# Classification supervisée de séquences biologiques, basée sur les motifs et les matrices de substitution

Rabie Saidi*, Mondher Maddouri**
Engelbert Mephu Nguifo***

*FSJEG, Université de Jendouba, rue de l'UMA, Tunisia
saidi@cril.univ-artois.fr
** Institut National des Sciences Appliquées and Technologies (INSAT),
Tunis-Carthage 2035, Tunisia
mondher.maddouri@fsegt.rnu.tn
*** CRIL – CNRS, Université d'Artois - IUT de Lens, France
mephu@cril.univ-artois.fr

**Résumé.** La classification des séquences biologiques est l'un des importants défis ouverts dans la bioinformatique, tant pour les séquences protéiques que pour les séquences nucléiques. Cependant, la présence de ces données sous la forme de chaînes de caractères ne permet pas de les traiter par les outils standards de classification supervisée, qui utilisent souvent le format relationnel. Pour remédier à ce problème de codage, plusieurs travaux se sont basés sur l'extraction des motifs pour construire une nouvelle représentation des séquences biologiques sous la forme d'un tableau binaire. Nous décrivons une nouvelle approche qui étend les méthodes précédentes par l'utilisation de matrices de substitution dans le cas des séquences protéiques. Nous présentons ensuite une étude comparative qui prend en compte l'effet de chaque méthode sur la précision de la classification mais aussi le nombre d'attributs générés et le temps de calcul.

## 1 Introduction

L'émergence de la bioinformatique, que nous témoignons durant les dernières années, trouve ses causes dans les progrès technologiques qui ont permis de conduire des projets de recherche à grande échelle. Le plus remarquable était le Projet du Génome Humain (PGH) [National Human Genome Research Institute, 2006] accompli en 13 ans depuis 1990 ; période qui s'avère très courte comparée avec la quantité de données extraites sur le génome humain : 3 milliards de bases qui constituent l'ADN humain. Ainsi, plusieurs problèmes sont ouverts :

- Comment le gène exprime-t-il sa protéine ?
- Où commence le gène et où finit-il ?
- Comment évoluent les familles de protéines et comment les classer ?
- Comment prédire la structure tridimensionnelle des protéines ?
- ….

Dans ce contexte, le besoin en fouille de données se fait de plus en plus pressant. Cependant, les techniques de fouille de données, qui traitent souvent des données sous le format relationnel, se trouvent confrontées au format inapproprié des séquences biologiques qui se

présentent sous la forme de chaînes de caractères. Ceci rend nécessaire la transformation de ces données avant de les analyser. Notre travail se situe dans le cadre du prétraitement des séquences biologiques à savoir leur codage sous un format standard et approprié à l'analyse qui est généralement le format relationnel utilisé couramment par les outils de fouille de données. Nous étudions et comparons quelques méthodes existantes de codage des séquences basées sur l'extraction des motifs. Nous proposons, aussi, une nouvelle approche qui étend les méthodes à base d'extraction de motifs pour le cas des séquences protéiques. Ces méthodes sont implémentées en langage C et regroupées dans une librairie permettant de les comparer en termes de précision, de nombre d'attributs générés et de temps de calcul.

L'introduction au problème et la motivation au codage des séquences biologiques font l'objet de la section 2. Dans la section 3, nous présentons un survol sur quelques méthodes de codage basées sur l'extraction des motifs. Nous proposons d'améliorer l'une de ces méthodes dans la section 4. Dans la section 5, nous effectuons une étude expérimentale. Puis, nous discutons les résultats obtenus dans la section 6. La section 7 conclut ce travail et indique quelques perspectives de futurs travaux.

## 2 Classification des séquences biologiques par les outils standard

La classification est l'un des problèmes ouverts les plus importants en bioinformatique. Ce problème se présente, aussi bien, pour les protéines que pour les ADN. En effet, les biologistes s'intéressent souvent à identifier la famille à laquelle appartient une protéine nouvellement séquencée. Ceci permet d'étudier l'évolution de cette protéine mais aussi de savoir ses fonctions biologiques. Pour les ADN, les biologistes cherchent par exemple à classifier des parties de séquences en zones codantes ou non codantes [Maddouri et al, 2002]. Ils utilisent des moyens biochimiques et des analyses *in vitro* pour effectuer ces tâches, ce qui s'avère très coûteux en termes de temps et d'argent, tandis que la quantité des séquences biologiques ne cesse de croître.

Dans ce contexte, l'utilisation des techniques de fouille de données se révèle un choix rationnel, puisqu'elles ont été efficaces dans diverses applications et particulièrement pour la classification supervisée. Cependant, sachant que les séquences biologiques sont représentées sous la forme de chaînes de caractères et que les outils de fouille de données traitent souvent les données sous le format relationnel, il n'est pas possible d'appliquer ces outils sur telles données. Ce qui fait que les séquences biologiques doivent être codées sous un autre format. Pour résoudre ce problème, [Maddouri et al, 2004] proposent un processus de fouille de données biologiques, dont le modèle est illustré par figure 1. Le modèle présente les trois étapes principales du processus de l'Extraction de Connaissances à partir de Données (ECD) appliqué au problème de la classification des séquences biologiques. Il consiste à l'extraction d'un ensemble de motifs à partir d'un ensemble de séquences. Ces motifs seront utilisés comme attributs pour construire un tableau binaire qui contient en ligne l'ensemble des séquences en question. La présence ou l'absence d'un attribut dans une séquence sont respectivement notées par 1 ou 0. Ce tableau binaire représente le résultat de la phase du prétraitement et le nouveau format du codage des séquences. Il est utilisé comme donnée pour la phase de fouille de données où un classifieur est appliqué pour générer les règles de classification. Ces règles sont utilisées pour classer d'autres séquences.

Ce travail se situe dans le cadre du prétraitement des séquences biologiques. Nous étudions l'effet de la méthode de codage sur la connaissance extraite par la mesure de la précision du classement.

Dans la section suivante, nous présentons et décrivons quelques méthodes de codage existantes basées sur l'extraction des motifs.

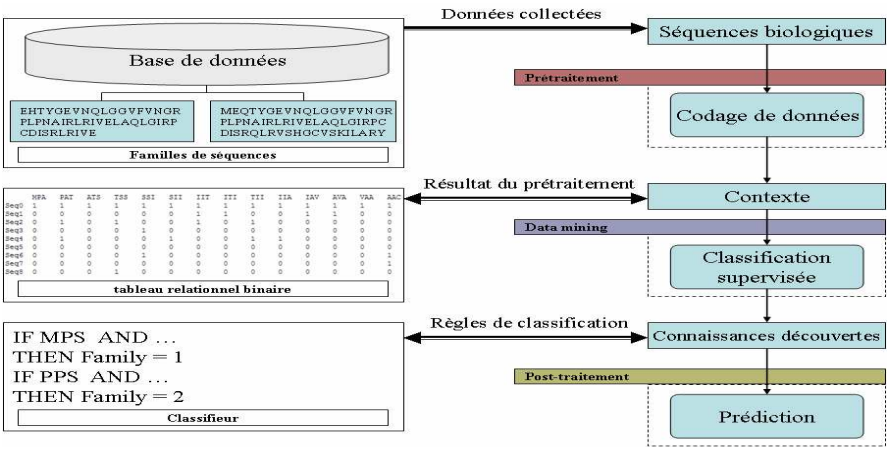


FIG. 1 – *Processus de classification supervisée des séquences biologiques*

# 3 Méthodes de codage existantes

Les séquences nucléiques et protéiques contiennent des patterns ou motifs qui ont été préservés tout au long de l'évolution, vue leur importance en terme de structure ou fonction de la molécule. La découverte de ces motifs peut aider à regrouper les séquences biologiques dans des familles structurelles ou fonctionnelles mais aussi à mieux comprendre les règles qui contrôlent l'évolution.

Les membres d'une famille de protéines sont souvent caractérisés par plus d'un motif : en moyenne chaque famille conserve 3 à 4 régions [Nevill-Manning et al, 1998]. Les motifs indiquent souvent des rapports fonctionnels évolutionnaires entre les protéines. Comme pour les protéines, la découverte des motifs peut être utilisée pour déterminer la fonction des séquences nucléiques comme l'identification des sites promoteurs ou des sites de jonction.

Nous présentons ci-après trois méthodes d'extraction de motifs qui sont les méthodes des N-Grammes (NG), des Motifs Actifs (MA) et des Descripteurs Discriminants (DD). Puis, nous proposons, dans la section 4, une extension des méthodes à base d'extraction de motifs par l'utilisation d'une matrice de substitution. Nous avons appliqué cette extension sur les Descripteurs Discriminants (DDMS).

## 3.1 N-Grammes

L'approche la plus simple est celle des N-Grammes (NG), dite aussi *N-Mots* ou *fenêtrage de longueur N* [Leslie et al, 2002]. Les motifs à construire sont de longueur fixe. Le N-gramme est une sous-séquence composée de N caractères, extraite d'une séquence plus lon-

gue. 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 une sous-séquence des N caractères est extraite. Ce processus est itéré pour toutes les séquences à analyser. Ensuite, on ne gardera que les motifs distincts.

### 3.2 Motifs Actifs

Cette méthode est fondée sur l'hypothèse que les régions importantes sont mieux conservées au cours de l'évolution et donc elles apparaissent plus fréquemment que prévu. En effet, elle permet d'extraire les motifs les plus fréquents, appelés encore *Motifs Actifs* (MA), dans un ensemble de séquences biologiques. L'activité du motif est le nombre de séquences qui le contiennent avec un nombre permis de mutation [Wang et al, 1999].

### 3.3 Descripteurs Discriminants

Etant donné un ensemble de chaînes, de même alphabet, affectées au préalable à P familles/classes F1 F2 .., FP, il s'agit de construire des sous-chaînes appelées Descripteurs Discriminants (DD) qui permettent de discriminer une famille Fi, pour i = 1..P, des autres familles [Maddouri et al, 2004].

Cette méthode est basée sur une adaptation de l'algorithme de Karp, Miller et Rosenberg (KMR) [Karp et al, 1972]. Cet algorithme permet d'identifier les mots répétés dans des chaînes de caractères, des arbres ou des tableaux. Donc, il sera appliqué, ici, sur les séquences biologiques. Les mots répétés, ainsi extraits, sont filtrés pour ne garder que ceux qui sont discriminants et minimaux.

Une sous chaîne $X$ est considérée comme étant discriminante entre la famille $F_i$ et les familles $F_j$, avec $i = 1..P$, $j = 1..P$ et $j \neq i$, ($\alpha$ et $\beta$ fixés) si et seulement si :

$$\frac{\text{nombre de séquences de } F_i \text{ où X apparaît}}{\text{numbre total de séquences de } F_i} * 100 \geq \alpha \cdot \quad (1)$$

$$\frac{\text{nombre de séquences de } F_j \text{ où X apparaît}}{\text{nombre total de séquences des } F_j} * 100 \leq \beta \cdot \quad (2)$$

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

## 4 Motivations biologiques de l'approche proposée

### 4.1 Phénomène de substitution pour les protéines

Pour le cas des protéines, les motifs extraits par la méthode de Descripteurs Discriminants permettent de discriminer entre familles distinctes. Mais, cette méthode néglige le fait que certains acides aminés ont des propriétés semblables et peuvent ainsi se substituer sans pourtant changer la structure ni la fonction de la protéine [Henikoff et al, 1992]. Ce qui fait qu'on peut trouver dans la liste des attributs générés par la méthode des Descripteurs Discri-

minants plusieurs motifs distincts mais qui sont similaires puisqu'ils peuvent se substituer entre-eux. De même lors de la construction du tableau binaire, on risque de perdre de l'information lorsque on marque par 0 l'absence d'un motif pouvant être substitué par un autre déjà présent.

La ressemblance entre les motifs est basée, comme déjà cité, sur la ressemblance des acides aminés qui les constituent. En effet, il existe différents degrés de similitude entre les acides aminés ; et puisqu'il existe 20 acides aminés, les possibilités de mutations entre eux sont mesurées par une matrice 20x20 appelée *matrice de substitution*.

### 4.2 Matrice de substitution des acides aminés

En bioinformatique, une matrice de substitution estime le taux auquel chaque résidu (acide aminé) dans une séquence change en un autre résidu dans le temps. Les matrices de substitution sont habituellement vues dans le contexte d'alignement des acides aminés, où la similitude entre les séquences dépend des taux de mutation comme représentés dans la matrice [Henikoff et al, 1992].

## 5 Extension du codage par les matrices de substitution

### 5.1 Terminologie

Soit $\mathcal{M}$ un ensemble de $n$ motifs, notés chacun par $\mathcal{M}[p]$, $p = 1..\ n$. $\mathcal{M}$ peut être divisé en $m$ groupes. Chaque groupe contient un *motif principal* $M^*$ et, probablement, d'autres motifs qui peuvent être substitués par $M^*$. Le *motif principal* est le motif ayant la plus grande probabilité, dans son groupe, de muter vers d'autres motifs. Pour un motif $M$ composé de $k$ acides aminés, cette probabilité, notée $P_m(M)$, est basée sur la probabilité $P_i$ ($i = 1..\ k$) que chaque acide aminé $M[i]$ de $M$ ne mute vers aucun autre acide aminé. Nous avons alors :

$$P_m = 1 - \prod_{i=1}^{k} P_i \ . \qquad (3)$$

Chaque $P_i$ est calculée en se basant sur la matrice de substitution utilisée selon la formule suivante :

$$P_i = S(M[i], M[i]) / \sum_{j=1}^{20} S^+(M[i], AA_j) \qquad (4)$$

$S(x, y)$ est le score de substitution de l'acide aminé $y$ par de l'acide aminé $x$ comme il est indiqué dans la matrice de substitution. $S^+(x, y)$ indique un score de substitution positif. Il est évident aussi que $S(x, x)$ est toujours positif. $AA_j$ est l'acide aminé d'indice $j$ parmi les 20 acides aminés

Nous considérons qu'un motif $M$ substitue un motif $M'$ si les conditions suivantes sont satisfaites:

- $M$ et $M'$ sont de même longueur $k$,
- $S(M[i], M'[i]) \geq 0$ , $i = 1..\ k$,
- $PS(M, M') \geq T$, où le seuil $T$ est un paramètre spécifié par l'utilisateur: $0 \leq T \leq 1$.

On note par $SP(M, M')$ la probabilité de substitution du motif $M'$ par le motif $M$ ayant la même longueur $k$. Il mesure la possibilité que $M$ mute vers $M'$ :

$$SP(M, M') = S_m(M, M') / S_m(M, M) . \quad (5)$$

$S_m(X, Y)$ est le score de substitution du motif $Y$ par le motif $X$. Il est calculé selon la formule suivante :

$$S_m(X, Y) = \sum_{i=1}^{k} S(X[i], Y[i]) \quad (6)$$

Il est clair, d'après les matrices de substitution, qu'il n'existe qu'un seul meilleur motif qui peut substituer un motif $M$, c'est évidemment lui-même, puisque les acides aminés qui le constituent sont mieux substitués par eux. Ceci montre que la probabilité de substitution d'un motif par un autre, s'ils vérifient les règles de substitution, est comprise entre 0 et 1.

## 5.2 Méthodologie

La modification de la méthode de Descripteurs Discriminants porte sur deux aspects. D'abord, le nombre des motifs extraits sera bien entendu réduit parce que nous allons garder un seul motif pour chaque groupe de motifs substituables de même longueur. Ensuite, nous allons modifier la règle de construction du tableau binaire mentionnée dans la section 2. En effet, nous allons marquer par 1 la présence du motif ou celle de l'un des motifs qu'il peut substituer. Le premier aspect peut être divisé en deux sous-parties : (1) identifier les motifs principaux des différents groupes (2) filtrer les motifs.

### 5.2.1 Identification des motifs principaux et filtrage

Le motif principal d'un groupe est le motif le plus susceptible, dans ce groupe, de muter vers d'autres motifs. Pour identifier tous les motifs principaux, nous trions $\mathcal{M}$ par ordre décroissant de la longueur du motif puis de la valeur de $P_m$. Pour chaque motif $M'$ de $\mathcal{M}$ nous cherchons le motif $M$ qui peut substituer $M'$ ayant la valeur de $P_m$ la plus élevée. Le regroupement est basé sur le calcul de la probabilité de substitution entre les motifs. Nous pouvons trouver un motif qui appartient à plus qu'un groupe. Dans ce cas, il doit être le motif principal de l'un d'eux.

Le filtrage consiste à conserver seulement les motifs principaux et à enlever tous les autres. Le résultat est un ensemble de motifs plus petit que l'ensemble initial mais qui peut représenter la même information de l'ensemble initial.

L'identification des motifs principaux et le filtrage sont effectués par l'algorithme simplifié suivant :

```
début
  M : ensemble de n motifs
  trier M par ordre décroissant de la longueur du motif;
  trier M par ordre décroissant de P_m;
  pour chaque motif M[i] de i=n à 1
    si P_m(M[i])=0 alors
      M[i] devient un motif principal;
```

```
    sinon
      x←position du premier motif de même longueur que 𝓜[i]
      pour chaque motif 𝓜[j] de j=x à i
        si 𝓜[j] substitue 𝓜[i] ou j=i alors
          𝓜[j] devient un motif principal;
          sortir pour
        fin si
      fin pour
    fin si
  fin pour
  pour chaque motif M de 𝓜
    si M n'est pas un motif principal alors
      supprimer M;
    fin si
  fin pour
fin.
```

La complexité en temps de cet algorithme est $O((n^2/2)*k)$ , avec $n$ est le nombre de motifs en question et $k$ est la longueur du plus grand motif.

**Exemple.** Etant donné la matrice de substitution BLOSUM62 et l'ensemble suivant (table 1) de motifs triés par leurs longueurs et par $P_m$, nous affectons chaque motif à un groupe représenté par son motif principal. Nous obtenons ainsi cinq groupes illustrés par le diagramme de la figure 2.

| $\mathcal{M}$ | LLK | IMK | VMK | GGP | RI | RV | RF | RA | PP |
|---|---|---|---|---|---|---|---|---|---|
| $P_m$ | 0.89 | 0.87 | 0.86 | 0 | 0.75 | 0.72 | 0.72 | 0.5 | 0 |
| Motif principal | LLK | LLK | LLK | GGP | RI | RI | RI | RV | PP |

TAB.1 – *Regroupement des motifs. La troisième ligne montre les motifs principaux.*

### 5.2.2 Construction du tableau binaire

Dans la phase de construction du tableau binaire, nous comparons chaque motif avec les k-grammes (k est la taille du motif) de chaque séquence jusqu'à trouver un substitut et marquer 1 ou parcourir toute la séquence sans trouver de substitut et marquer 0. Nous utilisons l'algorithme suivant :

```
début
  pour chaque séquence S
    pour chaque motif M de longueur k
      répéter
        extraire un k-gramme M'de S;
        si M substitue M'alors
          noter 1 dans le contexte pour S et M;
          aller à présence;
        fin si
      jusqu'à la fin de S
      noter 0 dans le contexte pour S et M;
      présence: continuer
    fin pour
  fin pour
fin.
```

La complexité en temps de cet algorithme est $O(m*n*l*k)$, avec $n$ est le nombre de motifs en question, $k$ est la longueur du plus grand motif, $m$ est le nombre des séquences et $l$ est la taille maximale d'une séquence.

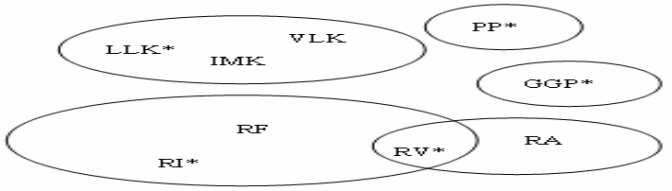


FIG.2 – *Regroupement des motifs. Le motif RV appartient à 2 groupes. Il est le motif principal de l'un d'eux.*

# 6 Expérimentations

Les méthodes de codage sont implémentées en langage C et regroupées dans une DLL. Les fichiers d'entrée sont des fichiers de séquences sous le format *FASTA*. La DLL génère des fichiers relationnels sous différents formats comme le format *ARFF* utilisé par la boite à outils WEKA [Witten et al, 2005] et le format DAT utilisé par le system DisClass [Maddouri et al, 2002].

## 6.1 Données expérimentales

Pour comparer les différentes méthodes de codage nous utilisons trois échantillons de données protéiques et nucléiques décrits par la table 2.

| Types | Echantillon | Familles/classes | Nombre | Source |
|---|---|---|---|---|
| Protéiques | Echantillon E1 | High-potential Iron-Sulfur Protein | 19 | SWISS-PROT |
| | | Hydrogenase Nickel Incorporation Protein HypA | 20 | |
| | | Hlycine Dehydrogenase | 21 | |
| | Echantillon E2 | TLR humaine | 14 | Entrepôt de l'université IRVINE |
| | | TLR non humaine | 26 | |
| Nucléiques | Echantillon E3 | Site promoteur | 53 | |
| | | Site non promoteur | 53 | |

TAB. 2 – *Données d'expériences*

L'échantillon E1 contient trois familles protéiques distinctes et distantes. Nous supposons que la classification dans ce cas sera relativement facile puisque chaque famille aura probablement des motifs conservés qui sont différents de ceux des autres familles [Nevill-Manning et al, 1998]. Cependant, l'échantillon E2 présente un problème de classification plus délicat.

Il s'agit de distinguer entre les séquences protéiques *Toll-like Receptors* (TLR) humaines des celles non humaines. La difficulté est due à la ressemblance structurelle et fonctionnelle de deux groupes. Nous examinons si la méthode de codage aide à améliorer la classification. L'échantillon E3 fait l'objet d'un problème de classification typique. Il s'agit de reconnaître les séquences nucléiques porteuses des sites promoteurs des celles qui ne le sont pas. Les promoteurs sont des courts segments d'ADN qui précèdent le début des gènes dont l'identification facilite la localisation des gènes dans les séquences nucléiques.

## 6.2 Processus expérimental

Dans nos expériences, nous utilisons la technique de validation croisée d'ordre 10 [Han et al, 2001]. Chacun des échantillons de données est aléatoirement et équitablement partitionné en 10 sous-ensembles mutuellement exclusifs. L'apprentissage et le test sont exécutés 10 fois. A chaque itération, un sous-ensemble est réservé au test et les autres sont utilisés ensemble pour l'apprentissage. Après avoir prétraité les séquences biologiques et construit les tableaux binaires d'apprentissage TBA et de test TBT, nous entamons l'étape de classification. En se servant du classifieur C4.5 de l'environnement WEKA [Witten et al, 2005], nous générons les règles de classifications à partir de TBA que nous testons sur TBT. La précision sera calculée comme étant la moyenne des précisions des 10 itérations. Le processus expérimental et de codage est illustré par la figure 3.

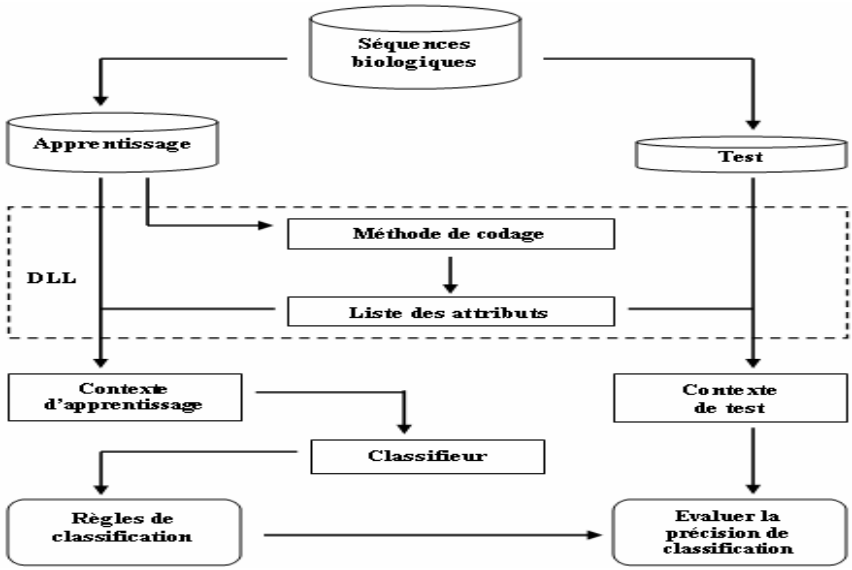


FIG. 3 – *Processus expérimental*

### 6.3 Résultats

Nous examinons, d'abord, chaque méthode à part en faisant varier ses différents paramètres pour rechercher leurs valeurs optimales (table 3). Puis, nous utilisons les meilleurs paramètres trouvés pour les comparer (table 4) en termes de précision (pourcentage des séquences correctement classées), nombre d'attributs et temps de calcul. Pour la méthode DDMS, nous avons testé les matrices de substitution BLOSUM62 et PAM250.

| Méthode | Paramètres | Echantillon | | |
|---|---|---|---|---|
| | | E1 | E2 | E3 |
| NG | N | 3 | 3 | 4 |
| MA | Long min | 3 | 3 | 3 |
| | Activité min | 50 % | 50 % | 25 % |
| DD | Alpha | 0 | 0 | 0 |
| | Beta | 0 | 0 | 0 |
| DDMS | Alpha | 0 | 0 | - |
| | Beta | 0 | 0 | - |
| | Matrice de substitution | BLOSUM62 | BLOSUM62 | - |
| | Seuil | 0.7 | 0.9 | - |

TAB. 3 – *Meilleurs paramètres utilisés pour l'expérimentation.*

| Données | Critères \ Méthodes | NG | AM | DD | DDSM |
|---|---|---|---|---|---|
| E1 | Précision | 90 % | 95 % | 95 % | 98.33 % |
| | Nb. d'attributs | 4777 | 1978 | 4709 | 2139 |
| | Temps de calcul (s) | 0.82 | 38 | 35 | 37 |
| E2 | Précision | 60 % | 55 % | 67.5 % | 77.5 % |
| | Nb. d'attributs | 5340 | 3458 | 6839 | 6562 |
| | Temps de calcul (s) | 1 | 91 | 921 | 954 |
| E3 | Précision | 73.58 % | 77.78 % | 77.78 % | - |
| | Nb. d'attributs | 244 | 314 | 701 | - |
| | Temps de calcul (s) | 0.05 | 2 | 1.57 | - |

TAB. 4 – *Résultats expérimentaux*

## 7 Discussion

D'après l'étude expérimentale, nous avons remarqué qu'il n'existe pas de valeurs optimales et uniques pour les paramètres des méthodes étudiées. En fait, ces valeurs dépendent de la nature des données en question. Ce qui fait que le réglage des paramètres nécessite une connaissance préalable des caractéristiques des données comme les longueurs des régions conservées ou le taux de mutation entre les séquences d'une famille.

Les résultats expérimentaux varient selon les données entrées. Pour l'échantillon E1, la classification était relativement facile puisque les trois familles de protéines sont complètement distinctes. Chacune d'elles a probablement ses propres motifs qui la caractérisent et la discriminent des autres. Ceci explique les précisions élevées pour toutes les méthodes avec un avantage pour la méthode des Descripteurs Discriminants avec Matrice de Substitution qui a permis d'atteindre une précision très élevée.

Les méthodes des Motifs Actifs et des Descripteurs Discriminants ont permis les meilleures précisions pour l'échantillon E3 (cet échantillon ne concerne pas notre extension puisqu'il contient des séquences nucléiques). L'échantillon E2 représente un vrai défi de classification puisque les TLR humaines et les TLR non humaines se ressemblent du point de la vue de la fonction et de la structure. En effet les deux classes partagent beaucoup de parties similaires ce qui explique la faible précision avec la méthode des Motifs Actifs. En effet, cette méthode, qui se base sur la fréquence des motifs pour les extraire, a construit des attributs qui appartiennent à la fois aux deux classes, ce qui augmente la possibilité de confusion. La méthode des N-Grammes a permis une meilleure précision, mais n'atteint pas la précision acceptable par défaut qui est 65% (si on affecte toutes les séquences à la classe des TLR non humaines).

La méthode des Descripteurs Discriminants a montré des résultats meilleurs que ceux de deux méthodes précédentes et elle a permis une meilleure distinction entre les TLR humaines et les TLR non humaines. Mais, pour améliorer encore la classification dans l'échantillon E2, il faut prendre en compte le phénomène de mutation des acides aminés. En effet, notre méthode DDSM a permis d'atteindre la précision la plus élevée avec la matrice de substitution BLOSUM62 tout en réduisant le nombre d'attributs générés.

# 8 Conclusion

Dans ce travail, nous avons présenté le codage des séquences biologiques comme étant une étape de prétraitement pour la classification supervisée. Nous avons décrit trois méthodes existantes qui sont les méthodes des N-Grammes (NG), des Motifs Actifs et des Descripteurs Discriminants (DD). Puis, nous avons proposé une amélioration de la méthode DD par l'utilisation d'une matrice de substitution.

Dans le but d'examiner l'effet de chaque méthode de codage sur la précision de classification, nous avons mené une étude expérimentale qui porte sur des données biologiques variées comportant des séquences protéiques et nucléiques. Nous avons aussi comparé les nombres d'attributs générés par ces méthodes et leurs temps de construction. Parmi les méthodes existantes, nous avons constaté que la méthode DD présente la meilleure précision. L'extension de cette méthode par l'utilisation d'une matrice de substitution a permis d'améliorer la précision de classification même dans un cas de classification relativement délicat. Cependant, nous avons noté que la méthode des Descripteurs Discriminants, ainsi que sa variante avec la matrice de substitution, sont les plus coûteuses en terme de temps de construction d'attributs surtout par rapport à la méthode NG.

En considérant le présent travail, plusieurs voies sont ouvertes. Il sera intéressant de concevoir une méthode hybride basée sur les N-grammes et qui utilise des filtres comme $\alpha$ et $\beta$, tout en prenant en compte la substitution et l'ordre des motifs extraits dans les séquences. Nous essayerons d'étendre notre approche pour traiter les séquences nucléiques par l'utilisation des tables de scores des bases d'ADN.

## Références

Altschul S. F., W. Gish, W. Miller, E. W. Myers, D. J. Lipman. (1990) *Basic local alignment search tool*. Journal of Molecular Biology, Vol. 215(3), pp. 403-413.

Han J., M. Kamber. (2001) *Data Mining: Concepts and Techniques*. ISBN 1-55860-489-8. Morgan Kaufmann Publishers: www.mkp.com.

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

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

Leslie, C., E. Eskin, et W. S. Noble. (2002) *The spectrum kernel: a string kernel for svm protein classification*. Pac Symp Biocomput, 564–575.

Maddouri M, Elloumi M. (2002) *A data mining approach based on machine learning techniques to classify biological sequences*. Knowledge Based Systems Journal.

Maddouri M. & M. Elloumi. (2004) *Encoding of primary structures of biological macromolecules within a data mining perspective*. Journal of Computer Science and Technology (JCST); VOL 19, num 1. Allerton Press: 78-88, USA.

Miller E., Shen D., Liu J. & Nicholas C. (1999) *Performance and scalability of a large-scale N-gram Based Information Retrieval System*. Journal of digital information.

National Human Genome Research Institute (June 2006). National Institute of Health. Available: http://www.nhgri.nih.gov.

Nevill-Manning, C. G., Wu, T. D., and Brutlag, D. L. (1998) *Highly specific protein sequence motifs for genome analysis*. Proceedings of the National Academy of Sciences of the United States of America, 95(11):5865-5871.

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

Witten I. H. & Eibe F. (2005) *Data Mining: Practical machine learning tools and techniques*. 2nd Edition. Morgan Kaufmann, San Francisco.

## Summary

The classification of biological sequences is one of the significant challenges in bioinformatics. However, the representation of this kind of data by strings of characters does not allow its processing by the standard classification tools, which often use the relational format. To remedy this problem of encoding, several works are based on the motifs extraction to build a new representation of the biological sequences under the form of a binary table. We describe a new approach which extends the previous methods by using substitution matrices in the case of the protein sequences. We present then a comparative study which takes into account the effect of each method on different criteria.