

BIF7000  
Introduction à la bioinformatique

Travail présenté à :  
Mme Anne Bergeron

Identification bayésienne des domaines conservés des  
protéines nucléolaires chez la levure *Saccharomyces*  
*cerevisiae*

par

Korka Ba  
BAXK03036302

Yannick Gingras  
GINY16027805

Philippe Simard  
SIMP12078308

Université du Québec à Montréal  
28 novembre 2005

## **Table des matières**

1	<b>Introduction</b>	3
1.1	<i>Nucléole et protéines nucléolaires</i>	3
1.2	<i>Protéines nucléolaire et maladies associées</i>	5
1.3	<i>La levure Saccharomyces cerevisiae comme modèle</i>	8
1.4	<i>Théorie bayésienne</i>	8
2	<b>Matériel et Méthodes</b>	12
2.1	<i>Récupération des données</i>	12
2.2	<i>Sélection des protéines nucléolaires et essentielles non connues</i>	12
2.3	<i>Création de jeu de donnée pour la sélection des protéines candidates</i>	12
2.4	<i>Préparation des données</i>	13
2.5	<i>Entraînement</i>	13
2.6	<i>Recherche de séquences intéressantes</i>	15
3	<b>Résultats et Discussion</b>	16
3.1	<i>Longueur des mots et pondération</i>	16
3.2	<i>Analyse des mots significatifs</i>	17
4	<b>Travaux futurs</b>	19
5	<b>Références</b>	20
6	<b>Annexes</b>	23
6.1	<i>Figures</i>	23

# 1 Introduction

## 1.1 Nucléole et protéines nucléolaires

Le séquençage complet de plusieurs organismes stimule grandement les recherches sur la protéomique par les industries biotechnologiques. À cause de leur non toxicité et de leur biodégradabilité, les protéines, en particulier les enzymes, sont de plus en plus utilisées dans les recherches biomédicales et pharmaceutiques.

Pour cela, il serait important de mieux comprendre les conformations des protéines et/ou leurs interactions qui sont généralement commandés par des unités fonctionnelles ou domaines. Selon leurs environnements, ces domaines pourraient influencer la stabilité des protéines ainsi que leurs interactions. Dans bien des cas, ces domaines joueraient un rôle particulier dans la classification des espèces selon leurs groupes (eucaryote, procaryote), embranchement, classe, famille, etc.

Chez les eucaryotes, la transcription des gènes de l'ARN ribosomique (ARNr) et la biosynthèse des grande et petite sous-unités ribosomiques ont lieu à l'intérieur du nucléole (Olson et al., 2000 ; Scheer et Hock, 1999). Ces gènes sont des unités répétées organisées en NOR (Nucleolar Organizer Region) localisés dans des zones de constriction bien visibles sur les chromosomes métaphasiques. Ces NOR seraient le centre de formation du nucléole (Shaw, 2001). Les gènes de l'ARNr sont transcrits par les ARN pol I et pol III sous forme de pré-ARNr 35S et pré-ARNr 5S, respectivement. Ces activités joueraient un rôle majeur pendant la croissance cellulaire et représenteraient près de 95% du total des transcriptions avec une consommation d'énergie supérieure à la moitié de l'énergie cellulaire. De plus, ces événements nécessitent une coordination de plusieurs processus biologiques comme la maturation des ARNr, l'assemblage de complexes ARN-protéines et/ou ARN-ARN.

Durant sa maturation, le 35S s'associerait à environ 200 facteurs pour former le pré-ARNr 90S dont la maturation entraînera la formation de 43S et 66S. Ces derniers évolueront respectivement en sous-unités 40S et 60S pour former un ribosome fonctionnel dans le cytoplasme (Fig.1 en annexe). Celui-ci constituera la machinerie centrale de la synthèse des protéines.

L'analyse de ces facteurs montre la présence de protéines ribosomiques, de petits ARN nucléolaires (snoRNA), des protéines nucléolaires, dont des endo et exo-nucléases, et un grand nombre de RNA-hélicases. Des études protéomiques ont montrées que plusieurs protéines nucléolaires présentent des domaines d'interaction protéine-protéine, des domaines WD40 et des domaines «HEAT repeat» (Dragon et al., 2002 ; Lee et Baserga, 1999).

Les domaines d'interaction protéine-protéine ou domaine « coiled-coil » sont caractérisés par une répétition d'une séquence consensus de sept résidus (abcdefg)<sub>n</sub>. Le premier et le quatrième résidu (a et d) sont généralement hydrophobes et non polaires. Par contre, les résidus en position e et g sont polaires ou chargés. Ces domaines impliqueraient deux ou cinq hélices alpha d'une même ou de protéines différentes. Chez les animaux et la levure les domaines coiled-coil ont été identifiés dans plusieurs protéines qui joueraient un rôle important dans l'association des protéines fonctionnelles avec les composantes cellulaires. Ces séquences autonomes présenteraient les mêmes interactions spécifiques aussi bien *in vitro* que *in vivo*. Par conséquent, elles seraient déterminantes dans l'identification des interactions protéine-protéine dans le système double hybride (Newman et al., 2000). Les prédictions de ces domaines permettraient une meilleure optimisation des recherches protéomiques.

Les RNA hélicases sont un autre groupe de protéines habituellement rencontrées dans le nucléole. Ces protéines ubiquitaires pouvant être subdivisées en deux groupes selon l'ordre de leurs résidus : la Superfamille 1 (SF1) et la Superfamille 2 (SF2). Les hélicases, membres de la SF2, sont les plus répandues dans les êtres vivants et sont également subdivisées en familles suivant la composition de leurs motifs conservés dans l'évolution. Chez *Saccharomyces cerevisiae*, les membres des familles «DEAD box» et « DEAH box» sont les plus représentés et sont majoritairement impliqués dans la biogenèse des ribosomes et dans l'épissage des pré-ARNm respectivement.

Les protéines de la famille «DEAD box» sont caractérisées par la présence d'une séquence interne catalytique renfermant neuf motifs (Fig.2 en annexe) incluant les acides aminés D, E, A et D tous bien conservés dans l'évolution (Rocak et Linder, 2004 ; Tanner et al., 2003). Les motifs sont encadrés par une extension N- et C-terminale faiblement conservée. Ces extensions joueraient un rôle fondamental dans la spécificité des ARN hélicases au moment de la reconnaissance de leurs substrats (de la Cruz et al. 1999). Les motifs I et II, ou motifs de Walker A et B, seraient des sites de fixation de l'ATP. Les motifs Q et VI favoriseraient l'hydrolyse de l'ATP en se fixant sur celui-ci. En effet, le motif Q serait caractérisé par une courte séquence de 9 résidus situés en amont du motif I dont le rôle serait prépondérant dans l'activité des enzymes (Tanner et al., 2003). Cependant, les motifs (Ia, Ib, IV, et V) et III seraient fortement impliqués dans la liaison à l'ARN et le changement de conformation de l'hélicase, respectivement (Tanner et Linder, 2001 ; Tanner et al., 2003).

Les «DEAD box» sont des protéines ubiquitaires qui seraient impliquées dans plusieurs processus métaboliques des ARN en particulier la maturation des ARNr, l'assemblage des ribosomes, l'épissage des ARNm, ainsi que la propagation virale. Par

conséquent, l'absence ou une déplétion de ces hélicases conduirait à la dégradation ou à l'accumulation des ARNr précurseurs intermédiaires, respectivement (Kressler et al., 1998 ; de la Cruz et al., 1999). Grâce à l'énergie générée par l'hydrolyse de leur ATP, ces protéines seraient en mesure de défaire des duplex ARN-ARN, ou encore dissocier des complexes ARN-protéines (de la Cruz et al., 1999 ; Tanner et Linder, 2001 ; Lorsch, 2002). De plus, ces hélicases joueraient un rôle important dans l'appariement de certaines paires de bases et la modulation de certaines conformations structurales des molécules d'ARN.

Tout récemment, il a été démontré qu'il existerait une relation étroite entre certaines «DEAD-box» (DDX1 par exemple) et la réplication du VIH-1 (Fang et al., 2004). Dbp2p est une autre «DEAD-box» qui serait impliquée dans la biogenèse des ribosomes et la dégradation de l'ARNm dans le sens (3'-5'). Curieusement, sur des souches ayant une délétion de DBP2, la fonction de la biogenèse des ribosomes est complétée par p68, l'homologue humain de DBP2 (Bond et al., 2001). Ainsi, dans bien des cas, p68 agirait comme un répresseur de la transcription (Wilson et al., 2004).

Dans le but de déterminer la présence d'un unique signal de localisation nucléolaire, Zachman et Nigg (1993) ont analysés la nucléoline, une des protéines majoritaires du nucléole. Contre toute attente, leurs travaux ont montré la présence de deux signaux : un signal NLS (pour Nucleus Localisation Signal) (KRKKEMANKSAPEAKKKK) et un domaine riche en RG spécifique à la liaison à l'ARN. Malgré tout, l'identification d'un unique signal de localisation nucléolaire reste toujours sans réponse.

## **1.2 Protéines nucléolaire et maladies associées**

Plusieurs maladies sont associées aux protéines nucléolaires. La plupart sont causées par des mutations dans la séquence codante. Ces mutations peuvent être des insertions ou des délétions qui peuvent induire l'apparition d'un codon stop. Ces mutations ont pour résultats de modifier le cadre de lecture, allant même jusqu'à la non transcription d'un exon complet. Bien évidemment, ceci cause des torts irréparables aux protéines qui sont traduites à partir de ces gènes, ce qui entraîne l'apparition de maladies. Bien qu'il existe beaucoup de maladies associées aux protéines nucléolaires, nous allons expliquer les causes et effets de seulement deux d'entre-elles : les syndromes de Werner et de Treacher Collins.

Premièrement, le syndrome de Werner est une maladie autosomique, c'est-à-dire qui n'est pas relié au sexe, récessive très rare qui est principalement caractérisée par l'apparition prématuré des signes associés au vieillissement et la prédisposition au cancer.

La cause de cette maladie est le gène WRN qui est un homologue des hélicases Blm humaine (Marcinaik et al., 1998).

Les personnes atteintes de ce syndrome se développent normalement jusqu'à la fin de leur première décennie. Par la suite, elles cessent de grandir, perdent leurs cheveux, ou bien, ils grisonnent. Dans la trentaine surviennent l'apparition de cataracte et autres signes du vieillissement. La longévité moyenne des personnes atteintes du syndrome de Werner est de seulement 48 ans.



Jeune Américano-asiatique atteinte du syndrome de Werner [3]

Comme dit précédemment, la cause de cette maladie est le gène WRN. Ainsi, le gène normal (non-muté) consiste en un gène de 35 exons codant pour une protéine multifonctionnelle de 1432 acides aminés (Yu et al., 1996) qui est un membre de la famille RecQ des hélicases à ADN. Cependant, il existe plusieurs variantes d'allèle codant pour la pathologie. En fait, le nombre de variants identifiés est de 35 (Moser et al., 1999) principalement causé par des délétions et des insertions. La variante la plus commune est attribuable à la mutation ponctuelle de la base 1336 qui passe d'un C à un T. Cette variante compte pour 20 à 25% des mutations chez les sujets caucasiens. Le produit de ce gène est donc anormal. En effet, en plus de perdre son signal de localisation nucléaire (Moser et al., 1999), les ARNm et les protéines « mutantes » en résultant possèdent un temps de demi-vie plus court que les protéines « normales ». De plus, plusieurs études tendent à indiquer que la protéine WRN serait impliquée dans la transcription des ARNr chez l'humain et que la diminution de la transcription par l'ARN polymérase 1 pourrait être un facteur du vieillissement prématuré des personnes atteintes (Shiratori et al., 2002). Par contre, puisque le gène WRN est seul le gène connu causant cette maladie, les mutations sur sa séquence codante peuvent être identifiées.

Deuxièmement, le syndrome de Treacher Collins est une autre maladie associée aux protéines nucléolaires. Tout comme le syndrome de Werner, elle est aussi une maladie autosomique récessive rare qui affecte le développement du crâne et de la face. Ce syndrome est causé par la mutation du gène TCOF1 qui code pour la protéine Treacle (Dixon et al., 1997). Cette protéine est localisée dans le composant fibrillaire dense du nucléole.

Les principales caractéristiques de ce syndrome sont entre autres : l'hypoplasie (réduction du volume d'un organe causée par une diminution du nombre des cellules dont sont constitués ses tissus) de l'os zygomatique et de la mandibule, d'un colobome (anomalie caractérisée par l'absence d'une portion de la structure de l'oeil) et chez 40 à 50% des sujets une perte auditive attribuée à une malformation des osselets.



Personne souffrant du syndrome de Treacher Collins [4]

Le gène TCOF1 a été isolé pour la première fois en 1996 par le Treacher Collins Syndrome Collaborative Group et contient 25 exons. Le transcrit « normal » est la protéine Treacle qui possède un poids moléculaire de 144kD et contient 4233 nucléotides, ou 1411 acides aminés. Cette protéine nucléolaire contient trois domaines (N-terminal, C-terminal et une 10 répétitions d'un motifs contenant la protéine kinase C et le site de phosphorylation de la caséine kinase 2 (Isaac et al., 2000)). De plus, cette protéine possède deux sites codant pour le signal de localisation nucléaire (Dixon, Hovanes et al., 1997). Elle est en plus impliquée dans la transcription des ADNr et de la biogenèse des ribosomes (Winokur et Shlang, 1998). Cependant, il existe des centaines de mutations documentées sur le gène TCOF1 causant la maladie (Gladwin et al., 1996; TCSCG, 1996). La majorité des mutations sont des insertions et des délétions causant des erreurs du cadre de lecture et induisant un codon stop. Ce qui en résulte que la protéine Treacle muté n'est plus nucléolaire. Par contre, un test prénatal permet de détecter la maladie. Il existe deux méthodes. En premier lieu, la méthode moléculaire consiste à vérifier sur l'ADN extrait des cellules (obtenue par amniocentèse) la présence d'une mutation sur le gène TCOF1. Cependant, même s'il y a présence d'une mutation, on ne peut prédire le type ou la sévérité des effets. En second lieu, l'examen par ultrasons, chez les grossesses à risques pour ce syndrome, afin de détecter des problèmes tel qu'une anomalie au niveau des structures de la face du fœtus (Milligan et al., 1994) peut être effectué afin de déceler la présence de ce syndrome.

Les maladies qui sont associées aux protéines nucléolaires sont donc très graves. Il est donc très important de mieux comprendre le fonctionnement du nucléole, ces protéines et leur implication dans la cellule afin de pouvoir aider les personnes souffrant de ces maladies.

### 1.3 La levure *Saccharomyces cerevisiae* comme modèle

Afin de trouver les domaines conservés des protéines nucléolaires, nous utiliserons la levure *Saccharomyces cerevisiae* comme modèle. Nous avons choisi cet organisme en particulier pour plusieurs raisons. Premièrement, son génome a été complètement séquencé et les outils pour la recherche sur son génome sont connus et bien maîtrisés. De plus, le site internet *Saccharomyces genome database* [1] est complètement dédié à cet organisme et la grande majorité des informations et outils de recherche génétiques sont accessibles. Deuxièmement, son nucléole ne se désorganise pas durant sa mitose (Carmo-Fonseca et al., 2000). En ne se désorganisant pas, le risque de mutation est grandement diminué. Finalement, la raison principale pour ce choix de modèle est que cet organisme est un eucaryote, il possède un noyau. Ainsi, puisque l'homme et la levure sont deux eucaryotes, le nucléoles et ses composantes (protéines, etc.) sont donc « comparables ». Donc, cet organisme a été choisi à cause de son type de cellule, la non désorganisation de son nucléole et à cause de la maîtrise de ses outils génétiques.

### 1.4 Théorie bayésienne

L'adjectif "bayésien" désigne ce qui se rapporte aux travaux du révérend Thomas Bayes [1702 - 1761]. Bayes a principalement travaillé dans le domaine des probabilités, il est connu pour le théorème dit "de Bayes" qui permet l'inversion des probabilités conditionnelles :

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

On désigne, pour une raison obscure, un système qui combine plusieurs probabilités conditionnelles d'évidences indépendantes par "classification bayésienne naïve". Formellement, on dit que la probabilité d'un évènement  $E$  sachant que plusieurs évidences  $e_1, e_2, \dots, e_n$  se sont produites est donnée par :

$$P(E|e_1, e_2, \dots, e_n) = \frac{\prod P(E|e_i)}{\prod P(E|e_i) + \prod (1 - P(E|e_i))}$$

Par exemple, si on sait qu'une personne mesurant plus de 2 mètres a 60% de chance d'être un joueur basket-ball et qu'une personne qui porte un ballon de basket a 72 % de chance d'être un joueur de basketball, on peut affirmer que la probabilité combiné est

$$P(E|e_1, e_2) = \frac{.6 \cdot .72}{.6 \cdot .72 + (1 - .6)(1 - .72)} \cong .7941$$

En théorie, la classification bayésienne naïve ne devrait pas fonctionner car il est très rare qu'on puisse accumuler plusieurs évidences parfaitement indépendantes. Pour tenir compte du fait que les évidences ne sont pas parfaitement indépendantes, il faudrait calculer les probabilités que chaque paire d'évidence se produisent simultanément, de même que chaque triplet, que chaque quadruplets, ... que la probabilité de toutes les évidences simultanément ce qui ramène à calculer les toutes les parties de l'ensemble des évidences. Pour  $n$  évidences, il y a en  $2^n$ .

L'avantage principal de la classification bayésienne naïve est que la formule est très simple à appliquer et très performante à calculer, en  $O(n)$ . Avec la classification bayésienne naïve, on sait ce qu'on calcule : une probabilité. On comprend bien que 50% de probabilités pour une évidence représente une absence de corrélation, que 99% représente une corrélation forte et que 30% représente une corrélation inverse plutôt faible. À l'opposé, on sait qu'avec BLAST on doit rechercher les alignements avec une "e-value" basse. Est-ce que 0.5 est assez bas ? Est-ce  $10^{-10}$  est un milliard de fois meilleur que 0.1 ? Et encore, la e-value a une définition probabiliste qu'on comprend bien : le nombre d'alignement qu'on peut espérer du au hasard. Si on parle de "score", on sombre rapidement dans l'incertitude la plus complètes. Un score de 200 est probablement moins bon qu'un score de 2000, mais où se situe la limite entre un bon est un mauvais score ?

Pour construire un classificateur bayésien naïf, on accumule des évidences et on leur assigne une probabilité. Cette table de la corrélation respective de chaque évidence est ensuite facile à consulter. On peut en tirer des information intéressantes, par exemple, trouver des choses qu'on en commun les évidences qui permettent une forte corrélation.

En 2001, Paul Graham publiait des résultats surprenant et très encourageants : la classification bayésienne naïve permettait de repérer 99.5 % du "spam", avec 0 faux positifs [Graham 01]. A première vue, cette approche est plutôt choquante : assumer que les mots d'un texte sont indépendants nous éloigne sans aucun doute de la réalité. Rappelons qu'on parle ici d'indépendance statistique, c'est à dire qu'on considère des événements indépendantes si leur probabilité conjointe est égale au produit de leur probabilités respectives :

$$P(EF) = P(E)P(F)$$

Un exemple classique est le jet de deux dés. Si on cherche la probabilité d’avoir une somme de 7, les 2 dés sont indépendants. Sinon, ils ne le sont pas.

Pour classifier les courriels entre spam et désirable, Paul Graham a considéré chaque mots comme une évidence. Avec un jeux de donné déjà catégorisé en deux classe, il est facile de calculer la corrélation de la présence d’un mot. La probabilité qu’un courriel soit un spam sachant qu’il contient un mot  $\omega$  est :

$$P(S|\omega) = \frac{\frac{b}{|B|}}{\frac{b}{|B|} + \frac{g}{|G|}}$$

où  $b$  est le nombre d’occurrence de  $\omega$  dans le multi-ensemble  $B$  des mots qu’on retrouve dans les spams. C’est-à-dire, sa fréquence chez les spams, divisé par la somme de sa fréquence chez les spams et de sa fréquence chez les courriels désirables. On voit facilement que c’est bien une probabilité :  $P(S|\omega) \in [0..1]$ . On se souvient que pour combiner les probabilités de  $n$  évidences dépendantes, on doit calculer  $2^n$  sous ensembles. Dans un courriel, il est commun d’avoir plus de 300 mots et on comprend que combiner la probabilité de chaque mot en tenant compte du fait qu’ils ne sont pas indépendants demande un calcul bien au-delà des capacités de tout ordinateur. Pour se donner un ordre de grandeur, il y a environ  $2^{270}$  particules élémentaires dans l’univers visible.

Explorons un exemple pour voir comment la classification bayésienne naïve permet de classer un texte. Imaginons un courriel contenant les mots “sex”, “montreal”, “tonight” et “security”. Avec un jeux d’entraînement on peut extraire les  $P(S|\omega)$  pour chacun de ces mots qui sont présentés dans le tableau 1.

TAB. 1: Probabilités conditionelles

mot	nb bon	nb spam	$P(S \omega)$
sex	3	304	.95
montreal	914	293	.09
tonight	19	48	.45
security	252	2202	.74

On voit que “sex” est un mot fort incriminant, que “montreal” est un mot plutôt favorable, que “tonight” plutôt neutre et que “security” n’est que légèrement incriminant. En combinant ces probabilités, on trouve 0.813983.

$$P(E|e_1, \dots, e_4) = \frac{.95 \cdot .09 \cdot .45 \cdot .74}{.95 \cdot .09 \cdot .45 \cdot .74 + (1 - .95)(1 - .09)(1 - .45)(1 - .74)} \cong .814$$

Ce qui nous permet d'affirmer qu'on a affaire à un spam. Par contre, si on ajoute un seul mot, "yannick", le prénom de l'individu nous ayant fourni les jeux d'entraînements, on arrive à une conclusion toute autre. Le dit mot apparaît 2571 fois dans les courriels désirables, mais seulement 3 fois dans les spams. Ceci lui donne une probabilité de .01. Évidemment, il est plus commun qu'un courriel désirable soit adressé à la bonne personne. En combinant les probabilités des mots dans ce nouveau courriel, on trouve 0.0423 ce qui nous permet d'affirmer avec 95.8% de certitude que ce courriel n'est pas un spam.

Ce résultat est fort intéressant. La nature adaptative du classificateur bayésien naïf permet de détecter par entraînement des mots qui d'apparence anodine sont en fait de très bons indices pour deviner la nature d'un texte. En effet, un filtre traditionnel basé sur des règles "codées en dure" reconnaîtra sûrement un mot comme "sex" mais il est hélas peu probable qu'il reconnaisse la forte corrélation entre bon courriel et les mots "montreal" et "yannick".

Un autre avantage de la classification bayésienne naïve est sa nature incrémentale. Au fur et à mesure qu'on connaît des évidences aidant la classification, on peut les ajouter au calcul pour affiner nos prédictions. Dans le cas des filtres à spam, le calcul de probabilité de chaque mot est lui aussi incrémental. Pour chaque nouveau courriel reçu, on peut mettre rapidement à jours les tables de fréquences des mots qu'il contient. De plus, l'approche des fréquences permet l'entraînement décremental. Si l'inspection manuelle repère un courriel mal classé, on peut soustraire les occurrences des mots qu'il contient de la table des fréquences de l'ancienne classe et les ajouter à la table de la nouvelle classe sans devoir refaire un entraînement complet du filtre.

Donc, le but de ce travail est de localiser les domaines nucléolaires chez la levure *Saccharomyces cerevisiae* à l'aide d'un filtre probabiliste bayésien. Pour ce faire, nous avons plusieurs sous objectifs à atteindre. Premièrement, nous devons créer un jeu de données contenant les séquences FASTA en nucléotides et en acides aminés de protéines nucléolaires et non-nucléolaires chez la levure et ensuite chez l'homme. Deuxièmement, nous allons utiliser un filtre probabiliste bayésien afin de classer une séquence selon sa localisation. Finalement, nous allons rechercher les séquences nucléolaires bien classées les plus fréquentes.

## **2 Matériel et Méthodes**

### **2.1 Récupération des données**

Afin d'utiliser un filtre bayésien, on doit avant tout lui fournir un jeu d'entraînement pour qu'il puisse calculer une table de probabilités. C'est pourquoi nous sommes partis à la chasse aux séquences dont la localisation est connue.

La construction du jeu de données a soulevé certains problèmes. La localisation d'une protéine n'est pas toujours connue. Souvent, cette information est disponible dans un format arbitraire et peu standardisé. Tous les services du NCBI qui sont disponibles aux usagers via une interface web le sont aussi aux programmes et aux scripts via une API ("application programming interface") simple et bien documenté. Il est donc facile d'automatiser la récupération d'un jeu de données à partir du NCBI. Hélas, la localisation des protéines n'est pas un champ indexé au NCBI.

Pour construire notre jeu de données, il a fallu utiliser des bases de données alternatives souvent spécialisée pour une espèce en particulier. Ces bases de données offrent souvent une interface web très conviviale, mais rarement l'équivalent de l'API du NCBI. L'essentiel du code écrit dans le cadre de ce projet sert à faire l'extraction des données de ces bases de données et à inférer les références croisées qui éventuellement, mènent au NCBI. Cette extraction est principalement faite en simulant une requête web et en utilisant des expressions régulières pour localiser les liens dans la page reçue.

### **2.2 Sélection des protéines nucléolaires et essentielles non connues**

Les différentes protéines de *Saccharomyces cerevisiae* ont été sélectionnées dans les bases de données du génome de *Saccharomyces* (SGD) (<http://www.Yeastgenome.org>) et dans MIPS (pour Munich Information center for Protein Sequence). Leur localisation a été déterminé à partir du lien <http://yeastgfp.ucsf.edu/getOrf.php> de l'interface SGD. Les protéines inconnues sont celles dont le processus biologique est indéterminé selon les données de SGD. De même, les protéines essentielles sont celles dont la délétion est létale pour la cellule. Pour ce qui concerne les protéines chez *Homo sapiens sapiens*, les nucléolaires ont été sélectionnés dans noPDB (pour Nucleolar Proteome Database), tandis que les non nucléolaires dans HPRD (pour Human Protein Research Database).

### **2.3 Création de jeu de donnée pour la sélection des protéines candidates**

Les fichiers FASTA de toutes les protéines candidates sont stockés dans un jeu de données afin de les soumettre à un filtre d'entraînement bayésien dans but d'identifier

les séquences nucléolaires. Dans cette étude, nous avons 5924 séquences de protéines nucléolaires et non nucléolaires répartie suivant le tableau 2.

TAB. 2: Nombre de protéines analysées suivant leur localisation et leur appartenance spécifique

Espèces	Nucléolaires	non-nucléolaires	Total
S.Cerevisiae	238	2945	3183
H. s.sapiens	748	2193	2941
Total	986	4938	5924

## 2.4 Préparation des données

La nature séquentielle de la bioinformation permet souvent d’interpréter celle ci de la même manière qu’on interprète un texte. Dans le cas présent, nous voulons lui donner l’aspect d’un texte composé de mots pour utiliser une table de fréquences dans l’as- signation de nos probabilités. Les séquences qui nous intéressent sont toute codantes. Comme elle sont toutes traduites en protéines, une division naturelle serait d’utiliser les domaines. PFam est une excellente base de données pour les domaines.

Hélas, la recherche sur PFam se fait soit par identifiant *uniprot*, soit par BLAST. La recherche par identifiant *uniprot* est très rapide, mais ces identifiants ne sont pas tou- jours disponibles. En fait, ils le sont même rarement pour les séquences de notre jeu de données. La recherche par BLAST quand à elle peut prendre une séquence arbitraire. Hélas, c’est une recherche assez longue ce qui pose problème étant donné la taille de notre jeux de données. De plus, l’interface web utilisée par PFam et le format de sortie se porte mal à l’automatisation.

Une option facile à implanter, quoique peu représentative de la réalité est le découpage systématique en sous séquences de longueur  $k$  en faisant glisser un “cadre de lecture” le long de la séquence. C’est l’approche que nous avons utilisé. Nous avons testé l’en- traînement autant avec les séquences nucléotidiques que les séquences peptidiques. Afin de garder un minimum de représentativité biologique, quand nous avons découpé les séquences nucléotidiques, nous avons fait glisser le cadre par sauts de trois. Les “textes” créés par ce découpage ont ensuite servis à entrainer un filtre à spam.

## 2.5 Entraînement

Certaines techniques d’apprentissage machine demandent l’utilisation de logiciels obs- curs aux fonctionnalités hautement théoriques. Par exemple, les moteurs de réseaux

neuronaux sont rarement conviviaux à utiliser et il faut souvent un large bagage en informatique théorique pour comprendre leur documentation. La classification bayésienne naïve en s'attaquant au spam est passer dans le domaine du grand public et il existe un grand nombre de filtres bayésiens facile à utiliser et à paramétrer. Par exemple, *spamoracle*, le filtre qu'on a utilisé permet l'entraînement incrémental et décrémental de même que l'interrogation de la table de fréquences à l'aide d'expressions régulières. Pour l'utilisation courante, nous dirons qu'une séquence non nucléolaire est un spam alors qu'une séquence nucléolaire est un message désirable.

Il y a un problème avec l'assignation des probabilités par les fréquence. En examinant attentivement la formule :

$$P(S|\omega) = \frac{\frac{b}{|B|}}{\frac{b}{|B|} + \frac{g}{|G|}}$$

on constate qu'un mot qui n'apparaît que dans le corpus des bonnes séquences aura une probabilité nulle. En combinant les probabilités, cette valeur nulle sera absorbante sur le produit. Ce qui aura pour effet de générer systématiquement une probabilité nulle pour tout message contenant un tel mot. Afin de contrer cet effet, on utilise des seuils frontière  $p$  et  $q$  ce qui donne la formule révisé

$$P(S|\omega) = \max \left( \min \left( \frac{\frac{b}{|B|}}{\frac{b}{|B|} + \frac{g}{|G|}}, q \right), p \right)$$

On peut ensuite faire varier ces seuils ce qui influence la sensibilité du filtre et permet de compenser la disymétrie dans la taille de chaque corpus.

Finalement, pour tester si le filtre est en mesure de faire une classification correcte, on retire 10 séquences, choisies de manière aléatoire, de chaque corpus avant de faire l'entraînement proprement dit. Ces séquences constituent notre jeu témoin et on vérifie si elle sont bien classées. Pour éviter tout biais dans le choix du jeu témoin, on relance l'entraînement avec un nouveau jeu témoin au moins 25 fois. Les résultats présentés sont la moyenne de ces multiples réentraînements.

Afin d'identifier les domaines conservés, quand une séquence est bien classée, on prend en note les "mots" qu'elle contient. Cette liste de mots sera ensuite classée par nombre d'apparitions et par  $P(S|\omega)$ .

## **2.6 Recherche de séquences intéressantes**

Puisque l'entraînement du filtre donne un tableau de la corrélation respective de chaque "mots" par rapport à la localisation, tout ce qui reste à faire est de trier ces mots par nombre d'occurrence. C'est-à-dire que les mots ayant été bien classifiés (les séquences nucléolaires sont classées comme étant nucléolaires) sont classées selon le nombre de fois qu'ils sont présent. Par la suite, les séquences ayant le plus d'occurrence sont examinées et comparées avec la littérature.

### 3 Résultats et Discussion

#### 3.1 Longueur des mots et pondération

En jouant sur la longueur des mots et sur les probabilités limites ( $p$  et  $q$  ci-haut) on obtient différent taux de classification. Un échantillon de ces résultats est présenté dans le tableau 3. Il y a deux points intéressants à souligner. Premièrement, la proportion de séquences nucléolaires bien classées est plutôt faible. Deuxièmement, la proportion de séquences non-nucléolaires classées comme nucléolaires est très faible. On résume ainsi ces résultats : on manque beaucoup de séquence, mais celle qu'on classe comme nucléolaires le sont probablement.

TAB. 3: Classification en fonction de la longueur des mots et des probabilités limites pour des séquences peptidiques.

longueur	prob min	prob max	% nucléolaire bien classé	% faux positifs
4	.01	.99	17	5
5	.01	.99	18	3
6	.02	.97	36	18
6	.01	.99	28	6
7	.02	.99	12	1
7	.01	.99	13	2
8	.04	.97	9	0
8	.01	.98	9	1
8	.01	.99	8	2

La classification bayésienne naïve n'est pas que binaire, elle émet aussi un degré de confiance. Nous avons considéré une séquence comme nucléolaire si la probabilité combinée de ses mots était de 30%. Certaines séquences nucléolaires non-reconnues ont été mal classées. D'autres ont obtenues une probabilité trop près de 0.5 pour qu'on puisse considérer leur classification comme valable. Plus les mots sont longs, moins on rencontre dans les séquences à classer des mots présents dans la table de probabilités. Et plus le nombre de séquences pour lesquelles la classification n'est pas valable augmente. Avec les séquences d'acides aminés, la proportion de séquences non classées approche 100% avec des mots de longueur 10.

La fiabilité de la classification des séquences nucléotidiques est assez semblable à celle des séquences peptidiques. Elle est légèrement moins bonne mais pas assez pour tirer des conclusions, les différences pourraient être dues au biais du choix du jeu témoin. On s'attendrait à ce que les résultats soient comparables avec des séquences nucléotidiques

environ trois fois plus longue que les séquences peptidiques. Il en est autrement, les résultats sont intéressants avec des séquences de longueur 9 et 12 mais chutent rapidement dépassé 15.

Coté performance, on se félicite du choix de la classification bayésienne. Sur un ordinateur de type *Pentium 4* cadencé à 3.00GHz, il est possible de faire l'entraînement complet du filtre en quelques minutes. Suite à quoi, classer une séquence se fait en moins d'une seconde. Il faut dire que le choix de *spamoracle* y est pour quelque chose. Il est reconnu par plusieurs comme un des filtres bayésiens les plus performants disponibles pour GNU/Linux [2]. Un autre filtre statistique très performant est *bogofilter*, mais son moteur n'est pas strictement bayésien.

### 3.2 Analyse des mots significatifs

L'analyse bayésienne des protéines candidates a permis d'identifier des « mots » dont les plus fréquents sont sélectionnés dans le tableau 4. Pour chacun des mots, nous avons déterminé le nombre d'occurrence et la probabilité de présence.

TAB. 4: Probabilité en pourcentage et nombre d'occurrence des mots identifiés après l'analyse bayésienne des séquences des protéines nucléolaires

Mot	Occurrence	Probabilité (%)
rgrgr	63	2
ggrgg	61	2
grgrg	53	3
deadr	49	2
grtar	43	2
rggrg	40	3
grggf	37	2
rfspd	11	19

L'analyse du tableau ci-dessus montre la présence de deux groupes de mots. Le premier, plus représentatif, est constitué des lettres R et G. Le second groupe est constitué de D, E, A, D et R. Ces résultats concordent bien avec les analyses biologiques antérieures (Zachman et Nigg, 1993) dans l'identification d'un signal de localisation nucléolaire. La richesse en domaine RG (arginine / glycine) indiquerait leur appartenance nucléolaire. Par ailleurs, le second groupe serait représenté par des RNA hélicases à cause de la présence du motif DEAD. Par conséquent le résidu R de ce groupe (arginine) indiquerait que la protéine aurait beaucoup plus d'affinité au nucléole.

Donc, nous voyons entre les résultats du filtre bayésien et des études en laboratoire, une assez bonne corrélation. En effet, la présence des acides aminés R et G ainsi que le groupement DEAD montre des séquences attribuable à certaines protéines. Cependant, avec ces résultats préliminaires, nous ne pouvons conclure à une séquence qui permettrait de localiser les protéines nucléolaire. Contrairement au noyau qui lui possède un signal de localisation nucléolaire (NLS), qui permet d'affirmer que les protéines seront localisée dans le noyau.

## 4 Travaux futurs

Bien qu'on n'arrive pas à des résultats aussi satisfaisant que Paul Graham, on constate que la classification bayésienne naïve peut dans une certaine mesure, identifier la localisation d'une protéine. Afin d'exploiter à fond les possibilités d'un tel système, plusieurs idées nous viennent en tête.

Une des meilleures façons d'améliorer la classification serait d'améliorer la table de probabilités. Ceci se fait automatiquement en raffinant le jeu d'entraînement.

Il serait aussi très pertinent d'examiner les séquences bien classifiées. Elles ont peut-être des choses en communs, comme une taille plus longue, une composition en acides aminés spécifiques (plus ou moins hydrophobes), etc.

Nous avons mis arbitrairement la limite de confiance d'une classification à 30%, il serait intéressant de voir si la variation de cette limite affecte la qualité des classifications.

Après la mise au point des paramètres du filtre, il serait évidemment intéressant de tester les prédictions du filtre sur des séquences inconnues, voir même sur un protéome entier, et de confirmer ces prédictions en laboratoire humide.

La classification bayésienne naïve n'est pas la seule méthode probabiliste qui permet des calculs à haute performance. Le filtre *bogofilter* est une implantation de la méthode de Fichier du  $\chi^{-2}$ , il serait intéressant de voir comment cette technique se comporte avec des séquences biologiques. De plus, il existe des filtres bayésiens à plus de 2 classes, par exemple *ifile*. Il est peut-être possible d'entraîner un tel filtre avec des séquences de plusieurs localisations et de prédire la localisation d'une séquence inconnue.

## 5 Références

[1] : SGD : <http://www.yeastgenome.org/>

[2] : <http://sam.holden.id.au/writings/spam2/>

[3] : William and Wilkens Publishing Inc.

[4] : <http://www.treachercollins.org>

[Graham 01] : <http://paulgraham.com/spam.html>

Bond, A.T., D.A. Mangus, F. He, A. Jacobson, 2001. Absence of Dbp2p alters both nonsense-mediated mRNA decay and rRNA processing. *Mol Cell Biol.* 21(21) :7366-7379.

Carmo-Fonseca M, Mendes-Soares L, Campos I. 2000. To be or not to be in the nucleolus. *Nat. Cell. Biol.* Jun ;2(6) :E107-12.

de la Cruz J, D.Kressler et P. Linder, 1999. Unwinding RNA in *Saccharomyces cerevisiae* : DEAD-box proteins and related families. *Trends Biochem Sci.* 24(5) :192-8.

Dixon J, Edwards SJ, Anderson I, Brass A, Scambler PJ, Dixon MJ. 1997. Identification of the complete coding sequence and genomic organization of the Treacher Collins syndrome gene. *Genome Res* ; 7 : 223-34.

Dixon J, Hovanes K, Shiang R, Dixon MJ. 1997 Sequence analysis, identification of evolutionary conserved motifs and expression analysis of murine *tcof1* provide further evidence for a potential function for the gene and its human homologue, TCOF1. *Hum Mol Genet.* 6 :727-37

Dragon, F., J.E. Gallagher, P.A. Compagnone-Post, B.M. Mitchell, K.A. Porwancher, K.A. Wehner, S. Wormsley, R.E. Settlage, J. Shabanowitz, Y. Osheim, A.L. Beyer, D.F. Hunt et S.J. Baserga, 2002. A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* 417 : 967-970.

Eichler, DC. N. Craig, 1994. Processing of eukaryotic ribosomal RNA. *Prog Nucleic Acid Res Mol Biol* 49 : 197-239. Fang, J., S. Kubota, B. Ying, N. Zhou, H. Zhang, R. Godbout et R. J. Pomerantz, 2004. A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology* 330 : 471-480.

Gladwin AJ, Dixon J, Loftus SK, Edwards S, Wasmuth JJ, Hennekam RC, Dixon MJ. 1996 Treacher Collins syndrome may result from insertions, deletions or splicing mutations, which introduce a termination codon into the gene. *Hum Mol Genet.* 5 :1533-8

- Hernandez-Verdun, D. Louvert, E. 2004. Le nucléole : structure, fonctions et maladies associées. *Medecine/sciences* 20 : 37-44
- Isaac C, Marsh KL, Paznekas WA, et al. 2000. Characterization of the nucleolar gene product, treacle, in Treacher Collins syndrome. *Mol Biol Cell* ; 11 : 3061-71.
- Kressler, D., J. de la Cruz, M. Rojo, P. Linder, 1998. Dbp6p is an essential putative ATP-dependant RNA helicase required for 60S ribosomal subunit assambly in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 18 : 1855-1865.
- Lee, S.J. et S.J. Baserga, 1999. Imp3p and Imp4p, Two Specific Components of the U3 Small Nucleolar Ribonucleoprotein That Are Essential for Pre-18S rRNA Processing. *Molecular and Cellular Biology* 19(8) : 5441-5452.
- Lorsch, JR., 2002. RNA chaperones exist and DEAD box proteins get a life. *Cell.* 109(7) : 797-800.
- Marciniak RA, Lombard DB, Johnson FB, Guarente L. 1998. Nucleolar localization of the Werner syndrome protein in human cells. *Proc Natl Acad Sci USA* ; 95 : 6887-92.
- Milligan DA, Harlass FE, Duff P, Kopelman JN. 1994. Recurrence of Treacher Collins' syndrome with sonographic findings. *Mil Med.*159 :250-2.
- Moser MJ, Oshima J, Monnat RJ Jr. 1999. WRN mutations in Werner syndrome. *Hum Mutat.*13 :271-9.
- Newman, JR., E. Wolf, PS. Kim, 2000. A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA.* 24 :13203-8.
- Olson, M. O., M. Dundr et A. Szebeni, 2000. The nucleolus : an old factory with unexpected capabilities. *Trends Cell Biol.* 10 : 189-196.
- Rocak, S., P. Linder, 2004. DEAD-box proteins : the driving forces behind RNA metabolism. *Nat. Rev Mol Cell Biol* 5 :232-241.
- Scheer, U., et R. Hock, 1999. Structure and function of nucleolus. *Curr. Opin. Cell Biol.* 11 : 385-390
- Shaw, P.J., 2001. Nucleolus. *ENCYCLOPEDIA OF LIFE SCIENCES/ Nature Publishing Group.*
- Shiratori M, Suzuki T, Itoh C, Goto M, Furuichi Y, Matsumoto T. 2002. WRN helicase accelerates the transcription of ribosomal RNA as a component of an RNA polymerase I associated complex. *Oncogene* ; 21 : 2447-54.

Tanner, N.K. O. Cordin, J. Banroques, M. Doere, P. Linder, 2003. The Q motif : a newly identified motif in DEAD box helicases may regulate ATP binding and hydrolysis. *Mol Cell*. 11(1) :127-38.

Tanner, N.K. et P. Linder, 2001. DExD/H Box RNA Helicase : From Genetic Motors to Specific Functions. *Mol. Cell*. 8 : 251-262.

Wilson, B.J., G.J. Bates, S.M. Nicol, D.J. Gregory, N.D. Perkins, et F.V. Fuller-Pace, 2004. The p68 and p72 DEAD box RNA helicases interact with HDAC1 and repress transcription in a promoter-specific manner. *BMC Mol Biol*. 5 : 11.

Winokur ST and Shiang R. 1998. The Treacher Collins syndrome (TCOF1) gene product, treacle, is targeted to the nucleolus by signals in its C-terminus. *Hum Mol Genet*. 7 :1947-52

Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, Nakura J, Miki T, Ouais S, Martin GM, Mulligan J, Schellenberg GD. 1996. Positional cloning of the Werner's syndrome gene. *Science* ; Apr 12 ;272(5259) :258-62.

## 6 Annexes

### 6.1 Figures

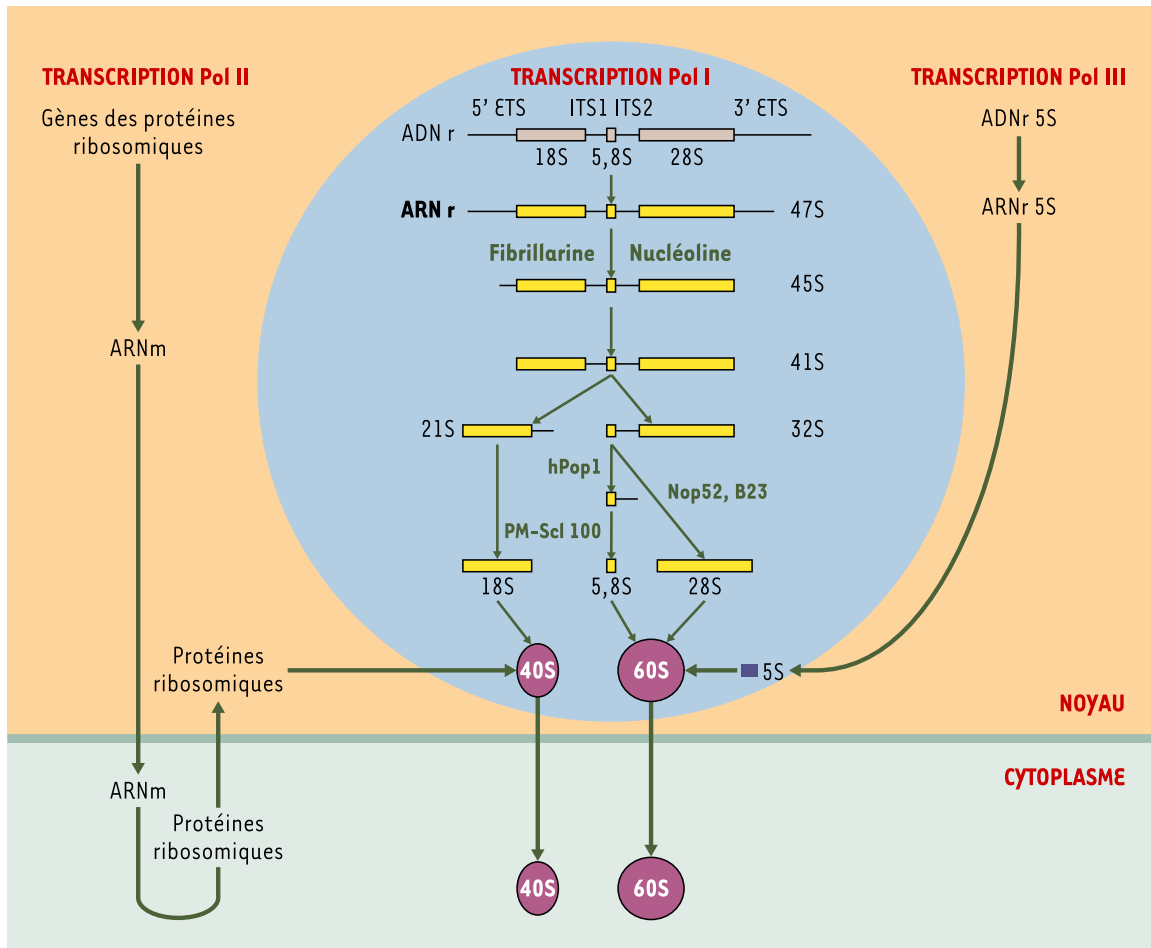


Figure1 : Formation des ribosomes chez *Saccharomyces cerevisiae*  
(Hernandez-Verdun et Louvert, 2004)

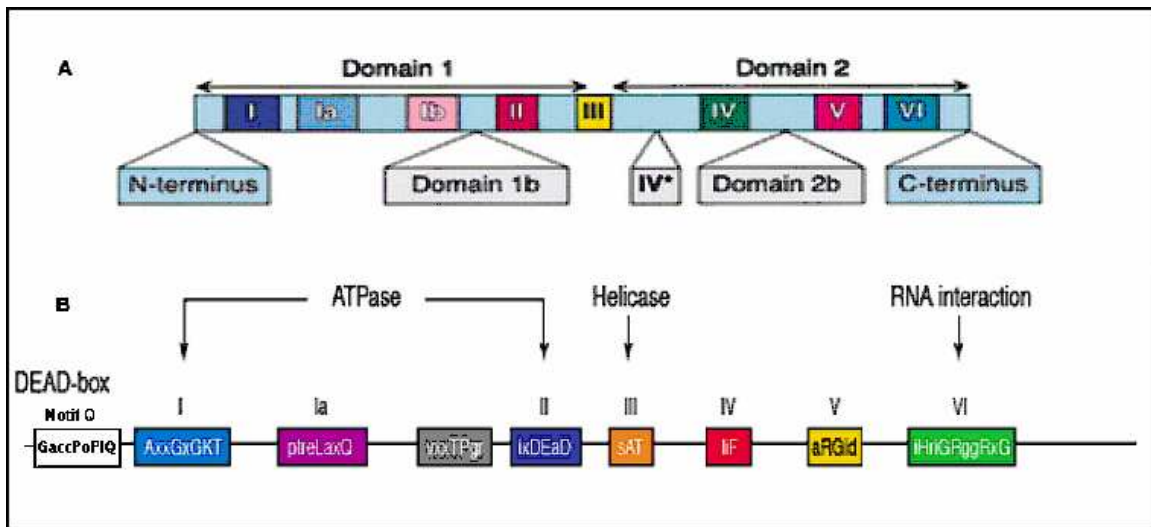


Figure 2 : Séquences des motifs conservés de l'ARN hélicase. (A) Position des motifs conservés dans eIF4A montrant leurs extensions amino et carboxy terminale au delà les domaines 1et 2. Ces domaines sont connectés par le motif III qui présenterait une activité hélicase. (B) Représentation schématique de la région centrale d'une hélicase à ARN montrant les principaux domaines fonctionnels (Tanner et al., 2001).