Recherche des gènes d'ARN non codant

Emmanuel Gothié, Yann Guermeur, Sébastien Muller, Christiane Branlant, Alexander Bockmayr

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Recherche des gènes d'ARN non codant

Emmanuel Gothié – Yann Guermeur – Sébastien Muller – Christiane Branlant – Alexander Bockmayr

N° 5057

Décembre 2003

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--5057--FR

Search for Non-Coding RNA Genes

Emmanuel Gothié*°, Yann Guermeur*, Sébastien Muller°,
Christiane Branlant°, Alexander Bockmayr*§

Thème 2 – Génie logiciel et calcul symbolique
Projet MODBIO

Rapport de recherche n° 5057 – Décembre 2003 - 19 pages

Abstract:

The considerable amount of data originating from genome sequencing programs requires innovative analysis techniques. The first step in annotating genomic sequences is the search for protein encoding regions or ORF (Open Reading Frame). However, non-coding RNA (ncRNA) genes, which produce functional RNAs instead of proteins, do not exhibit the signals used in the detection of ORF.
The systematic research of ncRNA genes hence requires the development of specifically designed tools, which represents an important challenge in the post-genomic era.
We propose a method based on statistical learning using Support Vector Machines (SVMs), which is applicable to all kind of genomic sequences. This approach has been validated by searching for C/D-box or H/ACA-box snoRNAs in the genomes of *S. cerevisiae* and of Archaea belonging to the *Pyrococcus* genus.

Keywords: bioinformatics, computational biology, inter-ORF sequences, ncRNA, genomes, multi-class SVM.

* Projet MODBIO, LORIA (UMR 7503 - CNRS, INPL, INRIA, Université Henri Poincaré Nancy 1, Université Nancy 2), BP 239, 54506 Vandoeuvre-lès-Nancy, France
° Laboratoire de Maturation des ARN et Enzymologie Moléculaire (MAEM), UMR 7567 CNRS-UHP, BP 239, 54506 Vandoeuvre-lès-Nancy, France
§ Une partie de ce projet a été réalisée avec un financement du Programme Bio-Informatique inter-EPST et un soutien de la Région Lorraine dans le cadre du CPER 2000-2006.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique

Recherche des gènes d'ARN non codant

Résumé : La masse considérable de données brutes extraite des programmes de séquençage nécessite de nouvelles techniques d'analyse. La première étape visant à annoter les séquences génomiques est la recherche de régions codant des protéines (ORF pour *Open Reading Frame*). Cependant les gènes d'ARN non codant (ARNnc), qui ne produisent pas de protéines mais des ARN fonctionnels en tant que tels, ne présentent pas les signaux utilisés pour la détection d'ORF.
La recherche systématique des gènes d'ARNnc requiert de ce fait le développement d'outils appropriés, ce qui représente un challenge de premier ordre dans l'ère post génomique.
Nous proposons ainsi d'utiliser une méthode issue de l'apprentissage statistique basée sur les machines à vecteurs support (SVM) qui est applicable à l'ensemble des séquences génomiques. Cette approche a été validée par la recherche de snoRNA à boîtes C/D ou H/ACA dans le génome de la levure *S. cerevisiae* et dans les génomes d'Archaea du genre *Pyrococcus*.

Mots clés : Bio-informatique, séquences inter-ORF, ARNnc, génomes, SVM multi-classes.

Sommaire

1 Introduction

Les données de séquençage de nombreux génomes sont l'objet d'analyses *in silico*, soit pour une annotation automatique précise, soit pour la recherche de séquences particulières : ORF, régions régulatrices, etc… La plupart des outils actuels sont orientés vers la recherche des séquences codant des protéines et sont inadaptés pour la recherche des ARN ne codant pas des protéines, les ARNnc, dont le rôle s'avère de plus en plus important **[1 à 4]**.

L'objectif de ce projet de recherche est de proposer un outil permettant l'identification dans les génomes de gènes codant pour des ARNnc. Le principe de cette recherche est basé sur l'apprentissage statistique qui se prête bien aux données déjà disponibles ou obtenues ultérieurement, à savoir des données de génomes très détaillées (ex. génomes de Levure, Drosophile etc…) et une masse croissante de séquences génomiques à annoter. En réalisant un apprentissage sur les données d'un génome modèle, il est possible d'envisager de tirer des connaissances intrinsèques et d'annoter les nouveaux génomes à étudier.

Dans une première étape, un modèle biologique de recherche, la levure *S. cerevisiae*, a permis de tester l'outil développé avant application à d'autres génomes. L'approche retenue est une approche hiérarchique avec, dans un premier temps, une recherche bas niveau réalisée grâce à l'utilisation de machines à vecteurs support (SVM **[5]**), un outil de la reconnaissance des formes ayant été utilisé avec succès dans plusieurs domaines comme la parole, la catégorisation de textes et récemment dans l'identification de sites d'épissage dans les ARN eucaryotes (**[9]** et références citées). Dans un second temps, un raffinement pourra être mis en place pour améliorer la détection en prenant en compte des informations symboliques (données de structures secondaires, de contexte etc…).

Le rapport est découpé en trois parties présentant dans un premier temps les contextes de l'étude au niveau biologique et vis-à-vis de la communauté nationale et internationale. La deuxième partie présentera brièvement le domaine de l'étude et l'approche utilisée. Les premiers résultats obtenus sont décrits dans la dernière partie du document.

2 Contexte de l'Etude

2.1 Séquences inter-ORF et ARN non codant

La plupart des outils actuels d'analyse de séquences sont orientés vers la recherche de séquences codant des protéines, cependant il existe de très nombreux ARN qui ne codent pas des protéines mais qui jouent des rôles essentiels dans la cellule (structural, régulateur, catalytique). Le développement d'outils d'analyse permettant dans les génomes nouvellement séquencés l'identification des gènes de ces ARN, mais aussi la découverte de nouveaux ARN de régulation aux fonctions inconnues, est donc nécessaire.

Ces ARN non codant sont présents dans tous les organismes (des bactéries aux mammifères). Beaucoup ont un rôle régulateur de l'expression génique au niveau traductionnel (via une action " antisens " ou par interaction avec des protéines) en particulier au niveau de la régulation du développement aussi bien chez les plantes, les invertébrés ou les insectes, que chez les vertébrés. D'autres sont impliqués dans la réplication, la sécrétion de protéines, l'épissage, les modifications post-transcriptionnelles des ARN (ARNribosomique (ARNr), ARN de transfert (ARNt), petits ARN nucléaires (snRNA) etc…)**[1]**. Une implication d'ARN a aussi été montrée dans l'étiquetage de protéines pour leur dégradation, dans l'activation de la traduction, et la modulation de l'activité de l'ARN Polymérase (pour revue, **[10]**).

Chez la bactérie *Escherichia c*oli, aux 4290 gènes codant pour des protéines, il a fallu ajouter 62 gènes de sRNA (*small RNA*) qui ont été découverts par combinaison de différentes approches (**[1]**, **[2]**, **[3]**, et références citées).

Des sRNA ont aussi été mis en évidence dans les Archaea (pour revue **[11]**), la levure ainsi que chez les mammifères comme la souris **[4]**. Leur recherche par les approches expérimentales ou informatiques classiques demeure difficile, mais les connaissances acquises à chaque nouvelle description, connaissances liées à la structure, la localisation de leurs gènes (majoritairement situés entres les gènes de protéines ou dans leurs introns), ainsi que la conservation des séquences entres les espèces, permettent d'envisager le développement d'outils d'analyse plus efficaces.

2.2 Etat de l'Art

Des recherches en bioinformatique pour la mise en évidence des ARNnc ont lieu maintenant au sein de plusieurs équipes dans le monde. La recherche d'ARNnc requiert la prise en compte de spécificités telles que la structure secondaire, l'existence de biais au niveau des nucléotides et des particularités qui ne sont pas prises en compte par les outils de recherche classiques tels que par exemple **Blast**, **Fasta** … qui initient leur recherche sur l'existence de mots exacts dans les séquences. Les algorithmes de recherche de motifs ne sont pas non plus adaptés.

Une première approche a pu être réalisée à l'aide d'algorithmes spécifiques basés sur des descripteurs comme c'est le cas par exemple pour les ARNt, une famille d'ARN très bien caractérisée (ex. tRNAscan). Pourtant cette approche nécessite une expertise qui la rend difficile à mettre en œuvre dans le cas des ARNnc encore mal définis. Cependant, des modèles stochastiques peuvent être utilisés comme les Chaînes de Markov Cachées (*Hidden Markov Models* ou HMM) permettant la modélisation de familles de séquences (distribuant les éléments en codant/non codant par exemple) ou des approches utilisant des grammaires hors-contexte stochastiques (*Stochastic Context Free Grammars ou* SCFG) permettant une modélisation des formations de structures secondaires. Ces méthodes ont l'avantage de fournir une description statistique des éléments (composition, probabilité).

Ces approches sont en cours de développement dans différentes équipes de recherche. Citons par exemple ***snoscan***, un outil de recherche de snoRNA guides de 2'-O-methylation basé sur la définition de caractéristiques clefs de ces gènes (**[14]** - USA), et plus récemment des prototypes de programmes de recherche d'ARNnc basés sur l'analyse par comparaison de séquences comme ***QRNA*** impliquant les *pair-HMMs* et une *pair-SCFG*, (**[15]** - USA), ou encore une utilisation du logiciel ***ERPIN*** basé sur l'alignement de séquences et la prise en compte de structures secondaires (Equipe D. Gautheret – France - **[16]**). Dans la plupart des cas, ces programmes nécessitent d'être encore développés/appliqués aux ARNnc ou améliorés et ne sont pas exhaustifs. Un autre programme de recherche de gènes codant pour des ARN structurés ***NCRNASCAN*** (**[17]** - USA), basé sur l'approche SCFG, a montré les limites d'une recherche basée sur les seules informations de structures secondaires. En effet, l'hypothèse de départ selon laquelle les ARN biologiquement fonctionnels ont en général un contenu en structures secondaires plus élevé qu'une séquence aléatoire avec la même composition, n'est pas vérifiée (**[17]**). Ainsi ce type de postulat ne peut être à la base d'un algorithme de recherche d'ARNnc, bien que la structure secondaire conservée reste un signal statistique prometteur pour la recherche de gènes d'ARNnc. Ces approches présentant des limites, étant encore en cours de développement, ou ne s'avérant pas forcément exhaustives, il est souhaitable d'envisager d'autres approches pour trouver une solution pour la recherche des ARNnc.

Nous avons choisi pour notre part de développer une approche basée sur l'apprentissage statistique : les machines à vecteurs support (SVM) qui possèdent comme caractéristique intéressante de trouver une séparation des données maximisant la distance aux enveloppes convexes des nuages de points associés aux différentes catégories (aucune connaissance de la structure sous-jacente n'étant requise). Les machines à vecteurs support ont comme autre avantage de pouvoir intégrer des sources de données hétérogènes. Enfin, dans certains cas, l'approche par SVM s'est montrée plus efficace que des approches par réseau de neurones ou HMM **[9]**. Il faut néanmoins

préciser que l'analyse discriminante et la modélisation stochastique présentent chacune leurs avantages et leurs inconvénients, et qu'il est possible de les combiner dans des systèmes hybrides.

3 Domaine de l'Etude et Approche

3.1 Domaine de l'Etude

L'étude a été initiée sur le génome de la levure *S. cerevisiae,* premier génome eucaryote séquencé entre 1992-1996. Sa taille est de 12 156 307 pb (paires de bases), il comporte 16 chromosomes nucléaires et 1 chromosome mitochondrial avec un total de 5651 gènes actuellement répertoriés. Les données de séquences ont été récupérées à partir du site SGD[1]. Les génomes de levures Hémiascomycètes constituent un matériel d'étude intéressant. En effet, les processus cellulaires de base sont similaires à ceux des cellules humaines (transcription de l'ADN, synthèse des protéines etc…) et la levure *S. cerevisiae* se cultive facilement, ce qui fait de cette dernière un bon outil pour les approches génétiques. De plus, ces génomes présentent l'avantage d'être de petite taille, et seuls les gènes essentiels à la machinerie cellulaire sont présents, ce qui facilite la recherche de nouveaux gènes (temps de calcul). Par ailleurs, la disponibilité de plusieurs génomes (5 génomes entièrement séquencés et une base de données conséquente de fragments d'ADN annotés dans le cadre du projet Génolevures (*Génoscope* - http://www.genoscope.cns.fr) **[12]** - portant sur 13 espèces représentatives de la classe des Hémiascomycètes) permet d'utiliser les stratégies de génomique comparative.
Le cas des ARNs nucléolaires snoRNA a été choisi dans le panorama des ARN non codant pour développer l'approche de recherche d'ARNnc. Il existe deux sous-classes de snoRNA impliquées dans les deux types de modifications majeures des ARN. Les snoRNA à boîtes C/D intervenant dans des 2'-O-méthylations des riboses et les snoRNA à boîtes H/ACA impliqués dans des pseudouridylations (isomérisation des résidus uridines en résidus Pseudouridine Ψ). Ces modifications participent à la maturation des ARNr qui peuvent, à l'issue de différentes étapes d'activation, fixer les protéines ribosomiques pour former une structure mature, le ribosome, qui est un élément essentiel dans la traduction des ARNm en protéines. Au début du projet, aucun logiciel ne permettait de rechercher les snoRNA de la classe H/ACA dans les génomes à l'inverse des snoRNA à boîtes C/D (logiciel *snoscan*). Utiliser le cas des snoRNA permet d'avoir un modèle d'étude d'ARNnc, de rechercher les snoRNA à boîtes C/D manquants (1/10 estimé chez la levure et chez la souris **[13]** mais également de proposer une technique de recherche systématique des snoRNA à boîtes H/ACA encore inexistante et mettre en évidence les snoRNA H/ACA manquants (20 snoRNA H/ACA connus pour 44 pseudouridines d'ARNr chez la levure, et 42 guides de pseudouridylation mis en évidence chez les mammifères sur les 91-93 pseudouridines d'ARNr connus **[13]**.
Par ailleurs, il semblerait qu'au delà des cibles sur les ARNr, les snoRNA peuvent modifier d'autres ARN augmentant ainsi l'intérêt de leur recherche systématique (exemple de guide de méthylation d'ARNt **[13]** et références citées).

3.2 Approche

3.2.1 Présentation de l'Apprentissage Statistique

L'apprentissage statistique permet d'effectuer des tâches d'estimation de fonctions ou de discrimination à partir d'un échantillon d'observations étiquetées. Dans le cas de la classification, réaliser un apprentissage consiste à sélectionner dans une famille de fonctions $\mathcal{F}$ donnée une fonction minimisant l'erreur en généralisation.

[1] **SGD** (Saccharomyces Genome Database) : http://genome-www.stanford.edu/Saccharomyces/

3.2.2 Présentation des Machines à Vecteurs Support (SVM)

Les machines à vecteurs support sont des systèmes d'apprentissage automatique introduits par Vapnik et ses collègues **[7,8,5]**. Elles permettent, à partir de la connaissance acquise sur un échantillon d'apprentissage, de classer de nouvelles observations.
L'algorithme des SVM est fondé sur le principe inductif SRM (Structural Risk Minimization) **[6]**. Il consiste à minimiser une borne supérieure sur le risque nommée risque garanti. En pratique, cela revient à chercher un bon compromis entre capacité et risque empirique. La SVM peut être bi-classe (calculer des dichotomies) ou multi-classe.
La SVM met en œuvre un séparateur linéaire à marge maximale (cf. Figure 1A) dans un espace de représentation de haute dimension où les données sont projetées par un opérateur Φ. Pour déterminer l'opérateur de projection, on choisit une fonction noyau k qui calcule le produit scalaire dans l'espace de représentation ($k(x,x') = < \Phi(x), \Phi(x')>$). Cette fonction k, et non l'opérateur de projection Φ, est la seule qui apparaît dans les calculs. La possibilité de gérer l'opérateur de projection de façon implicite est connue sous le nom de *kernel trick*.
L'intérêt de la projection est de rendre les données plus facile à séparer par un hyperplan. (cf. Figure 1B).
Le programme de SVM multi-classe (M-SVM **[25]**) utilisé pour cette étude est téléchargeable à partir de l'adresse http://www.loria.fr/~guermeur/

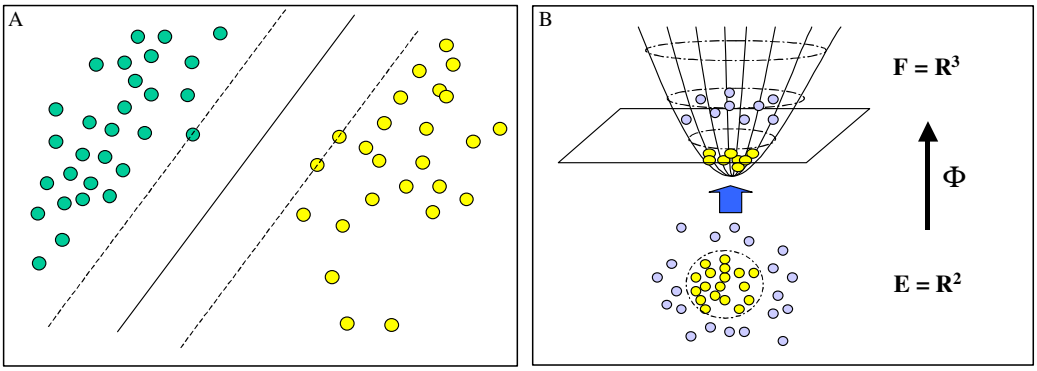


Figure 1 **(A)** Exemple de distribution de points en deux classes linéairement séparables (représentées selon deux couleurs vert et jaune). L'hyperplan de marge maximale est matérialisé par une ligne pleine, les marges par des lignes parallèles pointillées.

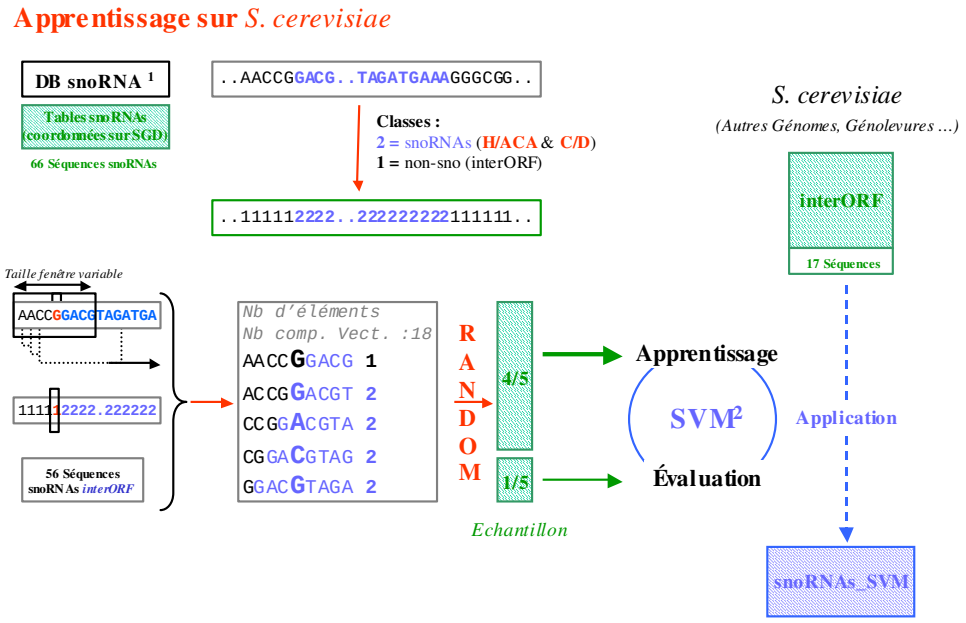
(B) Cas où les données ne peuvent pas être séparées linéairement. Un pré-traitement des données est réalisé (Φ : projection des données sur un paraboloïde) qui rend possible une séparation linéaire.

4 Résultats

4.1 Préparation des Données et Principe

L'apprentissage a été réalisé sur les données connues de *S. cerevisiae* pour laquelle 66 séquences snoRNA sont décrites (*Dmitry A. Samarsky and Maurille J. Fournier Database* **[18]**). Pour ce faire, une étape de Traduction/Assignation des séquences des chromosomes de *S. cerevisiae* en code numérique (alphabet à n chiffres, n étant le nombre de prédicteurs pris en compte) a été réalisée en vue du traitement par la SVM. Deux classes ont été dans un premier temps fixées : la **classe 2** dans le cas d'une séquence de snoRNA et la **classe 1** dans le cas d'une séquence non-snoRNA. Un fichier de classes est alors réalisé à l'aide d'un script Perl qui assigne les différentes classes selon les tables de coordonnées des séquences snoRNA. A partir du fichier de séquences et du fichier des classes assignées, nous générons les fichiers en entrée des SVM qui sont constitués d'une **fenêtre d'apprentissage de taille fixe** (pouvant être choisie) ainsi que la

classe du nucléotide central (cf. Figure 2). Il est possible de faire varier la taille de la fenêtre d'apprentissage (de façon systématique ou en incorporant des connaissances biologiques au problème) et de prendre ainsi en considération un nombre plus ou moins important de nucléotides permettant de déterminer l'appartenance du nucléotide central à une des classes répertoriées. Seules les séquences snoRNA dans un contexte de séquences inter-ORF sont conservées (56 séquences dont 17 snoRNA H/ACA et 39 snoRNA C/D).



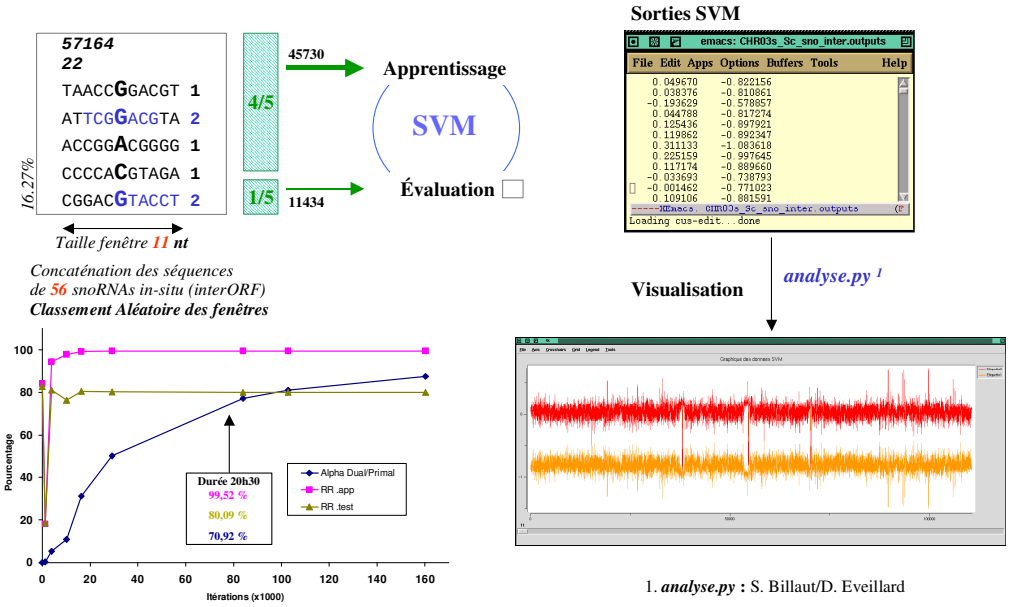
1 - Dmitry A. Samarsky and Maurille J. Fournier
2 - M-SVM1 : Yann Guermeur (train_SVM.c v1.1 & eval_SVM.c v1.0)

Figure 2 : Schéma de la préparation des données et analyse par SVM. A partir des coordonnées des snoRNA les séquences ADN sont annotées à l'aide de chiffres représentant la classe à laquelle appartient un nucléotide considéré (**classe 1** pour un nucléotide de séquence inter-ORF, et **classe 2** pour un nucléotide de snoRNA). A partir du fichier des séquences snoRNA placés dans leur contexte de séquence inter-ORF concaténées et du fichier correspondant des classes, un script génère un fichier d'entrée de la SVM constitué d'une succession de fenêtres de **n nucléotides** en vis-à-vis de la classe du nucléotide central correspondant. Les fenêtres de séquences sont transformées en code binaire puis mélangées de façon aléatoire. 4/5 de l'échantillon est utilisé pour faire un apprentissage par la SVM, 1/5 pour faire l'évaluation de l'apprentissage. Une fois cette évaluation validée, il est possible d'utiliser cette SVM entraînée pour rechercher des snoRNA dans des séquences inter-ORF.

Un apprentissage est alors réalisé sur cet ensemble constitué des séquences des snoRNA de *S. cerevisiae* situées dans leur contexte de séquences inter-ORF (4/5 de l'échantillon mélangé aléatoirement), et évalué sur une fraction de l'échantillon non utilisée dans l'apprentissage (1/5 restant). La validation de cette étape permet d'appliquer la méthode d'analyse par SVM à la prédiction de séquences snoRNA dans les fichiers de données des séquences inter-ORF générés pour *S. cerevisiae*, dans d'autres génomes complets de levures ou dans les données de Génolevures.

Les données en sortie de la SVM sont visualisées à l'aide d'un logiciel développé en langage python (analyse.py : S. Billaut/D. Eveillard – **[19]**). Ce programme permet de tracer deux signaux graphiques en fonction des scores donnés par la SVM (cf. Figure 3). Dans le cas d'une séquence snoRNA, le signal inter-ORF s'interchange avec le signal snoRNA à l'emplacement prévu pour l'ARNnc. Sur un échantillon d'apprentissage réalisé avec une fenêtre de 11 nucléotides (57164 fenêtres – 16.27% de Classe 2), et pour une durée d'apprentissage variable (jus-

qu'à 40h) le système apprend bien sur les données (Taux de Reconnaissance augmentant dans le temps). La SVM utilisée pour les analyses (entraînement sur 80000 itérations – Durée 20h30[2]) donne un taux de reconnaissance de 99.52% en apprentissage (4/5 de l'échantillon) et de 80.09% en test (1/5 de l'échantillon) (cf. Figure 3).



1. ***analyse.py*** **:** S. Billaut/D. Eveillard

Figure 3 : Paramètre d'apprentissage. Les 56 séquences de snoRNA situées dans leurs séquences inter-ORF sont concaténées, un fichier constitué d'une succession de fenêtres d'apprentissages de 11nt placées en vis-à-vis de la classe du nucléotide central correspondant est généré et les données sont mélangées de façon aléatoire. 4/5 est utilisé en apprentissage. La proportion finale **fenêtres de classe 2** pour cet apprentissage est de 16.27%. La SVM ainsi entraînée sur les 4/5 et appliquée sur le 1/5 restant est capable d'attribuer correctement la classe des nucléotides à 80.09% pour 20h30 d'apprentissage. Les sorties de la SVM, des fichiers contenant les scores inter-ORF (colonne 1) et snoRNA (colonne 2), sont visualisées à l'aide du logiciel **analyse.py** (signal inter-ORF : courbe rouge, signal snoRNA : courbe orange), le signal supérieur indiquant la classe probable du nucléotide central. Le signal peut être filtré en appliquant un lissage des différentes données (cf. [19]) permettant de mettre en évidence les signaux marqués en diminuant le bruit de fond.

4.2 Mise en œuvre – Analyse de *S. cerevisiae*

Le système entraîné selon la manière décrite précédemment (entraînement de 20h30 sélectionné) et testé sur le génome de *S. cerevisiae* est capable de retrouver au sein de l'ensemble des séquences inter-ORF de cette levure l'intégralité des snoRNA contenus dans les séquences inter-ORF (56 Séquences - cf. Figure 4). Avec les paramètres utilisés, très peu de bruit de fond est observé et les signaux sont très marqués. Des signaux supplémentaires sont également présents et permettent de penser que d'autres snoRNA peuvent être mis en évidence par cette approche. L'analyse fine de ces signaux, ainsi que des vérifications expérimentales permettront de valider ces candidats.

[2] : Pour le serveur de calcul utilisé (Cf. Données Techniques)

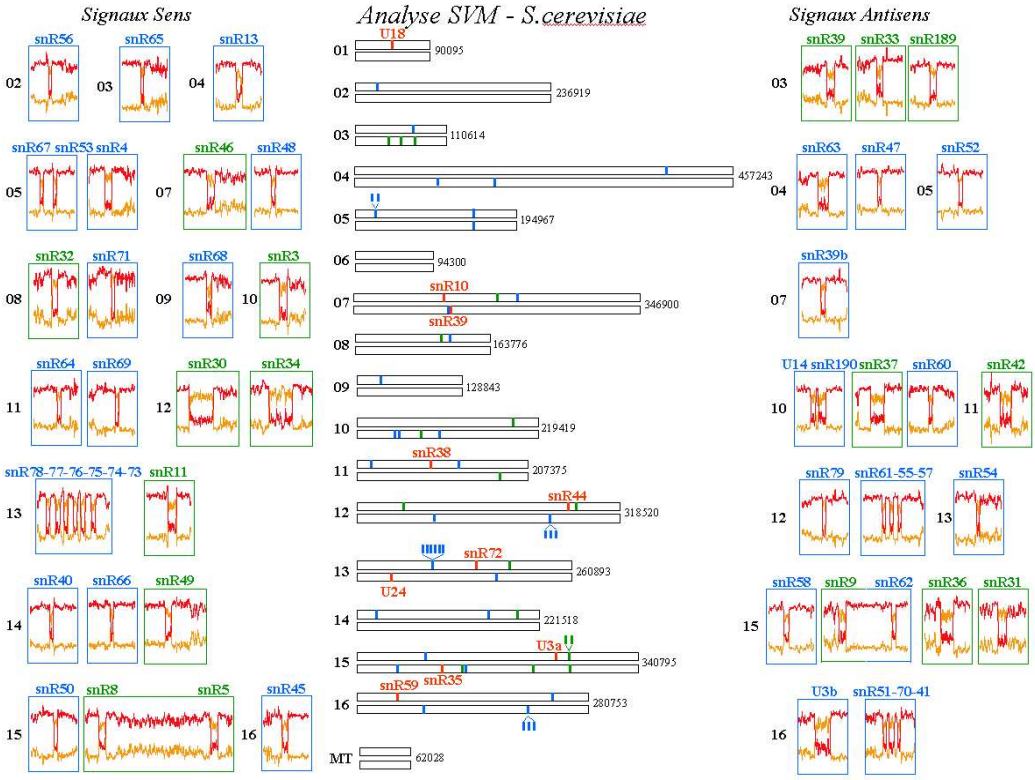


Figure 4 : Analyse de Saccharomyces cerevisiae. Recherche simultanée des snoRNA à boîtes H/ACA et C/D dans les séquences inter-ORF de *S. cerevisiae*. L'apprentissage a été réalisé sur 4/5 des 56 séquences snoRNA placées dans leur contexte de séquences inter-ORF (durée de l'apprentissage 20h30). Après validation de l'apprentissage (test sur 1/5), la recherche des snoRNA a été lancée sur les séquences inter-ORF de *S. cerevisiae* concaténées par chromosome. Les positions relatives des snoRNA (positions indiquées par des barres vertes pour les H/ACA et des barres bleues pour les C/D) sont indiquées sur ces séquences (centre du schéma). Les 10 séquences snoRNA contenues dans des ORF ou chevauchant des ORF, n'ayant pas été utilisées pour l'apprentissage, ont été laissées dans les séquences à analyser (positions indiquées par des barres rouges). Les 56 séquences snoRNA ont été retrouvées lors de l'analyse graphique des signaux (via *analyse.py* pour un lissage de 11, cf. [19]) et les signaux correspondant sont donnés (brin Sens de l'ADN à gauche et brin Antisens de l'ADN à droite). Avec les paramètres utilisés lors de cette étude, les snoRNA des ORFs ne sont pas détectés, soulignant l'extrême spécificité de l'apprentissage.

4.2.1 Détails de l'analyse du Chromosome 07

L'exemple de l'analyse réalisée sur le Chromosome 07 de *S. cerevisiae* est donné ci-dessous. Ce chromosome contient 5 séquences snoRNA dont 3 sont situées dans les séquences inter-ORF (cf. table de la Figure 5). Les séquences sont analysées dans les deux directions (Sens et Antisens) et les tracés de sortie des SVM sont étudiés afin de retrouver les positions des snoRNA. La Figure 5 présente les signaux des 3 snoRNA attendus retrouvés à leur position. Deux signaux supplémentaires peuvent également être mis en évidence : ils peuvent correspondre à des faux positifs ou à des snoRNA non encore mis en évidence. Aucun signal n'est détecté pour les snoRNA contenus dans les ORF (laissées dans le fichier de séquences à analyser). Ils sont donc considérés par la SVM comme ne faisant pas partie de la classe des snoRNA telle que définie par l'apprentissage.

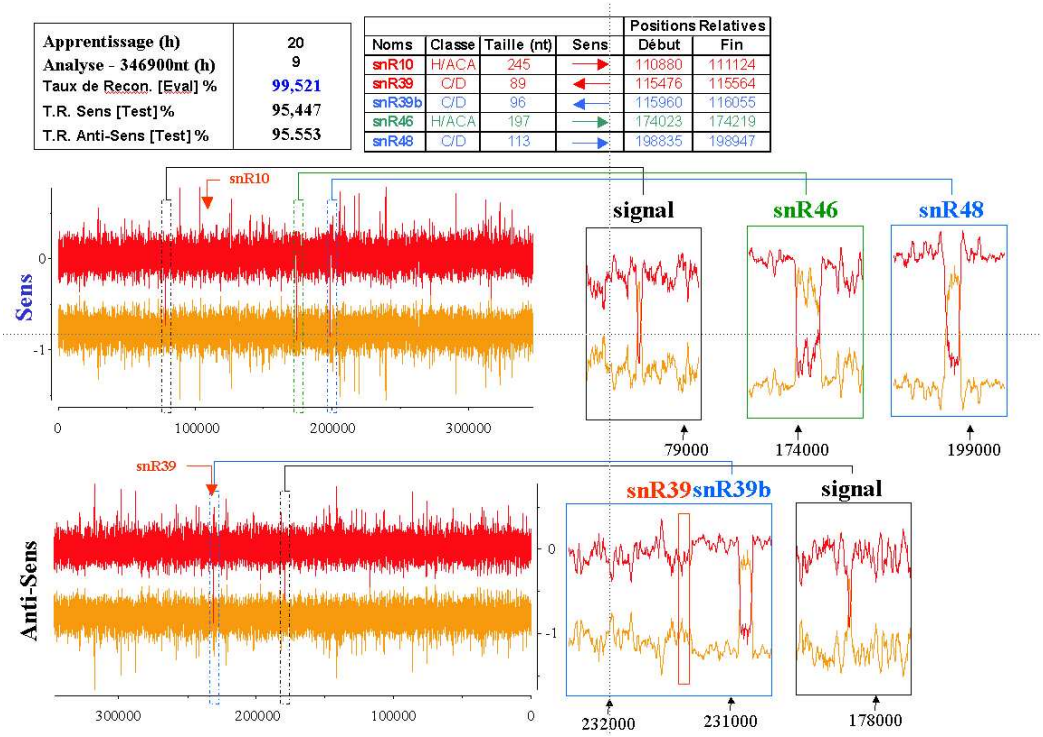


Figure 5 : Illustration : Séquences inter-ORF du Chromosome 07. La table en haut à gauche indique la durée de l'analyse SVM du Chromosome en Sens et Antisens ainsi que les taux de reconnaissance obtenus. La table de droite indique les paramètres des snoRNA contenus dans ce chromosome (en rouge, les snoRNA contenus dans des ORF non pris en compte dans l'apprentissage (mais laissés dans les séquences à analyser), en bleu snoRNA à boîtes C/D et en vert les snoRNA à boîtes H/ACA), ainsi que leurs positions relatives dans les séquences inter-ORF concaténées (position sur le brin sens). Les tracés SVM pour un lissage de 11 (signal filtré cf. [19]) sont présentés avec des agrandissements sur certaines portions présentant un signal positif. Les trois snoRNA snR39b, snR46 et snR48 sont retrouvés à leurs positions attendues. Deux signaux supplémentaires assez marqués peuvent être mis en évidence. Les snoRNA snR39 et snR10 ne sont pas détectés par la SVM entraînée sur les seuls snoRNA contenus dans les séquences inter-ORF (voir exemple donné pour le snR39 positionné à côté du snR39b).

4.2.2 Discussion des résultats

La SVM entraînée sur 4/5 des séquences des 56 snoRNA utilisés en apprentissage (snoRNA C/D ou H/ACA placés dans leur contexte de séquences inter-ORF) est capable de retrouver de façon spécifique les snoRNA au sein de toutes les séquences inter-ORF analysées chez cette même levure. Avec les paramètres utilisés pour la visualisation graphique (lissage de 11), leurs signaux sont très clairs et seuls quelques courts signaux supplémentaires (taille variant de 13nt à 54nt) sont détectés. Ces signaux peuvent correspondre soit à des faux positifs (séquences répétitives) soit à des signaux de séquences qui feront l'objet d'études ultérieures pour découvrir s'il s'agit de snoRNA non encore caractérisés.

Dans les séquences analysées, les snoRNA non utilisés en apprentissage (10 snoRNA chevauchant ou inclus dans des ORF) ont été laissés pour vérifier si la SVM entraînée pouvait les mettre en évidence. Les signaux ont été analysés aux positions attendues, mais ils ne répondent pas aux critères d'une séquence snoRNA tels que déterminés par la SVM. Il sont donc considérés comme différents de ceux utilisés pour l'apprentissage. Leurs séquences (contenues dans des ORF) sont sans doute trop différentes pour êtres détectées.

Dans un travail publié récemment, l'équipe de Eddy **[20]** a présenté trois nouveaux H/ACA potentiels chez *S. cerevisiae* (RUF1, RUF2 et RUF3) contenus dans les séquences inter-ORF.

L'analyse aux positions attendues n'a pas permis de les trouver. Le système est donc particulièrement spécifique aux séquences utilisées en apprentissage. Il sera donc nécessaire de relâcher les critères utilisés en apprentissage pour pouvoir détecter ces nouveaux représentants de la famille des H/ACA. La « propreté » des signaux actuellement obtenus permet d'envisager ce relâchement sans trop de craintes concernant les faux positifs. Cependant, une étude réalisée sur les seuls H/ACA doit être menée pour s'assurer que l'étude menée conjointement sur les snoRNA de deux classes différentes (C/D et H/ACA) ne diminue pas la possibilité de trouver des candidats d'une classe donnée.

4.3 Application à d'autres génomes – Analyse sur génomes d'Archaea

Afin de tester si la méthode est applicable à d'autre génomes, l'approche a été testée sur des fichiers de séquences inter-ORF issues des génomes d'Archaea du genre *Pyrococcus* pour lesquels 3 génomes sont entièrement séquencés : *P. abyssi*, *P. furiosus* et *P. horikoshii*. Ce n'est que très récemment que des sRNA de la classe des H/ACA ont été mis en évidence dans les génomes d'Archaea (**[21]** et **[22]**). Par une approche basée sur l'analyse de similarités de séquences inter-ORF entre les 3 génomes du genre *Pyrococcus* cités précédemment, S. Muller a pu identifier 6 sRNA de classe H/ACA retrouvés dans ces trois génomes (Muller *et al.* Article en préparation). C'est à partir de ce jeu de données connues pour ces trois génomes que la technique a pu être testée.
L'analyse a été réalisée dans un premier temps en faisant un apprentissage sur le génome de *P. abyssi* (6 séquences de sRNA de classe H/ACA placées dans leur contexte de séquence inter-ORF) et en testant les deux autres génomes de manière à vérifier la généralisation de la méthode entre génomes apparentés. Dans un deuxième temps, les séquences de tous les sRNA des 3 génomes (18 séquences) ont été regroupées dans l'échantillon d'apprentissage pour améliorer le système de reconnaissance et rechercher d'éventuels signaux supplémentaires dans ces 3 génomes (cf. Figure 6 et Tableau 1).

Apprentissage sur *P. abyssi*

Dans le cas d'un apprentissage réalisé sur 4/5 des 6 sRNA H/ACA de *P. abyssi*, ceux-ci sont bien retrouvés au sein de l'ensemble des séquences inter-ORF de *P. abyssi* et l'analyse des génomes de *P. furiosus* et *P. horikoshii* a permis de retrouver respectivement 6/6 et 5/6 des sRNA de classe H/ACA de ces deux génomes aux positions attendues (cf. Figure 6) et confirme ainsi les résultats obtenus par l'analyse comparative. Dans le cas de *P. horikoshii*, un des sRNA (équivalent de Pab91) ne peut être retrouvé puisqu'il chevauche une ORF contiguë non prise en compte dans cette analyse réalisée sur les seules séquences inter-ORF.
Il est à noter que les signaux correspondant aux sRNA équivalents de Pab91 et de Pab160 chez *P. furiosus*, et équivalents de 105 et de Pab160 chez *P. horikoshii* sont relativement peu marqués bien que significatifs. Cela est dû à leurs séquences généralement courtes et probablement à des différences de séquence plus marquées que pour les autres sRNA vis-à-vis de *P. abyssi*.
Dans les conditions de test et pour le serveur de calcul utilisé[3], la durée de l'apprentissage est de 10 mn tandis que la durée d'une analyse (sens et antisens simultanément) d'un génome est comprise entre 6 et 10 mn.

Apprentissage Multipyrococcus

Dans le cas d'un apprentissage réalisé sur 4/5 des 18 sRNA H/ACA des trois *Pyrococcus* regroupés (cf. Figure 6), il est observé une augmentation du signal sRNA en sortie des SVM qui correspond à un relâchement dans la définition des sRNA utilisés pour l'apprentissage (classe sRNA plus importante en apprentissage). Cela permet d'augmenter la performance de la recher-

[3] cf. données techniques.

che puisque les signaux de tous les sRNA sont plus fortement marqués (cas des sRNA peu visibles dans l'étude précédente) et par ailleurs d'amplifier les signaux d'autres régions pouvant s'avérer être des séquences intéressantes (cf. Tableau 1).
L'analyse fine de ces signaux sera réalisée en collaboration avec S. Muller du MAEM qui travaille spécifiquement sur ces génomes.
Dans les conditions de test et pour le serveur de calcul utilisé[3], la durée de l'apprentissage est d'environ 30 mn tandis que la durée d'une analyse (sens et antisens simultanément) d'un génome est comprise entre 20 et 30mn.

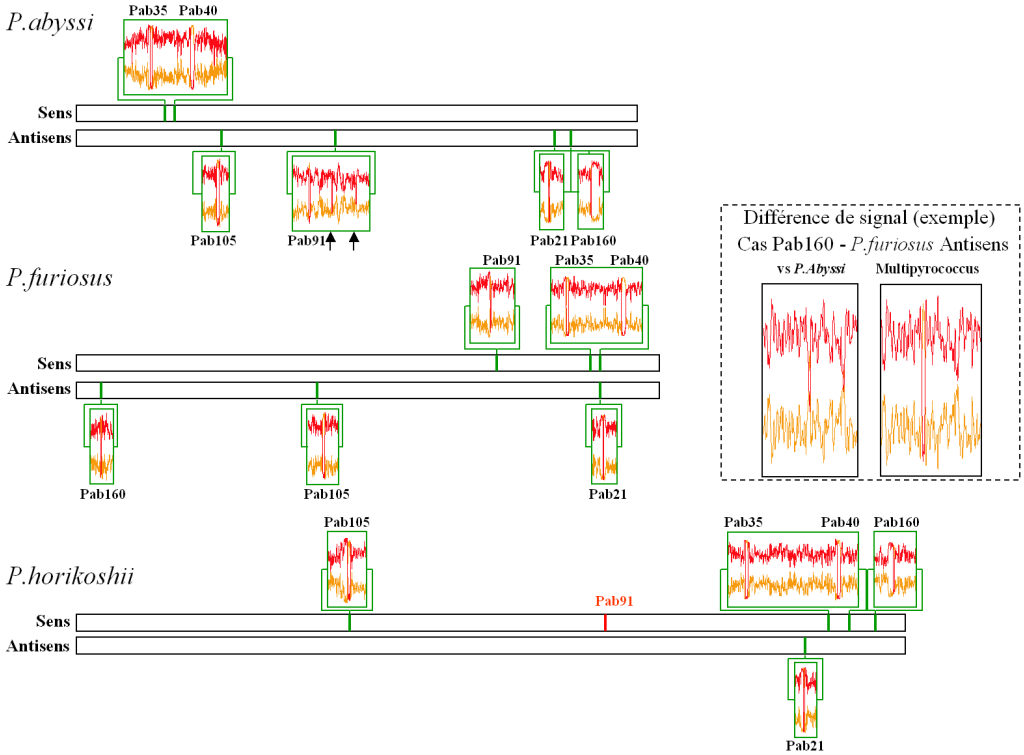


Figure 6 : Analyse des génomes d'Archaea du genre *Pyrococcus*. Les signaux détectés par SVM sont indiqués en vis-à-vis de leurs positions relatives sur les séquences Sens et Antisens (résultant de la concaténation de séquences inter-ORF) des génomes de 3 Archaea *P. abyssi*, *P. furiosus* et *P. horikoshii*.

Les signaux présentés sont ceux obtenus avec la SVM entraînée sur 4/5 des 18 séquences sRNA mis en évidence par analyse comparative des 3 Archaea (Muller *et al.* – Article en préparation). Une SVM entraînée sur les seules séquences de *P. abyssi* (4/5) est capable de détecter ces mêmes signaux sur les trois génomes. Un exemple de différence de signal est donné pour un sRNA (Pab160) à signal faible chez *P. furiosus* (se reporter au texte). L'équivalent de Pab91 chez *P. horikoshii* ne peut être détecté dans cette analyse réalisée sur les séquences inter-ORF (séquence chevauchant l'ORF contiguë). Des signaux supplémentaires détectés par la SVM et ne correspondant pas aux sRNA connus sont obtenus (voir par exemple sur la droite du signal du sRNA Pab91 chez *P. abyssi*).

	P.Abyssi		P.Furiosus		P.Horikoshii	
Nombre de Séquences interORF	912		1054		929	
Nombre de nucléotides	137058		142475		202421	
Nombre de fenêtres de 11 nucléotides	137048		142465		202411	
Signal sRNA recherché (%) (Sens et Antisens)	0,25	0,22	0,27	0,18	0,25	0,03

Paramètres des apprentissages	**Durée (mn)**	**Rapport Dual/Primal**	**Taux de reconnaissance (%)**
Apprentissage P.abyssi	10	99,45	78,96
Apprentissage Multipyrococcus	27	95,01	89,74

Résultats SVM

Vs P.abyssi

	P.Abyssi		P.Furiosus		P.Horikoshii		
Durée de l'analyse (mn)	6		7		10		
	Sens	**Antisens**	**Sens**	**Antisens**	**Sens**	**Antisens**	
Taux de reconnaissance (%) - Global	**96,58**	**96,64**	**97,11**	**97,28**	**96,84**	**96,77**	1
Erreur (%) - Global	3,42	3,36	2,89	2,72	3,16	3,23	2
(B) Erreur sur Classe 1 (%) - Global	3,40	3,34	2,83	2,67	3,08	3,23	3
Erreur sur Classe 1 (%)	3,41	3,35	2,84	2,68	3,09	3,23	4
Erreur sur Classe 2 (%)	**10,03**	**10,47**	**23,25**	**25,29**	**33,99**	**16,95**	5
(A) Taux de signal sRNA (%) - Global	3,62	3,53	3,04	2,81	3,24	3,25	6
Delta (A) - (B)	0,22	0,19	0,21	0,13	0,17	0,02	7

Vs Multipyroccocus

	P.Abyssi		P.Furiosus		P.Horikoshii		
Durée de l'analyse (mn)	20		22		33		
	Sens	**Antisens**	**Sens**	**Antisens**	**Sens**	**Antisens**	
Taux de reconnaissance (%) - Global	**95,70**	**95,83**	**96,54**	**96,57**	**96,19**	**96,21**	1
Erreur (%) - Global	4,30	4,17	3,46	3,43	3,81	3,79	2
(B) Erreur sur Classe 1 (%) - Global	4,29	4,15	3,45	3,42	3,80	3,79	3
Erreur sur Classe 1 (%)	4,30	4,16	3,46	3,43	3,81	3,79	4
Erreur sur Classe 2 (%)	**3,83**	**6,08**	**3,86**	**5,45**	**5,14**	**0,00**	5
(A) Taux de signal sRNA (%) - Global	4,53	4,36	3,72	3,59	4,04	3,82	6
Delta (A) - (B)	0,24	0,20	0,26	0,17	0,24	0,03	7

Tableau 1 : Résultats de l'analyse SVM des génomes d'Archaea. Le tableau présente les résultats obtenus lors des analyses SVM sur les génomes d'Archaea du genre *Pyrococcus* (*P. abyssi*, *P. furiosus* et *P. horikoshii*). En haut du tableau sont données les caractéristiques des séquences inter-ORF analysées (concaténations des séquences). Les paramètres des apprentissages (4/5 de l'échantillon de départ) sont indiqués ensuite pour les deux analyses SVM (apprentissage *P. abyssi* ou Multipyrococcus). Le **Rapport dual/primal** est un paramètre interne de la SVM indiquant l'état du système à la fin de l'apprentissage (tend vers 100). Le **Taux de reconnaissance** correspond à la reconnaissance réalisée sur l'échantillon sur lequel la SVM n'a pas appris (1/5 de l'échantillon de départ).

La partie basse du tableau donne les résultats chiffrés des SVM. **Le Taux de reconnaissance** correspond au résultat fourni par la SVM par rapport à la séquence analysée où les positions réelles des sRNA ont été reportées (lignes 1). **L'Erreur** est directement tirée de la valeur précédente (erreur de classification sur inter-ORF et sur sRNA - lignes 2). Le détail des erreurs sur les classes est donné dans les lignes 3 (ramené au nombre total de fenêtres analysées), 4 et 5 (ramené au nombre de fenêtres bien affectées). **Le Taux de signal sRNA** (lignes 6) correspond au signal sRNA connu ou inconnu détecté par la SVM (nombre de fenêtres affectées en classe 2 sur le nombre total de fenêtres). La valeur **Delta (A) - (B)** donne le signal sRNA obtenu par la SVM aux positions attendues et est à comparer aux valeurs théoriques (Signal sRNA recherché (Sens et Antisens) données dans le haut du tableau). La prise en compte de plus de séquences en apprentissage augmente l'erreur sur les séquences inter-ORF (augmentation du nombre de faux positifs ou découverte de signaux supplémentaires d'intérêt – Comparer les lignes 4) mais permet d'améliorer la reconnaissance des sRNA (diminution de l'erreur spécifique sur ces séquences - Comparer les lignes 5).

5 Discussion

Un des intérêts de l'approche par les SVM est qu'il s'agit d'une méthode souple et efficace permettant de rechercher non seulement les gènes de snoRNA mais également d'autres classes de gènes d'ARNnc. En effet, la technique permet de discriminer des classes (objets), sans faire d'a priori sur leur structure/organisation etc. Elle utilise actuellement les seules données de séquences (environnement immédiat d'un nucléotide permettant de lui affecter la classe à laquelle il appartient), bien qu'il soit envisageable de prendre également en compte des données de structures secondaires, ou d'envisager la conception d'un noyau dédié aux séquences.

Par ailleurs, d'après nos premiers tests, il a été montré que l'approche développée était exhaustive dans la recherche des snoRNA sur le génome complet *de S. cerevisiae* (snoRNA utilisés pour l'apprentissage de la SVM). Nous pouvons comparer ce résultat à celui obtenu avec un logiciel dédié aux snoRNA à boîtes H/ACA récemment proposé et testé sur ce même génome qui présente des faux négatifs pour un nombre de faux positifs raisonnable **[23]**.

Enfin, les temps de calcul sont raisonnables et du même ordre de grandeur que la technique citée ci-dessus prenant en compte les structures secondaires des ARNnc pour la recherche des gènes de snoRNA à boîtes H/ACA (15000nt/h/GHz pour l'étude sur *S. cerevisiae).* Cependant, dans le cas de l'approche par SVM la recherche des seuls snoRNA H/ACA (sans la classe des C/D) aurait permis de diminuer les temps de calcul. L'utilisation d'une version plus récente du logiciel de la M-SVM permettra également de diminuer de façon importante les temps de calcul.
Enfin, la technique est généralisable. Elle a ainsi été testée avec succès sur d'autres génomes (Archaea) et a permis de démontrer la possibilité de rechercher une classe de sRNA présente dans un génome de référence, dans d'autres génomes proches sur le plan phylogénétique. La prise en compte de nouvelles molécules découvertes dans ces génomes permet d'enrichir l'échantillon d'apprentissage, d'augmenter la signature des molécules et éventuellement d'en trouver de nouvelles.

6 Perspectives

La récente publication des génomes de 3 autres levures **[24]** va permettre d'appliquer la technique développée à la recherche de snoRNA dans ces levures. Par ailleurs, un fichier de données inter-ORF réalisé à partir des données de la banque Génolevures, permettra de faire une recherche de snoRNA dans les 13 espèces représentatives de la classe des Hémiascomycètes concernées par le projet Génolevures **[12]**.
Un développement à court terme de cette approche par les SVM est la prise en compte des données complètes des génomes (ORF et introns ajoutés aux séquences inter-ORF actuellement étudiées). Considérer la totalité d'un génome est en effet préférable pour son annotation puisqu'il est connu que l'organisation et l'expression des gènes des snoRNA peuvent être très différentes suivant les génomes analysés. Ainsi, la localisation principale chez les vertébrés des gènes de snoRNA correspond aux introns tandis que chez *S. cerevisiae*, la majorité de ces ARN sont codés par des séquences inter-ORF, plus rarement par des introns.
Pour répondre au problème posé, une classe supplémentaire devra être introduite (classe 3 pour les ORF), et ce problème de discrimination à catégories multiples exploitera pleinement les SVM multi-classes (M-SVM) conçues et implémentées au LORIA **[11]**. Ceci présentera plusieurs avantages parmi lesquels : (i) éviter un travail laborieux de constitution de banques de séquences inter-ORF utilisées pour l'approche SVM, (ii) retrouver/découvrir des ARNnc qui ne se situent pas dans les séquences inter-ORF, (iii) extraire directement des informations telles que la distribution d'une classe d'ARNnc dans les ensembles inter-ORF, ORF, introns.

7 Données techniques

Serveur de Calcul
DELL Precision 530MT - Bi-processeur Xéon HyperThreading (2 processeurs physiques et 2 processeurs logiques) : 2.8GHz - RDRAM : 4 Go PC800 ECC – Disque Dur : SCSI U160 – 15000rpm
Paramétrage SVM
La SVM est utilisée avec un noyau gaussien, le paramètre C de la fonction objectif fixé à 10.0, et la taille du *chunk* fixée à 20 (se reporter à la documentation technique de la SVM).
Le programme M-SVM **[25]** est disponible sur la page http://www.loria.fr/~guermeur/
Paramètres analyse.py [19]
Modifications du programme pour avoir une visualisation sur une fenêtre de 11nt.

8 Bibliographie

1 - Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S. **Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.** Curr Biol. 2001 Jun 26;11(12):941-50.

2 - Vogel J, Bartels V, Tang TH, Churakov G, Slagter-Jager JG, Huttenhofer A, Wagner EG. **RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria.** Nucleic Acids Res. 2003 Nov 15;31(22):6435-43.

3 - Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. **Identification of novel small RNAs using comparative genomics and microarrays.** Genes Dev. 2001 Jul 1;15(13):1637-51

4 - Huttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J. **RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse.** EMBO J. 2001 Jun 1;20(11):2943-53.

5 - Vapnik VN. **Statistical learning theory**. John Wiley & Sons, Inc., N.Y. – 1998

6 - Vapnik VN. **Estimation of dependences based on empirical data.** Springer-Verlag, New York, 1982.

7 - Boser BE, Guyon IM, Vapnik V. **A training algorithm for optimum margin classifiers**. COLT'92 - 1992:144-152

8 - Cortes C, Vapnik VN. **Support vector networks.** ***Machine Learning*** 1995, 20:1-25

9 - Sun YF, Fan XD, Li YD. **Identifying splicing sites in eukaryotic RNA: support vector machine approach**. - Comput Biol Med. 2003 Jan;33(1):17-29.

10 - Wassarman KM, Zhang A, Storz G. **Small RNAs in Escherichia coli.** Trends Microbiol. 1999 Jan;7(1):37-45. Review.

11 - Dennis PP, Omer A, Lowe T. **A guided tour: small RNA function in Archaea.** Mol Microbiol. 2001 May;40(3):509-19. Review.

12 - FEBS Letters - Numéro Spécial n°487, issue 1 - 22 Décembre 2000

13 - Bachellerie JP, Cavaille J, Huttenhofer A. **The expanding snoRNA world**. Biochimie. 2002 Aug;84(8):775-90.

14 - Lowe TM, Eddy SR. **A computational screen for methylation guide snoRNAs in yeast.** - Science. 1999 Feb 19;283(5405):1168-71.

15 - Elena Rivas and Sean R. Eddy **Noncoding RNA gene detection using comparative sequence analysis -** BMC Bioinformatics. 2001; 2 (1): 8

16 - Gautheret D, Lambert A. **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. -** J Mol Biol. 2001 Nov 9;313(5):1003-11.

17 - Rivas E, Eddy SR. **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. -** Bioinformatics 2000 Jul;16(7):583-605

18 - http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html

19 - Billaut S. **–Recherche de motifs de régulation de l'épissage alternatif : apprentissage statistique sur HIV-1** – Rapport de Maîtrise de Biologie Cellulaire et Physiologie - UHP - Nancy I - 2003

20 - McCutcheon JP, Eddy SR. **Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res.** 2003 Jul 15;31(14):4119-28.

21 - Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A**. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus.** Proc Natl Acad Sci U S A. 2002 May 28;99(11):7536-41.

22 - Rozhdestvensky TS, Tang TH, Tchirkova IV, Brosius J, Bachellerie JP, Huttenhofer A. **Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea.** Nucleic Acids Res. 2003 Feb 1;31(3):869-77.

23 - Edvardsson S, Gardner PP, Poole AM, Hendy MD, Penny D, Moulton V. **A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction.** Bioinformatics. 2003 May 1;19(7):865-73.

24 - Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. **Sequencing and comparison of yeast species to identify genes and regulatory elements.** Nature. 2003 May 15;423(6937):241-54.

25 - Guermeur Y. - **Combining discriminant models with new multi-class SVMs -** Pattern Analysis and Applications. 2002 - 5 (2):168-179.