REPUBLIQUE DE COTE D'IVOIRE
Union - Discipline – Travail

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Institut National Polytechnique
Félix HOUPHOUËT-BOIGNY de Yamoussoukro

École Doctorale Polytechnique

N° d'ordre :

THÈSE UNIQUE

Pour l'obtention du grade de

Docteur de l'Institut National Polytechnique Félix Houphouët Boigny de Yamoussoukro

Mention : Sciences et Technologies

Spécialité : Informatique

SUJET :

OUTILS ET TECHNIQUES INFORMATIQUES POUR LA PRÉDICTION D'INTERACTION PROTÉINE-PROTÉINE À PARTIR DES INFORMATIONS DE LA SÉQUENCE

Présentée et soutenue publiquement le 24/03/2022 par :

KOPOIN N'Diffon Charlemagne

JURY

M. OUMTANAGA Souleymane	Professeur Titulaire	Institut National Polytechnique Félix Houphouët Boigny, Côte d'Ivoire	Président
M. GUEYE Amadou Dahirou	Maître de Conférences	Université Alioune Diop de Bambey, Sénégal	Rapporteur
M. DOSSO Mouhamadou	Maître de Conférences	Université Félix Houphouët-Boigny, Côte d'Ivoire	Rapporteur
M. KONAN Kouadio Fransisco	Maître de Conférences	Ecole Normale Supérieure, Côte d'Ivoire	Examinateur
M. BABRI Michel	Professeur Titulaire	Institut National Polytechnique Félix Houphouët Boigny, Côte d'Ivoire	Directeur de Thèse

Dédicace

A ma très chère famille.

Remerciements

Cette thèse est le fruit de longues et périlleuses épreuves qui, sans le soutien inconditionnel et l'aide de certaines personnes, aurait pu ne pas aboutir. Nous tenons donc à remercier chacune de ces personnes.

Nos remerciements très chaleureux, sont adressés en premier lieu, à Monsieur BABRI Michel, Professeur à l'INP-HB, qui a accepté de diriger cette thèse mais également de nous transmettre des valeurs telles que : la rigueur dans le travail, le dépassement de soi, l'amour du prochain, la sérénité dans le travail. Ses conseils et ses orientations ont à chaque fois été un déclic pour l'aboutissement de cette thèse. Il a toujours été disponible et a partagé nos peines et nos souffrances.

Nos sincères remerciements vont à l'endroit de Monsieur YAO Kouassi Benjamin, Professeur à l'INP-HB et Directeur de l'Ecole Doctorale Polytechnique de l'INP-HB, ainsi qu'à toute son équipe de direction : Monsieur SORO Doudjo, Maître de Conférences à l'INP-HB, Docteur ABRO Koutouan Désiré, Attaché de Recherche à l'INP-HB, pour leurs ouverture et disponibilité à toutes les questions relatives à l'EDP.

Nous tenons à remercier Monsieur ZOUEU Thouakesseh Jérémie, Professeur à l'INP-HB et Directeur de l'UMRI 78 pour ses conseils et ses encouragements.

Nous remercions particulièrement Monsieur GUEYE Amadou Dahirou, Maître de Conférences à l'Université Alioune Diop de Bambey du Sénégal et Monsieur DOSSO Mouhamadou, Maître de Conférences à l'Université Félix Houphouët-Boigny de Côte d'Ivoire, qui malgré leurs charges de travail importantes, ont accepté de rapporter cette thèse.

Nos vifs remerciements vont à l'endroit de Monsieur KONAN Kouadio Fransisco, Maître de Conférences à l'Ecole Normale Supérieure de Côte d'Ivoire pour avoir accepté la lourde tâche d'en être l'examinateur.

Merci à Monsieur OUMTANAGA Souleymane, Professeur à l'INP-HB qui nous a accueilli sous son aile depuis notre dépôt de dossier à l'EDP de INPHB jusqu'à l'accomplissement de cette thèse. Il a su nous rappeler à l'ordre lorsque nous nous égarions et tombions dans la passivité. Merci Professeur, pour avoir mis en place le prestigieux Laboratoire de Recherche en Informatique et Télécommunication (LARIT) et aussi le Laboratoire de Mécanique et Informatique (LAMI).

Nous remercions spécialement Monsieur AKA Boko, Professeur à l'Université Nangui Abrogoua qui a été à la base de notre aboutissement. C'est lui qui a rendu possible notre inscription au MASTER et nous a conduit au LARIT.

Nous tenons à remercier Dr. N'TAKPE Tchimou, Dr. ATIAMPO Kodjo Armand, Dr. N'GUESSAN Gerard et Dr. SAHA Bernard, avec qui nous avons eu à collaborer durant cette thèse.

Merci à tous les membres du LARIT : Docteurs, doctorants et stagiaires de MASTER en particulier pour leur soutien indéfectible. Nous encourageons tous les doctorants à s'armer de courage pour atteindre leur objectif.

Nous adressons nos sincères remerciements aux membres de l'équipe MIABD du LARIT en particulier nos devanciers : Dr. COULIBALY Tiekoura, Dr. KONE Malick, Dr. MORIE Maho, Dr. KAYE-BI Bertin. Nous remercions également les devanciers des autres équipes du LARIT, Dr. N'TAKPE Tchimou, Dr. ANOH Georges, Dr. PANDRY Koffi Ghislain, Dr. MAMBE Digrais Moïse, Dr. OUATTARA Nouho, Dr. GNIMANSOU Edgar et Dr. ATTA Amanvon Ferdinand pour leur promptitude et leur accompagnement tout au long de ces années d'épreuves.

Nos sincères remerciements vont également à l'endroit des membres de notre famille, en particulier mon père N'DIFON Kopoin Jean et mon frère KOPOIN N'Dépo Johnson pour leur soutien sans retenue durant les études, aussi nous remercions tous nos amis pour leurs prières et leurs apports de tout genre.

Merci également à Mademoiselle SOTONGUI Richemonde et Monsieur YAO Kouadio Anselme, qui sans leur soutien, je n'aurais pas eu la force et l'équilibre suffisant pour terminer la rédaction de cette thèse.

Liste des tableaux

Liste des figures

Notations

IPP^+	IPP positive
IPP^-	IPP négative
Y2H	Levure à double hybride
L	La longueur de la séquence
R_i	Résidu d'acide aminé *i*
5.5	5,5

Abréviations

AAC	Amino Acid Composition
ADN	Acide DésoxyNucléique
ANN	Artificial Neural Network
APAAC	Amphiphilic Pseudo Amino Acid Composition
ARN	Acide RiboNucléique
BP	Bigram Physicochemical
FN	Faux Négatif
FP	Faux Positif
I.I.D	Indépendant et Identiquement Distribué
IPP	Interaction Protéine-Protéine
LARIT	LAboratoire de Recherche en Informatique et Télécommunication
ML	Machine Learning
PseAAC	Pseudo Amino Acid Composition
RG	Recherche sur Grille
RNA	Réseau de Neurones Artificiels
SVM	Support Vector Machine
SVOH	Sélection des Valeurs Optimales d'Hyperparamètres
TAS	Théorie de l'Apprentissage Statistique
VCK	Validation Croisée K-fois
VN	Vrai Négatif
VP	Vrai Positif

Résumé

La protéine est un composant essentiel de la cellule biologique. Les différentes interactions chimiques entre les protéines appelées interactions protéine-protéine (IPP) sont liées à certaines maladies ainsi qu'à certaines thérapies. De ce fait, leur identification a des implications importantes pour plusieurs processus parmi lesquels la prévention des maladies, l'annotation fonctionnelle et la conception de médicaments. De nombreuses IPP ont été détectées par des expériences biologiques au cours des dernières décennies, mais beaucoup d'entre elles restent encore non découvertes. En outre, ces expériences biologiques sont limitées en raison des contraintes de temps et de coûts. Par conséquent, le développement d'outils informatiques est vivement recommandé. Cela pourrait permettre d'accélérer la découverte de médicaments en réduisant les expériences biologiques lentes et coûteuses à l'aide de simulations informatiques plus rapides et moins coûteuses, et d'annoter la fonction des protéines à partir des séquences de protéines. Compte tenu de nombreuses IPP détectées expérimentalement et dont les informations sont disponibles dans des banques de données sur les protéines, plusieurs outils d'inférence et d'apprentissage automatique (*machine learning*) ont été proposés pour prédire les IPP. La conception de tels outils passe par deux étapes : l'extraction de caractéristiques (descripteurs) à partir des informations de la séquence et la classification des interactions à l'aide d'un algorithme d'apprentissage supervisé. Cependant, les caractéristiques extraites par les techniques existantes ne permettent pas aux algorithmes d'apprentissage supervisé d'être efficaces et d'obtenir des résultats idéaux. Ainsi, notre objectif est d'améliorer les performances prédictives par la conception de nouvelles techniques complémentaires permettant aux algorithmes d'apprentissage supervisé d'inférer correctement les interactions à partir des données de séquences de protéines et d'obtenir des résultats idéaux.

Dans cette thèse, nous avons premièrement proposé une technique d'extraction de caractéristiques notée BP (*Bigram-Physicochemical*). Cette technique permet d'extraire des caractéristiques bigrammes à partir des séquences de protéines pour un apprentissage automatique efficace. Un bigramme est un ensemble de deux lettres (ici les acides aminés) successives dans un document texte (ici la séquence de protéine). Pour une protéine donnée, BP calcule d'abord une matrice à partir d'informations de propriétés physicochimiques de la protéine. Cette matrice peut être obtenue soit à partir d'une distance (approche BP1) ou soit à

partir d'une fonction (approche BP). Ensuite, BP extrait des caractéristiques bigrammes en se servant de la matrice calculée. La technique BP ne produit pas de vecteurs strictement parcimonieux et ne dépend pas d'une base de données comme certaines techniques d'extraction de caractéristiques bigrammes existantes. Deuxièmement nous avons proposé une nouvelle approche de sélection de valeurs optimales d'hyperparamètres d'un modèle d'apprentissage automatique notée SVOH. Les hyperparamètres sont les paramètres influents du modèle d'apprentissage automatique. Généralement, la technique de la recherche sur gille (*Grid search*) est combinée à la technique de validation croisée *k*-fois pour la recherche des valeurs optimales d'hyperparamètres. Contrairement à la littérature qui fixe la valeur du nombre *k*, correspondant en fait au nombre de sous-ensembles de l'échantillon, à 5 ou 10 sur des bases a priori, SVOH fait un apprentissage du nombre *k* afin de déterminer une valeur optimale du nombre de sous-ensembles. L'approche développée permet ainsi de rechercher rigoureusement sur un nombre *k* ajusté de sous-ensembles de valeurs optimales d'hyperparamètres.

Les techniques décrites au sein de la thèse, combinées à une méthode des machines à vecteurs de supports (SVM) et formant ainsi les outils SVM-BP et SVM-BP1, sont testées et validées sur trois différents ensembles de données IPP réelles : les IPP humaines HPRD, les IPP de la levure Saccharomyces cerevisiae et les IPP de la bactérie Helicobacter pylori. Les résultats obtenus après certaines expériences comparatives ont montré que ces outils et particulièrement l'outil SVM-BP, ont obtenu des performances supérieures sur les trois différents ensembles de données IPP dans les métriques justesse, précision et sensibilité. Nous pouvons dire que les outils SVM-BP et SVM-BP1 améliorent bien les performances prédictives des IPP et constituent ainsi une véritable aide pour les biologistes dans l'identification des interactions protéine-protéine et la recherche médicamenteuse.

Mots clés : Apprentissage automatique, SVM, extraction de caractéristiques, sélection de modèles, prédiction d'interaction protéine-protéine.

Abstract

Computer Tools and Techniques for Protein-Protein Interaction Prediction from Sequence Information

Protein is an essential component of the biological cell. The different chemical interactions between proteins called protein-protein interactions (PPIs) are linked to certain diseases as well as to certain therapies. Therefore, their identification has important implications for several processes including disease prevention, functional annotation, and drug design. Many PPIs have been detected by biological experiments over the past decades, but many remain undiscovered. In addition, these biological experiments are limited due to time and cost constraints. Therefore, the development of computational tools is highly recommended. This could accelerate drug discovery by reducing slow and expensive biological experiments with faster and cheaper computer simulations and annotate protein function from protein sequences. Given many experimentally detected PPIs with information available in protein databases, several inference and machine learning tools have been proposed to predict PPIs. The design of such tools involves two steps: feature extraction (descriptors) from sequence information and classification of interactions using a supervised learning algorithm. However, the features extracted by existing techniques do not allow supervised learning algorithms to be efficient and to obtain ideal results. Thus, our goal is to improve predictive performance by designing new techniques that allow supervised learning algorithms to correctly infer interactions from protein sequence data and obtain ideal results.

In this thesis, we first proposed a feature extraction technique noted BP (Bigram-Physicochemical). This technique allows to extract bigram features from protein sequences for efficient machine learning. A bigram is a set of two successive letters (here the amino acids) in a text document (here the protein sequence). For a given protein, BP first computes a matrix from information about the physicochemical properties of the protein. This matrix can be obtained either from a distance (BP1 approach) or from a function (BP approach). Then BP extracts bigram features using the calculated matrix. The BP technique does not produce

strictly parsimonious vectors and does not depend on a database like some existing bigram feature extraction techniques. Secondly, we proposed a new approach for selecting optimal values of hyperparameters of a machine learning model noted SVOH. Hyperparameters are the influential parameters of the machine learning model. Generally, the grid search technique is combined with the k-fold cross-validation technique to find the optimal values of hyperparameters. Contrary to the literature which fixes the value of the number k, which corresponds to the number of subsets of the sample, to 5 or 10 on a priori bases, SVOH does a learning of the number k to determine an optimal value of the number of subsets. The developed approach thus allows to rigorously search for the optimal values of hyperparameters on an adjusted number k of subsets.

The techniques described in the thesis, combined with a support vector machine (SVM) method, and thus forming the SVM-BP and SVM-BP1 tools, are tested and validated on three different real PPI datasets: human HPRD PPI, yeast Saccharomyces cerevisiae PPI and Helicobacter pylori PPI. The results obtained after some comparative experiments showed that these tools and particularly the SVM-BP tool, obtained superior performances on the three different PPI datasets in the accuracy, precision, and sensitivity metrics. We can say that the SVM-BP and SVM-BP1 tools improve well the predictive performance of PPI and thus are a real help for biologists in protein-protein interaction identification and drug discovery.

Keywords: Machine learning, SVM, features extraction, model selection, protein-protein interaction prediction.

Table des matières

INTRODUCTION GENERALE

Contexte et motivation

Les protéines réalisent la plupart des fonctions biologiques dans une cellule et en général elles les réalisent rarement seules. Leurs interactions avec d'autres protéines, appelées interaction protéine-protéine (IPP), sont responsables de certaines pathologies, ainsi que les éventuelles thérapies. Par conséquent, l'identification des interactions protéine-protéine aide à plusieurs niveaux parmi lesquels l'annotation fonctionnelle (documentation sur la fonction), l'exploration de la parthénogenèse des maladies, la détection des cibles thérapeutiques. Une cible thérapeutique (ou cible médicamenteuse ou cible biologique) est en fait une protéine sur laquelle un médicament ou ligand peut se fixer et modifier sa fonction [Bakail et Ochsenbein 2016]. Certaines techniques expérimentales *in vivo* ou *in vitro* telles que la co-immunoprécipitation et la technique hybride à deux levures ont permis d'identifier de nombreuses interactions. Cependant, elles sont longues, coûteuses et ne peuvent détecter qu'une petite fraction de l'ensemble du réseau d'interaction protéine-protéine. Pour cette raison, la question de la prédiction d'interactions inconnues est désormais considérée comme difficile à résoudre uniquement par des méthodes expérimentales [Von Mering et al. 2002 ; Shoemaker and Panchenko 2007].

En plus des techniques expérimentales de détection des interactions entre les protéines, des outils informatiques sont donc développés, particulièrement ceux utilisant les informations de la séquence de protéine [Valente et al. 2013 ; You et al. 2017]. Ces outils, peuvent prédire des interactions qui pourront ensuite être vérifiées expérimentalement. Ils peuvent également s'intégrer dans le processus de conception d'un médicament en guidant le choix des composés ayant les meilleurs potentiels thérapeutiques et permettre d'annoter les fonctions de protéines [Nguyen et al. 2011].

Compte tenu des nombreuses informations sur les interactions protéine-protéine détectées expérimentalement et disponibles dans des banques de données sur les protéines [Patil 2019], la prédiction d'interactions entre les protéines peut être vue comme un problème d'inférence et d'apprentissage automatique dans un cadre supervisé (apprentissage supervisé) [Sathya et Abraham 2013]. Le développement d'outils informatiques pour la prédiction d'interactions protéine-protéine basés sur les séquences comportent généralement deux étapes. Nous avons l'étape d'extraction de caractéristiques et celle de la classification des échantillons à l'aide d'un algorithme d'apprentissage supervisé. L'extraction de caractéristiques à partir des informations de séquences de protéines est une tâche d'apprentissage de la représentation des données [Bengio et al. 2013]. Elle vise à extraire les attributs les plus représentatifs des

échantillons et à normaliser des séquences de protéines de longueurs différentes en vecteurs de même taille. Il faut noter qu'une séquence de protéines est une longue chaîne d'acides aminés symbolisés par des lettres de l'alphabet [Grantham 1974].

Une des problématiques les plus courantes dans la conception d'outils informatiques de prédiction d'interaction protéine-protéine basés sur la séquence concerne l'étape d'extraction de caractéristiques. En effet, le succès des algorithmes d'apprentissage automatique dépend généralement de la représentation des données, et nous émettons l'hypothèse que cela est dû au fait que différentes représentations peuvent enchevêtrer et masquer plus ou moins les différents facteurs explicatifs de variation derrière les données. Dans le cas de la prédiction d'interaction, les caractéristiques extraites devront refléter les corrélations intrinsèques à l'interaction. Cela voudrait encore dire que la technique d'extraction doit représenter les informations essentielles de la paire de protéines.

Une autre difficulté apparente est l'entrainement de l'algorithme d'apprentissage supervisé. Nous faisons face en fait à un problème de classification binaire où d'un côté nous avons la classe des protéines en interaction et de l'autre côté la classe des protéines sans interaction. L'algorithme d'apprentissage supervisé utilisé ou développé doit pouvoir apprendre une fonction capable de prédire correctement la classe associée à une nouvelle observation [Islam et al. 2018]. Cette fonction de prédiction est appelée un classifieur [Lemberger et al. 2015]. Pour réaliser un tel classifieur, une sélection rigoureuse des paramètres influents de cet algorithme, qui sont ses hyperparamètres, s'impose. Cette étape correspond au problème de sélection de modèles [Arlot et Celisse 2010].

Bien que plusieurs travaux ont été faits sur des techniques d'extraction de caractéristiques à partir des informations de la séquence, la précision de l'apprentissage demeure encore limitée [Guo *et al*. 2008 ; Huang *et al*. 2016 ; Göktepe et Kodaz 2018]. Les techniques existantes ne permettent pas au classifieur d'inférer correctement les interactions et de faire moins d'erreur pour classer une nouvelle observation. Le problème que nous voulons résoudre ici est comment construire un outil qui améliore les performances de la prédiction des interactions protéine-protéine ? Autrement dit, quelle technique algorithmique développer pour extraire des caractéristiques de la séquence de protéine permettant un apprentissage automatique efficace ? quelles informations de la séquence représenter ? Quelles techniques adopter pour la sélection rigoureuse des valeurs optimales d'hyperparamètres ?

Objectif et contributions

L'objectif de cette thèse est de proposer des outils et techniques informatiques pour améliorer la prédiction d'interaction protéine-protéine à partir des données de séquences de protéines et aider ainsi à la conception médicamenteuse. Cela passe par :

(1) Développer des techniques d'extraction de caractéristiques qui représentent les informations essentielles des paires de protéines ;

(2) Combiner les techniques d'extraction développées avec un algorithme d'apprentissage supervisé et construire des classifieurs ;

(3) Evaluer les classifieurs construits sur des ensembles de données IPP réelles ;

(4) Développer des techniques pour sélectionner rigoureusement les valeurs optimales d'hyperparamètres des classifieurs.

Les contributions de cette thèse portent sur la proposition d'outils de prédiction des interactions entre les protéines et sont décrites ci-dessous.

Nous avons tout d'abord développé une technique d'extraction de caractéristiques notée BP (*Bigram Physochemical*). Cette technique permet d'extraire les caractéristiques bigrammes contenues dans une séquence et les représente sous forme d'un vecteur de composantes bigrammes. Les bigrammes sont en effet les fréquences de deux lettres (acides aminés) successives de la séquence. Ils représentent des propriétés chimiques importantes liées à l'interaction de la protéine parmi lesquelles la reconnaissance des 'plis' des protéines ou du repliement des protéines qui fait l'objet de plusieurs études bio-informatiques [Yang et al. 2011 ; Tsubaki et al. 2017]. Il faut noter qu'à ce jour il existe deux techniques d'extraction des caractéristiques bigrammes. La première notée PF (*Pairwize Frequency*) proposé par [Ghanty and Pal 2009] applique la technique 2-gramme [Cavnar and Trenkle 1994], qui est une technique de *text mining* [Islam et al. 2018], directement sur la séquence primaire de la protéine. Or, pour une protéine donnée, nous n'avons pas toujours toutes les combinaisons de deux lettres successives. Par conséquent, le vecteur constitué comporte plusieurs composantes nulles et est qualifié de vecteur strictement parcimonieux [Sbai 2012]. Un tel vecteur rend la méthode d'apprentissage automatique moins performante et sa capacité à ne pas faire d'erreur sur de nouvelles observations, sa capacité de généralisation, reste faible. La deuxième technique est celle des probabilités bigrammes proposé par [Sharma *et al*. 2013]. Cette technique passe d'abord par une représentation matricielle de la séquence de protéine, en l'occurrence la matrice PSSM (*Position Specific Score Matrix*) [Dehzangi et al. 2017], et

applique ensuite la technique 2-gramme sur la matrice obtenue. Tous les éléments de la PSSM étant renseignés, cette deuxième technique permet d'éviter les composantes nulles dans le vecteur résultant. Cependant, la démarche d'obtention de la PSSM est longue car il faut comparer chaque séquence de l'échantillon à une gigantesque base de données de séquences à travers l'outil PSI-BLAST [McGinnis and Madden 2004]. De plus la significativité des valeurs PSSM, c'est-à-dire le fait que les valeurs PSSM ne soient pas obtenues par hasard, dépend d'une base de données qu'on aura choisie. Par conséquent la performance du classifieur dépendra également de cette base de données. L'approche de la technique BP développée est comparable à l'approche de la technique des probabilités bigrammes. BP calcule dans un premier temps de façon heuristique une matrice de scores à partir des informations de propriétés physicochimiques des acides aminés de la séquence. Cette matrice notée MSP (Matrice de Scores Physicochimiques) est obtenue en modélisant certaines informations permettant le repliement de la protéine à savoir la flexibilité et l'effet hydrophobe des acides aminés [Dunker et al. 2001 ; Martin 2008]. La MSP peut être calculée de deux manières : soit à partir d'une distance (approche BP1) ou soit à partir d'une fonction (approche BP). La technique BP permet de représenter ainsi une protéine par un vecteur de 400 composantes. Elle apporte plus d'informations utiles pour la reconnaissance des 'plis' des protéines et pour inférer efficacement les interactions entre les protéines. Elle ne dépend pas également d'une base de données de protéines et n'est pas longue en exécution. Par conséquent la technique BP améliore l'état de l'art sur les techniques d'extraction de caractéristiques à partir des informations de la séquence de protéine.

Après l'étape d'extraction de caractéristiques, nous avons utilisé la technique de l'analyse en composante principale (ACP) pour choisir les caractéristiques les plus pertinentes [Lorenz 1989]. En effet, pour modéliser la paire de protéine, les vecteurs de composantes issus des techniques d'extraction de caractéristiques sont généralement concaténés [Göktepe et Kodaz 2018]. L'on assiste le plus souvent à des redondances de données ou à des bruits dans les caractéristiques extraites. L'utilisation de l'ACP va permettre de ne retenir que les composantes essentielles pour l'inférence. Dans cette étude la technique de l'ACP a permis de retenir 471 composantes sur 800 composantes après application de l'approche BP et 384 composantes sur 800 composantes après application de l'approche BP1. Par la suite, nous avons introduit l'algorithme des SVM pour classifier les différentes paires d'interaction selon que les protéines formant une paire interagissent ou pas. Nous notons que l'algorithme du SVM est à l'origine une méthode de classification binaire [Cortes et Vapnik 1995] et est adapté aux observations non linéairement séparables, ce qui est notre cas. En fait, Il comporte

une fonction noyau qui permet de projeter les observations de l'espace original vers un espace beaucoup plus grand facilitant ainsi la séparation des différentes classes d'observations [Ben-Hur et Noble 2005; Brouard 2013]. Le noyau utilisé dans cette thèse est le noyau gaussien car il offre de meilleures performances de prédiction et son espace de projection est infiniment grand [Wei et al. 2016 ; Göktepe and Kodaz 2018].

Enfin nous avons proposé l'algorithme SVOH (Sélection des Valeurs Optimales d'Hyperparamètres), qui est une approche de sélection de modèles [Yang et Shami 2020] . Cet algorithme permet de rechercher de manière robuste les valeurs optimales d'hyperparamètres du modèle de SVM en vue d'améliorer les performances du classifieur utilisé. L'algorithme SVOH est une modification de l'algorithme de recherche sur grille avec validation croisée (*grid search-CV*) qui combine la technique de recherche sur grille (*grid search*) [Brito et al. 2005] avec une technique de validation croisée [Arlot et Celisse 2010], le plus souvent la validation croisée *K*-fois (VCK) pour la recherche des valeurs d'hyperparamètres. En effet, la VCK est une technique de sélection de modèles qui permet de subdiviser les échantillons d'apprentissage en *k* sous-ensembles où à tour de rôle, *k*-1 sous-ensembles sont réservés à l'entrainement pendant que l'autre sous-ensemble est réservé au test. La plupart des auteurs se basent sur des valeurs à priori pour le choix de la valeur du nombre *k* ($k = \{5, 10, 20\}$). Or, la valeur du nombre k joue sur l'erreur d'approximation et l'erreur d'estimation du modèle [Arlot et Celisse 2010] et peut donc influencer fortement les performances du modèle sélectionné et la qualité de l'erreur prédite. L'algorithme SVOH combine donc la technique de la *grid search* avec celle de la VCK sur un nombre k^* ajusté de sous-ensembles pour une recherche plus rigoureuse des hyperparamètres du classifieur du SVM [Hsu et al. 2003 ; Brito et al. 2005 ; Anguita et al. 2012]. SVOH permet donc de sélectionner les valeurs d'hyperparamètres qui permettent au classifieur du SVM de réaliser des performances supérieures.

Organisation du document

La suite du manuscrit est constituée de quatre chapitres dont les deux premiers chapitres sont des généralités, suivis d'une conclusion générale et des perspectives.

Le **chapitre 1** est dédié au contexte biologique de l'étude. Nous abordons dans ce chapitre l'intérêt de prédire les interactions chimiques entre les protéines. Après avoir décrit quelques techniques expérimentales, nous présentons certaines approches informatiques développées pour la prédiction d'interaction protéine-protéine tout en montrant les enjeux biologiques et

pharmacologiques. Nous montrons également dans ce chapitre que les techniques d'extraction de caractéristiques basées sur les informations de la séquence d'acides aminés présentent un grand intérêt pour de nombreux auteurs de la littérature.

Dans le **chapitre 2,** nous faisons un état de l'art des outils et techniques algorithmiques développés pour prédire les interactions à partir des informations de la séquence de protéine, tout en mettant un accent particulier sur les techniques d'extraction de caractéristiques basées sur la séquence d'acides aminés. Les techniques d'extraction des caractéristiques bigrammes étant d'un grand intérêt pour l'inférence des interactions, nous nous sommes beaucoup plus intéressés à ces techniques, les bigrammes étant les fréquences de deux acides aminés successifs dans la séquence d'une protéine. Nous présentons dans une des sections de ce chapitre, les techniques d'extraction de caractéristiques bigrammes existantes et leurs limites.

Le **chapitre 3** présente la nouvelle technique que nous avons développée pour extraire les caractéristiques bigrammes. Après une présentation générale de la technique, nous montrons les grandes étapes de calcul des caractéristiques avec la technique proposée. Ensuite, nous construisons deux classifieurs en combinant la technique proposée et un modèle des machines à vecteurs de supports (SVM). Les résultats obtenus après validation sur les ensembles de données IPP réelles de Saccharomyces Cerevisiae, Helicobacter pylori et les IPP humaines de la base de données HPRD montrent comment ces classifieurs peuvent constituer des outils efficaces de prédiction automatique d'interaction protéine-protéine.

Le **chapitre 4** est consacré à l'algorithme de sélection des valeurs optimales d'hyperparamètres, SVOH, que nous avons également développé. Nous présentons tout d'abord la démarche de sélection des valeurs optimales des hyperparamètres du SVM avec un nombre fixe de sous-ensembles de l'échantillon. Ensuite nous montrons comment nous pourrons sélectionner rigoureusement les valeurs optimales d'hyperparamètres en faisant varier le nombre de sous-ensembles de l'échantillon d'apprentissage. Pour finir nous évaluons la performance de l'approche proposée en comparant le modèle avec application de SVOH et le modèle avec les valeurs à priori du nombre k de sous-ensembles (k =5 ou k = 10). Les résultats obtenus indiquent que le modèle réalise des performances supérieures avec un nombre de sous-ensemble k = 7, différent des valeurs à priori.

La conclusion générale est un récapitulatif des travaux effectués. Elle conduit aux perspectives de cette thèse.

CHAPITRE 1. PREDICTION AUTOMATIQUE D'INTERACTION PROTEINE-PROTEINE

SOMMAIRE

Introduction

L'objectif de ce chapitre est de présenter le contexte biologique relatif à cette thèse. Le fonctionnement des cellules est assuré par les macromolécules biologiques, telles que les protéines, l'ADN, ou l'ARN. Les différentes fonctions de la cellule sont régulées par les interactions chimiques se produisant entre elles ainsi qu'avec d'autres petites molécules, appelées ligands ou médicaments. La compréhension de ce réseau d'interactions est nécessaire dans le processus de conception d'un médicament. Nous nous intéressons ici en particulier à la prédiction d'interactions entre les protéines. La section **1.1** aborde la notion d'interaction protéine-protéine et décrit brièvement son implication dans certaines pathologies et éventuellement certaines thérapies. La section **1.2** présente certaines techniques expérimentales de détection des interactions. Les différentes approches informatiques de prédiction d'interaction ainsi que les enjeux pour la biologie et la pharmacologie sont montrées à la section **1.3**. Nous terminons ce chapitre par la section **1.4** qui est consacrée à la démarche informatique de prédiction des interactions à partir des informations de la séquence.

1.1. Protéines et interactions chimiques

La cellule représente l'unité structurelle, fonctionnelle et biologique de tous les organismes vivants connus. Les constituants de la cellule peuvent être divisés en deux grands groupes en fonction de leur masse molaire. Le premier groupe comprend des molécules de faible masse molaire (inférieure à 2000 grammes par mole). Ce sont essentiellement les sucres, les acides gras, les acides aminés, les nucléotides et tous les précurseurs et intermédiaires du métabolisme. Le métabolisme étant l'ensemble des réactions chimiques se déroulant à l'intérieur d'un être vivant et lui permettant notamment de se maintenir en vie (fr.wikipedia.org/wiki/Métabolisme). Le second groupe comprend les molécules de grande masse molaire appelées macromolécules et sont les protéines, les acides nucléiques et les polysaccharides [Burley et al. 2017 ; Dunphy and Papin 2018].

1.1.1. Protéines

Les protéines sont les principales actrices au sein de la cellule, censées remplir les fonctions spécifiées par les informations codées dans les gènes. À l'exception de certains types d'ARN (Acide Ribonucléique), la plupart des autres molécules biologiques sont des éléments relativement inertes sur lesquels les protéines agissent [Voet and Voet 2004].

1.1.1.1. *Synthèses des protéines*

Une protéine, est un polymère d'acides aminés, c'est à dire une grande molécule formée d'une longue chaîne de plus petites, les acides aminés. Les acides aminés (figure 1-1) sont constitués d'un atome de carbone auquel sont liés un groupement amine (NH_2), un groupement acide (COOH) et une portion variable d'un acide aminé à l'autre, indiqué par la lettre **R** sur la figure 1-1 et indiqué en orange sur la figure 1-2 ; **R** pour Radical. Les acides aminés Glycine (Gly), Alanine (Ala) et Leucine (Leu) de la figure 1-2 diffèrent donc par leur radical **R** (en orange).

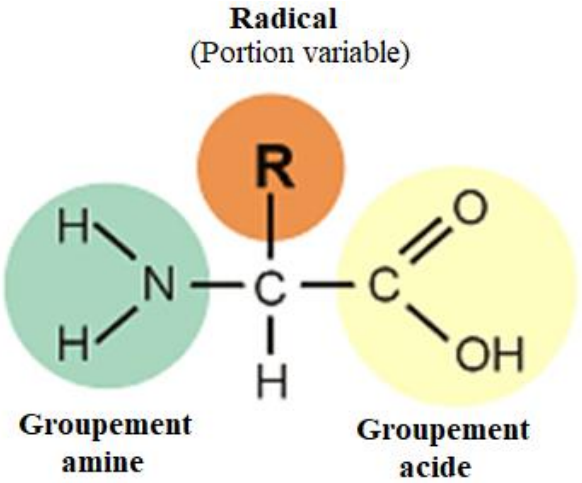


Figure 1-1: Molécule d'un acide aminé

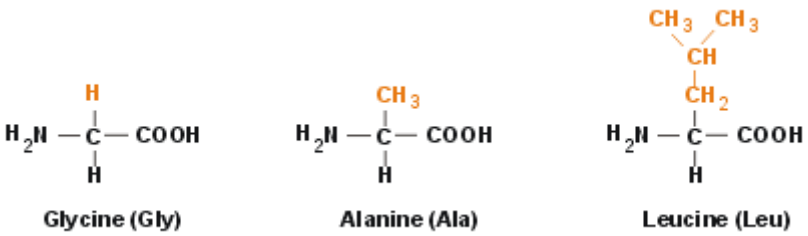


Figure 1-2: Trois des 20 acides aminés formant des protéines

Les acides aminés sont déterminés à partir du code génétique, qui, à chaque groupement de trois nucléotides, nommé codon fait correspondre un acide aminé à l'exception de trois d'entre eux qui sont nommés codons-stop [Lengyel et Söll 1969]. Nous avons par exemple AUG (Adénine-Uracile-Guanine) qui est le code de la méthionine (voir tableau 1-1). Il existe 20 acides aminés différents qui participent généralement à la synthèse des protéines. Par commodité, ils sont désignés par une abréviation de trois lettres ou par un symbole constitué

d'une lettre majuscule provenant dans la majorité de la première lettre du nom de l'acide aminé. Le tableau 1-1 nous renseigne sur les 20 acides aminés, leur abréviation et les lettres de l'alphabet qui les symbolisent.

Tableau 1-1: Les 20 acides aminés du code génétique

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

1.1.1.2. *Structure des protéines*

Les protéines peuvent être modélisées à plusieurs niveaux de granularité, appelés structure primaire, secondaire, tertiaire, et quaternaire et sont présentés ci-dessous [Anfinsen 1973].

- La structure primaire correspond à la séquence d'acides aminés. La figure 1-3 par exemple nous montre la structure primaire de la protéine antisens VIH1 ASP (ncbi.nlm.nih.gov/protein). Elle comporte 189 acides aminés dans sa chaîne et est présentée ici dans un format FASTA où le nom de la protéine est précédé du signe '>' suivi à partir de la ligne suivante de sa séquence primaire [Binz et al. 2019].
- La structure secondaire désigne l'organisation des groupes d'acides aminés en structures locales stabilisées par des liaisons hydrogène ;
- La structure tertiaire correspond au repliement de la chaîne polypeptidique dans l'espace. On parle plus couramment de structure tridimensionnelle et est intimement liée à la fonction de la protéine.
- La structure quaternaire désigne la structure formée par plusieurs molécules protéiques.

```
> YP_009028572.1 Asp [Virus de l'immunodéficience humaine 1]
MPQTVSCNRCCCASIALSKLFCCCTIPDNNCLACTVSVIEAAPIVLPAAPKNPRNKAPIPTALFSLCTTL
LFALVGATPNGSIFTTLYLYNSLLQLSLISPPPGLKISDSLLLLPPSLVNSSPVIFDEHLICPLMGGAYI
AFPTFCHMFIICFILHGRVIVSLPSVLFDPSVLQVLLNQVLLNSCVELQ
```

Figure 1-3: Séquence primaire de la protéine antisens VIH1 ASP

1.1.1.3. *Fonction des protéines*

Les protéines se retrouvent dans quasiment toutes les fonctions mises en œuvre dans la cellule. Plusieurs classes fonctionnelles sont attribuées aux protéines dont les principales selon [Lesk 2010] incluent :

- les protéines de structure, qui entrent dans la constitution des tissus ;
- les protéines de défense (par exemple les anticorps immunitaire) ;
- les protéines régulatrices, telles que les facteurs de transcription contrôlant l'expression des gènes ;
- les protéines de signalisation, qui détectent les signaux extérieurs et les transmettent dans la cellule ;
- les protéines de transport, qui contrôlent le trafic à l'intérieur comme à l'extérieur de la cellule ;
- les protéines motrices, qui permettent la motricité des cellules ou d'autres éléments de la cellule ;
- les protéines enzymatiques, qui catalysent les réactions chimiques du métabolisme.

Les enzymes sont généralement très spécifiques et n'accélèrent qu'une ou quelques réactions chimiques. Elles effectuent la plupart des réactions impliquées dans le métabolisme, ainsi que la manipulation de l'ADN dans des processus tels que la réplication de l'ADN, la réparation de l'ADN et la transcription. Certaines enzymes agissent sur d'autres protéines pour ajouter ou supprimer des groupes chimiques dans un processus connu sous le nom de modification post-traductionnelle. Nous soulignons que la modélisation structurale des protéines et la description des différents mécanismes fonctionnels ne font pas partie du cadre de notre étude ; cependant nous notons simplement que les différentes fonctions biologiques des protéines sont déterminées par leur structure tertiaire.

1.1.2. Interactions entre les protéines et cibles médicamenteuses

Parmi des milliers de protéines d'une cellule, la plupart réalisent des fonctions biologiques en s'associant avec d'autres protéines dans la cellule. Ces différentes associations sont illustrées par des contacts physiques qui ont lieu au sein de la cellule et sont appelées interactions protéine-protéine (IPP) [Brouard 2013]. Malheureusement, les interactions anormales qui perdent leur fonction ou se stabilisent à un moment ou à un endroit inapproprié sont associées à de nombreuses maladies, telle que le cancer [Chautard et al. 2009]. La connaissance des interactions reste un défi et fait l'objet d'une attention croissante de la part de la communauté scientifique [Ballone et al. 2018].

1.1.2.1. *Propriétés chimiques et fonctions biologiques*

Le dogme central de la biologie moléculaire avait comme objectif de décrire les transferts d'information entre les différentes macromolécules [Crick 1958, 1970]. En particulier, l'ADN (Acide Désoxynucléique) est vu comme le support de l'information génétique codé par la séquence des nucléotides. Les protéines assurent les fonctions biologiques de la cellule, et l'ARN a principalement pour rôle de transférer l'information de l'ADN afin de produire les protéines. Parfois, l'ARN réalise une fonction biologique propre au même titre que les protéines. Ce principe a depuis été partiellement corrigé pour tenir compte de nouvelles découvertes notamment concernant les fonctions de l'ARN, mais reste suffisant pour décrire la problématique de la prédiction d'interactions. Ce sont donc essentiellement l'ARN et les protéines qui assurent les fonctions biologiques de la cellule. Pour ce faire ces macromolécules interagissent entre elles ou avec d'autres molécules selon leurs propriétés géométriques et chimiques propres.

1.1.2.2. *Réseau d'interaction protéine-protéine*

L'ensemble des interactions entre protéines ayant lieu dans un organisme, un organe ou un type de cellules donné est appelé interactome [Lage 2014]. Il peut être représenté par un réseau d'interactions dans lequel chaque protéine correspond à un nœud, et où deux nœuds sont liés par un arc signifie l'existence d'une interaction entre les deux protéines correspondantes. La figure 1-4 illustre l'interactome de la schizophrénie [Ganapathiraju et al. 2016]. Sur cette figure, les gènes associés à la schizophrénie sont représentés sous forme de nœuds bleu foncé, de nouveaux interacteurs sous forme de nœuds de couleur rouge et les interacteurs connus sous forme de nœuds de couleur bleue. Les bords rouges sont les nouvelles interactions, tandis que les bords bleus sont des interactions connues.

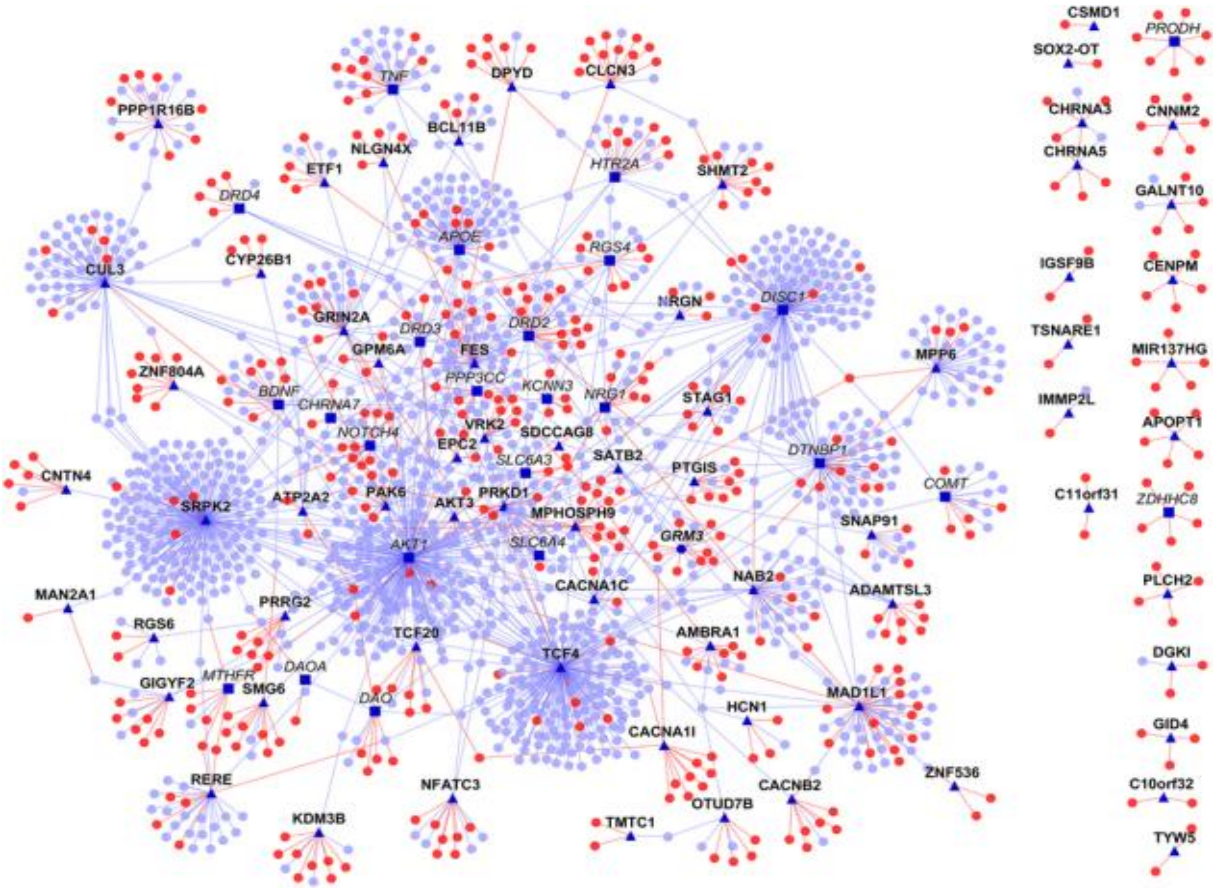


Figure 1-4: Interactome de schizophrénie [Ganapathiraju et al. 2016]

Dans un réseau d'interaction protéine-protéine, la connaissance des fonctions des protéines peut être obtenue par les études d'interactions entre protéines. Pour cela, l'on se base sur l'hypothèse que la fonction d'une protéine inconnue sera identifiée au travers de son interaction avec une protéine cible de fonction connue [Bakail et Ochsenbein 2016]. L'identification et la caractérisation de ces interactions est essentielles pour mieux comprendre les mécanismes des processus biologiques au niveau moléculaire et ainsi proposer des cibles thérapeutiques.

1.1.2.3. *Cible thérapeutique*

Une cible thérapeutique ou médicamenteuse ou encore cible biologique désigne la protéine dans le corps dont l'activité est modifiée par un médicament, une drogue ou un ligand endogène ou exogène entrainant un effet thérapeutique souhaitable ou un effet indésirable [Bakail and Ochsenbein 2016; Volet 2017]. La principale extension qu'on peut faire au dogme central de la biologie moléculaire qui soit nécessaire à la compréhension des mécanismes cellulaires est le rôle des petites molécules comme présenté sur la figure 1-5

[Schreiber 2005]. En effet des interactions, généralement non covalentes, peuvent avoir lieu entre des petites molécules et les macromolécules, ce qui peut modifier la fonction biologique des macromolécules. Ces petites molécules peuvent être endogènes, c'est à-dire produites par l'organisme, ou exogènes, issues du milieu extérieur et éventuellement artificielles comme dans le cas de certains médicaments. Ces petites molécules ont ainsi souvent un rôle de régulation sur la fonction des macromolécules en activant ou désactivant certaines fonctions.

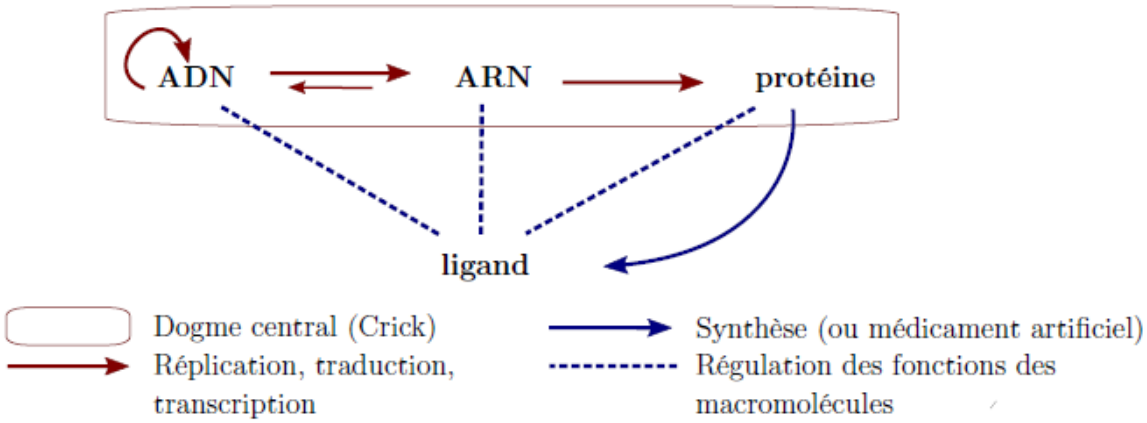


Figure 1-5 : Rôle des petites molécules dans le cadre du dogme central régissant les macromolécules [Volet 2017]

Le ciblage thérapeutique a été dans le passé le domaine des industries biotechnologiques [Ottman 2013]. Pour la plupart des tumeurs humaines, plusieurs cibles thérapeutiques ainsi que des médicaments potentiels sont connues [Bakail et Ochsenbein 2016; Makondi et al. 2018]. La pertinence de l'interaction protéine-protéine comme cibles thérapeutiques pour le développement de nouveaux traitements est particulièrement évidente dans le cancer, avec plusieurs essais cliniques en cours dans ce domaine. Le consensus parmi ces cibles prometteuses est néanmoins indiqué dans les médicaments déjà disponibles sur le marché pour traiter une multitude de maladies. Des exemples sont le tirofiban [McClellan and Goa 1998] utilisé comme médicament cardiovasculaire, et le maraviroc [Perry 2010], utilisé comme médicament anti-VIH.

1.2. Détection expérimentale d'interaction

Des méthodes à grande échelle ont été inventées pour détecter les interactions protéine-protéine, notamment les puces protéomiques, l'immunoprécipitation, la technique du double hybride chez la levure (Y2H), la spectrométrie de masse (MS), et bien d'autres [Ito et al.

2001; Liu et al. 2002 ; Waksman 2005]. Plusieurs de ces méthodes, comme le système Y2H, sont des approches binaires, c'est à dire qu'elles sont capables de mesurer une interaction directe entre deux protéines. Cependant, d'autres méthodes, comme l'immunoprécipitation, permettent d'identifier des complexes protéiques (interactions entre des groupes de protéines). Ce type de méthode ne permet pas de connaître les protéines étant en contact direct, mais apporte néanmoins une information sur les protéines qui sont trouvées dans un même réseau à un instant donné [Brouard 2013].

Les méthodes expérimentales se différencient selon le nombre d'interactions qu'elles arrivent à détecter. Ainsi nous pouvons distinguer les méthodes de détection à petite échelle, qui se concentrent spécifiquement sur un petit nombre d'interactions tandis que les méthodes de détection à grande échelle, comme le Y2H, permettent de détecter un très grand nombre d'interactions. Nous notons également que les méthodes expérimentales peuvent détecter les interactions *in vivo*, c'est à dire au sein d'un organisme, ou *in vitro*, c'est à dire en dehors d'un organisme vivant ou d'une cellule [Uetz et al. 2008]. Les principales méthodes de détection expérimentale et leurs caractéristiques sont listées dans les tableaux 1-2 et 1-3.

1.2.1. Méthodes de détection à petite échelle

Certaines méthodes de détection à petite échelle présentées dans le tableau 1-2, comme les méthodes de spectroscopie par résonance magnétique nucléaire et cristallographie aux rayons X permettent ainsi de caractériser finement les interactions et l'interface d'interaction entre deux protéines.

Tableau 1-2: Méthodes expérimentales de détection des interactions à petite échelle [Shoemaker et Panchenko 2007a]

Méthode	Condition	Type
Cristallographie aux rayons X	*in vitro*	complexe
Spectroscopie par résonance magnétique nucléaire	*in vitro*	complexe
FRET	*in vivo*	binaire
Co-immunoprécipitation	*in vitro / in vivo*	complexe

La co-immunoprécipitation et le FRET (Fluorescence Resonance Energy Transfert) sont deux méthodes de détection d'interactions *in vivo*. La première consiste à isoler un groupe de protéines (complexe protéique) en utilisant un anticorps dirigé contre un des membres du groupe, et à identifier ensuite les protéines obtenues. Pour ce qui est du FRET, c'est une technique qui permet de détecter la proximité immédiate de deux protéines de fluorescence.

Ainsi, lorsque ces deux protéines sont très proches l'une de l'autre, un transfert d'énergie de fluorescence est effectué entre elles, ce qui a pour conséquence de modifier leurs intensités de fluorescence respectives. La combinaison de cette approche avec la technique FLIM (Fluorescence Lifetime Imaging Microscopy) permet de détecter des interactions protéine-protéine directes.

Toutefois, les méthodes de détection expérimentale à petite échelle restent en général relativement coûteuses et peuvent nécessiter un temps important pour détecter une interaction protéine-protéine allant jusqu'à plusieurs mois [Brouard 2013].

1.2.2. Méthodes de détection à grande échelle

Des méthodes capables de détecter un très grand nombre d'interactions, qualifiées de méthodes à grande échelle, ont été développées (voir tableau 1-3). La méthode à grande échelle la plus utilisée est celle du système Y2H. Le principe de cette méthode est qu'un facteur de transcription est découpé en deux parties : un domaine de fixation sur l'ADN (BD ou *Binding Domain*) et un domaine d'activation (AD ou *Activation Domain*). Chaque protéine d'intérêt est couplée à l'un des domaines et si les deux protéines interagissent ensemble, le facteur de transcription devient actif et le gène rapporteur est transcrit. Cette méthode permet d'identifier à grande vitesse des interactions protéine-protéine. Des cartes d'interaction ont ainsi été identifiées chez la levure *Saccharomyces Cerevisiae* [Uetz et al. 2000], la bactérie *Helicobacter pylori* [Rain et al. 2001], le ver *Caenorhabditis elegans* [Dupuy et al. 2004] et l'homme [Rual et al. 2005]. Cette technique étant utilisée *in vivo*, elle permet de détecter des interactions transitoires et instables.

Tableau 1-3: Méthodes expérimentales de détection à grande échelle des interactions [Shoemaker et Panchenko 2007]

Méthode	Condition	Type
Double hybride chez la levure (Y2H)	*in vivo*	binaire
Purification par affinité couplée à la spectrométrie de masse	*in vitro*	complexe
Puces à protéine	*in vitro*	complexe
Phage display	*in vitro*	complexe

Malgré son utilité, le système Y2H présente des limites. Le principal inconvénient de cette méthode est les taux importants de faux positifs qui sont les interactions détectées expérimentalement mais qui n'existent pas en réalité et de faux négatifs qui sont les

interactions existantes qui n'ont pas été détectées par la méthode (estimés à environ 50% selon [Von Mering et al. 2002]). Ces erreurs s'expliquent en partie du fait que les interactions sont testées dans le noyau, qui est différent du compartiment d'origine de beaucoup de protéines. De ce fait, des interactions physiques peuvent être détectées entre des protéines qui ne sont jamais à proximité l'une de l'autre dans la cellule car elles ont des localisations cellulaires différentes ou bien parce qu'elles sont exprimées à des moments différents du cycle cellulaire.

Parmi les méthodes de détection expérimentale à grande échelle, se trouve les méthodes de purification par affinité couplées à la spectrométrie de masse (TAP-MS) [Domon et Aebersold 2006; Gavin et al. 2002], qui sont très utilisées pour l'identification de complexes protéiques. Cette technique consiste tout d'abord à marquer individuellement les protéines d'intérêt. Ces protéines sont utilisées pour récupérer par purification biochimique l'ensemble des protéines du complexe. Les différentes protéines du complexe sont ensuite séparées et identifiées par spectrométrie de masse. Nous avons également les puces à protéines qui permettent de détecter les IPP à grande échelle [Zhu et al. 2001 ; Sakanyan et Arnaud 2007]. Cependant, ces différentes techniques soufrent également de taux importants de faux positifs et faux négatifs.

Ces expériences au laboratoire ont permis d'identifier un nombre élevé d'interactions protéine-protéine, néanmoins beaucoup reste encore non découvertes [Shin et al. 2017].

1.3. Outils informatiques pour la prédiction d'interaction

Les approches *in vitro* et *in vivo* ont permis le développement à grande échelle d'outils utiles pour la détection des interactions protéine-protéine. Cependant, elles génèrent un grand coût avec une lenteur dans l'obtention des résultats. Pour mieux élucider le contexte global des interactions potentielles, il est préférable de développer des approches qui prédisent la gamme complète des interactions entre les protéines. Plusieurs approches informatiques ont donc été proposées pour enrichir les connaissances trouvées par les méthodes expérimentales. Dans cette section nous présentons quelques-unes des principales approches après avoir montré de façon générale les enjeux pour la biologie et la pharmacologie.

1.3.1. Enjeux pour la biologie et la pharmacologie

Le développement d'outils informatiques pour la prédiction d'interaction présente des enjeux importants pour la biologie et la pharmacologie. Les outils informatiques permettent

par exemple d'accélérer la découverte de médicaments en réduisant l'espace des cibles, en ce sens que les cibles identifiées par ces outils seront à priori celles sur lesquelles les protocoles expérimentaux devront être appliqués pour vérification. Cela va non seulement limiter l'espace de recherche mais aussi le temps et le coût de recherche. Les outils informatiques constituent également une véritable aide dans le processus de conception d'un médicament car ils guident le choix des composés ayant les meilleurs potentiels thérapeutiques. Le médicament est dans la plupart des cas une molécule organique qui vient se fixer sur une macromolécule biologique, qui est la cible, impliquée dans la pathologie et agit sur cette dernière afin d'en moduler les effets. Ces outils doivent donc permettre de décrire les interactions entre le ligand et la cible biologique, et permettre de tester rapidement plusieurs modes d'opération afin de guider la recherche.

Un *complexe ligand-cible* est défini comme l'interaction d'une petite molécule appelée ligand avec une macromolécule ou grosse molécule, généralement une protéine, appelée cible [Fischer 1894]. Ces interactions entre des ligands et leurs cibles interviennent naturellement dans les cellules en permettant notamment la régulation des fonctions des macromolécules. La pharmacologie vise à étudier ces interactions afin d'agir sur ces mécanismes en introduisant de nouveaux ligands dans l'organisme.

1.3.2. Principales approches informatiques dans la prédiction d'interaction

De nombreuses approches informatiques ont ainsi été développées pour résoudre le problème de la prédiction de liens physiques entre protéines à partir de diverses sources de données indirectes [Shen et al. 2007]. L'on peut classer ces sources d'information en plusieurs types dont les principaux sont : génomiques, structures et séquences. Dans la suite de cette section, nous détaillons les approches appartenant à ces différents types.

1.3.2.1. *Approches génomiques*

Un certain nombre de méthodes analysent le contexte génomique dans différentes espèces (voir figure 1-6) afin d'inférer des associations fonctionnelles entre des gènes, et donc potentiellement des interactions entre les protéines codées par ces gènes. Nous présentons ici deux d'entre elles.

La première approche est basée sur l'hypothèse que des protéines ayant co-évoluées sont susceptibles de présenter des fonctions similaires [Pellegrini et al. 1999]. Pour cela, cette approche s'intéresse aux protéines ayant des homologues dans les mêmes organismes. Chaque

protéine est représentée par un profil phylogénétique, qui correspond à un vecteur dont la longueur est égale au nombre d'espèces considérées. Ce vecteur binaire indique la présence ou l'absence d'un orthologue du gène associé à la protéine dans l'espèce correspondante. Deux gènes sont dits orthologues s'ils sont issus d'un ancêtre commun et ne résultent pas d'une duplication génétique (copie accidentelle d'un gène).

Une deuxième catégorie d'approches consiste à analyser la fusion des gènes [Enright et al. 1999; Ghanty et Pal 2009]. Ces méthodes recherchent ainsi des protéines ayant des homologues qui ont fusionnés en une seule protéine dans un autre génome.

D'autres méthodes se basent sur l'hypothèse que des gènes présentant une association fonctionnelle restent proches [Overbeek et al. 1999].

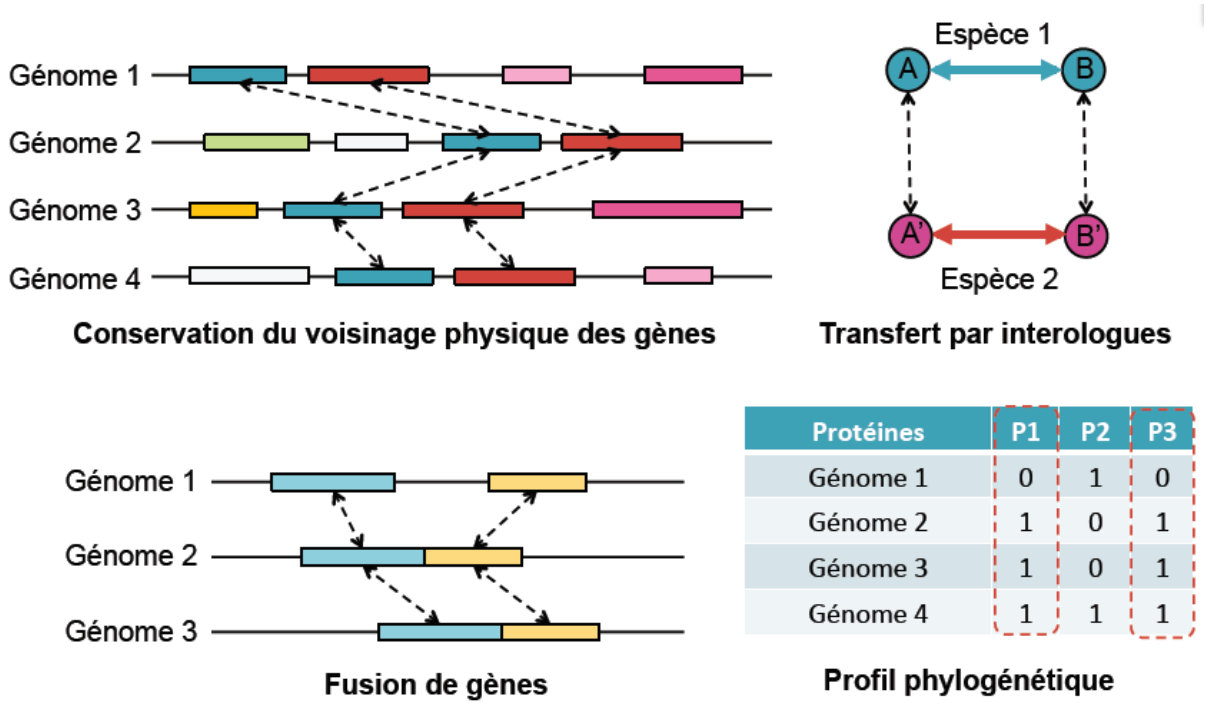


Figure 1-6: Méthodes basées sur l'analyse génomique [Brouard 2013]

1.3.2.2. *Approche basée sur la structure tertiaire d'une protéine*

L'approche basée sur la structure tertiaire prédit l'interaction selon que deux protéines ont une structure similaire. Cette approche utilise comme caractéristiques de prédiction la représentation structurelle d'une protéine, c'est à dire la forme, les surfaces réceptrices, les hélices, les informations structurelles 3D, et bien d'autres. [Smith et Sternberg 2002 ; Veber 2007].

Ces approches sont ainsi capables dans certains cas de déterminer les caractéristiques physiques d'une interaction ainsi que le site d'interaction à la surface des protéines [Gamblin et al. 2004]. Ce sont par exemple les méthodes de Docking, les méthodes par homologie.

Modélisation Docking

Les méthodes de Docking étudient la complémentarité entre les structures connues des protéines. Ces méthodes procèdent en deux étapes. La première consiste à générer un grand nombre de conformations (représentation tridimensionnelle) possibles pour l'association des deux protéines. Ensuite, une fonction de score est utilisée pour classer les différentes conformations [Smith et Sternberg 2002 ; Barradas-Bautista et al. 2018]. Les méthodes de Docking sont certes très précises, cependant elles sont très coûteuses en temps de calcul. Aussi, cette approche est limitée par le fait que les structures tertiaires résolues expérimentalement ne sont disponibles que pour une faible proportion des protéines [Zhang et al. 2012].

Modélisation par homologie

Pour contourner la limitation avec les méthodes de Docking, certaines méthodes utilisent les structures résolues expérimentalement pour modéliser les interactions entre des protéines, pour lesquelles la structure n'est pas connue [Blüthner et al. 2000 ; Martin 2005]. L'hypothèse sous-jacente à ces approches est que les protéines présentant une importante homologie de séquence ou de structure interagissent généralement de la même façon. Ces méthodes, contrairement au Docking, peuvent donc être appliquées à l'échelle d'un interactome.

Définition 1-1 : Homologie de séquences [Martin 2005]

Deux protéines sont dites homologues si elles dérivent d'un ancêtre moléculaire commun. Au cours de l'évolution, des actions de mutation et délétion s'opèrent sur les séquences d'ADN. Ces mutations sont conservées si les protéines codées sur les gènes conservent leurs fonctions et donc leur structure tridimensionnelle, en raison de la pression de sélection qui tend à maintenir la fonction. La conséquence de cette pression sélective est que des séquences différentes peuvent adopter la même structure. Par exemple sur la figure 1-7, en (a), le cas de deux enzymes qui assurent la même fonction chez la levure et chez le blé (structures PDB 1ayz et 2aak). Les deux séquences sont identiques à 63%. En (b), le cas de deux protéines constituant du muscle, chez le nématode et chez l'homme (structures PDB 1wit et 1tit). L'identité de séquences n'est que de 9%.

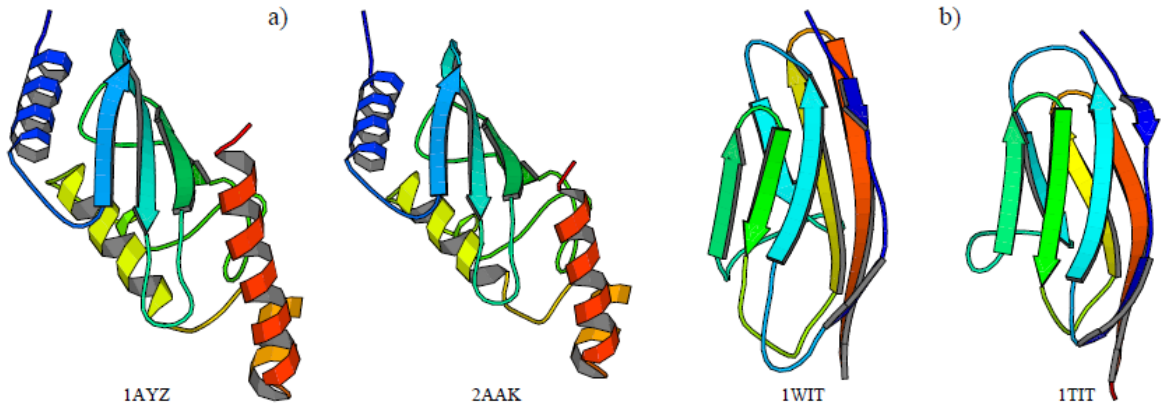


Figure 1-7: Example de paires de protéines homologues [Martin J. 2005]

PDB signifie Protein Data Bank ou Banque de données sur les protéines. Il faut souligner que chaque modèle moléculaire de la PDB possède un code d'accès ou d'identification unique. Ces codes comportent toujours 4 caractères. Le premier caractère est un chiffre compris entre 1 et 9, tandis que les trois derniers caractères peuvent être des chiffres (compris entre 0 et 9) ou des lettres (comprises entre A et Z dans l'alphabet latin) [Burley et al. 2017]

1.3.2.3. *Approches basées sur la séquence d'acides aminés*

Les approches basées sur la séquence primaire d'acides aminés des protéines utilisent les informations codées inhérentes aux séquences telles que la longueur de la séquence et le nombre d'acides aminés de la séquence puis à l'aide d'un algorithme d'apprentissage supervisé prédisent si deux protéines formant une paire interagissent ou non [Göktepe et Kodaz 2018 ; An et al. 2019 ; Ma et al. 2020]. En plus des informations de la séquence, certaines méthodes considèrent les informations des propriétés physico-chimiques des acides aminés, telles que l'hydrophobicité, l'hydrophilie et la polarité [Guo et al. 2008 ; Jia et al. 2016 ; Du et al. 2017], pendant que d'autres considèrent les informations de domaines protéiques [Zhang et al. 2016].

Un domaine protéique se défini généralement comme une partie conservée de la séquence d'acides aminés d'une protéine et d'une structure tridimensionnelle qui associe les fonctions biologiques d'une protéine tout en se pliant et en évoluant indépendamment [C.-H. Huang et al. 2015]. Les interactions entre domaines sont considérées comme des moteurs des interactions entre protéines [González et Liao 2010].

Plusieurs méthodes utilisent l'information des domaines pour prédire de nouvelles interactions protéine-protéine. Celles-ci se basent sur le fait que dans certains cas, les

protéines interagissent entre elles par l'intermédiaire d'interactions physiques entre domaines [Zhang et al. 2016]. Cependant, l'inconvénient de ce type d'approche est que le nombre d'interactions domaine-domaine ayant été détectées expérimentalement est limité. Par conséquent, une stratégie couramment utilisée consiste à commencer par identifier des paires de domaines susceptibles d'interagir ensemble à partir d'un ensemble d'interactions protéine-protéine connues. Les interactions prédites entre les domaines sont ensuite utilisées pour prédire de nouvelles interactions entre protéines [Shoemaker et Panchenko 2007b].

Les méthodes par association s'intéressent à des séquences ou des motifs structuraux caractéristiques permettant de faire la distinction entre les protéines qui interagissent et celles qui n'interagissent pas. Dans le cas particulier des domaines, ces méthodes recherchent les paires de domaines sur-représentées parmi les interactions protéine-protéine connues [Wan et al. 2002]. Pour cela, la fréquence de co-occurrence de chaque paire de domaines parmi les paires de protéines interagissant ensemble est calculée. Une méthode d'estimation du maximum de vraisemblance [Guo et al. 2018] a été par ailleurs proposée afin d'estimer les probabilités d'interaction entre domaines qui sont consistantes avec les interactions entre protéines observées. Cette méthode a été étendue dans un premier temps par Riley et al. [2005], puis dans un deuxième temps par [Lee et al. 2006].

Les approches citées précédemment considèrent uniquement des interactions entre deux domaines et supposent que les paires de domaines interagissant ensemble sont indépendantes les unes des autres. Comme les protéines peuvent contenir plusieurs domaines, [Han et al. 2003] ont proposé de considérer les interactions entre protéines comme le résultat d'interactions entre des groupes de domaines. La différence entre ces deux types d'approche est illustrée à travers la figure 1-8 où en (1) nous avons une illustration des méthodes de prédiction basées sur les interactions entre domaines et celle de droite en (2) la méthode basée sur les interactions entre combinaisons de domaines.

Les approches basées sur la séquence ne demandent pas une connaissance approfondie des séquences, contrairement aux autres approches. En outre, de nombreuses études ont montré que la prédiction basée sur les séquences, en l'occurrence, l'identification des protéases et de leurs types [Chou et Shen 2008b; Huang et al. 2014], la prédiction de la localisation subcellulaire des protéines [Chou et Shen 2006, 2008a; Rouillon et Cetau 2000], peut fournir des informations très utiles pour la recherche fondamentale et la conception de médicaments.

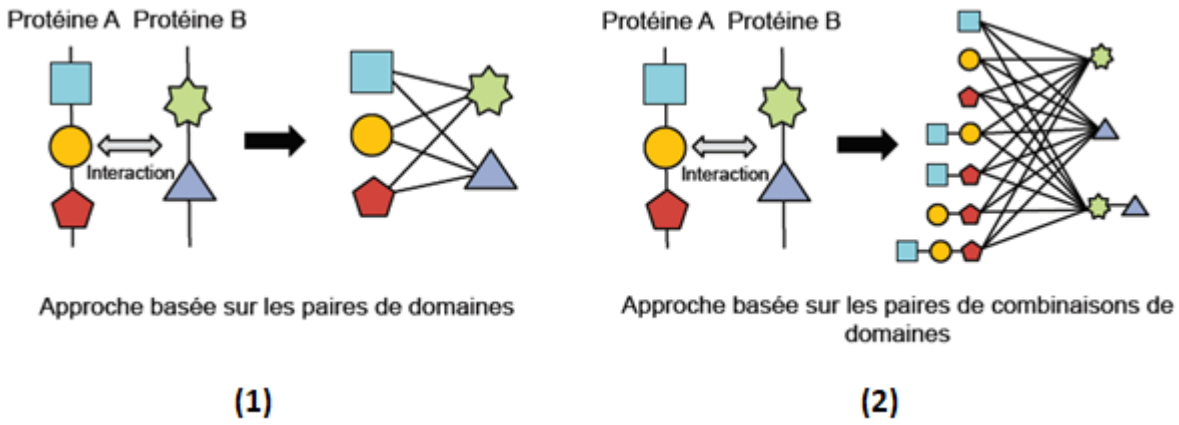


Figure 1-8: Modèles de prédiction basé sur l'information des domaines [Brouard 2013]

1.4. Méthodologie des approches informatiques de prédiction basées sur la séquence

Plusieurs méthodes de prédiction d'interactions protéine-protéine utilisent la séquence primaire des protéines [Bock et Gough 2001; Guo et al. 2008 ; Pan et al. 2010 ; You et al. 2014 ; Göktepe et Kodaz 2018 ; Ma et al. 2020]. Nous présentons dans cette section le fonctionnement général de telles approches, les techniques algorithmiques classiques feront l'objet du second chapitre.

1.4.1. Définitions et concepts autour de la séquence de protéine

En général, une séquence est une liste ordonnée d'événements. Un événement peut être représenté par une valeur symbolique, une valeur numérique réelle, un vecteur de valeurs réelles ou un type de données complexes. La séquence d'une protéine donnée est une longue chaîne d'acides aminés symbolisée par des lettres de l'alphabet (voir tableau 1-1) et peut-être assimilé à un mot. Par conséquent toutes les définitions relatives à un mot s'appliquent également à une séquence [Liefooghe 2008].

Définition 1-2 : **Séquence**

Soit Σ *un alphabet fini de lettres* $\{l_1, l_2, l_3, \dots, l_{20}\}$. *Une séquence sur l'alphabet* Σ *est une suite finie de lettres de* Σ.

Définition 1-3 : Longueur de séquence

Soit S une séquence. La longueur d'une séquence S est définie par la longueur de sa suite de lettres associées et est notée $L = |S|$,

Exemple 1.1. $|ACDDF| = 5$.

Définition 1-4 : Position dans une séquence

Soit S une séquence et $i = 1, 2, \ldots, |S|$, *pour une numérotation des indices commençant à* 1, *une position i de S est notée* S_i *et est la* $i^{ème}$ *lettre de S.*

Exemple 1.2. Pour *S =ADC*, $S_2 = D$.

Définition 1-5 : Concaténation

Soit S_1 *et* S_2 *deux séquences, la concaténation de ces deux séquences* S_1 *et* S_2 *est la séquence composée des lettres de* S_1 *puis celle de* S_2 *notée* S_1S_2.

Exemple 1.3. Soit $S_1 = ACD$ *et* $S_2 = MAC$, $S_1S_2 = ACDMAC$.

Sources de données

L'identification à grande échelle des interactions protéine-protéine par des protocoles expérimentaux a généré des centaines de milliers d'interactions. L'information de ces différentes interactions est stockée dans différentes bases de données biologiques spécialisées et en libre accès [Patil 2019]. Le tableau 1-4 indique les principales d'entre elles tout en donnant leur URL et une petite description.

Tableau 1-4: Bases de données biologiques

Bases de données	URL	Petite description
HPRD [Keshava Prasad et al. 2009] Human Protein Reference Database	https://www.hprd.org	Des millions de données sur les IPP humaines
BioGRID [Chatr-aryamontri et al. 2017] Biological General Repository for Interactions Datasets	https://www.thebiogrid.org/	Curation manuelle des interactions protéiques et génétiques validées expérimentalement contient 1.072.173 interactions protéiques
DIP [Xenarios et al. 2000] Database of Interacting Proteins	https://dip.doe-mbi.ucla.edu/	Gère les IPP déterminés expérimentalement
IntAct [Areta et al. 2010] The IntAct molecular interaction Database	https://www.ebi.ac.ek/intact/	Interactions moléculaires de plusieurs organismes

Définition 1-6 : **Ensemble de données** [Chou 2011]

Pour développer une méthode de prédiction statistique pour un attribut donné, la première chose importante est de construire un ensemble de données de référence (*benchmark dataset*) en fonction de sa classification possible comme suit :

$$\mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_m \cup \ldots \cup \mathbb{S}_M$$

où $\mathbb{S}_1$ représente le sous-ensemble de la catégorie 1 de l'attribut, $\mathbb{S}_2$ pour la catégorie 2 et ainsi de suite, et *M*, le nombre de catégories différentes pour l'attribut concerné. Dans le cas des interactions protéine-protéine, nous avons deux classes à savoir celle où les protéines interagissent (IPP positives) et celle où les protéines n'interagissent pas (IPP négatives). Les ensembles de données IPP constituent en fait les échantillons originaux d'IPP. Trois ensembles de données IPP de référence sont généralement utilisés pour la prédiction d'interaction à partir des séquences de protéines. Nous avons les IPP humaines HPRD [Keshava Prasad et al. 2009], les IPP de la levure Saccharomyce Cerevisiae (S. Cerevisiae) [Uetz et al. 2000] et les IPP de la bactérie Helicobacter Pylori (H. Pylori) [Rain et al. 2001], et sont détaillées ci-dessous.

- **HPRD**

Les ensembles de données IPP HPRD sont constituées à partir de la base de données de référence sur les protéines humaines en abrégé HPRD (*Human Protein Reference Database*). Nous notons que les bases de données biologiques indiquées dans le tableau 1-4 ne stockent que les paires positives, c'est-à-dire uniquement les paires où les protéines formant la paire interagissent entre elles [Patil 2019]. Etant donné que nous voulons prédire les interactions entre les protéines à l'aide d'un algorithme d'apprentissage supervisé, nous avons besoin autant de paires positives (échantillons positifs) que de paires négatives (échantillons négatifs). Les paires négatives n'étant pas stockées, elles sont créées en appariant des protéines situées dans des emplacements sous cellulaires distincts [Wang et al. 2014]. Ainsi, l'ensemble de données IPP HPRD issu des travaux de Pan et al. [2010] comporte au total 73110 paires de protéines reparties en 36630 paires positives et 36480 paires négatives.

- **S. Cerevisiae**

L'ensemble de données IPP S. cerevisiae est celui décrit par You et al. [2013], ensemble de données recueillies à partir du sous-ensemble de S. Cerevisiae dans la base de données des protéines en interaction DIP (*Database of Interacting Proteins*). Cet ensemble de données est

constitué de 5594 paires positives et 5594 paires négatives, combinées en un total de 11188 paires de protéines.

- **H. Pylori**

L'ensemble de données H. pylori décrit par Martin et al [2005], comprend 2916 paires de protéines, dont 1458 paires positives et 1458 paires négatives.

1.4.2. Mise en place d'un outil informatique de prédiction d'interaction

Un outil informatique de prédiction d'interaction protéine-protéine peut être défini comme un modèle informatique qui utilise les algorithmes d'apprentissage supervisé pour prédire ou classifier les interactions entre les protéines. Deux grandes étapes participent généralement à la construction de ce type d'outil. Nous avons d'une part l'étape de la représentation des données qui constitue l'extraction de caractéristiques. D'autre part nous avons l'étape de prédiction ou de classification à l'aide d'un algorithme d'apprentissage supervisé [Y.-A. Huang et al. 2015]. Nous pouvons ajouter une troisième étape qui permet d'évaluer la performance du classifieur (fonction de prédiction) construit à travers des techniques de validation croisé [Anguita et al. 2009]. Ces trois étapes constituent ainsi le cœur de l'outil de prédiction et sont schématisées à travers la figure 1-9.

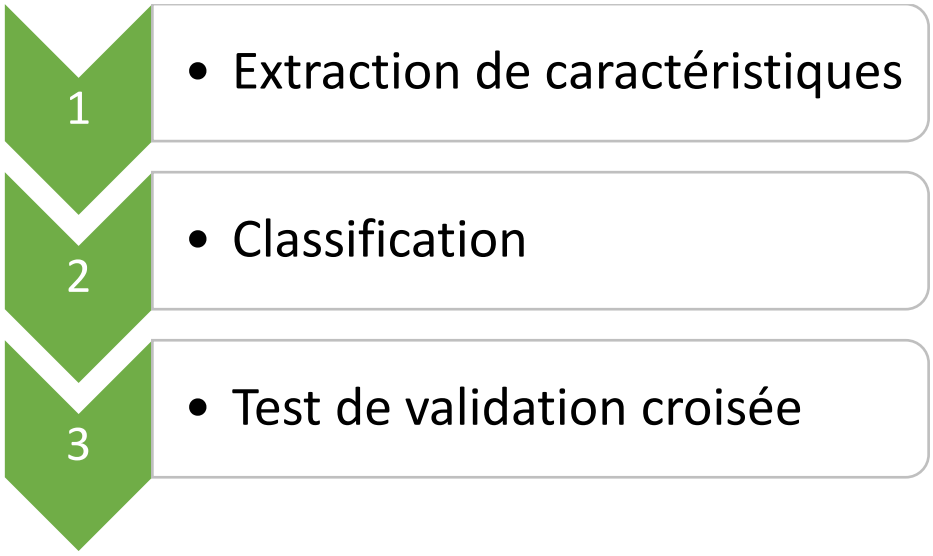


Figure 1-9: Les 3 niveaux du cœur d'un outil informatique de prédiction d'interaction basé sur les séquences

1.4.2.1. *Extraction de caractéristiques*

L'extraction de caractéristiques est une technique de l'apprentissage de la représentation des données [Bengio et al. 2013 ; Tsubaki et al. 2017]. Cette première étape des méthodes de calcul dans la prédiction d'IPP à partir des séquences, vise à extraire à l'aide d'une formulation mathématique les attributs les plus représentatifs des échantillons originaux et à les représenter sous forme de vecteurs de caractéristiques normalisés de même taille [Y.-A. Huang et al. 2015]. Le défi dans cette étape est que pour une protéine donnée, les caractéristiques extraites doivent être corrélées à l'information d'interaction de la protéine avec une autre protéine. Il s'agit ici de représenter les données de telles sortes à extraire les facteurs explicatifs de variations derrière les données.

1.4.2.2. *Classification*

Les échantillons utilisés dans la prédiction d'interaction entre les protéines peuvent être regroupés en deux classes dont celle où il y'a interaction et celle sans interaction. La classification revient donc à construire une fonction de prédiction qui sépare au mieux les deux classes à l'aide d'algorithmes d'apprentissage supervisé [Sathya et Abraham 2013]. Cette fonction appelée classifieur doit être à mesure de prédire correctement la classe associée à une nouvelle observation. Il faut souligner qu'une méthode efficace d'extraction de caractéristiques aide généralement le système de prédiction à améliorer ses performances.

Conclusion

Nous avons présenté dans ce chapitre le contexte de l'étude. Nous avons montré que la détection des interactions entre protéines est importante pour comprendre certains processus biologiques mais surtout pour proposer des cibles médicamenteuses. Les expériences en laboratoire étant longues et coûteuses, les outils informatiques viennent compléter les techniques expérimentales limitant ainsi le temps et les coûts des expériences. Plusieurs outils informatiques sont donc proposés, principalement ceux utilisant les informations de la séquence qui est l'approche utilisée dans cette thèse. Dans le chapitre suivant, nous présentons certaines techniques algorithmiques développées pour prédire les interactions à partir des informations de la séquence.

CHAPITRE 2. OUTILS ET TECHNIQUES INFORMATIQUES POUR LA PREDICTION D'INTERACTION

SOMMAIRE

Introduction

Ce chapitre est consacré à un état de l'art sur les outils et techniques informatiques développées pour la prédiction d'interactions entre les protéines à partir des informations de la séquence primaire. Ici, nous mettons un accent particulier sur les techniques d'extraction de caractéristiques à partir des informations de la séquence qui reste le grand défi dans le développement de ces outils et techniques. Plus précisément, nous nous intéressons à la technique d'extraction des bigrammes de la séquence. La section **2.1** présente les techniques algorithmiques d'apprentissage automatique, précisément d'apprentissage supervisé pour l'inférence des interactions. Les différentes approches de l'extraction de caractéristiques à partir des données de la séquence sont décrites dans la section **2.2**. Enfin, la section **2.3** aborde la problématique de l'extraction des caractéristiques bigrammes de la séquence avec les techniques existantes.

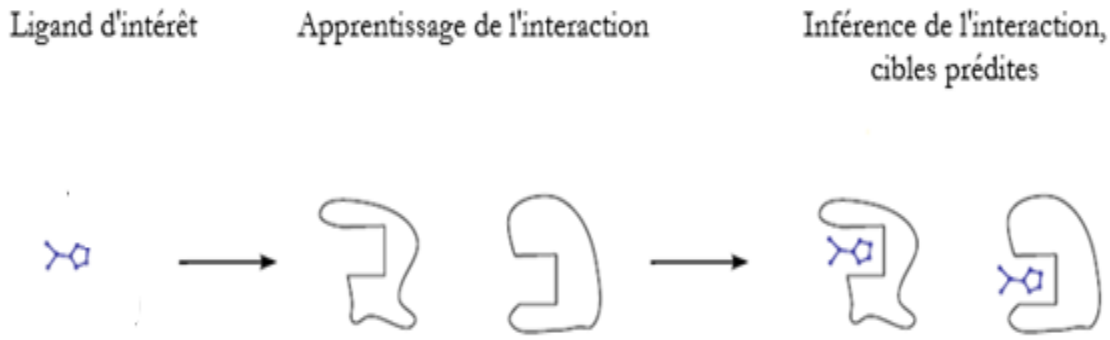
2.1. Apprentissage supervisé pour la prédiction entre protéines

L'apprentissage supervisé est l'une des tâches importantes dans la prédiction d'interaction protéine-protéine. Dans cette section, nous présentons le principe d'inférence de l'interaction ainsi que quelques techniques autour de l'apprentissage supervisé.

2.1.1. Inférence de l'interaction

L'objet de notre étude est le problème de la recherche d'une fonction *f* qui soit à mesure d'apprendre l'interaction entre deux protéines. Cette fonction est l'outil pour répondre à la problématique biologique et pharmacologique de la prédiction d'interaction protéine-protéine [Chautard et al. 2009 ; Bakail and Ochsenbein 2016]. Il existe d'autres applications biologiques ou pharmacologiques qui peuvent se traduire par un problème semblable à la recherche d'une fonction pour prédire l'interaction entre protéines, et certaines méthodes peuvent être utilisées pour répondre à différentes problématiques biologiques. Toutefois chaque problématique induit des choix généralement spécifiques dans les modèles et algorithmes, et chaque approche est souvent spécialisée pour répondre à une ou des problématiques biologiques précises. Ces problématiques et leurs liens avec la recherche d'une fonction par l'apprentissage de l'interaction sont illustrées à la figure 2-1.

Prédiction de cibles [Shin et al. 2017]**.** Il s'agit de prédire de potentielles cibles pour un ligand d'intérêt.



Annotation fonctionnelle [Saha et al. 2014]**.** Il s'agit de prédire de nouveaux ligands pour une protéine d'intérêt, à partir des autres cibles connues de ces ligands.

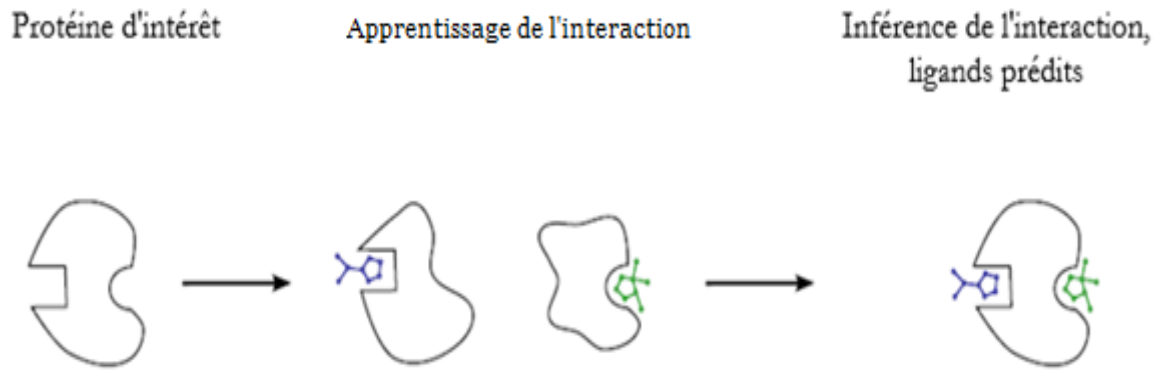


Figure 2-1: Apprentissage et inférence de l'interaction appliquée à la problématique de la prédiction de cibles et annotation fonctionnelle

Les méthodes expérimentales de détection des interactions protéine-protéine ont généré un nombre considérable d'informations d'interaction. Ainsi, la plupart des travaux utilisent des algorithmes d'apprentissage supervisé pour l'inférence de l'interaction [You et al. 2013].

2.1.2. Apprentissage supervisé et classification

Réaliser un apprentissage supervisé revient à fournir à la machine des données étiquetées ou labellisées. Par exemple pour les interactions protéine-protéine, les étiquettes sont : 'il y'a interaction' et 'il n'y a pas d'interaction'. Soit des observations $x_i \, \epsilon \, X$, associées à des sorties (éventuellement des labelles ou étiquettes) $y_i \, \epsilon \, Y$, avec $1 \leq i \leq n$. On considère que les n

couples (x_i, y_i) ont été générées d'après une distribution de probabilité jointe P sur $X \times Y$ et qu'ils sont indépendants. L'ensemble $Z = \{(x_i, y_i)_{1 \leq i \leq n}\}$ des couples entrée-sortie est appelé ensemble d'apprentissage [Yamanishi et al. 2004 ; Lemberger et al. 2015].

Le but d'un algorithme d'apprentissage supervisé est donc d'utiliser l'ensemble d'apprentissage Z afin d'apprendre une fonction $f: X \to Y$, qui soit capable de prédire correctement la sortie *y* associée à une nouvelle entrée *x*. Lorsque l'ensemble des valeurs de sorti est fini, on parle alors d'un problème de classification, qui revient à attribuer une étiquette à chaque entrée et la fonction de prédiction est donc appelée un classifieur. Le cas présent de la prédiction des interactions est un problème de classification car nous avons deux ensembles de valeurs de sorti : soit il y'a interaction ou soit il n'y a pas d'interaction.

Dans la suite, nous présentons les machines à vecteurs de support ou SVM (*Support Vector Machine*) et les réseaux de neurones artificiels (RNA), deux des algorithmes d'apprentissage automatique beaucoup utilisés pour la prédiction d'interaction protéine-protéine [Cai et al. 2001 ; Guo et al. 2008 ; You et al. 2014 ; Du et al. 2017 ; Ma et al. 2020].

2.1.3. Machines à vecteurs de support et réseaux de neurones artificiels

2.1.3.1. *Machines à vecteurs de support*

Appartenant à la classe de l'apprentissage supervisé et des méthodes à noyaux, les machines à vecteurs de supports ou SVM sont un algorithme à l'origine conçu pour la classification binaire ($y_i \in \{-1, +1\}$) [Aitchison et Aitken 1976; Cortes et Vapnik 1995]. Pour deux classes (ou deux groupes) d'exemples donnés, l'objectif du SVM est de construire une bande séparatrice (marge) non linéaire de largeur maximale qui sépare au mieux les deux classes. Il peut être vu comme un problème de recherche d'un hyperplan (ou frontière) d'équation $f(x) + b = 0$ permettant de séparer les exemples positifs des exemples négatifs. Nous soulignons que, les points les plus proches de part et d'autre de l'hyperplan sont appelés vecteurs de support comme sur la figure 2-2, où les exemples négatifs sont représentés par les carrés orange et les exemples positifs sont représentés par les carrés vert. Une description complète de la théorie des SVM pour la classification ou la reconnaissance des formes se trouve dans le livre de Vapnik [Cortes et Vapnik 1995].

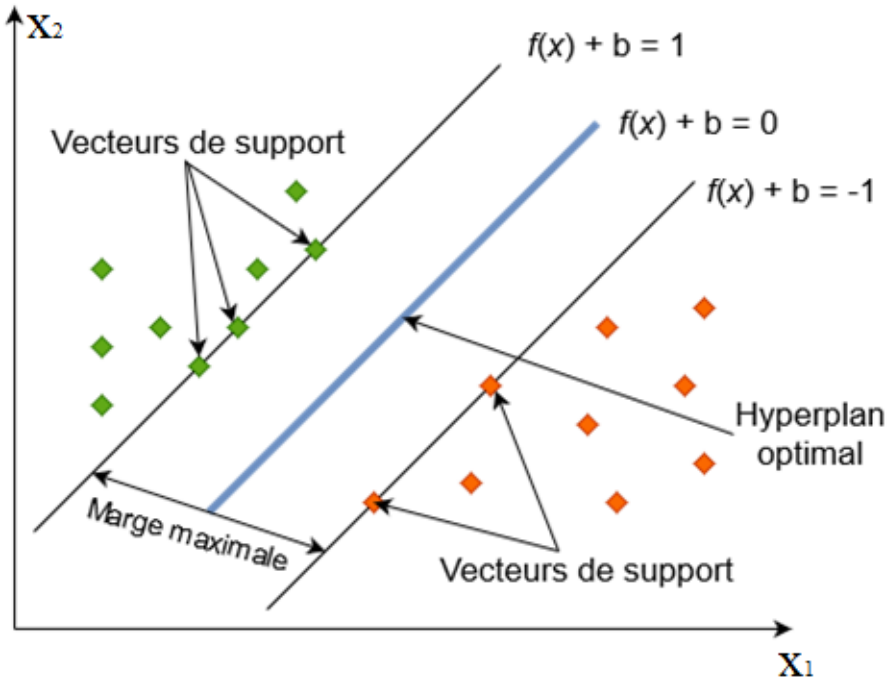


Figure 2-2: **SVM à marge douce**.

Apprentissage avec les SVM

L'idée de base de l'utilisation du modèle SVM pour la classification peut être énoncée brièvement comme suit. Tout d'abord, la mise en correspondance des données d'origine *X* dans un espace de caractéristiques $\mathcal{G}$ de haute dimensionnalité par une fonction de mise en correspondance linéaire ou non linéaire, qui est pertinente pour la sélection d'une fonction noyau. Ensuite, dans l'espace des caractéristiques de la première étape, on cherche une division linéaire optimisée, c'est-à-dire qu'on construit un hyperplan qui sépare les données en deux classes.

Considérons un ensemble de données d'apprentissage de paires instance-étiquette $(x_i, y_i), i \in [1, n]$ où à chaque vecteur d'entrée $x \in \mathbb{R}^p$ est associé une valeur de sortie $y \in \{-1, +1\}$. L'idée est alors de construire une fonction g (fonction discriminante) qui au vecteur d'entrée x fait correspondre la sortie $y = g(x)$. La fonction de décision de classification mise en œuvre par le SVM est représentée par l'équation 2-1 :

$$g(x) = signe[\sum_{i=1}^{n} y_i \theta_i \cdot K(x, x_i) + b] \qquad (2\text{-}1)$$

où les coefficients θ_i sont obtenus en résolvant le problème convexe de programmation quadratique suivant [Wang et Zhong 2014] :

$$Maximiser \qquad \sum_{i=1}^{n} \theta_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \theta_i \theta_j \cdot y_i y_j \cdot K(x, x_j)$$

$$sous\ réserve\ de \qquad 0 \leq \theta_i \leq \mathcal{C}, \ \ i = 1, .. n$$

$$\sum_{i=1}^{n} \theta_i y_i = 0$$

Dans l'équation 2-1, $\mathcal{C}$ est un paramètre de régularisation qui contrôle le compromis entre la marge et l'erreur de mauvais classement. Les vecteurs x_j ne sont appelés vecteurs de support que si les θ_j correspondants sont supérieurs à zéro.

2.1.3.2. *Réseaux de neurones artificiels*

Les réseaux de neurones artificiels (RNA) sont un modèle d'apprentissage basé sur le fonctionnement du neurone biologique. Ils sont constitués de plusieurs neurones artificiels repartis dans des couches. Ces couches sont classées en trois catégories : entrée, cachée et sortie (voir figure 2-3). Le nombre de neurones dans la couche d'entrée correspond au nombre de variables d'entrée dans les données traitées. En outre, chaque réseau possède une seule couche d'entrée et de sortie. Chaque neurone de la couche *L* est connecté à chaque neurone dans la couche *L*+1 et chaque neurone combine ses entrées pour produire une fonction *Z* puis applique une fonction d'activation *f* sur *Z* pour avoir une sortie $\hat{y} = f(z)$. Nous pouvons avoir plusieurs architectures RNA. La figure 2-3 par exemple présente une architecture perceptron multicouche ou MLP (*Multi-Layer Perceptron*) avec une couche d'entrée, deux couches cachées ($L_1\ et\ L_2$) et une couche de sortie formée par deux neurones. Toutefois, le modèle le plus simple pour illustrer les réseaux de neurones artificiels est le perceptron simple couche ou neurone formel [Gardner and Dorling 1998; Wira 2009; Daudt et al. 2018].

Nous soulignons que l'algorithme RNA peut être utilisé dans un cadre supervisé tout comme dans un cadre non supervisé. Dans le cas de l'apprentissage supervisé, l'algorithme RNA s'entraîne sur un ensemble de données étiquetées et se modifie jusqu'à être capable de traiter tout l'ensemble pour obtenir le résultat souhaité. Cependant, dans le cas de l'apprentissage non-supervisé, les données ne sont pas étiquetées. Le réseau de neurones analyse l'ensemble de données, et une fonction coût lui indique dans quelle mesure il est éloigné du résultat souhaité [Sathya et Abraham 2013].

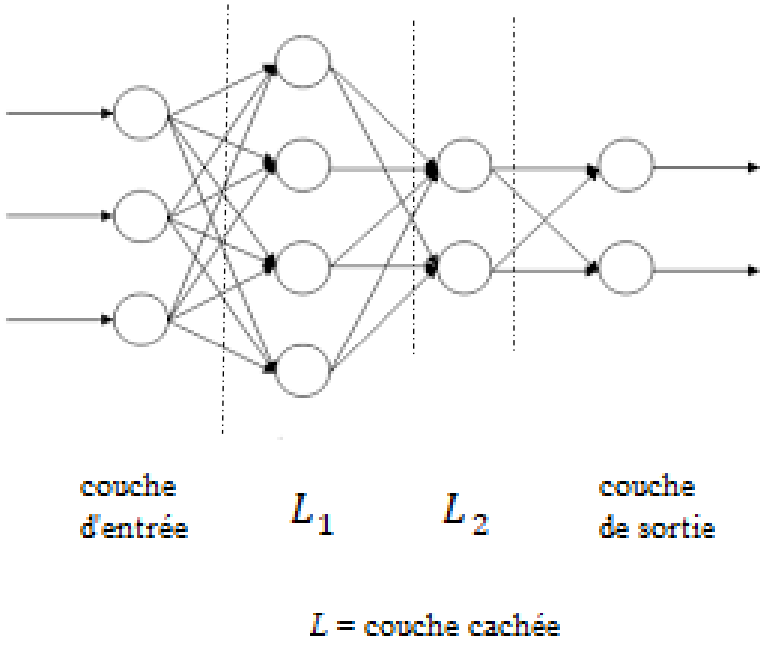


Figure 2-3: Exemple d'architecture d'un MLP

Apprentissage avec les RNA

Dans presque tous les problèmes de réseaux neuronaux (RN), l'apprentissage consiste à trouver les poids ***w*** et les biais ***b*** du réseau qui minimisent une fonction de coût généralement noté ***j***. Étant donné une fonction générique $j(w)$, qui est la fonction de coût, où w est un vecteur de poids, la valeur de w pour laquelle $j(w)$ a un minimum peut être trouvé avec un algorithme basé sur les étapes ci-dessous (algorithme de descente du gradient stochastique [Laura 2015]) :

1. Itération 0 : choisir initialement une estimation aléatoire w_0
2. Itération n + 1 (avec n commençant à 0) : Les poids à l'itération n + 1, w_{n+1} , seront mis à jour à partir des valeurs précédentes à l'itération n, w_n, à partir de la formule suivante :

$$w_{n+1} = w_n - \lambda \nabla j(w_n)$$

où $\nabla j(w_n)$ indique le gradient de la fonction de coût, qui est un vecteur dont les composants sont la dérivée partielle de la fonction de coût et λ représente le coefficient d'apprentissage (sa valeur est généralement comprise entre 0 et 1).

Pour décider quand s'arrêter, nous pouvons vérifier quand la fonction de coût $j(w)$ cesse de trop changer. En général, les gens laissent simplement l'algorithme fonctionner pendant un grand nombre η fixe d'itérations et vérifient les résultats finaux. Si le résultat n'est pas celui escompté, ils augmentent η [Wong et Hsu 2006].

2.1.4. Capacité de généralisation d'une méthode d'apprentissage

La capacité de généralisation d'une méthode d'apprentissage se rapporte à sa capacité de réaliser des taux de précision élevés sur des données de test indépendants. La problématique de l'évaluation de cette performance apparaît sous deux angles dont la sélection de modèles et l'évaluation du modèle [Hastie et al. 2009].

2.1.4.1. *Sélection de modèles*

Une méthode d'apprentissage est caractérisée par différents paramètres. La sélection de modèle revient à choisir les valeurs optimales des paramètres qui influent sur la performance du modèle. Ces paramètres sont appelés hyperparamètres. Plusieurs critères peuvent être utilisés pour faire ce choix, comme la stabilité ou une mesure de performance [Yang et Shami 2020].

2.1.4.2. *Evaluation du modèle*

L'évaluation d'un modèle 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 consiste à diviser les 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. Cependant, si l'on dispose de peu d'observations, l'approche de la validation croisée est privilégiée [Arlot et Celisse 2010]. Cette technique consiste à subdiviser les données en *n* parties égales et procéder par des itérations où chaque partie devra jouer le rôle de données test et de données d'entrainement. L'évaluation de l'erreur finale s'effectue en calculant le score de validation croisée. Il consiste à un calcul de l'erreur par itération en pourcentage puis à moyenner ces chiffres sur toutes les itérations. La limite naturelle de la validation croisée correspond au cas où *k* est égal au nombre d'exemples dans la base d'apprentissage. Il existe plusieurs types de validation croisée dont la "*leave-one-out*" [Kearns and Ron 1999], la "*k-fold*" [Bengio and Gretvalet 2004], la "*leave-v-out*" [Kearns et Ron 1999], et bien d'autres. La "*k-fold*" ou validation croisée *k*-fois (VCK) est la plus utilisée dans la prédiction d'interaction basée sur les séquences [Wang et al. 2018]. La figure 2-4 illustre le principe de la VCK où les parties grisées étant considérées pour la phase des tests pendant que les autres parties sont pour l'apprentissage. Le premier sous-ensemble (P_1) sera utilisé ici comme ensemble de validation et le reste, c'est-à-dire les $(k-1)$ autres sous-ensembles $(P_2, P_3, P_4, \dots, P_k)$, est

pour l'ensemble d'apprentissage. La fonction de prédiction est ensuite entrainée à l'aide de cet ensemble de données et un score de précision ou de perte est calculé. Ensuite, cet apprentissage est répété (*k*-1) fois mais en utilisant à chaque fois un sous-ensemble différent pour l'ensemble de validation. Le score de validation après cette procédure s'obtient en faisant la moyenne des scores de tous les apprentissages.

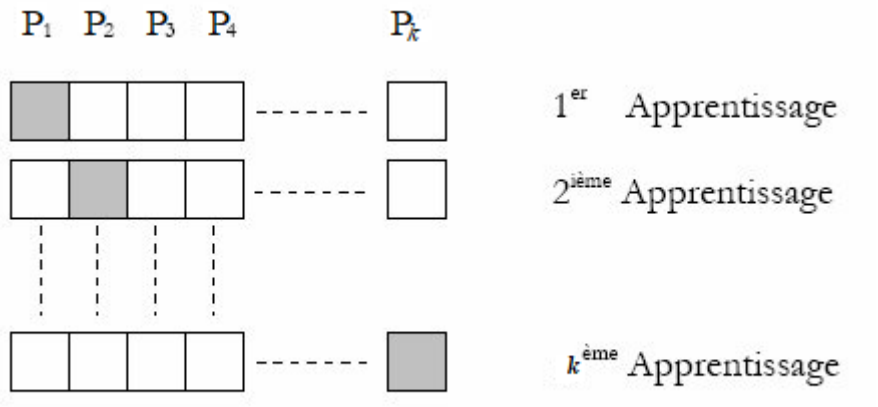


Figure 2-4: Principe de la validation croisée *k*-fois

2.2. Modèles d'extraction de caractéristiques à partir des données de séquences

Pour utiliser avec succès les méthodes d'apprentissage supervisé pour prédire les interactions protéine-protéine à partir de séquences de protéines, l'un des plus importants défis informatiques est de savoir comment représenter efficacement les données essentielles de la séquence. Plusieurs techniques de représentation des données à partir des informations de la séquence d'acides aminés dans le cas de la prédiction d'interactions entre les protéines ont été développées. Ces techniques peuvent être regroupées en deux types de modèles. L'un est le modèle séquentiel et l'autre le modèle discret [Chou 2011]. Dans la suite de cette section nous présentons ces deux types de modèles avec quelques approches de ces modèles.

2.2.1. Modèle séquentiel

Le modèle séquentiel représente une protéine par une série d'acides aminés selon l'ordre de leurs positions dans la chaîne de protéine. Par conséquent, ce modèle peut naturellement refléter toutes les informations sur l'ordre et la longueur de la séquence d'une protéine. Le modèle séquentiel le plus simple pour un échantillon de protéine est telle qu'exprimée dans l'équation 2-2 :

$$P = R_1R_2R_3R_4R_5 \dots R_L \quad (2\text{-}2)$$

où R_1 représente le 1er résidu (ou acide aminé) de la séquence d'une protéine P, R_2 le 2ème résidu, …, R_{L-1} le $(L\text{-}1)^{\text{ème}}$ résidu et R_L le $L^{\text{ème}}$ résidu, et chacun des résidus appartenant à l'un des 20 types d'acides aminés vus au chapitre précédent. L'approche des *Triade Conjointe* [Shen et al. 2007] est un modèle séquentiel qui utilise les données de classification des acides aminés. En effet, sur la base des propriétés de dipôles et des volumes des chaînes latérales, les 20 acides aminés peuvent être classés en sept groupes (voir tableau 2-1). Les acides aminés d'une même classe impliquant probablement des mutations semblables en raison de leurs caractéristiques similaires [Shen et al. 2007]. Ainsi plusieurs approches d'extraction de caractéristiques se basent sur les informations de groupe d'acides aminés pour représenter la séquence [Shen et al. 2007 ; Pan et al. 2010 ; Göktepe and Kodaz 2018].

Tableau 2-1: Classification des acides aminés en fonction de leurs dipôles et volumes des chaînes latérales

Acide aminé	Groupe
A, G, V	1
I, L, F, P	2
Y, M, T, S	3
H, N, Q, W	4
R, K	5
D, E	6
C	7

Dans ces approches, chaque acide aminé de la séquence de protéine définie à l'équation 2-2 est remplacée par son numéro de groupe. Par exemple la séquence $P = MLVASANFGD$ est remplacé par $P = 3211314216$.

Triade conjointe

La triade conjointe (TC) prend en compte les propriétés d'un acide aminé et de ses acides aminés voisins et considère les trois acides aminés continus comme une unité. Ainsi, les triades peuvent être différenciées en fonction des groupes d'acides aminés, c'est-à-dire que les triades composées de trois acides aminés appartenant aux mêmes groupes, telles que ART et VKS, peuvent être traitées comme des triades conjointes sur la base qu'elles jouent des rôles

similaires lors du traitement de l'interaction. En considérant que 3 acides aminés consécutifs forment un groupe, 343 combinaisons différentes de groupes peuvent être vues. Par conséquent, un vecteur de 343 composantes peut être obtenu à l'aide des fréquences de chaque groupe. Pour l'échantillon de séquence protéique "*P*" donné ci-dessus, la procédure de comptage de fréquence est illustrée à la figure 2-5 où f_{10} correspond à la triade formée des groupes 3, 2 et 1, f_2 formé des groupe 2, 1 et 1, etc.

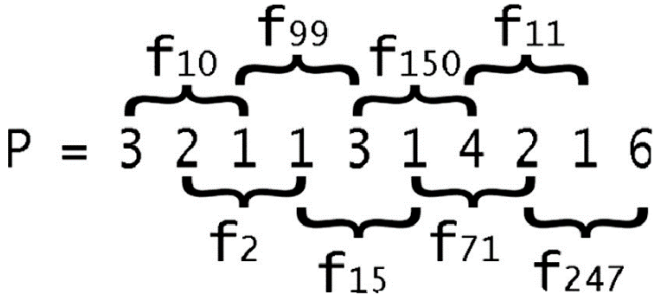


Figure 2-5: Procédure de comptage des fréquences des triades

Certains travaux récents tels que les travaux de Shin et al. [2017] et ceux de Göktepe et Kodaz [2018] ont utilisé l'approche des triades conjointes pour extraire les caractéristiques des paires de protéines sur des ensembles de données IPP HPRD, puis ont appliqué un modèle de SVM pour prédire les IPP. Ils ont obtenu un taux de justesse (*accuracy*) respectivement de 73,81% et 83,9%.

Selon certains auteurs comme [Chou 2000, 2001], les approches séquentielles ne prennent pas en compte la composition de la séquence en acide aminé et les relations existantes entre les différents acides aminés. Selon ces mêmes auteurs, ces informations sont pertinentes pour un apprentissage efficace. Face aux limites des modèles séquentiels, des modèles discrets furent proposés.

2.2.2. Modèle discret

Le modèle discret représente une protéine par un ensemble de nombres discrets ou un vecteur à dimensions multiples. Le modèle discret le plus simple pour représenter les données de la séquence d'acides aminés d'une protéine est sa composition en acide aminé AAC (*Amino Acid Composition*) [Nakashima et al. 1986]. Selon le modèle AAC, la protéine *P* de l'équation 2-2 peut s'écrire de la manière suivante :

$$P = [f_1, f_2, \dots, f_{20}]^T \quad (2\text{-}3)$$

où $f_1, f_2, \dots, f_{20}$ sont les occurrences de fréquences normalisées des 20 acides aminés dans *P* et *T*, l'opération de transposition. Certaines méthodes basées sur le modèle AAC ont été proposées pour encoder la séquence d'acides aminés d'une protéine. Cependant, comme le montre l'équation 2-3, la méthode AAC ne tient pas compte des effets d'ordre de la séquence et donc la qualité de la prédiction ainsi obtenue pourrait être limitée [Chou 2000, 2001, 2005]. Pour éviter de perdre complètement l'information sur les effets d'ordre de la séquence, un modèle discret complètement différent, appelé modèle PseAAC (Pseudo AAC), a été proposé pour les représenter et est formulé selon l'équation 2-4 [Chou 2001] :

$$P = [f_1, f_2, \dots, f_{20}, P_{20+1}, \dots, P_{20+\delta}]^T \quad (2\text{-}4)$$

où les 20 premiers éléments sont associés aux 20 éléments de l'équation 2-2 (qui sont les 20 composantes d'acides aminés de la séquence), et les δ facteurs supplémentaires sont utilisés pour intégrer des informations d'ordre séquentiel sur des niveaux variables comme indiqué sur la figure 2-8. Les valeurs d'intersection $J_{1,2}, J_{1,3}, \dots, J_{2,3}, \dots$ indiquent les relations ordre-séquence d'acides aminés d'une protéine. Le 1er niveau ($J_{1,2}, J_{2,3}, J_{3,4}, \dots$) correspond à l'effet ordre séquence entre un acide aminé à la position *i* et l'acide aminé contigu à la position *i+1*. Ici par exemple $J_{1,2}$ indique l'effet ordre-séquence entre les acides aminés R_1 et R_2. Le 2ème niveau ($J_{1,3}, J_{2,4}, J_{3,5}, \dots$) exprime l'ordre séquence entre un acide aminé et son 2ème voisin qui suit. $J_{1,3}$ indique donc l'effet ordre-séquence entre les acides aminés R_1 et R_3. Le 3ème niveau ($J_{1,4}, J_{2,5}, J_{3,6}, \dots$) est l'ordre séquence entre un acide aminé et son 3ème voisin qui suit.

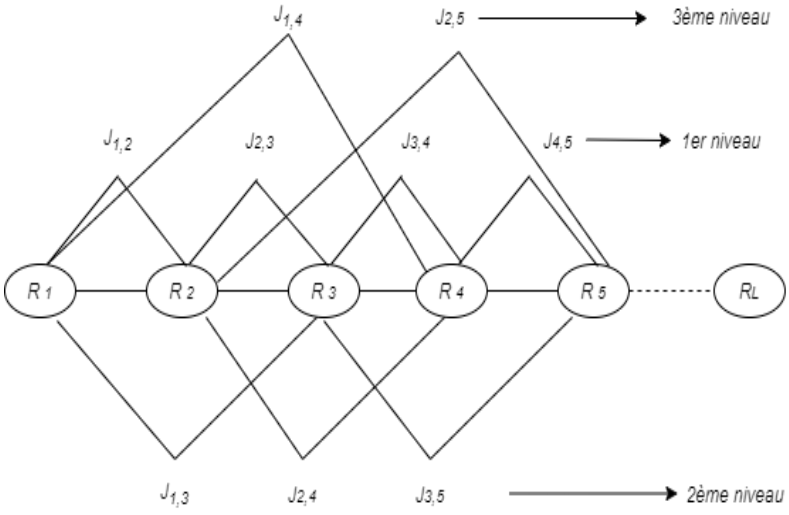


Figure 2-6: Différents niveaux ordre-séquence [Chou 2001]

Plusieurs approches utilisent le modèle discret pour extraire les informations de la séquence et deux d'entre elles, les plus utilisés, sont décrites ci-dessous [Guo et al. 2008 ; Ghanty and Pal 2009 ; Huang et al. 2016].

2.2.2.1. *Approches basées sur des propriétés physicochimiques des acides aminés*

Comme nous l'avons vu dans le chapitre 1, les acides aminés possèdent plusieurs propriétés physicochimiques. L'interaction protéine-protéine peut être définie selon quatre modes d'interaction [You et al. 2014]. Ces quatre modes sont : interaction électrostatique, interaction hydrophobe, interaction stérique et liaison hydrogène. Sept propriétés physicochimiques ont été sélectionnées pour refleter ces modes d'interaction chaque fois que possible. Nous avons l'hydrophobicité (H_1), l'hydrophilie (H_2), les volumes des chaînes latérales d'acides aminés (M), la polarité (P_1), la polarisabilité (P_2), la surface accessible aux solvants ($SASA$) et l'indice de charge net des chaînes latérales d'acides aminés (NCI) [Tanford 1962 ; Grantham 1974; Krigbaum et Komoriya 1979 ; Hopp et Woods 1981; Charton et Charton 1982; Rose et al. 1985 ; Peng et al. 2006]. Les valeurs originales des sept propriétés physicochimiques de chaque acide aminé sont indiquées dans le tableau 2-2.

Tableau 2-2: Valeurs originales des sept descripteurs physicochimiques des acides aminés [You, Yu, et al. 2014]

Acide aminé	H_1	H_2	VSC	P_1	P_2	$(SASA)$	NCI
A	0.62	-0.5	27.5	8.1	0.046	1.81	0.007187
C	0.29	-1.0	44.6	5.5	0.128	1.461	-0.03661
D	-0.9	3.0	40	13	0.105	1.587	-0.02382
E	-0.74	3.0	62	12.3	0.151	1.862	0.006802
F	1.19	-2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0.0	0.0	9.0	0.0	0.881	0.179082
H	-0.4	-0.5	79	10.4	0.23	2.025	-0.01062
I	1.38	-1.8	93.5	5.2	0.186	1.81	0.021631
K	-1.5	3.0	100	11.3	0.219	2.258	0.017708
L	1.06	-1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	-1.3	94.1	5.7	0.221	2.034	0.002683
N	-0.78	0.2	58.7	11.6	0.134	1.655	0.005392

P	.012	0.0	41.9	8	0.131	1.468	0.239531
Q	-0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	-2.53	3.0	105	10.5	0.291	2.56	0.043587
S	-0.18	0.3	29.3	9.2	0.062	1.298	0.004627
T	-0.05	-0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	-1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	-3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	-2.3	117.3	6.2	0.298	2.368	0.023599

Pour extraire des caractéristiques discriminantes et bien décrire la séquence d'acides aminés d'une protéine donnée, les auteurs utilisent différentes techniques statistiques basées sur une mesure de distance ou un scalaire à partir des différentes valeurs des différentes propriétés physicochimiques [Chou 2001, 2005; Guo et al. 2008].

2.2.2.2. *Approches basées sur le text mining*

La séquence d'acides aminés d'une protéine peut être traitée comme une chaîne de texte où des informations cachées sont déchiffrées par la mise en œuvre de techniques NLP (*Natural Langage Processing*) [Kobayashi et Aono 2004; Vyas et al. 2016 ; Yao et al. 2019]. De ce fait les techniques NLP sont utilisées pour représenter les séquences d'acides aminés des protéines, précisément la technique Word2Vec et la technique *N*-gramme.

Word2Vec

Word2vec est une technique très efficace pour l'apprentissage de la représentation de mots de manière non supervisée [Church 2017]. Tsubaki et al. [2017] ont adopté Word2vec pour apprendre des représentations de caractéristiques à partir d'une base de données de protéines afin de résoudre le problème de la reconnaissance des 'plis' de protéines.

Il existe deux modèles disponibles dans Word2vec : le modèle de sac continu de mots (CBOW) et le modèle de saut de gramme (SG). Selon des études antérieures, le modèle SG a l'avantage de créer des représentations vectorielles de meilleure qualité [Zohra et al. 2018 ; Smaili et al. 2019]. Yao et al. [2019] ont proposé l'outil Res2Vec basé sur le modèle SG de Word2Vec pour représenter les attributs de la séquence de protéines. Pour une séquence protéique, l'outil Res2vec transforme dans un premier temps chaque résidu d'acide aminé en

un vecteur propre de dimension fixe. Ensuite, les différents vecteurs des différents résidus sont concaténés et deviennent un seul vecteur. Ici chaque résidu est traité comme un mot pendant que la séquence protéine est traitée comme une phrase. Soit la séquence protéique $MKPGA$, la figure 2-9 montre les différentes étapes de codage de la séquence avec en (A) le processus de traitement de chaque résidu en vecteur de dimension fixe (2 dans l'exemple) et en (B) la transformation des vecteurs des résidus en un seul vecteur. Les résidus M, K, P, G, A produisent donc le vecteur [2.96,1.03,3.19,0.67,2.94,1.18,3.19,0.67,2.89,1.23].

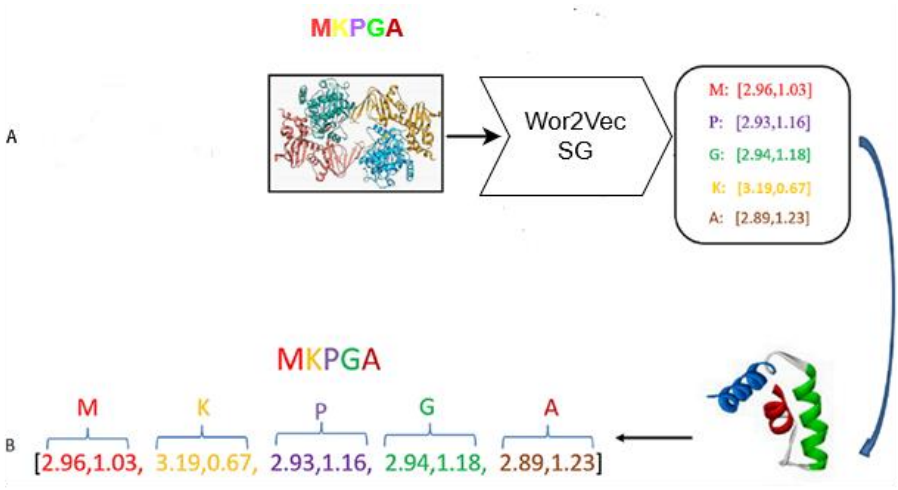


Figure 2-7: Apprentissage de la représentation de résidu par l'outil Res2Vec (adapté de [Yao et al. 2019])

Approche *N*-gramme

L'approche *N*-gramme (ou *N*-fréquence) consiste à utiliser un sous-ensemble de N éléments construits à partir d'un ensemble de données [Cai et al. 2001 ; Dehzangi et al. 2017]. En effet, le principe de cette approche est de construire une distribution de probabilité pour la prochaine lettre avec un historique de taille N à partir d'un corpus d'apprentissage. Cette modélisation est semblable à une chaine de Markov d'ordre N où sa matrice des transitions n'est pas connue et seule les N dernières observations sont utilisées pour la prédiction de la lettre suivante [Almagor 1983; Blaisdell 1985; Liefooghe 2008].

Un modèle de Markov d'ordre k suppose les variables aléatoires représentatives des lettres du texte T_i dépendantes des k variables précédentes. Un tel modèle est donc défini par les probabilités conditionnelles suivantes pour tout $u_{i-k} \dots u_i \in \Sigma$.

$$\mathbb{P}(u_i | u_{i-k} \dots u_{i-1})$$

Le paramètre de ce modèle est sa matrice des transitions $|\Sigma|^k \times |\Sigma|$ notée $\prod$ telle que pour toute lettre $l \in \Sigma$ et tout mot $u \in \Sigma^k$, l'élément d'indices l et u de $\prod$ notée $\prod_{u,l}$ est défini par

$$\mathbb{P}(l|u)$$

Lorsque la matrice de transitions n'est pas connue, ses éléments $\prod_{u,l}$ peuvent être estimés par la fréquence du mot *ul* relativement à la fréquence du mot *u* d'un texte observé. La probabilité de génération d'un mot $u = u_0 \ldots u_{L-1}$ selon le modèle de Markov *M* se calcule par la suite de multiplications suivantes.

$$\mathbb{P}(u|M) = \left[\sum_{v \in \Sigma^k} \prod_{v,u_0}\right] \times \left[\sum_{v \in \Sigma^{k-1}} \prod_{vu_0u_1}\right] \times \ldots \times \prod_{S_{L-k-1,\ldots,L-2,L-1}}$$

Une séquence de protéine peut donc être décrite comme un processus de Markov où les acides aminés sont les résultats d'un générateur de séquences. Soit $[\mathbb{P}_{j|i}]$ la matrice de probabilité de transition sur laquelle le générateur de séquence est basé. Soit $\mathbb{P}_{ij}$, la probabilité que le doublet *i,j* (*i,j* = Σ) soit généré et $\mathbb{P}_i$, la probabilité pour l'acide ainé *i*, les termes :

$$\mathbb{P}_i = \sum_{j=1}^{|\Sigma|} \mathbb{P}_{ij}$$

et

$$\mathbb{P}_{j|i} = \mathbb{P}_{ij}|\mathbb{P}_i$$

sont liés pour la probabilité conditionnelle de l'acide aminé j d'être généré juste à côté de l'acide aminé *i* le long de la chaîne. Ce générateur générera des chaînes dans lesquelles l'apparition d'un certain acide aminé à n'importe quelle étape ne dépend que de l'acide aminé de l'étape précédente. Les acides aminés seront corrélés au plus proche voisin. Ceci correspond à des corrélations de premier ordre (voir tableau 2-3). Un tel générateur est appelé source de Markov du premier ordre, et la chaîne qu'il génère est une chaîne de Markov du premier ordre [Almagor 1983].

Si l'on représente une séquence par la technique 2-gramme [Cavnar and Trenkle 1994], le vecteur produit aura |Σ|×|Σ| composantes, soit $20 \times 20 = 400$. Pour le 3-gramme, on aura $20 \times 20 \times 20 = 8000$ composantes. La technique 2-gramme appliquée sur la séquence primaire est appelée technique bigramme et permet ainsi d'extraire les fréquences de deux

acides aminés dans une séquence [Li et al. 2005 ; Sharma et al. 2013 ; Hayat et al. 2014 ; Dehzangi et al. 2017].

Tableau 2-3: Expressions de probabilité pour le générateur [Almagor 1983]

	Générateur d'ordre zéro	**Générateur d'ordre un**
$\boldsymbol{i} \in \boldsymbol{\Sigma}$	$\mathbb{P}_i$	$\mathbb{P}_i = \sum_{j=1}^{\lvert\Sigma\rvert} \mathbb{P}_{ij}$
Bigramme (doublet)	$\mathbb{P}_{i,j} = \mathbb{P}_i \mathbb{P}_j$	$\mathbb{P}_{i,j} = \mathbb{P}_i \mathbb{P}_{j\|i}$
Trigramme (Triplet)	$\mathbb{P}_{i,j,k} = \mathbb{P}_i \mathbb{P}_j \mathbb{P}_k$	$\mathbb{P}_{i,j,k} = \mathbb{P}_{ij} \mathbb{P}_{k\|j}$

2.3. Techniques d'extraction des caractéristiques bigrammes

Dans le domaine de la prédiction d'interactions entre les protéines, une des techniques *N*-gramme [Cavnar and Trenkle 1994] beaucoup utilisée est le 2-gramme (N = 2), appelée également technique bigramme (voir tableau 2-3) [Li et al. 2005]. Cette technique permet d'extraire les caractéristiques bigrammes dans une séquence de protéine. Un bigramme ici un ensemble de deux acides aminés successifs. Par exemple *AA* ou *AC* sont des bigrammes. Dans cette thèse, nous nous sommes particulièrement intéressés aux caractéristiques bigrammes. Dans la suite de cette section, nous présentons les motivations de l'intérêt des caractéristiques bigrammes puis nous montrons les techniques existantes pour extraire ces caractéristiques avec leurs limites.

2.3.1. Nécessité d'extraire les caractéristiques bigrammes

Plusieurs chercheurs selon [Almagor 1983] ont suggéré que certaines caractéristiques biologiques ou chimiques de base des acides aminés peuvent être exprimées par les bigrammes de la séquence d'une protéine. Comme caractéristiques biologiques exprimées nous avons les indices de séquence d'ARNm qui fournissent des indications sur les contraintes moléculaires existantes au cours de l'évolution des gènes [Grantham 1974]. Nous avons également la reconnaissance des 'plis' des protéines qui est une simulation du repliement des protéines [Sali et al. 1994 ; Bushmarina et al. 2005].

Le repliement est le passage d'une chaîne d'acides aminés (structure primaire) vers une structure tridimensionnelle (tertiaire) bien définie et permettant à la protéine d'exercer sa fonction biologique. Pour être biologiquement actives, toutes les protéines doivent adopter des structures tridimensionnelles pliées spécifiques [Dill et al. 2008]. La propriété chimique du repliement fait l'objet de nombreuses études bio-informatiques et constitue une étape importante dans la réalisation de la fonction des protéines [Dill et al. 2008; Tsubaki et al. 2017]. En outre, selon certaines études récentes [Su et al. 2019 ; Shao et al. 2021], la reconnaissance des 'plis' des protéines est l'une des techniques clés pour l'étude des structures et des fonctions des protéines et la conception de médicaments. En particulier, elle joue un rôle clé dans la prédiction des structures protéiques associées à la COVID-19 [Afify et Zanaty 2021]. Or justement selon certains auteurs comme [Sharma et al. 2013], les caractéristiques bigrammes peuvent représenter les points de 'plis' de la séquence d'acides aminés. Par conséquent les caractéristiques bigrammes simulent bien le repliement des protéines et l'interaction protéine-protéine.

Il existe dans la littérature deux techniques permettant d'extraire les caractéristiques bigrammes à partir de la séquence de protéine. Nous avons la technique *Pairwize Frequency* [Ghanty et Pal 2009] et la technique qui utilise une matrice de scores spécifiques à la position (PSSM) [Sharma et al. 2013] que nous notons dans la suite *Bi-gram*. Ces deux techniques sont détaillées ci-dessous.

2.3.2. Technique *Pairwize Frequency*

Les auteurs Ghanty et Pal [2009] ont proposé la technique PF (*Pairwize Frequency*) pour extraire les caractéristiques bigrammes. Dans la technique PF, Ils ont calculé ces caractéristiques en comptant les fréquences d'occurrences bigrammes de la séquence d'acides aminés représentant la structure primaire d'une protéine donnée. Par exemple si nous prenons la séquence exemple *AACEAAI*, la technique PF compte le nombre de fois d'observer un couple d'acides aminés comme exprimé à l'équation 2-5 :

$$\begin{cases} AA = 2 \\ AC = 1 \\ AE = 0 \\ \dots = \cdots \\ CA = 0 \\ CC = \cdots \\ \dots = \cdots \\ YW = 0 \\ YY = 0 \end{cases} \quad (2\text{-}5)$$

Puisque toutes les séquences primaires de protéines sont composées de 20 acides aminés, il y aura donc 400 couples différents d'acides aminés. Le vecteur ainsi calculé comporte 400 composantes bigrammes pour une protéine donnée. Cependant, le nombre d'acides aminés dans une séquence de protéine est limité car tous les acides aminés ne sont pas toujours représentés dans la séquence. Ici par exemple, nous n'avons pas les acides aminés tels que $Y, V, G, W, F, \ldots$ En outre, la dimensionnalité du vecteur obtenu est comparativement grande (400). Par conséquent, de nombreuses composantes du vecteur calculé deviennent égales à zéro.

Limite de la technique PF

La technique PF extrait les caractéristiques bigrammes en comptant la fréquence des différents bigrammes à partir de la structure primaire de la protéine. Cependant, le vecteur de caractéristiques bigrammes obtenu à partir de cette technique est strictement parcimonieux [Sbai 2012], c'est-à-dire la majorité des composantes sont nulles. Un tel vecteur n'apporte pas suffisamment d'informations au classifieur pour inférer correctement les interactions entre les protéines. En effet, la nature statistique du modèle *N*-gramme pose ainsi le problème de la représentativité des données fournies au système pour le calcul des probabilités. Ce problème est plus connu sous le nom d'éparpillement des données [Schadle et al. 2001] où un grand nombre de données est concentré sur un nombre de cas limité. Ce qui veut dire encore qu'un grand nombre d'acides aminés ne sont pas observés et se voient affectés d'une probabilité nulle ou d'une fréquence nulle. Or un vecteur où la plupart des composantes sont nulles, est un vecteur strictement parcimonieux [Sbai 2012]. Un tel vecteur selon [Sharma et al. 2013] ne permet pas au classifieur utilisé pour l'inférence, de réaliser de bonnes performances de prédiction et de présenter une grande capacité de généralisation. Ainsi, l'application de la technique 2-gramme sur la séquence primaire pour représenter la protéine n'est pas un moyen efficace de capturer l'information. Par conséquent, les performances de classification devraient être faibles.

2.3.3. Technique utilisant la matrice de scores spécifiques à la position

Face au problème de vecteur strictement parcimonieux dans la technique PF, l'application de la technique 2-gramme sur la matrice de scores spécifiques à la position notée PSSM (*Position Specific Score Matrix*) [cheol Jeong et al. 2010] fut proposée par Sharma [2013]. Dans cette technique que nous notons *Bi-gram*, le vecteur assorti est calculé en comptant les

fréquences d'occurrences bigrammes obtenues à partir de la PSSM. Etant donné que la PSSM fournit des informations sur la probabilité de 20 acides aminés à chaque emplacement de la séquence d'acides aminés d'une protéine, les composantes nulles dans le vecteur résultant sont évitées. Généralement les auteurs utilisent l'outil PSI-BLAST (*Position-Specific Iterated Basic Local Alignment Search Tool*) [Altschul et al. 1997] accessible via https://blast.ncbi.nlm.nih.gov/Blast.cgi pour le calcul de scores PSSM. La PSSM est une technique basée sur l'alignement de séquences. Dans la suite, nous présentons la matrice PSSM et le mode d'obtention des scores PSSM avec l'outil PSI-BLAST tout en présentant d'abord le concept de l'alignement de séquences.

2.3.3.1. *Alignement de séquences*

L'alignement est un processus par lequel deux ou plusieurs séquences sont comparées afin d'obtenir le plus de correspondances (identités ou substitutions conservatives) possibles entre les lettres qui les composent [Parry-Smith et Attwood, 1991]. Nous énumérons ci-dessous certains termes utilisés dans le cadre de l'alignement de séquences.

- **Similarité de séquences** [Nikhila et Nair 2018]

La similarité correspond au pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Sur la figure 2-8, la partie (a) représente un cas d'identité où les lettres *A*, *C* et *G* se retrouvent sur deux séquences différentes. En revanche sur la partie (b) nous avons un cas de substitution où la lettre *A* est remplacée par la lettre *G*.

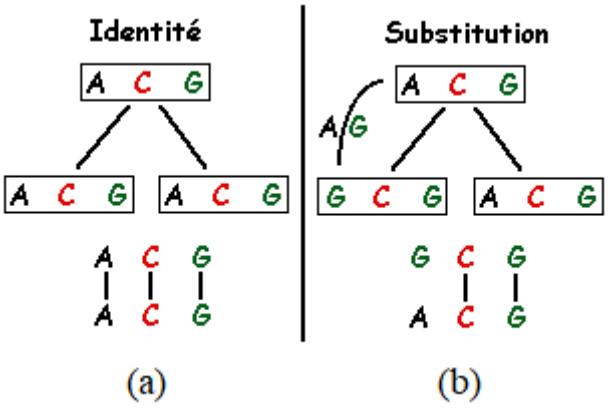


Figure 2-8 : Similarité par Identité et Substitution

- **Homologie de séquences**

L'homologie est généralement mise en évidence par la comparaison de séquences, quand l'ancêtre commun est encore assez proche (voir chapitre précédent). Cependant, si les

séquences ont trop divergé, la comparaison de séquences ne permet pas de détecter l'homologie.

- **Alignement multiple de séquences** [Liefooghe 2008]

On distingue différents types d'alignements de séquences parmi lesquels nous avons l'alignement local, l'alignement global et l'alignement multiple. Lorsque l'alignement concerne une partie de la séquence, l'on parle d'alignement local. Cependant, s'il s'agit de toute la longueur de la séquence on parle d'un alignement global. L'alignement multiple quant à lui correspond à un alignement global de trois séquences ou plus. Il permet de connaître la variabilité en base de chaque position et de différencier les zones à haute similarité, porteuses de sens, des zones observées par hasard. Sur la figure 2-9 par exemple, les acides aminés dans les colonnes en bleu sont conservés sur chaque séquence de l'alignement (haute similarité) et ceux dans les colonnes en vert sont conservés sur certaines séquences.

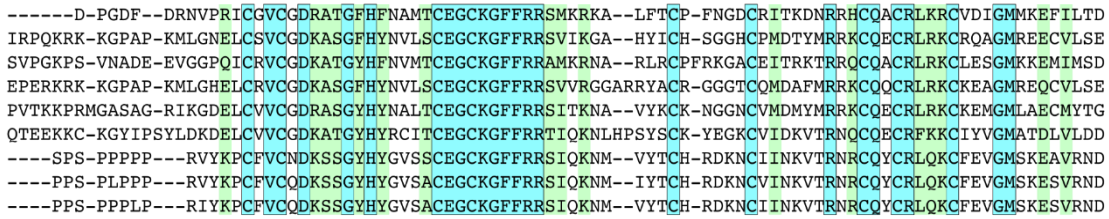


Figure 2-9: Exemple d'alignement de séquences.

2.3.3.2. *Matrice de scores spécifiques à la position ou PSSM*

La modélisation de courtes séquences d'acides aminés à l'aide de matrices est une caractérisation précise des alignements multiples. En effet, si l'on considère les positions indépendantes les unes des autres, les matrices permettent de restituer l'ensemble des informations contenues dans un alignement. La PSSM est l'une des analyses quantifiées utilisées pour l'alignement de séquences multiples [Altschul et al. 2009]. C'est une approche basée sur le profil, c'est-à-dire un tableau des fréquences observées des acides aminés à chaque position dans un alignement multiple (voir figure 2-10). La figure 2-10 illustre un cas de construction d'un profil à partir d'un alignement de 5 séquences.

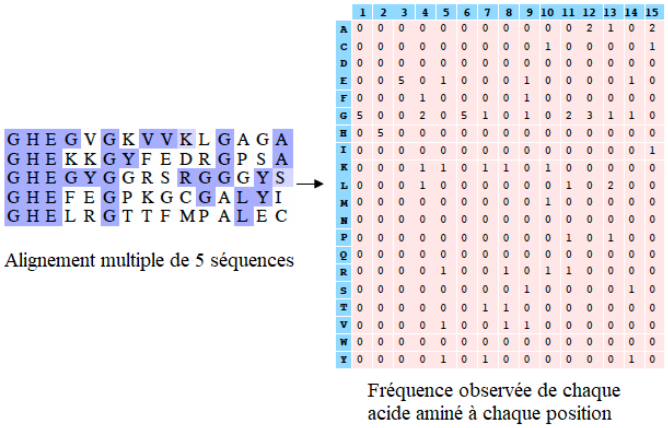


Figure 2-10 : Profil construit à partir d'un alignement de 5 séquences

La construction de ces profils est basée sur la fréquence de chaque acide aminé à une position spécifique de l'alignement multiple. En suivant la figure 2-10, on aura :

- Colonne 1 : $f\ (A, 1)\ =\ 0/5\ =\ 0$; $f\ (G, 1)\ =\ 5/5\ =\ 1$; …
- Colonne 2 : $f\ (A, 2)\ =\ 0/5\ =\ 0$; $f\ (H, 2)\ =\ 5/5\ =\ 1$; …
- …
- Colonne 15 : $f\ (A, 15)\ =\ 2/5\ =\ 0{,}4$; $f\ (C, 15) =\ 1/5\ =\ 0{,}2$; …

Avec *f*, la fréquence. Pour la colonne 1, on lira la fréquence de l'acide aminé *A* à la position 1 est égal à 0/5 = 0 ; fréquence de l'acide G à la position 1 est égal à 5/5 = 1.

Nous pouvons constater que certaines fréquences sont nulles du fait du nombre de séquence dans l'alignement multiple. Une telle fréquence pourrait entrainer une exclusion de l'acide aminé concerné à cette position. Pour contourner cela, une petite valeur appelée '*e-value*' ou 'pseudo-count' est ajoutée à toutes les fréquences observées. La valeur du pseudo-count est toujours comprise entre 0 et 1 [Parry-Smith and Attwood 1991]. Par défaut, la valeur du pseudo count est 0,05. Si nous reprenons les calculs en considérant un pseudo-count de 1, on obtient :

- Colonne 1 : $f'\ (A, 1)\ =\ \frac{0+1}{5+20} =\ 0{,}04$; $f'(G, 1) = \frac{5+1}{5+20} =\ 0{,}24$; …
- Colonne 2 : $f'(A, 2) = \frac{0+1}{5+20} =\ 0{,}04$; $f'(H, 2) = \frac{5+1}{5+20} =\ 0{,}24$; …
- Colonne 15 : $f'(A, 15) = \frac{2+1}{5+20} =\ 0{,}12$; $f'(C, 15) = \frac{1+1}{5+20} =\ 0{,}08$; …

La fréquence de chaque acide aminé déterminée à chaque position est comparée à la fréquence à laquelle chaque acide aminé est attendu dans une séquence au hasard. On fait l'hypothèse que chaque acide aminé est observé avec une fréquence identique dans une séquence au hasard. Le score PSSM est calculé à partir du logarithme du rapport (fréquences observées) / (fréquences attendues) de la manière suivante :

$$S_{ij} = \log\left(\frac{f'_{ij}}{q_i}\right)$$

avec S_{ij}, le score pour l'acide aminé *i* à la position *j*, f'_{ij} est la fréquence relative pour l'acide aminé *i* à la position *j*, corrigée par les pseudo-count et q_i est la fréquence relative attendue pour l'acide aminé *i* dans une séquence au hasard.

2.3.3.3. *Outil PSI-BLAST pour le calcul des scores*

PSI-BLAST est une méthode de recherche de profils de séquences protéiques qui s'appuie sur les alignements générés par une exécution du programme BLAST (*Basic Local Alignment Search Tool*) pour les protéines appelé BLASTP [Delaney et al. 2000]. BLAST est une méthode heuristique de recherche de similarité de séquences, dans laquelle une séquence de protéines ou de nucléotides [Tatusova et Madden 1999] appelée séquence requête est comparée à des séquences de protéines ou de nucléotides dans une base de données cible (voir tableau 2-4) , afin d'identifier les régions d'alignement local et de signaler les alignements dont le score est supérieur à un seuil donné [Altschul et al. 1997]. Le tableau 2-4 nous donne quelques différentes bases de données cibles utilisées pour la comparaison de séquences dans le processus BLAST, une brève description et le nombre de séquences à la date du 22 Février 2022.

Tableau 2-4: Bases de données de comparaison (blast.ncbi.nlm.nih.gov)

Bases de données	Description	Nombre de séquences
nr	Toutes les traductions CDS non redondantes de GenBank+PDB+SwissProt+PIR+PRF, à l'exception des échantillons environnementaux des projets WGS.	> 450 million
RefSeq Select proteins	Cette base de données contient des séquences de protéines RefSeq de NCBI provenant d'humains, de souris et de procaryotes, restreintes à l'ensemble de protéines RefSeq Select.	> 25 million
UniProtKB/SwissProt	Séquences UniProtKB/SwissProt non redondantes.	478091
Reference proteins	Séquences de référence des protéines du NCBI	> 215 million

PSI-BlAST construit un profil à partir de l'alignement multiple des séquences qui ont obtenu les meilleurs scores avec la séquence requête. Ce profil est comparé à la banque cible (banque interrogée) et est affiné au fur et à mesure des itérations. Ainsi, la sensibilité du programme est augmentée. La sensibilité ici c'est l'aptitude à détecter toutes les similarités considérées comme significatives et donc à générer le minimum de faux-négatifs. Un résultat est considéré comme significatif si la probabilité de l'obtenir par hasard est très faible [Altschul et al. 1997]. Les différentes étapes dans la construction des matrices PSSM à travers l'outil PSI-BLAST sont listées ci-dessous :

1. Une recherche standard BLAST est effectuée contre une base de données en utilisant une matrice de substitution [Eddy 2004], obtenue par la construction de profils.
2. Une matrice PSSM est construite automatiquement à partir d'un alignement multiple des séquences ayant le plus haut score ("*hits*") dans cette première recherche BLAST.
    - Positions très conservées : scores élevés
    - Positions faiblement conservées : scores faibles
3. La matrice PSSM remplace la matrice initiale et on effectue une 2ème recherche BLAST.
4. Les étapes 3 et 4 sont répétées et à chaque fois, les séquences nouvellement trouvées sont ajoutées afin de construire une nouvelle matrice PSSM.
5. On considère que le programme PSI-BLAST a convergé quand aucune nouvelle séquence n'est ajoutée.

Plusieurs travaux récents sur les techniques d'extraction utilisent ainsi l'outils PSI-BLAST pour calculer des scores PSSM dans la démarche de la technique [Zhang et al. 2012 ; Dehzangi et al. 2017; Li et al. 2017; Göktepe et Kodaz 2018 ; An et al. 2019].

2.3.3.4. *Limites de la technique Bi-gram*

La technique *Bi-gram* applique la technique 2-gramme [Cavnar and Trenkle 1994] sur une matrice de scores PSSM obtenue à partir de l'outil PSI-BLAST pour extraire les caractéristiques bigrammes et ne présente pas le problème de vecteur strictement parcimonieux [Sbai 2012]. Toutefois, l'efficacité de cette technique dépend d'un certains nombres de paramètres tel que la base de données cible pour comparer la séquence requête.

En effet, la base de données de cible ou de comparaison doit contenir des protéines d'attributs connus qui présentent une homologie (forte similarité) avec la séquence requête. Selon Chou [2001] et Garg *et al*. [2005] une telle démarche présente un inconvénient. Le problème est que lorsque la protéine de requête n'a pas d'homologie ou de similarité significative avec des protéines aux caractéristiques connues, les scores attribués ne sont pas significatifs et donc ne reflètent pas le score d'évolutivité recherché de la séquence. Il faut savoir en effet qu'à l'origine, les outils basés sur la recherche de similarité ont été utilisés pour les annotations fonctionnelles (documentations sur les fonctions) des protéines [Nguyen et al. 2011]. Or, les fonctions des protéines sont étroitement liées à leurs attributs cellulaires [Chou 2001] ; par conséquent, les protéines apparentées doivent être localisées dans le même compartiment cellulaire pour avoir une fonction commune [Garg et al. 2005]. Pour résumé, un score attribué sera vraiment significatif si la protéine de requête et les protéines qui servent de comparaison sont apparentées, ce qui n'est pas toujours le cas en choisissant une base de données pour la comparaison. Un autre problème avec cette technique est le temps d'exécution pour le calcul des scores PSSM. Pour augmenter la significativité des valeurs de scores PSSM, la plupart des auteurs choisissent la base de données nr (voir tableau 2-4) [Sharma et al. 2013; An et al. 2019]. Or cette base de données comporte des millions de séquence (plus de 437 millions), ce qui rend long le temps d'exécution pour le calcul des scores PSSM des échantillons de protéines comme le mentionne [Yao et al. 2019]. Par exemple pour la séquence de la figure 2-11 dont la longueur est de 195 (195 acides aminés), le temps de calcul des scores PSSM est de 220 secondes soit 3 minutes 40 secondes. Ce calcul est effectué sur la base du choix de la base de données 'nr'.

```
TNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVI
RGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEI
YQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGP
```

Figure 2-11: Séquence primaire de la Chaîne B, domaine de liaison du récepteur du SRAS-CoV-2

Si nous extrapolons ce calcul à un jeux de données d'IPP comportant 2000 échantillons, on se retrouve avec un temps d'exécution minimum de 480.000 secondes. Cela correspond à 133,33 heures ou 5,55 jours. Ce qui est extrêmement long.

Conclusion

Ce chapitre a présenté un état de l'art des outils d'apprentissage pour la prédiction d'interaction entre protéines. Les performances des méthodes d'apprentissage supervisé dépendent fortement des caractéristiques extraites pour réaliser l'inférence. Dans cette étude nous nous sommes beaucoup plus intéressés aux caractéristiques bigrammes. En effet, les caractéristiques bigrammes simulent le repliement des protéines, propriété chimique importante dans l'interaction protéine-protéine. Les techniques existantes pour extraire ces caractéristiques présentent certaines limites à savoir le problème de vecteur strictement parcimonieux dans le cas de la technique *Pairwize Frequency* et le problème de dépendance d'une base de données de comparaison pour la technique *Bi-gram*. Dans le prochain chapitre, nous proposons une nouvelle technique pour extraire les caractéristiques bigrammes des protéines en contournant les problèmes cités ci-dessus.

CHAPITRE 3. NOUVELLE TECHNIQUE D'EXTRACTION DE CARACTERISTIQUES BIGRAMMES

SOMMAIRE

Introduction

Dans ce chapitre, une nouvelle technique d'extraction des caractéristiques de la séquence, notée BP (*Bigram Physicochemical*), est proposée. La technique BP est une approche de représentation de la séquence par les bigrammes et repose sur une modélisation du repliement de la protéine. Les bigrammes sont les fréquences de deux lettres (ou acides aminés) successives. Cette représentation permettant de définir le problème de l'extraction de caractéristiques pour un apprentissage automatique efficace. La technique BP développée dans cette thèse vient pallier certaines limites constatées dans les approches PF (*PairWize Frequency*) et *Bi-gram* de représentation d'une protéine par les bigrammes d'acides aminés contenus dans sa séquence. Cette technique, contrairement aux techniques existantes construit un vecteur où non seulement la plupart des composantes ne sont pas nulles et les caractéristiques bigrammes extraites ne dépendent d'aucune base de données.

La technique BP comporte deux étapes : la première calcule de manière heuristique une matrice de scores à partir des informations de propriétés amphiphiles (hydrophobes et hydrophiles) des acides aminés [Chou 2005] et de la flexibilité des acides aminés. La seconde étape utilise la technique 2-gramme [Cavnar and Trenkle 1994] sur la matrice calculée à la première étape pour extraire des caractéristiques bigrammes d'acides aminées. Les scores de la matrice sont obtenus selon deux approches : soit à partir d'une distance ou soit à partir d'une fonction. La section **3.1** fait une présentation générale de la nouvelle technique développée. Nous rappelons le problème avec les techniques existantes par une illustration. Ensuite nous présentons le fonctionnement de la nouvelle technique. La section **3.2** est dédiée aux approches heuristiques de calcul de la matrice de scores utilisée par la technique développée. La section **3.3** aborde l'extraction des caractéristiques bigrammes par la nouvelle technique. Dans la section **3.4**, il est question du matériel et des différentes techniques utilisées pour former un outil de prédiction d'interaction protéine-protéine à partir de la nouvelle technique développée. La section **3.5** présente les résultats d'expérimentation obtenus avec la nouvelle technique proposée. Enfin, la section **3.6** propose une discussion sur la technique développée.

3.1. Présentation générale de la nouvelle technique

Du point de vue du problème de l'extraction de caractéristiques ou de la représentation des données [Bengio et al. 2013], l'objectif de la technique BP (*Bigram Physicochemical*) est de

proposer une représentation de la protéine par les différents bigrammes à partir de la séquence d'acides aminés. Cette technique calcule de façon heuristique une matrice de scores relatifs à la flexibilité et à la propriété amphiphile (effet hydrophobe, effet hydrophile) des acides aminés puis extrait les différentes caractéristiques bigrammes à travers la matrice calculée.

Les bigrammes d'une protéine, qui sont les fréquences de deux acides aminés successifs, révèlent en effet plusieurs propriétés biochimiques de la protéine [Almagor 1983]. L'une d'entre elles est le repliement de la protéine, propriété qui est liée à la réalisation de la fonction d'une protéine ainsi qu'à son interaction avec d'autres protéines [Bushmarina et al. 2005]. En effet, le repliement des protéines est un procédé par lequel la chaîne d'acides aminés des protéines se plie pour adopter une structure tridimensionnelle (structure tertiaire) et permet aux protéines d'être biologiquement actives [Sali et al. 1994]. Une façon de simuler le repliement est de pouvoir reconnaître les parties de la séquence de protéine pouvant être pliées. Extraire les caractéristiques bigrammes d'une protéine répond à la problématique de la reconnaissance des 'plis' de la protéine [Sharma et al. 2013]. Ainsi, des approches d'extraction des caractéristiques bigrammes de la protéine ont été développées. Cependant, les approches existantes présentent certaines limites dans la procédure d'extraction de ces caractéristiques.

3.1.1. Problème avec les techniques existantes

L'approche PF (*Pairwize Frequency*) [Yang et al. 2011] utilise la technique 2-gramme directement sur la structure primaire d'une protéine donnée pour extraire les caractéristiques bigrammes. Le vecteur résultant de cette technique pour une protéine donnée comporte donc 400 composantes en raison des 20 acides aminés standards qui forment la protéine. Cependant, compte tenu du nombre limité de combinaisons de bigrammes pour une séquence de protéine, de nombreuses composantes du vecteur formé sont égales à zéro. Cela pose ainsi le problème de vecteur strictement parcimonieux [Sbai 2012] que nous illustrons à travers l'exemple ci-dessous.

Supposons que dans la littérature nous ayons trois acides aminés naturels que sont *ACT* (au lieu de 20). Considérons la séquence primaire d'une protéine exemple *CCACA* de longueur $L = 5$. Les différents bigrammes formés avec la séquence exemple sont donnés comme suit :

$$Caractéristiques\ BAA\ avec\ PF \begin{cases} AA = 0 \\ AC = 1 \\ AT = 0 \\ CA = 2 \\ CC = 1 \\ CT = 0 \\ TA = 0 \\ TC = 0 \\ TT = 0 \end{cases}$$

Le vecteur résultant V par l'application de cette approche est obtenu par l'ensemble des caractéristiques bigrammes calculées, c'est à dire $V = \{0,1,0,2,1,0,0,0,0\}$. Ce vecteur formé est constitué de 67% de zéros, soit six zéros sur neuf. Un tel vecteur est qualifié de vecteur strictement parcimonieux et peut rendre moins performant tout classifieur car n'apportant pas suffisamment d'informations utiles pour un meilleur apprentissage de l'algorithme d'apprentissage supervisé. Ici, toutes les informations pour la reconnaissance des 'plis' de la protéine ne sont pas renseignées.

Pour résoudre ce problème de vecteur strictement parcimonieux, Sharma et al. [2013] proposent l'approche *Bi-gram*, qui applique la technique 2-gramme sur une représentation matricielle de la séquence, en l'occurrence la PSSM [cheol Jeong et al. 2010] plutôt que directement sur la séquence primaire. L'outil de recherche de similarité PSI-BLAST [Altschul et al. 2009] est généralement utilisé. Cependant, l'efficacité de cette technique dépend uniquement de la base de données cible choisie pour effectuer l'alignement de séquences dans le processus de calcul des scores PSSM. En effet, la technique *Bi-gram* échoue lorsque la base de données cible ne contient pas de séquences d'attributs connus qui présentent une homologie (forte similarité) avec la séquence requête. En outre, le temps de calcul des scores PSSM est long comme l'indique [Yao et al. 2019].

3.1.2. Technique proposée

La technique BP adopte une approche similaire à l'approche de la technique des probabilités bigrammes (*Bi-gram*) pour représenter la protéine et extraire les caractéristiques bigrammes. Pour une protéine donnée, au lieu de calculer les bigrammes directement sur sa structure primaire ou sur sa PSSM, BP calcule les bigrammes sur une matrice de scores notée MSP (Matrice de Scores Physicochimiques) et obtenue à partir de certaines informations des acides aminés de la protéine. La MSP est calculée à partir du multiplicateur d'une fonction de rang lié à la flexibilité des acides aminés [Dunker et al. 2001] et d'une fonction de

dépendance simulant la distribution des acides aminés hydrophobes et hydrophiles respectivement à travers les propriétés physicochimiques hydrophobicité (H_1) et hydrophilie (H_2). La MSP a les mêmes dimensions que la PSSM ($L\_lignes \times 20\_colonnes$). Elle fournit également des informations (ici des scores) sur la probabilité de 20 acides aminés à chaque emplacement de la séquence d'une protéine permettant ainsi d'éviter les composantes nulles dans le vecteur de caractéristiques résultant. Contrairement au temps de calcul des scores PSSM qui est long (plus de deux minutes pour une séquence de 200 acides aminés par exemple), le temps de calcul des scores MSP est court (environ 42 secondes pour une séquence de 200 acides aminés).

3.1.2.1. *Informations de la séquence représentées dans la nouvelle technique*

Une donnée extraite par rapport à une entité *u* sera fortement représentative à condition d'extraire les informations inhérentes à *u*. En partant de cela, la technique développée combine bien des informations de la séquence pour extraire des caractéristiques fortement liées au repliement de la protéine. En effet, la technique proposée fait l'apprentissage de la propriété chimique du repliement de la protéine à travers deux propriétés fonctionnelles de la protéine à savoir l'effet hydrophobe et la flexibilité des acides aminés que nous détaillons ci-dessous.

- **Effet hydrophobe** [Martin 2008]

Notons que chaque acide aminé possède des caractéristiques physicochimiques propres. Certains sont hydrophobes pendant que les autres sont hydrophiles (caractéristiques amphiphiles [Chou 2005]). Les acides aminés hydrophobes ont plus d'affinité entre eux qu'avec les molécules d'eau entourant la protéine [Tanford 1962]. Par conséquent, lors du repliement d'une protéine, la chaine a tendance à se courber de façon à les regrouper entre eux au centre de la molécule sans contact direct avec l'eau (voir figure 3-1). Inversement, les acides aminés hydrophiles ont tendance à se disposer à la périphérie de façon à être en contact avec l'eau. Selon [Chou 2005], ces deux propriétés des acides aminés constitutifs d'une protéine jouent un rôle très important dans son repliement, son interaction avec l'environnement et d'autres molécules, ainsi que dans son mécanisme catalytique. Les propriétés physicochimiques hydrophobicité (symbolisé par H_1) et hydrophilie (symbolisé par H_2) sont utilisées ici pour représenter les caractéristiques amphiphiles de la protéine. Nous signalons que les autres types d'interactions (liaison ionique, liaison hydrogène et pont disulfure) de la figure 3-1 ne sont pas pris en compte dans cette étude.

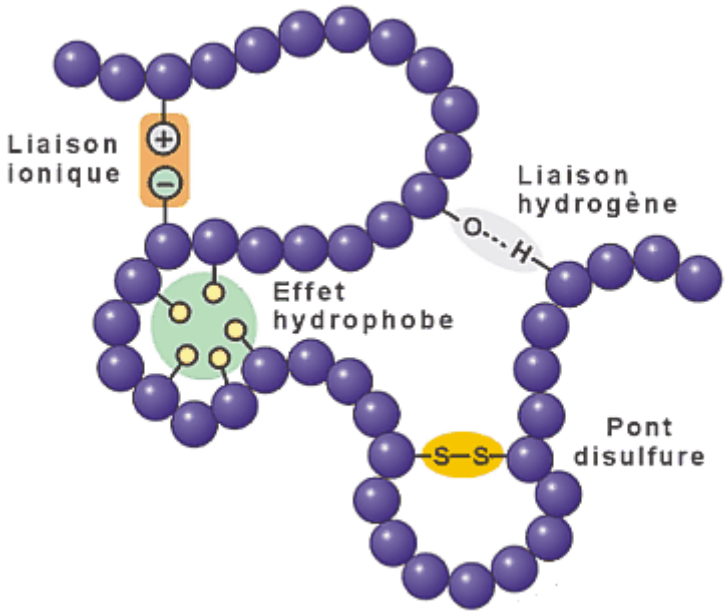


Figure 3-1: Effet hydrophobe (adapté de babel.cegep-ste-foy.qc.ca)

- **Flexibilité des acides aminés** [Dunker et al. 2001]

Selon les fonctions qu'elles accomplissent dans les cellules, les protéines se présentent sous une forme rigide ou flexible. La flexibilité est une caractéristique intrinsèque des protéines qui doivent, dès le moment de leur synthèse, passer d'un état de chaîne linéaire à un état de structure tridimensionnelle (tertiaire) repliée et enzymatiquement active. Soit $\Sigma = \{W, C, F, I, Y, V, L, H, M, A, T, R, G, Q, S, N, P, D, E, K\}$, avec $|\Sigma| = 20$, représentant l'ensemble des 20 acides aminés. Ici, les acides aminés sont rangés dans un certain ordre où les plus rigides sont à gauche et les plus flexibles sont à droite (voir figure 3-2) comme dans [Dunker et al. 2001]. Selon Dunker et al. [2001] cet ordre de rangement des acides aminés de la figure 3-2 est basé sur l'échelle établie par [Vihinen et al. 1994] où par exemple les acides aminés W, C et F sont les trois plus rigides, par-contre les acides aminés D, E et K sont les trois plus flexibles.

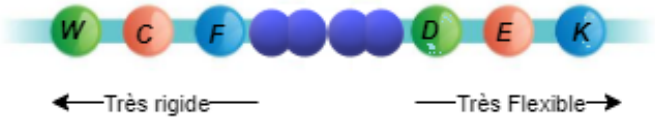


Figure 3-2: Disposition des acides aminés du moins au plus flexible

3.1.2.2. *Etapes de calcul des caractéristiques bigrammes*

Les caractéristiques bigrammes avec la technique BP proposée sont calculées en deux étapes :

(1) **Calcul de la matrice de scores physicochimiques (MSP)** : Une séquence de protéine peut être vue comme une suite finie de m lettres où m est le nombre d'éléments de l'alphabet fini Σ. La séquence peut être représentée à l'aide d'une matrice, qui restitue le contenu informationnel de la séquence, où le nombre de ligne correspond à la longueur L de la séquence, le nombre de colonnes correspond au nombre $m = |\Sigma|$ et tout élément $e_{ij,(1\leq i\leq L;1\leq j\leq m)} \in \mathbb{R}$ représente un score relatif à la dépendance (indépendance) fonctionnelle entre la $j^{ème}$ lettre de Σ et la $i^{ème}$ lettre de la séquence. Le score défini ici met en relief la caractéristique du repliement de la protéine.

(2) **Calcul des caractéristiques bigrammes** : les bigrammes sont calculés par l'application de la technique 2-gramme [Cavnar and Trenkle 1994] sur la MSP de l'étape (1). Ce calcul permet de formuler une matrice carrée (matrice 20x20) de caractéristiques bigrammes. Cette matrice est ensuite transposée en un vecteur de 400 composantes bigrammes.

La figure 3-3 montre le logigramme du calcul des caractéristiques bigrammes par la technique BP. Pour une séquence de longueur L où les acides aminés sont représentés par leur radical $R_1R_2R_3R_4R_5 \ldots R_L$, nous calculons la matrice de scores MSP. Ensuite, nous calculons les caractéristiques bigrammes notées BP à partir de la matrice MSP. Enfin nous représentons les différentes caractéristiques bigrammes calculées sous forme d'un vecteur de caractéristiques bigrammes notée V_{BP}.

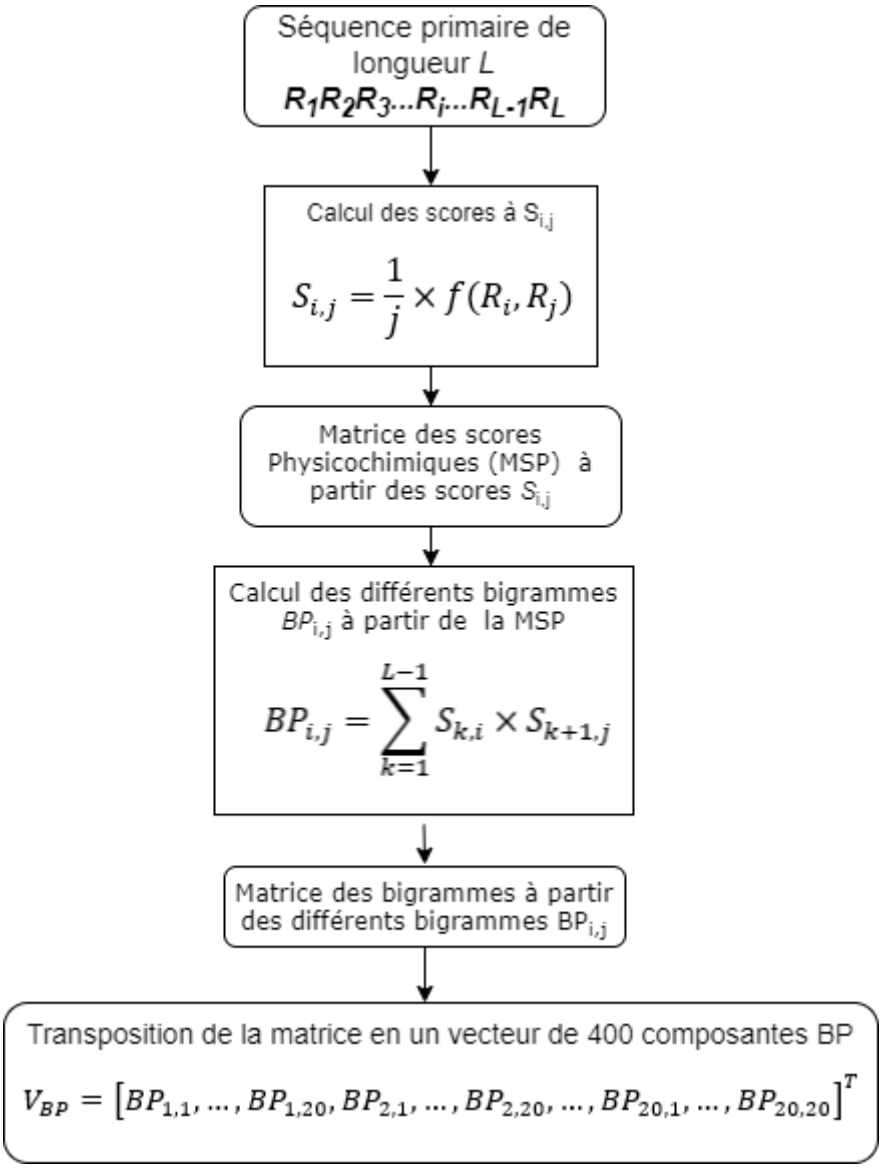


Figure 3-3: Logigramme de calcul des caractéristiques bigrammes

3.2. Calcul de la matrice de scores physicochimiques

La matrice de scores physicochimiques (MSP) est calculée à partir des informations de propriétés physicochimiques des acides aminés. Le score calculé ici représente la probabilité d'information de chacun des 20 acides aminés à une position *i* dans la séquence. Dans cette section, nous donnons les détails de calcul de la matrice de scores physicochimiques.

Définition 3-1 : Matrice de scores physicochimiques

Soit S une séquence de longueur L et Σ *l'ensemble des 20 acides aminés. La matrice de scores physicochimiques notée MSP représentative de la séquence S est une matrice L* × |Σ|

telle que pour toute position $i = 1, \dots, L$ et $j = 1, \dots, |\Sigma|$, *l'élément d'indices* i *et* j *de la MSP noté* $S_{i,j}$ *est définie par l'équation* 3-1 *:*

$$S_{i,j} = \frac{1}{j} \times f(R_i, R_j) \tag{3-1}$$

où $\frac{1}{j}$, avec $1 \leq j \leq 20$ est une fonction inverse prise ici comme une fonction de pondération de rang [Besson 1975; Roy 2007]. En effet, nous considérons l'information relative à la flexibilité des acides aminés. Pour cela, nous avons disposé les éléments de Σ, c'est-à-dire les 20 acides aminés sur la base de la flexibilité des résidus d'acides aminés, où les plus rigides sont à gauche et les plus flexibles à droite (voir figure 3-2). Cette fonction modélise un comportement décroissant. Si nous considérons le plus rigide comme le niveau supérieur et le plus flexible comme le niveau inférieur, la décroissance est donc simulée à travers un certain ordre de rangement des 20 acides aminés naturels. Une modélisation plus complète de la flexibilité peut être trouvée dans [Bornot 2009].

L'expression $f(R_i, R_j)$ avec $1 \leq i \leq L$ est une fonction qui représente la distribution de l'hydrophobicité et de l'hydrophilie des acides aminés le long de la séquence de longueur *L*. nous pouvons calculer cette fonction selon deux approches. Une première approche utilise la distance et une autre approche utilise une fonction. Le fait de proposer deux approches de calcul de cette fonction est motivé par le fait que ces deux approches permettent de représenter la distribution des acides aminés hydrophobes et hydrophiles le long de la séquence de protéine. En effet, nous pouvons exprimer la dépendance entre les 20 acides aminés naturels et les acides aminés de la protéine à travers une distance (nous parlerons dans ce cas de dissimilarité) comme dans [Chou 2001] ou à travers une fonction comme dans [Chou 2005]. Dans cette étude nous avons utilisé le carré de la distance euclidienne [Perlibakas 2004] pour l'approche de la distance. Cette mesure de distance est l'une des plus utilisées en ce qui concerne les nombres discrets comme mentionné dans [Sherali and Tuncbilek 1992]. Dans la suite de cette section, nous donnons les détails de chaque approche de calcul de la matrice de scores MSP.

3.2.1. Approche de calcul à partir d'une distance

L'indépendance ou la dissimilarité est exprimée dans l'approche BP par une distance. Nous utilisons ici le carré de la distance euclidienne [Perlibakas 2004] pour mesurer la dissimilarité

entre les acides aminés d'une séquence de longueur L et ceux de l'alphabet fini |Σ|. Soit la protéine P définie comme suit :

$$P = R_1R_2R_3R_4R_5 \dots R_L$$

où R_i et R_j sont respectivement l'acide aminé à la position i dans la séquence et le j ^{ème} acide aminé de l'alphabet fini |Σ|. La dissimilarité entre R_i et R_j est calculée de la manière suivante :

$$f(R_i, R_j) = \frac{1}{2}\left\{\left[H_1(R_j) - H_1(R_i)\right]^2 + \left[H_2(R_j) - H_2(R_i)\right]^2\right\}$$

où $H_1(R_i)$ et $H_2(R_i)$ sont respectivement la valeur normalisée de l'hydrophobicité de l'acide aminé i et de l'hydrophilie de l'acide aminé i. Considérons H_1^0 et H_2^0, les valeurs originales respectives des propriétés hydrophobicité et hydrophilie. Les valeurs H_1 et H_2 sont calculées à partir de l'équation 3-2 :

$$\begin{cases} H_1(R_i) = \dfrac{H_1^0(R_i) - \varphi_1}{\sqrt{\sum_{i=1}^{20}\left[H_1^0(R_i) - \varphi_1\right]^2 / 20}} \\ H_2(R_i) = \dfrac{H_2^0(R_i) - \varphi_2}{\sqrt{\sum_{i=1}^{20}\left[H_2^0(R_i) - \varphi_2\right]^2 / 20}} \end{cases} \quad (3\text{-}2)$$

où φ_1 et φ_2 , sont respectivement la moyenne des valeurs d'hydrophobicité des 20 acides aminés de |Σ| et la moyenne des valeurs d'hydrophilie des 20 acides aminés de |Σ| exprimées comme suit :

$$\begin{cases} \varphi_1 = \sum_{k=1}^{20} H_1^0(R_k) / 20 \\ \varphi_2 = \sum_{k=1}^{20} H_1^0(R_k) / 20 \end{cases}$$

Compte tenu que la variance dépend de l'échelle des variables, il est d'usage de normaliser d'abord chaque variable pour qu'elle ait une moyenne de zéro et un écart-type de un.

3.2.2. Approche de calcul par une fonction

Nous exprimons dans cette approche la dépendance entre deux résidus d'acides aminés à travers une fonction définie comme suit :

$$f(R_i, R_j) = H_1^*(R_i) \times H_1^*(R_j) + H_2^*(R_i) \times H_2^*(R_j) \; ; 1 \leq i \leq L, \;\; 1 \leq j \leq 20$$

où H_1^* et H_2^* sont respectivement les valeurs normalisées d'hydrophobicité et d'hydrophilie calculée dans l'équation 3-2.

La matrice ainsi constituée comporte *L* lignes et 20 colonnes où *L* correspond à la longueur de la séquence. Chaque élément à la *i*ème ligne et la *j*ème colonne noté $S_{i,j}$ peut être interprétée comme le score relatif du *j*ème acide aminé à la *i*ème position de la séquence primaire.

3.3. Calcul des caractéristiques bigrammes par la nouvelle technique

Le calcul de la matrice de scores physicochimiques constituant la première étape de la technique BP, nous présentons dans cette section la deuxième étape de la technique qui concerne le calcul des caractéristiques bigrammes à travers la matrice calculée.

3.3.1. Matrice d'occurrences et vecteur de caractéristiques bigrammes

3.3.1.1. *Matrice d'occurrences bigrammes*

La valeur du bigramme entre les acides aminés *i* et *j*, représentée par la fréquence d'occurrence de la transition de l'acide aminé se trouvant à la position *i* à l'acide aminé se trouvant à la position *j* est calculée selon l'équation 3-3:

$$BP_{i,j} = \sum_{k=1}^{L-1} C_{k,i} \times C_{k+1,j} \;\; , \; 1 \leq i \leq 20; 1 \leq j \leq 20 \qquad (3\text{-}3)$$

où *L* représente la longueur de la séquence, $C_{k,i}$ est la valeur de la MSP à la ligne *k* et à la colonne *i* et $C_{k+1,j}$ est la valeur de la MSP à la ligne *k* +1 et à la colonne *j*. L'équation 3-3 donne ainsi 400 $BP_{r,s} \; ; 1 \leq r \leq 20 \; ; 1 \leq s \leq 20$ occurrences de fréquences pour 400 transitions bigrammes (20×20). Nous appelons la matrice ***BP,*** la matrice d'occurrences bigramme et ses 400 éléments et peuvent être rangés sous forme d'un vecteur de 400 composantes.

3.3.1.2. *Vecteur de caractéristiques bigrammes*

La matrice BP et ses 400 éléments constituent le vecteur caractéristiques bigrammes noté V_{BP} et peut- être formulé comme suit :

$$V_{BP} = \left[BP_{1,1}, BP_{1,2}, \dots, BP_{1,20}, BP_{2,1}, BP_{2,2}, \dots, BP_{2,20}, \dots, BP_{20,1}, BP_{20,2}, \dots, BP_{20,20}\right]^T$$

avec *T*, l'opérateur de transposition du vecteur.

Le vecteur de caractéristiques bigrammes construit permet de capturer les caractéristiques essentielles et centrales d'une protéine car il peut également être écrit sous la forme PseAAC (voir chapitre 2) comme suit [Chou, 2011] :

$$V_{BP} = [\Phi_1, \Phi_2, \Phi_3, \dots, \Phi_\mu, \dots, \Phi_\Psi]^T$$

où $\Psi = r \times s = 400$ est la dimensionnalité du vecteur caractéristique V_{BP}. Les composantes du vecteur V_{BP} peuvent être exprimées comme les caractéristiques des pseudo-acides aminés de la manière suivante :

$$\Phi_\mu \begin{cases} BP_{1,\mu} & (\mu \in [1,20]) \\ BP_{2,\mu-20} & (\mu \in [21,40]) \\ BP_{3,\mu-40} & (\mu \in [41,60]) \\ & \dots \\ & \dots \\ BP_{20,\mu-380} & (\mu \in [381,400]) \end{cases}$$

Etant donné que dans le calcul du vecteur caractéristique V_{BP}, toutes les informations de la probabilité de MSP formée ont été intuitivement utilisées, V_{BP} contient plus d'informations utiles pour la tâche de reconnaissance des 'plis' de la protéine que le calcul du bigramme directement à partir de la séquence protéique. Un diagramme de flux montrant les étapes de calcul du vecteur caractéristique bigramme est représenté à la figure 3-1. Dans la suite nous illustrons le calcul du vecteur bigramme formé.

3.3.2. Illustration du calcul des caractéristiques bigrammes

Pour illustrer le calcul des caractéristiques bigramme par la technique BP, nous considérons le même cas d'exemple que dans la section 3.1, c'est-à-dire la séquence primaire *CCACA* avec les acides aminés naturels *A*, *C* et *T*. Le vecteur généré en calculant le bigramme directement à partir de la séquence exemple donne $V = \{0; 1; 0; 2; 1; 0; 0; 0; 0\}$, soit 67% de zéros. Nous donnons dans la suite le calcul par l'approche de dissimilarité suivi du calcul par l'approche de similarité.

3.3.2.1. *Exemple de calcul par l'approche de distance*

Les tableaux 3-1 et 3-2 renseignent respectivement sur les valeurs d'hydrophobicité et hydrophilie des acides aminés *ACT* et la matrice de scores physicochimiques obtenue après calcul de tous les scores en appliquant l'équation 3-2. La matrice d'occurrences bigramme obtenue après le calcul des différents bigrammes par application de l'équation 3-3 est indiquée dans le tableau 3-3.

Tableau 3-1: Valeurs originales des propriétés Hydrophobicité et Hydrophilie des acides aminés *ACT*

Acide aminé	H_1	H_2
A	0.62	-0.5
C	0.29	-1.0
T	-0.05	-0.4

Tableau 3-2: Matrice de scores physicochimiques dans le cas de la distance

Séquence primaire	*C*	*A*	*T*
C	2.542	0	1.128
C	2.542	0	1.128
A	0	1.271	1.024
C	2.542	0	1.128
A	0	1.271	1.024

Tableau 3-3: Matrice des occurrences bigramme avec la séquence *CAT* dans le cas de la distance

Acides aminés	*C*	*A*	*T*
C	5.02	6.46	1.29
A	3.23	0	2.06
T	5.51	2.5	2.81

Le calcul du bigramme $BP_{A,C}$, par application de l'équation 3-3 par exemple est donné comme suit :

$$BP_{A,C} = (2.542 \times 0) + (2.542 \times 1.271) + (0 \times 0) + (2.542 \times 1.271)$$

$$= \mathbf{6.46}$$

La valeur 6.46 calculée est obtenue en faisant la somme des produits des valeurs de de l'acide aminé *A* à la ligne *i* par les valeurs de l'acide aminé *C* à la ligne *i* +1 de la matrice de scores physicochimiques renseignées dans le tableau 3-2. L'ensemble des différentes caractéristiques bigrammes sont listées comme suit :

$$Caractéristiques\ bigramme\ BP \begin{cases} BP_{A,A} = 5.01 \\ BP_{A,C} = 6.46 \\ BP_{A,T} = 1.29 \\ BP_{C,A} = 3.23 \\ BP_{C,C} = 0 \\ BP_{C,T} = 2.06 \\ BP_{T,A} = 5.51 \\ BP_{T,C} = 2.50 \\ BP_{T,T} = 2.81 \end{cases}$$

Le vecteur de caractéristiques bigrammes résultant est le suivant :

$$V_{BP} = (5.01; \mathbf{6.46}; 1.29; 3.23; 0; 2.06; 5.51; 2.5; 2.81)$$

Nous pouvons constater que le vecteur de caractéristiques bigramme obtenu avec la première approche de calcul contient moins de 11% de 0 (soit un zéro sur neuf) et donc ne souffre pas de problème de vecteur strictement parcimonieux comme avec l'approche PF vu dans la section 3-1.

3.3.2.2. *Exemple de calcul par l'approche de la fonction*

La matrice de scores physicochimiques obtenue par application de l'équation 3-2 donne les scores renseignés dans le tableau 3-4. La matrice d'occurrences bigramme résultant en utilisant l'équation 3-3 est présentée à travers le tableau 3-5. Le vecteur caractéristique résultant est :

$$V_{BP} = (-3.185; 1.509; 0.049; 2.962; -0.076; -0.938; -0.931; -0.452; 0.617)$$

Nous pouvons dire que le vecteur de caractéristiques bigramme obtenu par cette deuxième approche contient 0% de 0, et donc contient moins de 0 que dans la première approche.

Tableau 3-4: Matrice des scores physicochimiques dans la deuxième approche

Séquence primaire		*C*	*A*	*T*
1	*C*	-0.7	0.97	-0.42
2	*C*	-0.7	0.97	-0.42
3	*A*	1.75	-0.35	-0.35
4	*C*	-0.7	0.97	0.42
5	*A*	1.75	-0.35	-0.35

$$BP_{A.C} = (-0.7 \times 0.97) + (-0.7 \times (-0.35)) + (1.75 \times 0.97) + (-0.7 \times (-0.35)) = 1.509$$

La dernière ligne du tableau 3-5 nous montre le calcul qui a permis d'obtenir la valeur du bigramme *AC*.

Tableau 3-5: Matrice des occurrences de bigrammes dans le cas de la fonction

Acides aminés	*C*	*A*	*T*
C	-3.185	1.509	0.049
A	2.962	-0.076	-0.938
T	-0.931	-0.452	0.617

3.3.2.3. *Représentation de la paire de protéines*

Le calcul des caractéristiques bigrammes par l'approche de la distance et celle de la fonction ont produit dans chacune des approches un vecteur dense qui apporte beaucoup plus d'informations sur la reconnaissance des 'plis' de la protéine. Par conséquent, la technique BP extrait représente mieux les bigrammes de la séquence que celle d'appliquer directement la

technique 2-gramme sur la séquence primaire de la protéine. Pour représenter la paire formée par deux protéines, nous concaténons les vecteurs de caractéristiques V_{BP} de chacune des deux protéines [Göktepe et Kodaz 2018] formant la paire. Le vecteur final obtenu est constitué alors de 800 composantes.

3.4. Matériel et méthodologie pour la prédiction des interactions

Après l'étape d'extraction des caractéristiques des séquences d'acides aminés par la nouvelle technique développée, l'étape suivante est la phase d'inférence de l'interaction par l'introduction d'un classifieur. Dans cette section, nous décrivons le classifieur ainsi que les ensembles de données IPP utilisés pour constituer un outil d'inférence des interactions entre les protéines.

3.4.1. Ensembles de données IPP de référence

Les ensembles de données de séquences sur lesquelles la performance de la technique BP a été évaluée sont les données IPP HPRD [Pan et al. 2010]. Pour une meilleure comparaison et pour éviter le déséquilibre des données, nous avons considéré dans cette étude un même nombre d'observations pour les données HPRD que la plupart des auteurs qui ont évalué leur technique sur les ensembles de données HPRD. Nous avons donc considéré 10000 échantillons repartis en 5000 paires IPP positives et 5000 paires IPP négatives comme [Zhou et al. 2011 ; You et al. 2013 ; Göktepe and Kodaz 2018 ; An et al. 2019 ; Ma et al. 2020] au lieu de 36600 IPP positives et 36400 IPP négatives comme vu au chapitre 1. Les ensembles de données IPP S. Cerevisiae [You, Zhu, et al. 2014] constitués de 11188 échantillons (avec 5594 paires positives et 5594 paires négatives) et IPP H. Pylori [Martin et al. 2005] constitué de 2496 échantillons (1458 paires positives et 1458 paires négatives) ont également été utilisés. Nous soulignons également que les différents ensembles de données IPP utilisés sont au format FASTA [Binz et al. 2019]. La simplicité du format FASTA facilite la manipulation et l'analyse des séquences à l'aide d'outils de traitement de texte et de langages de script tels que le langage de programmation R, Python, Ruby et Perl.

3.4.2. Réduction de l'espace de caractéristiques

Dans le domaine du codage des séquences d'une protéine, il est commode de rencontrer des redondances de caractéristiques et du bruit dans les composantes du vecteur formé. Pour sélectionner les caractéristiques pertinentes pour l'apprentissage, la technique de l'analyse en

composante principale (ACP) est appliquée [Lorenz 1989]. L'idée de base de l'ACP est la réduction de la dimension de l'espace des caractéristiques en ne retenant que les composantes de forte variance qualifiées de principales.

Soit $x_i \in \mathbb{R}^d$, un ensemble d'observations constituées de vecteurs lignes avec $i \in [1; \sigma]$; les composantes principales sont déterminées par les valeurs propres $\lambda > 0$ et les vecteurs propres v non nul comme suit :

$$\lambda v = Mv$$

où M représente la matrice de covariance des observations et exprimée de la manière suivante :

$$M = \frac{1}{\sigma}\sum_{i=1}^{\sigma} x_i . {x_i}^T$$

ainsi, nous avons donc

$$\lambda v = \frac{1}{\sigma}\sum_{i=1}^{\sigma} (x_i . v) x_i$$

Nous en déduisons donc que tous les vecteurs v avec $\lambda \neq 0$ sont des éléments du sous espace généré par les observations x_i.

Pour le choix du nombre de composantes, nous retenons le critère de Kaiser, aussi connu sous le nom de critère des valeurs propres, qui est l'un des critères beaucoup utilisés pour l'application de l'ACP [Yeomans et Golder 1982 ; Yao et al. 2019]. Le critère de kaiser consiste tout simplement à sélectionner un nombre ζ de composantes où ζ correspond au nombre de valeurs propres plus grand que 1.

La technique BP produisant 400 composantes pour une protéine donnée, la concaténation des deux protéines formant une paire donne 800 composantes. Désignons par BP1 l'approche utilisant la distance et BP l'approche utilisant une fonction. Les résultats de l'application du critère de Kaiser nous donnent 384 valeurs propres avec la plus grande variance expliquée (soit 85% de l'information) pour un vecteur obtenu par BP1. Concernant BP, nous enregistrons 471 valeurs propres avec la plus grande variance expliquée (soit 90% de l'information).

3.4.3. Classifieur SVM

Le plus souvent, les données issues du codage des protéines sont non linéairement séparables [Wei et al. 2016]. La figure 3-4 présente dans notre cas une répartition des données d'informations sur deux composantes après application de la technique BP. Nous avons ici une répartition de 100 observations par classe : la classe 1 correspond à la classe des paires positives, représentées par les croix bleues et la classe -1 correspond à la classe des paires négatives, représentées par les carrés rouges. Pour pouvoir séparer convenablement de telles données, les méthodes à noyaux comme les SVM sont généralement utilisés.

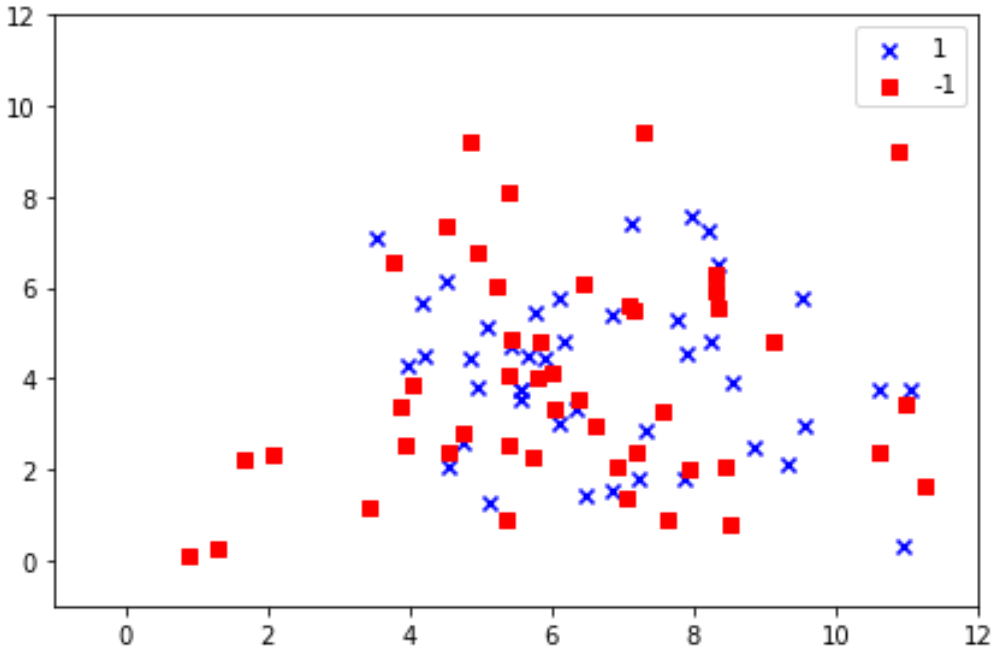


Figure 3-4: Répartition de 100 observations par classe avec les données HPRD

Considérons notre problème de classification (P) où $\mathcal{D}_n = \{(x_i, y_i), i \in [1, n]\}$ est notre jeu de données composé de n paires i.i.d telles que $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \{-1, +1\}$. D'après l'équation 2-1 vu au chapitre 2, la fonction de décision g des SVM dans le cas non linéaire s'écrit de la manière suivante :

$$g(x) = \sum_{i=1}^{n} \theta_i y_i k(x_i, x) + b$$

avec b le biais et $k(x_i, x)$ une fonction noyau. La méthode des noyaux consiste à projeter les données de l'espace $\mathbb{R}^d$ dans un autre espace de dimension plus élevée où les données qui étaient non linéairement séparables peuvent le devenir [Ben-Hur et Noble 2005]. Nous avons utilisé le noyau gaussien, compte tenu des performances réalisées avec ce noyau dans certains travaux de la littérature et aussi du fait qu'il permet une projection dans un espace indéfiniment grand [Göktepe et Kodaz 2018].

3.5. Expérimentation et résultats obtenus avec la nouvelle technique

Tout au long de cette thèse, les expériences ont été réalisées avec les langages python dans sa version 3.7 et R dans sa version 4.1, sur une machine possédant un processeur i7 avec une fréquence minimum de 2,56 GHz avec 8 Go de RAM. Nous présentons dans cette section tout d'abord les métriques d'évaluation utilisées. Ensuite nous montrons les résultats de performances obtenus avec la technique proposée sur les différents ensembles de données. Après cela, nous comparons les résultats obtenus par la technique proposée avec les techniques PF et *Bi-gram*. Nous comparons également les résultats des performances avec celle d'autres techniques qui codent autres types d'informations de séquences. Nous fermons cette section en nous comparant avec certains travaux de la littérature en termes d'outil formé et en termes de capacité de généralisation.

3.5.1. Métriques d'évaluation et expérimentation

3.5.1.1. *Métriques d'évaluation*

Les métriques d'évaluation utilisées sont celles généralement utilisées dans la littérature pour évaluer un outil; à savoir la justesse (taux d'exactitude), la précision, la sensibilité (rappel) et la courbe ROC (*Receiver Operating Curve*) et la valeur AUC (*Area Under Curve*) [An et al. 2019; Ma et al. 2020]. Certaines de ces métriques sont obtenues à partir de la matrice de confusion (voir tableau 3-9).

- **Matrice de confusion**

La matrice de confusion (MC) illustrée par le tableau 3-7 est utilisée pour évaluer la qualité de la méthode BP à travers un modèle de classification. Dans le tableau 3-7, les éléments diagonaux (les cases en bleu) représentent le nombre de points pour lesquels la classe de l'IPP

prédite est égale à sa vraie classe (prédiction correcte), tandis que les éléments hors diagonale (les cases en gris) sont les IPP mal regroupées par le classifieur (erreur de classification ou de prédiction). Plus les valeurs diagonales de la MC sont élevées, mieux c'est, car cela indique que les prédictions correctes sont nombreuses.

Soient IPP^+ et IPP^-, deux paires de protéines désignant respectivement une interaction positive (présence d'interaction) et une interaction négative (absence d'interaction). Ici, est qualifié de vrais positifs (VP) toute paire de protéine où les deux protéines formant la paire sont IPP^+ et que le modèle classifie comme une paire IPP^+. De même, sera qualifiée de vrais négatifs (VN) toute paire de protéine dans laquelle les deux protéines formant la paire sont IPP^- et que le modèle classifie comme une paire IPP^-. Nous notons également que toute paire de protéine où les deux protéines formant la paire sont IPP^- mais qui est classifiée par le modèle comme IPP^+ par erreur donne alors un résultat de faux positifs (FP). De la même façon, toute paire de protéine dans laquelle les deux protéines formant la paire sont liées et qui est classifiée comme IPP^- par erreur donne alors un résultat de faux négatifs (FN). Ainsi, en fonction des résultats de la MC, différentes métriques sont calculées pour évaluer la performance du modèle de prédiction. Nous donnons ci-dessous l'expression des métriques justesse, précision et sensibilité utilisées dans cette thèse.

Tableau 3-6: Matrice de confusion utilisée pour l'évaluation du modèle de classification

Vrai Positif (VP) : ▪ Réalité : **IPP^+** ▪ Prédiction du modèle : **IPP^+**	Faux Positif (FP) : ▪ Réalité : **IPP-** ▪ Prédiction du modèle : **IPP^+**
Faux Négatif (FN) : ▪ Réalité : **IPP^+** ▪ Prédiction du modèle : **IPP^-**	Vrai Négatif (VN) : ▪ Réalité : **IPP^-** ▪ Prédiction du modèle : **IPP^+**

- **Métriques**
  - Justesse

La justesse désigne la proportion de prédiction correcte (à la fois vrais positifs et varis négatifs) effectuée par le modèle, représente ici le nombre de prédictions d'IPP correctes par rapport au nombre total de prédictions et peut être exprimée selon l'expression ci-dessous :

$$\boldsymbol{Justesse} = \frac{\boldsymbol{VP+VN}}{\boldsymbol{VP+VN+FP+FN}}$$

- Précision

La précision permet de mesurer la proportion d'identification d'IPP$^+$ effectivement correcte. La précision peut être calculée selon l'expression suivante :

$$\boldsymbol{Précision} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FP}} \ ;$$

ce qui revient à dire que la précision soit égale à 1 ou 100% si aucun cas de FP n'est observé.

Une précision élevée (proche de 1) indique que moins de paires IPP$^-$ sont considérés comme des paires IPP$^+$ et l'erreur de classification dans ce cas est mineure. Nous soulignons que l'un des objectifs de la détection des IPP$^+$ est d'associer une fonction particulière à une protéine dont on ignorait sa fonction par le biais de son interaction avec une protéine dont on connait sa fonction. Cela permet de classer certaines protéines comme des potentielles cibles thérapeutiques sur lesquelles les médicaments peuvent agir. Ainsi, l'erreur produite par la classification erronée (erreur de précision) est désavantageuse puisque dans ce cas une paire IPP$^-$ sera considérée comme IPP$^+$. La conséquence est que cette paire sera classée parmi les paires pouvant contribuer à l'élaboration d'une solution médicamenteuse. Or cette paire étant IPP$^-$, les protéines formant la paire ne constituent pas de potentielles cibles thérapeutiques.

- Sensibilité

La sensibilité ou le rappel permet de connaître la proportion de résultats IPP$^+$ réels qui a été identifiée correctement et peut être définie selon l'équation suivante :

$$\boldsymbol{Sensibilité} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FN}}$$

La sensibilité est une mesure de l'exhaustivité ou de la quantité. Une sensibilité élevée (proche de 1) indique que moins de paires IPP$^+$ sont considérés comme des paires IPP$^-$. Toutefois, l'erreur produite par la classification erronée dans un tel cas est aussi désavantageuse que dans le cas de la précision. En effet, une erreur dans un ce cas est qu'une paire IPP$^+$ est considérée comme IPP$^-$. La conséquence est que cette paire sera ignorée dans la prise en compte des paires pouvant contribuer à la mise en place d'une solution médicamenteuse.

Nous utilisons également comme métrique la courbe ROC et la valeur AUC.

- Courbe ROC et valeur AUC

La courbe ROC et la valeur AUC illustrent la performance d'un système de classification binaire par un graphique. Il s'agit d'un graphique du taux de faux positifs ou taux de fausses alarmes (axe des abscisses) qui indique le taux de vrais positifs (axe des ordonnées) pour plusieurs valeurs de seuil candidates différentes comprises entre 0.0 et 1.0. En d'autres termes, il compare le taux de se tromper au taux de réussite. La figure 3-2 illustre le tracé de la courbe ROC qui est représentée par la courbe en bleu avec différentes valeurs de seuil.

AUC signifie *Area Under Curve* ou aire sous la courbe en français. Cette valeur mesure en effet l'intégralité de l'aire à deux dimensions situées sous l'ensemble de la courbe ROC (par calculs d'intégrales) de (0.0) à (1.1). On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire. Un modèle dont 100 % des prédictions sont erronées à un AUC de 0.0. Si toutes ses prédictions sont correctes, son AUC est de 1.0.

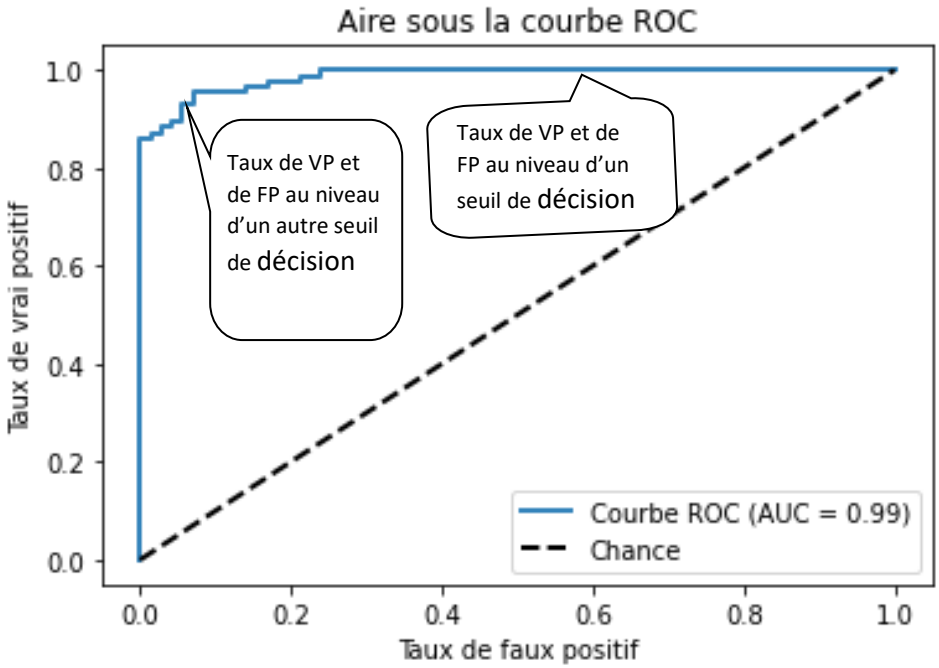


Figure 3-5: Illustration de la courbe ROC et la valeur AUC

3.5.1.2. *Expérimentation*

Nous avons considéré le classifieur SVM combiné à la nouvelle technique pour former un outil de prédiction d'interaction. L'entrainement du classifieur SVM a consisté à retrouver les meilleures valeurs de paramètres qui garantissent de meilleures performances de prédiction et une plus grande capacité de généralisation. Dans notre cas d'étude, nous avons subdivisé l'ensemble des observations IPP HPRD en données d'apprentissage pour la procédure interne (75%) et en données de tests pour la procédure externe (25%) comme dans [Wong et Hsu 2006 ; Du et al. 2017]. L'ensemble des données d'apprentissage a été également subdivisé en deux autres parties dont une partie en données d'entrainement (80%) et une deuxième partie en données de validation (20%).

Nous déterminons les meilleures valeurs de paramètres du SVM en utilisant la technique de la recherche sur grille (RG) combinée à une validation croisée 5-fois sur les données d'entrainement et de test dans la procédure interne comme dans [Brito et al. 2005 ; Bao et Liu 2006]. Pour un SVM avec noyau gaussien, deux paramètres influent sur la performance du modèle. Le paramètre de régularisation $\mathcal{C}$ qui contrôle le compromis entre la marge et l'erreur de mauvais classement et le paramètre gamma (γ) qui détermine l'étendue de l'influence d'un seul exemple d'entraînement [Huang et al. 2018]. La RG s'est effectuée sur les valeurs $\gamma \in \{1;\ 0.1;\ 0.01;\ 0.001;\ 0.0001\}$ et $C \in \{100; 50; 32; 10; 3; 1; 0.8\}$ comme dans la plupart des travaux récents [You et al. 2014 ; Wei et al. 2016 ; Göktepe et Kodaz 2018 ; Ma et al. 2020]. Les valeurs d'hyperparamètres obtenues sont $(C, \gamma) = (100; 0.01)$ et $(C, \gamma) = (100; 0.001)$ respectivement pour l'approche de la distance BP1 et pour l'approche de la fonction BP. La figure 3-6 nous montre que nous ne sommes pas en sous apprentissage ou en surapprentissage pour une valeur de $\gamma = 0.01$ (cas BP1). En effet, nous constatons que pour des valeurs de γ très inférieures à 0.01, les scores d'entrainement (courbe orange) et les scores de validation (courbe vert) sont faibles. Par-contre, pour $\gamma = 0.01$, nous avons les scores les plus élevés pour l'entrainement, tout comme pour la validation, ce qui traduit un bon apprentissage. Cependant, au-delà de valeurs de $\gamma > 0.01$, nous sommes en sur-apprentissage car les scores de validation deviennent de plus en faibles pendant que les scores d'entrainement restent inchangés.

De même, nous vérifions dans le cas BP que nous ne sommes pas en surapprentissage pour une valeur de $\gamma = 0.001$.

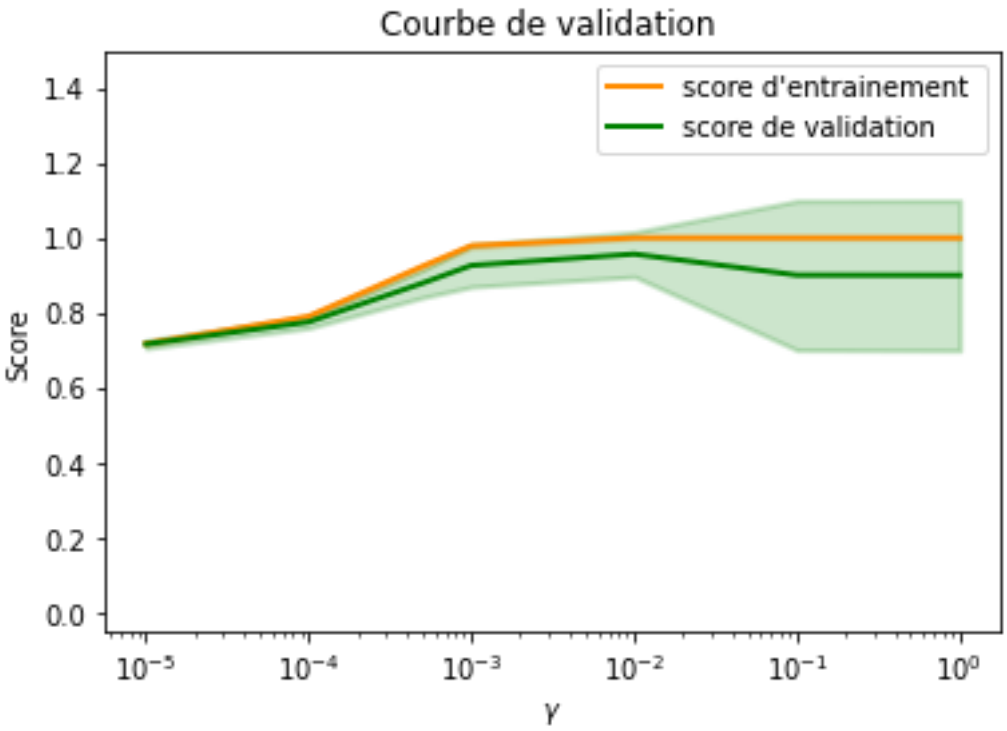


Figure 3-6: Courbe de validation et d'entrainement

3.5.2. Résultats obtenus sur les différents ensembles de données IPP

Nous montrons ici les résultats de performance obtenus sur les trois ensembles de données IPP que sont les ensembles de données HPRD, S. Cerevisiae et H. Pylori par la nouvelle technique BP.

3.5.2.1. *Résultats sur les données d'entrainement HPRD*

Nous comparons dans cette partie les performances moyennes de prédiction de la technique BP sur les données HPRD entre les approches utilisant la distance (BP1) et celle utilisant une fonction sur toutes les métriques citées plus haut après application de la technique de la validation croisée [Arlot and Celisse 2010]. Ici, nous avons appliqué la validation croisée 5-fois comme la plupart des auteurs dans ce domaine [You et al. 2013 ; An et al. 2019].

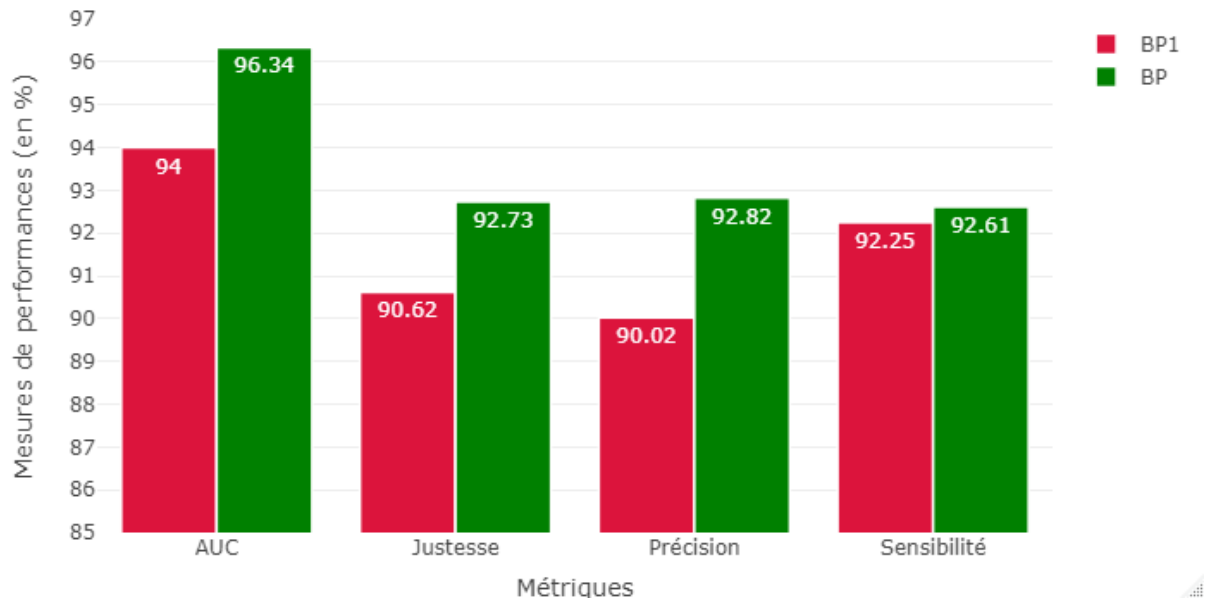


Figure 3-7: Comparaison des performances des approches de la technique BP

La figure 3-7 montre que l'approche utilisant une fonction (approche BP) en vert présente des taux élevés dans chaque métrique utilisée par rapport à celle utilisant une distance (approche BP1) en rouge. Nous enregistrons ainsi 96.34% contre 94% dans la métrique AUC, 92.73% contre 90.62% pour la justesse, 92.83% contre 90.02% pour la précision et 92.61% contre 92.25% pour la sensibilité. En effet, l'approche BP utilise une fonction pour calculer la matrice de scores physicochimiques. Cette approche obtient un vecteur caractéristique qui apporte plus d'informations utiles avec 0% de zéros dans ses composantes pour la reconnaissance des 'plis' des protéines. L''approche BP1 est en fait basée sur une distance. Or, le cas exemple avec l'approche BP1 vu à la section 3.3 a montré que le calcul de la matrice de scores physicochimiques avec la distance produit des composantes nulles dans le vecteur de caractéristiques bigrammes (environ 11% de zéros), contrairement à la fonction qui ne produit aucune composante nulle. Dans la suite du manuscrit, nous présentons uniquement les résultats obtenus avec l'approche de la fonction, c'est-à-dire l'approche BP.

3.5.2.2. *Résultats sur les données tests S. Cerevisiae et H. Pylori*

Dans le but de vérifier la capacité de généralisation de l'outil SVM-BP formé, nous avons évalué le modèle sur des ensembles de données non vus, c'est-à-dire qui n'ont pas été utilisés

dans la phase d'entrainement du modèle. Nous enregistrons, pour les données S. Cerevisiae, des performances moyennes de 90.68% ; 90.83%, 90.74% et 94.56% respectivement dans les métriques justesse, précision, sensibilité et AUC. En ce qui concerne les données H. Pylori, nous enregistrons plutôt une performance de 87.77% en justesse, 88.62% en précision, 87.43% en sensibilité et 91.27% en AUC.

Les performances moyennes obtenues sur les données S. Cerevisiae et H. Pylori dans les métriques justesse, précision, sensibilité et AUC sont respectivement toutes supérieures à 90% et 87%. Ces résultats indiquent que l'outil formé de la combinaison du classifieur SVM avec la technique d'extraction des bigrammes à partir de la matrice de scores physicochimiques arrive à garder des performances élevées sur les données tests. Aussi, le fait que les valeurs moyennes dans les différentes métriques sur les deux ensembles de données de façon générale soient supérieures à 87% traduit une bonne capacité de généralisation du modèle SVM-BP.

3.5.3. Comparaison des résultats avec les méthodes PF et *Bi-gram*

Dans les figures 3-7 et 3-8, nous comparons sur les données IPP HPRD les résultats de performance de prédiction obtenus en appliquant la technique 2-gramme sur la MSP (approche BP en vert) avec ceux obtenus en l'appliquant directement sur la séquence primaire d'une protéine (approche PF en rouge) ou sur la PSSM d'une protéine (approche *Bi-gram* en bleu). Pour l'implémentation de la technique *Bi-gram*, nous avons utilisé l'outil PSI-BLAST en ligne. Nous avons également choisi la même base de données de comparaison (Swiss-prot) que la plupart des travaux récents pour l'obtention des scores PSSM de chacune des séquences de protéine considéré dans notre ensemble de données HPRD [Dehzangi et al. 2017 ; An et al. 2019]. Les résultats de l'application de l'ACP pour la réduction des caractéristiques ont donné respectivement pour la méthode PF et *Bi-gram* 175 et 289 composantes. Pour ce qui est des valeurs optimales d'hyperparamètres, l'application de la RG a donné respectivement $(C, \gamma) = (8; 0.001)$ et $(C, \gamma) = (100; 0.001)$ pour les techniques PF et *Bi-gram*.

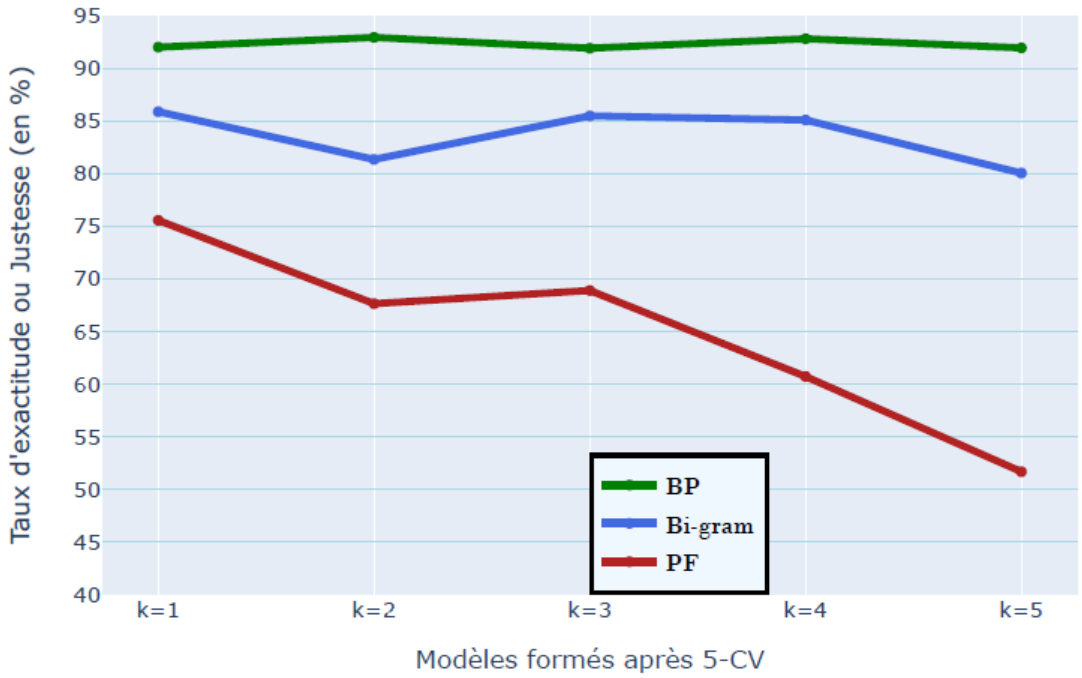


Figure 3-8: Comparaison de la justesse après une validation croisée 5-parties sur les IPP HPRD

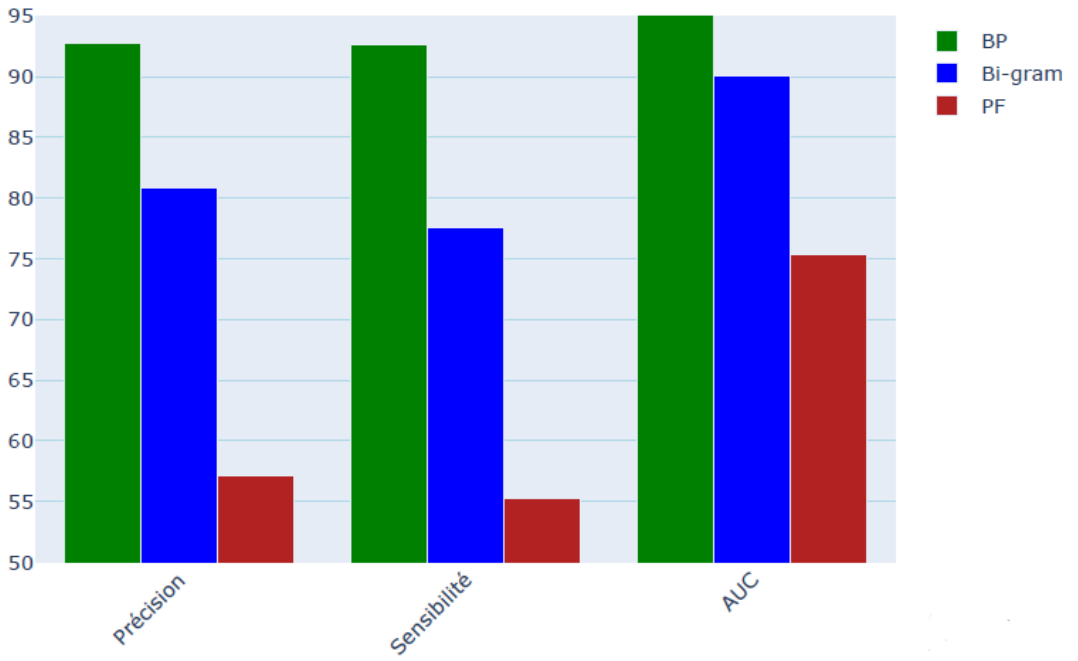


Figure 3-9: Comparaison des taux moyens de précision, sensibilité et AUC après une validation croisée 5-parties sur les IPP HPRD

La figure 3-8 indique que la méthode *Bi-gram* qui utilise les PSSM présente en moyenne une justesse qui se situe entre 80% et 85% dans chacune des phases de la validation croisée ($K = \{1 ; 2 ; 3 ; 4 ; 5\}$). Ces mêmes taux sont également observés sur la figure 3-9 dans les

métriques précision et sensibilité, excepté le taux dans la métrique AUC qui avoisine les 90%. En effet, le sens biologique du PSSM est d'exprimer l'évolutivité des séquences par des actions d'insertion ou de délétion des acides aminés [Sharma et al. 2013]. En outre, les paires de protéines présentant une grande similarité sont plus susceptibles d'interagir entre elles, par conséquent la méthode avec PSSM utilise les informations relatives à l'évolution pour prédire les interactions protéine-protéine. Pour ce qui concerne l'approche du bigramme directement sur la séquence primaire d'acides aminés (approche PF), les taux de justesse sont compris entre 50 et 75%. Nous constatons également des taux moyens de précision et sensibilité inférieur à 60% (figure 3-9). Ces faibles performances s'expliquent par le fait que le vecteur de caractéristiques bigramme généré par cette approche est strictement parcimonieux (compte tenu des nombreux zéros qu'il peut comporter) lorsqu'il est codé directement à partir de la séquence protéique. Cela joue sur les performances du classifieur utilisé pour prédire les interactions. Par-contre, pour la méthode BP, tout comme celle avec les PSSM, le vecteur formé ne présente pas le problème de vecteur strictement parcimonieux et apporte beaucoup plus d'informations pour inférer efficacement les interactions entre les protéines. En outre, nous pouvons constater sur les deux figures (8 et 9) que la technique BP présente des taux de justesse élevés sur chaque phase de la validation croisée, éventuellement des taux moyens supérieurs par rapport aux deux autres approches. Ces bonnes performances s'expliquent par le fait que notre technique BP ne représente pas seulement l'information sur la reconnaissance des 'plis' des protéines, mais conserve également suffisamment d'informations préalables provenant de la matrice de scores physicochimiques, qui a été générée à partir des propriétés amphiphiles (hydrophobicité et hydrophilie) et de la flexibilité des acides aminés. Les propriétés H_1 et H_2, représentant respectivement l'hydrophobicité et l'hydrophilie, utilisées dans la technique proposée sont des propriétés d'ordre fonctionnelles d'après [Chou 2005]. Les différents scores exprimés à partir de ces propriétés aideront à révéler si pour une paire de protéines, les deux protéines interagissent l'une avec l'autre. Ainsi, l'outil SVM-BP proposée ici utilise les informations relatives aux différentes fonctionnalités des protéines et prédit avec précision les interactions protéine-protéine.

Nous comparons dans la suite les résultats de performance dans toutes les métriques avec d'autres méthodes de codage de la séquence d'acides aminés d'une protéine utilisées dans la littérature.

3.5.4. Comparaison avec d'autres méthodes de codage sur les données HPRD

Nous comparons ici les performances de la technique développée avec d'autres méthodes d'extraction que nous avons implémentées. Il s'agit plus précisément des méthodes PseAAC, APAAC, AC et CTD [Chou 2001, 2005; Guo et al. 2008; You, Zhu, et al. 2014]. Les codes sources de ces différentes méthodes sont disponibles sur le serveur csbio[1]. Les méthodes APAAC et PseAAC utilisent également les propriétés physicochimiques des acides aminés et prennent un paramètre λ, qui indique le niveau ordre-séquence (voir figure 2-3 chapitre 2) [Chou 2001]. Ces deux méthodes sont intensives en calcul lorsque le paramètre λ est élevé. Nous avons choisi λ = 15 dans l'expérience comme dans [Du et al. 2017 ; Göktepe et Kodaz 2018]. La méthode AC utilise également les valeurs des propriétés physico-chimiques des acides aminés. Dans notre cas, nous avons choisi six propriétés physicochimiques comme dans [You, Yu, et al. 2014], notamment l'hydrophobicité, l'hydrophilie, la polarité, la polarisabilité, la surface accessible aux solvants. Les méthodes PseAAC, APAAC, AC et CTD produisent respectivement un vecteur de 35, 50, 180 et 120 composantes. Les résultats de la RG des valeurs optimales des hyperparamètres de ces différentes méthodes sont renseignés dans le tableau 3-7.

Tableau 3-7: Valeurs d'hyperparamètres des différentes méthodes

Méthodes	$(\boldsymbol{C};\ \boldsymbol{\gamma})$
PseAAC	(100; 0.001)
APAAC	(100; 0.001)
AC	(50; 0.001)
CTD	(1; 0.1)

Tableau 3-8: Comparaison avec différentes méthodes d'extraction sur les données HPRD

Méthode	Justesse	Précision	Sensibilité
PseAAC	87.84%	89.97%	86.30%
APAAC	91.07%	92.52%	91.53%
AC	74.96%	73.91%	77.15%
CTD	70.01%	69.31%	91.76%
BP (notre étude)	**92.73%**	**92.83%**	**92.60%**

[1] www.csbio.sjtu.edu.cn/bioinf/

Les résultats indiqués dans le tableau 3-8 montrent que la méthode APAAC obtient les meilleures performances dans toutes les métriques par rapport aux autres méthodes proposées avec une justesse égale à 91.07%, une précision égale à 92.52% et une sensibilité égale à 90.53%. Cependant, les performances de prédiction affichées par la technique BP surpassent celles de l'approche APAAC dans toutes les métriques. Nous enregistrons donc un taux d'environ 0.66% de plus pour la justesse (92.73% contre 91.07%), environ 0.33% de plus pour la précision (92.83% contre 91.52%) et environ 1.07% de plus en sensibilité (92.60% contre 90.53%).

La figure 3-10 indique également que nous réalisons une performance AUC supérieure aux performances affichées par les autres méthodes. En revanche, la méthode APAAC (en vert) affiche une performance AUC proche de la nôtre avec un taux de 95%. Ceci est dû au fait que la méthode BP utilise les mêmes informations de la séquence que la méthode APAAC pour représenter la protéine. La méthode APAAC combine la technique de composition en acide aminé (AAC) avec une fonction de corrélation linéaire définie également à partir des propriétés amphiphiles (hydrophobicité et hydrophilie) des acides aminés. Par conséquent les informations codées peuvent présenter les mêmes caractéristiques que la méthode BP.

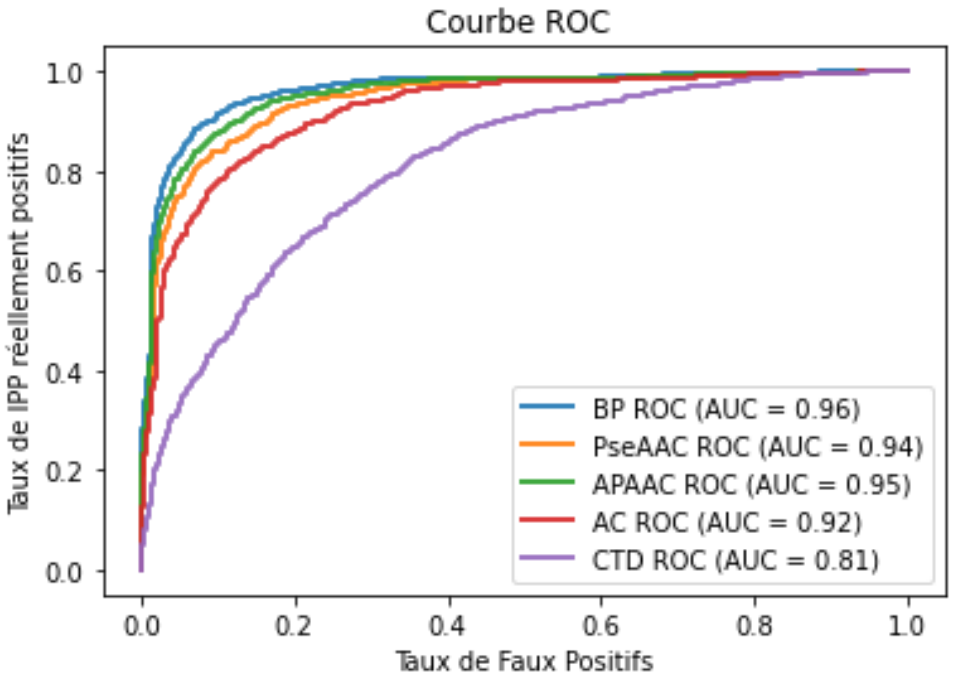


Figure 3-10: Comparaison des courbes ROC de différentes méthodes sur les données HPRD

Nous pouvons conclure que comparé à certaines méthodes d'extraction de caractéristiques existantes, la technique BP qui code les informations de bigrammes de la séquence d'une protéine par la technique 2-gramme sur la matrice de scores MSP présente des performances supérieures sur l'ensemble des métriques utilisées. Par conséquent, la technique BP proposée code bien qualitativement les informations essentielles de la paire de protéines à partir des informations de la séquence que les méthodes existantes.

3.5.5. Comparaison en termes d'outil de prédiction formé avec d'autres auteurs

Nous présentons ici les résultats de comparaison de l'outil SVM-BP avec des outils existants. Mise à part la métrique AUC, toutes les autres métriques sont utilisées pour la comparaison. Cela est justifié par le fait que la plupart des auteurs comparés ne présentent pas de résultats dans la métrique AUC.

3.5.5.1. *Comparaison sur les données IPP HPRD*

Le tableau 3-9 indique les résultats comparaison de performances obtenus par nos travaux avec ceux obtenus par les auteurs [You et al. 2013], [Göktepe et Kodaz 2018], [An et al. 2019] et [Ma et al. 2020] les données IPP HPRD .

Tableau 3-9: Comparaison des performances sur les données HPRD avec d'autres auteurs

Méthode	Justesse	Précision	Sensibilité
[You et al. 2013]	84.80%	85.47%	84.08%
[Göktepe et Kodaz. 2018]	73.81%	74.11%	73.24%
[An et al. 2019]	90.40%	88.03%	93.54%
[Ma et al. 2020]	75.82%	78.24%	72.74%
SVM-BP (nos travaux)	**92.73%**	**92.83%**	**92.60%**

Les résultats du tableau montrent que nos travaux obtiennent des performances supérieures dans toutes les métriques que ceux des auteurs cités plus haut. Nous obtenons une justesse de 92.73%, une précision de 92.83% et une sensibilité de 92.60%. cependant, nous pouvons observer que les travaux de [An et al. 2019] obtiennent le taux le plus élevé en sensibilité avec 93.54%. Toutefois, la méthode proposée par [An et al. 2019] est lente en exécution et

nos travaux présentent des taux supérieurs dans les métriques justesse et précision par rapport aux taux affichés dans les travaux de [An et al. 2019].

3.5.5.2. *Les résultats de performance obtenus sur les données IPP S. Cerevisiae*

La figure 3-11 donne les résultats de performances sur les données IPP S. Cerevisiae comparés aux outils formés avec les méthodes LD (Local Descriptor) [Zhou et al. 2011], ELM (*Extreme Learning Machine*) [You et al. 2013], MCD (*Multiscale Continuous Discontinuous*) [You et al. 2014] , PSSM [Wang et al. 2017] et LCPSSMMF (*Local Coding Position Scoring Specific Matrix with Multifeatures Fusion*) [An et al. 2019]. Les métriques utilisées pour la comparaison sont la justesse, la précision et la sensibilité.

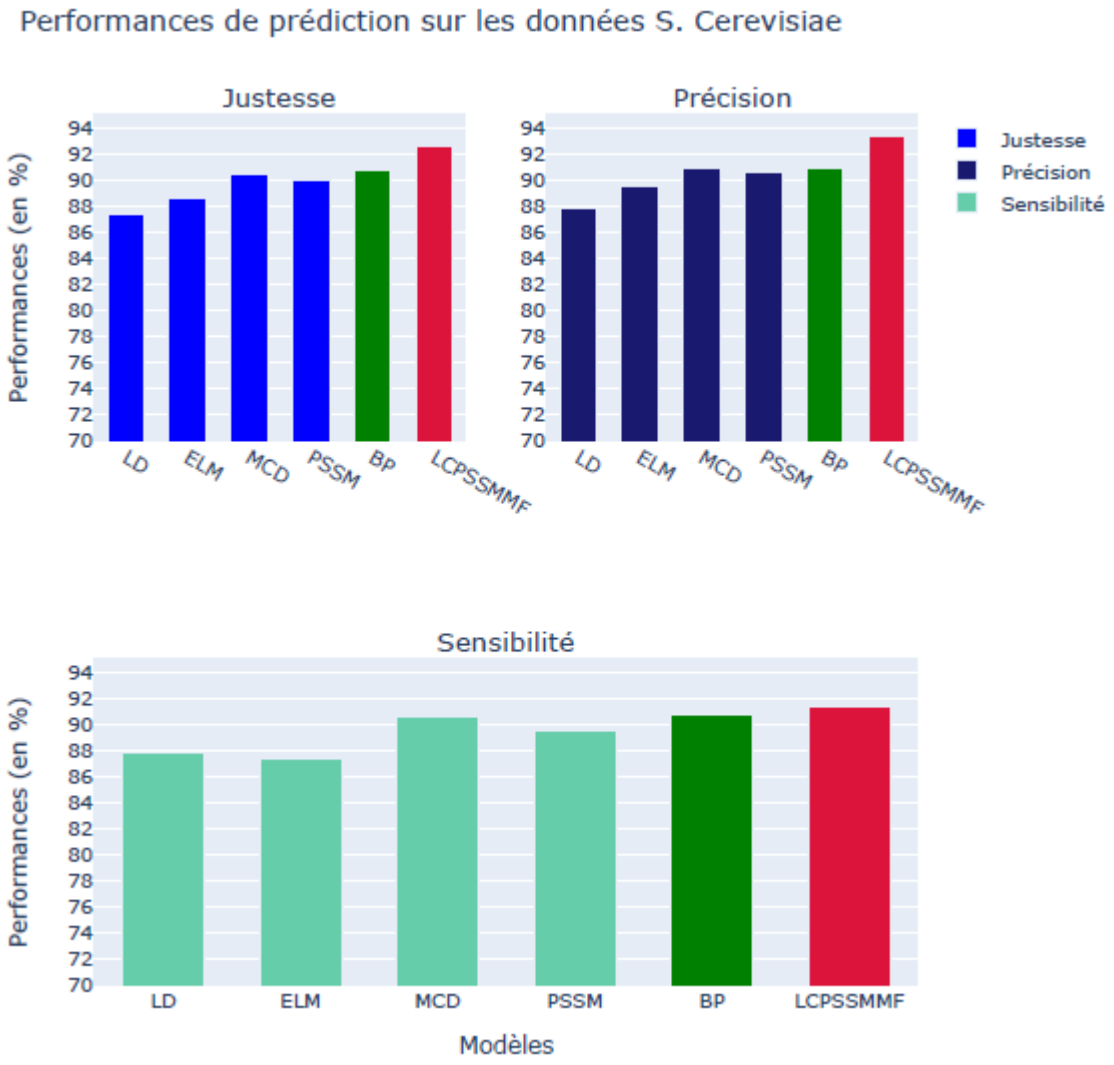


Figure 3-11: Comparaison des performances avec d'autres auteurs sur les données S. Cerevisiae

Nous notons sur cette figure que les taux affichés par la méthode BP sont supérieurs aux taux présentés par la plupart des méthodes, excepté la méthode LCPSSMMF, qui présente des performances de 92.54%, 93.40% et 91.40% respectivement dans les métriques justesse, précision et sensibilité. Il faut souligner que cette méthode est une autre version de la méthode *Bi-gram* qui calcule plusieurs PSSM avant d'appliquer la technique 2-gramme pour coder la protéine. Les auteurs [An et al. 2019] utilisent l'outil PSI-BLAST pour avoir la PSSM de chaque protéine. Par la suite, ils divisent la PSSM en 4 autres PSSM constitué des premiers 75%, les derniers 75% et les 75% du milieu du PSSM originale qu'ils nomment parties A, B et C. La matrice PSSM entière est représentée dans la quatrième partie D. Enfin, une nouvelle PSSM (CPSSM) est créée en fusionnant les sous-PSSM A, B, C et D où la longueur de la CPSSM est plus longue que celle de la PSSM originale. Pour finir, ils calculent les caractéristiques bigrammes sur la nouvelle PSSM. Les taux supérieurs observés avec méthode peuvent s'expliquer par le fait que les séquences de la base de données choisie pour avoir les PSSM des protéines présentent une forte homologie avec les séquences requêtes de l'ensemble de données S. Cerevisiae. Cependant, une telle démarche est lente dans l'exécution et les performances de prédiction dépendent de la base de données choisie. En effet, la base de données choisie dans la technique LCPSSMMF pour la comparaison des séquences à travers l'outil PSI-BLAST est la base de données 'nr'. Cette base de données comporte des millions de séquences (> 467 millions). Pour obtenir la PSSM d'une séquence de longueur 200 par exemple (200 acides aminés), cette méthode met environ 220 secondes, soit 3 minutes 40 secondes. Or notre technique BP ne met que 41 secondes pour extraire les caractéristiques bigrammes de cette même séquence. Nous rappelons que nous travaillons avec une machine core i7 tournant au minimum à 256 GHz. En outre, les résultats de performances sur les données HPRD présentées par notre technique sont supérieurs dans l'ensemble aux résultats de performances affichés par la technique LCPSSMMF.

3.5.5.3. *Comparaison avec d'autres auteurs sur les données IPP H. Pylori*

Les résultats de tests sur l'ensemble de données IPP H. Pylori sont donnés dans le tableau 3-10. Les performances de prédiction obtenues par la combinaison du classifieur SVM avec l'approche d'extraction des caractéristiques bigramme à partir de la MSP sont comparées à celles obtenues par les modèles de certains auteurs à savoir [Zhou et al. 2011], [You et al. 2013], [You, Zhu, et al. 2014], [Göktepe et Kodaz 2018] et [Ma et al. 2020]. Nous pouvons noter que la plupart des outils de la littérature affichent des performances moyennes de 80%. Cela s'explique par le fait que les performances des modèles sont affectées par le peu

d'observations pour ce qui est des données H. Pylori (seulement 2896 observations). Nous pouvons constater également que les résultats de performance affiché dans l'ensemble des métriques par la méthode BP sont supérieurs que ceux affichés dans la plupart des travaux existants.

Tableau 3-10: Comparaison de performances sur les données H. Pylori avec d'autres auteurs

Modèles	Justesse	Précision	Sensibilité
[Zhou et al. 2011]	84.20%	86.30%	85.10%
[You et al. 2013]	87.50%	86.15%	88.65%
[You, Zhu, et al. 2014]	84.91%	86.12%	83.24%
[Göktepe et Kodaz 2018]	86.23%	84.32%	89.44%
[Ma et al. 2020]	88.96%	86.86%	91.86%
BP	**87.77%**	**88.02%**	**87.43%**

Les différents résultats de performance de prédiction présentés tout au long de cette section montre que la technique BP proposée combinée au classifieur SVM sépare bien les différentes classes d'IPP que la plupart des méthodes existantes. De plus, les performances moyennes obtenues dans les données de tests que sont les données S. Cerevisiae et H. Pylori sont tout aussi élevées que celles observées sur les données d'entrainement. Par conséquent, l'outil SVM-BP formé présente une bonne capacité de généralisation.

3.6. Discussion

Les principales techniques utilisées dans cette étude sont BP (*Bigram Physicochemical*) pour la représentation de la protéine par des caractéristiques bigrammes et SVM pour la classification. Les résultats de prédiction sur différents ensembles de données indiquent que BP fait mieux que d'autres techniques d'extraction de caractéristiques de la séquence d'acides aminés d'une protéine. En outre, les outils formés SVM-BP et SVM-BP1 constituent une aide dans le problème de prédiction des interactions protéine-protéine.

Un gros avantage de la technique BP est la possibilité de coder ou extraire les caractéristiques de la séquence sous divers angles en fonction des propriétés physicochimiques utilisées. En effet, comme vu au chapitre 2, nous avons plusieurs propriétés physicochimiques

des acides aminés et certaines sont d'ordre fonctionnelles comme les propriétés hydrophobicité et hydrophilie considérées dans l'approche BP. Nous pouvons par exemple codées les caractéristiques des protéines qui sont d'ordre structurelles en ajoutant aux propriétés hydrophobicité et hydrophilie d'autres propriétés vues au chapitre précédent comme la polarité, la polarisabilité, le volume de la chaîne latérale, et bien d'autres.

Comparés aux méthodes d'extraction des bigrammes que sont la technique PF (*Pairwuze Frequency*) [Yang et al. 2011] qui extrait les bigrammes directement à partir de la séquence primaire d'une protéine et la technique *Bi-gram* [An et al. 2019] qui extrait les bigrammes sur une PSSM, la technique BP à travers l'outil SVM-BP a présenté des taux supérieurs de prédiction sur tous les ensembles de données de formation (voir figures 3-8 et 3-9). Par conséquent, nous pouvons dire que la technique BP qui extrait les bigrammes à partir de la matrice de scores physicochimiques arrive à mieux représenter les caractéristiques bigrammes (reconnais mieux les 'plis' d'une protéine) que les techniques PF et *Bi-gram*.

Pour ce qui est de la performance de l'outil formé SVM-BP, nous retenons que sur les trois ensembles de données IPP, la technique BP a présenté des performances de prédiction supérieures avec un taux de justesse moyen de 92.73% sur les données HPRD, 90.68% sur les données S. Cerevisiae et 87.77% sur les données H, Pylori. Nous notons cependant que sur les données S. Cerevisiae, les travaux de [An et al. 2019] ont présenté des taux élevés sur toutes les métriques par rapport aux résultats avec la technique BP. Toutefois, les étapes pour représenter la protéine par la technique proposée par [An et al. 2019] sont longues dans l'exécution (environ 4 minutes pour une séquence de 200 acides aminés contre 41 secondes avec la technique BP) et les taux affichés sur les données HPRD sont nettement inférieurs aux taux affichés par la technique BP. Nous retenons que dans l'ensemble les différents résultats obtenus par la technique BP proposée sont supérieurs aux résultats présentés par certains travaux récents (voir tableau 3-9). Par conséquent, la technique BP représente efficacement les informations essentielles de la paire de protéine et améliore la prédiction d'interaction protéine-protéine.

L'outil SVM-BP a montré de bonnes performances en générales dans toutes les métriques sur les données de formation HPRD ainsi que les données de tests S. Cerevisiae et H. Pylori (voir figure 3-10 et tableau 3-10). En outre, la précision de l'outil formé se situe entre 90% et 93%. Nous pensons que nous pouvons améliorer les performances de l'outil formé avec le classifieur SVM en recherchant de manière plus rigoureuse les valeurs optimales des paramètres influents du classifieur que sont ses hyperparamètres comme suggéré par [Anguita et al. 2012].

Conclusion

L'extraction de caractéristiques à partir des informations de la séquence d'acides aminés d'une protéine est une étape cruciale dans la mise en place d'un outil de prédiction d'interaction protéine-protéine. Dans ce chapitre, nous avons proposé une nouvelle approche d'extraction des caractéristiques de séquences de protéines combinée à l'algorithme d'apprentissage SVM pour prédire les interactions protéine-protéine. Nous avons montré qu'au lieu de calculer les bigrammes à partir de la séquence primaire de protéine, ou à partir de la PSSM, nous pouvons les calculer à partir de la matrice de scores physicochimiques. La technique proposée a été testée et validée sur trois ensembles de données IPP différents. L'efficacité de la technique BP a été évaluée particulièrement à travers l'outil SVM-BP dans un premier temps par rapport à d'autres techniques d'extraction des caractéristiques bigrammes. Les résultats présentés dans ce chapitre indiquent que la technique BP proposée améliore jusqu'à plus de 10% la justesse de reconnaissance des 'plis' des protéines et également la reconnaissance des classes d'interactions des paires de protéines. Deuxièmement, l'efficacité de la technique BP a été évaluée par rapport à plusieurs techniques de représentation d'une protéine et des résultats très prometteurs ont été obtenus. Bien vrai que la technique BP présente des taux de réussite supérieurs à certaines techniques de la littérature, comme mentionné dans la discussion, nous pouvons améliorer les performances de l'outil SVM-BP. Cela passe par une recherche rigoureuse des valeurs optimales des paramètres du classifieur SVM qui fait l'objet du prochain chapitre.

CHAPITRE 4. NOUVELLE APPROCHE DE SELECTION DES VALEURS OPTIMALES D'HYPERPARAMETRES

SOMMAIRE

Introduction

Dans la littérature sur l'apprentissage automatique, la technique de la recherche sur grille combinée à celle de la validation croisée *k*-fois sont le plus souvent utilisées pour retrouver les valeurs optimales de paramètres influents ou hyperparamètres du classifieur utilisé [Yang et Shami 2020]. La valeur *k* représentant le nombre de sous ensemble de l'ensemble d'apprentissage. Dans le processus de la validation croisée, la valeur de *k* du nombre de sous ensemble est choisie et fixée de manière aprioristique (sans aucune expérience). Cependant, selon [Arlot et Celisse 2010], la valeur de *k* agit sur le choix du meilleur compromis entre l'erreur d'estimation et l'erreur d'approximation du modèle. Ainsi, la valeur *k* du nombre de sous-ensembles peut influencer sévèrement les valeurs optimales d'hyperparamètres et donc par conséquent agir sur la performance du modèle sélectionné et sa capacité de généralisation [Hastie et al. 2009 ; Hsu et al. 2003].

Dans ce chapitre, nous proposons une approche rigoureuse de recherche d'hyperparamètres noté SVOH (Sélection de Valeurs Optimales d'Hyperparamètres). L'approche proposée considère la valeur *k* du nombre de sous-ensembles comme un paramètre influent du modèle et fait donc l'apprentissage pour retrouver une valeur optimale de *k*. La section **4.1 p**résente deux des techniques de recherche des valeurs d'hyperparamètres. Nous montrons dans cette section comment les valeurs d'hyperparamètre agissent sur la performance du classifieur SVM. En effet, pour le noyau gaussien qui est celui considéré dans notre étude, deux principaux hyperparamètres sont à optimiser pour que la fonction de décision puisse présenter de bonnes performances et faire moins d'erreur sur de nouvelles observations [Anguita et al. 2012]. La section **4.2** est consacrée à l'approche SVOH que nous avons développée pour rechercher de façon rigoureuse les valeurs d'hyperparamètres. Les résultats expérimentaux sont exposés à la section **4.3.** Dans la section **4.4** nous présentons d'autres résultats de validation de l'approche proposée. Enfin, nous discutons l'approche proposée dans la section **4.5.**

4.1. Les machines à vecteurs de supports : recherche de paramètres

4.1.1. Hyperparamètres des machines à vecteurs de supports

Les machines à vecteurs de supports (SVM) appartiennent au domaine des réseaux de neurones artificiels (RNA) [Scholkopf et al. 1997], mais se caractérisent par les solides fondations de la théorie de l'apprentissage statistique (TAS) [Vapnik 2013]. Dans la littérature sur l'apprentissage automatique, le processus de recherche des meilleurs hyperparamètres (paramètres influents) est généralement appelé phase de sélection de modèles [Yang et Shami

2020]. Cette phase est strictement liée à l'évaluation de la capacité de généralisation du classifieur [Anguita et al. 2009] ou, en d'autres termes, au taux d'erreur que le classifieur peut atteindre sur de nouvelles observations. En fait, un SVM optimal est obtenu en sélectionnant les hyperparamètres optimaux, c'est-à-dire ceux qui permettent au SVM de présenter l'erreur de généralisation la plus faible.

Si nous Considérons un ensemble d'apprentissage $\mathcal{Z} = \{(x_i, y_i), i \in [1, n]\}$ où à chaque vecteur $x \in \mathbb{R}^p$ est associé une valeur $y \in \{-1, +1\}$. La relation entre $\mathcal{X}$ et $\mathcal{Y}$ est encapsulée dans une distribution inconnue $P(\mathcal{X}, \mathcal{Y})$, qui est à l'origine des données. Le but de l'apprentissage est de trouver une fonction f: $\mathbb{R}^d \to \mathcal{Y}_f \subset \mathbb{R}$ qui se rapproche de cette relation. L'algorithme SVM [Vapnik 1999] peut être exploité à cette fin, où le classificateur est identifié pendant la phase de recherche des hyperparamètres en résolvant le problème quadratique convexe suivant :

$$Maximiser \qquad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \cdot y_i y_j \cdot h(x, x_j)$$

$$sous\ réserve\ de \qquad 0 \leq \alpha_i \leq \mathcal{C}, \qquad i = 1, .. n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

où les α_i sont les multiplicateurs de Lagrange et $\mathcal{C}$ un des hyperparamètres, qui contrôle le compromis entre la marge et l'erreur de mauvais classement et $h(x_i, x_j)$, la fonction noyau. Le noyau considéré ici est le noyau gaussien. Le noyau gaussien est dérivé de la fonction RBF (*Radial Basic Function*) et dépend de la distance euclidienne entre les vecteurs x_i, x_j dans l'espace de départ. Il est défini de la manière suivante :

$$h(x_i, x_j) = \exp\left(\frac{\| x_i - x_j \|^2}{2\gamma^2}\right)$$

avec γ un hyperparamètre additionnel qui détermine l'étendue de l'influence d'un seul exemple d'entraînement [Göktepe and Kodaz 2018]. La résolution du problème quadratique convexe permet d'avoir un classifieur comme défini à l'équation 2-1 du chapitre 2 de la manière suivante :

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i h(x_i, x_j) + b$$

où b représente le biais.

Les deux hyperparamètres $\mathcal{C}$ et γ sont donc les paramètres influents du classifieur SVM lui permettant d'estimer l'erreur de généralisation.

Pour montrer comment le choix des valeurs de $\mathcal{C}$ et γ jouent sur la performance du modèle SVM, considérons une répartition de 35 observations d'interaction protéine-protéine (IPP) dont 15 IPP négatives (carrés rouges) et 20 IPP positives (croix bleues). Nous présentons à travers la figure 4-1 quatre situations qui résument bien les comportements asymptotiques du classificateur SVM en fonction des valeurs de $\mathcal{C}$ et γ comme décrits dans [Keerthi et Lin 2003].

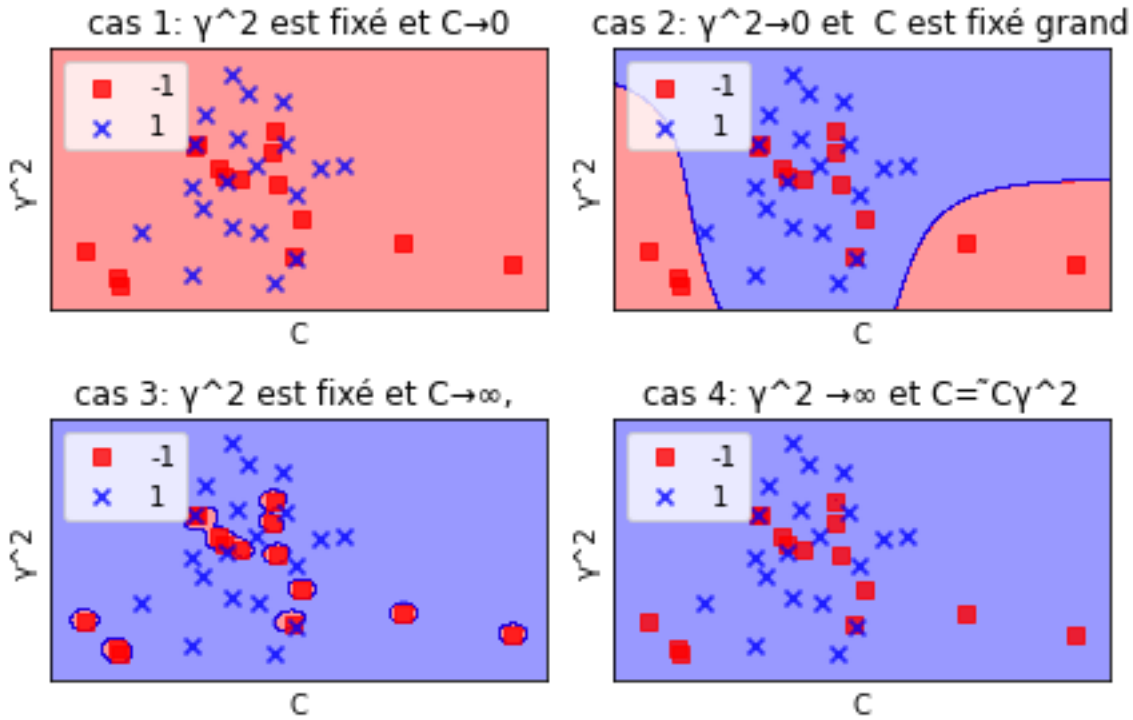


Figure 4-1: Figuration du comportement asymptomatique du SVM Gaussien pour différentes valeurs de $\mathcal{C}$ et γ

- Cas 1 : Sous apprentissage (*underfiting*), c'est lorsque l'espace de données entier est attribué à la classe minoritaire et peut se produire dans les scénarios suivants :
  - γ^2 est fixé et $\mathcal{C} \to 0$;

- $\gamma^2 \to 0$ et $\mathcal{C}$ est fixé à une valeur suffisamment petite ;
- $\gamma^2 \to \infty$ et $\mathcal{C}$ est fixé.

Les trois scénarios possibles du cas 1 vous sont présentés en annexe B.

- Cas 2 : Surapprentissage sévère (*severe overfiting*), c'est lorsque de petites régions autour des exemples de formation de la classe minoritaire sont classées comme étant cette classe, tandis que le reste de l'espace de données est classé comme étant la classe majoritaire, se produit dans le cas où $\gamma^2 \to 0$ et $\mathcal{C}$ est fixé à une valeur suffisamment grande.

- Cas 3 : Si γ^2 est fixé et $\mathcal{C} \to \infty$, le classificateur SVM sépare strictement des exemples de formation des deux classes. Il s'agit d'un cas de surapprentissage, si le problème considéré comporte du bruit.

- Cas 4 : Si $\gamma^2 \to \infty$ et $\mathcal{C} = \tilde{\mathcal{C}}\gamma^2$, avec $\tilde{\mathcal{C}}$ fixé, alors le classifieur SVM converge vers le classifieur SVM linéaire avec le paramètre de pénalité $\tilde{\mathcal{C}}$.

Le cas 1 montre qu'une faible valeur de $\mathcal{C}$ et une valeur de γ élevée entraînent un sous-apprentissage (sous-ajustement). Inversement, une valeur $\mathcal{C}$ élevée et une valeur γ faible entraîneront un surapprentissage (surajustement) représenté par le cas 2. Plus la valeur de $\mathcal{C}$ est petite, plus elle ignorera le point caractéristique en tant que vecteur de support existant près de l'hyperplan ; ainsi que l'augmentation de la marge maximale (cas 1). Plus la valeur de γ est élevée, plus le vecteur de support augmente et plus la flexibilité de la frontière de décision (hyperplan) est importante.

La paire d'hyperparamètres ($\mathcal{C}$, γ) doit être réglée pendant la phase de sélection de modèles. Bien que certaines méthodes empiriques aient été proposées pour rechercher les hyperparamètres d'une manière très simple et efficace (par exemple dans [Milenova et al. 2005]), la procédure la plus utilisée et la plus efficace est la recherche sur grille combinée à la validation croisée *k*-fois, où le problème quadratique convexe est résolu pour plusieurs hyperparamètres et les modèles sont comparés en évaluant leur performance sur des données non vues [Anguita et al. 2012 ; Yang et Shami 2020].

4.1.2. Recherche de valeurs d'hyperparamètres

Nous venons de voir que la sélection des valeurs des hyperparamètres $\mathcal{C}$ et γ joue sur la performance du classifieur SVM gaussien. La recherche sur grille est souvent combinée à la validation croisée *k*-fois pour retrouver les valeurs d'hyperparamètre d'un modèle d'apprentissage automatique [Ranjan et al. 2019].

La recherche sur grille (RG) est l'une des méthodes les plus couramment utilisées pour explorer l'espace de configuration des hyperparamètres [Brito et al. 2005 ; Yang and Shami 2020]. La RG peut être considérée comme une recherche exhaustive ou une méthode de force brute qui évalue toutes les combinaisons d'hyperparamètres données à la grille de configurations [Injadat et al. 2020]. Elle fonctionne en évaluant le produit cartésien d'un ensemble fini de valeurs spécifiées par l'utilisateur. Pour chaque combinaison possible de paramètres, les modèles de classification correspondants sont formés et évalués. Considérons notre fonction de SVM avec les hyperparamètres $\mathcal{C}$ et γ , mathématiquement, le fonctionnement de la RG peut être formulé comme suit :

$$\max_{\mathcal{C},\gamma} \not{f}(\mathcal{C},\gamma)$$

où $\not{f}$ est une fonction objective à maximiser, généralement le taux de justesse du modèle et $(\mathcal{C},\gamma)$ est le couple d'hyperparamètres à régler. Pour identifier les optimums globaux en testant différents ensembles d'apprentissage, la méthode de validation croisée *k*-fois est appliquée dans le processus de recherche sur grille [Arlot and Celisse 2010; Berrar 2019]. Le principe de la validation croisée *k*-fois est que l'ensemble d'apprentissage $\mathcal{Z}$ est divisé en *k* sous-ensembles de manière aléatoire, chacun d'eux étant constituée de $\mathcal{Z}/k$ échantillons, où, *k*-1 sous-ensembles sont utilisés, à tour de rôle, comme ensemble d'entrainement et la partie restante est utilisée comme ensemble de validation. Pour *k* sous-ensembles, le test est effectué *k* fois par la méthode *leave-one-out* [Kearns and Ron 1999], c'est-à-dire qu'une partie est utilisée de manière interchangeable dans l'ensemble de données de test et les (*k*-1) autres ensembles sont utilisés comme ensemble de données d'entraînement. Dans chaque test, la combinaison des paramètres $\mathcal{C}$ et γ est effectuée *k* fois avec *k* jeux de données d'entraînement et de test différents. Ceci afin de s'assurer qu'il n'y a pas de surajustement excessif sur le test avec des données de test différentes. L'erreur réalisée par le SVM entraîné sur l'ensemble de validation peut être utilisée pour estimer de manière fiable l'erreur de généralisation (erreur sur de nouvelles observations), car elle n'a pas été utilisée pour l'entraînement du modèle. L'estimation du taux d'erreur est la proportion globale d'erreur encourue sur tous les sous-ensembles de test. La validation croisée *leave-one-out* est le cas particulier de la validation

croisée *k*-fois où $k = \mathcal{Z}$, et son coût de calcul devient plus élevé en proportion de la taille de l'échantillon d'entraînement.

Obtenir une estimation précise et rigoureuse de l'erreur de généralisation d'un classifieur est un moyen de garantir, au sens statistique, la fiabilité du modèle [Yang et Shami 2020]. Les paramètres $\mathcal{C}$ et γ avec les bonnes valeurs peuvent maintenir un faible biais (une mesure de la contribution de l'erreur) et une faible variance (mesure des déviations) lors de l'utilisation d'un ensemble d'apprentissage différent avec la méthode de validation croisée *k*-fois [Budiman 2019]. La plupart des auteurs utilisent un nombre fixe de sous-ensembles [Bengio and Grandvalet 2004; Budiman 2019]. D'après [Hsu et al. 2003], les méthodes empiriques en effet suggèrent de fixer les valeurs de *k* à 5, 10 ou 20. Cependant, un seul fractionnement n'est souvent pas suffisant pour obtenir des modèles fiables et les estimations d'erreur correspondantes, comme cela a été rapporté dans la littérature [Anguita et al. 2012 ; Arlot and Celisse 2010]. Le problème que nous voulons traiter ici est la sélection des valeurs optimales d'hyperparamètres du modèle SVM qui passe par le choix du nombre *k* de sous-ensembles permettant au modèle de SVM de minimiser l'erreur de généralisation non pas en se basant sur des valeurs fixées, mais plutôt en faisant un apprentissage de valeurs pour le nombre *k* de sous-ensembles. Dans la section suivante nous proposons une recherche du nombre de subdivision *k* pour la phase de validation croisée afin d'augmenter les chances d'obtenir des modèles fiables.

4.2. Nouvelle approche de sélection de valeurs optimales d'hyperparamètres

4.1.3. Problème à résoudre

Le choix d'une valeur fixe de sous-ensembles pour la validation croisée peut produire un modèle avec un biais et une variance élevée. En particulier, il peut être montré que la variance de l'erreur obtenue en utilisant les *k* sous-ensembles pendant la phase de recherche de paramètres, peut être importante dans certains cas [Arlot and Celisse 2010]. La validation croisée fait en effet la moyenne de plusieurs estimateurs du risque de retenue correspondant à différents fractionnements de données. Dans [Anguita et al. 2009] nous pouvons vérifier que la valeur *k* influence la stabilité de la moyenne des erreurs. Toujours selon [Arlot and Celisse 2010], les performances de sélection de modèles avec la validation croisée sont généralement optimales lorsque la variance est aussi faible que possible. Cette variance diminue généralement lorsque le nombre *k* de sous-ensembles augmente, avec une taille d'échantillon d'entraînement fixe *n*. Lorsque *k* est fixe, la variance de la validation croisée dépend également de *n*. en effet, dans [Bengio and Grandvalet 2004], nous pouvons voir que la valeur de γ

dépend fortement de l'ensemble d'apprentissage utilisé. Le choix de k influe donc sur la variance de l'estimateur de la validation croisée et selon [Hsu et al. 2003 ; Hastie et al. 2009 ; Anguita et al. 2012], elle peut avoir un impact significatif sur la recherche des valeurs optimales d'hyperparamètres. Comme alternative, nous proposons d'utiliser la procédure qui consiste à considérer un certain nombre de combinaisons possibles de sous-ensembles dans lesquels l'ensemble d'apprentissage original peut être divisé. L'objectif est de choisir une meilleure procédure d'estimation de la validation croisée, celle qui présente le biais et la variance les plus faibles, permettant d'identifier une bonne combinaison d'hyperparamètres $(\mathcal{C}, \gamma)$ afin que le classificateur puisse présenter une faible erreur de généralisation et prédire des données inconnues avec un taux de justesse supérieur.

Pour l'approche proposée, nous considérerons le nombre k en tant qu'hyperparamètre comme dans [Anguita et al. 2012], qui peut prendre n'importe quelle valeur dans l'ensemble $k \in \{i, \ldots, 10\}$, $i \geq 3$. La plus petite valeur de k est fixée à 3 car pour chaque sous-ensemble, les données d'entrainement doivent être supérieures à 60% de l'ensemble d'apprentissage comme le montre [Laura 2015]. Ici, nous fixons la plus grande valeur de k à 10 pour rester dans la marge fixée par les méthodes empiriques. Ce choix limité des valeurs test de k à 10 permet également, dans les cas où l'ensemble d'apprentissage est large, d'éviter que la technique soit beaucoup gourmande en calcul. En effet, en supposant qu'il y a q paramètres, et que chacun d'entre eux a m valeurs distinctes, sa complexité de calcul augmente exponentiellement à un taux de $\mathcal{O}(m^q)$ [Brito et al. 2005 ; Yang et Shami 2020]. De plus, dans [Anguita et al. 2012], nous pouvons voir que plus de 10 bases de données différentes ont produit une valeur optimale k inférieure à 10. Le nombre de paramètres à optimiser pour notre cas devient donc le triplé $(\mathcal{C}, \gamma, k)$, étant donné que notre fonction de décision f utilise un noyau gaussien qui lui-même fonctionne avec le couple $(\mathcal{C}, \gamma)$.

4.1.4. Fonctionnement de l'algorithme SVOH

Soient $\{\mathcal{C}\}$ et $\{\gamma\}$ correspondant respectivement à l'ensemble des valeurs pour le paramètre $\mathcal{C}$ et l'ensemble des valeurs pour le paramètre γ. Posons D_Z notre ensemble d'apprentissage de Z observations et f notre modèle de SVM obtenu avec les hyperparamètres $(\mathcal{C}, \gamma)$, D_{Z_E}, les $\frac{Z(k-1)}{Z}$ sous-ensembles de l'ensemble d'apprentissage après subdivision en k-sous-ensembles et D_{Z_S} , les $\frac{Z}{k}$ sous-ensemble restant réservé pour le test après subdivision. L'algorithme prend en entrée D_Z, $\{\mathcal{C}\}$ et $\{\gamma\}$. Pour chaque k subdivision ($k \in \{3,10\}$) de l'ensemble d'apprentissage

D_{Z},, l'algorithme entraine un classifieur $\mathbf{f}$ à l'aide des valeurs de $\{\mathcal{C}\}$ et $\{\gamma\}$ sur D_{Z_E}, puis évalue sur D_{Z_S} le taux de justesse de $\mathbf{f}$. Pour finir, l'algorithme sélectionne le meilleur triplé $(\mathcal{C}, \gamma, k)$ ayant donné un taux de justesse supérieur. Le tableau 4-1 nous donne un pseudo-code de l'algorithme SVOH.

Tableau 4-1: Pseudo code de l'algorithme SVOH

Algorithme 4.1: SVOH	
Entrées :	D_Z: ensemble d'apprentissage $\{\mathcal{C}\}$: ensemble des valeurs pour $\mathcal{C}$, $\{\gamma\}$: ensemble des valeurs pour γ
Sorties :	$\{k^*, \mathcal{C}^*, \gamma^*\}$
1 :	**pour tout** $\mathcal{C} \in \{\mathcal{C}\}, \gamma \in \{\gamma\}, k \in \{3,10\}$ **faire** :
2 :	$\mathbf{f} = \emptyset$
3 :	$(D_{Z_E}, D_{Z_S}) =$ subdivision (D_Z, k)
4 :	$f_E = SVM(D_{Z_E}, \mathcal{C}, \gamma)$
5 :	$E_r =$ Evaluer le taux de justesse (f_E, D_{Z_S})
6 :	$\mathbf{f} = \mathbf{f} \cup \{E_r\}$
7 :	**fin pour**
8 :	$\{k^*, \mathcal{C}^*, \gamma^*\} =$ le meilleur taux de justesse de $\mathbf{f}$
9 :	**Retourner** $\{k^*, \mathcal{C}^*, \gamma^*\}$

4.2. Validation de l'approche proposée

Dans cette section, nous décrivons le matériel utilisé et nous donnons les résultats d'expérimentation de l'approche SVOH.

4.2.1. Matériel d'expérimentation utilisé

Nous nous situons dans les mêmes conditions que dans les expériences réalisées au chapitre précédent. Nous avons utilisé les mêmes ensembles de données IPP, c'est-à-dire HPRD, S. Cerevisiae et H. Pylori, où les données HPRD ont servi de données d'apprentissage pendant que les deux autres ont servi de tests. Nous utilisons les caractéristiques extraites par la

technique BP, précisément celle utilisant la fonction, vue au chapitre précédent. Nous rappelons que cette technique permet de représenter pour chaque séquence un vecteur comportant 400 caractéristiques.

4.2.2. Résultats d'entrainement

L'entrainement a été mené sur les données HPRD et a consisté à rechercher à l'aide de l'algorithme SVOH une valeur pour k^*, C^*, γ^* parmi une grille de valeurs potentielles renseignées dans le tableau 4-2. Nous nous sommes servis du taux de justesse comme métrique d'évaluation de performance pour retrouver les valeurs optimales d'hyperparamètres. La capacité de généralisation du modèle formé est évaluée sur les ensembles de données IPP S. Cerevisiae et H. Pylori.

Tableau 4-2: Rangée des valeurs d'hyperparamètres pour la procédure de recherche avec SVOH

Hyperparamètres	Valeurs de grille
$\boldsymbol{C}$	$\{1; 3; 10; 32; 50; 100\}$
$\boldsymbol{\gamma}$	$\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$
$\boldsymbol{k}$	$\{3,4,5,6,7,8,9,10\}$

L'application de l'algorithme SVOH a donné les valeurs optimales d'hyperparamètres suivantes : $(C, \gamma, k) = (32; 0.01; 7)$. Le tableau 4-3 présente les meilleures valeurs de performances du taux de justesse moyen pour différentes combinaisons du triplé (C, γ, k).

Tableau 4-3: Taux de justesse pour des différentes valeurs de *K*

k	(C^*, γ^*)	Justesse (%)
3	(10 ; 0.1)	91.92
4	(50 ; 0.01)	92.36
5	(100 ; 0.001)	92.70
6	(10 ; 0.001)	91.13

7	**(32 ; 0.01)**	**93.69**
8	(32 ; 0.001)	92.36
9	(100 ; 0.1)	92.21
10	(100 ; 0.01)	92.49

Les résultats du tableau montrent que pour des valeurs de $k \in \{3; 4; 5; 6\}$, les taux de justesse se situent entre 92% et 93%. À partir de k = 7, les taux de justesse sont nettement supérieurs et se situent entre 92% et 94%. Dans l'ensemble, les taux de justesse sont sensiblement égaux, cependant, pour un nombre k = {5 ; 7 ; 10} où les valeurs 5 et 10 représentent les valeurs à priori, le modèle formé avec un nombre k = 7 de sous-ensembles obtient le meilleur score de justesse avec 93.69% contre 92.70% pour k = 5 et 92.49% pour k = 10. Ces premiers résultats démontrent que les meilleures performances du modèle SVM sont obtenues sur le triplé $(k, \mathcal{C}, \gamma) = (7, 32, 0.01)$.

Dans le tableau 4-4, nous comparons sur les autres métriques (précision, sensibilité et AUC) vues au chapitre précédent les scores obtenus pour les valeurs de subdivision $k = 7$ déterminé par l'approche SVOH contre ceux obtenus pour les valeurs $k = \{5; 10\}$ qui sont les valeurs généralement appliquées.

Tableau 4-4: Performances obtenues après application de SVOH pour différentes valeurs de subdivision

k	Précision (%)	Sensibilité (%)	AUC (%)
5	92.90	92.15	96.36
7	**94.09**	**93.15**	**97.88**
10	92.87	92.67	95.58

Les scores obtenus pour des valeurs à priori du nombre de sous-ensembles sont sensiblement les mêmes dans toutes les métriques. Pour une subdivision $k = 5$ de l'ensemble d'apprentissage, nous obtenons comme valeurs d'hyperparamètres $(\mathcal{C}, \gamma) = (100; 0.001)$. Les scores obtenus dans les métriques précision, sensibilité et AUC sont respectivement 92.90% ; 92.15% et 96.36%. Pour une subdivision $k = 10$, nous obtenons comme valeurs

d'hyperparamètres $(\mathcal{C}, \gamma) = (10; 0.01)$. Les scores obtenus dans les différentes métriques sont respectivement 92.87% ;92.67% et 95.38%. Par-contre, les taux obtenus pour une subdivision $k = 7$ avec pour valeurs d'hyperparamètres $(\mathcal{C}, \gamma) = (32; 0.01)$ sont respectivement 94.09% ; 93.15% et 97.88%. En outre, bien vrai que l'écart entre les différents taux ne soient pas très grand, nous retenons tout de même qu'une subdivision de l'ensemble d'apprentissage en 7 sous-ensembles améliore le taux de justesse d'environ 1% par rapport aux taux obtenus avec les valeurs de subdivision à priori (voir tableau 4-3). Aussi, nous constatons également un meilleur score dans les métriques précision et sensibilité avec une performance moyenne supérieure à 0.7% que celles obtenues par les valeurs à priori. Les résultats montrent que les meilleurs taux sont obtenus avec une subdivision k = 7, c'est-à-dire celle déterminée par l'approche SVOH.

4.3.3. Résultats de validation

Nous utilisons ici la technique de la validation croisée 5-fois pour évaluer les performances du modèle SVM-BP en considérant les valeurs d'hyperparamètres obtenues après application de l'approche SVOH. Il faut noter que la technique de validation croisée *k*-fois peut être utilisée également pour l'évaluation de modèles tout comme la sélection de modèles [Rodriguez et al. 2010]. Cinq modèles après une validation croisée 5-fois ont donné les résultats du tableau 4-5. Les performances moyennes obtenues dans les métriques justesse, précision, sensibilité et AUC sont respectivement 94.77% ; 94.79% ; 94.69% et 97.38%. De même les écart-types obtenus sont respectivement de $\pm 0.31\%$; $\pm 0.23\%$, $\pm 0.91\%$ et $\pm 0.6\%$. Nous constatons que ces écart-types sont tous inférieurs à 1%, ce qui traduit une bonne fiabilité du modèle.

Tableau 4-5: Résultats après validation croisée 5-parties

k	Justesse (%)	Précision (%)	Sensibilité (%)	AUC (%)
1	94.74	94.90	95.09	96.92
2	94.97	95.94	95.50	96.82
3	94.24	94.67	94.02	95.68
4	94.84	94.01	93.70	97.80
5	94.29	95.24	94.44	97.74
Moyenne	**94.77% ±0.31%**	**94.79% ± 0.23%**	**94.69% ± 0.91%**	**97.38% ± 0.6%**

4.3.3.1 Comparaison avec l'outil SVM-BP

Nous comparons ici les résultats de performances obtenus après une validation croisée 5-fois sur les données de validation entre l'outil SVM-BP obtenu dans le chapitre précédent et le nouvel outil SVM-BP obtenu à partir de l'approche SVOH. La figure 4-2 présente les taux de justesse obtenus pour les cinq modèles SVM après 5 validation croisée. Il faut rappeler que les performances obtenues dans la métrique justesse par le modèle SVM-BP durant les cinq phases de la validation croisée sont respectivement 92.04%, 92.97%, 91.94%, 92.84% et 91.97%. Nous constatons sur la figure 4-2 que les performances affichées par le nouvel outil SVM-BP (SVOH) sur l'ensemble des cinq phases de la validation croisée sont supérieures aux performances affichées par l'outil SVM-BP.

La figure 4-3 montre les taux moyens affichés par les outils SVM-BP et SVM-BP (SVOH) dans les métriques précision, sensibilité et AUC. Tout comme dans la métrique justesse, les taux moyens affichés par l'outil constitué avec l'approche SVOH sont supérieurs. De manière générale, par rapport à l'approche BP combiné au classifieur SVM que nous avons proposé précédemment, SVOH améliore ici la justesse de plus de 1.17%, la précision de 1.32%, la sensibilité de 1.93 %, et l'AUC de 0.25%. Nous pouvons dire que les valeurs d'hyperparamètres $(\mathcal{C}, \gamma) = (32; 0.01)$ déterminés en appliquant l'approche SVOH donne des performances supérieures par rapport aux performances obtenues avec les valeurs d'hyperparamètres $(\mathcal{C}, \gamma) = (100; 0.001)$ déterminés dans le chapitre précédent.

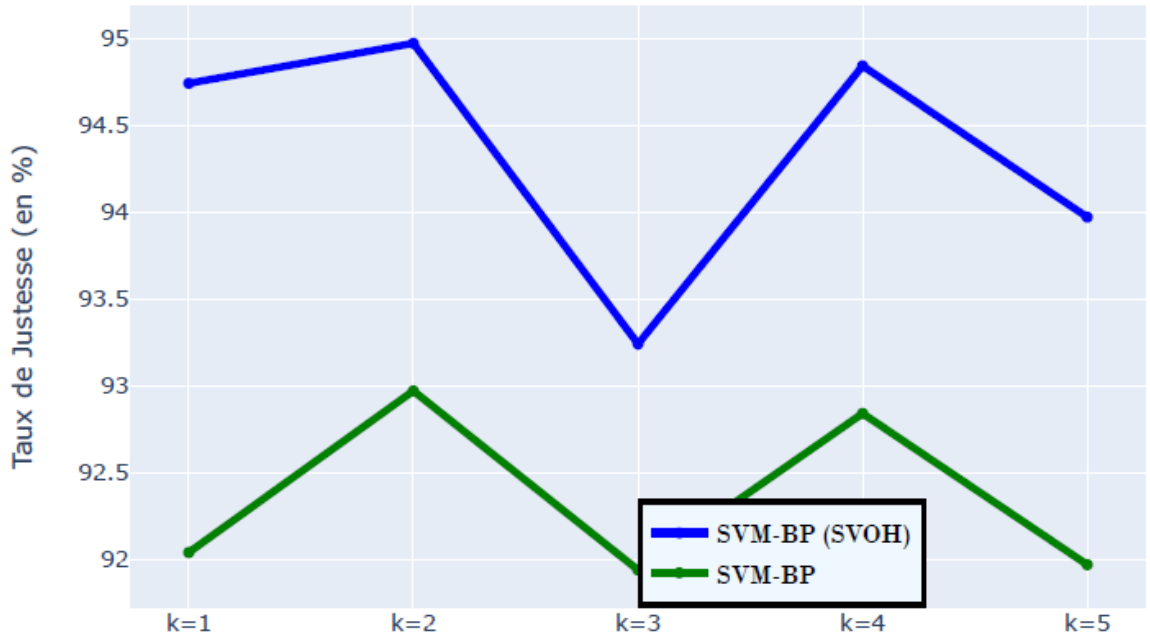


Figure 4-2: Comparaison du taux de justesse entre SVM-BP et SVM-BP-SVOH après 5-VC

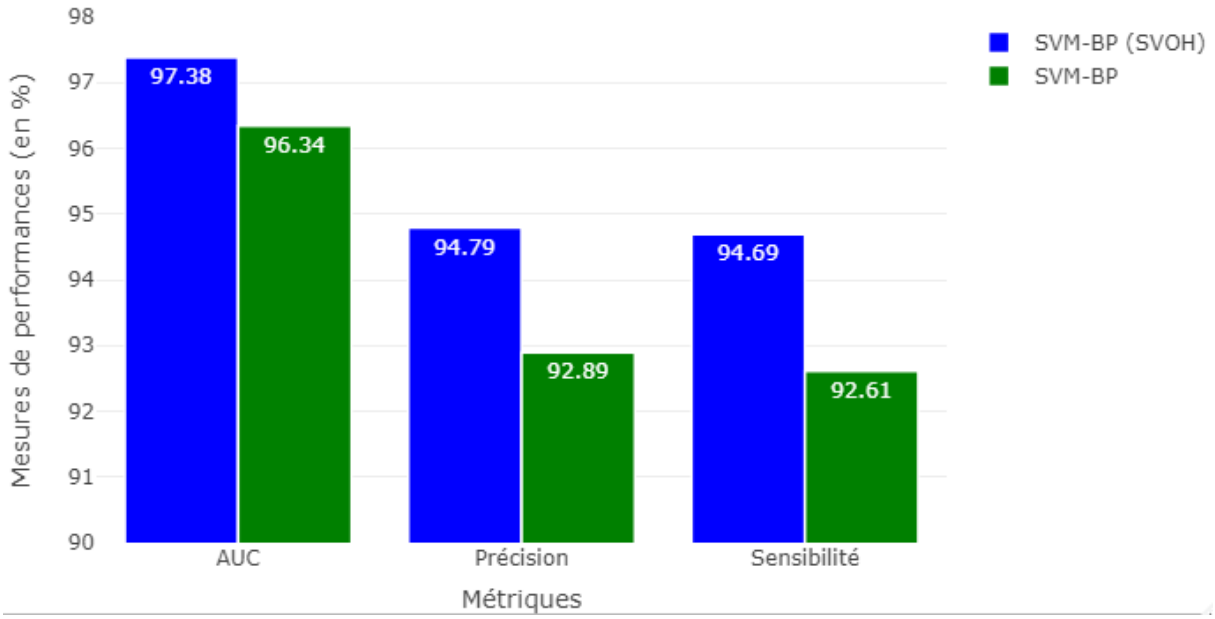


Figure 4-3: Comparaison des taux moyens dans les métriques AUC, précision et sensibilité

4.3.3.2 Comparaison avec d'autres outils de prédiction d'interaction de la littérature

Le tableau 4-6 indique les résultats de performance obtenus sur les données IPP HPRD avec ceux obtenus par les auteurs [You et al. 2013], [Göktepe et Kodaz 2018], [An et al. 2019] et [Ma et al. 2020].

Tableau 4-6: Comparaison des performances sur les données HPRD avec d'autres auteurs

Méthode	Justesse	Précision	Sensibilité
[You et al. 2013]	84.80%	85.47%	84.08%
[Göktepe et Kodaz. 2018]	73.81%	74.11%	73.24%
[An et al. 2019]	90.40%	88.03%	93.54%
[Ma et al. 2020]	75.82%	78.24%	72.74%
SVM-BP (SVOH)	**94.71%**	**94.79%**	**94.69%**

Nous pouvons constater que contrairement à l'outil SVM-BP construit dans le chapitre précédent, l'outil SVM-BP formé en appliquant l'approche SVOH obtient des taux de performance élevés dans toutes les métriques par rapport aux autres auteurs.

4.3.4 Résultats de prédiction sur les données S. Cerevisiae et H. Pylori

Nous évaluons ici le taux d'erreur de l'outil obtenu avec l'approche SVOH sur les ensembles de données non vus.

4.3.4.1 Résultats obtenus sur les données S. Cerevisiae

Les résultats obtenus par l'outil SVM-BP construit à partir de l'approche SVOH sont 91.10% ; 91.65% ; 90.74% et 94.90% respectivement dans les métriques justesse, précision, sensibilité et AUC. Ces résultats démontrent que le nouvel outil formé fait moins d'erreur sur des données qui n'ont pas été utilisées pour l'entrainement.

La figure 4-4 indique les résultats de comparaison sur les données S. Cerevisiae avec le modèle d'outil construit sans l'approche SVOH. Nous constatons que dans l'ensemble, le nouvel outil SVM-BP avec l'approche SVOH (en bleu) réalise des performances supérieures aux performances de l'outil SVM-BP en vert. Ces résultats montrent également que la capacité de généralisation avec l'approche SVOH de l'outil SVM-BP est améliorée.

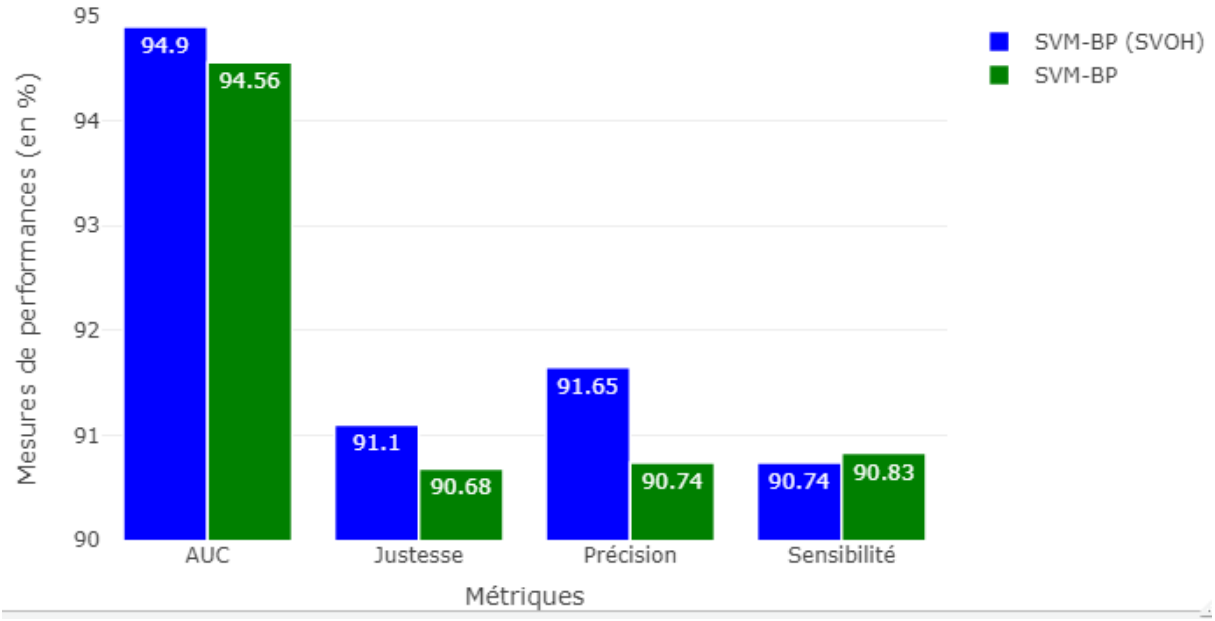


Figure 4-4: Comparaison entre SVM-BP et SVM-BP (SVOH) sur les données S. Cerevisiae

Nous comparons sur la figure 4-5 les résultats de performance obtenus avec d'autres auteurs sur les données S. Cerevisiae. Nous pouvons voir sur cette figure que les taux affichés par le nouvel outil SVM-BP avec l'approche SVOH (en bleu) sont supérieurs aux taux présentés par la plupart des méthodes, excepté l'outil construit avec la technique LCPSSMMF

(en rouge clair). Nous rappelons que cette méthode est une autre version de la méthode *Bi-gram* qui calcule plusieurs PSSM avant d'appliquer la technique 2-gramme pour extraire les bigrammes d'acide aminé. Les taux supérieurs observés avec cette méthode s'expliquent par le fait que les séquences de la base de données choisie pour avoir les PSSM des protéines présentent une forte similarité avec les séquences requêtes de l'ensemble de données IPP S. Cerevisiae. Cependant, une telle démarche est lente dans l'exécution et les performances de prédiction dépendent de la base de données choisie. En outre nous constatons que l'écart de performance est réduite avec le nouvel outil SVM-BP.

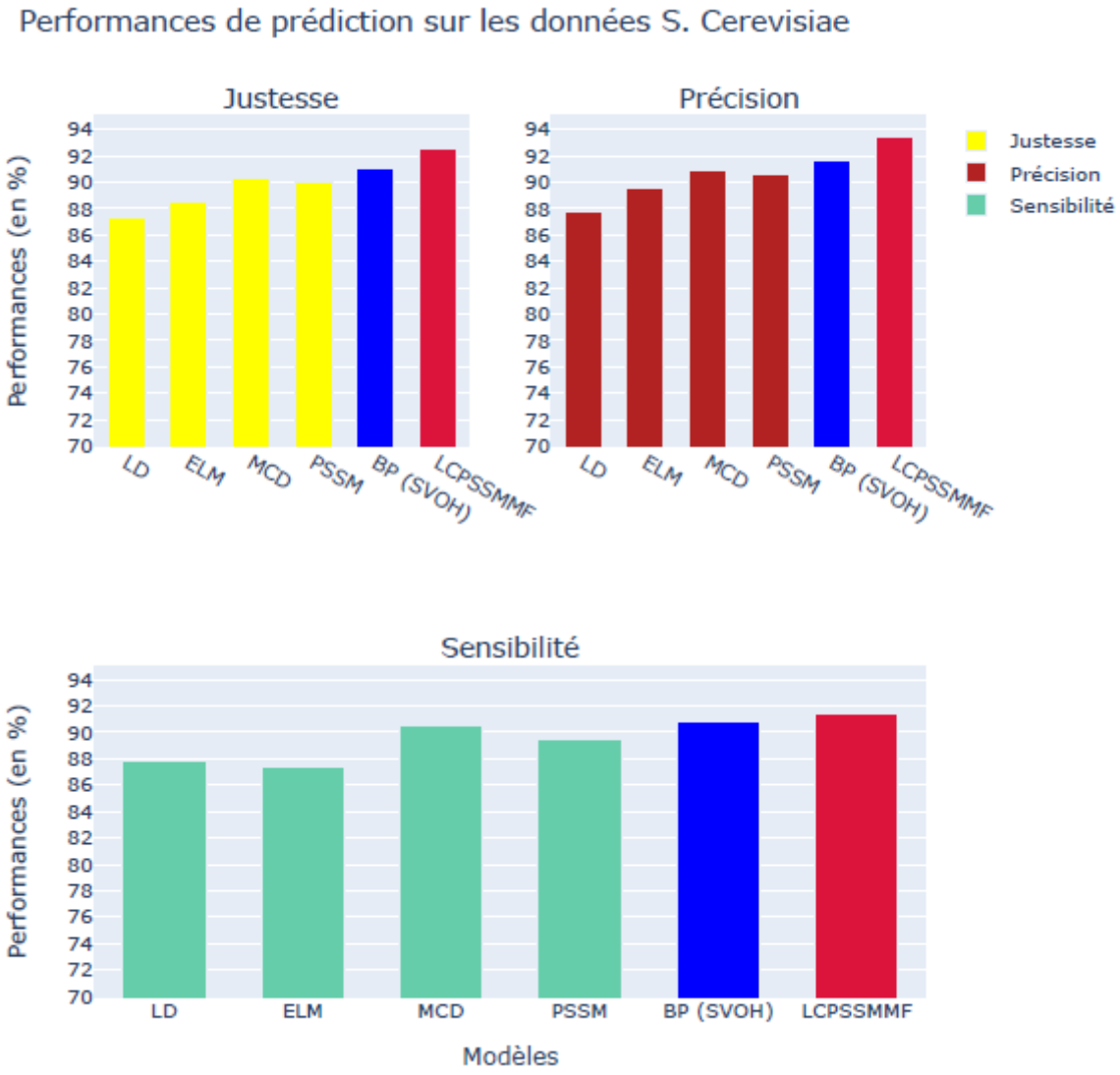


Figure 4-5: Comparaison avec d'autres auteurs sur les données S. Cerevisiae

4.3.4.2 Résultats obtenus sur les données H. Pylori

La figure 4-6 montre les résultats de comparaison avec le modèle d'outil construit sans l'approche SVOH. Les résultats obtenus par l'outil SVM-BP construit à partir de l'approche SVOH (en bleu) sont 89.07% ; 90.52% ; 90.43% et 92.10% respectivement dans les métriques justesse, précision, sensibilité et AUC. Nous constatons que cet outil réalise des performances supérieures aux performances de l'outil SVM-BP (en vert). Ces résultats montrent que l'approche SVOH permet à l'outil SVM-BP de réaliser des performances supérieures.

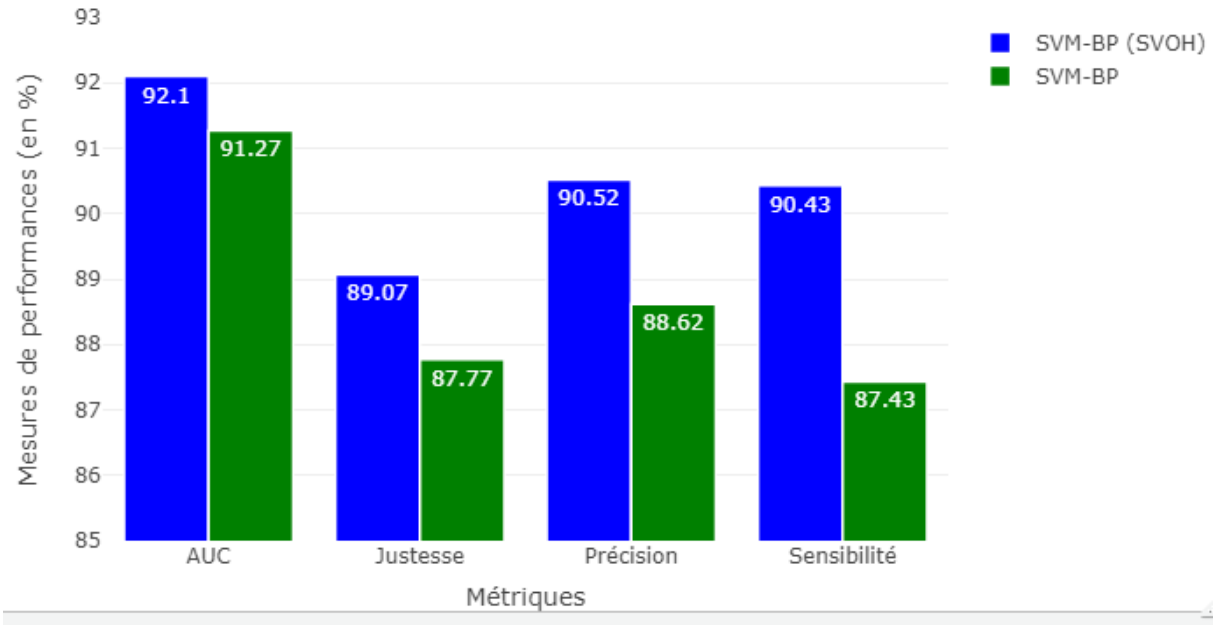


Figure 4-6: Résultats de prédiction sur les données de tests

Dans le tableau 4-7, les performances de prédiction obtenues par le outil SVM-BP avec l'approche SVOH sont comparées à celles obtenues par les auteurs [Zhou et al. 2011], [You et al. 2013], [You, Zhu, et al. 2014], [Göktepe et Kodaz 2018] et [Ma et al. 2020]. Nous pouvons constater que les résultats de performance affiché dans l'ensemble des métriques par l'outil SVM-BP (SVOH) sont supérieurs que ceux affichés dans la plupart des travaux existants.

Tableau 4-7: Comparaison de performances sur les données H. Pylori avec d'autres auteurs

Modèles	Justesse	Précision	Sensibilité
[Zhou et al. 2011]	84.20%	86.30%	85.10%

[You et al. 2013]	87.50%	86.15%	88.65%
[You, Zhu, et al. 2014]	84.91%	86.12%	83.24%
[Göktepe et Kodaz 2018]	86.23%	84.32%	89.44%
[Ma et al. 2020]	88.96%	86.86%	91.86%
SVM-BP (SVOH)	**89.07%**	**90.52%**	**90.43%**

Les différents résultats de performance de prédiction présentés tout au long de cette section montre que l'approche de sélection d'hyperparamètres SVOH permet à l'outil SVM-BP de réaliser des performances supérieures contrairement à l'approche de sélection classique des valeurs d'hyperparamètres. De plus, les performances moyennes obtenues dans les données de tests que sont les données S. Cerevisiae et H. Pylori sont tout aussi élevées que celles observées sur les données d'entrainement HPRD.

4.4. Autres résultats de validation avec la nouvelle approche

Pour montrer la robustesse de l'approche proposée, nous avons appliqué les expériences sur d'autres ensembles de données IPP et avons évaluer dans la métrique justesse. Nous avons également testé l'approche avec le classifieur des réseaux de neurones sur les données HPRD.

4.4.1. Validation sur d'autres ensembles de données IPP

Quatre autres ensembles de données IPP utilisés également pour la prédiction des interactions ont servi à tester l'approche SVOH. Le premier est l'ensemble de données Homo Sapiens (H. Sapiens). Cet ensemble de données est collecté à partir de la base de données HPRD comme le décrit [Y.-A. Huang et al. 2015]. Il comporte 8161 paires de protéines dont 3899 paires IPP positives (paires où il y'a interaction) et 4262 paires IPP négatives (paires où il n'y a pas d'interaction). Le deuxième est l'ensemble de données de la bactérie *Escherichia coli* (E. coli) constitué uniquement de paires positives au nombre de 6594 [Riley 1993]. Le troisième est l'ensemble de données nommé C. elegans [X.-T. Huang et al. 2016] qui contient 4013 paires positives. Enfin, le quatrième ensemble est nommé M. musculus et contient 313 paires positives [Zhou et al. 2011].

Tableau 4-8: Résultats sur différents ensembles de données IPP

Ensemble de données IPP	(k^*, C^*, γ^*)	Justesse (%)
H. sapiens	$(4; 32; 0.01)$	90.92
E. coli	$(5; 10; 0.001)$	90.36
C. elegans	$(7; 50; 0.001)$	88.49
M. musculus	$(6; 10; 0.1)$	74.43

Les résultats du tableau 4-8 indique que les valeurs de triplés d'hyperparamètres (k^*, C^*, γ^*) qui permettent de réaliser les meilleures performances sur les ensembles de données H. sapiens, E. coli, C. elegans et M. musculus sont respectivement (4 ;32 ;0.01), (5 ;10 ;0.001), (7 ; 50 ; 0.001) et (6 ; 10 ; 0.1). Nous pouvons constater que mis à part les données E. coli où les performances sont obtenues avec une valeur à priori de subdivision de l'ensemble d'apprentissage ($k = 5$), les autres ensembles de données affichent des performances pour des valeurs de subdivisions différentes des valeurs habituelles. Ces résultats démontrent que le nombre de subdivision de l'ensemble d'apprentissage est important pour retrouver les valeurs optimales d'hyperparamètres du classifieur SVM.

4.4.2. Validation avec le classifieur des réseaux de neurones artificiel

L'architecture d'un réseau de neurone artificiel (RNA ou ANN) [Wira 2009] est un empilement multicouche de simples modules. La couche d'entrée reçoit les données, puis les informations des données sont transformées de manière non linéaire à travers plusieurs couches cachées. Le gradient [Bottou 2012] moyen est calculé et les poids sont ajustés en conséquence, avant que les sorties finales ne soient calculées dans la couche de sortie. Par exemple, considérons l'apprentissage d'un réseau neuronal artificiel avec λ-couches cachées, où chaque couche calcule $H^\alpha, \alpha \in [1, \dots, \lambda]$. La première couche prend en compte les entrées du réseau, tandis que la dernière couche renvoie les sorties H^λ sous forme de probabilité a posteriori. Soit $\{N^1, \dots, N^\alpha, \dots, N^\lambda\}$, le nombre de neurones pour chaque couche. Les couches intermédiaires renvoient $H^\alpha = \{h_i^\alpha\}$ où h_i^α représente la sortie du $i^{\text{ème}}$ neurone de la couche H^α. Cette sortie est déterminée selon l'expression suivante :

$$h_i^{\alpha} = f^{\lambda}(\sum_{i=1}^{N^{\alpha-1}} \omega_{i,j}^{\alpha} \times h_j^{\alpha-1} + b^{\alpha-1}) \;\; ; \forall i \in \{1, \dots, N^{\alpha}\}, \forall j \in \{1, \dots, N^{\alpha-1}\}$$

où $\omega_{i,j}^{\alpha}$ représentent les poids et $b^{\alpha-1}$, le biais (un par couche) et f^{λ}, une fonction non linéaire appliquée sur la somme des poids.

L'architecture RNA que nous avons adoptée ici est une architecture de réseaux avec deux couches cachées (voir figure 4-4).

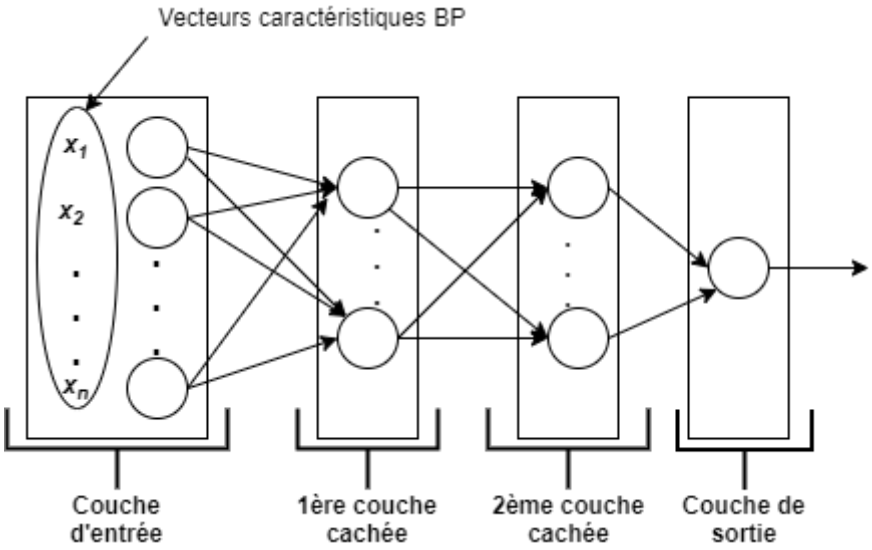


Figure 4-7: Architecture du réseau neuronal

Dans ce travail, N = 800 est pris comme entrée du réseau neuronal. Le processus d'entrainement du réseau neuronal a consisté à ajuster les paramètres du réseau (les poids) selon l'algorithme d'apprentissage jusqu'à ce que la fonction d'erreur du réseau atteigne un minimum. Nous avons utilisé la fonction sigmoïde comme fonction d'activation du réseau, fonction recommandée dans les cas d'une classification binaire [Cao et al. 2018]. Elle est définie selon l'équation suivante :

$$h(x) = \frac{1}{1 + e^{-x}}$$

où x représente l'entrée de la couche avale. Une telle fonction fait varier les valeurs des évaluations de 0 à 1 et est généralement utilisée pour produire une distribution de Bernoulli.

Pour une sortie $h(x) < 0.5$, le réseau la classera comme une interaction négative et pour une sortie $h(x) > 0.5$, elle sera classée comme une interaction positive.

Divers hyperparamètres d'un RNA à savoir le nombre de couches cachées, le nombre de nœuds cachés, les fonctions de transfert, le taux d'apprentissage, la taille de lot, et bien d'autres, peuvent affecter le taux de convergence et donc la qualité de la solution. Comme le nombre de configurations et d'hyperparamètres augmente de façon exponentielle, il est impossible de les essayer tous dans la pratique. Il est donc recommandé d'optimiser les hyperparamètres les plus importants, tels que le taux d'apprentissage et la taille du lot. Cela revient à explorer différentes valeurs tout en maintenant constants tous les autres hyperparamètres. Pour notre test, nous retenons que le taux d'apprentissage et la taille de lot comme paramètre à tuner. La technique de recherche par grille s'est également appliquée sur une plage de valeurs utilisée par d'autres auteurs pour retrouver les valeurs optimales des paramètres de taux d'apprentissage et de la taille par lot. Si nous désignons par τ le taux d'apprentissage et par ϑ la taille de lot, les différentes valeurs retenues sont celles utilisées dans la littérature [Cai et al. 2001 ; Du et al. 2017]: $\tau \in \{0.5;\ 0.1;\ 0.01;\ 0.001\}$ et $\vartheta \in \{128; 100; 64; 50; 32; 16\}$.

Tableau 4-9: Performances obtenues après application de SVOH dans les cas des ANN

k	(τ, ϑ)	Justesse (%)
3	(0.5 ; 0.01)	90.28
4	**(0.01 ; 32)**	**92.74**
5	(0.01 ; 100)	91.88
6	(0.001 ; 100)	91.58
7	(0.01 ; 50)	90.49
8	(0.001 ; 128)	87.75
9	(0.01 ; 64)	90.45
10	(0.001 ; 32)	88.79

Le tableau 4-9 indique que contrairement aux valeurs classiques de subdivision de l'ensemble d'apprentissage, $k = 5$ ou $k = 10$, la meilleure performance dans la métrique justesse (92.74%) du classifieur RNA (ANN) sur les données HPRD est obtenue avec le triplé $(k, \tau, \vartheta) = (4\ ;\ 0.01\ ;\ 32)$. Nous pouvons dire que le classifieur RNA obtient les meilleures

valeurs d'hyperparamètres sur un nombre de subdivision de l'ensemble d'apprentissage différent des valeurs habituelles.

4.5. Discussion

La principale technique utilisée dans cette étude est SVOH pour la recherche rigoureuse des valeurs optimales d'hyperparamètres du classifieur SVM gaussien. La particularité de cette approche est qu'elle considère le nombre *k* de subdivision de l'ensemble d'apprentissage comme un hyperparamètre. Les résultats d'expérimentation avec le classifieur SVM tout comme le classifieur RAN (ANN) ont confirmé la pertinence du choix de la valeur du nombre *k* car, après des tests sur les ensembles de données IPP HPRD, H. Pylori et S. Cerevisiae, le taux de justesse affiché en utilisant SVOH s'est avérée supérieur aux taux de justesse affichés en utilisant les valeurs habituelles (5 ou 10). Enfin, l'outil SVM-BP formé dans les données HPRD en utilisant l'approche SVOH présente de meilleurs taux dans la plupart des métriques utilisées comparés à l'outil SVM-BP formé avec une valeur de $k = 5$ dans le chapitre précédent. Ces résultats démontrent bien que l'approche développée permet de retrouver de façon rigoureuse les valeurs optimales d'hyperparamètres du classifieur utilisé.

Conclusion

Dans ce chapitre, nous avons proposé une approche modifiée de recherche sur grille combinée à la validation croisée à *k*-fois. Cette approche permet d'arbitrer automatiquement entre le pourcentage de données utilisées pour entraîner un classifieur et la rigueur de l'erreur estimée, en considérant le nombre de sous-ensembles comme un hyperparamètre à ajuster pendant la phase de sélection de modèles. Alors que le nombre de sous-ensembles *k*, dans la pratique est généralement fixé, nous avons montré, par le biais de tests sur des ensembles de données de référence bien connus, que l'approche proposée permet d'obtenir des limites d'erreur de généralisation plus légère sur l'ensemble de données de test.

CONCLUSION GENERALE
ET PERSPECTIVES

Conclusion générale

Les travaux présentés dans cette thèse portent sur le problème de prédiction d'interaction entre les protéines (IPP). La prédiction d'interaction se situe dans un contexte de classification binaire où nous avons d'une part les protéines qui interagissent ensemble ou liées et de l'autre part les protéines qui n'interagissent pas ensemble ou non liées. L'objectif visé ici était donc la mise en place d'un outil informatique qui améliore la prédiction des interactions protéine-protéine. Pour atteindre cet objectif, il est nécessaire de bien représenter les informations de la protéine par des caractéristiques inhérentes à l'intersection entre les protéines. Notre étude a révélé que la plupart des études informatiques de prédiction d'interaction s'oriente vers les séquences de protéines du fait que les données de séquences de protéines sont largement disponibles contrairement aux données sur les structures de protéines. Toutefois, les recherches dans ce domaine sont souvent confrontées à deux problèmes majeurs dont le problème d'extraction de caractéristiques à partir des séquences de protéines et le problème de sélection de modèles.

Nos contributions majeures dans cette étude se situent à ces deux niveaux : une technique d'extraction de caractéristiques à partir d'informations de séquences de protéines notée BP (*Bigram Physicochemical*) et une approche de recherche de paramètres, notée SVOH.

Pour ce qui est de l'extraction de caractéristiques des protéines, la technique BP proposée extrait des caractéristiques bigrammes qui au sens biologique simulent la reconnaissance des 'plis' de la protéine. Les bigrammes sont les fréquences de deux lettres (acides aminés) successives. Ces caractéristiques sont obtenues en appliquant la technique 2-gramme sur une matrice de scores calculée à partir d'informations liées au repliement des protéines. Précisément, la matrice est calculée à partir du multiplicateur d'une fonction de rang modélisant la flexibilité des acides aminés et d'une distance (approche BP1) ou d'une fonction (approche BP) obtenue à partir des valeurs de propriétés physicochimiques hydrophobicité et hydrophilie des acides aminés. Il faut souligner que la technique proposée vient résoudre le problème de vecteur strictement parcimonieux obtenu en appliquant la technique 2-gramme directement sur la séquence primaire d'une protéine donnée. Elle résoud également le problème de lenteur d'exécution et le problème de valeurs significatives dans le cas de l'extraction des caractéristiques bigrammes à partir de la PSSM (*Position Specific Score Matrix*). Un grand avantage de l'approche proposée est le fait qu'elle peut générer différents

types de caractéristiques d'une part en fonction des informations de propriétés physicochimiques fournies, d'autre part à partir de la fonction de calcul de la matrice de scores.

En ce qui concerne la deuxième contribution, nous avons proposé une approche de sélection de modèle en recherchant les valeurs optimales d'hyperparamètres du classifieur SVM. L'approche est une modification de la technique de recherche sur grille combinée à la technique de la validation croisée *k*-fois où *k* désigne le nombre de sous-ensembles de l'ensemble d'apprentissage. La validation croisée *k*-fois permet en effet de diviser l'ensemble d'apprentissage en *k* sous-ensembles, où, à tour de rôle, (*k* - 1) sous-ensembles sont utilisés pour la phase d'entrainement et l'autre est exploitée pour la phase de test. Cependant, le choix de la valeur de *k* est dans le plus souvent basé sur des choix à priori qui suggèrent de prendre une valeur $k = \{5; 10\}$. Or, le choix de *k* peut influencer sur les performances du modèle sélectionné. L'approche SVOH fait un apprentissage du nombre *k* plutôt que de le fixer comme cela se fait dans la littérature. Les résultats de test menés ont montré que l'approche SVOH permet de retrouver rigoureusement les valeurs optimales d'hyperparamètres sur un nombre *k* optimal plutôt que les valeurs habituelles (5ou 10) du nombre *k*.

Dans cette thèse, nous avons utilisé particulièrement l'algorithme d'apprentissage SVM avec noyau gaussien. L'utilisation des SVM avec noyau se justifie par le fait que les données d'apprentissage sont non linéairement séparables. L'objectif de l'utilisation du noyau est de pouvoir projeter les données dans un nouvel espace de caractéristiques suffisamment grand pour bien classifier les données. Or justement le noyau gaussien se caractérise par le fait qu'il peut projeter les données dans un espace de dimension infinie et est plus performant que le noyau polynomial ou le noyau linéaire selon les résultats de plusieurs études antérieures.

Les performances de prédiction des différents outils formés à partir des approches proposées ont été illustrées à travers les résultats numériques après validation sur des ensembles de données IPP réelles et après comparaison avec d'autres auteurs de la littérature.

Perspectives

Dans la continuité de ces travaux de thèse, nous envisageons différentes perspectives à différents niveaux. Certaines d'entre elles concernent la méthode d'extraction des caractéristiques à partir des informations de protéines, tandis que d'autres portent sur la méthode d'apprentissage automatique.

Extraction de caractéristiques

La technique d'extraction BP que nous avons développée peut être améliorée de deux manières :

(1) Introduire d'autres propriétés physicochimiques autres que les propriétés amphiphiles : comme vu au chapitre 2, concernant les acides aminés, nous avons plusieurs autres propriétés physicochimiques fonctionnelles telles que la polarité, le volume des chaines latérales, la polarisabilité, et bien d'autres. En incorporant plus de propriétés dans le calcul de la matrice de scores physicochimiques, nous pourrions ainsi obtenir des scores encore plus significatifs pour mieux modéliser la reconnaissance des 'plis' de la protéine. En effet, mise à part l'interaction hydrophobe qui participe au repliement de la protéine, nous avons trois autres types d'interactions à savoir l'interaction hydrogène, l'interaction ionique et l'interaction disulfure. En incorporant les propriétés physicochimiques autour de ces différents types d'interaction, nous exprimerons mieux le repliement de la protéine.

(2) Développer d'autres fonctions de calcul de la matrice : une des particularités de la matrice de scores utilisée dans la méthode BP est qu'elle est calculée à partir d'une distance. Dans cette thèse nous avons considéré le carré de la distance euclidienne comme mesure de dissimilarité pour représenter la distribution des acides aminés hydrophobes et hydrophiles le long de la séquence. La distance euclidienne fait partir de la distance de Minkowski normalisée [Merigó et Casanovas 2011]. Soit deux acides aminés *A* et *C*, une distance de Minkowski normalisée de dimension *n* est un mapping $d: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ comme suit :

$$d(A,C) = \frac{1}{n}\sum_{i=1}^{n} |a_i - c_i|^{\frac{1}{\rho}}$$

où a_i et c_i sont les *i*-ème propriétés physicochimiques des acides respectivement des acides aminés A et C, et le paramètre ρ est tel que $\rho \in (-\infty, \infty)$.

Si nous donnons différentes valeurs au paramètre ρ, nous pouvons obtenir un large éventail de cas particuliers. Par exemple, si ρ = 1, nous obtenons la distance de Hamming normalisée. Si ρ = 2, nous obtenons la distance euclidienne normalisée.

D'autres distances telle que la distance de Mahalanobis [De Maesschalck et al. 2000] peut être également exploitée ici pour mesurer la dépendance fonctionnelle entre les différents acides aminés. Nous pouvons exprimer aussi la dépendance fonctionnelle à travers différentes mesures de similarité comme celles décrites dans [Goshtasby 2012].

Modèle de classification

Bien que les outils SVM-BP et SVM-BP1 développés dans cette thèse arrivent à traiter des problèmes de classification d'interaction protéine-protéine, nous pouvons accroître les performances de prédiction et la capacité de généralisation de ces outils. Une solution est d'entrainer le classifieur SVM sur des ensembles de données beaucoup plus grand. Cependant, même si le classifieur SVM obtient de bonnes performances de prédiction sur des ensembles de données moyennes, il peut présenter de faibles performances sur des ensembles de données beaucoup larges. En outre, les modèles basés sur les réseaux de neurones ont montré de bonnes performances dans les domaines comme l'imagerie, la bio-informatique, la reconnaissance de la parole, etc. et sont robustes face aux bruits dans les grandes données [Du et al. 2017 ; Cao et al. 2018 ; Zhang et al. 2018 ; Yao et al. 2019]. Une combinaison du classifieur SVM et des réseaux de neurones pourra nous permettre de résoudre ce problème. Les algorithmes d'apprentissages profonds avec des machines à vecteurs de support permettent la formation d'un apprentissage couche par couche [Kim et al. 2013]. En empilant les SVM-BP, nous pouvons donc extraire des caractéristiques fortement discriminantes avec des vecteurs de supports qui maximisent la marge et garantissent les performances de généralisation.

Les couches multiples de l'architecture profonde ont pour rôle d'extraire des caractéristiques de pertinentes pour la reconnaissance de formes. Soit $\{x_1, \dots, x_n\}$ un ensemble d'observations. À partir des données d'apprentissage, nous pouvons obtenir l vecteurs de support $\{s_1, \dots, s_l\}$ à m dimensions avec $\{\varepsilon_1, \dots, \varepsilon_l\}$ les multiplicateurs de Lagrange associés et $\{y_1, \dots, y_l\}$ les cibles. L'activation de la couche suivante sera donc calculée comme suit :

$$h^1(i) = \varepsilon_i y_i k(s_i, x)$$

où $h^1(i)$ est le $i^{ème}$ élément de la première couche cachée. La dimensionnalité de h^1 est *l*, qui représente donc le nombre de vecteurs de support de la couche d'entrée. Cette première étape correspond donc à l'entraînement du poids reliant la couche d'entrée à la première couche cachée. Les poids des couches suivantes seront entraînés couche par couche de la même manière.

Une autre perspective pour améliorer la classification est de proposer un modèle basé sur les techniques ensemblistes comme l'algorithme de forêts aléatoires. Pour un tel modèle, nous allons tout d'abord construire plusieurs classifieurs SVM. L'étape suivante consistera à former parallèlement les classifieurs construits sur des sous-ensembles contenant des données différentes. La prédiction finale sera donc la moyenne de toutes les prédictions proposées par chaque classifieur SVM. Pour classer une nouvelle observation $\boldsymbol{x}$ par exemple, chaque *i* classifieur SVM-sera utilisé, et nous sélectionnerons la classe majoritaire parmi les *i* SVM.

REFERENCES

Afify, H. M., and Zanaty, M. S. [2021] "Computational predictions for protein sequences of COVID-19 virus via machine learning algorithms," *Medical & biological engineering & computing*, Springer, Vol. 59, No.9, pp. 1723–1734.

Aitchison, J., and Aitken, C. G. [1976] "Multivariate binary discrimination by the kernel method," *Biometrika*, Oxford University Press, Vol. 63, No.3, pp. 413–420.

Almagor, H. [1983] "A Markov analysis of DNA sequences," *Journal of Theoretical Biology*, Vol. 104, No.4, pp. 633–645.

Altschul, S. F., Gertz, E. M., Agarwala, R., Schäffer, A. A., and Yu, Y.-K. [2009] "PSI-BLAST pseudocounts and the minimum description length principle," *Nucleic Acids Research*, Vol. 37, No.3, pp. 815–824.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. [1997] "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, Vol. 25, No.17, pp. 3389–3402.

An, J.-Y., Zhou, Y., Zhao, Y.-J., and Yan, Z.-J. [2019] "An efficient feature extraction technique based on local coding PSSM and multifeatures fusion for predicting protein-protein interactions," *Evolutionary Bioinformatics*, SAGE Publications Sage UK: London, England, Vol. 15, p. 1176934319879920.

Anfinsen, C. B. [1973] "Principles that govern the folding of protein chains," *Science*, JSTOR, Vol. 181, No.4096, pp. 223–230.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., and Ridella, S. [2012] "The'K'in K-fold Cross Validation.," *ESANN*, pp. 441–446.

Anguita, D., Ghio, A., Ridella, S., and Sterpi, D. [2009] "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines.," *DMIN*, pp. 291–297.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. [2010] "The IntAct molecular interaction database in 2010," *Nucleic Acids Research*, Vol. 38, No.Database issue, pp. D525-531.

Arlot, S., and Celisse, A. [2010] "A survey of cross-validation procedures for model selection," *Statistics Surveys*, Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, Vol. 4, No.none, pp. 40–79.

Bakail, M., and Ochsenbein, F. [2016] "Targeting protein–protein interactions, a wide open field for drug design," *Comptes Rendus Chimie*, Emerging Chemistry in France, Vol. 19, No.1, pp. 19–27.

Ballone, A., Centorrino, F., and Ottmann, C. [2018] "14-3-3: a case study in PPI modulation," *Molecules*, Multidisciplinary Digital Publishing Institute, Vol. 23, No.6, p. 1386.

Bao, Y., and Liu, Z. [2006] "A Fast Grid Search Method in Support Vector Regression Forecasting Time Series," *Intelligent Data Engineering and Automated Learning – IDEAL 2006*, Lecture Notes in Computer Science, E. Corchado, H. Yin, V. Botti, and C. Fyfe, eds., Springer, Berlin, Heidelberg, pp. 504–511.

Barradas-Bautista, D., Rosell, M., Pallara, C., and Fernández-Recio, J. [2018] "Chapter Seven - Structural Prediction of Protein–Protein Interactions by Docking: Application to Biomedical Problems," *Advances in Protein Chemistry and Structural Biology*, Protein-Protein Interactions in Human Disease, Part A, R. Donev, ed., Academic Press, pp. 203–249.

Bengio, Y., Courville, A., and Vincent, P. [2013] "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, IEEE, Vol. 35, No.8, pp. 1798–1828.

Bengio, Y., and Grandvalet, Y. [2004] "No unbiased estimator of the variance of k-fold cross-validation," *Journal of machine learning research*, Vol. 5, No.Sep, pp. 1089–1105.

Ben-Hur, A., and Noble, W. S. [2005] "Kernel methods for predicting protein–protein interactions," *Bioinformatics*, Vol. 21, No.suppl_1, pp. i38–i46.

Berrar, D. [2019] "Cross-validation," *Encyclopedia of bioinformatics and computational biology*, Academic, Vol. 1, pp. 542–545.

Besson, M. [1975] "Rang moyen et agrégation de classements," *Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle*, EDP Sciences, Vol. 9, No.V1, pp. 37–58.

Binz, P.-A., Shofstahl, J., Vizcaíno, J. A., Barsnes, H., Chalkley, R. J., Menschaert, G., Alpi, E., Clauser, K., Eng, J. K., and Lane, L. [2019] "Proteomics standards initiative extended FASTA format," *Journal of proteome research*, ACS Publications, Vol. 18, No.6, pp. 2686–2692.

Blaisdell, B. E. [1985] "Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding," *Journal of Molecular Evolution*, Vol. 21, No.3, pp. 278–288.

Blüthner, M., Mahler, M., Müller, D. B., Dünzl, H., and Bautz, F. A. [2000] "Identification of an α-helical epitope region on the PM/Scl-100 autoantigen with structural homology to a region on the heterochromatin p25β autoantigen using immobilized overlapping synthetic peptides," *Journal of Molecular Medicine*, Vol. 78, No.1, pp. 47–54.

Bock, J. R., and Gough, D. A. [2001] "Predicting protein–protein interactions from primary structure," *Bioinformatics*, Vol. 17, No.5, pp. 455–460.

Bornot, A. [2009] "Analyse et prédiction de la relation séquence-structure locale et flexibilité au sein des protéines globulaires," *PhD Thesis*, Université Paris-Diderot-Paris VII.

Bottou, L. [2012] "Stochastic gradient descent tricks," *Neural networks: Tricks of the trade*, Springer, pp. 421–436.

Brito, J. A. A., McNeill, F. E., Webber, C. E., and Chettle, D. R. [2005] "Grid search: an innovative method for the estimation of the rates of lead exchange between body compartments," *Journal of Environmental Monitoring*, The Royal Society of Chemistry, Vol. 7, No.3, pp. 241–247.

Brouard, C. [2013] "Inférence de réseaux d'interaction protéine-protéine par apprentissage statistique."

Budiman, F. [2019] "SVM-RBF parameters testing optimization using cross validation and grid search to improve multiclass classification," *Научная визуализация*, Федеральное государственное автономное образовательное учреждение высшего …, Vol. 11, No.1, pp. 80–90.

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. [2017] "Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive," *Protein Crystallography: Methods and Protocols*, Methods in Molecular Biology, A. Wlodawer, Z. Dauter, and M. Jaskolski, eds., Springer New York, New York, NY, pp. 627–641.

Bushmarina, N. A., Blanchet, C., Vernier, G., and Forge, V. [2005] "Repliement des protéines : exemple de l'${\alpha}$-lactalbumine," *Journal de Physique IV (Proceedings)*, EDP Sciences, Vol. 130, pp. 209–228.

Cai, Y.-D., Liu, X.-J., and Chou, K.-C. [2001] "Artificial neural network model for predicting membrane protein types," *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, Vol. 18, No.4, pp. 607–610.

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., and Xie, Z. [2018] "Deep Learning and Its Applications in Biomedicine," *Genomics, Proteomics & Bioinformatics*, Vol. 16, No.1, pp. 17–32.

Cavnar, W. B., and Trenkle, J. M. [1994] "N-gram-based text categorization," *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, Citeseer.

Charton, M., and Charton, B. I. [1982] "The structural dependence of amino acid hydrophobicity parameters," *Journal of Theoretical Biology*, Vol. 99, No.4, pp. 629–644.

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. [2017] "The BioGRID interaction database: 2017 update," *Nucleic Acids Research*, Vol. 45, No.D1, pp. D369–D379.

Chautard, E., Thierry-Mieg, N., and Ricard-Blum, S. [2009] "Interaction networks: From protein functions to drug discovery. A review," *Pathologie Biologie*, Médecine régénératrice : cellules souches et matrice extracellulaire, Vol. 57, No.4, pp. 324–333.

Chou, K.-C. [2000] "Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect," *Biochemical and Biophysical Research Communications*, Vol. 278, No.2, pp. 477–483.

Chou, K.-C. [2001] "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, Vol. 43, No.3, pp. 246–255.

Chou, K.-C. [2005] "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, Vol. 21, No.1, pp. 10–19.

Chou, K.-C. [2009] "Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology," *Current Proteomics*, Vol. 6, No.4, pp. 262–274.

Chou, K.-C. [2011] "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, Vol. 273, No.1, pp. 236–247.

Chou, K.-C., and Shen, H.-B. [2006] "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization," *Biochemical and Biophysical Research Communications*, Vol. 347, No.1, pp. 150–157.

Chou, K.-C., and Shen, H.-B. [2008a] "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, Vol. 3, No.2, pp. 153–162.

Chou, K.-C., and Shen, H.-B. [2008b] "ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information," *Biochemical and Biophysical Research Communications*, Vol. 376, No.2, pp. 321–325.

Church, K. W. [2017] "Word2Vec," *Natural Language Engineering*, Cambridge University Press, Vol. 23, No.1, pp. 155–162.

Cortes, C., and Vapnik, V. [1995] "Support-vector networks," *Machine Learning*, Vol. 20, No.3, pp. 273–297.

Crick, F. [1970] "Central dogma of molecular biology," *Nature*, Nature Publishing Group, Vol. 227, No.5258, pp. 561–563.

Crick, F. H. [1958] "On protein synthesis," *Symp Soc Exp Biol*, p. 8.

Daudt, R. C., Le Saux, B., Boulch, A., and Gousseau, Y. [2018] "Détection dense de changements par réseaux de neurones siamois," *Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*, Marne-la-Vallée, France.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. [2000] "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, Elsevier, Vol. 50, No.1, pp. 1–18.

Dehzangi, A., López, Y., Lal, S. P., Taherzadeh, G., Michaelson, J., Sattar, A., Tsunoda, T., and Sharma, A. [2017] "PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction," *Journal of Theoretical Biology*, Vol. 425, pp. 97–102.

Delaney, S., Butler, G., Lam, C., and Thiel, L. [2000] "Three improvements to the BLASTP search of genome databases," *Proceedings. 12th International Conference on Scientific and Statistica Database Management*, pp. 14–24.

Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. [2008] "The protein folding problem," *Annu. Rev. Biophys.*, Annual Reviews, Vol. 37, pp. 289–316.

Domon, B., and Aebersold, R. [2006] "Mass spectrometry and protein analysis," *science*, American Association for the Advancement of Science, Vol. 312, No.5771, pp. 212–217.

Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., and Zhang, Y. [2017] "DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks," *Journal of Chemical Information and Modeling*, Vol. 57, No.6, pp. 1499–1510.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. [2001] "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, Vol. 19, No.1, pp. 26–59.

Dunphy, L. J., and Papin, J. A. [2018] "Biomedical applications of genome-scale metabolic network reconstructions of human pathogens," *Current Opinion in Biotechnology*, Systems biology • Nanobiotechnology, Vol. 51, pp. 70–79.

Dupuy, D., Li, Q.-R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., and Hope, I. A. [2004] "A first version of the Caenorhabditis elegans Promoterome," *Genome research*, Cold Spring Harbor Lab, Vol. 14, No.10b, pp. 2169–2175.

Eddy, S. R. [2004] "Where did the BLOSUM62 alignment score matrix come from?," *Nature Biotechnology*, Vol. 22, No.8, pp. 1035–1036.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. [1999] "Protein interaction maps for complete genomes based on gene fusion events," *Nature*, Vol. 402, No.6757, pp. 86–90.

Fischer, E. [1894] "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte der deutschen chemischen Gesellschaft*, Wiley Online Library, Vol. 27, No.3, pp. 2985–2993.

Gamblin, S. J., Haire, L. F., Russell, R. J., Stevens, D. J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D. A., Daniels, R. S., and Elliot, A. [2004] "The structure and receptor binding properties of the 1918 influenza hemagglutinin," *Science*, American Association for the Advancement of Science, Vol. 303, No.5665, pp. 1838–1842.

Ganapathiraju, M. K., Thahir, M., Handen, A., Sarkar, S. N., Sweet, R. A., Nimgaonkar, V. L., Loscher, C. E., Bauer, E. M., and Chaparala, S. [2016] "Schizophrenia interactome with 504 novel protein–protein interactions," *NPJ Schizophrenia*, Vol. 2, p. 16012.

Gardner, M. W., and Dorling, S. R. [1998] "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, Elsevier, Vol. 32, No.14–15, pp. 2627–2636.

Garg, A., Bhasin, M., and Raghava, G. P. [2005] "Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order,

and similarity search," *Journal of biological Chemistry*, Elsevier, Vol. 280, No.15, pp. 14427–14432.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. [2002] "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, Vol. 415, No.6868, pp. 141–147.

Ghanty, P., and Pal, N. R. [2009] "Prediction of Protein Folds: Extraction of New Features, Dimensionality Reduction, and Fusion of Heterogeneous Classifiers," *IEEE Transactions on NanoBioscience*, Vol. 8, No.1, pp. 100–110.

Göktepe, Y. E., and Kodaz, H. [2018] "Prediction of Protein-Protein Interactions Using An Effective Sequence Based Combined Method," *Neurocomputing*, Vol. 303, pp. 68–74.

González, A. J., and Liao, L. [2010] "Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines," *BMC Bioinformatics*, Vol. 11, No.1, p. 537.

Goshtasby, A. A. [2012] "Similarity and dissimilarity measures," *Image registration*, Springer, pp. 7–66.

Grantham, R. [1974] "Amino Acid Difference Formula to Help Explain Protein Evolution," *Science*, Vol. 185, No.4154, pp. 862–864.

Guo, L., Peng, J., and Xie, Q. [2018] "Maximum likelihood estimation based regression for multivariate calibration," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 189, pp. 316–321.

Guo, Y., Yu, L., Wen, Z., and Li, M. [2008] "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Research*, Vol. 36, No.9, pp. 3025–3030.

Han, D., Kim, H.-S., Seo, J., and Jang, W. [2003] "A domain combination based probabilistic framework for protein-protein interaction prediction," *Genome Informatics*, Japanese Society for Bioinformatics, Vol. 14, pp. 250–259.

Hastie, T., Tibshirani, R., and Friedman, J. [2009] *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Hayat, M., Tahir, M., and Khan, S. A. [2014] "Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces," *Journal of Theoretical Biology*, Vol. 346, pp. 8–15.

Hopp, T. P., and Woods, K. R. [1981] "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the National Academy of Sciences*, Vol. 78, No.6, pp. 3824–3828.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. [2003] *A practical guide to support vector classification*, Taipei.

Huang, C.-H., Peng, H.-S., and Ng, K.-L. [2015] "Prediction of Cancer Proteins by Integrating Protein Interaction, Domain Frequency, and Domain Interaction Data Using Machine Learning Algorithms," *BioMed Research International*, Vol. 2015, pp. 1–15.

Huang, G., Zhang, Y., Chen, L., Zhang, N., Huang, T., and Cai, Y.-D. [2014] "Prediction of Multi-Type Membrane Proteins in Human by an Integrated Approach," *PLoS ONE*, Vol. 9, No.3.

Huang, H., Wei, X., and Zhou, Y. [2018] "Twin support vector machines: A survey," *Neurocomputing*, Vol. 300, pp. 34–43.

Huang, X.-T., Zhu, Y., Chan, L. L. H., Zhao, Z., and Yan, H. [2016] "An integrative C. elegans protein–protein interaction network with reliability assessment based on a probabilistic graphical model," *Molecular BioSystems*, Royal Society of Chemistry, Vol. 12, No.1, pp. 85–92.

Huang, Y.-A., You, Z.-H., Chen, X., Chan, K., and Luo, X. [2016] "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, Vol. 17, No.1, p. 184.

Huang, Y.-A., You, Z.-H., Gao, X., Wong, L., and Wang, L. [2015] "Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence," *BioMed research international*, Vol. 2015.

Injadat, M., Moubayed, A., Nassif, A. B., and Shami, A. [2020] "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, Springer, Vol. 50, No.12, pp. 4506–4528.

Islam, S. M. A., Heil, B. J., Kearney, C. M., and Baker, E. J. [2018] "Protein classification using modified n-grams and skip-grams," *Bioinformatics*, Vol. 34, No.9, pp. 1481–1487.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. [2001] "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences*, Vol. 98, No.8, pp. 4569–4574.

cheol Jeong, J., Lin, X., and Chen, X.-W. [2010] "On position-specific scoring matrix for protein function prediction," *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE, Vol. 8, No.2, pp. 308–315.

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.-C. [2016] "Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition," *Journal of Biomolecular Structure and Dynamics*, Vol. 34, No.9, pp. 1946–1961.

Kearns, M., and Ron, D. [1999] "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural computation*, MIT Press, Vol. 11, No.6, pp. 1427–1453.

Keerthi, S. S., and Lin, C.-J. [2003] "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," *Neural Computation*, Vol. 15, No.7, pp. 1667–1689.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. [2009] "Human Protein Reference Database--2009 update," *Nucleic Acids Research*, Vol. 37, No.Database, pp. D767–D772.

Kim, S., Kavuri, S., and Lee, M. [2013] "Deep Network with Support Vector Machines," *Neural Information Processing*, Lecture Notes in Computer Science, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, eds., Springer, Berlin, Heidelberg, pp. 458–465.

Kobayashi, M., and Aono, M. [2004] "Vector Space Models for Search and Cluster Mining," *Survey of Text Mining: Clustering, Classification, and Retrieval*, M. W. Berry, ed., Springer New York, New York, NY, pp. 103–122.

Krigbaum, W. R., and Komoriya, A. [1979] "Local interactions as a structure determinant for protein molecules: II.," *Biochimica et biophysica acta*, Vol. 576, No.1, pp. 204–248.

Lage, K. [2014] "Protein–protein interactions and genetic diseases: The interactome," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, From genome to function, Vol. 1842, No.10, pp. 1971–1980.

Laura, G. A. Y. [2015] "Algorithme de descente du gradient stochastique."

Lee, H., Deng, M., Sun, F., and Chen, T. [2006] "An integrated approach to the prediction of domain-domain interactions," *BMC Bioinformatics*, Vol. 7, No.1, p. 269.

Lemberger, P., Batty, M., Morel, M., Raffaëlli, J.-L., and Delattre, M. [2015] *Big data et machine learning: manuel du data scientist*, Dunod, Paris.

Lengyel, Pe., and Söll, D. [1969] "Mechanism of protein biosynthesis," *Bacteriological Reviews*, Am Soc Microbiol, Vol. 33, No.2, pp. 264–301.

Lesk, A. [2010] *Introduction to protein science: architecture, function, and genomics*, Oxford university press.

Li, M., Chen, Z., Wenyin, L., and Zhang, H.-J. [2005] "Statistical bigram correlation model for image retrieval."

Li, Z.-W., You, Z.-H., Chen, X., Li, L.-P., Huang, D.-S., Yan, G.-Y., Nie, R., and Huang, Y.-A. [2017] "Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier," *Oncotarget*, Vol. 8, No.14, pp. 23638–23649.

Liefooghe, A. [2008] "Matrices score-position, algorithmes et propriétés," *phdthesis*, Université des Sciences et Technologie de Lille - Lille I.

Liu, X. S., Brutlag, D. L., and Liu, J. S. [2002] "An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nature biotechnology*, Vol. 20, No.8, p. 835.

Lorenz, G. [1989] "Principal Component Analysis in Technology," *CIRP Annals*, Vol. 38, No.1, pp. 107–109.

Ma, W., Cao, Y., Bao, W., Yang, B., and Chen, Y. [2020] "ACT-SVM: Prediction of Protein-Protein Interactions Based on Support Vector Basis Model," *Scientific Programming*, Hindawi, Vol. 2020, p. e8866557.

Makondi, P. T., Lee, C.-H., Huang, C.-Y., Chu, C.-M., Chang, Y.-J., and Wei, P.-L. [2018] "Prediction of novel target genes and pathways involved in bevacizumab-resistant colorectal cancer," *PLOS ONE*, Vol. 13, No.1, p. e0189582.

Manning, C., and Schutze, H. [1999] *Foundations of Statistical Natural Language Processing*, MIT Press.

Martin, C. [2008] "Sélection immersive et guidée par des motifs géométriques spécifiques de sites d'intérêt pour l'amarrage protéine-protéine," *PhD Thesis*, Université Paris Sud-Paris 11; Université Paris Sud-Paris 11.

Martin, J. [2005] "Prédiction de la structure locale des protéines par des modeles de chaˆınes de Markov cachées," *PhD Thesis*, Citeseer.

Martin, S., Roe, D., and Faulon, J.-L. [2005] "Predicting protein–protein interactions using signature products," *Bioinformatics*, Vol. 21, No.2, pp. 218–226.

McClellan, K. J., and Goa, K. L. [1998] "Tirofiban," *Drugs*, Springer, Vol. 56, No.6, pp. 1067–1080.

McGinnis, S., and Madden, T. L. [2004] "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, Vol. 32, No.suppl_2, pp. W20–W25.

Merigó, J. M., and Casanovas, M. [2011] "A New Minkowski Distance Based on Induced Aggregation Operators," *International Journal of Computational Intelligence Systems*, Taylor & Francis, Vol. 4, No.2, pp. 123–133.

Milenova, B. L., Yarmus, J. S., and Campos, M. M. [2005] "SVM in oracle database 10g: removing the barriers to widespread adoption of support vector machines,"

Proceedings of the 31st international conference on Very large data bases, pp. 1152–1163.

Nakashima, H., Nishikawa, K., and Ooi, T. [1986] "The folding type of a protein is relevant to the amino acid composition," *The Journal of Biochemistry*, Oxford University Press, Vol. 99, No.1, pp. 153–162.

Nguyen, C. D., Gardiner, K. J., and Cios, K. J. [2011] "Protein annotation from protein interaction networks and Gene Ontology," *Journal of Biomedical Informatics*, Vol. 44, No.5, pp. 824–829.

Nikhila, K. S., and Nair, V. V. [2018] "Protein Sequence Similarity Analysis Using Computational Techniques," *Materials Today: Proceedings*, Vol. 5, No.1, pp. 724–731.

Ottman, C. [2013] "Protein–protein interactions: an overview," *Protein–Protein Interactions in Drug Discovery. Singapore, Wiley-VCH Verlag GmbH & Co*, Wiley Online Library, pp. 1–19.

Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D., and Maltsev, N. [1999] "The use of gene clusters to infer functional coupling," *Proceedings of the National Academy of Sciences*, National Acad Sciences, Vol. 96, No.6, pp. 2896–2901.

Pan, X.-Y., Zhang, Y.-N., and Shen, H.-B. [2010] "Large-Scale Prediction of Human Protein−Protein Interactions from Amino Acid Sequence Based on Latent Topic Features," *Journal of Proteome Research*, Vol. 9, No.10, pp. 4992–5001.

Parry-Smith, D. J., and Attwood, T. K. [1991] "SOMAP: a novel interactive approach to multiple protein sequences alignment," *Bioinformatics*, Oxford University Press, Vol. 7, No.2, pp. 233–235.

Patil, A. [2019] "Protein–Protein Interaction Databases," *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, eds., Academic Press, Oxford, pp. 849–855.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. [1999] "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles," *Proceedings of the National Academy of Sciences*, Vol. 96, No.8, pp. 4285–4288.

Peng, Z., Tian, F., Li, B., Wu, S., and Li, Z. [2006] "Genetic algorithm-based virtual screening of combinative mode for peptide/protein," *ACTA CHIMICA SINICA-CHINESE EDITION-*, Vol. 64, No.7, p. 691.

Perlibakas, V. [2004] "Distance measures for PCA-based face recognition," *Pattern Recognition Letters*, Vol. 25, No.6, pp. 711–724.

Perry, C. M. [2010] "Maraviroc," *Drugs*, Springer, Vol. 70, No.9, pp. 1189–1213.

Rain, J.-C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A., and Legrain, P. [2001] "The protein–protein interaction map of Helicobacter pylori," *Nature*, Vol. 409, No.6817, pp. 211–215.

Ranjan, G. S. K., Kumar Verma, A., and Radhika, S. [2019] "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–5.

Riley, M. [1993] "Functions of the Gene Products of Escherichia coli," *MICROBIOL. REV.*, Vol. 57, p. 91.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. [2005] "Inferring protein domain interactions from databases of interacting proteins," *Genome biology*, BioMed Central, Vol. 6, No.10, pp. 1–17.

Rodriguez, J. D., Perez, A., and Lozano, J. A. [2010] "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No.3, pp. 569–575.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. [1985] "Hydrophobicity of amino acid residues in globular proteins," *Science*, Vol. 229, No.4716, pp. 834–838.

Rouillon, J. D., and Candau, R. [2000] "La fatigue périphérique: sites subcellulaires et mécanismes biologiques," *Science & Sports*, Vol. 15, No.5, pp. 234–241.

Roy, B. [2007] "Double pondération pour calculer une moyenne : pourquoi et comment ?," *RAIRO - Operations Research - Recherche Opérationnelle*, Vol. 41, No.2, pp. 125–139.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., and Ayivi-Guedehoussou, N. [2005] "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, Nature Publishing Group, Vol. 437, No.7062, pp. 1173–1178.

Saha, I., Zubek, J., Klingström, T., Forsberg, S., Wikander, J., Kierczak, M., Maulik, U., and Plewczynski, D. [2014] "Ensemble learning prediction of protein–protein interactions using proteins functional annotations," *Molecular BioSystems*, Royal Society of Chemistry, Vol. 10, No.4, pp. 820–830.

Sakanyan, V., and Arnaud, M.-C. [2007] "Puces à protéines et perspectives d'applications médicales," *IRBM*, NUMERO SPECIAL BIOPUCES, Vol. 28, No.5, pp. 187–193.

Sali, A., Shakhnovich, E., and Karplus, M. [1994] "How does a protein fold?," *nature*, Vol. 369, No.6477, pp. 248–251.

Sathya, R., and Abraham, A. [2013] "Comparison of supervised and unsupervised learning algorithms for pattern classification," *International Journal of Advanced Research in Artificial Intelligence*, Citeseer, Vol. 2, No.2, pp. 34–38.

Sbai, S. M. A. [2012] "Traitement des signaux parcimonieux et applications," p. 130.

Schadle, I., Le Pévédic, B., Antoine, J.-Y., and Poirier, F. [2001] "Prédiction de lettre pour l'aide à la saisie de texte," *Actes des 3e Journées de l'Informatique Messine, JIM'2001*, Citeseer.

Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. [1997] "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE transactions on Signal Processing*, IEEE, Vol. 45, No.11, pp. 2758–2765.

Schreiber, S. L. [2005] "Small molecules: the missing link in the central dogma," *Nature chemical biology*, Nature Publishing Group, Vol. 1, No.2, pp. 64–66.

Shao, J., Yan, K., and Liu, B. [2021] "FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network," *Briefings in Bioinformatics*, Vol. 22, No.3.

Sharma, A., Lyons, J., Dehzangi, A., and Paliwal, K. K. [2013] "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of Theoretical Biology*, Vol. 320, pp. 41–46.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. [2007] "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, Vol. 104, No.11, pp. 4337–4341.

Sherali, H. D., and Tuncbilek, C. H. [1992] "A squared-euclidean distance location-allocation problem," *Naval Research Logistics (NRL)*, Wiley Online Library, Vol. 39, No.4, pp. 447–469.

Shin, W.-H., Christoffer, C. W., and Kihara, D. [2017] "In silico structure-based approaches to discover protein-protein interaction-targeting drugs," *Methods*, Systems Approaches for Identifying Disease Genes and Drug Targets, Vol. 131, pp. 22–32.

Shoemaker, B. A., and Panchenko, A. R. [2007a] "Deciphering protein–protein interactions. Part I. Experimental techniques and databases," *PLoS Comput Biol*, Public Library of Science, Vol. 3, No.3, p. e42.

Shoemaker, B. A., and Panchenko, A. R. [2007b] "Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners," *PLoS Comput Biol*, Public Library of Science, Vol. 3, No.4, p. e43.

Smaili, F. Z., Gao, X., and Hoehndorf, R. [2019] "Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction," *Bioinformatics*, Oxford University Press, Vol. 35, No.12, pp. 2133–2140.

Smith, G. R., and Sternberg, M. J. E. [2002] "Prediction of protein–protein interactions by docking methods," *Current Opinion in Structural Biology*, Vol. 12, No.1, pp. 28–35.

Su, R., Liu, X., Wei, L., and Zou, Q. [2019] "Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response," *Methods*, Elsevier, Vol. 166, pp. 91–102.

Tanford, Charles. [1962] "Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins," *Journal of the American Chemical Society*, Vol. 84, No.22, pp. 4240–4247.

Tatusova, T. A., and Madden, T. L. [1999] "BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences," *FEMS Microbiology Letters*, Vol. 174, No.2, pp. 247–250.

Tsubaki, M., Shimbo, M., and Matsumoto, Y. [2017] "Protein fold recognition with representation learning and long short-term memory," *IPSJ Transactions on Bioinformatics*, Information Processing Society of Japan, Vol. 10, pp. 2–8.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. [2000] "A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*," *Nature*, Vol. 403, No.6770, pp. 623–627.

Uetz, P., Titz, B., and Cagney, G. [2008] "Experimental methods for protein interaction identification and characterization," *Protein-protein interactions and networks*, Springer, pp. 1–32.

Valente, G. T., Acencio, M. L., Martins, C., and Lemke, N. [2013] "The development of a universal in silico predictor of protein-protein interactions," *PloS one*, Public Library of Science San Francisco, USA, Vol. 8, No.5, p. e65587.

Vapnik, V. [2013] *The Nature of Statistical Learning Theory*, Springer Science & Business Media.

Vapnik, V. N. [1999] "An overview of statistical learning theory," *IEEE transactions on neural networks*, IEEE, Vol. 10, No.5, pp. 988–999.

Veber, P. [2007] *Modélisation grande échelle de réseaux biologiques : vérification par contraintes booléennes de la cohérence des données*, Rennes 1.

Vihinen, M., Torkkila, E., and Riikonen, P. [1994] "Accuracy of protein flexibility predictions," *Proteins: Structure, Function, and Bioinformatics*, Vol. 19, No.2, pp. 141–149.

Voet, D., and Voet, J. G. [2004] "Biochemistry. Hoboken," *John Wiley & Sons*, Vol. 1, p. 591.

Voland, M. [2017] "Algorithmes pour la prédiction in silico d'interactions par similarité entre macromolécules biologiques," *PhD Thesis*, Université Paris-Saclay.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. [2002] "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, Nature Publishing Group, Vol. 417, No.6887, pp. 399–403.

Vyas, R., Bapat, S., Jain, E., Karthikeyan, M., Tambe, S., and Kulkarni, B. D. [2016] "Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis," *Computational Biology and Chemistry*, Vol. 65, pp. 37–44.

Waksman, G. (Ed.). [2005] *Proteomics and protein-protein interactions: biology, chemistry, bionformatics, and drug design*, Protein reviews, Springer, New York, NY.

Wan, K. K., Park, J., and Suh, J. K. [2002] "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair," *Genome Informatics*, Japanese Society for Bioinformatics, Vol. 13, pp. 42–50.

Wang, D. D., Wang, R., and Yan, H. [2014] "Fast prediction of protein–protein interaction sites based on Extreme Learning Machines," *Neurocomputing*, Vol. 128, pp. 258–266.

Wang, K., and Zhong, P. [2014] "Robust non-convex least squares loss function for regression with outliers," *Knowledge-Based Systems*, Elsevier, Vol. 71, pp. 290–302.

Wang, L., You, Z.-H., Xia, S.-X., Liu, F., Chen, X., Yan, X., and Zhou, Y. [2017] "Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier," *Journal of Theoretical Biology*, Vol. 418, pp. 105–110.

Wang, T., Li, L., Huang, Y.-A., Zhang, H., Ma, Y., and Zhou, X. [2018] "Prediction of Protein-Protein Interactions from Amino Acid Sequences Based on Continuous and Discrete Wavelet Transform Features," *Molecules*, Vol. 23, No.4, p. 823.

Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. [2016] "Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests," *Neurocomputing*, Vol. 193, pp. 201–212.

Wira, P. [2009] "Réseaux de neurones artificiels: architectures et applications," *Cours en ligne, Université de Haute-Alsace*.

Wong, W.-T., and Hsu, S.-H. [2006] "Application of SVM and ANN for image retrieval," *European Journal of Operational Research*, Elsevier, Vol. 173, No.3, pp. 938–950.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. [2000] "DIP: the Database of Interacting Proteins," *Nucleic Acids Research*, Vol. 28, No.1, pp. 289–291.

Yamanishi, Y., Vert, J.-P., and Kanehisa, M. [2004] "Protein network inference from multiple genomic data: a supervised approach," *Bioinformatics*, Oxford University Press, Vol. 20, No.suppl_1, pp. i363–i370.

Yang, L., and Shami, A. [2020] "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, Elsevier, Vol. 415, pp. 295–316.

Yang, T., Kecman, V., Cao, L., Zhang, C., and Huang, J. Z. [2011] "Margin-based ensemble classifier for protein fold recognition," *Expert Systems with Applications*, Elsevier, Vol. 38, No.10, pp. 12348–12355.

Yao, Y., Du, X., Diao, Y., and Zhu, H. [2019] "An integration of deep learning with feature embedding for protein–protein interaction prediction," *PeerJ*, PeerJ Inc., Vol. 7, p. e7126.

Yeomans, K. A., and Golder, P. A. [1982] "The Guttman-Kaiser criterion as a predictor of the number of common factors," *The Statistician*, JSTOR, pp. 221–229.

You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J., and Wang, B. [2013] "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, Vol. 14, No.S8, p. S10.

You, Z.-H., Li, X., and Chan, K. C. [2017] "An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers," *Neurocomputing*, Advanced Intelligent Computing: Theory and Applications, Vol. 228, pp. 277–282.

You, Z.-H., Yu, J.-Z., Zhu, L., Li, S., and Wen, Z.-K. [2014] "A MapReduce based parallel SVM for large-scale predicting protein–protein interactions," *Neurocomputing*, Vol. 145, pp. 37–43.

You, Z.-H., Zhu, L., Zheng, C.-H., Yu, H.-J., Deng, S.-P., and Ji, Z. [2014] "Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set," *BMC Bioinformatics*, Vol. 15, No.15, p. S9.

Zhang, L., Yu, G., Xia, D., and Wang, J. [2018] "Protein-Protein Interactions Prediction based on Ensemble Deep Neural Networks," *Neurocomputing*.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., and Honig, B. [2012] "Structure-based prediction of protein–protein interactions on a genome-wide scale," *Nature*, Vol. 490, No.7421, pp. 556–560.

Zhang, S., Ye, F., and Yuan, X. [2012] "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM," *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, Vol. 29, No.6, pp. 1138–1146.

Zhang, X., Jiao, X., Song, J., and Chang, S. [2016] "Prediction of human protein–protein interaction by a domain-based approach," *Journal of Theoretical Biology*, Vol. 396, pp. 144–153.

Zhou, Y. Z., Gao, Y., and Zheng, Y. Y. [2011] "Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence," *Advances in Computer Science and Education Applications*, Communications in Computer and Information Science, M. Zhou and H. Tan, eds., Springer Berlin Heidelberg, pp. 254–262.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M., and Snyder, M. [2001] "Global Analysis of Protein Activities Using Proteome Chips," *Science*, Vol. 293, No.5537, pp. 2101–2105.

Zohra Smaili, F., Gao, X., and Hoehndorf, R. [2018] "OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction," *arXiv e-prints*, p. arXiv-1804.

LISTE DES PUBLICATIONS ET CONFERENCES SCIENTIFIQUES

Les travaux réalisés au cours de cette thèse ont permis de publier deux **(2)** articles scientifiques.

Articles scientifiques

C. N. Kopoin, Tchimou N'Takpé, B. K. Saha, et M. Babri, "A Feature Extraction Method in Large Scale Prediction of Human Protein-Protein Interactions using Physicochemical Properties into Bi-gram," in *2020 IEEE International Conf on Natural et Engineering Sciences for Sahel's Sustainable Development - Impact of Big Data Application on Society et Environment (IBASE-BF)*, Feb. 2020, pp. 1–7, doi: 10.1109/IBASE-BF48578.2020.9069594.

Kopoin, Charlemagne N'Diffon, Armet Kodjo Atiampo, Behou Gerard N'Guessan, et Michel Babri. 2021. « Prediction of Protein-Protein Interactions from Sequences Using a Correlation Matrix of the Physicochemical Properties of Amino Acids ». *International Journal of Computer Science et Network Security* 21 (3): 41-47. https://doi.org/10.22937/IJCSNS.2021.21.3.6

Conférence

Natural et Engineering Sciences for Sahel's Sustainable Development - Impact of Big Data Application on Society et Environment (IBASE-BF), IEEE International Conf, Feb. 2020.

Séminaires

CYBER SECURITY 2018 : Workshop on CYBERSECIRITY organized by the World Bank sponsored Africa Center of Excellence OAU-ICT Driven Knowledge Park (OAK-PARK) from 3rd to 7th September, 2018 at Institut National Polytechnique Felix Houphouet- Boigny.

IndabaX 2021: Conférence IndabaX sur L'Intelligence Artificielle (C2IA). 15 Juillet 2021, Université Félix Houphouët-Boigny, Abidjan, CÔTE D'IVOIRE.

ANNEXES

A-1. Normalisation des valeurs hydrophobicité et hydrophilie

Le tableau A-1 nous donné l'algorithme de calcul des valeurs normalisées d'hydrophobicité et d'hydrophilie.

Tableau A-1: Algorithme de normalisation des valeurs H_1 et H_2

Algorithme 4.1: Normalisation_H_1H_2	
Entrées :	$\boldsymbol{H_1^0}$: valeur originale de l'hydrophobicité $\boldsymbol{H_2^0}$: valeur originale de l'hydrophilie $\mathbb{R}_{\boldsymbol{k}}$: les 20 résidus d'acides aminés natifs (k = 1, …, 20)
Sorties :	H_1: valeur normalisée de l'hydrophobicité H_2 : valeur normalisée de l'hydrophilie
1 :	$\varphi_1, \varphi_2 \leftarrow 0$ // φ: une variable
2 :	**Pour** $k \leftarrow$ 1 **à** 20 **faire** :
3 :	$\varphi_1 \leftarrow \varphi_1 + H_1^0(\mathbb{R}_k)/20$; $\varphi_2 \leftarrow \varphi_2 + H_2^0(\mathbb{R}_k)/20$
4 :	**Fin pour**
5 :	$E_1 \leftarrow H_1^0 - \varphi_1$; *// Variance hydrophobicité*
6 :	$E_2 \leftarrow H_2^0 - \varphi_2$; *// variance hydrophilie*
7 :	$EC_1 \leftarrow \sqrt{\sum {E_1}^2 / 20}$; *//Ecart type hydrophibicité*
8 :	$EC_2 \leftarrow \sqrt{\sum {E_2}^2 / 20}$; *// Ecart type hydrophilie*
9:	$H_1 = \frac{E_1}{EC_1}$; *// Valeur normalisée d'hydrphobicité*
10:	$H_2 = \frac{E_2}{EC_2}$ *// Valeur normalisée d'hydrophilie*
11 :	**Retourner** $H_1,\ H_2$

A-2. Calcul de la distance physicochimique

Cette distance $D(R_i, R_j)$ est calculée sur la base du calcul de distance euclidienne au carré [Perlibakas 2004 p.]. L'algorithme pour le calcul de la distance est donné dans le tableau A-2 :

Tableau A-2: Calcul de la fonction de distance

Algorithme 4.2 : *Distance BP*	
Entrées :	R_i:l'acide aminé *i* R_j : l'acide aminé *j*

	H_1: valeur normalisée de l'hydrophobicité H_2 : valeur normalisée de l'hydrophilie
Sorties :	$\boldsymbol{D_{ij}}$: valeur de distance entre le résidu *Ri* et le résidu *Rj*
1 :	$D_{i,j} \leftarrow \frac{1}{2}[(H_1(R_j) - H_1(R_i))^2 + (H_2(R_j) - H_2(R_i))^2]$
2 :	**Retourner** $D_{i,j}$

A.3. Matrice de scores

Le tableau A-3 ci-dessous représente l'écriture algorithmique de la matrice de scores physicochimiques.

Tableau A-3: Algorithme de la matrice de scores physicochimiques

Algorithme 4.3 : ***MSP***	
Entrées :	$\boldsymbol{L}$: longueur de la séquence, $\mathbb{R}_{\boldsymbol{k}}$: les 20 résidus d'acides aminés natifs
Sorties :	$\boldsymbol{C}$: matrice de scores
1 :	$C \leftarrow$ initialisation de la matrice à L lignes et 20 colonnes
2 :	**Pour** i ← 1 **à** L faire :
3 :	**Pour** k ← 1 **à** 20 f**aire** :
4 :	total $\leftarrow \frac{1}{k}\ D_{i,k}$
5 :	$C\ [i, k]\ \leftarrow total$
6 :	**Fin pour**
7 :	**Fin pour**
8 :	**Retourner** C

Dans la phase suivante, nous passons au calcul du bigramme avec les valeurs de la matrice de corrélation C.

A.4. Caractéristiques physicochimiques bigrammes

Dans le tableau A-4 Ci-dessous nous donnons l'algorithme associé au calcul des vecteurs *BP*.

Tableau A-4: Algorithme de calcul de la matrice des bigrammes

Algorithme 4.4 : ***BP***	
Entrées :	*L : la longueur de la séquence* *C : la matrice de scores*
Sorties :	BP : un vecteur de 400 éléments
2 :	*BP* ← initialisation de la matrice 20 lignes, 20 colonnes
3 :	**Pour** *i* ←1 **à** 20 **faire** :
4 :	**Pour** *j* ← 1 *à* 20 **faire** :
5 :	Tab ←initialisation d'une liste vide
6 :	**Pour** k ← 1 **à** *L* **faire** :
7 :	Tab← ajouter à tab (*C*[k, i] * *C*[k+1,j])
8 :	**Fin pour**
9 :	*Bigram* [i, j] ← $\sum tab$
10:	**Fin pour**
11 :	**Fin pour**
12 :	$BP \leftarrow Bigram^T$
13 :	**Retourner** *BP*

Pour représenter la paire d'interaction entre deux protéines, nous concaténons le vecteur *BP* de chaque protéine en interaction, ce qui donne un vecteur final de $2 \times 400 = 800$ variables *BP*.

L'algorithme présenté dans le tableau A-5 suivant permet d'avoir le vecteur *BP* de la paire *A-B* constituée des protéines *A* et *B*.

Tableau A-5: Algorithme de représentation de la paire de séquences

Algorithme : paire_*BP*	
Entrées :	BP_A : vecteur *BP* d'une protéine *A* BP_B : vecteur *BP* d'une protéine *B*
Sorties :	BP_{AB} : vecteur de la paire *A-B*
1 :	$BP_{AB} \leftarrow concatener\ BP_A\ et\ BP_B$
2 :	**Retourner** BP_{AB}

L'objectif recherché dans le calcul des caractéristiques des protéines est d'avoir un jeu de données constitué de vecteurs numériques caractéristiques devant servir à l'apprentissage d'une fonction de prédiction. Supposons que nous avons notre ensemble de données d'IPP qui

est l'ensemble de données HPRD dans laquelle la première ligne et la deuxième ligne constituent les protéines en interaction formant une paire et ainsi de suite. L'algorithme suivant permet de constituer un jeu de données HPRD constituées des vecteurs BP de chaque paire de protéines en interaction.

Tableau A-6: Algorithme pour la constitution du jeu de données d'apprentissage

Algorithme : frame_BP	
Entrées :	L'ensemble des données HPRD (*dataset* HPRD)
Sorties :	***JD_BP* : Un jeu de données**
1 :	***tab***← *initialisation d'un tableau vide*
2 :	**pour** $i \leftarrow 1$ **à** fin du dataset HPRD, par **pas de** 2 **faire** :
3 :	$tab \leftarrow BP_{i,i+1}$
4 :	**fin pour**
5	*JD_BP* ← convertir tab en un *dataframe*
6 :	**Retourner** *JD_BP*

La taille du *dataset* HPRD étant de 72000 lignes, (36000 lignes positives et 36000 lignes négatives), nous obtenons ainsi un jeu de données de 36000 lignes constituées des paires BP positives et des paires BP négatives.

Tableau B-1 : Recherche de valeurs optimales des hyperparamètres

Temps mis (s)	Paramètres	K=3	K=4	K=5	K=6	Moyenne	EC
0,913438	{'C': 3, 'gamma': 3}	0,779163	0,779725	0,798744	0,794580035	0,78335371	0,011426288
0,572368	{'C': 10, 'gamma': 0.0001}	0,841907	0,830269	0,832186	0,823452617	0,820795578	0,013083536
0,463411	{'C': 10, 'gamma': 0.001}	0,920374	0,94053	0,918852	0,916859059	0,9145347	0,011421506
0,468358	{'C': 10, 'gamma': 0.001}	0,920374	0,94053	0,918852	0,916859059	0,9145347	0,011421506
0,695377	{'C': 10, 'gamma': 0.01}	0,952145	0,960127	0,957385	0,956993459	0,950666913	0,007648291
0,882275	{'C': 10, 'gamma': 1}	0,782401	0,782404	0,79974	0,800049472	0,786309186	0,011228835
0,930256	{'C': 10, 'gamma': 3}	0,779163	0,779725	0,798744	0,794580035	0,78335371	0,011426288
0,545897	{'C': 32, 'gamma': 0.0001}	0,856216	0,855245	0,8499	0,844897482	0,840068476	0,013792561
0,433437	{'C': 32, 'gamma': 0.001}	0,932717	0,946749	0,936785	0,927523087	0,927884212	0,008834045
0,437605	{'C': 32, 'gamma': 0.001}	0,932717	0,946749	0,936785	0,927523087	0,927884212	0,008834045
0,671191	{'C': 32, 'gamma': 0.01}	0,952145	0,960127	0,957385	0,956993459	0,950666913	0,007648291
0,885067	{'C': 32, 'gamma': 1}	0,782401	0,782404	0,79974	0,800049472	0,786309186	0,011228835
0,928447	{'C': 32, 'gamma': 3}	0,779163	0,779725	0,798744	0,794580035	0,78335371	0,011426288
0,513138	{'C': 100, 'gamma': 0.0001}	0,873609	0,887423	0,868885	0,869963445	0,863388368	0,01434827
0,415671	{'C': 100, 'gamma': 0.001}	0,931271	0,937837	0,933377	0,921373406	0,926756415	0,006715082
0,407	{'C': 100, 'gamma': 0.001}	0,931271	0,937837	0,933377	0,921373406	0,926756415	0,006715082
0,671642	{'C': 100, 'gamma': 0.01}	0,952145	0,960127	0,957385	0,956993459	0,950666913	0,007648291
0,847271	{'C': 100, 'gamma': 1}	0,782401	0,782404	0,79974	0,800049472	0,786309186	0,011228835
0,972371	{'C': 100, 'gamma': 3}	0,779163	0,779725	0,798744	0,794580035	0,78335371	0,011426288

B.1. Cas de sous apprentissage avec le noyau Gaussien

La figure B-1 présente les trois scenarios de sous apprentissage sévère vu au chapitre 5 à la section 5-1. Nous pouvons constater que l'espace de données entier est attribué à la classe majoritaire qui n'est rien d'autre que la classe 1 représentée par les croix en bleu.

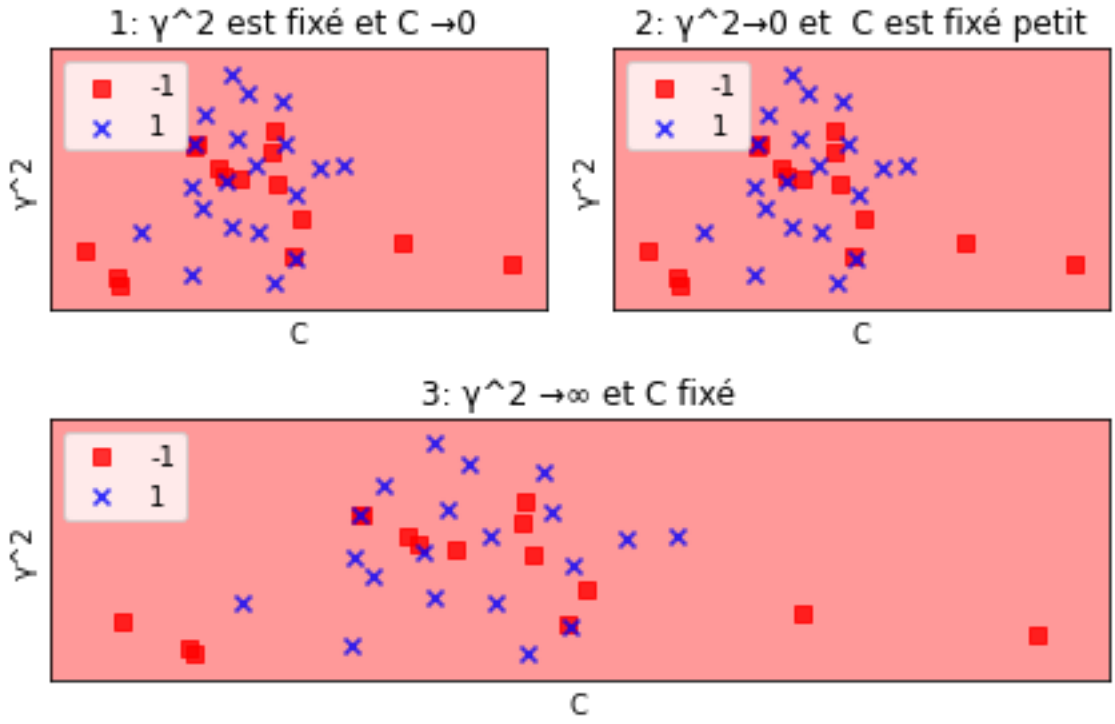


Figure B-1 : Les trois scénarios du sous apprentissage sévère avec le noyau Gaussien (voir chapitre 4)

B.2. Frontière de décision des SVM avec différents noyaux

La figure B-2 ci-dessous montre les frontières de décision obtenues en appliquant les hyperparamètres optimaux obtenus pour les noyaux linéaire, polynomial et gaussien. Le test est mené sur un échantillon de 20 IPP positives et 20 IPP négatives en considérant uniquement que les deux premières caractéristiques. Nous pouvons voir que le noyau Gaussien sépare mieux les observations que dans le cas linéaire ou polynomial.

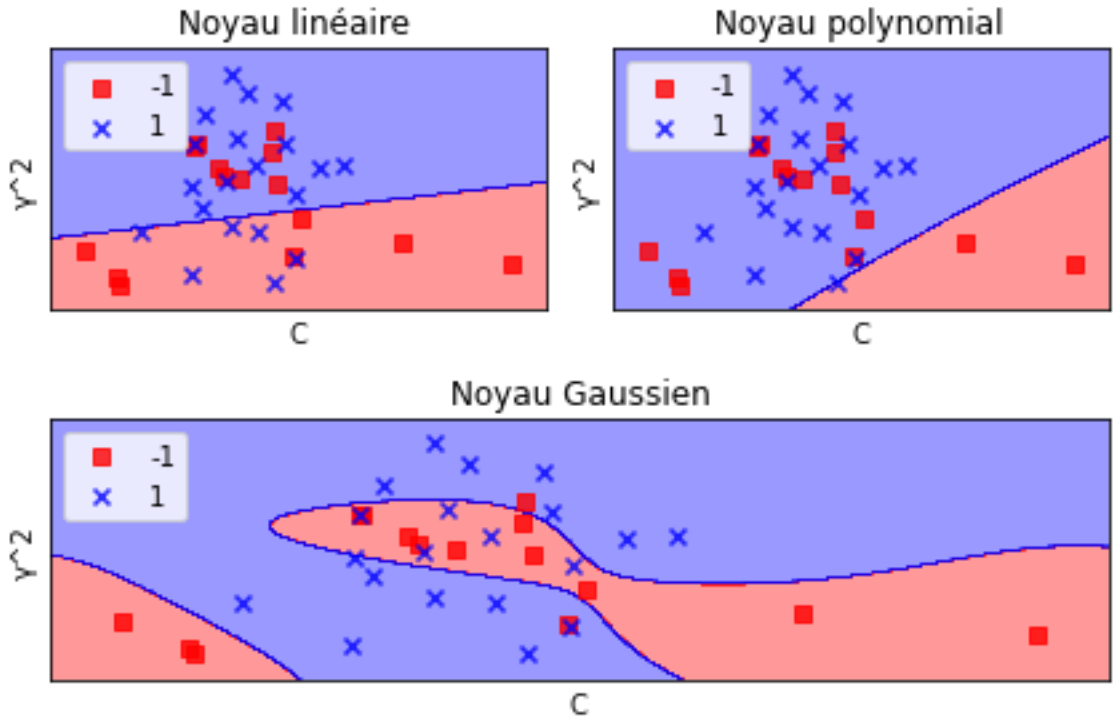


Figure B-2 : Frontière de décision avec les noyaux linéaire, polynomial et Gaussien