Doctorat en informatique cognitive

Examen de projet

Analyse de motifs protéiques par méthode hybride réseau de neurones artificiel et Modèle de Markov Caché

Directeur : Anne Bergeron
Codirecteur : Fathey Sarhan

DIC9410 Examen de projet
par Guylaine Poisson POIG31576500
Avril 2004

TABLE DES MATIERES

I. Projet

1. Résumé

Un ancrage GPI est une structure d'ancrage membranaire complexe mais commune chez les protéines eucaryotes extracellulaires. Cette structure a été très bien conservée durant l'évolution de la cellule de la levure jusqu'à celle des mammifères. La fonction précise de ce type d'attachement n'est pas bien définie, mais cette conservation élevée dans l'évolution des cellules eucaryotes, laisse facilement présumer un rôle fonctionnel important (Nosjean et al, 97). Toutefois quelques caractéristiques sont connues. Par exemple les ancres GPI sont souvent définies comme des cibles ou des signaux positionnés à la surface des cellules (Ali et al, 1996). Les banques de séquences protéiques tel que Swiss-Prot (EBI, 2004) proposent peu de séquences ayant cette modification car leur présence n'est pas connue depuis longtemps (1980) (low et Zilversmit, 1980) et peu d'outils permettent l'annotation automatique des nouvelles séquences. Les différents projets de séquençages de génomes, amènent une profusion de nouvelles séquences qu'il faut annoter. De plus la prédiction de modification post-traductionnelle des protéines fait partie intégrante d'une étude approfondie permettant la compréhension des fonctions biologiques. Elle se révèle être une étape importante, non seulement pour l'annotation des protéomes mais aussi pour l'étude des systèmes biologiques à grande échelle. Des outils qui pourront aider à l'annotation des signaux dans les séquences sont donc une nécessité, surtout pour des structures récemment découvertes comme les ancres GPI.

Ce projet a pour but de développer un système hybride qui se base sur l'utilisation d'un réseau de neurones artificiel (RNA) et d'un Modèle de Markov Caché (MMC). Le RNA sélectionne les séquences protéiques ayant un signal GPI potentiel et le MMC structure le signal. La combinaison des deux techniques d'apprentissage machine révèle un pouvoir prédictif intéressant car elle exploite les propriétés physicochimiques de la molécule ainsi que la nature séquentielle de sa représentation. Le système permet de prédire 93% des séquences protéiques annotées comme protéine à ancrage GPI dans la base de données Swiss-Prot. Une caractéristique importante du système hybride que nous proposons est qu'il cible uniquement la partie C-terminale de la protéine. Cette particularité le rend moins sensible aux bruits si repandus dans les bases de données de séquences. De plus ce système n'est pas spécifique à un seul groupe taxonomique. Il peut être utilisé pour prédire la présence de protéines à ancrage GPI chez tous les eucaryotes (plantes, animaux, champignons, protozoaires etc.). Finalement une technique d'annotation selon une échelle de qualité permet de combiner une très grande sensibilité ainsi qu'une annotation informative de chaque prédiction du système hybride.

2. Problématique

En 1866 un moine tchèque du nom de Gregor Mendel établissait les premières lois de l'hérédité grâce à son étude sur l'hybridation des plantes (Mendel, 1866). De ces travaux est née la génétique classique. Cette découverte ouvrait la porte à l'étude du transfert de l'information dans le matériel vivant. Plusieurs domaines de recherche sont nés suite aux travaux de Mendel, notons entre autre la biologie moléculaire et la bioinformatique. Depuis plus de 130 ans, les projets d'études génétiques et moléculaires se sont multipliés.

L'importance des gènes n'est plus un secret pour personne. De nos jours, des termes tel que ADN (acide désoxyribonucléique) et protéine ne sont plus des termes techniques connus que par les experts. Les séquences biologiques ont maintenant une place capitale dans la recherche sur le vivant. Ces séquences sont représentées par une suite de lettres provenant d'un alphabet de 4 lettres pour les acides nucléiques de l'ADN, et de 20 lettres pour les acides aminés des protéines. Depuis 1955 lors de la publication de la première séquence protéique (Insuline bovine) (Sanger et al, 1955.) le nombre de séquences protéiques et nucléiques rendues publiques ne fait qu'augmenter. La nécessité de trouver des moyens d'entreposer et surtout d'analyser toute cette information fut vite un sujet de discussion et de recherche. La première base de données de séquences biologiques (Dayhoff et al, 1965) et les premiers algorithmes d'analyses de ces données ont donc vu le jour quelques années plus tard donnant ainsi, par la même occasion, naissance à un nouveau domaine de recherche : La bioinformatique. Depuis maintenant plusieurs années, des milliers de projets scientifiques reliés de près ou de loin à la découverte d'un moine tchèque ayant vécu il y a plus de 150 ans, se concentrent sur l'analyse de ces séquences et des mystères qu'elles renferment.

2.1 La recherche de motifs de séquences biologiques

Deux types principaux de séquences biologiques existent : les séquences d'acides nucléiques et les séquences d'acides aminées. La première catégorie forme l'ADN et l'ARN (acide ribonucléique) tandis que la seconde forme la protéine. Tel un langage humain, un alphabet de base est associé à ces séquences biologiques. L'alphabet de l'ADN et l'ARN est composé de 4 lettres {A, C, G, T}. Pour la protéine on remarque une plus grande diversité avec 20 symboles { a, r, n, d, c, q, e, g, h, i, l, k, m, f, p, s, t, w, y, v }.

Ces séquences de lettres ont une structure primaire, secondaire et tertiaire (Figure 1). Dans le cas de ce projet seul la structure primaire est étudiée. Il est maintenant accepté qu'un motif ou patron de séquence ou de structure représente une caractéristique importante pour la fonction de la molécule. Un motif de structure primaire agit comme signal ou comme base pour la structuration de la molécule dans sa forme tertiaire. Par

exemple un motif représentant une série de répétitions dans la structure primaire du prion fut conservé au cours de l'évolution dans l'ARN messager. Ce motif pourrait possiblement former un pseudonoeud lors de la structuration tridimensionnelle et être responsable, à un certain niveau, de la traduction défectueuse de la protéine prion dans le cas de la forme héréditaire de la maladie de Creutzfeldt-Jakob (Barrette et al, 1999).

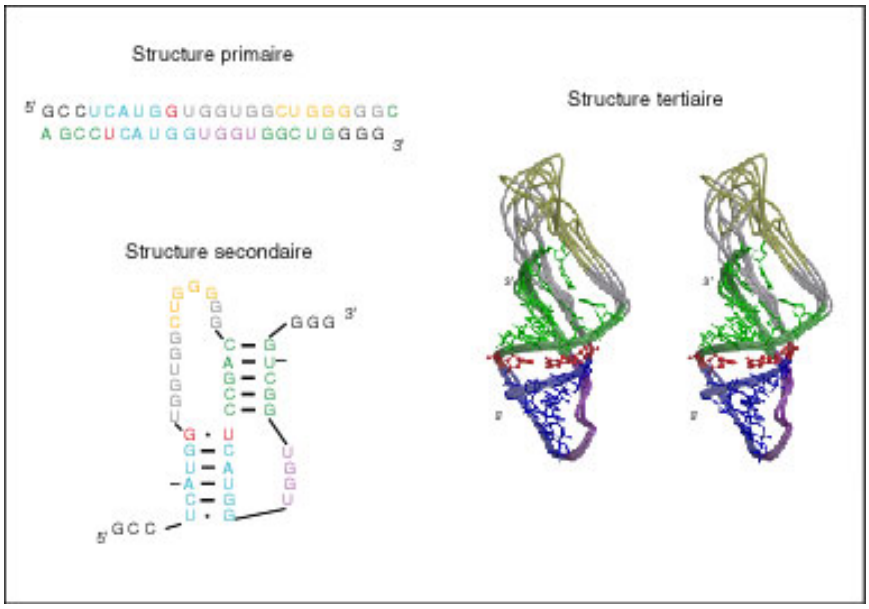


Figure 1 : Structure primaire, secondaire et tertiaire de l'ARNm du prion (Barrette et al, 1999).

La bioinformatique se penche depuis longtemps sur la recherche de motif dans les séquences. En analyse de séquences protéiques la recherche de motif est un problème des plus important. Ces motifs peuvent être de puissants outils permettant la classification des séquences protéiques. Comme spécifié plus haut, la découverte de motif dans des séquences biologiques peut nous donner un indice sur la fonction de cette séquence, mais elle peut aussi nous informer sur la relation entretenue entre deux séquences au cours de l'évolution (Berjova et al, 2000). Plusieurs algorithmes sont utilisés pour traiter les motifs de séquences biologiques. On peut les classer selon l'approche algorithmique utilisée (recherche exhaustive, apprentissage machine, méthode heuristique itérative etc.) (Berjova et al, 2000). On peut aussi classer les méthodes utilisées selon la tâche accomplie; c'est-à-dire la recherche de motif pour classification de séquences ou la recherche de nouveaux motifs. Dans le cadre de ce projet, la tâche à accomplir est celle de la classification.

3. La nature « linguistique » des séquences biologiques

Afin de choisir la bonne technique de classification à utiliser il faut bien comprendre la nature de nos données. Les molécules sont des objets ou des corps simples qui existent dans un monde tridimensionnel.

Toutefois notre façon de les représenter, d'entreposer l'information et surtout notre façon de les analyser demande un formalisme beaucoup plus malléable. La représentation des séquences biologiques étant faite sous la forme d'un texte composé à partir de différents alphabets, il n'est pas étonnant qu'un parallèle fut vite proposé : celui de la linguistique.

L'analogie entre la linguistique et la biologie moléculaire est plus qu'une simple métaphore. La grande similitude entre le langage humain et celui de la cellule offre de grandes opportunités comme par exemple l'utilisation des méthodes d'analyses linguistiques afin de mieux comprendre et décomposer le langage cellulaire. Plusieurs techniques bioinformatiques ont leurs racines en linguistique même si leurs développements ont été indépendants. L'approche « mathématique » de la linguistique a permis une avancée importante dans le développement de la théorie du langage formel qui sera par la suite un des piliers de la recherche de motifs et de structures dans les séquences biologiques.

3.1 Les grammaires

Comme le définit si bien Searls dans son article de 1993 « *Formally, a language is simply a set of strings of caracters drawn from some alphabet*… ». Le langage est donc, d'un point de vue formel, un simple groupement de chaînes de symboles appartenant à un alphabet. Les travaux de Chomsky (1957) ont eu pour but de fournir des méthodes formelles de définition de langages ou, plus simplement, de donner une théorie du modelage des chaînes de symboles présentes dans un langage. Le but de la représentation formelle des langages est l'économie d'expression, c'est-à-dire qu'il s'agit de ne pas avoir à énumérer exhaustivement toutes les chaînes possibles dans un langage. La puissance octroyée par cette représentation «économique» est bien réelle. Un autre avantage est celui de pouvoir généraliser l'information structurelle d'un système linguistique (Searls et Dong, 1993). La structuration syntaxique de Chomsky sera amplement utilisée en informatique sous la forme de la linguistique computationelle. Si on transpose cette définition du langage aux protéines, l'alphabet sera composé de 20 symboles que sont les acides aminés. La protéine sera ainsi représentée par une chaîne de symboles. Toutes les différentes compositions de chaînes formeront un langage.

Mais comment une théorie de modelage de symboles peut-elle servir en analyse de séquences protéiques? Il est important de comprendre la théorie de la grammaire générative de Chomsky pour bien pouvoir répondre à cette question.

3.2 La hiérarchie de Chomsky

Chomsky a spécifié 4 types de grammaires basées sur les restrictions des règles de production. Plus on monte dans la hiérarchie, plus on a la possibilité de construire des règles générales. Pour bien comprendre les niveaux hiérarchiques de Chomsky, il faut d'abord définir quelques principes de notation. Il y a deux types de symboles : les variables abstraites ***non-terminales*** (lettres majuscules) et les symboles concrets ***terminaux*** (lettres minuscules). Les ***règles*** de production du langage seront notées A → *a* ou la partie de gauche contient au moins un symbole non-terminal qui sera transformé dans la partie de droite en une chaîne terminale et/ou non-terminale.

Une grammaire génère les chaînes composant son langage en prenant un symbole de départ (A) et en le réécrivant. Cette réécriture se fait en recherchant, à répétition, une règle ayant un côté gauche correspondant à un symbole non-terminal de la chaîne courante et en y substituant le côté droit de la règle. Tout ceci jusqu'à ce que la chaîne ne contienne que des symboles terminaux.

Voici un exemple de grammaire régulière. La grammaire G = (*N, E, P, S*)

S = ***Symbole initial*** : {A}
E = ***Symboles terminaux*** : {a, b}
N= ***Symboles non terminaux*** : {A, B}
P= ***Règles*** : {A→a, A→aB, B→b, B→bB}

À partir de cette grammaire, on peut dériver toutes les chaînes qui composeront le langage en partant du symbole non terminal A.

Appliquons la première. On obtient la suite de mots [a]. On ne peut aller plus loin.
Appliquons la seconde. On obtient la suite [aB]
Appliquons la troisième règle à [aB]. On obtient la suite [ab]. On ne peut aller plus loin.
Appliquons la quatrième règle à [aB]. On obtient la suite [abB].
Appliquons la quatrième règle à [abB]. On obtient [abbB].
Appliquons la troisième règle à [abbB]. On obtient [abbb]. On a une suite de constantes. On ne peut aller plus loin. On peut représenter cet exemple de grammaire grâce à un arbre de dérivation (Figure2).

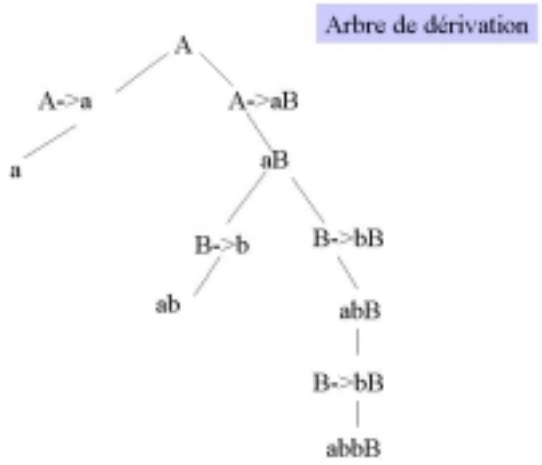


Figure 2 : Arbre de dérivation relié à la grammaire décrite ci-dessus (Habrias, 2002).

Les types de grammaires :

1. La grammaire régulière (RG) : Seulement les règles de production de type A→*a* ou A → *a*A sont permises. Les chaînes peuvent donc grandir que dans une seule direction. Exemple : Voir ci haut.
2. La grammaire hors contexte (CFG) : Toute règle de production de type A → β est accepté ou β représente n'importe quelle chaîne terminale et/ou terminale excluant la chaîne nulle. Cette grammaire permet la description de boucles (aabaabaa) mais n'autorise pas les croisements, par exemple deux copies de *aab* séparés par la chaîne *cc* (aabccaab) (Searls et Dong, 1993). Exemple : avec les règles P = {A → aAa, A→bAb, A→aa, A→bb }, une des dérivations obtenues est A→aAa→aaAaa→aabAbaa→**aabaabaa**.
3. La grammaire sensible au contexte (CSG) : Les CSG répondent au problème des copies en autorisant plus d'un symbole du côté gauche de la règle. Cette grammaire autorise les copies (aabccaab). Le côté droit de la règle est au moins aussi long que le côté gauche. Il y aura, par exemple, présence de règles de réorganisation de symboles non terminaux et de génération de symbole terminaux. Aucun algorithme fonctionnant en temps polynomial n'existe pour leur analyse : problème NP-complet. Les CSG sont donc en pratique non considérées. Exemple : Les règles sont ici beaucoup plus complexes. Voici un exemple de règles. ÂB→BA ou (Â) représente

la génération d'un (a) du côté droit de la séquence et (A) représente la génération d'un (a) du côté gauche.

4. La grammaire sans restriction (UnresG) : Dans une UnresG, n'importe quel symbole peut se retrouver des deux côtés de la règle. C'est la grammaire la plus générale. Chaque dérivée possible est énumérée. Aucun algorithme ne peut garantir qu'une chaîne est une dérivation valable de la grammaire dans un temps fini.

3.3 La linguistique et la biologie moléculaire

Comme la séquence biologique est une séquence de lettres, il n'est pas étonnant de penser à l'utilisation des grammaires pour son analyse. La plus importante utilisation des grammaires chomskyennes en bioinformatique est la recherche de motifs dans les séquences biologiques via les grammaires régulières (Betel et Hogues, 2002; Xuan et al, 2002). Les grammaires régulières sont aussi très utilisées pour la recherche de motifs dans les bases de données protéiques et nucléiques (Gattiker et al 2002) et la prédiction de gènes dans les séquences génomiques (Burge et Karlin, 1997; Kulp et al, 1996). La plupart des algorithmes d'analyse de structure primaire des séquences biologiques sont donc des modèles se situant au niveau de base de la hiérarchie de Chomsky (Durbin et al 1998). Mais l'utilisation des grammaires hors contexte est de plus en plus courante lors de problèmes impliquant la structure secondaire et tertiaire des séquences. L'utilisation des principes de la grammaire générative de plus haut niveau est moins commune, peu d'auteurs l'utilisent.

4. La classification de motifs protéiques

La recherche de motifs dans les séquences protéiques sert comme base pour la classification des familles protéiques. Un motif fonctionnel caractéristique à une famille peut donc servir comme outil de recherche pour sélectionner de nouveau membres de cette famille (Sonnhammer et al, 1997; Bateman et al, 2004). Ces motifs servent aussi à répertorier des caractéristiques particulières, tel des signaux, rencontrer chez certaines protéines (Nielson et al, 1997A; Nielson et al, 1997B). Un fois un motif ou un signal identifié il reste à trouver une méthode adéquate de l'utiliser. Certains motifs de séquences sont très clairs et bien définis. Dans ces cas de simples outils de recherche de similarité entre les séquences sont nécessaires. Toutefois la complexité des séquences biologiques est beaucoup plus grande dans plusieurs cas. Pour ces motifs ou signaux nous avons besoin de méthodes beaucoup plus performantes dans des conditions bruitées.

4.1 Deux problèmes, deux méthodes, une solution

Lorsqu'on a un motif complexe que l'on veut utiliser pour la classification ou l'annotation, on se retrouve devant une quantité très importante de données et d'informations. Une première étape dans cette classification est de bien résoudre le problème de nettoyage de nos données. Les bases de données contenant les séquences biologiques ainsi que les séquences elles-mêmes sont souvent incomplètes ou bruitées. Même si ces séquences sont déterminées expérimentalement avec un grande précision elles subissent plusieurs manipulations avant d'être accessibles pour fins d'analyses. Le taux d'erreur devient donc beaucoup plus important que l'erreur initiale que l'on retrouve normalement durant le processus expérimental (Brunak et al, 1990). L'utilisation de techniques d'apprentissage machine tel que les réseaux de neurones artificiels (RNA) est une alternative intéressante car ils sont performants face à des problèmes de classification impliquant une grande quantité d'exemples ayant de l'information bruitée. (Wu et McLarty, 2000). Les données présentées à un RNA doivent être encodées. Cet encodage nous donne la possibilité d'incorporer beaucoup d'informations, tel que des propriétés physicochimiques des acides aminés de la protéine.

On retrouve différentes architectures de RNA dans l'étude des séquences biologiques. Au niveau des protéines seulement, plusieurs études utilisent les réseaux neuroniques en prédiction de structure protéique et en analyse de séquence. Notons entre autres les travaux de Quian et Sejnowski (1988) qui ont étudié la prédiction de structure secondaire de protéine à l'aide d'un algorithme de rétropopagation ou ceux de Rost et Sander (1994) sur la prédiction de l'accessibilité des solvants. Dans le domaine de l'analyse de séquences on peut citer les travaux de Nielsen et al (1997A et 1997B) qui ont développé un système de prédiction d'un site protéique combinant la recherche du signal et sa composition. D'autres travaux effectués dans le domaine de l'analyse de séquences protéiques sont ceux de Nakata (1995) sur l'utilisation d'un algorithme de rétropopagation pour prédire les sites d'attachements à l'ADN de certaines protéines. Il est aussi important de noter ceux de Wu et al (1992) qui utilisent l'apprentissage supervisé pour classifier de nouvelles séquences dans des familles protéiques déjà existantes.

Une fois les données nettoyées et de bons candidats bien ciblés, il faut raffiner notre classification et la structurer. Ici le caractère séquentiel et linguistique de nos séquences incite à l'utilisation d'une autre technique d'apprentissage automatique, qui est aussi une méthode d'analyse linguistique: le Modèle de Markov Caché (MMC). Pour bien classifier un objet dans une classe il faut « structurer ou modéliser » cette classe. Un modèle commun aux différents éléments d'un même classe sera donc proposé.

L'utilisation d'un MMC et de sa capacité de structuration du langage régulier des séquences biologiques, va permettre de mieux définir la classe ciblée. Le MMC est, tel le RNA, amplement utilisé en analyse de séquences biologiques. Cette méthode est utilisé pour la prédiction de signaux chez les protéines (Nielson et al 1997B), pour construire des modèles de structures secondaires (Francesco et al, 19997) ou comme outil d'alignement de séquences permettant de créer un profil spécifique à un groupe (Eddy, 1995). On retrouve aussi le MMC dans les outils de recherche de gènes (Krogh et al, 1994).

Nous voici donc devant une solution au problème de classification faisant appel à deux techniques d'apprentissage déjà individuellement abondamment utilisées en analyse de séquences biologiques et qui ont des forces complémentaires. Le RNA pour la fouille de données et pour une première épuration basée sur des caractéristiques physicochimiques de la protéine. Et finalement le MMC, qui lui, structure le classement.

5. Les Réseaux de Neurones Artificiels et les Modèles de Markov

5.1 Les Réseaux de Neurones artificiels

La connaissance se retrouvant dans les RNA est, comme dans les réseaux de neurones biologiques, acquise par un processus d'apprentissage par l'exemple. Un autre parallèle est celui de l'entreposage des connaissances. C'est la force des interconnexions (poids synaptiques) qui entrepose la connaissance (Haykin 1994).

Un RNA de type perceptron multicouche est un réseau interconnecté composé, d'au moins trois couches ayant un nombre varié de neurones (unités) : une couche d'entrée, une ou des couches cachées et une couche de sortie. Il existe plusieurs architectures de réseau et de nombreux algorithmes d'apprentissage. Ces modèles de RNA sont utilisés de différentes manières : comme modèle de l'intelligence « humaine » ou comme méthode d'analyse de données. Une « faculté » intéressante des RNA est celle d'être de très bons classificateurs. La nature de cette reconnaissance de formes utilisée pour la classification est statistique car les patrons sont représentés par des points dans un espace de décision multidimensionnel (Haykin, 1994). Les RNA seront donc souvent considérés comme un outil d'implantation de méthodes statistiques (Sarle, 1994; Holmström et al, 1996; Jain et al, 2000).

Au point de vue des caractéristiques, on peut dire que les RNA sont des outils apprennent à partir d'une base de connaissances qui montrent une grande capacité de généralisation. La qualité, la diversité ainsi que la quantité d'information présente dans cette base est directement reliée à la performance du réseau

face aux problèmes qui lui sont exposés. Ils sont capables de construire de façon non paramétrique des frontières de décisions séparant différentes classes : ils offrent donc la possibilité de résoudre des problèmes de classification extrêmement complexes (Lemaire, 1999). Toutefois, ils ont un faible pouvoir explicatif, car la structure interne du réseau reste inconnue.

Une des particularités des RNA est leur apprentissage *online*. Un algorithme *online* effectue la mise à jour des poids après chaque passage des données plutôt que d'accumuler les valeurs des gradients pour chaque donnée (apprentissage par *batch*). L'apprentissage *online* a plusieurs avantages dont la possibilité de traiter de grande quantité de données, d'effectuer un apprentissage plus rapide lorsque les données sont redondantes et de permettre l'entrée de nouvelles données lors de l'apprentissage (Orr et Cummins, 1999).

5.1.1 Le perceptron multicouche

Le perceptron multicouche est une architecture très utilisée dans le domaine de la bioinformatique. L'architecture du perceptron se résume en au moins trois couches composées de neurones. Chaque neurone de la première couche est relié à la suivante par une connectivité totale et ceci, jusqu'à la couche de sortie. (Figure 3). La première couche est celle d'entrée, la dernière est la couche de sortie tandis que les couches intermédiaires sont des couches cachées. Ce type d'architecture permet la résolution de problème non linéaire.

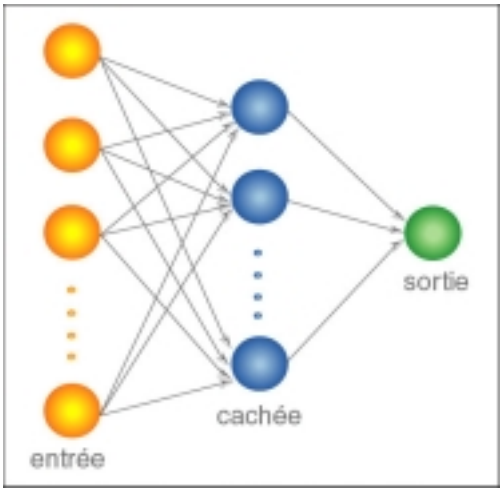
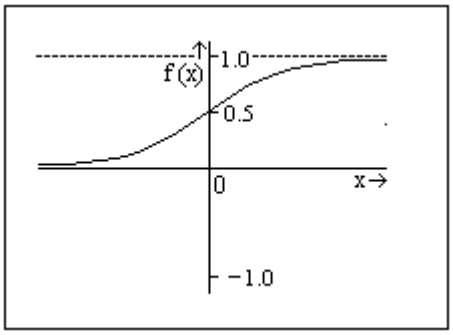


Figure 3 : Architecture de RNA.

Le fonctionnement du perceptron multicouche est simple : chaque neurone de la couche d'entrée génère un signal de sortie qui sera envoyé vers les neurones de la ou les couche(s) cachée(s). Ces derniers neurones génèrent un signal de sortie qui sera reçu par les neurones de la couche de sortie qui eux engendrent le résultat de la prédiction.

Numériquement chaque neurone est le résultat d'un calcul impliquant différentes fonctions et sommations. La valeur de sortie est une fonction pondérée des valeurs d'entrée qui est modulée par une autre fonction généralement sigmoïde (Figure 4).



$$a(x) = 1/1 + e^{-\Sigma}$$

Figure 4 : Représentation de la fonction sigmoïde. Dans cette fonction Σ est la sommation de tous les poids affectés aux neurones de la couche précédente.

Le type d'entraînement d'un RNA décrit en fait la façon selon laquelle les poids seront déterminés. Différents algorithmes d'apprentissage existent. Nous décriront ici l'algorithme de rétropopagation « resilent » RPROP (Riedmiller et Braun 1992, 1993).

L'algorithme de rétropopagation proposé en 1986 par Rumelhart est un algorithme couramment utilisé en analyse de séquences biologiques et a démontré la pertinence de son utilisation. Il est très efficace pour des tâches de classification dans un nombre précis de classes. Ce type d'apprentissage de correction de l'erreur est dit supervisé. Lors d'un apprentissage supervisé, à chaque entrée présentée correspond une sortie précise. Le réseau apprend à bien reconnaître la structure des différentes classes de sortie qu'on lui présente. Si une nouvelle classe est présentée au réseau entraîné, elle sera dirigée dans une des classes existantes avec une faible valeur de correspondance. Aucune nouvelle classe n'est formée avec ce type d'apprentissage.

L'algorithme de rétropopagation de l'erreur utilise une descente de gradient de la surface de l'erreur. Un algorithme de descente de gradient repose sur une fonction de coût quadratique C qui doit être minimisée au cours de l'apprentissage.

$$C = \sum_{x \in P} \sum_{i \in O} (d_i^x - s_i^x)^2, \tag{5.1}$$

L'équation 5.1 montre le calcul de gradient global « batch ». Dans cette fonction, *P* est l'ensemble des exemples d'apprentissage, *O* est l'ensemble des cellules de sortie, s_i^x est la valeur du neurone de sortie *i* après la présentation de l'exemple *x* et d_i^x est la valeur désirée pour le neurone correspondant. Cette fonction de coût quadratique n'est pas la seule possible, toute fonction dérivable en *S* et *d* peut être utilisée.

Lors de la présentation d'un exemple de couple (entrée-sortie), le réseau produira des valeurs de sorties utilisées dans le calcul de la fonction de coût. L'erreur est ensuite propagée aux couches antérieures (Siveton, 2002). La particularité du RPROP comparativement au rétropopagation traditionnel est que seulement le signe de la dérivé est pris en compte pour permettre la direction de la mise à jour des poids. Cet algorithme permet une convergence plus rapide comparativement à la rétropopagation classique. (Riedmiller et Braun 1992, 1993). Son fonctionnement peut se définir ainsi : il commence par une petite valeur de mise à jour et ensuite il augmente cette valeur si le gradient présent à la même direction (signe) que le gradient précédent. Toutefois si la direction est opposée, il diminue la valeur. Cette mise à jour est ajoutée au poids si le gradient est positif et soustrait du poids, s'il est négatif.

Les RNA tel que le perceptron multicouche sont donc une tentative d'utiliser un modèle voulant mimer la capacité « humaines » d'apprentissage pour chercher à classifier des données ayant des motifs possiblement cachés et ayant des ramifications trop complexes pour l'œil d'un expert. Il devient donc très intéressant de chercher à voir si ces modèles artificiels peuvent trouver un lien entre des données diverses et faire ressortir quelques conclusions sur les liens utilisés.

5.2 Les Modèles de Markov

Lorsque l'apprentissage s'effectue sur des séquences d'évènements la méthode utilisée doit représenter la nature des évènements ainsi que la manière dont ils s'enchaînent : une méthode efficace pour ces séquences d'évènements est le MMC (Cornuéjols et Miclet, 2002). Pour bien décrire le modèle de Markov caché il faut tout d'abord décrire un Modèle de Markov plus simple ; le modèle de Markov observable.

5.2.1 Le Modèle de Markov observable

Les modèles de Markov observables sont des automates probabilistes à états finis. Ils se basent sur l'hypothèse de Markov « Le futur ne dépend que du présent et non du passé ». Un modèle de Markov

observable est un graphe d'états dotés de transitions probabilistes. À chaque étape le système se déplace de l'état présent vers l'état suivant selon des probabilités de transitions. Dans ce cas les états sont observables, la chaîne produite sera donc une suite d'états observés. La figure 5 représente un exemple d'un modèle de Markov observable pour un modèle de prédiction météorologique. Ici la question est « En se basant sur la température d'aujourd'hui, quel sera la température de demain? ». Comme on peut voir dans l'exemple, si la journée est « Sunny » ensoleillé nous avons une probabilité de transition vers une autre journée ensoleillée de 0.8 avec toutefois seulement 0.1 de probabilité d'avoir une journée de pluie « Rainy ». Dans ce cas, nous avons une séquences d'états (Sunny, Rainy ou Foggy) car à chaque état correspond un seul évènement observable. Cependant dans certains cas les états ne peuvent être observés. On utilise alors, les modèles de Markov cachés.

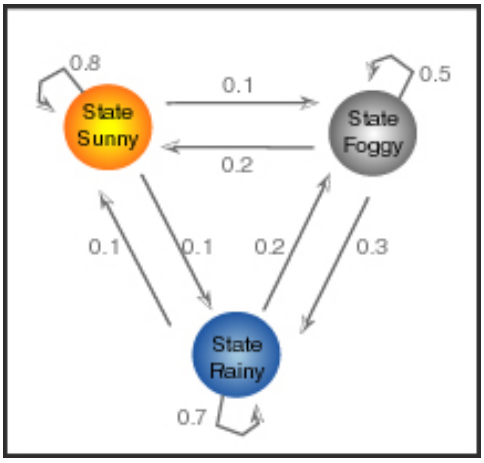


Figure 5 : Modèle de Markov observable. Les états sont Sunny pour ensoleillé; Fuggy pour brumeux et Rainy pour pluvieux.

5.2.1 Le Modèle de Markov Caché

Le modèle de Markov caché est un modèle de Markov ou les états ne sont pas des évènements observables. Dans ce cas les états ont aussi des probabilités d'émissions des évènements observables. On se retrouve dans un processus stochastique double puisque nous avons la probabilité de transition entre les états et la probabilité d'émission d'évènements provenant de ces états. La figure 6 montre le même modèle de prévision météorologique que la figure 5 toutefois ici les évènements ne sont pas observables ; ils sont cachés. Les états sont bon, mauvais ou variable « Good, Bad ou Variable ». Il y a une probabilité d'émission des évènements Sunny, Rainy et Foggy pour chaque état. La séquence produite par ce modèle est une séquence d'évènements. La question est de savoir quelle séquence d'états a produit cette séquence d'évènements.

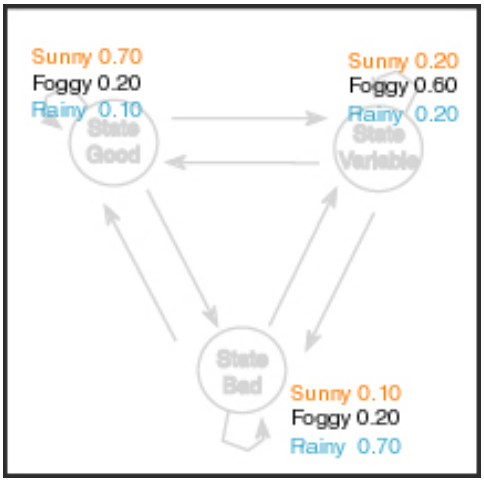


Figure 6 : Modèle de Markov caché. Les états sont Good pour bon; Variable et Bad pour mauvais. Les évènements Sunny, Foggy et Rainy sont émis par les états.

Lorsqu'on a un modèle de Markov caché correspondant à un ensemble de séquences, il y a trois problèmes d'intérêts.

1. **Le problème d'évaluation**: Étant donné les paramètres d'un MMC particulier et une séquence d'évènements, quel est la probabilité que cette séquence ait été générée par ce modèle. Ce problème est résolu par l'algorithme « Forward (Baldi et Brunak, 2001)».
2. **Le problème de décodage** : Étant donné les paramètres d'un MMC particulier et une séquence d'évènements, quel est la séquence d'états cachés la plus probable, pouvant générer cette séquence. Ce problème est résolu par l'algorithme « Viterbi (Forney, 1973) ».
3. **Le problème d'apprentissage** : Étant donné les paramètres d'un MMC particulier et une séquence d'évènements, quel ajustement devons-nous faire aux probabilités d'émissions d'évènements et de transitions d'états pour que le modèle corresponde le plus possible à la dite séquence d'évènements. Ce problème est résolu par l'algorithme « Baum-Welch algorithm (Baldi et Brunak, 2001)».

Les MMC peuvent aussi être décrits comme des grammaires stochastiques régulières. Dans ce cas, on doit simplement remplacer la probabilité de transition de l'état S_i vers l'état S_j pour l'émission de l'évènement X par la règle de production $S_j \rightarrow X\ S_i$ avec les probabilités associées (Baldi et Brunak, 2001).

5.3 Les hybrides RNA/MMC

Comme notre solution potentielle au problème de classification de signaux protéiques implique deux méthodes d'apprentissage, il est intéressant de voir les différentes applications de leur hybridation. Les

hybrides RNA/MMC sont assez communs dans la littérature scientifique depuis plus de 10 ans. Ils sont surtout vus dans des domaines comme la reconnaissance de l'écriture manuscrite (Bengio et al, 1995 ; Senior, 1994) et dans celui de la reconnaissance de la parole (Rigoll et Willett, 1998). Cette hybridation n'est pas étonnante car dans plusieurs cas il est intéressant de pouvoir combiner le pouvoir discriminant des RNA avec la capacité de modéliser des séquences détenue par les MMC.

L'hybridation est utilisée pour différentes raisons et sous différentes formes. Dans quelques architectures hybrides, les deux systèmes sont inséparables. Dans ces cas le RNA est utilisé pour paramétrer et moduler le MMC. Dans ces architectures l'apprentissage des deux systèmes est unifié (Baldi et Chauvin, 1996). Dans d'autres architectures les deux systèmes sont entraînés séparément. Ici le RNA peut être utilisé pour classifier les patrons de probabilités d'une séquence d'évènements produits par plusieurs MMC (Cho et Kim, 1995). Un autre exemple est l'utilisation du RNA pour l'estimation des probabilités à priori (Senior, 1994). En bioinformatique des travaux comme ceux de Martelli et collaborateurs (Martelli et al, 2002) utilise ce type de combinaison.

Pour notre projet, nous avons combiné les deux méthodes RNA et MMC entraînées séparément dans une cascade d'évènements.

6. Notre projet : Méthode hybride RNA/MMC de recherche de motifs protéiques

6.1 Méthodologie de recherche

Les efforts de séquençages de génomes ont amené une croissance exponentielle des données de séquences DNA, ARN et protéines. Il devient donc important de développer des outils pouvant annoter ces séquences de façon automatique et à grande échelle. La structure primaire des protéines, provenant de la traduction des gènes, n'est pas suffisante pour indiquer toute la complexité de leurs fonctions. Des modifications post-traductionnelles peuvent amener, par exemple, des changements d'activités, de localisation cellulaire et d'interaction avec d'autres protéines (Seo et Lee, 2003). Les modifications post-traductionnelles tel que l'ancrage GPI, ont une grande importance dans le processus de compréhension des fonctions biologiques toutefois leur étude souffre d'un manque de méthodes valables permettant l'étude à grande échelle (Mann et Jensen, 2003). La prédiction de modification post-traductionnelle des protéines fait partie intégrante d'une étude approfondie permettant la compréhension des fonctions biologiques. Elle se révèle être une étape importante non seulement pour l'annotation de protéomes mais aussi pour l'étude des systèmes biologiques à grande échelle.

Différents types de modification post-traductionnelle existent. Notons entre autres, la phosphorylation, l'acétylation et la glycosylation. L'ancrage GPI est une forme spéciale de glycosylation. Ce type d'attachement à la membrane cellulaire a jusqu'à maintenant été identifié seulement chez les eucaryotes et quelques archéobactéries (Ikezawa, 2002).

6.1.1 Structure des protéines à ancrage GPI

Les protéines attachées à la membrane par un ancrage GPI ne sont pas facilement identifiables par des méthodes d'analyses de séquences traditionnellement utilisées en bioinformatique. L'absence de constance dans les motifs composant le signal font que les analyses de similarités ne donnent pas de bons résultats.

Une protéine ayant un ancrage GPI est caractérisée par deux signaux. Le premier se retrouve dans la partie N-terminale (NH2) de la protéine (Figure 7a). Ce signal permet de diriger la molécule vers le réticulum endoplasmique pour la biosynthèse. Le second signal se retrouve dans la partie C-terminale (COOH) (Figure 7a). Ce signal, qui sera clivé lors de l'attachement de la protéine, se divise en quatre régions (Eisenhaber et al, 1998). (Figure 7b)

1. Une région non structurée d'environ 10 acides aminés.
2. Une région de petits d'acides aminés incluant le site d'attachement GPI.
3. Une région intermédiaire d'environ 7 acides aminés ayant au moins 3 résidus hydrophiles.
4. Une région terminale hydrophobe.

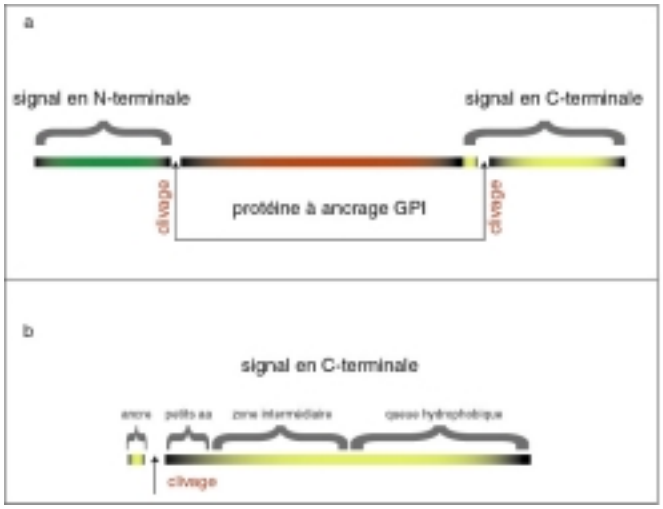


Figure 7 : Structure des signaux d'une protéine à ancrage GPI.
7a : signaux dans la partie N-terminale et en C-terminale.
7b : Structure du signal en C-terminale.

Ces règles semblent claires et précises, toutefois quelques séquences à ancrage GPI ne respectent pas ces règles. On retrouve aussi la possibilité de biosynthétiser la protéine sans le signal en

N-terminale (Howell et al, 1994). De plus on retrouve quelques d'exemples ou la région intermédiaire et la queue hydrophobe se chevauchent.

6.1.2 Méthodes de prédiction existantes

Il existe présentement deux outils de prédiction de protéines à ancrage GPI disponibles publiquement. Le plus utilisé est incontestablement big-π développé par le groupe du Dr Eisenhaber du IMP Bioinformatics Group de Vienne en Autriche (Eisenhaber et al, 1998; Eisenhaber et al, 1999; Sunyaev et al, 1999; Eisenhaber et al, 2000). Notons aussi DGPI développé par une équipe de l'université de Genève (Buloz et Kronegg, 1999). Big-π est très spécifique, toutefois cette forte spécificité diminue sa capacité de généralisation. De plus les deux outils demandent la présence du signal en N-terminale pour qualifier une protéine comme potentiellement GPI.

6.1.3 Matériel et méthode

6.1.3.1 L'architecture hybride RNA/MMC

Nous avons construit une architecture hybride réunissant deux techniques d'apprentissage: Premièrement un RNA pour prédire le potentiel signal GPI. Deuxièmement un MMC servant à raffiner la prédiction en la structurant. Cet outil ne cible que la partie C-terminale de la protéine, facilitant ainsi son utilisation pour des bases de données ayant des séquences plus ou moins complètes. De plus il n'est pas spécifique à un groupe taxonomique en particulier, rendant son utilisation plus universelle.

6.1.3.2 Le RNA

Les données d'entraînement utilisées sont composées des 50 dernières acides aminés en partie C-terminale de 79 séquences de protéines annotées, dans la base de données Swiss-Prot, comme ayant un ancrage GPI ainsi que 79 segments C-terminal de protéines n'ayant pas cette annotation. Dans le groupe d'entraînement, les séquences de protéine à ancrage GPI ont été sélectionnées principalement pour la qualité de leur annotation. Nous avons enlevé les séquences ayant une annotation ambiguë ainsi que les séquences ayant une partie C-terminale tronquée. Nous avons aussi prêté un grande attention à ne pas choisir des séquences ayant une forte similarité entre elles pour éviter des biais d'apprentissage en leurs faveur. Pour l'encodage des données l'importance de l'hydropathie et du poids moléculaire pour les protéines à ancrage GPI, nous a fait choisir ces caractéristiques comme valeurs numériques des neurones. Pour l'hydropathie l'échelle de Kyte et Doolitle a été sélectionnée (Kyte et Doolittle, 1982).

L'architecture du RNA est de type perceptron multicouche. L'apprentissage est de type RPROP (Resilent back propagation) (Riedmiller et Braun 1992, 1993). La couche d'entrée est composée de 100 neurones,

correspondant aux deux valeurs octroyées aux 50 acides aminés. La couche cachée est composée de 150 neurones tandis que la couche de sortie contient un seul neurone (Figure 8). Le processus d'apprentissage consiste à graduellement ajuster le poids des connexions en vu d'atteindre le score optimal affecté aux séquences ayant un ancrage GPI. Un score plus grand ou égale à 0.90 indique que le RNA a identifié une séquence potentiellement à ancrage GPI. Le seuil d'acceptation du score est sélectionné à l'aide d'une courbe ROC (Receiver Operating Characteristic) (Maloof, 2002). Le RNA a été simulé grâce au simulateur développé par l'Université de Stuttgart (JavaNNS) (Zell et al, 2002)

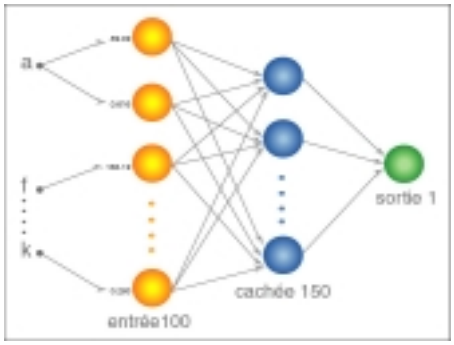


Figure 8 : Architecture du perceptron multicouche construit pour la classification de protéine à ancrage GPI.

6.1.3.3 Le MMC

Le MMC [1] reflète notre connaissance de la structure des protéines à ancrage GPI. L'architecture de l'automate peut être divisée en trois parties (Figure 9). Les trois premiers états correspondent aux petits résidus près du site de clivage. Par la suite une transition vers les 10 prochains états est possible (état 4 à 13). Cette section correspond au début de la partie intermédiaire. Chaque état situé dans cette partie a une possibilité de transition vers les autres états. L'état 13 représente le début d'une zone linéaire de transition composée de 3 états (14 à 16). Cette zone ferme la partie intermédiaire du signal. Cette zone intermédiaire peut ainsi être composée de 4 à 13 acides aminés. Finalement à partir de l'état 16, des transitions sont possibles pour 20 états (17 à 36) représentant le début de la partie hydrophobe du signal. Chacun de ces 20 états ont une possibilité de transition vers les autres états. L'état 36 correspond au début d'une autre zone linéaire de 5 états (37-41) composant la fin de la zone hydrophobe. Cette zone hydrophobe peut donc être composée de 6 à 35 acides aminés. L'état final est un état particulier qui n'émet pas d'aide aminé. Pour pouvoir terminer, chaque séquence doit passer par ce stade. De cette façon la longueur de la séquence est prise en compte dans le processus d'affectation du score permettant ainsi de ne pas biaiser les résultats en faveur des séquences de courtes longueurs.

[1] L'architecture du MMC a été développée en collaboration avec Anne Bergeron et Cedric Chauve. La programmation ainsi que l'optimisation ont été réalisées par Cedric Chauve.

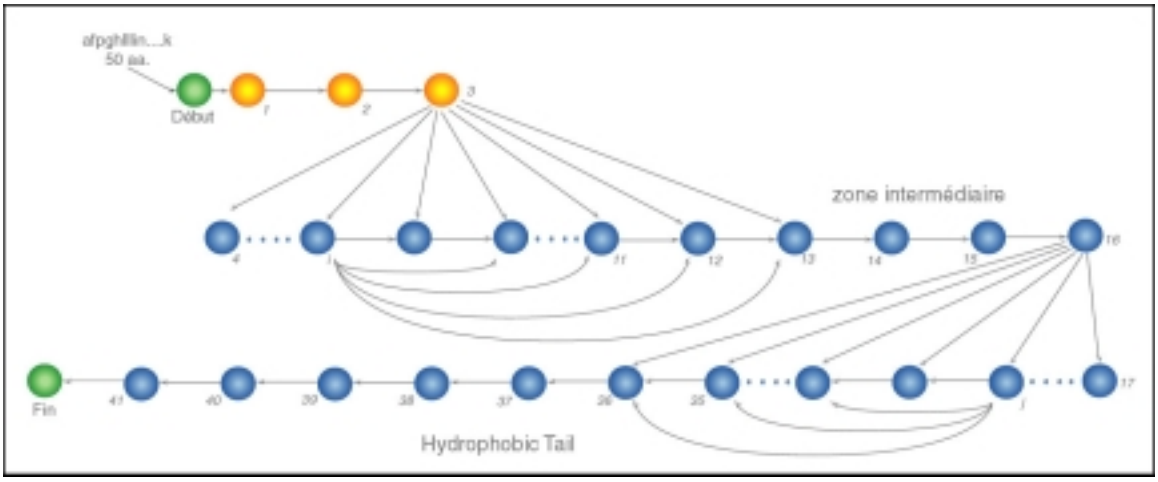


Figure 9 : Structure du modèle de Markov cachée.

Les probabilités de transition initiales sont définies selon le nombre de transitions possibles pour chaque état. Si un état à *k* transitions vers d'autres états, chaque transition aura une probabilité de *1/k*. Les probabilités d'émissions des états ont été affectées à chaque acide aminé selon une analyse de la répartition des acides aminés dans les séquences connues de protéines à ancrage GPI. Pour l'apprentissage un groupe de séquences d'entraînement composé de 87 séquences annotées comme ayant un ancrage GPI a été constitué. Le processus d'apprentissage a nécessité 100 itérations de l'algorithme de Baum-Welch.

Pour l'affectation des scores les sections d'une séquence débutant par au moins 3 acides aminés de faible poids moléculaire et ayant au minimum 15 acides aminés de longueur ont été présenté au MMC et évalué selon l'algorithme Forward et le principe du « log-odd » (Krogh, 1998). Le meilleur score obtenu est gardé comme le score du segment. Par la suite, l'algorithme Viterbi a été utilisé pour trouver le chemin le plus probable et ainsi nous donner la structure du signal (site d'ancrage, zone intermédiaire et queue hydrophobe). Le seuil d'acceptation d'une protéine à ancrage GPI est sélectionné à l'aide d'une courbe ROC (seuil 0.539). Pour la prédiction du site d'ancrage, les 3 meilleurs scores obtenus pour chaque séquence, représentent les sites d'ancrage potentiels.

6.1.4 Méthode de validation

6.1.4.1 Validation du RNA

La validation du RNA a été effectué en utilisant un groupe de données test contenant 134 séquences de protéines à ancrage GPI ayant dans la majorité des cas une qualité d'annotation supérieure et 134

séquences de protéines autres (cytoplasmiques, transmembranaires, aléatoires, etc.) Une méthode d'évaluation de la performance de généralisation d'un classificateur est d'effectuer une mesure de qualité des prédictions. Pour un classificateur à deux classes (vrai ou faux) la performance peu être évaluée grâce à la sensibilité, la spécificité, la valeur de prédiction positive, la valeur de prédiction négative, la précision et le coefficient de corrélation (tableau 1) (Wu et McLarty, 2000). Un autre test d'évaluation de l'efficacité de plus en plus utilisé en apprentissage machine est l'analyse de la courbe ROC. Ce type d'analyse permet de connaître le taux de vrai positif ainsi que le taux de faux positif à différents seuils de confiance. Une façon de connaître une mesure de la performance des prédictions est de donner une approximation de l'aire sous la courbe en utilisant la règle du trapèze. Plus l'aire tend vers 1 plus la performance du modèle est élevée.

TN = Vrai négatif, **TP** = Vrai positif, **FP** = Faux positif, **FN** = Faux négatif	
Précision	TP+TN/Total
Sensibilité	TP/(TP+FN)
Specificité	TN/(FP+TN)
Valeur de prédiction positive	TP/(TP+FP)
Valeur de prediction négative	TN/(FN+TN)
Coefficient de corrélation	(TPxTN-FPxFN)/√(TP+FP)x(FP+TN)x(TN+FN)x(FN+TP)

Tableau 1 : Tests de performance.

6.1.4.2 Validation du MMC

Un groupe test de 69 séquences annotées comme ayant un ancrage GPI ainsi que 69 séquences de protéines n'ayant pas cette annotation ont servi à valider le MMC. De plus nous avons effectué 500 expériences de bootstrap non paramétrique sur notre groupe de données d'apprentissage. Pour chacune des 500 expériences, un échantillonnage de 87 séquences (répétitions permises) a servi à construire un nouveau groupe d'apprentissage. Le MMC a ensuite été entraîné avec ces nouveaux groupes de données d'apprentissage selon la même démarche que pour le modèle initiale. Chaque « nouveau » MMC a par la suite été testé à l'aide du groupe test.

6.1.5 Le système hybride

Dans le système hybride les candidats sélectionnés par le RNA sont présentés au MMC. Le score obtenu pour chaque candidat sert à qualifier la prédiction selon une échelle de probabilité. Cette échelle s'échelonne de la catégorie hautement probable jusqu'à la catégorie faiblement probable ou potentiellement faux positif. Les catégories sont définies à l'aide d'une courbe ROC. L'analyse du chemin le plus probable obtenue grâce à l'algorithme Viterbi, nous donne aussi la structure potentielle du signal (Figure 10).

Figure 10 : Exemple de structure du signal GPI prédite par le système hybride.

6.2 Résultat et Discussion

6.2.1 Le RNA

Une fois le RNA bien entraîné, nous avons soumis le modèle au groupe de séquences tests contenant 134 séquences d'ancre GPI ainsi que 134 séquences Non GPI. La performance du RNA est démontrée dans le tableau 2.

TN = Vrai négatif, **TP** = Vrai positif, **FP** = Faux positif, **FN** = Faux négatif	
Précision	0.93
Sensibilité	0.90
Specificité	0.97
Valeur de prédiction positive	0.97
Valeur de prediction négative	0.90
Coefficient de corrélation	0.87

Tableau 2 : Test de performance du RNA

Ce test donne une sensibilité de 90% et une spécificité de 97%. Une analyse ROC donne une aire sous la courbe de 0.97. Ce résultat indique une bonne performance de notre classificateur.

Comme l'échantillonnage du test est plutôt restreint comparativement à un protéome entier nous avons effectué d'autre tests ciblant la totalité des protéines annotées comme étant à ancre GPI ainsi que des groupes particuliers ayant une très faible probabilité de contenir des séquences de protéines à ancrage GPI. Comme les protéines à ancrage GPI sont uniquement extracellulaires, des protéines cytoplasmiques et nucléaires sont de bon choix de séquences Non GPI. Comme l'hydrophobicité est une caractéristique importante chez les protéines à ancrage GPI, nous avons voulu tester des séquences ayant aussi cette particularité. Nous avons donc sélectionné des séquences protéiques transmembranaires et quelques protéines de transport pour nos groupes tests Non GPI (tableau 3).

a: Séquences Non GPI

Tests	Nombre de séquences	RNA Prédiction
Cytoplasmic_Nuclear_SP	111	0,02
Transmembrane_SP	182	0,06
Transport_Protein_SP	83	0,07
Random	2445	0,04

b: Séquences GPI

Tests	Nombre de séquences	RNA Prédiction
GPI_Swiss-Prot_TOTAL	468	0,93

Tableau 3 : Résultats du RNA aux des tests **3a** non GPI. **3b** GPI.

Il devient donc intéressant de noter que le classificateur issue uniquement d'un modèle de réseau de neurone est très sensible au signal d'ancrage GPI avec une sensibilité de 93% mais la spécificité du modèle est un peu trop basse avec une moyenne de 95.25% donc 4.75% de faux positif . Notons ici que cette spécificité diminue chez des séquences hautement hydrophobiques comme certaines séquences transmembranaires (7% de faux positif). Ceci n'est pas étonnant puisque la région hydrophobe du signal GPI est similaire au domaine transmembranaire hydrophobe (Dalley et Bulleid, 2003). Des tests de mutations dans la partie du site d'ancrage ainsi que dans la région de transition démontrent que le RNA est peu sensible à ces modifications. Ceci explique le plus haut taux de faux positif retrouve chez des séquences hautement hydrophobes. Toutefois le RNA a su caractériser une hydrophobicité GPI, c'est-à-dire que le RNA ne sélectionne pas toutes les séquences hautement hydrophobes. Un problème toutefois se pose, la structure opaque de notre RNA rend impossible la structuration du signal GPI. Il nous est donc impossible, à ce stade, de donner l'information du positionnement du site d'ancrage. Le haut taux de faux positif ainsi que l'opacité du système amène la nécessité de mieux structurer la prédiction à l'aide d'une méthode qui tirera avantage de la nature séquentielle des données.

6.2.2 Le MMC

Une fois le MMC entraîné nous lui avons présenté les 69 séquences annotées comme ayant une ancre GPI ainsi que les 69 séquences n'ayant pas cette annotation (tableau 4). Ce test montre une sensibilité de 97% et une spécificité de 95 %. Une analyse ROC donne une aire sous la courbe de 0.98. Cette valeur indique que le système de prédiction a une performance élevée. De plus le test de bootstrap donne en moyenne une sensibilité de 91% et une spécificité de 98%.

TN = Vrai négatif, **TP** = Vrai positif, **FP** = Faux positif, **FN** = Faux négatif	
Précision	0.96
Sensibilité	0.97
Specificité	0.95
Valeur de prédiction positive	0.96
Valeur de prediction négative	0.97
Coefficient de corrélation	0.93

Tableau 4 : Tests de performance du MMC

Vu le petit échantillonnage de notre test nous avons, comme pour le RNA, présenté au MMC les tests à grande échelle (tableau 5). En moyenne le MMC donne un moins grande sensibilité que le RNA avec 88% mais la spécificité elle augmente avec 98%.

a: Séquences Non GPI

Tests	Nombre de séquences	MMC seuil 0.539
Cytoplasmic_Nuclear_SP	111	0,00
Transmembrane_SP	182	0,05
Transport_Protein_SP	83	0,02
Random	2445	0,002

b: Séquences GPI

Tests	Nombre de séquences	MMC seuil 0.539
GPI_Swiss-Prot_TOTAL	468	0,88

Tableau 5 : Résultats du MMC aux tests **5a** non GPI. **5b** GPI.

Avec le MMC il nous est possible de prédire la position de l'ancrage GPI. Pour 347 séquences ayant un site annoter le MMC a pu prédire correctement 75% des sites d'ancrage. La plupart des sites incorrectement annotés sont toutefois en moyenne un ou deux acides aminés de distance du vrai site d'ancrage.

6.2.3 Comparaison entre RNA et MMC

Comme l'apprentissage du MMC et du RNA ne requière pas le même échantillonnage, les tests de performances n'étaient pas équivalents pour les deux systèmes. Ces tests de performances ne peuvent donc être comparés. De plus la taille restreinte de ces tests amène un biais non négligeable dans leur évaluation. Il est donc beaucoup plus justifiable d'utiliser les tests à grande échelle pour comparer les deux systèmes.

Pour bien comparer les deux modèles. Le seuil d'acceptation du MMC a été diminué pour augmenter la sensibilité du modèle. Pour obtenir une sensibilité générale de 93% nous avons du accepter un seuil de -4.50 (tableau 6). Ce seuil négatif n'est toutefois pas idéal car il indique un score inférieur à une hypothèse nulle.

Avec un seuil de -4.50 la spécificité du MMC se voit diminué en moyenne à 94.75% donc un taux de faux positif de 5.25% ce qui est plus élevé que le RNA. Fait étonnant, le MMC est beaucoup moins sensible à l'hydropathie particulière du signal GPI avec un taux de faux positif de 11% chez les transmembranaires. Le RNA a donc un pouvoir discriminant beaucoup plus intéressant que celui du MMC quand à l'hydropathie du signal GPI. Cependant le pouvoir structurant du MMC doit être exploité.

a: Séquences Non GPI

Tests	Nombre de séquences	MMC seuil 0.539	MMC seuil -4.5	RNA seuil 0.9
Cytoplasmic_Nuclear_SP	111	0,00	0,02	0,02
Transmembrane_SP	182	0,05	0,11	0,06
Transport_Protein_SP	83	0,02	0,06	0,07
Random	2445	0,002	0,02	0,04

b: Séquences GPI

Tests	Nombre de séquences	MMC seuil 0.539	MMC seuil -4.5	RNA seuil 0.9
GPI_Swiss-Prot_TOTAL	468	0,88	0,93	0,93

Tableau 6 : Comparaison entre le RNA et le MMC. **6a** Résultats des tests non GPI. **6b** Résultat des tests GPI.

6.2.4 Système hybride

Les scores des candidats à un ancrage GPI obtenus à la sorite du système hybride, s'échelonnent de -147.56 à 34.24. Les séquences ayant les plus haut scores auront une annotation comme séquences à ancrage GPI fortement probable tandis que celles ayant les plus bas scores seront annotées comme séquences potentiellement faux positifs. Cette étape reste à faire.

6.2.5 Comparaison avec big-π et DGPI

La comparaison entre notre système hybride et les autres outils reste à être finaliser toutefois une analyse préliminaire donne un outil combinant le pouvoir prédictif des deux autres systèmes. Dans la catégorie supérieure de notre annotation, qui n'accepte que les candidats respectant la structure connue du signal GPI, on retrouve une sensibilité supérieure et une spécificité comparable aux deux autres systèmes combinés. Les catégories suivantes permettent de cibler les candidats moins typiques qui échappent aux deux autres outils. De plus la présence du signal en N-terminale est une contrainte qui nous avons éliminée contrairement à DGPI qui demande une protéine complète et à big-π qui lui demande de vérifier la présence du signal en N-terminale avant de présenter les séquences au système. Cette contrainte n'est pas justifiée vu la découverte de séquences GPI n'ayant pas ce signal, tel que la P137, (Ellis et Lazio, 1995) ainsi que la capacité de biosynthèse de la protéine sans ce signal (Howell et al, 1994).

6.2.6 Une grammaire du signal GPI

Grâce à notre MMC final il nous sera possible de proposer une grammaire régulière stochastique représentant le signal GPI. Ce travail reste à faire

6.2.7 Une mise à jour de nos connaissance sur le signal GPI

L'analyse des résultats ainsi que l'analyse de notre modèle MMC pourra permettre de réévaluer nos connaissances relatives à la structure des protéines à ancrage GPI. Ce travail reste à faire.

7. Conclusion

Certains voient les séquences biologiques tel un langage : le langage des gènes. La séquence de lettres représentant une protéine peut donc être vu tel un texte ou les motifs et les signaux sont des mots clés permettant de bien comprendre le sens de ce texte. Chercher à comprendre le texte d'une protéine consiste à comprendre sa fonction et à « visualiser » son état. La complexité des séquences biologiques rend l'annotation de séquences biologiques, à grande échelle, difficile et demande beaucoup de ressources et de temps à un expert. L'utilisation de modèles d'apprentissage artificiels rend cette tâche réalisable.

Dans ce projet nous avons voulu utiliser la capacité d'apprentissage du RNA principalement pour son pouvoir discriminant mais aussi pour nous aider à trouver ou confirmer un lien commun entre les séquences de protéines à ancrage GPI. Il est maintenant très clair que l'hydropathie de ces séquences est très importante, ce qui était déjà bien accepté. Toutefois notre analyse suggère qu'une hydropathie particulière aux GPI existe car le RNA obtient des résultats satisfaisants dans la tâche de classification de ces séquences. Il reste maintenant à caractériser cette hydropathie particulière. Comme les séquences biologiques sont représentées sous forme de séquences de lettres, l'utilisation de modèle de grammaire régulière stochastique, tel que les MMC, se voit très efficace pour trouver une structure du langage régulier représentant le signal d'ancrage GPI. Cette étape nous permet de cibler la zone d'ancrage et la partie la plus hydrophobe du signal. Le MMC correspondant au signal GPI nous permettra de réévaluer nos connaissances sur la structure de ce signal.

L'annotation de séquences biologiques, selon la qualité de la structuration du signal, permet de cibler des candidats plus typiques mais aussi ceux qui sortent un peu des limites de la structure du MMC. Ce nouveau type de classificateur de signaux protéiques s'avère très efficace car dans les catégories supérieures il atteint la spécificité des outils les plus strictes et dans les catégories inférieures il laisse place à la

découverte de protéines à ancrage GPI ayant un signal moins spécifique. Cette combinaison de reconnaissance et de structuration du signal s'avère très efficace et pourra être utilisée ultérieurement pour la prédiction de différents types de modifications post-traductionnelles ou autre signaux protéiques.

8. Volet cognitif

Ce projet de recherche se situe dans le domaine des techniques informatiques pour l'extraction des connaissances. Les séquences biologiques contiennent des connaissances qui sont le plus souvent cachées voir bruitées. L'utilisation de technique d'apprentissage tel que les RNA offre un énorme avantage dans la phase de « nettoyage » des données. Cette méthode de modélisation nous offre la possibilité d'automatiser le travail d'un expert en simulant la classification qu'il ferait conte tenu de l'information existante tout en augmentant la performance de la tâche par l'utilisation d'informations inaccessibles à l'œil de l'expert.

De plus notre système mise sur la structure retrouvée dans ces séquences. Une particularité des séquences biologiques est leur représentation sous forme de texte fait à partir d'un alphabet bien précis. Cette « nature » linguistique pointe vers l'utilisation des grammaires tel que les grammaires régulières stochastique (MMC) pour l'extraction des connaissances qu'elles contiennent sous forme de motifs ou signaux. Grâce à une forme certaine de langage cellulaire, ces motifs nous renseignent sur le rôle joué par ces séquences dans le fonctionnement de la cellule ou nous indique son état à un moment précis face à une situation particulière. La définition d'une grammaire les représentant pourra sûrement aider à l'établissement d'un langage cellulaire global plus précis qui, dans l'avenir, servira à mieux comprendre le fonctionnement de tout être vivant.

9. Contribution originale

La contribution originale de ce projet est la conception d'un outil de prédiction d'un signal protéique important pour l'annotation des protéomes et pour l'étude fonctionnelle. Nous proposons aussi une grammaire relative au signal GPI ainsi qu'une réévaluation de nos connaissances sur sa structure. L'utilisation de la méthode d'apprentissage neuronale permet une bonne fouille préliminaire des données. L'utilisation de la nature régulière du langage des séquences biologiques sert à structurer les prédictions mais aussi à annoter chaque prédiction selon la qualité de structuration du langage. Un système hybride RNA/MMC donne donc un outil plus complet que ceux déjà existant et ouvre les portes à d'autres applications en analyses de signaux protéiques.

10. Avancement des travaux

10.1 Travail effectué

Échéancier :

Automne 2001 :

-Début du doctorat.

-Cours préparatoires.

Hiver 2002 :

-Premier prototype du système.

-Récolte des données nécessaires au projet. Ces données furent sélectionnées parmi les séquences disponibles publiquement dans les bases de données de séquences biologiques.

-Développement du premier prototype constitué d'un unique réseau de neurones artificiel.

Été 2002 :

-Stage chez Warnex inc.

Automne 2002 à Hiver 2004 :

-Poursuite des cours.

-Amélioration du classificateur.

-Collaboration avec le laboratoire du Dr. Fathey Sarhan (département de Biologie UQAM) et le laboratoire du Dr. Patrick Gulick (département de Biologie Université Concordia) pour un projet d'analyse de données de micropuces d'ADN. Ma collaboration fut dans l'analyse des données de micropuces d'ADN à l'aide de méthodes statistiques.

-Présentation des résultats par les gens de biologie au « 7 th International Congress of Plant Molecular Biology, Barcelona, Spain, June 2003» [2] (Voir Annexe A)

-Soumission d'un article au journal Plant Physiology [3].

-Collaboration avec Cédric Chauve pour le développement d'un modèle de Markov Caché effectuant la même tâche de classification.

-Travail avec deux stagiaires durant l'été 2003 et l'automne 2003.

-Étude comparative des classificateurs.

-Présentation des résultats de l'étude comparative au congres «Human Proteome Organisation, HUPO 2003 » [4]. (Voir Annexe B)

---

[2] A. Monroy, G. Poisson, S. Drouin, F. Sarhan, P. Gulick (2003). *Gene Expression Profiling During Cold Acclimatation in Wheat.,* 7 th International Congress of Plant Molecular Biology, Barcelona, Spain, June 2003

[3] P. Gulick, J. Danyluk, S. Drouin, A. Monroy, G. Poisson, A. Bergeron, and F. Sarhan. (2003). *Genotypic comparison of low temperature responsive genes in wheat using microarray analysi*s.

-Publication de la comparaison des deux classificateurs dans le journal Molecular & Cellular Proteomics [5]. (Voir Annexe C)

-Design d'un prototype de système hybride RNA/MMC.

-Confection du système hybride

-Préparation de la publication des résultats.

-Invitation pour présentation des résultats au « Information and Computer Science department » de l'université d'Hawaii à Manoa [6]

10.2 Travail à venir

Échéancier :

Hiver 2004- Été2004:

-Pointage flou.

-Comparaison avec big-π.

-Construction de la grammaire régulière stochastique représentant le signal GPI et mise à jour de la structure connue du signal.

-Analyse du protéome connus d'*Arabidopsis taliana*.

-Tests en laboratoire de certaines prédictions.

-Soumission d'un article portant sur le système hybride.

-Rédaction de thèse et dépôt.

[4] G. Poisson, A. Bergeron, C. Chauve et P. Simard (2003). *Prediction of post-translational GPI anchor modification of protein by machine learnin*g., Human Proteome Organisation, HUPO 2003

[5] G. Poisson, A. Bergeron, C. Chauve et P. Simard (2003). *Prediction of post-translational GPI anchor modification of protein by machine learnin*g., Molecular & Cellular Proteomics. Human Proteome Organisation, HUPO 2003 Special, 2(9):826

[6] G. Poisson (2004) *Artificial Neural Network and Hidden Markov Model for GPI_anchored Protein Prediction* ,Information and Computer Sciences Department University of Hawaii at Manoa.

II. Bibliographie

Ali S., Hall J., Hazelwood G.P., Hirst B.H. and Gilbert H.J. (1996) *A protein targeting signal that functions in polarized epithelial cells in vivo* Biochem. J., 315: 857-862

Baldi P. and Chauvin Y. (1996) *Hybrid modeling, HMM/NN architectures, and protein applications*. Neural Computation., 8(7):1541--1565

Baldi P. and Brunak S. (2001) *Bioinformatics, The Machine Learning Approach* – 2nd ed., The MIT Press.

Barrette I., Poisson G., Gendron P., and Major F. (2001) *Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching.* Nucl. Acids Res., 29(3):753-8.

Bateman A., Coin L., Durbin R., Finn R.D., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E.L.L., Stud holme D.J., Yeats C. and Eddy S.R. *The Pfam proteins Families database*. Nucl. Acids Res., (2004) Database Issue 32:D138-D141

Bengio, Y., LeCun, Y., Nohl, C. and Burges, C.A C. (1995) *LeRec: A NN/HMM Hybrid for On-Line Handwriting Recognition*. Neural Computation., 7(6):1289-1303

Betel, d. et Hogues, C.W.V. (2002) *Kangaroo – A pattern-matching program for biological sequences.* BMC Bioinformatics., 3(1):20

Brejova,B., DiMarco,C., Vinar,T., Hidalgo,S.R., Holguin,C. and Patten,C. (2000) *Finding patterns in biological sequences*. Project Report for CS798g. University of Waterloo.

Brunak S. J. Engelbrecht et S. Krudsen (1990) *Cleaning up gene database*. Nature., 343 :123

Buloz, D. et Kronegg, J. (1999) *Détection/prédiction de site de clivage GPI (GPI-anchor) dans une protéine.* http://129.194.185.165/dgpi/DGPI_demo_en.html

Burge, C. et Karlin, S. (1997) *Prediction of complete gene structures in human genomic DNA*. Journal of Molecular Biology., 268 : 78-94

Chomsky , N. (1957) *Syntaxic Structures*. Moutn, The Hague.

Cornuéjols A. et Miclet L.(2002) Apprentissage artificiel : Concept et algorithmes. Eyrolles

Dalley J.A. et Bulleid N.J. (2003) *How does the translocon differentiate between hydrophobic sequences that form part of either a GPI (glycosylphosphatidylinositol)-anchor signal or a stop transfer sequence?* Biochem. Soc. Trans., 31:1257–1259

Dayhoff, M.O., Eck, R.V., Chang, M.A., and Sochard, M.R. (1965). *Atlas of Protein Sequence and Structure Vol. 1.* National Biomedical Research Foundation, Silver Spring, MD

Durbin, R., Eddy, S., Krogh, A., et Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press.

EBI (European Bioinformatics Institute) (2004) *SWISS-PROT/TrEMBL database*. http://www.ebi.ac.uk/swissprot/access.html

Eddy. S.R. (1995*) Multiple alignment using hidden Markov models.* Proc Int Conf Intell Syst Mol Biol. 3:114-20.

Eisenhaber, B., Bork, P. et Eisenhaber, F. (1998). *Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase.* Protein Engineering., 11(12): 1155-1161

Eisenhaber, B., Brok, P. et Eisenhaber, F. (1999) *Prediction of potentiel GPI-modification Sites in Proprotein Sequences.* J. Mol. Biol., 292: 741-758

Eisenhaber, B., Brok, P., Yuan, Y., Löffler, G. et Eisenhaber, F. (2000) *Automated annotation of GPI anchor sites:case study C. elegans.* TIBS., 25: 340-341

Ellis J.A. and Luzio J.P. (1995) *Identification and Characterization of a Novel Protein (p137) Which Transcytoses Bidirectionally in Caco-2 Cells.,* 270(35): 20717-20723

Forney, G.D. Jr.(1973) *The Viterbi algorithm* Proc. of the IEEE., 61(3): 268-278.

Francesco V.D., Garnier J. and Munson P.J. (1997) *Protein topology recognition from secondary structure sequences-Applications of the hidden Markov models to the alpha class proteins*. J.Mol. Biol., 267:446-463.

Gattiker A., Gasteiger E. and Bairoch A.(2002) *ScanProsite: a reference implementation of a PROSITE scanning tool.* Applied Bioinformatics., 1: 107-108

Habrias, H (2002) *Génie logiciel Module de spécification 2*. IUT Université Nantes.

Haykin, S. (1994) *Neural Networks, A comprehensive foundation.* MacMillan College Publishing New York.

Holmström, L., Koistinen, P., Laaksonen, J. Et Oja, E. (1996) *Comparison of Neural and Statistical Classifiers—Theory and Practice..* Rolf Nevanlinna Institute Research Reports A13, Helsinki

Howell S, Lanctot C, Boileau G, Crine P. (1994) *A cleavable N-terminal signal peptide is not a prerequisite for the biosynthesis of glycosylphosphatidylinositol-anchored proteins*. J Biol Chem., 269(25):16993-6.

Ikezawa, H. (2002) *Glyscosylphosphatidylinositol (GPI)-Anchored Proteins*. Boil. Pharm. Bull., 25(4) : 409-417

Jain, A.K., Duin, R.P.W. et Mao, J. (2000) *Statistical Pattern Recognition : A Review.* IEEE Transactions on Pattern Analysis and Machine Intelligence., 2(1): 4-37.

Krogh A., Mian I.S. and Haussler D. (1994) *A hidden Markov model that finds genes in E.coli DNA*. Nucl. Acid. Res., 22:4768-4778.

Krogh A. (1998) *An Introduction to Hidden Markov Models for Biological Sequences*. in S.L. Salzberg et al., eds., Computational Methods in Molecular Biology., 45-63 Elsevier.

Kulp D, Haussler D, Reese MG, Eeckman FH. (1996) *A generalized hidden Markov model for the recognition of human genes in DNA*. Proc Int. Conf. Intell. Sys. Mol. Biol., 4: 134-42

Kyte, J., and Doolittle, R. F. (1982) *A simple method for displaying the hydropathic character of a protein*. J. Mol. Biol., 157 :105-132

Lemaire, V. (1999) *Une nouvelle fonction de coût régularisante dans les réseaux de neurones artificiels : Application à l'estimation des temps de blocage dans un nœud ATM*. Thèse de doctorat, Université Paris VI

Low, M.G. et Zilversmit, D.B. (1980) *Role of phosphatidylinositol in attachment of alkaline phosphatase to membranes.* Biochemistry., 19: 3913-3918.

Maloof, MA. (2002) *On machine learning, ROC analysis, and statistical tests of significance.* Proceedings of the Sixteenth International Conference on Pattern Recognition., 204-207, Los Alamitos, CA: IEEE Press.

Mann M. and Jensen O. (2003) *Proteomic analysis of post-translational modifications*. Nature Biotechnology., 21:255-261

Martelli PL, Fariselli P, Malaguti L, Casadio R. (2002) *Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks.* Protein Eng., 15(12):951-3

Mendel, Gregor. (1866). Versuche über Plflanzen-hybriden. *Verhandlungen des naturforschenden Ver-eines in Brünn, Bd. IV für das Jahr 186*5, Abhand-lungen., 3–47.

Nakata, K. (1995) *Prediction of zinc finger DNA binding protein.* Comput Appl Biosci., 11 : 125-131

Nielsen H., Engelbrecht J., Brunak S. and von Heijne G. (1997A) *A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites*. Int. J. Neural Sys., 8: 581-599

Nielsen H., Engelbrecht J., Brunak S. and von Heijne G. (1997B) *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites*. Protein Engineering., 10, 1-6

Nosjean, O., Briolay, A. et Roux B. (1997) *Mammalian GPI proteins : sorting, membrane residence and functions.* Biochim. et Biophys. Acta., 1331: 153-186.

Orr, G. et Cummins, F. (1999) *Neural network : lecture notes*. Willamate University Oregon.

Riedmiller, M. et Braun, H. (1992) *RPROP A Fast Adaptive Learning Algorithm.* Proceeding of ISCIS VII

Riedmiller, M. et Braun, H. (1993) *A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm* Proc. of the IEEE Intl. Conf. on Neural Networks

Rigoll G. and Willett D. (1998) A NN/HMM Hybrid for Continuous Speech Recognition with a Discriminant Nonlinear Feature Extraction. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 9-12, Seattle

Ross, B. et Sander, C. (1994) *Conservation and prediction of solvent accessibility in protein families.* Proteins., 20: 216-226

Rumelhart, DE. et McClelland, JL (1986) *Parallel distributed processing exploration in the microstructure of cognition.* A Bradford Book MIT Press, Cambridge (MA) USA

Sanger, F., Thompson, E. O. P., and Katai, R.(1955) *The amide groups of insulin*. Biochem J., 59:509-514,

Sarle, W.S. (1994), *Neural Networks and Statistical Models*. Proceedings of the Nineteenth Annual SAS Users Group International Conference, Cary, NC, SAS Institute., pp:1538-1550

Searls, D.B. (1993) *The Computational Linguistics of Biological Sequences.* In Artificial Intelligence and Molecular Biology (L. Hunter, ed.), AAAI Press., chapter 2: 47-120.

Searls, D.B. and Dong, S. (1993) *A Syntactic Pattern Recognition System for DNA Sequences* in Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis (H.A. Lim, J. Fickett, C.R. Cantor, and R.J. Robbins, eds.), World Scientific., pp: 89-101.

Senior A. (1994). *Off–Line Cursive Handwriting Recognition using Recurrent Neural Network*s. PhD thesis, University of Cam-bridge, Cambridge, England

Seo J. and Lee K-J (2004) *Post-translational Modifications and Their Biological Functions: Proteomic Analysis ans Systematic Approches.* J. Biochemistry and Molécular Biolgy., 37(1):35-44

Siveton V. (2002) *Extraction de connaissances hydriques au moyen de diverses méthodes supervisées* Mémoire de maîtrise, UQAM

Sonnhammer E.L.L., Eddy S.R., and Durbin R. (1997) *Describes the Pfam database of multiple sequence alignments and HMMs, and its use in large scale genome analysis*. Proteins., 28:405-420.

Sung-Bae Cho and Jin H. Kim. (1995) *An HMM/MLP architecture for sequence recognition*. Neural Comp., 7:358-369

Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G., and Kuznetsov, E.N. (1999) *PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations*. Protein Engineering., 12(5) : 387-394

Quian, N. et Sejnowski, TJ. (1988) *Predicting the secondary structure of globular proteins using neural network models.* J. Mol. Biol., 202 : 865-884

Wu, C., Whitson, G., McLarty, J., Ermongkonchai, A. et Chang, TC. (1992) *Protein classification artificial neural system.* Protein Sci., 1(5) : 667-677

Wu, C. et McLarty, JW. (2000) *Neural networks and genome informatics* Methods in computational biology and biochemistry 1, Elsevier publishing NewYork

Xuan, Z., McCombie, W.R. et Zhang, M.Q. (2002) *GFScan: A Gene Family Search Tool at Genomic DNA Level*. Genome Research., 12 : 1142-1149

Zell A., et al (2002) JavaNNS 2002, *Stuttgart Neural Networks Simulator (SNNS).* Stuttgart University http://www-ra.informatik.uni-tuebingen.de/

III. Annexe

Annexe A

Présentation des résultats au « 7 th International Congress of Plant Molecular Biology, Barcelona, Spain, June 2003»

Annexe B

Présentation des résultats de l'étude comparative au congres «Human Proteome Organisation, HUPO 2003 »

Annexe C

Publication de la comparaison des deux classificateurs dans le journal Molecular & Cellular Proteomics.