# Prédiction de la localisation des protéines membranaires

Sami Laroum

université angers

Année : 2011
*N°* d'ordre : 1188

# Prédiction de la localisation des protéines membranaires : méthodes méta-heuristiques pour la détermination du potentiel d'insertion des acides aminés.

## Thèse de doctorat

Spécialité : Informatique

## École Doctorale d'Angers

Présentée et soutenue publiquement

Le 25 Novembre 2011

À Angers

Par **SAMI LAROUM**

**Devant le jury ci-dessous :**

| | | |
|---|---|---|
| *Rapporteurs :* | Clarisse Dhaenens, | Professeur à l'Université de Lille 1 |
| | François Jacquenet, | Professeur à l'Université de Saint-Etienne |
| *Examinateurs :* | Dominique Lavenier, | Directeur de recherche CNRS IRISA-INRIA |
| *Directeur de thèse :* | Jin-Kao Hao, | Professeur à l'Université d'Angers |
| *Co-encadrant de thèse :* | Béatrice Duval, | Maître de Conférences à l'Université d'Angers |
| *Co-encadrant de thèse :* | Dominique Tessier, | Ingénieur de recherche à l'INRA de Nantes |

Laboratoire d'Etude et de Recherche en Informatique d'Angers
2, Bd Lavoisier, 49045 Angers Cedex 01

# Remerciements

Tout d'abord, je tiens à remercier Dominique Tessier, Béatrice Duval et Jin-Kao Hao, mes directeurs de thèse, pour m'avoir accueillie et encadrée pendant ces trois années. Merci pour leur patience et leurs conseils avisés tout au long de cette thèse. Merci d'avoir été si disponible.

J'adresse mes sincères remerciements à Clarisse Dhaenens et François Jacquenet pour avoir accepté d'être rapporteurs de ma thèse, ainsi que Dominique Lavenier pour avoir accepté d'être membre de mon jury de thèse. Un grand merci à vous.

Mes remerciements vont aussi aux deux laboratoires m'ayant accueilli et plus particulièrement je remercie l'équipe Bioinfo dans leur totalité, ainsi que tous les membres du LERIA. Ce fut un véritable plaisir de travailler avec ces personnes.

Je remercie aussi les membres du personnel administratif et technique pour leur aide dans les différents besoin que j'avais à leur demandé. Je n'oublie pas non plus mes collègues et amis thésards pour les bons moments passés en leur compagnie.

Ma famille et mes amis qui ont toujours été présent d'une manière ou d'une autre pour moi. Mes parents pour le soutien et encouragement durant ces trois années.

Merci du fond du cœur.

# Table des matières

# Introduction Générale

## Contexte de travail

La prédiction de la localisation des protéines au sein des différents compartiments de la cellule est un problème important en bioinformatique. De nombreuses méthodes de localisation de protéines ont été développées au cours de ces 30 dernières années. Ces méthodes utilisent des biais de composition en acides aminés ou la détection de signaux d'adressage pour identifier la destination finale des protéines. Cette identification permet d'obtenir des informations sur la fonction de la protéine au sein de la cellule. Dans notre travail, nous nous intéressons plus particulièrement à la localisation des protéines transportées à travers la membrane du réticulum endoplasmique, avec l'objectif de distinguer les protéines qui traversent la membrane de celles qui vont s'y insérer. Ces dernières années, de nombreux travaux ont conduit à une meilleure compréhension des mécanismes de transport des protéines au travers de la membrane [Cheng, 2010; Rapoport, 2008]. Ces protéines sont traitées par un complexe protéique appelé le translocon, qui décide de leur destination. Si la protéine nouvellement synthétisée ne possède qu'une suite d'acides aminés à son extrémité appelée peptide signal, cette protéine va traverser le translocon et elle est reconnue comme une protéine sécrétée. Par contre, si la protéine posséde des suites d'acides aminés appelées segments transmembranaire, ceux-ci vont s'insérer dans la membrane, car un segment transmembranaire correspond à une sorte de « code d'ouverture » reconnu par le translocon. Un peptide signal et un segment transmembranaire partagent des propriétés biochimiques très proches et il est très difficile de les distinguer. Il existe de nombreuses méthodes de prédiction dédiées à la localisation des protéines qui transitent par la membrane. Certaines de ces méthodes ont été uniquement développées pour la localisation des protéines avec un peptide signal [Klee and Ellis, 2005; Bendtsen *et al.*, 2004], tandis que d'autres méthodes sont dédiées uniquement à la localisation des protéines avec des segments transmembranaires [Tusnady and Simon, 2001; Krogh *et al.*, 2001a]. Néanmoins, en dépit de leurs bonnes performances de prédiction, dans certains cas, les méthodes de prédiction ont tendance à confondre le peptide signal et le segment transmembranaire. En effet, les méthodes de prédiction des protéines avec des segments transmembranaires ont tendance à prédire la région hydrophobe du peptide signal comme un segment transmembranaire [Chen *et al.*, 2002], alors que les méthodes de prédiction du peptide signal prédisent la partie hydrophobe du segment transmembranaire au début de la protéine comme élément du peptide signal [Nielsen and Krogh, 1998; Nielsen *et al.*, 1999]. De nombreux travaux ont été réalisés au cours de ces dernières années pour mieux comprendre la structure du translocon, et les mécanismes par lesquels le translocon sélectionne et insére les protéines membranaires dans la membrane du réticulum endoplasmique. Les travaux réalisés par Hessa *et al* [Hessa *et al.*, 2005a; Hessa *et al.*, 2005b] supposent que la capacité d'insertion d'un segment dans la membrane est uniquement dépendante de la composition en acides aminés de ce segment, et que

le potentiel d'insertion d'un segment est la somme des énergies d'insertion de ses acides aminés. Les expérimentations semblent montrer que l'énergie d'insertion d'un acide aminé est dépendante de sa position dans le translocon.

En nous appuyant sur ces hypothèses, nous concevons une méthode de classification des peptides signaux et des segments transmembranaires fondée sur le potentiel d'insertion de chaque acide aminé. Notre objectif est de déterminer « in silico » les courbes représentant le potentiel d'insertion de chacun des acides aminés. Les courbes sont définies de maniére à obtenir les meilleures performances pour notre méthode de classification. Si nous arrivons à obtenir de bon résultats de discrimination entre les peptides signaux et les segments transmembranaires, nous espérons ensuite déterminer la position des segments transmembranaires tout le long de la protéine. A plus long terme, cela devrait permettre de déterminaer avec une meilleure précision la structure des protéines membranaires.

Nous considérons la détermination des courbes comme un problème d'optimisation que nous abordons grâce aux méthodes méta-heuristiques comme les algorithmes de recherche locale et les algorithmes génétiques. On remarque que c'est la première fois que de telles approches sont appliquées à ce problème de prédiction difficile. L'évaluation de nos méthodes a été réalisée sur des jeux de données spécifiques que nous avons construits à partir de base de données de protéines. Les principaux résultats obtenus indiquent une bonne performance de discrimination entre les séquences peptides signaux et les séquences segments transmembranaires.

## Organisation de la thèse

Nous introduisons tout d'abord dans le Chapitre 1 des notions de biologie importantes pour la compréhension de cette thèse. Dans le Chapitre 2, nous effectuons un zoom sur les protéines adressées vers le réticulum endoplasmique en présentant les éléments essentiels à la compréhension de l'adressage de ces protéines. Par la suite, nous présentons trois familles de méthodes de prédiction des protéines membranaires. Nous y détaillons aussi les connaissances biologiques nécessaires à la compréhension du fonctionnement de ces méthodes. Nous nous attardons ensuite plus longuement sur de nouvelles hypothèses d'insertion des protéines membranaires. Nous introduisons le principe qu'elles mettent en œuvre afin d'expliquer le phénomène d'adressage des protéines à travers la membrane du réticulum endoplasmique. Nous y décrivons également les méthodes employant ces connaissances dans leur prédiction. Dans le Chapitre 3, nous présentons différents jeux de données qui vont nous servir dans la phase d'apprentissage et d'évaluation de notre méthode. Nous présentons ensuite le principe d'une nouvelle méthode de discrimination entre le peptide signal et le segment transmembranaire. Il s'agit d'une méthode qui se fonde sur les dernières connaissances biologiques connues sur le phénomène d'adressage des protéines membranaires. Enfin, différents critères d'évaluation seront détaillés dans ce chapitre. Dans le Chapitre 4, nous introduisons trois algorithmes se basant sur une approche par recherche locale afin d'optimiser les courbes des acides aminés : LSTranslocon [Laroum *et al.*, 2010], MN-LS [Laroum *et al.*, 2011], et BioGLS. Nous proposons ensuite, dans le Chapitre 5, une méthode fondée sur un algorithme génétique pour déterminer les

courbes d'insertion des acides aminés. Enfin, nous concluons en rappelant les principaux résultats obtenus et en proposant des perspectives de recherche.

# Chapitre 1

# Notions de biologie

Ce chapitre présente les notions de biologie essentielles pour la compréhension de cette thèse. Nous rappelons tout d'abord quelques éléments de biologie cellulaire en décrivant la cellule et ses compartiments cellulaires, puis nous présentons une brève description du processus de la synthèse des protéines ainsi que le mécanisme de leur adressage vers les différents compartiments cellulaires.

## Sommaire

## 1.1 Structure de la cellule

C'est avec l'avènement du microscope et à Robert Hooke que l'on doit la première observation d'une cellule d'une mince tranche de liège (figure 1.1).

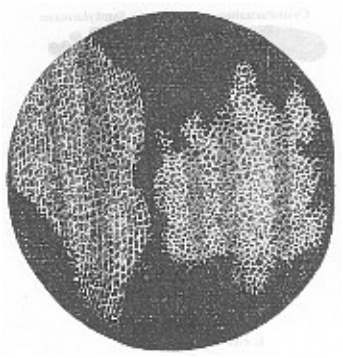

FIGURE 1.1 – Structure microscopique d'une coupe de liège dessinée par Robert Hooke et publiée en 1665 dans son traité « Micrographia ».

Une cellule est l'unité structurale et fonctionnelle d'un être vivant. Elle coordonne un ensemble de réactions biochimiques nécessaires à la vie. Au sein de la cellule, les protéines exécutent les principales fonctions cellulaires et assurent la construction et la maintenance de son architecture. Il existe deux grands types de cellule dans le monde vivant : les cellules *eucaryotes* qui comportent un noyau et les cellules *procaryotes* sans noyau (par exemple les bactéries).
Pour bien fonctionner, les cellules ont compartimenté leur processus biochimique entre le cytoplasme et les différents organites cellulaires (ou organelles). Nous présentons dans la figure 1.2 le schéma d'une cellule végétale avec ses principaux organites. Parmi les éléments principaux de la cellule, nous trouvons la membrane, le cytoplasme et le noyau.

La membrane, ensemble complexe constitué principalement de lipides, de protéines et de glucides, fonctionne comme une barrière entre l'intérieur et l'extérieur de la cellule ou entre deux compartiments cellulaires permettant l'échange des molécules entre le milieu intérieur et extérieur à l'aide de transporteurs. Le cytoplasme est un compartiment cellulaire délimité par la membrane. Chez les eucaryotes, le cytoplasme est le compartiment le plus volumineux et il contient de nombreux organites, tandis que chez les procaryotes, le cytoplasme est le seul compartiment de la cellule. Seules les cellules eucaryotes possèdent un noyau qui contient le matériel génétique de la cellule. Dans les cellules eucaryotes, nous devons également noter la présence du réticulum endoplasmique (RE). Comme nous le verrons dans la suite de cette thèse, cet organite joue un rôle essentiel durant la synthèse de certaines protéines.

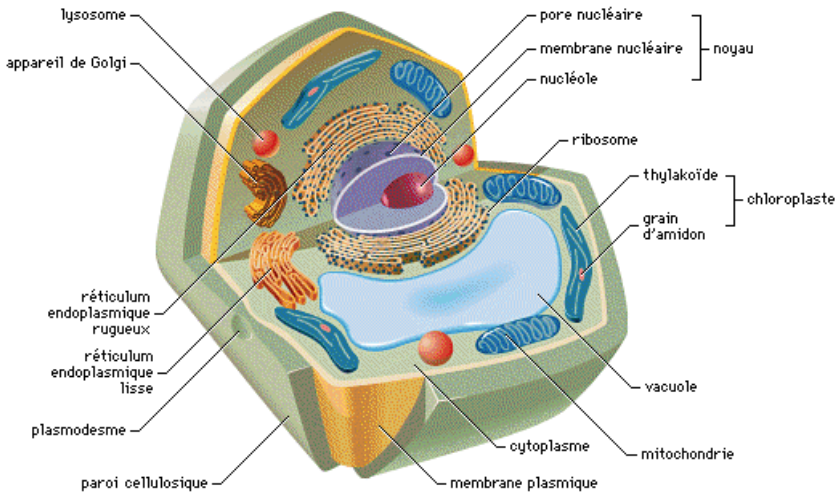


FIGURE 1.2 – Représentation schématique d'une cellule végétale. Source : http ://images.encarta.msn.com/xrefmedia/fencmed/targets/illus/ilt/T059293A.gif.

## 1.2 Synthèse des protéines

### 1.2.1 Rôle des protéines

L'étude des protéines est fondamentale en biologie car ce sont elles qui assurent les fonctions biochimiques nécessaires pour la vie des cellules et des organismes. En effet, les raisons pour lesquelles les protéines sont importantes :

- elles jouent le rôle de catalyseur qui maintient les processus métaboliques dans la cellule ;
- elles jouent le rôle des récepteurs qui transmettent des informations entre le milieu extérieur et intérieur ;
- elles sont impliquées dans la manipulation de l'ADN et l'ARN par des processus tels que la réplication de l'ADN ;
- elles servent comme élément structurel à l'intérieur et à l'extérieur de la cellule ;
- etc.

En fait, la plupart des fonctions cellulaires sont assurées par des protéines. L'étude et la caractérisation des fonctions des protéines est une tâche très importante qui permet de déterminer des « signatures » de protéines ayant la même fonction ou l'identification de protéines clefs de certaines pathologies humaines.

### 1.2.2 Dogme de la biologie moléculaire

Le dogme de la biologie se réfère à un principe biologique introduit la première fois par Francis Crick en 1958. La figure 1.3 présente le principe du dogme de la biologie moléculaire.

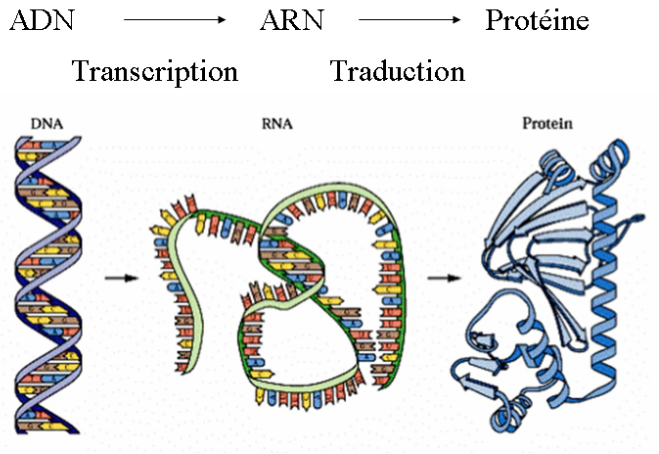


FIGURE 1.3 – Dogme de la biologie moléculaire. Source : http ://www.svt.ac-aix-marseille.fr/expoconf/genetiques/p3.htm.

Le principe se résume ainsi : l'ADN [1] (acide désoxyribonucléique) est le support de l'information génétique qui définit les fonctions biologiques d'un organisme. Toute molécule d'ADN est constituée d'un enchaînement réalisé à partir de 4 bases différentes : A, G, C, et T. La synthèse des protéines se déroule en deux étapes. La première étape est la transcription qui s'effectue à l'intérieur du noyau d'une cellule eucaryote ou dans le cytoplasme des procaryotes. Durant la transcription, une séquence d'ADN, qu'on appelle gène, va être transcrite en une molécule d'ARN messager (acide ribonucléique). L'ARN produit est une copie très proche chimiquement de l'ADN, qui va ensuite être traduit en une protéine. En effet, l'étape suivante est l'étape de traduction qui consiste à lire la suite de l'ARN messager pour produire les acides aminés dont l'assemblage donnera une protéine. Cette étape est réalisée au niveau d'un complexe protéique appelé ribosome se trouvant dans le cytoplasme. La figure 1.4 illustre le principe de la transcription de l'ADN en ARNm puis de la traduction de ce dernier en protéine.

De nouvelles avancées biologiques ont permis de montrer que les mécanismes de synthèse des protéines sont plus complexes et par conséquent certains points dans le principe du dogme sont à modifier. Parmi les découvertes, un gène peut être composé de plusieurs séquences et un même ARN peut produire plusieurs protéines en subissant un processus d'épissage ou de régulation.

### 1.2.3 Les acides aminés

Un acide aminé est une molécule constituée d'un carbone auquel sont liés un groupement amine (-NH2), un groupement acide (-COOH), et une chaîne latérale variable d'un acide aminé à l'autre. La figure 1.5 représente la structure schématique d'un acide aminé

1. L'ADN est une molécule présente dans toutes les cellules vivantes et qui renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme.

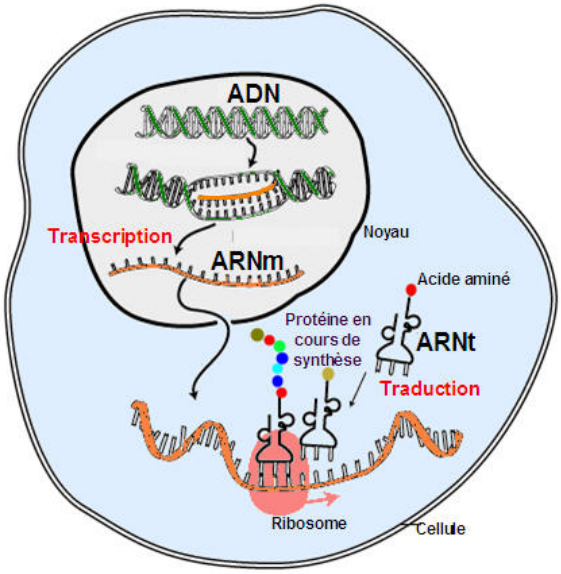


FIGURE 1.4 – Schéma illustrant les étapes de synthèse d'une protéine. Source : http ://www.journaldunet.com/science/biologie/dossiers/06/0609-adn/adn2/transcription-traduction.jpg

quelconque avec les groupes amine et acide, et la chaîne latérale symbolisée par la lettre R. Il existe dans la nature 20 acides aminés communs à l'ensemble des espèces. Ces acides

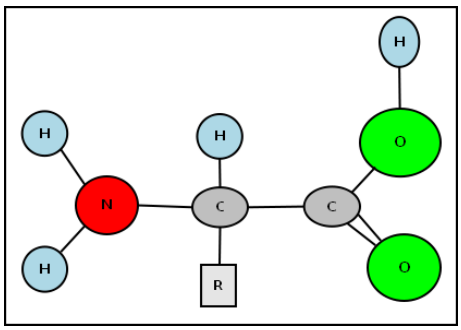


FIGURE 1.5 – Structure commune à tous les acides aminés.

aminés sont représentés par un code à une lettre ou par un code à trois lettres défini par l'IUPAC (*International Union of Pure Applied Chemistry*) et présenté dans le tableau 1.1. La chaîne latérale d'un acide aminé lui confère des propriétés physico-chimiques particuliéres. On distingue ainsi les acides aminés aromatiques (His, Phe, Trp, Tyr), les acides aminés chargés (Arg, Asp, Glu, Lys), les acides aminés neutres (Ser, Cys), les acides aminés polaires ou hydrophiles (Asn, Gln,Gly, Met, Pro, The) et les acides aminés apolaires ou hydrophobes (Ala, Ile, Val, Leu).

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

TABLE 1.1 – Nomenclature des acides aminés.

### 1.2.4 Les liaisons peptidiques

Nous avons pu voir que les acides aminés sont au nombre de 20. Il existe un seul type de liaison permettant de relier deux acides aminés : c'est la liaison peptidique.

La liaison peptidique est une liaison formée durant la synthèse de la protéine. Elle est formée par une liaison covalente (liaison chimique) entre la groupement $\alpha-$aminé d'un premier acide aminé et le groupement carboxylique d'un second acide aminé 1.6. Suite à cette liaison une molécule d'eau est libérée.

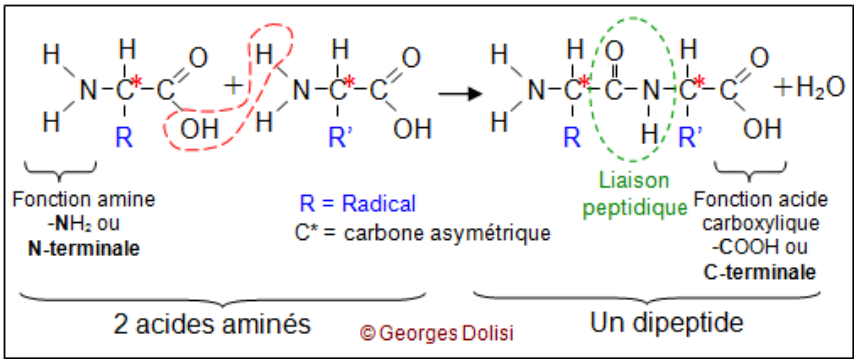


FIGURE 1.6 – Liaison peptidique entre deux acids aminés. Source : http ://www.medicalorama.com/encyclopedie/17369

Nous pouvons avoir différentes chaînes portant différents noms. Une chaîne de taille allons jusqu'à 5 acides aminés portera le nom de **tag**, alors qu'une composée de 50 acides aminés sera nommée **peptide**. Une **protéine** sera composé d'un ou plusieurs polypeptides ayant un nombre d'acides aminés très variable qui peut aller de moins de cent jusqu'à plusieurs milliers.

### 1.2.5 La structure des protéines

Les protéines sont donc composées par l'association d'acides aminés. Ces acides aminés sont placés dans un ordre précis qui caractérise la protéine, et que l'on appelle la structure primaire de la protéine (voir figure 1.7). Le début de la protéine dispose d'un acide aminé avec une extrémité amine, cette région est connue sous le nom de région N-terminale. De manière symétrique, l'autre extrémité possède un acide aminé qui présente une extrémité carboxylate, connue sous le nom de région C-terminale.

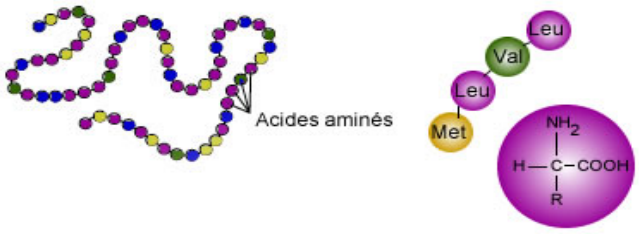


FIGURE 1.7 – Structure primaire d'une protéine. Source : http ://www.colvir.net/prof/chantal.proulx/701/chap2_contenu.htm.

Lors de sa synthèse, la protéine va acquérir une structure tridimensionnelle en plusieurs étapes. Tout d'abord, la chaîne protéique se replie pour constituer les structures secondaires avec la formation des hélices alpha et les feuillets bêta. La structure tertiaire dont un exemple est présenté dans la figure 1.8 correspond au repliement de la structure secondaire dans l'espace.

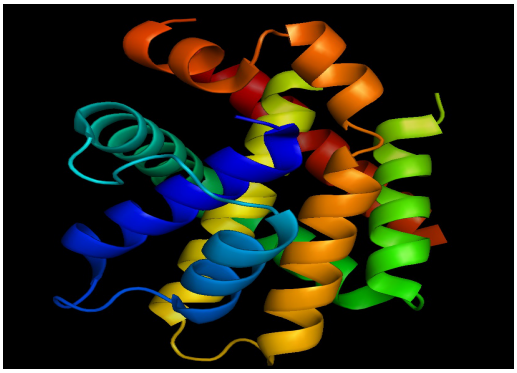

FIGURE 1.8 – Structure tertiaire d'une protéine [Kvansakul *et al.*,2010].

Enfin, la structure quaternaire se réfère à ce qu'on appelle les complexes protéiques qui correspondent à l'association d'au moins deux chaines de séquences protéiques (identiques ou différentes).

### 1.2.6 Les bases de données de protéines

Différentes bases de données proposent un regroupement des protéines selon différents critères. On trouve des bases de données fonctionnelles et structurales. Leur objectif est de regrouper les connaissances acquises sur les différentes protéines (structures 3D, annotations fonctionnelles, références bibliographiques, etc.) et de rendre accessible ces informations afin de pouvoir annoter et identifier de nouvelles séquences.

La construction de banques de données fonctionnelles nécessite tout d'abord de rassembler l'ensemble des séquences de la famille des protéines considérée selon leur fonction. Ensuite, il faut annoter les séquences en fonction des informations se trouvant dans la littérature et/ou des méthodes de prédictions.

Concernant les banques de données structurales, nous citons la plus connue des bases de données la Protein Data Bank (PDB) [Berman *et al.*, 2000]. Celle-ci dispose d'informations sur la structure 3D de nombreuses molécules biologiques, acides nucléiques et protéines. Afin de déterminer les structures 3D des protéines constituant la PDB, on utilise le plus souvent la méthode de cristallographie et la méthode de diffraction des rayons X. Cependant, l'accroissement exponentiel des données protéiques issues de séquençage de génomes [2] complets rend impossible l'identification de la fonction et l'étude de la structure 3D des protéines par des expérimentations biologiques. On ne dispose que de séquences primaires pour extraire le maximum d'informations. De plus, il est très difficile de déterminer la structure de certaines protéines car celles-ci sont difficiles à exprimer et à cristalliser ; c'est le cas des protéines membranaires [Tusnady *et al.*, 2005]. Le recours à des méthodes bio-informatiques pour traiter ces données représente une solution rapide qui permet une annotation fonctionnelle et une détermination de la structure de ces protéines.

## 1.3 Adressage des protéines

C'est en 1971 que le prix Nobel de médecine Gunter Blobel découvrit que la protéine a un signal qui gouverne son transport ainsi que sa localisation dans la cellule [Blobel and Dobberstein, 1975]. La destination finale d'une protéine est ainsi déterminée par un ensemble d'aiguillages successifs guidés par la présence de signaux d'adressage le long de la séquence. Chez les eucaryotes, le premier aiguillage a lieu en même temps que la traduction de la protéine, c'est-à-dire, de manière co-traductionnelle. En effet, la traduction de l'ARNm en protéine peut avoir lieu soit sur un ribosome libre dans le cytoplasme, soit sur un ribosome lié à la membrane du RE. La traduction des protéines sur un ribosome libre va produire les protéines destinées au cytosol, à la mitochondrie, aux chloroplastes, aux peroxysomes, ainsi qu'â l'intérieur du noyau, tandis qu'une traduction avec un ribosome lié à la membrane du RE fournira des protéines qui vont s'insérer dans la membrane du RE ou être délivrées vers la voie sécrétoire (RE, appareil Golgi, lysosomes, endosomes, membranes nucléaires, extérieur de la cellule).

2. Le génomes représente l'ensemble du matériel génétique d'un individu ou d'une espèce.

## 1.4 Importance de la localisation subcellulaire des protéines

Il est presque impossible d'estimer le nombre total des protéines qui existent dans la nature car chaque organisme fabrique quelque chose comme 100 000 protéines différentes les unes des autres. Chacune de ces protéines a une forme tridimensionnelle (section 2.5) et possède des propriétés chimiques qui lui sont propres afin de remplir de nombreux rôles dans la cellule (section 1.2.1). Nous avons aussi pu voir dans la section précédente que chaque protéine va être adressée vers un compartiment de la cellule afin d'exercer sa fonction. Il existe un lien très étroit entre la fonction de la protéine et sa localisation dans les différents compartiments de la cellule.

Parmi les types de protéines, nous nous intéressons particulièrement aux protéines membranaires. Ces protéines représentent 25-30% des protéines constituant le génome d'une espèce [Krogh *et al.*, 2001b; Cuthbertson *et al.*, June 2005]. Malgré leur nombre et leur importance, les protéines membranaires avec une structure tridimensionnelle connue représentent seulement 1% des protéines de la Protein Data Bank (PDB) [3] [Berman *et al.*, 2000]. Parmi ces protéines qui constituent la PDB et qui possèdent une structure connue, la moitié représentent des protéines membranaires insérées dans la membrane du RE. La difficulté de déterminer leur structure est due au fait que les protéines membranaires sont difficile à exprimer et à cristalliser [Tusnady *et al.*, 2005; Lukas, 2010]. De plus, les protéines membranaires jouent un rôle très important dans beaucoup de processus biologiques et elles sont la cible de nombreux développements pharmaceutiques. En effet, 50% de ces protéines sont ciblées par des médicaments [Cuthbertson *et al.*, June 2005; Terstappen and Reggiani, 2001].

Au niveau biologique les mécanismes d'adressage des protéines dans la cellule sont de mieux en mieux connus et s'avèrent plus complexes que prévu. Plusieurs équipes de recherche en Bio-Informatique s'intéressent à ces mécanismes et tentent de développer des méthodes pour prédire le site final de la synthèse des protéines. La prise en compte des nouvelles connaissances dans de nouveaux algorithmes doit permettre d'améliorer la qualité des prédiction. Dans le cadre de cette thèse, nous étudions les protéines qui transitent par le RE, en cherchant à distinguer celle qui vont s'insérer dans la membrane de celles qui vont être relâchées au niveau de la lumière du RE.

3. PDB : Base de données de collecte de protéines.

# Chapitre 2

# Localisation subcellulaire : difficulté à discriminer entre un peptide signal et un segment transmembranaire

Il existe à ce jour plus d'une centaine de méthodes pour la localisation cellulaire des protéines. Ces méthodes mettent en œuvre principalement des méthodes d'apprentissage supervisé telles que les réseaux de neurones, les machines à vecteurs de support, et les modéles de chaines de Markov cachées. Plusieurs dizaines de ces méthodes s'intéressent tout particulièrement à la prédiction des protéines transmembranaires ou des peptides signaux. Ces méthodes exploitent principalement des biais de composition en acides aminés ou la détection de signaux. Leurs performances sont parfois améliorées en exploitant des informations d'homologie ou en combinant plusieurs méthodes.

À partir de 2005, de nouvelles expérimentations biologiques ont permis de mieux comprendre les mécanismes biologiques d'insertion des protéines dans la membrane. De toutes nouvelles méthodes de prédiction s'appuyant sur ces connaissances fraîchement acquises sont alors apparues. Ce sont également ces nouvelles connaissances que nous cherchons à exploiter dans le cadre de cette thèse.

**Sommaire**

## 2.1 Introduction

Nous définissons la localisation subcellulaire comme l'identification du compartiment fonctionnel d'une protéine. Plusieurs techniques expérimentales permettent de déterminer la localisation d'une protéine. La méthode la plus courante est le marquage avec un rapporteur fluorescent (GFP pour « Green Fluorescent Protein ») à l'une des deux extrémités N-terminale ou C-terminale de la protéine. La localisation de la protéine est ensuite observée au microscope optique fluorescent. D'autres méthodes comme l'utilisation de la spectrométrie de masse sont également disponibles. Cependant, l'ensemble de ces méthodes présentent des limites : faible résolution de la microscopie optique, difficulté à analyser certaines protéines comme les protéines membranaires en spectrométrie de masse. De plus, la réalisation de ces expérimentations est très lourde à mettre en œuvre.

La disponibilité de méthodes informatiques de prédiction subcellulaire fiable offre des possibilités de traitement à grande échelle avec un coût très réduit. Dans le cadre de cette thèse, nous allons nous intéresser à la prédiction de la localisation des protéines qui transitent par le RE en abordant des problémes qui restent encore ouverts : une meilleure compréhension du mécanisme d'insertion des protéines dans les membranes d'une part, et une meilleure discrimination entre le signal d'adressage appelé le peptide signal et les segments transmembranaires.

Dans la première partie de ce chapitre, nous présentons le mécanisme général d'adressage des protéines qui transitent par le RE, puis nous présentons les principes sur lesquels se fondent les différentes méthodes de prédiction de la localisation de ces protéines. Dans la section 2.3.3, nous étudions plus en détail les méthodes qui cherchent à distinguer les protéines transmembranaires des protéines sécrétées. La section 2.4 introduit de nouvelles connaissances biologiques et présente les différentes méthodes s'appuyant sur ces connaissances. La section 2.5 conclut ce chapitre.

## 2.2 Translocation des protéines de la voie de sécrétion

Les protéines sont synthétisées à l'aide des ribosomes qui sont des complexes protéiques. Chez les eucaryotes, les ribosomes sont sous forme libre dans le cytoplasme ou liés aux membranes du RE. Les protéines synthétisées par les ribosomes libres sont par exemple les protéines du noyau, de la mitochondrie, des peroxysomes tandis que les protéines synthétisées sur des ribosomes liés sont des protéines destinées au milieu extracellulaire, les protéines membranaires ainsi que les protéines intermédiaires (RE, Golgi, vésicules sécrétoires...).

C'est une courte séquence d'acides aminés qui se trouve à l'extrémité N-terminale des protéines qui est responsable de l'adressage des protéines synthétisées depuis le cytosol vers le RE [Osborne *et al.*, 2005]. Cette courte séquence d'environ 20 à 30 acides aminés [Ng *et al.*, 2007] est appelée **peptide signal** (PS). Elle est reconnue par le SRP (Signal Recognition Particle) qui va fixer le ribosome sur la membrane du RE au niveau d'un canal de translocation appelé **translocon**. L'ensemble du processus d'adressage est schématisé dans la figure 2.1.

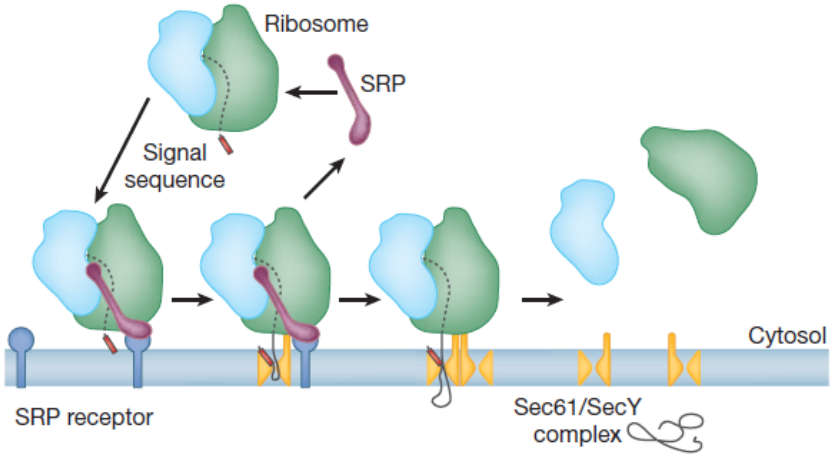


FIGURE 2.1 – Modèle de la translocation co-traductionnelle [Rapoport, 2007].

[van den Berg *et al.*, 2004] ont étudié et cristallisé la structure du translocon montrant un canal constitué de plusieurs protéines transmembranaires ménageant un pore, qui permet le passage de la chaine protéique en cours de synthèse du ribosome vers la lumière du RE (figure 2.2). On parle ainsi de translocation co-traductionnelle. Lorsque le peptide signal atteint la lumière du RE, il est coupé et libéré [Osborne *et al.*, 2005].

Le translocon présente aussi une ouverture latérale qui, lorsqu'elle est ouverte, permet à certaines parties de la protéine de s'insérer dans la membrane. Les protéines qui traversent entièrement le translocon pour rejoindre la lumière du RE sont les **protéines solubles**. Les protéines dont certaines parties sont insérées dans la membrane sont les **protéines membranaires**. Les parties de la protéine qui sont insérées dans la membrane sont appelés les **segments transmembranaires** ou **segments TM**.

### 2.2.1 Signal peptide et signal anchor

Le peptide signal est la courte séquence d'acides aminés située en partie N-terminale de la protéine qui permet l'adressage au réticulum endoplasmique. Ce signal est donc nécessaire pour les protéines de la voie de sécrétion et les protéines membranaires. Depuis très longtemps déjà, des études comparatives ont permis de mettre en évidence l'organisation structurelle des peptides signaux en trois régions (figure 2.3) différentes : une région chargée positivement appelée n-région, suivie d'une région hydrophobe - h-région

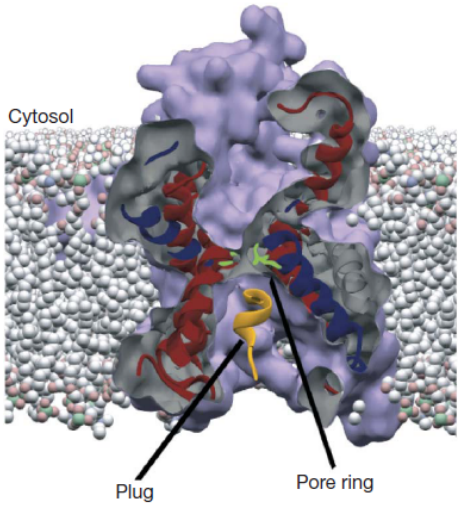


FIGURE 2.2 – Canal de translocation [Rapoport, 2007].

-, et enfin la c-région contenant le site de coupure [Dalbey *et al.*, 1997; Ng *et al.*, 2007]. En général, la taille de la région hydrophobe varie entre 7-15 acides aminés, tandis que la taille de la n-région varie de 1-12 acides aminés, et enfin la taille de la c-région varie de 3-8 acides aminés [Kall *et al.*, 2004]. On notera cependant que certains peptides signaux ont des longueurs plus importantes [Hiss and Schneider, 2009]. Un peptide signal est considéré comme étant un long peptide signal si sa taille est supérieure à 40. Ils sont beaucoup plus difficiles à détecter par les logiciels de prédiction du peptide signal.

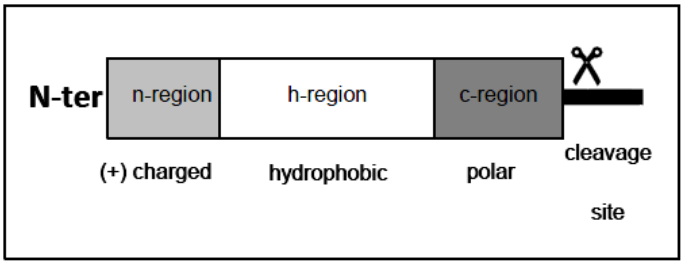


FIGURE 2.3 – Structure d'un peptide signal avec 3 régions.

Bien que sans peptide signal, certaines protéines ont une séquence d'acides aminés hydrophobes en partie N-terminale qui initie leur translocation de la même manière qu'un peptide signal. Cependant, à la différence du peptide signal qui est coupé quand il atteint la lumière du RE, cette séquence reste ancrée dans la membrane [Nielsen and Krogh, 1998]. Ainsi, cette séquence joue un double rôle : elle est responsable à la fois de l'adressage au réticulum endoplasmique et de l'ancrage de la protéine dans la membrane. Elle ne posséde pas de site de coupure et elle est généralement plus longue qu'un peptide signal (18 à 25 résidus) pour former une hélice transmembranaire [Goder and Spiess, 2001]. De plus, elle

se trouve généralement un peu plus à l'intérieur de la séquence protéique plutôt qu'â sa toute extrémité N-terminale. On l'appelle « **Signal Anchor** » ou aussi Signal d'Ancrage (SA).

### 2.2.2 Protéines membranaires et segments transmembranaires

Les protéines membranaires représentent 25 à 30% des protéines dans les génomes complets actuellement séquencés [Krogh *et al.*, 2001b]. Ces protéines membranaires jouent un rôle très important dans plusieurs processus cellulaires. Elles sont impliquées dans le transport de molécules et la signalisation entre les cellules [Cuthbertson *et al.*, June 2005; Tusnady *et al.*, 2005]. De plus, ces protéines ont un intérêt au niveau pharmaceutique, puisque 50% de ces protéines sont ciblées par un médicament [Terstappen and Reggiani, 2001].

Les protéines membranaires appelées **protéines bitopiques** (traversant une seule fois la membrane) ont un seul segment transmembranaire (segment TM), tandis que les **protéines polytopiques** (traversant plusieurs fois la membrane) ont plusieurs segments transmembranaires [Tan *et al.*, 2008]. Jusqu'â récemment, on pensait que dans le cas des protéines ayant plusieurs segments transmembranaires, l'insertion dans la membrane se faisait de manière ordonnée en progressant à travers le translocon séquentiellement [Sadlish *et al.*, 2005]. Cependant, on sait maintenant que pour un certain nombre de protéines, certains segments entrent dans la membrane de manière groupée [Skach, 2007].

Comme toutes les membranes, la membrane du RE est majoritairement constituée de lipides qui ont tendance à orienter leur partie hydrophobe vers l'intérieur de la membrane et leur partie hydrophile vers l'extérieur. Ainsi, la membrane est souvent schématisée par un coeur hydrophobe et deux zones d'interfaces avec le RE d'une part et le cytosol d'autre part. à l'intérieur de la membrane, les segments transmembranaires d'une protéine peuvent prendre la forme d'hélice alpha (HTM) ou la forme de brin béta (BTM). La structure en hélice alpha (figure 2.4) est de très loin la plus fréquente [Reynolds *et al.*, 2008]. Généralement la taille des segments TM est de l'ordre de 15 à 30 acides aminés avec une très grande région hydrophobe [Kall *et al.*, 2004].

## 2.3 Principe général des méthodes de prédiction

La plupart des méthodes de prédiction se fondent sur l'information donnée par la séquence primaire, c'est-à-dire la suite d'acides aminés de la protéine. Dans tous les cas, les différentes méthodes exploitent des biais de composition en acides aminés et/ou la détection d'un signal -motif particulier- afin d'effectuer leur prédiction. Il existe de très nombreuses méthodes de prédiction. Certaines ont une couverture très large en terme de localisation prise en compte, tandis que d'autres sont focalisées sur la recherche d'un motif ou d'une localisation spécifique. Dans le cadre de cette thèse, nous allons limiter notre présentation à trois catégories de méthodes de prédiction : les méthodes de prédiction du peptide signal, les méthodes de prédiction des segments TM, et les méthodes avec double prédiction du peptide signal et du segment TM.

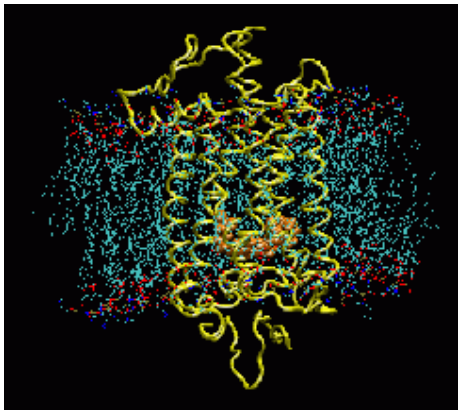

FIGURE 2.4 – Représentation d'une protéine transmembranaire. Source : http ://www.ks.uiuc.edu/Research/rhodopsin/

### 2.3.1 Détection du peptide signal

Généralement, la prédiction des protéines sécrétées consiste à reconnaître la présence d'un peptide signal, et dans la plupart des cas à positionner son site de coupure. Les premières méthodes de prédiction utilisaient une matrice de poids afin d'identifier le peptide signal [Mcgeoch, 1985; Von Heijne, 1986a]. Ces approches donnaient de bonnes performances surtout limitées à la reconnaissance des sites de coupure, car il n'est pas possible d'obtenir une séquence consensus à partir de l'alignement de peptides signaux, ceux-ci étant trop variables.

Les approches basées sur les réseaux de neurones [Nielsen *et al.*, 1997a], les machines à vecteurs de support [Vert, 2002], et les modèles de chaines de Markov cachées [Nielsen *et al.*, 1999] ont rapidement permis d'améliorer les performances de prédiction en tirant partie de son organisation en 3 régions de longueur variable.

Aujourd'hui, de très nombreux autres outils existent pour la localisation du peptide signal et de son site de coupure. Certains de ces outils sont spécifiques à certains organismes qu'ils soient eucaryotes, procaryotes, limités aux plantes, etc. Une description et une évaluation de certains de ces outils est disponible dans [Klee and Sosa, 2007; Klee and Ellis, 2005; Zhang *et al.*, 2009].

Parmi les outils les plus populaires, on retrouve toujours le logiciel SignalP. Ce prédicteur est l'un des premiers programmes à avoir utilisé les réseaux de neurones pour l'identification du peptide signal et de son site de coupure [Nielsen *et al.*, 1997b]. Plusieurs améliorations de SignalP ont été proposées au fil du temps avec des versions basées sur des réseaux de neurones et une version développée avec des chaînes de Markov cachées (Hidden MM). La version du prédicteur utilisant les HMM possède l'avantage de mieux discriminer le signal anchor du peptide signal. La dernière version de SignalP - version 3.0 - [Bendtsen *et al.*, 2004] modélise de manière plus précise la composition en acides aminés du site de coupure. Les outils signalP en versions V2 et V3 fournissent les meilleurs taux de prédiction du peptide signal et de son site de coupure [Klee and Ellis, 2005].

### 2.3.2 Détection du segment TM

Il existe de nombreux outils de prédiction des protéines membranaires. Le principe de prédiction est fondé sur l'idée de discriminer les parties transmembranaires des parties non transmembranaires et ainsi de déterminer combien de domaines transmembranaires la protéine possède. Par ailleurs, d'autres informations sont susceptibles d'être ajoutées comme l'orientation des segments et leurs positions précises.

La plupart des méthodes de prédiction de la topologie membranaire sont seulement spécialisées dans la prédiction de certains types de protéines membranaires. Il existe des méthodes pour prédire les segments TM structurés en hélices alpha [Sonnhammer *et al.*, 1998a; Ahmed *et al.*, 2010] et d'autres pour les segments TM structurés en feuillets bêta [Bagos *et al.*, 2004; Pagos *et al.*, 2005]. Les méthodes de prédiction des segments TM structurés en hélices alpha TM sont de loin les plus nombreuses en raison de la forte prédominance de ce type de structure d'une part et d'une identification plus aisée d'autre part. On dispose en effet de plus de données sur leur sujet, ce qui facilite l'élaboration de méthodes de prédiction.

Une des premières méthodes de prédiction de protéines membranaires utilise l'information se trouvant dans la propriété d'hydrophobie des acides aminés. Cette méthode élaborée par Kyte et Doolittle (KD) [Kyte and Doolittle, 1982] utilise une échelle d'hydrophobie des 20 acides aminés (voir tableau 2.1). Chaque acide aminé est représenté par un score lié à sa préférence pour l'eau ou les lipides.

À l'époque, on a observé que les acides aminés les plus hydrophobes (Ile, Leu, Val, Phe) étaient surreprésentés dans les hélices TM. Grâce à ce biais de composition, les segments transmembranaires peuvent être distingués des segments non TM.

| Acide aminé | | Acide aminé | | Acide aminé | | Acide aminé | |
|---|---|---|---|---|---|---|---|
| Ala (A) : | +1.8 | Gln (Q) : | -3.5 | Leu (L) : | +3.8 | Ser (S) : | -0.8 |
| Arg (R) : | -4.5 | Glu (E) : | -3.5 | Lys (K) : | -3.9 | Thr (Y) : | -0.7 |
| Asn (N) : | -3.5 | Gly (G) : | -0.4 | Met (M) : | +1.9 | Trp (W) : | -0.9 |
| Asp (D) : | -3.5 | His (H) : | -3.2 | Phe (F) : | +2.8 | Tyr (T) : | -1.3 |
| Cys (C) : | +2.5 | Ile (I) : | +4.5 | Pro (P) : | -1.6 | Val (V) : | +4.2 |

TABLE 2.1 – Échelle d'hydrophobie de Kyte et Doolittle.

La méthode développée par Kyte et Doolittle consiste à faire glisser une fenêtre de taille fixe (taille du segment TM) le long de la séquence protéique. Pour chaque sous-séquence correspondant à la fenêtre glissée, on calcule une moyenne d'hydrophobie en utilisant les valeurs de l'échelle d'hydrophobie du tableau 2.1 où une valeur numérique représentant la propriété d'hydrophobie est associée à chaque acide aminé. Si la valeur de la moyenne calculée est supérieure à un certain seuil alors la sous-séquence d'acides aminés correspond à un segment TM autrement la sous-séquence est un segment non TM.

La figure 2.5 est obtenue à l'aide de l'outil ProtScale[1] [Gasteiger *et al.*, 2005]. Elle

1. http ://expasy.org/tools/protscale.html

présente le profil d'une protéine calculé en faisant glisser une fenêtre de taille 19 (segment TM de 19 aicdes aminés) et en utilisant l'échelle d'hydrophobie de Kyte et Doolittle.

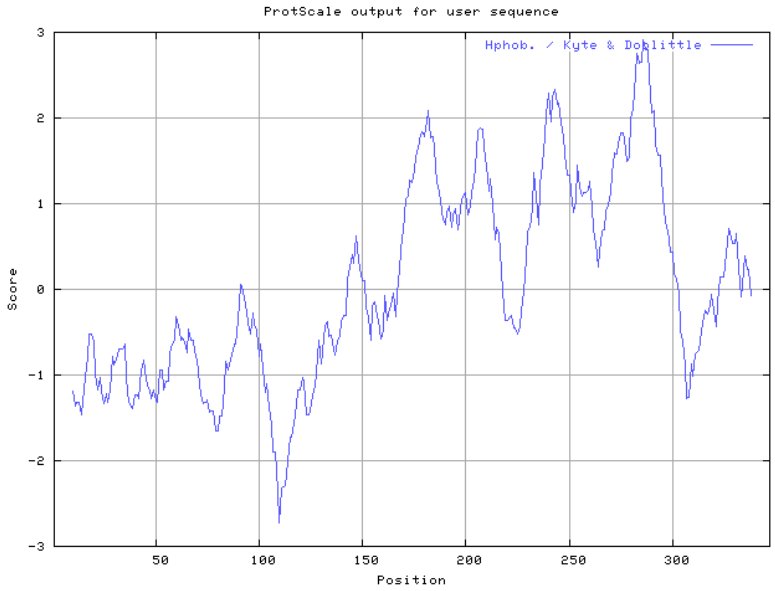


FIGURE 2.5 – Profil d'hydrophobie d'une protéine. L'axe des ordonnées montre les scores correspondant à la moyenne d'hydrophobie calculée sur chaque sous fenêtre de longueur 19 acides aminés et l'axe des abscisses montre les différentes positions de la protéine.

Les méthodes qui utilisent lors de la prédiction une échelle de valeurs sont les plus simples à mettre en oeuvre. Elles ont l'avantage d'être visuelles, et peu consommatrices en temps de calculs. Cependant, ces méthodes sont limitées dans la reconnaissance des segments transmembranaires.

Une autre méthode utilise la comparaison de séquences. DAS (Dense Alignment Surface) est fondée sur le principe du « dot-plot [2] » de deux protéines [Cserzo *et al.*, 1997].

Si deux segments d'une certaine taille de deux protéines sont identifiés avec une forte similarité et dépassant un seuil, alors les régions correspondantes du dot plot seront marquées. La similarité utilisée par DAS est calculée grâce à une matrice de substitution, qui est traduite par un score de différentes mutations possibles d'un acide aminé à un autre. La prédiction d'une protéine par DAS s'effectue par la comparaison de celle-ci par dot-plot à l'ensemble des protéines ayant une topologie déjâ caractérisée expérimentalement. Par la suite, un profil est construit en sommant les scores du dot-plot pour chaque position de la séquence. Le profil moyen qui dépasse un seuil est considéré comme zone membranaire. Le développement d'une telle méthode est très dépendant du choix de la base de données de protéines, le choix de la matrice de substitution et le choix du seuil de similarité.

Tout comme dans le cas des outils de prédiction des peptides signaux, l'accroissement

2. Le dot-plot est une méthode graphique qui permet la comparaison de deux séquences afin d'identifier les régions de forte similarité.

des données structurales et biochimiques sur les protéines membranaires a donné naissance à plusieurs méthodes fondées sur une machine d'apprentissage. De nombreux travaux utilisent les réseaux de neurones [Rost *et al.*, 1995] ou les Machines à vecteurs de support [Garg *et al.*, 2005]. D'autres comme [Krogh *et al.*, 2001a; Nilsson *et al.*, 2000; Tusnady and Simon, 2001] sont fondées sur les HMM pour la prédiction des segments TM.

Dans un modèle de Markov, une séquence de protéine peut être vue comme un automate où les transitions sont étiquetées par des probabilités. Ainsi, un HMM est défini par un quintuplet $(S, \Sigma, T, G, \pi)$ où S est un ensemble d'états, $\Sigma$, est un alphabet de symboles, T la matrice indiquant les probabilités de transition d'un état à un autre, G la matrice indiquant les probabilités de génération associées aux états, $\pi$ le vecteur de génération de probabilités initiales de visites.

Nous présentons l'outil TMHMM [Sonnhammer *et al.*, 1998b], qui se base sur le modèle de Markov caché en construisant une architecture proche du système biologique. TMHMM définit un ensemble d'états, chacun modélisant une région dans les protéines à modéliser.

Le modéle TMHMM définit 7 régions différentes composées d'un ou plusieurs états qui sont :

- la région coeur de l'hélice TM,
- l'hélice TM coté cytoplasme,
- l'hélice TM coté non-cytoplasme,
- la boucle cytoplasmique (celle-ci représente les acides aminés entre 2 segments TM cotés cytosol),
- la boucle courte non cytoplasmique (celle-ci représente les acides aminés entre 2 segments TM cotés non cytosol),
- la boucle longue non cytoplasmique,
- en dernier la partie segment non TM.

L'architecture TMHMM est présentée en figure 2.6.

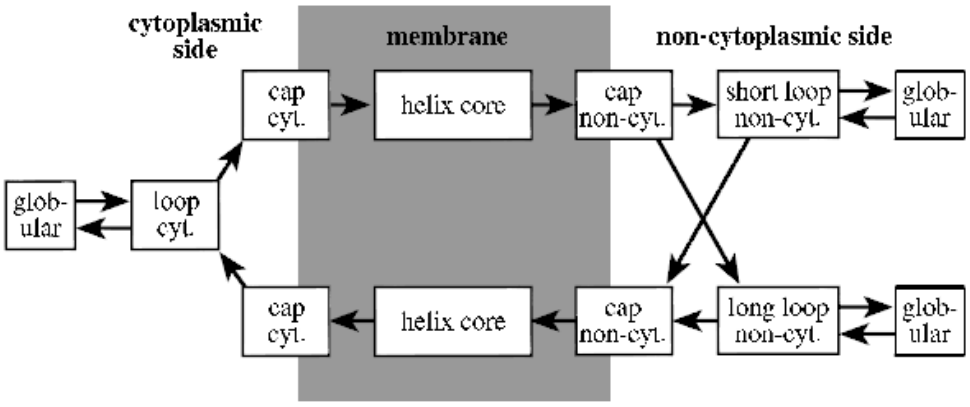


FIGURE 2.6 – Architecture du modéle TMHMM [Sonnhammer *et al.*, 1998b].

Cependant, la construction d'un tel modèle a besoin d'entrainement, appelé phase d'apprentissage pour déterminer les probabilités d'émission des états c'est-à-dire la pro-

babilité de rencontrer chacun des acides aminés dans cet état ainsi que les probabilités de transition d'un état à un autre.

De nombreuses méthodes utilisent le modèle de Markov caché pour la prédiction de la structure des protéines membranaires. Cependant, l'apprentissage d'un tel modèle nécessite un jeu de données très important et sans biais où la topologie membranaire est connue à l'avance afin de pouvoir identifier les régions transmembranaires. Parmi les méthodes disponibles, HMMTOP2 [Tusnady and Simon, 2001] et TMHMM2 [Krogh *et al.*, 2001a] sont les plus performantes [Cuthbertson *et al.*, June 2005; Punta *et al.*, 2007].

### 2.3.3 Double prédiction peptide signal et segment transmembranaire

Les segments transmembranaires de type hélice alpha contiennent une région hydrophobe tout comme la H-région du peptide signal. La similarité entre ces deux régions induit souvent une confusion : les méthodes de prédiction des segments TM ont tendanceà prédire la région hydrophobe du peptide signal comme un segment TM [Chen *et al.*, 2002], alors que les méthodes de prédiction du peptide signal prédisent la partie hydrophobe du segment TM en région N-terminale comme élément du peptide signal [Nielsen and Krogh, 1998; Nielsen *et al.*, 1999].

Les travaux de [Lao *et al.*, 2002] montrent que la présence du peptide signal dans les segments TM affecte significativement la prédiction de la topologie de protéines membranaires. Pour remédier à ce problème, différentes méthodes tentent de différentier le peptide signal d'un segment transmembranaire.

[Nielsen and Krogh, 1998] implémentent un modèle fondé sur des HMM pour modéliser les différentes régions du peptide signal et du signal anchor. Un modèle dédié à la prédiction du peptide signal implémente trois états correspondants aux 3 régions (n-région, h-région, et c-région), le modèle du signal anchor implémente seulement deux régions (n-région et h-région). L'estimation de la méthode montre une spécificité de 81% et une sensibilité de 71% dans la discrimination entre le peptide signal et le signal anchor.

[Zheng *et al.*, 2003] proposent également une méthode de discrimination entre le peptide signal en région N-terminale et les segments transmembranaires. Les auteurs se limitent aux 70 premiers acides aminés de la séquence protéique. En effet, le peptide signal a rarement été observé après les 70 premiers acides aminés [Nielsen *et al.*, 1997a] et le segment TM après les 70 premiers acides aminés ne peut être confondu avec un PS.

La discrimination entre la classe contenant des peptides signaux et la classe contenant des segments TM est réalisée en utilisant une fonction qui combine trois propriétés. La première propriété mesure la fréquence d'apparition de chaque acide aminé dans la séquence de 70 acides aminés, la deuxième propriété mesure l'hydrophobie d'une fenêtre glissante de taille 22 acides aminés (taille du segment TM) en utilisant les valeurs d'hydrophobie de l'échelle de Kyte et Doolittle, et enfin la position de la fenêtre ayant la plus grande valeur d'hydrophobie.

L'évaluation des performances de la méthode est menée selon le principe du test de

jack-knife [Efron, 1979] et annonce un taux de bonne prédiction de 91.5% sur un jeu de données constitué d'un ensemble de 943 peptides signaux et d'un ensemble de 272 protéines membranaires.

Une autre méthode dédiée à la double prédiction du PS et du segment TM est Phobius [Kall *et al.*, 2004; Kall *et al.*, 2007]. Cette méthode s'appuie également sur un modèle HMM qui combine une double prédiction du peptide signal et de la topologie de la protéine, si celle-ci est transmembranaire. Le modéle utilisé par Phobius est présenté comme la combinaison de deux modèles : le premier est utilisé pour la détection des segments transmembranaires, tandis que le deuxième est dédié à la modélisation du peptide signal. Chaque modéle est découpé en états correspondants aux régions permettant de représenter au mieux le modèle biologique. Une transition entre les états est arrangée de telle sorte que la position du PS est toujours localisée en partie N-terminale de la séquence. La région hydrophobe du PS est rarement localisée après les 30 premiers acides aminés, tandis que les segments TM peuvent être localisés à n'importe quelle position de la séquence.

L'évaluation des performances est conduite sur quatre collections différentes : une collection avec des protéines qui contiennent seulement des segments TM, une collection avec des protéines qui n'ont qu'un PS, une collection avec des protéines qui contiennent des segments TM et un PS, et enfin une collection avec des protéines qui ne contiennent ni de PS ni de segments TM. Pour la première collection Phobius présente un taux de 63.6% de bonne prédiction, la deuxième collection 96.1% de bonne prédiction, la troisième collection 91.1% de bonne prédiction, et enfin 98.2% de bonne prédiction pour la dernière collection.

Dans la même idée de la méthode Phobius, [Reynolds *et al.*, 2008] proposent la méthode Philius. Cette méthode combine la prédiction de la topologie membranaire et la localisation du peptide signal en utilisant des réseaux Bayesian dynamiques (DBN). La méthode imite Phobius en construisant un modéle de prédiction pour chaque région de la topologie telle que la région cytoplasmique, la région segment membranaire, la région non cytoplasmique, et la région peptide signal. Dans l'étude expérimentale, Philius se compare à Phobius. Les résultats montrent des performances proches de son concurrent.

Philius identifie donc quatre types de segments : le peptide signal, le segment TM, l'intérieur et l'extérieur de la boucle. Pour le segment TM, le segment prédit doit présenter un chevauchement avec le segment annoté dans la base de données par au moins 5 acides aminés pour être identifié correctement. Afin d'identifier correctement un peptide signal, Philius doit prédire seulement son existence en partie N-terminale de la protéine. Pour la prédiction des boucles, le modèle utilise un chevauchement de seulement 1 acide aminé, car la taille des boucles est trop courte.

La dernière méthode, SPOCTOPUS [Viklund *et al.*, 2008] est une méthode combinant prédiction du peptide signal et la détermination de la topologie des protéines membranaires. SPOCTOPUS est une extension d'un outil de prédiction de la topologie membranaire appelée OCTOPUS [Viklund and Elofsson, 2008], avec l'ajout de la prédiction du peptide signal. L'outil combine un modèle de réseaux de neurones et des chaînes de Markov cachées. SPOCTOPUS prend comme données d'entrées des séquences profils qui sont créés

en exécutant l'outil BLAST [3]. Ainsi, l'utilisation de profil permet une recherche beaucoup plus sensible de séquences homologues [4] que l'utilisation d'une séquence seule, car le profil contient de l'information sur la variabilité des différentes positions des acides aminés. L'évaluation des performances de SPOCTOPUS est conduite par un processus de 10 validations croisées. Bien que la méthode présente une bonne performance de prédiction, il est très difficile d'interpréter le phénomène biologique d'insertion des protéines membranaires par les résultats obtenus de SPOCTOPUS.

### 2.3.4 Les limites des approches actuelles

Les méthodes de prédiction de la topologie membranaire annoncent des taux de bonne prédiction de 70 à 85 % de toutes les protéines. Cependant, les études sur l'ensemble des données du génome montrent qu'il y a une sur estimation de leurs performances [Kall and Sonnhammer, 2002; Melen *et al.*, 2003]. Certaines méthodes manquent la prédiction de plusieurs segments TM le long de la protéine, tandis que d'autres détectent des segments TM qui n'en sont pas (faux positifs) et présentent des segments TM qui n'existent pas dans la protéine [Klammer *et al.*, 2009].

Malgré les bonnes performances des méthodes de prédiction, il est souvent très difficile de lier les résultats obtenus à la compréhension du mécanisme biologique d'insertion des protéines dans la membrane. Dans les méthodes de prédiction comme Phobius et Philius, il existe toujours une confusion dans la prédiction du peptide signal et du signal anchor qui se trouvent tous les 2 dans la région N-terminale de la protéine.

## 2.4 Nouvelles connaissances sur l'insertion des protéines membranaires dans la membrane du RE

### 2.4.1 Reconnaissance des protéines membranaires par le translocon

En 2005, Hessa *et al* ont mis au point un système d'expression *in vitro* [5] [Saaf *et al.*, 1998], qui permet une évaluation quantitative de l'efficacité d'insertion d'un segment polypeptidique, dit H-segment de taille 19 acides aminés dans une membrane.

Le H-segment est inséré dans une protéine connue, la protéine LEP présentée figure 2.7. Cette protéine contient nativement 2 segments transmembranaires et lorsque sa traduction dans un système in vitro est faite en présence de microsomes formés par de la membrane de RE, elle s'insére dans cette membrane. Un repérage basé sur la glycosylation [6] permet alors de savoir quelle est la proportion de segments H insérés (ésglycosylation f1g) ou non (glycosylation f2g) dans la membrane. Cette proportion peut être facilement convertie en

3. BLAST (basic local alignment search tool) permet de réaliser un alignement de séquences protéiques afin de calculer la similarité entre deux ou plusieurs séquences.

4. homologie : 2 séquences sont homologues si elles ont un ancêtre commun.

5. in vitro signifie une expérimentation menée en dehors de la cellule, elle est réalisée principalement en éprouvette.

6. La glycosylation est la modification post-traductionnelle qui ajoute des sucres. Les protéines glycosylées sont destinées à être sécrétées ou intégrées à la membrane plasmique.

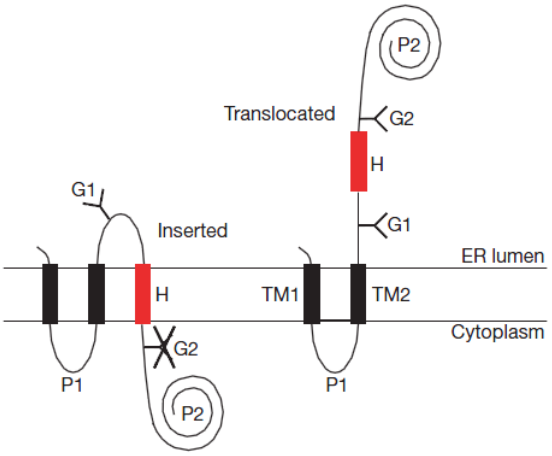


FIGURE 2.7 – Système d'évaluation de l'efficacité d'insertion du H-segment dans la membrane [Hessa *et al.*, 2005a]. La peptidase d'Escherichia Coli (Lep) a deux hélices transmembranaires (TM1 et TM2) et un grand domaine luminal (P2). Le H-segment (rouge) montre le segment TM dans le domaine P2 avec deux emplacements de glycosylations (G1, G2). Dans la partie gauche de la figure, on observe un segment inséré dans la membrane du RE (une seule glycosylation à l'emplacement G1), tandis que, dans la partie droite le H-segment est transféré à travers la membrane (double glycosylation sur les emplacements G1 et G2.)

énergie libre apparente grâce à l'équation : $\Delta G_{app} = -RT \ln K_{app}$ où R est la constante des gaz, T est la température de l'expérience et $K_{app} = f1g/f2g$. On remarquera que lorsque la proportion de segments insérés est en équilibre avec la proportion de segments non insérés, $\Delta G_{app} = 0$. On considérera que l'énergie d'insertion donnée par le $\Delta G_{app}$ représente la somme des contributions de chaque acide aminé $\Delta G_{app} = \sum_{i=1}^{l} \Delta G_{app}^{aa(i)}$.

On dispose donc d'un système mettant en œuvre la machinerie du translocon, très proche de la réalité biologique pour mesurer l'efficacité d'insertion de différents segments H. Tous les segments H ont été construitsà partir de la séquence $GGPG-(L_nA_{19-n})-GPGG$.

Dans un premier temps, Hessa *et al* ont étudié l'insertion de segments composés de proportions variables d'alanines (A) et de leucines (L). Puis à partir d'une proportion déterminée, ils ont successivement remplacé l'acide aminé central par l'un des 20 acides aminés dont on veut mesurer l'efficacité d'insertion. Ces mesures ont permis d'élaborer une échelle d'hydrophobie appelée « biological hydrophobicity scale » qui quantifie la contribution de chacun des 20 acides aminés. Cette échelle donnée en mesure d'énergie $\Delta G_{app}$ est présentée en figure 2.8.

Dans une nouvelle étude Hessa *et al* [Hessa *et al.*, 2007] ont déterminé pour chaque acide aminé testé, le nombre respectif d'acides aminés L et A de telle sorte que l'on on ait un équilibre entre le nombre de segments H insérés et non-insérés dans la membrane avec l'acide aminé testé en position centrale. Puis, la position de cet acide aminé a été successivement déplacé sur les 19 positions du H segment (voir figure 2.9).

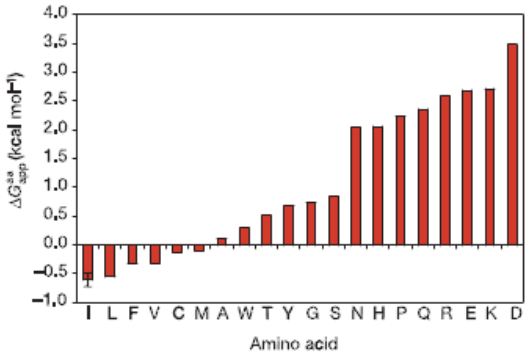


FIGURE 2.8 – Échelle d'hydrophobie présentée en énergie [Hessa *et al.*, 2005a].

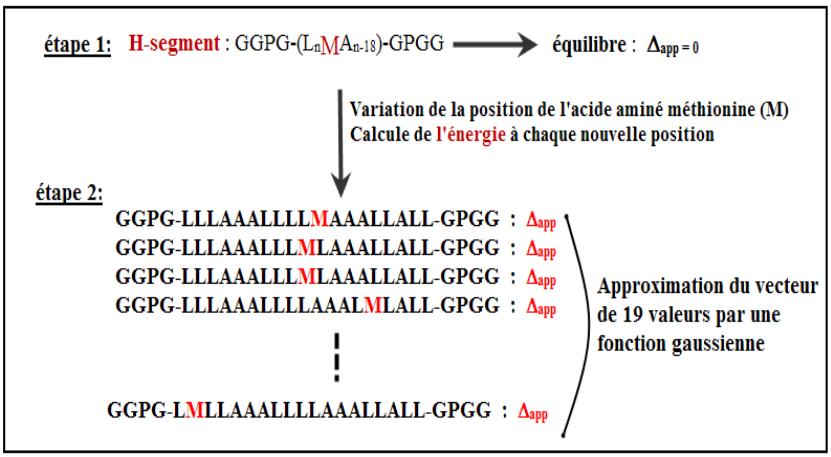


FIGURE 2.9 – Illustration du systéme utilisé pour calculer la contribution des acides aminés.

La mesure de la proportion des segments insérés par rapport aux segments non insérés permet d'obtenir une matrice de valeurs du $\Delta G_{app}$. Cette matrice est représentée par 20 lignes correspondant aux 20 acides aminés et par 19 colonnes correspondant à la taille du segment TM. Hessa *et al* utilisent 324 H-segments de taille 19 acides aminés pour calculer les valeurs du $\Delta G_{app}$. À l'exception du tryptophan (W) et de la tyrosine (Y), les profils des acides aminés obtenus par les 19 valeurs sont alors approximés par une fonction gaussienne simple. Les profils du tryptophan (W) et de la tyrosine (Y ) sont approximés par des doubles gaussiennes. Les courbes obtenues sont présentées en figure 2.10.

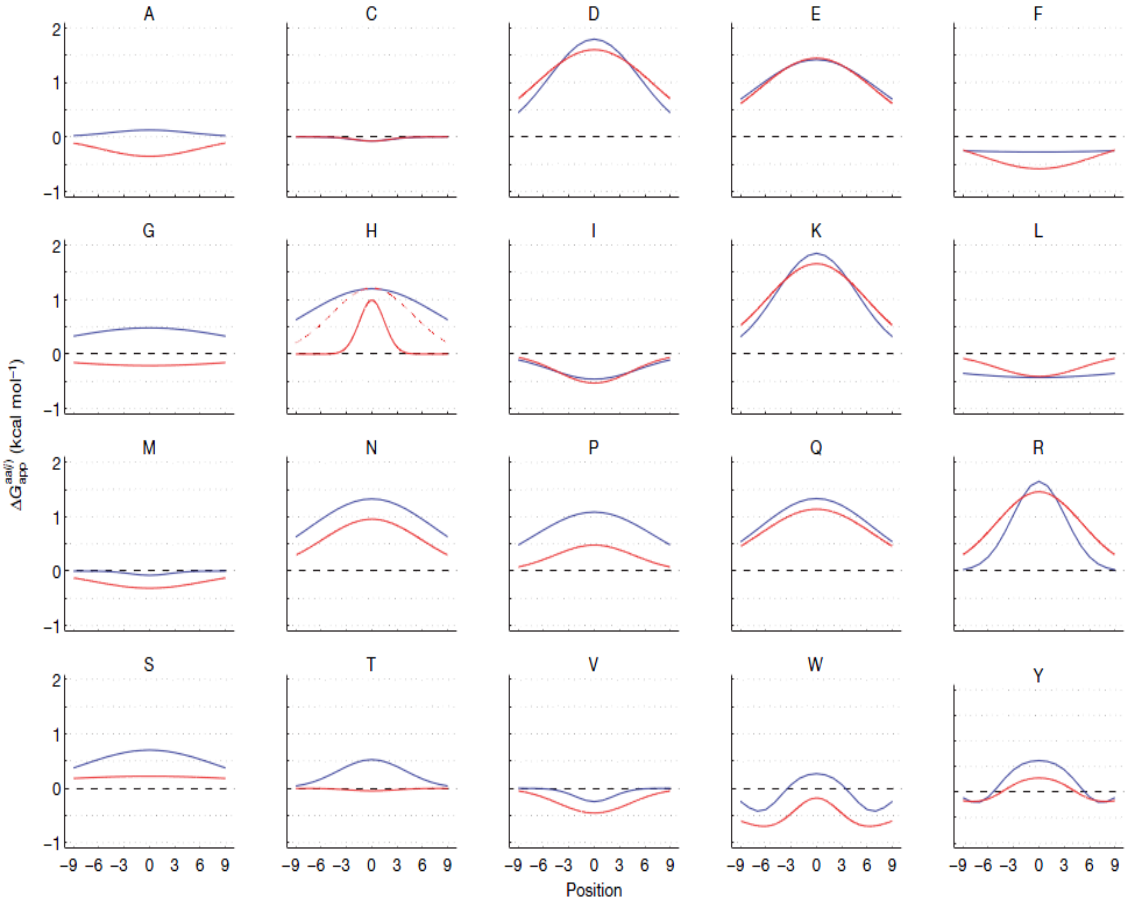


FIGURE 2.10 – Représentation des courbes d'énergie de chacun des 20 acides aminés. [Hessa *et al.*, 2007]. Les gaussiennes en bleu décrivent la matrice $\Delta G_{app}^{aa}$ qui représente la contribution de chaque acide aminé selon sa position au sein d'un segment de taille 19 acides aminés. Les gaussiennes en rouge représentent la distribution statistique des acides aminés calculés à partir de protéines avec une structure 3D connue.

### 2.4.2 Méthodes de prédiction récentes dérivées d'une meilleure connaissance des mécanismes d'insertion

Les travaux de Hessa *et al* ont permis un regain d'intérét pour les méthodes de prédiction basées sur les propriétés physico-chimiques, méthodes qui avaient été délaissées avec l'utilisation de méthodes d'apprentissage plus sophistiquées.

Une des méthodes de prédiction, connue sous le nom de $TopPred^{\Delta G}$ [Bernsel *et al.*, 2008], propose une nouvelle version de l'algorithme TopPred [Von heijne, 1992] intégrant l'énergie d'insertion $\Delta G_{app}$ pour prédire la topologie des protéines membranaires. Cette méthode glisse une fenêtre d'une taille de 21 acides aminés le long de la séquence pour tracer le profil d'énergie d'une séquence protéique, puis en utilisant deux seuils ($\Delta G_{low}$, $\Delta G_{high}$) la méthode identifie les segments TM (voir figure 2.11 (A)). Tous les segments avec une énergie inférieure au seuil $\Delta G_{low}$ sont identifiés et marqués comme des segments TM **certain**, alors que, tous les segments TM avec une énergie supérieure à $\Delta G_{low}$ mais inférieure à $\Delta G_{high}$ sont marqués comme des segments TM **putatifs**. À partir de ces 2 ensembles de TM, toutes les topologies possibles sont générées prenant tous les segments TM surs et incluant ou excluant chacun des TM putatifs. La topologie qui respecte le mieux la régle « positive-inside » fondée sur l'observation que les acides aminés chargés R et K sont beaucoup plus localisés dans le cytoplasme est retenue comme la bonne prédiction.

Les deux seuils ($\Delta G_{low}$, $\Delta G_{high}$) sont optimisés sur un benchmark de protéines membranaires dont la topologie est connue. Le benchmark est constitué de 123 protéines ayant une structure déterminée par cristallographie avec une grande qualité, et de 146 protéines ayant une structure déterminée par cristallographie avec une basse qualité. Les résultats expérimentaux montrent que l'équation $\Delta G_{app}$, qui calcule l'énergie d'insertion des segments dans la membrane, présente des performances comparables aux meilleures méthodes de prédiction de la topologie membranaire, en présentant un taux de 78% de bonne prédiction pour les séquences protéiques avec seulement un seul segment TM et 84% de taux de bonne prédiction pour les séquences protéiques avec plusieurs segments TM.

Pour comparer la méthode $TopPred^{\Delta G}$, les auteurs [Bernsel *et al.*, 2008] ont développé une deuxième méthode de prédiction de la topologie des protéines membranaires, connue sous le nom de SCAMPI (voir figure 2.11 (B)). Cette méthode est similaire à un modèle markovien dans le sens où SCAMPI définit un ensemble d'états et une matrice de probabilités de transition pour prédire la topologie.

L'architecture du modèle définie quatre types de prédictions :

- la partie intérieure de la boucle [7] ($I$) représentée par un seul état avec une probabilité d'émission constante pour tous les acides aminés.
- la partie extérieure de la boucle ($O$) représentée par un seul état avec une probabilité d'émission constante pour tous les acides aminés.
- la partie intérieure de la boucle et proche de la membrane ($i$) représentée par 24 états avec une probabilité d'émission constante pour tous les acides aminés sauf

7. La boucle représente la séquence d'acides aminés entre deux segments TM.

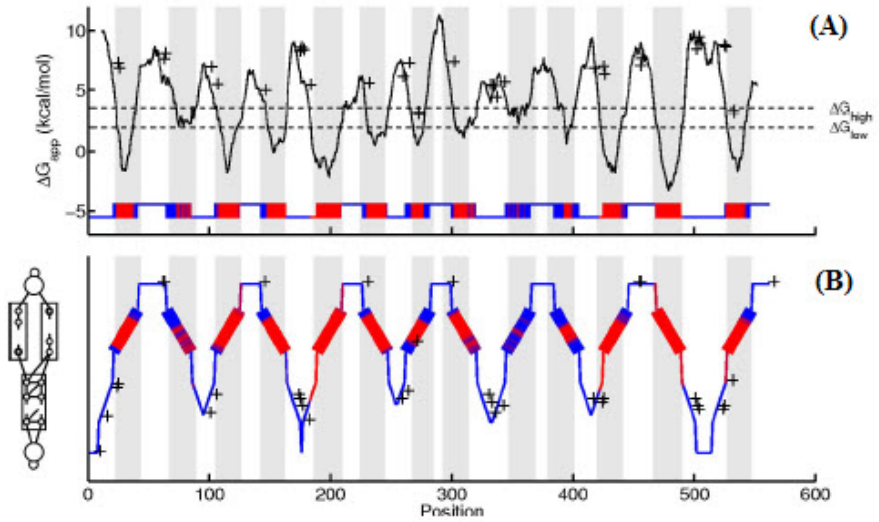


FIGURE 2.11 – Topologie d'une protéine en utilisant TOPPRED (figure (A)) et SCAMPI (figure (B) [Bernsel et al., 2008].

pour les deux acides aminés L et R où la probabilité d'émission est optimisée sur le jeu d'apprentissage. L'idée de cette prédiction est de modéliser la sur-représentation des acides aminés chargés positivement, i.e., la règle « positive-inside ».

– la partie membrane ($M$) représentée par 21 états. Dans les 20 premiers états, une probabilité d'émission constante est utilisée pour tous les acides aminés. Cependant pour le dernier état, le modèle utilise l'énergie libre $\Delta G_{app}$ pour prédire l'énergie d'insertion du segment TM de taille 21 acides aminés.

L'évaluation de SCAMPI est menée sur la même collection de protéines que TOPPRED. SCAMPI présente 79% de taux de bonne prédiction pour les séquences protéiques avec un seul segment TM et 85% pour les séquences protéiques avec plusieurs segments TM.

MINS [Park and Helms, 2008b] et MINS2 [Park and Helms, 2008a] sont des méthodes développées avec l'objectif de déterminer l'énergie d'insertion des protéines.

MINS cherche à obtenir une matrice de 19*20 valeurs correspondant aux énergies d'insertion de chacun des 20 acides aminés sur chacune des 19 positions d'un segment TM. La détermination de cette matrice se fait à partir de 13836 segments de taille 19 extraits de 73 protéines ayant une structure 3D connue. Pour augmenter l'information disponible, un alignement multiple est généré pour chacune des 73 protéines de manière à obtenir non pas une séquence d'acides aminés pour chacun des 13836 segments de longueur 19, mais un profil qui représente la fréquence de chaque acide aminé à chaque position de l'alignement.

L'hypothèse de départ de la méthode est que l'énergie d'insertion des segments TM peut être approximée par la mesure de la distance entre le milieu de la membrane et

l'acide aminé se trouvant au milieu du segment. Ainsi, il est possible d'associer une énergie d'insertion à chacun des 13836 profils obtenus grâce aux structures 3D des protéines. À partir de ces éléments, les auteurs ajustent les paramètres pour trouver une matrice MINS qui expliquera au mieux les valeurs d'énergie associées aux 13836 segments.

L'évaluation de la matrice MINS est donnée par une comparaison entre l'énergie d'insertion prédite avec les énergies d'insertion de 357 segments connues expérimentalement. Le calcul du coefficient de corrélation donne un taux de 0.74% entre les énergies prédites et celles des 357 segments.

De plus, MINS reconnait correctement la position et le nombre de segments TM pour 21 protéines bitopiques parmi 22, 33 protéines polytopiques parmi 51, et 54 protéines parmi les 73 protéines du jeu de données. La méthode reconnait aussi 299 segments TM expérimentalement observés parmi 316 segments TM, et prédit correctement 286 segments TM parmi 299 segments TM.

MINS2 suit le même principe que MINS, mais l'énergie d'insertion n'est plus évaluée de la même manière. En effet, plutôt que partir de l'hypothèse que cette énergie est liée à la distance entre un segment et la membrane, MINS2 se base sur les énergies mesurées par Hessa *et al* en 2007. Ces données concernent 357 peptides de longueur 19 acides aminés [Hessa *et al.*, 2007].

En utilisant ces données, MINS2 cherche à réduire les paramètres de la matrice MINS (19*20). La méthode suggére un profil symétrique pour les courbes des acides aminés, conduisant à un modèle linéaire de 201 paramètres. Pour réduire encore le nombre de paramètres à déterminer, 2 positions qui se succédent ont la même valeur d'énergie. Le nombre de paramètres est alors réduit conduisant à un modèle linéaire de 101 paramètres à déterminer en utilisant les 357 segments.

Les résultats présentés par MINS2 suggérent que l'utilisation de cet ensemble de données permet de capturer au mieux l'énergie d'insertion des séquences protéiques, ainsi que la prédiction des segments TM dans la protéine. Cependant, la prédiction des segments TM dans les protéines bitopiques semble performante que la prédiction des segments TM des protéines polytopiques. L'insertion des segments dans les protéines polytopiques nécessite une énergie moins importante que l'insertion des segments TM des protéines bitopiques.

## 2.5 Conclusion

Dans la première partie de ce chapitre, nous avons présenté le mécanisme par lequel les protéines membranaires transitent à travers la membrane du RE (réticulum endoplasmique). Nous avons aussi présenté différentes notions en décrivant le PS (peptide signal), le segment TM (segment transmembranaire), etc. Par la suite, nous avons introduit différentes méthodes de prédiction. Certaines méthodes s'intéressent seulement à la prédiction du peptide signal et de son site de coupure, tandis que d'autres méthodes sont développées pour la prédiction de la topologie membranaire, et enfin, certaines méthodes font de la double prédiction PS et segments TM.

Dans un deuxième temps, nous avons présenté de nouvelles hypothèses d'insertion de protéines membranaires. Nous avons vu que l'insertion des protéines est liée à la contribution de chaque acide aminé sous forme d'énergie de contribution ou d'insertion. Cependant, [Shental-Bechor *et al.*, 2006] discutent les résultats des expérimentations produits par Hessa *et al.* en soulignant que les conditions dans lesquelles elles ont été réalisées sont très spécifiques. Les courbes trouvées n'expliquent pas tous les cas d'insertion du segment TM dans la membrane. Hessa *et al.* définissent seulement deux états d'insertion, la forme insérée et non insérée du segment dans la membrane. Ils ne prennent pas en compte l'orientation des hélices alpha côté lumen ou côté cytosol. Par conséquent, les courbes ne peuvent être généralisées.

À la fin, nous avons présenté des méthodes qui tente de prédire la topologie membranaire en utilisant l'énergie de contribution des acides aminés et d'autres méthodes qui essayent de mesurer l'énergie d'insertion des segments TM en utilisant aussi l'énergie $\Delta G_{app}$. Ces différentes méthodes introduisent les connaissances biologiques d'insertion des protéines membranaires dans leur méthode de prédiction, néanmoins elles n'effectuent pas de discrimination entre le segment TM et le peptide signal.

Dans cette thèse, nous nous intéressons aux nouvelles connaissances biologiques présentées par Hessa *et al.* et qui décrivent le mécanisme d'insertion des protéines dans la membrane. Nous allons utilisé ces informations pour mimer le comportement du translocon afin de distinguer le peptide signal du segment transmembranaire. La méthode que nous construirons devra utiliser les hypothèses de Hessa *et al.* pour optimiser les courbes des acides aminés qui maximisent la classification entre une collection de PS et une collection de TM.

# Chapitre 3

# Détermination des courbes d'insertion des acides aminés dans la membrane

Il serait tout à fait intéressant de disposer d'une méthode de prédiction nous permettant de déterminer précisément si une partie de la protéine est susceptible de s'insérer dans une membrane de réticulum endoplasmique ou pas. De telles méthodes existent déja et donnent des résultats satisfaisants, mais nous pensons qu'il y a encore une marge d'amélioration possible en utilisant les dernières connaissances biologiques acquises sur les mécanismes d'insertion des segments transmembranaires dans la membrane.

Nous cherchons donc à développer une nouvelle méthode de détermination des segments transmembranaires qui s'appuie sur ces dernières avancées biologiques. Ainsi, comme dans les travaux de Hessa *et al.* présentés au chapitre 2, nous cherchons à définir le potentiel d'insertion dans une membrane de chaque acide aminé.

Après avoir décrit les hypothèses de développement de notre méthode, nous présentons l'approche retenue dans son ensemble intégrant à la fois une problématique de classification et une problématique d'optimisation. Après la description des jeux de données utilisés dans le cadre de cette thèse, la problématique de classification est détaillée. Les algorithmes d'optimisation que nous avons développés sont ensuite exposés dans les chapitres 4 et 5.

## Sommaire

## 3.1 Introduction

Dans la section 2.4 du chapitre 2, nous avons présenté les expérimentations in vitro menées par Hessa *et al.* dont le but est de simuler la machinerie du translocon. A l'issue de ces expériences, Hessa *et al.* proposent une nouvelle échelle pour mesurer le potentiel d'insertion - ce qu'ils appellent l'énergie d'insertion - des acides aminés dans une membrane. Cette nouvelle échelle est nommée échelle biologique d'hydrophobie (*biological hydrophobicity scale*).

Même si la plupart des échelles d'hydrophobie ont été jusqu'â présent obtenues expérimentalement, nous supposons que nous pouvons élaborer in silico une échelle fondée sur l'étude du phénomène d'insertion à partir de deux ensembles de peptides : un ensemble de peptides signaux et un ensemble de segments transmembranaires. Rappelons qu'un peptide signal va traverser le translocon pour rejoindre la lumière du RE, tandis que le segment TM va s'insérer dans la membrane par une porte latérale. Si notre échelle est suffisamment précise, elle est capable de distinguer un peptide signal d'un segment TM bien que ces peptides partagent des propriétés très proches. Le segment TM possède **le code** qui lui permet d'ouvrir la porte latérale du translocon, tandis que le peptide signal ne le possède pas. La comparaison de comportement des 2 ensembles de données devraient ainsi nous permettre d'élaborer l'échelle recherchée. Cette échelle pourrait bénéficier d'une plus grande quantité de données stockées dans les bases de données protéiques et pourrait par conséquent être beaucoup plus précise que les échelles déterminées expérimentalement. Dans cette échelle, chaque acide aminé est défini par un vecteur qui représente son index d'hydrophobie pour différentes positions à travers les positions de la séquence protéique.

La section 3.3 décrit les jeux de données utilisés pour l'apprentissage et l'évaluation de notre approche. Le principe de notre méthode est présenté dans la section 3.4. Cette approche est validée grâce à la capacité de discriminer entre le peptide signal et le segment transmembranaire. La section 3.4.2, présente différentes mesures d'évaluation des performances de notre méthode. La section 3.5 conclut le chapitre.

## 3.2 Approche retenue

L'hypothèse fondamentale de notre approche, qui semble confirmée par les expérimentations de Hessa *et al.* est que la capacité pour un segment de s'insérer dans une membrane est uniquement dépendante de la composition en acides aminés de ce segment. L'épaisseur d'une membrane du RE est évaluée à environ 30 Angstrom, ce qui correspond environ à une vingtaine d'acides aminés. Les travaux de Hessa *et al* semblent indiquer que la capacité d'insertion d'un segment - son énergie d'insertion - peut être évaluée à partir de la somme des énergies d'insertion de chacun des acides aminés. De plus, l'énergie d'insertion d'un acide aminé serait dépendante de sa position dans le translocon.

Pour chaque acide aminé $a$, nous cherchons donc à déterminer un vecteur $V[a]$ de taille $l$, où $V[a, j]$ correspond à l'influence de $a$ lorsque il est en position $j$ dans une fenêtre d'insertion de taille $l$. Si on estime que pour toute position $j$, il existe une certaine continuité entre les valeurs des index aux positions $j-1$, $j$ et $j+1$, plutôt que de chercher une vingtaine de valeurs discrètes, nous allons chercher à déterminer des courbes de valeurs. En fait, nous allons chercher les $l$ valeurs de $V[a]$ pour qu'elles correspondent à une courbe continue. Notons que dans ce qui suit, nous parlerons de courbes à la place de vecteurs d'acides aminés.

A partir d'une solution initiale, notre approche va donc consister à optimiser les 20 courbes correspondant aux 20 acides aminés de telle sorte que l'utilisation de ces courbes améliore la classification entre un jeu de données de peptide signaux et un jeu de données de segments TM. Pour cela, nous devons construire des jeux de données appropriés, définir un classifieur et des méthodes d'optimisation efficaces. L'ensemble de la démarche est présentée dans la figure 3.1.

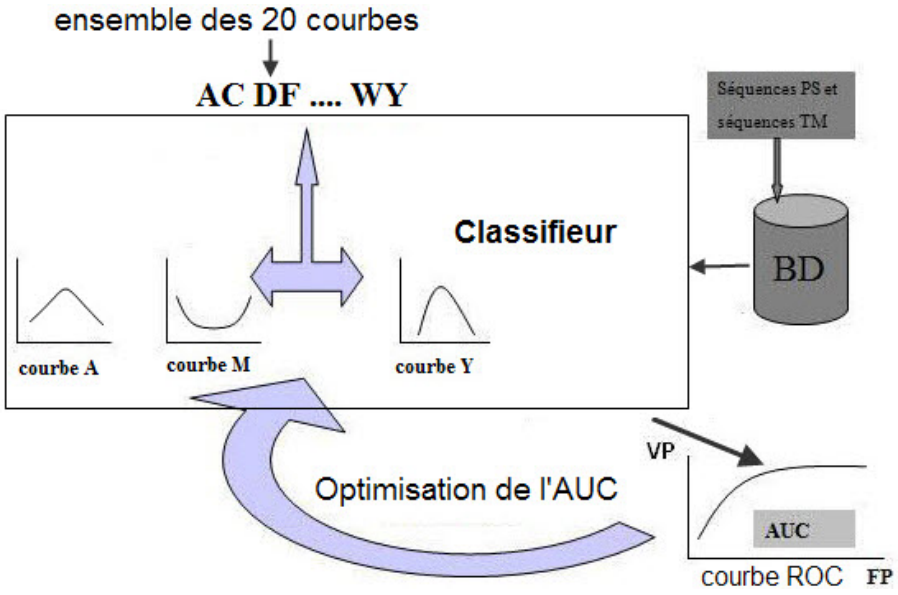


FIGURE 3.1 – Démarche retenue pour notre méthode.

## 3.3 Jeux de données

### 3.3.1 Jeu de données pour l'apprentissage des courbes : nommé SWP

L'apprentissage de notre méthode nécessite de disposer d'un jeu de données composé de séquences PS et de séquences TM. L'idéal serait bien sûr de disposer de jeux de données de protéines appropriée et prête à l'utilisation pour ce problème de localisation subcellulaire de protéines membranaires. Malheureusement, de tels jeux de données ne sont pas disponibles.

Plusieurs bases de données dédiées à la collecte des protéines membranaires sont disponibles, la plus connue de ces bases de données de stockage est la PDB [Berman *et al.*, 2000] (*Protein Data Bank*) qui contient des protéines ayant une structure 3D identifiée. La base de données PDBTM (*Protein Data Bank of Transmembrane Proteins*) [Tusnady *et al.*, 2004b; Tusnady *et al.*, 2005], construite à partir des protéines de la PDB, dénombre 1484 transmembranes protéines. Parmi ces protéines, il y a 1276 protéines structurées en hélice alpha et 207 protéines structurées en brins bêta. Notons aussi qu'il est souvent difficile de déterminer précisément la localisation les segments TM dans les séquences protéiques [Tusnady *et al.*, 2004b], ceci complique l'apprentissage ou l'évaluation de notre méthode.

Face à ce manque de données, nous avons construit un jeu de données qui permettra la mise au point de notre classifieur à partir de la base de données de protéines « Uniprot » (*Universal Protein Resource*) [Apweiler *et al.*, 2004], plus précisément la base de données Swiss-Prot [Junker *et al.*, 1999]. Nous utilisons la version de la base de données publiée en mars 2010, en suivant les étapes décrites ci-dessous :

1. Les protéines sélectionnées sont seulement celles marquées à la ligne OC (organism classification) par « eukaryota » or « eukaryotic ». Nous n'étudions que les protéines eucaryotes, car les mécanismes d'adressage que nous étudions sont différents dans les organismes procaryotes.
2. Nous extrayons à partir des protéines de la précédente étape toutes celles marquées comme « peptide signal » et « transmen » dans la ligne FT (Feature Table).
3. Pour les protéines avec un PS, nous sélectionnons seulement celles marquées à la ligne CC (subcellular localization) par « secreted » (protéines sécrétées à travers la membrane du RE).
4. Pour les protéines avec un segment TM, nous sélectionnons seulement celles marquées à la ligne CC par « membrane » ou « endoplasmic reticulum ». Le but est de de s'assurer de n'avoir que des protéines adressées vers la membrane du RE.
5. Les protéines constituant la base Uniprot sont annotées selon leur degré de certitude : « potential [1] », « probable [2] », « By similarity [3] », ou pas d'annotation dans le cas d'une protéine avec une structure connue.

   Nous supprimons toutes les protéines PS et annotées avec les mots clé « potential, probable », or « By similarity ». Cependant, pour les protéines avec un segment TM nous supprimons seulement les protéines annotées probable, ou by similarity. En effet, nous retenons les protéines annotées « potential » dans lesquelles la position des segments TM est connue avec une meilleure certitude par rapport aux protéines annotées « probable », ou « by similarity ». Ceci est dans l'objectif d'avoir un jeu de données plus grand.
6. Nous appliquons sur le jeu de données obtenu une réduction d'homologie en utilisant le programme CD-HIT [Li and Godzik, 2006]. L'objectif est de supprimer toutes les

1. Potential signifie que l'information est obtenue par un programme de prédiction.
2. Probable indique que l'information peut être trouvée dans l'environnement naturel.
3. By similarity signifie que l'information est obtenue par expérimentation sur une protéine et a été transférée à d'autres familles de protéines.

séquences redondantes à 50% d'identité de manière à êter la redondance forte et à conserver un échantillon de taille suffisante.

Il est reconnu que le peptide signal est localisé en partie N-terminale, et sa longueur varie selon les protéines. Dans le cas des protéines eucaryotes la moyenne de la longueur du peptide signal est de 22 à 32 acides aminés [Bendtsen *et al.*, 2004]. En conséquence, la séquence appelée PS qui représente une protéine avec un PS sera formée des 35 premiers acides aminés de la protéine.

Pour les séquences avec des segments TM, nous extrayons seulement le premier segment TM selon son annotation dans la base de données. En effet, l'insertion dans la membrane des protéines bitopiques (un seul segment TM) est différente de l'insertion des protéines polytopiques (plusieurs segments TM) [Park and Helms, 2008a]. Notons aussi que dans notre travail, nous considérons le signal anchor comme un segment TM.

Tous les segments TM avec une taille inférieure à 17 acides aminés sont supprimés. Par contre, pour les segments TM dont la taille est supérieure à 17 et inférieure à 19, nous élargissons la fenêtre d'extraction avant et après la position annotée pour obtenir un segment TM de taille supérieure à 19 acides aminés. Nous avons supprimé les segments TM avec une taille inférieure à 17 car ces segments TM sont considérés comme des segments de courte taille et il est très difficile de les déterminer. De plus, la taille de la membrane est d'environ 30 Angström ce qui correspond à une vingtaine d'acides aminés.

En résumé, le jeu de données final appelé SWP contient 1046 séquences de longueur 35 acides aminés représentant un peptide signal et 684 séquences de longueur entre 19 et 25 acides aminés représentant un segment TM.

**Représentation des segments TM :** Le jeu de données SWP représente une toute petite partie de la base protéique Uniprot. L'annotation des segments TM dans la base de données UniProt est fondée sur des informations biochimiques complétées par des informations données par des articles [Junker *et al.*, 1999]. Dans notre jeu de données, nous remarquons que plusieurs segments TM sont annotés par le qualificatif « potential ». Cette annotation est le résultat de l'application de logiciels de prédiction TMHMM [Sonnhammer *et al.*, 1998b], Memsat [Jones *et al.*, 1994], et Phobius [Kall *et al.*, 2004]. En conséquence, l'annotation de la position des segments TM dans la séquence protéique biaisée par l'utilisation de ces logiciels. Nous décidons donc d'élargir la séquence représentant un segment TM dans le jeu SWP par l'ajout d'acides aminés avant et après la position annotée dans la base de données. Nous avons testé plusieurs tailles pour cette représentation et nous avons choisi d'élargir les segments TM par l'ajout de 10 acides aminés de chaque côté du segment TM. Au final, un segment TM sera représenté par une séquence d'acides aminés de taille variant entre 39 et 45 acides aminés, 39 pour les segments TM de taille 19 et 45 pour les segments TM de taille 25.

**Distribution des acides aminés dans le jeu de données SWP :** Il est très intéressant d'observer la distribution statistique des acides aminés dans le jeu de données qui servira à l'optimisation des courbes, le jeu de données SWP.

La figure 3.2 montre la distribution statistique des acides aminés dans le jeu SWP.

Nous observons que certains acides aminés comme l'histidine(H), la tryptophane (W), et l'asparagine (N) sont moins présents dans le jeu de données SWP, alors que d'autres acides aminés comme la leucine (L), l'alanine (A), la valine (V), et l'isoleucine (I), phenylalanine (F), et la serine (S) sont très fréquents dans le jeu de données SWP. Notons que la plupart des acides aminés fréquents ont une grande valeur d'hydrophobie.

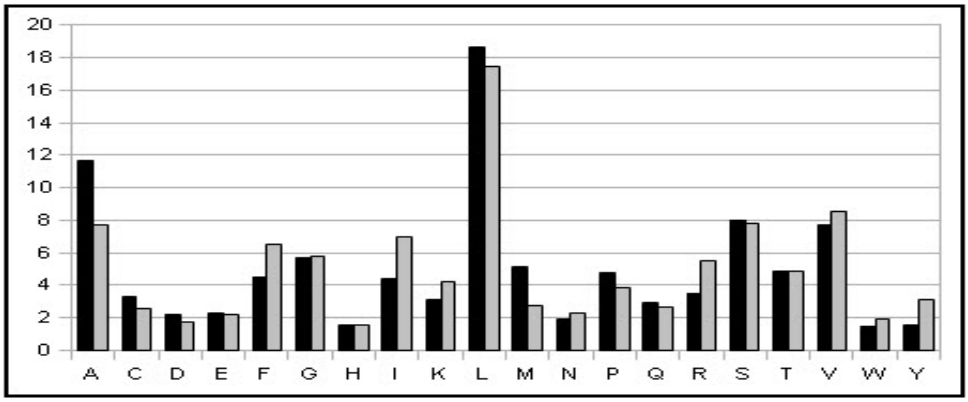


FIGURE 3.2 – Distribution statistique de chaque acide aminé dans le jeu de données SWP. Les segments TM sont **élargis avec 10 acides aminés avant et aprés la position annotée** dans la base de données. Les barres noires indiquent la distribution des séquences avec un peptide signal alors que les barres grises indiquent la distribution des séquences avec un segment TM (ou SA). L'axe des abscisses présente les 20 acides aminés. L'axe des ordonnées montre la fréquence d'apparition de chaque acide aminé, laquelle est calculée en divisant le nombre de chaque acide aminé dans le jeu SWP par le nombre total des acides aminés constituant le jeu SWP * 100.

Nous observons aussi la sur-représentation des deux acides aminés arginine (R), et lysine (K) dans la figure 3.2, puisque si l'on regarde l'échelle d'hydrophobicité de Kyte et Doolittle (2.1), l'arginine est l'acide aminé ayant la plus faible valeur d'hydrophobie de tous les acides aminés avec une valeur R = -4,5, tandis que la lysine a une valeur d'hydrophobie très faible (K= -3,5). En plus, il y a un net déséquilibre dans la distribution de ces acides aminés en faveur des segments TM.

Afin de vérifier cette observation, nous réalisons la distribution des acides aminés dans le jeu de données SWP avec des segments TM sans élargissement des 10 acides aminés sur chaque côté du segment (voir figure 3.3). Les segments TM sont représentés par une chaîne d'acides aminés correspondant à l'annotation dans la base de données Uniprot. Nous observons que la distribution des acides aminés Arginine et Lysine est moins présente dans la figure 3.3 comparée avec la distribution de ces mêmes acides aminés dans la figure 3.2. De plus, l'arginine et la lysine sont plus présentes dans le jeu peptide signal que dans le jeu segments TM.

En conclusion, il semble que les segments TM soient entourés par des acides aminés moins favorables à l'insertion des segments TM dans la membrane.

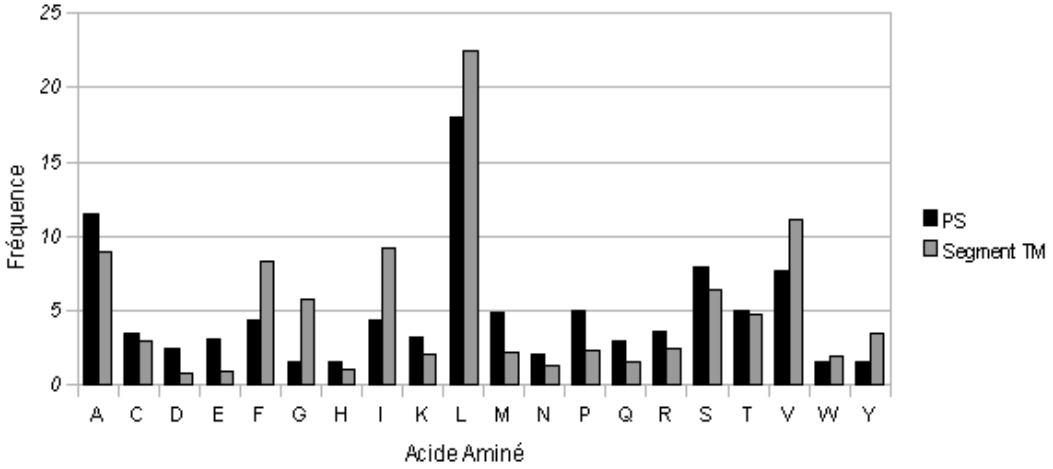


FIGURE 3.3 – Distribution de chaque acide aminé dans le jeu SWP sans élargissement des segments TM.

### 3.3.2 Jeu de test pour le classifieur : nommé SCAMPI

Le jeu SWP décrit ci-dessus est un jeu qui sera utilisé pour l'apprentissage de notre méthode. Dans cette section ainsi que la section 3.3.3, nous décrivons deux autres jeux de données qui seront utilisés uniquement pour le test de notre méthode.

Le premier jeu a été utilisé par le logiciel SCAMPI. Ce jeu est constitué seulement de séquences avec des segments TM mais il est trés intéressant, car la position des segments TM est connue avec une meilleure certitude que dans le cas des protéines d'Uniprot.

Les données de SCAMPI sont divisées en deux collections de protéines dont on connaét la structure 3D. La premiére collection est une collection de 123 protéines transmembranaires de « haute résolution [4] » avec une réduction d'homologie de 40%, tandis que la seconde collection est constituée de 146 protéines transmembranaires de « basse résolution [5] » avec une réduction d'homologie de 40%.

Nous appelons les segments TM qui sont obtenus é partir de la collection SCAMPI de haute résolution « ScampiHigh », tandis que les segments TM qui sont obtenus à partir de la collection SCAMPI de basse résolution sont appelés « ScampiLow ».

À partir de la collection ScampiHigh, nous construisons 2 nouvelles collections. La première collection nommée « ScampiHighFirst » contient seulement le premier segment TM de chaque protéine de la collection ScampiHigh selon son annotation dans la base de données. Par contre, la deuxième collection nommée « ScampiHighALL » contient tous les segments TM de la collection ScampiHigh selon leur annotation dans la base de données. Dans ces deux collections, nous supprimons tous les segments TM de taille inférieure à 19. Finalement, ScampiHighFirst contient 100 segments TM et ScampiHighALL contient 419 segments TM. Nous construisons aussi deux collections à partir de ScampiLow. Scam-

4. protéines ayant une structure cristallographiée avec une grande qualité
5. protéines ayant une structure cristallographiée avec une basse qualité

piLowFirst contient 122 segments TM, alors que ScampiLowALL contient 712 segments TM.

L'objectif de la construction d'une collection avec seulement le premier segment TM (cas « first ») et une collection avec tous les segments TM (cas « All ») est d'évaluer le comportement d'insertion de ces différents segments TM. En effet, des travaux [Park and Helms, 2008a] ont constaté que l'insertion des protéines bitoypiques (un seul segment TM) est différente de l'insertion des protéines polytopiques (plusieurs segments TM).

Notons aussi que ces deux collections fournissent des données protéiques différentes des données protéiques de SWP. La plupart des données de SWP sont constituées de protéines avec un seul segment TM, tandis que les données de SCAMPI contiennent plusieurs segments TM. De plus, les protéines constituant les données SWP sont des protéines d'organisme eucaryotes adressées vers la membrane du réticulum endoplasmique. Alors que les protéines de SCAMPI sont des protéines qui peuvent être eucaryotes ou procaryotes qui sont adressées à la membrane plasmique.

### 3.3.3 Jeu de test pour le classifieur : nommé PDB

Nous présentons un autre jeu de données que nous nommons PDB qui sera utilisé pour évaluer les performances de notre méthode. Dans ce jeu la structure 3D est connue avec une meilleure certitude, cependant il ne contient pas de séquences peptide signal.

Nous utilisons la base de données PDBTM [6] [Tusnady *et al.*, 2004b]. Cette base de données est construite à partir des protéines de la Protein Data Bank [Berman *et al.*, 2000] en utilisant l'algorithme TMDET [Tusnady *et al.*, 2004a]. L'objectif est de collecter toutes les protéines transmembranaires déposées dans la PDB et déterminer leurs régions d'insertion (segments TM) en utilisant seulement les informations structurelles des protéines. En d'autres termes, TMDET utilise la structure 3D des protéines pour calculer la position des segments TM. Ce jeu de données va nous permettre d'intégrer les protéines dont la structure a été récemment déposée dans la PDB, permettant ainsi d'intégrer de nouvelles protéines par rapport aux jeux SCAMPI

La construction du jeu PDB est obtenue en utilisant la liste des protéines non redondante à hélices alpha donnés par la PDBTM. Nous appliquons une réduction d'homologie de 50% sur ces protéines pour obtenir une collection de 229 protéines. Nous supprimons toutes les protéines ayant un premier segment de taille inférieure à 17. Le jeu PDB ainsi obtenu contient 180 séquences protéiques. À partir de ce jeu, nous construisons 2 collections. La premiére collection nommée « PDBFirst » contient seulement le premier segment TM du jeu PDB selon son annotation dans la base de données, la collection contient 180 segments TM. Par contre, la deuxième collection nommée « PDBAll » contient tous les segments TM du jeu PDB selon leur annotation dans la base de données, et contient donc 569 segments TM.

6. http ://pdbtm.enzim.hu/

### 3.3.4 Résumé de nos jeux de données

Nous avons construit un jeu SWP contenant 1046 PS et 684 TM. Nous disposons par ailleurs de 2 jeux ne contenant que des TM, SCAMPI et PDB. Afin d'utiliser ces deux jeux en phase de test, nous les complétons avec des séquences PS tirées du jeu SWP.

Nous divisons le jeu de données SWP en deux parties. La première partie constituée de 684 séquences PS sera utilisée pour l'apprentissage. Le reste des 362 séquences PS qui n'ont pas servi à l'apprentissage sont ajoutées aux jeux SCAMPI et PDB.

Nous présentons dans le tableau 3.1 un résumé de la taille des différents jeux de données.

| | SWP | ScampiHigh First | ScampiHigh ALL | ScampiLow First | ScampiLow All | PDBFirst | PDBAll |
|---|---|---|---|---|---|---|---|
| PS | 684 | 362 | 362 | 362 | 362 | 362 | 362 |
| TM | 684 | 100 | 419 | 122 | 712 | 180 | 569 |

TABLE 3.1 – Résumé de la taille des jeux de données.

## 3.4 Un classifieur pour la discrimination entre le PS et le segment TM

### 3.4.1 Système de classification

Après avoir présenté les jeux de données qui serviront à l'apprentissage des courbes d'insertion des acides aminés, nous présentons dans cette section notre classifieur. Celui-ci permettra de déterminer si un peptide est un peptide signal ou un segment transmembranaire en utilisant les courbes d'insertion de chacun des acides aminés.

La figure 3.4 présente le principe de notre classifieur qui utilise la fonction **f** laquelle discrimine entre deux classes, la classe de séquences avec un PS et la classe de séquences avec un segment TM.

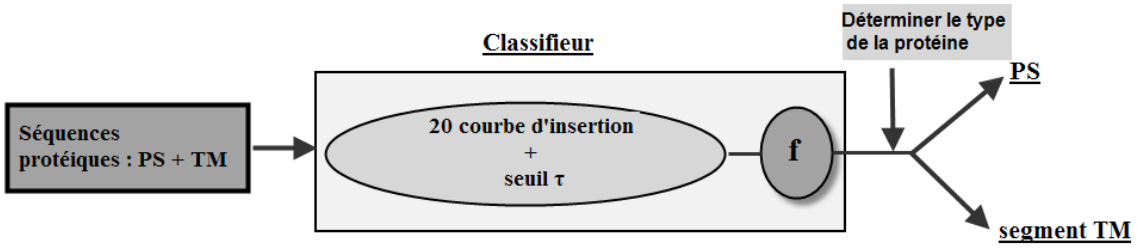


FIGURE 3.4 – Schéma illustrant le classifieur du PS et du segment TM.

Les données d'entrées de notre classifieur sont des séquences protéiques, issues de protéines solubles ou des protéines membranaires. Dans le cas d'une protéine soluble,

cette séquence correspond au peptide signal et dans le cas d'une protéine membranaire la séquence correspond au premier segment transmembranaire.

La sortie du classifieur consiste à déterminer le type de la séquence ce qui correspond donc à un problème de classification binaire entre deux classes que nous notons **PS** pour la classe avec des peptides signaux et **TM** pour la classe avec des segments transmembranaires.

Nous définissons par $a$ un des 20 acides aminés et nous lui associons un vecteur $V$ de valeurs. Le vecteur $V[a]$ définit la contribution de l'acide aminé $a$ selon sa position lors du processus d'insertion des protéines dans la membrane. La longueur de ce vecteur est définie sur un intervalle de taille $l$ représentant les positions pertinentes pour l'insertion d'un segment TM. Ainsi, chaque vecteur d'un acide aminé a la même taille $l$.

Étant donné $V[a^i]$ le vecteur de l'acide aminé $a^i$, ces valeurs sont notées par $V[a^i, j] = V[a^i](j)$ où $j$ représente une position, $j \in [1, l]$. Pour une séquence $Seq$ d'acides aminés de taille $l$, nous utilisons la notation $Seq = < a_1 a_2 \dots a_l >$. Nous définissons la **moyenne d'insertion** de la séquence $Seq$ par la moyenne d'index d'insertion des acides aminés pour $j$ variant de 1 à $l$ :

$$E(Seq) = \frac{\sum_{j=1}^{l} V[a^j, j]}{l} \tag{3.1}$$

Dans le cas d'une séquence plus longue $Seq = < a_1 a_2 \dots a_n >$ de taille $n > l$, nous glissons une fenêtre de taille fixe $l$ sur toute la séquence. Nous calculons la valeur de $E$ sur cette fenêtre et finalement nous retenons la valeur maximale de $E$ trouvée.

La **fonction d'insertion** de la séquence $Seq$ est définie comme étant le maximum de la moyenne d'insertion calculée sur une sous-séquence de taille $l$ définie par :

$$E_{max}(Seq) = \max_{1 \leq k \leq n-l+1} \{E(Seq_k)\} \tag{3.2}$$

oé $Seq_k = < a_k a_{k+1} \dots a_{k+l-1} >$.

**Classifieur PS/TM :** La distinction entre la classe PS et la classe TM est donnée par la fonction d'insertion $E_{max}(Seq)$ et un seuil $\tau$. La fonction $E_{max}(Seq)$ correspond à la valeur maximale d'hydrophobie de la séquence.

Un segment TM est généralement très hydrophobe [Kyte and Doolittle, 1982]. Par conséquent, notre règle de classification est la suivante :

$$\begin{cases} Seq \in classe\ PS & \text{si} \quad E_{max}(Seq) < \tau \\ Seq \in classe\ TM & \text{sinon} \end{cases} \tag{3.3}$$

La figure 3.5 montre un schéma illustrant le principe de notre classifieur. À partir d'une sous-séquence de la séquence complète, on glisse une fenêtre de taille $l$ et on calcule la fonction d'insertion. L'ensemble des vecteurs $(V[a_i])_{i=A}^{i=Y}$ et le seuil $\tau$ déterminent notre

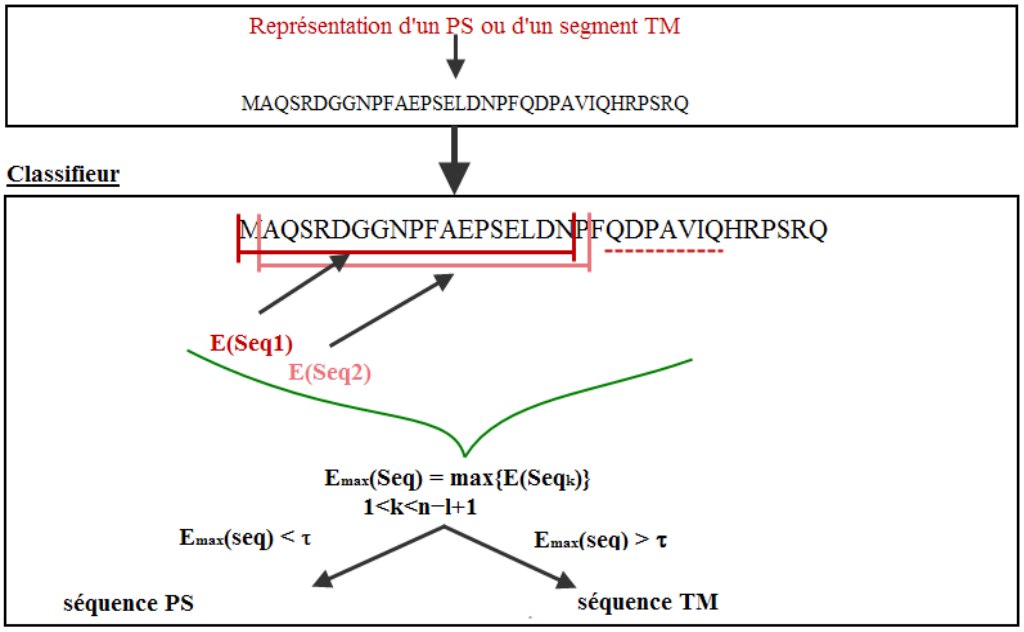


FIGURE 3.5 – Prédiction d'une séquence protéique.

système de discrimination entre une séquence PS et une séquence segment TM. Nous utiliserons différents jeux pour la construction et l'évaluation de notre modèle de classification. Le jeu de données SWP est le jeu d'apprentissage de notre méthode : il va être utilisé pour la détermination des courbes d'insertion des acides aminés et la détermination du seuil de discrimination $\tau$.

Notre objectif est de déterminer les valeurs des paramètres (vecteurs et seuil) permettant d'optimiser la discrimination entre une séquence PS et une séquence segment TM. Le classifieur que nous proposons ne correspond pas à un classifieur classique, notamment car l'énergie de la séquence doit être déterminée en glissant une fenêtre le long de la séquence.

### 3.4.2 Mesure d'évaluation des performances de classification

La construction d'un classifieur par un processus d'apprentissage nécessite souvent la construction de trois différents ensembles de données. Le premier ensemble de données est connu sous le nom de l'ensemble d'apprentissage. Celui-ci servira pour apprendre le classifieur, c'est-à-dire, les règles de décisions. A côté de l'ensemble d'apprentissage, on est souvent amené à construire un deuxiéme ensemble indépendant du premier, l'ensemble de test qui servira pour l'estimation de l'erreur. Dans certains cas et pour certaines méthodes, nous construisons un troisiéme ensemble appelé ensemble de validation qui servira au réglage des paramètres comme le critère d'arrét. L'objectif de cet ensemble de validation est d'éviter que le réglage des paramètres soit dépendant du jeu d'apprentissage.

Il est donc préférable de diviser l'ensemble de données initial en trois sous-ensembles disjoints (voir figure 3.6) : un ensemble d'apprentissage, un ensemble de test, et un en-

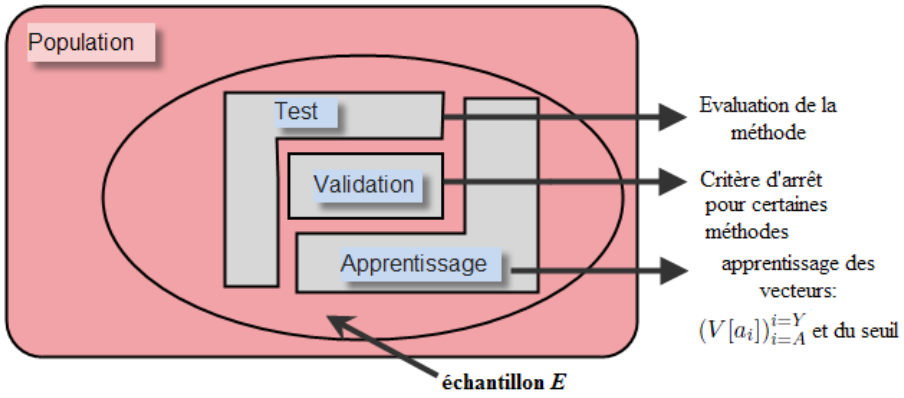


FIGURE 3.6 – Partitionnement d'une population en trois collections : ensemble d'apprentissage, ensemble de test, et ensemble de validation.

semble de validation.

Le partitionnement des données initiales n'est possible que si le nombre d'observations dont on dispose est suffisamment grand. Dans le cas contraire, on utilise le principe de la méthode de validation croisée [Stone, 1974] (*cross-validation*) qui consiste à diviser l'ensemble des données en $N$ sous-ensembles (N-fold cross validation) de taille égale. Le système apprend sur $N-1$ sous-ensembles et estime l'erreur sur le sous-ensemble n'ayant pas participé dans l'apprentissage (voir figure 3.7). On répète $N$ fois le même processus en faisant varier le sous-ensemble qui joue le rôle de jeu de test. L'erreur retournée est la moyenne des erreurs obtenues lors de ces N expériences. Très souvent on fixe $N$ à 10. Ce système permet d'obtenir une estimation non biaisée de l'erreur.

Un cas particulier de la validation croisée est la méthode du leave-one-out. Elle est principalement utilisée quand on dispose de peu de données d'apprentissage. L'idée est de pousserà l'extrème la validation croisée en laissant à chaque fois un seul exemple de données pour le test et utiliser tout le reste des données pour l'apprentissage. Cette méthode est très peu utilisée, car elle est coûteuse en temps de calcul.

Afin de mesurer les performances d'un classifieur binaire, plusieurs techniques sont utilisées.

**Critéres de performances :** Pour une téche de classification, l'indice de qualité est souvent défini par le taux de classification correcte. Dans notre cas, la tâche consiste à discriminer entre une classe de séquences avec un peptide signal et une classe de séquences avec un segment TM.

Dans le problème de classification binaire, on distingue une classe appelée la classe positive l'autre étant appelée négative. Dans notre problématique, la classe positive sera la classe PS et la classe négative sera la classe TM.

Le regroupement des quatre situations de prédiction du classifieur est donné dans une matrice, dite matrice de confusion (tableau 3.8). Celle-ci permet le calcul d'indices ca-

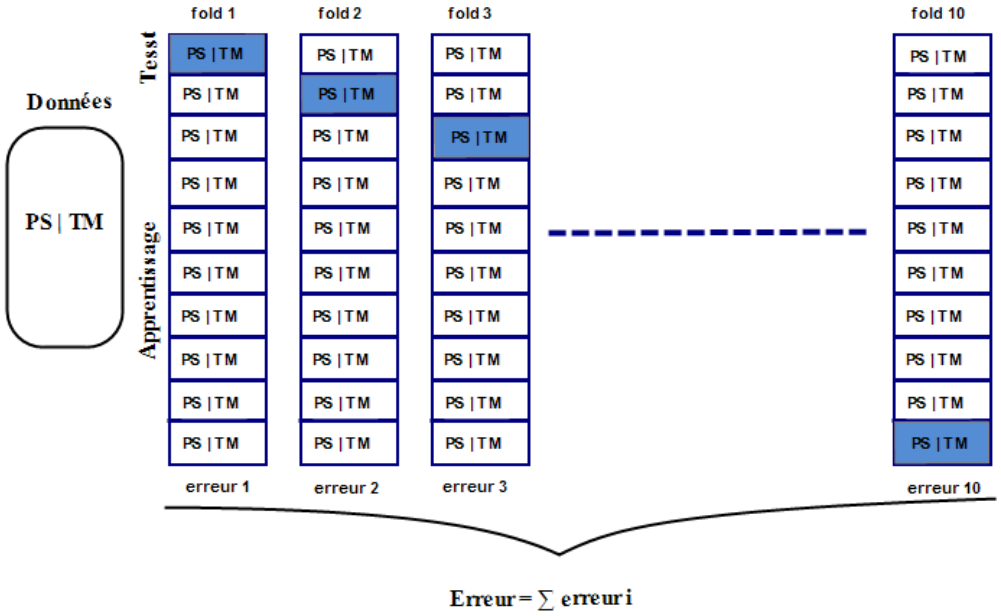


FIGURE 3.7 – Illustration du système de validation croisée. Le jeu de données est divisé en 10 sous ensembles de taille égale. Chaque sous-ensemble est constitué d'un ensemble de séquences PS et de séquences TM. La méthode apprend sur 9 sous-ensembles et estime l'erreur sur le jeu laissé pour le test. Ce processus est répété 10 fois en modifiant à chaque fois le jeu de test, puis l'erreur retournée est la moyenne des 10 erreurs obtenues.

| | *Réel positif* | *Réel négatif* |
|---|---|---|
| *Prédit positf* | **Vrai Positif (VP)** | **Faux Positif (FP)** |
| *Prédit négatif* | **Faux Négatif (FN)** | **Vrai Négatif (VN)** |

FIGURE 3.8 – Matrice de confusion.

ractérisant la classification, les plus connus sont la sensibilité et la spécificité. La sensibilité (équation 3.4) sera dans notre cas définie comme le pourcentage de séquences prédites PS parmi celles ayant réellement un PS, tandis que la spécificité (équation 3.5) sera définie comme le pourcentage de séquences prédites TM parmi celles ayant réellement un segment TM.

$$sensibilité = \frac{VP}{VP + FN} \tag{3.4}$$

$$spécificité = \frac{VN}{VN + FP} \tag{3.5}$$

On utilise aussi l'exactitude (*accuracy*) qui définit le taux de bonne prédiction, dit aussi le taux de classification (équation 3.6).

$$taux\ de\ bonne\ prédiction = \frac{VN + VP}{VN + FP + VP + FN} \tag{3.6}$$

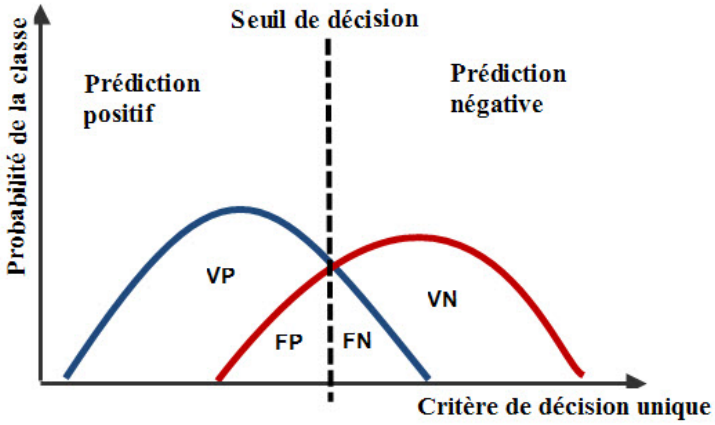


FIGURE 3.9 – Les quatre états possibles lors de la prédiction d'un modèle de classification binaire. Le seuil de décision choisi détermine les états correspondants aux VP, FN, FP, et VN.

Cependant, l'utilisation seule de ces indices présente des limites, puisque, souvent on est amené à réaliser un compromis entre spécificité et sensibilité. De plus, le taux de bonne prédiction n'offre pas à lui seul une réponse, car un même taux de bonne prédiction peut correspondre à des compromis sensibilité/spécificité bien différents.

Dans notre probléme à deux classes, nous avons à déterminer un seuil de décision qui permettra la distinction entre la classe positive et la classe négative (figure 3.9). Dans le cas où le seuil de décision est faible, la valeur de la sensibilité va être élevée et celle de la spécificité sera faible et inversement. Il est clair que les résultats en terme de sensibilité, spécificité et taux de prédiction dépendent totalement du seuil de décision pour une distribution donnée.

Il existe d'autres indices qui peuvent remplacer ou être associés aux mesures présentées ci-dessus, tel que l'aire sous la courbe ROC que nous présentons dans la section qui suit.

**Courbes de ROC :** L'approche *Receiver Operating Characteristic* (ROC) a été développée pour la théorie de détection de signaux [Egan, 1975] en relation avec les signaux radars. Elle est très utilisée en apprentissage automatique et en fouille de données comme méthode d'évaluation [Yan *et al.*, 2003; Fawcett, 2004].

L'idée principale de la courbe ROC est d'évaluer la performance d'un classifieur indépendamment du choix du seuil de décision. La courbe ROC représente les valeurs de la sensibilité en fonction de (1-spécificité), en faisant varier les valeurs du seuil de décision et en joignant ces points par une courbe. Par conséquent, la courbe ROC résume tous les compromis sensibilité/spécificité pour toutes les valeurs du seuil.

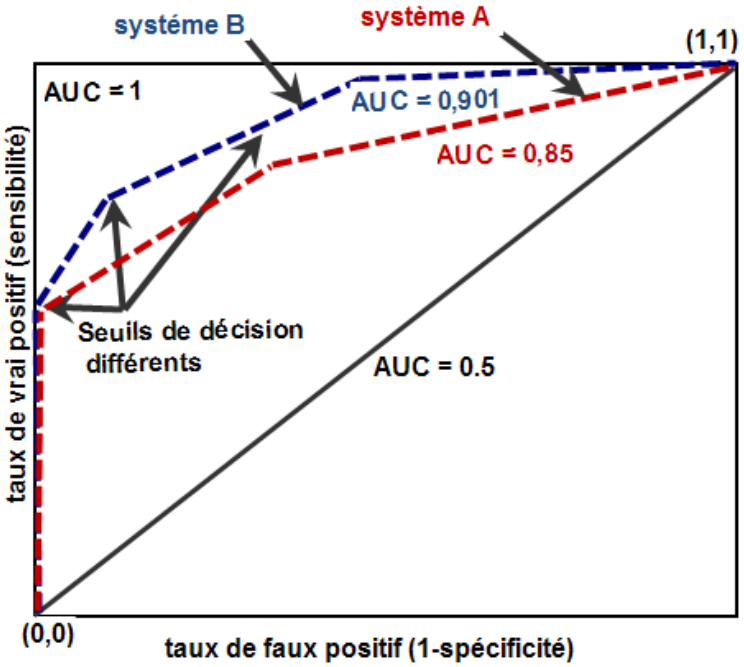


FIGURE 3.10 – Exemple de courbes ROC.

Il est possible de calculer un indice permettant d'évaluer numériquement la courbe ROC. Cet indice définit l'aire sous la courbe ROC appelée *Area Under the Roc Curve* (AUC) [Corinna and Mohri, 2004]. L'AUC est la capacité du classifieur à ordonner les instances de la classe positive avant les instances de la classe négative. La figure 3.10 montre les cas suivants :

- une valeur de l'AUC égale à 0,5 signifie que le classifieur est non discriminant.
- une valeur de l'AUC égale à 1 correspond à un système parfaitement discriminant. Dans ce cas, il existe un seuil qui sépare totalement les deux classes.

Dans la pratique, la valeur de l'AUC est entre [0.5, 1]. Par exemple, les courbes cor-

respondantes au système A et au système B de la figure 3.10. Selon l'aire sous la courbe ROC, on est capable de déterminer numériquement que le système B a globalement de meilleures performances de discrimination que le système A.

Dans notre étude, la méthodologie ROC permettra d'évaluer la capacité du classifieur à discriminer les PS des TM et indirectement elle permet d'évaluer la capacité des courbes d'insertion à modéliser le phénomène d'adressage des protéines membranaires.

Un processus de validation croisée sera appliqué sur le jeu de données SWP pour évaluer la qualité du classifieur obtenu. La détermination du seuil de discrimination $\tau$ est réalisée en faisant varier la valeur du seuil tout en calculant la valeur du taux de bonne prédiction (section 3.4.2) entre la classe PS et la classe TM du jeu de données SWP. La valeur du seuil qui donne le meilleur score du taux de prédiction est choisi pour être le seuil de discrimination $\tau$ entre le PS et le TM. Les deux jeux de données SCAMPI et PDB seront utilisés comme des jeux de test indépendants pour l'évaluation du classifieur appris sur le jeu SWP.

**Validation croisée :** Nous présentons dans la figure 3.11 le schéma général d'apprentissage de notre méthode sur le jeu de données SWP.

Nous appliquons une validation croisée de 10 expérimentations. Pour chaque expérimentation, nous construisons deux types d'ensemble de données : un ensemble d'apprentissage et un ensemble de test. Nous apprenons les courbes et le seuil de discrimination $\tau$ sur 9 ensembles et nous testons sur l'ensemble n'ayant pas servi à l'apprentissage. Nous répétons 10 fois ce processus puis nous estimons l'erreur moyenne.

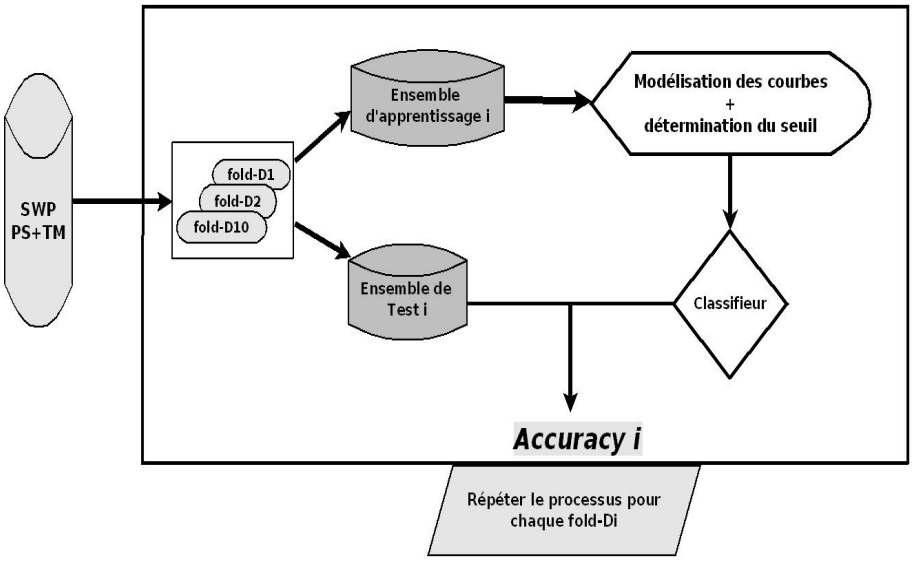


FIGURE 3.11 – Schéma général du processus d'apprentissage.

**Test sur le jeu SCAMPI et PDB :** La figure 3.12 illustre le schéma général du processus de test sur les jeux de données SCAMPI et PDB. Nous utilisons les 20 courbes

ainsi que le seuil qui sont appris sur le jeu d'apprentissage pour discriminer les séquences PS et les séquences TM dans les différents jeux de test.

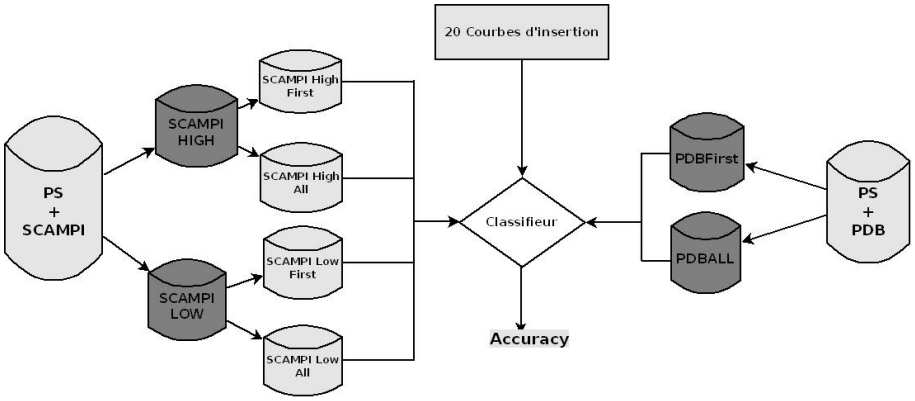


FIGURE 3.12 – Schéma général du processus de test sur les jeux SCAMPI et PDB.

## 3.5 Conclusion

La première partie du chapitre a été consacrée à la présentation des jeux de données qui seront utilisés d'un côté pour la construction du classifieur et d'un autre côté pour l'estimation de ses performances.

Nous avons ensuite introduit le principe de notre classifieur qui prend en compte la position des acides aminés lors de l'adressage des protéines par le translocon afin de discriminer les séquences PS des séquences TM. La méthode définit aussi pour chaque acide aminé une courbe qui représente sa contribution dans le processus d'insertion des segments dans la membrane. La qualité des courbes est estimée selon la capacité de discrimination entre les séquences PS et les séquences TM.

Nous avons enfin présenté différentes mesures d'évaluation. L'utilisation de ces dernières est primordiale, car elle nous permettra de mesurer la qualité du classifieur et la pertinence des courbes d'insertion que nous proposerons.

# Chapitre 4

# Développement d'une approche par recherche locale

Dans ce chapitre nous allons présenter trois algorithmes fondés sur une approche par recherche locale afin d'optimiser les courbes d'insertion ainsi que le seuil qui déterminent le classifieur que nous proposons pour discriminer les séquences PS des séquences TM. Ces algorithmes sont validés grâce à des expériences sur les jeux de données présentées dans le chapitre précédent. Une partie des travaux de ce chapitre ont été publiés dans [Laroum *et al.*, 2010], et [Laroum *et al.*, 2011].

**Sommaire**

## 4.1 Recherche locale

### 4.1.1 Fonctionnement général des algorithmes de recherche locale

Un problème d'optimisation est caractérisé par le couple $(\zeta, f)$ tel que $\zeta$ est l'ensemble des configurations candidates et $f$ est une fonction associant un coût à chaque configuration, l'objectif étant de trouver une configuration $s^* \in \zeta$ telle que $f(s^*)$ soit maximale.

Notre problème de détermination des courbes des acides aminés est considéré comme un problème d'optimisation dans lequel nous cherchons à optimiser un ensemble de courbes. Chaque ensemble de courbes est évalué en utilisant une fonction qui estime la qualité des courbes à discriminer les séquences PS des séquences TM. Dans notre cas, nous cherchons à maximiser cette fonction d'évaluation en utilisant un algorithme de recherche locale.

La recherche locale (RL) est une famille de métaheuristiques fondée sur la notion de voisinage [Aarts and Lenstra, 1997; Hoos and Stutzle, 2004]. Le principe de base est d'explorer l'espace de recherche en se déplaçant d'un voisin à un autre jusqu'â ce qu'un voisin de meilleur qualité soit trouvée. À chaque itération de la recherche locale, seuls les voisins les plus proches de la solution courante sont étudiés.

Un algorithme de recherche locale est défini par trois éléments :

1. Un voisinage $N$ qui génére des configurations voisines.
2. Une fonction d'évaluation $f$ qui estime la qualité de chaque configuration.
3. Une stratégie de mouvement qui permet de passer d'une configuration courante à une configuration voisine.

Formellement, un algorithme de recherche locale débute avec une configuration $s$ issue de l'espace de recherche $\zeta$ et par transformations successives construit des suites de solutions. À chaque transformation ou itération, des voisins de la configuration $s$ sont générés dans le voisinage $N$ à l'aide d'un opérateur de transformation (appelé opérateur de mouvement), et évalués par $f$. Une stratégie consiste à évaluer chacun des voisins de $s$, $\forall s' \in N$, dans le cas $f(s') \geq f(s)$, la configuration courante $s$ est remplacée par la configuration $s'$, autrement dans le cas où $f(s') < f(s)$ la configuration $s'$ est rejetée.

L'algorithme de recherche locale s'arrête lorsqu'on ne peut plus améliorer la solution courante ou lorsqu'un nombre maximal d'itérations est atteint. L'algorithme retourne

alors la meilleure solution trouvée durant la recherche. Cette meilleure solution n'est pas nécessairement une solution optimale au problème, mais une solution de bonne qualité dite solution approchée.

### 4.1.2 La recherche locale pour l'optimisation des index d'insertion des acides aminés

**Espace de recherche :** La méthode de recherche locale que nous utilisons emploie un espace de recherche que l'on notera $\zeta$, qui représente toutes les configurations possibles pour la résolution du problème de prédiction du PS et du segment TM.

Une configuration $s$ ou un individu est un ensemble de $n-uplet$ de courbes représentées par $(V[a_i])_{i=A}^{i=Y}$. Dans le cas $n = 20$, chacun des 20 acides aminés sera représenté par une courbe distincte, par contre dans le cas $n < 20$ certains acides aminés seront représentés par une même courbe. Par exemple, on peut attribuer une même courbe à deux acides aminés différents si ces derniers exercent le même comportement durant l'insertion des segments dans la membrane.

**Fonction d'évaluation :** Pour évaluer la qualité d'une configuration $s$ de l'espace de recherche $\zeta$, la méthode la plus simple consiste à utiliser une fonction qui retourne la capacité du classifieur à discriminer les séquences PS des séquences TM. Les fonctions d'évaluation que nous utilisons sont l'AUC ou le taux de bonne prédiction (*accuracy*).

**Voisinage :** Un algorithme de recherche locale effectue un mouvement qui consiste à modifier la configuration $s$ en changeant un ou plusieurs des éléments qui la définisse. Une configuration $s'$ est dite voisine de $s$ si elle s'obtient en appliquant à $s$ un mouvement donné. L'ensemble des configurations accessibles à partir de $s$ avec les mouvements que l'on autorise est appelé **voisinage** de $s$.

Rappelons qu'une configuration est un $n-uplet$ de courbes. Un mouvement consisteà modifier une ou plusieurs courbes et les modifications envisagées nous donneront également différents types de mouvements.

Nous parlons d'un voisinageà **1 dimension** dans le cas ou nous modifions la courbe d'un seul acide aminé et d'un voisinage à $K$ **dimensions** si nous changeons les courbes de $K$ acides aminés en même temps.

**Stratégie de recherche :** La stratégie de recherche locale que nous utilisons suit le principe de la **descente** (ou amélioration itérative). À chaque itération, l'algorithme que nous avons implémenté parcourt l'ensemble des voisins jusqu'à ce qu'un voisin de qualité supérieure soit sélectionné. La descente accepte uniquement des voisins de meilleure qualité, et termine lorsqu'un optimum local est atteint, c'est-à-dire, qu'il n'existe plus aucun voisin de la solution courante dont la qualité lui soit strictement supérieure. Dans notre cas, cela signifie qu'une configuration $s'$ remplace la configuration courante $s$ si et seulement si $f(s') > f(s)$.

## 4.2 LSTranslocon : un algorithme de recherche locale

Nous présentons dans cette section, la première méthode nommée LSTranslocon qui est développée pour apprendre des courbes d'insertion pour les 20 acides aminés.

### 4.2.1 Présentation de la méthode

Afin d'ajuster les courbes d'insertion des acides aminés, nous commençons par générer une configuration initiale, puis de manière itérative nous allons nous déplacer vers une solution voisine selon le voisinage donné. En effet, à chaque itération, l'exploration du voisinage nous permettra de sélectionner une solution (ensemble de courbes) de meilleure qualité que la solution courante. L'optimisation des courbes s'arrête lorsque la condition d'arrét est satisfaite : soit un nombre d'itérations fixé au départ est atteint, soit une solution jugée de bonne qualité est trouvée. Enfin, l'approche retourne la solution trouvée.

Les résultats des expérimentations d'Hessa *et al.* suggérent que le profil des vecteurs d'énergie d'insertion des acides aminés prend une forme symétrique sous forme de gaussienne. Cette forme symétrique (figure 4.1 (A)) signifie que chaque acide aminé exerce la même influence sur les interfaces de la membrane et une influence inverse au milieu de la membrane.

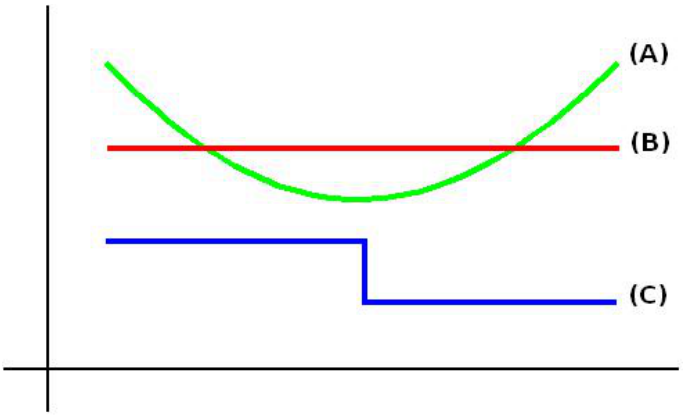


FIGURE 4.1 – Différentes formes de courbes. (A) représente une forme symétrique, (B) représente une droite, et (C) représente une forme en escalier.

Nous proposons dans LSTranslocon de déterminer pour chaque acide aminé une courbe qui détermine son aptitude à s'insérer dans la membrane sous forme d'une énergie d'interaction entre l'acide aminé et la membrane. L'apparence de cette courbe prend une forme de courbe symétrique.

### 4.2.2 Représentation d'une courbe

En nous inspirant des résultats des expérimentations d'Hessa *et al* [Hessa *et al.*, 2007], nous avons choisi dans LSTranslocon de représenter les courbes d'insertion des acides

aminés par une parabole. Chaque courbe sera représentée par une équation de la forme :

$$H = \alpha(x - X_0)^2 + \beta \tag{4.1}$$

où $x$ varie entre $X_{min} = 1$ et $X_{max} = l$ (la taille de la fenêtre considérée) et $X_0$ est le point central de cette fenêtre.

La fonction 4.1 peut être exprimée en fonction du couple de paramètres $(\alpha, \beta) \in \Re^2$. Le paramètre $\alpha$ détermine la courbure et l'orientation de la courbe. Un $\alpha$ positif correspond à une courbe convexe, i.e. la valeur au centre de la fenêtre est plus grande que la valeur aux extrémités. Un $\alpha$ négatif correspond à une courbe concave. L'allure de la courbe s'interprète de la manière suivante en termes biologiques : Si un acide aminé a une courbe convexe, cela signifie que sa contribution dans l'insertion des segments est plus grande au milieu de la courbe, tandis que si un acide aminé a une courbe concave, cela signifie que sa contribution est plus grande sur les extrémités de la courbe.

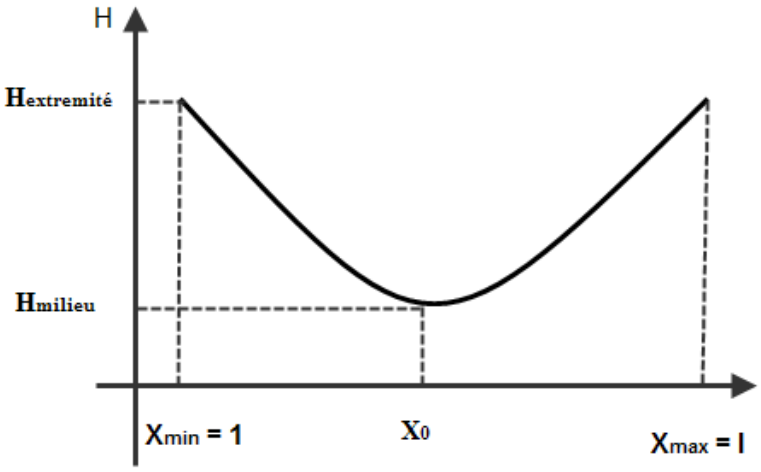


FIGURE 4.2 – Représentation d'une courbe définie par le couple de paramètres $(H_{extremite}, H_{milieu})$.

Nous pouvons aussi définir cette courbe par le couple de paramètres $(H_{extremite}, H_{milieu})$ (figure 4.2). La valeur $H_{extremite}$ est la valeur de la fonction aux extrémités pour $X_{min} = 1$ et $X_{max} = l$, sachant que $l$ représente la taille de la fenêtre. La valeur $H_{milieu}$ est la valeur de la fonction au milieu de la fenêtre.

Pour modifier la courbe on peut agir sur $(\alpha, \beta)$ ou sur $(H_{extremite}, H_{milieu})$. Les modifications s'interprètent plus aisément si on agit sur $(H_{extremite}, H_{milieu})$. Modifier $H_{extremite}$ signifie que l'on agit sur le comportement de l'acide aminé aux interfaces de la membrane et agir sur $H_{milieu}$ signifie que l'on agit sur le comportement de l'acide aminé au milieu de la membrane.

Le lien entre $(\alpha, \beta)$ et $(H_{extremite}, H_{milieu})$ est donné par les formules $H_{milieu} = \alpha(X_0 - X_0)^2 + \beta$, soit $H_{milieu} = \beta$ et $H_{extremite} = \alpha(\frac{X_{min} - X_{max}}{2})^2 + \beta$ soit $\alpha = 4\frac{(H_{extremite} - H_{milieu})}{(X_{min} - X_{max})^2}$.

Notons que les courbes manipulées peuvent être des droites (figure 4.1 (B)). Il s'agit de paraboles particuliéres définies par $(\alpha = 0, \beta)$ ou $H_{extremite} = H_{milieu} = \beta$.

Dans notre travail, nous cherchons des courbes avec une fenêtre de taille $l = 21$. Les travaux d'Hessa *et al* utilisent une fenêtre de taille 19 acides aminés, mais les résultats présentés dans [Pasquier *et al.*, 1999] indiquent qu'une taille de 21 est plus appropriée. En effet, la distribution statistique des segments TM dans des protéines avec une structure 3D connue montre que la plupart des segments ont une taille qui varie entre 21 et 30 acides aminés sachant que 60% de ces segments ont une taille de 21 acides aminés.

### 4.2.3 Construction d'une configuration initiale

Nous décidons dans notre travail de commencer la recherche locale avec une configuration initiale de bonne qualité qui permettra à la recherche locale de converger rapidement vers une meilleure solution.

Par conséquent, nous considérons des courbes prenant les valeurs d'hydrophobie données par l'échelle de Kyte et Doolittle (voir section 2.3.2 chapitre 2). Ces valeurs sont connues pour avoir de bonnes performances de discrimination des peptides signal [Bannai *et al.*, 2002].

Nous considérons que les index de la solution initiale sont indépendants de la position des acides aminés dans la séquence. Pour un acide aminé $a^i$, la courbe est une droite représentée par la valeur $h_i$. La configuration initiale est $S_0 = (V[a^i])_{i=A}^{i=Y}$ tel que $\forall j \in [1, l]$, $V[a^i][j] = h_i$.

### 4.2.4 Espace de recherche

Nous avons expliqué dans la section 4.1.2 que $\zeta$ représente l'ensemble des configurations. Dans cet espace de recherche un individu ou une configuration est un $n - uplet$ de courbes représentée par $(V[a_i])_{i=A}^{i=Y}$. Dans LSTranslocon, nous considérons $n = 20$ et par conséquent chaque acide aminé est représenté par une courbe distincte. Comme chaque courbe dépend de deux paramètres $(H_{extremite}, H_{milieu})$, l'espace de recherche $\zeta$ est défini par $[(H_{extremite_i}, H_{milieu_i})]_{i=A}^{i=Y}$.

### 4.2.5 Voisinage

La recherche locale effectue un mouvement d'une solution $s$ vers une solution voisine $s'$ qui améliore la solution courante $s$. Dans LSTranslocon, un mouvement consiste à modifier la courbe d'un seul acide aminé choisi aléatoirement. Nous rappelons qu'une courbe d'un acide aminé est définie par le couple $(H_{extremite}, H_{milieu})$ (section 4.2.2). La génération d'une nouvelle courbe est donnée par le couple $(H_{extremite} \pm \epsilon, H_{milieu} \pm \delta)$, $\epsilon$ et $\delta$ sont deux paramètres de notre algorithme. On peut dire que $\epsilon$ et $\delta$ définissent la largeur du voisinage et qui consisterait à bouger la courbe de + $\epsilon$ ou - $\epsilon$ (+ ou - $\delta$).

Dans notre cas, nous avons considéré le voisinage correspondant à $(\epsilon, \delta) \in \{-0.5, 0, 0.5\}^2$ ou $(\epsilon, \delta) \in \{-0.3, 0, 0.3\}^2$, cela conduit é un total de 18 combinaisons. Nous supprimons les deux cas triviaux avec $(\epsilon, \delta) = (0, 0)$, ce qui permet de générer 16 courbes différentes à

partir d'une courbe existante. En d'autres termes, chaque solution $s$ a 16 solutions voisines, dans lesquelles une seule courbe de $s$ a été modifiée.

### 4.2.6 Fonction d'évaluation

À chaque itération de la recherche locale, tous les voisins de la solution courante $s$ sont évalués. Selon le principe de la **descente**, la meilleure solution est choisie pour remplacer la solution actuelle $s$ et le processus de recherche locale est itéré à partir de cette nouvelle solution. La qualité d'une solution voisine $s'$ est évaluée par la fonction d'évaluation $AUC(s')$ qui estime la capacité de la solution à obtenir une bonne discrimination entre le PS et le segment TM, en utilisant le système de classification fondé sur les courbes de $s'$. Notons qu'il est suffisant de calculer l'AUC du système de classification associé à l'ensemble de courbes pour évaluer une solution. La détermination d'un seuil spécifique $\tau$ de discrimination entre le PS et le segment TM n'est pas nécessaire à ce moment.

### 4.2.7 Opérateur de perturbation

La recherche locale est une méthode d'optimisation itérative qui améliore la solution courante en explorant le voisinage. Il arrive parfois que l'exploration du voisinage ne permette pas de trouver un voisin de meilleur qualité que la solution courante, c'est-à-dire, $\forall s' \in N(s)$, $f(s) > f(s')$. Dans ce cas-la, l'exploration identifie un optimum local. Cet optimum local est une solution pour laquelle il n'existe pas de solution de meilleure qualité.

Pour contourner un optimum local, une solution consiste à appliquer un opérateur de perturbation [LourenÃ§o *et al.*, 2000]. En effet, notre recherche locale explore à chaque itération le voisinage pour sélectionner un voisin améliorant la solution courante. Le processus s'arrête lorsqu'elle la recherche locale atteint un optimum local. Á ce moment, nous perturbons la meilleure solution afin de générer une nouvelle solution initiale pour une prochaine relance de notre recherche locale.

Notre opérateur de perturbation sélectionne un acide aminé et modifie sa courbe selon l'opérateur de mouvement décrit en haut. Notons que la courbe choisie peut dégrader la qualité de la solution qui va être perturbée.

### 4.2.8 Expérimentations et discussions

Les expérimentations que nous présentons dans cette section sont obtenues en appliquant notre algorithme LSTranslocon sur les différents jeux de données décrits dans la section 3.3 du chapitre 3. Nous rappelons que les courbes d'insertion sont apprises sur le jeu de données SWP. Par la suite, nous utilisons ces courbes afin de discriminer les données des jeux SCAMPI et PDB.

Tout d'abord, le tableau 4.1 montre une évaluation de notre classifieur basée sur la configuration initiale. Notre configuration initiale est unique et elle est représentée par des droites prenant les valeurs de l'échelle d'hydrophobie de Kyte et Doolittle. Le seuil de discrimination entre le PS et le segment TM est appris sur le jeu d'apprentissage SWP et représente le meilleur score qui donne le meilleur taux de bonne prédiction.

| | SWP Test | SCAMPI HighFirst | SCAMPI HighAll | SCAMPI LowFirst | SCAMPI LowAll | PDB First | PDB All |
|---|---|---|---|---|---|---|---|
| AUC | 0.826 | 0.684 | 0.645 | 0.765 | 0.675 | 0.642 | 0.580 |
| Accuracy | 0.747 | 0.764 | 0.593 | 0.788 | 0.564 | 0.702 | 0.516 |

TABLE 4.1 – Résultats de l'AUC et de l'accuracy obtenu avec LSTranslocon sur la configuration initiale.

On observe dans le tableau de faibles performances de discrimination sur les différents jeux de tests. À partir de cette configuration initiale, nous exécutons plusieurs fois notre algorithme LSTranslocon. Le tableau 4.2 montre les résultats de 5 exécutions en présentant les valeurs de l'AUC et de l'*accuracy* obtenus sur le jeu SWP par le processus de validation croisée décrit dans la section 3.3.4 du chapitre 3.

Nous remarquons que nous avons amélioré considérablement les résultats de l'AUC et du taux de bonne prédiction sur le jeu SWP. De plus, nous observons que les résultats sont stables dans les différentes exécutions de LSTranslocon. Nous obtenons une moyenne de l'AUC égale à 0.920 (0.826 avec la solution initiale ) et une moyenne de l'accuracy égale à 0.854 (0.747 avec la solution initiale).

| | exe1 | | exe2 | | exe3 | | exe4 | | exe5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| fold 1 | 0.905 | 0.836 | 0.912 | 0.833 | 0.919 | 0.848 | 0.924 | 0.850 | 0.903 | 0.845 |
| fold 2 | 0.935 | 0.865 | 0.936 | 0.855 | 0.939 | 0.862 | 0.907 | 0.869 | 0.933 | 0.867 |
| fold 3 | 0.914 | 0.867 | 0.925 | 0.838 | 0.911 | 0.852 | 0.940 | 0.823 | 0.911 | 0.845 |
| fold 4 | **0.946** | **0.862** | 0.942 | 0.845 | 0.946 | 0.867 | 0.912 | 0.896 | 0.934 | 0.865 |
| fold 5 | 0.936 | 0.852 | 0.925 | 0.852 | 0.949 | 0.852 | 0.946 | 0.838 | 0.912 | 0.838 |
| fold 6 | 0.918 | 0.826 | 0.908 | 0.839 | 0.902 | 0.840 | 0.935 | 0.836 | 0.900 | 0.852 |
| fold 7 | 0.941 | 0.867 | 0.942 | 0.860 | 0.945 | 0.865 | 0.915 | 0.867 | 0.948 | 0.867 |
| fold 8 | 0.923 | 0.852 | 0.918 | 0.845 | 0.925 | 0.867 | 0.940 | 0.860 | 0.919 | 0.838 |
| fold 9 | 0.946 | 0.867 | 0.941 | 0.869 | 0.946 | 0.848 | 0.929 | 0.830 | 0.932 | 0.833 |
| fold 10 | 0.924 | 0.845 | 0.928 | 0.855 | 0.928 | 0.860 | 0.918 | 0.860 | 0.931 | 0.867 |
| moyenne | 0.929 | 0.854 | 0.926 | 0.858 | 0.931 | 0.856 | 0.927 | 0.850 | 0.922 | 0.852 |
| écart type | 0.016 | 0.022 | 0.021 | 0.031 | 0.003 | 0.020 | 0.015 | 0.026 | 0.015 | 0.015 |

TABLE 4.2 – Résultats de 5 exécutions sur le jeu SWP sans perturbation.

Nous comparons ces résultats avec Phobius (décrit en section 2.3.3 du chapitre 2) l'une des meilleures méthodes de double prédiction du PS et du segment TM. Pour une comparaison équitable, nous évaluons Phobius sur chaque fold du jeu SWP en calculant le taux de bonne prédiction. Puis, nous calculons la moyenne des 10 évaluations. Phobius obtient une moyenne égale à 0.856 (0.854 avec LSTranslocon) avec un écart type de 0.032 (0.031 avec LSTranslocon). Nous constatons que la moyenne du taux de bonne prédiction obtenue avec LSTranslocon est similaire à la moyenne du taux de bonne prédiction obtenue avec Phobius.

L'objectif premier de notre méthode est de déterminer les courbes qui représentent la contribution des acides aminés lors de l'insertion des segments dans la membrane. Puis, ces courbes sont utilisées pour discriminer les séquences PS et les séquences TM. Par contre, l'objectif principal de Phobius est la discrimination entre les séquences PS et les séquences TM. Notre méthode présente des performances comparables à l'une des meilleures méthodes de prédiction.

Le tableau 4.3 montre l'évaluation de LSTranslocon sur les autres jeux de tests ainsi que les valeurs de la matrice de confusion sans l'application de l'opérateur de perturbation. Nous observons une amélioration des performances de discrimination de tous les jeux de données utilisés pour le test en comparaison avec les résultats obtenus avec la configuration initiale (tableau 4.1).

| | | | | Matrice de confusion | |
|---|---|---|---|---|---|
| | | | | PS | TM |
| SWP Test | 0.946 | 0.862 | Prédit PS | 64 | 15 |
| | | | Prédit TM | 4 | 53 |
| SCAMPI HighFirst | 0.844 | 0.796 | Prédit PS | 293 | 25 |
| | | | Prédit TM | 69 | 75 |
| SCAMPI HighALL | 0.837 | 0.754 | Prédit PS | 293 | 123 |
| | | | Prédit TM | 69 | 296 |
| SCAMPI LowFirst | 0.850 | 0.809 | Prédit PS | 293 | 23 |
| | | | Prédit TM | 69 | 99 |
| SCAMPI LowAll | 0.831 | 0.729 | Prédit PS | 293 | 222 |
| | | | Prédit TM | 69 | 490 |
| PDB First | 0.780 | 0.750 | Prédit PS | 293 | 66 |
| | | | Prédit TM | 69 | 114 |
| PDB All | 0.729 | 0.638 | Prédit PS | 293 | 268 |
| | | | Prédit TM | 69 | 301 |

TABLE 4.3 – Résultats de l'évaluation de LSTranslocon sur les autres jeux de tests avec le classifieur du fold 4 de la première exécution.

Nous observons aussi que nous obtenons de meilleurs résultats en termes d'AUC et du taux de bonne prédiction lorsque les données de test sont constituées seulement du premier segment TM (SCAMPIHighfirst, SCAMPIlowfirst, et PDBfirst). Dans le cas contraire, lorsque les jeux de tests contiennent tous les segments TM d'une protéine, nous observons des moins bons résultats. La première explication à ce phénomène est que l'insertion du premier segment TM se comporte différemment de l'insertion des segments TM se trouvant plus loin dans la séquence protéique. Cette observation est conforme à ce qui a déjâ été signalé dans la littérature [Park and Helms, 2008a]. L'insertion des segments TM se trouvant à l'intérieure de la protéine nécessite une énergie plus faible que l'insertion du premier segment. La deuxième explication est que nous avons concentré notre apprentissage seulement sur les premiers segments TM des protéines, ce qui permet de mieux prédire le premier segment.

Notons aussi dans la matrice de confusion que la prédiction du PS donne toujours de bonne performance. La différence dans les résultats ce joue principalement sur les segments TM.

Nous avons souligné en section 4.2.7, que les méthodes fondées sur un algorithme de recherche locale peuvent être piégées dans un optimum local. Pour contourner ce problème,

nous appliquons un opérateur de perturbation afin de diversifier notre recherche.

Le tableau 4.4 montre les résultats obtenus en appliquant l'opérateur de perturbation décrit dans la section 4.2.7 lors de l'apprentissage des courbes. Les résultats obtenus avec perturbation montrent une **légère amélioration** sur le jeu SWP et une amélioration plus marquante sur les autres jeux de tests, par rapport aux résultats obtenus sans perturbation (tableau 4.3). En d'autres termes, notre opérateur de perturbation a un effet modéré.

| | AUC | Accuracy | | Matrice de confusion | |
|---|---|---|---|---|---|
| SWP Test | 0.947 | 0.875 | Prédit PS | 63 | 12 |
| | | | Prédit TM | 5 | 56 |
| SCAMPI HighFirst | 0.861 | 0.818 | Prédit PS | 324 | 27 |
| | | | Prédit TM | 38 | 73 |
| SCAMPI HighALL | 0.858 | 0.765 | Prédit PS | 324 | 145 |
| | | | Prédit TM | 38 | 274 |
| SCAMPI LowFirst | 0.870 | 0.821 | Prédit PS | 324 | 29 |
| | | | Prédit TM | 38 | 93 |
| SCAMPI LowAll | 0.858 | 0.712 | Prédit PS | 324 | 271 |
| | | | Prédit TM | 38 | 441 |
| PDB First | 0.799 | 0.784 | Prédit PS | 324 | 79 |
| | | | Prédit TM | 38 | 101 |
| PDB All | 0.757 | 0.648 | Prédit PS | 324 | 289 |
| | | | Prédit TM | 38 | 280 |

TABLE 4.4 – Résultats obtenus avec LSTranslocon et un l'opérateur de perturbation.

Nous présentons dans la figure 4.3 un exemple de courbes d'insertion obtenues pour les 20 acides aminés. Nous pouvons observer que la forme des courbes est différente selon l'acide aminé. Ces formes de courbes suggérent que certains acides aminés tels que les acides aminés proline (P) ou méthionine (M) facilitent davantage l'insertion lorsque ces acides aminés sont intégrés à l'intérieur de la membrane au milieu de la courbe, alors que d'autres acides aminés tels que la tryptophane (W) ou la glutamine (Q) préférent les positions d'interfaces aux extrémités des courbes. Nous constatons aussi que certains acides aminés comme la tyrosine (Y) et la valine (V) sont représentés par une droite ce qui signifie que ces acides aminés exercent la même influence au milieu et aux interfaces de la membrane.

Nous avons constaté dans nos différentes expérimentations que les courbes de certains acides aminés changent de forme (de concavité) d'une exécution à l'autre. Cela est obtenu en particulier pour les acides aminés peu fréquents dans notre jeu de données. Une explication possible de cette observation est que ces acides aminés n'ont pas une grande influence dans le processus d'insertion, ce qui explique leur manque de représentation dans le jeu de données et la difficulté à ajuster correctement leur courbe d'insertion.

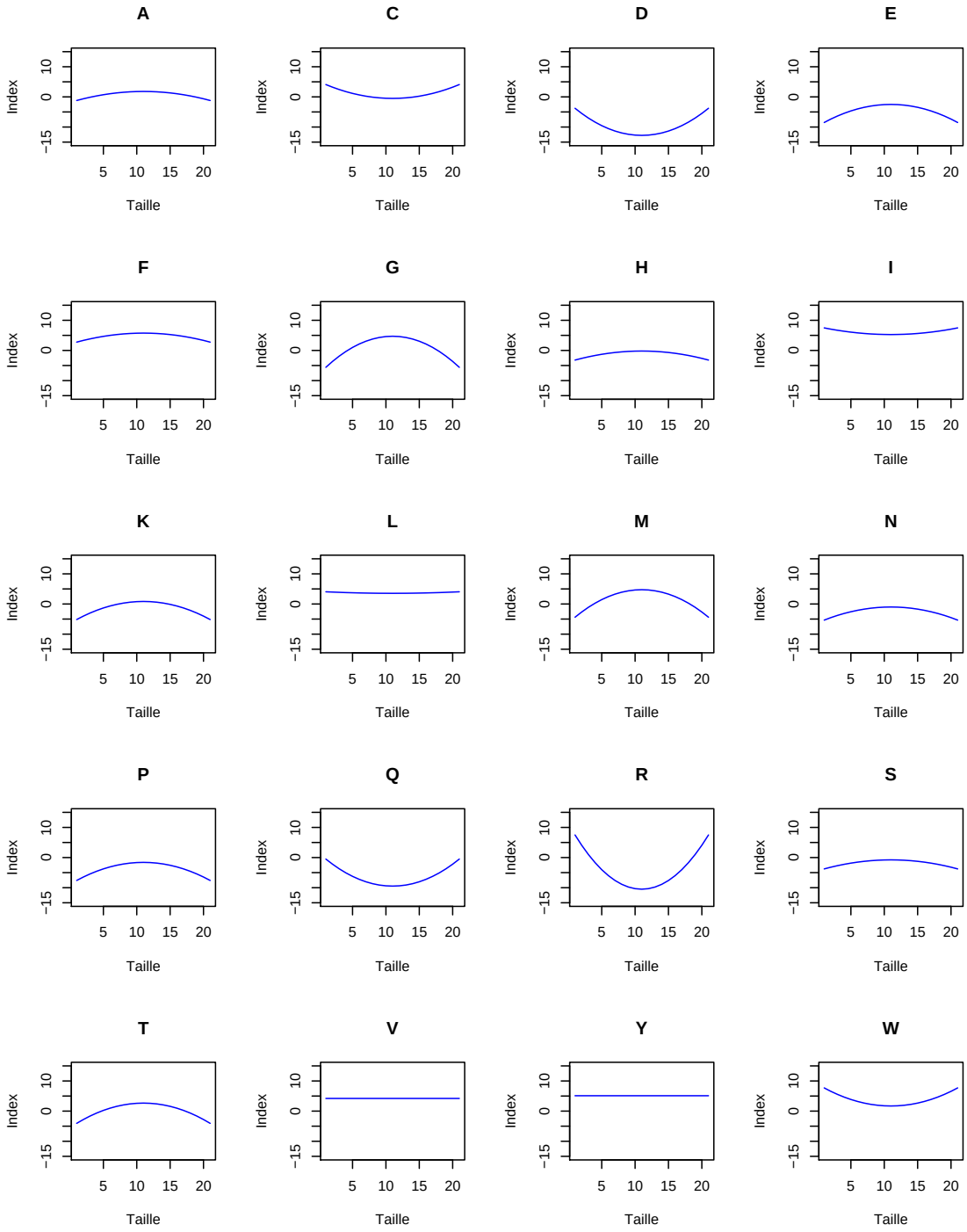


FIGURE 4.3 – Courbes d'insertion des acides aminés correspondante au classifieur du fold 4 de la première exécution. Les courbes sont définies sur une fenêtre de longueur 21 acides aminés.

### 4.2.9 Synthèse

Nous avons présenté LSTranslocon, notre première méthode d'optimisation des courbes des acides aminés. LSTranslocon emploie une approche par recherche locale afin d'améliorer de manière itérative les courbes d'insertion. L'application de l'opérateur de perturbation a montré un effet modéré dans l'amélioration de la méthode.

Les expérimentations réalisées sur les différents jeux de tests, nous ont permis d'observer que les courbes de certains acides aminés changent de forme. Face à ce problème, nous présentons dans la section ci-dessous une approche pour résoudre le problème d'instabilité de nos courbes.

## 4.3 MN-LS : un algorithme de recherche locale avec un nouvel espace de recherche

Nous présentons dans cette section, une amélioration de la méthode LSTranslocon nommée MN-LS. La méthode explore un espace modifié et intégre un voisinage multiple pour l'optimisation des courbes.

### 4.3.1 Motivation

Afin de corriger le problème d'instabilité de nos courbes, nous proposons dans MN-LS d'explorer un nouvel espace de recherche. [Kyte and Doolittle, 1982] suggérent que certains acides aminés ont une valeur d'hydrophobie faible, tandis que d'autres acides aminés ont une grande valeur d'hydrophobie. Pour les acides aminés ayant une faible valeur d'hydrophobie, il est observé qu'ils jouent un rôle minime lors de l'insertion des segments TM. De plus, certains acides aminés sont peu présents dans les segments TM [Ulmschneider *et al.*, 2005a]. Nous supposons donc qu'il est suffisant de représenter ces acides aminés par une droite (figure 4.1 (B)), ce qui signifie que l'acide aminé exerce la même influence au milieu et aux interfaces de la membrane. Pour l'ensemble des 20 acides aminés, nous pouvons combiner 2 formes de courbes. Le classifieur pourra utiliser des droites pour certains acides aminés peu fréquents et peu influents et des courbes symétriques pour d'autres acides aminés qui sont fréquents dans le jeu de données.

Dans MN-LS, nous optimisons des droites et des courbes symétriques. Nous associons une droite aux acides aminés peu fréquents et donc un seul paramètre à déterminer, tandis que pour les acides aminés fréquents dans le jeu d'apprentissage nous associons une courbe symétrique et donc deux paramètres à déterminer.

De plus, nous avons pu constater dans LSTranslocon que la perturbation apporte seulement de légères améliorations dans la stratégie de recherche locale employée. Par conséquent, nous décidons d'utiliser un autre mécanisme pour échapper à un optimum local. Nous utilisons dans cette nouvelle stratégie un voisinage multiple qui consiste à explorer plusieurs voisinages de manière itérative.

### 4.3.2 Nouvel espace de recherche

Dans ce nouvel algorithme, nous proposons de considérer que l'index d'insertion peut être défini par une droite (valeur de l'index constante sur la fenêtre) ou par une courbe symétrique. La courbe symétrique est définie par les deux paramètres ($H_{extremite}$, $H_{milieu}$), tandis que la droite est définie par un seul paramètre $H_{milieu}$.

Selon la répartition statistique des acides aminés dans notre ensemble de données (voir la section 3.3.1 du chapitre 3) et selon les propriétés d'hydrophobie des acides aminés (échelle de Kyte et Doolittle), nous proposons de considérer deux groupes d'acides aminés.

D'une part, les acides aminés alanine (A), phenylalanine (F), isoleucine (I), leucine (L) et valine (V) sont très fréquents dans notre jeu de données. D'autre part, nous remarquons qu'ils sont connus pour avoir des valeurs d'hydrophobie élevées et jouent donc un rôle important dans l'insertion des segments dans la membrane. Par conséquent, nous pensons disposer de suffisamment d'informations dans notre jeu d'apprentissage pour ajuster les courbes de ces acides aminés. Nous formons donc un groupe noté $\mathcal{C}$, pour lequel chaque courbe d'index d'insertion sera une courbe symétrique. Le deuxième groupe, noté groupe $\mathcal{D}$, contient les quinze autres acides aminés pour lesquels chaque index d'insertion est défini par une droite.

Ces hypothèses définissent un nouvel espace de recherche pour notre nouvel algorithme (voir figure 4.4), où une solution $s$ est un ensemble de 15 droites et 5 courbes symétriques, chaque courbe étant définie sur l'intervalle $[1, 21]$. Une droite est définie par un paramètre unique (voir section 4.2.2), ceci réduit le nombre de paramètres qui caractérisent une solution et réduit donc la taille de l'espace de recherche.

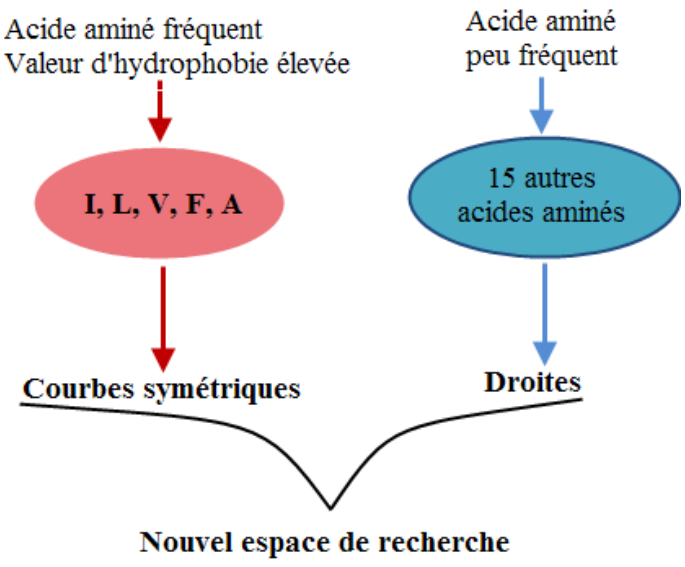


FIGURE 4.4 – Un nouvel espace de recherche.

Pour explorer cet espace de recherche, nous adoptons une stratégie en deux temps. Nous commenéons avec une solution initiale $s_0$ définie par 20 valeurs d'hydrophobie constantes.

Notre algorithme optimise d'abord les valeurs des 15 droites. Puis, quand un optimum est atteint pour les droites, nous optimisons les 5 courbes symétriques.

Dans chaque cas, nous explorons le voisinage de la solution actuelle afin de choisir la solution qui offre la meilleure amélioration de l'AUC.

### 4.3.3 Voisinage multiple

Pour mener la tâche de détermination des courbes d'insertion de manière efficace, nous introduisons dans cette section une recherche à voisinage multiple. Notre algorithme est conéu pour explorer de maniére séquentielle les deux espaces de recherche afin d'optimiser 15 droites associées aux acides aminés du groupe $\mathcal{D}$ puis l'ensemble des 5 courbes symétriques du groupe $\mathcal{C}$.

Pour ce faire, notre algorithme MN-LS emploie différents voisinages et utilise une stratégie pour explorer ces voisinages de manière séquentielle.

**Voisinage pour les droites (groupe $\mathcal{D}$) :** Comme chaque droite est définie par le paramètre $H_{milieu}$ (voir section 4.2.2), la solution du premier espace de recherche (groupe $\mathcal{D}$ des 15 acides aminés) est identifiée par un vecteur de 15 valeurs. Étant donné une telle solution $s$, nous choisissons aléatoirement un acide aminé dans la solution $s$ et nous modifions sa courbe en ajoutant ou en soustrayant une valeur $\epsilon$, soit $(H_{milieu} \pm \epsilon)$. Des grandes valeurs pour $\epsilon$ ménent à d'importants changements de la droite, tandis que des valeurs petites pour $\epsilon$ ne donnent que de légéres modifications. Dans ce qui suit, nous considérons le voisinage correspondant à $\epsilon \in \{0.3, 0.7\}$, soit 4 voisins.

L'algorithme commence son exploration en examinant ce premier espace de recherche. Ceci est simplement réalisé en appliquant la stratégie de la descente pour explorer le voisinage donné. Chaque voisin est évaluée selon le score de l'AUC (section 4.2.6), le meilleur voisin améliorant la solution courante est choisi pour remplacer la solution $s$. La recherche s'arrête si aucune solution voisine n'améliore la solution courante. Lorsque cette phase est terminée, l'algorithme passe à la phase suivante de l'optimisation des courbes symétriques.

**Voisinage pour les courbes symétriques (groupe $\mathcal{C}$) :** Nous rappelons que chaque courbe symétrique est définie par le couple $(H_{extremite}, H_{milieu})$ (voir section 4.2.2). Une solution est représentée par 5 courbes symétriques, définie par 5 couples $(H_{extremite}, H_{milieu})$. Étant donné une solution $s$, nous générons une solution voisine en ajoutant ou en soustrayant une valeur $\epsilon$ d'un couple $(H_{extremite} \pm \epsilon, H_{milieu} \pm \epsilon)$ de $s$. Ainsi, pour chaque couple $(H_{extremite}, H_{milieu})$ de $s$, nous avons huit voisins. Dans ce qui suit, $\epsilon$ est expérimentalement fixé à 0,7, 0,5 et 0,3, soit $\epsilon \in \{0.3, 0.5, 0.7\}$.

Pour explorer l'espace de recherche des courbes symétriques, l'algorithme opère en examinant d'abord le voisinage défini par la plus grande valeur $\epsilon = 0.7$ . Après avoir atteint un optimum local, l'algorithme passe au voisinage suivant défini par $\epsilon = 0.5$ jusqu'â trouver un autre optimum local. La recherche continue avec le voisinage défini par $\epsilon = 0.3$. Nous justifions ces explorations successives par le fait qu'il est préférable de faire des changements importants afin d'assurer une large exploration de l'espace au début de la recherche et de limiter les changements pour une exploration plus fine vers la fin de la

recherche.

### 4.3.4 Expérimentations et discussions

Dans cette section, nous présentons les résultats de l'évaluation de MN-LS sur les différents jeux de tests. Dans ces premières expérimentations, nous utilisons l'échelle d'hydrophobies de Kyte et Doolittle comme configuration initiale.

Le tableau 4.5 montre les résultats de 5 exécutions en présentant les valeurs de l'AUC et du taux de bonne prédiction (*accuracy*) obtenue sur le jeu SWP par le processus de validation croisée. Effectivement, nous constatons une stabilité dans les résultats entre les différentes exécutions de notre méthode. Plus particulièrement, nous observons une amélioration des performances de discrimination en comparaison avec la méthode LSTranslocon. Nous obtenons une moyenne d'*accuracy* égale a 0.877 pour MN-LS comparé à 0.854 pour LSTranslocon.

| | exe1 | | exe2 | | exe3 | | exe4 | | exe5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| fold 1 | 0.917 | 0.845 | 0.917 | 0.852 | 0.915 | 0.852 | 0.917 | 0.845 | 0.919 | 0.860 |
| fold 2 | 0.935 | 0.889 | 0.940 | 0.882 | 0.937 | 0.867 | 0.936 | 0.882 | 0.941 | 0.889 |
| fold 3 | 0.912 | 0.852 | 0.918 | 0.852 | 0.916 | 0.860 | 0.915 | 0.852 | 0.916 | 0.845 |
| fold 4 | 0.969 | 0.891 | 0.963 | 0.894 | 0.961 | 0.889 | 0.961 | 0.875 | 0.966 | 0.882 |
| fold 5 | 0.951 | 0.860 | 0.948 | 0.882 | 0.944 | 0.882 | 0.959 | 0.882 | 0.954 | 0.882 |
| fold 6 | 0.910 | 0.828 | 0.908 | 0.846 | 0.907 | 0.836 | 0.908 | 0.826 | 0.904 | 0.826 |
| fold 7 | 0.959 | 0.882 | 0.956 | 0.882 | 0.956 | 0.889 | 0.906 | 0.897 | 0.954 | 0.887 |
| fold 8 | 0.935 | 0.882 | 0.937 | 0.889 | 0.935 | 0.897 | 0.935 | 0.894 | 0.930 | 0.889 |
| fold 9 | 0.958 | 0.889 | 0.963 | 0.899 | 0.960 | 0.891 | 0.963 | 0.889 | 0.962 | 0.889 |
| fold 10 | 0.938 | 0.882 | 0.935 | 0.897 | 0.932 | 0.867 | 0.938 | 0.875 | 0.935 | 0.860 |
| MN-LS moyenne | 0.938 | 0.870 | 0.939 | 0.877 | 0.936 | 0.873 | 0.938 | 0.872 | 0.938 | 0.871 |
| écart type | 0.020 | 0.029 | 0.019 | 0.029 | 0.019 | 0.027 | 0.020 | 0.026 | 0.020 | 0.025 |
| LSTranslocon moyenne | 0.929 | 0.854 | 0.926 | 0.858 | 0.931 | 0.856 | 0.927 | 0.850 | 0.922 | 0.852 |
| écart type | 0.016 | 0.022 | 0.021 | 0.031 | 0.003 | 0.020 | 0.015 | 0.026 | 0.015 | 0.015 |

TABLE 4.5 – Résultats de 5 exécutions sur le jeu SWP en appliquant la méthode MN-LS ainsi que les résultats de LSTranslocon (voir tableau 4.2).

Dans le tableau 4.6, nous comparons les résultats obtenus avec MN-LS aux résultats obtenus avec LSTranslocon sur les différents jeux de tests. Les deux méthodes débutent avec une même configuration initiale (échelle d'hydrophobie de Kyte et Doolittle). Nous remarquons que MN-LS permet une amélioration des performances de discrimination sur les différents jeux de tests, et particulièrement sur les jeux SCAMPIHighFirst et SCAMPILowFirst.

Le voisinage multiple permet à la méthode MN-LS d'explorer plusieurs voisinages. En début de recherche, nous avons appliqué des modifications importantes dans le voisinage, ce qui nous permet d'apporter des modifications marquantes sur la solution et donc d'améliorer sensiblement la qualité des solutions. Par contre en fin de recherche, nous appliquons un voisinage avec de petites valeurs permettant de légères modifications pour affiner les courbes.

Nous constatons aussi que l'utilisation du nouvel espace de recherche qui est de dimension réduite garantit une stabilité des courbes d'insertion entre différentes exécutions de notre méthode. MN-LS ajuste des courbes symétriques seulement pour les acides aminés

qui sont très fréquents dans le jeu de données et optimise une droite pour ceux qui sont moins fréquents dans le jeu de données (voir figure 4.5).

| | LSTranslocon | | MN-LS | |
|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy |
| SWP Test | 0.947 | 0.875 | 0.969 | 0.891 |
| SCAMPI HighFirst | 0.861 | 0.818 | 0.851 | 0.841 |
| SCAMPI HighALL | 0.858 | 0.765 | 0.829 | 0.747 |
| SCAMPI LowFirst | 0.870 | 0.821 | 0.878 | 0.831 |
| SCAMPI LowAll | 0.858 | 0.712 | 0.857 | 0.770 |
| PDB First | 0.799 | 0.784 | 0.794 | 0.782 |
| PDB All | 0.757 | 0.648 | 0.745 | 0.656 |

TABLE 4.6 – Comparaison entre les résultats d'évaluation de LSTranslocon (tableau 4.4) et les résultats d'évaluation de MN-LS correspondants au classifieur du fold 4 de la première exécution (tableau 4.5).

**étude de l'influence de la configuration initiale :** Dans cette expérimentation, nous étudions si une échelle particulière est mieux adaptée pour initier notre processus de recherche. Nous évaluons les deux échelles Eisenberg [Eisenberg *et al.*, 1982] et Engelman [Engelman *et al.*, 1986]. Ces échelles ne sont pas normalisées et leurs valeurs sont assez différentes. Par conséquent, nous donnons différents temps d'exécution pour le processus de recherche locale dans l'objectif d'avoir une comparaison équitable.

Le tableau 4.7 montre les résultats de l'initialisation de MN-LS avec les trois échelles d'hydrophobie. Nous observons que les résultats obtenus suivant les différentes échelles d'hydrophobies sont similaires. La configuration initiale n'a donc peu d'effet sur la recherche locale, en d'autres termes MN-LS est indépendante de la solution initiale.

| | MN-LS (Kyte et Doolittle) | | MN-LS (Eisenberg) | | MN-LS (Engelamn) | |
|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| SWP Test | 0.969 | 0.891 | 0.967 | 0.899 | 0.965 | 0.88 |
| SCAMPI HighFirst | 0.851 | 0.841 | 0.839 | 0.844 | 0.841 | 0.847 |
| SCAMPI HighALL | 0.829 | 0.747 | 0.824 | 0.745 | 0.823 | 0.746 |
| SCAMPI LowFirst | 0.878 | 0.831 | 0.858 | 0.847 | 0.868 | 0.831 |
| SCAMPI LowAll | 0.857 | 0.770 | 0.819 | 0.762 | 0.827 | 0.756 |
| PDB First | 0.794 | 0.782 | 0.756 | 0.776 | 0.755 | 0.787 |
| PDB All | 0.745 | 0.656 | 0.712 | 0.650 | 0.713 | 0.647 |

TABLE 4.7 – Évaluation de MN-LS en utilisant trois différentes configurations initiales.

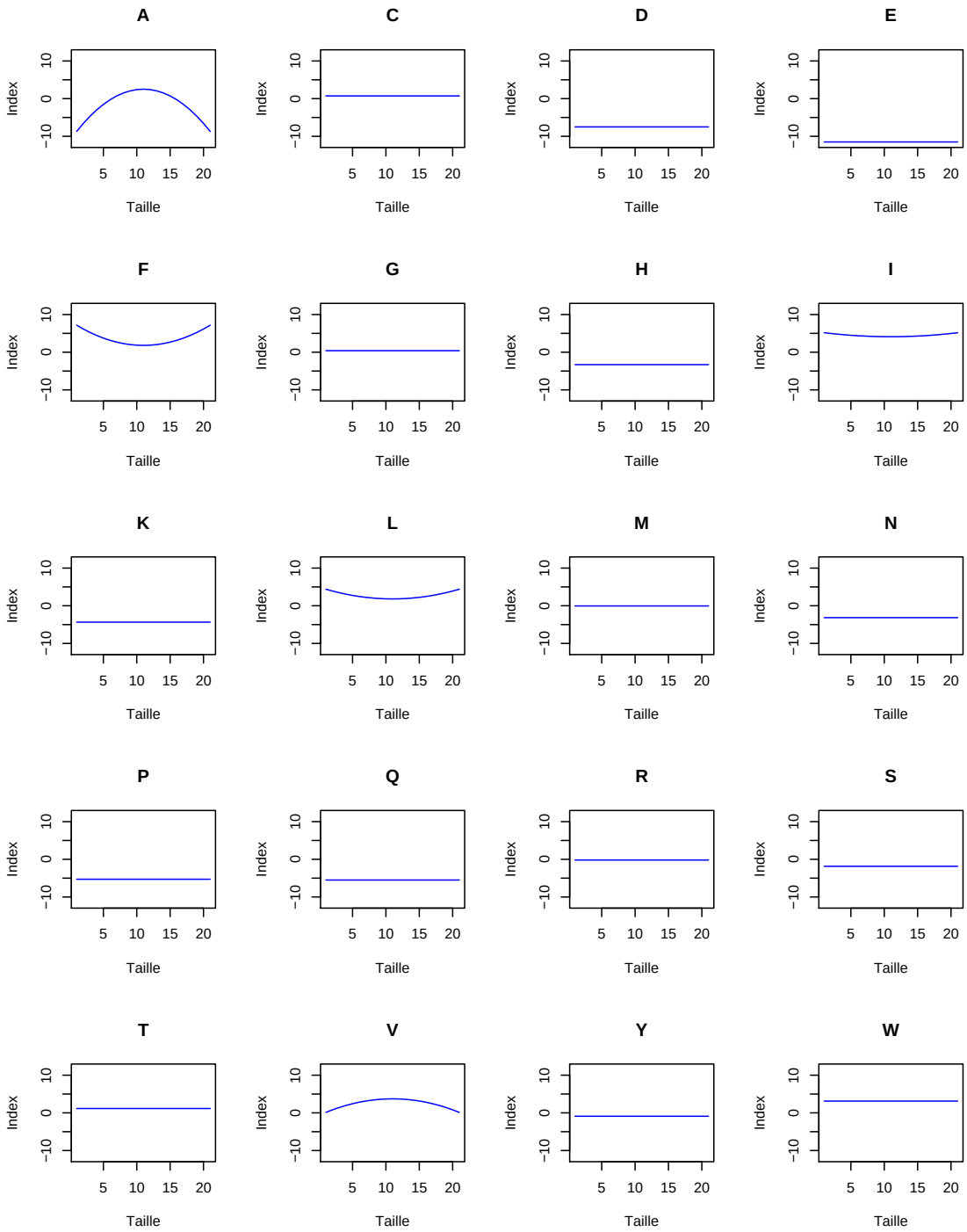


FIGURE 4.5 – Courbes d'insertion des acides aminés correspondant au classifieur du fold 4 de la première exécution. Les courbes sont définies sur une fenêtre de taille 21 acides aminés.

### 4.3.5 Synthèse

Dans cette nouvelle méthode, nous avons introduit deux fonctions pour améliorer l'approche précédente. En s'appuyant sur des propriétés connues des acides aminés et sur leur fréquence dans les données étudiées, la nouvelle méthode optimise d'abord un ensemble de 15 droites correspondant à l'ensemble des acides aminés avec une faible valeur d'hydrophobie, puis l'apprentissage cherche à optimiser des courbes symétriques pour les 5 acides aminés restants qui sont fortement hydrophobes.

Pour ces explorations en 2 étapes (en premier les droites et en deuxième les courbes symétriques), la méthode proposée examine de manière séquentielle et exhaustive différents voisinages. Les résultats expérimentaux montrent que la méthode donne de meilleurs résultats par rapport à la précédente méthode LSTranslocon en termes d'AUC et de taux de bonne classification. De plus, nous obtenons une stabilité dans les courbes d'insertion.

Malgré l'amélioration des performances de la méthode et la stabilité des courbes, nous observons que la stratégie de recherche que nous employons permet seulement d'intensifier la recherche de solution.

Nous présentons dans la section ci-dessous une méthode qui utilise un voisinage élargi afin d'explorer au mieux l'espace de recherche. De plus, nous examinons une nouvelle forme de courbe.

## 4.4 BioGLS : un algorithme de recherche locale à voisinage élargi

Nous présentons BioGLS, un algorithme intégrant de nouvelles connaissances biologiques et utilisant un voisinage élargi pour apprendre les courbes d'insertion.

### 4.4.1 Motivation

Dans les deux précédentes méthodes présentées en section 4.2 (pour la méthode LSTranslocon) et en section 4.3 (pour la méthode MN-LS), nous avons utilisé un voisinage à 1 dimension, c'est-à-dire, nous modifions la courbe d'un seul acide aminé dans une solution. La taille réduite de ce voisinage conduit souvent à des voisins de méme qualité de discrimination et présente une très faible capacité d'amélioration de la solution. Le nombre d'itérations de la recherche locale sans amélioration est important, ce qui nous conduit é une exploration très locale de l'espace de recherche.

En effet, en examinant le processus de recherche locale nous avons constaté que le positionnement de la fenêtre ayant la plus grande moyenne d'hydrophobie lors de la configuration initiale varie seulement (modification de la position) de quelques acides aminés en comparaison avec la fenêtre ayant la plus grande valeur d'hydrophobie calculée à la fin du processus de recherche. En d'autres termes, la fenêtre a été décalée de seulement quelques acides aminés. Par conséquent, les courbes optimisées é la fin du processus de recherche locale sont liées aux valeurs de la configuration initiale.

Afin d'exploiter au maximum la puissance de la recherche locale et en même temps de se libérer de la configuration initiale, nous proposons d'utiliser une recherche locale qui

explore un voisinage élargi de $K$ dimensions. Contrairement aux deux premières méthodes, l'utilisation d'un voisinage de $K$ dimensions modifie plusieurs courbes d'acides aminés lors d'un seul mouvement, ce qui nous permettra de faire bouger la fenêtre avec d'importantes modifications pour ne plus dépendre des valeurs initiales.

La deuxième contribution dans cette méthode est l'introduction de nouvelles connaissances biologiques pour mieux ajuster nos courbes. En particulier, nous proposons d'affecter la même courbe à certains acides aminés qui sont considérés comme ayant le même comportement d'insertion dans la membrane (voir section ci-dessous).

Nous décidons aussi d'examiner dans cette méthode, une nouvelle forme de courbe. [Chamberlain *et al.*, 2004; Park and Helms, 2008b] suggèrent que l'influence des acides aminés est asymétrique dans la séquence. En effet, certains acides aminés sont beaucoup plus présents sur une extrémité de la membrane et en même temps présentent une plus grande influence pour l'insertion des segments dans la membrane. Dans notre cas, nous supposons que le profil des acides aminés prend une forme de courbe en escalier (figure 4.1 (C)) défini par 2 valeurs : une pour le côté lumen et l'autre pour le côté du cytosol.

### 4.4.2 Contraintes sur les acides aminés

Plusieurs travaux ont été réalisés pour étudier les profils de distribution des acides aminés en calculant la fréquence d'apparition des acides aminés dans les protéines membranaires [Ulmschneider and Sansom, 2001; Nilsson *et al.*, 2005] et les énergies d'insertion des différents acides aminés au long de la membrane [Ulmschneider *et al.*, 2005b; Johansson and Lindahl, 2008]. L'objectif de ces méthodes est d'identifier des caractéristiques communes et d'en extraire une échelle pour les acides aminés afin de prédire les protéines membranaires [Pilpel *et al.*, 1999; Pellegrini-Calace *et al.*, 2003].

Ces travaux montrent que certains acides aminés ont des préférences distinctes pour différentes régions de la membrane [Ulmschneider *et al.*, 2005a; Senes *et al.*, 2007]. Le profil de distribution montre que chaque acide aminé est plus présent en une région particulière de la membrane (interface ou milieu). Selon la position de l'acide aminé au milieu ou sur les interfaces de la membrane, il contribue favorablement ou défavorablement à l'insertion des protéines. Par exemple, nous pouvons constater que les acides aminés tels que la tyrosine (Y) et le tryptophane (W) ont une fréquence d'apparition plus grande sur chaque extrémité de la membrane. On constate aussi que les acides aminés tels que l'aspartique (D) et le glutamique (E) ou la cystéine (C) et la méthionine (M) présentent la même distribution avec un comportement assez similaire lors de l'insertion des protéines [Johansson and Lindahl, 2008; von Heijne, 2006; MacCallum *et al.*, 2008].

Dans notre méthode, nous décidons d'affecter la même courbe aux acides aminés qui présentent un même profil de distribution et un même comportement dans la membrane. Par conséquent, nous attribuons la courbe de l'acide aminé tyrosine (Y) à l'acide aminé tryptophane (W), la courbe de l'acide aminé aspartique (D) à l'acide aminé glutamique (E), et la courbe de l'acide aminé cystéine (C) à l'acide aminé méthionine (M). Les expériences que nous avons menèes montrent que la représentation d'autres acides aminés avec une même courbe dégrade les performances de la méthode. Par conséquent, nous nous limitons à ces trois représentations.

### 4.4.3 Fonction d'évaluation

Dans les systèmes précédents, nous avons utilisé l'AUC comme mesure d'évaluation. L'AUC évalue la qualité globale du classifieur à discriminer les PS des TM, mais nous avons constaté que plusieurs voisins présentent la même valeur de l'AUC.

Nous proposons d'utiliser dans BioGLS le taux de bonne prédiction (*accuracy*) comme fonction d'évaluation et qui sera notée *ACC*. L'utilisation du taux de bonne prédiction nous oblige à choisir un seuil de discrimination. Celui-ci sera déterminé en faisant varier le seuil tout en calculant la valeur de l'ACC. Le seuil final correspond au meilleur score présentant la plus grande valeur du taux de bonne prédiction.

### 4.4.4 Voisinage à $K$ dimensions

Nous avons pu constater que le voisinage représente une importante caractéristique d'un algorithme de recherche locale. Un voisinage à 1 dimension permet de choisir un seul acide aminé puis modifier sa courbe. Dans le cas d'un voisinage de $K$ dimensions, nous modifions $K$ acides aminés en même temps $N_K(s)$. Par conséquent, la taille du voisinage peut être représentée par le $(nombre\ de\ voisins)^K$, $K$ étant le nombre d'acides aminés dont nous allons changer la courbe. La taille de ce voisinage devient très grande dans le cas d'un nombre de voisins important et augmentant le temps de calcul.

### 4.4.5 Exploration du voisinage en deux temps

Comme l'algorithme MN-LS (section 4.3), l'algorithme que nous proposons explore le voisinage en deux étapes. La première étape consiste à apprendre des droites pour tous les acides aminés avec l'objectif de bien positionner la fenêtre de calcul de la fonction d'insertion (équation 3.2). Ensuite, la deuxième étape consiste à ajuster des courbes en escalier pour tous les acides aminés dans l'objectif d'affiner la discrimination entre le PS et le segment TM.

Comme indiqué auparavant, l'utilisation d'un voisinage élargi conduit à un temps de calcul très important. En effet, le temps de calcul est passé de minutes à des heures. Afin de minimiser le temps de la recherche, nous proposons :

- de réduire le nombre de modifications qui peuvent être appliquées sur les courbes.
- de diviser la recherche sur 3 groupes d'acides aminés.

**Réduire le nombre de modifications :** Comme chaque droite est définie par le paramètre $H_{milieu}$, nous considérons seulement le voisinage suivant $V1 \in \{0, \epsilon, -\epsilon\}$ avec une taille $t = (nombre\ de\ voisins\ de\ V1)^K = (3)^K$. Le premier voisin signifie que la courbe reste elle même (pas de modification), par contre le deuxième voisin consiste à déplacer la droite vers le haut d'une valeur égale à $\epsilon$. Le dernier voisin consiste à déplacer la droite vers le bas avec une valeur égale à $\epsilon$. Des grandes valeurs pour $\epsilon$ conduisent à faire bouger les droites avec d'importants changements, par contre des petites valeurs pour $\epsilon$ font bouger la droite avec de légers changements.

Dans le cas d'une courbe en escalier, nous considérons le voisinage $V2 \in \{(0,0), (\epsilon, -\epsilon), (-\epsilon, \epsilon)\}$ avec une taille $t = (nombre\ de\ voisins\ de\ V2)^K = (3)^K$. Le voisin (0.0) veut dire

que la courbe ne change pas de forme, par contre les deux autres voisins modifient la forme de la courbe.

**Diviser les acides aminés en groupe :** La deuxième solution que nous proposons est de diviser les 20 acides aminés en 3 groupes. Le premier groupe (groupe $1 = \{F, I, L, V, Y, W\}$) est principalement constitué des acides aminés les plus hydrophobes, tandis que le deuxième groupe est essentiellement constitué d'acides aminés polaires (groupe $2 = \{C, S, M, N, P, Q\}$). Le dernier groupe est constitué des acides aminés aromatiques et chargés (groupe $3 = \{A, T, D, E, R, G, H, K\}$). Nous apprenons des courbes symétriques pour les 20 acides aminés.

En attribuant la même courbe pour certains acides aminés, l'espace de recherche du groupe 1 est $\zeta = \{F, I, L, V, Y\}$ avec la contrainte que $V[Y] = V[W]$, ainsi la taille du voisinage $t_1$ est égale à $(3)^5$. Pour le groupe 2, l'espace de recherche est $\zeta = \{S, N, P, Q, C\}$ avec la contrainte que $V[C] = V[M]$, et une taille du voisinage $t_2$ égal à $(3)^5$. Le dernier groupe 3 est représenté par un espace de recherche $\zeta = \{A, T, R, G, H, K, D\}$ avec la contrainte $V[D] = V[E]$ et une taille du voisinage $t_3$ égal à $(3)^7$.

L'algorithme commence son exploration en examinant de manière séquentielle le groupe 1, puis le groupe 2, et enfin le groupe 3. L'ordre dont lequel les groupes sont optimisés est détermié par expérimentations. Dans ce premier passage, BioGLS optimise les droites des acides aminés. Lorsque cette étape est terminée, l'algorithme passe à la phase d'optimisation des courbes symétrique.

Nous appliquons aussi dans BioGLS l'opérateur de perturbation décrit en section 4.2.7 afin d'éviter à l'algorithme d'être retenu dans un optimum local.

Rappelons que nous optimisons des courbes en escalier pour tous les acides aminés.

### 4.4.6 Expérimentations et discussions

Cette section présente les résultats obtenus sur les différents jeux de tests en appliquant BioGLS. Nous initions notre méthode par les valeurs d'hydrophobie de Kyte et Doolittle.

Le tableau 4.8 montre les résultats de 5 exécutions en présentant les valeurs de l'AUC et l'*accuracy*. Nous présentons aussi les résultats de la méthode MN-LS à titre de comparaison.

Nous constatons que les différentes exécutions présentent des résultats similaires. De même que nous obtenons des scores de l'AUC similaire aux résultats de la méthode MN-LS avec un léger avantage pour la méthode MN-LS en terme du taux de bonne classification.

Le tableau 4.9 présente une comparaison entre un classifieur de la méthode BioGLS et un classifieur de la méthode MN-LS. Nous constatons que nous obtenons un meilleur score de l'AUC et du taux de prédiction avec le classifieur BioGLS sur le jeu SWP. Par contre, pour les autres jeux de tests, nous obtenons la méme AUC avec un léger meilleur score en terme du taux de prédiction en faveur de MN-LS. En d'autres mots, les deux méthodes sont comparables en termes de performances.

La figure 4.6 présente les courbes en escalier correspondant à un classifieur de la méthode BioGLS. La valeur 1 de la courbe correspond au côté cytosol, tandis que la valeur

| | exe1 | | exe2 | | exe3 | | exe4 | | exe5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| fold 1 | 0.917 | 0.845 | 0.911 | 0.967 | 0.910 | 0.816 | 0.910 | 0.838 | 0.953 | 0.889 |
| fold 2 | 0.926 | 0.860 | 0.939 | 0.875 | 0.926 | 0.852 | 0.914 | 0.823 | 0.923 | 0.852 |
| fold 3 | 0.920 | 0.867 | 0.914 | 0.860 | 0.943 | 0.852 | 0.942 | 0.860 | 0.928 | 0.838 |
| fold 4 | 0.958 | 0.889 | 0.927 | 0.845 | 0.955 | 0.882 | 0.939 | 0.845 | 0.914 | 0.830 |
| fold 5 | 0.957 | 0.882 | 0.956 | 0.875 | 0.920 | 0.867 | 0.927 | 0.867 | 0.935 | 0.875 |
| fold 6 | 0.915 | 0.838 | 0.908 | 0.838 | 0.930 | 0.882 | 0.939 | 0.875 | 0.961 | 0.882 |
| fold 7 | 0.951 | 0.882 | 0.957 | 0.867 | 0.961 | 0.904 | 0.936 | 0.867 | 0.951 | 0.882 |
| fold 8 | 0.923 | 0.845 | 0.913 | 0.852 | 0.953 | 0.882 | 0.940 | 0.867 | 0.917 | 0.860 |
| fold 9 | 0.958 | 0.882 | **0.972** | **0.911** | 0.960 | 0.860 | 0.967 | 0.919 | 0.942 | 0.867 |
| fold 10 | 0.928 | 0.860 | 0.921 | 0.882 | 0.969 | 0.882 | 0.969 | 0.911 | 0.926 | 0.860 |
| BioGLS moyenne | 0.935 | 0.865 | 0.932 | 0.867 | 0.943 | 0.868 | 0.938 | 0.867 | 0.935 | 0.863 |
| écart type | 0.018 | 0.018 | 0.005 | 0.020 | 0.019 | 0.024 | 0.018 | 0.029 | 0.016 | 0.019 |
| MN-LS moyenne | 0.938 | 0.870 | 0.939 | 0.877 | 0.936 | 0.873 | 0.938 | 0.872 | 0.938 | 0.871 |
| écart type | 0.020 | 0.029 | 0.019 | 0.029 | 0.019 | 0.027 | 0.020 | 0.026 | 0.020 | 0.025 |

TABLE 4.8 – Résultats de 5 exécutions de BioGLS sur le jeu SWP, ainsi que les résultats de l'évaluation de la méthode MN-LS (voir tableau 4.5).

| | MN-LS | | BioGLS | |
|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy |
| SWP Test | 0.969 | 0.891 | 0.972 | 0.911 |
| SCAMPI HighFirst | 0.851 | 0.841 | 0.851 | 0.826 |
| SCAMPI HighALL | 0.829 | 0.747 | 0.850 | 0.752 |
| SCAMPI LowFirst | 0.878 | 0.831 | 0.885 | 0.836 |
| SCAMPI LowAll | 0.857 | 0.770 | 0.850 | 0.753 |
| PDB First | 0.794 | 0.782 | 0.786 | 0.771 |
| PDB All | 0.745 | 0.656 | 0.739 | 0.649 |

TABLE 4.9 – Comparaison entre les résultats obtenus avec la méthode MN-LS (tableau 4.6) et la méthode BioGLS (classifieur du fold 9 de la deuxième expérimentation).

21 de la courbe correspond au côté lumière de membrane. Nous constatons par exemple que les acides aminés comme le tryptophane (W) et la tyrosine (Y) favorisent l'insertion des protéines côté lumière de la membrane. Inversement, les acides aminés comme l'histidine (H) et l'arginine (R) favorisent l'insertion des protéines dans la membrane côté cytosol. Pour les acides aminés ayant simplement une droite comme l'acide aminé glutamine (S) signifie que cet acide aminé présente la même contribution d'insertion sur les 2 côtés de la membrane.

Contrairement à la méthode MN-LS qui optimise les courbes des acides aminés en modifiant la courbe d'un seul acide aminé, la méthode BioGLS optimise des courbes pour les acides aminés en prenant en considération la modification de plusieurs courbes d'acides aminés en même temps. Cette stratégie permet à BioGLS de prendre en considération l'influence du changement d'une courbe d'un acide aminé par rapport au changement d'une courbe d'un autre acide aminé, puisque les modifications des courbes s'effectuent en groupe. Les courbes apprises avec la méthode BioGLS ont donc l'avantage d'être plus

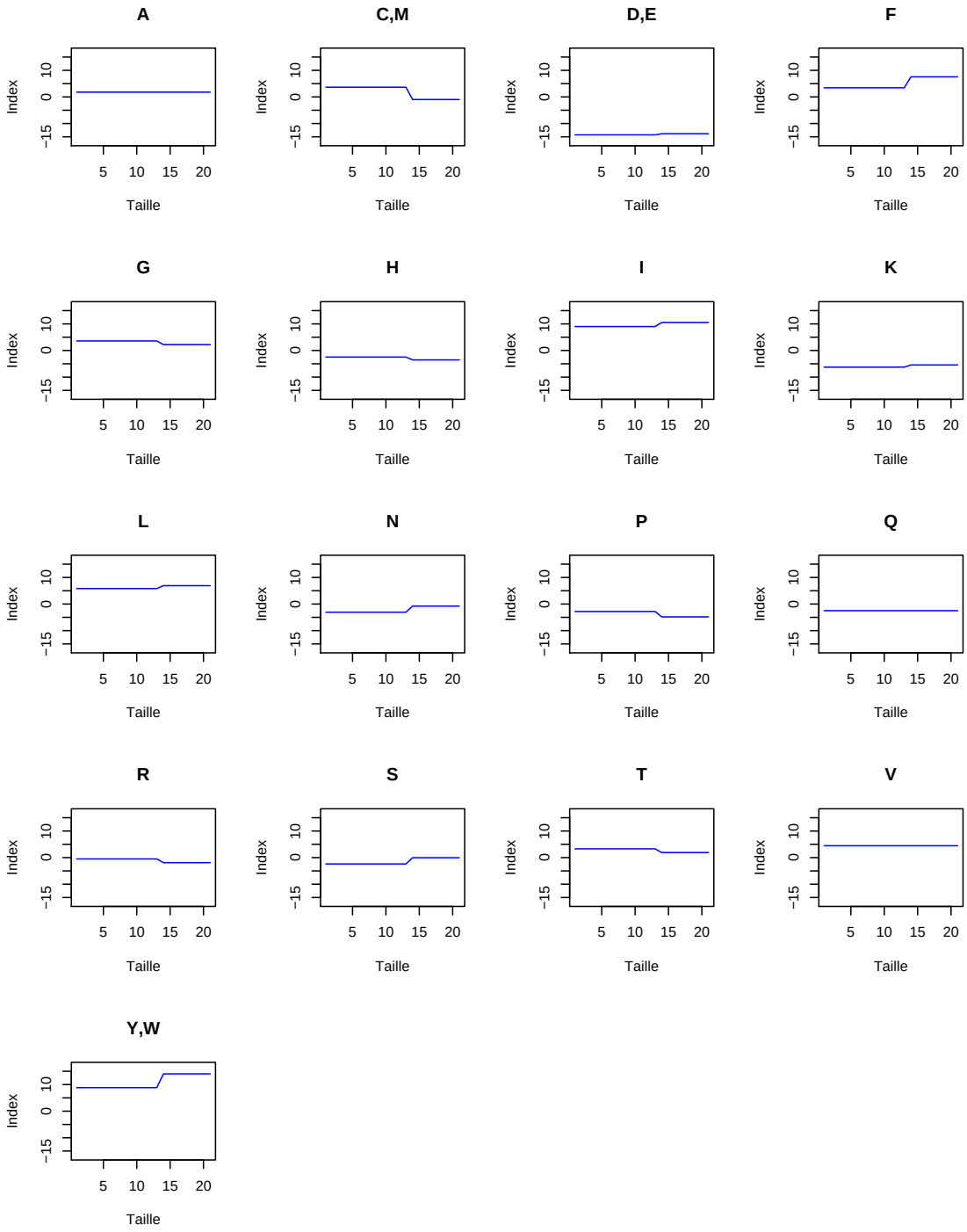


FIGURE 4.6 – Courbe d'insertion des acides aminés apprises avec BioGLS ayant une forme de courbe en escalier.

précises et de ne plus dépendre des valeurs de la configuration initiale.

### 4.4.7 Synthèse

Nous avons introduit dans cette section, une variante des deux premiers algorithmes. Tout d'abord, BioGLS explore un voisinage de $K$ dimensions qu'il lui permet de modifier la courbe de plusieurs acides aminés en même temps. Cette stratégie permet à notre algorithme d'examiner un plus grand nombre de solutions afin d'améliorer la solution courante et au même temps d'avoir d'important changements sur la solution initiale dans l'objectif de ne plus dépendre de ces valeurs initiaux. L'évaluation des performances de la méthode montre des bons résultats de discrimination des séquences avec PS et des séquences avec segment TM.

La méthode intégre de nouvelles observations biologiques en attribuant à certains acides aminés la même courbe d'insertion. En effet, les acides aminés présentant un même profil de distribution dans la protéine ainsi qu'un même comportement seront défini pas une même courbe.

Nous avons aussi exploré avec BioGLS une nouvelle forme de courbe à savoir une courbe en escalier pour les acides aminés. L'idée est que les acides aminés n'emploient pas la même énergie d'insertion sur les extrémités de la membrane. Nous pensons que cette nouvelle forme de courbe permet de mieux capturer l'information à partir des données pour mieux expliquer la contribution des acides aminés.

## 4.5 Conclusions

Dans ce chapitre, nous avons présenté trois algorithmes pour l'optimisation des courbes d'insertion des acides aminés. La plus importante contribution de ces algorithmes est l'intégration de connaissances biologiques fondées sur le potentiel de contribution des acides aminés pour discriminer deux séquences partageant une même propriété physico-chimique.

Ces algorithmes utilisent une stratégie de recherche locale en intégrant des mécanismes tels que la perturbation ou le voisinage multiple pour diversifier la recherche. L'évaluation de ces méthodes sur les différents jeux de tests montre de bonnes performances de discrimination entre les séquences PS et les séquences TM.

Nous avons aussi examiné différentes formes de courbes : des courbes symétriques pour les acides aminés dans la méthode LSTranslocon, une droite pour certains acides aminés et une courbe symétrique pour d'autres acides aminés dans la méthode MN-LS, et des courbes en escalier dans la méthode BioGLS.

# Chapitre 5

# Développement d'une approche avec un algorithme à population

La conception d'un algorithme performant nécessite un compromis entre l'intensification et la diversification pour explorer l'espace de recherche. Le second point « diversification » est l'objet de ce chapitre en présentant une adaptation d'un algorithme génétique pour la détermination des courbes d'insertion des acides aminés. L'algorithme génétique permet de diversifier la recherche en vue de déterminer les courbes des acides aminés qui présentant les meilleures performances pour notre méthode de classification.

**Sommaire**

## 5.1 Principe d'un algorithme génétique

Les algorithmes génétiques (GA) sont des méthodes d'optimisation inspirées par des mécanismes d'évolution de la nature. On parle généralement d'algorithmes évolutionnaires impliquant une population. Ces algorithmes ont été introduits la premiére fois de manière formelle par John Holland [Holland, 1962], et vulgarisé par l'ouvrage de David Goldberg [Goldberg, 1989].

L'utilisation d'un algorithme génétique nécessite la définition de plusieurs éléments :
- Le codage des individus de la population qui consiste à représenter chaque solution par une structure de données.
- La génération de la population initiale $P$ avec la contrainte de produire une population d'individus non homogéne qui servira de base pour les générations futures.
- La fonction d'évaluation $f$ (appelée fitness) qui consiste à évaluer un individu de la population et lui associer une valeur. La fonction d'évaluation permet de conserver les individus avec une bonne performance dans la population.
- L'utilisation d'opérateurs qui permettent d'apporter des modifications sur la population. En général, on peut trouver deux types d'opérateurs différents : les opérateurs de mutation et ceux de croisement. Les opérateurs de mutation sont des opérateurs qui prennent un individu et le modifient afin de diversifier la population. Par contre, les opérateurs de croisement prennent deux individus et génèrent un ou plusieurs nouvel individu qui contient idéalement le meilleur des parents. Ces différents opérateurs sont souvent appliqués selon une probabilité d'application.
- Les paramètres de l'algorithme génétique comme la taille de la population, le nombre total de générations ou le critère d'arrêt.

De manière générale, un algorithme génétique (AG) suit une procédure qui commence par générer aléatoirement une population initiale de $n$ individus. Le passage d'une génération $k$ à la génération $k+1$ implique une phase de sélection qui consiste à choisir sur la population $k$ des individus pour construire la génération $k+1$. Généralement, les individus sont sélectionnés selon leur fitness, puis un opérateur de croisement leur est appliqué pour générer des enfants. D'autres individus de la population sont choisis pour subir une mutation. Tous les nouveaux enfants ayant subi une mutation ou un croisement peuvent être évalués avant d'être insérés dans la nouvelle population. On réitère ensuite le procédé à partir de cette population jusqu'à obtenir une solution que l'on juge satisfaisante ou lorsqu'un critère d'arrêt de l'algorithme est satisfait. Très souvent, on utilise le nombre de générations qui peut être fixé à priori.

## 5.2 Adaptation de l'algorithme génétique au problème de discrimination du PS et du segment TM

Notre algorithme suit le schéma classique d'un algorithme génétique en faisant évoluer une population d'individus (figure 5.1). Au début de l'algorithme, une population de taille

$n$ à chaque individu représente un ensemble de 20 courbes, est initialisée aléatoirement. La population évolue au cours des générations en appliquant à chaque génération un processus de sélection. Les individus ayant une bonne qualité de discrimination sont reproduits, tandis que les moins bons sont supprimés. Les opérateurs de croisement et de mutation sont appliqués ensuite sur la population sélectionnée. Ce processus est répété jusqu'à ce qu'un nombre prédéfini de générations soit atteint, ou qu'un autre critère (qualité d'une solution) soit réalisé.

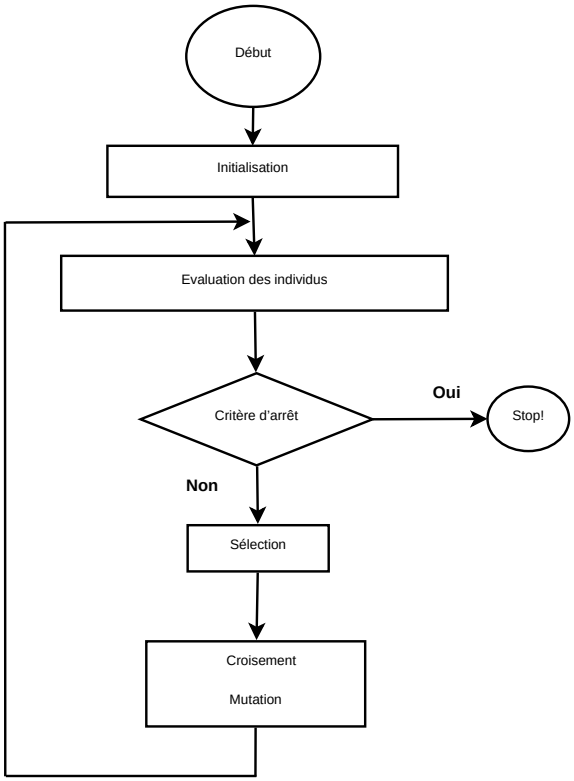


FIGURE 5.1 – Schéma classique d'un algorithme génétique.

### 5.2.1 Représentation d'un individu

Dans notre algorithme génétique, un individu est représenté par un ensemble de 20 courbes où chaque courbe définit la contribution d'un acide aminé. Afin de générer une courbe, nous utilisons la fonction 4.1 décrite en section 4.2.2 du chapitre 4. Rappelons dans le cas d'une courbe de forme symétrique, la fonction est définie par le couple $(H_{extremite}, H_{milieu})$, alors que dans le cas d'une droite la fonction est définie seulement par le paramètre $H_{milieu}$.

### 5.2.2 Population initiale

La population initiale peut influencer la convergence de l'algorithme. Dans notre cas, nous décidons de générer une population initiale à partir de modifications apportées sur une échelle d'hydrophobie et choisissons d'utiliser l'échelle d'hydrophobie de Kyte et Doolittle.

Dans un premier temps, nous initions un nombre aléatoire entre 1 et 20 qui nous permet de choisir le nombre d'acides aminés pour lesquelles nous modifions la courbe. Ensuite, nous modifions les valeurs de ces acides aminés par l'ajout d'un couple $(\Delta_{Hmid}, \Delta_{Hext})$ de valeurs réelles sélectionnées aléatoirement entre $[-3, 3]$. Ce processus est répété pour générer le nombre d'individus nécessaires dans la population initiale.

Dans le cas de l'ajustement de droites pour les acides aminés, nous suivons la même démarche décrite dans cette section, cependant nous modifions seulement la valeur du $H_{milieu}$ par l'ajout d'un $\Delta_{Hmid}$ sélectionné aléatoirement entre $[-3, 3]$.

### 5.2.3 Opérateur de sélection

L'opérateur de sélection a la charge de définir quels seront les individus de la population $P$ qui vont être transmis dans la nouvelle population $P'$ afin de servir de parents pour l'application de l'opérateur de croisement. L'algorithme génétique que nous avons développé combine deux types de méthodes de sélection :

1. Sélection par élitisme : cette méthode consiste à sélectionner $m$ individus à partir de la population $P$ aprés l'avoir triée de manière décroissante selon la fonction d'évaluation. Ces individus seront copiés directement dans la prochaine population $P'$. Cette méthode permet d'améliorer considérablement l'algorithme génétique en s'assurant d'avoir toujours les meilleurs individus dans la population suivante.
2. Sélection par roulette : dans cette méthode les parents sont sélectionnés en fonction de leur performance. L'idée est d'attribuer à chaque individu une zone proportionnelle à sa « fitness ». Cette méthode ressemble à une sorte de roulette de casino sur laquelle sont placés tous les individus constituant la population. On fait tourner la roulette et quand elle s'arréte sur une zone, on sélectionne l'individu correspondant à cette zone.

Dans notre algorithme, nous appliquons un principe d'élitisme pour recopier 10% des meilleurs individus dans la population suivante. Les autres individus de la nouvelle population (90%) sont générés par croisement entre 2 parents sélectionnés dans la population courante en suivant le principe de la roulette.

Afin d'évaluer la qualité des individus, nous utilisons le taux de bonne prédiction comme fonction d'évaluation (présenté en section 3.4.2),

$$taux\ de\ bonne\ prédiction = \frac{VN + VP}{VN + FP + VP + FN}$$

Le taux de bonne prédiction permettra de mesurer la capacité d'un individu à discriminer entre les séquences PS et les séquences TM.

### 5.2.4 Opérateur de croisement

L'opérateur de croisement que nous employons consiste tout simplement à réaliser un échange de courbes entre les deux parents (voir figure 5.2) pour générer deux nouveaux individus. Nous sélectionnons aléatoirement des acides aminés entre un nombre 1 et 15 (interval détermié expérimentalement) dans un premier parent, sur ces acides aminés nous effectuons un changement de courbe avec les courbes des mêmes acides aminés dans un deuxième parent. En d'autres termes, nous échangeons la courbe d'un acide aminé par la courbe d'un autre acide aminé dans l'objectif d'obtenir un nouveau jeu de courbe.

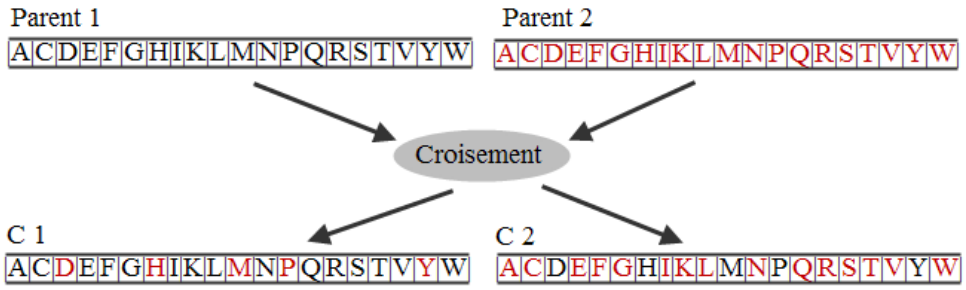


FIGURE 5.2 – Opérateur de croisement entre deux parents en modifiant les courbes des acides aminés aspartique (D), histidine (H), méthionine (M), proline (P), tyrosine (Y).

### 5.2.5 Opérateur de mutation

L'opérateur de mutation que nous appliquons consiste à apporter des modifications à un individu (voir figure 5.3) en choisissant aléatoirement un acide aminé parmi les 20 acides aminés et nous modifions sa courbe par l'ajout d'un couple $(\Delta_{Hmid}, \Delta_{Hext})$ de valeur réelle sélectionnée aléatoirement entre $[-2, 2]$. Ce processus est répété selon un

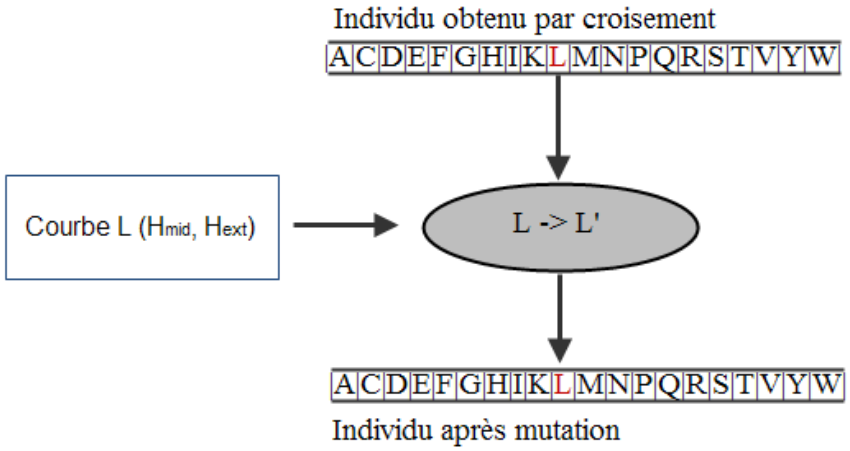


FIGURE 5.3 – Opérateur de mutation modifiant la courbe de l'acide aminé leucine (L).

nombre sélectionné aléatoirement entre 1 et 15 (interval détermié expérimentalement), en s'assurant à chaque fois de sélectionner un nouvel acide aminé.

En d'autres termes, nous apportons une modification de la courbe de l'acide aminé dans l'objectif de tester une nouvelle courbe. L'opérateur de mutation est appliqué toutes les 10 générations sur les 90% des individus ayant subi un croisement.

### 5.2.6 Expérimentations et discussions

Dans cette section, nous présentons les expérimentations effectuées en utilisant l'algorithme génétique sur les différents jeux de données. Concernant la taille de la population et le nombre d'itérations, ils sont fixés expérimentalement à 100 pour la taille de la population et 120 itérations pour l'évolution de la population.

Notre algorithme est lancé plusieurs fois, nous rapportons dans le tableau 5.1 les résultats de 5 exécutions de l'application de notre algorithme sur le jeu SWP. Le tableau 5.1 montre une stabilité entre les différentes exécutions avec une moyenne d'AUC égale 0.934 et un taux de bonnes prédictions égale 0.871.

| | exe1 | | exe2 | | exe3 | | exe4 | | exe5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| fold 1 | 0.902 | 0.848 | 0.921 | 0.839 | 0.908 | 0.838 | 0.908 | 0.836 | 0.905 | 0.836 |
| fold 2 | 0.935 | 0.867 | 0.940 | 0.878 | 0.923 | 0.862 | 0.934 | 0.875 | 0.931 | 0.911 |
| fold 3 | 0.923 | 0.875 | 0.931 | 0.867 | 0.931 | 0.852 | 0.933 | 0.882 | 0.924 | 0.845 |
| fold 4 | 0.932 | 0.862 | 0.917 | 0.860 | 0.948 | 0.877 | 0.947 | 0.852 | 0.948 | 0.852 |
| fold 5 | 0.950 | 0.911 | 0.923 | 0.875 | 0.956 | 0.929 | 0.962 | 0.914 | 0.962 | 0.895 |
| fold 6 | 0.898 | 0.886 | 0.965 | 0.848 | 0.894 | 0.848 | 0.892 | 0.816 | 0.907 | 0.843 |
| fold 7 | 0.950 | 0.855 | 0.907 | 0.849 | 0.953 | 0.882 | 0.948 | 0.882 | 0.958 | 0.919 |
| fold 8 | 0.904 | 0.847 | 0.948 | 0.889 | 0.913 | 0.840 | 0.878 | 0.838 | 0.909 | 0.840 |
| fold 9 | 0.964 | 0.897 | 0.890 | 0.838 | 0.971 | 0.914 | 0.938 | 0.873 | 0.956 | 0.889 |
| fold 10 | 0.942 | 0.853 | 0.954 | 0.867 | 0.949 | 0.867 | 0.952 | 0.875 | 0.903 | 0.870 |
| moyenne | 0.930 | 0.871 | 0.929 | 0.861 | 0.934 | 0.870 | 0.929 | 0.864 | 0.930 | 0.870 |
| écart type | 0.022 | 0.021 | 0.0005 | 0.017 | 0.024 | 0.030 | 0.027 | 0.028 | 0.023 | 0.031 |

TABLE 5.1 – Résultats de 5 exécutions sur le jeu SWP obtenus avec l'algorithme génétique.

Nous comparons notre algorithme génétique avec les résultats de l'échelle d'hydrophobie de Kyte et Doolittle ainsi que les résultats des deux méthodes BioGLS et MN-LS avec lesquelles nous avons obtenu les meilleurs résultats en utilisant une approche fondée sur un algorithme de recherche locale.

Le tableau 5.2 montre que notre méthode présente de bonnes performances en comparaison avec l'échelle d'hydrophobie de Kyte et Doolittle, en particulier sur les jeux de tests SCAMPI et PDB.

Le tableau 5.2 montre aussi que notre algorithme génétique présente des résultats similaires aux résultats obtenus avec les algorithmes MN-LS et BioGLS sur les différents jeux de tests. Le principal avantage de l'algorithme génétique est l'exploration d'un plus grand espace de recherche afin déterminer la solution qui présente le meilleur taux de discrimination entre le PS et le segment TM.

La figure 5.4 et la figure 5.5 montrent la répartition des observations du jeu SWP. La figure 5.5 montre la répartition en utilisant des droites prenant les valeurs de l'échelle de Kyte et Doolittle. Nous observons une trés grande zone d'intersection entre le données PS et les données TM. Par contre, en utilisant les courbes apprises avec l'algorithme

| | échelle Kyte et Doolittle | | MN-LS | | BioGLS | | GA | |
|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| SWP Test | 0.826 | 0.747 | 0.969 | 0.891 | 0.972 | 0.911 | 0.956 | 0.924 |
| SCAMPI HighFirst | 0.684 | 0.764 | 0.851 | 0.841 | 0.851 | 0.826 | 0.881 | 0.811 |
| SCAMPI HighALL | 0.645 | 0.593 | 0.829 | 0.747 | 0.850 | 0.752 | 0.868 | 0.764 |
| SCAMPI LowFirst | 0.765 | 0.788 | 0.878 | 0.831 | 0.885 | 0.836 | 0.889 | 0.814 |
| SCAMPI LowAll | 0.675 | 0.564 | 0.857 | 0.770 | 0.850 | 0.753 | 0.854 | 0.734 |
| PDB First | 0.642 | 0.702 | 0.794 | 0.782 | 0.786 | 0.771 | 0.801 | 0.758 |
| PDB All | 0.580 | 0.516 | 0.745 | 0.656 | 0.739 | 0.649 | 0.780 | 0.668 |

TABLE 5.2 – Comparaison entre l'algorithme génétique, échelle Kyte et Doolittle, MN-LS et BioGLS.

génétique nous réduisons la zone d'intersection (figure 5.5). Cela montre bien que l'algorithme génétique sépare mieux les deux classes.

Nous effectuons une deuxième expérimentation, dans celle-ci nous apprenons seulement des droites pour les acides aminés. Le tableau 5.3 montre une stabilité des résultats obtenus sur le jeu SWP. Nous obtenons une AUC égale à 0.910 (0.930 pour les courbes symétriques) et un taux de bonne prédiction égale à 0.840 (0.870 pour les courbes symétriques). En effet, l'ajustement des droites permet d'atteindre une très bonne qualité de discrimination entre le PS et le segment TM. Dans certains folds, nous obtenons des résultats au même niveau des meilleurs résultats obtenus en terme d'AUC et de taux de bonne prédiction pour l'algorithme génétique optimisant des courbes symétriques.

| | exe1 | | exe2 | | exe3 | | exe4 | | exe5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy | AUC | Accuracy |
| fold 1 | 0.900 | 0.823 | 0.895 | 0.838 | 0.896 | 0.816 | 0.905 | 0.816 | 0.896 | 0.830 |
| fold 2 | 0.939 | 0.875 | 0.910 | 0.875 | 0.934 | 0.875 | 0.939 | 0.875 | 0.936 | 0.867 |
| fold 3 | 0.910 | 0.816 | 0.916 | 0.838 | 0.915 | 0.857 | 0.922 | 0.838 | 0.921 | 0.867 |
| fold 4 | 0.924 | 0.838 | 0.920 | 0.816 | 0.910 | 0.860 | 0.910 | 0.838 | 0.929 | 0.860 |
| fold 5 | 0.835 | 0.875 | 0.912 | 0.860 | 0.936 | 0.845 | 0.936 | 0.823 | 0.931 | 0.975 |
| fold 6 | 0.878 | 0.794 | 0.890 | 0.814 | 0.870 | 0.799 | 0.919 | 0.838 | 0.877 | 0.801 |
| fold 7 | 0.939 | 0.845 | 0.937 | 0.860 | 0.920 | 0.872 | 0.879 | 0.816 | 0.935 | 0.860 |
| fold 8 | 0.916 | 0.845 | 0.928 | 0.860 | 0.902 | 0.845 | 0.919 | 0.860 | 0.925 | 0.860 |
| fold 9 | 0.928 | 0.838 | **0.933** | **0.884** | 0.926 | 0.845 | 0.935 | 0.882 | 0.924 | 0.875 |
| fold 10 | 0.920 | 0.860 | 0.921 | 0.852 | 0.913 | 0.867 | 0.923 | 0.897 | 0.910 | 0.875 |
| moyenne droites | 0.912 | 0.841 | 0.917 | 0.849 | 0.910 | 0.848 | 0.914 | 0.848 | 0.916 | 0.857 |
| écart type droites | 0.020 | 0.925 | 0.020 | 0.030 | 0.026 | 0.030 | 0.020 | 0.028 | 0.022 | 0.023 |
| moyenne symétrique | 0.930 | 0.871 | 0.929 | 0.861 | 0.934 | 0.870 | 0.929 | 0.864 | 0.930 | 0.870 |
| écart typesymétrique | 0.022 | 0.021 | 0.0005 | 0.017 | 0.024 | 0.030 | 0.027 | 0.028 | 0.023 | 0.031 |

TABLE 5.3 – Résultats de 5 exécutions en utilisant l'algorithme génétique qui apprend des droites pour les acides aminés.

Le tableau 5.4 montre une comparaison entre un classifieur des courbes droites et un classifieur des courbes symétriques. Nous remarquons que les résultats sur les différents jeux de tests sont similaires. Nous pouvons supposer que dans certains cas l'ajustement des droites et des courbes symétriques sont équivalant en termes de performance de discrimination. Cela ne veut pas dire que l'apprentissage des droites est suffisant, car la moyenne du taux de prédiction (0.840) des droites reste inférieure à la moyenne du taux de prédiction (0.870) des courbes symétriques. Mais, nous pouvons dire que l'ajustement

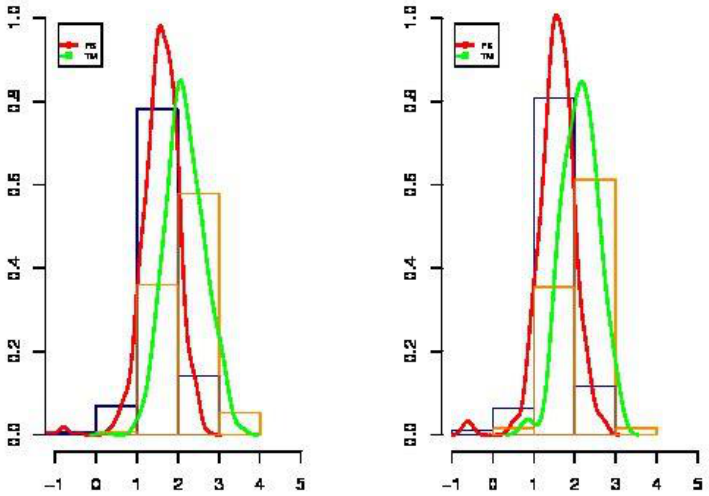


FIGURE 5.4 – Profil de répartition des données en utilisant les courbes de la configuration initiale. La figure à gauche représente la répartition avec les données d'apprentissage, tandis que la figure à droite repésente la répartition des données de test. La courbe rouge présente la répartition des PS et le courbe verte présente la répartition de TM. En bleu l'histogramme de distribution des données PS et en orange l'histogramme de distribution des données TM.

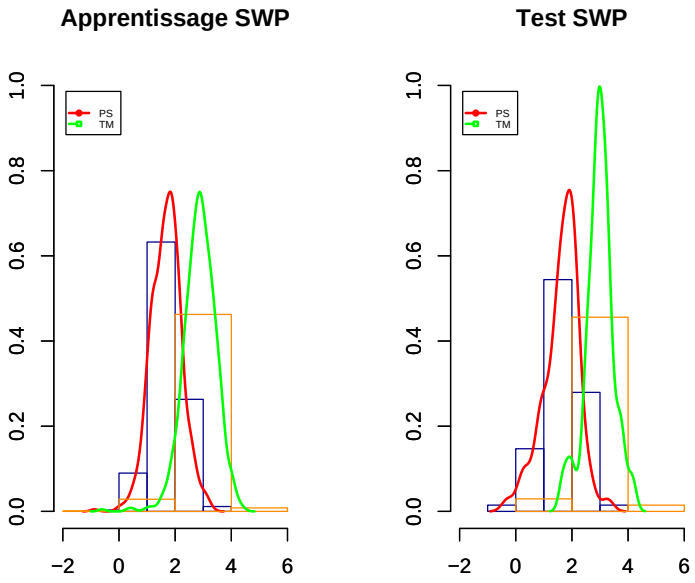


FIGURE 5.5 – Profil de répartition des données en utilisant les courbes apprises avec l'algorithme génétique. La figure à gauche représente la répartition avec les données d'apprentissage, tandis que la figure à droite repésente la répartition des données de test. La courbe rouge présente la répartition des PS et le courbe verte présente la répartition de TM. En bleu l'histogramme de distribution des données PS et en orange l'histogramme de distribution des données TM.

des droites permet une bonne discrimination entre le PS et le segment TM.

| | GA Symétriques | | GA Drtoies | |
|---|---|---|---|---|
| | AUC | Accuracy | AUC | Accuracy |
| SWP Test | 0.956 | 0.924 | 0.933 | 0.884 |
| SCAMPI HighFirst | 0.881 | 0.811 | 0.867 | 0.820 |
| SCAMPI HighALL | 0.868 | 0.764 | 0.856 | 0.760 |
| SCAMPI LowFirst | 0.889 | 0.814 | 0.870 | 0.802 |
| SCAMPI LowAll | 0.854 | 0.734 | 0.853 | 0.751 |
| PDB First | 0.801 | 0.758 | 0.799 | 0.780 |
| PDB All | 0.780 | 0.668 | 0.769 | 0.678 |

TABLE 5.4 – Comparaison entre les résultats des courbes symétriques et les résultats de droites.

Nous présentons dans la figure 5.6 des courbes d'insertion apprises avec l'algorithme génétique. Rappelons que notre population initiale est le résultat de modification apportée sur l'échelle d'hydrophobie de Kyte et Doolittle. L'algorithme génétique applique l'opérateur de croisement et de mutation pour un certain nombre de générations afin d'améliorer la population initiale. Ceci permet de changer les valeurs de la population initiale et d'apprendre de nouvelles valeurs qui discriminent au mieux les séquences PS des séquences TM. En effet, les courbes apprises avec l'algorithme génétique ne dépendent plus des valeurs initiales. L'algorithme génétique explore un espace de recherche large et diversifié menant à la détermination de courbes de bonne qualité de discrimination.

## 5.3 Conclusion

Nous avons présenté dans ce chapitre un algorithme génétique pour apprendre les courbes des 20 acides aminés. L'utilisation de l'algorithme génétique explore de manière plus diversifiée l'espace de recherche et en même temps permet d'améliorer les courbes d'insertion. L'évaluation de cette méthode sur les différents jeux de tests montre de bonnes performances de discrimination entre les séquences PS et les séquences TM.

Nous avons aussi examiné l'apprentissage de droites pour les 20 acides aminés en utilisant l'algorithme génétique. L'évaluation des droites sur les différents jeux de données a montré que dans certains cas les courbes symétriques et les droites sont équivalant en termes de performance de discrimination.

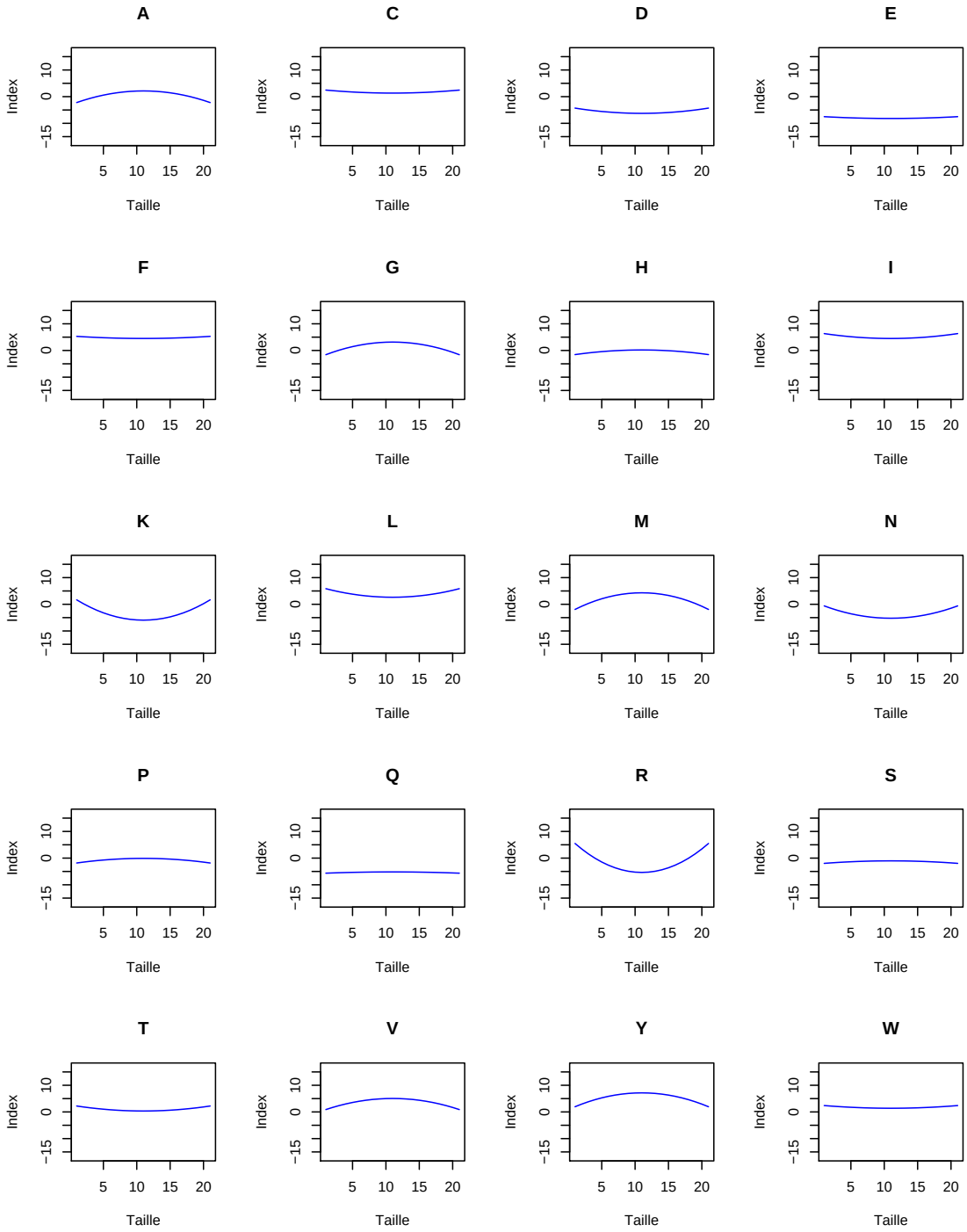


FIGURE 5.6 – Courbes d'insertion apprises avec l'algorithme génétique.

# Conclusion Générale

## Principales contributions

Dans cette thèse, nous nous sommes intéressés à la prédiction de la localisation des protéines qui transitent par la membrane du réticulum endoplasmique. Nos travaux se sont particulièrement dirigés vers la création d'une méthode permettant la discrimination entre un jeux de séquences de peptides signaux et un jeu de séquences de segments TM. Afin de réaliser cette distinction, nous nous appuyons sur des connaissances biologiques portant sur le mécanisme d'insertion des protéines dans la membrane du RE. La principale hypothèse prend en considération le potentiel d'insertion de chaque acide aminé dans la membrane.

Tout d'abord, nous avons développé la méthode LSTranslocon. Cette méthode utilise une approche par recherche locale et optimise les courbes d'insertion des acides aminés. Notre méthode a l'avantage d'intégrer les hypothèses d'insertion des protéines dans la membrane présentées par Hessa *et al.* De plus, elle offre la possibilité de distinguer entre deux séquences partageant la même propriété d'hydrophobie : séquence PS et séquence TM. Cependant, nous avons constaté une difficulté d'ajustement pour les courbes de certains acides aminés.

Par la suite, nous avons développé des stratégies visant à améliorer les performances de l'algorithme LSTranslocon, que ce soit en terme de qualité de distinction ou de stabilité des courbes. Ces améliorations sont présentées dans la méthode MN-LS. La méthode MN-LS restreint l'espace de recherche grâce à des informations pertinentes sur les acides aminés en associant une droite pour les acides aminés peu fréquents et une courbe symétrique pour les acides aminés fréquents dans le jeu de données. Cette stratégie assure à notre méthode une stabilité des courbes, puisque nous cherchons à optimiser un seul paramètre pour les droites et deux paramètres pour les courbes symétriques. Nous avons aussi utilisé un voisinage multiple dans l'objectif d'éviter à notre méthode d'être capturée dans un optimum local. La méthode MN-LS obtient sur les différents jeux de tests de meilleurs résultats de discrimination que la méthode LSTranslocon, aussi bien en terme de qualité des solutions que de stabilité des courbes.

Afin d'améliorer la localisation des segments TM de la recherche, nous avons défini un voisinage à $K$ dimensions où nous modifions en même temps les courbes de plusieurs acides aminés. Cette nouvelle stratégie est décrite dans un nouvel algorithme nommé BioGLS. Dans BioGLS, nous appliquons un opérateur de perturbation pour éviter à l'algorithme d'être retenu dans un optimum local. Nous avons aussi introduit une nouvelle forme de courbes en escalier. Les résultats de BioGLS montrent de bonnes performances de discrimination tout en modélisant au mieux la contribution des acides aminés lors de l'insertion des protéines membranaires.

Dans un second temps, nous avons conçu un algorithme génétique pour l'optimisation des courbes des acides aminés. Cet algorithme a pour point fort de ne pas dépendre des

valeurs de la solution initiale. En effet, les courbes apprises avec l'algorithme génétique bénéficient d'un plus large espace de recherche pour leur détermination.

Enfin, dans cette thèse, nous avons eu comme objectif principal la détermination des courbes d'insertion des acides aminés. Pour cela, nous avons examiné trois différentes formes de courbes : des courbes symétriques, des droites, des courbes en escalier. Nous notons qu'il faut être prudent dans l'hypothése que les acides aminés présentent un profil de courbe symétrique. Malgré les bonnes performances de discrimination entre les séquences PS et les séquences TM en utilisant les courbes symétriques, nous avons pu constater que même l'ajustement des droites pour les acides aminés présente une bonne qualité de discrimination.

**Travaux en cours :** Une tâche très importante lors de la discrimination entre le peptide signal et le segment TM est **la localisation de la position des segments TM** au sein de la protéine. Cette tâche consiste à mesurer la capacité des méthodes à situer les segments TM sur la séquence protéique.

L'hypothèse que nous avançons est que si les courbes d'insertion obtenues en utilisant une des méthodes (algorithme génétique ou recherche locale) permettent de distinguer entre les séquences PS et les séquences TM, alors ces mêmes courbes doivent permettre de reconnaitre la position des segments TM dans une protéine.

Pour évaluer cette hypothèse, nous suivons la démarche présentée par [Bernsel *et al.*, 2008]. En utilisant le seuil et les courbes obtenues avec une des méthodes d'optimisation que nous avons développées, nous commençons par faire glisser une fenêtre de taille fixe $l$ à travers toute la séquence protéique. Pour chaque fenêtre, nous calculons la moyenne d'insertion selon l'équation 3.1 présentée en section 3.4.1 du chapitre 3. Nous localisons la fenêtre ayant la plus grande valeur de la fonction d'insertion et en méme temps supérieure au seuil de discrimination. Nous masquons cette fenêtre de la séquence protéique. Puis, nous répétons le même processus pour la séquence protéique masquée jusqu'â ce qu'il ne reste plus de fenêtre avec une valeur de la fonction d'insertion supérieure au seuil de discrimination. Au final, toutes les fenêtres (séquences d'acides aminés) masquées sont considérées comme des segments TM. Les segments TM que nous obtenons auront tous une taille fixe égale à la taille de courbe d'un acide aminé, c'est-à-dire, une taille de 21 acides aminés. Notons que le système développé est principalement dédiéà la localisation du premier segment TM, car l'apprentissage des courbes est réalisé sur un jeu de données constitué seulement du premier segment TM.

L'évaluation de ce système sur le jeu de données PDB de 180 protéines présente une précision de 0.81% dans la localisation de la position du premier segment TM. On remarque que la stratégie employée dans ce système manque la localisation de certains premiers segments TM. Une recherche plus approfondie permettra d'améliorer la stratégie utilisée dans ce système de localisation.

Une extension logique de cette localisation est de pouvoir disposer d'un système qui reconnait la position de tous les segments TM composant une protéine.

## Perspectives de recherches

Nous avons présenté ici plusieurs méthodes pour l'optimisation des courbes des acides aminés et la résolution du problème de discrimination du peptide signal et du segment TM. Nos méthodes présentent des résultats satisfaisants. Cependant, il reste de nombreuses perspectives de recherche qui peuvent être envisagées pour une éventuelle amélioration de la discrimination des protèines membranaires et pour une meilleure détermination et ajustement des courbes.

**Proposition d'un algorithme mimétique :** Une des premères contributions qui peut être réalisé est l'hybridation de l'algorithme génétique avec une méthode de recherche locale. L'algorithme génétique s'attache à générer une population d'individus solutions, tandis que la recherche locale consiste à améliorer chaque individu de la population ou un groupe d'individus par l'exploration d'un voisinage. La recherche locale pourra être utilisée à la place de l'opérateur de mutation. L'avantage d'une telle approche est l'intensification dans la solution par l'utilisation de la recherche locale, tandis que l'algorithme génétique permet d'avoir une diversification dans la solution.

**Comparaison approfondie des différentes types de courbes :** Dans cette thèse, nous avons expérimenté plusieurs types de courbes : droites, symétriques, et courbes en escalier. Une comparaison approfondie entre ces différentes formes permettra une meilleure compréhension de la contribution des acides aminés lors de l'insertion des protéines dans la membrane. De plus, cette étude permettra de déterminer la meilleure forme de contribution de chacun des acides aminés.

**Enrichir le jeu de données :** Nous pourrions utiliser la parenté entre les protéines -l'homologie- pour enrichir notre jeu de données. L'homologie pourrait contribuer à l'ajout d'informations comme le profil des séquences afin d'améliorer la prédiction des segments TM des protéines. En effet, [Viklund and Elofsson, 2004] montrent que l'utilisation de séquences homologues améliore jusqu'â 10% la précision de la topologie des segments TM.

**Introduction de connaissances biologiques pour guider la recherche :** La sélection de voisins pour la recherche locale et le croisement pour l'algorithme génétique décrit dans cette thèse ont été développés de manière générale. On peut alors introduire des éléments spécifiques pour développer des opérateurs adaptés au problème de discrimination du PS et du segment TM. Par exemple, dans le cas de l'enrichissement du jeu de données par homologie, nous pouvons introduire des connaissances biologiques sur ces protéines pour construire un croisement spécifique afin qu'il exerce pleinement son rôle. Concernant la recherche locale, l'introduction de connaissances spécifiques permet effectivement de choisir un voisin parmi d'autres dans le cas où ces derniers présentent la même performance de discrimination.

# Liste des figures

# Listes des tables

# Références bibliographiques


[Aarts and Lenstra, 1997] cité page 54
E. Aarts and J. K. Lenstra, editors. *Local Search in Combinatorial Optimization*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1997.

[Ahmed *et al.*, 2010] cité page 21
R. Ahmed, H. Rangwala, and G. Karypis. Toptmh : Topology predictor for transmembrane alpha-helices. *J. Bioinformatics and Computational Biology*, 8(1) :39–57, 2010.

[Apweiler *et al.*, 2004] cité page 38
R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H.Z. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S.L. Yeh. UniProt : the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(Sp. Iss. SI) :D115–D119, 2004.

[Arkin and Brunger, 1998] IT. Arkin and AT. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochimica et Biophysica ACTA-Potein Structure and Molecular Enzymology*, 1429(1) :113–128, DEC 8 1998.

[Bagos *et al.*, 2004] cité page 21
P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, and S.J. Hamodrakas. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, 5, 2004.

[Bannai *et al.*, 2002] cité page 58
H. Bannai, Y. Tamada, O. Maruyama, and S. Nakai, K.and Miyano. Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics*, 18(2) :298–305, 2002.

[Bendtsen *et al.*, 2004] cité page 1, 20, 39
J.D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides : SignalP 3.0. *Journal of Molecular Biology*, 340(4) :783–795, 2004.

[Berman *et al.*, 2000] cité page 12, 13, 37, 42
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1) :235–242, 2000.

[Bernsel *et al.*, 2008] cité page 30, 30, 90
A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, and A. Elofsson. Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences of the Unites States of America*, 105(20) :7177–7181, 2008.

[Blobel and Dobberstein, 1975] cité page 12
G. Blobel and B. Dobberstein. Transfer of proteins across membranes .1. Presence of proteolytically processed and unprocessed nascent immunoglobulin light-chains on membrane-bound ribosomes of murine myeloma. *Journal of Cell Biology*, 67(3) :835–851, 1975.

[Chamberlain *et al.*, 2004] cité page 71
A. K. Chamberlain, Y. Lee, S. Kim, and Bowien J. U. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *Journal of Molecular Biology*, 339(2) :471 – 479, 2004.

[Chen *et al.*, 2002] cité page 1, 24
C.P. Chen, A. Kernytsky, and B. Rost. Transmembrane helix predictions revisited. *Protein Science*, 11(12) :2774–2791, DEC 2002.

[Cheng, 2010] cité page 1
Z. Cheng. Protein translocation through the Sec61/SecY channel. *Bioscience Reports*, 30(3) :201–207, JUN 2010.

[Corinna and Mohri, 2004] cité page 49
C. Corinna and M. Mohri. Auc optimization vs. error rate minimization. In *in Advances in Neural Information Processing Systems*. MIT Press, 2004.

[Cserzo *et al.*, 1997] cité page 22
M. Cserzo, E. Wallin, I. Simon, G. vonHeijne, and A. Elofsson. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins : the dense alignment surface method. *Protein Engineering*, 10(6) :673–676, 1997.

[Cuthbertson *et al.*, June 2005] cité page 13, 13, 19, 24
J. M. Cuthbertson, D. A. Doyle, and M. S.P. Sansom. Transmembrane helix prediction : a comparative evaluation and analysis. *Protein Engineering Design and Selection*, 18(6) :295–308, June 2005.

[Dalbey *et al.*, 1997] cité page 17
R.E. Dalbey, M.O. Lively, S. Bron, and J.M. VanDijl. The chemistry and enzymology of the type I signal peptidases. *Protein Science*, 6(6) :1129–1138, 1997.

[Efron, 1979] cité page 24
B. Efron. Bootstrap Methods : Another Look at the Jackknife. *The Annals of Statistics*, 7(1) :1–26, 1979.

[Egan, 1975] cité page 49
J. P. Egan. *Signal detection theory and ROC analysis*. Series in Cognition and Perception. Academic Press, New York, NY, 1975.

[Eisenberg *et al.*, 1982] cité page 68
D. Eisenberg, R.M. Weiss, and T.C. Terwilliger. The helical hydrophobic moment : a measure of the amphiphilicity of a helix. *Nature*, 299(5881) :371–374, 1982.

[Emanuelsson *et al.*, 2000] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 300(4) :1005–1016, 2000.

[Engelman *et al.*, 1986] cité page 68
D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. 1986.

[Fawcett, 2004] cité page 49
F. Fawcett. ROC Graphs : Notes and Practical Considerations for Researchers. Technical report, 2004.

[Ferri *et al.*, 2002] C. Ferri, P. Flach, and J. Hern ?ndez-Orallo. Learning decision trees using the area under the roc curve. In *Proceedings of the 19th International Conference on Machine Learning*, pages 139–146. Morgan Kaufmann, 2002.

[Garg *et al.*, 2005] cité page 22
A. Garg, M. Bhasin, and G.P.S. Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of Biological Chemistry*, 280(15) :14427–14432, 2005.

[Gasteiger *et al.*, 2005] cité page 21
E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch. Protein identification and analysis tools on the ExPASy server. In J. M. Walker, editor, *The Proteomics Protocols Handbook*, chapter 52, pages 571–607. Humana Press, Totowa, NJ, 2005.

[Goder and Spiess, 2001] cité page 18
V. Goder and M. Spiess. Topogenesis of membrane proteins : determinants and dynamics. *Febs Letters*, 504(3, Sp. Iss. SI) :87–93, 2001.

[Goldberg, 1989] cité page 78
D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[Hessa *et al.*, 2005a] cité page 1
T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S.H. White, and G. von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024) :377–381, 2005.

[Hessa *et al.*, 2005b] cité page 1
T. Hessa, S.H. White, and G. von Heijne. Membrane insertion of a potassium-channel voltage sensor. *Science*, 307(5714) :1427, 2005.

[Hessa *et al.*, 2007] cité page 27, 32, 56
T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, and G. von Heijne. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, 450(7172) :1026–U2, 2007.

[Hiss and Schneider, 2009] cité page 17
J. A. Hiss and G. Schneider. Architecture, function and prediction of long signal peptides. *Briefings in Bioinformatics*, 10(5) :569–578, 2009.

[Holland, 1962] cité page 78
J.H. Holland. Outline for a logical theory of adaptive systems. *J. ACM*, 9 :297–314, July 1962.

[Hoos and Stutzle, 2004] cité page 54
H. H. Hoos and T. Stutzle. *Stochastic Local Search—Foundations and Applications*. Morgan Kaufmann Publishers, San Francisco, CA, 2004.

[Japkowicz and Stephen, 2002] N. Japkowicz and S. Stephen. The class imbalance problem : A systematic study. *Intell. Data Anal.*, 6 :429–449, October 2002.

[Johansson and Lindahl, 2008] cité page 71, 71
A.C.V. Johansson and E. Lindahl. Position-resolved free energy of solvation for amino acids in lipid membranes from molecular dynamics simulations. *Proteins-Structure Function and Bioinformatics*, 70(4) :1332–1344, 2008.

[Jones *et al.*, 1994] cité page 39
D.t. Jones, W.r. Taylor, and J.m. Thorton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33(10) :3038–3049, 1994.

[Junker *et al.*, 1999] cité page 38, 39
V.L. Junker, R. Apweiler, and A. Bairoch. Representation of functional information in the SWISS-PROT data bank. *Bioinformatics*, 15(12) :1066–1067, 1999.

[Kall and Sonnhammer, 2002] cité page 26
L. Kall and E.L.L. Sonnhammer. Reliability of transmembrane predictions in whole-genome data. *FEBS Letters*, 532(3) :415–418, 2002.

[Kall *et al.*, 2004] cité page 17, 19, 25, 39
L. Kall, A. Krogh, and E.L.L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5) :1027–1036, 2004.

[Kall *et al.*, 2007] cité page 25
L. Kall, A. Krogh, and E.L.L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. *Nucletic Acids Research*, 35(Suppl. S) :W429–W432, 2007.

[Klammer *et al.*, 2009] cité page 26
M. Klammer, D.N. Messina, T. Schmitt, and E.L.L. Sonnhammer. MetaTM - a consensus method for transmembrane protein topology prediction. *BMC bioinformatics*, 10(1) :314, 2009.

[Klee and Ellis, 2005] cité page 1, 20, 20
E.W. Klee and L.B. Ellis. Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 6, 2005.

[Klee and Sosa, 2007] cité page 20
E. W. Klee and C. P. Sosa. Computational classification of classically secreted proteins. *Drug Discovery Today*, 12(5-6) :234–240, 2007.

[Krogh *et al.*, 2001a] cité page 1, 22, 24
A. Krogh, B. Larsson, G. von Heijne, and E.L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model : Application to complete genomes. *Journal of Molecular Biology*, 305(3) :567–580, 2001.

[Krogh *et al.*, 2001b] cité page 13, 19
A. Krogh, B. Larsson, G. von Heijne, and E.L.L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model : application to complete genomes. *Journal of Molecular Biology*, 305(3) :567 – 580, 2001.

[Kvansakul *et al.*, 2010] M. Kvansakul, A. H. Wei, J. I. Fletcher, S. N. Willis, L. Chen, A. W. Roberts, D. C. S. Huang, and P. M. Colman. Structural basis for apoptosis inhibition by epstein-barr virus bhrf1. *PLoS Pathog*, 6(12) :e1001236, 12 2010.

[Kyte and Doolittle, 1982] cité page 21, 44, 64
J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1) :105–132, 1982.

[Landolt-Marticorena *et al.*, 1993] C. Landolt-Marticorena, K.A. Williams, C.M. Deber, and R.A.F. Rrithmeier. Nonrandom Distribution of Amino-Acids in the Transmembrane Segements of Humain Type-I Single Span Membrane-Proteins. *Journal of Molecular Biology*, 229(3) :602–608, FEB 5 1993.

[Lao *et al.*, 2002] cité page 24
D.M. Lao, M. Arai, M. Ikeda, and T. Shimizu. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics*, 18(12) :1562–1566, 2002.

[Laroum *et al.*, 2010] cité page 2, 53
S. Laroum, D. Tessier, B. Duval, and J-K. Hao. A local search appproach for transmembrane segment and signal peptide discrimination. In *EvoBIO*, pages 134–145, 2010.

[Laroum *et al.*, 2011] cité page 2, 53
S. Laroum, B. Duval, D Tessier, and J-K. Hao. Multi-neighborhood search for discrimination of signal peptides and transmembrane segments. In *EvoBio*, pages 111–122, 2011.

[Li and Godzik, 2006] cité page 38
W. Li and A. Godzik. Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13) :1658–1659, July 2006.

[LourenÃ§o *et al.*, 2000] cité page 59
H.R. LourenÃ§o, O.C. Martin, and T. Stutzle. Iterated local search. Economics Working Papers 513, Department of Economics and Business, Universitat Pompeu Fabra, November 2000.

[Lukas, 2010] cité page 13
L. Lukas. Prediction of transmembrane topology and signal peptide given a protein's amino acid sequence. 673 :53–62, 2010.

[MacCallum *et al.*, 2008] cité page 71
J. L. MacCallum, W. F. D. Bennett, and D. P/ Tieleman. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophysical Journal*, 94(9) :3393–3404, 2008.

[Mcgeoch, 1985] cité page 19
D.J. Mcgeoch. On the predictive recognition of signal peptide sequences. *Virus Research*, 3(3) :271–286, 1985.

[Melen *et al.*, 2003] cité page 26
K. Melen, A. Krogh, and G. von Heijne. Reliability measures for membrane protein topology prediction algorithms. *Journal of Molecular Biology*, 327(3) :735–744, 2003.

[Ng *et al.*, 2007] cité page 16, 17
S. Y. M. Ng, B. Chaban, D. J. VanDyke, and K. F. Jarrell. Archaeal signal peptidases. *Microbiology-SGM*, 153(Part 2) :305–314, 2007.

[Nielsen and Krogh, 1998] cité page 1, 18, 24, 24
H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden markov model. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, ISMB '98, pages 122–130. AAAI Press, 1998.

[Nielsen *et al.*, 1997a] cité page 20, 24
H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10(1) :1–6, 1997.

[Nielsen *et al.*, 1997b] cité page 20
H. Nielsen, J. Engelbrecht, S. Brunak, and G. vonHeijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10(1) :1–6, 1997.

[Nielsen *et al.*, 1999] cité page 1, 20, 24
H. Nielsen, S. Brunak, and G. von Heijne. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 12(1) :3–9, 1999.

[Nilsson *et al.*, 2000] cité page 22
J. Nilsson, B. Persson, and G. von Heijne. Consensus predictions of membrane protein topology. *Febs Lettres*, 486(3) :267–269, 2000.

[Nilsson *et al.*, 2005] cité page 71
J. Nilsson, B. Persson, and G. von Heijne. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins-Structure Function and Bioinformatics*, 60(4) :606–616, 2005.

[Nugent and Jones, 2009] T. Nugent and D. Jones. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10(1) :159, 2009.

[Osborne *et al.*, 2005] cité page 16, 17
A.R. Osborne, T.A. Rapoport, and B. van den Berg. Protein translocation by the Sec61/SecY channel. *Annual Review of Cell and Developmental Biology*, 21 :529–550, 2005.

[Pagos *et al.*, 2005] cité page 21
G.G. Pagos, T. Liakopoulos, and S.J. Hamodrakas. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, 6 :7, 2005.

[Park and Helms, 2008a] cité page 31, 39, 42, 61
Y. Park and V. Helms. MINS2 : Revisiting the molecular code for transmembrane-helix recognition by the Sec61 translocon. *Bioinformatics*, 24(16) :1819–1820, 2008.

[Park and Helms, 2008b] cité page 31, 71
Y. Park and V. Helms. Prediction of the translocon-mediated membrane insertion free energies of protein sequences. *Bioinformatics*, 24(10) :1271–1277, MAY 15 2008.

[Pasquier *et al.*, 1999] cité page 58
C. Pasquier, V.J. Promponas, G.A. Palaios, J.S. Hamodrakas, and S.J. Hamodrakas. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the swissprot database : the pred-tmr algorithm. *Protein Engineering*, 12(5) :381–385, 1999.

[Pellegrini-Calace *et al.*, 2003] cité page 71
M. Pellegrini-Calace, A. Carotti, and DT. Jones. Folding in lipid membranes (FILM) : A novel method for the prediction of small membrane protein 3D structures. *Proteins-Structure Function and bioinformatics*, 50(4) :537–545, 2003.

[Pilpel *et al.*, 1999] cité page 71
Y. Pilpel, N. Ben-Tal, and D. Lancet. kPROT : A knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *Journal of Molecular Biology*, 294(4) :921–935, 1999.

[Punta *et al.*, 2007] cité page 24
M. Punta, L.R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, and B. Rost. Membrane protein prediction methods. *Methods*, 41(4) :460 – 474, 2007.

[Rapoport, 2007] T. A. Rapoport. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170) :663–669, 2007.

[Rapoport, 2008] cité page 1
T. A. Rapoport. Protein transport across the endoplasmic reticulum membrane. *Febs Journal*, 275(18) :4471–4478, 2008.

[Reynolds *et al.*, 2008] cité page 19, 25
S.M. Reynolds, L. Kaell, M.E. Riffle, J.A. Bilmes, and W.S. Noble. Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks. *Plos Computational Biology*, 4(11), 2008.

[Rost *et al.*, 1995] cité page 22
B. Rost, R. Casadio, P. Fariselli, and C. Sander. Transmembrane helices predicted at 95-percent accuracy. *Protein Science*, 4(3) :521–533, 1995.

[Saaf *et al.*, 1998] cité page 26
A. Saaf, E. Wallin, and G. Von Heijne. Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *European Journal of Biochemistery*, 251(3) :821–829, 1998.

[Sadlish *et al.*, 2005] cité page 19
H. Sadlish, D. Pitonzo, A.E. Johnson, and W.R. Skach. Sequential triage of transmembrane segments by Sec61 alpha during biogenesis of a native multispanning membrane protein. *Nature Structural & Molecular Biology*, 12(10) :870–878, 2005.

[Senes *et al.*, 2007] cité page 71
A. Senes, D. C. Chadi, P. B. Law, R. F. S. Walters, V. Nanda, and W. F. DeGrado. E-z,

a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes : Derivation and applications to determining the orientation of transmembrane and interfacial helices. *Journal of Molecular Biology*, 366(2) :436–448, 2007.

[Shental-Bechor *et al.*, 2006] cité page 32
D. Shental-Bechor, S.J. Fleishman, and N. Ben-Tal. Has the code for protein translocation been broken ? *Trends in Biochemical Sciences*, 31(4) :192 – 196, 2006.

[Skach, 2007] cité page 19
W.R. Skach. The expanding role of the ER translocon in membrane protein folding. *Journal of Cell Biology*, 179(7) :1333–1335, 2007.

[Sonnhammer *et al.*, 1998a] cité page 21
E.L.L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6 :175–182, 1998.

[Sonnhammer *et al.*, 1998b] cité page 23, 39
E.L.L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, ISMB '98, pages 175–182. AAAI Press, 1998.

[Stone, 1974] cité page 46
M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(1) :111–147, 1974.

[Tan *et al.*, 2008] cité page 19
S. Tan, H. T. Tan, and M. C. Chung. Membrane proteins and membrane proteomics. *Proteomics*, 8(19) :3924–3932, 2008.

[Terstappen and Reggiani, 2001] cité page 13, 19
G.C. Terstappen and A. Reggiani. In silico research in drug discovery. *Trends in Pharmacological Sciences*, 22(1) :23 – 26, 2001.

[Tusnady and Simon, 2001] cité page 1, 22, 24
G.E. Tusnady and I. Simon. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17(9) :849–850, 2001.

[Tusnady *et al.*, 2004a] cité page 42
G.E. Tusnady, Z. Dosztanyi, and I. Simon. Tmdet : web server for detecting transmembrane regions of proteins by using their 3d coordinates. *Bioinformatics*, 2004.

[Tusnady *et al.*, 2004b] cité page 37, 37, 42
G.E. Tusnady, Z. Dosztanyi, and I. Simon. Transmembrane proteins in protein data bank : identification and classification. *Bioinformatics*, 20 :2964–2972, 2004.

[Tusnady *et al.*, 2005] cité page 12, 13, 19, 37
G.E. Tusnady, Z. Dosztanyi, and I. Simon. Pdbtm : selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Research*, 33(suppl 1) :D275–D278, 2005.

[Ulmschneider and Sansom, 2001] cité page 71
M.B. Ulmschneider and M.S.P. Sansom. Amino acid distributions in integral membrane protein structures. *Biochimical et Biophysica Acta-Biomembranes*, 1512(1) :1–14, MAY 2 2001.

[Ulmschneider *et al.*, 2005a] cité page 64, 71
M.B. Ulmschneider, M.S.P. Sansom, and A. Di Nola. Properties of integral membrane protein structures : Derivation of an implicit membrane potential. *Proteins-structure Function and Bioinformatics*, 59(2) :252–265, 2005.

[Ulmschneider *et al.*, 2005b] cité page 71
M.B. Ulmschneider, M.S.P. Sansom, and A. Di Nola. Properties of integral membrane protein structures : Derivation of an implicit membrane potential. *Proteins-Structure Function and Bioinformatics*, 59(2) :252–265, MAY 1 2005.

[van den Berg *et al.*, 2004] cité page 17
B. van den Berg, W.M. Clemons, I. Collinson, Y. Modis, E. Hartmann, S.C. Harrison, and T.A. Rapoport. X-ray structure of a protein-conducting channel. *Nature*, 427(6969) :36–44, 2004.

[Vert, 2002] cité page 20
J.P. Vert. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauerdale, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 649–660. World Scientific, 2002.

[Viklund and Elofsson, 2004] cité page 91
H. Viklund and A. Elofsson. Best a-helical transmembrane protein topology predictions are achieved using hidden markov models and evolutionary information. *Protein Science*, 13(7) :1908–1917, 2004.

[Viklund and Elofsson, 2008] cité page 25
H. Viklund and A. Elofsson. OCTOPUS : improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15) :1662–1668, 2008.

[Viklund *et al.*, 2006] H. Viklund, E. Granseth, and A. Elofsson. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins : application to complete genomes. *Journal of Molecular Biology*, 361(3) :591–603, 2006.

[Viklund *et al.*, 2008] cité page 25
H. Viklund, A. Bernsel, M. Skwark, and A. Elofsson. SPOCTOPUS : a combined predictor of signal peptides and membrane protein topology. *Bioinfrmatics*, 24(24) :2928–2929, 2008.

[Von Heijne, 1986a] cité page 19
G. Von Heijne. A new method for predicting signal sequence clavage sites. *Nucleic Acids Research*, 14(11) :4683–4690, 1986.

[von Heijne, 1986b] G. von Heijne. Net n-c charge imbalance may be important for signal sequence function in bacteria. *Journal of Molecular Biology*, 192(2) :287 – 290, 1986.

[Von heijne, 1992] cité page 30
G. Von heijne. Membrane Protein Structure Prediction : Hydrophobicity Analysis and the 'Positive Inside' Rule. *Journal of Molecular Biology*, 225(2) :487–494, MAY 20 1992.

[von Heijne, 1999] G. von Heijne. Recent advances in the understanding of membrane protein assembly and structure. *Quartely Reviews of Biophysics*, 32(4) :285–307, 1999.

[von Heijne, 2006] cité page 71
Gunnar von Heijne. Membrane-protein topology. *ature Reviews Molecular Cell Biology*, 7(12) :909–918, 2006.

[Wallin and von Heijne, 1998] E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science*, 7(4) :1029–1038, 1998.

[Yan *et al.*, 2003] cité page 49
L. Yan, R. Dodier, M.C. Mozer, and R. Wolniewicz. Optimizing classifier performance via the wilcoxon-mann-whitney statistic, 2003.

[Yau *et al.*, 1998] W.M. Yau, W.C. Wimley, K. Gawrisch, and S.H. White. The preference of tryptophan for membrane interfaces. *Biochemistry*, 37(42) :14713–14718, OCT 20 1998.

[Zhang *et al.*, 2009] cité page 20
X. Zhang, Y. Li, and Y. Li. Evaluating signal peptide prediction methods for gram-positive bacteria. *Biologia*, 64 :655–659, 2009. 10.2478/s11756-009-0118-3.

[Zheng *et al.*, 2003] cité page 24
Y. Zheng, J. D. Melissa, Z. Fasheng, and D.T. Rohan. Computational differentiation of n-terminal signal peptides and transmembrane helices. *Biochemical and Biophysical Research Communications*, 312(4) :1278 – 1283, 2003.

Prédiction de la localisation des protéines membranaires : méthodes méta-heuristiques pour la détermination du potentiel d'insertion des acides aminés.


## Résumé

Dans ce travail, nous nous intéressons à la localisation des protéines adressées vers la membrane du réticulum endoplasmique, et plus spécifiquement à la reconnaissance des segments transmembranaires et des peptides signaux. En utilisant les dernères connaissances acquises sur les mécanismes d'insertion d'un segment dans la membrane, nous proposons une méthode de discrimination de ces deux types de séquences basée sur le potentiel d'insertion de chaque acide aminé dans la membrane. Cela amène à rechercher pour chaque acide aminé une courbe donnant son potentiel d'insertion en fonction de sa place dans une fenêtre correspondant à l'épaisseur de la membrane. Notre objectif est de déterminer « in silico » une courbe pour chaque acide aminé, afin d'obtenir les meilleures performances pour notre méthode de classification. L'optimisation, sur des jeux de données construits à partir des banques de données de protéines, des courbes est un problème difficile que nous abordons grâce aux méthodes méta-heuristiques.
Nous présentons tout d'abord un premier algorithme de recherche locale permettant d'apprendre un ensemble de courbes. Son évaluation sur les différents jeux de données montre de bons résultats de classification. Cependant, nous constatons une difficulté d'ajustement pour les courbes de certains acides aminés. La restriction de l'espace de recherche grâce à des informations pertinentes sur les acides aminés et l'introduction d'un voisinage multiple nous permettent d'améliorer les performances de notre méthode et en même temps de stabiliser les courbes apprises. Nous présentons également un algorithme génétique développé afin d'explorer de manière plus diversifiée l'espace de recherche de ce problème.



**Mots-clés : optimisation, recherche locale, algorithme génétique, position des acides aminés, insertion des segments TM, peptide signal**


Membrane proteins prediction : Metaheuristic methods for the determination of potential insertion of amino acids.


## Abstract

In this work, we are interested in the localization of proteins transported towards the endoplasmic reticulum membrane, and more specifically to the recognition of transmembrane segments and signal peptides. By using the last knowledges acquired on the mechanisms of insertion of a segment in the membrane, we propose a discrimination method of these two types of sequences based on the potential of insertion of each amino acid in the membrane. This leads to search for each amino acid a curve giving its potential of insertion according to its place in a window corresponding to the thickness of the membrane. Our goal is to determine "in silico" a curve for each amino acid to obtain the best performances for our method of classification. The optimization, on data sets constructed from data banks of proteins, of the curves is a difficult problem that we address through the meta-heuristic methods.
We first present a local search algorithm for learning a set of curves. Its assessment on the different data sets shows good classification results. However, we notice a difficulty in adjusting the curves of certain amino acids. The restriction of the search space with relevant information on amino acids and the introduction of multiple neighborhood allow us to improve the performances of our method and at the same time to stabilize the learnt curves. We also developed a genetic algorithm to explore in a more diversified way the space of search for this problem.



**Keywords :optimization, local search, genetic algorithm, amino acid position, TM segment insertion, signal peptide**