الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

présenté par

BENNAI KHEIRA

&

MAHDI DJAZIA

Pour l'obtention du diplôme master deux en Électronique option Traitement de l'information & Système électronique

Thème

Recherche de motifs protéiques par les profils des modèles de Markov cachés

Proposé par : Mme Ait Abdesslam Houria

Année Universitaire 2014-2015

Remerciements

Nous tenons avant tout de remercier le bon DIEU qui nous a données la volonté et le courage pour la réalisation de ce travail.

Nous remercions vivement M_{me} Ait abdesslam Houria notre promotrice pour la précieuse assistance, sa disponibilité et son soutien qu'elle nous a accordé tout au long de ce projet.

Nos remerciement les plus vifs s'adressent aussi a messieurs le président et les membres de jury d'avoir accepté d'examiner et d'évaluer notre travail.

Nous tenons à exprimer notre reconnaissance a tout le enseignants de la faculté de technologie plus particulièrement le département d'électronique qui nous ont attribués

Sans omettre bien sûr de remercie profondément tous ceux qui ont contribué de près ou de loin à la réalisation du présent travail.

Dédicace

En premier lieu, je remercie Allah de m'avoir permis d'acquérir

Une infime partie de sa science sans limites

Je dédie ce modeste travail à :

Ma chère mère pour son sacrifice et ses prières ;

Mon père pour son soutient absolu ;

A mon fiancé pour son aide ;

A mes frères ;

A leurs conjoints ;

Ma grande famille ;

A ma promo de TISE;

A tous mes amies sans exceptions ;

Que Allah nous aide tous à faire que le bien

BENNAI KHEIRA

Je dédie ce modeste travail :

A mes très chères mère et grand-mère qui n'ont cessé de me donner courage, volonté et de me pousser à toujours faire au mieux

A Mon père

A mes sœurs et mes frères

A mes oncles (surtout Nasser) et mes tantes (Hassina)

A tous mes ami(e)s et collègues

A ma cousine Abla

DJAZIA

ملخص:

تبيان وظيفة البروتين هو مشكل كبير في البيولوجيا ، لهذا هو مهم إعطاء طرق صحيحة لتبيانها و التي لديها خصائص موحدة .

في إطار موضوع مذكرتنا ، سندرس طريقة التحليل التي تعتمد على استعمال النماذج ماركوف المخفية انماط لأنّ هذا التحليل المستعمل على وحدات من البروتين من نفس العائلة (التي تتقاسم جدّ موحد) لبحث العائلة البروتينية. لهذا سنستعمل أنواع النماذج ماركوف المخفية.

كلمات المفاتيح: **المجين، السلاسل البيولوجية، نمط البروتين، سلاسل ماركوف، نماذج ماركوف المخفية.**

Résumé: Déterminer la fonction d'une protéine est un problème majeur en biologie, c'est pourquoi il est important de proposer des méthodes efficaces pour identifier les protéines partageant des caractéristiques communes. Dans le cadre de notre thèse, nous allons étudier une méthode d'analyse qui se base sur l'utilisation d'un modèle de Markov caché (HMM).

En effet, cette analyse est faite sur les séquences protéiques de la même famille (qui partagent un ancêtre commun) pour la recherche des motifs des familles de protéines. Pour cela, nous allons utiliser les profils des Modèles de Markov cachés.

Mots clés: Génome ; séquences biologiques ; motif protéique ; chaines de Markov; modèles de Markov cachés.

Abstract: Determine the function of a protein is a major problem in biology is why it is important to offer effective methods to identify proteins that share common characteristics. As part of our thesis we study a method of analysis that is based on the use of a hidden Markov model (HMM).

Indeed this analysis is made on the protein sequences of the same family (who share a common ancestor) For search of patterns of protein families, we will use for this the profiles Hidden Markov Models.

Keywords: Genome; biological sequences; protein motif; Markov chains; Hidden Markov Models.

Listes des acronymes et abréviations

A: Adénine.

C: Cytosine.

T: Thymine.

G: Guanine.

ADN: l'acide désoxyribonucléique.

ARN: l'acide ribonucléique.

EMBO: European Moleculary Biology Organisation.

EMBL: European Moleculary Biology Library.

EBI: European Bioinformatics Institute.

NIH: National Institute of Health

NCBI: National Center of Biotechnologie Information

HSSP: Homology-derived Secondary Structure of Proteins

STRIDE : Secondary STRucture IDEntification

PFAM : Proteine FAMilies.

Fasta : Foundation for Analytical Science & Technology in Africa

Blast : Basic Local Alignment Search Tool.

MMC: modèle de Markov caché.

HMM: Hidden Markov Model.

MIPS: Martinsried Institute for Proteine Séquences.

JIPID : Japon International Protéine Information Data base.

NBRF : National Biomedical Research Foundation.

MSA : Multiple Sequence Alignment.

PAM : Percent Accepted Mutation.

BLOSUM : Blocks Substitutions Matrices.

PSSMs: Position Specific Scoring Matrices.

$\mathbf{d_{i,j}}$: Indique la distance entre la séquence i et la séquence j.

Table des matières :

Chapitre II: Analyse des séquences protéiques

Chapitre III : Modélisation des familles de protéines par les MMC

Chapitre IV : Réalisation et interprétation des résultats

Liste des Figures

- **Chapitre I :** Définitions et généralités sur les protéines

- **Chapitre II : Analyse des séquences protéiques**

- **Chapitre III :** Deux structures différentes de HMM

- **Chapitre IV: Réalisation et interprétation des résultats**

Liste des tableaux :

- **Chapitre I:** Définitions et généralités sur les protéines

- **Chapitre II :**

- **Chapitre IV :**

Introduction générale

Il y a plus d'un siècle, Mendel comprit que le gène était une entité distincte. Ceci mit en évidence un fait maintenant bien accepté : l'information nécessaire à la construction d'un nouvel organisme se transmet d'une génération à l'autre [1].

Trois grandes classes de molécules sont impliquées dans ce processus d'entreposage, de conversion et de transmission d'information, l'acide désoxyribonucléique (ADN), l'acide ribonucléique (ARN) et les protéines. Les protéines sont les molécules les plus diversifiées des trois.

L'importance des gènes n'est plus un secret pour personne. De nos jours, des termes tels que l'ADN (acide désoxyribonucléique) et protéine ne sont plus des termes techniques connus seulement des 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, l'insuline bovine (Sanger, Thompson et Kitai, 1955), le nombre de séquences protéiques et nucléiques rendues publiques ne fait qu'augmenter.

La première base de données de séquences biologiques (Dayhoff et al., 1965) et les premiers algorithmes d'analyse de ces données ont donc vu le jour quelques années plus tard, donnant, par la même occasion, naissance à un nouveau domaine de recherche : la bioinformatique.

La bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. Plusieurs champs d'application ou sous-disciplines de la bioinformatique se sont constitués :

- La bioinformatique des séquences, qui traite de l'analyse de données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines.
- les modèles de Markov cachés (plus souvent désigné par l'acronyme HMM pour Hidden Markov Models), constituant une classe de modèles à l'origine de quelques-unes de plus spectaculaires avancées de ces dernières années.

Le principe central de cette famille de modèles est de supposer l'existence d'un état caché (latent) évoluant suivant une dynamique markovienne, les observations étant des fonctions (déterministe ou aléatoire) de cet état caché. Les modèles HMM sont très flexibles du fait de l'introduction de variables latentes (non observées), qui permettent de modéliser des structures de dépendance temporelles complexes. La structure markovienne des variables latentes permes l'utilisation de procédures numérique de traitement dont la complexité de calcul reste compatible avec les moyens informatiques actuels. Au cours de ces dernières années, de nombreuses variantes de ce paradigme de base ont été étudiées dans des applications aussi variées que la reconnaissance de paroles, la poursuite en environnement complexes, la détection et le diagnostic de pannes, la bioinformatique, la modélisation et le contrôle de réseaux, les communications numériques, etc.

Notre travail s'inscrit dans le cadre de la modélisation de protéines à l'aide de modèles de motifs protéiques par les profils des modèles de Markov cachés qui offrent une description générale du comportement des séquences. Ainsi, à l'aide des propriétés statistique de ces modèles, différentes propriétés biologiques des séquences protéiques ou d'ADN peuvent être soulignées.

Ce projet présente le travail réalisé. Il contient trois chapitres, organisés comme suit :

- Le premier chapitre contient des définitions et des généralités sur les protéines.
- Le deuxième chapitre portera sur l'Analyse des séquences protéiques.
- Le troisième chapitre présentera la modélisation des familles de protéines par modèles de Markov caché et leurs applications au domaine biologique.
- Le quatrième chapitre présentera notre modèle HMM ainsi que les résultats obtenus
    - Enfin nous terminerons par une conclusion.

Chapitre I

I.1 Introduction

Toutes les cellules ont besoin de protéine pour faciliter les réactions chimiques. Ces larges molécules déterminent les propriétés physiques et chimiques de la cellule. Une protéine est construite par une collection d'un grand nombre de molécules plus petites appelées acides aminés. Les caractéristiques spatiales des protéines sont la clé de leurs fonctions. La structure des protéines est définie à plusieurs niveaux.

Ce chapitre portera sur les prés requis biologiques nécessaires pour appréhender les données manipulées ainsi que quelques notions de biologie moléculaire. Ceci permettra de mieux comprendre en quoi consistent les séquences biologiques, en particulier les protéines, support et sujet de notre travail.

I.2 Le gène

Le gène est l'unité d'hérédité contrôlant un caractère particulier. Cet élément génétique correspondant à un segment d'ADN ou d'ARN (virus), situé à un endroit bien précis (locus) sur un chromosome. Chaque région de l'ADN qui produit une molécule d'ARN fonctionnelle est un gène.

Un gène est destiné à être transcrit en acide ribonucléique (ARN), si c'est le cas, la séquence est dite « codante ». La plupart du temps, un gène commence par une séquence de nucléotides appelée *promoteur*, dont le rôle est de permettre l'initiation mais surtout la régulation (tous les gènes ne sont pas exprimés dans toutes les cellules) de la transcription de l'ADN en ARN, et se termine par une séquence terminatrice appelée terminateur, qui marque la fin de la transcription. La molécule d'ARN ainsi produite peut soit être traduite en protéine (elle est dans ce cas appelée *ARN messager*), soit être directement fonctionnelle (c'est le cas pour les ARN ribosomaux ou les ARN de transfert). Il y a environ 13 000 gènes dans l'ADN des cellules d'une drosophile et 200 gènes chez l'Homme [2].

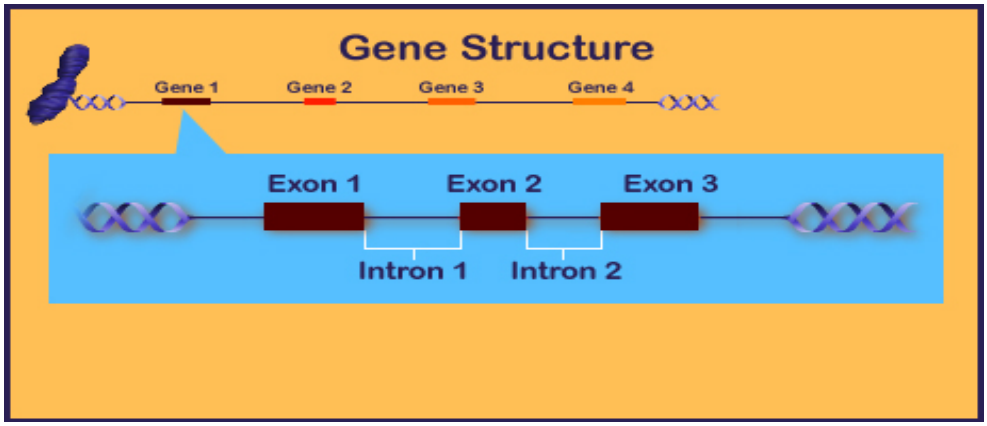


Figure I.1: Structure d'un gène [3].

I.3 L'ADN

Un acide nucléique est composé de petites molécules appelées *nucléotides*. Deux types d'acides nucléiques existent : l'ADN, qui entrepose l'information génétique d'un individu et l'ARN qui sert le plus souvent de vecteur de l'information. Un nucléotide est constitué d'une base azotée de type purine (adénine **A** ou guanine **G**) ou de type pyrimidine (cytosine **C** ou thymine/uracile **T/U),** d'un sucre désoxyribose (ADN) ou ribose (ARN) et d'un groupe phosphate.

Lorsque qu'une base azotée est attachée à un sucre, on la nomme un *nucléoside*. L'adénine devient une adénosine, la guanine devient une guanosine, la cytosine devient une cytidine et la thymine devient une thymidine. Une fois ces nucléosides attachés à un groupe de phosphate, on les nomme des nucléotides. Ces nucléotides forment des acides nucléiques de type ADN. Les bases azotées formant une chaîne de nucléotides ont des affinités avec, ou une attirance vers, un autre membre du groupe. Cette affinité les incitent à former des liens hydrogènes entre elles, lorsque mises en contact. **L'adénine A** est ainsi dite complémentaire à la **thymine T** et la **cytosine C** est complémentaire à la **guanine.**

L'ADN typique d'un organisme consiste en une molécule formée de deux chaînes de nucléotides entrelacées entre elles (fig. I.2). Des liens hydrogènes les unissent, des liens forts (trois ponts hydrogènes) entre les nucléotides **C** et **G** et des liens plus faibles (deux ponts hydrogènes) entre **A** et **T** [1].

Pour un brin d'ADN possédant vingt nucléotides comme dans l'exemple suivant, on peut retrouver la séquence du brin complémentaire et reconstituer la double séquence de la double hélice.

Exemple d'une séquence de nucléotides :

5'-ATTGCCGTATGTATTGCGCT-3'

3'-TAACGGCATACATAACGCGA-5'

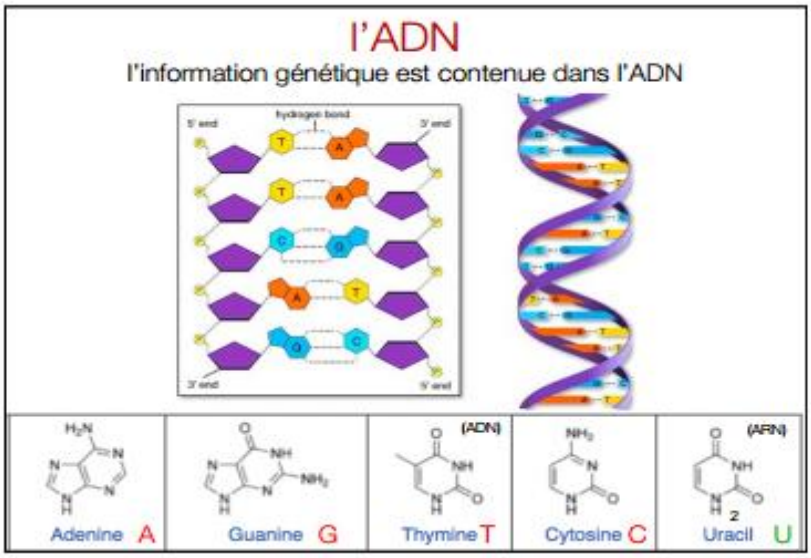


Figure I.2: Structure de la molécule d'ADN **[3].**

I.4 Les protéines

Les protéines sont des macromolécules essentielles pour la structuration et le fonctionnement des cellules vivantes.

En fait, l'immense majorité des fonctions cellulaires est assurée par des protéines. De nos jours, la caractérisation de la fonction des protéines est donc une des tâches essentielles en bioinformatique dans une visée fondamentale (avancée des connaissances) ainsi que dans une visée appliquée (identifier par exemple les protéines clefs de certaines pathologies humaines) [4].

I.5 Différents niveaux de structures :

On distingue communément plusieurs niveaux de structure pour décrire une protéine :

a) Structure primaire : L'ordre dans lequel les acides aminés s'enchaînent constitue la structure primaire de la protéine. On parle alors de séquence d'acides aminés ou séquence protéique.Les acides aminés issus du code génétique sont au nombre de 22 (cf. Tableau I.1).

Exemple : Un exemple en protéique du gène Antennipedia de Drosophile :

1 mtmstnncesmtsyftnsymgadmhhghypgngvtdldaqqmhhysqnanhqgnmpypr
61 ppydrmpyyngqgmdqqqqhqvysrpdspssqvggvmpqaqtngqlgvpqqqqqqqqqps
121 qnqqqqqaqqapqqlqqqlpqvtqqvthpqqqqqqpvvyascklqaavgglgmvpeggsp
181 plvdqmsghhmnaqmtlphhmghpqaqlgytdvgvpdvtevhqnhhnmgmyqqqsgvppv
241 gappqgmmhqgqgppqmhqghpgqhtppsqnpnsqssgmpsplypwmrsqfgkcqgk

b) Structure secondaire : Le concept de structure secondaire a été introduit par [4] Linderstrøm-Lang .1952]. La structure secondaire décrit le repliement de segment court de la structure primaire. Ce sont des structures locales, stabilisées par des liaisons hydrogènes entre les groupements amide (-NH) et carbonyle (-CO) du squelette peptidique [4] [Pauling et al., 1951)]. L'existence de structures secondaires vient du fait que les repliements énergétiquement favorables de la séquence protéique sont limités et que seules certaines conformations sont possibles. Les principaux types de structures secondaires régulières sont les hélices α et les feuillets β. Ils se trouvent fréquemment dans les protéines mais en proportions et combinaisons variables. On trouve aussi des coudes (également nommés retour en arrière ou tournants — traductions de l'anglais turns), ainsi que des boucles (coil). Ces structures ne sont pas moins ordonnées que les hélices ou les feuillets ; elles sont plus irrégulières et donc plus difficile à décrire.

Les hélices, feuillets, coudes et boucles comptent pour environ 90% en moyenne dans les protéines standards (dont 31% d'hélices et 28% de feuillets) (Voet et Voet, 2004). Il est connu que certains acides aminés favorisent la formation d'une structure secondaire plutôt qu'une autre (Blout et al., 1960).

De nombreuses approches ont été décrites pour prédire la structure secondaire d'une protéine à partir de sa séquence primaire, basées sur les propriétés physico-chimiques, de simples statistiques linéaires, ainsi que de nombreuses méthodes d'apprentissage (réseaux neuronaux, k-plus proches voisins, arbres phylogénétiques, etc.) (Cuff et Barton, 2000) [4].

Il est également possible de calculer la structure secondaire d'une protéine à partir de sa structure tertiaire (voir ci-dessous), par exemple grâce au dictionnaire de structure secondaire de protéines (DSSP) (Kabsch et Sander, 1983), ou grâce à des algorithmes tels que HSSP (Homology-derived Secondary Structure of Proteins) (Sander et Schneider, 1991) ou STRIDE (secondary STRucture IDEntification) (Frishman et Argos, 1995) [4].

Exemple :

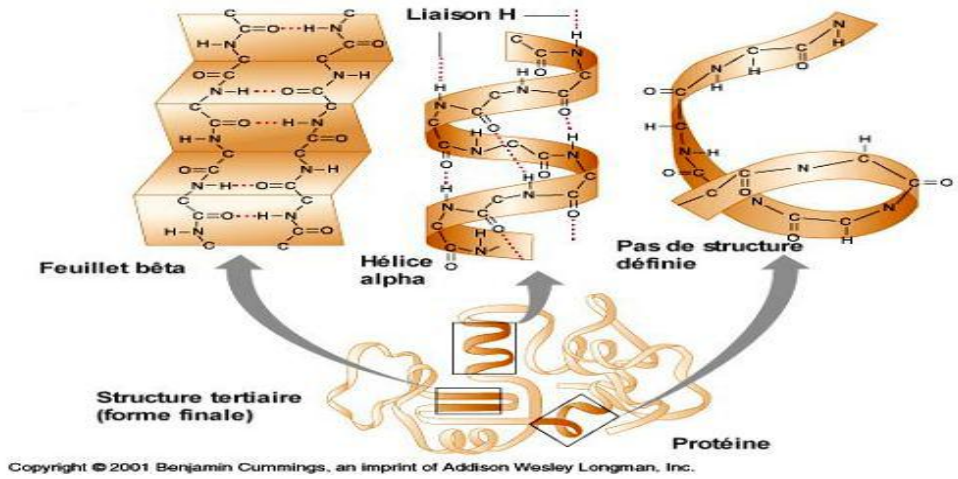


Figure I.3 : structure tridimensionnelle (3D) finale qu'adopte la chaîne d'acides aminés[3].

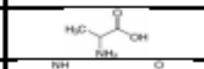
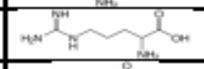
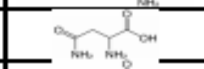
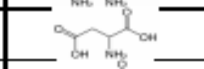
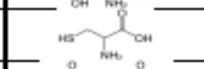
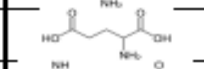
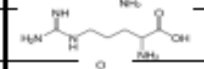
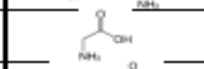
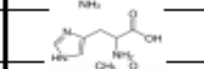
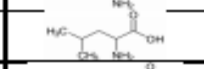
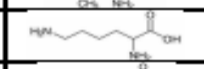
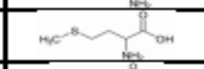
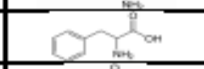
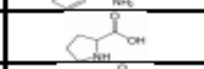
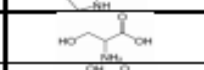
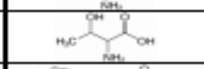
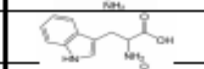
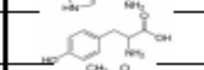
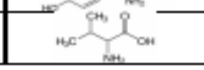
Acide aminé	Code à 3 lettres	Code à 1 lettre	Formule	Traduit par les codons
Alanine	Ala	A		GCU, GCC, GCA, GCG
Arginine	Arg	R		CGU, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn	N		AAU, AAC
Acide aspartique	Asp	D		GAU, GAC
Cystéine	Cys	C		UGU, UGC
Acide Glutamique	Glu	E		GAA, GAG
Glutamine	Gln	Q		CAA, CAG
Glycine	Gly	G		GGU, GGC, GGA, GGG
Histidine	His	H		CAU, CAC
Isoleucine	Ile	I		AUU, AUC, AUA
Leucine	Leu	L		UUA, UUG, CUU, CUC, CUA, CUG
Lysine	Lys	K		AAA, AAG
Méthionine	Met	M		AUG
Phénylalanine	Phe	F		UUU, UUC
Proline	Pro	P		CCU, CCC, CCA, CCG
Sérine	Ser	S		UCU, UCC, UCA, UCG, AGU, AGC
Thréonine	Thr	T		ACU, ACC, ACA, ACG
Tryptophane	Trp	W		UGG
Tyrosine	Tyr	Y		UAU, UAC
Valine	Val	V		GUU, GUC, GUA, GUG

Tableau I.1 : Table des Structure des 20 acides aminés [4]

c) Structure tertiaire : L'arrangement spatial des structures secondaires locales aboutit à une forme globale spécifique de la protéine maintenue par des interactions qui peuvent être de nature différente. La structure tertiaire d'une protéine est le paramètre fondamental dont dépend l'expression de ses fonctions biologiques qu'elles soient structurales ou dynamiques (protéines des membranes ou du cytosquelette ou etc., récepteurs, enzymes, etc.). On les subdivise en groupes par rapport à la dominance des motifs de structure secondaire :

- protéines à domaine α ;
- protéines à domaine β ;
- protéines α/β : elles comprennent des domaines α et des domaines β (les plus nombreuses).

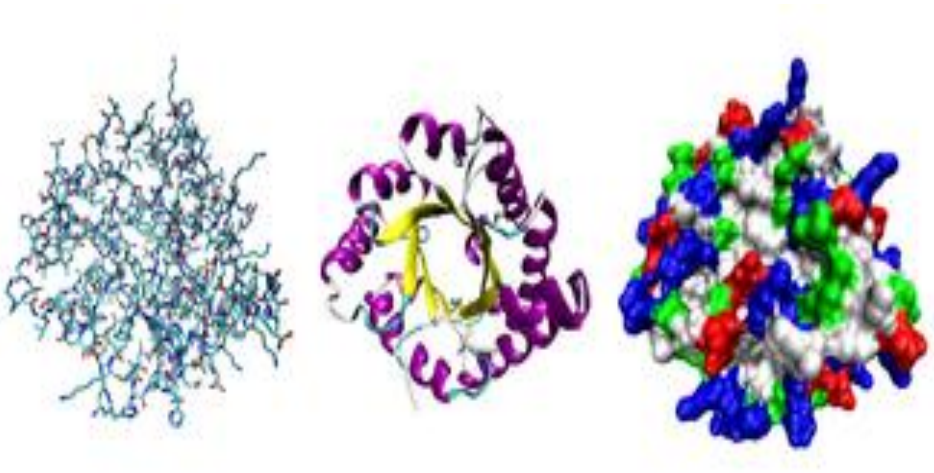


Figure I.4 : Trois représentations possibles de la structure tertiaire de la protéine triose phosphate isomerase [4].

d) Structure quaternaire : C'est l'organisation des protomères qui définit la structure quaternaire d'une protéine formée de multimètres. De manière très succincte, indiquons que :

- l'association ou la dissociation des protomères permettent à ces protéines d'avoir des fonctions de signalisation ou de contrôle (l'enzyme protéine-kinase dépendante de l'AMP cyclique)
- l'interaction entre les protomères (sous-unités) est un moyen très précis de régulation : phénomène coopératif de fixation de ligands, enzymes allostériques, récepteurs membranaires
- les isoformes des protéines : elles ont les mêmes fonctions, toutefois leur structure est dépendante du tissu soit au niveau du nombre de protomères dans l'association soit une différence au niveau du protomère.

I.6 Domaine et motif protéique

On distingue également au sein des protéines des motifs structuraux ou motifs protéiques. Un motif est une séquence courte associée à des interactions bien précises : site actif ou d'ancrage par exemple (Doolittle)[4].

La différence entre domaine et motif est assez mince, et porte souvent à confusion. Elle tient principalement au fait qu'un domaine contenant plusieurs sites d'ancrage est composé de plusieurs motifs. De plus, les motifs n'ont pas forcément de repliement propre. Ils contiennent des résidus essentiels à la fonction et à l'interaction, éventuellement entrecoupés de résidus non-essentiels : il s'agit d'un pattern. Cependant, participant à un même site, les acides aminés essentiels peuvent être assez proches dans la structure 3D.

Les motifs ont été définis comme des groupes d'acides aminés extrêmement bien conservés entre des séquences globalement différentes. Lesk [4] décrit l'émergence des motifs comme issue des contraintes imposées par l'évolution à certaines portions des séquences pour la conservation de la fonction des sites d'ancrage. Le premier exemple fut le motif appelé "doigt de zinc" (car il fixe l'ion Zn2+) impliqué dans des interactions spécifiques avec l'ADN.

I.7 les bases de données biologiques

Au début des années 80 que les premières banques de séquence sont apparues sous l'initiative de quelques équipes comme celle du professeur Grantham à Lyon. Il existe un grand nombre de bases de donnes d'intérêt biologique [5].

- En Europe : financée par l'EMBO (European Moleculary Biology Organisation), une équipe c'est constituées pour développer une banque de séquence nucléique **EMBL** (European Moleculary Biology Library) **banque européenne financée par l'EMBO .Elle est aujourd'hui diffusé par l'EBI (**European Bioinformatiques Institute, Cambridge).

- En Amérique: soutenue par le NIH (National Institute of Health) une banque nucléique nommée **GenBank** a été créée**.** Cette base de données est diffusée maintenant par le **NCBI** (National Center for Biotechnologie Information).

Parallèlement, pour les protéines, deux banques sont créées :

- La première sous l'influence du **NBRF** (National Biomedical Research Foundation) à Washington, produit maintenant une association de donnée issue du **MIPS** (Martinsried Institute for Protéine Séquences) ; de la base Japonaise **JIPID** (Japon International Protéine Information Data base) et des données propres de la **NBRF**. Elle se nomme la **PIR-NBRF** (Protéine Identification Ressource, 1986).
- La deuxième, **SWISSPROT** a été constitué à l'université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la **PIR-NBRF** ainsi que des séquences codantes traduites de **l'EMBL**.

- **PRODOM, PFAM**: banques de domaines associes à des familles de protéines.

  PFAM : Pfam est une base de données conservées, formée des familles de domaines de protéines, largement utilisée par les biologistes pour annoter et classer les protéines. La base de données comprend deux classes d'entrées:

  a) des familles Pfam-A, qui se composent d'un alignement de séquences, un modèle de Markov caché, alignements complets, annotation associée, références bibliographiques, et des liens de base de données;

  b) les familles Pfam-B, qui sont générés automatiquement par des alignements de grappes de séquence, provenant de la base de données du domaine de décomposition algorithme automatique (ADDA), sans annotation, qui complètent les familles Pfam-A. Les utilisateurs sont en mesure de rechercher leurs séquences avec les bibliothèques des deux ensembles d'entrées afin de mieux comprendre fonctionnement de leurs protéines d'intérêt[6].

On sait que certains segments d'ADN ou de protéines sont déterminants dans l'analyse des séquences car ils correspondants à des sites précis d'activité biologique comme par exemple les éléments de régulation des gènes ou signature peptidiques. C'est pourquoi des bases spécialisées se sont naturellement constituées autour de ces séquences L'utilisation de ces base de motifs est devenu un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéine inconnus ou savoir a quelle famille appartient une séquence non encore caractérisée. Il existe principalement deux types de bases de motifs qui permettent de recenser des signatures protéiques liées des activités biologiques :

- la base de motifs protéiques **PROSITE** :
  La base PROSITE peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.
- **la base de motifs protéiques BLOCK**

La base Block est également basée sur un système qui détecte et assemble les régions conservées de protéine apparentées.

Autres exemple des banques :

• PDB: banque de structures cristallographiques/RMN.

• ENZYME: banque des EC numbers, reactions enzymatiques associées, pointeurs sur SwissProt

• Metacyc: encyclopedie de chemins métaboliques (réaction enzymatiques) .

• Medline: catalogue d'articles scientifiques (en biologie).

La différence entre base de données et une banque de données n'est pas liée à la structure de la base mais la nature du contenu même [6].

Une base de données biologique concerne des données spécifiques à une discipline, une maladie, une espèce vivante, une molécule,…..elle est donc spécifique. On peut recenser plus de 250 BD spécialisées.

Une banque de donnée biologique, est une base de données généralisée on y trouve des informations sur nombreuses espèces nombreuses molécules, en même temps son utilisation concerne plusieurs ombreuses domaines à la fois.

I.8 Les formats des séquences biologiques :

Le format Fasta : est sans doute le plus répondu et l'un des plus pratique. La séquence est décrite sous forme de lignes de 80 caractères maximum est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">". Cela permet de mettre plusieurs séquences dans un même fichier [7].

```
>gi|22777494|dbj|BAC13766.1| glutamate dehydrogenase [Oceanobacillus iheyensis]
MVADKAADSSNVNQENMDVLNTTQTIIKSALDKLGYPEEVFELLKEPMRILTVRIPVRMDDGNV
LGGSHGRESATAKGVTIVLNEAAKKKGIDIKGARVVIQGFGNAGSFLAKFLHDAGAKVVAISDA
YGALYDPEGLDIDYLLDRRDSFGTVTKLFNNTISNDALFELDCDII
>EM|U03177|FL03177 FELINE LEUKEMIA VIRUS CLONE FELV-69TTU3-16.
AGATACAAGGAAGTTAGAGGCTAAAACAGGATATCTGTGGTTAAGCACCTG
GCCAGCAGTCTCCAGGCTCCCCA
```

Figure I.5: exemple du format FASTA d'une séquence protéique [7]

CODE	SIGNIFICATION
">"	Début de séquence.
gi|22777494	GenInfo Identifier
dbj|BAC13766.1|	Un enregistrement de séquence peut être enregistré dans plusieurs banques de données donc il y aura un identifiant dans la banque de données dans cet exemple c'est DNA Database of Japan sous le n° dbj|BAC13766.1
BAC13766.1|	". 1" la séquence a été révisée une fois
"glutamate dehydrogenase"	nom de la protéine
[Oceanobacillus iheyensis]	nom de l'organisme à partir duquel elle a été déterminée.

Tableau I.2 : description du fichier FASTA de l'exemple de figure I.3 [7].

Autre format les plus utilisé :

- STADEEN
- PIR(spécifique à la Bdd PIR)

1.9 La Bioinformatique

La bioinformatique c'est une science qui conceptualise la biologie en termes de molécules (dans le sens de la chimie-physique) et applique des " techniques d'informatiques " pour comprendre e et organiser L'information liée à ces molécules, sur une grande échelle. En bref, la bioinformatique est un système intégré de gestion pour la biologie moléculaire e et a beaucoup d'applications pratiques.

Le mot « bioinformatique » découle donc de l'analyse par ordinateur des données biologiques. Ces données représentent l'information stockée dans le code génétique, mais également des résultats expérimentaux de diverses sources et des statistiques, … etc.

La bioinformatique est une science récente qui évolue rapidement et qui est fortement interdisciplinaire, elle conjugue plusieurs sciences telles que la biologie moléculaire, l'informatique, et les mathématiques (statistiques)… etc. Le but de la recherche dans la bioinformatique est l'organisation et l'extraction des données, la mise en application des algorithmes complexes et le développement des outils de visualisation afin d'atteindre une compréhension exhaustive et une exploitation des informations contenues dans les séquences d'un génome.

L'histoire du calcul dans la biologie moléculaire n'est pas récente mais date des années 20 où les scientifiques pensaient déjà à établir des lois biologiques par induction. Cependant, seulement le développement des ordinateurs puissants, et la disponibilité des données expérimentales qui peuvent être aisément traitées par calcul (par exemple, les séquences d'ADN ou d'acide aminé et des structures tridimensionnelles des protéines) ont lancé la bioinformatique comme un domaine indépendant. Aujourd'hui, les applications pratiques de la bioinformatique sont aisément disponibles sur le Web, et sont largement répondues dans la recherche biologique et médicale.

Le rapport entre l'informatique et la biologie moléculaire est normal pour plusieurs raisons. D'abord, le taux phénoménal de données biologiques produites

fournit des défis: des quantités massives de données doivent être stockées, analysées, et doivent être rendues accessibles[8].

L'analyse de séquences

Les outils d'*analyse de séquences* sont développé à fin de pouvoir déterminer leurs propriétés :

- Recherche de protéines à partir de la traduction de séquences nucléiques connues. Celle-ci passe par la détermination des phases ouvertes de lecture d'une séquence nucléique et de sa ou ses traduction(s) probables.
- Recherche de séquences dans une banque de données à partir d'une autre séquence ou d'un fragment de séquence. Les logiciels les plus fréquemment utilisés sont de la famille BLAST (blastn, blastp, blastx, tblastx et leurs dérivés).
- Alignement de séquences : pour trouver les ressemblances entre deux séquences et déterminer leurs éventuelles homologies. Les alignements sont à la base de la construction de parentés suivant des critères moléculaires, ou encore de la reconnaissance de motifs particuliers dans une protéine à partir de la séquence de celle-ci.
- Recherche de motifs ou structures consensus pour caractériser les séquences.

1.10 Conclusion

Dans ce chapitre nous avons exposé les définitions et les notions concernant notre domaine de travail. En particulier, nous avons défini ce qui est une protéine et ce qui permet de la localiser dans une séquence biologique.

Chapitre II

II.1 Introduction

La recherche de similitudes entre séquences est un élément fondamental qui constitue souvent la première étape des analyses de séquence. Elle permet de révéler des régions proches dans leur séquence primaire. On peut alors dégager :

- Des informations importantes sur la structure ;
- La fonction ou l'évolution des biomolécules ;
- Une recherche d'homologie ; [9].

Plusieurs méthodes basées sur l'alignement multiples des séquences et le calcul de score de similitude ont été développées.

Dans ce chapitre, nous présenterons les principes de base liés à ces méthodes

II.2 Les principes de bases pour identifier les ressemblances entre deux séquences

L'identité, la similitude et l'alignement

Les programmes de comparaison de séquences ont pour but de repérer les endroits où se trouvent des régions identiques ou très proches entre deux séquences et d'en déduire celles qui sont significatives et qui correspondent à un sens biologique de celles qui sont observées par hasard. En générale les algorithmes fonctionnent sur des segments de séquences (on parle de fenêtres, de motifs ou de mots) sur lesquels on regarde s'il existe ou pas une similitude significative. Si on ne prend en compte que des analogies entre sous-séquences sans traiter la possibilité d'insertion ou de délétion, on parlera alors de segments similaires. On distingue pour cette catégorie deux classes précises de similitude : la ressemblance parfaite ou identité et la ressemblance non parfaite que l'on qualifie de similitude (voir figure II.1) [10].

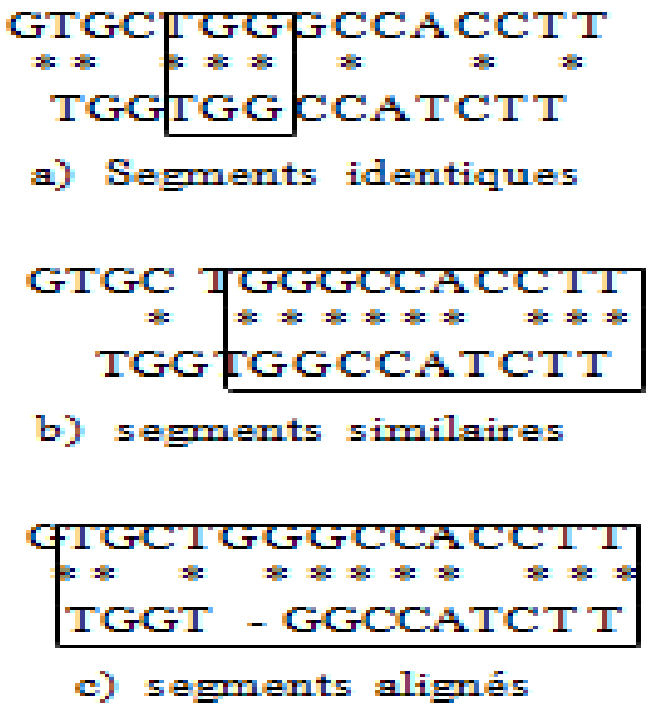


Figure II.1 : Les trois grandes classes de ressemblance issues de la comparaison de séquences [10].

II.3 Motifs dans les séquences

II.3.1 Pourquoi la recherche de motifs ?

Les motifs sont souvent recherchés dans des séquences car ils sont généralement impliqués dans des systèmes de régulation ou ils définissent des fonctions biologiques comme la détermination de la fonction d'une nouvelle séquence (par exemple en localisant un ou plusieurs motifs répertoriés dans des bases de motifs), l'identification dans une séquence nucléique de régions codantes, ou bien l'extraction à partir des banques de données des séquences possédant le même signal de régulation ou la même signature protéique pour effectuer des études comparatives ultérieures [11].

II.3.2 Qu'est-ce qu'un motif ?

Un "motif" (ou «pattern» en Anglais) est un segment court dans une séquence, il est continu et non ambigu. Il peut représenter une structure plus complexe lorsque lui-même est composé de différents "motifs" qui peuvent être plus ou moins éloignés les uns des autres et sa définition peut comporter des exclusions ou des associations de "motifs" [12].

Plusieurs méthodes ont été imaginées pour identifier des éléments fonctionnels en utilisant leur conservation en séquence. La recherche de motifs conservés peut se faire à partir d'alignements multiples par recherche de blocs conservés dans l'alignement ou directement à partir de la séquence par des méthodes qui à la fois recherchent et déterminent des consensus. Ces dernières méthodes sont à la base des techniques d'alignement multiple "par blocs" (Sagot 1997) [11].

Dans notre travail, nous nous intéressons à la recherche de motif par la technique des MMC (HMM).

Exemple de recherche de motif:

Séquence : C T G T G T G T A C A T G T G de longueur 15

Motif : T G T G de longueur 4

Position : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Séquence : C T G T G T G T A C A T G T G

Solution**:** Ensemble de positions: *{2, 4, 12}.*

II.3.3 Qu'est- ce qu'un consensus ?

Chaîne de caractère indiquant les résidus conservés à chaque colonne d'un alignement multiple.

Le consensus est obtenu en retenant, pour chaque colonne d'un alignement multiple, soit un seul résidu (on parle alors de consensus strict, soit une combinaison de résidus représentatifs (consensus dégénéré). Les consensus dégénérés peuvent être représentés par des expressions régulières, combinées avec les spécifications IUPAC pour les résidus ambigüs.

Un consensus fournit une représentation compacte d'un motif séquentiel. Les consensus sont par exemple utilisés :

- pour les séquences nucléiques, afin de représenter les motifs de liaison de facteurs transcriptionnels sur les séquences d'ADN,
- pour des séquences peptidiques, afin de représenter les caractéristiques de domaines conservés au sein de familles de protéines homologues.
- Le consensus fournit une représentation compacte et intuitive d'un motif, mais souffre de quelques limitations.
- Les règles appliquées pour décider si l'on représentera une colonne de l'alignement par un résidu unique ou une combinaison de résidus sont floues, et varient selon les auteurs.
- En cas de positions ambigües, le consensus indique quels résidus sont acceptés, mais n'indique pas la fréquence relative de ces résidus dans les alignements initiaux (contrairement aux profils, ou matrices position-poids) [9].

II.3.4 Expressions régulières

Une *expression régulière* est une chaîne de caractères qui décrit un *motif* (ou *pattern*) composé de différents types d'éléments :

- Caractères parfaitement déterminés (composés dans l'alphabet des acides aminés ou nucléotides selon le cas).
- Une série de lettres entre crochets [] indique une liste de résidus alternatifs acceptés à cette position (résidus partiellement déterminés).
- Un nombre entre accolades représente la répétition d'un caractère. Le nombre de répétitions peut être fixe (ex: A {4}=AAAA) ou variable (N {3,20} signifie "un nombre de nucléotides variant entre 3 et 20").
- Pour les motifs peptidiques, une série de résidus entre accolades signifie "tout sauf ces résidus". Par exemple {A, L} signifie "n'importe quel acide aminé sauf A et L" [9].

Exemples:

- L'expression régulière: GAT [AT] AG signifie "la séquence GAT suivie soit d'un A soit d'un T, lui-même suivi de AG".
- Pour des séquences nucléiques, l'expression régulière CCGn {11} CCG signifie "la séquence CCG suivie de 11 résidus indifférents (A, C, G ou T), suivis d'une séquence CGG "
- Pour des séquences peptidiques, l'expression régulière [AV]-x-L {2} décrit un motif qui commence par soit Alanine soit Valine, ([AV]) suivi par n'importe quel acide aminé (x), suivi par 2 Leucines (L {2}).

II.4 L'alignement des séquences protéiques

II.4.1 Alignement de deux séquences

L'alignement de deux séquences permet d'identifier les mutations qui ont eu lieu lors de l'évolution. Ces évènements sont à l'origine de la divergence des séquences [11].

Un alignement de deux séquences (appelé souvent Alignement deux à deux) est une mise en correspondance entre les résidus avec une possible insertion des espaces (gaps) afin d'obtenir des séquences de longueur égales. Toutes les correspondances sont autorisées à condition que l'ordre des résidus soit respecté.

Trois situations sont possibles pour une position donnée de l'alignement:

- Les caractères sont les mêmes : identité
- Les caractères ne sont pas les mêmes : substitution
- L'une des positions est un gap (espace) : Insertion\Délétion [13].

Exemple d'alignement de deux séquences:

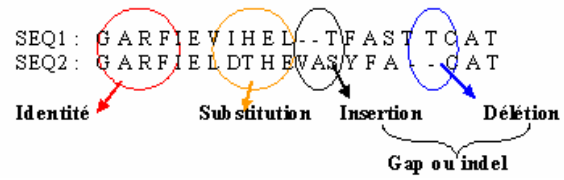


Figure II.2: Alignement de deux séquences

II.4.2 Les méthodes d'alignement de deux séquences

Il existe deux types d'alignements de séquences: global et local.

Le premier prend en considération l'ensemble des résidus de chacune des séquences. Si les longueurs des séquences sont différentes, alors la plus courtes va subir des insertions de gaps afin d'arriver à aligner les deux séquences d'une extrémité à l'autre. Cependant dans un alignement global, si uniquement des segments courts sont très similaires entre deux séquences, les autres parties des séquences risquent de diminuer le poids de ces régions. C'est pourquoi d'autres algorithmes d'alignements, dits locaux, basés sur la localisation des zones de similarité sont nés. Le but de ces alignements locaux est de trouver sans prédétermination de longueur les zones les plus similaires entre deux séquences. L'alignement local comporte donc une partie de chacune des séquences et non la totalité des séquences comme dans la plupart des alignements globaux [11].

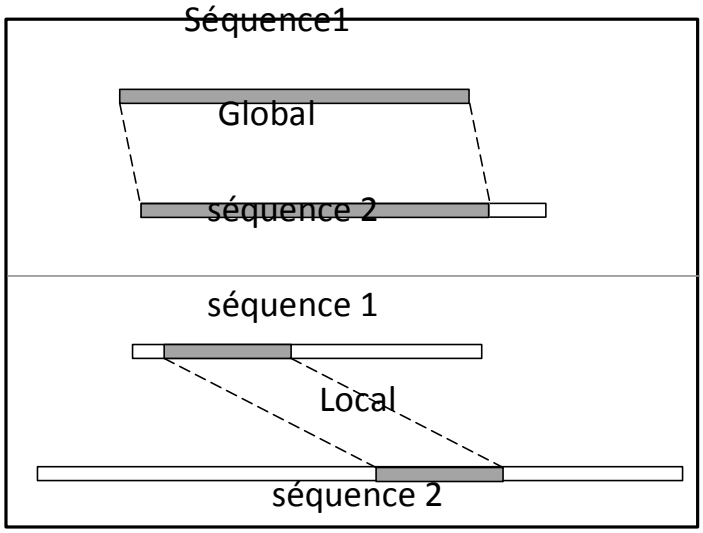


Figure II.3 : Alignement local et global [13].

Cependant, il est clair que pour deux séquences données quel conques il y a plusieurs alignements possibles. Il est devenu alors nécessaire de pouvoir déterminer quel est le meilleur alignement ou plutôt l'optimal si possible.

II.4.3 Les alignements multiples

L'alignement multiple de séquences (Multiple Sequence Alignment: MSA) est une tâche cruciale et très importante en biologie moléculaire. MSA offre aux biologistes un moyen pour analyser des séquences d'ADN ou de protéines et de déterminer par la suite leur degré d'homologie ou de divergence. MSA est utilisé dans la construction des arbres phylogénétiques et identifier les motifs dans des familles de protéines, ceci permet de prédire leur aspect structurel et fonctionnel [11].

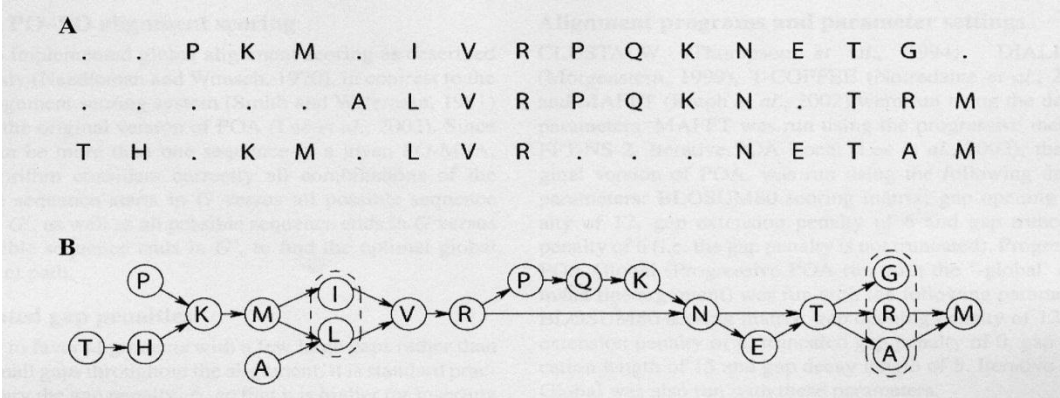


Figure II.4.: Représentation d'un alignement multiple sous forme linéaire (A) et sous forme de graphe (B). figure issue de [Lee etal.,2002] [14] .

II.4.4 Intérêt de l'alignement multiple de séquence :

L'alignement multiple de séquences est un outil fondamental pour de nombreuses analyses en biologie. Il permet de comparer un groupe de protéines ou de gènes apparentés, afin d'établir des relations évolutives. Si deux séquences ont une similarité significative, il est fait l'hypothèse qu'elles partagent un ancêtre commun, elles sont donc homologues .Si deux séquences ont des motifs communs, il est fait l'hypothèse qu'elles sont soumises à une pression de sélection qui empêchent les mutations de se fixer, probablement parce que le motif est important pour assurer une fonction.

L'alignement multiple est principalement utilisé pour :

- Trouver des caractéristiques communes à une famille de protéines soit des régions conservées (des motifs), soit des acides aminés strictement conservés permettant de relier une séquence à une structure et à une fonction ;
- Construire l'arbre phylogénétique des séquences homologues considérées;
- Déduire des contraintes de structures pour les ARN [1].

Pour caractériser les régions conservées dans les séquences, il est souvent plus efficace d'utiliser plusieurs séquences et d'effectuer un alignement multiple. Récemment sont apparues des méthodes basées sur des stratégies itératives de raffinement d'un alignement initial, en utilisant soit des alignements locaux par programmation dynamique (Morgenstern et al. 1996), soit des alignements globaux par utilisation de chaînes de Markov cachés (Morgenstern et al. 1996) ou des algorithmes génétiques (Notredame & Higgins 1996). Les algorithmes itératifs sont capables d'une plus grande précision, mais ils sont plus gourmands en temps de calcul. La nature heuristique de ces programmes recommande la prudence dans l'interprétation des résultats et de préférence leur validation par l'utilisation de plusieurs programmes.

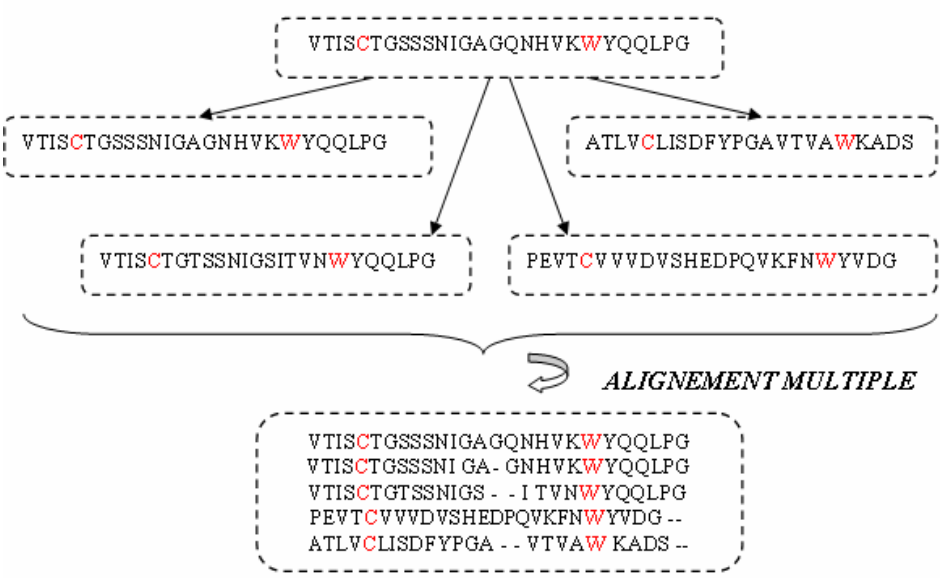


Figure II.5: L'Alignement multiple de séquences protéiques [13].

II.4.5 L'Alignement Progressif

La méthode la plus utilisée pour aligner plusieurs protéines est l'*alignement progressif*. Elle se décompose en plusieurs étapes:

1. Calcul d'une matrice de distance entre toutes les paires de séquences.
2. Construction d'un arbre-guide à partir de cette matrice de distances
3. Construction de l'alignement sur base de l'arbre-guide

1) Matrice de distances entre paires de séquences

La première étape d'un alignement progressif consiste à aligner chaque paire de séquences, et à calculer leur distance. On regroupe les résultats dans une matrice de distances, où :

- chaque ligne correspond à une séquence
- chaque colonne correspond à une séquence
- la valeur $d_{i,j}$ indique la distance entre la séquence i et la séquence j.

	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

Tableau II.1 : Distance entre les séquences à alignées

Les alignements par paires peuvent être effectués en utilisant la programmation dynamique (algorithme de Needleman-Wunsch) ou une heuristique plus rapide (fasta, blast).

2) Construction de l'arbre-guide

A partir de la matrice de distance, on peut construire un arbre-guide par la méthode du Neighbourjoining (*NJ*).

Le principe est d'établir en premier lieu un branchement qui relie les deux séquences les plus proches (celles qui ont la distance minimale dans la matrice de distances), puis les séquences un peu moins proches, et ainsi de suite jusqu'à avoir branché toutes les séquences.

Il s'agit uniquement d'un outil utilisé temporairement pour déterminer l'ordre d'incorporation des séquences dans l'alignement entre séquences multiples.

L'inférence phylogénétique nécessite des analyses plus poussées, qui ne pourront être effectuées qu'après avoir obtenu l'alignement multiple.

3) Construction de l'alignement progressif

Après avoir calculé la matrice de distance et construit l'arbre-guide, on construit l'alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.

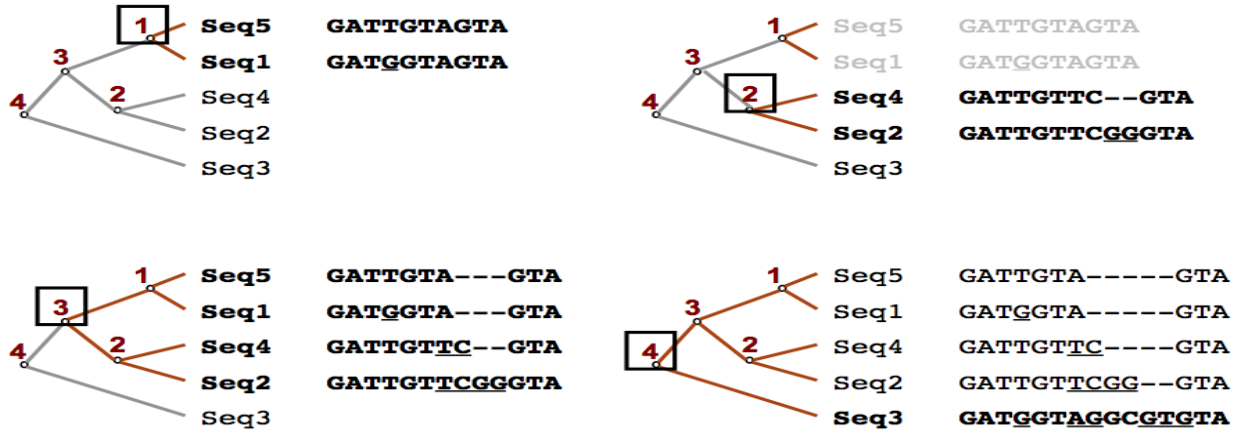


Figure II.6 : Construction d'alignement progressif

II.5 Score d'un alignement multiple

Le score d'un alignement multiple doit rendre compte de la qualité de l'alignement. Les algorithmes utilisés cherchent à maximiser ce score, qui est une indication de l'alignement optimal.

Quelle que soit la méthode d'alignement multiple, le problème de la méthode de calcul du score se pose. La plus utilisée est le score somme des paires (SP)" sum of pairs ": somme sur chaque colonne de tous les scores entre acides aminés pris deux à deux

(selon une matrice de substitution).En faisant la moyenne par paires ou la somme sur l'ensemble des colonnes, on obtient un score pour l'alignement

En outre, chaque algorithme implémente son propre calcul de score selon plusieurs critères, notamment:

- les modalités de prise en compte des pénalités de gap: ouverture, extension, fermeture;
- la prise en compte de la région concernée: dé favoriser les gaps dans les régions hydrophobes et les favoriser dans les régions hydrophiles. [9]

II.6 Matrices de score pour les protéines

II.6.1 Les matrices de substitutions

Deux grandes familles de matrices (log odds matrix)

➢ Matrices PAM

Les matrices PAM pour «Percent Accepted Mutation/Accepted Point Mutation », sont construites par étude de segments pris dans des séquences protéiques homologues (moins de 15% de différences).

PAM*x* : x% de mutations acceptées entre les séquences qui ont servi à construire la matrice.

Les fréquences de substitutions observées (ou probabilité conditionnelle: appelée "odd") sont transformées en logarithme de probabilité, normalisé en unité d'évolution. Le logarithme est utilisé pour que dans les programmes de recherche de ressemblance, la somme de ces éléments donne le logarithme de la probabilité pour la séquence entière (le modèle étant Markovien: indépendance de fréquences de substitution).

Les éléments diagonaux de la matrice indiquent une évolution sans substitution.

Pour PAM1, leur somme est telle qu'elle correspond à une probabilité de 99/100 (1 mutation pour 100 résidus: d'où le nom PAM : accepted point mutation)

L'indépendance des fréquences et les éléments de la matrice étant des logarithmes de fréquence,

On peut calculer PAM(N) en élevant PAM 1 à la puissance N, par exemple pour PAM 120, il faut multiplier PAM 1 par elle-même 120 fois.

➢ Matrice BLOSUM

Ces matrices BLOSUM (Blocks Substitutions Matrices) sont construites par analyse de séquences de protéines. Les séquences sont découpées en blocs (2000 résidus au total) par rapport au pourcentage d'acides aminés inchangés.

BLOSUM*x* : matrice obtenue à partir de séquences présentant au minimum x % d'identité (similitude) entre elles.

Une matrice "d'odds" est calculée à partir des blocs d'alignement pour chaque valeur de similitude, et ensuite chaque élément est transformé en unité d'information en prenant le logarithme du rapport de la valeur observée à la valeur qu'on obtiendrait au hasard. Cette matrice est ensuite normalisée. Les correspondances entre BLOSUM et PAM, basées sur la théorie de l'information sont:

- PAM250--->BLOSUM45
- PAM160--->BLOSUM62
- PAM120--->BLOSUM8

Les matrices : sont calculées selon le principe du log-odd-ratio.

$$\textbf{odd-ratio} = \frac{\Pr(\text{observé})}{\Pr(\text{attendu})}$$

odd-ratio (a,b) = $\frac{Pa,b}{Pa \times Pb}$ pour deux lettres a, b avec :

Pa,b : probabilité d'observer a aligné à b sur deux séquences homologues.

Pa × Pb : probabilité attendu d'aligner a à b sur deux séquences non-homologues.

log-odd-ratio = log (odd-ratio)

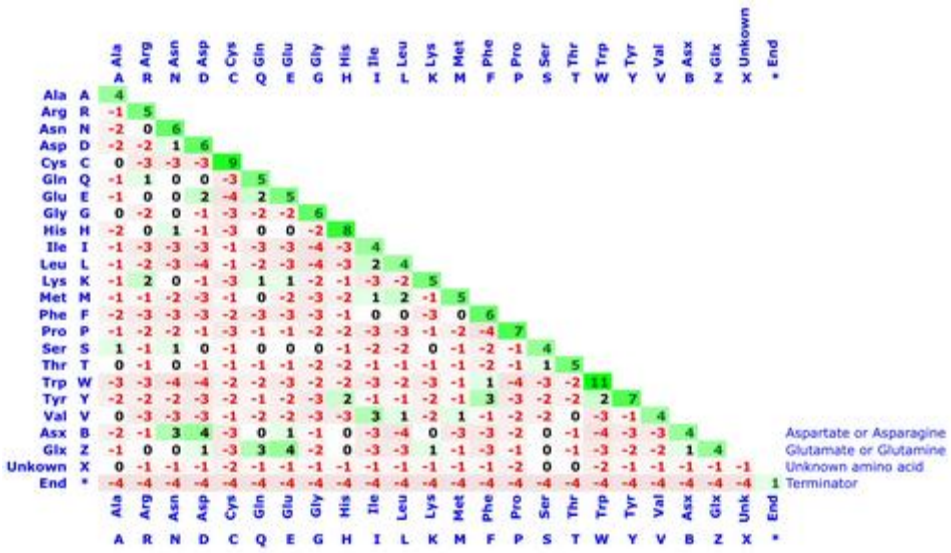


Figure II.7 : Représentation d'un exemple de la matrice BLOSUM

II.6.2 Choix de la Matrice Protéique

Le choix d'une matrice dé pend du type d'analyse que l'on veut faire. Il n'y a pas une matrice idéale et un grand nombre d'études comparatives sur les matrice sont mis en évidence (de manière schématique) que :

- Pour des séquences similaires et courtes, il est préférable d'utiliser une matrice BLOSUM élevée ou PAM faible.
- Pour des séquences divergentes et longues, il est préférable d'utiliser une matrice BLOSUM faible ou PAM élevée.
- La matrice BLOSUM 62 semble être la matrice la plus utilisée pour la comparaison avec les banques de données, et pour un grand nombre de logiciels d'alignement de séquence, elle semble être la matrice par défaut [13].

II.6.3 Comparaison des matrices PAM et BLOSUM

Plusieurs études comparatives ont été réalisées entre ces matrices performantes, largement utilisés pour réaliser des recherches dans les banques ou des alignements. La figure3.3 établit la correspondance entre ces deux types de matrice. La plupart des études ont montré que les matrices BLOSUM donnaient de meilleurs résultats que les matrices PAM. Deux raisons expliquent ce résultat:

- ces matrices sont réalisées à partir d'alignements locaux (régions structurellement conservées) alors que les matrices PAM incluent des régions très divergentes.
- les matrices BLOSUM sont plus et donc ont bénéficié d'un plus grand nombre de données à disposition.

Les algorithmes de recherche dans les banques FASTA [LipmanandPearson, 1985] utilise par défaut la matrice BLOSUM 62 [14].

Évaluation d'un Alignement

Évaluer un alignement revient alors à mesurer sa qualité en déterminant la distance qui sépare les deux séquences. Le score d'un alignement est la somme des scores de toutes les positions de bases (résidus) prises deux à deux [13].

***Exemple d'évaluation**:*

On peut attribuer une valeur positive à des symboles alignés identiques et une pénalité (valeur négative) à une substitution ou à un gap.

Si l'on considère l'exemple précédent:

Score (identité)=2

Score (substitution)=-1

Score (gap)= -2

Le score de cet alignement serait alors:

SEQ1: G A R F I E V H E L - - T F A T T C A T

SEQ2: G A R F I E L T H E V A S Y F - - C A T

score total = 2+2+2+2+2+2-1-1-1-1-2-2-1-1-1-2 -2+2+2+2 =**+3** [13]

Nous présenterons tout d'abord les méthodes permettant de mesurer, de donner un score à ces mutations. Ce score nous permettra alors de déterminer le score d'un alignement, défini comme la somme des scores des évènements mutationnels (illustrés sur la figure II.1) étant survenus entre deux séquences. Ensuite, nous décrirons les alignements de deux séquences et enfin l'alignement multiple de séquences [14].

II.7 La fonction profil

Cette fonction a pour objectif de calculer le profil d'un alignement A.

Ce profil est une représentation numérique d'un MSA qui représente les caractéristiques communes d'une famille de protéines. La fonction Profil est utilisée pour déterminer le degré d'appartenance d'une protéine à une famille. On peut signaler qu'il est utile dans l'alignement des séquences pas trop divergentes .Il permet de déterminer des régions conservées dans une séquence ou plusieurs. C'est la somme des fréquences d'apparition de chaque résidu dans chaque colonne de l'alignement.

II.8 Conclusion

Pour identifier la ressemblance entre les séquences protéiques, il est nécessaire de faire un alignement de ces séquences. Lorsque cet alignement est fait sur plus que deux séquences, est dit un alignement multiple des séquences (MSA). Le MSA détectent les régions qui ont été conservées lors de l'évolution Très souvent des domaines associés à une fonction clé de la molécule.

Chapitre III

III.1 Introduction

Les modèles de Markov cachés(MMC) sont des outils statistiques permettant de modéliser des phénomènes stochastiques. Ces modèles sont utilisés dans de nombreux domaines tels que la reconnaissance et la synthèse de la parole, la biologie, l'ordonnancement, l'indexation de documents, la reconnaissance d'images, la prédiction de séries temporelles, Pour pouvoir utiliser ces modèles efficacement, il est nécessaire d'en connaitre les principes.

Cette partie du chapitre a pour objectif d'établir les principes, les notations utiles et les principaux algorithmes qui constituent la théorie des modèles de Markov cachés(MMC).

A cet effet, nous commençons en présentant un historique des étapes les plus marquantes dans la construction de cette théorie, nous verrons que pour mieux modéliser les phénomènes étudiés, il est nécessaire de considérer un modèle ayant un pouvoir d'expression supérieur. Les modèles de Markov cachés en font partie. Nous présentons alors les profils MMC.

III.2 Définition d'un HMM

Un modèle de Markov caché (HMM) est un modèle probabiliste qui décrit en termes d'automate d'états la structure d'une séquence biologique. Il se compose d'états i et de transitions ai→j. A chacun des états correspond une distribution (P(X|i)) donnant la probabilité pour l'automate de produire la lettre X lorsqu'il se trouve dans l'état i.

Les transitions ai→j représentent la probabilité de passer de l'état i à l'état j. Les paramètres constitutifs du modèle sont donc de trois sortes : (1) la topologie de l'automate (états et arêtes), (2) les distributions de probabilité dans chacun des états, (3) les probabilités de transition entre états. Les lettres émises par les états seront ici les 20 acides aminés, mais le principe s'étend `a d'autres alphabets (ADN, codons, . . .).

Une fois un HMM établi, l'utilisation principale consiste à l'utiliser pour lire des séquences et leur donner un score. Ce score est la probabilité de générer la séquence au moyen du modèle considéré.

On définit la structure d'un HMM comme l'ensemble de ses états, son graphe de transition et son alphabets du point de vue de la théorie des langages, une protéine est donc un mot sur l'alphabet des acides aminés.

Un modèle de famille de protéines est composé :

- d'information décrivant les séquences protéiques de la famille (par exemple, une matrice);
- d'une règle d'acceptation permettant de décider l'appartenance ou non d'une séquence à la famille (par exemple, un système de score et un seuil).

III.3 Topologie du Modèle

La génération de la topologie d'un HMM est une étape importante et délicate car elle doit être conforme à la représentation biologique que l'on veut modéliser. Une représentation de plus en plus fine du modèle biologique entraine une capture de l'information de plus en plus importante (Eddy, 1998 ;Krogh*etal.*,1998).En générale on trouve un modèle linéaire (gauche-droite) ou un modèle ergodique(figure III.1)

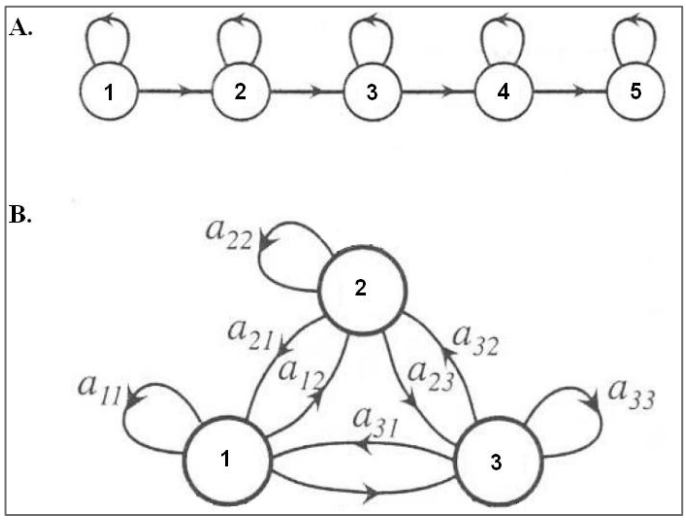


Figure III.1 : Deux structures différentes de HMM.

- A.HMM linéaire ou «gauche-droite» minimal, les transitions possibles sont d'un état caché vers l'état caché suivant ou sur lui-même. A partir de l'état caché suivant, il n'est pas possible de revenir dans l'état caché précédent. Nous utilisons ce modèle pour établir une estimation initiale de densités utilisées dans un modèle ergodique
- B.HMMergodique:touteslestransitionsentreétatssontpossibles.NousutilisonsceHMMpourse gmenterle génome.

Une structure couramment utilisée est celle des modèles de Bakis, dits également modèles gauche-droite (figure III.2), ainsi appelés parce qu'ils n'autorisent aucune transition d'un état vers un autre d'indice inférieur : les états qui se succèdent ont donc des indices égaux ou supérieurs aux précédents. Une fois dans le dernier état, le système est condamné à y rester : c'est pourquoi la probabilité initiale du premier état est posée égale à 1. [15]

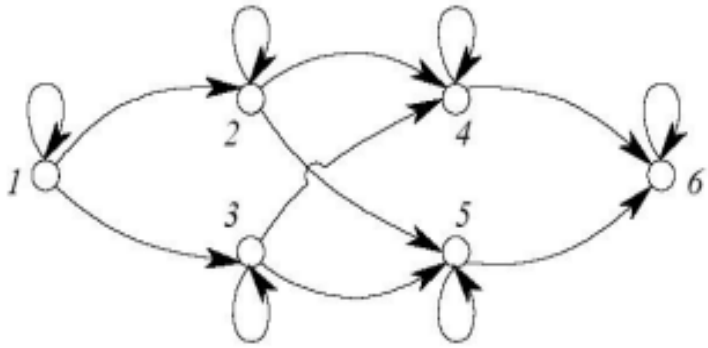


Figure III.2 : modèle de Bakis

III.4 Les Profils de Markov cachés (HMM-profils)

Ce sont les mieux adaptés pour modéliser les alignements multiples. A partir d'une famille de protéines, un HMM-profil peut-être réalisé pour chercher d'autres membres de cette famille. La figure III.3 représente un HMM-profil [16] .

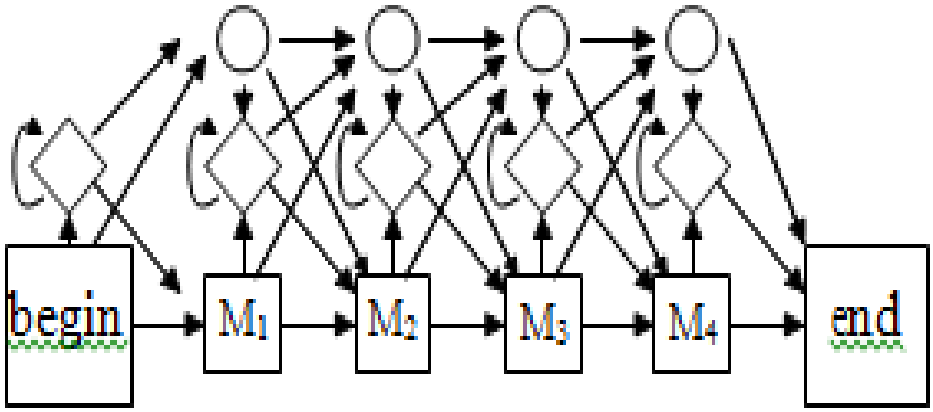


Figure III.3:Schéma d'un profil HMM

On a dans cet exemple :

4 états 'match' (carrés), 5 états 'insert' (losanges) et 4 états 'delete'(cercles)

Avec :

– Mj : état 'match' j représentant l'identité du résidu lors de l'alignement multiple

– Ij : état 'insert' j illustrant une insertion de résidus (la transition-boucle signifie que plusieurs insertions peuvent subvenir)

– Dj : état 'delete' permettant l'apparition de gaps entre deux résidus : cet état n'émet pas de résidus Score d'appartenance à une classe de protéine ;

Un HMM doit donc permettre d'évaluer si une séquence donnée appartient ou non à la famille qu'il modélise ; il faut calculer la probabilité que la séquence soit créée par un HMM élaboréà partir d'une famille.

III.5 Exemple de Profil MMC (HMM)

Le profil HMM que nous allons présenter est une topologie standard de modélisation de motifs des séquences biologiques. Il a été proposé par Krogh et Haussler en 1994.

III.5.1 Notation

1. ***N*** nombres d'états $(S_1,\ldots,S_N)$

2. ***M*** symboles de lettre dans l'alphabet

3. Les paramètres, λ:

   1. Distribution initial des états $\pi(i)$
   2. Probabilité de transition $a_{ij} = P(q_t = S_i \mid q_{t-1} = S_j)$. Tell que: $\sum_{i=1}^{N} a_{ij} = 1, \ \forall j$
   3. Probabilité d'émission $\mathbf{e_i(a)}$ probabilité que l'état ***i*** émette ***a***

4. Séquence de symboles: $O = O_1, O_2, \ldots, O_T$

5. Séquence d'états: $Q = q_1, q_2, \ldots, q_T$

III.5.2 Exemple de profil HMM de taille 5

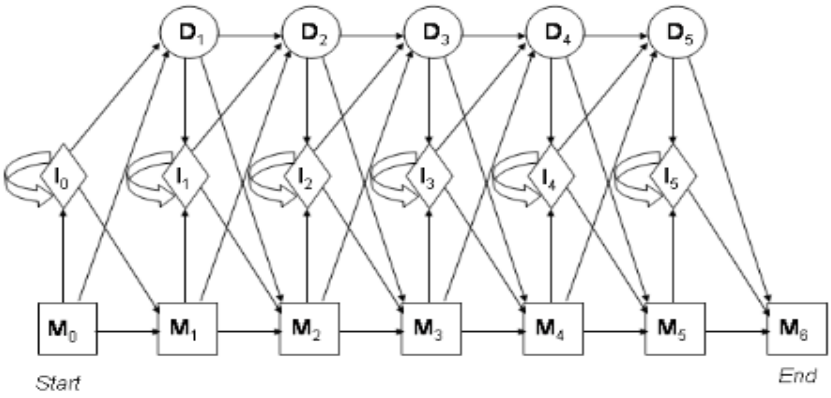


Figure III.4 : Exemple de profil MMC (HMM) à 5 états

Les états d'insertion (I_i) et de similitude (M_i) émettent les 20 acides aminés. L'état de délétion émet un gap (-). Les probabilités d'émission et de transition sont calculées à partir des données utilisées en phase d'apprentissage.

a) Estimation des paramètres :

$$\mathbf{e_k(\sigma)} = \frac{\mathbf{E_i(\sigma) + b}}{\sum_{\alpha} \mathbf{E_i(\sigma) + 20b}}$$

$$\mathbf{a(i,j)} = \frac{\mathbf{A(i,j)} + \mathbf{b}}{\sum_{\mathbf{l}}[\mathbf{A(i,l)} + \mathbf{b}]}$$

Avec $E_i(\sigma)$ le nombre d'occurrences, en phase d'apprentissage, où le symbole σ a été émis lorsqu'on se trouve dans l'état i, $A(i,j)$ le nombre de transition de l'état i vers l'état $j$et b est un pseudo compte qui tient compte des transitions non observées.

Dans l'exemple de la figure III.4, on a :

$$\mathbf{e_{M_6}(a)} = \mathbf{e_{M_0}(a)} = \mathbf{0}$$

$$\mathbf{e_{I_k}(a)} = \mathbf{p_a}$$

$$\mathbf{e_{D_k}(a)} = \mathbf{0} \qquad \mathbf{e_{D_k}(" - ")} = \mathbf{1}$$

$$\mathbf{e_{M_k}(a)} = \frac{\mathbf{E_i(a)} + \mathbf{b}}{\sum_{\alpha} \mathbf{E_i(a)} + \mathbf{20b}}$$

Avec $p_a$est la fréquence du résidu $a \in \sum$.

b) Construction du profile

- Cas de données étiquetées : Si les séquences sont alignées, nous avons déjà l'étiquetage des données. Sinon, nous pouvons le déterminer à partir de l'alignement, c.-à-d. repérer quel état est associé à un symbole dans chaque séquence. Ainsi, nous devons définir le nombre d'états match, soit la topologie du profil, et calculer ses paramètres à partir des données étiquetées.

Exemple : soit l'alignement des 4 séquences protéiques suivantes :

```
VG--H
V---N
VE--D
IAADN
```

La longueur du profil MMC peut être prise comme la moyenne des longueurs des séquences. Les séquences ont des longueurs respectives 3, 2, 3 et 5, soit une moyenne de 3.25. Le profil possède un état silencieux start M_0, 3 états match M_1, M_2, M_3, 4 états d'insertion I_0, I_1, I_2, I_3, I_4, 3 états de délétion D_1, D_2, D_3 et un état silencieux end M_4.

Pour estimer les paramètres du profil, nous devons étiqueter les données en se basant sur l'alignement multiple. Les positions présentant moins de 50% de gaps seront considérées comme les états match. Celles présentant plus de 50% de gaps correspondront à des états d'insertion :

V	G	-	-	H
V	-	-	-	N
V	E	-	-	D
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

Ce qui conduit à l'étiquetage suivant des séquences :

V	G	H		
M_1	M_2	M_3		
V	-	N		
M_1	D_2	M_3		
V	E	D		
M_1	M_2	M_3		
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

A partir de cet étiquetage, nous pouvons estimer les paramètres. Si nous prenons b=1 comme pseudo compte, nous obtenons :

$$e_{M_1}(V) = \frac{3+1}{4+20}$$

$$a_{M_2 I_2} = \frac{1+1}{(2+1)+(1+1)+(0+1)}$$

Les trois sommes du dénominateur correspondent à toutes les transitions possibles à partir de l'état M_2. Dans ce cas d'entrainement, nous avons 2 transitions de M_2 vers M_3, une transition de M_2 vers I_2 et aucune transition de M_2 vers D_3.

- Cas de séquences non étiquetées : étant données des séquences non étiquetées partageant un modèle. Nous pouvons utiliser le profil MMC pour déterminer ce modèle, étiqueter les données et construire l'alignement multiple de ces séquences. Pour cela, il faut suivre les étapes suivantes :

1. Estimer la taille du profil : La longueur du modèle est calculée comme la moyenne des longueurs des séquences rentrant dans la détermination du profil. Si une séquence contient le motif pour lequel on veut calculer le profil mais qu'elle est trop longue par rapport au modèle, on utilisera d'autres précisions biologiques pour réduire la taille de celle-ci (repérer l'emplacement du motif).

2. Topologie : construire un profil HMM avec L+2 états Match dont M_0 et M_{L+1} sont silencieux.
3. Apprentissage des paramètres : poser les paramètres de départ tels que $a_i(M_j) >> a_i(I_j)\ or\ a_i(D_j)$.
4. Entrainer le modèle en utilisant l'algorithme de (Baum Welch.)
5. Déterminer le motif : utiliser l'algorithme de Viterbi pour trouver le meilleur chemin qui pourra générer chacune des séquences. Pour le cas du profil MMC, l'algorithme de Viterbi

est utilisé sous sa formulation logarithmique (log odds : The odds ratio of A is P(A)/P(¬A)) pour réduire l'effet de la taille du modèle.

6. Alignement multiple des séquences : le motif déterminé en 4, peut être utilisé pour réaliser un alignement multiple de séquences. Si $\boldsymbol{O_t^d}$ et $\boldsymbol{O_u^c}$ sont émises par un même état match, on pourra aligner les positions $\boldsymbol{t}$ et $\boldsymbol{u}$.
7. Amélioration du modèle : la topologie du modèle peut être raffinée. Si plus de 50% des séquences comprennent un gap D_i, supprimer l'état match M_i, D_i et I_i de la topologie. Si plus de 50% des séquences font entrer un état d'insertion I_i, ajouter des états match, délétion et insertion entre Mi et Mi+1, en nombre égal à la moyenne de la longueur de ces insertions.
8. Ré-estimation des paramètres : si le modèle est retouché, par l'étape 6, ses paramètres doivent être ré-estimés. Pour cela, re-étiqueter les séquences alignées avec le nouveau modèle et recalculer les probabilités d'émission et de transition. Si le pourcentage des états retouchés est important par rapport à tout le modèle, il faut refaire l'entrainement avec Baum Welch.

III.6 Position Specific Scoring Matrices(PSSMs)

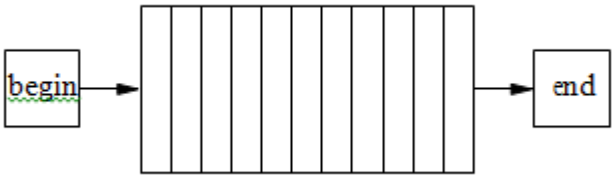


Figure III.5–Schéma d'une PSSM.

Les matrices poids-position, position specific scoring matrices (PSSM), décrivent des positions consécutives caractéristiques de la famille dans la séquence protéique.

Les colonnes de la matrice représentent les positions consécutives; les lignes représentent les acides aminés. Les coefficients de la matrice sont des poids représentant la probabilité d'observation des acides aminés aux positions correspondantes. Pour noter une séquence protéique, on cherche à aligner au mieux une partie de la séquence de la protéine avec la matrice. Le score de l'alignement est la somme des poids correspondants. On fixe alors un seuil au-delà duquel les protéines sont reconnues comme membres de la famille.

Le problème des PSSMs est qu'elles reconnaissent mal des protéines de la famille ayant subi des mutations de type insertion ou suppression car elles décrivent des positions successives. Des alignements non consécutifs (recherche de sous mots au lieu de facteurs), avec une pénalité de gap affine et dépendant de la position, ont été utilisés mais un modèle adapté à ces types de mutations est préférable. Les PSSMs restent tout de même un bon moyen d'identifier des candidats pour la famille.

III.7 Apprentissage des modèles

La question au centre de ce paragraphe est celle de la construction d'un HMM paramétré de manière à modéliser de façon satisfaisante les séquences à traiter. Dans le cas le plus favorable, le HMM recherché peut être construit directement à partir des connaissances a priori dont on dispose sur le domaine. Dans la plupart des applications, le HMM doit être construit à l'aide d'un algorithme d'apprentissage. Ces algorithmes sont appliqués sur un ensemble de séquences représentatives des séquences que l'on souhaite modéliser et appelées séquences d'apprentissage. On peut distinguer dans le problème de l'apprentissage d'un HMM deux cas de figure distincts, suivant que la structure (nombre d'états du HMM et transitions autorisées) est connue ou ne l'est pas. Lorsque la structure est connue, le problème se réduit à un problème d'entrainement consistant à estimer les paramètres numériques (distributions de probabilité de première visite, de transition et de génération) de manière à expliquer au mieux les séquences d'apprentissage. Pour certaines applications, on ne dispose pas de connaissances suffisantes pour inférer naturellement la structure du HMM. L'apprentissage devient alors encore plus difficile. Il ne suffit plus de paramétrer une structure mais il faut également déduire cette structure des exemples fournis.

III.8 Calcul d'un score

Une fois le modèle construit, la question est de savoir, étant donné une protéine étudiée, si le HMM profil reconnait cette séquence. Cette reconnaissance peut être évaluée, dans un premier temps, par le calcul d'un score, aussi appelé log odds ratio.

On a vu dans la section précédente, la probabilité de génération d'une séquence S par un HMM H, notée P(S /H). Cette probabilité est comparée à la probabilité de générer cette même séquence S par un modèle dit nul, notée P(S/nul).le score d'une séquence S étant donné un HMM profil H s'apparente au test du rapport vraisemblance et s'obtient par la formule :

$$\text{Score}\,(S|H) = \log \frac{P(S\,/H)}{P(S/\text{nul})}$$

III.9 Phylogénies

Là où les modèles de Markov fournissent une vision longitudinale des séquences, les modèles phylogénétiques apportent des informations supplémentaires basées sur les liens évolutifs entre séquences. Ces dernières sont reliées par un arbre dont les feuilles constituent les séquences actuelles et les nœuds internes les séquences ancestrales. La longueur des branches correspond au nombre de substitutions par site entre les deux nœuds reliés. Les évènements de susbtitution sont modélisés par Une approche phylo-HMM pour la recherche de séquences [17].

Un arbre phylogénétique est défini par plusieurs paramètres :

1). la topologie de l'arbre (généralement binaire).

2). les longueurs de branches.

3). la ou les matrices de substitution qui agissent le long des branches de l'arbre.

III.10 Conclusion

Dans ce chapitre nous avons présenté les profils MMC qui ont pour objectif de modéliser une famille protéique, c.à.d. un ensemble de séquences homologues. Et dans les précédents modèles qu'un alignement de séquences est la donnée clef dans ce genre de modélisation. Ces profils modélisent les familles de protéines par une séquence significative (conservées) d'une fonction commune de ces séquences. La séquence significative est donc le motif recherché.

Chapitre IV

IV.1 Introduction

Les familles de séquences de protéines connues de la base de données PFAM sont apprises pour former des modèles de Markov cachés. Chaque modèle représente une généralisation de la façon dont les acides aminés sont distribués dans un groupe de séquence (famille de protéine). Par la suite, les séquences consensus sont prédites par comparaison avec ces différents modèles. Un score est associé à chaque séquence comparée à un modèle donné ; le modèle donnant le plus haut score à un fragment prédit alors la séquence consensus.

Dans ce chapitre, nous allons chercher les motifs protéiques (séquence consensus) en utilisant le modèle de Markov caché. Pour cela, nous avons utilisé dix modèles (familles protéiques) prise de la base de données PFAM. Nous allons aussi présenter dans ce chapitre les résultats obtenus par nos expérimentations et notre simulation sur Matlab.

IV.2 Description des Modèles de test

Nous avons téléchargé les modèles suivants :

1) Modèle 1 (Mitochondria Localisation Sequence)

Cette famille contient une protéine trouvée dans les eucaryotes. Les protéines de cette famille sont typiquement entre 240 et 613 acides aminés de longueur. Elle est considérée comme importante pour la fonction de la membrane et est exprimée dans l'endothélium de la cornée et de l'ovaire.

2) Modèle 2 (Adipogenin)

Cette famille de protéines est impliquée dans la stimulation de la différenciation et le développement des adipocytes.

3) Modèle 3 (LAT)

Un lymphocyte T protéine adaptateur qui couple le récepteur de l'antigène à des voies de signalisation en aval.

4) Modèle 4 (AKAP2_C) [Une protéine kinase-2 d'ancrage C-terminale]

Cette famille comprend la terminaison C de la protéine kinase A-2 d'ancrage (AKAP2). Il comprend le site où les sous-unités régulatrices (RII) de la protéine kinase sont toutes liées.

5) Modèle 5 (ALMS motif)

Ce domaine se trouve à l'extrémité C-terminale de la protéine de syndrome Alstrom 1 (ALMS1), KIAA1731 et C10orf90 [1-2].

6) Modèle 6 (40S_S4_C)

Ce domaine se trouve à l'extrémité C-terminale de ribosomique 40S des protéines de S4.

7) Modèle 7 (AbLIM_anchor)

AbLIM_anchor est un domaine situé entre le domaine des protéines LIM liant l'actine et la vilin- tête LIM. Il est probable que ce domaine est impliqué dans l'ancrage abLIMs à faisceaux d'actine circonférentielles dans des types cellulaires spécifiques.

8) Modèle 8 (ATF7IP-BD)

Est une région conservée à court d'activation du facteur de transcription 7-interacting protein 1 trouvé chez les eucaryotes supérieurs. Ce domaine semble se lier plusieurs protéines clés tels que

TFIIE-alpha et bêta-TFIIE ainsi le régulateur transcriptionnel Sp1 qui font partie de la machinerie transcriptionnelle.

9) Modèle9 (zf-RRPl_C4)

Ribonucléoprotéique putative doigt de zinc de type PF C4 vos commentaires

Ceci est une famille de protéines en grande partie spécifique de microsporidies . Un membre est annoté comme étant un ribonucléoprotéique . La famille porte deux paires de résidus CXXC suggérant qu'il y ait liaison à l'ADN.

10) Modèle10 (AIM5) (Altered héritage des mitochondries 5)

AIM5 mitochondriale est une protéine de la membrane intérieure fongique. Elle est une composante du système mitochondrial d'organisation interne de la membrane (MINOS / MITOS), qui favorise la morphologie mitochondriale normale.

Le tableau suivant IV.1 représente les modèles à étudier dans notre travail.

Modèles	Numéro de Succession	longueur	Nbre de séquences	Nom de la famille des protéines	Déscription
Modèle1	PF14962	192	13	AIF-MLS	Mitochondria Localisation Sequence
Modèle2	PF15202	79	3	Adipogenin	Adipogenin
Modèle3	PF15234	240	12	LAT	Linker for activation of T-cells
Modèle4	PF15304	348	17	AKAP2_C	A-kinase anchor protein 2 C-terminus
Modèle5	PF15309	136	20	ALMS_motif	ALMS motif
Modèle6	PF16121	48	86	40S_S4_C	40S ribosomal protein S4 C-terminus
Modèle7	PF16182	371	51	AbLIM_anchor	Putative adherens-junction anchoring region of AbLIM
Modèle8	PF16788	222	31	ATF7IP_BD	ATF-interacting protein binding domain
Modèle9	PF17026	108	14	zf-RRPl_C4	DESC Putative ribonucleoprotein zinc-finger pf C4 type
Modèle10	PF17050	76	72	AIM5	Altered inheritance of mitochondria 5

Tableau IV.1 : Description des 10 modèles

IV.3 Alignement des séquences des familles de protéines

Pour construire le profile lié à chaque famille de protéine, nous procédons à l'alignement des séquences qui vont servir pour sa construction.

Pour cela nous avons utilisé la boite à outil « Bioinformatique » de Matlab. La fonction « multialign » nous a permis de réaliser les alignements des séquences de chacune des familles de protéines testés et de le comparer avec les alignements de la base de donnés PFAM.

Les figures IV.1 et IV.2 montrent respectivement les alignements obtenus par PFAM et celui qui est fait par la fonction de Matlab. L'alignement fait par matlab est un alignement progressif tel que décrit au chapitre 2.

```
H0WJV0_OTOGA/1-192 MYLRRAVSKTLALPLRAPPSPAP-LGKDASLRRMSSNKFPGSSGSNMIYYLLVGVTVSAGGYYTYKTVTSDQAKHTEHITN
I3NAZ6_SPETR/1-191 MYLRRVVSKTLALPLRAPPNPAP-LGKDASLRWMSSNKFPGTSGSNMIYYLVVGVTVSAGGYYTYKTVTSKQAKHTEHITN
G3TW02_LOXAF/1-192 MYLRRVVSKTLALPLRAPPGSTP-FRKDASLHRLSSNKLPGSSGSNLMYYLVVGVTVSAGGYYTYKAVRSEQARHTEHTTN
L9L670_TUPCH/1-187 MYLRRAVSRTLALPLRAPPSAAP-LGKDASLRRMSSSKFPGPSGSNMIYYLVVGVTVSAGGYYTYKAVASEPAKHTEHVTT
F6UJ52_MONDO/1-188 MYLRRAVGKTLTLPLRTAPSSPPLLWKHASLREMSSKKAPGSAGSNMIYYLFVGVTFSAGGYYAYRTVTSDQARYTERINN
G1SR45_RABIT/1-186 MYLRRAVSKTLALPLRAPPSPAP-LGKDASLRRVSSIKFPGASGSNMIYYLVVGVTVSAGGYYTYKSVTSEQAKHTEHITN
L5L146_PTEAL/1-192 MYLRRAVSKTLALPLRAPPSPAP-LGKDASLRWMSSNKFPGSSGSNMLYYLVVGVTVSAGGYYTYRTVTSEQARHTEHITN
G1PCV0_MYOLU/1-193 CTSAGPVSRSLALPLRAPPSPAP-LRKDASLRWMSSNKSPGSSGSNMIYYLVVGVTVSAGGYYTYRTVTSEQVKHTGHIKN
HUMMR_BOVIN/1-192  MYLRRAVSKTLALPLRAPPGPAP-LRKDASLRWISSNKFPGSSGSNMIYYLVVGVTVSAGGYYTYKRVTSGKAKRSDHVTD
D3K5J1_PIG/1-190   MYLRRAVSKTLALPLRAPPGPAP-LRKDASLRWMSSNKFPGTSGSNMMYYLLVGVTVSAGGYYTYKTV-KRQAKHSEHTAN
F7HQ57_MACMU/1-191 MYLRRAVSKTLALPLRAPPNPAP-LGKDASLRRMSSNRFPGSSGSNMIYYLVVGVTVSAGGYYAYKTVTSDQAKHTEHITD
M3YXG5_MUSPF/1-192 MYLRRAVSKTLALPLRAPPGPAP-LRKDASLRWMSSNKWFGSSGSNMIYYLVVGVTVSAGGYYTYRTITSEQGKQTERVTN
D4A4W7_RAT/1-193   MYLRRAVSKTLALPRRAPPGPAP-LGKDASLRRVSSSKFPGPSGSNMIYYLVVGVTVSAGGYYTYKAFTSKQVRHTEHVTD
```

Figure IV.1 : Séquences alignés fournis par la base des données PFAM (Modèle 1)

```
H0WJV0_OTOGA/1-192 MYLRRAVSKTLALPLRAPPSPAPL-GKDASLRRMSSNKFPGSSGSNMIYYLLVGVTVSAGGYYTYKTVTSDQAKHTEHITNM
I3NAZ6_SPETR/1-191 MYLRRVVSKTLALPLRAPPNPAPL-GKDASLRWMSSNKFPGTSGSNMIYYLVVGVTVSAGGYYTYKTVTSKQAKHTEHITNL
G3TW02_LOXAF/1-192 MYLRRVVSKTLALPLRAPPGSTPF-RKDASLHRLSSNKLPGSSGSNLMYYLVVGVTVSAGGYYTYKAVRSEQARHTEHTTNL
L9L670_TUPCH/1-187 MYLRRAVSRTLALPLRAPPSAAPL-GKDASLRRMSSSKFPGPSGSNMIYYLVVGVTVSAGGYYTYKAVASEPAKHTEHVTTL
F6UJ52_MONDO/1-188 MYLRRAVGKTLTLPLRTAPSSPPLLWKHASLREMSSKKAPGSAGSNMIYYLFVGVTFSAGGYYAYRTVTSDQARYTERINNM
G1SR45_RABIT/1-186 MYLRRAVSKTLALPLRAPPSPAPL-GKDASLRRVSSIKFPGASGSNMIYYLVVGVTVSAGGYYTYKSVTSEQAKHTEHITNL
L5L146_PTEAL/1-192 MYLRRAVSKTLALPLRAPPSPAPL-GKDASLRWMSSNKFPGSSGSNMLYYLVVGVTVSAGGYYTYRTVTSEQARHTEHITNL
G1PCV0_MYOLU/1-193 CTSAGPVSRSLALPLRAPPSPAPL-RKDASLRWMSSNKSPGSSGSNMIYYLVVGVTVSAGGYYTYRTVTSEQVKHTGHIKNL
HUMMR_BOVIN/1-192  MYLRRAVSKTLALPLRAPPGPAPL-RKDASLRWISSNKFPGSSGSNMIYYLVVGVTVSAGGYYTYKRVTSGKAKRSDHVTDL
D3K5J1_PIG/1-190   MYLRRAVSKTLALPLRAPPGPAPL-RKDASLRWMSSNKFPGTSGSNMMYYLLVGVTVSAGGYYTYKTVKR-QAKHSEHTANM
F7HQ57_MACMU/1-191 MYLRRAVSKTLALPLRAPPNPAPL-GKDASLRRMSSNRFPGSSGSNMIYYLVVGVTVSAGGYYAYKTVTSDQAKHTEHITDL
M3YXG5_MUSPF/1-192 MYLRRAVSKTLALPLRAPPGPAPL-RKDASLRWMSSNKWFGSSGSNMIYYLVVGVTVSAGGYYTYRTITSEQGKQTERVTNL
D4A4W7_RAT/1-193   MYLRRAVSKTLALPRRAPPGPAPL-GKDASLRRVSSSKFPGPSGSNMIYYLVVGVTVSAGGYYTYKAFTSKQVRHTEHVTDP
```

Figure IV.2 : Séquences alignés par une fonction de Matlab (Modèle 1)

Les alignements paraissent proches. Pour estimer le degré de leur ressemblance, nous avons alignés les séquences consensus des deux alignements par la fonction « nwalign », le score obtenu est de 370, valeur incluse dans l'intervalle des scores (Tableau IV.2). Ceci dénote d'une bonne ressemblance entre les deux alignements.

Les figures IV.3 et IV.4 nous donnent l'alignement des séquences ainsi que les séquences consensus obtenues respectivement par l'alignement de PFAM et celui fait par Matlab, en utilisant la fonction « multialignviewer ».

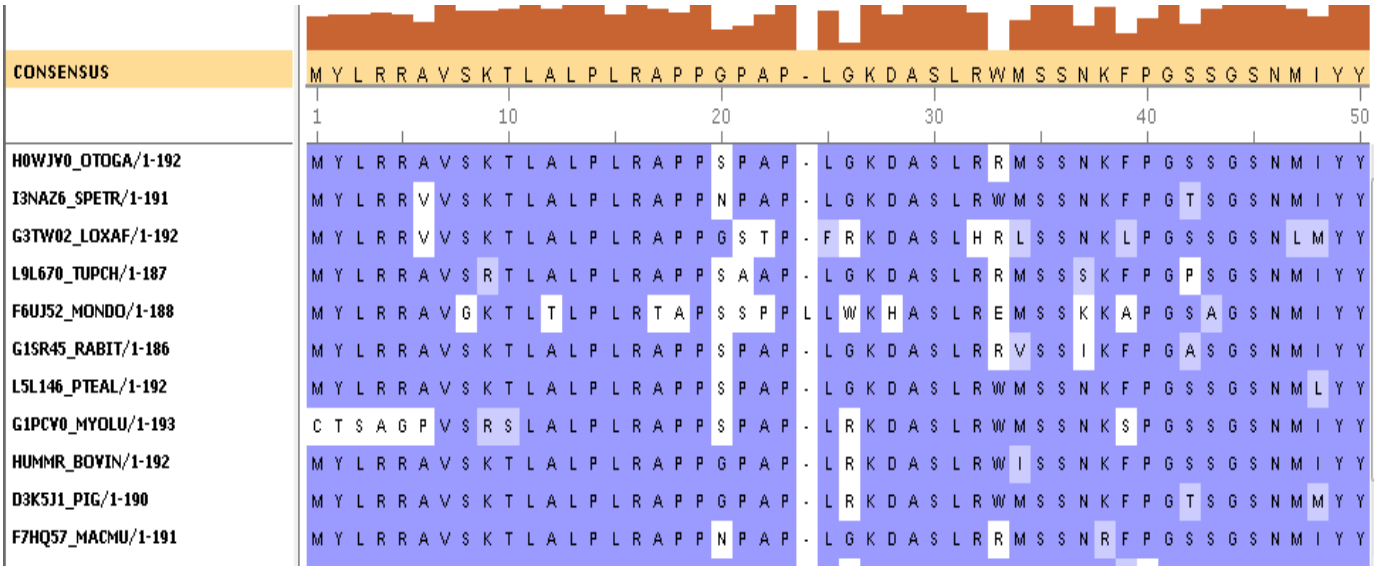


Figure IV.3: Séquences téléchargé alignés par PFAM (Modèle 1)

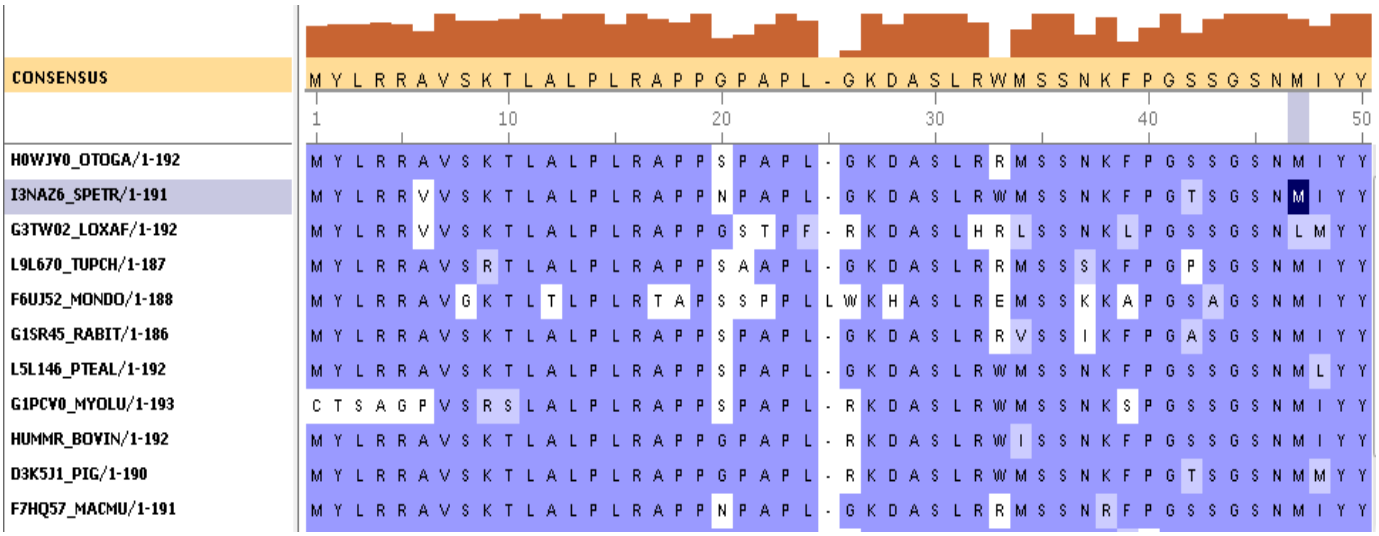


Figure IV.4 : séquences alignés par Matlab (Modèle 1)

Cette fonction permet d'afficher l'alignement avec une possibilité de régler interactivement en donnant la position d'un symbole voulu. La séquence consensus comme un résultat de l'alignement est affichée en haut de la figure.

IV.4 Construction du Profil

```
Model1 =

        ModelLength: 192
           Alphabet: 'AA'
      MatchEmission: [192x20 double]
     InsertEmission: [192x20 double]
       NullEmission: [1x20 double]
             BeginX: [193x1 double]
             MatchX: [191x4 double]
            InsertX: [191x2 double]
            DeleteX: [191x2 double]
    FlankingInsertX: [2x2 double]
              LoopX: [2x2 double]
              NullX: [2x1 double]
```

Figure IV.5 : Construction du profil 1 du modèle 1

La figure ci-dessus décrit les paramètres de notre modèle de Markov de base utilisé pour déterminer le profile 1 de la famille de protéine référencée sous modèle 1.

La taille (longueur) du modèle correspond à celle donnée par le tableau IV.1. Si celle-ci n'est pas connue, on prendra la moyenne des longueurs des séquences de la famille de protéine.

Ce modèle est construit par la fonction « hmmstruct » suivant les paramètres de la figure IV. Présentés en annexe.

Pour estimer ce profil, nous avons utilisé la fonction « hmmprofestimate », en utilisant le modèle précédent et l'alignement tel que réalisé au paragraphe IV.3 comme paramètres de cette fonction.

Ainsi, nous avons réalisé cette opération pour la construction des profils des 10 familles de protéines étudiés (Tableau IV.1).

IV.5 L'arbre phylogénétique

La figure suivante indique l'apparenté entre les séquences du modèle 1.

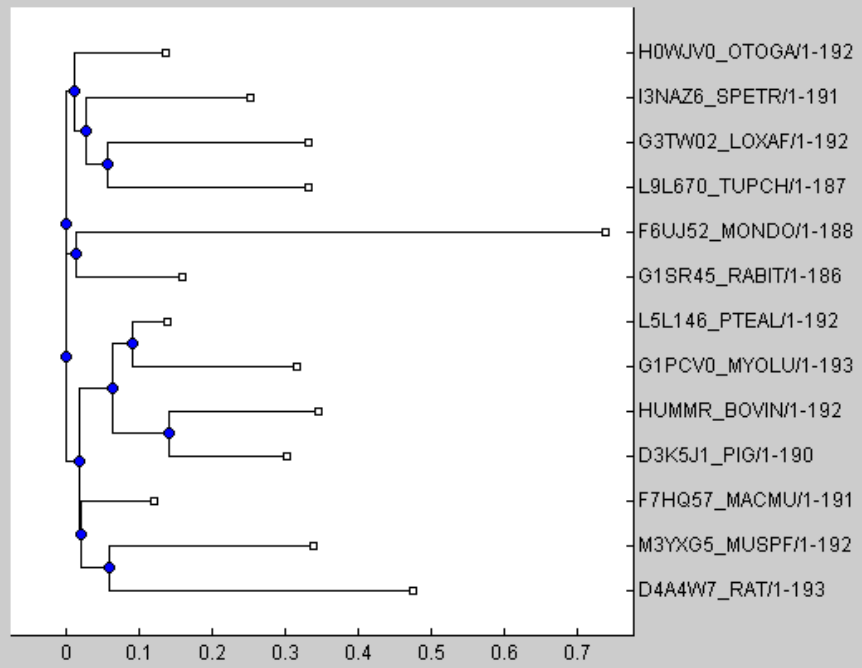


Figure IV.6 : Arbre phylogénétique des séquences du profile 1

IV.6 Probabilité d'émission des symboles et de transition entre les états

Les probabilités d'émission des symboles et de transition entre les états est données par les figures suivantes (IV.7 à IV.9)

L'Échelle sur la droite des figures (IV.7 et IV.8) indique les probabilités logarithmiques. Cette échelle est représentée avec une palette des couleurs, cette dernière lorsqu'elle est d'une couleur clair indique que la probabilité d'émettre tel symbole est grande et lorsqu'elle est dégradée en allant vers une couleur foncée indique que la probabilité est faible et cela d'après un alignement fait précédemment.

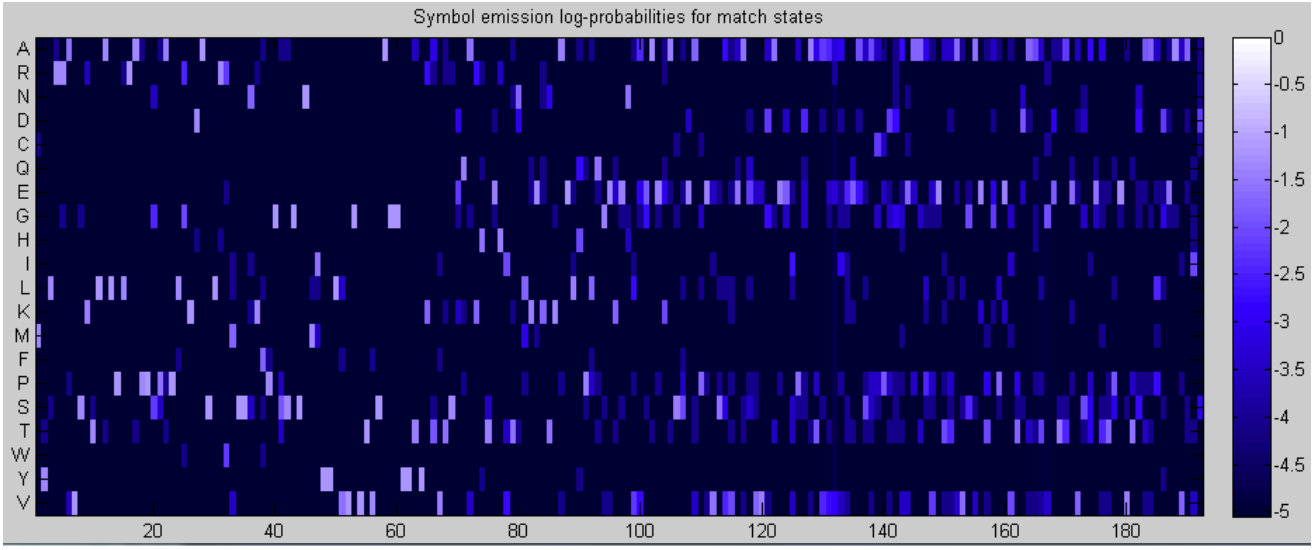


Figure IV.7 : Probabilités logarithmiques d'émission des symboles pour les états matchs.

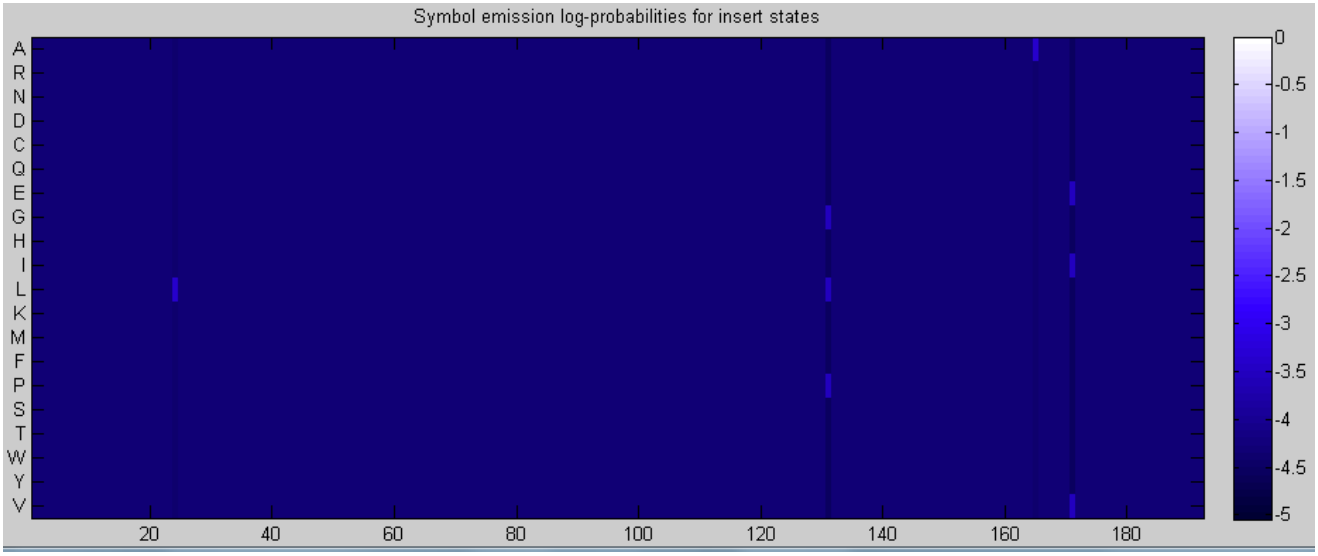


Figure IV.8 : Probabilités logarithmiques d'émission des symboles pour les états d'insert

La Figure IV.9 représente les probabilités logarithmiques de transition possible pour les états du profil 1.

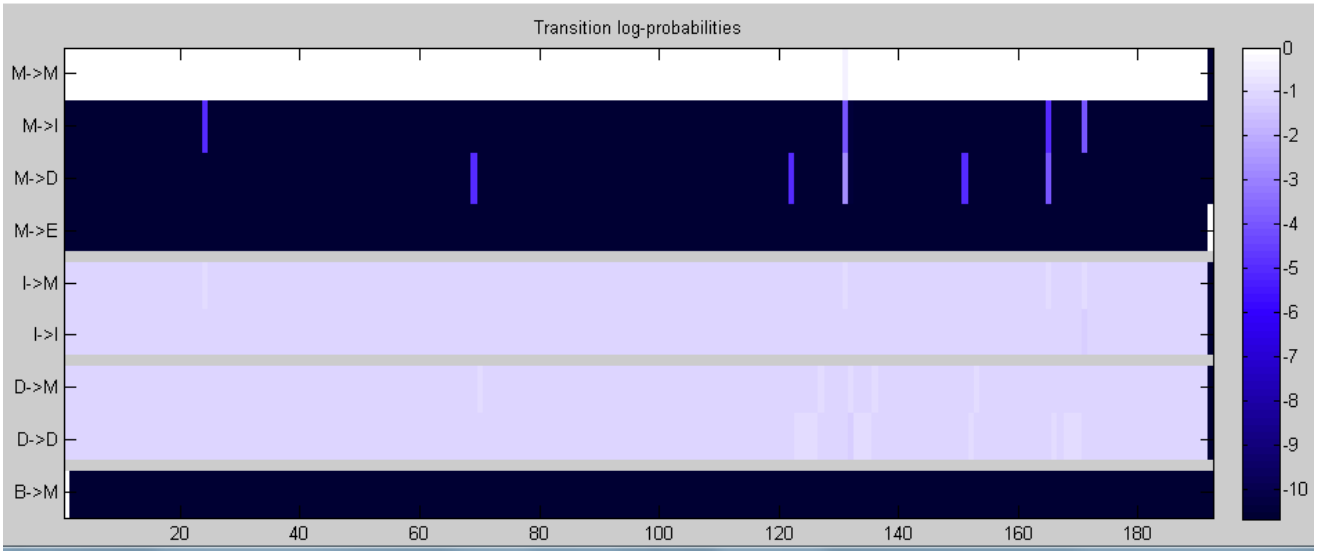


Figure IV.9 : Probabilités logarithmiques de transition entre les différents états

Les états possibles que nous avons :

M>M Transition d'un symbole depuis un état match vers un autre état match.

M>I Transition d'un symbole depuis un état match vers un état insertion.

M>D Transition d'un symbole depuis un état match vers un état de délétion.

M>E Transition d'un symbole depuis un état match vers l'état end.

I>M Transition d'un symbole depuis un état d'insertion vers un état match.

I>I Transition d'un symbole depuis un état d'insertion vers un autre état d'insertion.

D>M Transition d'un symbole depuis un état de délétion vers un état match.

D>D Transition d'un symbole depuis un état de délétion vers un autre état.

B>M Transition d'un symbole depuis un état begin vers un état match.

IV.7 Estimation des intervalles de scores

Afin de déterminer si une séquence donnée correspond à l'un des modèles de test, nous devons estimer l'intervalle des scores relatif à chaque profil d'une famille de protéines. Ceci définira le seuil d'appartenance qui permettra de savoir si une séquence protéique requête appartienne ou pas à la famille de protéines représenté par ce profil.

Modèles	Limitation du score(Pfam) (seq,Modèles)	Limitation du score(Matlab) (seq,Modèles)	Longueur du consensus (PFAM)	Longueur du consensus (Matlab)
Modèle1	[298.0980 484.1346]	[302.7085 478.1568]	202	197
Modèle2	[113.6605 124.5612]	[113.6605 123.9763]	79	79
Modèle3	[291.1559 531.6407]	[325.3485 564.8855]	305	289
Modèle4	[48.1219 735.7644]	[85.8784 731.8871]	425	397
Modèle5	[104.9965 278.3956]	[106.6648 276.2902]	168	154
Modèle6	[64.0664 153.9222]	[62.4821 153.5227]	51	52
Modèle7	[327.4625 928.3352]	[334.2781 876.6709]	885	602
Modèle8	[186.8468 440.4550]	[173.6804 422.4333]	292	232
Modèle9	[110.2806 196.5566]	[113.5268 188.7612]	129	125
Modèle10	[53.9551 174.6974]	[48.8707 175.6301]	181	105

Tableau IV.2 : Résultats de la manipulation 1 sur les 10 profils (modèles)

Nous faisons remarquer que les longueurs des séquences consensus obtenues des profils de PFAM et celles obtenues par les profils construit par matlab sont proches.

IV.8 Score de vraisemblance

IV.8.1 Scores entre les séquences des profils et les profils eux même

Le tableau suivant représente les scores entre des échantillons de séquences utilisés pour la construction des profils avec les différents profils obtenus. Ces scores sont calculés par la fonction de Matlab « hmmprofalign » en donnant comme paramètres, la séquence requête et le profil construit.

Meilleur Score	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10
Seq1	450.8676	125.9378	-9.8435	-23.6685	-87.4046	-152.4087	-26.5523	-34.2755	-112.4672	-111.3160
Seq2	3.2018	116.5854	9.2018	4.2018	-0.5530	-37.5152	6.0319	6.0319	0.8620	6.4470
Seq3	-44.6798	-153.4873	535.9915	-30.5479	-128.7233	-199.8542	-21.7069	-62.1135	-145.1304	-151.2969
Seq4	-130.3299	-229.4582	-79.0618	465.5478	-177.9981	-259.9068	1.4215	-112.1653	-224.4335	-206.3733
Seq5	-20.3775	-64.8706	-8.4170	-13.1046	171.2341	-85.3840	-13.2296	-16.0712	-46.7650	-43.5445
Seq6	-26.3422	-27.9720	- 30.4481	-23.0074	-18.4733	153.5227	-18.9106	-22.2505	-32.7739	-26.7710
Seq7	-150.4588	-234.9406	-127.0572	-101.1665	-184.2289	-252.0763	870.2914	-147.8894	-214.1442	-203.2869
Seq8	-53.7873	-147.7991	-45.7296	-40.7856	-101.7290	-169.2871	-39.3006	367.6993	-134.2222	-136.5436
Seq9	-15.5943	-44.2112	-9.9806	-6.6420	3.9243	-64.8148	-8.7009	7.7317	145.3727	-33.5753
Seq10	-10.0899	-37.6150	-4.2626	-5.8970	-13.2025	-43.2707	-12.8774	-10.7776	-29.2239	169.0201

Tableau IV.3 : Scores entre les profils et les séquences échantillons

Il est clair dans le tableau précédent que le score a une grande valeur dans la diagonale, nous pouvons interpréter ce résultat par l'appartenance de la séquence au profil étudié. En effet, le

numéro de la séquence indique le modèle auquel elle appartient. Les scores obtenus dans les cas de reconnaissance appartiennent aux intervalles du score calculés précédemment (tableau 4.2). Par contre, le reste des résultats sont plus petits et n'appartiennent pas à l'intervalle du score.

IV.8.2 Scores entre les séquences consensus et les profils utilisés

Le tableau suivant représente les résultats des scores entre les séquences consensus et les profils étudiés, ce score est calculé par la fonction de matlab « hmmprofalign ».

Meilleur Score	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8	Model9	Model10
Css1	498.9989	-130.0167	2.4626	-14.9416	-93.6913	-151.6557	-31.8537	-42.0770	-125.3302	-124.3025
Css2	2.2018	126.9011	8.2018	6.4470	1.0319	-37.9774	6.4470	6.0319	0.4470	-2.7740
Css3	-48.1651	-146.1958	618.7233	-17.4072	-126.1959	-196.1400	-30.0392	-67.4294	-157.0951	-162.5926
Css4	-121.4221	-227.5916	-85.4842	639.2039	-174.0999	-258.3218	-34.6628	-126.6888	-228.2622	-227.7912
Css5	-21.7269	-64.4955	-12.5869	-6.2054	285.0947	-83.9040	-29.3864	-5.3363	-33.3021	-63.7416
Css6	-22.8232	-19.8400	-29.0557	-24.6814	-26.4033	158.5196	-27.2134	-15.5909	-29.2227	-27.3643
Csst7	-160.1957	-237.6832	-118.0849	-88.3817	-176.2379	-250.0429	757.9498	-159.0471	-223.8493	-224.8878
Css8	-47.5274	145.9407	-48.6983	-28.3916	90.5740	-172.5944	-61.1438	470.9823	-115.7817	-139.7768
Css9	-13.7327	-40.1475	-13.0574	-8.8141	4.0190	-63.9971	-2.0484	3.7501	202.9334	-37.7477
Css10	-5.9516	-32.0531	9.7352	-9.6351	-22.2563	-43.3934	-28.7162	-27.7281	-42.4830	193.3325

Tableau IV.4 : Scores entre les profils et les séquences consensus

IV.9 Conclusion

Notre travail a consisté à élaborer des profils de famille de protéines en utilisant les MMC. Ces derniers représentent en fait les alignements des séquences protéiques en question. Dix modèles ont été ainsi construis pour 10 familles de protéines puisées dans la base de données en ligne PFAM. Celle-ci a constitué aussi une référence pour l'évaluation et la validation de nos résultats, ce qui nous conduit à conclure que les résultats obtenus sont encourageants.

Toutefois, il faut élargir notre base de données pour une meilleure appréciation.

Conclusion Générale

Dans ce travail, on a réalisé un programme qui a consisté à élaborer des profils de famille de protéines en utilisant les MMC afin de détecter les zones qui ont une forte probabilité de contenir une séquence protéique .

La réalisation de ce programme nous a permis de nous familiariser avec le monde biologique, le monde de l'analyse mathématique et la programmation. Ce projet nous a permis aussi de voir la relation entre les différents mondes (la biologie, la probabilité et la programmation). Nous avons utilisé l'environnement Matlab pour la programmation. Les résultats obtenus même s'ils doivent être améliorés restent très encourageants.

Nous pouvons enrichir cette application dans le futur, d'une part par l'élargissement de la base de données de test et d'autre part en utilisant des séquences biologiques autres que celles issues de la base de données PFAM.

A la fin, espérons donc que notre projet va ouvrir des portes pour les autres étudiants pour travailler sur des sujets dans ce domaine.

Bibliographie

[1] GUYLAINE POISSON : « Analyse de Motifs protéiques par méthodes hybride réseaux de neurones artificiels et modèle de Markov caché' thèse de doctorat »2004.

[2] BENYGZER Adel et BRIHI Tarek," Analyse de Séquences biologiques à l'aide de Modèles de cachés Markov", Thèse d'ingénieur en électronique de l'Université de Blida, Algérie, p 07-08. 2011.

[3] image,http://www.google.dz/url?sa=i&rct&j&q=&esrc=s&source=images&cd=&cad=rja&uact=&&ved=0CAcQjRw&url=https%3A%2F%2Ffr.wikipedia.org%2Fcid_d%25C3%25A9soxiribonucl%25C3%25A9ique&ei=yLCXVfSNNOW17gadpqOIAw&bvm=bv.96952980,d.ZGU&psig=AFQjCNGQD4IfbiqGtptuHEAT3WJN7UE9Pw&ust=1436090864769253.

[4] Nicolas Terrapon : « Recherche de domaines protéiques divergents à l'aide de modèles de Markov cachés »,2013.

cachés : application `a Plasmodium falciparum. Bioinformatiques. Université Montpellier II - Sciences et Techniques du Languedoc, 2010. French. <tel-00811835\|

[5] MELBOUS Mohamed.les banque de données en ligne accès et problématique (cas des données biologiques).2012 /2013.

[6] PFAM (protéine Family database) (http ://www.sanger.ac.uk/Pfam/)

[7] Abdelhak Mansoul, thèse, "Fouille de donnée biologiques : Etude comparative et expérimentation", 2009/2010.

[8] bioinformatique et données biologiques .cours d'introduction à la bioinformatique et de présentation des banques de séquences. équipe Bonsai (2014).

[9] Jean-Baka Domelevo-Entfellner1, 2 et Olivier Gascuel1 : « Une approche phylo-HMM pour la recherche de séquences »,2008

[10] Guillaume Chakroun : « Recherche d'homologies » ,2004.

[11] E.P.C.Rocha «Analyse exploratoire des génomes bactériens» Thèse de doctorat », Université de Versailles,2000.

[12] Guillaume Chakroun : «prédiction de la structure d'une protéine » ,2004

[13] Nadira Benlahrache : « Optimisation Multi-Objectif pour l'Alignement Multiple de Séquences »mémoire du magister ,2007

[14] Laure Vescovo : «Outils et méthodes pour la classification pyramidale de données biologiques »

[15] Vincent Barra : « Apprentissage de modèles de Markov cachées »,2005/2006.

[16] S.Batzoglou « SequenceAlignment'I:CS262Winter2004:LectureII,2004.

[17] Jacques Van Helden : «Bioinformatique Appliquée définition et concept» TAGC, Aix-Marseille Université,2012

[18] Sandrine PERRIN : « Calcul de score d'alignements multiples de séquences»,2010.

[19] Vincent Picard « Pondération d'automates modélisant des familles de protéines et significativité des scores»

[20] Anne-Muriel Arigon Chifolleau : « GMIN220 – Alignement », 2014-2015.

[21] Laurent Noé, Martin Figeac : « Introduction à la comparaison de séquences »

Annexe

Field	Description
`ModelLength`	Integer specifying the length of the profile (number of MATCH states).
`Alphabet`	String specifying the alphabet used in the model. Choices are `'AA'` (default) or `'NT'`. **Note** `AlphaLength` is 20 for `'AA'` and 4 for `'NT'`.
`MatchEmission`	Symbol emission probabilities in the MATCH states. Either of the following: • A matrix of size `ModelLength`-by-`AlphaLength`, where each row corresponds to the emission distribution for a specific MATCH state. Defaults to uniform distributions. • A structure containing residue counts, such as returned by `aacount` or `basecount`.
`InsertEmission`	Symbol emission probabilities in the INSERT state. Either of the following: • A matrix of size `ModelLength`-by-`AlphaLength`, where each row corresponds to the emission distribution for a specific INSERT state. Defaults to uniform distributions. • A structure containing residue counts, such as returned by `aacount` or `basecount`.
`NullEmission`	Symbol emission probabilities in the MATCH and INSERT states for the NULL model. Either of the following: • A 1-by-`AlphaLength` row vector. Defaults to a uniform distribution. • A structure containing residue counts, such as returned by `aacount` or `basecount`. **Note** The NULL model is used to compute the log-odds ratio at every state and avoid overflow when propagating the probabilities through the model.

Field	Description
	Note NULL probabilities are also known as the background probabilities.

`BeginX`	BEGIN state transition probabilities.
	
	Format is a 1-by-`(ModelLength + 1)` row vector:
	
	`[B->D1 B->M1 B->M2 B->M3 .... B->Mend]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from the BEGIN state equals 1:
	
	`sum(`*`Model`*`.BeginX) = 1`
	
	For fragment profiles:
	
	`sum(`*`Model`*`.BeginX(3:end)) = 0`
	
	Default is `[0.01 0.99 0 0 ... 0]`.

`MatchX`	MATCH state transition probabilities.
	
	Format is a 4-by-`(ModelLength - 1)` matrix:
	
	`[M1->M2 M2->M3 ... M[end-1]->Mend;`
	` M1->I1 M2->I2 ... M[end-1]->I[end-1];`
	` M1->D2 M2->D3 ... M[end-1]->Dend;`
	` M1->E  M2->E  ... M[end-1]->E  ]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from every MATCH state equals 1:
	
	`sum(`*`Model`*`.MatchX) = [ 1 1 ... 1 ]`
	
	For fragment profiles:
	
	`sum(`*`Model`*`.MatchX(4,:)) = 0`
	
	Default is `repmat([0.998 0.001 0.001 0],ModelLength-1,1)`.

`InsertX`	INSERT state transition probabilities. Format is a 2-by-`(ModelLength - 1)` matrix: `[ I1->M2 I2->M3 ... I[end-1]->Mend;`

Field	Description
	I1->I1 I2->I2 ... I[end-1]->I[end-1]] **Note** If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from every INSERT state equals 1: `sum(`*Model*`.InsertX) = [ 1 1 ... 1 ]` Default is `repmat([0.5 0.5],ModelLength-1,1)`.

`DeleteX`	DELETE state transition probabilities.
	
	Format is a 2-by-`(ModelLength - 1)` matrix:
	
	`[ D1->M2 D2->M3 ... D[end-1]->Mend ;`
	`  D1->D2 D2->D3 ... D[end-1]->Dend ]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from every DELETE state equals 1:
	
	`sum(`*Model*`.DeleteX) = [ 1 1 ... 1 ]`
	
	Default is `repmat([0.5 0.5],ModelLength-1,1)`.

`FlankingInsertX`	Flanking insert states (N and C) used for LOCAL profile alignment.
	
	Format is a 2-by-2 matrix:
	
	`[N->B  C->T ;`
	`N->N  C->C]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from Flanking Insert states equals 1:
	
	`sum(`*Model*`.FlankingInsertsX) = [1 1]`
	
	Note To force global alignment use:
	
	Model`.FlankingInsertsX = [1 1; 0 0]`
	
	Default is `[0.01 0.01; 0.99 0.99]`.

`LoopX`	Loop states transition probabilities used for multiple hits alignment. Format is a 2-by-2 matrix:

Field	Description

	`[E->C  J->B ;`
	` E->J  J->J]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from Loop states equals 1:
	
	`sum(`*`Model`*`.LoopX) = [1 1]`
	
	Default is `[0.5 0.01; 0.5 0.99]`.

`NullX`	Null transition probabilities used to provide scores with log-odds values also for state transitions.
	
	Format is a 2-by-1 column vector:
	
	`[G->F ; G->G]`
	
	Note If necessary, `hmmprofstruct` will normalize the data such that the sum of the transition probabilities from Null states equals 1:
	
	`sum(`*`Model`*`.NullX) = 1`
	
	Default is `[0.01; 0.99]`.

`IDNumber`	Optional. User-assigned identification number.
`Description`	Optional. User-assigned description of the model.