5	-2	-1	-3	-4	-2	2	3	-2	8	0	-4	-2	-1	-2	-1	1
F	-3	-4	-4	-5	-4	-5	-5	-5	-2	0	1	-5	0	9	-5	-4	-3	1	4	-1
P	-1	-3	-3	-2	-4	-2	-2	-3	-3	-4	-4	-2	-4	-5	11	-1	-2	-5	-4	-4
S	2	-1	1	0	-1	0	0	0	-1	-4	-4	0	-2	-4	-1	6	2	-4	-3	-2
T	0	-2	0	-2	-1	-1	-1	-2	-3	-1	-2	-1	-1	-3	-2	2	7	-4	-2	0
W	-4	-4	-6	-6	-3	-3	-4	-4	-4	-4	-2	-4	-2	1	-5	-4	-4	16	3	-4
Y	-3	-3	-3	-5	-4	-2	-3	-5	3	-2	-2	-3	-1	4	-4	-3	-2	3	10	-2
V	0	-4	-4	-5	-1	-3	-4	-5	-5	4	1	-3	1	-1	-4	-2	0	-4	-2	6

Figure 4.2 :- Matrice SM_BLOSOM62.

4.3 Système de prédiction d'interaction protéine-protéine

Nous avons développée une application implémentant la proposition présentée dans le chapitre précédent pour son test et validation. L'application est composée de deux parties représentant chacune une module de la méthode proposée :

- Préparation des données.
- Prédiction.

4.3.1 Base de données utilisé

L'ensemble de données a été téléchargé à partir du sous-ensemble de base de S.cerevisiae de la base de données des protéines (DIP). Notre base de données contient 7798 chaînes protéiques connues avec son interaction (Positive) divisé en deux groupes (Positive_a contient 3899 chaînes protéiques, et Positive_b contient 3899 chaînes protéiques). Et aussi 8524 chaînes protéiques connues pour ne pas interagir (Négative) divisé en deux groupes (Négative_a contient 4262 chaînes protéiques, et Négative_b contient 4262 chaînes protéiques).

La figure suivante représente à quoi ressemble notre base de données lorsque nous l'ouvrons avec le MATLAB:

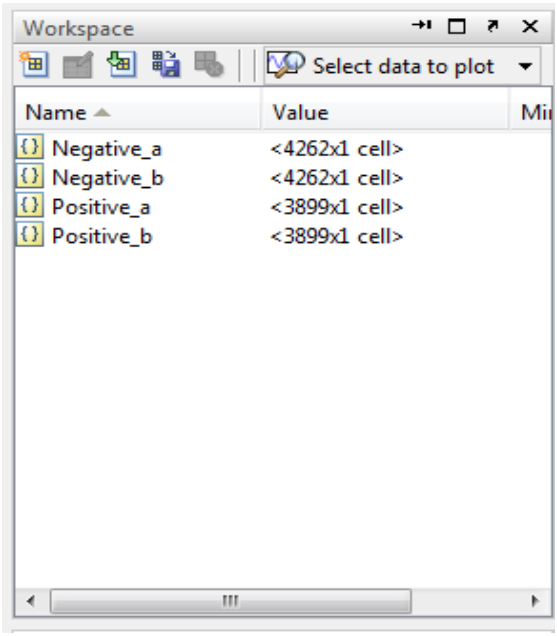


Figure 4.3:- Les quatre groupes qui composent notre base de données.

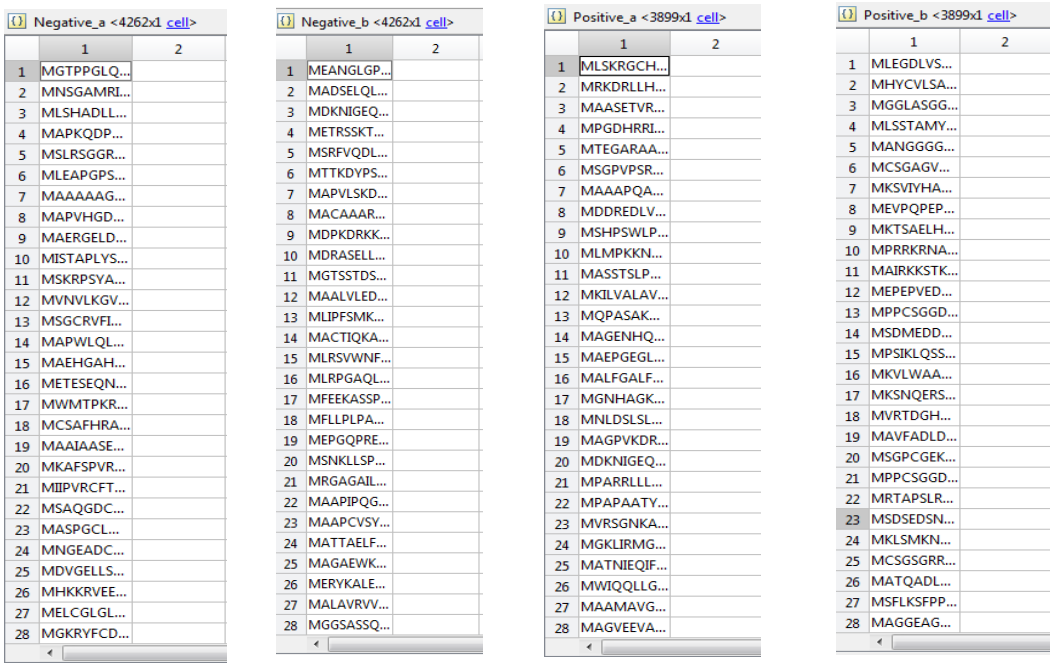


Figure 4.4:- Le contenu de chaque groupe dans la base de données.

Dans notre base de données, chaque ligne de chaque groupe représente une chaîne d'acides aminés d'une protéine.

4.3.2 Préparation des données

Pour pouvoir utiliser la base de données pour la prédiction. Nous avons appliqué un prétraitement pour adapter les types des attributs au format utilisé, en suivant les étapes suivantes :

- Conversion des chaînes protéiques avec la matrice SM_BLOSUM62.
- Extraction de paires de protéines (positives ensemble et négatives ensemble).
- La réduction des données avec le Pattern de reconnaissance 2D_LDA qui réduit la taille de la base de données.
- Enfin, nous faisons une validation croisée, nous voudrons faire 5 modèles de prédiction. Nous répétons donc cette étape 5 fois et à chaque fois nous divisons la base de données en deux parties. L'un est destiné au trainning (80% de la base) et l'autre aux tests (20% de la base).

Les résultats de ce module sont un ensemble de bases de données pour l'entrainement (5 bases de données contenant les caractéristiques et 5 contenants la classification

correspondante), et un ensemble de bases de données pour le test (5 bases de données contenant les caractéristiques et 5 contenants la classification correspondante).

4.3.3 Prédiction

Dans ce module, nous construire 5 modèles de prédiction en utilisant :

- Les bases d'entrainement (pour chaque modèle on utilise une base contient les caractéristiques avec son base qui contient la classification correspondante).
- Les machines à vecteurs supports (**svmtrain**).
- **-C** (coût)= 0.3.
- **-g** (gamma)=1.3.

Après avoir obtenu 5 modèles. Nous l'avons validé en utilisant :

- Les bases de test (pour chaque modèle on utilise une base contient les caractéristiques avec son base qui contient la classification correspondante).
- Les machines à vecteurs supports (**svmpredict**).
- Les modèles obtenus.

Après cette étape, nous obtenons une prédiction pour les bases de données du test. Et pour voir à quel point cette prédiction était vraie, nous avons calculé les performances du modèle : la précision, Rappel (Recall), et taux de reconnaissance (Accurecy) pour chaque modèle. Enfin on calcule le taux d'erreur de ces modèles.

4.4 Interface de l'application de la prédiction

Une Interface Homme-Machine (IHM) désigne la manière dont est présenté un logiciel à l'écran permettant à un utilisateur d'interagir avec une machine. Ces interfaces doivent être faciles à utiliser et compréhensibles par les utilisateurs pour garantir un bon degré de fiabilité lors des interactions ainsi qu'un temps d'apprentissage réduit.

Pour L'application que nous avons développée. On utilise le GUI (Graphical User Interface) qui a différents objets graphiques (boutons, menus…). Qui permettre

d'implémenter une interface utilisateur sous MATLAB. Nous présentons ici l'interface de notre système :

Figure 4.5:- L'interface principale du système prédiction d'interaction protéine-protéine.

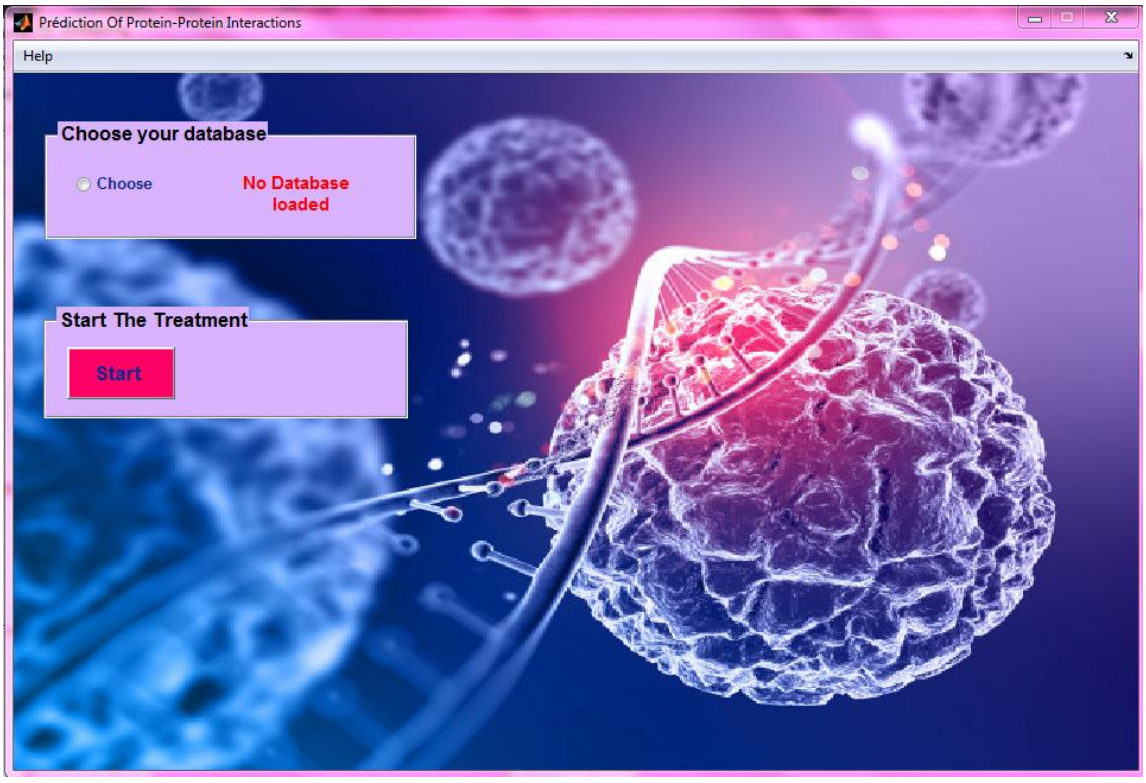


Figure4. 6:- L'interface de traitement pour le système de prédiction d'interaction protéine-protéine

Cette interface permet de :

- Sélectionner une base de données sur l'ordinateur de l'utilisateur.
- Traiter la base saisie, crée des modèles et faire la prédiction (Prédire la classe en utilisant les données de test).
- Afficher les résultats de la prédiction.
- Dans la barre de menu, il y a '**Help**' qui vous permet d'ouvrir une nouvelle fenêtre d'aide dans l'application.

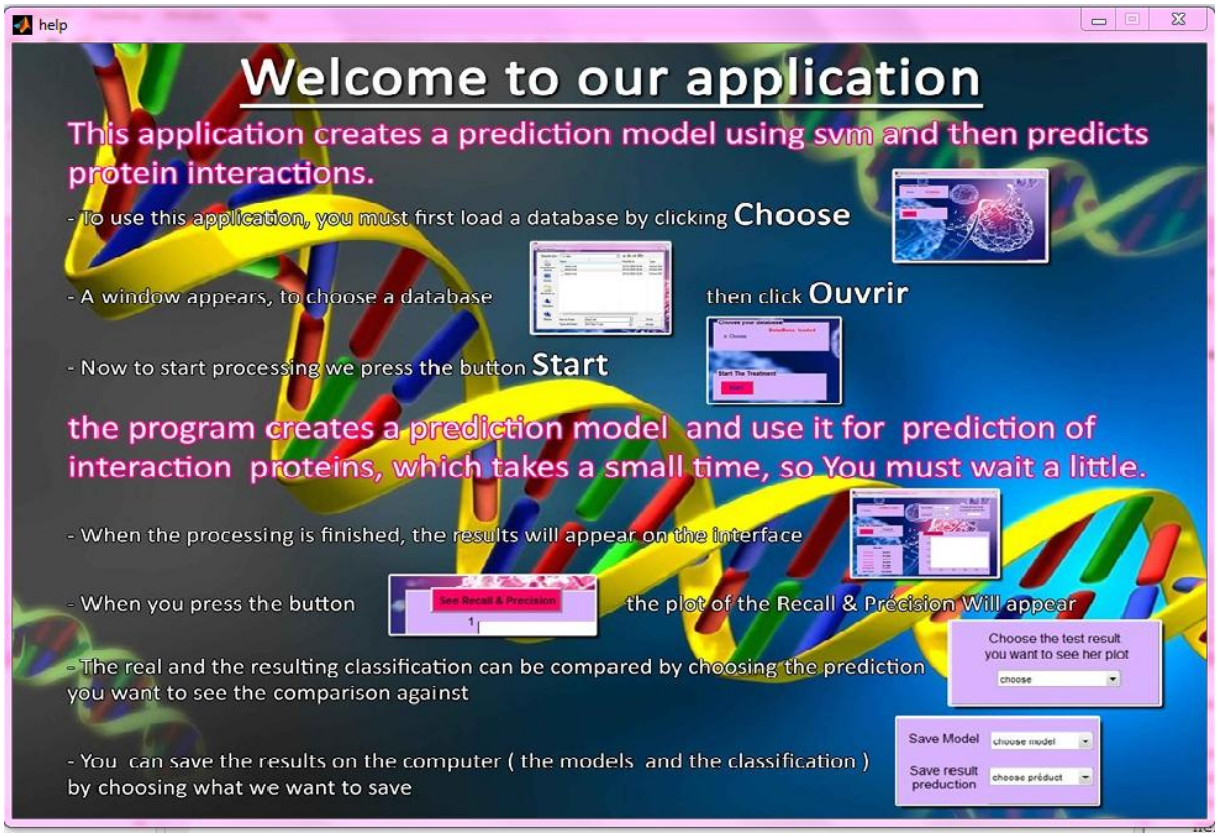


Figure 4.7:- L'interface Help.

Cette interface permet de :

- Expliquer les étapes et comment utiliser l'application.

4.5 Expérimentations et résultats

Pour tester notre méthode, nous avons utilisé cinq expérimentations (chacun pour tester un modèle). Les résultats obtenus sont présentés dans les matrices de confusion. La matrice de confusion montre comment les prédictions sont faites par le modèle. Les lignes correspondent à la classe connue des données et les colonnes correspondent aux prévisions faites par le modèle. Ainsi, les éléments diagonaux montrent le nombre de classifications correctes faites pour chaque classe, et les éléments hors-diagonaux montrent les erreurs commises. Et a chaque fois on calcule l'accurecy (indique la bonne prédiction et Aussi appelé taux de reconnaissance).

Cette matrice permet d'identifier 4 catégories de résultats :

TP (True Positives) : les cas ou la prédiction est positive, et ou la valeur réelle est effectivement positive.

TN (True Négatives) : les cas ou la prédiction est négative, et ou la valeur réelle est effectivement négative.

FP (False Positives) : les cas ou la prédiction est positive, et ou la valeur réelle est effectivement négative.

FN (False Négatives) : les cas ou la prédiction est négative, et ou la valeur réelle est effectivement positive.

4.5.1 Première expérimentation

Nous avons utilisé le modèle1 avec la base du test1 (1632 paires protéines). La matrice de confusion suivante en détaille les résultats :

	Classe prédite	
Classe Réelle	**Courriel**	**Pourriel**
Courriel	TP = 866	FN = 7
Pourriel	FP = 11	TN = 748

Table 4.1: Matrice de confusion de la première expérimentation sur les données de test.

Le modèle a pu atteindre un taux de reconnaissance (Accurecy) de 98.89 % sur les exemples de test.

4.5.2 Deuxième expérimentation

Nous avons utilisé le modèle2 avec la base du test2 (1632 paires protéines). La matrice de confusion suivante en détaille les résultats :

	Classe prédite	
Classe Réelle	**Courriel**	**Pourriel**
Courriel	TP = 837	FN = 26
Pourriel	FP = 19	TN = 750

Table 4.2: Matrice de confusion de la deuxième expérimentation sur les données de test.

Le modèle a pu atteindre un taux de reconnaissance (Accurecy) de 97.24 % sur les exemples de test.

4.5.3 Troisième expérimentation

Nous avons utilisé le modèle3 avec la base du test3 (1632 paires protéines). La matrice de confusion suivante en détaille les résultats :

	Classe prédite	
Classe Réelle	**Courriel**	**Pourriel**
Courriel	TP = 800	FN = 24
Pourriel	FP = 28	TN = 780

Table 4.3: Matrice de confusion de la troisième expérimentation sur les données de test.

Le modèle a pu atteindre un taux de reconnaissance (Accurecy) de 96.81 % sur les exemples de test.

4.5.4 Quatrième expérimentation

Nous avons utilisé le modèle4 avec la base du test4 (1632 paires protéines). La matrice de confusion suivante en détaille les résultats :

	Classe prédite	
Classe Réelle	**Courriel**	**Pourriel**
Courriel	TP = 851	FN = 14
Pourriel	FP = 14	TN = 753

Table 4.4: Matrice de confusion de la quatrième expérimentation sur les données de test.

Le modèle a pu atteindre un taux de reconnaissance (Accurecy) de 98.28 % sur les exemples de test.

4.5.5 Cinquième expérimentation

Nous avons utilisé le modèle5 avec la base du test5 (1632 paires protéines). La matrice de confusion suivante en détaille les résultats :

	Classe prédite	
Classe Réelle	**Courriel**	**Pourriel**
Courriel	TP = 824	FN = 13
Pourriel	FP = 23	TN = 773

Table 4.5: Matrice de confusion de la cinquième expérimentation sur les données de test.

Le modèle a pu atteindre un taux de reconnaissance (Accurecy) de 97.79 % sur les exemples de test.

Nous avons également calculé la précision qui représente le pourcentage du nombre d'exemple correctement classé dans les exemples de test tandis. Et le rappel (Recall) représente le pourcentage des exemples positifs correctement classé par rapport au nombre total des exemples positifs. Ce dernier permet de faire le compromis entre les erreurs positives (une paire de protéines positive sain classé négative) et les erreurs négatives (une paire de protéines négative classé positive).

La figure suivante représente l'évolution de la précision et le rappel avec notre méthode :

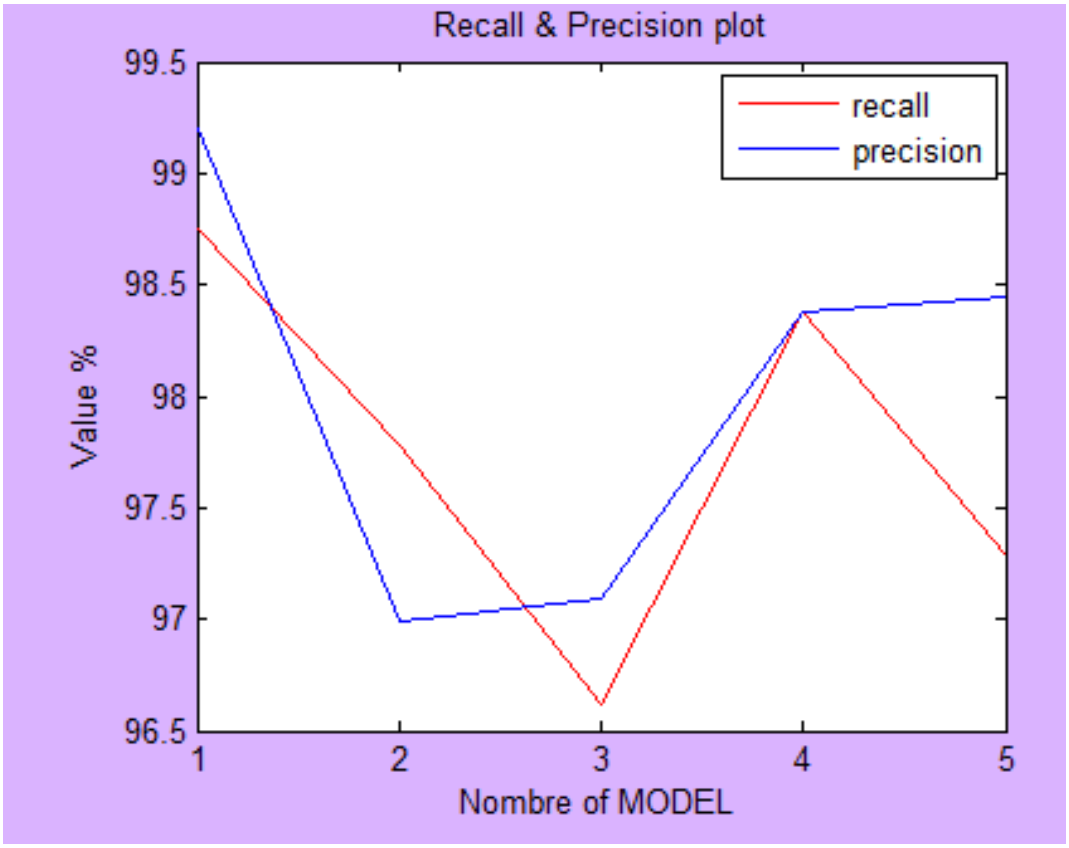


Figure 4.8:- Précision et Rappel de notre méthode.

4.5.6 Après l'exécution

Après l'exécution du programme, et à la fin du processus de traitement, nous obtenons les résultats suivants qui apparaissent sur l'interface de l'application :

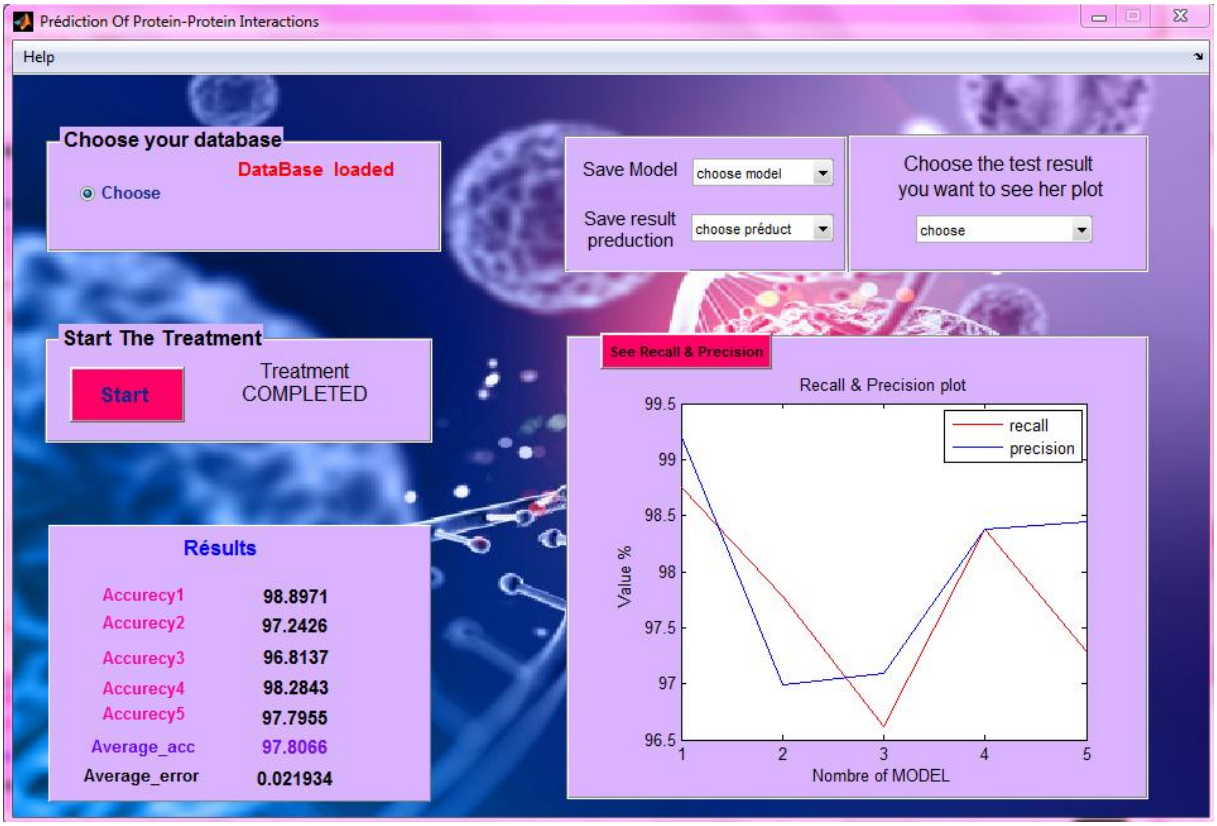


Figure 4.9 :- L'interface de l'application après la fin du processus de traitement et l'apparition des résultats.

Cette interface permet de :

- Afficher le taux de reconnaissance (Accurecy) de chaque modèle.
- Afficher le moyen du taux de reconnaissances (**Average_acc**).
- Afficher le taux d'erreur (**Average_error**).
- Afficher la courbe graphique de Rappel (Recall) et Précision calculées.
- Afficher un graphique qui montre la différence entre la classification réelle et la classification prédictive des bases de test. (un exemple dans la figure 10).
- Vous pouvez également enregistrer les modèles crées et les bases de test avec leur classification prédictive sur l'ordinateur utilisé.

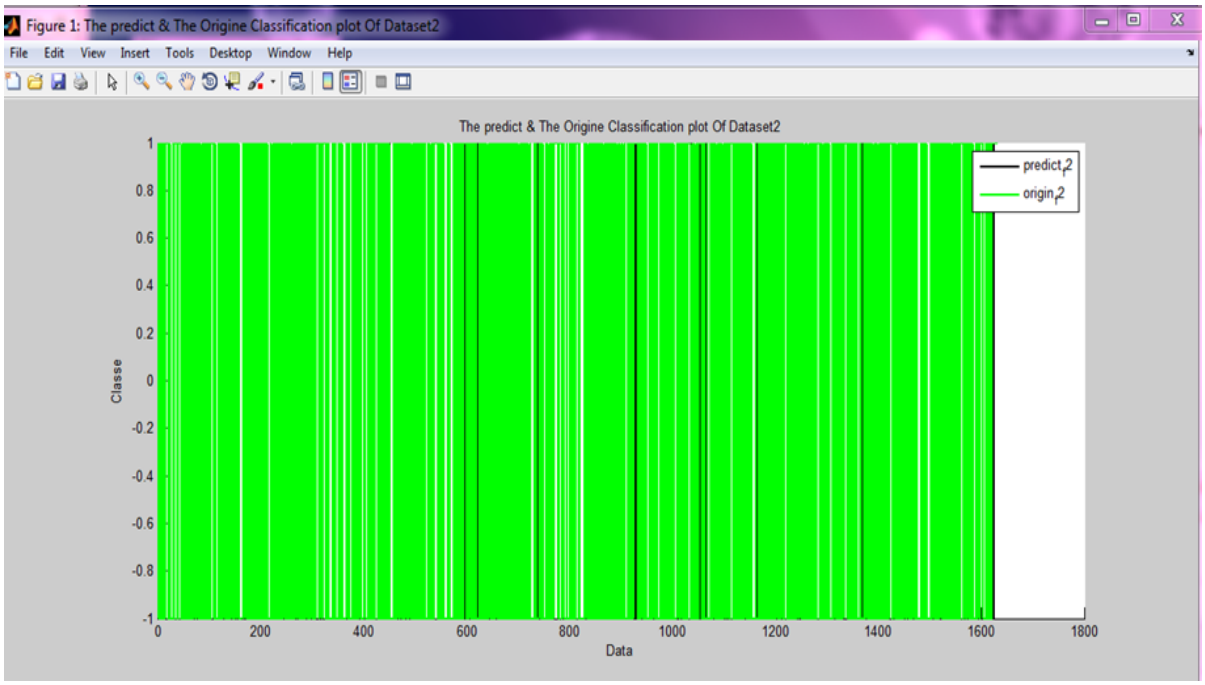


Figure 4.10 :- la classification réelle et la classification prédictive de base de test2.

Pour chaque graphique concerné à leur base du test, la classification de prévision n'apparaîtra que lorsqu'elle est différente de la classification réelle. Autrement dit, il apparaîtra en vert lorsque la classification réelle (l'origine) correspond à la classification prédictive (prédict), et il apparaîtra noir lorsqu'il est différent. (Exemple : la classification du base de test2 dans la figure 10).

4.6 Discussion des résultats

Les résultats obtenus dans les différentes expérimentations où notre méthode donne une moyenne de taux de reconnaissance (accurecy) de 97.80 %, précision 98.02% et du Rappel 97.76%.et donne aussi un taux d'erreur de 0.02. Cela montre l'efficacité de notre proposition et la possibilité de l'utiliser dans ce domaine.

A travers toutes nos expérimentations et résultats, nous pouvons dire que notre système proposé est capable de prédire des interactions protéines correctement de 97 %.

4.7 Conclusion

Dans ce dernier chapitre, nous avons représenté l'implémentation de notre système proposé : l'environnement, le langage de programmation et les outils de développement. Ensuite nous avons présenté quelques expérimentations effectuées et présenté les résultats obtenu et discuté.

Conclusion générale

Il est maintenant bien connu que les protéines interagissent spécifiquement les unes avec les autres pour remplir leurs fonctions, donc la fonction peut être étudiée par l'analyse des interactions protéiques.

La biologie offre des techniques expérimentales pour analyser ces interactions à l'échelle d'une cellule, d'un tissu ou d'un organisme. La compréhension de l'ensemble de ces interactions est un enjeu majeur de la compréhension de la fonction des protéines et du fonctionnement cellulaire.

Malgré le développement des techniques expérimentales, elles restent coûteuses et prennent du temps. Pour résoudre ce problème, un certain nombre de techniques de calcul ont été proposées basés sur les informations des protéines pour prédire leurs interactions. Et elle a rencontré un grand succès dans son domaine.

Dans ce travail, nous avons étudié ces interactions protéiques. Nous avons proposé une méthode basée sur l'apprentissage automatique et plus particulièrement sur la méthode SVM.

Dans cette méthode, une protéine est représentée par une chaine des acides amminées, qui est converti en une forme utilisable. Les paires protéine-protéine doivent être extraites pour être utilisées dans le traitement.

Un ensemble de paires obtenues est sélectionnées et leurs caractéristiques sont extraites, enregistrées dans des bases de données et utilisées pour construire des modèles de prédiction permettant de reconnaitre les interactions protéine-protéine pendant la phase de test.

Pour valider notre proposition, nous avons préparés des bases du test contient des paires de protéines. Les accurecy (taux de reconnaissance) obtenus sur ces données dépassent les 97% ce qui est encourageant et démonte l'efficacité de la méthode proposé.

Un tel système offre des avantages certains, particulièrement :

- On na pas besoin de plusieurs informations sur les protéines à cause de la suffisance en chaînes protéiques pour prédire l'interaction protéine-protéine.

- La possibilité de prédire un grand nombre d'interactions protéine-protéine en peu de temps.
- Réduction des coûts de la prédiction des interactions protéine-protéine.

Nous avons rencontré des difficultés, parmi lesquelles :

- La difficulté de trouver une base de données adaptée à notre étude.
- Puisque notre méthode repose uniquement sur des chaînes d'acides aminés, il est nécessaire de trouver une forme d'expression appropriée pour les acides aminés à utiliser dans les calculs et les traitements, dans laquelle nous avons rencontré des difficultés.

Pour les travaux futur, nous suggérons quelques idées qui peuvent améliorer notre système telles que :

- Appliquer le système à d'autres bases de données afin de construire des bases d'entrainement pour la création d'autres modèles.
- Renforcer le système de prédiction par d'autres méthodes de machine Learning pour améliorer sa précision.
- Ajouter une étape pour étudier la fonction du résultat de l'interaction protéine-protéine.

Bibliographie

[1] Gautamb.singh « fundaments of bioinformatics and computational biology methods and exercises in matlab» . Département of Computer Science and Engineering USA.

[2] Daniel Gautheret ESIL « initiation à la bioinformatique », université de méditerranée. 2004 .

[3] Jerome waldispuhl « modélisation et prédiction de la structure des protéines transmembranaires » Mémoire de doctorat. FRANCE 2004.

[4] Dr L.Djerou «Calcul à l'ADN et Bio-informatique (M2-IA) » cour de master2 université de Biskra.

[5] Benoît Robisson «Méthodes de classification de réseaux d'interactions protéine-protéine et évaluations pour l'étude de la fonction. » Université d'Aix-Marseille. École Doctorale des sciences de la vie et de la santé 2013.

[6] Magali Michaut. «Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds». Sciences du Vivant [q-bio]. Université Paris Sud - Paris XI, 2008.

[7] A. J. M. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal, «Protein interaction mapping in c. elegans using proteins involved in vulval development» Science, vol. 287, pp. 116–122, July 2000.

[8] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M.Vidal, «Identification of potential interaction networks using sequence-based searches for conserved proteinprotein interactions or "Interologs", » Genome Research, vol. 11, pp. 2120–2126, Jan. 2001. PMID :11731503.

[9] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, «Assigning protein functions by comparative genome analysis : Protein phylogenetic profiles, » Proceedings of the National Academy of Sciences, vol. 96, pp. 4285–4288, Apr. 1999. PMID : 10200254.

[10] Céline Brouard. « Inférence de réseaux d'interaction protéine-protéine par apprentissage statistique. Application au réseau d'interaction autour de la protéine CFTR, impliquée dans la mucoviscidose » l'Université d'Évry Val d'Essonne. Ecole doctorale des génomes auxorganismes.le 14 février 2013.

[11] PHILIPPE BESSE « Apprentissage Statistique » cours. Département Génie Mathématique et Modélisation Institut National des Sciences Appliquées de Toulouse— 31077 – Toulouse cedex 4.

[12] Mathwork.com .

[13] Y. Kodratoff Cornuejols L. Miclet. «Apprentissage artificiel ». Eyrolles.

[14] Fomani Boris Mohamadally Hasan. «SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges ». Versailles St Quentin, France, 2006.

[15] Steve R. Gunn. «Support Vector Machines for Classication and Regression». Faculty of Enginnering, Science and Mathematics School of Electronics and Computer Science, May 1998.

[16] Marref Nadia. «Apprentissage Incrémental &Machines à Vecteurs Supports». Thèse de doct. 2013.

[17] Yiming Ying Colin Campbell. « Learning with support vector machines synthesis lectures on articial intelligence and machine learning. » Morgan and Claypool publishers, 2011.

[18] Alexander J. Smola Bernhard Scholkopf. « Learning with kernels, support vector machines, regularization, optimization, and beyond. » the MIT Press, 2002.

[19] Abdelhamid DJEFFAL « Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données ».thèse de doctorat .université du biskra .2011.

[20] Mohamadally Hasan et Fomani Boris « SVM : Machines à Vecteurs de Support». Université du Versailles St Quentin. France .2006.

[21] Yann Guermeur « SVM Multiclasses, Théorie et Applications » thèse de doctorat. Ecole doctorale IAEM Lorraine. Université Nancy I.2007.

[22] Bernhard Scholkopf, Alexander J. Smola « Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond », the MIT Press 2002.

[23] S. Knerr, L. Personnaz, and G. Dreyfus. « Single-layer learning revisited : A stepwise procedure for building and training a neural network. In F. Fogelman-Soulié and J. Hérault, editors, Neurocomputing : Algorithms, Architectures and Applications», volume F68 of NATO ASI Series, pages 4150. Springer-Verlag, 1990.

[24] I. M. A. Nooren and J. M. Thornton, «Diversity of protein–protein interactions» TheEMBO Journal,vol. 22, pp. 3486–3492, July 2003.

[25] E. Sprinzak and H. Margalit,«Correlated sequence-signatures as markers of protein-protein interaction, » Journal of Molecular Biology, vol. 311, pp. 681–692, Aug. 2001.

[26] J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb, and P. Uetz, «MPIDB : the microbial protein interaction database, » Bioinformatics, vol. 24, pp. 1743–1744, June 2008.

[27] O. Souiai, E. Becker, C. Prieto, A. Benkahla, J. De Las Rivas, and C. Brun, «Functional integrative levels in the human interactome recapitulate organ organization, » PLoS ONE, vol. 6, p. e22051, July 2011.

[28] Madam Djarmouni Meriem. COURS DE BIOCHIMIE STRUCTURALE . Universiré FERHAT ABBAS –SETIF 1. faculté des science de la nature et de la vie.2016.

[29] Vincent Derrien. « Heuristiques pour la résolution du problème d'alignement multiple. PhD thesis», Université d'Angers, 2008.

[30] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. « Molecular biology of the cell ».

[31] Chih-Chung Chang and Chih-Jen Lin « A Library for Support Vector Machines» Department of Computer Science National Taiwan University, Taipei, Taiwan.

[32] Imane el hassani. «Svr avec boosting pour la prévision à long terme. » Thèse de doctorat. Ecole polytechnique de l'université de tours, 2011-2012.

[33] S Henikoff et J Henikoff, *«* Amino acide substitution matrices from protein blocks *»,* Proceedings of the National Academy of Sciences of the United States of America, vol. 89, n^{o} 22, 1992, p. 10915–9.