Construction d'attributs pour l'extraction de connaissances à partir de séquences biologiques

M. Maddouri[*] et F. Mhamdi [**]
Unité de Recherche en Programmation, Algorithmique et Heuristique

[*] Institut National des Sciences Appliquées et de Technologie,
Université 7 Novembre à Carthage, Centre Urbain Nord,
BP. 676, 1080 Tunis, Tunisie
mondher.maddouri@fst.rnu.tn

[**] Institut Supérieur des Langues Appliquées et d'Informatique de Béja,
Université de Jendouba, Av. Habib Bourguiba, 9000, Béja, Tunisie
faouzi.mhamdi@ensi.rnu.tn

Résumé. Dans cet article nous étudions un problème de prétraitement de données : la construction d'attributs décrivant des séquences biologiques. Afin d'assurer l'extraction de connaissances à partir de séquences biologiques (ADN, ARN et protéines), tout système de fouille de données (datamining) se confronte à la représentation non habituelle de ce type de données. Une séquence biologique est représentée, en structure primaire, par une chaîne de caractères. La construction d'attributs décrivant les séquences biologiques est une étape de prétraitement inévitable. Dans cet article, nous étudions les méthodes existantes de construction d'attributs décrivant des séquences biologiques, notamment, celles qui se basent sur les n-grammes, l'arbre de suffixes généralisés et les modèles de Markov cachés. Notre contribution dans cet axe a été la proposition de la méthode des descripteurs discriminants et la présentation d'une étude comparative approfondie de ces méthodes en les appliquant à des problèmes biologiques typiques comme la reconnaissance de sites promoteurs des gènes de *E. Coli*, la reconnaissance de sites de jonction de *Primate* et la classification des protéines. Une confrontation des résultats de chaque méthode avec la banque de motifs Pfam sera aussi présentée.

1 Introduction

La plupart des méthodes de datamining, traitent des données représentées sous forme d'une table relationnelle (tableau attributs/valeurs). Il existe toutefois quelques travaux qui portent sur une représentation de données plus complexes [Cornuéjols *et al.* 2002, Zighed *et al.* 2000]. Une problématique supplémentaire s'ajoute lorsqu'on veut utiliser ces approches pour analyser des données ayant des représentations atypiques : séquences, hiérarchies, image, son, vidéo, etc [Mitra *et al.* 2003]. La *construction d'attributs*, qui consiste à inventer des attributs pour décrire ces données en format relationnelle (attributs/valeurs), permet d'apporter une réponse au problème de données atypiques [Liu *et al.* 1998, Liu *et al.* 2001]. Elle peut être vue comme l'une des tâches de prétraitement dans le procédé de l'Extraction de connaissances à partir de données [Fayyad *et al.* 1996].

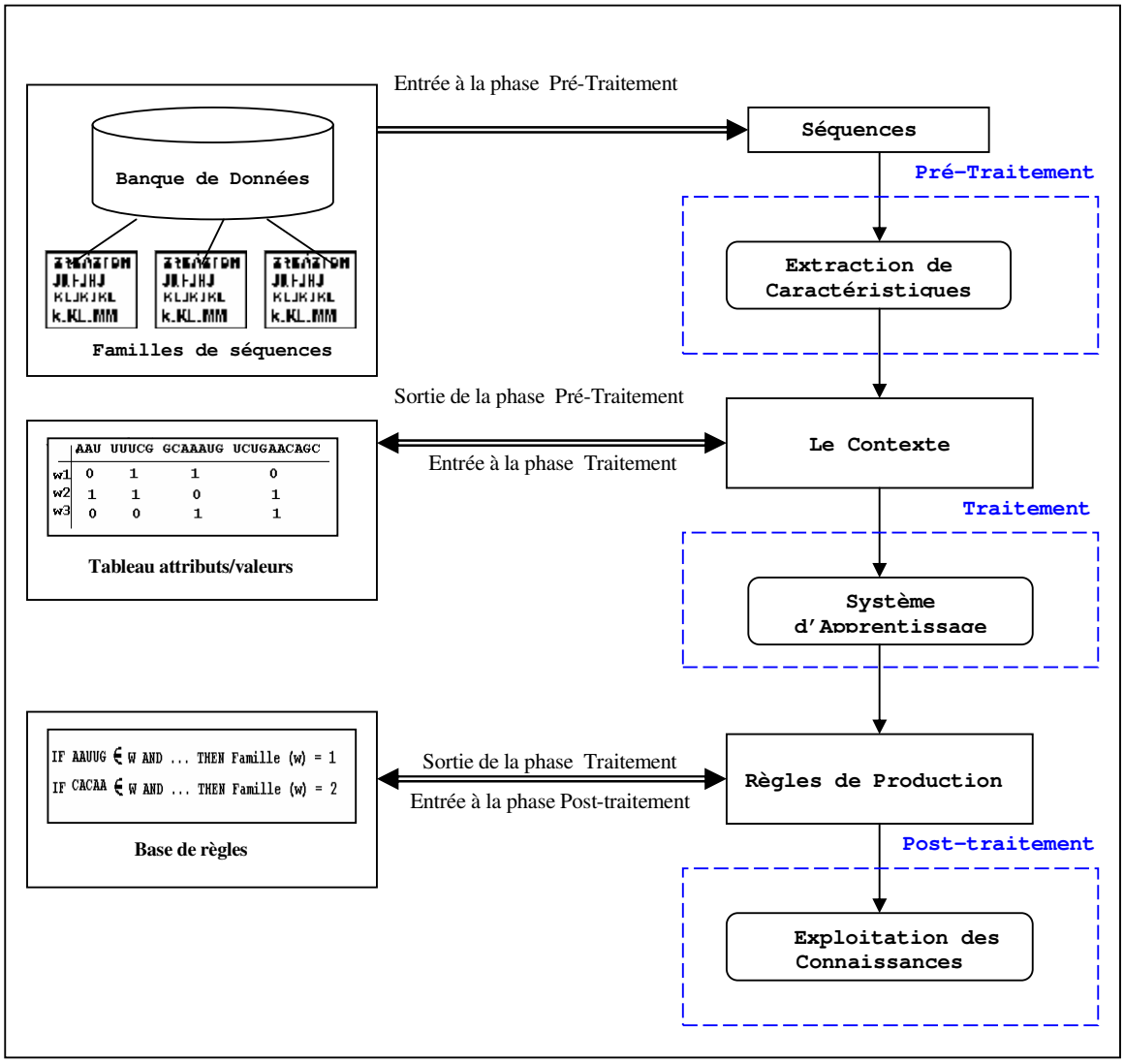

FIG. 1 - Principales étapes de fouille de séquences biologiques

La fouille de séquences biologiques se situe dans ce cadre et représente l'un des axes de recherche récents et prometteurs en fouille de données [Mitra *et al.* 2003, Delaplace 2003, Bes 2002]. En effet, en Biologie Moléculaire, les scientifiques se sont penchés durant la dernière décennie, sur la collection des séquences biologiques (ADN, ARN, protéines), leur stockage, et leur utilisation pour expérimenter leurs théories. Notant que ces données biologiques ont plusieurs formes de représentations. On peut les représenter par des structures primaires,

secondaires ou tertiaires. De nos jours, des centaines de banques de données hétérogènes et volumineuses, sont disponibles dans les centres de recherche et accessibles via Internet [Perrière 2000].

L'analyse de ces données par les experts biologistes devient de plus en plus difficile et coûteuse. Le recours aux systèmes de fouille de données devient de plus en plus nécessaire. En effet, la combinaison de méthodes spécifiques, permettra de déchiffrer les génomes par un ordinateur, et donc de prédire de nouvelles connaissances dans ce domaine [Yu *et al.* 2003].

En réalité, la plupart de ces systèmes n'analyse que des données relationnelles (attributs/valeurs). Alors que les structures primaires des données biologiques sont sous forme de séquences représentées par des chaînes de caractères appartenant à un alphabet particulier. Les attributs nécessaires à la description de ces données ne sont pas définis au préalable. Ils doivent, donc, être construits/inventés à partir des séquences brutes. D'où l'utilité du développement d'un système de construction d'attributs qui sera exécuté dans la phase de pré-traitement des données.

La figure 1 illustre les trois principales étapes du processus de fouille de séquences biologiques : le pré-traitement, la fouille des données et le post-traitement. Dans l'étape de pré-traitement, un système de construction d'attributs est requis. Les attributs construits servent à produire une représentation relationnelle des séquences biologiques (tableau attributs/valeurs). Dans l'étape principale du traitement, un système d'apprentissage symbolique est utilisé pour découvrir des règles de productions. Dans l'étape de post-traitement, les règles extraites sont exploitées pour des fins de classification, de regroupement, d'association ou de prédiction.

[Dardel *et al.* 2002] ont montré que ce problème de construction d'attributs a une complexité combinatoire. Plusieurs approches heuristiques de construction d'attributs ont été proposées. Dans un travail antérieur [Maddouri *et al.* 2004], nous avons étudié les approches qui utilisent les propriétés bio-chimiques comme attributs fixés par les biologistes [Fu 2001]. Dans cet article, nous nous intéressons aux approches qui extraient des motifs (sous-séquences) significatifs biologiquement. La présence/absence dans la séquence de chaque motif constitue un attribut binaire. Certes, la méthode de construction des n-grammes était la première à attirer l'attention des chercheurs pour sa simplicité [Dumas *et al.* 1982]. La méthode basée sur les modèles de Markov cachés est aussi largement utilisée par les bio-informaticiens [Krogh *et al.* 1994]. [Wang *et al.* 2001] ont montré l'intérêt de la méthode basée sur les arbres de suffixes généralisés.

Dans la deuxième section, nous présentons ces trois méthodes de construction de motifs décrivant les séquences biologiques, ainsi que nous détaillons notre méthode de construction de descripteurs discriminants. Nous présentons aussi, une étude de complexité algorithmique de ces différentes méthodes. La section 3 est dédiée à l'étude expérimentale comparative de ces méthodes. Pour ce faire, nous avons choisi de comparer le pouvoir classificatoire de ces attributs par le calcul des taux d'erreurs avec la méthode validation croisée [Kohavi 1995] de C4.5 [Quinlan 1993]. Nous avons aussi effectué une étude comparative sur la signification biologique des motifs construits, en les confrontant aux motifs reconnus par des études d'experts biologistes, notamment la banque de motifs Pfam [Bairoch *et al.* 1994].

Le cadre applicatif de cette étude expérimentale a porté sur des problèmes biologiques typiques comme la reconnaissance de sites promoteurs des gènes de *E. Coli*, la reconnaissance de sites de jonction de *Primate* et la classification des protéines *Toll-Like Receptor* et des protéines *Arginine*.

2 Construction d'attributs décrivant des séquences biologiques

Les structures primaires ou séquences des macromolécules d'ADN, d'ARN et de protéines sont des chaînes dont les caractères appartiennent, respectivement, aux alphabets $\mathcal{A}_{ADN}$={*A, C, G , T*}, $\mathcal{A}_{ARN}$={*A, C, G, U*} et $\mathcal{A}_{PRN}$={*A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y*}.

Soit $\mathcal{A}$ un alphabet fini. Une *chaîne de caractères* est la concaténation d'éléments de $\mathcal{A}$. La longueur d'une chaîne *w*, notée |*w*|, est le nombre de caractères qui la composent. Le $i^{ème}$ caractère d'une chaîne *w* de longueur *m*, $1 \leq i \leq m$, sera noté *w*[*i*]. Une portion de *w* qui commence à la position *i* et se termine à la position *j*, $1 \leq i \leq j \leq m$, est appelée *sous-chaîne* de *w* et sera notée *w*[*i, j*]. La chaîne *w* peut être notée par *w*[1, *m*].

Soient w_1, w_2, …, w_N, *N* chaînes appartenant au même alphabet et *S* l'ensemble de ces chaînes. Le problème de construction d'attributs décrivant ces chaînes, consiste à analyser les chaînes pour déterminer un ensemble de descripteurs $\Delta=\{\delta_1, \delta_2, \ldots, \delta_K\}$. Soit *R* une *relation binaire* définie entre *S* et Δ telle que $(w, \delta) \in R$, si et seulement si l'attribut δ décrit la chaîne *w*. Le triplet (S, Δ, R) forme une représentation relationnelle (tableau attributs/valeurs) de l'ensemble des chaînes à analyser.

En bioinformatique, deux types d'attributs ont été considérés. L'approche la plus ancienne utilise des propriétés biochimiques prédéfinies par les experts biologistes [Dickerson *et al.* 1969]. Il ne s'agit pas d'une approche automatique de construction d'attributs. De nouvelles méthodes relevant de cette approche ont été proposées par De La Maza et Gracy [Fu *et al.* 2004]. L'approche récente consiste à déterminer des attributs indiquant la présence (ou la fréquence d'apparition) de sous-chaînes particulières dans l'ensemble des chaînes analysées.

Récemment, F. Dardel [Dardel *et al.* 2002] a évalué la complexité algorithmique exacte pour extraire des sous-chaînes communes à *N* chaînes, dont chacune est de longueur inférieure ou égale à *m*, comme :

$$(N/2)*(N-1)*m^N*(2^N-1) \qquad (1)$$

Si, pour aligner 2 séquences protéiques de 100 acides aminés il faut une seconde de temps machine, il faudra 3 jours pour en aligner 4. A partir de 9 séquences, le temps de calcul dépasse l'âge de l'univers [Dardel *et al.* 2002]. Il est donc impensable d'employer une approche exacte. Nous devrons avoir recours à des approches heuristiques.

Dans cet article, nous étudions la méthode de construction des n-grammes [Dumas *et al.* 1982], la méthode basée sur les arbres de suffixes généralisés [Wang *et al.* 1994, Molla *et al.* 2004], la méthode basée sur les modèles de Markov cachés [Krogh *et al.* 1994] et La méthode heuristique que nous avons proposé, celle des descripteurs discriminants [Maddouri *et al.* 2002].

2.1 Méthode de construction des n-grammes

La méthode la plus utilisée, est celle des *n-grammes*, dite aussi *n-uplets* ou fenêtrage de longueur *n* [Dumas *et al.* 1982]. Elle adopte une description par motif d'alignement simple. Les sous-séquences envisageables sont de longueur fixée au préalable. Un n-gramme et une sous-séquence de *n* caractères. Pour une séquence quelconque, l'ensemble des n-grammes pouvant être générés est obtenu en déplaçant une fenêtre de *n* caractères sur la séquence entière. Ce

déplacement s'effectue caractère par caractère. A chaque déplacement la sous-séquence des *n* caractères est extraite. Cette méthode est très utilisée dans la fouille de données textuelles [Sebastiani 2005, Radwan *et al.* 2003]. La figure 2 illustre ce principe. L'ensemble de ces sous-séquences constitue les n-grammes pouvant être générés à partir d'une seule séquence [Dardel *et al.* 2002]. Ce processus sera itéré pour toutes les séquences analysées.

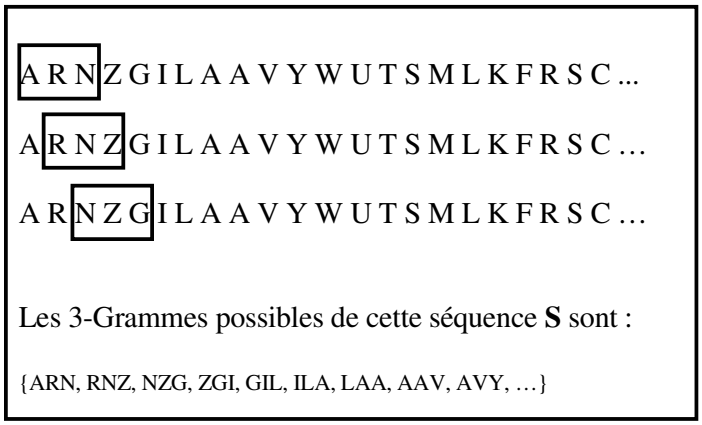

FIG. 2 – Illustration de l'extraction des n-grammes

Dans les travaux antérieurs sur l'extraction de motifs à partir de textes, plusieurs algorithmes d'extraction de n-grammes ont été développés [Dumas *et al.* 1982]. Leur complexité algorithmique est égale à $n*m*N$, avec n la taille d'un n-gramme, m la taille d'une séquence et N le nombre de séquences analysées. Par ailleurs, comme il existe 20 acides aminés dans l'alphabet des protéines[1], le nombre de n-grammes possibles pour les séquences protéiques est égal à 20^n. Ainsi, le handicap majeur de cette méthode est ce nombre d'attributs très élevé. En pratique, l'utilisation de cette approche se limite à l'analyse de deux séquences, en comparant leurs sous-séquences.

Le logiciel BLAST [Altschul *et al.* 1990] adopte une méthode d'alignement local pour repérer les fragments similaires de longueur fixe (valeurs par défaut[1]: 3 pour les séquences protéiques et 11 pour les séquences génomiques), dits aussi HSP (High scoring Segment Pair). Le logiciel FASTA [Pearson *et al.* 1988] se base aussi sur des fragments de longueur fixe et construit une matrice numérique de similarité entre les deux séquences en question. Généralement, ces logiciels sont utilisés pour la comparaison d'une séquence nouvellement identifiée avec une séquence figurant dans une banque de données.

2.2 Méthode basée sur les arbres de suffixes généralisés

La notion d'arbre de suffixes généralisés (General Suffix Tree : GST) est utilisée pour découvrir tous les sous-séquences possibles dans une séquence donnée. Le système SDiscover

[1] Nous rappelons qu'il y a 4 caractères dans l'alphabet codant des séquences génomiques {A, C, G, T}, et qu'il y a 20 caractères dans l'alphabet codant des séquences protéiques (chacun constitué par la succession de 3 caractères parmi A, C, G, et T).

de [Wang *et al.* 1999] utilise cette méthodes pour découvrir des motifs actifs dans un ensemble de séquences biologiques. Cette approche se réalise sur deux phases. La première consiste à construire des motifs simples de type*__X*__ à partir de l'échantillon de séquences. La deuxième consiste à combiner ces motifs simples pour former des motifs plus complexes de type *X*Y* (motifs qui présentent des gaps). A la fin du traitement, on évalue l'activité de ces motifs dans toutes les séquences pour déterminer ceux qui vérifient certaines conditions spécifiées. Ces conditions sont exprimées à travers trois variables : *la taille minimale du motif*, *le nombre minimal d'occurrences* et *le nombre de mutations.*

La construction des motifs se base sur la construction de l'arbre de suffixes généralisés. Puisqu'un motif représente un suffixe. Soit *CC* une chaîne de longueur *m*, un suffixe *S* est la sous-chaîne de *CC* qui commence à la position *i* et se termine à la fin de *CC*, soit *w*[*i*, *m]*.

Supposant qu'on a N séquences, pour construire l'arbre qui correspond à cet ensemble de séquences, on les concatène en une seule chaîne de caractères *CC*. Puis, on construit l'arbre de suffixes généralisés à partir de la chaîne *CC*. On construit le nœud racine de l'arbre. La racine a autant de fils qu'il y a de suffixes dans la chaîne *CC*. Chaque nœud fils possède au minimum deux fils. Un arc entre deux nœuds correspond à un caractère de la chaîne *CC*. Les notations des arcs sortant d'un même nœud doivent correspondre à des caractères différents. Chaque feuille *i* de l'arbre correspond à un suffixe de *CC* qui commence à la position *i*. C'est la concaténation des notations des arcs, dans le chemin allant de la racine au nœud *i*.

Pour illustrer cet approche sur un exemple, [Wang *et al.* 1994] considère les trois séquences suivantes : w1=FFRR, w2=MRRM et w3=MTRM. La concaténation de ces trois séquences donne la chaîne *CC*=FFRRMRRMMTRM. La figure 3 présente l'arbre de suffixes généralisés correspondant à cette chaîne. Chaque feuille est indexée par le numéro de séquence qui contient son suffixe. Chaque nœud non feuille contient le nombre de différents indexes associés avec les feuilles du sous-arbre.

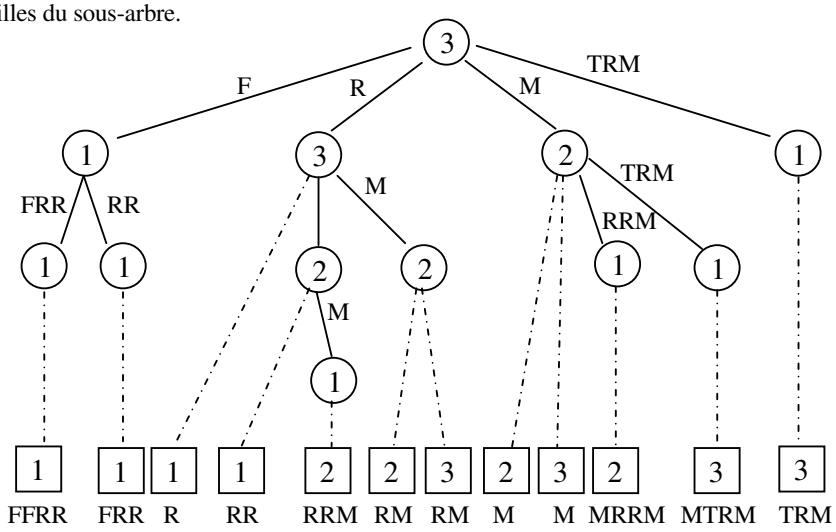

FIG. 3 – Exemple de construction d'un arbre de suffixes généralisés

Dans la littérature, il existe plusieurs algorithmes de construction des arbres de suffixes généralisés. La complexité de construction d'un arbre GST est de O(*m***N*), avec *m* la taille

d'une séquence et *N* le nombre de séquences concaténées [Wang *et al.* 1994]. Pour extraire tous les motifs qui respectent les conditions de la recherche, il faut parcourir tout l'arbre en comparant un motif en question avec l'ensemble des séquences. Soit une complexité de $O((m*N)^2)$.

2. 3 Méthode basée sur les Modèles de Markov Cachés

Un modèle de Markov est un modèle de séquences d'évènements où les probabilités associées à un évènement dépendent de l'évènement présent et non du passé [Krogh *et al.* 1994]. Il débute par un état initial. A chaque étape, le système se déplace de l'état présent vers l'état suivant selon une probabilité de transitions associée aux états. Le modèle devient caché si on se trouve devant un scénario où les états ne peuvent être directement observés. Donc, chaque état a une probabilité d'émission servant à déterminer l'évènement observable.

Le modèle de Markov caché (Hidden Markov Model : HMM) a été largement utilisé pour l'analyse des séquences biologiques [Dardel *et al.* 2002]. Pour construire un modèle de Markov pour des séquences biologiques, il faut passer par un alignement multiple de ces séquences. La figure 4 présente un alignement multiple de 5 séquences :

A	-	-	T	G	T
A	-	-	G	A	T
A	C	T	G	G	T
A	-	-	G	-	T
A	-	-	C	G	T

FIG. 4 – Exemple d'alignement multiple de plusieurs séquences

La figure 5 présente le modèle de Markov caché correspondant à cet alignement de 5 séquences. Dans cette figure, l'état M (match) correspond aux positions dans un alignement multiple de séquences (appelé profil). Il représente les différentes fréquences pour chaque caractère dans chaque position. L'état I (insertion) correspond à l'insertion de caractères entre états. Il contient les différentes fréquences pour chaque insertion dans chaque position. Il est possible d'avoir plusieurs insertions dans le même site. L'état D (délitions) permet de sauter plusieurs positions si nécessaire. A titre illustratif, nous considérons l'exemple de la sous-séquence "AGT". Le modèle permet de calculer la probabilité de ce motif dans le modèle, à savoir p=1*1*0.8*0.6*0.2*1*1, ainsi que son score comme $\log_2(p)$.

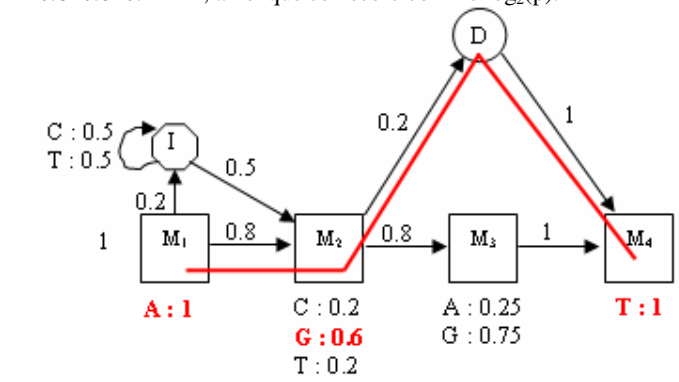

FIG. 5 – Illustration de la modélisation de séquences par HMM

2.4 Méthode de construction de descripteurs discriminants

Cette méthode à été introduite dans [Maddouri *et al.* 2002]. Cette méthode consiste à construire un ensemble de sous-chaînes discriminantes et minimales tolérant une certaine ambiguïté reflétant celle des données. Les chaînes de caractères analysées peuvent être regroupées au préalable en P familles/classes $f_1, f_2, \ldots, f_P$, de chaînes du même alphabet. Il s'agit de construire des sous-chaînes qui, d'une part, ne contiennent aucune autre sous-chaîne discriminante. Elles sont donc *minimales*. D'autre part, elles apparaissent dans la majorité des chaînes de la famille en question f_i, pour $1 \le i \le P$, mais n'apparaissent que dans une minorité des chaînes des autres familles $f_1, \ldots, f_j, \ldots, f_P$, avec $j \ne i$. Donc, elles sont *discriminantes*. Nous appelons cet ensemble de sous-chaînes : les *Descripteurs Discriminants* (*DD*). Cet ensemble de DD constitue l'ensemble des attributs construits $\Delta = \{\delta_1, \delta_2, \ldots, \delta_K\}$.

Nous notons par $\Delta_i(f_1, f_2, \ldots, f_P)$, $1 \le i \le P$, l'ensemble des descripteurs discriminants de la famille f_i, $1 \le i \le P$. La notation $(f_1, f_2, \ldots, f_P)$ s'explique par le fait suivant: nous pouvons utiliser toutes les familles $f_1, f_2, \ldots, f_P$, pour construire un descripteur de la famille f_i, $1 \le i \le P$. Ainsi, nous avons :

$$\Delta = \Delta_1(f_1, f_2, \ldots, f_P) \cup \ldots \cup \Delta_P(f_1, f_2, \ldots, f_P) \qquad (2)$$

Le *k-Descripteur Discriminant* (*k*-DD, noté $k\_\Delta_i(f_1, f_2, \ldots, f_P)$, est le sous-ensemble de $\Delta_i(f_1, f_2, \ldots, f_P)$ constitué par les sous-chaînes minimales et discriminantes de longueur k. Comme m est la longueur maximale d'une séquence, nous avons alors :

$$\Delta_i(f_1, f_2, \ldots, f_P) = 1\_\Delta_i(f_1, f_2, \ldots, f_P) \cup \ldots \cup m\_\Delta_i(f_1, f_2, \ldots, f_P) \qquad (3)$$

2.4.1 Identification de sous-chaînes répétées

Cette étape permet de trouver toutes les sous-chaînes d'une longueur donnée dans une chaîne de caractères. Soit w une chaîne de longueur m et soient i, j et k trois entiers tels que $1 \le k \le m$ et $1 \le i \le j \le m-k+1$. Nous disons que les positions i et j appartiennent à la même *classe d'équivalence* pour la relation d'équivalence E_k, si et seulement si, nous avons $w[i, i+k-1] = w[j, j+k-1]$. Nous disons aussi que les positions i et j sont k-équivalentes. La figure 6 illustre un cas de 7-équivalence entre les positions $i=5$ et $j=26$. Nous remarquons qu'il s'agit d'une sous-chaîne répétée « AGCAAUA » de longueur $k=7$ identifiée dans les positions $i=5$ et $j=26$. Cette sous-chaîne répétée constitue l'une des classes d'équivalences de la relation E_7.

Une relation d'équivalence E_k, $1 \le k \le m$, peut être représentée par un vecteur $V_k[1 .. m-k+1]$. Chaque composante $V_k[i]$, $1 \le i \le m-k+1$, de ce vecteur représente le numéro de la classe d'équivalence à laquelle appartient la position i pour la relation d'équivalence E_k . L'algorithme est basé sur le théorème suivant [Karp *et al.* 1972].

FIG. 6 : Illustration de k-équivalence entre les positions i=5 et j=26

Théorème 1: Soit *w* une chaîne de longueur *m* et soient *i, j, k,* et *l* quatre entiers tels que : $l \le k$, $1 \le k+l \le m$, $1 \le i < j \le m-k-l+1$. On a alors :

$$i\ E_{k+l}\ j \quad \text{si et seulement si} \quad i\ E_k\ j \ \text{ et } \ (i+l)\ E_k\ (j+l) \qquad (4)$$

La méthode proposée est composée de deux étapes. La première étape d'*Initialisation* consiste à générer le vecteur V_1. Ce dernier représente la relation d'équivalence E_1 dont les classes d'équivalences sont les sous-chaînes de longueur un. A travers cette étape, nous parcourons la chaîne étudiée caractère par caractère. Si le caractère est rencontré pour la première fois, nous créons une nouvelle classe d'équivalence, nous insérons son numéro dans l'indice courant de V_1, et nous marquons le caractère comme étant rencontré. Si le caractère est déjà rencontré, nous insérons dans la case courante de V_1 le numéro de la classe d'équivalence à laquelle appartient ce caractère.

La deuxième étape de *Déduction* consiste à générer des relations d'équivalences supérieures en utilisant le théorème précédemment énoncé. Au premier lieu, nous regroupons les positions correspondantes aux débuts des classes d'équivalences dans la chaîne étudiée. Puis, nous vérifions pour chaque élément si le théorème est vérifié.

Pour déterminer les sous-chaînes d'une longueur donnée *n* dans une chaîne, nous effectuons l'étape d'*initialisation*, puis nous répétons l'étape de *déduction K* fois tel que $n=2^K$ ou $n=2^K-1$.

2.4.2 Discrimination et minimalité

Dans le cas où les chaînes appartiennent à des familles distinctes $f_1, f_2, \ldots, f_P$, la construction des attributs décrivant ces chaînes $\Delta(f_1, f_2, \ldots, f_P)$, se fait en construisant d'une manière simultanée les P descripteurs associés, respectivement, aux P familles $f_1, f_2, \ldots, f_P$ (équation 1.2). La construction d'un descripteur discriminant $\Delta_i(f_1, f_2, \ldots, f_P)$ entre une famille f_i, $1 \le i \le P$, et les autres familles *f1,...,fi-1,fi+1,...,fp*, résulte d'une adaptation de la méthode précédente.

Nous concaténons les chaînes de $f_1, f_2, \ldots, f_P$ en une seule chaîne ζ. Nous effectuons l'étape d'initialisation, pour déterminer les sous-chaînes de longueur un $\Delta_1(f_1, f_2, \ldots, f_P)$ (les classes d'équivalence de E_1). Puis, nous répétons l'étape de déduction plusieurs fois. A chaque fois, l'étape de déduction construit des sous-chaînes répétées (les classes d'équivalence de E_{k+1}). Ces sous-chaînes sont plus longues que celles de l'étape précédente ($\Delta_k(f_1, f_2, \ldots, f_P)$: les classes d'équivalence de E_k). Ceci, en appliquant le théorème 1. Ensuite, nous filtrons ces nouvelles sous-chaînes répétées (les classes d'équivalence de E_{k+1}) de façon à ne garder que les sous-chaînes *discriminantes* et *minimales* : $\Delta_{k+1}(f_1, f_2, \ldots, f_P)$.

Une sous-chaîne est dite *discriminante*, si elle apparaît fréquemment dans la $i^{\text{ème}}$ portion de ζ (dans des chaînes de f_i), et n'apparaît pas ailleurs (dans les chaînes de *f1,...,fi-1,fi+1,...,fp*). Cet algorithme utilise deux paramètres α et β qui définissent le degré d'ambiguïté de la discrimination entre exemples et contre-exemples. Le paramètre α représente le pourcentage des exemples (les chaînes de f_i contenant la sous-chaîne en question). Le paramètre β représente le pourcentage des contre-exemples (les chaînes n'appartenant pas à f_i et qui contiennent la sous-chaîne en question).

Définition 1 : sous-chaîne discriminante

Une sous-chaîne ambiguë *x* est considérée comme étant *discriminante* entre la famille f_i, $1 \le i \le P$, et les autres familles *f1,...,fi-1,fi+1,...,fp*, si et seulement si :

$$\begin{cases} \dfrac{\text{nombre des chaînes de } f_i \text{ où } x \text{ apparaît}}{\text{nombre total des chaînes de } f_i} * 100 \ge \alpha \\ \dfrac{\text{nombre des chaînes de } \bigcup_{j \ne i} f_j \text{ où } x \text{ apparaît}}{\text{nombre total des chaînes de } \bigcup_{j \ne i} f_j} * 100 \le \beta \end{cases} \quad (5)$$

Définition 2 : sous-chaîne minimale

Une sous-chaîne est dite *minimale* si elle ne contient pas d'autres sous-chaînes discriminantes.

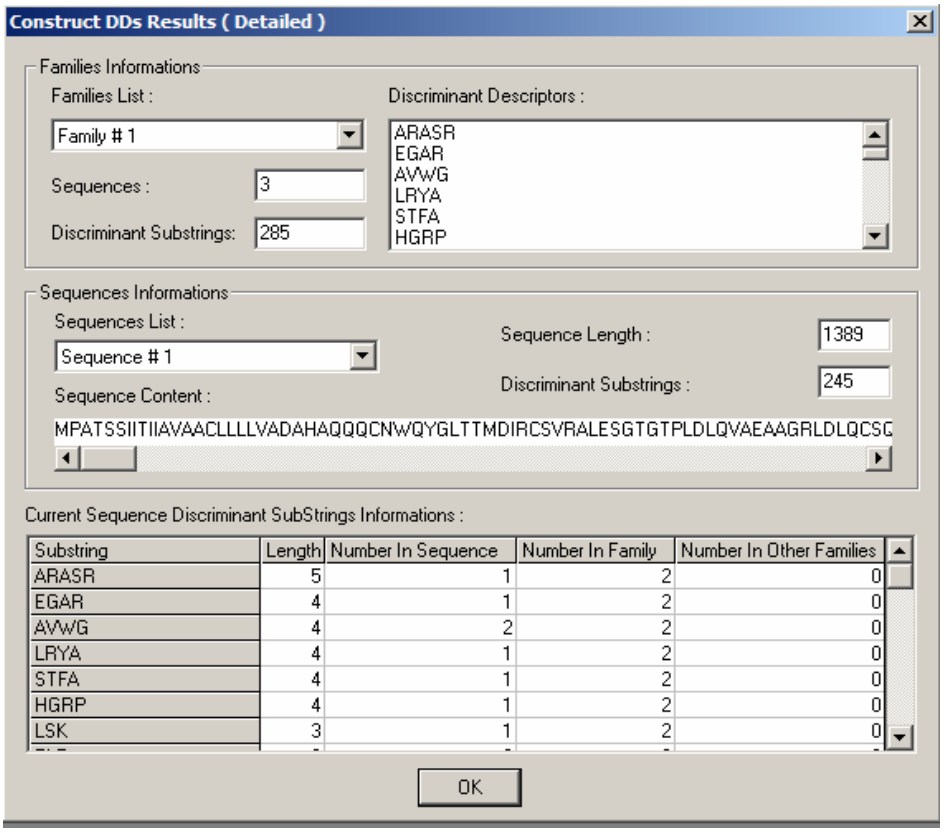

FIG. 7 : Exemple d'extraction de descripteurs discriminants décrivant la protéine "TLR-1 Humain"

Pour illustrer la notion de minimalité, nous considérons l'exemple suivant. Soient ch_1 et ch_2 deux sous-chaînes, tel que ch_1 = "A**DCG**T" et ch_2 = "**DCG**". L'algorithme ne peut pas considérer ch_1 et ch_2 comme deux descripteurs différents d'une même famille de chaînes. Car, ch_2 est inclus dans ch_1. D'où, ch_1 n'est pas minimale. La figure 7 un deuxième exemple concrêt. Il s'agit des descripteurs discriminants extraits à partir de la séquence protéique *"TLR 1 humaine"*.

La méthode proposée dispose de deux conditions d'arrêts se basant sur des critères heuristiques : la condition de discrimination et la condition de minimalité. La méthode commence par la construction des sous-chaînes de longueur k (les classes d'équivalence de E_k). Puis, elle les combine pour déterminer des sous-chaînes plus longues (les classes d'équivalence de E_{k+1}). Dès qu'elle obtient une sous-chaîne ne vérifiant pas le critère du paramètres α, elle arrête de construire avec des sous-chaînes plus longues, puisqu'elles ne seront plus discriminantes. Dès qu'elle obtient une sous-chaîne discriminante, elle arrête de construire avec des sous-chaînes plus longues, puisqu'elles ne seront plus minimales. Elle ne considère dans l'étape k+1, que les sous-chaînes vérifiant le critère du paramètre α, et ne vérifiant pas le critère du paramètre β.

La complexité de l'algorithme est limitée par la construction des relations d'équivalence E_{k+1} à partir de la relation E_k. Comme ce calcul se fait sur la chaîne totale ζ, la complexité est en $O(m*N)$. La chaîne totale ζ est formée par N chaînes dont chacune est de taille maximale m. Ainsi le problème d'extraction des sous-chaînes de longueur inférieure à n est résolu en $O(m*N*log(n))$. Dans le pire des cas, on a $n=m$. Mais en pratique, n est très inférieur à m. La vérification des paramètres α et β, est de complexité $n*m*N$, ce qui fait une complexité globale de $n*log(n)*m^2*N$, avec N est le nombre de séquences analysées.

2.5 Discussion

Les méthodes adoptant une description par motif émergent dans la recherche sur le génome [Mitra *et al.* 2003]. Un motif représente une ou plusieurs sous-séquences continues ayant une signification biologique particulière. Ce qui peut conduire à une expression régulière caractéristique d'une famille de séquences. Les méthodes de recherche de motifs doivent localiser des sous-séquences satisfaisant certaines contraintes de taille, de maximalité ou de fréquence d'apparition.

Plusieurs méthodes comme celle de construction des n-grammes [Dumas *et al.* 1982] ou celle basée sur les arbres de suffixes généralisés [Yin *et al.* 2004, Wang *et al.* 2000] incluent ces contraintes sous forme de paramètres. Il est difficile de se prononcer sur les valeurs optimales de ces paramètres. En plus, ces méthodes construisent aussi un nombre très grand d'attributs.

D'autres approches adoptent une description par des motifs plus complexes (alignements multiples). Dans ce cadre, les modèles de Markov cachés ont été largement exploités par les logiciels HMMer et SAM [Krogh *et al.* 1994]. Ces logiciels ont permis le développement de la banque de motifs protéiques PROSITE [Bairoch *et al.* 1994].

Vu que les motifs présentent une grande importance dans l'analyse des séquences biologiques et des données textuelles. Les chercheurs continuent leurs travaux pour améliorer les performances des méthodes et techniques existantes. [Keinduangjum *et al.* 2005] proposent une approche statistique pour découvrir les signatures des motifs (ceux les plus significatifs). Ils extraient tous les n-grammes possibles d'un ensemble de séquences d'ADN ensuite ils appliquent six fonctions statistiques sur les n-grammes pour identifier comment un n-gramme

est relié à un ensemble de séquences. Enfin ils sélectionnent les meilleurs n-grammes. Pour les GST, [Huang et al. 2003] proposent une méthode effective de construction des arbres pour extraire des motifs séquentiels. Pour extraire un motif spécifique (TFBSs) à partir des expérimentations biologiques *ChIP-chip* [Li et al. 2005] proposent une nouvelle méthode basée sur les HMM.

Outre que les méthodes adoptant une description par motif, nous connaissons la description par propriétés bio-chimique [Geis *et al.* 1969, Fu 2001, Fu *et al.* 2004a]. Les travaux de [Maddouri *et al.* 2004] ont montré les limites de la description des séquences biologiques par les propriétés bio-chimiques. Le nombre des attributs construits est très élevé surtout dans le cas des séquences protéiques. Le tableau de données résultant est volumineux, creux et rend l'application de systèmes de fouille de données difficile. Les taux d'erreurs obtenues par cette méthode sont assez élevés. Cette approche garde l'avantage de rapidité, puisqu'elle ne construit pas les attributs. Ces derniers sont définis au préalable par des biologistes.

3 Étude expérimentale

Pour comparer les méthodes de constructions d'attributs, nous avons effectué une étude expérimentale. Dans la première section, nous présentons les échantillons de données utilisés dans cette étude ainsi que leurs cadres applicatifs. Par la suite, nous étudions l'effet des attributs construits sur la classification des séquences biologiques correspondantes. A la fin, nous procédons à la comparaison des attributs découverts avec les motifs reconnus par les biologistes, notamment ceux de la banque de motifs Pfam[Bairoch *et al.* 1994] pour les attributs protéiques et ceux de l'étude de Towell [Towell *et al.* 1990, Towell 1991] pour les attributs nucléiques.

3.1 Cadre applicatif

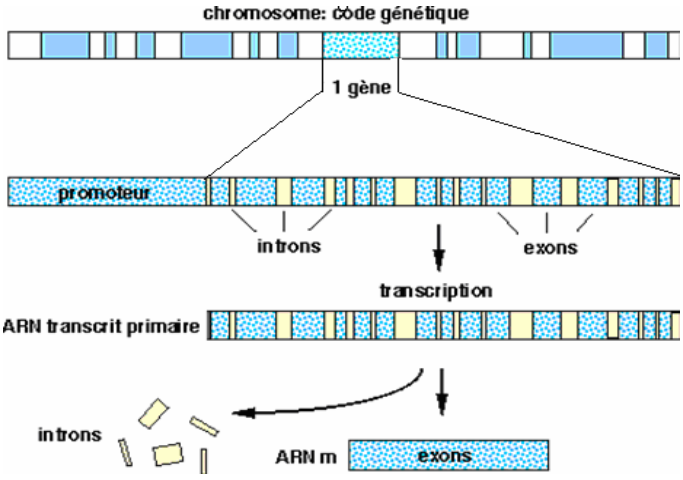

FIG. 8- phase de transcription de l'ADN (ADN→ARN)

Le processus de synthèse de protéines est un processus intracellulaire, Il consiste à transformer une séquence d'ADN en une protéine. Ce processus est composé de deux étapes. La première est celle de transformation d'un ADN en ARNm. Elle s'appelle phase de transcription (Figure 8). La deuxième est celle de transformation d'un ARNm en protéine. Elle s'appelle phase de traduction. En fait, un ADN est une séquence de gènes. Chaque gène est caractérisé par un motif connu sous le nom de site promoteur. Il précède directement le gène. Un gène est une suite de sous-séquences (intron et exon). Les frontières intron/exon et exon/intron sont intitulées site de jonction.

La phase de transcription, nécessite, en premier lieu la détection du site promoteur et en deuxième lieu la détection des sites de jonctions pour bien séparer les introns et les exons. La phase de traduction nous fourni une protéine qui appartient à une famille particulière.

Dans notre processus d'évaluation des méthodes d'extraction de motifs, nous avons abordé par ordre, les différentes étapes du processus de synthèse d'une protéine, à savoir la détection des sites promoteurs, la détection des sites de jonction et la classification de protéines. Dans ce contexte, nous avons étudié quatre problèmes réels d'analyse de séquences biologiques. D'une part, nous avons étudié le problème d'identification de sites promoteurs de E. coli [Harley *et al.* 1987, Towell *et al.* 1990], ainsi que le problème d'identification de sites de jonction de Primate [Towell 1991]. D'autre part, nous avons étudié le problème de classification des protéiques TLR [Chuang *et al.* 2001], ainsi que le problème de classification des protéines Arginine [Sandmeier *et al.* 1994].

3.1.1 Problème d'identification des sites promoteur

Les promoteurs qui sont de courts segments de séquences d'ADN précédant le début des gènes, initient le processus d'expression des gènes (FIG. 9). Généralement, le promoteur est le motif "TATAAT", il est connu chez les biologiste par *la boite TATA*. La reconnaissance de sites promoteurs donc, revient à l'extraction du motif "TATAAT". Cependant ce motif seul n'assure pas la bonne identification des gènes dans les longues séquences d'ADN [O'Neill *et al.* 1989]. Dans ce cadre, nous avons analysé des données publiées sur Internet par l'université IRVINE de Californie [Towell 1991]. Il s'agit de 53 séquences contenant des sites promoteurs et 53 séquences représentant des sites non promoteurs. Chaque séquence est formée de 57 caractères pris sur le début de gènes de *"E. Coli"* (déjà identifiés par les biologistes dans des structures d'ADN) [Harley *et al.* 1987].

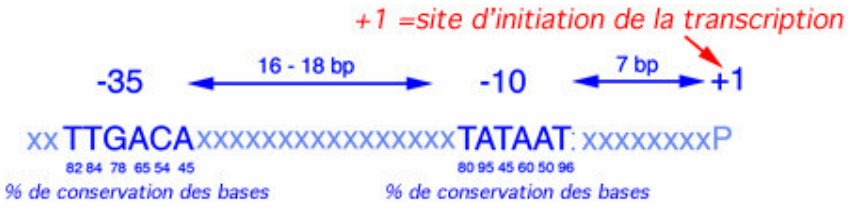

FIG. 9- Composition d'un promoteur procaryote

3.1.2 Problème d'identification des sites de jonction

Lors de la maturation, qui est une phase de passage du Pré-ARN à l'ARNm, certaines zones (zones non codantes : les introns) sont enlevées. Les fragments d'ADN qui sont conservés dans l'ARNm sont appelés exons. La frontière entre un intron et un exon est un site de jonction accepteur noté "IE". Alors que la frontière entre un exon et un intron est un site de jonction donneur noté "EI". Comme le montre la figure 10, le site de jonction est formé par 4 éléments: site donneur "GT", site accepteur "AG", un point de branchement "TACTAAC" et région riche en pyrimidines. Pour pouvoir éliminer les introns lors de la maturation, il suffit d'identifier les sites de jonction donneur et les sites de jonction accepteur. Cependant, les motifs "GT" et "AG" seules n'assurent pas la bonne identification des introns. Dans ce cadre, nous avons analysé des données publiées sur Internet par l'université IRVINE de Californie [Towell *et al.* 1990]. Nous avons considéré 200 séquences contenant des sites de jonction accepteur, 200 séquences contenant des sites de jonction donneur et 200 séquences qui ne forment ni des sites de jonction accepteur, ni des sites de jonction donneur.

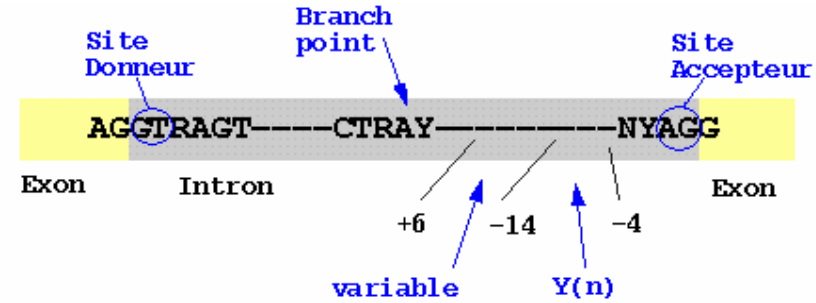

FIG. 10 – Les sites de jonction : site donneur, site accepteur

3.1.3 Problème de classification de protéines

Les protéines sont les molécules les plus complexes et les plus variées des êtres vivants. Les protéines ont plusieurs fonctions biologiques. Il existe par exemple, des enzymes, des protéines de transport, des protéines de défense, etc. Les biologistes regroupent les protéines dans des superfamilles et des familles selon ces fonctions. Généralement, les protéines d'une même famille ont des structures similaires. Donc pour classer une protéine nouvellement découverte dans la bonne famille, on peut se baser sur sa structure primaire. En plus de ça, chaque famille de protéines est caractérisée par un domaine. Le domaine d'une famille est un ensemble de motifs de tailles variables dispersés tout au long des séquences de cette famille [Gibas *et al.* 2002]. Ces motifs sont sauvegardés par les biologistes dans les banques biologiques. Nous allons comparer les motifs extraits par les méthodes, présentées dans la section 2, avec les motifs des banques pour identifier la signification biologique des motifs extraits. Nous avons extrait les familles de protéines de la banque SCOP [Murzin *et al.* 1995].

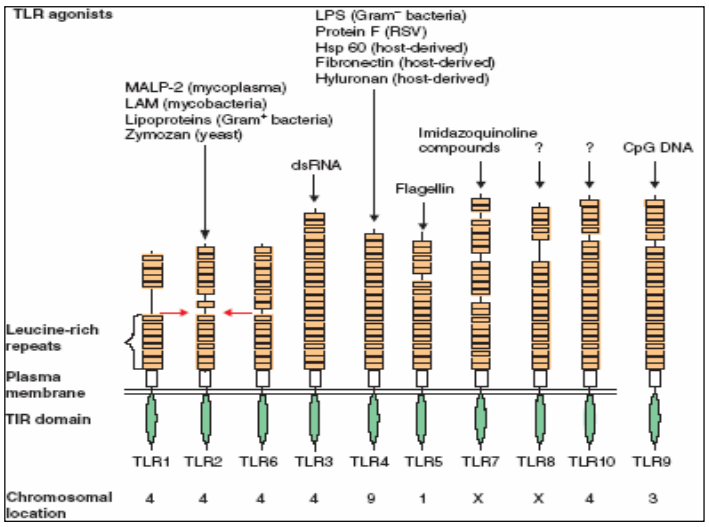

FIG. 11- Structure des protéines TLR chez l'homme

- **Les protéines Toll-Like Receptors (TLRs)**

Les chercheurs du Laboratoire de Biologie Moléculaire de l'Institut Pasteur de Tunis (IPT) s'intéressent aux maladies infectieuses. En Tunisie, les infections parasitaires constituent un problème de santé public (les leishmanioses et la tuberculose). Ils s'intéressent particulièrement à l'analyse de la famille : *Toll Like Receptors* (*TLR*). Ces protéines sont des récepteurs qui jouent un rôle critique dans la réponse immunitaire. Ces récepteurs sont des éléments essentiels de la défense de l'hôte contre les pathogènes grâce à l'activation de l'immunité innée qui est un pré-requis à l'induction de l'immunité adaptative. La figure 8 présente les sous-familles de protéines TLR chez l'Homme. Cette famille possède un domaine intitulé TIR [Ivison *et al.* 2007]. L'ensemble de motifs de ce domaine existe dans la banque Pfam.

Notre étude porte sur la discrimination des TLR humaines de celles non humaines (deux classes). La première classe représente la famille des protéines Toll-Like Receptor (TLR) d'origine humaine. La deuxième représente des protéines TLR non humaines. Nous avons pu construire les attributs (motifs) décrivant les séquences TLR. Ces descripteurs peuvent servir à identifier de nouveaux membres de la famille des TLR. Les résultats de cette application ont été bien accueillis par nos partenaires biologistes [Mhamdi *et al.* 2003].

- **Les protéines Arginine**

Les arginines sont des protéines de type enzymes. Ils ont le rôle de catalyser d'autres protéines. On prend par exemple l'enzyme *superoxyde dismutase* (SOD) qui intervient dans l'élimination de l'anion superoxyde (O2-). Elle est impliquée dans le vieillissement. Dans la deuxième problématique de classification de protéines, nous avons utilisé deux familles. La première, est la famille PAD (Porphyromonas Type Peptidyl Arginine Deiminase) . C'est une enzyme de Deiminase qui catalyse les groupes de guanidines des résidus d'arginine de

l'extrémité carboxylique de divers peptides pour produire l'ammoniaque. La deuxième, est la famille AD (Arginine Décarboxylase) . Ces enzymes catalysent la décarboxylation de l'ornithine, ou l'arginine, ou la lysine

Dans le cadre d'un travail antérieur [Hmaidi *et al.* 2006], nous avons considéré un échantillon de protéines PAD et un échantillon de protéines AD. Nous avons étudié également le problème de caractérisation de ces protéines par construction de motifs. Les motifs découverts ont été confrontés à la banque de motifs Pfam. Dans le tableau 1 nous présentons, les échantillons des données réelles sur lesquelles nous sommes basés dans les expérimentations.

Type de séquences	Échantillons	Familles de séquences	Nombre de séquences
Protéiques	TLR	TLR Humain	14
		TLR Non Humain	26
	Arginine	Arginine PAD (F2)	47
		Arginine AD (F3)	54
Nucléiques	Promoteur	Site Promoteur	53
		Site Non Promoteur	53
	Jonction	Site jonction accepteur	200
		Site jonction donneur	200
		Site Non jonction	200

TAB. 1 – Présentation des échantillons de séquence biologiques et leurs tailles

3.2 Étude de la précision en classification des attributs

Pour étudier la précision en classification des attributs découverts par les méthodes étudiées, nous avons essayé de discriminer les familles de séquences deux à deux. Nous avons extrait les motifs par les méthodes n-grammes, arbre de suffixes généralisés et celle des descripteurs discriminants. Nous avons utilisé, par la suite, la méthode d'apprentissage C4.5 [Quinlan 1993] pour extraire des règles de classification se basant sur ces motifs. nous avons trouvé qu'il est incomplet de comparer ces méthodes en considérant une seule valeur de paramètres, à savoir la longueur des n-grammes (la valeur de *n*), les valeurs de α et β relatives aux descripteurs discriminants, ainsi que les valeurs de longueur et d'occurrence minimales pour l'arbre de suffixes généralisés. Nous avons opté pour l'estimation du taux d'erreurs par validation croisée (10 répétitions, 2 paquets) [Kohavi 1995].

3.2.1 Etude de la précision des n-grammes

Valeur de « n »	TLR		Promoteur		Jonction donneur		Jonction accepteur	
	Nombre	Erreur	Nombre	Erreur	Nombre	Erreur	Nombre	Erreur
2	400	0.35	16	0.5472	16	0.5045	16	0.5193
3	5563	**0.38**	64	0.3057	64	**0.3988**	64	0.3288
4	15484	0.39	256	**0.2585**	256	0.4265	256	0.2740
5	-	-	974	0.3472	1019	0.4393	1024	**0.2217**
6	-	-	2330	0.3811	3723	0.4392	3790	0.24
7	-	-	3240	0.4981	6912	-	9816	-
8	-	-	3577	0.5142	-	-	-	-

TAB 2- Variation du nombre et du taux d'erreurs des n-grammes selon la taille n

Pour calculer la taille optimale des *n*-grammes, nous avons expérimenté plusieurs valeurs sur différents échantillons de données : TLR, Promoteur, Jonction donneur et Jonction accepteur. Le tableau 2 montre que la valeur optimale varie entre 3, 4 et 5. Ces valeurs donnent les taux d'erreurs minimaux. Il est à noter qu'une valeur optimale universelle n'existe pas. Elle varie avec les données et avec la méthode d'apprentissage. Théoriquement, le nombre des n-grammes possibles est très grand. Pour les protéines, il est égal à 20^n. Pour n=4, on a 20^4=160.000 n-grammes possibles. Nous avons remarqué que le nombre des 5-grammes, figurant dans les données, est de l'ordre de soixantaine de milliers pour le cas de protéines TLR. Ce nombre est trop élevé pour être traité par les méthodes d'apprentissage artificiel (notamment, le logiciel SIPINA que nous avons utilisé [Zighed *et al.* 2000]).

		TLR		**Promoteur**		**Jonction donneur**		**Jonction accepteur**	
α	**β**	**Nombre**	**Erreur**	**Nombre**	**Erreur**	**Nombre**	**Erreur**	**Nombre**	**Erreur**
0	**100**	20	0.35	4 (a, c, g, t)	--	5	0.225	5	**0.2333**
100	**0**	0	--	0	0	0	--	0	--
0	**0**	7127	--	730	0.2453	4208	0.315	4044	0.4641
10	**20**	2696	0.34	194	0.2094	206	0.3168	200	0.3162
10	**40**	2615	0.4150	144	0.2670	131	0.3248	150	0.3138
10	**60**	2130	0.4	54	0.2896	58	0.1655	54	0.2467
10	**80**	1142	0.45	64	0.3028	38	0.168	43	0.2449
20	**20**	1026	0.3475	125	0.2311	51	0.217	58	0.3344
20	**40**	1246	0.38	115	0.2623	114	0.306	125	0.3215
20	**60**	1207	0.42	54	0.3113	55	0.1678	54	0.2554
20	**80**	811	0.47	64	0.3132	38	0.1625	43	0.2467
30	**20**	268	0.3675	46	0.1877	8	0.188	6	0.3549
30	**40**	562	0.33	60	0.2538	59	0.3493	56	0.3628
30	**60**	682	0.3575	52	0.2877	50	0.1715	48	0.2397
30	**80**	573	0.425	60	0.2887	38	0.1687	43	0.2418
40	**20**	74	0.3475	14	**0.1717**	1	0.3075	0	--
40	**40**	221	0.37	26	0.2509	21	0.3115	17	0.3764
40	**60**	361	0.38	51	0.2934	38	0.1685	36	0.2433
40	**80**	403	0.4025	63	0.2887	38	0.169	43	0.2441
50	**20**	21	0.2875	0	--	0	--	0	--
50	**40**	108	0.345	8	0.3510	5	0.3610	2	0.43
50	**60**	226	0.375	46	0.3019	25	0.1692	29	0.2397
50	**80**	306	0.4275	58	0.3245	37	0.1645	40	0.2497
60	**20**	1	0.3325	0	--	0		0	--
60	**40**	23	0.365	5	0.3311	1	0.3795	0	--
60	**60**	77	0.3675	31	0.3009	15	0.1618	17	0.2346
60	**80**	170	0.39	43	0.2877	30	0.164	33	0.2392
70	**20**	1	0.3125	0	--	0	--	0	--
70	**40**	8	**0.31**	1	0.3991	0		0	
70	**60**	32	0.405	16	0.3255	4	0.1622	5	0.2364
70	**80**	100	0.4175	23	0.3170	19	0.1642	17	0.2546
80	**20**	0	--	0	--	0	--	0	--
80	**40**	0	--	0	--	0	--	0	--
80	**60**	5	0.3725	6	0.3075	2	**0.16**	**1**	**0.2333**
80	**80**	49	0.4325	8	0.2962	7	0.1685	6	0.2544
90	**20**	0	--	0	--	0	--	0	--
90	**40**	0	--	0	--	0	--	0	--
90	**60**	0	--	0	--	1	0.225	**1**	**0.2333**
90	**80**	11	0.355	0	--	1	0.225	**1**	**0.2333**

TAB 3- Variation du nombre et du taux d'erreurs des descripteurs discriminants selon α et β

3.2.2 Etude de la précision des descripteurs discriminants

Nous avons expérimenté plusieurs valeurs de α et β sur différents échantillons de données : TLR, Promoteur, Jonction donneur et Jonction accepteur. Pour chaque couple de valeurs de α et β, nous avons calculé le taux d'erreurs de classification avec le classifieur C4.5 de [Quinlan 1993].

Le tableau 3 présente les résultats de ces expérimentations. Ce tableau montre, dans un premier lieu, que des valeurs universellement optimales n'existent pas. Elles varient avec les données. Nous remarquons aussi que le nombre de descripteurs discriminants est considérablement réduit par toutes les valeurs non nulles de α et β. D'ailleurs, les meilleurs taux d'erreurs sont obtenus avec des nombres d'attributs très réduits : 8, 14, 2 et 1.

3.2.3 Etude de la précision des arbres de suffixes généralisés

, Nous avons expérimenté plusieurs valeurs de taille et d'occurrence minimales correspondant à la méthode des arbres de suffixes généraliséssur différents échantillons de données : TLR, Promoteur, Jonction donneur et Jonction accepteur. Pour chaque couple de valeurs, nous avons calculé le taux d'erreurs de classification avec le classifieur C4.5 de [Quinlan 1993].

Le tableau 4 présente les résultats de ces expérimentations. Ce tableau montre, dans un premier lieu, que les meilleurs taux d'erreurs sont obtenus souvent avec une occurrence minimale de 25%. Nous remarquons aussi que des valeurs universellement optimales de taille minimale n'existent pas. Elles varient avec les données : 8, 2, 3 et 4.

Echantillon de données	**Taille minimale**	**Nombre minimal d'occurrences**					
		25%		**50%**		**75%**	
		Nombre d'attributs	**Taux d'erreurs**	**Nombre d'attributs**	**Taux d'erreurs**	**Nombre d'attributs**	**Taux d'erreurs**
TLR	**2**	1864	0.4775	427	0.4525	236	0.37
	3	1481	0.3925	147	0.47	20	0.3825
	4	748	0.425	9	0.4175	1	0.3850
	5	508	0.4	1	0.3850	-	-
	6	383	0.3875	-	-	-	-
	7	298	0.4175	-	-	-	-
	8	238	**0.35**	-	-	-	-
	9	190	0.3975	-	-	-	-
Promoteur	**2**	148	**0.2868**	71	0.3047	21	0.4538
	3	132	0.2792	55	0.3189	5	0.4009
	4	69	0.3028	0	-	0	-
Jonction donneur	**2**	137	0.4880	56	0.4885	18	0.532
	3	121	**0.4575**	40	0.498	3	0.5245
	4	61	0.485	0	0	0	-
Jonction accepteur	**2**	148	O.31	55	0.335	20	0.5263
	3	132	0.323	39	0.313	5	0.5325
	4	68	**0.3025**	0	-	0	-

TAB. 4 - Variation du nombre et du taux d'erreurs des motifs générés par le GST selon la taille minimale et le nombre minimal d'occurrences

3.3 Étude de la signification biologique des attributs

Afin d'étudier la signification biologique des motifs découverts par les méthodes présentées dans la section 2, nous avons procédé à leurs comparaison avec les motifs découverts par les biologistes [Hmaidi *et al.* 2006]. Dans le cas de l'analyse des séquences protéiques (TLR et Arginine), nous avons eu recourt à la banque de motifs protéiques Pfam [Bairoch *et al.* 1994]. A l'aide de cette banque, nous avons obtenu les motifs de chacune de ces familles. Dans le cas de l'analyse des séquences nucléiques (promoteur et jonction), nous avons eu recourt à des études sur ces sujets préalablement établies par les biologistes (respectivement [Harley *et al.* 1987] et [Towell 1991]).

3.3.1 Étude des n-grammes

Pour calculer la taille optimale des *n*-grammes. Nous avons expérimenté plusieurs valeurs de *n* (la taille du n-grammes) sur différents échantillons de données : TLR, Arginine, Promoteur et Jonction. Le tableau 5 présente, pour chaque valeur, le nombre d'attributs construits et le nombre d'attributs protéiques qui figurent dans Pfam, ainsi que le nombre d'attributs nucléiques qui figure dans l'étude de [Towell 91].

	TLR		**ARGININE**		**SITES PROMOTEURS**		**SITES DE JONCTION**		
n	**Humain**	**Non Humain**	**PAD**	**AD**	**Promoteur**	**Non promoteur**	**Non jonction**	**Jonction donneur**	**Jonction accepteur**
1	0	0	16	12	0	0	0	0	0
2	4	10	36	41	0	0	0	0	0
3	0	5	88	87	0	0	2	0	0
4	1	3	48	89	0	0	0	0	2
5	5	2	3	10	1	1	0	0	0
6	6	1	0	39	180	61	0	0	0
7	3	0	0	12	8	1	0	0	0
8	1	2	0	11	0	0	0	0	0
9	0	0	0	12	0	0	0	4	0
10	1	0	0	2	0	0	0	0	0

TAB. 5 - Nombre de n-grammes en commun avec Pfam et l'étude de [Towell 91])

3.3.2 Étude des descripteurs discriminants

Nous avons fait varier les valeurs de α et β. Le tableau 5 présente, pour chaque couple de valeurs de ces paramètres, le nombre total de descripteurs discriminants protéiques figurant dans Pfam, ainsi que le nombre de descripteurs discriminants dans l'étude de [Towell 91].

Nous remarquons que les valeurs α=20 et β=80 peuvent être considérés comme meilleurs paramètres dans l'extraction de motif protéiques. Cependant, pour les séquences nucléiques (cas des séquences promotrices ou des séquences de jonction), les motifs obtenus avec toutes les valeurs non nulles sont des faux motifs. Aucun d'eux ne figure dans l'étude de [Towell 91].

PARAMETRES		TLR		ARGININE		SITES PROMOTEURS		SITES DE JONCTION		
α	β	Humain	Non Humain	PAD	AD	Promoteur	Non promoteur	Non jonction	Jonction donneur	Jonction Accepteur
0	0	0	3	48	38	40	1	0	0	0
0	100	0	0	0	0	0	0	0	0	0
20	100	0	0	0	0	0	0	0	0	0
40	100	0	0	0	0	0	0	0	0	0
60	100	0	0	0	0	0	0	0	0	0
0	80	0	7	23	22	0	0	0	0	0
0	60	0	5	24	33	0	0	0	0	0
0	40	0	4	42	47	0	0	0	0	0
20	80	0	6	22	22	0	0	0	0	0
40	60	0	2	6	6	0	0	0	0	0
60	40	0	0	0	0	0	0	0	0	0
80	20	0	0	0	0	0	0	0	0	0

TAB 6- Nombre de DD en commun avec Pfam et l'étude de [Towell 91]

3.3.3 Étude des motifs basés sur l'arbre de suffixes généralisés

Nous avons fait varier les tailles et les occurrences des motifs de l'arbre de suffixes généralisés. En premier lieu, nous fixons une taille et nous varions l'occurrence. En second lieu, nous fixons l'occurrence tout en variant la taille. Le tableau 7 présente, pour chaque couple de valeurs de ces paramètres, le nombre d'attributs protéiques figurant dans Pfam, ainsi que

PARAMETRES		TLR		ARGININE		SITES PROMOTEURS		SITES DE JONCTION		
Taille	Occr	Humain	Non Humain	PAD	AD	Promoteur	Non promoteur	Non jonction	Jonction donneur	Jonction Accepteur
2	2	20	27	170	215	189	63	2	2	3
2	6	7	17	171	221	189	62	2	2	3
2	10	4	16	171	218	183	60	2	2	3
2	14	4	16	170	217	**178**	**49**	2	2	3
2	24	0	**15**	**153**	**199**	80	18	**2**	**2**	**3**
5	2	15	24	--	--	189	63	0	0	2
10	2	4	0	--	--	0	0	0	0	0
5	6	**3**	1	0	4	189	62	--	--	--
10	6	0	1	0	0	0	0	--	--	--
5	10	--	--	0	3	--	--	0	0	2
10	10	--	--	0	0	--	--	0	0	0

TAB. 7 - Nombre de motif GST en commun avec Pfam et l'étude de [Towell 91]

Le nombre d'attributs figurant dans l'étude de [towell 91]. Les meilleurs paramètres engendrent le nombre le plus important de motifs en commun avec la banque Pfam ou l'étude de [Towell 91]. Pour les séquences nucléiques, il faut considérer en plus que le nombre de motif en commun avec l'étude de [Towell 91] correspondant aux contre exemples ("non jonction" et "non promoteur") soit le plus petit possible. En effet, les données correspondantes à

la ligne **"non jonction"** (respectivement **"non promoteur"**), ne doivent pas contenir les motifs reconnus par les biologistes comme descripteurs des sites de jonctions (respectivement des sites promoteurs). Pour cette raison, les meilleurs paramètres sont ceux qui ont le pourcentage le plus petit de motifs identifiés par [Towell 90] comme descripteurs des sites de jonctions (respectivement des sites promoteurs).

Nous remarquons que les paramètres taille 2 et occurrence 24 peuvent être considérés comme meilleurs paramètres dans l'extraction de motifs. Cependant, pour les contre-exemples (cas des séquences "non promotrices" ou des séquences "non jonction"), nous obtenons avec ces mêmes paramètres un nombre important de faux motifs.

3.4 Synthèse et interprétation des expérimentations

Nous avons trouvé que les trois méthodes (n-gramme, GST et la notre) sont paramétrées. Par conséquent, la qualité des résultats de chaque méthode dépend de la valeur du paramètre considéré. Il nous semble incomplet de comparer ces méthodes en considérant une seule valeur de paramètres. Nous étions obligé donc d'étudier la variation de leurs taux d'erreurs selon différentes valeurs de chaque paramètre.

Pour comparer la qualité des attributs construits par les méthodes présentées, nous considérons les meilleures valeurs des paramètres de chaque approche. Le tableau 8 résume les meilleurs pourcentages de motifs significatifs atteints par chaque approche, ainsi que les valeurs des paramètres correspondants. Le tableau 9 résume les taux d'erreurs optimaux atteints par chaque approche, ainsi que les valeurs des paramètres correspondants.

	Arbre de suffixes généralisés		**Descripteurs discriminants**		**N-grammes**	
Échantillons de données	**Paramètres optimaux (T, Oc)**	**% motifs significatifs**	**Paramètres optimaux (α, β)**	**% motifs significatifs**	**Paramètres optimaux (N)**	**% motifs significatifs**
TLR humaines	5, 6	0.189%	Tous	0%	**2**	**1.002%**
TLR non humaines	2, 24	0.202%	(20,80)	1%	**2**	**2.5%**
Arginine PAD	2, 24	1.051%	(20,80)	9.482%	**1**	**80%**
Arginine AD	2, 24	1.191%	(20.80)	7.508%	**1**	**60%**
Pas de jonction	2, 24	0.023%	**Tous**	**0%**	3	3.125%
Jonction donneur	2, 24	0.015%	Tous	0%	**9**	**0.038%**
Jonction accepteur	2, 24	0.034%	Tous	0%	**4**	**0.781%**
Promoteur	2, 14	6.5%	(0,0)	10.23%	**6**	**10.33%**
Non promoteur	2, 14	2.45%	**(0,0)**	**0.30%**	6	5.859%

TAB. 8 – Comparaison entre les différentes méthodes selon la signification biologique des motifs extraits en utilisant leurs paramètres optimaux

Le tableau 8 montre que l'approche des n-grammes permet d'obtenir des motifs plus proches de ceux obtenus par les experts biologistes. Seulement, même avec cette approche le pourcentage des motifs significatifs reste très faible (à l'exception des protéines Arginines).

Nous remarquons ainsi que ces meilleurs pourcentages ne correspondent pas à une même longueur des n-grammes (2, 1, 4, 6 puis 9)

	Arbre de suffixes généralisés		Descripteurs discriminants		N-grammes	
Échantillons de données	**Paramètres optimaux (T, Oc)**	**Taux d'erreurs**	**Paramètres optimaux (α, β)**	**Taux d'erreurs**	**Paramètres optimaux (N)**	**Taux d'erreurs**
TLR	8, 10	35%	**70, 40**	**0.31**	3	0.38
Jonction donneur	3, 10	45.75%	**80, 60**	**0.16**	3	0.3988
Jonction accepteur	4, 10	30.25%	80, 60	0.2333	**5**	**0.2217**
Promoteur	2, 10	28.68%	**40, 20**	**0.1717**	4	0.2585

TAB. 9 - Comparaison entre les différentes méthodes selon les taux d'erreurs de classification en utilisant leurs paramètres optimaux

Le tableau 9 montre que l'approche des descripteurs discriminants garde l'avantage d'être l'approche la plus précise pour le classifieur C4.5. Pour les sites de jonction donneur et les sites promoteurs, les taux d'erreurs des descripteurs discriminants sont nettement meilleurs que ceux des autres approches.

4 Conclusions et Perspectives

Pour pouvoir extraire des connaissances à partir des séquences biologiques (ADN, ARN et protéines), nous devons procéder à une étape de construction d'attributs. Dans un travail antérieur, nous avons étudié l'approche qui utilise les caractéristiques biochimiques [Maddouri *et al.* 2004]. Cette approche est simple à déployer car les caractéristiques sont prédéfinies par le biologiste. Seulement, elle est difficile à utiliser en fouille de données car elle produit un nombre très élevé d'attributs symboliques (surtout dans le cas des protéines, où l'alphabet est étendu).

Dans cet article, nous avons étudié les solutions existantes pour la construction de motifs contenus dans les séquences biologiques. Nous avons étudié la méthode basée sur les n-grammes, largement utilisée pour la recherche d'homologies dans les banques de séquences. Nous avons présenté, également, la méthode basée sur les arbres de suffixes généralisés et la méthode basée sur les modèles de Markov cachés. Nous avons présenté également la méthode de construction de descripteurs discriminants [Maddouri *et al.* 2002].

La méthode de construction de descripteurs discriminants détermine les sous-séquences (motifs) minimales répétées dans une famille de séquences. Les sous-séquences obtenues peuvent être de tailles différentes. Au contraire, l'approche des n-grammes se limite à une taille *n* fixée au préalable. En plus, la méthode de construction de descripteurs discriminants permet d'extraire seulement des sous-séquences figurant plus que α% de fois dans les exemples et au plus β% de fois dans les contre-exemples. Elle permet ainsi de refléter la réalité biologique relativement complexe et ambiguë.

L'étude expérimentale est faite sur deux échantillons de séquences protéiques et deux échantillons de séquences nucléiques. Cette étude a montré que la méthode de construction de descripteurs discriminants a deux avantages : une meilleure précision en classification par C4.5 et un nombre d'attributs plus réduit. L'approche des n-grammes a l'avantage de reproduire des motifs plus reconnus par les études biologiques. Il est à noter que ces résultats restent

dépendants du classifieur C4.5. Actuellement, nous sommes en train de tester d'autres classifieurs linéaires (SVM linéaire, PLS et Naïve bayesien) et non linéaires (SVM RBF, CART et 1-PPV).

Nous avons remarqué que les systèmes BLAST et FASTA adoptent l'approche des n-grammes utilisant les *"matrices de substitution"* pour prendre en considération l'aspect d'imprécision au niveau de l'apparition des motifs [Altschul *et al.* 1997]. Ce qui leurs permet de prendre en charge les insertions, les suppressions et les mutations qui surgissent souvent durant l'évolution génétique. L'approche des arbres de suffixes généralisés a l'avantage de prendre en considération cet aspect. Actuellement, nous essayons d'améliorer la méthode de construction de descripteurs discriminants par l'utilisation de matrices de substitution comme PAM [Dayhoff *et al.* 1979] ou BLOSUM [Henikoff *et al.* 1992].

Également, nous avons remarqué que le nombre d'attributs construits par l'approche proposée reste élevé, bien qu'il demeure inférieur à celui produit par les autres approches. Par ailleurs, dans le cadre d'un autre travail en cours, nous avons procédé à une étape de sélection d'attributs [Jourdan *et al.* 2002]. Ce travail à pour but d'éliminer, en postériori, les attribut non pertinents qui dégradent les résultats.

Remerciement

Les auteurs remercient les évaluateurs anonymes pour leurs remarques pertinentes qui ont contribuées à l'amélioration de la qualité de cet article. L'auteur principal remercie, également, ses étudiants F. Mhamdi et S. Hmaidi pour la participation aux expérimentations.

Références

[Altschul *et al.* 1990] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. *Basic local alignment search tool.* Journal of Molecular Biology, Vol. **21**5(3), pp. 403-413, 1990.

[Altschul *et al.* 1997] S. F. Altschul, T. L. Madden, A. A. Scha_er, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. *Gapped Blast and PSI-Blast: A new generation of protein database search programs*. Journal of Nucleic Acids Research, Vol. 25(17), pp. 3389-3402, 1997.

[Bairoch et al. 1994] A. Bairoch, P. Bucher. *PROSITE: recent developments*. Nucleic Acids Research, Vol. 22, page : 3583-3589, 1994.

[Sandmeier *et al.* 1994] E. Sandmeier, T.I. Hale, P.Christen. *Multiple evolutionary origin of pyridoxal -5'- phosphate – dependent amino acid decarboxylases*. Eur. J. Biochen . **221**: pp. 997--1002, 1994.

[Chuang *et al.* 2001] T. Chuang, R. J. Ulevitch. *Identification of hTLR10: a novel human Toll-like receptor preferentially expressed in immune cells*. Biochim Biophys Acta, v. 1518, pp. 157-161, 2001

[Bes 2002] M. Bes. *Extraction de règles d'association sur les données de puces à ADN*. Mémoire de Diplôme d'Etudes Approfondies de Génomique et Informatique, Université de Rennes-1, Rennes - France, Juin 2002.

[Cornuéjols *et al.* 2002] A. Cornuéjols, L. Miclet. *Apprentissage artificiel : concepts et algorithmes*. Editions Eyrolles, pages : 591, France, 2002.

[Dardel *et al.* 2002] F. Dardel, F. Képès. *Bioinformatique, Génomique et Post-génomique.* Editions de l'Ecole Polytechnique, Paris - France, Septembre 2002.

[Delaplace 2003] F. Delaplace. *Analyse statique : du calcul haute performance à la bioinformatique*. Mémoire d'Habilitation à Diriger des Recherches, Université d'Evry Val d'Essone, France, 28 novembre 2003.

[Dayhoff *et al.* 1979] M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt. *A model for evolutionary change*. Atlas of Protein Sequence and Sructure, Vol. 5(3), pp. 345-352, 1979.

[Dickerson *et al.* 1969] R.E. Dickerson, I. Geis. *The Structure and Actions of Proteins*. Harper and Row Publishers, New York - USA, 1969.

[Dumas *et al.* 1982] J. P. Dumas, J. Ninio. *Efficient algorithms for folding and comparing nucleic acid sequences*. Nucleic Acids Research, Vol. 10, page : 197-206, 1982.

[Fayyad *et al.* 1996] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, 1996.

[Fu *et al.* 2004] H. FU, M. N. Engelbert. *Clustering Binary Codes to Express the Biochemical Properties of Amino Acids*. International Conference on Intelligent Information Processing (ICIIP), October 2004.

[Fu 2001] H. FU, *Intelligence Artificielle et codage de séquences de protéines*. Mémoire de Mastère en informatique, Université des Sciences et de Technologie de Lille, Lille - France, Juillet 2001.

[Harley *et al.* 1987] C. Harley and R. Reynolds. *Analysis of E. Coli Promoter Sequences*. Journal of Nucleic Acids Research, Vol. 15, pp. 2343-2361, 1987.

[Henikoff *et al.* 1992] S. Henikoff, J. G. Henikoff. *Amino acid substitution matrices from protein blocks*. National Academy of Sciences, USA, 89, pp. 10915-10919, 1992.

[Hmaidi *et al.* 2006] S. Hmaidi, M. S. BenMarzouk. *Etude des logiciels d'extraction de motifs à partir de séquences biologiques*, Mémoire de Mastère spécialisé en Bio-informatique, Ecole Nationale des Sciences de l'Informatique, Tunis - Tunisie, Avril 2006.

[Jourdan *et al.* 2002] L. Jourdan, C. Dhaenens, E.G. Talbi and S. Gallina. *A Data Mining Approach to Discover Genetic and Environmental Factors involved in Multifactorial Diseases*. Journal of Knowledge Based Systems, 15(4) pp. 235-242, May 2002.

[Krogh *et al.*, 1994] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, D. Haussler. *Hidden markov models in computational biology: Applications to protein modeling*. Journal of Molecular Biology, 235, pp.1501-1531, February 1994.

[Karp *et al.* 1972] R. Karp, R. E. Miller, A. L. Rosenberg. *Rapid Identification of Repeated Patterns in Strings, Trees and Arrays*. 4th Symposium of Theory of Computing, pp.125-136, 1972.

[Kohavi 1995] R. Kohavi. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. International Joint Conference on Artificial Intelligence, IJCAI-95, pp.1137-1143, 1995. Morgan Kaufmann.

[Liu *et al.* 1998] H. Liu, H. Motoda. *Feature Extraction, Construction and Selection : A Data Mining Perspective*. Volume 453, Kluwer Academic Publishers, Boston, July 1998.

[Liu *et al.* 2001] H. Liu and H. Motoda. *Instance Selection and Construction for Data Mining*. Volume 608. Kluwer Academic Publishers, Boston, February 2001.

[Maddouri *et al.* 2002] M. Maddouri, M. Elloumi, *A Data Mining Approach based on Machine Learning Techniques to Classify Biological Sequences*, Journal of Knowledge Based Systems, Vol. 15, Issue 4, Elsevier Publishing Co., Amsterdam, North-Holland, 2002, p217-223.

[Maddouri *et al.* 2004] M. Maddouri, M. Elloumi, *Encoding of Primary Structures of Biological Macromolecules Within a Data Mining Perspective*, Journal of Computer Science and Technology (JCST), Volume 19, Number 1, p.78-88, January 2004, Science Press (China) and Allerton Press (USA).

[Mhamdi *et al.* 2003] F. M'hamdi, M. Elloumi, M. Maddouri, K. Bsaies and S. Abd Elhak. *Adaptation du Système DisClass à la Présence des Gaps dans des motifs de Séquences Protéiques*. Congrès Annuel de la SFBBM sur le Post-génome : de la protéine aux molécules bio-actives, Lyon-France, 4 et 5 novembre 2003.

[Mhamdi *et al.* 2005] F. M'hamdi, M. Maddouri, M. Elloumi. *Extraction et comparaison entre n-grammes et descripteurs discriminants pour la classification de protéines*. Conférence Francophone sur l'Apprentissage Automatique (Cap2005) de l'AFIA, Atelier Apprentissage et Bioinformatique, Nice-France, du 31 mai au 3 juin 2005.

[Mitra *et al.* 2003] S. Mitra, T. Acharya. *Data Mining: Multimedia, Soft Computing and Bioinformatics*. John Wiley & Sons Inc. Publication, New Jersey - USA, 2003.

[Molla *et al.* 2004] M. Molla, M. Waddell, D. Page and J. Shavlik. *Using Machine Learning to Design and Interpret Gene-Expression Microarrays*. AI Magazine, 25, pp. 23-44, 2004.

[O'Neill *et al.* 1989] M.C. O'Neill, F. Chiafari. *Escherichia coli promoters. II: A spacing class-dependent promoter search protocol*. Journal of Biological Chemestry, 264, pp.5531-5534, 1989.

[Perrière 2000] Guy Perrière. *Bases de données et outils d'analyse pour la génomique bactérienne*. Mémoire d'Habilitation à Diriger des Recherches, Université Claude Bernard – LYON 1, Lyon - France, 27 juin 2000.

[Pearson *et al.* 1988] W. R. Pearson, D. J. Lipman. *Improved tools for biological sequence comparison*. National Academy of Sciences, U S A, 85(8), pp.2444-8, April 1988.

[Quinlan 1993] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Towell 1991] G. Towell. *Symbolic Knowledge and Neural Networks: Insertion, Refinement and Extraction*. Thèse de Doctorat (Ph.D.), Université de Wisconsin-Madison, Wisconsin – USA, 1991.

[Towell *et al.* 1990] G. Towell, J. Shavlik and M. Noordewier. *Refinement of Approximate Domain Theories by Knowledge-Based Artificial Neural Networks*. 8th National Conference on Artificial Intelligence (AAAI-90), 1990.

[Wang *et al.* 1994] Wang, J. T. L., Marr, T. G., Shasha, D., Shapiro, B. A., and Chirn, G. *Discovering active motifs in sets of related protein sequences and using them for classification*. Journal of Nucleic Acids Research, 22(14), pp. 2769-2775, 1994.

[Wang *et al.* 1999] J. T. L. Wang, S. Rozen, B. A. Shapiro, D. Shasha, Z. Wang, and M. Yin. *New techniques for DNA sequence classification*. Journal of Computational Biology, 6 (2), pp. 209-218, 1999.

[Wang *et al.* 2000] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. *Application of neural networks to biological data mining: A case study in protein sequence classification*. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.305-309, 2000.

[Wang *et al.* 2001] J. T. L. Wang, Qicheng Ma, Dennis Shasha and Cathy H. Wu. *New Techniques for Extracting Features from Protein Sequences*. IBM Systems Journal, 40(2), pp. 426-441, 2001.

[Yin *et al.* 2004] M. M. Yin and J. T. L. Wang. *GeneScout: A Data Mining System for Predicting Vertebrate Genes in Genomic DNA Sequences*. Journal of Information Sciences, 163(1-3), pp. 201-218, June 2004.

[Yu *et al.* 2003] L. T. H. Yu, F. Chung, S. C.F. Chan. *Emerging Pattern Based Projected Clustering for Gene Expression Data*. European Workshop on Data Mining and Text Mining for Bioinformatics, Held in Conjunction with ECML / PKDD- 2003, Dubrovnik – Croatia, September 2003.

[Zighed *et al.* 2000] D. A. Zighed, R. Rakotomalala. *Graphes d'induction: apprentissage et data mining*. Hermes Sciences Publications, pages : 475, France, 2000.

[Sebastiani 2005] F. Sebastiani, *Text Mining and its Applications to Intelligence, CRM and Knowledge Managemen*t, p. 109–129, WIT Press, Southampton, UK, 2005.

[Radwan *et al.* 2003] J. Radwan, C. Jérémy, R. Rakotomalala. *Un cadre pour la catégorisation de textes multilingues*. In Proc of 7èmes Journées internationales d'Analyse statistique des Données Textuelles. p 650-660, 2003

[Murzin *et al.* 1995] G.A. Murzin, E.S. Brenner, T. Hubbard and C. Chothia. *SCOP: a structural classification of proteins database for the investigation of sequences and structures*, J. Mol. Bio., v.247, pp. 536--540, 1995.

[Gibas *et al.* 2002] C. Gibas, P. Jambeck. *Introduction à la bioinformatique*, Oreilly 2002.

[Li *et al.* 2005] Wei Li,Clifford A.Meyer, X.Shirley Liu. *A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences*. Bioinformatics Journal, Vol.21 Suppl.1, pp- i274 -– i282, 2005

[Huang *et al.* 2003] Y. F. Huang, S. Y. Lin. *Mining sequential patterns using graph search techniques*. Proceedings of the 27th Annual International Computer Software and Applications Conference, COMPSAC 2003, pp. 4--9, November 2003

[Keinduangjum *et al.* 2005] J. Keinduangjun, P. Piamsa-nga, Y. Poovorawan. *Models for Discovering Signature Patterns in DNA Sequences*. From Proceeding (458) Biomedical Engineering, 2005.

[Ivison *et al.* 2007] S.M. Ivison, M.A. Khan, N.R. Graham, C.Q. Bernales, A. Kaleem, C.O. Tirling, A. Cherkasov, T.S. Steiner. *A phosphorylation site in the Toll-like receptor 5 TIR domain is required for inflammatory signalling in response to flagellin.* Biochem Biophys Res Commun, V. 352(4), pp. 936—41, 2007

Construction d'attributs pour l'extraction de connaissances à partir de séquences biologiques

Abstract. In this paper, we handle a problem of data pre-processing: construction of features describing biological sequences. In order to discover knowledge from biological sequences (DNA, RNA or proteins), all data-mining systems face the difficulty of unusual data representation/format. In fact a biological sequence is represented, in its primary structure, as a string. To transform this data format to a relational table, a step of feature construction is required.

This paper outlines existent methods allowing the construction of features describing biological sequences: the method based on Discriminant Descriptors, the method based on N-Grams, the method based on Generalised Suffix Tree and the method based on Hidden Markov Models. Our main contribution is the theoretical and experimental comparative study of these methods. Particularly, we used these methods to handle typical biological problems: the recognition of promoter sites in E. Coli genes, the recognition of junction splice sites for Primate and the classification of protein families. A comparison of the features built by these in-silico methods with those in the motif data bank Pfam is also presented.

.