UNE APPROCHE BIO-INFORMATIQUE INTÉGRÉE POUR L'IDENTIFICATION DES CIBLES ARN DE L'ENDORIBONUCLÉASE III CHEZ LA LEVURE

par

Jules Gagnon

thèse présentée au Département de biologie en vue
de l'obtention du grade de docteur ès sciences (Ph.D.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, octobre 2014

Le 28 octobre 2014

le jury a accepté la thèse de Monsieur Jules Gagnon dans sa version finale.

Membres du jury

Professeur Daniel Lafontaine
Directeur de recherche
Département de biologie

Professeur Sherif Abou Elela
Codirecteur de recherche
Département de microbiologie et d'infectiologie

Professeur Gabriel Girard
Codirecteur de recherche
Département d'informatique

Professeur Mathieu Blanchette
Évaluateur externe
Université McGill

Professeur Ryszard Brzezinski
Président-rapporteur
Département de biologie

SOMMAIRE

Les endoribonucléases III sont conservées chez tous les eucaryotes. Ils jouent un rôle important dans le cycle de vie des acides ribonucléiques (ARN) que ce soit au niveau de leur maturation, leur régulation ou leur dégradation. Cependant, seul un petit nombre d'ARN ciblés par les ribonucléases (RNases) III sont connus et les motifs reconnus par l'enzyme sont encore mal caractérisés. Actuellement, la découverte de nouvelles cibles repose principalement sur la validation *in vitro* de gènes individuels. Il est important d'avoir une vue d'ensemble des cibles des RNases III pour comprendre le rôle de la dégradation spécifique des ARN dans le métabolisme cellulaire.

Ainsi, cette thèse a comme objectif de développer des approches haut débit pour permettre une identification plus rapide des cibles tout en minimisant l'utilisation des ressources expérimentales. Elle présente l'utilisation combinée d'approches bio-informatiques, d'étude génétique de l'expression et de traitement *in vitro* dans le but d'avoir un portrait global des cibles de l'endoribonucléase III. Elle a aussi comme but d'identifier les motifs d'ARN non codants qui guident la reconnaissance par l'endoribonucléase III.

Les principaux accomplissements de cette recherche sont : le développement de nouveaux algorithmes de prédiction des cibles de la RNase III, la détection de deux nouvelles classes de transcrits dont l'expression est dépendante de la RNase III, l'identification de quelques centaines de nouvelles cibles de la RNase III, la production d'un catalogue plus complet des motifs coupés par la RNase III et l'identification de nouvelles catégories de motifs coupées par la RNase III. Le tout procure un portrait global de l'impact de la RNase III sur le transcriptome et le cycle de vie des ARN.

De plus, ce travail montre comment une approche intégrée incluant la recherche *in silico*, *in vivo* et *in vitro* permet de mieux comprendre le rôle d'un enzyme dans la cellule et comment chaque approche peut pallier les déficiences des autres approches et fournir globalement des résultats plus complets.

Mots-clés: ARN ; ARNnc ; ARNsno ; RNT1 ; Rnt1p ; RNase III ; Apprentissage machine ; Puces à ADN

REMERCIEMENTS

Je voudrais tout d'abord remercier Sherif Abou Elela de m'avoir accordé sa confiance et m'avoir permis d'accomplir ce travail. Je tiens également à remercier Daniel Lafontaine d'avoir accepté d'encadrer mon cheminement et Gabriel Girard pour son support pour tout ce qui a trait à la bio-informatique. La contribution de Hervé Philippe et François Major à titre de conseillers a également joué un rôle important dans la progression de mes recherches. La contribution financière du Fonds de Recherche en Santé du Québec (FRSQ) fut aussi importante pour la réalisation de ce projet.

Je tiens aussi à dire un merci tout spécial à Ghada Ghazal et Mathieu Lavoie pour leur excellent travail de validation et pour les nombreux projets communs que nous avons eux. Sans eux, mes travaux seraient restés théoriques et leur impact n'aurait pas été le même.

Finalement, je voudrais remercier tous les étudiants présents et passés du laboratoire de même que les étudiants stagiaires et l'équipe bio-informatique du laboratoire de génomique fonctionnel pour leur participation, leurs questions, leurs commentaires et leur amitié.

Je dédie ce travail à la mémoire de mon père, Lionel Gagnon (1937-2014), pour tous ses sacrifices et tous ses efforts qui m'ont permis d'arriver là où j'en suis.

TABLE DES MATIÈRES

ABRÉVIATIONS

ADN Acide déoxiribonucléique

ADNc ADN complémentaire

ARN Acide ribonucléique

ARNmi micro ARN

ARNpi ARN interagissant avec Piwi

ARNr ARN ribosomal

ARNsi petit ARN interférant

ARNsno petit ARN nucléolaire

ARNt ARN de transfert

ARNtasi ARNsi agissant en trans

CUT Transcrit cryptique instable

PCR Réaction de polymérisation en chaîne

RISC Complexe de répression médié par l'ARN

RNase Ribonucléase

RT-PCR PCR avec transcription inverse

SCFG Grammaire stochastique sans contexte

sqRT-PCR RT-PCR semi-quantitatif

SUT Transcrit stable non annoté

SVM Machine à vecteurs de support

XUT Transcrit instable sensible à XRN1

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1

INTRODUCTION

Le dogme central de la biologie moléculaire (Figure 1.1) (Crick, 1958) postule que le transfert de l'information génétique s'effectue de façon unidirectionnelle de l'acide désoxyribonucléique (ADN) vers l'acide ribonucléique (ARN) puis de l'ARN vers les protéines. Ainsi, selon ce dogme, l'ADN a principalement un rôle de stockage de l'information génétique et l'ARN a plutôt comme fonction de transporter cette information. Plusieurs découvertes des dernières décennies remettent en question ce dogme et montre que l'ARN est plus qu'un simple messager. Ainsi, l'étude du cycle de vie de l'ARN est d'une importance capitale pour parvenir à une meilleure connaissance de la biologie cellulaire et des mécanismes de la vie.

figure 1.1 – Dogme central de la biologie moléculaire.

1.1 ARN

L'ARN, comme l'ADN, est un polymère de nucléotides. Les bases qui le composent sont l'adénine, la guanine, la cytosine et l'uracile. Il est synthétisé par l'ARN polymérase à partir d'un gabarit ADN. La première extrémité synthétisée par l'ARN polymérase est appelée extrémité 5' alors que la fin de la transcription s'effectue à l'extrémité 3'. Contrairement à l'ADN qui est habituellement présent sous forme d'une hélice constituée de deux brins complémentaires, l'ARN est le plus souvent retrouvé sous forme d'une molécule simple brin. Cependant, tout comme l'ADN, les nucléotides de l'ARN ont la capacité de s'apparier. Ainsi, les molécules d'ARN sont généralement repliées sous une forme thermodynamiquement stable.

La fonction la plus connue de l'ARN est la transmission de l'information génétique des chromosomes vers les ribosomes pour effectuer la synthèse des protéines par l'intermédiaire des ARN messagers (ARNm). Cependant, dès les années 60, plusieurs chercheurs ont démontré que l'ARN pouvait aussi catalyser certaines réactions chimiques, donc que certains ARN peuvent avoir une fonction autre que d'être un simple messager. Cela a amené Carl Woese à postuler qu'un monde où la vie était basée sur l'ARN a pu exister avant l'apparition du monde basé sur les protéines que l'on connaît (Carl, 1968). Il est maintenant connu que plusieurs types d'ARN existent avec des rôles très variés et des fonctions bien plus complexes que la simple transmission d'informations génétiques.

1.1.1 ARN messagers

Les ARN messagers forment la classe d'ARN la plus diversifiée en fonction du nombre de molécules différentes. Chez un organisme aussi simple que la levure du boulanger (*Saccharomyces cerevisiæ*), on compte plus de 6000 transcrits ARNm différents, alors que chez l'homme ce nombre dépasse 20 000.

Chez les eucaryotes, les ARNm sont synthétisés dans le noyau par la polymérase II. La polymérase est recrutée au promoteur par des facteurs de transcription. L'extrémité 5' du transcrit est protégée par une coiffe 7-méthylguanosine. La transcription se poursuit jusqu'à une région terminatrice où d'autres facteurs coupent le transcrit ARN en formation et, ainsi, provoquent la libération de la polymérase. Une queue de polyadénine est ajoutée à l'extrémité 3' produite par cette coupure.

Pour les ARNm contenant des introns, la maturation de l'ARNm est complétée par le complexe d'épissage qui permet d'enlever les introns. Une fois l'ARNm mature est exporté vers le cytoplasme et les ribosomes par les pores nucléaires.

1.1.2 ARN non codant

Il existe une grande variété d'ARN non codants (ARNnc). Ils sont aussi appelés ARN fonctionnels, car malgré le fait qu'ils ne codent pas pour une protéine, ils ont quand même la capacité d'exercer une fonction sous leur forme ARN. Plusieurs de ces ARN effectuent directement une fonction catalytique. D'autres ont plutôt un rôle structural dans un complexe ARN-protéine. Finalement, plusieurs agissent comme guide pour des protéines grâce à leur capacité à s'hybrider avec des nucléotides complémentaires sur une molécule d'ARN ou d'ADN.

Les ARN les plus abondants en termes de contenu total sont les ARN ribosomaux. Ils forment environ 90 % des ARN totaux de la cellule. Composant des ribosomes, il effectue la synthèse des protéines (Cech, 2000). Ils sont synthétisés par la polymérase I à partir de nombreuses répétitions chromosomiques. Leur maturation s'effectue à partir d'un grand transcrit (35S) en plusieurs plus petits transcrits (25S, 18S, 5.8S). Cette maturation comporte plusieurs étapes et plusieurs transcrits intermédiaires (Fromont-Racine $\mathrm{et\ al.}$, 2003).

Une autre classe d'ARN essentiels au métabolisme de base de la cellule sont les ARN de transfert (ARNt). Les ARN de transfert transportent les acides aminés aux ribosomes et assurent la traduction fidèle du code génétique (Felsenfeld et Cantoni, 1964). Ils sont présents en plusieurs copies dans le génome. Ce sont de petits ARN qui sont transcrits par la polymérase III et maturés par la ribonucléase P (RNase P) avant d'être exportés dans le cytoplasme.

Au cours de la maturation des ARNr, plusieurs nucléotides sont modifiés à des positions spécifiques grâce à des ARN guides appelés petits ARN nucléolaires (ARNsno). Deux types d'ARNsno existent : à boîtes C/D et à boîtes H/ACA qui guident respectivement la méthylation et la pseudourydilation de l'ARNr. Ces ARN existent chez tous les eucaryotes. Bien que les modifications guidées par ces ARN ne soient pas essentielles, elles peuvent donner un avantage à la cellule dans certaines conditions (Esguerra $\mathrm{et\ al.}$, 2008). Les ARNsno à

boîtes C/D sont caractérisés par une courte tige d'ARN double brin reliant les boîtes C et D qui ont respectivement 7 et 4 nucléotides et sont séparés par une région peu structurée. La séquence guide se situe en amont de la boîte D ou d'une boîte D' (Kiss-László $\mathrm{et\ al.}$, 1998). La structure des ARNsno à boîtes H/ACA est différente. Elle consiste en deux tiges-boucles séparées par la boîte H et suivies par la boîte ACA. La ou les séquences guides se situent à l'intérieur de boucles internes comprises dans les tiges (Ganot $\mathrm{et\ al.}$, 1997). En plus des modifications de l'ARNr, certains ARNsno sont connus pour guider des modifications sur des ARNsn (Jády et Kiss, 2001) et même des ARN messagers (ARNm) (Kishore et Stamm, 2006). L'expression des ARNsno peut se faire par un transcrit indépendant, par un transcrit comptant plusieurs ARNsno ou à l'intérieur d'un intron.

Il existe plusieurs autres types d'ARN non codants. Il y a les ARN des ribonucléases P et MRP qui participent respectivement à la maturation des ARN de transfert et des ARN ribosomaux (Pannucci $\mathrm{et\ al.}$, 1999; Woodhams $\mathrm{et\ al.}$, 2007). L'ARN de la télomérase est nécessaire au maintien des télomères et de l'intégrité des chromosomes (Lustig, 1999). Les petits ARN nucléaires (ARNsn) sont nécessaires à l'épissage des ARN messagers. Il existe aussi plusieurs classes d'ARN fonctionnels qui sont impliquées dans la régulation génétique (ARN antisens, ARNsi, ARNmi, ARNtasi, ARNpi). Leur importance est telle que les petits ARN furent nommés découverte de l'année en 2002 par le magazine Science. Chez l'humain, un bon nombre d'ARNnc n'a pas de fonction connue et plusieurs restent à découvrir. De plus, l'expression de plusieurs ARNnc est altérée dans certaines maladies, dont le cancer.

Actuellement, la découverte de nouveaux ARNnc repose principalement sur la validation de gènes individuels. Il est important d'avoir une vue d'ensemble du cycle de vie des ARN pour utiliser toute l'information disponible pour faciliter l'identification de nouveaux ARNnc. De plus en plus, des techniques d'identification haut débit sont utilisées.

1.2 Identification haut débit d'ARN non codants

Une grande variété de techniques ont été utilisées pour identifier de nouveaux ARN non codants. Chacune de ces méthodes a démontré son potentiel pour l'identification de nouveaux ARN non codants. Cependant, elles ont aussi des limitations importantes. Ces méthodes peuvent être regroupées en trois grandes catégories. Les approches transcriptomiques détectent directement les ARNnc parmi tous les transcrits produits par la cellule. Les approches d'enrichissement utilisent différentes approches pour obtenir un échantillon d'ARN contenant plus d'ARNnc qu'un extrait d'ARN total. Les approches *in silico* prédisent des ARNnc d'après leurs caractéristiques de séquence et de structure.

1.2.1 Approches transcriptomiques

Les approches de cartographie du transcriptome, que ce soit par puces à ADN couvrant tout le génome ou par séquençage à haut débit, permettent d'identifier de nombreux nouveaux ARN non codants (David $\mathrm{et\ al.}$, 2006; Neil $\mathrm{et\ al.}$, 2009). Cependant, ces approches ne fournissent pas d'information précise sur la fonction et l'origine de ces transcrits. De plus, elles ont révélé qu'une grande proportion des régions non codantes du génome sont transcrites (Johnson $\mathrm{et\ al.}$, 2005). De plus, ces approches sont limitées par les niveaux d'expression des transcrits et par les conditions de culture.

Puces à ADN

Il existe plusieurs technologies de puces à ADN qui varient tant par le processus de synthèse que par la longueur des oligonucléotides et leur densité. La technologie utilisée par Affymetrix est celle qui permet la plus grande densité : une surface d'un peu plus de 1 cm^2 peut contenir plus de 6 millions de sondes de 25 nucléotides chacune. Les sondes sont syn-

thétisées directement sur la surface par un procédé de photolithographie. Les puces à ADN couvrant le génome de *S. cerevisiæ* comptent 3.2 millions de sondes parfaitement complémentaires à 25 nucléotides du génome. Elles couvrent tout le génome avec un décalage moyen de quatre nucléotides et sont complémentaires au brin Crick du génome.

À la suite de l'extraction de l'ARN de la culture à étudier, une réaction de transcription inverse est effectuée pour synthétiser un brin d'ADN complémentaire à l'ARN source. Ensuite, si nécessaire, une étape de polymérisation de l'ADN peut être effectuée si l'on désire obtenir les deux brins d'ADN. L'ARN est ensuite dégradé et l'ADN est fragmenté pour avoir une hybridation plus uniforme. Les fragments d'ADN sont marqués à une extrémité avec une molécule de biotine. L'ADN marqué est hybridé avec la puce à ADN et un anticorps qui cible la biotine est finalement ajouté pour permettre la lecture de la puce à ADN par un laser. Une image numérique de la puce à ADN est obtenue et est convertie en données brutes d'intensité pour chacune des sondes.

Chaque puce ne peut être utilisée que pour un seul échantillon. Comme chaque échantillon ne contient pas exactement la même concentration d'ADN et que les conditions d'hybridation peuvent varier, il est nécessaire de faire une normalisation des intensités obtenues avant de pouvoir comparer les résultats provenant de plusieurs puces. De plus, les sondes ont des propriétés différentes selon leurs compositions. Pour corriger les variations dues aux sondes, une hybridation de l'ADN génomique est faite. Pour chaque sonde, un facteur de correction est calculé pour rendre le niveau des sondes égal. Une étape de soustraction du bruit de fond est aussi nécessaire pour éliminer les effets provenant de l'hybridation non spécifique.

L'analyse des puces couvrant tout le génome demande aussi une étape de segmentation. Cette étape permet d'identifier des segments où l'expression est uniforme, ce qui devrait correspondre à un transcrit. L'approche préconisée pour la segmentation se base sur une modélisation de changements structuraux pour identifier les points de changement, les extrémités des segments. Ce modèle a comme paramètre une taille maximale d'un segment et le nombre maximal de segments.

D'autres fabricants de puces à ADN possèdent des technologies ayant des caractéristiques différentes comme sondes plus longues, des puces réutilisables ou des sensibilités différentes. Pour la tâche d'identification d'ARNnc, la technologie d'Affymetrix est la plus appropriée et la plus utilisée compte tenu de sa haute résolution.

Séquençage à haut débit

De nouvelles technologies ont permis de diminuer de beaucoup le coût et le temps demandé par le séquençage d'ADN. Ces technologies permettent de séquencer de quelques millions à quelques milliards de molécules d'ARN en 1 à 10 jours. La disponibilité d'un aussi grand nombre de séquences permet de nouvelles applications.

Par exemple, pour l'étude de l'ARN, l'ARN cellulaire total peut être séquencé. Avec une faible couverture du génome, il est possible d'estimer le niveau d'expression des transcrits. Avec une plus forte couverture, il serait possible de faire une cartographie précise de tous les transcrits et aussi de comparer les niveaux d'expression des ARN entre deux échantillons. La précision et la sensibilité sont seulement dépendantes du nombre de séquences obtenues et il n'y pas d'effet de saturation ou d'hybridation croisée.

Il existe cependant un certain nombre de limitations. Puisqu'environ 90 % de l'ARN total est constitué d'ARN ribosomaux, le nombre de séquences nécessaires pour obtenir un bon niveau de couverture avec de l'ARN total est énorme. Il est donc nécessaire de diminuer la quantité d'ARN ribosomaux dans l'échantillon. Pour cela, deux techniques peuvent être utilisées : la dégradation de l'ARNr et l'enrichissement en séquences polyadénylées. Dans les deux cas, le processus affecte aussi certains autres ARN et introduit certains biais dans les résultats. D'autres biais peuvent aussi être introduits par la préparation de l'échantillon. Lors de la transcription inverse, certaines séquences peuvent être favorisées au dépens d'autres.

1.2.2 Approches d'enrichissement

Il existe d'autres approches qui ciblent plus spécifiquement les courts ARN ou bien les ARN ayant une extrémité 5' phosphate. Ces méthodes révèlent aussi de nombreux ARN non codants, mais encore une fois il est difficile de distinguer les ARN fonctionnels, les produits de dégradation et le bruit de fond transcriptionnel. Aussi, ce processus d'enrichissement ne fournit pas ou peu d'information sur le niveau d'expression des transcrits détectés.

Une de ces techniques consiste à séparer sur gel les transcrits selon leur taille. Puisque les ARNnc sont généralement plus courts que les ARNm, un enrichissement est obtenu. Il est aussi possible de chercher à identifier les transcrits dont l'expression change lors de l'inactivation de voies métaboliques spécifiques à la production d'ARNnc.

Une autre technique profite du fait que plusieurs ARNnc ne possèdent pas de coiffe 5' comme les ARNm. Une ligase ARN est utilisée pour attacher une séquence spécifique à l'extrémité 5' et cette séquence est utilisée pour faire un enrichissement spécifique des ARN ayant une extrémité 5' libre. Il est aussi possible d'utiliser une protéine recombinante qui se lie spécifiquement à une classe d'ARN pour en faire l'enrichissement.

Pour toutes ces méthodes d'enrichissement, il est ensuite nécessaire d'employer une méthode d'identification haut débit comme les puces à ADN ou le séquençage pour obtenir les résultats finaux. L'inconvénient majeur des approches d'enrichissement est d'être très ciblées pour une classe de transcrits spécifiques. Elles ne produisent donc pas une identification globale de tous les ARNnc.

1.2.3 Approches *in silico*

Toute prédiction *in silico* de transcrits implique l'utilisation d'une ou plusieurs techniques de classification. Il existe une grande variété de techniques qui peuvent être employées. Bien qu'il soit possible de créer un classificateur à partir d'un modèle bien défini, la plupart du temps les classificateurs sont basés sur un algorithme d'apprentissage machine. Pour l'identification de nouveaux ARNnc, l'apprentissage supervisé est le choix habituel puisque des exemples ARNnc sont disponibles.

En effet, les algorithmes d'apprentissage supervisé doivent être entraînés. Pour les entraîner, il est nécessaire de pouvoir diviser les données en classes connues, c'est-à-dire un groupe de données positives et un groupe de données négatives. Ainsi, elles ne sont pas adaptées pour tous les problèmes biologiques, car il est souvent difficile d'obtenir un ensemble représentatif de données, surtout de données négatives.

Un des problèmes les plus courants en apprentissage machine est le manque de capacité de généralisation, aussi appelé variance élevée ou surapprentissage. Dans ce cas, le modèle fait une bonne prédiction pour l'ensemble d'entraînement, mais est beaucoup moins efficace sur l'ensemble de validation. Cela est causé par un modèle trop complexe ou un poids trop élevé assigné à certains attributs non représentatifs.

Pour pallier ce problème, la plupart des algorithmes incluent des paramètres qui permettent de limiter la complexité du modèle en pénalisant les poids élevés. Il est aussi possible de réduire le nombre d'attributs. Cependant, lorsque cela est possible, la solution idéale est d'augmenter la taille de l'ensemble de données d'entraînement.

Le problème inverse peut aussi se présenter, soit un modèle qui a peu de succès sur les données d'entraînement ou sousapprentissage. Dans ce cas, la conclusion est que l'algorithme manque d'information pour effectuer la classification. Si possible, on voudra alors ajouter des attributs ou bien en créer de nouveaux, soit en augmentant l'ordre, soit en utili-

sant un algorithme qui permet la séparation non linéaire comme les réseaux neuronaux ou les SVM avec kernel non linéaires.

Classificateur bayésien

L'apprentissage d'un classificateur bayésien consiste à identifier les paramètres de distribution des variables pour chaque classe. Ensuite, en fonction de la proportion de chacune des classes, chaque élément sera assigné à la classe qui correspond à la plus forte probabilité postérieure.

Par exemple, pour deux classes équiprobables, l'une avec une moyenne de 5 et un écart type de 1, l'autre avec respectivement 8 et 2 comme moyenne et écart type. Une nouvelle donnée de valeur 6 aura une probabilité postérieure de 0.24 pour la première classe et 0.12 pour la deuxième classe selon la loi normale. Elle sera donc classée dans la première classe.

L'efficacité de cette technique est très dépendante de la qualité de l'estimation des paramètres. Les classificateurs bayésiens ont l'avantage de fournir une estimation statistique de la qualité de la prédiction c'est-à-dire la probabilité que la classification soit exacte. L'exactitude de cette estimation est fortement influencée par la taille de l'ensemble d'entraînement.

Le plus souvent, un classificateur naïf, qui assume l'indépendance des variables, est utilisé. Malgré cette supposition, souvent fausse, les classificateurs bayésiens naïfs sont très performants. De plus, cette supposition permet de diminuer la nécessité d'avoir un nombre important d'exemples lorsque le nombre de variables augmente. Plusieurs types de distribution peuvent être choisis pour modéliser les variables : distributions gaussiennes, distributions multinomiales, distributions de Bernoulli.

Les classificateurs bayésiens naïfs sont quelques fois utilisés pour l'identification de petits ARN non codants. Cependant, lorsque utilisé seul, ils génèrent un grand nombre de candidats faux positifs. Ils doivent donc être combinés avec d'autres approches comme la conservation interespèce (Yousef $\mathrm{et\ al.}$, 2006). De plus, puisqu'ils assument l'indépendance des variables, ils s'apparentent aussi aux classificateurs linéaires et aux régressions logistiques dans le sens où ils ne permettent pas directement l'utilisation de relations d'ordres supérieurs entre les variables.

Machine à vecteurs de support

Les "Support Vector Machines" (SVM) (Cortes et Vapnik, 1995) ont des similitudes avec les régressions et avec les réseaux neuronaux. Il est possible d'utiliser un SVM de façon linéaire. Dans le cas de deux classes séparables linéairement, la régression standard donnera une droite (n'importe laquelle) séparant les deux classes. Une SVM donnera la droite qui maximise la distance entre chaque élément et cette droite (droite = hyperplan, distance = marge). Plus la marge est grande, meilleure est la capacité théorique de généralisation.

Les SVM peuvent aussi créer des attributs pour effectuer une séparation non linéaire. La fonction permettant de créer ces attributs est appelée "kernel". La fonction "kernel" la plus courante est appelée "kernel" radial et est basée sur la distribution gaussienne. Dans ce cas, les nouveaux attributs sont les probabilités dans des distributions gaussiennes centrées sur des points de l'ensemble d'entraînement, ces points étant choisis de façon à obtenir une séparation des classes.

Les SVM sont à la base des classificateurs binaires non probabilistes, donc ils ne peuvent pas être utilisés pour plus de deux classes d'objets et ne donnent pas d'information sur la fiabilité de la classification. Il est cependant possible d'appliquer successivement plusieurs classificateurs SVM pour identifier plusieurs classes. Les classificateurs SVM, comme les réseaux neuronaux, peuvent tenir compte des relations entre les variables dans la sépara-

tion des classes. Un inconvénient de ces deux types de classificateurs est la difficulté d'en extraire les paramètres pour constituer un modèle biologique simple.

Un classificateur SVM permet de distinguer avec une grande efficacité les transcrits codants des transcrits non codants (Liu $\mathrm{et\ al.}$, 2006). Cependant, ce classificateur utilise un grand nombre de variables qui ne sont pas nécessairement applicables pour tous les types recherches, comme le nombre d'homologues et la fréquence des nucléotides. Par exemple, pour certaines classes d'ARNnc, la fréquence des nucléotides diffère très peu de celle des régions intergéniques non exprimées.

Modèle de covariance

Les modèles de covariance sont un type de grammaire stochastique sans contexte (stochastic context-free grammar, SCFG). Les SCFG fonctionnent de façon similaire aux modèles de Markov cachés, mais offrent la possibilité de représenter des structures 2D plutôt que simplement des séquences linéaires.

Les modèles de covariance sont déterminés à partir d'un alignement structural de plusieurs ARN orthologues (Eddy et Durbin, 1994). Les modèles ainsi produits représentent l'ARN selon les probabilités de chaque nucléotide et chaque appariement. Ils utilisent l'information provenant des changements compensatoires des bases qui préservent la structure pour établir l'importance des appariements. Ils peuvent ensuite être utilisés pour trouver d'autres membres de la même famille d'ARN et cela même dans le cas d'ARN où seule la structure serait importante et où il n'y aurait que très peu de conservation de séquence.

Cette technique est utilisée couramment pour l'identification des ARNt avec l'outil tRNAscan-SE (Lowe et Eddy, 1997). Cet algorithme détecte environ 99 % des ARNt et a un très faible taux de faux positifs. La difficulté principale de cette technique est la création d'un modèle représentatif de la famille. Il est nécessaire d'avoir un ARN dont la structure est bien conser-

vée parmi les membres. Bien qu'il soit en théorie possible de créer un modèle à partir de séquences non alignées et sans prédiction de structure, il est souhaitable d'avoir une prédiction correcte de la structure de l'ARN pour construire un modèle efficace.

La principale limitation de ce type de classificateur est la nécessité d'avoir une base de données fiable des membres de la famille ARN dont on veut établir le modèle. Il est impossible d'identifier de nouveaux types d'ARNnc par cette technique. Aussi, le modèle est très sensible à l'introduction de faux membres dans l'ensemble d'entraînement. Finalement, la construction du modèle peut être fastidieuse puisqu'il est préférable de valider le modèle à chaque introduction d'un nouveau membre.

Prédiction de structures d'ARN

La prédiction des structures secondaires de l'ARN à partir de sa séquence primaire est une tâche complexe. La prédiction de la structure secondaire se divise en deux étapes indépendantes : le repliement de l'ARN et l'évaluation de la structure. Le repliement de l'ARN est généralement effectué par un algorithme de programmation dynamique qui parcourt l'espace des structures possibles à la recherche de la structure optimale. Plusieurs algorithmes de programmation dynamiques peuvent être utilisés. La limitation principale de ces algorithmes est la taille maximale des boucles pouvant être évaluées ce qui rend la prédiction des structures des longs ARN moins fiable.

L'évaluation des structures est faite principalement en utilisant deux techniques. La première (Hofacker, 2003; Zuker, 2003) se base sur des paramètres de thermodynamique de stabilité de l'ARN. Elle fut la première technique développée. Les paramètres thermodynamiques mettent l'emphase sur les nucléotides voisins. Il s'agit encore aujourd'hui du type d'algorithme le plus utilisé malgré sa faible précision. Les paramètres thermodynamiques pour des structures complexes sont souvent absents et doivent être estimés. L'acquisition de nouveaux paramètres thermodynamiques demande des mesures calorimétriques fastidieuses.

La somme de travail demandée pour obtenir de nouveaux paramètres thermodynamiques a mené au développement des techniques d'évaluation probabilistes. Ces techniques (Knudsen et Hein, 2003) furent d'abord développées pour la prédiction de structures homologues grâce aux modèles de covariance. Elles furent ensuite utilisées pour générer des paramètres à partir de vastes bases de données de structures d'ARN. Ce type d'algorithme est susceptible au surentraînement et au manque de diversité à l'intérieur des ensembles d'entraînement. La précision de ces algorithmes augmentera avec le nombre de structures d'ARN cristallisées disponibles.

En somme, la précision des prédictions de structures secondaires est faible. Lorsque cela est possible, il est souhaitable d'utiliser d'autres sources d'information comme la covariation ou des tests *in vitro* en combinaison avec la prédiction *in silico*. De plus, en solution, la structure des ARN est dynamique et il est important de tenir compte de l'ensemble des structures possibles. Cela est particulièrement important dans le cas de long ARN où la prédiction d'une structure optimum est peu fiable. *In vivo*, la plupart des mécanismes impliquant l'ARN ont lieu grâce à des complexes ribonuléoprotéiques (RNP), mais ces interactions protéines-ARN sont encore peu comprises et souvent les prédictions de structure ne correspondent pas à ce qui est observé lorsque la structure 3D est déterminée expérimentalement.

Dans le cadre de recherche *in silico* couvrant tout le génome, la prédiction de structure peut devenir longue. En plus, dans certains cas, la structure prédite comme la meilleure ne correspond pas nécessairement à celle de l'ARNnc recherché. Pour pallier ces inconvénients, il est souvent souhaitable d'effectuer d'abord une recherche de motifs d'ARN et d'utiliser ensuite la prédiction de structure pour évaluer ces motifs. La recherche de motifs consiste à créer un descripteur pour la structure recherchée en spécifiant des contraintes dans un langage propre à l'outil utilisé. Un de ces outils est RNAmotif (Macke $\mathrm{et\ al.}$, 2001). Il permet d'obtenir tous les motifs ayant la possibilité de correspondre avec la structure décrite, peu importe leurs stabilités. Il permet aussi d'obtenir une évaluation de chacun des motifs selon des critères préétablis.

1.2.4 Travaux précédents

Plusieurs travaux ont tenté de découvrir de nouveaux ARNnc en utilisant des approches *in silico*. Chez la levure, la problématique d'identifier l'ensemble des ARNsno a fait l'objet de plusieurs publications. Une première approche (Lowe et Eddy, 1999) qui a été utilisée est une approche probabiliste basée sur un modèle de type HMM (modèle de Markov caché). Cette approche a été utilisée pour la recherche d'ARNsno de type C/D. Ce type d'ARNsno a peu d'éléments structuraux fixes. Le modèle utilisé évaluait la présence de séquences consensus séparées par des régions variables. En plus des longueurs de ces régions et des séquences consensus, le modèle est très dépendant dans son évaluation de la présence d'une région complémentaire à l'ARN ribosomal. Ainsi, bien que cet algorithme ait identifié 22 nouveaux ARNsno, il laisse 4 sites prédits comme cible sans ARNsno correspondant. De plus, il ne permet pas de détecter des ARNsno qui ne cibleraient pas l'ARNr ou qui auraient une organisation différente dans ARNsno à boîte C/D classique.

Les ARNsno de type H/ACA ont une structure plus complexe et peu d'éléments de séquence consensus. Une approche (Edvardsson $\mathrm{et\ al.}$, 2003) se basant sur la structure d'énergie minimum a été développée pour chercher ce type d'ARNsno. Une première étape identifie les éléments de séquence pouvant faire partie d'un ARNsno de type H/ACA. Ces éléments de séquence incluent des régions complémentaires à l'ARNr. La structure secondaire d'énergie minimum est ensuite prédite pour cette région et des filtres sont appliqués pour ne conserver que les candidats dont la structure correspond avec la structure classique d'un ARNsno de type H/ACA. Malgré ces filtres, cet algorithme identifie environ 4000 candidats dans le génome complet de la levure. L'ajout de critères supplémentaires et une sélection manuelle a permis de réduire cette liste et d'identifier 3 candidats particulièrement prometteurs.

Une approche (McCutcheon et Eddy, 2003) moins spécifique a aussi été tentée. Dans cette approche, les génomes de 5 espèces de levure ont été utilisés pour identifier de nouveaux ARNnc par génomique comparative. Les candidats ont été sélectionnés en fonction de la présence d'une structure conservée entre les espèces démontrée par de la covariation. 92

candidats ont été obtenus. De ceux-ci, 13 transcrits ont pu être détectés. Cette approche a permis l'identification de trois nouveaux ARNsno de type H/ACA, malgré qu'aucun type d'ARNnc n'était ciblé en particulier. Cette approche est prometteuse, mais implique que les ARNnc soient conservés entre les espèces.

En somme, la recherche *in silico* d'ARNnc est prometteuse, mais il existe encore plusieurs problèmes majeurs. Les critères de recherche doivent être très stricts de façon à limiter le nombre de faux positifs. Malgré cela, une sélection manuelle et une validation poussée sont nécessaires pour confirmer les résultats.

1.3 Cycle de vie des ARN

Contrairement à l'ADN, les ARN sont relativement instables et doivent être détruits et remplacés pour ne pas nuire au métabolisme. Cela implique que les ARN ont un cycle de vie commençant par la transcription et se terminant par la dégradation et pouvant comporter plusieurs autres étapes de maturation, de modification et de transport. Chacune de ces étapes est associée à différents facteurs et mécanismes de contrôle et régulation. Les ribonucléases jouent un rôle important dans le cycle de vie des ARN à la fois au niveau de la maturation et du contrôle de la stabilité.

1.3.1 Ribonucléases

Les ribonucléases (RNases) sont des enzymes qui coupent l'ARN en fragments plus petits. Il en existe deux grands groupes. Les exoribonucléases enlèvent des nucléotides à partir de l'extrémité d'un ARN alors que les endoribonucléases coupent à l'intérieur de l'ARN. Il existe plusieurs RNases et elles sont regroupées en plusieurs types selon leurs mécanismes d'action et leur homologie avec les autres RNases du même groupe. Chez la levure, parmi les

RNases qui participent à la maturation de certains ARN fonctionnels il y a la RNase MRP, la RNase P, RNT1 et RRP6. La RNase MRP et la RNase P sont en fait des ribozymes, c'est-à-dire que leur activité catalytique est due à un ARN. Elles sont respectivement impliquées dans la maturation des ARN ribosomaux et des ARN de transfert.

Exoribonucléases

Les exoribonucléases sont des RNases qui ont la capacité d'enlever de manière non spécifique un nucléotide à l'extrémité d'une molécule d'ARN. Il en existe plusieurs types et elles sont impliquées à la fois dans la maturation des ARN et dans leur dégradation.

L'exosome est un complexe dégradant les ARN. Il est présent à la fois dans le noyau et le cytoplasme. Sa principale activité catalytique provient de DIS3, une ribonucléase possédant à la fois une activité exoribonucléase 3'-5' et endoribonucléase. Sa fonction principale est la dégradation des ARN. RRP6 est une exoribonucléase de la famille des RNases D. Il dégrade les ARN par leur extrémité 3' tout comme DIS3. Il fait partie de l'exosome nucléaire et contribue aussi à la maturation de plusieurs ARN fonctionnels en plus de dégrader certains transcrits aberrants (Davis et Ares, 2006).

XRN1 est une exoribonucléase principalement cytoplasmique qui dégrade les ARN par leur extrémité 5'. Il est un composant des corps P où il agit avec les enzymes enlevant la coiffe pour dégrader les ARNm. RAT1 est une exoribonucléase 5'-3' nucléaire impliquée dans la maturation des ARNr et des ARNsn. Elle est aussi requise pour la terminaison de la transcription des ARNm.

Les exoribonucléases, bien qu'elles soient en elles-mêmes peu spécifiques, agissent par l'intermédiaire des complexes auxquels elles sont associées sur des transcrits spécifiques. Leur inactivation peut être précieuse dans l'identification d'ARNnc dont elles provoquent

l'accumulation des précurseurs. Leur inactivation révèle aussi l'existence de nombreux transcrits aberrants (Davis et Ares, 2006) qui ont normalement une courte durée de vie.

Ribonucléases III

Les ribonucléases III sont des endoribonucléases qui coupent les ARN double brin. Elles sont présentes chez les procaryotes et les eucaryotes. Selon les organismes, leurs rôles peuvent varier. La plupart des eucaryotes possèdent au moins deux RNases III, l'une d'elles est localisée au noyau alors que l'autre est cytoplasmique. En règle générale, les RNases III cytoplasmiques sont impliquées dans la régulation des ARNm et dans la dégradation des ARN double brin exogènes, alors que les RNases III nucléaires participent à la maturation de plusieurs ARNnc.

Bien que toutes les RNases III coupent des ARN double brin, leurs spécificités varient (Lamontagne $\mathrm{et\ al.}$, 2001). La plupart des RNases III cytoplasmiques, comme Dicer, sont peu spécifiques et coupent tout ARN double brin en petits fragments. La RNase III chez les procaryotes est un peu plus restrictive, l'enzyme reconnait certains antidéterminants (Lamontagne et Elela, 2004) qui empêchent la coupure et elle ne coupe qu'en leur absence. Les RNases III nucléaires des eucaryotes sont plus spécifiques et reconnaissent une tige d'ARN double brin avec certaines boucles simple brin qui positionne l'enzyme et guide la coupure à une position précise. Cette précision est nécessaire dans leur fonction de maturation des ARNnc.

Chez l'humain, Drosha, une RNase III nucléaire, participe à la maturation des ARN ribosomaux et des ARNmi primaires. Les ARNmi précurseurs produits sont ensuite maturés par Dicer, une RNase III cytoplasmique. Les ARNmi matures sont recrutés par le complexe RISC pour la régulation de l'expression de certains ARNm. La régulation par les ARNmi est d'une importance majeure dans le cycle de vie des ARNm chez les eucaryotes et les niveaux d'expression des ARNmi sont modifiés dans plusieurs types de cellules cancéreuses.

Chez la levure *S. cerevisiæ*, il n'existe pas de RNase III cytoplasmique, orthologue de Dicer. Seul Rnt1p, une RNase III nucléaire, est présente. Elle est un homologue proche de la RNase III bactérienne. Rnt1p, de même que Pac1 et Drosha, reconnaît une structure en tige-boucle. La tige-boucle reconnue par Rnt1p a la particularité d'être habituellement coiffée d'une tétraboucle ayant une guanine en deuxième position (NGNN). Comme les autres RNases III, Rnt1p reconnait les ARN double brin, mais la reconnaissance spécifique de la tétraboucle et des nucléotides de la boucle positionnerait l'enzyme et renforcerait la liaison. Dans un deuxième temps, le domaine nucléase catalyse la coupure à 14 et 16 nucléotides de la tétraboucle. L'efficacité de cette coupure varie selon la composition dans la région de la coupure. Cela en fait un enzyme beaucoup plus spécifique que la RNase III bactérienne tout en étant beaucoup plus simple que les autres RNases III eucaryotes. De plus, elle joue un double rôle métabolique puisqu'elle est responsable de la régulation directe de l'expression de certains ARNm (Ge $\mathrm{et\ al.}$, 2005) en plus des rôles de maturation des ARNnc.

L'absence de RNase III non spécifique chez *S. cerevisiæ* peut s'expliquer par l'existence du virus L-A, un virus ARNdb d'environ 46 000 bp. Ce virus produit une toxine qui tue les levures non infectées. Ainsi, l'infection par ce virus constitue un avantage compétitif pour la levure, d'où l'importance pour la levure de posséder seulement une RNase III suffisamment spécifique pour ne pas dégrader ce virus ARN (Drinnenberg $\mathrm{et\ al.}$, 2011).

À ce jour, la plupart des cibles des RNases III ont été identifiées par isolation de petits ARN ou par génomique comparative (Lim $\mathrm{et\ al.}$, 2003). Chez la levure, plusieurs cibles ont pu être détectées par une approche combinée de prédiction *in silico* et de détection de variation de l'expression sur puces à ADN (Ghazal $\mathrm{et\ al.}$, 2005). L'isolation de petits ARN est un processus fastidieux et souvent, elle ne détecte que les ARN le plus abondants. La génomique comparative ne permet pas de détecter des cibles qui seraient apparues spécifiquement dans une espèce pour remplir un besoin particulier. La combinaison de plusieurs approches permet de palier ces limitations, mais les techniques utilisées pourraient être raffinées pour élargir le champ de recherche.

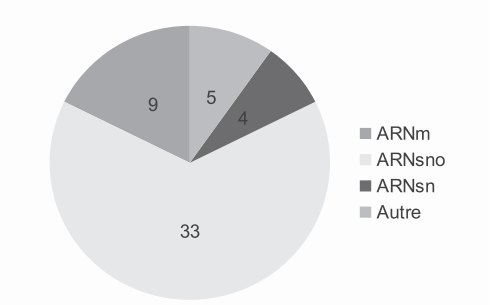


figure 1.2 – Les types de transcrits ciblés par les substrats confirmés de Rnt1p

Rnt1p est connu pour participer à la maturation de l'ARN ribosomal 25S, de certains ARNsn et de plusieurs ARNsno. Il participe aussi à la régulation de certains ARN messagers. Son substrat habituel est une tige coiffée d'une tétraboucle NGNN (Lamontagne et Elela, 2004). Il coupe des deux côtés de la tétraboucle à 14 et 16 nucléotides respectivement. À ce jour, il existe 51 substrats confirmés de Rnt1p. La majorité d'entre eux sont situés dans la région 5' du transcrit précurseur d'ARNsno (1.2, annexe A pour la liste complète).

1.4 Problématiques

Le cycle de vie de l'ARN est d'une importance primordiale dans le métabolisme cellulaire. Les RNases III interviennent dans les étapes de maturation et de régulation de plusieurs ARN et contribuent ainsi au contrôle de leur cycle de vie. La spécificité des RNases III pour leurs substrats et la façon dont l'enzyme reconnaît ses substrats sont mal connues. L'identification de nouveaux substrats est un processus long et fastidieux. Le modèle actuel du rôle et du mécanisme d'action des RNases III est basé sur un nombre limité de substrats et ne représente probablement pas la gamme complète de cibles de l'enzyme. Ainsi, le rôle métabolique des RNases III n'est que partiellement connu.

1.5 Objectifs

Cette thèse a comme objectif de développer des approches haut débit pour permettre une identification plus rapide des cibles de RNase III tout en minimisant l'utilisation des ressources expérimentales. Elle présente l'utilisation combinée d'approches bio-informatiques, d'étude génétique de l'expression et de traitement *in vitro* dans le but d'avoir un portrait global des cibles de l'endoribonucléase III. Elle a aussi comme but d'identifier les motifs d'ARN non codants qui guident la reconnaissance par l'endoribonucléase III.

La RNase III chez la levure *S. cerevisiæ*, Rnt1p, a été choisie comme modèle d'étude. Il s'agit d'une RNase III eucaryote parmi les plus simples. Elle est disponible sous forme de recombinante ce qui permet d'effectuer des validations *in vitro*. La génétique simple de cette levure permet de faire des validations *in vivo* de façon rapide et simple.

Les approches bio-informatiques qui ont été utilisées dans le passé étaient limitées par le très petit nombre de substrats utilisés pour la création du modèle. L'ensemble complet de substrats connus sera utilisé pour raffiner le modèle. De nouveaux algorithmes d'apprentissage machine seront testés pour créer un outil de recherche plus performant.

Actuellement, la majorité des substrats de Rnt1p ont été détectés par une augmentation de leurs niveaux d'expression lorsque Rnt1p est inactivé. Cependant les puces à ADN qui ont été utilisées ne mesurent que l'expression globale des gènes connus. Puisque Rnt1p participe à la maturation de certains transcrits, l'utilisation de puces à ADN à haute résolution couvrant tout le génome pourrait permettre de révéler de nouvelles cibles. Ces cibles n'auraient qu'un changement mineur de niveau d'expression, mais auraient un changement de taille de transcrit. Des cibles situées dans des régions habituellement non exprimées pourraient aussi être détectées.

La plupart des substrats connus de Rnt1p sont des précurseurs d'ARNnc où Rnt1p coupe dans le cadre de la maturation du transcrit. Il est possible que cet ensemble de substrats

ne soit pas représentatif de la spécificité de Rnt1p. Tester individuellement des substrats *in vitro* dans le but de mieux comprendre quels motifs d'ARN sont reconnus par Rnt1p est fastidieux. En combinant les essais de coupure *in vitro* de l'ARN avec les techniques d'analyse d'ARN haut débit, il devrait être possible d'identifier rapidement un grand nombre de motifs reconnus par Rnt1p et ainsi de mieux comprendre comment l'enzyme reconnait ses substrats. De plus, en utilisant tout l'ARN de la levure pour effectuer l'identification, les cibles potentielles de Rnt1p dans la cellule seront révélées et il sera possible de mieux comprendre le rôle métabolique global des RNases III dans la cellule.

L'utilisation combinée de ces trois types de recherche (*in silico*, *in vivo* et *in vitro*) permettra une meilleure confiance dans les résultats obtenus et fournira un portrait plus global du fonctionnement et des rôles des RNases III. De plus, les avantages et limitations de chacune des techniques seront plus clairement révélés.

CHAPITRE 2

CARACTÉRISATION ET IDENTIFICATION DES SUBSTRATS RECONNUS PAR RNT1P PAR L'UTILISATION DE DIFFÉRENTS ALGORITHMES DE CLASSIFICATION

La recherche *in silico* de motifs reconnus par Rnt1p permet l'identification non biaisée de substrats potentiels indépendamment des conditions de culture et des limites des mesures expérimentales. Le développement d'un algorithme de recherche *in silico* permet aussi d'obtenir un modèle des motifs pouvant théoriquement être reconnus par l'enzyme.

2.1 Introduction

2.1.1 Rnt1p

Chez la levure *S. cerevisiæ*, il n'existe pas de RNase III cytoplasmique, orthologue de Dicer. Seul Rnt1p, une RNase III nucléaire, est présente. Elle est un homologue proche de la RNase III bactérienne. Rnt1p de même que ses homologues, Pac1 et Drosha, reconnaissent une structure en tige-boucle. La tige-boucle reconnue par Rnt1p a la particularité d'être habituellement coiffée d'une tétraboucle ayant une guanine en deuxième position (NGNN). Cela en fait un enzyme beaucoup plus spécifique que la RNase III bactérienne tout en étant beaucoup plus simple que les autres RNase III eucaryotes. De plus, elle joue un double rôle puisqu'elle est responsable de la régulation directe de l'expression de certains ARNm (Ge $\mathrm{et\ al.}$, 2005) en plus des rôles de maturation des ARNnc qui sont habituellement attribués aux RNases III nucléaires.

2.1.2 Substrats de Rnt1p

À ce jour, 45 substrats de Rnt1p sont connus (voir Annexe A). La majorité des substrats de Rnt1p sont associés à des transcrits précurseurs d'ARN non codants (38), dont l'ARNr 25S, 4 ARNsn et 33 ARNsno. On compte aussi 2 substrats situés dans des introns, un situé dans la région non traduite en 5' d'un ARNm et 4 situés dans les séquences codantes d'ARNm.

Pour la majorité des substrats de Rnt1p, la coupure s'effectue par la reconnaissance d'une structure d'ARN comportant une tige d'ARN coiffée d'une tétraboucle AGNN. Il a été montré que l'adénosine en première position n'est pas essentielle, mais le remplacement de la guanine par un autre nucléotide abolit généralement la liaison. La coupure s'effectue généralement à 14 et 16 nt de chaque côté de la tétraboucle.

Il existe cependant de nombreuses exceptions à cette structure consensus (Voir figure 2.1). D'abord, la coupure d'un précurseur d'ARNsno est guidée par une tétraboucle AAGU plutôt que AGNN ou NGNN. Ensuite, plusieurs substrats comportent des boucles dans la tige d'ARN qui sont exclues lors de la liaison. Ainsi, pour certains substrats, la coupure est détectée à plus de 100 nt de la tétraboucle guide. Finalement, il existe aussi un substrat, le précuseur de l'ARNsno snR18, dont la reconnaissance serait dépendante de la présence d'une protéine chaperonne (Giorgi $\mathrm{et\ al.}$, 2001).

Une étude sommaire de la conservation des substrats de Rnt1p a révélé que la maturation par une enzyme ayant une spécificité similaire à Rnt1p est présente chez la plupart des hémiascomycètes à l'exception de *Y. lipolytica*, l'espèce la plus distante parmi celles étudiées (Chanfreau, 2003). Le plus haut niveau de conservation est présent pour les substrats de l'ARNr, alors qu'il est plus faible pour les substrats ARNsno. Des mécanismes alternatifs de maturation sont envisageables pour les ARNsno.

2.1.3 Recherche de substrats

La plupart des substrats connus ont été identifiés par examen manuel des séquences et validation par coupure *in vitro*. Trois techniques de validation *in vitro* sont couramment utilisées. Les trois utilisent la protéine Rnt1p purifiée. La réaction de coupure peut être effectuée sur un ARN synthétique ou sur de l'ARN total extrait d'une culture de levure. Les produits de coupure peuvent être observés par buvardage Northern ou par extension d'amorce. La réaction de coupure en ARN total permet de mieux reproduire la coupure *in vivo* se faisant sur un transcrit complet en compétition pour l'enzyme avec les autres substrats. La séparation des produits de coupure permet de quantifier plus facilement l'efficacité de la coupure, alors que l'extension d'amorce identifie au nucléotide près la position du site de coupure.

Un seul algorithme de recherche *in silico* a été développé (Ghazal $\mathrm{et\ al.}$, 2005). Il a été basé sur un ensemble d'entraînement de 18 substrats connus. Il utilise trois critères : la similitude avec une matrice de séquence créée à partir de l'ensemble d'entraînement, la similitude avec un des substrats de l'ensemble d'entraînement et la stabilité de la structure.

Les paramètres de cet algorithme ont été fixés de façon arbitraire en se basant sur les substrats connus et des tests *in vitro*. Lorsqu'il est utilisé sur l'ensemble du génome de la levure, cet algorithme génère plusieurs dizaines de milliers de candidats ce qui le rend inutilisable pour une recherche génomique. Il a cependant été utilisé avec succès pour l'identification plusieurs nouveaux substrats précurseurs d'ARNsno (Ghazal $\mathrm{et\ al.}$, 2005) et d'un substrat ARNm (Ge $\mathrm{et\ al.}$, 2005).

Pour l'identification de précurseur d'ARNsno, il a été utilisé en limitant la région de recherche à 500 nt autour de l'ARNsno. Pour l'ARNm, plusieurs des meilleurs candidats ont été testés par buvardage Northern jusqu'à trouver un résultat positif. Dans ces cas, la détection d'une accumulation du transcrit par puces à ADN a été utilisée pour accroître le niveau de confiance dans les candidats.

2.1.4 Objectifs

L'algorithme de recherche *in silico* actuel génère un grand nombre de faux positifs et est difficilement utilisable seul pour une recherche génomique.

Ce travail vise à développer un algorithme de recherche plus performant pour la recherche *in silico* de substrats de Rnt1p. Il sera ensuite utilisé, conjointement avec d'autres critères, pour identifier de nouveaux substrats.

La première étape sera de mieux définir les substrats en utilisant l'ensemble complet des substrats connus pour établir un modèle des éléments de séquence et de structure présents dans l'ensemble des substrats. La conservation de ces éléments chez d'autres espèces de levure sera étudiée pour obtenir une meilleure estimation de l'importance de ces éléments.

Ce modèle servira de base pour le développement d'un algorithme de recherche. L'algorithme actuel sera raffiné en utilisant les éléments du nouveau modèle et en l'optimisant pour réduire le nombre de faux positifs. L'algorithme modifié sera comparé avec des approches d'apprentissage machine standards et avec un approche qui s'adresse plus spécifiquement aux ARN structurés, les modèles de covariance .

L'identification de nouveaux substrats de Rnt1p dans les régions intergéniques du génome de la levure pourrait permettre de révéler de nouveaux ARN non codants qui auraient échappé aux recherches précédentes. Finalement, en combinant la recherche de substrats de Rnt1p avec la présence d'éléments typiques d'ARNsno ou encore avec la présence d'orthologues, le nombre de candidats ARN sera limité et les candidats les plus intéressants pourront être validés.

2.2 Méthodes

2.2.1 Extraction des séquences des substrats

Les séquences des substrats connus (45 publiés avant le début de l'étude et 6 validés au cours de l'étude) de Rnt1p ont été extraites des séquences de références du génome de *Saccharomyces cerevisiæ* souche S288C version R56-1-1. Une région de 50 nucléotides où la tétraboucle guide est placée en position centrale a été utilisée.

2.2.2 Prédiction des structures d'ARN

La prédiction des structures secondaires et le calcul de l'énergie libre de Gibbs (ΔG) a été effectuée à l'aide de la suite Vienna RNA version 1.8.5.

2.2.3 Identification des éléments surreprésentés

Les éléments de séquence surreprésentés dans les substrats de Rnt1p ont été identifiés par la comparaison avec un ensemble de 10 000 séquences de même longueur générées aléatoirement en conservant une composition en nucléotide et en dinucléotide similaire aux substrats connus. Une différence de composition avec les séquences aléatoires est définie comme une valeur $p < 0.05$ pour un test exact de Fisher avec correction de Holm-Bonferroni (Holm, 1979).

Les éléments de structure surreprésentés dans les substrats de Rnt1p ont été identifiés en comparant l'appariement ou le non-appariement pour la même position dans l'ensemble de séquences générées aléatoirement pour la structure correspondante prédite avec l'énergie

minimale. Une normalisation a été appliquée pour tenir compte de la différence de composition en nucléotides. Le pourcentage moyen d'appariement pour chaque nucléotide à chaque position a été calculé et le pourcentage d'appariement pour une distribution aléatoire a été calculé à partir du pourcentage de chaque nucléotide observé et du pourcentage d'appariement pour ce nucléotide. Un élément de structure surreprésenté est défini comme ayant une valeur $p < 0.05$ pour test binomial à deux échantillons avec correction de Holm-Bonferroni.

2.2.4 Préparation des données pour la classification

Les 51 substrats connus de Rnt1p ont été utilisés comme ensemble positif pour la classification. Un ensemble de 2512 séquences non coupées a été obtenu à partir de fenêtres de 50 nt tirées des séquences des ARNsno matures dont la maturation dépend de Rnt1p. Cet ensemble a été utilisé comme ensemble négatif. Une validation croisée de type 10-fold a été effectuée. Chaque nucléotide de chaque élément à classifier est représenté par trois attributs binaires : deux qui représentent la base et un, l'appariement. L'expérience a aussi été répétée en utilisant quatre attributs pour la base.

2.2.5 Algorithme d'identification par classification bayésienne

Les ensembles d'entraînement ont été utilisés pour produire les tables de fréquence pour chacun des attributs dans l'ensemble positif et l'ensemble négatif. Les probabilités postérieures ont été calculées pour chaque attribut et additionnées en utilisant différentes valeurs pour la probabilité d'appartenir à la classe positive. Pour l'ensemble de validation croisé, la valeur optimale de .49 a été sélectionnée pour maximiser le coefficient de corrélation de Matthews (Matthews, 1975). L'algorithme a ensuite été testé sur une région de 10 000 nucléotides provenant du chromosome 1 et débutant à la position 50 000.

2.2.6 Algorithme d'identification par classification SVM

La classification par SVM a été effectuée à l'aide de la bibliothèque e1071 (Dimitriadou $\mathrm{et\ al.}$, 2008). Des noyaux ("kernel") de type radial ont été utilisés. Les valeurs de coût et de gamma ont été optimisées respectivement à 4 et 0.004 pour obtenir une valeur maximale du coefficient de corrélation de Matthews. L'algorithme a ensuite été testé sur une région de 10 000 nucléotides provenant du chromosome 1 et débutant à la position 50 000.

2.2.7 Identification des orthologues des substrats

Les orthologues des substrats connus ont été identifiés en utilisant l'alignement multiple de génomes du genre *Saccharomyces* obtenu de UCSC (http ://hgdownload.cse.ucsc.edu/-goldenPath/sacCer1/bigZips/multizYeast.zip). Lorsque disponibles, les séquences correspondantes aux substrats connus chez les autres espèces ont été testées à l'aide du classificateur bayésien précédemment décrit et seules les séquences classées positives pour la coupure par Rnt1p ont été retenues comme orthologues.

2.2.8 Algorithme d'identification par modèles de covariance

Les modèles de covariance et la recherche les utilisant ont été effectués à l'aide de l'implémentation nommée Infernal (Nawrocki $\mathrm{et\ al.}$, 2009) version 1.0.2. Un modèle de covariance a été généré pour chaque ensemble d'orthologues précédemment identifié. Une valeur de limite de 7 pour le pointage a été déterminée en maximisant le coefficient de corrélation de Matthews pour un ensemble de validation contenant 10 substrats connus et 502 non substrats.

2.2.9 Algorithme d'identification par similitude

Un algorithme simple de recherche par similitude a précédemment été développé (Ghazal $\mathrm{et\ al.}$, 2005). Cet algorithme a été optimisé pour assigner différents poids aux nucléotides et aux appariements en fonction du niveau de conservation de chaque position chez d'autres espèces de levure *sensu stricto*.

Quatre valeurs normalisées de 0 à 1 sont générées par cet algorithme :
– La similitude de séquence avec les substrats connus ;
– La similitude de structure avec les substrats connus ;
– La stabilité de structure (ΔG) ;
– La stabilité de la tétraboucle.
Les poids accordés à chacune de ces valeurs ont été optimisés pour réduire au maximum le nombre de candidats obtenant un pointage supérieur au substrat connu ayant le plus bas pointage.

2.2.10 Recherche de candidats conservés

Parmi tous les candidats comme substrats de Rnt1p identifiés par l'algorithme de recherche par similitude, ceux qui sont situés dans une région non transcrite du génome ont été vérifiés pour la conservation chez trois autres espèces de levures *sensu stricto*. Les régions non transcrites sont définies comme n'ayant pas d'annotation assignée à elles. La conservation a été définie comme la conservation du G en deuxième position de la tétraboucle et de la paire de bases fermant la tétraboucle dans les séquences orthologues d'après l'alignement multiespèces décrit précédemment.

2.3 Résultats

2.3.1 Modèle d'un substrat

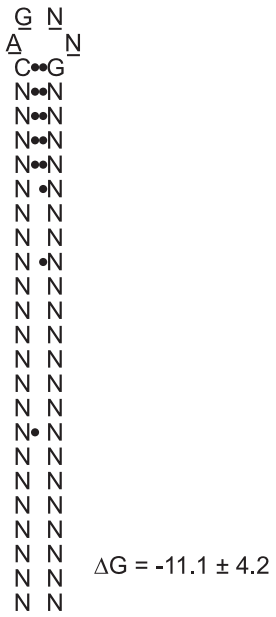


figure 2.1 – Modèle d'un substrat de Rnt1p.

Construit d'après 51 substrats validés et publiés. Les nucléotides indiqués sont significativement enrichis par rapport à une distribution aléatoire. Les points noirs indiquent un pourcentage d'appariement significativement plus élevé qu'attendu alors que les nucléotides soulignés sont plus rarement appariés qu'attendu. Le ΔG est la médiane des ΔG des substrats connus avec la déviation absolue médiane.

Tous les substrats validés et publiés ont été utilisés pour produire une structure consensus. Tous ces substrats ont la capacité de se replier pour former une tétraboucle centrale. Cependant, dans 6 cas sur 51, il ne s'agit pas de la structure optimale (énergie libre minimum) pour la séquence retenue. Malgré ce fait, la formation d'une tétraboucle et l'appariement des six premières paires de bases sont fortement enrichis par rapport à des séquences de même composition générées aléatoirement. Les structures reconnues par Rnt1p sont aussi plus stables (ΔG inférieur) que des séquences de même composition en dinucléotides (test

de Mann-Whitney, p < 1e-15) ce qui n'est pas le cas pour la plupart des ARN non codants (Rivas et Eddy, 2000).

Du point de vue de la séquence, la préférence pour une tétraboucle AGNN est confirmée. On remarque aussi une préférence pour la paire C-G comme paire fermant la tétraboucle (Voir Figure 2.2). Il est possible que cela confère une plus grande stabilité à la tétraboucle, mais il est aussi connu que la paire U-A fermant la tétraboucle inhibe la coupure par Rnt1p (Sam et al., 2005) ce qui est confirmé par la rare présence dans les substrats connus (Voir Figure 2.2). Du côté de la structure, le haut de la tige est pairé à près de 100 %. Un pourcentage élevé d'appariement est aussi observable dans la région près du site de coupure (Voir Figure 2.3).

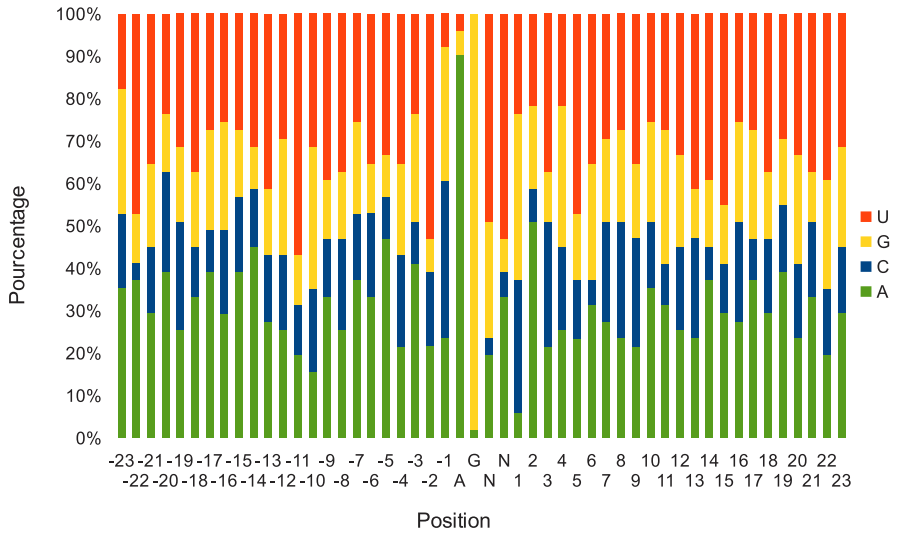


figure 2.2 – Composition en nucléotides des substrats connus.

D'après 51 substrats validés et publiés. La composition en nucléotides pour chacune des positions étudiées est indiquée.

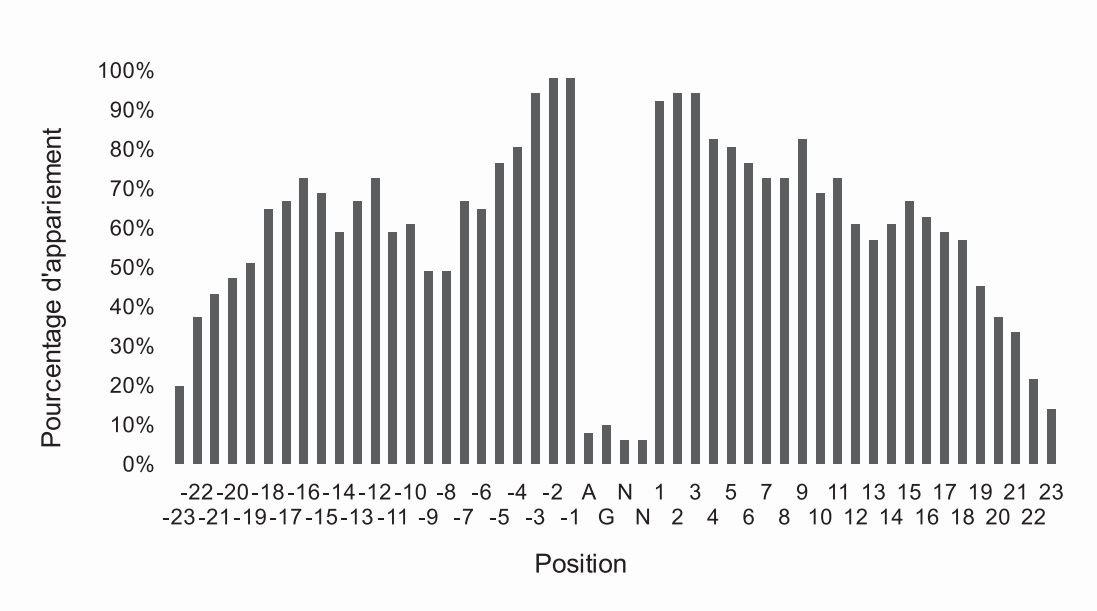


figure 2.3 – Positions appariées des substrats connus.

D'après 51 substrats validés et publiés. Le pourcentage d'appariement pour chacune des positions étudiées est indiqué.

2.3.2 La prédiction des substrats par classificateur bayésien

Un classificateur bayésien naïf a été développé pour classifier les séquences comme faisant partie de la classe des substrats de Rnt1p ou non. Pour les données de nucléotides encodées sur deux attributs, ce classificateur a un coefficient de Matthews moyen de 0.74 sur l'ensemble de validation ce qui confirme que la classification fonctionne. Sa spécificité moyenne est de 99.9 %, alors que sa sensibilité moyenne est de 61 %. Pour les données de nucléotides encodées sur quatre attributs, ce classificateur a un coefficient de Matthews moyen de 0.77 sur l'ensemble de validation. Sa spécificité moyenne est de 99.8 %, alors que sa sensibilité moyenne est de 67 %.

Pour une région de 10 000 nucléotides du génome, 9 régions de 50 nt sont classifiées comme substrats potentiels. En extrapolant pour tout le génome de la levure *S. cerevisiæ*, le nombre de candidats à titre de substrats peut être estimé à environ 11 000.

2.3.3 La prédiction des substrats par classificateur SVM

Un modèle de classification SVM a été défini pour classifier les séquences comme faisant partie de la classe des substrats de Rnt1p ou non. Ce classificateur, lorsqu'utilisé sur des données où les nucléotides sont encodés comme deux attributs, a un coefficient de Matthews moyen de 0.73 sur l'ensemble de validation ce qui confirme que la classification fonctionne. Sa spécificité moyenne est de 99.9 %, alors que sa sensibilité moyenne est de 57 %. Lorsque les nucléotides sont encodés comme quatre attributs, son coefficient de Matthews moyen est de 0.73, sa spécificité moyenne est de 99.9 % et sa sensibilité moyenne est de 61 %.

Pour une région de 10 000 nucléotides du génome, 18 régions de 50 nt sont classifiées comme substrats potentiels. En extrapolant pour tout le génome de la levure *S. cerevisiæ*, le nombre de candidats à titre de substrats peut être estimé à environ 22 000. Le fait d'entraîner un classificateur avec moins d'attributs (30 nt) augmente le nombre de candidats potentiels à 38 pour la même région de 10 000 nucléotides.

2.3.4 Les substrats sont conservés

Il a déjà été établi que plusieurs substrats de Rnt1p sont conservés chez d'autres espèces (Chanfreau, 2003). Ici, la présence d'un orthologue dans au moins une autre espèce est confirmée pour 42 des 51 substrats connus et plus de la moitié des substrats connus (26) sont conservés dans au moins quatre autres espèces de levure. Les substrats de Rnt1p semblent en général conservés, car il est probable que plusieurs des orthologues man-

quants n'aient pas été identifiés dû à des défauts dans l'alignement des séquences. Il est important de noter que les substrats associés à des régions non codantes sont aussi plus conservés que ceux situés dans des régions codantes. Tous les substrats dont un homologue est détecté dans au moins deux autres espèces de levure sont situés dans des régions non codantes.

Pour ce qui est des quatre substrats situés dans la séquence codante d'un ARNm, aucun ne semble conservé chez les autres espèces de levure étudiées. Cependant, chez les différentes souches de levure *Saccharomyces cerevisiæ*, on peut observer différents niveaux de conservation selon le substrat. Le substrat ARN2-2 est conservé chez toutes les souches de *S. cerevisiæ* comportant une région orthologue. Le substrat MIG2 est conservé chez toutes les souches, sauf une. Le substrat ARN2-1 est conservé chez toutes les souches et en plus, une covariation de la paire UA pour la paire CG est observable à 18 paires de bases de la boucle. Le substrat ADI1 n'est conservé que chez 75 % des souches. La conservation des substrats situés dans les séquences codantes est spécifique à l'espèce et même pour certains substrats à certaines souches.

Ensemble les orthologues des substrats fournissent 137 nouveaux exemples de substrats de Rnt1p. Cependant, ces exemples sont peu représentatifs en eux-mêmes, ils ne devraient pas être utilisés lors de l'apprentissage avec un algorithme sensible aux doublons. Ils permettent quand même d'obtenir de l'information sur l'importance de chacune des variables et dans un contexte d'apprentissage dirigé, ils permettent d'assigner un poids à chaque variable ce qui réduit la nécessité d'avoir un grand nombre d'exemples.

2.3.5 La recherche de substrat à l'aide de modèle de covariance

En utilisant des modèles de covariance indépendants pour chacune des familles d'orthologues, il a été possible de développer une technique de recherche pour les substrats de Rnt1p. Cette technique obtient un coefficient de corrélation de de Matthews de 0.193 sur un

ensemble de 10 substrats et 502 non substrats n'ayant pas été utilisés pour son entraînement. Sa spécificité est de 94 %, alors que sa sensibilité est de 40 %.

Pour une région de 10 000 nucléotides du génome, 128 régions ont été identifiées comme substrats potentiels. En extrapolant pour tout le génome de la levure *S. cerevisiæ*, le nombre de candidats à titre de substrats peut être estimé à environ 155 000.

2.3.6 La conservation de séquence et d'appariement identifie les régions importantes de la structure

L'étude de la conservation chez les orthologues permet de valider le consensus (Figure 2.1). À la figure 2.4, on observe la très forte conservation des deux premières bases de la tétraboucle de même que de la paire fermant la tétraboucle. En fait, les trois premières paires fermant la tétraboucle sont conservées à près de 100 % au niveau de la séquence et de la structure. Les deux paires suivantes sont conservées au niveau de la structure, mais beaucoup moins au niveau de la séquence. Une conservation accrue est aussi observée près du site de coupure, mais elle est moins claire que dans la région voisine de la tétraboucle.

2.3.7 La recherche de substrat par similitude

L'étude de la conservation a permis le développement d'un algorithme de recherche de substrats comparant la séquence et la structure des candidats avec celles des substrats. La principale différence que cet algorithme a par rapport aux autres algorithmes d'apprentissage machine est que les poids accordés à chacune des positions de la séquence a été fixé en fonction du degré de conservation chez les orthologues. De plus, il fournit un pointage de similitude normalisé de 0 à 1 pour chaque candidat évalué qui permet ensuite de faire le classement des candidats.

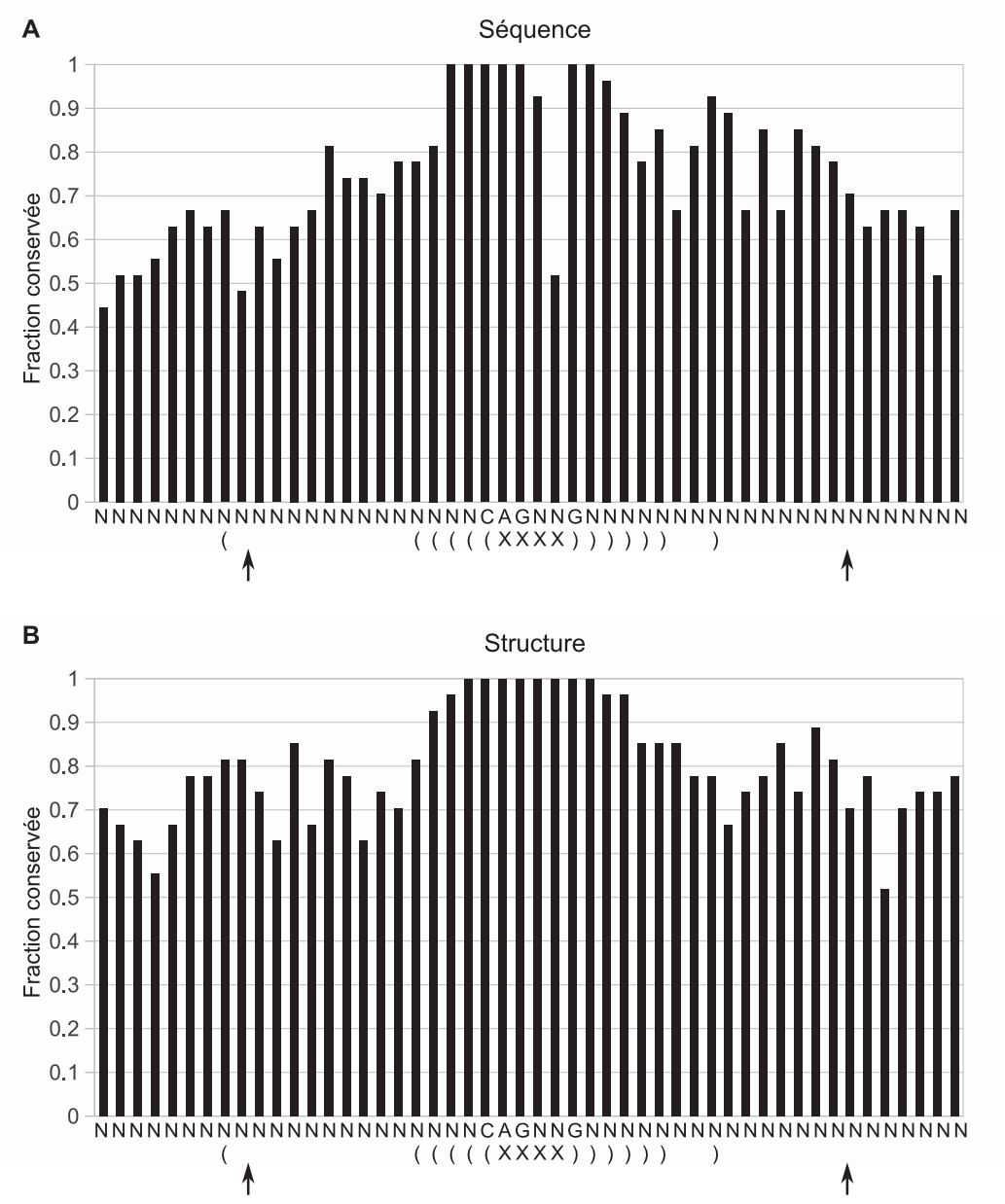


figure 2.4 – Conservation moyenne pour 27 substrats et leurs orthologues chez *S. bayanus*.

En abscisse, la position selon la structure consensus de la figure 2.1. Les flèches indiquent la position théorique de la coupure, à 14 et 16 nt de la tétraboucle. A Conservation du nucléotide à la position spécifiée. B Conservation de l'appariement ou du non-appariement à la position spécifiée.

En somme, l'algorithme de recherche dirigé utilise trois critères : similitude de séquences, similitude de structures et stabilité thermodynamique, mais contrairement aux méthodes non supervisées, les poids accordés à chaque position dans la structure ont été assignés en fonction du degré de conservation chez les orthologues plutôt qu'optimisés avec un ensemble d'entraînement. L'approche a donc des similitudes avec d'autres approches utilisées dans la littérature (Edvardsson et al., 2003; Lowe et Eddy, 1999) qui utilise des motifs provenant de la littérature combinés avec un classificateur. Cependant, le plus grand nombre de substrats et d'homologues permet de procéder de façon plus systématique dans l'assignation des poids.

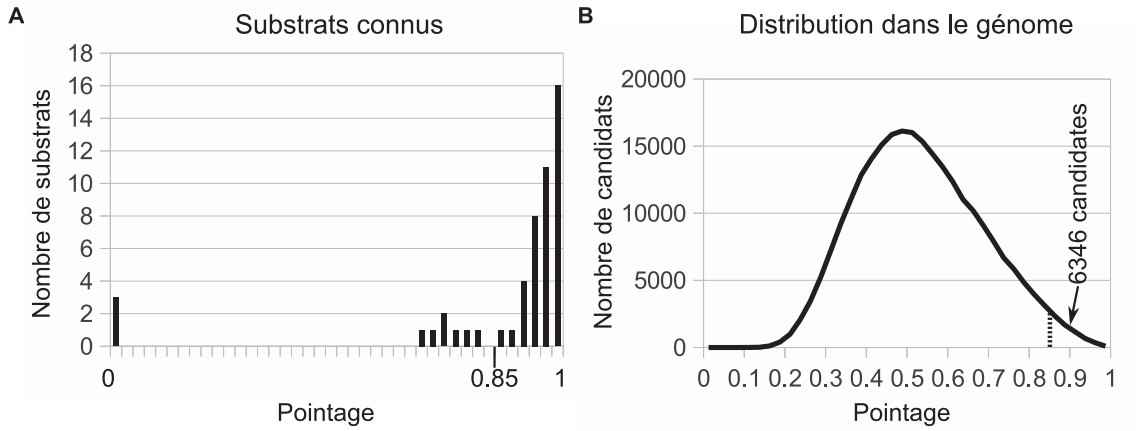


figure 2.5 – Résultats de l'évaluation par l'algorithme de recherche par similitude.

A Distribution des pointages pour les substrats connus. B Distribution des pointages pour toutes les tétraboucles NGNN présentes dans le génome de *S. cerevisiæ*.

Le pointage de 0.85 a été choisi comme valeur seuil, car 80 % des substrats connus obtiennent un pointage supérieur. Pour cette valeur seuil, il existe dans le génome 6346 tétraboucles candidates ayant un pointage supérieur.

2.3.8 Des structures reconnues par Rnt1p impliquées dans la terminaison de la transcription

Maintenant qu'il est établi que plusieurs substrats sont conservés chez plusieurs espèces de levures, il est possible d'ajouter ce critère pour augmenter le niveau de confiance dans les candidats qui ont été identifiés. Parmi les candidats trouvés par l'algorithme de recherche par similitude, 31 sont situés dans des régions non transcrites et sont conservés. Parmi ces 31, 18 sont des substrats connus associés à des ARNsno, trois sont des substrats connus associés à des ARNsn et deux sont des substrats connus situés dans des introns. Cette recherche a identifié un nouveau substrat associé à un ARNsno, snR87, lequel avait été supposé (Davis et Ares, 2006), mais n'avait jamais été validé *in vitro*. Cette recherche a aussi permis d'identifier un substrat situé entre deux ARNm, NPL3 et GPI17, lequel a mené à la découverte d'un nouveau mode de terminaison de la transcription chez la levure (Ghazal $\mathrm{et\ al.}$, 2009). Un autre substrat impliqué dans la terminaison de la transcription a aussi été validé dans cette étude. Finalement, un nouveau substrat associé à un ARN non codant non caractérisé a été validé (résultat non publié).

2.4 Discussion

2.4.1 Caractéristiques des substrats de Rnt1p

Malgré le petit nombre de substrats connus (51), il est possible de définir une structure consensus représentant le substrat idéal : une tétraboucle AGNN stable fermée par une paire C-G et cinq autres paires avec un site de coupure apparié. Cependant, plusieurs substrats ne répondent pas à ce consensus. Quelques substrats ne forment pas une tétraboucle dans leur structure d'énergie minimum prédite. Un substrat est coiffé d'une tétraboucle AAGU (Ghazal et Elela, 2006) plutôt que AGNN et de nombreux substrats n'ont pas d'adénine comme première base de leur tétraboucle. Ainsi, aucun nucléotide ni aucun

élément de structure ne sont essentiels à la reconnaissance d'un substrat par Rnt1p. Pourtant, la coupure par Rnt1p est spécifique et lors de tests *in vitro* très peu de candidats sont coupés.

Donc, il est raisonnable de penser que la reconnaissance par Rnt1p s'effectue par l'intermédiaire d'une combinaison de déterminants et d'antidéterminants. Ainsi, il existe probablement différentes classes de substrats dont la reconnaissance dépend de différentes combinaisons d'éléments de séquence et de structure. Cependant, le nombre restreint de substrats connus empêche de faire une analyse *in silico* de leurs caractéristiques puisque l'étude des relations de second ordre n'a pas permis d'identifier des relations significatives. Pour obtenir plus d'information sur le mode de liaison d'un substrat, plusieurs expériences *in vitro* sont possibles (Ghazal et Elela, 2006; Hartman $\mathrm{et\ al.}$, 2013; Lavoie et Elela, 2008). Cependant, le travail nécessaire est considérable et il ne permet que de valider le mode de liaison pour un seul substrat. Ce type de travail a permis de montrer que le substrat AAGU est reconnu différemment des substrats AGNN classique (Gaudin $\mathrm{et\ al.}$, 2006), mais des différences sont aussi à prévoir à l'intérieur même des substrats AGNN. Il serait donc important d'identifier un plus grand nombre de substrats.

Récemment, la structure 3D de Rnt1p lié avec un substrat typique a été déterminée (Liang $\mathrm{et\ al.}$, 2014). Cette structure est compatible avec les éléments observés comme étant enrichis parmi les structures coupées par Rnt1p. Le domaine N-terminal interagit avec les deux premiers nucléotides de la boucle. Le domaine de liaison à l'ARN 0 interagit avec les deux derniers nucléotides de la boucle. Le domaine de liaison à l'ARN 1 interagit avec les nucléotides situés après la boucle et le domaine de liaison à l'ARN 3 interagit avec les nucléotides situés près du site de coupure. Ces résultats confirment le bien fondé d'accorder plus de poids aux positions les plus conservées qui correspondent aussi avec les sites d'interaction avec la protéine selon la structure 3D.

2.4.2 Différents rôles pour les substrats de Rnt1p

De plus, tous les substrats ne sont pas égaux, car on peut observer parmi les substrats connus une grande variabilité au niveau de l'efficacité de la liaison et de la coupure par Rnt1p. Il est connu que la composition de la boîte de stabilité de la liaison (6 paires de bases fermant la tétraboucle) et la composition de la boîte d'efficacité de la coupure peuvent grandement influencer l'activité de Rnt1p (Babiskin et Smolke, 2011). Ainsi, la conservation de séquence et de structure qui a été observée chez les substrats orthologues n'est possiblement pas due uniquement à l'importance de conserver un substrat de Rnt1p, mais aussi à l'importance de garder une activité spécifique à chaque substrat. Cette modulation de l'activité de Rnt1p selon la cible offre une meilleure flexibilité dans l'action de Rnt1p qui peut être impliquée autant dans la maturation que dans la dégradation ou la régulation d'un ARN.

Malgré qu'environ la moitié des substrats connus soient conservés chez quatre espèces ou plus, parmi les 9 substrats ARN messagers, aucun n'est conservé chez plus d'une espèce. Il semble donc que la régulation spécifique d'un ARNm par Rnt1p ne soit pas un mécanisme conservé entre les espèces, mais plutôt une adaptation spécifique de *S. cerevisiæ* à son environnement de croissance. Il serait intéressant de valider la présence des substrats ARNm chez d'autres souches de *S. cerevisiæ* et de valider l'influence de la présence ou de l'absence d'un substrat sur l'efficacité de croissance dans différentes conditions de culture. La façon dont Rnt1p régule le niveau des ARNm est aussi variée. Certains transcrits sont régulés de façon constitutive, d'autres de façon conditionnelle (Lavoie $\mathrm{et\ al.}$, 2012) et d'autres selon la phase du cycle cellulaire (Larose $\mathrm{et\ al.}$, 2007). Aussi, Rnt1p affecte le niveau d'expression de certains ARNm par l'intermédiaire du promoteur sans qu'il y ait de coupure et la coupure par Rnt1p peut être cotranscriptionnelle (Ghazal $\mathrm{et\ al.}$, 2009). Il serait donc important pour l'étude de la régulation des ARNm d'établir des modèles permettant d'isoler chacun de ces modes de régulation.

L'identification de nouveaux substrats conservés dans les régions intergéniques a permis de révéler un nouveau rôle de Rnt1p (Ghazal $\mathrm{et\ al.}$, 2009). Ainsi, Rnt1p serait associé à l'ARN polymérase II durant la transcription et dans certains cas, il agirait comme terminateur. Dans

les deux cas répertoriés, NPL3 et RPL8A, il s'agit de gènes ayant un niveau d'expression élevé et possédant aussi au moins un autre site de terminaison classique en amont du site de terminaison dépendante de Rnt1p. Dans la terminaison classique (type torpedo), l'ARNm en cours de synthèse est coupé par un facteur de clivage et l'extrémité 5' produite est dégradée par l'exoribonucléase Rat1p. Lorsque Rat1p rejoint la polymérase, celle-ci se détache de l'ADN. Dans la terminaison dépendante de Rnt1p, Rnt1p reconnait un de ses substrats et effectue la coupure qui produit une extrémité 5'. Il est possible que ce mode de terminaison soit plus courant que les deux exemples détectés. Ces deux exemples possèdent un substrat de Rnt1p très conservé et un site de terminaison classique faible en amont du site de coupure. Dans ces cas, la terminaison par Rnt1p peut jouer un rôle de protection du promoteur du gène situé en aval.

2.4.3 Limites de la classification

Toutes les approches de classification ont échoué à identifier un nombre raisonnable de candidats. Le génome de la levure compte environ 6 000 gènes et la régulation par les RNases III chez d'autres espèces est une régulation spécifique de certains gènes. Il serait donc souhaitable d'obtenir beaucoup moins que 6 000 candidats. Dans différentes études (Catala $\mathrm{et\ al.}$, 2012; Lavoie $\mathrm{et\ al.}$, 2012), plusieurs dizaines d'ARNm contenant des candidats ont été testés et le taux de coupure obtenu a été inférieur à 10 % des candidats testés. Pourtant plusieurs candidats ont été synthétisés sous forme d'ARN courts et sont coupés *in vitro*. Les conditions de coupure, ARN total ou candidat seul, jouent un rôle important dans les résultats de coupure *in vitro* dû à la compétition entre les substrats de différentes efficacités. Cette différence peut aussi être due à un repliement différent à l'intérieur d'un transcrit long ou à une question d'accessibilité pour Rnt1p.

Il est aussi probable que l'accessibilité dans un long ARNm soit importante pour la détection de la coupure. La prédiction de cette accessibilité par les outils bio-informatiques de prédiction de structure est difficile et imprécise. Il a aussi été montré que pour un même ARNm tous les substrats n'ont pas la même efficacité (Meaux $\mathrm{et\ al.}$, 2011). Ainsi, certains

substrats guident la coupure lorsque insérés à l'intérieur d'un transcrit artificiel, alors que d'autres insérés à l'intérieur du même transcrit à la même position ne guident pas la coupure dans des conditions physiologiques.

Les algorithmes de classification binaire courants (classificateur bayésien naïf et SVM) ont obtenu des résultats de 10 000 à 25 000 candidats. Il est probable que cela soit dû à la petite taille des ensembles d'entraînement. Le nombre de substrats connus est faible (51) et l'ensemble d'entraînement négatif contient peu d'exemples similaires à des substrats, mais non coupés. De plus, le grand nombre d'attributs utilisés, 50 nt, ne semble pas être la cause de la difficulté de classification. Augmenter la taille de l'ensemble d'entraînement permettrait d'augmenter la performance des classificateurs.

Pour cette raison, un algorithme dirigé a été conçu. Il réduit considérablement le nombre de candidats, mais il fait aussi plusieurs hypothèses qui ne permettent pas de découvrir de nouvelles classes de substrats. Selon sa conception actuelle, les candidats ne possédant pas un G en deuxième position de la tétraboucle et trois paires de bases fermantes sont rejetés. Si les mêmes critères étaient appliqués à l'algorithme de classification bayésien naïf, seulement six candidats seraient trouvés dans une région de 10 000 nt du génome. Cela amène l'estimation pour le génome complet à environ 7 500 soit un nombre comparable avec l'algorithme de recherche par similitude. La seconde amélioration importante que comporte l'algorithme dirigé est l'assignation de poids à chacune des positions en fonction de leur degré de conservation. Cela a un effet semblable à la réduction du nombre d'attributs. Il est donc possible de supposer que l'assignation manuelle des poids a permis de réduire d'environ la moitié le nombre de candidats.

Pour la recherche par modèle de covariance, le faible nombre d'espèces (5) est une cause probable des résultats obtenus. La production d'un modèle de covariance unique regroupant tous les substrats et leurs orthologues permettrait probablement d'obtenir un meilleur résultat, mais il n'a pas été possible de produire un alignement satisfaisant représentant toute la variété des substrats.

Une autre approche à vérifier serait d'utiliser tous les orthologues pour entraîner les classificateurs binaires. Il est prévisible que cela permettrait au classificateur d'accorder plus de poids aux positions conservées grâce au nombre accru d'exemples. Cependant, pour le classificateur bayésien, cela violerait l'hypothèse selon laquelle les exemples d'entraînement sont indépendants et cela pourrait provoquer un surentraînement pour les substrats les plus conservés. Quant à lui, le classificateur SVM est moins sensible à ce type de violation. Pour valider cette avenue, un classificateur SVM a été entrainé sur tous les orthologues disponibles (190). Le classificateur a une performance grandement améliorée. Son coefficient de corrélation de Matthews moyen est de 0.94. Sa spécificité moyenne est de 99.8 % et sa sensibilité est de 91.6 %. Sur une région génomique aléatoire, il identifie 24 candidats ce qui permet d'estimer à 30 000 le nombre de candidats dans le génome. Bien qu'il identifie plus de candidats que la recherche dirigée par similitude, son excellente performance en fait une alternative à ne pas négliger.

L'utilisation de la conservation comme critère de sélection additionnel pour les candidats classés positifs a été un succès. Sur huit nouveaux candidats retenus, cinq ont été confirmés comme substrats véritables. De plus, une nouvelle classe de cibles de Rnt1p a été établie, les terminateurs, et un nouvel ARNnc a été identifié.

Bien que ce soit un succès pour les régions non codantes, le critère de conservation interespèces n'est pas applicable pour les ARNm. D'abord, il est difficile de séparer la conservation due à la fonction de la protéine de la conservation due à la structure ARN. Ensuite, pour les substrats ARNm connus la structure reconnue par Rnt1p n'est pas conservée. Le problème de l'identification de nouveaux substrats ARNm reste donc très difficile. Cependant, la prédiction *in silico* permet tout de même de réduire le nombre de candidats à tester et elle a été utilisée dans plusieurs travaux comme premier outil de sélection des cibles potentielles (Catala et al., 2012; Larose et al., 2007; Lavoie et al., 2012).

2.5 Contributions

L'auteur a sélectionné les algorithmes utilisés, a implémenté le code nécessaire à leur application, a conçu les sondes pour la validation et a analysé les résultats. L'algorithme de classification par similitude est basé sur le travail de Julien Gervais-Bird. La validation *in vitro* des substrats a été effectuée par Ghada Ghazal. Gabriel Girard a participé au choix des algorithmes. Sherif Abou Elela a participé aux choix des critères et à l'analyse des résultats.

2.6 Résumé de l'impact

Le travail de ce chapitre a été utilisé dans un article soumis pour publication (Gagnon et al., soumis à PLoS Genetics). Les résultats de la prédiction *in silico* sont couramment utilisés pour la recherche de nouveaux substrats potentiels. Ils ont été utilisés dans la production de plusieurs artciles publiés (Catala $\mathrm{et\ al.}$, 2012; Larose $\mathrm{et\ al.}$, 2007; Lavoie $\mathrm{et\ al.}$, 2012). L'identification de substrats intergéniques conservés a mené à une article publié (Ghazal $\mathrm{et\ al.}$, 2009) et un second article est en préparation.

CHAPITRE 3

IDENTIFICATION DE NOUVEAUX TRANSCRITS ARN INFLUENCÉS RNT1P GRÂCE À L'UTILISATION DE PUCES À ADN COUVRANT TOUT LE GÉNOME

3.1 Introduction

Malgré les efforts pour trouver un algorithme plus performant, la technique d'identification de substrats par la recherche *in silico* génère un nombre important de candidats. Il donc nécessaire de se tourner vers une approche haut débit de validation qui mesure l'effet de l'enzyme sur les transcrits.

3.1.1 Puces à ADN

Il existe plusieurs technologies de puces à ADN qui varient tant par le processus de synthèse que par la longueur des oligonucléotides et leur densité. La technologie utilisée par d'Affymetrix est celle qui permet la plus grande densité : une surface d'un peu plus de 1 cm^2 peut contenir plus de 6 millions de sondes de 25 nucléotides chacune. Les sondes sont synthétisées directement sur la surface par un procédé de photolithographie. Les puces à ADN couvrant le génome de *S. cerevisiæ* comptent 3.2 millions de sondes parfaitement complémentaires à 25 nucléotides du génome. Elles couvrent tout le génome avec un décalage moyen de quatre nucléotides et sont complémentaires au brin Crick du génome.

À la suite de l'extraction de l'ARN de la culture à étudier, une réaction de transcription inverse est effectuée pour synthétiser un brin d'ADN complémentaire à l'ARN source. Ensuite, si nécessaire, une étape de polymérisation de l'ADN peut être effectuée si l'on désire obtenir les deux brins d'ADN. L'ARN est ensuite dégradé et l'ADN est fragmenté pour avoir une hybridation plus uniforme. Les fragments d'ADN sont marqués à une extrémité avec une molécule de biotine. L'ADN marqué est hybridé avec la puce à ADN et un anticorps qui cible la biotine est finalement ajouté pour permettre la lecture de la puce à ADN par un laser. Une image numérique de la puce à ADN est obtenue et est convertie en données brutes d'intensité pour chacune des sondes.

D'autres fabricants de puces à ADN possèdent des technologies ayant des caractéristiques différentes comme des sondes plus longues, des puces réutilisables ou des sensibilités différentes. Pour la cartographie du transcriptome, la technologie d'Affymetrix est la plus appropriée et la plus utilisée compte tenu de sa haute résolution. Pour la mesure d'expression différentielle, une très haute résolution n'est pas essentielle et la spécificité des sondes est plus importante. Certains fabricants produisent donc des sondes plus longues, alors que d'autres utilisent plusieurs sondes par transcrits.

3.1.2 Rôles de Rnt1p

Rnt1p est une endoribonucléase de type III nucléaire. La coupure en 3' du transcrit de l'ARNr 25S constitue l'activité principale de Rnt1p dans la cellule. Il s'agit d'une des nombreuses étapes menant à l'assemblage des ribosomes. Cet assemblage s'effectue dans un sous-compartiment du noyau appelé nucléole. Ainsi, Rnt1p est principalement localisée dans le nucléole où il participe à la maturation des ARNr.

Le second rôle le plus connu de Rnt1p est la maturation des transcrits précurseurs de certains ARNnc. Le plus grand nombre de ses substrats se trouvent dans la catégorie des ARNsno, des ARNnc qui guident la modification des ARNr. Rnt1p participe à la maturation

de la majorité des ARNsno. Cependant, certains ARNsno peuvent être maturés indépendamment de Rnt1p par des exoribonucléases. Rnt1p participe aussi à la maturation des ARNsn. Encore une fois, Rnt1p n'est pas toujours essentiel à cette maturation.

Rnt1p est aussi impliqué dans la progression du cycle cellulaire. Son absence provoque un délai dans la phase G1 du cycle et des problèmes de division cellulaire (Catala $\mathrm{et\ al.}$, 2004). Durant le cycle cellulaire, la localisation de Rnt1p change. Durant la phase G2/M, il passe du nucléole vers le nucléoplasme. Ce changement de localisation implique que Rnt1p peut cibler n'importe quel transcrit et non seulement ceux accessibles au nucléole.

Il a été montré que Rnt1p joue un rôle régulateur chez des transcrits ARNm impliqués dans l'adaptation de la cellule aux conditions de croissance. En effet, il participe à la répression dépendante du glucose du gène MIG2 (Ge $\mathrm{et\ al.}$, 2005) et il joue un rôle dans le maintien du niveau intracellulaire de fer et dans la prévention de l'intoxication par le fer (Lee $\mathrm{et\ al.}$, 2005). L'absence de Rnt1p provoque aussi de nombreux autres effets phénotypiques. La croissance des cellules *rnt1*Δ est très lente et cette souche est sensible à la température de croissance. Plusieurs cellules ont aussi des formes anormales lorsqu'observées au microscope. Il est donc probable que Rnt1p agisse sur de nombreuses voies métaboliques.

3.1.3 Utilisation de puces à ADN pour l'identification de substrats

L'utilisation de puces à ADN détectant la variation du niveau d'expression des transcrits est une application courante. L'inactivation du gène d'intérêt permet de détecter les transcrits ciblés par ce gène. Dans le cas de Rnt1p, trois études ont utilisé les puces à ADN pour détecter des transcrits ciblés par Rnt1p.

Les deux premières études (Ge $\mathrm{et\ al.}$, 2005; Ghazal $\mathrm{et\ al.}$, 2005) ont combiné l'emploi de puces à ADN avec la prédiction *in silico* de substrats pour identifier respectivement un substrat ARNm et des substrats ARNsno. Dans la troisième étude (Lee $\mathrm{et\ al.}$, 2005), un

groupe de 9 ARNm a été identifié comme surexprimé en l'absence de Rnt1p. Cependant, seulement deux de ces ARNm ont été confirmés comme étant des substrats. Les puces à ADN utilisées lors de ces études fournissent uniquement de l'information sur le niveau d'expression du transcrit mature et ne permettent pas de savoir si l'expression de tout le transcrit varie de la même façon ou si seulement certaines régions s'accumulent. Il est possible qu'en l'absence de Rnt1p certaines régions non coupées par Rnt1p s'accumulent sans que le niveau d'expression global du transcrit soit affecté de façon importante. De plus, la préparation des échantillons pour ce type de puces à ADN fait généralement appel à une étape d'amplification des transcrits polyadénylés. Les transcrits non codants matures n'étant généralement pas polyadénylés, cette étape rend plus difficile l'identification de nouveaux substrats ARNnc.

L'apparition des puces à ADN couvrant tout le génome a permis de faire une cartographie complète de tout le transcriptome de la levure (Huber $\mathrm{et\ al.}$, 2006). Cette technique a permis de déterminer la taille des régions non traduites des ARNm et a permis l'identification de nombreux nouveaux transcrits ARNnc. Cette technique a aussi été utilisée en combinaison avec l'inactivation d'une exoribonucléase, Rrp6p, pour identifier un grand nombre de ses substrats et mieux comprendre son rôle dans la surveillance des transcrits aberrants.

3.1.4 Objectifs

Puisque l'approche de prédiction *in silico* des substrats génère un grand nombre de faux positifs et que les puces à ADN traditionnelles ne permettent pas d'avoir une résolution suffisante pour détecter l'accumulation de régions non maturées, des puces à ADN couvrant tout le génome seront utilisées pour identifier *in vivo* les transcrits affectés par l'absence de Rnt1p. L'ARN utilisé ne sera pas enrichi en transcrits polyadénylés de façon à pouvoir détecter sans biais les transcrits non codants. L'ARN de la souche inactivée pour Rrp6p sera testé pour confirmer la capacité de détecter l'accumulation de transcrits affectés par cette exoribonucléase. L'ARN de la souche inactivée pour Xrn1p sera testé pour confirmer la capacité de détecter l'effet d'une ribonucléase ciblant les transcrits ARNm.

Ces résultats permettront à la fois d'identifier les transcrits surexprimés, les transcrits ayant un défaut de maturation et des nouveaux transcrits non codants. De plus, le portrait global de l'effet de Rnt1p sur le transcriptome permettra de mieux comprendre le rôle biologique de Rnt1p dans le métabolisme de la cellule.

3.2 Méthodes

3.2.1 Extraction de l'ARN

L'ARN total a été extrait par extraction phénol-chloroforme à partir de cellules de levure *Saccharomyces cereviæ* des souches BY4703 et BY4741 en phase de croissance exponentielle sur milieu riche YPD. La qualité de l'ARN a été validée sur un appareil Agilent 2100 Bioanalyzer (Agilent technologies, Santa Clara, CA).

3.2.2 Synthèse de l'ADNc

L'ADNc double brin a été synthétisé à partir d'ARN total en utilisant le protocole de marquage terminal à la biotine Affymetrix. L'ADNc simple brin a été synthétisé à partir d'ARN total par transcription inverse en utilisant des hexamères aléatoires et la transcriptase inverse SuperScript II. Le gabarit d'ARN a été dégradé par les RNases A et H. L'ADNc a été fragmenté grâce à la DNase I et marqué à la biotine à son extrémité.

3.2.3 Hybridation aux puces à ADN

L'ADNc marqué à la biotine a été hybridé sur des puces Affymetrix GeneChip S. cerevisiæ Tiling 1.0R Array (PN : 900645).

3.2.4 Annotation des sondes

Les sondes ont été annotées d'après le génome de *Saccharomyces cerevisiæ* souche S288c provenant de la "Saccaromyces Genome Database" (yeastgenome.org) version du 10 août 2007. L'annotation a été faite telle que décrite dans David $\mathrm{et\ al.}$ (2006)

Il a été noté que les sondes situées sur le chromosome 11 ont globalement un niveau d'expression supérieur au reste du génome dans la souche *xrn1*Δ. Elles n'ont pas été utilisées pour la suite des analyses par crainte d'une polyploïdie du chromosome 11 dans cette souche. Le même phénomène a aussi été observé dans les résultats d'un autre groupe ayant utilisé la même souche.

3.2.5 Normalisation

L'étude des variations des réponses des sondes a permis de montrer que la distribution des réponses n'est pas aléatoire par rapport à la stabilité thermodynamique prédite (ΔG) pour leur duplex ADN-ADN (Voir annexe B, Figure B.1A). Les méthodes utilisées pour l'analyse de données de puces à ADN corrigent partiellement pour cette corrélation. Cependant, même après ces corrections, il existe encore une faible corrélation résiduelle. Elle est négligeable pour la plupart des analyses, mais doit être corrigée pour réduire le taux de faux positifs dus à des variations de composition locales dans le génome. Ainsi, le 5 % des sondes ayant un ΔG le plus élevé a été retiré des données. Pour les autres sondes, le biais

de réponse est linéaire par rapport au ΔG calculé. Ainsi, une étape de correction pour égaliser ce biais a été ajoutée au traitement des données standard (David $\mathrm{et\ al.}$, 2006). Après l'ajout de ces étapes à la procédure de traitement des données, le ratio signal bruit a été réduit (Annexe B, Figure B.1B).

3.2.6 Segmentation

Après la normalisation, les données de puces à ADN provenant de la souche de type sauvage ont été soustraites des données de chaque souche mutante. La différence a été segmentée comme cela est décrit dans Huber $\mathrm{et\ al.}$ (2006) en utilisant la bibliothèque tilingArray du langage R. Le seul paramètre de la segmentation est le nombre de segments. Le nombre de segments a été sélectionné en utilisant le critère d'information Bayésien (BIC). La segmentation a été répétée à plusieurs reprises avec un nombre de segments attendus différent. La valeur BIC a été calculé pour chaque valeur du nombre de segments et la valeur optimale a été choisie pour chaque chromosome.

3.2.7 Mesure du niveau d'expression des transcrits connus

Le niveau d'expression correspond au niveau médian de toutes les sondes contenues entre les positions de l'annotation selon la "Saccharomyces Genome Database" (SGD) du 10 août 2007. Les niveaux d'expression de chacune des souches ont été ajustés de façon à ce que trois régions absentes de chacune des souches aient une variation nulle.

3.2.8 Identification des régions intergéniques non annotées

Les régions intergéniques sont toutes les régions qui ne sont pas annotées comme région codante, comme ARNnc ou comme intron, ainsi que les 50 nt attenants à ces annotations. Les transcrits intergéniques montrant une variation d'expression sont des segments dans des régions intergéniques ayant un changement de niveau d'expression médian supérieur à la somme de la médiane et de 1.96 fois la variation absolue médiane ($p < 0.05$) de tous les segments de plus de 48 nt.

3.2.9 Validation

L'expression des transcrits non codants a été validée par PCR semi-quantitatif radioactif. De l'ARN total a été extrait de souches *rnt1*Δ et de type sauvage. L'ARN a été traité avec DNase pour éliminer la contamination par l'ADN génomique. Une réaction de transcription inverse a été effectuée avec des hexamères aléatoires et une amplification PCR a été effectuée avec des amorces spécifiques aux transcrits à valider. Les produits PCR ont subi une migration sur gel et ont été quantifiés d'après leur niveau de radioactivité. Le niveau d'expression ainsi mesuré a été normalisation avec un gène exprimé constitutivement, l'actine (ACT1).

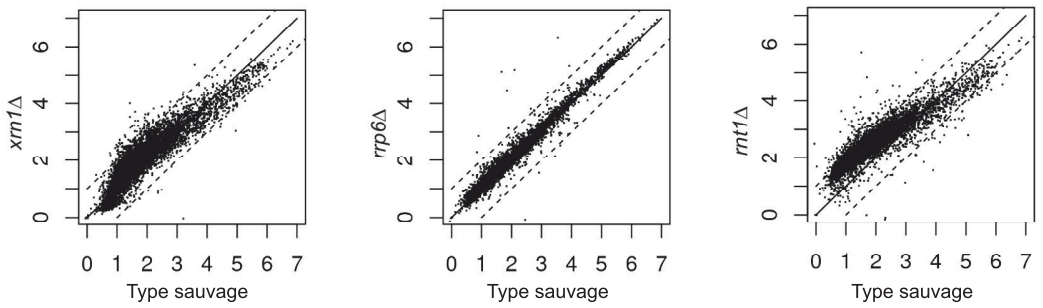


figure 3.1 – Comparaison des niveaux d'expression des ARNm dans trois souches mutantes pour une ribonucléase avec la souche de type sauvage.

Chaque point représente les niveaux d'expression d'un exon.

3.3 Résultats

3.3.1 Les RNases influent sur le niveau d'expression de nombreux transcrits

La comparaison des niveaux d'expression des ARNm montre un effet très différent de chacune des ribonucléases. L'absence de XRN1 provoque une augmentation générale du niveau des ARNm et plus particulièrement des ARNm d'expression moyenne. Dans la souche où RRP6 est absent, le niveau des ARNm est en général peu affecté, mais quelques ARNm spécifiques sont surexprimés. Dans la souche *rnt1*Δ, une surexpression globale est aussi observée. Cependant, contrairement à la souche *xrn1*Δ, même les transcrits qui sont normalement peu ou pas exprimés dans la souche sauvage s'accumulent en absence de RNT1.

Les transcrits surexprimés par un facteur supérieur à deux font partie de certaines voies métaboliques particulières. La majorité de ces voies sont associées à la respiration mitochondriale (Voir Table 3.1).

tableau 3.1 – Liste des voies métaboliques surexprimées dans la souche *rnt1*Δ.

Voie métabolique	Valeur p
Respiratory electron transport chain	1.93E-006
ATP synthesis coupled electron transport	1.93E-006
Mitochondrial ATP synthesis coupled electron transport	1.93E-006
Oxidative phosphorylation	3.52E-006
Mitochondrial electron transport, ubiquinol to cytochrome c	5.44E-006
Electron transport chain	2.53E-004
Cellular respiration	1.41E-003
Aerobic respiration	1.36E-002
Energy derivation by oxidation of organic compounds	1.84E-002

Les voies métaboliques (ontologies géniques de type processus biologique) significativement enrichies (valeur p < 0.05 après correction de Holm-Bonferroni) en gènes surexprimés dans la souche ne contenant pas RNT1 sont présentées.

3.3.2 Les ARNsno maturés par Rnt1p sont surexprimés

Conformément aux attentes, les segments où le plus de surexpression est détecté sont associés à des ARNsno substrats de Rnt1p. Plus de 80 % des ARNsno substrats de Rnt1p sont associés à des segments surexprimés au moins deux fois. La plupart des ARNsno non détectés sont situés dans des introns ou font partie de transcrits polycistroniques.

Cependant, dans la plupart des cas, le niveau d'expression du transcrit mature varie peu. Le segment qui est détecté comme surexprimé est un segment situé en 5' du transcrit mature et couvrant la région où se situe la structure reconnue par Rnt1p. Ainsi, l'absence de Rnt1p provoque une accumulation du transcrit précurseur, mais ne provoque pas une surexpression de l'ARNsno.

Une surexpression est aussi détectés pour quatre ARNsno qui ne sont pas connus comme substrats de Rnt1p. Les quatres possèdent une tige-boucle NGNN en 5' du transcrit mature. snR85 a été testé et est confirmé comme substrat (Voir Figure 3.2).

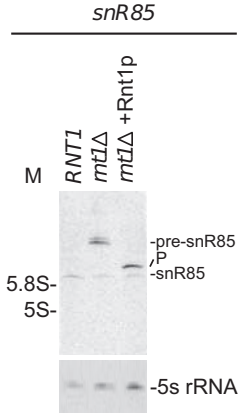


figure 3.2 – Validation de snR85 comme nouveau substrat de Rnt1p.

Validation par buvardage Northern de la coupure du précurseur de snR85 par Rnt1p purifié.

3.3.3 Des transcrits non annotés sont surexprimés en absence de certaines RNases

L'absence des ribonucléases n'a pas seulement un effet sur le niveau des ARNm, mais provoque aussi l'accumulation de certains transcrits intergéniques non codants. Ainsi, environ 35 transcrits surexprimés sont détectés dans les souches *xrn1*Δ et *rrp6*Δ, alors que dix le sont dans la souche *rnt1*Δ (Figure 3.3A). Beaucoup plus de transcrits avaient été détectés dans les études précédentes sur *rrp6*Δ. La différence peut être expliquée par l'utilisation d'une amplification des transcrits polyadénylés par ces études. Les transcrits non codants aberrants sont souvent polyadénylés avant d'être dégradés par l'exosome.

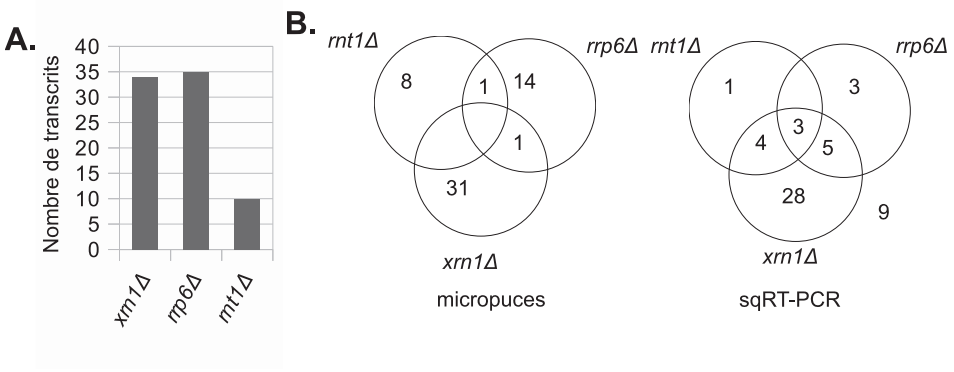


figure 3.3 – Accumulation de certains transcrits intergéniques non codants dans les souches déficientes en ribonucléases.

A Nombre de transcrits intergéniques surexprimés dans chacune des souches. B Validation des transcrits intergéniques par réaction PCR semi-quantitative. Le diagramme micropuces indique les transcrits qui ont été sélectionnés pour validation, alors le diagramme sqRT-PCR indique le résultat de la validation.

3.3.4 L'expression de certains transcrits peut être confirmée par d'autres techniques

Parmi les transcrits intergéniques surexprimés, 55 ont été choisis pour être validés. L'existence de 53 d'entre eux a été confirmée et la surexpression dans l'une ou l'autre des souches mutantes a aussi été confirmée à l'exception de neuf cas (Figure 3.3B). Xrn1p joue un rôle dans le niveau d'expression de la plupart des ARN intergéniques comme c'est le cas pour les ARNm.

3.3.5 Absence de structure reconnue par Rnt1p dans les transcrits non annotés régulés par Rnt1p

Puisque plusieurs ARNnc sont régulés par Rnt1p, la présence de structures ressemblant aux substrats de Rnt1p a été vérifiée. Seulement deux transcrits intergéniques contiennent une structure ayant un pointage supérieur au seuil de 0.85. La coupure par Rnt1p n'a pas pu être confirmée pour un d'entre eux, mais l'autre est un ARNnc qui avait été identifié par recherche de substrats conservés et il a été confirmé (Voir Figure 3.4).

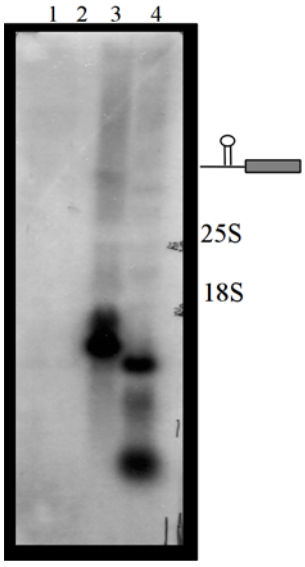


figure 3.4 – Validation de la coupure d'un transcrit non codant associé au gène CHD1

Détection par buvardage Northern de la coupure d'un transcrit non codant associé au gène CHD1. Une sonde couvrant la région de la tige-boucle prédite a été utilisée. La piste 1 montre l'ARN de la souche de type sauvage. La piste 2 montre l'ARN de la souche de type sauvage traité avec Rnt1p purifié. La piste 3 montre l'ARN de la souche *rnt1*Δ. La piste 4 montre l'ARN de la souche *rnt1*Δ traité avec Rnt1p purifié.

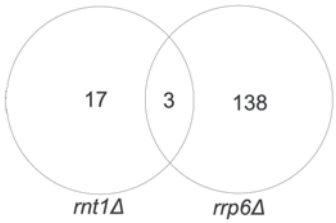


figure 3.5 – Détection de transcrits non codants exprimés à partir du brin opposé à des ARNm

3.3.6 Des transcrits qui proviennent du brin opposé à des ARNm

L'utilisation de puces à ADN spécifiques au brin Watson a permis l'identification de 158 transcrits non codants qui proviennent de la transcription du brin opposé à un transcrit ARNm connu qui sont surexprimés en absence d'une ribonucléase (Figure 3.5). La majorité de ces transcrits sont détectés dans la souche *rrp6*Δ, mais il en existe 20 qui sont surexprimés dans la souche *rnt1*Δ.

3.4 Discussion

3.4.1 Effets des ribonucléases sur les ARN connus

L'absence de XRN1 provoque l'accumulation de la plupart des ARNm et des ARNnc connus. Parmi les rôles de XRN1, la dégradation des ARNm est connue comme sa fonction principale et les variations des niveaux d'expression des ARNm correspondent à ce qui est attendu pour un enzyme qui participe à la dégradation de la plupart des ARNm. Cependant, des mécanismes redondants existent et expliquent que tous les ARNm ne sont pas affectés au même niveau. Cependant, la dégradation d'un ARN par Xrn1p peut s'effectuer seulement en présence d'une extrémité 5' phosphate. Certains ARNnc ont naturellement une

extrémité 5' phosphate et doivent être protégés par leur localisation ou un complexe ribonucléoprotéique de la dégradation par Xrn1p. Cependant, la majorité des transcrits codants sont protégés de la dégradation par Xrn1p par la présence d'une coiffe à leur extrémité 5'. Ils doivent donc avoir subi une étape d'enlèvement de la coiffe avant de pouvoir être dégradés par Xrn1p. Cette étape supplémentaire combinée avec les mécanismes redondants de dégradation explique la grande variabilité de l'effet de l'absence de XRN1 sur le niveau des ARN.

RRP6 a été principalement impliqué dans la dégradation de transcrits aberrants et dans la maturation d'ARN non codants. Il est donc logique qu'un très petit nombre d'ARNm soient affectés par son absence. Cependant, l'expression de plusieurs ARNnc est augmentée en absence de RRP6. Rrp6p est impliqué dans la formation de l'extrémité 3' de plusieurs ARNnc. De plus, en absence de RRP6 plusieurs transcrits aberrants associés au ARNnc s'accumulent. Ces transcrits aberrants semblent associés à des défauts de la terminaison qui provoquent une extension 3' pouvant s'étendre sur plusieurs centaines de nucléotides. Ces transcrits aberrants constituent probablement des cibles pour Rrp6p et l'exosome nucléaire auquel il peut être associé. Il est possible que les quelques ARNm surexprimés dans la souche *rrp6*Δ soient aussi des ARN ayant un défaut dans leurs terminaisons.

La fonction principale de RNT1 dans la cellule est de participer à la maturation de l'ARN ribosomal 25S. Il est aussi impliqué dans la maturation d'autres ARNnc et dans la régulation de certains ARNm. Il est possible que l'effet observé sur le niveau global d'expression des ARNm soit indirect, c'est-à-dire qu'il s'agit d'un effet provenant d'une des cibles de Rnt1p. Cependant, il a aussi été observé que Rnt1p est associé à la polymérase (Ghazal $\mathrm{et\ al.}$, 2009). Donc, en plus d'avoir un effet spécifique sur certains transcrits, son absence pourrait avoir une influence globale sur la transcription puisqu'il a été observé que la présence de polymérase II dans les régions intergéniques est augmentée en absence de RNT1. La plupart des ARNnc qui sont maturés par Rnt1p ont une extension à leur extrémité 5' laquelle contient une structure ciblée par Rnt1p. Cependant, pour certains ARNnc, la forme mature est aussi détectée même en absence de Rnt1p. Il est donc probable qu'il existe deux sites d'initiation de la transcription ou sinon un mécanisme alternatif de maturation 5' existe. Dans

un cas comme dans l'autre, la régulation du niveau de l'ARN mature pourrait se faire par l'intermédiaire de la maturation 5'.

3.4.2 Transcrits non codants de fonction inconnue

Les nouvelles techniques de cartographie haut débit de l'ARN ont permis l'identification d'un grand nombre de transcrits intergéniques non codants dont la fonction n'a pas été déterminée. L'étude des variations de niveaux d'expression de ces ARN dans différentes conditions permet d'augmenter nos connaissances sur ces ARN.

Les transcrits non codants détectés dans la souche sauvage et surexprimés dans la souche *rrp6*Δ sont appelés SUT (Davis et Ares, 2006) et ceux qui sont seulement détectés dans la souche *rrp6*Δ sont appelés CUT. Les transcrits surexprimés dans la souche *xrn1*Δ sont quant à eux appelés XUT (van Dijk $\mathrm{et\ al.}$, 2011). Ces deux types de transcrits ont été détectés dans notre étude. Cependant, la distinction entre SUT et CUT est plutôt arbitraire et dépendante de plusieurs facteurs comme la sensibilité de la méthode, la souche ou les conditions de culture. Dans notre cas, certains CUT ont été détectés dans la souche de type sauvage, alors que certains SUT ont été détectés seulement dans la souche *rrp6*Δ.

Il est important de noter que la plupart des études transcriptomiques utilisent des échantillons enrichis pour les ARN polyadénylés de façon à augmenter la sensibilité et réduire la quantité d'ARNr dans l'échantillon. Cependant, plusieurs transcrits aberrants subissent une étape de polyadénylation avant d'être dégradés par l'exosome. Dans cette étude, il n'y a pas eu d'enrichissement pour les transcrits polyadénylés de façon à éviter un biais pour les transcrits aberrants. Ainsi, pour les souches *rrp6*Δ et *xrn1*Δ, le nombre de transcrits de fonction inconnue détectés est inférieur aux autres études transcriptomiques utilisant ces souches. Cette différence peut aussi être due à la technologie utilisée et à l'analyse.

Finalement, une nouvelle classe de transcrits non codants a été identifiée : des transcrits non codants surexprimés dans la souche *rnt1*Δ (Voir Table 3.2). Cependant, à l'exception d'un de ces transcrits, il n'a pas été possible de montrer une régulation directe de Rnt1p sur ces ARNnc. Il est possible que la surexpression soit causée par un effet indirect, mais il peut aussi s'agir d'une régulation indépendante de la coupure par Rnt1p ou bien un site de coupure inhabituel peut être présent. Il est important de noter que ces transcrits peuvent être détectés dans la souche de type sauvage. Il ne s'agit donc pas de transcrits aberrants présents seulement dans la souche *rnt1*Δ, mais bien de transcrits présents en tout temps dans la cellule à différents niveaux d'expression.

tableau 3.2 – Liste des transcrits intergéniques surexprimés dans la souche *rnt1*Δ.

Chromosome	Position de début	Position de fin	Taille (nt)	Variation d'expression
chrI	199404	200132	729	6.3
chrII	643793	644974	1182	4.0
chrV	503987	504781	795	13.6
chrIX	93076	93249	174	19.3
chrIX	387167	388345	1179	8.9
chrXII	1052456	1053208	753	8.0
chrXIII	396472	396614	143	3.6
chrXIII	480855	481185	331	3.9

La colonne variation d'expression indique le ratio de l'expression dans la souche *rnt1*Δ par rapport à la souche de type sauvage obtenu par PCR semi-quantitatif.

3.4.3 Transcrits antisens

L'utilisation d'ADN complémentaire préparé de façon à conserver l'information sur l'orientation initiale du transcrit permet d'identifier des transcrits provenant du brin opposé à un autre ARN déjà connu. Ces ARN sont appelés ARN antisens.

De nombreux transcrits antisens sont surexprimés dans la souche *rrp6*Δ comme cela a déjà été montré (Neil $\mathrm{et\ al.}$, 2009). Ces transcrits proviendraient de transcription aberrante

dans la direction opposée de promoteurs connus ou de régions libres en nucléosomes. Des transcrits antisens sont aussi détectés dans la souche *rnt1*Δ. La source de ces transcrits n'est pas établie. Les puces à ADN utilisées ont permis la détection sur seulement un des deux brins du génome, donc d'autres transcrits antisens pourraient être détectés en utilisant une autre technologie.

La préparation de l'ADNc peut produire des ADN qui semblent provenir d'un antisens, mais sont en fait des artéfacts. Il est possible de modifier le protocole de préparation de l'ADNc pour empêcher ces artéfacts (Perocchi $\mathrm{et\ al.}$, 2007) et il sera important de le faire lors de la validation de ces transcrits. Malgré qu'il soit possible que certains transcrits détectés dans la souche *rnt1*Δ soient des artéfacts de la préparation des échantillons, au moins un antisens a été confirmé par le travail d'un autre groupe (Camblong $\mathrm{et\ al.}$, 2007).

Cet antisens associé au gène sens PHO84 est aussi surexprimé dans la souche *rrp6*Δ. Il a été montré que son expression a un effet répressif sur l'expression du gène sens. Cette répression serait due à la modification des histones. Cet effet de répression est aussi observé dans la souche *rnt1*Δ au point où la présence de l'ARN sens n'est pas détectée, alors qu'il l'est dans la souche *rrp6*Δ. Cette étude a donc permis la détection d'une deuxième classe de transcrits non codants surexprimés en absence de Rnt1p, mais le lien avec la fonction de la protéine n'a pas été établi.

Malgré la détection de deux nouvelles classes de transcrits non codants dont le niveau d'expression augmente en absence de RNT1, il n'a pas été possible d'identifier des nouveaux substrats directs de Rnt1p. De plus, pour les ARNm, l'augmentation du niveau d'expression est globale et non spécifique aux cibles de Rnt1p.

3.5 Contributions

L'auteur a participé au design expérimental, a effectué l'analyse complète des puces à ADN, a conçu les sondes pour la validation et a analysé les résultats. L'extraction d'ARN pour les puces à ADN a été effectuée par Mona Wu. La synthèse et le marquage de l'ADNc ont été effectués comme service au Wisconsin Gene Expression Center. L'hybridation aux puces à ADN a été effectuée comme service par le Centre for Applied Genomics at University of Toronto. La validation *in vitro* a été effectué par Sabrina Bossé et Francis Malenfant. Sherif Abou Elela a participé au design expérimental et à l'analyse des résultats.

3.6 Résumé de l'impact

Le travail de ce chapitre a été utilisé dans un article soumis pour publication (Gagnon et al., soumis à PLoS Genetics). Les résultats des analyses ont été utilisés dans deux articles publiés (Catala $\mathrm{et\ al.}$, 2012; Ghazal $\mathrm{et\ al.}$, 2009) et sont couramment utilisés comme référence dans l'étude de la variation d'expression des transcrits dans la souche *rnt1*Δ.

CHAPITRE 4

IDENTIFICATION DE NOUVELLES CIBLES DE RNT1P DANS LE TRANSCRIPTOME PAR ESSAI *IN VITRO*

4.1 Introduction

La recherche bio-informatique de cibles de Rnt1p a identifié plusieurs milliers de candidats dans le génome, à un tel point que la grande majorité des transcrits seraient ciblés. L'étude du transcriptome dans la souche *rnt1*Δ a montré une dérégulation globale de l'expression. Pourtant, les tests effectués sur des transcrits individuels ont eu un faible pourcentage de succès. La détection des ARNm coupés par Rnt1p reste un problème non résolu. Il est donc nécessaire de développer une approche permettant de détecter directement la coupure par Rnt1p.

4.1.1 Substrats de Rnt1p

Pour la majorité des substrats de Rnt1p, la coupure s'effectue par la reconnaissance d'une structure d'ARN comportant une tige d'ARN coiffée d'une tétraboucle AGNN. Il a été montré que l'adénosine en première position n'est pas essentielle, mais le remplacement de la guanine par un autre nucléotide abolit généralement la liaison. La coupure s'effectue généralement à 14 et 16 nt de chaque côté de la tétraboucle.

Il existe cependant de nombreuses exceptions à cette structure consensus (Voir figure 2.1). D'abord, la coupure d'un précurseur d'ARNsno est guidée par une tétraboucle AAGU plutôt

que AGNN ou NGNN. Ensuite, plusieurs substrats comportent des boucles dans la tige d'ARN qui sont exclues lors de la liaison. Ainsi, pour certains substrats, la coupure est détectée à plus de 100 nt de la tétraboucle guide. Finalement, il existe aussi un substrat dont la reconnaissance serait dépendante de la présence d'une protéine chaperonne.

En somme, une coupure standard par Rnt1p génère trois molécules d'ARN : le produit 5', la tige-boucle et le produit 3'. Le produit 5' possède une coiffe si l'ARN originel en possédait une et il se termine par un groupement 3'-OH. La tige-boucle a généralement 34 nucléotides de longueur, débute avec un 5'-phosphate et se termine avec un 3'-OH. Le produit 3' débute avec un 5'-phosphate et se termine comme le transcrit originel avec une queue polyA s'il s'agit d'un ARNm.

4.1.2 Isolation des produits

Les caractéristiques différentes des produits de réaction permettent d'envisager leur isolation sélective. Il existe des enzymes qui reconnaissent de façon spécifique les extrémités 5'-phosphate ou 3-'OH. De plus, la taille de l'ARN peut aussi être utilisée comme caractère discriminant.

Ainsi, l'exoribonucléase Xrn1p reconnaît de façon spécifique les transcrits ayant une extrémité 5'-phosphate et elle les dégrade de l'extrémité 5' vers l'extrémité 3' en enlevant chaque nucléotide successivement. Il est possible de concevoir que dans un échantillon d'ARN total traité par Rnt1p, Xrn1p ne laisserait que les transcrits non coupés protégés par une coiffe 5' et les produits de coupure 5' possédant aussi une coiffe. Les produits de coupure 3' et les tiges-boucles seraient dégradés. La disparition des produits 3' devrait être facilement détectable en utilisant des puces à ADN couvrant tout le génome ou tout autre technique de cartographie du transcriptome.

Les tiges-boucles coupées ont une très petite taille comparable à celle de certains petits ARNnc. Des techniques d'isolation des petits transcrits devraient permettre d'isoler ces produits de coupure pour ensuite permettre leur identification à haut débit. Cependant, puisque certains substrats ne correspondent par à la structure idéale, cette technique utilisée seule risque de ne détecter que les substrats idéaux.

4.1.3 Séquençage à haut débit

Les techniques de séquençage à haut débit permettent d'obtenir la séquence des molécules d'ARN contenus dans un échantillon. Elles offrent la possibilité de séquencer des millions de molécules d'ARN en une seule réaction. Ces techniques ont été utilisées pour faire la cartographie du transcriptome (Nagalakshmi $\mathrm{et\ al.}$, 2008). La résolution obtenue est fortement dépendante du nombre de molécules séquencées. Puisqu'environ 90 % de l'ARN total est constitué d'ARN ribosomaux, lorsque de l'ARN total est utilisé, il est nécessaire d'obtenir un très grand nombre de séquences pour obtenir une représentativité suffisante des ARN moins abondants.

Le séquençage à haut débit a aussi été utilisé pour la détection de petits ARNnc (Hafner $\mathrm{et\ al.}$, 2008; Lu $\mathrm{et\ al.}$, 2005). Dans ce cas, il existe de nombreuses techniques permettant d'enrichir l'échantillon en transcrits de la taille désirée. La technique la plus courante est l'isolation sur gel de polyacrylamide dénaturant. Cette technique permet de sélectionner les transcrits de la taille désirée d'après leur distance de migration.

Il est aussi possible de sélectionner des transcrits en utilisant les protéines qui les lient de façon spécifique. Cette technique consiste à lier de façon covalente l'ARN à la protéine, à isoler la protéine et ensuite à libérer l'ARN pour le séquencer (Sugimoto $\mathrm{et\ al.}$, 2012). Par exemple, la détection de microARN est possible en utilisant la protéine Argonaute qui lie les microARN lors de la coupure des ARNm cibles (Chi $\mathrm{et\ al.}$, 2009).

4.1.4 Objectifs

Les cibles de Rnt1p seront identifiées directement par une réaction de coupure *in vitro* qui subira ensuite une étape d'enrichissement de certains produits de coupure. Dans le but confirmer la spécificité des produits détectés, une partie de l'échantillon d'ARN ne sera pas traité par Rnt1p, mais subira toutes les autres étapes du traitement.

Deux techniques seront employées et comparées. La première consistera à enrichir l'échantillon en produits de coupure 5' en soumettant l'échantillon après coupure par Rnt1p purifié à un traitement par l'exoribonucléase Xrn1p qui devrait dégrader les tiges-boucles coupées et les produits de coupure 3'. Les ARN cibles seront ensuite identifiés grâce à des puces à ADN couvrant tout le génome. Cette technique devrait pouvoir identifier tous les types de substrats incluant ceux où une structure plus complexe que le substrat idéal est coupée.

La deuxième technique tentera d'identifier les substrats idéaux produisant une tige-boucle coupée d'une taille d'environ 34 nucléotides. Les produits de coupure seront séparés selon leur taille en utilisant le même procédé que pour l'enrichissement des microARN. Ensuite, l'ARN enrichi en petits ARN sera séquencé pour connaître les produits de coupure.

Ensemble ces techniques permettront d'obtenir un ensemble complet des ARN substrats de Rnt1p. En particulier, les substrats ARNm pourront être identifiés. De plus, puisque la réaction de coupure se fera en ARN total, le portrait obtenu sera plus représentatif qu'un essai fait sur des ARN synthétiques. Aussi, ces résultats permettront de connaître si certaines voies métaboliques contiennent plus de substrats potentiels et si Rnt1p a une action globale ou plus spécifique à certains processus cellulaires. Finalement, un nombre plus imposant de substrats permettra de mieux comprendre les éléments de structure et de séquence reconnus par l'enzyme lors de la liaison et de la coupure.

4.2 Méthodes

4.2.1 Identification de substrats de Rnt1p par puces à ADN

L'ARN total est extrait d'une culture de levure *rnt1*Δ. La moitié de cet ARN est soumise à une coupure *in vitro* par Rnt1p recombinant, alors que l'autre moitié est incubée de la même façon, mais sans l'ajout de Rnt1p. Ensuite, les deux échantillons sont soumis à une dégradation par Xrn1p (une exoribonucléase 5'->3') qui dégrade le produit de coupure 3' tout en laissant intact le produit de coupure 5' protégé par sa coiffe 5'. Une validation sur gel avec un substrat coupé (MIG2) et un transcrit non coupé (ACT1) a été'effectuée. Les deux échantillons sont finalement traités tels que décrit précédemment et hybrider à des puces à ADN couvrant tout le génome.

L'analyse des données de puces à ADN a été faite selon la méthode décrite dans David $\mathrm{et\ al.}$ (2006) à l'exception d'une étape de prétraitement qui corrige l'intensité d'hybridation des sondes en fonction du ΔG d'hybridation de la sonde. L'algorithme de segmentation a été appliqué sur la variation d'intensité entre l'échantillon traité et l'échantillon non traité. Le paramètre du nombre maximal de segments a été déterminé en utilisant la valeur optimale selon le critère d'information bayésien (BIC). Les régions avec des valeurs négatives correspondent aux régions où l'ARN a été dégradé par Xrn1p.

Les petits segments et les segments ayant une faible densité de sondes uniques ont été enlevés compte tenu de leur faible degré de fiabilité. La majorité des segments trouvés ont un niveau proche de zéro et les segments qui dévient de zéro ont majoritairement un niveau négatif comme attendu. La médiane et la déviation absolue médiane (MAD) ont été utilisées pour choisir une valeur seuil appropriée. Cette valeur est de -0.2425 ce qui correspond à une valeur p de 0.05 (la médiane moins 1.96 fois la MAD). Les segments voisins ayant un niveau se situant sous le seuil ont été fusionnés et les régions de moins de 125 nt ont été enlevées. Les 238 régions résultantes ont été assignées aux transcrits annotés.

Pour chaque segment, les variations des sondes situées à l'intérieur du segment et dans une région de 250 nt de chaque côté ont été adoucies en utilisant la moyenne mobile avec une fenêtre de 151 nt. Les valeurs obtenues pour chaque paire de segments ont été alignées en utilisant une version modifiée dans le cadre de ce travail de l'algorithme de Needleman–Wunsch (Needleman et Wunsch, 1970). Cet algorithme, plutôt que de maximiser la similarité d'acides aminés, minimise le carré de la différence entre les deux segments. Les positions non alignées sont pénalisées en utilisant le carré de la différence entre la valeur non alignée et la moyenne des deux valeurs adjacentes dans l'autre segment. Les positions terminales non alignées sont pénalisées en utilisant le carré de la différence maximale entre les deux segments.

La distance entre deux profils de dégradation (deux segments) est la somme du carré des différences entre toutes les positions alignées et 10 fois les nombres de positions non alignées divisée par le total des longueurs des segments non alignés. Cette mesure de distance est utilisée pour effectuer un regroupement hiérarchique des profils similaires en utilisant la méthode de Wards. Le choix a été fait de conserver dix groupes, car à ce point la prochaine séparation ne produit pas d'enrichissement significatif par rapport au groupe parent pour ce qui est de la longueur des profils, de l'intensité des profils ou du contenu en transcrits de différents types.

4.2.2 Identification de substrats de Rnt1p par séquençage à haut débit

L'ARN total est extrait d'une culture de levure *rnt1*Δ. La moitié de cet ARN est soumise à une coupure *in vitro* par Rnt1p recombinant, alors que l'autre moitié est incubée de la même façon, mais sans l'ajout de Rnt1p. Ensuite, les deux échantillons sont soumis à un enrichissement en court ARN à l'aide du protocole mirVana de Invitrogene. Puis les deux échantillons ont été préparés et séquencés sur un séquenceur Ion Torrent.

Les séquences des adaptateurs 5' ont été enlevées à l'aide de l'outil cutadapt (Martin, 2011) version 1.2rc2 et les séquences restantes ayant moins de 16 nt ont été enlevées des analyses subséquentes. Les séquences ont été alignées au génome de référence de *Saccharomyces cerevisiæ* révision R64-1-1 en utilisant Rsubread 1.1.1 (Liao $\mathrm{et\ al.}$, 2013).

Seuls les produits de séquençage de 32 à 38 nt ont été considérés pour l'analyse d'enrichissement. Les produits de séquençage présents en quatre copies ou plus dans l'échantillon traité ont été dénombrés dans les deux échantillons et un test binomial avec correction de Holm-Bonferroni a été appliqué avec une valeur seuil de 0.05. Les produits de séquençage enrichis ayant plus de 50 % de chevauchement ont été fusionnés et ensuite assignés aux transcrits annotés. Pour chaque groupe de produits de séquençage, la structure secondaire la plus stable a été prédite et une boucle centrale avec au moins 10 nt de chaque côté a été utilisée pour la classification.

4.2.3 Validation

Les nouveaux types de boucles ont été validés par traitement *in vitro*. Un transcrit synthétique marqué par un phosphate radioactif en 5' en a été incubé avec Rnt1p purifié. Les produits de réaction ont été séparés sur gel d'acrylamide. Des transcrits mutés ont aussi été testés pour valider la spécificité de la coupure et les éléments nécessaires à la reconnaissance par Rnt1p.

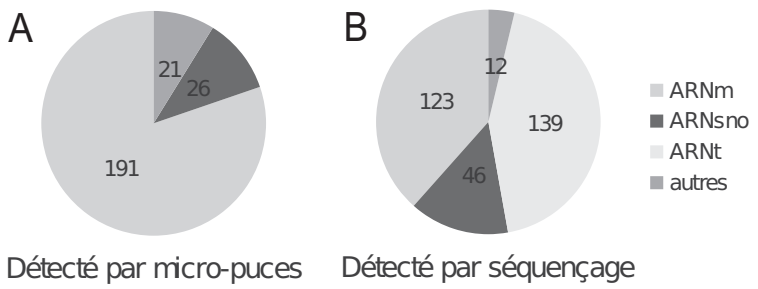


figure 4.1 – Nombre de transcrits coupés par Rnt1p dans chaque catégorie de transcrits.

A Résultats obtenus par puces à ADN. B Résultats obtenus par séquençage à haut débit des courts ARN.

4.3 Résultats

4.3.1 Rnt1p cibles des centaines de transcrits dans le transcriptome

Deux méthodes ont été employées pour détecter directement la coupure produite par Rnt1p dans une réaction *in vitro* sur l'ARN total extrait d'une souche ne contenant pas RNT1. Comme attendu, les deux méthodes identifient une grande proportion des ARNsno ciblés par Rnt1p. Les deux méthodes identifient plus d'une centaine de cibles dans des ARNm (Figure 4.1). Pour les ARNm, le chevauchement entre les deux méthodes est plutôt faible : seulement 30 ARNm cibles sont identifiés par les deux méthodes (Voir Figure 4.7).

L'approche de séquençage des petits fragments d'ARN a aussi détecté de nombreuses cibles potentielles dans des ARN de transfert (ARNt) lesquels ne sont pas connus comme étant ciblés par Rnt1p.

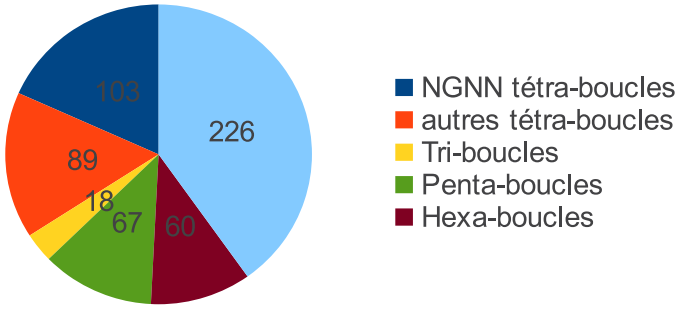


figure 4.2 – Nombre de régions enrichies selon le type de structure.

4.3.2 Rnt1p reconnaît plusieurs types de structure

À ce jour, toutes les structures reconnues par Rnt1p comportaient une tétraboucle NGNN. Par séquençage, plusieurs autres structures montrent un enrichissement significatif dans l'échantillon coupé par Rnt1p (Figure 4.2). Dix structures ont été testées parmi les structures non NGNN. La coupure par Rnt1p a été confirmée pour sept de ces structures. La coupure a été confirmée dans les types de structure suivants :

– triboucle ;
– tétraboucle non NGNN ;
– pentaboucle.

4.3.3 Nouveau modèle de substrats NGNN

L'ensemble des substrats connus de Rnt1p (51) et des nouveaux substrats ayant une tétraboucle NGNN identifiés par séquençage (75) a permis d'établir un nouveau modèle pour les substrats NGNN de Rnt1p (Figure 4.3). Ce modèle fournit plus d'information que le modèle précédent (Figure 2.1) qui a été construit en utilisant seulement les substrats publiés.

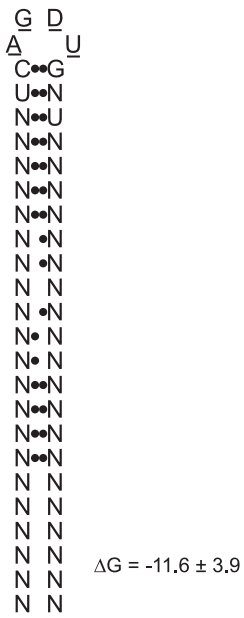


figure 4.3 – Structure consensus obtenue par recherche d'enrichissement parmi l'ensemble des structures ciblées par Rnt1p connues et les structures à tétraboucle NGNN identifiées par séquençage.

Tous les nucléotides indiqués montrent un enrichissement ($p < 0.05$). D indique significativement moins de cytosine qu'attendu ($p < 0.05$). Les points noirs indiquent que le nucléotide adjacent est plus souvent apparié qu'attendu, alors que les nucléotides soulignés sont moins souvent appariés qu'attendu ($p < 0.05$)

Il montre que les quatre nucléotides de la tétraboucle jouent un rôle dans la reconnaissance des substrats. Les cytosines en troisième position sont rares, alors que le quatrième nucléotide est souvent un uracile. Les trois premières paires de bases fermant la tétraboucle ont aussi certains nucléotides enrichis.

Ce modèle montre aussi que l'appariement des nucléotides fermant la tétraboucle est important jusqu'à la septième paire de bases. Aussi, il est maintenant clair que les nucléotides près du site de coupure habituel, à 14 et 16 nt de la tétraboucle, sont préférentiellement appariés. L'appariement semble moins important dans la région intermédiaire : les paires 8 à 13 à partir de la tétraboucle.

4.3.4 Différentes classes de substrats NGNN existent

Compte tenu du nombre de substrats maintenant identifiés (126), il est possible d'identifier des groupes de substrats comportant des différences soit dans la séquence, la structure ou la stabilité de la structure.

Les premiers groupes de substrats qui ont été comparés sont les substrats provenant des régions codantes et ceux provenant des régions non codantes. Pour ces deux groupes, une différence significative a été identifiée dans la composition de la troisième base de la tétraboucle. Les substrats provenant des régions codantes des ARNm ont dans plus de 40 % des cas une guanine comme troisième base de la tétraboucle, alors que les substrats provenant des régions non codantes ont un uracile à cette position dans environ la moitié des cas (Figure 4.4). Cette différence de composition n'est pas due à une composition différente des séquences codantes puisque le troisième nucléotide suivant les bases AG dans les séquences codantes a une composition en guanine très différente des substrats provenant des ARNm.

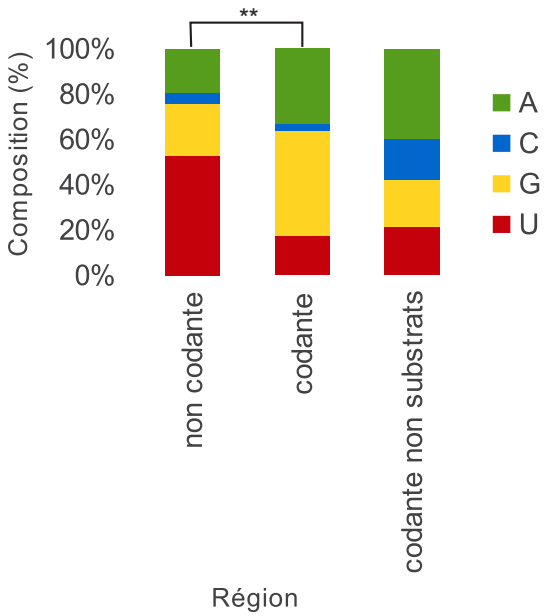


figure 4.4 – Composition de la troisième base de la tétraboucle.

La composition de la troisième base est comparée entre les substrats provenant de régions non codantes, les substrats provenant de régions codantes et tous les trinucléotides débutant par AG provenant des régions codantes du génome. Les deux étoiles indiquent la différence de composition est significative avec $p < 0.01$.

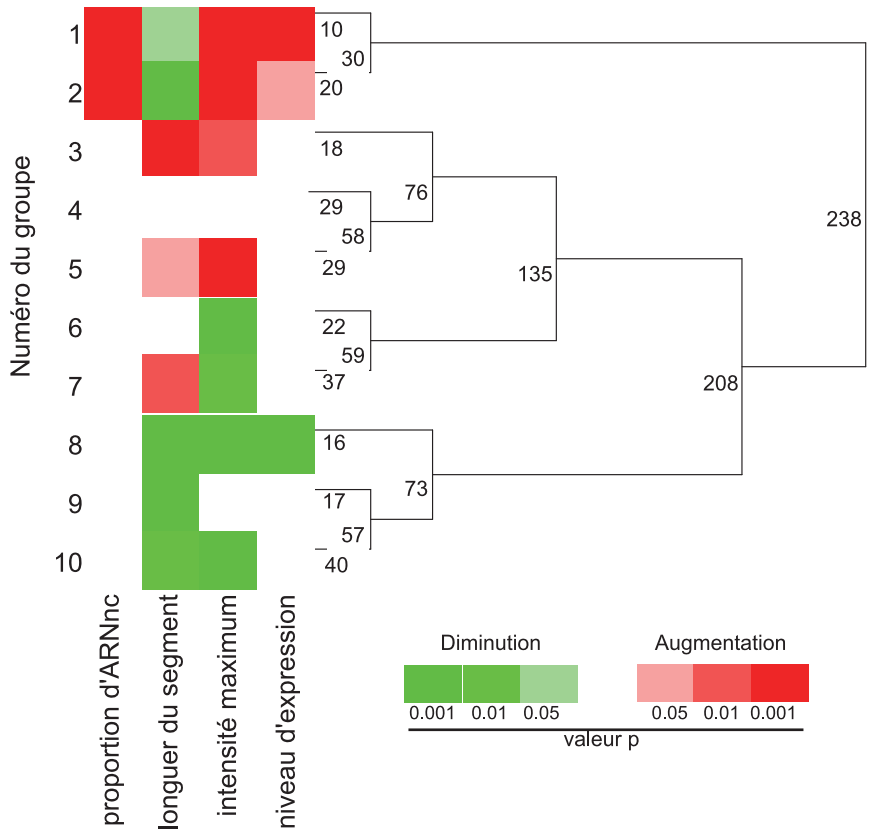


figure 4.5 – Regroupement hiérarchique des substrats de Rnt1p identifiés par puces à ADN.

À chaque noeud est indiqué le nombre de membres. Le graphique en couleur indique les attributs qui diffèrent entre les groupes.

Dans le but d'identifier d'autres groupes de substrats similaires, un algorithme a été conçu pour aligner les patrons de dégradation obtenus par puces à ADN et pour mesurer la similarité entre les patrons de dégradation. Cet algorithme a permis d'obtenir dix groupes de substrats par regroupement hiérarchique (Figure 4.5). Ce regroupement produit des groupes similaires pour différents attributs : la longueur du segment dégradé par Xrn1p, l'intensité maximale de la dégradation et le niveau d'expression du transcrit. Ainsi, certains groupes sont enrichis en ARNnc puisque leur niveau d'expression est élevé, que leur efficacité de coupure est élevée et qu'il s'agit de petits transcrits comparativement aux ARNm. Il aurait été souhaitable de faire un regroupement basé uniquement sur la coupure et non sur la longueur du gène ou son expression. Lorsque des corrections pour normaliser ces éléments sont effectuées, les artéfacts de la technique sont amplifiés et influent sur les groupes formés.

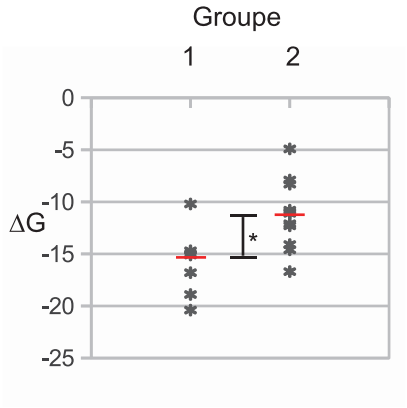


figure 4.6 – Les groupes 1 et 2 ont une stabilité de structure différente.

Les étoiles désignent la stabilité (ΔG) de chacun des substrats connus et uniques dans les deux groupes. Les lignes rouges indiquent les valeurs des médianes pour chacun des deux groupes. Les lignes noires indiquent la différence significative de stabilité entre les deux groupes ($p < 0.05$).

La plupart des groupes comportent peu de membres dont la région reconnue par Rnt1p est connue. Cependant, les groupes 1 et 2 contiennent respectivement 7 et 11 membres ayant une seule région identifiée comme étant reconnue par Rnt1p. Pour ces deux groupes, il a été possible d'identifier une différence significative dans la stabilité de la structure (ΔG). Les membres du groupe 1 étant plus stables que ceux du groupe 2. Ces deux groupes sont

composés principalement d'ARNsno. La principale différence qui a pu être observée entre les deux groupes à l'exception de la stabilité de la boucle est que les ARNsno du groupe 1 ont en général un niveau d'expression un peu plus élevé que ceux du groupe 2. Leurs substrats plus stables pourraient ainsi favoriser la coupure des transcrits plus abondants.

4.4 Discussion

4.4.1 Particularités de chacune des méthodes

Chaque méthode d'identification *in vitro* des cibles de Rnt1p a des spécificités différentes. D'abord, les méthodes d'analyse standards des puces à ADN ne permettent pas d'obtenir des résultats fiables pour les régions présentes en plusieurs copies dans le génome. Les résultats obtenus par puces à ADN dépendent aussi de la taille du produit de coupure 3'. Même dans le cas d'une coupure peu efficace, un long produit 3' permettra la détection. Par opposition, certaines coupures intenses n'ont pas pu être détectées dû à un produit 3' trop court. De plus, les puces à ADN ne permettent pas d'identifier avec précision la position de la coupure. Il semble y avoir une dégradation 3'-5', variable selon les gènes, qui masque la position précise de la coupure. La source de cette dégradation est inconnue. La technique par séquençage est quant à elle sensible à la création d'artéfacts par les ARN fortement exprimés. La validation de la coupure des substrats ARNt n'a pas été possible. Il est donc probable que la technique décèle des fragments résultant d'une dégradation partielle. La quantité de fragments produits est fonction du niveau d'expression initiale. Il est possible que certains autres gènes ne soient pas réellement coupés par Rnt1p et aient été identifiés dû à des artéfacts semblables.

Les résultats obtenus par séquençage haut débit sont en fonction du nombre de séquences obtenues. Pour la technologie utilisée ici, environ 5 millions de séquences ont été obtenues et environ 3 millions ont pu être alignées sur le génome. Une part importante de ces sé-

quences se situe dans les ARN les plus abondants : les ARNr et les ARNt. Ainsi, le nombre de séquences obtenues dans le reste du transcriptome est une faible proportion de l'ensemble des séquences. Pour notre étude, la sensibilité de détection dans les ARNm a été plus faible pour les résultats de séquençage que pour les résultats de puces à ADN. De plus, grâce à la dégradation par Xrn1p, l'approche par puces à ADN permet la détection des coupures moins efficaces. Ainsi, les ARNm identifiés par les deux méthodes sont des coupures intenses dans des ARNm ayant un bon niveau d'expression.

La capacité de la technique utilisant le séquençage à détecter les substrats peut être mise en perspective. Parmi tous les substrats identifiés par puces à ADN, seule une faible proportion a aussi été identifiée par séquençage (Voir Figure 4.7). Près de la moitié des substrats identifiés par les deux méthodes sont des substrats ARNnc ayant généralement un niveau d'expression et de coupure élevé. Un nombre plus important de séquences pourrait permettre d'identifier plus de substrats, mais la technique souffre aussi des certaines limitations. D'abord, seules les coupures produisant de petits produits peuvent être détectées. Aussi, pour obtenir un nombre suffisant de séquence, la coupure doit être efficace. De plus, la coupure doit se faire de chaque côté de la structure. Il existe des substrats de Rnt1p qui produisent des produits de coupure plus grands que 38 nt et il est aussi possible que Rnt1p coupe d'un seul côté de la tige boucle. Finalement, la coupure doit se faire le plus possible aux mêmes positions dans la séquence, car la variabilité diminue la sensibilité de la détection. Toutes ces particularités dues à la technique font qu'il a été possible d'identifier un grand nombre de substrats, mais il est possible qu'il en existe beaucoup d'autres. Aussi, les substrats identifiés sont possiblement biaisés pour certains types de coupure par exemple les coupures les plus efficaces.

Le principal avantage de la méthode par séquençage est l'identification directe des sites de coupure ce qui permet de mieux définir les régions reconnues par Rnt1p. Cependant, cette méthode ne permet pas de détecter tous les types de coupure. Parmi les cibles détectées par puces à ADN, mais non détectées par séquençage, plusieurs montrent une dégradation non spécifique dans les données de séquençage. Il est possible que Rnt1p puisse dégra-

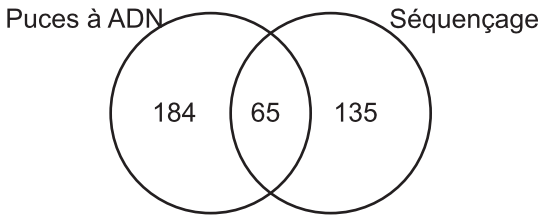


figure 4.7 – Résumé du nombre de substrats détectés par chaque méthode.

Substrats détectés par chacune des deux méthodes excluant les ARNt et autres ARN répétitifs.

der de façon non spécifique certaines cibles, mais il est aussi envisageable que l'enzyme recombinante comporte une activité secondaire.

L'utilisation de ces deux méthodes a permis pour la première fois d'obtenir un catalogue fiable des cibles de Rnt1p dans la cellule. Cela permettra d'étudier de façon ciblée les voies métaboliques ciblées par Rnt1p et de mieux comprendre le rôle de Rnt1p dans la régulation de l'expression. Maintenant, il est aussi confirmé que le nombre de substrats de Rnt1p dans le génome est de l'ordre de quelques centaines. Ainsi, contrairement aux estimations obtenues par recherche bio-informatique de substrats semblables, les coupures par Rnt1p sont relativement rares et l'enzyme est plus spécifique que ce à quoi on pourrait s'attendre en observant la diversité des séquences et structures des substrats connus. De plus, pour plusieurs transcrits, la région reconnue par Rnt1p est maintenant connue et peut être utilisée pour mieux comprendre comment Rnt1p reconnaît ses substrats.

Il est aussi possible d'envisager l'utilisation d'autres méthodes pour obtenir un ensemble plus complet des substrats de Rnt1p. Certaines techniques de séquençage du dégradome permettent d'identifier directement les extrémités 5'. Il serait donc possible d'identifier les sites de coupure dans les cas où le produit de coupure est long ou dans le cas où il n'y a qu'une seule coupure.

4.4.2 Substrats de Rnt1p

L'emploi du séquençage de petits ARN a permis d'identifier un total de 563 régions où se trouvent de séquences de 32 à 38 nt enrichies dans l'échantillon traité par Rnt1p. Pour une centaine de ces régions, la structure prédite comme étant la plus stable correspond à une tige coiffée d'une tétraboucle NGNN comme la majorité des substrats publiés. Ces régions ont été considérées comme étant constituées majoritairement de réelles cibles de Rnt1p. À partir de ces régions, il a été possible de raffiner les modèles de substrats de Rnt1p et d'identifier plusieurs nouveaux éléments importants pour la reconnaissance des substrats.

Malgré le plus grand nombre de substrats NGNN identifiés, il n'a pas été possible d'identifier des relations de second ordre entre les éléments de structure et de séquence. Cependant, il est quand même possible d'utiliser le nouvel ensemble de substrats pour faire des hypothèses simples à valider expérimentalement. Tel que décrit précédemment dans Lamontagne et Elela (2004), trois régions semblent impliquées dans la reconnaissance des substrats : la boucle, le haut de la tige et le site de coupure. Le substrat idéal aurait une boucle AGUU fermée par les bases C-G, U-A et A-U. Il est aussi possible de noter que l'appariement n'est pas symétrique de chaque côté de la boucle(voir Figure 4.3). Il est possible que le côté 3' doive être plus rigide, alors que le côté 5' peut contenir plus de nucléotides non appariés.

Il a aussi été possible d'observer des différences importantes entre des groupes de substrats. D'abord, une différence entre les substrats provenant des régions codantes et ceux provenant des régions non codantes a été observée (Voir Figure 4.8 et 4.9). Cette différence n'est pas due à la composition particulière des régions codantes, donc il semble que Rnt1p reconnaît différemment les substrats situés à l'intérieur des séquences codantes. Pour les ARNnc, le rôle de Rnt1p est de participer à la maturation de l'ARN en enlevant des régions ne faisant pas partie de l'ARN mature. Pour les ARNm, Rnt1p serait plutôt impliqué dans la régulation du niveau d'expression. Il est donc possible que les substrats situés à l'intérieur des ARNm doivent être moins efficaces pour prévenir la dégradation complète des transcrits. Une autre possibilité est que lorsque Rnt1p cible un ARNm, il le fasse à l'inté-

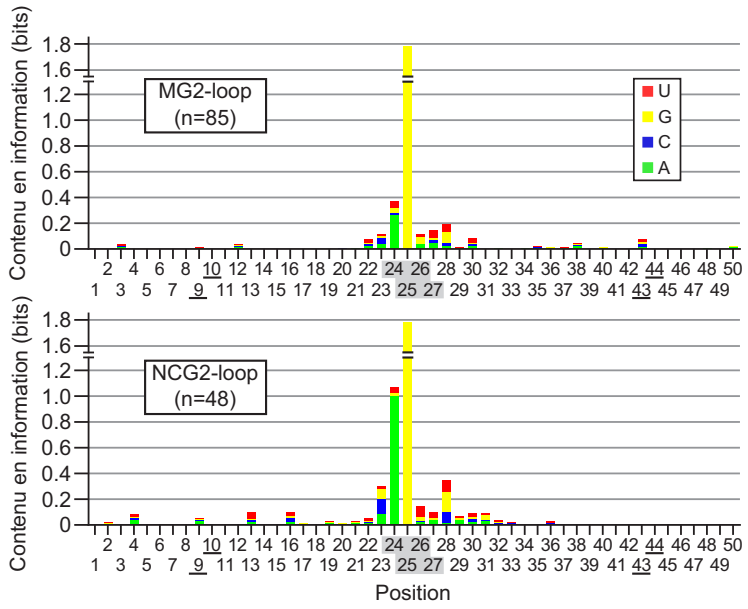


figure 4.8 – Contenu en information selon la position.

Le contenu en information pour chacune des positions a été calculé (hauteur de la barre) et la proportion de chaque nucléotide dans les boucles NGNN est représentée. MG2 indique les substrats ARN messagers, alors que NCG2 indique les substrats ARN non codant.

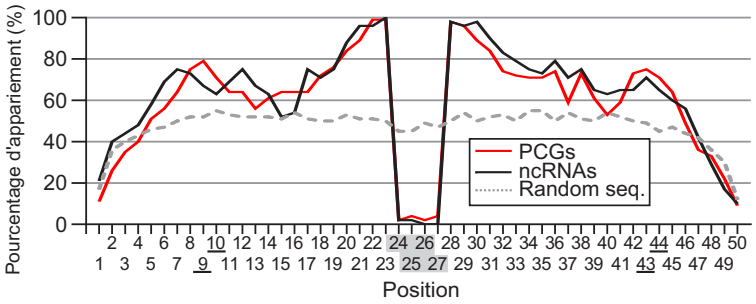


figure 4.9 – Pourcentage d'appariement en fonction de la position.

Le pourcentage d'appariement pour chaque position a été calculé pour les substrats ARNm (PCGs), pour les substrats ARNnc (ncRNAs) et pour des séquences de même composition générées aléatoirement (Random seq).

rieur d'un complexe impliquant d'autres protéines et que ce complexe modifie légèrement la reconnaissance par l'enzyme.

Une autre différence a été observée entre deux groupes de substrats ayant des coupures très fortes. Il s'agit d'une différence au niveau de la stabilité de la structure. Une autre différence entre ces deux groupes est le niveau d'expression du transcrit. Il est possible que les transcrits les plus exprimés possèdent des structures plus stables de façon à maximiser l'efficacité de la coupure. Il est aussi possible que pour les transcrits ayant une expression plus faible, la maturation ne soit pas le seul rôle de Rnt1p. Ainsi, les modifications guidées par les ARNsno facilitent la croissance dans certaines conditions. Donc, il est possible que les niveaux de ces ARNsno doivent être modulés selon les conditions de croissance et Rnt1p pourrait jouer un rôle dans cette régulation grâce à des cibles imparfaites.

Il a déjà été montré que la composition de la tige peut être variée pour produire différentes efficacités de coupure (Babiskin et Smolke, 2011). Les techniques utilisées ne permettent pas d'avoir une mesure précise de l'efficacité de la coupure, car le niveau initial de l'ARN n'est pas connu précisément et certains biais font que certains produits de coupure sont détectés plus facilement. Cependant, il serait intéressant d'obtenir cette mesure pour certains substrats dans le but de développer une prédiction d'efficacité en fonction de la séquence et de la structure.

La principale nouveauté qui a été révélée par ce travail est la variété des structures coupées par Rnt1p. À l'exception d'un substrat ayant une tétraboucle AAGU, tous les substrats publiés ont une tétraboucle NGNN. Ici, non pas un, mais six nouveaux types de substrats ont été identifiés (Voir Figure 4.11). Certains ont des tétraboucles AUGU ou AUUU qui pourraient être reconnues de façon similaire à la tétraboucle AAGU. D'autres ont des boucles de trois ou cinq nucléotides ce qui implique un mode de reconnaissance différent. De plus, seulement dix candidats ont été testés parmi environ 450 candidats potentiels et le taux de succès a été de 60 %. Donc, il existe potentiellement plusieurs autres types de substrats. Bien que la majorité des substrats identifiés n'aient pas de boucle NGNN, il est probable qu'un grand nombre de ces substrats soient dus à des artéfacts. Il y a d'abord les arté-

facts provenant d'ARN abondants dégradés comme ceux provenant des ARNt. Il y a aussi plusieurs exemples de substrats dont la structure prédite comme optimale n'est pas une boucle NGNN, mais qui peuvent adopter cette structure. Lorsque les ARN abondants et répétés comme les ARNt sont enlevés de l'analyse, la distribution des types de boucles change (Voir Figure 4.10).

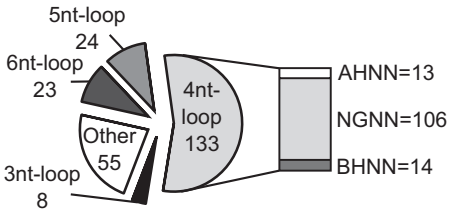


figure 4.10 – Nombre de régions enrichies selon le type de structure.

Distribution des types de structures identifiées excluant celles provenant des ARN abondants et répétés.

4.5 Contributions

L'auteur a participé au design expérimental, à l'analyse des données de puces à ADN et de séquençage, à la conception des sondes pour la validation, à l'adaptation des algorithmes et à l'analyse des résultats. Mathieu Lavoie a participé au design expérimental, a réalisé les extractions d'ARN et les réactions de coupure *in vitro* et a effectué la validation *in vitro*. La synthèse, le marquage de l'ADNc et l'hybridation aux puces à ADN a été effectué comme service par le Centre for Applied Genomics at University of Toronto. Francis Malenfant a effectué la validation *in vitro*. Mathieu Catala a effectué la validation *in vivo*. Sherif Abou Elela a participé au design expérimental et à l'analyse des résultats.

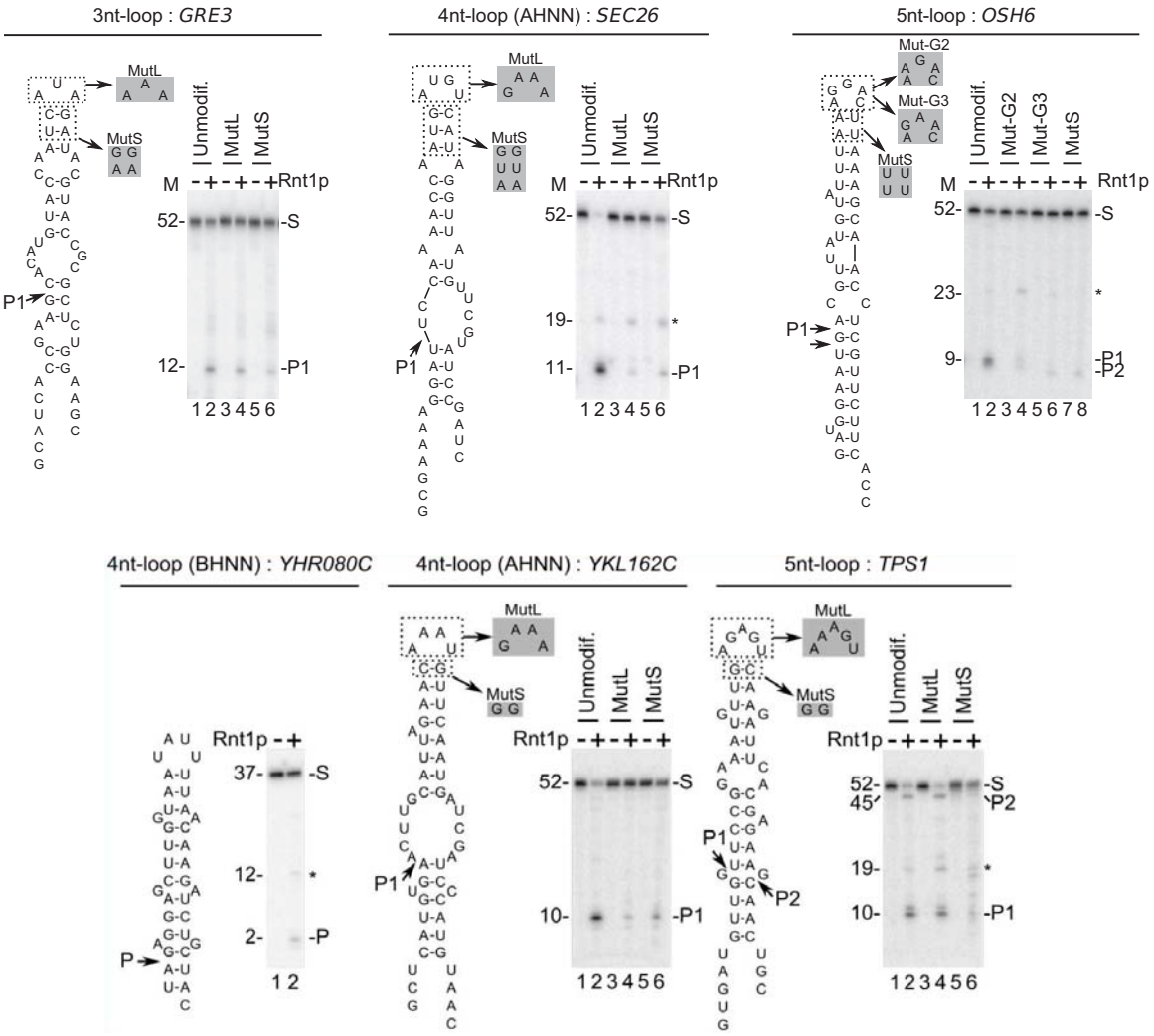


figure 4.11 – Six structures atypiques validées.

Parmi les structures atypiques identifiées, certaines ont été validées. Des transcrits synthétiques marqués en 5' ont été coupés par Rnt1p purifié *in vitro*. S indique l'ARN non coupé et le P le produit de la coupure. Les flèches indiquent les sites de coupure.

4.6 Résumé de l'impact

Le travail de ce chapitre a mené à l'écriture d'un article soumis pour publication (Gagnon et al., soumis à PLoS Genetics).

CHAPITRE 5

DISCUSSION ET CONCLUSION GÉNÉRALE

5.1 Discussion

5.1.1 Limites de techniques utilisées

Trois approches différentes et complémentaires ont été utilisées pour identifier les substrats de l'endoribonucléase Rnt1p. L'approche de prédiction *in silico* a utilisé les substrats déjà connus pour identifier des structures similaires pouvant être reconnues par l'enzyme. L'approche *in vivo* a détecté les transcrits qui s'accumulent lorsque l'enzyme est inactive. L'approche *in vitro* a identifié les transcrits coupés par l'enzyme purifié. Chaque méthode a permis d'en apprendre plus sur l'enzyme, mais prises séparément, elles ne fournissent qu'une vue partielle de la spécificité de l'enzyme.

Approche *in silico*

L'approche *in silico* a produit une liste de plusieurs milliers de substrats potentiels. La majorité de ces candidats ne semblent pas être des cibles réelles de Rnt1p dans la cellule. Pourtant, ils ne doivent pas tous être considérés comme des faux positifs pour autant. Lorsqu'un substrat synthétique court est produit à partir de ces prédictions, la majorité du temps il peut être coupé par l'enzyme dans un essai de coupure *in vitro*. Il semble que la prédiction en soi n'est pas mauvaise, mais il y aurait certaines caractéristiques qui seraient manquantes pour faire une prédiction plus juste. Compte tenu du faible contenu en information qui a pu

être détecté dans la séquence de la tige-boucle, il semble que ces caractéristiques soient situées à l'extérieur de la séquence évaluée.

Une de ces caractéristiques pourrait être l'accessibilité de la structure par l'enzyme. Il est probable que des structures compactes puissent empêcher l'enzyme de se lier à un substrat potentiel. Pour inclure cette caractéristique à la prédiction de substrats *in silico*, il serait important d'avoir une prédiction fiable de la structure du transcrit complet. Actuellement, ce type de prédiction est peu fiable. De plus, il est difficile de connaître la flexibilité de cette structure.

Il semble aussi que cet effet du contexte de l'ARN puisse provenir d'une relation entre la tige-boucle et le reste du transcrit. Il a été montré que différentes tiges-boucles placées à la même position dans un transcrit artificiel montrent des différences importantes dans l'efficacité de coupure (Meaux $\mathrm{et\ al.}$, 2011). Ces différences proviennent en partie des différences d'efficacité de coupure inhérentes à la tige-boucle elle-même, mais ces différences ne suffisent pas à expliquer toutes les variations observées. Il faut envisager la possibilité que Rnt1p reconnaisse des éléments à l'extérieur de la tige-boucle. Cela est concevable puisqu'il est proposé que Rnt1p agisse sous forme de dimère. Sous cette forme, un deuxième site de liaison à l'ARN serait libre dans le complexe enzyme-substrat et pourrait être utilisé pour stabiliser le complexe dans certaines conditions.

En somme, bien que des ensembles d'entraînement beaucoup plus complets et représentatifs soient maintenant disponibles, le problème de l'approche *in silico* se situé plutôt au niveau des variables d'apprentissage que des données d'entraînement et de l'algorithme. Une caractérisation biochimique plus poussée utilisant des substrats synthétiques longs permettrait de mieux orienter le choix des variables à considérer dans l'apprentissage.

Approche *in vivo*

Bien que l'analyse du transcriptome ait permis d'identifier quelques nouveaux substrats, elle a surtout mis en évidence l'effet profond que l'absence de Rnt1p a sur le cycle de vie de l'ARN. Ainsi, en plus des rôles de maturation et de régulation de l'ARN, Rnt1p influence globalement l'expression des gènes.

Rnt1p affecte plus spécifiquement certaines voies métaboliques comme l'absorption du fer (Lee $\mathrm{et\ al.}$, 2005), le métabolisme du glucose (Ge $\mathrm{et\ al.}$, 2005; Lavoie $\mathrm{et\ al.}$, 2012), la stabilité des parois cellulaires (Catala $\mathrm{et\ al.}$, 2012) et la respiration (Gagnon et al., soumis à PLoS Genetics). Plusieurs ARNm faisant partie de ces voies sont affectés par l'absence de Rnt1p, mais la plupart ne sont pas des substrats directs de Rnt1p. En effet, la dérégulation d'un seul gène peut entrainer un effet domino sur plusieurs dizaines de gènes. Il devient alors difficile de différencier les substrats directs des cibles indirectes.

En plus de ces effets indirects, un effet global sur la transcription a été observé. Les régions intergéniques habituellement non exprimées et les gènes faiblement exprimés sont surexprimés en l'absence de Rnt1p, alors que les gènes fortement exprimés subissent une diminution de leur niveau d'expression. Il est probable que les défauts et le retard dans l'assemblage des ribosomes causés par l'absence de Rnt1p puisse avoir une telle influence. Cependant, il a aussi été observé que Rnt1p est associé à la polymérase lors de la transcription (Ghazal $\mathrm{et\ al.}$, 2009). La quantité et la localisation de Rnt1p ne permettent certainement pas à Rnt1p d'être associé à la polymérase en permanence, mais il est tout de même possible que cette association influe sur la transcription globale.

Il a aussi été montré que l'effet de Rnt1p sur le cycle de vie de l'ARN n'est pas limité à la coupure des transcrits. Rnt1p peut aussi influencer le niveau d'expression d'un ARNm sans qu'il y ait coupure (Lavoie $\mathrm{et\ al.}$, 2012). Cette régulation est dépendante du promoteur et de la présence de Rnt1p, mais indépendante de la présence d'un site de coupure ou non. Il pourrait s'agir d'un autre effet cotranscriptionnel de Rnt1p.

Pour pouvoir utiliser une approche de détection *in vivo*, il serait nécessaire de différencier les effets directs, les effets cotranscriptionels et les effets indirects. Pour cela, il serait utile d'avoir accès à des données d'immunoprécipitation de la chromatine à haute résolution pour mesurer la présence de Rnt1p lors de la transcription. Il serait aussi utile de mesurer l'effet sur le transcriptome d'un mutant de Rnt1p ne pouvant pas couper l'ARN, mais pouvant remplir ses autres fonctions. Un portrait de l'effet de l'absence de Rnt1p sur le transcriptome un court instant après son inactivation permettrait aussi de mieux différencier les effets directs et indirects.

Approche *in vitro*

L'approche de coupure *in vitro* a permis d'identifier un grand nombre de nouveaux substrats de Rnt1p. La principale limite théorique de cette approche est l'importance *in vivo* des substrats identifiés. Cette approche ne tient pas compte de la structure *in vivo* de l'ARN, de l'accessibilité du substrat pour Rnt1p et de la régulation possible par des cofacteurs. Les substrats identifiés par cette approche sont des substrats potentiels, mais l'action réelle de Rnt1p sur le substrat dans la cellule n'est pas démontrée. Pourtant les transcrits faisant partie de certaines voies métaboliques semblent comporter plus de coupure que d'autres voies métaboliques. Ainsi, les coupures qui ont été détectées ne sont pas distribuées au hasard et représentent, du moins en partie, des cibles réelles de Rnt1p.

Pour éclaircir ce point, il serait intéressant de lier de façon stable Rnt1p à ses substrats alors que la cellule est intacte. Cela permettrait d'identifier plus directement les substrats *in vivo* de Rnt1p. Cependant, Rnt1p a la capacité de lier certains ARN sans qu'il y ait coupure et en plus, la coupure peut être un processus très rapide. Il donc possible qu'une approche de ce type ne détecte que des substrats peu efficaces et peu représentatifs.

Les résultats obtenus par cette approche ont aussi été limités par les techniques utilisées. Bien que sensible, la détection par puces à ADN ne permet pas de situer avec précision le site de coupure. La technique de séquençage haut débit qui a été utilisée ne permettait

pas d'identifier les substrats ayant un site de coupure atypique. De plus, le nombre de séquences obtenu a limité la détection de certains substrats faiblement exprimés ou faiblement coupés.

D'autres techniques pourraient être envisagées pour pallier ces limitations. Rnt1p nécessite des ions magnésium pour effectuer la coupure, mais en absence de magnésium il peut tout de même lier l'ARN. Ainsi, si dans une première étape les transcrits non liés sont retirés de l'échantillon, la détection des substrats pourrait être rendue plus sensible.

Une autre technique consisterait à détecter les extrémités 5'-phosphate produites par la coupure de Rnt1p. Cette technique de détection des extrémités des ARN consiste à lier ARN synthétique à toutes les extrémités 5'-phosphate et à utiliser cet ARN synthétique comme guide pour l'isolation de l'extrémité de l'ARN pour être ensuite séquencé (German $\mathrm{et\ al.}$, 2008). Cette technique pourrait détecter des coupures atypiques et serait plus sensible que le séquençage des tiges-boucles coupées.

Combinaison des approches

L'utilisation de plusieurs approches différentes permet d'augmenter le niveau de confiance en chacun des résultats individuels. Ainsi, lorsque les tige-boucles identifiées par séquençage sont évaluées par l'algorithme de classification par similitude avec les substrats connus, 75 % d'entre elles ont une évaluation supérieure au seuil de 0.85. Cela confirme que l'approche par séquençage fonctionne et en plus, permet d'avoir un haut niveau de confiance que ces tige-boucles sont des substrats réels et non des artéfacts. Il est de même pour les cibles identifiés à la fois par séquençage et par détection des produits de coupure sur puces à ADN.

L'utilisation de plusieurs approches permet aussi de réévaluer certains résultats. il existe en effet des exemples où un faible niveau de coupure rend la détection de la coupure sur

puces à ADN ambiguë, mais combiné avec la détection de quelques séquences sur le même transcrit par séquençage, cela permet d'identifier de nouveaux substrats.

5.1.2 Caractéristiques des substrats de Rnt1p

L'identification d'un grand nombre de substrats a permis d'avoir une meilleure idée de la spécificité de Rnt1p pour l'ARN double brin. Jusqu'ici, Rnt1p était considéré comme une RNase III très spécifique qui ne reconnaissait que des tiges stables d'ARN coiffés d'une tétraboucle NGNN et ayant un site de liaison particulier près de cette boucle. On sait maintenant que Rnt1p peut reconnaitre des boucles de différentes tailles et séquences. Rnt1p est tout de même très spécifique comparativement à la RNase III bactérienne. Il ne coupe qu'un sous-ensemble restreint de structure d'ARN double brin, mais ce sous-ensemble est plus vaste qu'attendu.

Récemment, la structure 3D de Rnt1p lié avec un substrat typique a été déterminée par Liang $\mathrm{et\ al.}$ (2014). Cette structure est compatible avec les éléments observés comme étant enrichis parmi les structures coupées par Rnt1p. Le domaine N-terminal interagit avec les deux premiers nucléotides de la boucle. Le domaine de liaison à l'ARN 0 interagit avec les deux derniers nucléotides de la boucle. Le domaine de liaison à l'ARN 1 interagit avec les nucléotides situés après la boucle et le domaine de liaison à l'ARN 3 interagit avec les nucléotides situés près du site de coupure. Toutes ces régions ont montré des enrichissements significatifs pour certaines caractéristiques de séquence et de structure.

La découverte de substrats ne comportant pas une tétraboucle est pour le moins surprenante. Jusqu'ici tous les substrats de Rnt1p possédaient une tétraboucle. Des études de mutagenèse ont montré la boucle est tout de même reconnue spécifiquement par l'enzyme et qu'elle peut guider la coupure lorsque placée sur une tige différente. De plus, lorsque certains nucléotides sont mutés pour d'autres, la coupure est grandement réduite. Cela porte

à croire que lorsque ces substrats atypiques sont liés par l'enzyme, ils peuvent adopter une structure similaire aux tétraboucles et sont correctement reconnus par l'enzyme.

Le faible nombre de substrats possédant des triboucles ou des pentaboucles n'a pas permis d'obtenir un modèle clair des caractéristiques nécessaires à leur reconnaissance par l'enzyme. Il serait important de tenter d'identifier plus de ces types de substrats et leurs homologues. De plus, une structure des ces substrats liés par l'enzyme permettrait de comprendre la liaison des substrats avec boucles atypiques.

5.1.3 Fonctions des substrats de Rnt1p

Il est maintenant clair que la majorité des substrats de Rnt1p sont des ARN codants pour des protéines et non des ARN non codants. Ainsi, Rnt1p a la capacité de couper environ 5 % des ARNm. Il semble donc que chez la levure, il joue le rôle que Dicer et Argonaute jouent chez d'autres eucaryotes. Il existe cependant une différence importante. Rnt1p reconnaît directement ses cibles, alors qu'Argonaute utilise un microARN produit par Dicer comme guide. Rnt1p a la possibilité d'utiliser un ARN guide (Lamontagne et Elela, 2007) pour cibler un transcrit, mais à ce jour, aucun ARN guide de ce type n'a pu être identifié *in vivo* chez la levure.

Les substrats des ARNnc et des ARNm ne sont pas équivalents. Plusieurs substrats atypiques ont été identifiés à l'intérieur des séquences codantes des ARNm. Pour les substrats typiques, une différence de composition en nucléotides a aussi été observée entre les substrats des ARNm et des ARNnc. De plus, certains ARNm comptent plusieurs sites de coupure par Rnt1p. Ces différences contribuent à renforcer l'idée d'une fonction différente pour les substrats des ARNnc et ceux des ARNm. Les substrats des ARNnc seraient des substrats plus optimaux du point de vue de la cinétique enzymatique, alors que les substrats des ARNm seraient moins efficaces de façon a permettre une régulation plutôt qu'une dégradation systématique.

La régulation du niveau d'expression des ARNm par l'intermédiaire de substrats a été réalisée de façon artificielle (Babiskin et Smolke, 2011). Dans cette étude, 16 substrats artificiels ont été identifiés pour couvrir une vaste échelle d'efficacité de coupure. Le tout a été fait en modifiant uniquement la boîte de stabilité de la liaison, c'est-à-dire les nucléotides de la tige à proximité de la boucle. Plus de données de ce type pourraient permettre d'estimer l'efficacité de coupure à partir de la séquence de la tige-boucle. Pour ce faire, d'autres études portant sur les autres régions interagissant avec l'enzyme seraient nécessaires.

Rnt1p n'affecte pas de façon égale toutes les voies métaboliques. Il serait intéressant de vérifier si les voies ciblées chez la levure *Saccharomyces cerevisiæ* sont les mêmes chez d'autres organismes ou s'il s'agit d'une adaptation spécifique à cette espèce. On pourrait définir quatre grandes catégories de fonction pour ces ARNm coupés par Rnt1p : le métabolisme du glucose, le bourgeonnement, la transcription et les protéines liant l'adénosine triphosphate (ATP). Ces catégories ne sont certainement pas indépendantes et on peut remarquer certains chevauchements.

Le point commun entre ces catégories semble être la gestion de l'énergie. En effet, le glucose constitue la principale source d'énergie de la cellule. La transcription influence le niveau de métabolisme de la cellule. Le bourgeonnement et la division cellulaire constituent la principale dépense d'énergie de la cellule. Les protéines liant l'adénosine triphosphate (ATP) sont des enzymes qui utilisent activement de l'énergie sous forme d'ATP.

Ainsi, Rnt1p pourrait être un régulateur global du métabolisme. Il agit au niveau de la transcription, de la synthèse des ribosomes, de la reproduction et des activités catalytiques utilisant de l'énergie.

Ce rôle de Rnt1p dans le métabolisme de l'énergie est supporté par plusieurs observations biologiques. D'abord, il a été observé qu'avec le temps les souches où RNT1 est inactivé perdent leurs mitochondries lorsqu'elles croissent dans un milieu riche. Au microscope, des défauts morphologiques des mitochondries sont observables (Voir Figure 5.1). De plus, lorsque les levures croissent en milieu riche, elles oscillent entre la production d'énergie

par fermentation et par respiration. Ces oscillations peuvent être mesurées par la détection de la fluorescence du NAD/NADH un marqueur important de l'activité mitochondrial. Après synchronisation des cellules, les cellules où RNT1 est inactif n'ont que de faibles oscillations et deviennent rapidement désynchronisées (Voir Figure 5.2). Cela indique que les métabolisme énergétique de la levure est perturbé par l'absence de Rnt1p.

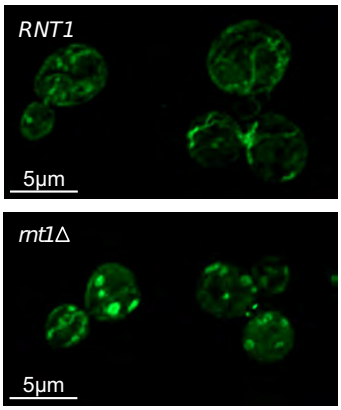


figure 5.1 – Morphologie des mitochondries.

Les mitochondries des souches de type sauvage et *rnt1*Δ ont été colorées avec le marqueur MitoTracker green FM.

5.1.4 Rôles de Rnt1p

On possède maintenant un portrait beaucoup plus représentatif des rôles de Rnt1p dans la cellule. Ce rôle est beaucoup plus vaste que le simple rôle de maturation des ARN non codants qui lui avait d'abord été attribué.

Sa localisation cellulaire dans le nucléole montre bien que son rôle principal est la maturation de l'extrémité 3' de l'ARNr 25S. Cette coupure est une étape importante de la maturation des ARNr et de l'assemblage des ribosomes (Elela $\mathrm{et\ al.}$, 1996). Cette tâche constitue la principale activité catalytique de Rnt1p.

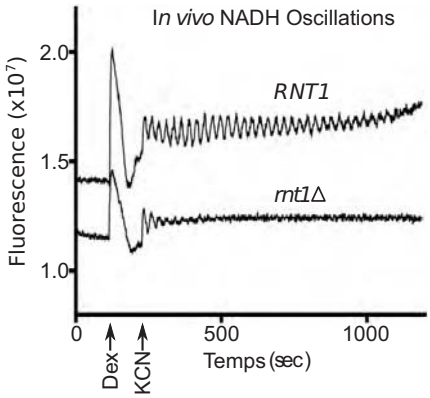


figure 5.2 – Oscillations métaboliques.

Des levures placés en incubateur en présence de dextrose sont synchronisées par l'injection de KCN. La fluorescence du NADH est mesurée pour observer les oscillations métaboliques.

Il semble que Rnt1p joue aussi un rôle important dans la régulation de la transcription. Sa présence sur la chromatine suit un patron similaire à celui de la polymérase II (Ghazal $\mathrm{et\ al.}$, 2009). Ce patron similaire indique que Rnt1p s'associe au complexe transcriptionnel. Sa présence pourrait influencer le niveau de transcription particulièrement celui des transcrits faiblement exprimés. Puisque les régions intergéniques montrent un niveau d'expression supérieur lorsque Rnt1p est absent, il est possible qu'il influence la spécificité du complexe transcriptionnel. Il n'existe cependant aucune information sur son mode d'action. Il n'est pas clair si la fonction catalytique de Rnt1p est importante pour cette fonction ni si Rnt1p agit directement ou via un ou plusieurs intermédiaires.

Le fait d'être lié au complexe de transcription permet aussi à Rnt1p d'effectuer la terminaison de la transcription par sa coupure (Ghazal $\mathrm{et\ al.}$, 2009). Ce nouveau mode de terminaison semble être utilisé pour des gènes fortement exprimés où un défaut de la terminaison standard provoquerait des effets néfastes aux gènes en aval. Il est aussi possible que ce mode de terminaison puisse servir à la régulation du cycle de vie du transcrit. En effet, les gènes dont la transcription est terminée par la coupure de Rnt1p ont une région non traduite plus longue et une queue polyA assemblée par un mécanisme différent. Le tout peut influer sur l'exportation au cytoplasme et la durée de vie du transcrit.

Rnt1p régule aussi de nombreux transcrits d'ARN messagers par coupure à l'intérieur de la région codante. Ainsi, Rnt1p peut réprimer conditionnellement certains gènes (Ge $\mathrm{et\ al.}$, 2005; Lavoie $\mathrm{et\ al.}$, 2012). Cette répression par une endoribonucléase nucléaire offrirait un mécanisme d'action plus rapide que d'affecter le niveau de transcription par l'intermédiaire du promoteur. De plus, l'association de Rnt1p au complexe de transcription permet une action localisée et efficace. Cette fonction en soi n'est pas essentielle, mais elle confère probablement un avantage sélectif important.

Rnt1p coupe plusieurs transcrits précurseurs d'ARNnc. Ces substrats sont parmi les plus efficaces et les mieux conservés. Dans le cas des ARNsno, ce ne sont qu'une partie de ceux-ci qui sont maturés par Rnt1p. De plus, pour plusieurs d'entre eux, il existe un mécanisme de maturation alternatif qui permet à ces ARNsno d'être maturés correctement même en l'absence de Rnt1p. Puisque le rôle des ARNsno est la modification des ARNr et que ces modifications non essentielles influencent le fonctionnement des ribosomes, il est possible que Rnt1p joue un rôle dans la régulation de ces modifications et dans la création d'une diversité parmi les ribosomes.

La régulation de Rnt1p dans l'accomplissement de ses différentes fonctions reste mal comprise. Le niveau d'expression de RNT1 varie peu d'une condition à l'autre, donc sa régulation ne semble pas dépendante de son promoteur. Il est possible que Rnt1p comportent un ou plusieurs sites de phosphorylation qui pourraient affecter sa localisation, ses partenaires ou sa reconnaissance des substrats. La validation de ses hypothèses reste à faire.

5.2 Conclusion

Le travail présenté ici a permis plusieurs découvertes importantes. Il a d'abord permis de mieux connaître la spécificité de l'enzyme Rnt1p. Des outils bio-informatiques ont été développés pour identifier des candidats potentiels en les comparant aux substrats connus. Ces outils ont l'avantage de pouvoir identifier des cibles que les méthodes *in vivo* ou *in vitro*

ne peuvent pas détecter comme des cibles à l'intérieur de transcrits faiblement exprimés ou exprimés seulement dans certaines conditions. Parmi les candidats identifiés, quelques uns ont pu être validés et publiés et d'autres sont en cours de publication.

Plus d'information sur l'influence de Rnt1p sur la transcription a été obtenue par l'analyse de l'expression par puces à ADN. Une augmentation globale du niveau de transcription a été observée et la surexpression spécifique de certains transcrits a été validée. De plus, certains de ces transcrits sont des transcrits non codants dont la fonction est encore inconnue et l'information obtenue pourra fournir des pistes quant à leur régulation. Ces transcrits sont des candidats intéressants pour être caractérisés plus en détail de façon à identifier leurs fonctions. Une nouvelle technique de correction des données de puces à ADN a aussi été développée pour réduire le bruit dû à la variabilité entre les sondes et ainsi réduire le nombre d'artéfacts dus à la composition locale du génome.

Plusieurs centaines de nouvelles cibles de Rnt1p ont été identifiées et parmi celles-ci, plusieurs sont ciblées grâce à une région qui ne ressemble pas aux autres substrats connus. De plus, il a été montré que tous les substrats ne sont pas égaux pour Rnt1p et que certains groupes de substrats sont reconnus différemment. Tous ces nouveaux transcrits permettent d'avoir un catalogue des cibles de Rnt1p dans le génome et de mieux comprendre quelles sont les voies métaboliques régulées par cet enzyme. Les nouvelles classes de transcrits pourront être testées *in vitro* pour identifier les déterminants de leur structure qui permettent à Rnt1p de les lier. Aussi, les algorithmes de classification pourront utiliser cet ensemble de substrats pour augmenter leur capacité à identifier de façon fiable de nouveaux substrats. Il sera aussi important de rechercher des caractéristiques communes à l'intérieur des nouvelles classes de substrats de façon à pouvoir identifier des pistes sur la façon dont elles sont reconnues.

Finalement, ce travail montre comment une approche intégrée incluant la recherche *in silico*, *in vivo* et *in silico* permet de couvrir de façon plus approfondie l'influence d'un enzyme dans la cellule et comment chaque approche peut pallier les déficiences des autres approches et fournir globalement des résultats plus complets.

ANNEXE A

PREMIÈRE ANNEXE

tableau A.1: Liste des cibles validées et publiées de Rnt1p.

Nom	Cible	Réference	ARN testé	Orthologues	Pointage
ADI1	ARNm	Zer et Chanfreau (2005)	Synthétique	0	0,683
ARN2-1	ARNm	Lee et al. (2005)	Synthétique	0	0,000
ARN2-2	ARNm	Lee et al. (2005)	Synthétique	0	0,766
MIG2	ARNm	Ge et al. (2005)	Total	1	0,988
FIT2	ARNm 5' UTR	Lee et al. (2005)	Synthétique	1	0,802
U1	ARNsn	Seipelt et al. (1999)	Synthétique	4	0,947
U2	ARNsn	Elela et Ares (1998)	Synthétique	3	0,998
U4	ARNsn	Allmang et al. (1999)	Synthétique	5	0,991
U5	ARNsn	Chanfreau et al. (1997)	Synthétique	4	0,986
snR17a	ARNsno C/D	Kufel et al. (2000)	Synthétique	5	0,988
snR190	ARNsno C/D	Chanfreau et al. (1998b)	Synthétique	0	0,728
snR39	ARNsno C/D	Ghazal et al. (2005)	Total	2	0,958
snR39B	ARNsno C/D	Chanfreau et al. (1998a)	Synthétique	4	0,934
snR40	ARNsno C/D	Chanfreau et al. (1998a)	Synthétique	4	0,978
snR47	ARNsno C/D	Chanfreau et al. (1998a)	Synthétique	4	0,995

snR48	ARNsno C/D	Ghazal et al. (2005)	Total	4	0,000
snR50	ARNsno C/D	Lee et al. (2003)	Synthétique	1	0,867
snR51	ARNsno C/D	Chanfreau et al. (1998a)	Synthétique	4	0,962
snR52	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,939
snR55	ARNsno C/D	Ghazal et al. (2005)	Total	3	0,912
snR56	ARNsno C/D	Ghazal et al. (2005)	Total	4	0,902
snR57	ARNsno C/D	Ghazal et al. (2005)	Total	5	0,949
snR58	ARNsno C/D	Lee et al. (2003)	Synthétique	3	0,986
snR59	ARNsno C/D	Ghazal et al. (2005)	Total	2	0,705
snR60	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,988
snR62	ARNsno C/D	Lee et al. (2003)	Synthétique	0	0,973
snR63	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,987
snR64	ARNsno C/D	Lee et al. (2003)	Synthétique	3	0,939
snR65	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,985
snR66	ARNsno C/D	Lee et al. (2003)	Synthétique	2	0,952
snR67	ARNsno C/D	Ghazal et al. (2005)	Total	4	0,978

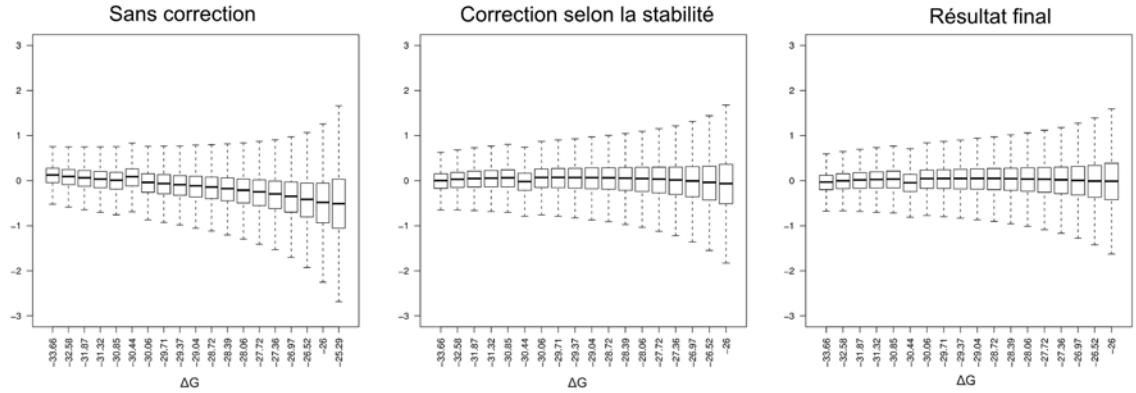
snR68	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,994
snR69	ARNsno C/D	Lee et al. (2003)	Synthétique	3	0,938
snR71	ARNsno C/D	Lee et al. (2003)	Synthétique	4	0,959
snR73	ARNsno C/D	Qu et al. (1999)	Synthétique	2	0,978
snR75	ARNsno C/D	Qu et al. (1999)	Synthétique	3	0,898
snR76	ARNsno C/D	Qu et al. (1999)	Synthétique	5	0,969
snR78	ARNsno C/D	Qu et al. (1999)	Synthétique	4	0,974
snR79	ARNsno C/D	Chanfreau et al. (1998a)	Synthétique	4	0,941
snR36	ARNsno H/ACA	Chanfreau et al. (1998a)	Synthétique	3	0,956
snR43	ARNsno H/ACA	Chanfreau et al. (1998a)	Synthétique	4	0,985
snR46	ARNsno H/ACA	Chanfreau et al. (1998a)	Synthétique	1	0,000
RPL18A	Intron	Danin-Kreiselman et al. (2003)	Synthétique	4	0,915
RPS22B	Intron	Danin-Kreiselman et al. (2003)	Synthétique	4	0,961
25S	rRNA	Elela et al. (1996)	Synthétique	6	0,988

La colonne référence renvoie à l'article ayant fait la première validation *in vitro*. La colonne orthologues indique le nombre d'orthologues trouvés parmi les espèces de Saccharomyces tel que décrit au chapitre 1. Le pointage est celui assigné par l'algorithme d'identification par similitude.

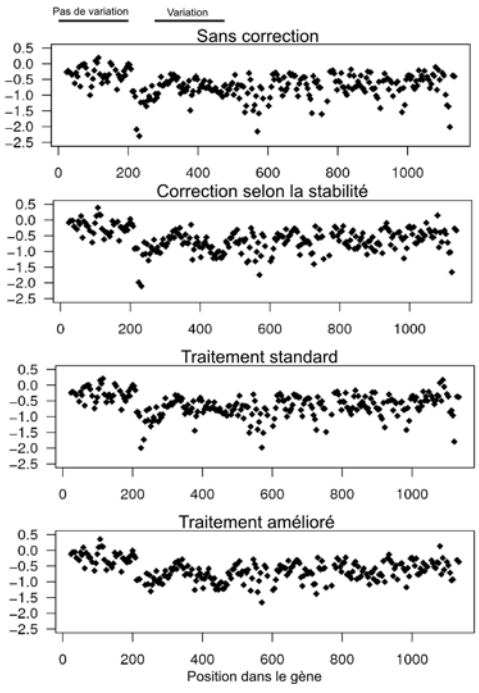
ANNEXE B

DEUXIÈME ANNEXE

A



B



	Signal(μ)	Bruit (σ)	$\frac{\text{Signal}(\mu)}{\text{Bruit}(\sigma)}$
Sans correction	0.3738	0.2389	1.5649
Correction selon la stabilité	0.5477	0.2645	2.0709
Traitement standard	0.4660	0.2214	1.8344
Traitement amélioré	0.5644	0.2492	2.2650

figure B.1 – L'ajustement selon la stabilité thermodynamique (ΔG) réduit le bruit.

A Variation dans le niveau des sondes selon la stabilité. B Exemple pour un gène contrôle

BIBLIOGRAPHIE

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E. and Tollervey, D. (1999). Functions of the exosome in rRNA, snoRNA and snRNA synthesis. EMBO J *18*, 5399–5410.

Babiskin, A. H. and Smolke, C. D. (2011). Synthetic RNA modules for fine-tuning gene expression levels in yeast by modulating RNase III activity. Nucleic Acids Res *39*, 8651–8664.

Camblong, J., Iglesias, N., Fickentscher, C., Dieppois, G. and Stutz, F. (2007). Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. Cell *131*, 706–717.

Carl, W. (1968). The Genetic Code. Harper & Row.

Catala, M., Aksouh, L. and Elela, S. A. (2012). RNA-dependent regulation of the cell wall stress response. Nucleic Acids Res *40*, 7507–7517.

Catala, M., Lamontagne, B., Larose, S., Ghazal, G. and Elela, S. A. (2004). Cell cycle-dependent nuclear localization of yeast RNase III is required for efficient cell division. Mol Biol Cell *15*, 3015–3030.

Cech, T. R. (2000). Structural biology. The ribosome is a ribozyme. Science *289*, 878–879.

Chanfreau, G. (2003). Conservation of RNase III processing pathways and specificity in hemiascomycetes. Eukaryot Cell *2*, 901–909.

Chanfreau, G., Elela, S. A., Ares, M. and Guthrie, C. (1997). Alternative 3'-end processing of U5 snRNA by RNase III. Genes Dev *11*, 2741–2751.

Chanfreau, G., Legrain, P. and Jacquier, A. (1998a). Yeast RNase III as a key processing enzyme in small nucleolar RNAs metabolism. J Mol Biol *284*, 975–988.

Chanfreau, G., Rotondo, G., Legrain, P. and Jacquier, A. (1998b). Processing of a dicistronic small nucleolar RNA precursor by the RNA endonuclease Rnt1. EMBO J *17*, 3726–3737.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. Machine Learning *20*, 273–297.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol *12*, 138–163.

Danin-Kreiselman, M., Lee, C. Y. and Chanfreau, G. (2003). RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns. Mol Cell *11*, 1279–1289.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. and Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A *103*, 5320–5325.

Davis, C. A. and Ares, M. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *103*, 3262–3267.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2008). Misc Functions of the Department of Statistics (e1071), TU Wien.

Drinnenberg, I. A., Fink, G. R. and Bartel, D. P. (2011). Compatibility with killer explains the rise of RNAi-deficient fungi. Science *333*, 1592.

Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Res *22*, 2079–2088.

Edvardsson, S., Gardner, P. P., Poole, A. M., Hendy, M. D., Penny, D. and Moulton, V. (2003). A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. Bioinformatics *19*, 865–873.

Elela, S. A. and Ares, M. (1998). Depletion of yeast RNase III blocks correct U2 3' end formation and results in polyadenylated but functional U2 snRNA. EMBO J *17*, 3738–3746.

Elela, S. A., Igel, H. and Ares, M. (1996). RNase III cleaves eukaryotic preribosomal RNA at a U3 snoRNP-dependent site. Cell *85*, 115–124.

Esguerra, J., Warringer, J. and Blomberg, A. (2008). Functional importance of individual rRNA 2'-O-ribose methylations revealed by high-resolution phenotyping. RNA *14*, 649–656.

Felsenfeld, G. and Cantoni, G. L. (1964). Use of thermal denaturation studies to investigate the base sequence of yeast serine sRNA. Proc Natl Acad Sci U S A *51*, 818–826.

Fromont-Racine, M., Senger, B., Saveanu, C. and Fasiolo, F. (2003). Ribosome assembly in eukaryotes. Gene *313*, 17–42.

Ganot, P., Caizergues-Ferrer, M. and Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev *11*, 941–956.

Gaudin, C., Ghazal, G., Yoshizawa, S., Elela, S. A. and Fourmy, D. (2006). Structure of an AAGU tetraloop and its contribution to substrate selection by yeast RNase III. J Mol Biol *363*, 322–331.

Ge, D., Lamontagne, B. and Elela, S. A. (2005). RNase III-mediated silencing of a glucose-dependent repressor in yeast. Curr Biol *15*, 140–145.

German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., Paoli, E. D., Lu, C., Schroth, G., Meyers, B. C. and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol *26*, 941–946.

Ghazal, G. and Elela, S. A. (2006). Characterization of the reactivity determinants of a novel hairpin substrate of yeast RNase III. J Mol Biol *363*, 332–344.

Ghazal, G., Gagnon, J., Jacques, P.-E., Landry, J.-R., Robert, F. and Elela, S. A. (2009). Yeast RNase III triggers polyadenylation-independent transcription termination. Mol Cell *36*, 99–109.

Ghazal, G., Ge, D., Gervais-Bird, J., Gagnon, J. and Elela, S. A. (2005). Genome-wide prediction and analysis of yeast RNase III-dependent snoRNA processing signals. Mol Cell Biol *25*, 2981–2994.

Giorgi, C., Fatica, A., Nagel, R. and Bozzoni, I. (2001). Release of U18 snoRNA from its host intron requires interaction of Nop1p with the Rnt1p endonuclease. EMBO J *20*, 6856–6865.

Hafner, M., Landgraf, P., Ludwig, J., Rice, A., Ojo, T., Lin, C., Holoch, D., Lim, C. and Tuschl, T. (2008). Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods *44*, 3–12.

Hartman, E., Wang, Z., Zhang, Q., Roy, K., Chanfreau, G. and Feigon, J. (2013). Intrinsic dynamics of an extended hydrophobic core in the S. cerevisiae RNase III dsRBD contributes to recognition of specific RNA binding sites. J Mol Biol *425*, 546–562.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res *31*, 3429–3431.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics *6*, 65–70.

Huber, W., Toedling, J. and Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. Bioinformatics *22*, 1963–1970.

Johnson, J. M., Edwards, S., Shoemaker, D. and Schadt, E. E. (2005). Dark matter in the genome : evidence of widespread transcription detected by microarray tiling experiments. Trends Genet *21*, 93–102.

Jády, B. E. and Kiss, T. (2001). A small nucleolar guide RNA functions both in 2'-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. EMBO J *20*, 541–551.

Kishore, S. and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science *311*, 230–232.

Kiss-László, Z., Henry, Y. and Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. EMBO J *17*, 797–807.

Knudsen, B. and Hein, J. (2003). Pfold : RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res *31*, 3423–3428.

Kufel, J., Allmang, C., Chanfreau, G., Petfalski, E., Lafontaine, D. L. and Tollervey, D. (2000). Precursors to the U3 small nucleolar RNA lack small nucleolar RNP proteins but are stabilized by La binding. Mol Cell Biol *20*, 5415–5424.

Lamontagne, B. and Elela, S. A. (2004). Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J Biol Chem *279*, 2231–2241.

Lamontagne, B. and Elela, S. A. (2007). Short RNA guides cleavage by eukaryotic RNase III. PLoS One *2*, e472.

Lamontagne, B., Larose, S., Boulanger, J. and Elela, S. A. (2001). The RNase III family : a conserved structure and expanding functions in eukaryotic dsRNA metabolism. Curr Issues Mol Biol *3*, 71–78.

Larose, S., Laterreur, N., Ghazal, G., Gagnon, J., Wellinger, R. J. and Elela, S. A. (2007). RNase III-dependent regulation of yeast telomerase. J Biol Chem *282*, 4373–4381.

Lavoie, M. and Elela, S. A. (2008). Yeast ribonuclease III uses a network of multiple hydrogen bonds for RNA binding and cleavage. Biochemistry *47*, 8514–8526.

Lavoie, M., Ge, D. and Elela, S. A. (2012). Regulation of conditional gene expression by coupled transcription repression and RNA degradation. Nucleic Acids Res *40*, 871–883.

Lee, A., Henras, A. K. and Chanfreau, G. (2005). Multiple RNA surveillance pathways limit aberrant expression of iron uptake mRNAs and prevent iron toxicity in S. cerevisiae. Mol Cell *19*, 39–51.

Lee, C. Y., Lee, A. and Chanfreau, G. (2003). The roles of endonucleolytic cleavage and exonucleolytic digestion in the 5'-end processing of S. cerevisiae box C/D snoRNAs. RNA *9*, 1362–1370.

Liang, Y.-H., Lavoie, M., Comeau, M.-A., Elela, S. A. and Ji, X. (2014). Structure of a Eukaryotic RNase III Postcleavage Complex Reveals a Double-Ruler Mechanism for Substrate Selection. Mol Cell *54*, 431–444.

Liao, Y., Smyth, G. K. and Shi, W. (2013). The Subread aligner : fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res *41*, e108.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B. and Bartel, D. P. (2003). The microRNAs of Caenorhabditis elegans. Genes Dev *17*, 991–1008.

Liu, J., Gough, J. and Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet *2*, e29.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955–964.

Lowe, T. M. and Eddy, S. R. (1999). A computational screen for methylation guide snoRNAs in yeast. Science *283*, 1168–1171.

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C. and Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. Science *309*, 1567–1569.

Lustig, A. J. (1999). Crisis intervention : the role of telomerase. Proc Natl Acad Sci U S A *96*, 3339–3341.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res *29*, 4724–4735.

Martin, M. (2011). Cutadapt removes adapter sequences from highthroughput sequencing reads. EMBnet.journal *17*, pp–10.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta *405*, 442–451.

McCutcheon, J. P. and Eddy, S. R. (2003). Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res *31*, 4119–4128.

Meaux, S., Lavoie, M., Gagnon, J., Elela, S. A. and van Hoof, A. (2011). Reporter mRNAs cleaved by Rnt1p are exported and degraded in the cytoplasm. Nucleic Acids Res *39*, 9357–9367.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Nawrocki, E. P., Kolbe, D. L. and Eddy, S. R. (2009). Infernal 1.0 : inference of RNA alignments. Bioinformatics *25*, 1335–1337.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443–453.

Neil, H., Malabat, C., d'Aubenton Carafa, Y., Xu, Z., Steinmetz, L. M. and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Pannucci, J. A., Haas, E. S., Hall, T. A., Harris, J. K. and Brown, J. W. (1999). RNase P RNAs from some Archaea are catalytically active. Proc Natl Acad Sci U S A *96*, 7803–7808.

Perocchi, F., Xu, Z., Clauder-Munster, S. and Steinmetz, L. M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucl. Acids Res. *35*, e128+.

Qu, L. H., Henras, A., Lu, Y. J., Zhou, H., Zhou, W. X., Zhu, Y. Q., Zhao, J., Henry, Y., Caizergues-Ferrer, M. and Bachellerie, J. P. (1999). Seven novel methylation guide small nucleolar RNAs are processed from a common polycistronic transcript by Rat1p and RNase III in yeast. Mol Cell Biol *19*, 1144–1158.

Rivas, E. and Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics *16*, 583–605.

Sam, M., Henras, A. K. and Chanfreau, G. (2005). A conserved major groove antideterminant for Saccharomyces cerevisiae RNase III recognition. Biochemistry *44*, 4181–4187.

Seipelt, R. L., Zheng, B., Asuru, A. and Rymond, B. C. (1999). U1 snRNA is cleaved by RNase III and processed through an Sm site-dependent pathway. Nucleic Acids Res *27*, 587–595.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol *13*, R67.

van Dijk, E. L., Chen, C. L., d'Aubenton Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Né, P., Loeillet, S., Nicolas, A., Thermes, C. and Morillon, A. (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature *475*, 114–117.

Woodhams, M., Stadler, P., Penny, D. and Collins, L. (2007). RNase MRP and the RNA processing cascade in the eukaryotic ancestor. BMC Evolutionary Biology *7*.

Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L. C. and Showe, M. K. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics *22*, 1325–1334.

Zer, C. and Chanfreau, G. (2005). Regulation and surveillance of normal and 3'-extended forms of the yeast aci-reductone dioxygenase mRNA by RNase III cleavage and exonucleolytic degradation. J Biol Chem *280*, 28997–29003.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res *31*, 3406–3415.