Étude du décalage de phase de lecture dans le génome de Saccharomyces cerevisiæ

Michaël Bekaert

Université Paris VI - Pierre et Marie Curie

Thèse

Étude du décalage de phase de lecture dans le génome de *Saccharomyces cerevisiæ*

Présenté devant
L'université Paris VI – Pierre et Marie Curie

Pour obtenir
Le grade de docteur

Formation doctorale
Physiologie et Génétique des Microorganismes

École doctorale
Logique du vivant : génome, cellule, développement

Par
Michaël Bekaert

Soutenue le 18 novembre 2004 devant le jury composé de :

Pr. B. Dujon	Président
Pr. C. Gaillardin	Examinateur
Pr. J.-P. Rousset	Directeur de thèse
Dr. M. Springer	Examinateur
Dr. C. Thermes	Rapporteur
Pr. E. Westhof	Rapporteur

Université Paris VI - Pierre et Marie Curie

Thèse

Étude du décalage de phase de lecture dans le génome de *Saccharomyces cerevisiæ*

Présenté devant
L'université Paris VI – Pierre et Marie Curie

Pour obtenir
Le grade de docteur

Formation doctorale
Physiologie et Génétique des Microorganismes

École doctorale
Logique du vivant : génome, cellule, développement

Par
Michaël Bekaert

Soutenue le 18 novembre 2004 devant le jury composé de :

Pr. B. Dujon	Président
Pr. C. Gaillardin	Examinateur
Pr. J.-P. Rousset	Directeur de thèse
Dr. M. Springer	Examinateur
Dr. C. Thermes	Rapporteur
Pr. E. Westhof	Rapporteur

Avant propos

Cette thèse a constitué une partie importante de mes quatre dernières années. De cette belle aventure académique, je retire le plaisir de m'être laissé guider par un sujet au gré des questions qui se sont posées à moi. Si les périodes d'hésitation et de découragement n'ont pas manqué, cette expérience de recherche sera toujours restée stimulante par sa diversité.

Par ailleurs, on ne passe pas de 24 à 27 ans en un lieu sans en être profondément empreint, et j'ai la sensation que mes yeux se sont ouverts un peu plus grand au cours de ces années. Des multiples rencontres que j'ai pu faire et dont je ne mesure sans doute pas encore toute l'importance, je ressens la plus grande gratitude.

Si une thèse ne porte qu'un nom, il s'agit bien d'une œuvre collective. Outre mes co-auteurs, plusieurs personnes reconnaîtront certainement le fruit de discussions que nous avons pu avoir.

- Jean-Pierre Rousset m'a accueilli dans son laboratoire. Depuis lors, il n'a cessé d'être disponible pour m'écouter, discuter, m'aider et me suggérer des conseils sans jamais chercher à m'imposer ses points de vue ;
- Les résultats que je présente proviennent du travail d'une équipe. Chacun de ses membres a participé à rendre son fonctionnement efficace, agréable et toujours amical. Plus particulièrement Isabelle Hatin, pour son épaule amicale, toujours prête à écouter mes états d'âme et mes difficultés scientifiques, ainsi que Céline Fabret, Laure Bidou, Marta Kwapisz (notre feu follet polonais), Agnès Baudin-Baillieu, Bruno Cosnier et Olivier Namy ;
- Les discussions avec des collègues, entre deux portes, au fond du couloir ou quelquefois dans des réunions un peu plus structurées ont été scientifiquement essentielles. Merci à Alain Denise, Christine Froideveau, Bernard Prum et Michel Termier ;
- Les remarques des membres de mon jury de thèse, Bernard Dujon, Claude Gaillardin, Mathias Springer, Claude Thermes et Eric Westhof, et leur gentillesse pour avoir accepté de participer à mon jury, ont été scientifiquement et humainement importantes pour moi ;
- Enfin mes rapports stimulants avec mes condisciples Jean-Paul Forest et Hugues Richard ont été très précieux.

Sommaire

Introduction

1 Introduction

1.1 La traduction

Comme le suggèrent les théories actuelles, le monde vivant aurait basculé d'une composition essentiellement ARN à une composition mixte, où le vivant dépend de la combinaison acides nucléiques / protéines. Cette étape a exigé l'apparition d'une machinerie moléculaire, le ribosome, permettant le transfert, par traduction, de la majorité des capacités catalytiques des ARN aux protéines, laissant aux premiers les propriétés de

support génétique. Cependant l'intégrité des protéines, et ainsi des fonctions cellulaires, exige une étape de traduction du code plus fiable que ne le permettent les simples réactions physico-chimiques concernées.

La traduction est l'étape de l'expression génique la plus coûteuse en énergie, puisque dans le cas de la levure *Saccharomyces cerevisiæ*, en phase exponentielle de croissance, 40% des protéines appartiennent à la machinerie de traduction. Cette machinerie permettant l'enchaînement des acides aminés est très complexe et fait intervenir de nombreux facteurs aussi bien protéiques que nucléiques.

Les règles conventionnelles de la lecture du code génétique sont très bien connues et décrites dans tout manuel de biologie moléculaire. Cependant la situation réelle est plus complexe de par la nature sophistiquée de la traduction. Bien que le code génétique soit considéré universel, la signification de certains codons, dans des organelles et dans un nombre restreint d'organismes, a été réassignée. Dans ces situations, la signification d'un codon a été changée pour tous les ARN messager (ARNm). En outre, les règles conventionnelles de lecture du code génétique peuvent être altérées spécifiquement pour certains ARNm. De telles extensions du code génétique sont appelées événements de *recodage*. Le recodage est souvent en concurrence avec le décodage conventionnel, et seule une faible proportion des ribosomes fait alors du recodage à un locus particulier. Ces événements peuvent être utilisés pour une régulation spécifique de la synthèse protéique.

Le travail présenté se concentre sur ces altérations transitoires de lecture du code génétique. En particulier le déphasage du cadre de lecture, les mécanismes sous-jacents de ce contrôle précis, ainsi que l'identification de gènes présentant ce type de régulations.

1.1.1 Le ribosome

Chez les eucaryotes, et en particulier la levure *S. cerevisiæ*, le ribosome cytoplasmique se compose de 77 protéines et de 4 ARN ribosomiques (ARNr). Les ARNr 25S (3 392 nt), 5,8S (158 nt), 5S (121 nt) et 45 protéines font partie de la grande sous-unité 60S. L'ARNr 18S (1 798 nt) est le seul composant nucléique de la petite sous-unité 40S qui compte 32 protéines. Lorsque ces deux particules sont assemblées et forment un ribosome fonctionnel (80S), trois cavités permettent l'allongement de la chaîne polypeptidique : Le site A (pour Aminoacyl-ARNt), le site P (pour Peptidyl-ARNt) et le site E (pour Exit). Lors de l'élongation, les ARN de transfert (ARNt) passent successivement par chacun de ces trois sites. Ces molécules jouent un rôle essentiel dans la traduction, puisqu'elles permettent la conversion du message nucléotidique en protéines. Les ARNt

ont une structure secondaire tout à fait particulière, qui leur permet d'une part d'être reconnus par des enzymes pour être spécifiquement chargés d'un acide aminé en fonction du codon reconnu, et d'autre part de s'apparier avec les codons présents sur l'ARNm grâce à l'anticodon et d'interagir avec le ribosome. Le ribosome joue aussi un rôle dans d'autres processus biologiques cellulaires, tels que le recodage (Gesteland *et al.*, 1992), le transport et le repliement de protéines (Beckmann *et al.*, 1997; Beckmann *et al.*, 2001) et probablement dans l'adressage aux ports nucléaires (Ho *et al.*, 2000).

La compréhension des mécanismes de la traduction passe par l'étude des interactions entre le ribosome et ses substrats. Celles entre le ribosome et l'ARNt ou l'ARNm sont cruciales puisqu'elles sont à la base des processus de décodage. Des avancées importantes ont été réalisées depuis quelques années sur la structure du ribosome procaryote et des facteurs associés par les études de cryo-microscopie électronique (Frank, 2001; Stark *et al.*, 2000) et de diffractions des rayons X (Ban *et al.*, 2000; Ramakrishnan, 2002; Yonath, 2002; Yusupov *et al.*, 2001). Ces études ont permis de caractériser le réseau complexe d'interactions existant entre les ARNt aux sites A, P et E, le ribosome et les ARNm (Yusupova *et al.*, 2001). L'obtention de la structure du ribosome 80S de *S. cerevisiæ* complexé avec un ARNt au site P par cryo-microscopie électronique et modélisation avec les structure cristallographique des sous-unités 30S de *Thermus thermophilus* (Wimberly *et al.*, 2000) et 50S de l'archaebactérie *Haloarcula marismortui* (Ban *et al.*, 2000) a montré que le ribosome est un complexe remarquablement conservé entre les différents règnes, notamment en son sein où se situent les centres catalytiques (Spahn *et al.*, 2001).

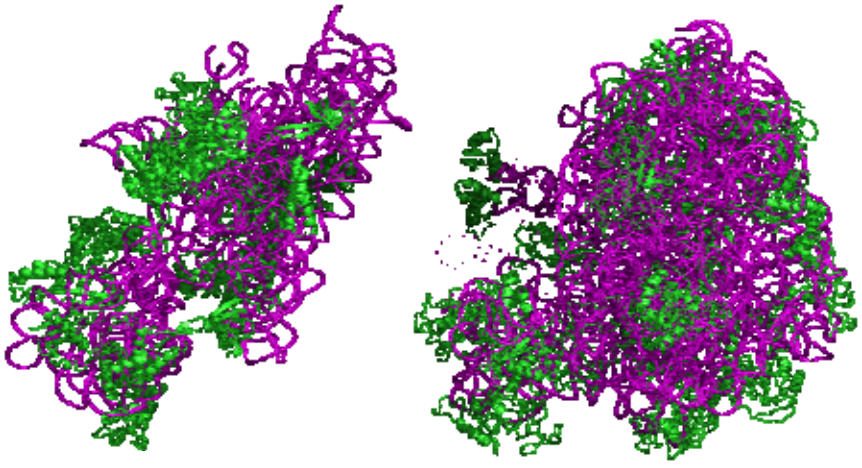


Figure 1 : Ribosome de la levure *S. cerevisiæ*. De gauche à droite, la sous-unité 40S et la sous-unité 60S. En rose figure le trajet de l'ARNr en vert la structure des protéines. (Données cryo-em 1S1H et 1S1I ; Spahn *et al.*, 2004)

1.1.2 Modifications des ARN

Les ARNt comme les ARNr sont modifiés de manière post-transcriptionnelle. Deux types de modifications sont prédominantes : les méthylations (qui affectent en majorité le ribose) et les pseudouridylations. Il est intéressant de noter que ces modifications sont deux fois plus nombreuses chez les vertébrés que chez la levure *Saccharomyces carlsbergensis* qui est proche de *Saccharomyces cerevisiæ* (Maden, 1990). Il n'est pas toujours bien défini si certaines de ces modifications se déroulent avant ou conjointement aux premières étapes de maturation (excision des introns). Si les modifications entre les ARNt et les ARNr sont semblables, les mécanismes mis en jeu sont différents. Pour les ARNt, la reconnaissance de la séquence à modifier est effectuée directement par l'enzyme. Dans le cas des ARNr, la reconnaissance de la séquence à modifier s'effectue généralement par l'intermédiaire de petits ARN guides (snoRNA), de 70 à 100 nt. Ils créent localement une région double hélice qui va être reconnue par la méthylase.

1.1.3 La machinerie traductionnelle eucaryote

1.1.3.1 L'initiation

L'initiation de la traduction chez les eucaryotes implique à la fois l'ARN messager, la sous-unité 40S et de nombreux facteurs protéiques. Le processus nécessite que l'ARNm présente à la fois une queue polyadénylée en 3' et une coiffe (m7GpppNm) en 5'. Si la polyadénylation a lieu immédiatement à la fin de la transcription via la poly(A) polymérase activée après clivage de l'extrémité 3' du messager, trois étapes successives sont nécessaires pour obtenir la coiffe. Chez *S. cerevisiæ*, ces trois étapes sont catalysées successivement par la protéine Cet1p (une ARN triphosphatase), la protéine Ceg1p (une ARN guanyltransférase) et par la protéine Abd1p (une ARN 7-guanidine méthyltransférase) (Mao *et al.*, 1995; Shibagaki *et al.*, 1992; Tsukamoto *et al.*, 1997).

Une vingtaine de protéines (eukaryotic translation Initiation Factor, eIF) interviennent dans le complexe d'initiation (Tableau 1). Le facteur eIF4E est celui qui reconnaît la structure de la coiffe et s'y fixe (Altmann *et al.*, 1985; Quiocho *et al.*, 2000). Il interagit avec un deuxième partenaire, le facteur eIF4G (Gross *et al.*, 2003), ce dernier va avoir pour fonction essentielle de recruter d'autres facteurs tels que les ARN hélicases eIF4A et eIF4B (Dominguez *et al.*, 1999; Matsuo *et al.*, 1997), nécessaires

pour résoudre les structures secondaires de l'ARNm. L'activité ARN hélicase semble être portée principalement par eIF4A, et eIF4B aurait pour fonction principale de stimuler l'activité ATPase de eIF4A (Rogers *et al.*, 2001). L'interaction entre les facteurs eIF4E et eIF4G est nécessaire afin de stabiliser le complexe eIF4E-coiffe (von Der Haar *et al.*, 2000). Une fois le complexe eIF4F (eIF4G, eIF4E et eIF4A) formé, la sous-unité 40S du ribosome est recrutée grâce à eIF3 par interaction avec eIF4G. Il a cependant été monté que eIF4A est absent du complexe eIF4F chez la levure (Goyer *et al.*, 1989). La protéine PABP (Pab1p chez la levure) se fixe sur la queue polyA des ARNm eucaryotes et régule ainsi la stabilité des ARNm (Bernstein *et al.*, 1989; Wang *et al.*, 1999; Wang and Kiledjian, 2000). Il a été montré que cette protéine joue un rôle prépondérant dans l'efficacité de formation du complexe d'initiation eIF4F. Effectivement la protéine PABP se lie au facteur eIF4G, cette interaction augmente alors l'affinité du facteur eIF4E pour la coiffe (von Der Haar *et al.*, 2000) et permet aussi une circularisation de l'ARNm (Wells *et al.*, 1998).

L'interaction entre eIF4G et eIF3 permet le transfert du complexe ternaire $ARNt^{Met}_i$-eIF2-GTP au sein de la sous-unité 40S au site P (Chaudhuri *et al.*, 1999). Une fois ce complexe 43S formé à l'extrémité 5' de l'ARNm, il va balayer la région 5' non traduite (5' UTR) de l'ARNm jusqu'à trouver un codon initiateur, généralement AUG. Il faut noter que chez *S. cerevisiæ* la taille des régions 5' non codantes est en moyenne de 25 à 32 nucléotides, que la taille minimum permettant une initiation efficace est de 15 nucléotides, et qu'enfin le premier codon AUG présent sur le transcrit est le codon initiateur. Plusieurs des facteurs déjà décrits interviennent lors du balayage du complexe de pré-initiation 43S, ainsi que dans la reconnaissance du site d'initiation. Les facteurs eIF4A et eIF4B ont un rôle important dans ces étapes puisqu'ils permettent la dissociation des structures secondaires. Il a été montré qu'*in vitro*, en absence des facteurs eIF1A et eIF1, le complexe 43S se formait mais était incapable de se déplacer (Pestova *et al.*, 1998). eIF1 joue aussi un autre rôle essentiel en permettant le recrutement du facteur eIF3 (Fletcher *et al.*, 1999). Le facteur eIF3 va permettre de stabiliser le complexe eIF1A-eIF2-$ARNt^{Met}_i$-40S, qui sinon se dissocie à l'arrivée de la sous-unité 60S (Chaudhuri *et al.*, 1999). Des résultats obtenus *in vitro* avec des extraits de germes de blé indiquent que la protéine PABP augmente l'activité hélicase du complexe eIF4A / eIF4B permettant un déplacement plus rapide du complexe 43S (Bi and Goss, 2000). Lorsque le complexe rencontre le codon initiateur, l'appariement entre le codon AUG et anti-codon de l'$ARNt^{Met}_i$ provoque une pause (Cigan *et al.*, 1988). Cette pause permet l'action du facteur eIF5 qui, en interagissant avec eIF3 (Phan *et al.*, 1998), provoque l'hydrolyse du

GTP porté par le facteur eIF2 (Asano *et al.*, 2000). Cette hydrolyse entraîne la dissociation de tous les facteurs eIF; la sous-unité 60S, jusque là protégée par eIF3A (Groft *et al.*, 2000), peut alors se fixer au complexe 40S-ARNt$^{Met}_{i}$

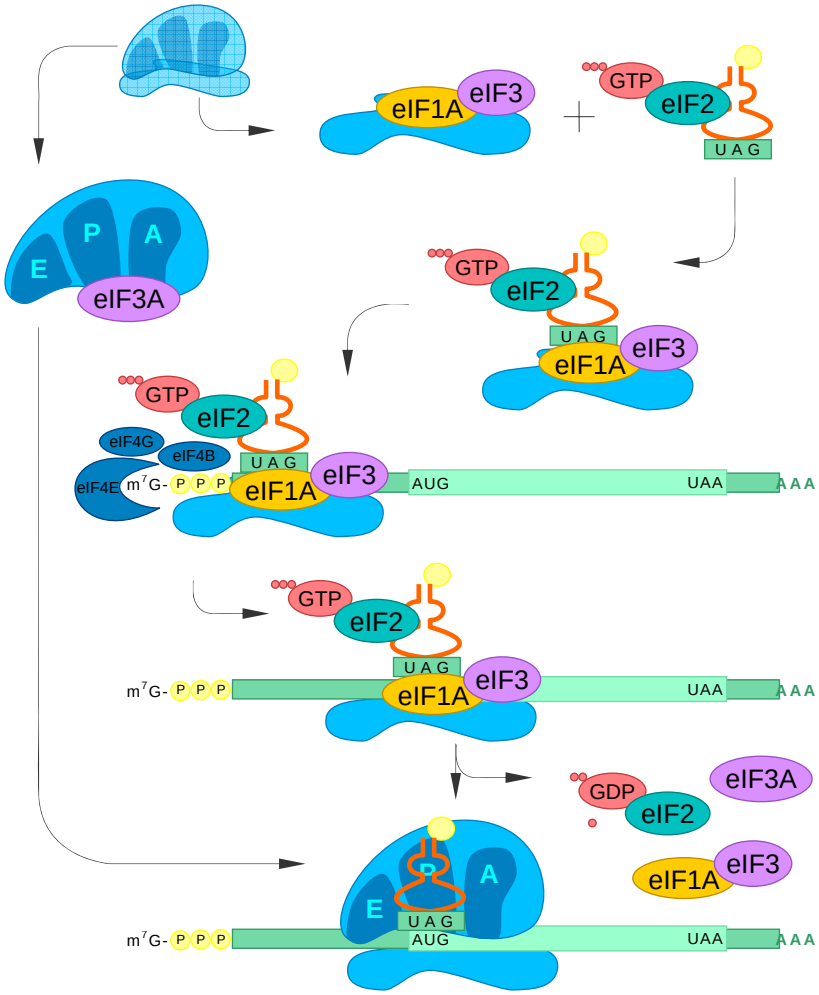


Figure 2 : Formation du complexe d'initiation de la traduction

L'interaction ARNt$^{Met}_{i}$-AUG est essentielle, mais pas suffisante pour promouvoir efficacement le démarrage de la traduction. Les facteurs eIF1 et eIF2 sont impliqués dans la fidélité de sélection du site d'initiation (Yoon and Donahue, 1992), cependant les mécanismes moléculaires demeurent mal compris. Le contexte nucléotidique du codon initiateur joue aussi un rôle important dans la reconnaissance du codon AUG par le complexe d'initiation 48S. Kozak a démontré que dans les cellules de mammifère la séquence optimale pour la sélection d'un codon AUG faisait intervenir les 6 bases précédant et la base suivant le codon initiateur

(CC[AG]CCAUGG) (Kozak, 1987; Kozak, 1997). Chez *S. cerevisiæ* la situation est sensiblement différente dans la mesure où la principale caractéristique de la région précédant le codon AUG est d'être riche en A, avec tout de même un biais au niveau du nucléotide -3, permettant de définir des contextes optimaux : [AG]NN AUG permettant quasiment 100% d'initiation, et des contextes non optimaux [CU]NN AUG permettant de 50% à 70% d'initiation (Yun *et al.*, 1996). Le fait que le premier codon AUG soit dans un contexte non optimal peut être un moyen d'exprimer, à partir du même gène, deux protéines avec des séquences N-terminales différentes.

Facteur	Fonction(s)	Masse	Sous-unités	Géne
eIF1	Stimule la liaison de l'ARNt$^{met}_{i}$ et de l'ARNm avec le ribosome 40S	12 kDa		SUI1
eIF1A	Stimule la liaison de l'ARNt$^{met}_{i}$ et de l'ARNm avec le ribosome 40S	17 kDa		TIF11
eIF2	Liaison de l'ARNt$^{met}_{i}$ au ribosome 40S	-	α (34 kDa) β (32 kDa) γ (58 kDa)	SUI2 SUI3 GCD11
eIF2B	Facteur d'échange GDP/GTP de eIF2	-	α (34 kDa) β (46 kDa) γ (66 kDa) δ (71 kDa) ε (81 kDa)	GCN3 GCD7 GCD1 GCD2 GCD6
eIF3	Dissociation du ribosome. Stabilise le complexe ternaire. Stimule la liaison avec l'ARNm	-	a (110 kDa) b (88 kDa) c (93 kDa) i (38 kDa) g (30 kDa)	TIF32 PRT1 NIP1 TIF34 TIF35
Ded1	Liaison à l'ARNm. ARN hélicase	65 kDa		DED1
p20	Liaison à l'ARNm.	20 kDa		CAF20
PADB	Reconnaissance du poly(A)	64 kDa		PAB1
eIF4A	Liaison à l'ARNm. ARN hélicase	45 kDa		TIF1/2
eIF4B	Liaison à l'ARNm. ARN hélicase	48 kDa		TIF3
eIF4E	Liaison à l'ARNm. Reconnaissance de la coiffe	24 kDa		CDC33
eIF4G	Liaison à l'ARNm. Protéine ancre	107 kDa		TIF4632 TIF4631
eIF5	Hydrolyse de eIF2-GTP	45 kDa		TIF5
eIF5A	Formation de la première liaison peptidique	17 kDa		HYP2 ANB1

Tableau 1 : Facteurs d'initiation de la traduction chez la levure.

L'association des deux sous-unités fait intervenir les ARNr. Les structures impliquées dans cette association ont été déduites des structures atomiques de basses résolutions obtenues pour le ribosome de levure. En plus des nombreux ponts entre les ARNr 18S et 25S, les protéines RPS13, RPL19 et RPL42 semblent également impliquées dans les associations entre les deux sous-unités (Spahn *et al.*, 2001). Le ribosome une fois formé est prêt à allonger la chaîne polypeptidique.

Ces données nous permettent de comprendre à quel point l'initiation de la traduction est complexe. De nombreuses zones d'ombre existent encore sur le rôle exact de certains facteurs. L'initiation peut aussi s'effectuer au niveau d'une structure très particulière l'IRES (Internal Ribosome Entry Site) et non au niveau du 1er codon d'initiation (Hellen and

Sarnow, 2001) ; et parfois certaines de ces structures utilisent une initiation au site A du ribosome (au lieu du site P). Cette initiation indépendante de la coiffe se produit à un codon non conventionnel (GCU) (Wilson *et al.*, 2000). L'initiation est l'étape de la traduction où le plus grand nombre de régulations a été identifié jusqu'à présent, c'est aussi la plus étudiée (Dyer and Sossin, 2000; Gil *et al.*, 2000; Harding *et al.*, 2000; Raught and Gingras, 1999). Le facteur eIF4G est notamment la cible de nombreuses protéases virales, ce qui provoque une inhibition de la traduction des gènes cellulaires lors de l'infection virale (Svitkin *et al.*, 1999). L'initiation traductionnelle est fortement régulée au cours du développement, puisque certains ARNm (dits maternels) sont transcrits puis stockés, leur traduction ne sera initiée que plus tard au cours de l'embryogenèse (Fahrenkrug *et al.*, 2000; Keiper and Rhoads, 1999; Wassarman and Kinloch, 1992). L'initiation est donc un point de contrôle clef dans l'expression des gènes; de son bon déroulement dépend toute la suite de la traduction.

1.1.3.2 L'élongation

La fin du processus d'initiation laisse l'ARNt initiateur aminoacylé au site P du ribosome et un site A vide, ce qui permet de débuter un cycle d'élongation.

Facteur	Fonction(s)	Masse	Géne
eEF1A	S'associe avec un aminoacyl-ARNt et une molécule de GTP. Amène l'ARNt dans le site A du ribosome	50 (kDa)	TEF2
eEF1β	Facteur d'échange GDP/GTP de eEF1	22 (kDa)	EFB1
eEF1γ	Facilite l'association eEF1α et aminoacyl-ARNt	46 (kDa)	TEF4
eEF2	Permet la translocation du peptidyl-ARNt du site A au site P via l'hydrolyse de GTP	93 (kDa)	EFT1/2
eEF3	S'associe avec les ribosomes. Permet le recyclage des ARNt. Activité ATPase	115 (kDa)	YEF3

Tableau 2 : Facteurs d'élongation de la traduction eucaryote. Le Facteur eEF3 est spécifique de la levure.

Les ARNt sont les éléments clefs de cette étape. Avant de pouvoir être utilisés par le ribosome ils doivent subir de nombreuses modifications, ensuite ils sont chargés de l'acide aminé leur correspondant par les aminoacyl-ARNt synthétases. Une fois chargés les ARNt sont très rapidement complexés avec le facteur eEF1A (eukaryotic elongation factor). Ce complexe formé, l'ARNt est prêt à être incorporé par la machinerie traductionnelle. Les facteurs d'élongation interviennent à chacune des différentes étapes, pour permettre l'incorporation de l'ARNt, pour vérifier que cet ARNt correspond bien au codon présent au site A,

pour provoquer la *translocation* du ribosome et enfin pour éjecter l'ARNt dé-acylé.

Le facteur eEF1A est chargé avec une molécule de GTP par le facteur eEF1β. L'aminoacyl-ARNt prend ensuite la place du facteur eEF1β. Le facteur eEF1α ne possède pas pour unique fonction de transporter l'aminoacyl-ARNt vers le ribosome, mais joue aussi un rôle essentiel dans la fidélité de la traduction (Dinman and Kinzy, 1997). Une fois constitué, le complexe ternaire aminoacyl-ARNt-eEF1A-GTP peut alors interagir avec le ribosome en coopération, chez la levure, avec le facteur eEF3 (Uritani and Miyazaki, 1988). L'interaction correcte codon / anti-codon entraîne un changement de conformation du ribosome qui stabilise la liaison à l'ARNt et permet l'hydrolyse du GTP par eEF1A. Le complexe bascule alors vers le site peptidyl transférase du ribosome. Ce processus est appelé accommodation. La liaison peptidique, qui implique la dé-acylation de l'ARNt du site P et le transfert de la chaîne peptidique à l'ARNt du site A, est réalisée par l'ARNr. Le ribosome présente alors un ARNt dé-acylé au site P et un ARNt portant la chaîne peptidique au site A. La translocation des ARNt et de l'ARNm est facilitée par eEF2, qui est également une GTPase. Le résultat est un ribosome prêt pour un nouveau cycle d'élongation, avec un ARNt dé-acylé au site E, un ARNt portant la chaîne peptidique au site P et un site A vide prêt à recevoir le prochain complexe ternaire cognat.

1.1.3.3 Le décodage

L'appariement entre le codon de l'ARNm et l'anti-codon porté par l'ARNt est à la base de la sélection correcte de l'ARNt qui participera à l'addition d'un nouvel acide aminé à la chaîne polypeptidique naissante. Cependant la différence d'énergie entre l'appariement d'un ARNt cognat, qui présente un appariement parfait, et un ARNt proche-cognat, qui ne présente généralement qu'un seul mésappariement, est trop faible pour expliquer l'exactitude du choix (taux d'erreur de 10^{-3} à 10^{-4}). L'énergie libre pour la formation par exemple d'une paire GU à la première position du codon, est pratiquement identique à celle d'une paire AU, pourtant le ribosome peut les distinguer.

Il a été suggéré que le ribosome contiendrait un site de décodage identifiant la géométrie de l'appariement codon / anti-codon, d'une façon comparable au mécanisme de reconnaissance enzyme / substrat (Hypothèse de l'*identification stérique* ; Davies *et al.*, 1964; Potapov, 1982). Une autre hypothèse, nommée *kinetic proofreading* fait appel à une étape de sélection suivie d'une étape de correction (Hopfield, 1974; Ninio, 1975), qui sont

séparées par une étape irréversible, l'hydrolyse du GTP par eEF1A. Dans cette hypothèse, l'ARNt peut se dissocier soit lors du choix initial, soit après l'hydrolyse du GTP. Théoriquement, ce processus peut avoir comme conséquence une sélectivité qui est le produit de la sélectivité de chaque étape. Mais dans la pratique, les vitesses relatives d'associations et de dissociation des ARNt à chaque étape déterminent la qualité de la sélection.

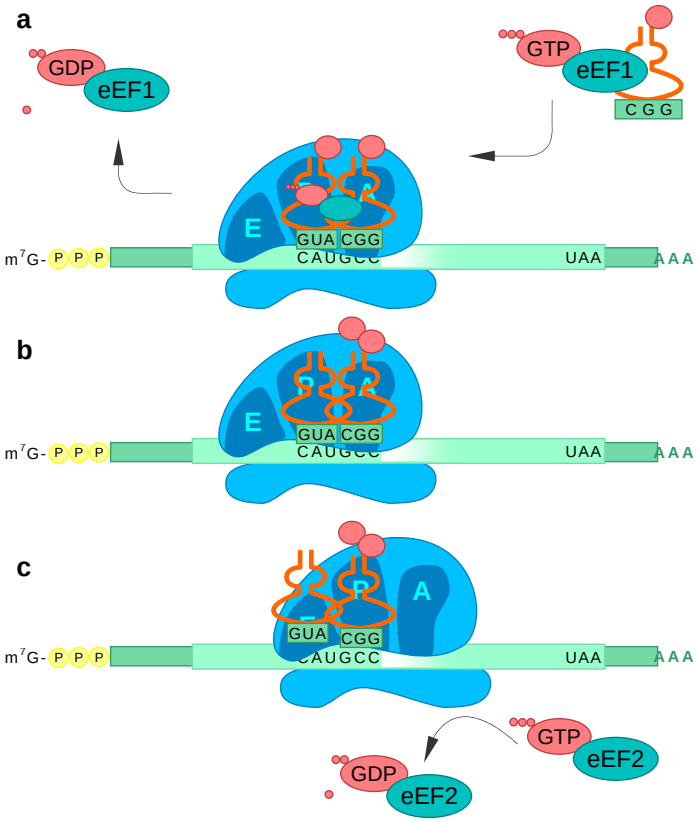


Figure 3 : Etapes de l'élongation de la traduction. a. Arrivée de l'ARNt guidé par eEF1 ; b. Formation de la liaison peptidique ; c. Déplacement.

Bien que l'identification stérique et le « *kinetic proofreading* » soient souvent considérées comme deux possibilités distinctes, les travaux récents prouvent qu'elles fonctionnent certainement de concert.

Le processus de sélection a été étudié en détail (Pape *et al.*, 1999). Un résultat étonnant est qu'en plus de vitesses de dissociation plus faibles, les ARNt cognats ont également des vitesses d'activation de la GTPase plus élevées que les ARNt proche-cognat. Basé sur ce résultat, il a été proposé que les ARNt cognats induisent un changement de conformation du ribosome. Un changement de conformation est aussi un élément du modèle à trois sites de Nierhaus (1990). Dans ce modèle, les affinités des ARNt des sites A et E sont réciproquement couplées. En présence de l'ARNt au site E, seul le complexe ternaire cognat a assez d'affinité pour le site A pour

induire un changement de conformation du ribosome, menant au dégagement de l'ARNt du site E.

1.1.3.4 La translocation

Après la transpeptidation, l'ARNt du site P est dé-acylé et l'ARNt du site A porte la chaîne peptidique avec un résidu additionnel. Pour préparer le ribosome à un nouveau cycle d'élongation, les ARNt doivent être déplacés : l'ARNt dé-acylé doit être déplacé du site P au site E pour ensuite sortir du ribosome, alors que l'ARNt porteur du polypeptide doit se déplacer du site A au site P. Ce déplacement doit être précis, et maintenir la phase de lecture sur l'ARNm.

Il a été proposé que la translocation implique un mouvement relatif des deux sous-unités et pourrait se produire lors d'étapes distinctes pour chacune des sous-unités (Bretscher, 1968). Ceci aurait pour conséquence des états intermédiaires où les ARNt seraient liés à la fois au site A dans la sous-unité 40S et au site P dans la sous-unité 60S. Ce modèle a l'avantage de ne faire se déplacer qu'une partie du complexe ARNt-ARNm à la fois, tandis que l'autre reste fixe et sert d'ancre à l'ensemble. Cette hypothèse crédite également l'existence universelle de deux sous-unités dans toutes les espèces. Mais il n'y a aucune donnée précise sur laquelle des sous-unités se déplace en premier. Récemment, un mouvement relatif entre les sous-unités a été suggéré en comparant des formes libres du ribosome procaryote à des formes associées à EF-G (équivalent de eEF2 chez les eucaryotes) (Frank and Agrawal, 2000).

Les données expérimentales sur la nature du mouvement des ARNt pendant la translocation proviennent essentiellement des travaux effectués sur les procaryotes. Les empreintes (*footprints*) caractéristiques des ARNt dans chaque site ont été utilisées pour suivre le mouvement de l'ARNt pendant le cycle d'élongation (Moazed and Noller, 1989). Lorsque la puromycine (qui mime un aminoacyl-ARNt au site A) est ajoutée aux ribosomes avec un ARNt au site P, l'empreinte de l'ARNt sur le 50S disparaît et est remplacée par une empreinte au site E, alors que l'empreinte dans la sous-unité 30S demeure inchangée, suggérant que l'ARNt est dans un état intermédiaire P/E. De même l'ARNt du site A se trouve initialement dans un état A/T (T pour EF-TU, équivalent de eEF1A chez les eucaryotes), où le complexe ternaire est au site A du ribosome, avec l'extrémité aminoacylée de l'ARNt attachée à EF-Tu. Après le départ d'EF-Tu, le bras accepteur de l'ARNt bascule dans le site peptidyl transférase du 50S, ayant pour résultat l'empreinte caractéristique de l'ARNt dans l'état du non-hybride A/A.

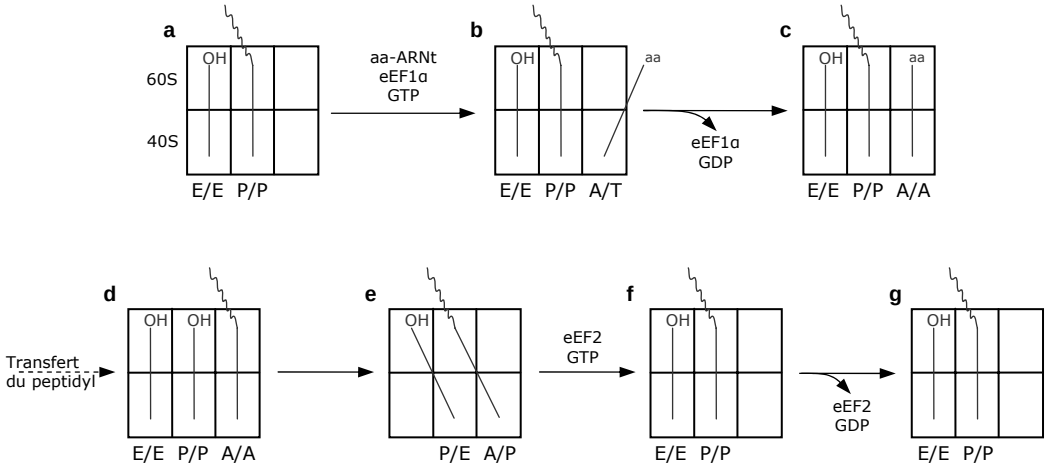


Figure 4 : Représentation schématique des étapes principales du cycle de translocation du modèle *états-hybrides*. Le ribosome 80S est schématisé par des rectangles, divisé en sous-unités 40S et 60S, dont chacune a un emplacement A, P et E. Les ARNt sont les lignes verticales, et l'ARNm n'est pas visualisé.

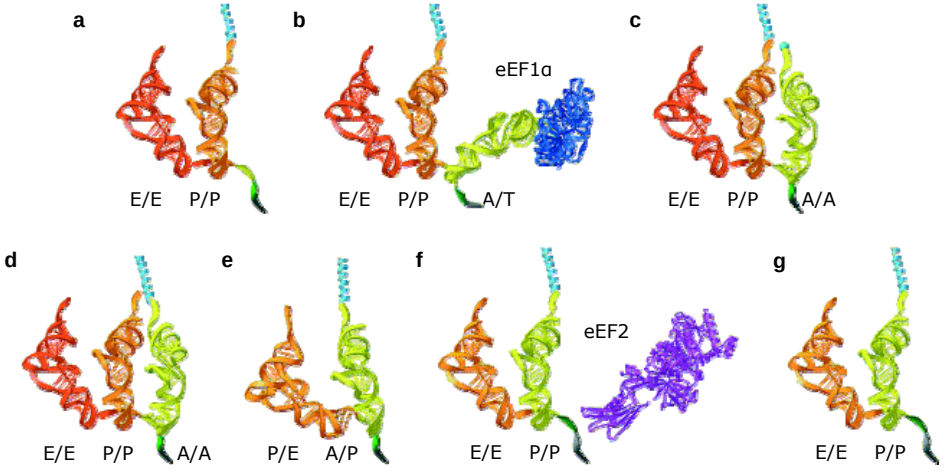


Figure 5 : Représentation tridimensionnelle du cycle de translocation montré dans la figure 4 (extrapolation des données structurales procaryotes ; Noller *et al.*, 2002).

Cependant, après le transfert du peptide, les empreintes des ARNt restent aux sites A et P dans la sous-unité 30S, alors que les bras accepteurs se sont déplacés respectivement aux sites P et E sur la sous-unité 50S. Ceci confirme les états hybrides A/P et P/E. Ce modèle de mouvement des ARNt distincts dans chaque sous-unité reste compatible avec l'idée de mouvement relatif des sous-unités elles-mêmes, comme cela a été vu en cryo-microscopie électronique (Noller *et al.*, 2002).

Un modèle « α-∈ » alternatif a été proposé (Spahn and Nierhaus, 1998). Le profil de protection du ribosome sur des ARNt, ne change pas pendant la translocation (Dabrowski *et al.*, 1998), suggérant que les contacts ribosomiques des ARNt aux sites A et P ne sont pas modifiés. Ceci

implique qu'il y ait des domaines mobiles qui transportent les ARNt à travers le ribosome.

1.1.3.5 La terminaison

Le processus de terminaison débute par l'entrée d'un codon stop au site A du ribosome. Chez les eucaryotes, le facteur eRF1 (pour eukaryotic translation release factor) identifie chacun des trois codons stop. Le facteur eRF3, une GTPase, s'associe à eRF1 et forme un complexe avec le ribosome.

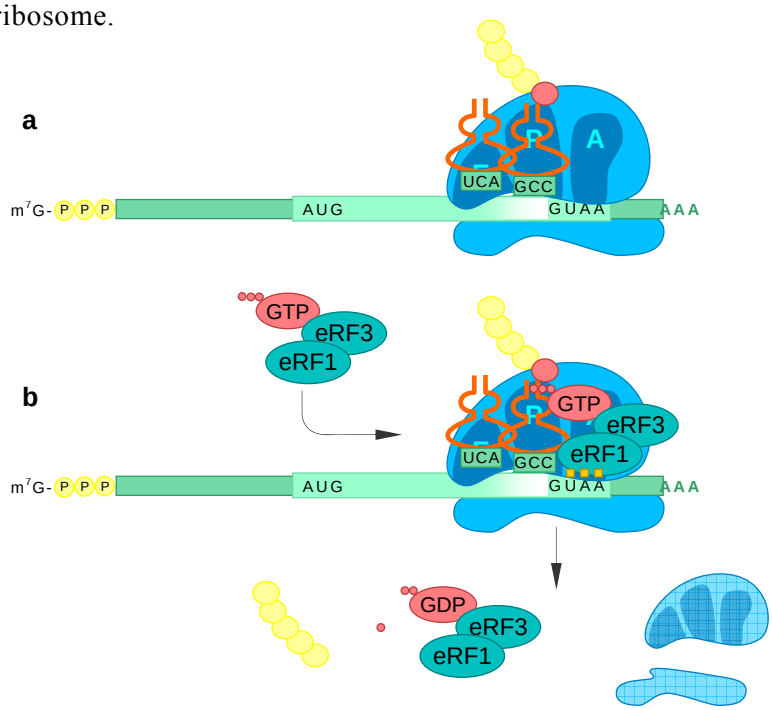


Figure 6 : Etapes de la terminaison de la traduction.

L'association de eRF1 avec le ribosome au site A, déclenche l'hydrolyse et le dégagement de la chaîne peptidique de l'ARNt au site P. La structure cristalline d'eRF1 humain suggère en effet que la protéine pourrait imiter un ARNt (Song *et al.*, 2000). Elle pourrait être modifiée de sorte qu'un motif hautement conservé `GGQ` soit près de l'extrémité `CCA` de l'ARNt du site P. Ceci est cohérent avec le fait que ce motif soit requis pour la libération du peptide (Frolova *et al.*, 1999; Song *et al.*, 2000). eRF3 favorise la dissociation rapide de eRF1. Initialement, il a été supposé que la liaison de eRF3 au ribosome déclenchait son activité de GTPase et le dégagement concomitant de eRF1.

1.1.3.6 Le recyclage

Après la libération de la chaîne polypeptidique, le ribosome reste avec l'ARNm et un ARNt dé-acylé engagé dans le site P. Ce complexe doit être désassemblé pour préparer le ribosome à un nouveau cycle. Actuellement aucun facteur de recyclage du ribosome n'a été identifié dans le cytoplasme des cellules eucaryotes. Pour expliquer cette différence avec les procaryotes où il existe un facteur de recyclage du ribosome (RRF), il a été suggéré (Buckingham *et al.*, 1997) que le caractère essentiel de eRF3 était dû à son rôle dans le recyclage à la fois des facteurs de terminaison (eRF1) et du ribosome. Bien que cette hypothèse semble être cohérente avec la taille d'eRF3 et ses propriétés structurales, elle n'a jamais été vérifiée directement.

1.2 Le recodage

La machinerie traductionnelle a évolué en tendant vers un équilibre satisfaisant pour la vie cellulaire dans un milieu donné, entre vitesse et fidélité de traduction. Des isolats naturels ont d'ailleurs des temps de génération, ainsi que des fidélités de traduction très variables (Kurland, 1992). La résolution du conflit entre deux contraintes en opposition, nous donne aujourd'hui à observer une traduction qui semble relativement fidèle, puisqu'on estime à environ 10^{-4} les erreurs faux sens (Parker, 1989). Malgré ce taux d'erreurs relativement faible : la traduction d'une phase ouverte de lecture de 500 acides aminés produit 25% de protéines erronées (mais pas forcément inactives). Pas un seul ribosome (2,5 MDa) de la cellule n'est strictement identique à un autre (Kurland, 1992). Ces exemples montrent l'importance du contrôle de la fidélité de la traduction et ses conséquences potentielles en cascade.

Depuis que le code génétique et son décodage ont été élucidés, les biologistes moléculaires ont trouvé de plus en plus d'exemples de sa non-universalité. Certains codons sont utilisées de façon différente par quelques organismes. Ou encore, des séquences d'ARNm sont capables de subvertir les machines de traduction cellulaires. Certains gènes, des virus aux eucaryotes supérieurs, ont évolué pour exploiter cette plasticité de la traduction afin de réguler leur propre expression génique.

Une faible proportion de gènes, probablement chez tous les organismes, utilise le recodage pour la traduction de leurs ARNm. Le recodage peut impliquer un changement de phase de lecture spécifique de quelques ribosomes, en réponse aux signaux de l'ARNm, ou l'incorporation d'un acide aminé (standard ou non) à la place d'un codon stop. Dans

d'autres cas, les ribosomes ignorent une partie du messager et reprennent leur lecture plus loin. Dans plusieurs cas, le recodage remplit une fonction de régulation. En général le produit issu d'un décodage conventionnel et celui recodé, les deux partageant la même séquence amino-terminale, ont des rôles distincts.

1.2.1 Une alternative ponctuelle au code génétique

1.2.1.1 Le décalage de phase de lecture

Aux emplacements définis de décalage sur l'ARNm, les ribosomes peuvent être programmés pour changer efficacement de phase de lecture (*frameshift*). Généralement, il y a des signaux stimulateurs dans l'ARNm, distincts du site de décalage lui-même, qui augmentent considérablement le niveau de décalage de phase de lecture. En conséquence, les ribosomes débutant au même codon initiateur synthétisent deux produits différents, l'un étant le produit d'une traduction conventionnelle et l'autre le produit d'un événement de décalage. La plupart des exemples connus présente un produit recodé plus long mais ce peut être un biais inhérent, soit au faible nombre d'exemples découverts jusqu'à présent, soit à la plus grande facilité d'identifier un produit plus long qu'un produit plus court, assimilable à un produit de dégradation.

1.2.1.2 La redéfinition du code

Dans la redéfinition, la signification d'un codon est changée d'une manière spécifique et ponctuelle sur un ARNm (par opposition aux réassignations du code génétique « universel » qui sont espèces ou organelles spécifiques). Le décodage d'un acide aminé à la place d'un codon stop (encore appelé *translecture*) a pour conséquence la production d'une protéine plus longue que le produit conventionnel.

Le décodage de la sélénocystéine par des codons UGA spéciaux, ou de la pyrrolysine par UAG, permet l'incorporation des 21ème et 22ème acides aminés, prolongeant de ce fait les possibilités du code génétique.

1.2.1.3 Le saut de ribosome

Lors du saut de ribosome (*hopping*), une fraction des ribosomes *saute* une partie de l'ARNm sans insérer d'acide aminé. Le décodage normal reprend plus loin en aval. Comme avec le décalage de phase et la

redéfinition du code, le saut de ribosome a lieu au sein de séquences définies sur l'ARNm.

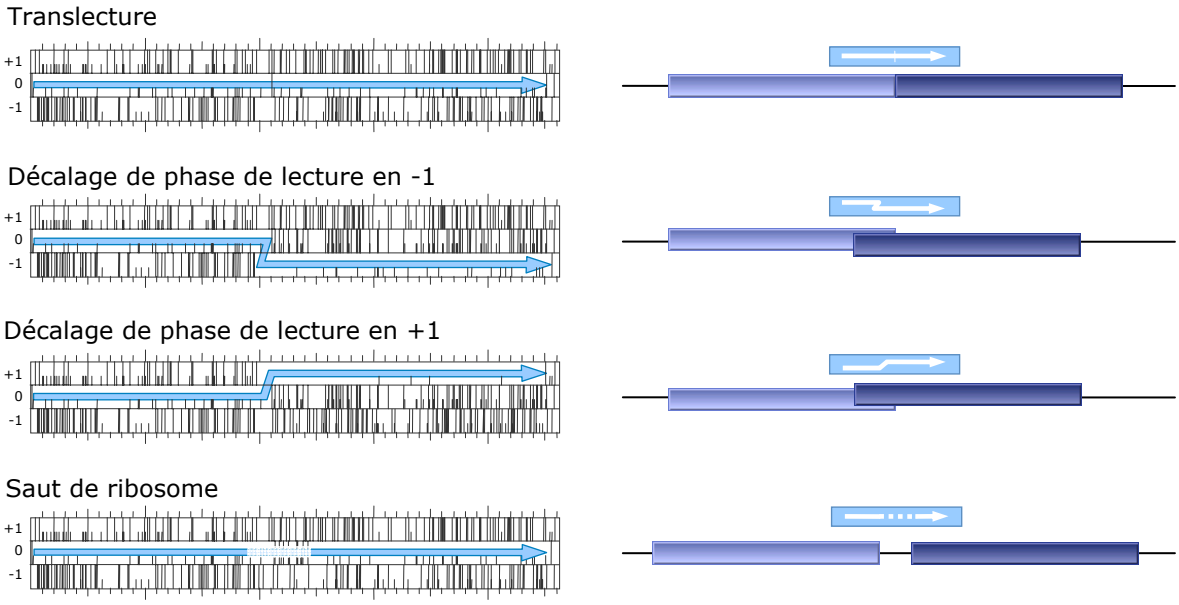


Figure 7 : Différents types de recodage. A gauche sont schématisées les trois phases de lecture. Un trait représente un codon stop, un demi trait un codon d'initiation. A droite une schématisation plus symbolique.

1.2.2 Mécanique du recodage

Les erreurs spontanées (10^{-4} par codon) et les événements de recodage (10^{-2} à 10^{-1} à un site spécifique) se distinguent donc d'un point de vue quantitatif. De plus dans, le cas du recodage, les signaux stimulateurs ont été sélectionnés au cours de l'évolution et le produit synthétisé est utilisé. Ces mécanismes placent ainsi les erreurs spontanées et le recodage dans des catégories qualitativement différentes. L'événement de recodage est toujours en compétition directe avec le décodage conventionnel. En conséquence, les signaux spécifiques de l'ARNm qui perturbent la traduction normale stimulent l'événement de recodage. Les signaux de recodage sont constitués de plusieurs éléments, le plus important étant le site de recodage sur l'ARNm. Des régions de taille variable et d'au moins trois nucléotides sont nécessaires et suffisantes pour l'induction de l'événement de recodage (excepté dans l'incorporation de la sélénocystéine et de le pyrrolysine), mais en général, pour une efficacité significative, des éléments accessoires en *cis* sont nécessaires aussi bien en 5' qu'en 3' du site de recodage. Dans certains cas, les éléments *cis* agissent par leur

séquence primaire, d'autres fois, le stimulateur est une structure secondaire. De nombreux éléments *cis* ont été identifiés, cependant, le mécanisme exact de leur action n'est pas encore compris.

En dépit de leur grande diversité, la majorité des événements de recodage peut être regroupée en deux classes: les décalages de phase par glissement des ARNt en tandem au site A et P (*tandem slippage*) et les événements où n'intervient que le site P du ribosome. Des mécanismes impliquant le site P peuvent être évoqués pour expliquer un grand nombre d'événements de recodage dont le saut de ribosome et le décalage de phase en +1. Ce que tous ces événements, ainsi que la translecture, ont en commun est qu'ils se produisent quand le site P est occupé par un peptidyl ARNt et que le site A adjacent est inoccupé. Un site A vide produit un ralentissement dans le décodage standard, facilitant l'événement de recodage, le plus simple étant le mécanisme de translecture.

1.2.3 Saut de ribosome

1.2.3.1 Site de saut de ribosome

Le saut de ribosome est assimilable à un cas extrême de décalage de phase de lecture. Son mécanisme peut être regardé comme une variation du dérapage au site P ayant pour résultat un décalage de phase de lecture. La majeure partie de notre connaissance du saut de ribosome provient du travail effectué sur le gène 60 du bactériophage T4, qui code une sous-unité de topo-isomérase (Huang *et al.*, 1988; Weiss *et al.*, 1990). A la fin de l'ORF1 du gène 60, la moitié des ribosomes peuvent ignorer 50 nucléotides et reprendre la traduction normale sur l'ORF2. (Maldonado and Herr, 1998). Trois étapes définissent cet événement. Pendant la première étape (décollage, *take-off*), l'ARNt au site P se dissocie du codon GGA. Dans l'étape suivante (balayage, *scanning*), le ribosome traverse la région ignorée jusqu'à ce qu'un codon s'apparie au site P. Lors de la dernière étape (atterrissage, *landing*), l'appariement ARNt / ARNm au site P est rétabli et la traduction conventionnelle reprend. Le codon initialement au site A est un codon stop UAG (décodé lentement). Plusieurs mutants qui réduisent le saut de ribosome du gène 60 ont été isolés et sont généralement situés sur le gène de l'$ARNt^{2}_{Gly}$ qui décode les codons GGA au site P (Herr *et al.*, 1999). Ces mutants semblent parfaitement capables de se dissocier des codons du site P, mais sont incapables de trouver un site d'atterrissage (Herr *et al.*, 2000; Herr *et al.*, 2001).

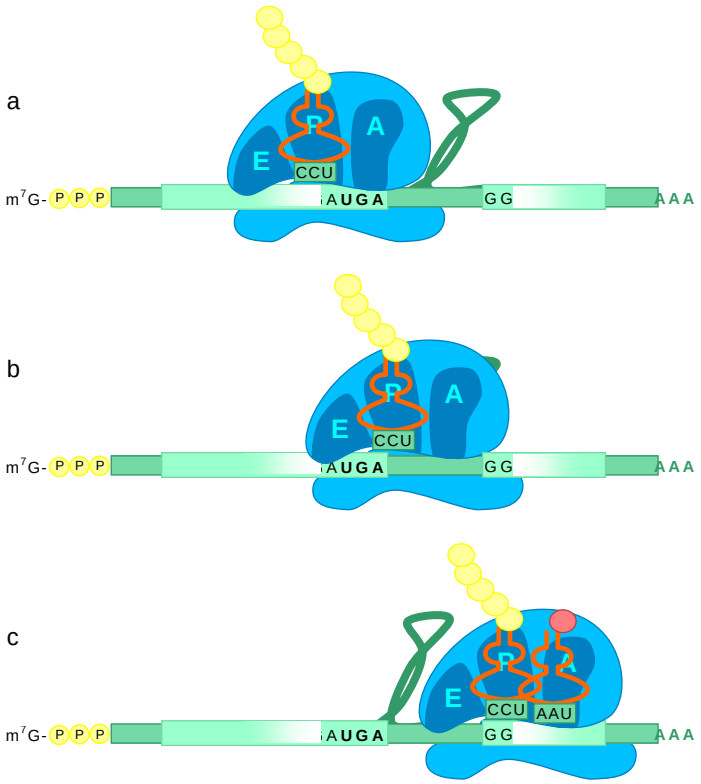


Figure 8 : Saut de ribosome (gène 60). Cinquante nucléotides entre les codons 47 et 48 de la phase codante sont ignorés par la moitié des ribosomes. Un codon stop, inclus dans une tige boucle, est directement situé après le site de décollage. a. décollage ; b. balayage ; c. atterrissage.

1.2.3.2 Séquences *cis* des ARNm stimulant le saut de ribosome

Plusieurs éléments *cis* stimulent le saut de ribosome dans le gène 60. L'un d'eux est une tige-boucle en aval qui recouvre partiellement le site de décollage (Weiss *et al.*, 1990). Il s'avère que cette structure interfère avec le décodage normal du codon stop au site A (Herr *et al.*, 2000). Quand la structure sauvage est substituée par des éléments plus stables, l'efficacité du recodage chute, indiquant que l'énergie d'ouverture de cette structure n'est pas la seule en jeu. La séquence en 5' du codon GAG du site d'atterrissage pourrait aussi fonctionner comme les séquences Shine-Dalgarno (SD) mais ceci n'a pas été encore démontré. La longueur de la région ignorée semble également affecter l'efficacité de recodage mais la raison n'en est pas claire (Herr *et al.*, 2001).

Finalement, un autre élément stimulateur peu commun existe aussi dans le gène 60. Ce stimulateur fonctionne par l'intermédiaire du peptide naissant qu'il code plutôt qu'à travers sa séquence nucléotidique (Weiss *et*

al., 1990). La propriété la plus importante de ce peptide est sa charge positive fournie par plusieurs arginines (Larsen *et al.*, 1995) et son rôle est d'induire la dissociation de l'ARNt et de l'ARNm au site P (Herr *et al.*, 2000; Herr *et al.*, 2001)

1.2.4 Translecture

1.2.4.1 Site de translecture

Avant que les premiers exemples de translecture ne soient connus, les études sur les suppressions de codons non-sens ont préparé le terrain en impliquant la compétition au niveau des codons stop entre l'arrêt de la traduction et l'incorporation d'un acide aminé. Les suppresseurs de non-sens sont des mutants qui permettent la traduction des ARNm contenant un stop (codons non-sens) prématuré dans leur séquence. De tels mutants entrent dans deux catégories. La plupart des suppresseurs ont des mutations dans la boucle de l'anticodon d'un ARNt, permettant à ce dernier de reconnaître le stop et d'introduire un acide aminé (Eggertsson and Soll, 1988; Hatfield *et al.*, 1990). Les suppresseurs non ARNt sont des mutations dans les gènes codant les protéines de terminaison, *prfA* (Oeschger *et al.*, 1980; Zhang *et al.*, 1998) et *prfB* (Karow *et al.*, 1998; Kawakami *et al.*, 1988) chez *E. coli*, *SUP35* (eRF3) et *SUP45* (eRF1) chez *S. cerevisiæ* (Stansfield and Tuite, 1994). Ces mutations induisent une réduction de l'efficacité de terminaison de la traduction, et favorisent donc l'insertion d'un acide aminé en altérant l'équilibre.

Aux sites de translecture, les signaux sur l'ARNm « neutralisent » partiellement l'arrêt de la traduction. Les stimulateurs *cis* connus de la translecture sont situés en aval du site de recodage. Certaines séquences immédiatement en 5' et en 3' d'un codon stop sont le recodage plus favorable que d'autres à une terminaison efficace (Namy *et al.*, 2001; Poole *et al.*, 1998; Tork *et al.*, 2004). Ces « contextes » sont utilisés dans certains sites de translecture afin d'atteindre une efficacité optimale de recodage (McCaughan *et al.*, 1995).

1.2.4.2 Quelques cas avérés

Le premier cas de translecture authentique a été observé chez *E. coli* dans une cellule infectée par le virus Qβ (Weiner and Weber, 1971). Un $ARNt^{Trp}_{CCA}$ suppresseur naturel stimule la translecture sur le codon stop UGA de la protéine d'enveloppe. Ceci conduit à la production d'une protéine longue nécessaire à la formation d'un virus infectieux.

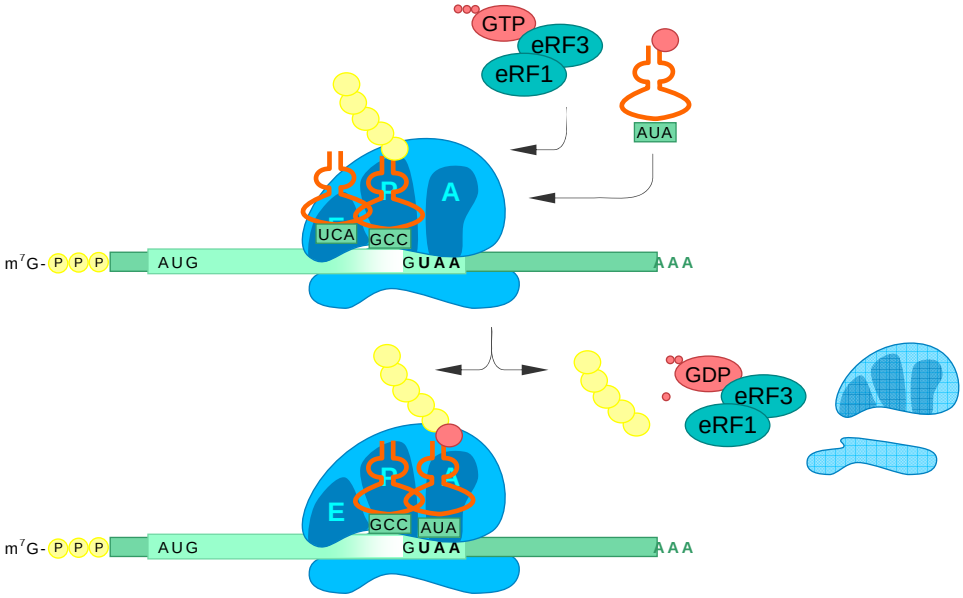


Figure 9 : Translecture.

Des gènes cellulaires qui utilisent un événement de translecture ont été récemment identifiés (Namy *et al.*, 2002). Dans le génome de la levure *S. cerevisiæ,* trois gènes connus possèdent un motif permettant de translecture. Ces gènes sont *PDE2* qui code la phosphodiesterase de l'AMPc, *RCK2* qui décode une protéine kinase et *CST6* codant une protéine impliquée dans la stabilité des chromosomes. Des gènes cellulaires, soumis à la translecture ont été identifiés chez la drosophile. Le gène *oaf* s'exprime dans les cellules pendant l'oogenèse et est nécessaire pour le développement de l'embryon. Il est exprimé dans les gonades de l'embryon, de la larve et des adultes des deux sexes. L'inactivation de ce gène est létale (Bergstrom *et al.*, 1995). Le gène *kelch* est impliqué dans la formation d'œufs viables dans les ovaires de la drosophile. Le gène *hdc* (headcase) est également impliqué dans le développement (Steneberg *et al.*, 1998).

1.2.4.3 Séquences *cis* des ARNm stimulant la translecture

Le contexte nucléotidique localisé en amont et en aval du codon stop et le codon stop lui- même affectent l'efficacité de terminaison de la traduction. Le contexte nucléotidique encadrant le codon stop n'est pas aléatoire. Le nucléotide localisé immédiatement après le codon stop possède un très fort biais qui affecte fortement l'efficacité de terminaison

de la traduction, in vivo chez la bactérie, la levure et l'humain (Bonetti *et al.*, 1995; Major *et al.*, 1996; McCaughan *et al.*, 1995). Il a ainsi été suggéré que le signal de terminaison est un quadruplet et pas un triplet (Shabalina *et al.*, 2004; Tate and Mannering, 1996). Cette hypothèse a été confortée par le pontage du facteur RF2 bactérien avec le nucléotide à la position +4 (Poole *et al.*, 1998). Chez l'homme et la souris, le signal d'arrêt UGAG est le signal le plus abondant tandis que le tétranucléotide UAAA est le moins représenté. Chez la levure *S. cerevisiæ*, il a été observé que les tétranucléotides UAAG et UGAG sont les plus abondants. Le nucléotide +4 affecte la compétition entre la terminaison de la traduction et la translecture, probablement en modulant la reconnaissance du codon stop par le facteur de terminaison ou par un ARNt suppresseur (Bonetti *et al.*, 1995). Pour les codons stop UAA/UAG, il y a un biais du nucléotides +4 (G>U=A>C), alors que le codon stop UGA, le plus efficace, est suivi par le nucléotide U=A>C>G. Chez *E. coli*, les tétranucléotides UAAU et UAAG sont les signaux majeur de terminaison de la traduction, alors que UGAU et UGAA sont les moins communs dans les gènes hautement exprimés (Bossi, 1983; Brown *et al.*, 1990; Tate and Brown, 1992). Ces données révèlent que la base +4, localisée après le codon stop, affecte la fidélité de terminaison de la traduction en favorisant la reconnaissance d'un codon stop efficace.

Le nucléotide qui suit directement le codon stop n'est pas le seul déterminant en 3' du codon stop. D'autres nucléotides situés plus loin en 3' sont aussi impliqués dans la terminaison de la synthèse protéique. Chez le virus de la mosaïque du tabac (TMV), des mutations dans les 6 nucléotides suivant le codon stop influencent l'efficacité de terminaison de la traduction (Skuzeski *et al.*, 1991). Ces mutations ont permis de déterminer le consensus CA[AG][UC][UC]A qui favorise un taux élevé de franchissement du codon stop dans les protoplastes de tabac. A l'inverse, chez *E. coli,* ce consensus permet une bonne efficacité de terminaison de la traduction. Chez la levure *S. cerevisiæ*, l'analyse des trois nucléotides suivant et précédant le codon stop indique un rôle déterminant de ces nucléotides dans la terminaison de la traduction (Bonetti *et al.*, 1995). Dans notre laboratoire, nous avons déjà étudié et identifié le motif localisé en aval du codon stop chez la levure *S. cerevisæ*. Une séquence consensus CA[AG]N[UCG]A est associée à une mauvaise efficacité de terminaison. Cette analyse met en évidence l'importance des nucléotides +8 et +9 dans l'efficacité de terminaison. L'effet du contexte ne dépend ni de la nature de l'acide aminé positionné en +1 du codon stop, ni de l'ARNt associé avec ce codon. Une analyse d'un grand nombre de virus (91 séquences) nécessitant le passage d'un codon stop interne pour produire ses protéines, montre que le contexte 3' du codon stop est très biaisé et module fortement l'efficacité

de terminaison de la traduction (Harrell *et al.*, 2002). Il a été proposé qu'il existe une interaction entre une région spécifique de l'ARNr et le contexte nucléotidique suivant le codon stop. Chez les procaryotes, cette interaction est observée entre le codon stop et le nucléotide C1054 d'hélice 34 de l'ARNr 16S (Prescott *et al.*, 1991). Ceci permet de proposer un modèle d'interaction entre le contexte nucléotidique 3' du codon stop et une des hélices de l'ARNr 18S de la levure *S. cerevisiæ* (Namy *et al.*, 2001). Une complémentarité parfaite est observée entre la séquence 3' du codon stop impliquée dans l'efficacité de terminaison et deux régions de l'ARNr 18S. La première région proposée (479-510) se localise proche de l'hélice 17 dans l'épaule de la petite sous-unité. Cette hélice contrôle le décodage au site A du ribosome procaryote (Van Ryk and Dahlberg, 1995). Un motif déjà identifié comme un stimulateur de décalage du cadre de lecture +1 sur la séquence Ty3 interagit directement avec l'hélice 18 de l'ARNr (Li *et al.*, 2001). La deuxième région (1 305-1 318) se localise dans la région 1 310 de l'ARNr 18S. Cette région se trouve dans la tête de la sous-unité 40S, le domaine le plus flexible du ribosome qui est impliqué dans les mouvements des ARNt et favorise l'approchement des hélices impliquées dans le décodage. Cette interaction proposée entre le contexte 3' du codon stop et les hélices de l'ARNr pourrait induire des modifications structurales dans le ribosome. Cette modification influencerait ensuite l'association du facteur de terminaison de classe 1 avec l'ARNm et le site actif du ribosome en favorisant la translecture.

L'implication du contexte nucléotidique précédant le codon stop dans la fidélité de terminaison de la traduction a également fait l'objet de plusieurs études. L'analyse statistique des 9 nucléotides 5' du codon stop de 748 gènes des plantes révèle que ces nucléotides sont biaisés (Angenon *et al.*, 1990). Les codons GCU (alanine) et UAC (tyrosine) sont sur-représentés à la position -1 du codon stop. D'autre part, il y a sous-représentation de l'arginine et du tryptophane. À la position -2, il y a sur-représentation de la sérine codée par les codons UCU ou UCC. A la position -3, il y a encore sur- représentation de l'alanine. Le biais de ces trois positions résulte probablement de la composition des derniers acides aminés dans la chaîne néosynthétisée (Angenon *et al.*, 1990). De même chez *B. subtilis*, *E. coli*, *S. typhimurium* et *S. cerevisiæ* le contexte situé en 5' du codon stop influence la terminaison de la traduction via les derniers acides aminés de la chaîne polypeptidique. Chez *E. coli,* une corrélation entre l'efficacité de terminaison et la charge de l'acide aminé en position -2 a été observée. Aussi, la présence d'une lysine à la position -1 permet une meilleure efficacité de terminaison, alors que la présence d'une proline ou

d'une thréonine entraîne une mauvaise efficacité de terminaison (Mottagui-Tabar *et al.*, 1994; Mottagui-Tabar and Isaksson, 1997).

Enfin un pseudonœud en 3' est important pour la translecture du gène de Gag du virus de la leucémie murine (MuLV) et une minorité d'autres rétrovirus (ten Dam *et al.*, 1990; Wills *et al.*, 1994). Le pseudonœud naturel du MuLV ne peut pas être substitué par le pseudonœud qui stimule le décalage de phase de lecture en -1 chez le virus de la tumeur mammaire de souris (MMTV ; Wills *et al.*, 1994). Induire une pause n'est vraisemblablement pas la caractéristique principale de ce pseudonœud, il interférerait plus probablement avec la terminaison. Un autre stimulateur de la translecture d'une nature encore incertaine est présent dans le gène d'enveloppe du *Barley Dwarf Yellow Virus PAV*. Dans ce cas, l'élément stimulateur est situé presque 700 bases en 3' du codon stop translu (Brown *et al.*, 1996).

1.2.5 Incorporation de Sélénocystéine

Un exemple remarquable de redéfinition de codon est le décodage de la sélénocystéine. Cette dernière est le 21ème acide aminé incorporé directement aux chaînes polypeptidiques pendant la traduction dans un certain nombre de bactéries, d'archaebactéries et d'eucaryotes. La majorité des sélénoprotéines étudiées sont des enzymes impliquées dans des réactions d'oxydoréductions et contiennent une sélénocystéine dans leur site actif. La sélénocystéine est codée par un codon UGA, qui marque habituellement une fin de traduction (codon opale). Dans chacun des trois ordres, la traduction d'un UGA en sélénocystéine exige la présence de signaux sur l'ARNm (nommés éléments SECIS pour Selenocysteine Insertion Sequence), d'un ARNt portant l'anticodon UCA et chargé d'une sélénocystéine, d'un facteur d'élongation spécifique pour cet ARNt et de plusieurs enzymes essentielles à la biogenèse du Sec-ARNt^{Sec} (Kohrl *et al.*, 2000).

Le mécanisme a été étudié chez *E. coli* et il semble être comparable dans d'autres bactéries. Le codon UGA spécifique des sélénocystéines est suivi d'une structure en tige-boucle dans l'ARNm (Zinoni *et al.*, 1990). Hormis quelques nucléotides conservés dans la boucle apicale et dans le mésappariement (*bulge*) de la tige, seule la structure secondaire est importante (Heider *et al.*, 1992). Les nucléotides conservés sont responsables de l'interaction avec le facteur d'élongation spécifique, *SelB*, qui lie Sec-ARNt^{Sec}. *SelB* est un homologue du facteur d'élongation EF-Tu (Forchhammer *et al.*, 1989) qui amène tous les autres ARNt au ribosome (Forster *et al.*, 1990). A la différence de EF-Tu, SelB possède une extension à son extrémité C-terminale (Kromayer *et al.*, 1996) qui se lie à

la tige-boucle de l'ARNm. L'attachement du complexe SelB avec Sec-ARNt^{Sec} et du GTP à la tige-boucle positionne le complexe à proximité du site A du ribosome et facilite l'incorporation de la sélénocystéine dans la chaîne polypeptidique (Ringquist *et al.*, 1994). Un facteur d'élongation EF^{Sec} a été identifié chez l'archaebactérie *Methanococcus jannaschii* (Rother *et al.*, 2000) et chez les eucaryotes, de *Caenorhabditis elegans* à l'homme (Fagegaltier *et al.*, 2000; Tujebajeva *et al.*, 2000).

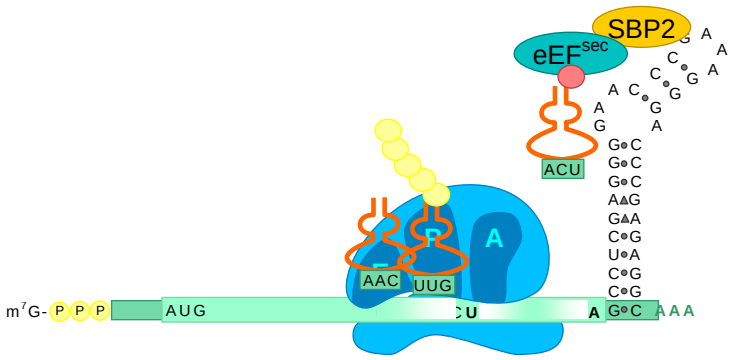


Figure 10 : Incorporation de la sélénocystéine. SBP2 se lie aux éléments de SECIS et recrute EFSec lié au Sec-ARNt^{Sec}.

Les mécanismes d'incorporation retrouvés chez les eucaryotes et les archaebactéries sont plus complexes. Bien que les éléments SECIS aient différentes structures secondaires, ils partagent une propriété commune entre les archaebactéries et les eucaryotes : à la différence de *E. coli*, ils sont placés dans le 3' UTR (Berry *et al.*, 1991; Rother *et al.*, 2001; Wilting *et al.*, 1997).

Comment l'élément SECIS en 3' UTR conditionne-t-il les UGA, parfois plusieurs kilobases en amont, pour incorporer une sélénocystéine ? Plusieurs protéines se liant aux éléments SECIS ont été identifiées (Fujiwara *et al.*, 1999; Hubert *et al.*, 1996), cependant, seule l'une d'entre elles, SBP2, identifiée chez le rat, se lie spécifiquement à un ARN portant SECIS (Copeland *et al.*, 2000). D'autres expériences montrent que SBP2 fait partie d'un complexe ; il a été proposé que SBP2 se lie aux éléments SECIS en 3' UTR et recrute d'autres composants de la machinerie d'insertion de sélénocystéine.

Les éléments SECIS eucaryotes sont de longues tiges-boucles (Berry *et al.*, 1993). Les séquences [AG]UGA et GA respectivement en 5' et 3' de la tige SECIS sont conservés. Il a été proposé que ces séquences forment un quartet de paires non-Watson-Crick, avec les paires en tandem G-A et A-G (Walczak *et al.*, 1996). Un autre élément conservé est une

répétition de deux ou trois adénosines dans la boucle apicale ou le *bulge*. La distance entre cette répétition et le tandem G-A, A-G est toujours de 9 à 11 nucléotides, ce qui correspond approximativement à un tour d'hélice A de l'ARN.

L'analyse de la structure de l'ARN ribosomique a révélé l'existence de paires G-A, A-G semblables à celles observées dans des éléments SECIS. Elles jouent un rôle dans la formation d'une structure commune dans l'ARNr, qui se nomme *kink-turn*, ou *K-turn*, et qui interagit avec les protéines possédant un motif de liaison à l'ARN L7Ae (Klein *et al.*, 2001). SBP2 possède ce motif, et sa mutation abolit toute liaison avec l'élément SECIS (Copeland *et al.*, 2001). Par conséquent, il est possible que les interactions de SBP2 avec SECIS soient semblables aux interactions entre un certain nombre de protéines ribosomiques et l'ARNr.

Une autre question est de savoir si la traduction des ARNm des sélénoprotéines est efficace et si elle est processive. L'efficacité de recodage d'un UGA est d'environ 37% (Berry *et al.*, 1992) et l'introduction d'un deuxième UGA réduit fortement ce niveau (Nasim *et al.*, 2000). Cependant, l'ARNm de la sélénoprotéine P (SelP) contient 10 codons UGA chez l'homme et le rat (Hill *et al.*, 1991; Himeno *et al.*, 1996), 12 chez le bovin, 17 chez le poisson zèbre *Danio rerio* (Kryukov and Gladyshev, 2000; Tujebajeva *et al.*, 2000) et tous semblent être traduits *in vivo* en sélénocystéines. Si l'incorporation de sélénocystéine à chaque UGA est peu efficace, alors les ARNm codant pour des UGA multiples, comme pour SelP, sont peu susceptibles d'être traduits à un niveau mesurable. Par conséquent, ou l'incorporation peut être efficace sur certains UGA uniquement, ou des ribosomes traduisant l'ARNm des sélénoprotéines sont modifiés d'une façon ou d'une autre au démarrage de traduction pour lire UGA comme sélénocystéine au lieu de s'arrêter. Néanmoins, la capacité de tels ribosomes modifiés d'insérer des sélénocystéines dépendrait du contexte nucléotidique des UGA puisque des isoformes multiples de la sélénoprotéine P ont été trouvés et résultent de l'arrêt à un certain UGA (Himeno *et al.*, 1996).

Le contexte nucléotidique du codon UGA est important pour l'insertion efficace de sélénocystéine. Il a été montré que l'identité des deux codons en 5', comme de la base en 3' influencent le rapport entre l'incorporation de sélénocystéine et la terminaison (Grundner-Culemann *et al.*, 2001; Liu *et al.*, 1999; McCaughan *et al.*, 1995). Il est possible que quelques autres éléments inconnus sur l'ARNm affectent l'incorporation de sélénocystéine. Ces éléments sont présents dans les ARNm endogènes pour réaliser une traduction optimale, mais pourraient être absents dans les constructions expérimentales. Tujebajeva et collaborateurs en 2000, ont

réussi à exprimer une sélénoprotéine P recombinante entière (Tujebajeva *et al.*, 2000) ; mais il convient de noter que l'expression a été réalisée seulement dans une des trois lignées de cellules où cela avait été tenté. Par conséquent, il est possible que les expériences dans ces cellules transfectées ne reflètent pas toujours la situation réelle *in vivo*.

1.2.6 Incorporation de Pyrrolysine

La pyrrolysine a été identifié plus d'une décennie après la découverte de la sélénocystéine et constitue un autre exemple de redéfinition de codons. Le codon UAG peut déclencher l'incorporation de pyrrolysine dans le géne de la monométhylamine méthyltransférases chez *Methanosarcina barkeri* (Hao *et al.*, 2002; Srinivasan *et al.*, 2002). Les substrats de cette archaebactérie méthanogène doivent être activés par une méthyltransférase avant de produire du méthane. L'efficacité de décodage de l'UAG du gène *mtmB1* du *M. barkeri* a été estimée à plus de 97%. La présence dans cet organisme d'un $ARNt^{Pyl}$ rare et de la lysyl-$ARNt^{Pyl}$ synthétase (PylS) correspondante (Polycarpo *et al.*, 2003; Srinivasan *et al.*, 2002) plaident pour l'incorporation co-traductionnelle de la pyrrolysine. De plus la présence de structures secondaires conservées, chez trois souches de *Methanosarcina*, 5 à 6 nucléotides en aval de l'UAG pourrait jouer le rôle d'une structure PYLIS (pyrrolysine insertion sequence) équivalente aux élements SECIS des sélénoprotéines (Namy *et al.*, 2004).

Il a été établi que la L-pyrrolysine synthétique est attaché à partir de molécules libres sur l'$ARNt^{Pyl}$ par la proteine PylS. Celle-ci active la pyrrolysine avec l'adénosine triphosphate et lie la pyrrolysine à $ARNt^{Pyl}$ *in vitro* dans les réactions spécifiques à la pyrrolysine. De plus, l'addition de pyrrolysine à des *E. coli* exprimant l'$ARNt^{Pyl}$ et l'ARNm de pylS, permet la traduction *in vivo* du gène *mtmB1* avec l'incorporation à la place du codon UAG de pyrrolysine (Blight *et al.*, 2004).

1.2.7 Décalage de phase de lecture en +1

1.2.7.1 Site de décalage de phase de lecture en +1

Peu de cas de gènes chromosomiques impliquant un décalage de phase de lecture en +1 sont connus et la plupart s'expliquent par un mécanisme impliquant le site P. Ceci inclut le gène *prfB* (RF2) de *E. coli* (Craigen and Caskey, 1986; Major *et al.*, 1996), *EST3* et *ABP140* de *S. cerevisiæ* (Asakura *et al.*, 1998; Morris and Lundblad, 1997), et l'antizyme chez les métazoaires (Howard *et al.*, 2001; Matsufuji *et al.*, 1995). Les

données initiales proviennent de l'analyse du décalage de phase de lecture de *prfB*. Comme discuté ci-dessus, le décalage se produit sur le dernier codon sens de l'ORF1, sur la séquence CUU-U. Le dernier U appartient au codon UGA. Le décalage se produit quand le peptidyl-ARNt glisse de CUU à UUU dans la phase +1. La stabilité de l'appariement à cette nouvelle position est cruciale (Curran, 1993). Une pause en phase 0, lorsque le site A est vide, est également essentielle car la quantité de facteurs de terminaison module directement le taux de recodage. Un niveau élevé de RF2, qui augmente l'efficacité de terminaison, réduit le temps de pause sur le codon stop UGA et ainsi le niveau de décalage de phase de lecture. L'inverse est également vrai : un faible niveau de RF2, réduit l'efficacité de terminaison, augmente le temps de pause et accroît le taux de translecture.

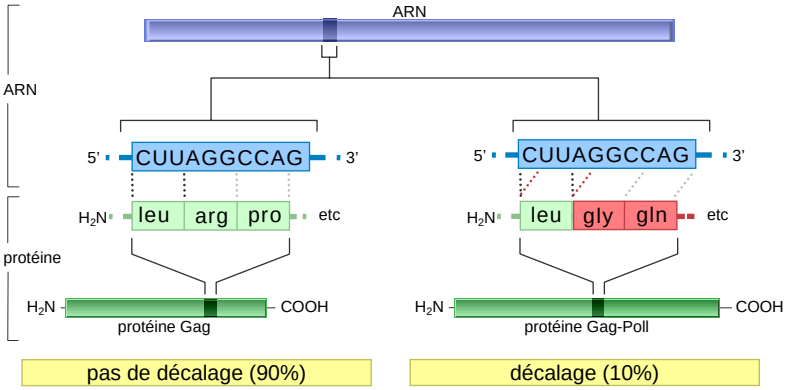


Figure 11 : Décalage de phase de lecture +1. Exemple de Ty1.

Le décalage de phase de lecture se produisant pendant la traduction de Gag-Pol du rétrotransposon Ty1 de levure implique, lui, le site P (Belcourt and Farabaugh, 1990; Weiss and Gallant, 1983). L'événement de décalage de phase en +1 se produit sur la séquence CUU-AGG-C. L'ARNt décodant CUU est au site P et le site A est alors vide. Là encore, le peptidyl-ARNt glisse en aval d'un nucléotide sur l'ARNm pour former deux appariements avec le codon UUA. À la différence de *prfB*, la pause est induite par un codon sens « affamé » plutôt qu'un codon stop. L'ARNt qui décode AGG chez *S. cerevisiæ* est peu efficace, alors que le codon AGG lui-même n'est que faiblement sous-représenté. Ceci mène à un ralentissement du décodage. La surexpression de l'ARNt reconnaissant le codon AGG réduit de manière significative le niveau de décalage de phase de lecture du Ty1 (Belcourt and Farabaugh, 1990) et sa délétion l'augmente (Kawakami *et al.*, 1993).

Un mécanisme différent a été proposé pour Ty3 (Farabaugh *et al.*, 1993). Le décalage de phase de lecture en +1 se produit au niveau de la séquence GCG-AGU-U. Le premier codon (GCG) est situé dans le site P et le site A (AGU) est vide. C'est la pause du ribosome sur le codon rare AGU qui est ici importante; cependant le mécanisme du décalage lui-même a été interprété différemment. L'ARNt cognat qui décode le codon GCG ne peut pas former d'appariement classique avec le codon en phase +1 (CGA) et un glissement de l'ARNt ne semble pas impliqué. Il a été proposé que l'ARNt au site P interfère d'une façon ou d'une autre avec l'appariement normal au site A, ce qui permettrait à des ARNt d'entrer en contact avec le codon en phase +1 GUU. Cette hypothèse est soutenue par le fait que la surexpression de l'ARNt décodant GUU accroît le niveau de décalage de phase de lecture en +1 (Pande *et al.*, 1995).

Cependant, plusieurs résultats récents laissent à penser que ces deux exemples de décalage de phase de lecture (Ty1 et Ty3) font en fait appel aux mêmes mécanismes et suggèrent que l'appariement codon / anticodon au site P est proche-cognat (Sundararajan *et al.*, 1999). La surexpression des gènes des ARNt au site P qui utilisent un appariement cognat avec les codons impliqués, réduit nettement le taux de décalage de phase de lecture en +1 des Ty1 et Ty3. En supprimant les ARNt cognats du site sur un site Ty1 modifié, le décalage est sensiblement accru. Pour le Ty1, ces résultats s'expliquent par un contact ARNt / ARNm instable, combiné avec une réduction de la vitesse du ribosome sur les codons « affamés » au site A, conduisant à un décalage de l'ARNt. De ces expériences, il a été proposé que l'ARNt situé au site A ne se dissocie pas lors du décalage de phase du Ty3 mais que la nature de l'interaction codon / anticodon au site P influence la base immédiatement en 3' du codon, de sorte qu'elle soit indisponible pour un appariement avec l'ARNt au site A (Farabaugh *et al.*, 1993; Sundararajan *et al.*, 1999). Au site A, les bases accessibles (2,3 et 4) sont alors dans la phase +1. Si cette hypothèse est juste, alors un seul type de modèle de dissociation / re-appariement peut s'appliquer à Ty1 et Ty3, et à tous les autres cas connus de décalage de phase de lecture en +1.

1.2.7.2 Séquences *cis* des ARNm stimulant le décalage en +1

Les stimulateurs en *cis* du décalage de phase de lecture en +1 présentent une grande variété. Un stimulateur en 5' du site de recodage peut être une séquence SD, comme c'est le cas en amont du site de décalage du gène *prfB*, premier exemple d'un stimulateur 5' (Weiss *et al.*, 1988). Cette séquence est située 3 nucléotides en amont du site de décalage

CUU-U (beaucoup plus près que pour une stimulation de décalage de phase de lecture en -1). La distance est cruciale. Déplacer la séquence SD d'une base réduit considérablement le décalage de phase de lecture en +1. Dans ce cas-ci, on pense que la distance réduite entre la séquence SD et le site de recodage est optimale pour physiquement « pousser » le ribosome situé sur le site de décalage dans la phase +1 pas l'intermédiaire de l'ARNt du site E (Marquez *et al.*, 2002). Un autre stimulateur en 5' est une séquence de 40 à 50 nucléotides placés en amont des gènes 1 et 2 de l'antizyme du rat qui stimule le décalage de phase d'un facteur 2,5 (Ivanov *et al.*, 2000; Matsufuji *et al.*, 1996).

Au moins deux type de stimulateurs participent à cette régulation en 3'. Le plus étudié est un pseudonœud dans l'ARNm, quelques nucléotides en aval du site de décalage (Matsufuji *et al.*, 1995). Il existe deux versions de ce stimulateur dans les orthologues de l'antizyme 1 et les orthologues de l'antizyme 2 des vertébrés. Comme avec les pseudonœuds qui interviennent dans le décalage de phase de lecture en -1, la distance au site de recodage est importante (voir ci-dessous). En outre, chez *S. cerevisiæ*, où la séquence d'antizyme mammifère produit généralement un décalage de phase de lecture en -2, le déplacement du pseudonœud de trois nucléotides en aval a pour conséquence une augmentation importante de la proportion de ribosomes décalant en phase +1 (Matsufuji *et al.*, 1996). Ceci montre encore une fois que les pseudonœuds ARN font bien plus que ralentir les ribosomes sur le site de recodage et exerce un rôle supplémentaire. Le stimulateur endogène de l'antizyme de *Schizosaccharomyces pombe* n'est pas un pseudonœud reconnaissable et, bien que sa nature soit obscure, il est susceptible d'être un nouveau type d'élément stimulateur (Ivanov *et al.*, 2000).

Un autre stimulateur intéressant est une séquence de 7 à 15 nucléotides immédiatement en aval du site de décalage du Ty3 qui stimule recodage d'environ 7 fois (Farabaugh *et al.*, 1993). Cette séquence pourrait interférer avec le site A en formant un appariement avec l'hélice 18 de l'ARNr 18S (Li *et al.*, 2001).

1.2.8 Décalage de phase de lecture en -1

La plupart des exemples de décalage de phase de lecture programmée qui ont été identifiés sont des événements de décalage en -1. Les exemples connus sont observés en général à la jonction entre les gènes *gag* et *pol*, ou leurs équivalents, chez les retrovirus, des coronavirus, des virus de plantes (Brierley, 1995; Farabaugh, 1996), des éléments insertionnels (IS) bactériens, des bactériophages (Atkins *et al.*, 1979; Dunn and Studier, 1983) et quelques gènes cellulaires (Baranov *et al.*, 2001;

Namy *et al.*, 2002; Namy *et al.*, 2004). Jacks et collaborateurs en 1988 ont proposé le premier modèle de glissement en tandem pour expliquer le décalage de phase de lecture en -1 du virus du sarcome de Rous (Jacks *et al.*, 1988). Selon ce modèle, le déphasage se produit sur un heptamère glissant en phase 0, X-XXY-YYZ. Lorsque le ribosome arrive sur cette séquence, les ARNt glissent simultanément d'un nucléotide en arrière.

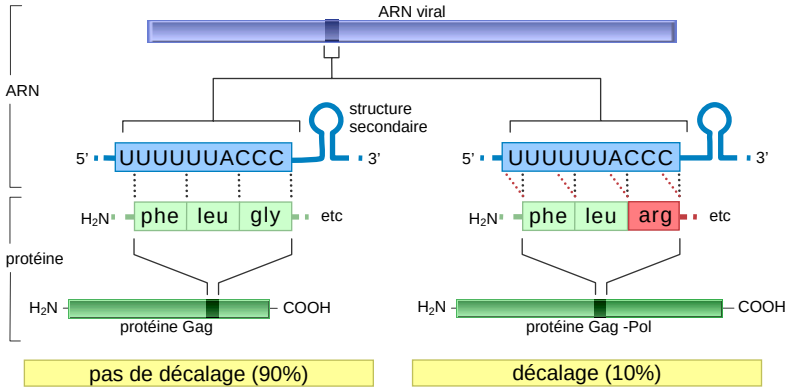


Figure 12 : Un décalage de phase de lecture -1 est nécessaire à la production de la transcriptase inverse et de l'intégrase du HIV-1. La transcriptase inverse et l'intégrase sont produites par clivage d'une protéine de fusion Gag-Pol, tandis que les protéines de la capside sont produites par le clivage de la protéine plus abondante Gag.

1.2.9 Quelques exemples

Le recodage représente un grand potentiel pour la régulation de l'expression génique, qui est activement exploité par divers organismes (Gesteland and Atkins, 1996). La capacité de faire deux protéines ou plus à partir d'un même ARNm est parfois très utile, particulièrement pour les organismes à génome compact. Il n'est alors pas étonnant que la plupart des exemples de recodage aient été trouvés chez les virus et les éléments transposables. En outre, la production de deux protéines à partir du même ARNm permet l'établissement d'un rapport stœchiométrique précis. Soit les événements de recodage permettent de maintenir une proportion relative entre formes recodée et non recodée invariable, soit ils maintiennent constante une quantité absolue de la forme recodée (voir plus bas).

Chez la plupart des virus et éléments rétrotransposables, les événements de recodage associent des polypeptides structuraux (par exemple la structure rétrovirale Gag) et catalytiques (par exemple, la polymérase Pol ou la polyprotéine protéase-polymérase, Pro-Pol) (Brierley, 1995; Farabaugh, 1996). La traduction conventionnelle de l'ARN

génomique des rétrovirus ne produit que la protéine Gag (dont l'ORF est localisée en 5' sur l'ARNm), alors que Gag-Pol (ou Gag-Pro-Pol) est produit sous la forme d'un polypeptide unique par recodage (Jacks *et al.*, 1987). Les virus ont besoin de beaucoup plus de molécules de protéines structurales que catalytiques. Pour cette raison, dans la plupart des cas, les taux de recodage sont de l'ordre de 1 à 10% en comparaison du décodage conventionnel. Ainsi la plupart des ribosomes synthétisent uniquement les sous-unités structurales, tandis qu'une minorité, ayant réalisé l'événement de recodage, synthétise les polyprotéines porteuses des activités catalytiques. La fréquence avec laquelle les événements de recodage se produisent change d'un virus à l'autre. L'ARNm du *Human immunodeficiency virus 1* (HIV1) induit un décalage de phase de lecture en -1 de 5% dans des cellules de mammifères (Bidou *et al.*, 1997). Le décalage de phase de lecture en +1 sur l'ARNm du rétrotransposon Ty3 de *S. cerevisiæ* présente, lui, un taux de 11% (Farabaugh *et al.*, 1993). Dans chaque cas, l'efficacité du recodage a vraisemblablement évolué pour convenir au *style de vie* du virus ou de l'élément rétro transposable. Les expériences ont montré que le rapport entre Gag et Gag-Pol, donc l'efficacité du recodage, peut être crucial pour la propagation des rétrovirus et rétrotransposons. Un rapport accru ou diminué, dû à des changements dans l'efficacité de recodage, entrave l'assemblage des particules virales et le caractère infectieux des virus L-A et HIV (Dinman and Wickner, 1992; Hung *et al.*, 1998; Ma *et al.*, 2002; Shehu-Xhilaga *et al.*, 2001). Cela mène également à une transposition réduite de Ty3 (Kawakami *et al.*, 1993).

Le gène de *dnaX* de *E. coli* est autre exemple d'événement de recodage (décalage de phase de lecture en -1) qui présente un rapport fixe entre les deux polypeptides codés, les sous-unités τ et γ de l'ADN polymérase III (Flower and McHenry, 1990; Tsuchihashi and Kornberg, 1990). Ces deux sous-unités sont présentes avec un rapport 1:1 dans l'holoenzyme. La forme longue τ est synthétisée par traduction conventionnelle. La forme courte γ partage la même séquence d'acides aminés que celle des deux premiers tiers de τ, puis ne possède qu'un seul acide aminé additionnel à l'extrémité carboxylique. Sous l'influence des signaux stimulateurs de l'ARNm (Larsen *et al.*, 1997; Larsen *et al.*, 1994) 50% des ribosomes reculent d'un nucléotide (glissent vers la phase de lecture -1) sur le motif glissant A-AAA-AAG, situé aux deux tiers de l'ORF. Les ribosomes qui changent de phase décodent ensuite un seul codon dans la nouvelle phase avant de s'arrêter sur un codon stop (UGA). Ils synthétisent ainsi la sous-unité γ (Flower and McHenry, 1990; Tsuchihashi and Kornberg, 1990) qui est amputée des deux domaines protéiques de τ leur permettant de se lier à la protéine DnaB ainsi qu'à la sous-unité *αεΘ* de

la polymérase III (Gao and McHenry, 2001). Les deux polypeptides ont donc différentes activités biochimiques et les deux sous-unités jouent différents rôles dans l'ADN polymérase III.

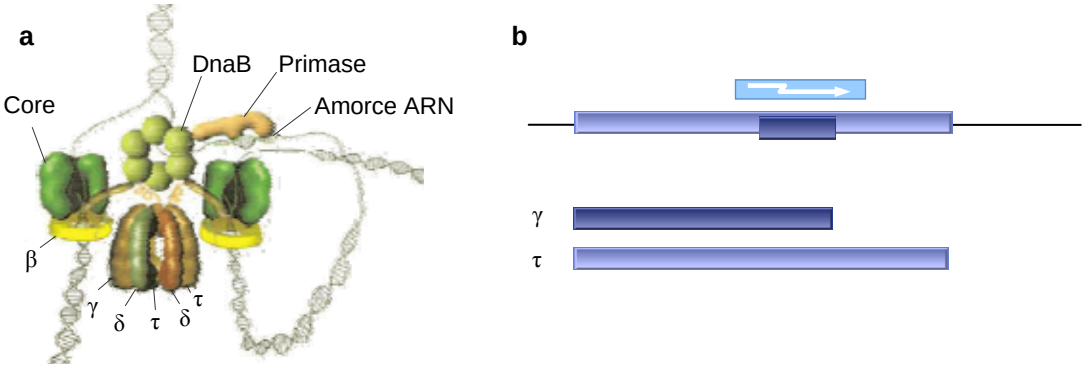


Figure 13 : Exemple de décalage de phase de lecture en -1. a. Organisation structurale de l'ADN polymérase III. Les sous-unité τ (traduction conventionnelle) et γ (recodage) sont deux produits du gène *dnaX* (Namy *et al.*, 2004) ; b. Organisation génomique des deux phases de lecture de d*naX*.

De nombreuses conditions physiologiques peuvent changer l'activité de l'appareil traductionnel. En conséquence, quelques gènes ont évolué pour permettre à un événement de recodage d'être utilisé comme voie d'autorégulation affectant la traduction elle-même. L'exemple le mieux étudié de cette régulation potentielle du recodage est le décalage de phase de lecture en +1 nécessaire à l'expression du facteur RF2 de *E. coli* (Craigen and Caskey, 1986; Major *et al.*, 1996). RF2 est un facteur de terminaison de la traduction permettant l'identification des codons stop UGA et UAA. La protéine RF2 est codée par deux phases de lecture partiellement chevauchantes. La première code seulement un peptide de 25 acides aminés, sans fonction biochimique connue, qui est rapidement dégradé. Un décalage de phase de lecture en +1 au niveau du dernier codon sens de la première ORF sur la séquence CUU-U, produit la protéine RF2 complète. Le codon stop de l'ORF1 est un UGA, qui n'est reconnu que par RF2; une pause du ribosome à cet emplacement, provoquée par un faible niveau de RF2, contribue au décalage de phase, et augmente la synthèse de RF2, fermant la boucle d'autorégulation. De cette façon, le site de décalage de phase perçoit le niveau de RF2 dans la cellule. Une analyse des séquences disponibles du gène *prfB* a montré que ce mécanisme est utilisé par un grand nombre de bactéries (Baranov *et al.*, 2002).

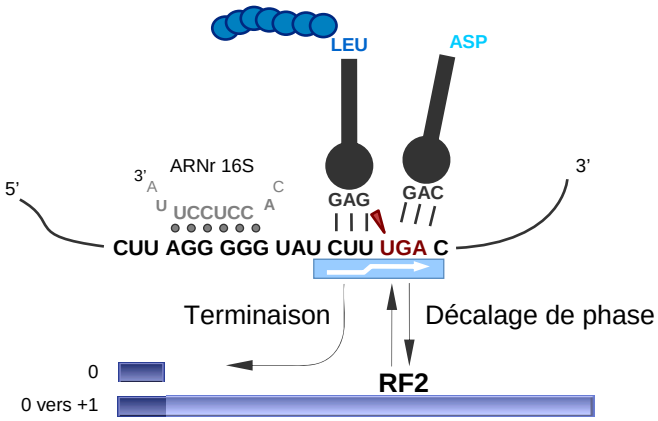


Figure 14 : Exemple de décalage de phase de lecture en +1. Autorégulation de RF2 chez *E. coli*.

L'antizyme de l'ornithine décarboxylase a initialement été défini comme une activité biochimique qui inhibe l'ornithine décarboxylase (ODC) et est induite en présence de concentrations élevées de polyamines. La décarboxylation de l'ornithine est la première étape limitante dans la biosynthèse des polyamines dans la cellule. Les polyamines ont diverses activités biologiques documentées, dont celle d'affecter le taux et la fidélité de la traduction (Atkins *et al.*, 1975). Le clonage du premier gène de l'antizyme chez les mammifères (rat) a révélé que l'ORF qui codait la partie biochimiquement fonctionnelle de la protéine (ORF2) débutait sa traduction en 75 nucléotides amont du premier codon AUG (Miyazaki *et al.*, 1992). En fait, la traduction débute dans une ORF partiellement chevauchante en amont (ORF1) et un décalage de phase de lecture en +1 produit la protéine entière. Des expériences *in vivo* ont montré qu'un niveau élevé de polyamines augmentait le taux de décalage en +1, formant ainsi une boucle d'autorégulation (Howard *et al.*, 2001; Matsufuji *et al.*, 1995). Ce mécanisme a évolué pendant plus d'un milliard d'années chez l'ancêtre commun des mycètes, des échinodermes, des nématodes, des insectes, et des vertébrés. Il est conservé chez tous les descendants qui sont connus pour posséder un homologue de l'antizyme (Ivanov *et al.*, 2000).

1.3 Le décalage de phase de lecture en -1

Le décalage programmé du ribosome d'un nucléotide en amont est apparemment le mécanisme de recodage le plus conservé et le plus répandu dans la nature. Bien que les évènements de décalage de phase de lecture en -1 soient rares, certains virus, possèdent deux sites de décalage successifs. C'est le cas du *Mouse mammary tumor virus* (MMTV) pour lequel un ribosome qui commence la traduction au codon d'initiation de l'ORF *gag*

synthétise la protéine Gag-Pro-Pol avec une efficacité de 2%. Après avoir subi un premier décalage à la jonction Gag-Pro, le ribosome en subit un second à la jonction Pro-Pol (Jacks *et al.*, 1987).

Parmi tous les sites de décalage de phase de lecture -1 fonctionnels chez les eucaryotes, deux points communs ont été identifiés : une séquence glissante et, plus en aval, une structure secondaire.

1.3.1 Séquence glissante

Selon le modèle proposé par Jacks (Jacks *et al.*, 1988), le déphasage se produit sur un heptamère glissant en phase 0, `X-XXY-YYZ` (où `X` pourrait être n'importe quel nucléotide, `Y` est un nucléotide à appariement faible et `Z` est espèce spécifique). Les ARNt qui sont aux sites P et A (décodant `XXY` et `YYZ` respectivement) glissent simultanément d'un nucléotide en arrière (`XXX-YYY`). Dans la nouvelle phase, au moins deux appariements sur trois par ARNt sont préservés avec l'ARNm (dans certains cas l'appariement dans la nouvelle phase est plus fort que dans la phase d'origine). L'élément important de cette séquence est sa nature répétitive. Une mutation qui perturbe les répétitions `XXX` ou `YYY` diminue sévèrement l'efficacité du déphasage. L'utilisation d'appariements forts dans la répétition `YYY` réduit aussi le déphasage (Brierley *et al.*, 1992; Dinman *et al.*, 1991). Cependant, des variations naturelles peuvent intervenir sur la répétition `XXX` tout en restant compatibles avec un décalage de phase efficace. Il existe, en effet, des séquences glissantes de la forme `N-XXY-YYZ` C'est le cas de l'*Equine arthritis virus* (EAV ; den Boon *et al.*, 1991). Un tel polymorphisme n'existe pas sur la répétition `YYY`.

Une amélioration du modèle proposé par Weiss en 1989 a postulé que le décalage avait lieu après le transfert du polypeptide (expliquant pourquoi l'ARNt au site A, a inséré son acide aminé dans le produit final) mais avant, ou plus probablement pendant, la translocation (Weiss *et al.*, 1989). Vraisemblablement, quand le ribosome est dans cet état et quand l'appariement entre l'ARNt du site A et le messager est faible, les ARNt aux sites P et A peuvent glisser simultanément en arrière d'un nucléotide sur l'ARNm (Kim *et al.*, 2001; Plant *et al.*, 2003). L'appariement plus favorable après décalage bloque le ribosome dans cette nouvelle phase.

1.3.2 Structure secondaire

La séquence glissante est nécessaire mais non suffisante pour rendre compte des taux élevés de décalage de phase en -1 mesurés. Dans tous les exemples publiés, il existe une séquence nucléotidique capable d'adopter une structure secondaire en aval de l'heptamère. Cette structure

secondaire est invoquée pour expliquer ces taux de décalage de phase élevés. L'influence de la structure secondaire a été vérifiée par des mutations ponctuelles et des délétions. Dans le cas de l'*Infectious bronchitis virus* (IBV), une délétion de cette structure abolit pratiquement le décalage de phase *in vitro* (2% à 3% de décalage de cadre de lecture, par rapport aux 25 % du sauvage) (Brierley *et al.*, 1989; Brierley *et al.*, 1991). Cet ARNm est en effet capable de prendre une conformation tertiaire complexe, soit sous la forme d'une tige-boucle, soit, plus généralement, sous la forme d'un pseudonœud de type H (H pour *hairpin-like*) : deux tiges coaxiales *empilées* reliées par une boucle (Dam *et al.*, 1992).

Les données cristallographiques obtenues sur le pseudonœud du virus *Beet western yellows virus* (BWYV), définissent cette structure dans un cube de 3 nm de côté (Su *et al.*, 1999) alors que le ribosome, lui, s'inscrit dans un cube de 20 nm de côté. Il semble qu'un pseudonœud stimule le décalage de phase plus qu'une tige-boucle, à stabilités thermodynamiques théoriques égales (Brierley *et al.*, 1991; Chamorro *et al.*, 1992).

L'existence de ces structures secondaires et leur implication dans le phénomène du décalage de phase en -1 ont été vérifiées par des mutations ponctuelles ou compensatoires dans plusieurs systèmes : IBV (Brierley *et al.*, 1989; Brierley *et al.*, 1991), L-A (Dinman *et al.*, 1991), MMTV (Chamorro *et al.*, 1992), *Feline Immunodeficiency Virus* (FIV) (Morikawa and Bishop, 1992), *Simian Retrovirus* (SRV1) (Dam *et al.*, 1992; ten Dam *et al.*, 1995) et HIV (Baril *et al.*, 2003; Bidou *et al.*, 1997; Dulude *et al.*, 2002; Leger *et al.*, 2004). Dans la majorité des mutants analysés, on peut corréler le taux de décalage de phase avec la présence d'une structure secondaire. Cependant, des mutants compensatoires ne restaurent pas toujours un taux de décalage de phase identique au sauvage (Brierley *et al.*, 1991; Chamorro *et al.*, 1992; Jacks *et al.*, 1988; ten Dam *et al.*, 1994; ten Dam *et al.*, 1995). De plus cette corrélation entre la stabilité de la structure secondaire et le taux de décalage de phase n'est pas toujours évidente (Bidou *et al.*, 1997; Brierley *et al.*, 1989; Chen *et al.*, 1995; Kollmus *et al.*, 1996; Kollmus *et al.*, 1994). Certains décalages de cadre de lecture semblent se faire sans structure secondaire (ten Dam *et al.*, 1990), et à l'inverse, quelques structures secondaires caractérisées s'avèrent beaucoup plus complexes qu'un simple pseudonœud, telles les « oreilles de lapin », les double tige-boucles, les « tiges boucles s'embrassant » (*kissing loop*) (Farabaugh, 1996) ou les interactions triples (Su *et al.*, 1999).

D'autres expériences menées pour analyser la structure spatiale des pseudonœuds par des sondes chimiques ont également suggéré que la stabilité thermodynamique des structures secondaires n'est pas seule en jeu.

L'étude d'un variant du pseudonœud du MMTV, permettant un fort taux de décalage de phase en -1, a montré une structure coudée. Une adénine à la jonction entre les deux tiges empêche celles-ci de se positionner de façon coaxiale. Ce type de structure est prédit pour être moins stable que deux tiges empilées coaxialement, et pourtant le décalage de phase en -1 est mieux stimulé par ce type de pseudonœud (Chen *et al.*, 1995). Les auteurs ont confirmé ces prédictions en étudiant précisément la structure tridimensionnelle de différents pseudonœuds par Résonance Magnétique Nucléaire (RMN) (Chen *et al.*, 1995; Sung and Kang, 1998). La stabilité thermodynamique de la structure secondaire ne semble donc pas être le seul déterminant de forts taux de décalage de phase en -1. Plus récemment, il a été montré que des liaisons hydrogène pouvaient se former entre les sucres des nucléotides, stabilisant la structure de manière non prédictible (Giedroc *et al.*, 2003). De même, la présence d'ions métalliques permettant de compenser les charges des groupements phosphates, tend à rendre la structure plus compacte (Csaszar *et al.*, 2001; Egli *et al.*, 2002).

1.3.3 Distance heptamère/structure secondaire

Des délétions de la région entre la séquence glissante et la structure secondaire ont montré que l'on ne peut pas ou peu faire varier la distance entre les deux éléments *cis* sans affecter fortement le décalage de phase en -1. Bien que la taille de cet espaceur varie d'un virus à l'autre, cet élément semble optimisé : la délétion, ou insertion, d'une base ou deux réduit, ou élimine, complètement la capacité de décalage (Brierley *et al.*, 1992).

En ce qui concerne l'IBV, l'ajout d'un codon aux 5 nucléotides de l'espaceur abolit le décalage de phase; un codon en moins le fait baisser d'un facteur 10 par rapport à la situation sauvage (Brierley *et al.*, 1989). Cette distance semble donc déterminée pour une cible particulière, mais par contre elle est variable en fonction des cibles étudiées : de 3 à 14 nucléotides en fonction des virus (Brierley, 1995; ten Dam *et al.*, 1990). Si les différents pseudonœuds sont capables de prendre des formes différentes, plus ou moins coudées, cela pourrait expliquer que la distance optimale varie d'une cible à l'autre.

1.3.4 Pause du ribosome

Pour expliquer le rôle de la structure secondaire et l'importance de l'espacement entre la séquence glissante et la structure secondaire, il a été invoqué un système de barrière physique, énergétique, précisément placée, où le ribosome viendrait buter, permettant aux ARNt présents aux sites A et

P de se recaler un nucléotide en arrière (des modèles plus précis sont discutés à la fin de ce paragraphe). Les premières preuves expérimentales de cette pause du ribosome ont été obtenues durant l'étude *in vitro* du décalage de phase en -1 nécessaire à l'expression de la sous-unité γ de l'ADN polymérase III de *E. coli* sur le messager du gène *dnaX*. Lors d'expériences de traduction *in vitro*, il était apparu des intermédiaires dont la taille correspondait exactement à celle obtenue si la traduction était bloquée juste en amont de la structure secondaire. Ces intermédiaires sont détectés quand la traduction est interrompue, mais disparaissent si on laisse la réaction se poursuivre. L'auteur a interprété ce résultat comme un arrêt transitoire, une *pause* du ribosome en cours de traduction au niveau de la structure secondaire (Tsuchihashi, 1991).

Cette pause du ribosome a été étudiée de manière approfondie en ce qui concerne l'IBV (Somogyi *et al.*, 1993) et le L-A (Tu *et al.*, 1992). Les deux approches utilisent une technique différente, et apportent des conclusions complémentaires. Les travaux sur l'IBV ont été réalisés de manière similaire à celle évoquée ci-dessus pour *dnaX* : les auteurs ont réalisé des expériences de traduction *in vitro* sur des ARNm ne comportant pas la séquence glissante. Des intermédiaires de traduction sont visualisables à 26ºC, et en présence d'édéine (utilisée pour synchroniser l'initiation de la traduction). L'un des intermédiaires de traduction, de taille conforme à un arrêt juste en amont du pseudonœud, est dépendant de la présence de cette structure. De plus, les auteurs ont fait varier les conditions expérimentales afin de modifier la stabilité de la structure secondaire : une hausse de la température de 26°C à 35ºC diminue le taux de produit intermédiaire, alors qu'une hausse de la concentration en Mg^{2+} de 1,5 à 2 mM diminue la vitesse de traduction, mais augmente la quantité de produit de traduction résultant d'une pause du ribosome. Il y a une corrélation entre stabilité du pseudonœud et fréquence/durée de la pause (Somogyi *et al.*, 1993).

Les travaux réalisés sur le virus L-A consistent à visualiser des fragments d'ARN protégés de la digestion nucléasique micrococcale par les ribosomes présents sur le messager (Wolin and Walter, 1988). Cette technique permet de cartographier très précisément la zone de protection : *in vitro*, il y a protection préférentielle de la zone située juste en amont du pseudonœud, à deux positions séparées de 3 nucléotides. Les auteurs proposent que ces deux positions correspondent à une pause avant et après le décalage de phase et invoquent pour expliquer que ces deux pauses soient séparées par 3 nucléotides, et non un seul, que la nucléase agit à des sites préférentiels (Tu *et al.*, 1992). Aucune expérience n'a été réalisé pour

déterminer si une de ces deux positions était dépendante de la présence de l'heptamère, et donc du décalage de phase.

Pour les travaux évoqués ci-dessus, les structures secondaires étudiées (sauvages, mutants ponctuels, compensatoires, délétions) montrent une corrélation entre pause du ribosome et décalage de phase en -1, même s'il n'y a pas forcément causalité. Mais la corrélation inverse n'est pas absolue : quelques structures secondaires provoquent une pause (parfois de moindre importance), bien que le décalage de phase ne soit plus du tout décelable. Par exemple, une délétion de la seconde partie du pseudonœud de l'IBV, laissant seule une tige-boucle, provoque quand même une pause, mais il n'y a plus décalage de phase (Somogyi *et al.*, 1993). De même, la délétion de 2 bases dans la deuxième tige du pseudonœud du L-A ne change pas l'empreinte observée, alors qu'on ne détecte plus du tout de décalage de phase (Tu *et al.*, 1992). Il peut donc y avoir pause sans décalage de phase. De même la structure secondaire seule n'assure pas le décalage de phase de lecture, la structure glissante (l'heptamère) est toujours nécessaire.

1.3.5 Modèle mécanistique

Le décalage se produit grâce à un déroulement/ré-enroulement partiel de la structure secondaire : ceci suppose qu'une activité hélicase associée au ribosome est responsable du déroulement des structures secondaires sur le messager. Au moment où le ribosome est bloqué en amont d'une structure secondaire, cette activité hélicase permettrait d'ouvrir la double hélice ARN pour poursuivre l'élongation. Si la structure secondaire est particulièrement stable, elle peut avoir tendance à se reformer, et donc à tirer l'ARNm par rapport au ribosome et aux ARNt. Ce modèle, en plus d'une explication énergétique (la force du ré-appariement de la structure secondaire donne l'énergie nécessaire au décalage de phase), expliquerait des résultats jusqu'alors difficilement interprétables :

- Si on élimine le premier appariement du pseudonœud de l'IBV, le taux de décalage de phase ne varie pratiquement pas (Brierley *et al.*, 1991). Si effectivement le pseudonœud est partiellement défait avant décalage des ARNt, le rôle des premiers appariements peut être moins critique ;
- Si on fait varier l'espaceur (entre l'heptamère et la structure secondaire) du FIV par addition ou délétion de nucléotides, le décalage de phase diminue de 6 à 15 fois (Morikawa and Bishop, 1992) ;
- Dans les expériences réalisées pour mettre en évidence les pauses du ribosome en amont d'un pseudonœud, les résultats de cartographie précise montrent deux endroits de pause séparés par 3 nucléotides (Tu *et al.*, 1992). Deux pauses distinctes sont difficilement explicables s'il y a une seule

structure secondaire précisément positionnée par rapport au ribosome. On pourrait envisager que le ribosome, bloqué par le pseudonœud à sa base (première pause), puisse ensuite avancer encore de 3 nucléotides en aval (deuxième pause) lors d'un déroulement partiel de la structure secondaire. Après cette seconde pause, le pseudonœud se reformerait complètement. Si des appariements supplémentaires sont défaits par le ribosome, jusqu'à atteindre la distance optimale pour le décalage, ces résultats deviennent interprétables simplement.

Ce modèle peut également expliquer pourquoi des structures secondaires de type pseudonœud sont plus efficaces à induire le décalage de phase que des tige-boucles. Cette hypothèse, évoquée sous une autre forme, et appelée « *Torsional Resistance Model* » (Dinman and Wickner, 1995) suggère que toute tentative de déroulement d'un pseudonœud maintenu à sa base par le ribosome, et à l'autre extrémité par sa boucle engagée dans l'interaction avec la deuxième tige, provoquerait la formation de supertours, qui opposeraient une forte résistance à l'avancement du ribosome, et augmenteraient la quantité d'énergie disponible pour faire reculer le ribosome. Ceci n'est pas applicable aux tige-boucles dont l'extrémité est libre.

1.3.6 Modèle temporel

Les ARNt peuvent de se décaler d'un nucléotide en amont sur le messager à différents moments :

- Décalage avant le transfert peptidique. Il a été proposé que le glissement du peptidyl-ARNt et de l'aminoacyl-ARNt se faisait après qu'ils se soient positionnés dans les sites P et A, mais avant qu'il y ait eu transfert du peptide d'un ARNt à l'autre (Jacks *et al.*, 1988). Cette hypothèse s'appuie sur le fait que le décalage de phase du HIV-1 *in vitro* (Jacks *et al.*, 1988), ou celui d'autres cibles en expression hétérologue chez *E. coli* (Weiss *et al.*, 1989; Yelverton *et al.*, 1994), se produit parfois grâce au glissement d'un seul d'ARNt au site P. Ceci indique que le décalage avant de transfert du peptide est possible; mais il ne s'agit pas là de décalage de phase en -1 en tandem, et ce décalage en singleton peut aussi avoir lieu juste après la translocation.
- Décalage après le transfert peptidique. Par la suite, il a été proposé que le glissement se fasse d'une façon postérieure au transfert du peptide, mais avant la translocation de l'ARNt déacylé et du peptidyl-ARNt vers les sites E et P (Weiss *et al.*, 1989). Le transfert du peptide serait extrêmement rapide une fois le site A occupé (Farabaugh, 1996). Le décalage des ARNt se ferait donc entre le transfert du peptide et la translocation.

1.3.7 Facteurs interagissant avec la structure secondaire

Finalement, on ne peut pas exclure l'existence de facteurs liant la structure secondaire. Il a été prouvé qu'une protéine se liant à une structure secondaire peut influer sur le décalage de phase au niveau du site de décalage de cadre du *Human immunodeficiency virus 1* (HIV-1) (Kollmus *et al.*, 1996). Cette étude montre qu'un facteur *trans* de ce type peut être envisagé, et non qu'il existe. Ce facteur, s'il existe, est difficile à mettre en évidence :

- Un excès d'oligonucléotides présentant la structure secondaire du SRV1 ou du MMTV ne modifie pas *in vitro* le taux de décalage de phase en -1 (Chen *et al.*, 1995; ten Dam *et al.*, 1994) ;
- Des expériences de retard sur gel ne révèlent pas la liaison d'un facteur protéique sur le pseudonœud du L-A (Dinman, 1995) ;
- Des études génétiques ont été entreprises pour isoler des facteurs faisant varier le taux de décalage de phase du L-A, mais aucun des mutants isolés n'étaient spécifique de la structure secondaire (Cui *et al.*, 1998; Cui *et al.*, 1996; Dinman, 1995; Dinman and Wickner, 1994).

Le pseudonœud pourrait aussi interagir directement avec le ribosome, et y modifier une activité biochimique :

- Interaction avec un ARNr ;
- Interaction avec une protéine du ribosome : S15 chez *E. coli* autorégule sa propre traduction en se fixant en 5' de son messager sur un pseudonœud possédant une adénine intercalée très semblable au pseudonœud utilisé par les rétrovirus (Benard *et al.*, 1998) ; de même S4, qui est impliquée dans la fidélité de la traduction, utilise ce type de régulation (Tang and Draper, 1989). De même les protéines S3, S4 et S5 situées à l'entrée du canal de passage de l'ARNm sont probablement en interaction avec les structures secondaire de l'ARNm (Yusupova *et al.*, 2001) ;
- Interférence avec eEF1A : il a été proposé que les structures observées par RMN pour certains pseudonœuds ressemblent fortement à une structure codon/anticodon appariés (Chen *et al.*, 1995).

1.3.8 État de l'art

Quelques nouveaux exemples de recodage sont découverts tous les ans. La majorité est retrouvée dans les génomes compacts des virus et des éléments transposables. Il est certain qu'une minorité de gènes chromosomiques utilise le recodage pour leur expression. Cependant, ces événements de recodage sont difficiles à identifier et seule une fraction d'entre eux est connue à ce jour. Actuellement, des événements de

recodage sont découverts quand la synthèse d'une protéine avec une activité biochimique connue ne peut s'expliquer par le décodage standard. Un exemple est le décalage de phase de lecture en +1/-2 décrit dans le virus de l'hépatite C (Xu *et al.*, 2001). Un polypeptide plus court que celui résultant du décodage standard a été découvert. Initialement, ce polypeptide avait été écarté en tant que produit de maturation post-traductionnelle du virus. Ceci limite la découverte de nouveaux exemples de recodage. On sait depuis longtemps qu'une synthèse protéique à partir d'un unique ARNm peut mener à la production d'un certain nombre de produits (apparaissant comme des bandes faibles sur des gels SDS), parfois plus grands, parfois plus courts, que le produit principal. Il est habituellement supposé qu'ils résultent soit du décrochage du ribosome à des sites préférentiels, soit de modifications post-traductionnelles (dégradation, phosphorylation, etc.). Il se pourrait pourtant qu'une partie de ces produits soit le résultat de recodages. Dans la plupart des cas, un produit mineur n'est pas étudié s'il constitue moins de 10% du produit principal. Les événements de recodage viraux montrent qu'un décalage de phase de lecture ou une translecture avec un taux inférieur ou égal à 10% peut être physiologiquement significatif. Il est donc possible qu'un grand nombre d'événements de recodage soient actuellement cachés par des limitations technologiques ou par des *a priori*.

Avec les années nos connaissances et notre compréhension du recodage ont beaucoup avancé, jusqu'au point où les premières tentatives de recherches systématiques ont été faites. De ces approches initiales aucun nouvel exemple de recodage authentique n'a été découvert. Les modélisations informatiques basées sur les données empiriques ont été ensuite utilisées lors d'autres tentatives de découverte de nouveaux recodages. Cette approche semble être plus prometteuse, mais probablement plus biaisée. La disponibilité de génomes entiers améliore considérablement les possibilités. La recherche dans les bases de données utilisant uniquement la séquence du site d'insertion de la sélénocystéine a été couronnée de succès. Les programmes développés permettent de parcourir des génomes entiers à la recherche d'éléments SECIS et analysent leur position relative dans l'ORF. Des gènes candidats peuvent être soumis à une vérification expérimentale. Cette approche a aidé à découvrir plusieurs sélénoprotéines de mammifères (Kryukov *et al.*, 1999; Lescure *et al.*, 1999), de la drosophile, *Drosophila melanogaster* (Castellano *et al.*, 2001; Martin-Romero *et al.*, 2001) et du poisson zèbre, *Danio rerio* (Kryukov and Gladyshev, 2000).

Les avancées récentes en bioinformatique et en protéomique, combinant la détection de peptides avec des algorithmes informatiques

sophistiqués, offrent de grands espoirs pour la recherche des recodages. Ces techniques ont l'avantage d'être de moins en moins biaisées, puisqu'elles ne se fondent pas sur la connaissance des sites de recodage précédemment identifiés. Les défis techniques restent cependant élevés.

1.3.9 Recherches Bioinformatiques

L'accumulation des génomes séquencés pose le problème du traitement automatisé de l'information. Un des éléments critiques dans la chaîne de décryptage des données génomiques se situe au niveau de la traduction en protéine du message codé sous forme d'acides nucléiques. Ainsi, si on ne sait pas définir une séquence codante ou si le code que l'on y définit est incomplet, on néglige une partie des informations. Cependant, il apparaît que d'un organisme à l'autre, les signaux présentent à la fois des propriétés communes et des dissemblances. Pour permettre la recherche systématique des gènes soumis à recodage dans les génomes séquencés deux approches bioinformatiques ont déjà été utilisées pour détecter des sites de décalage de phase de lecture en -1.

1.3.9.1 Hammell *et al.*, 1999

La première approche bioinformatique est le fruit du travail de Hammell et collaborateurs en 1999. Dans cette étude, les sites de décalage de phase de lecture en -1 sont caractérisés par la présence d'un motif consensus issu de la modélisation de sites avérés suivis d'un pseudonœud. Ce modèle a été utilisé pour chercher des sites candidats dans les génomes de *Saccharomyces cerevisiæ*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Sus scrofa* et *Drosophila melanogaster*. Parmi les candidats, deux seulement ont été testés biologiquement. *RAS1* chez l'homme et *CCR5* chez la levure. Ces deux séquences présentent, *in vivo* chez la levure dans des constructions artificielles ne respectant pas la site de décalage naturel, un décalage de phase de lecture de 4,4% et 0,2% respectivement. Puis des expériences ultérieures ont montré que *RAS1 in vivo* ne présente pas de décalage de phase de lecture (Hammell *et al.*, 1999).

Cette approche est fortement contrainte par les longueurs des différents éléments du pseudonœud mais pas par la composition nucléotidique elle-même. A l'inverse il n'y a aucune contrainte sur la phase décalée en -1.

1.3.9.2 Lipdardt, 1999

La seconde approche informatique a été réalisée par Lipdardt au cours de sa thèse. La méthode utilisée est un filtrage initial des séquences et une sélection des candidats en fonction d'une valeur seuil après une étape d'apprentissage à partir de sites avérés.

La recherche a été effectuée chez *Saccharomyces cerevisiæ* et *Caenorhabditis elegans*. Le modèle identifie respectivement 9 et 1 candidats. Chez la levure où les candidats ont été identifiés après avoir abaissé le seuil, l'auteur ne valide aucun de ces derniers (Liphardt, 1999).

Cette approche n'a donc pas permis d'identifier de site chez la levure. Le modèle semble effectivement biaisé car il ne considère qu'une tige boucle pour sélectionner les candidats alors que l'apprentissage est fait sur la première tige d'un pseudonœud dont les propriétés en tant qu'élément d'une plus grande structure ne sont probablement pas les mêmes.

1.3.10 Décalages de phase de lecture identifiés

Le nombre de cas documentés de décalage de phase de lecture reste réduit en dépit du grand nombre de séquences disponibles. Le Tableau 3 reprend l'ensemble des cas de déphasages eucaryote répertoriés jusqu'à présent.

ORFs	***Organismes***	**Type**	**Évaluation**
edr	*Homo sapiens*	cellulaire	Avéré
edr	*Mus musculus*	cellulaire	Avéré
pol-pol	*Avian infectious bronchitis virus*	virus	Avéré
39K-pol	*Barley yellow dwarf virus*	virus	-
pro-VPg-pol	*Beet western yellows virus*	virus	Avéré
orf1a-orf1b	*Berne virus*	virus	-
gag-pro-pol	*Bovine leukemia virus*	virus	-
ORF1-ORF2 (pol)	*Carrot mottle mimic virus*	virus	-
Pro-VPg-Pol	*Cereal yellow dwarf virus-RPV*	virus	Avéré
orf2a-orf2b[2]	*Cocksfoot mottle virus*	virus	-
ORF3	*Cucurbit aphid-borne yellows virus*	virus	-
gag-pro-pol	*Enzootic nasal tumor virus*	virus	-
gag-pol	*Equine infectious anemia virus*	virus	Avéré
gag-pol	*Feline immunodeficiency virus*	virus	Avéré
gag-pol	*Giardiavirus*	virus	-
orf1a-orf1b	*Gill-associated virus*	virus	-
ORF1-ORF2(pol)	*Groundnut rosette virus*	virus	-
1a-1b	*Human astrovirus*	virus	-
gag-pol	*Human coronavirus*	virus	Avéré
gag-pol	*Human immunodeficiency virus type 1*	virus	Avéré
gag-pol	*Human immunodeficiency virus type 2*	virus	-
gag-pro-pol	*Human T-cell lymphotropic virus type 1*	virus	-
gag-pro-pol	*Human T-cell lymphotropic virus type 2*	virus	-
gag-pro-pol	*Mason-Pfizer monkey virus*	virus	-
gag-pro-pol	*Mouse mammary tumor virus*	virus	Avéré
1A-1B	*Murine hepatitis virus*	virus	Avéré
ORF1-ORF2 (pol)	*Pea enation mosaic virus RNA 2*	virus	-
orf1a-orf1b	*Porcine reproductive and respiratory syndrome virus*	virus	-
protease-VPg-polymerase	*Potato leafroll virus*	virus	Avéré
p27-p57	*Red clover necrotic mosaic virus*	virus	Avéré
gag-pol	*Rous sarcoma virus*	virus	Avéré

gag-pol	*Saccharomyces cerevisiae virus L-A*	virus	Avéré
gag-pol	*SARS coronavirus*	virus	-
gag-pol	*Simian immunodeficiency virus*	virus	-
gag-pro-pol	*Simian retrovirus type 2*	virus	-
gag-pro-pol	*Simian T-cell lymphotropic virus type 1*	virus	-
gag-pro-pol	*Simian type D virus 1*	virus	Avéré
gag-pol	*Trichomonas vaginalis virus II*	virus	-
gag-pro pol	*Visna virus*	virus	-
gag-pol	*Drosophila ananassae Tom retrotransposone*	transposon	-
gag-pol	*Drosophila buzzatii Ossvaldo retrotransposone*	transposon	-
gag-pol	*Drosophila melanogaster gypsy transposable element*	transposon	-
gag-pol	*D. melanogaster retrotransposon 1731*	transposon	Avéré
gag	*D. melanogaster telomeric retrotransposon Het-A*	transposon	-
orf1-orf2-orf3	*Drosophila transposable genetic element 17.6*	transposon	-
gag-pol	*Intracisternal A-type particle IAP*	transposon	-

Tableau 3 : Liste des décalages de phase de lecture -1 eucaryotes. Seul le gène *edr* est un gène cellulaire. Les autres sont des ORFs de virus ou de retro-transposons. La dernière colonne indique si les événements de décalage de phase de lecture en -1 sont biologiquement avérés ou identifiés par comparaison de séquences.

1.4 La problématique

Comme il est rappelé ci-dessus, l'essentiel des événements non conventionnels de lecture du message génétique ont été observés jusqu'à présent chez des virus. On peut avancer deux explications : soit ces mécanismes sont réservés aux éléments transposables et aux virus, soit l'information disponible sur les génomes viraux étant beaucoup plus complète que sur les génomes d'organismes complexes, les gènes cellulaires ainsi contrôlés n'ont pas encore été identifiés.

Mon travail de thèse a eu pour objet la recherche de gènes contrôlés par décalage de phase de lecture en -1 chez la levure *S. cerevisiæ*. J'ai été amené à développer en parallèle des approches de biologie expérimentale (biologie moléculaire, génétique, microbiologie) et de bioinformatique (analyse de séquences, développement de base de données, HMM).

Afin de mener une étude systématique, des collaborations ont été établies avec le groupe de Bioinformatique des Génomes de l'Institut de Génétique et Microbiologie, avec l'équipe de Bioinformatique du Laboratoire de Recherche en Informatique de l'Université Paris-Sud et avec le Laboratoire Statistique et Génome de l'Université d'Evry.

Une première approche a consisté à concevoir un modèle de site de décalage en -1 qui prenne en compte les différents signaux, et qui permette de détecter de tels sites dans les génomes séquencés. Elle allie méthodes d'apprentissage de concepts, algorithmiques des séquences et expérimentations biologiques (Bekaert *et al.*, 2003).

Parallèlement, j'ai élaboré une approche fondée sur un principe totalement différent. Elle a consisté à rechercher, non pas des signaux

spécifiques associés au décalage de cadre, mais des organisations génomiques compatibles avec un tel événement. A partir des nombreuses séquences issues de ce premier crible, j'ai développé deux méthodes pour identifier les candidats les plus pertinents. L'une consiste à rechercher des motifs protéiques dans les extensions codantes, l'autre utilise des modèles de Markov cachés pour reconnaître des zones potentiellement traduites.

Au cours de ces travaux j'ai abordé d'autres points d'importance pour la connaissance de la mécanique du décalage de phase de lecture. C'est ainsi que j'ai identifié plusieurs nouveaux sites viraux en utilisant des approches statistiques basées sur la présence de motifs distinctifs. J'ai ensuite validé ces sites biologiquement. Mes recherches m'ont également permis de mettre en évidence le rôle dans l'efficacité de changement de cadre de lecture en -1, de la modification des ARNt présents au site E du ribosome au moment du décalage (Bekaert *et al.*, 2005).

1.5 Références

Altmann, M., Edery, I., Sonenberg, N., and Trachsel, H. (1985). Purification and characterization of protein synthesis initiation factor eIF-4E from the yeast Saccharomyces cerevisiae. Biochemistry *24*, 6085-6089. [3910088]

Angenon, G., Van Montagu, M., and Depicker, A. (1990). Analysis of the stop codon context in plant nuclear genes. FEBS Lett *271*, 144-146. [2226798]

Asakura, T., Sasaki, T., Nagano, F., Satoh, A., Obaishi, H., Nishioka, H., Imamura, H., Hotta, K., Tanaka, K., Nakanishi, H., and Takai, Y. (1998). Isolation and characterization of a novel actin filament-binding protein from Saccharomyces cerevisiae. Oncogene *16*, 121-130. [9467951]

Asano, K., Clayton, J., Shalev, A., and Hinnebusch, A. G. (2000). A multifactor complex of eukaryotic initiation factors, eIF1, eIF2, eIF3, eIF5, and initiator tRNA(Met) is an important translation initiation intermediate in vivo. Genes Dev *14*, 2534-2546. [11018020]

Atkins, J. F., Gesteland, R. F., Reid, B. R., and Anderson, C. W. (1979). Normal tRNAs promote ribosomal frameshifting. Cell *18*, 1119-1131. [391405]

Atkins, J. F., Lewis, J. B., Anderson, C. W., and Gesteland, R. F. (1975). Enhanced differential synthesis of proteins in a mammalian cell-free system by addition of polyamines. J Biol Chem *250*, 5688-5695. [167021]

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science *289*, 905-920. [10937989]

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002). Release factor 2 frameshifting sites in different bacteria. EMBO Rep *3*, 373-377. [11897659]

Baranov, P. V., Gurvich, O. L., Fayet, O., Prere, M. F., Miller, W. A., Gesteland, R. F., Atkins, J. F., and Giddings, M. C. (2001). RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res *29*, 264-267. [11125107]

Baril, M., Dulude, D., Steinberg, S. V., and Brakier-Gingras, L. (2003). The frameshift stimulatory signal of human immunodeficiency virus type 1 group O is a pseudoknot. J Mol Biol *331*, 571-583. [12899829]

Beckmann, R., Bubeck, D., Grassucci, R., Penczek, P., Verschoor, A., Blobel, G., and Frank, J. (1997). Alignment of conduits for the nascent polypeptide chain in the ribosome-Sec61 complex. Science *278*, 2123-2126. [9405348]

Beckmann, R., Spahn, C. M., Eswar, N., Helmers, J., Penczek, P. A., Sali, A., Frank, J., and Blobel, G. (2001). Architecture of the protein-conducting channel associated with the translating 80S ribosome. Cell *107*, 361-372. [11701126]

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics *19*, 327-335. [12584117]

Belcourt, M. F., and Farabaugh, P. J. (1990). Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. Cell *62*, 339-352. [2164889]

Benard, L., Mathy, N., Grunberg-Manago, M., Ehresmann, B., Ehresmann, C., and Portier, C. (1998). Identification in a pseudoknot of a U.G motif essential for the regulation of the expression of ribosomal protein S15. Proc Natl Acad Sci U S A *95*, 2564-2567. [9482926]

Bergstrom, D. E., Merli, C. A., Cygan, J. A., Shelby, R., and Blackman, R. K. (1995). Regulatory autonomy and molecular characterization of the Drosophila out at first gene. Genetics *139*, 1331-1346. [7768442]

Bernstein, H. D., Poritz, M. A., Strub, K., Hoben, P. J., Brenner, S., and Walter, P. (1989). Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature *340*, 482-486. [2502718]

Berry, M. J., Banu, L., Chen, Y. Y., Mandel, S. J., Kieffer, J. D., Harney, J. W., and Larsen, P. R. (1991). Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. Nature *353*, 273-276. [1832744]

Berry, M. J., Banu, L., Harney, J. W., and Larsen, P. R. (1993). Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. Embo J *12*, 3315-3322. [8344267]

Berry, M. J., Maia, A. L., Kieffer, J. D., Harney, J. W., and Larsen, P. R. (1992). Substitution of cysteine for selenocysteine in type I iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation. Endocrinology *131*, 1848-1852. [1396330]

Bi, X., and Goss, D. J. (2000). Wheat germ poly(A)-binding protein increases the ATPase and the RNA helicase activity of translation initiation factors eIF4A, eIF4B, and eIF-iso4F. J Biol Chem *275*, 17740-17746. [10748132]

Bidou, L., Stahl, G., Grima, B., Liu, H., Cassan, M., and Rousset, J. P. (1997). In vivo HIV-1 frameshifting efficiency is directly related to the stability of the stem-loop stimulatory signal. Rna *3*, 1153-1158. [9326490]

Blight, S. K., Larue, R. C., Mahapatra, A., Longstaff, D. G., Chang, E., Zhao, G., Kang, P. T., Green-Church, K. B., Chan, M. K., and Krzycki, J. A. (2004). Direct charging of tRNA(CUA) with pyrrolysine in vitro and in vivo. Nature. [15329732]

Bonetti, B., Fu, L., Moon, J., and Bedwell, D. M. (1995). The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. J Mol Biol *251*, 334-345. [7650736]

Bossi, L. (1983). Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J Mol Biol *164*, 73-87. [6188841]

Bretscher, M. S. (1968). Translocation in protein synthesis: a hybrid structure model. Nature *218*, 675-677. [5655957]

Brierley, I. (1995). Ribosomal frameshifting viral RNAs. J Gen Virol *76 (Pt 8)*, 1885-1892. [7636469]

Brierley, I., Digard, P., and Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell *57*, 537-547. [2720781]

Brierley, I., Jenner, A. J., and Inglis, S. C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. J Mol Biol *227*, 463-479. [1404364]

Brierley, I., Rolley, N. J., Jenner, A. J., and Inglis, S. C. (1991). Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. J Mol Biol *220*, 889-902. [1880803]

Brown, C. M., Dinesh-Kumar, S. P., and Miller, W. A. (1996). Local and distant sequences are required for efficient readthrough of the barley yellow dwarf virus PAV coat protein gene stop codon. J Virol *70*, 5884-5892. [8709208]

Brown, C. M., Stockwell, P. A., Trotman, C. N., and Tate, W. P. (1990). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Res *18*, 6339-6345. [2123028]

Buckingham, R. H., Grentzmann, G., and Kisselev, L. (1997). Polypeptide chain release factors. Mol Microbiol *24*, 449-456. [9179839]

Castellano, S., Morozova, N., Morey, M., Berry, M. J., Serras, F., Corominas, M., and Guigo, R. (2001). In silico identification of novel selenoproteins in the Drosophila melanogaster genome. EMBO Rep *2*, 697-702. [11493597]

Chamorro, M., Parkin, N., and Varmus, H. E. (1992). An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. Proc Natl Acad Sci U S A *89*, 713-717. [1309954]

Chaudhuri, J., Chowdhury, D., and Maitra, U. (1999). Distinct functions of eukaryotic translation initiation factors eIF1A and eIF3 in the formation of the 40 S ribosomal preinitiation complex. J Biol Chem *274*, 17975-17980. [10364246]

Chen, X., Chamorro, M., Lee, S. I., Shen, L. X., Hines, J. V., Tinoco, I., Jr., and Varmus, H. E. (1995). Structural and functional studies of retroviral RNA pseudoknots involved in ribosomal frameshifting: nucleotides at the junction of the two stems are important for efficient ribosomal frameshifting. Embo J *14*, 842-852. [7882986]

Cigan, A. M., Feng, L., and Donahue, T. F. (1988). tRNAi(met) functions in directing the scanning ribosome to the start site of translation. Science *242*, 93-97. [3051379]

Copeland, P. R., Fletcher, J. E., Carlson, B. A., Hatfield, D. L., and Driscoll, D. M. (2000). A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. Embo J *19*, 306-314. [10637234]

Copeland, P. R., Stepanik, V. A., and Driscoll, D. M. (2001). Insight into mammalian selenocysteine insertion: domain structure and ribosome binding properties of Sec insertion sequence binding protein 2. Mol Cell Biol *21*, 1491-1498. [11238886]

Craigen, W. J., and Caskey, C. T. (1986). Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature *322*, 273-275. [3736654]

Csaszar, K., Spackova, N., Stefl, R., Sponer, J., and Leontis, N. B. (2001). Molecular dynamics of the frame-shifting pseudoknot from beet western yellows virus: the role of non-Watson-Crick base-pairing, ordered hydration, cation binding and base mutations on stability and unfolding. J Mol Biol *313*, 1073-1091. [11700064]

Cui, Y., Dinman, J. D., Kinzy, T. G., and Peltz, S. W. (1998). The Mof2/Sui1 protein is a general monitor of translational accuracy. Mol Cell Biol *18*, 1506-1516. [9488467]

Cui, Y., Dinman, J. D., and Peltz, S. W. (1996). Mof4-1 is an allele of the UPF1/IFS2 gene which affects both mRNA turnover and -1 ribosomal frameshifting efficiency. Embo J *15*, 5726-5736. [8896465]

Curran, J. F. (1993). Analysis of effects of tRNA:message stability on frameshift frequency at the Escherichia coli RF2 programmed frameshift site. Nucleic Acids Res *21*, 1837-1843. [8493101]

Dabrowski, M., Spahn, C. M., Schafer, M. A., Patzke, S., and Nierhaus, K. H. (1998). Protection patterns of tRNAs do not change during ribosomal translocation. J Biol Chem *273*, 32793-32800. [9830024]

Dam, E., Pleij, K., and Draper, D. (1992). Structural and functional aspects of RNA pseudoknots. Biochemistry *31*, 11665-11676. [1280160]

Davies, J., Gilbert, W., and Gorini, L. (1964). Streptomycin, Suppression, and the Code. Proc Natl Acad Sci U S A *51*, 883-890. [14173007]

den Boon, J. A., Snijder, E. J., Chirnside, E. D., de Vries, A. A., Horzinek, M. C., and Spaan, W. J. (1991). Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily. J Virol *65*, 2910-2920. [1851863]

Dinman, J. D. (1995). Ribosomal frameshifting in yeast viruses. Yeast *11*, 1115-1127. [8619310]

Dinman, J. D., Icho, T., and Wickner, R. B. (1991). A -1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. Proc Natl Acad Sci U S A *88*, 174-178. [1986362]

Dinman, J. D., and Kinzy, T. G. (1997). Translational misreading: mutations in translation elongation factor 1alpha differentially affect programmed ribosomal frameshifting and drug sensitivity. Rna *3*, 870-881. [9257646]

Dinman, J. D., and Wickner, R. B. (1992). Ribosomal frameshifting efficiency and gag/gag-pol ratio are critical for yeast M1 double-stranded RNA virus propagation. J Virol *66*, 3669-3676. [1583726]

Dinman, J. D., and Wickner, R. B. (1994). Translational maintenance of frame: mutants of Saccharomyces cerevisiae with altered -1 ribosomal frameshifting efficiencies. Genetics *136*, 75-86. [8138178]

Dinman, J. D., and Wickner, R. B. (1995). 5 S rRNA is involved in fidelity of translational reading frame. Genetics *141*, 95-105. [8536994]

Dominguez, D., Altmann, M., Benz, J., Baumann, U., and Trachsel, H. (1999). Interaction of translation initiation factor eIF4G with eIF4A in the yeast Saccharomyces cerevisiae. J Biol Chem *274*, 26720-26726. [10480875]

Dulude, D., Baril, M., and Brakier-Gingras, L. (2002). Characterization of the frameshift stimulatory signal controlling a programmed -1 ribosomal frameshift in the human immunodeficiency virus type 1. Nucleic Acids Res *30*, 5094-5102. [12466532]

Dunn, J. J., and Studier, F. W. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol *166*, 477-535. [6864790]

Dyer, J. R., and Sossin, W. S. (2000). Regulation of eukaryotic initiation factor 4E phosphorylation in the nervous system of Aplysia californica. J Neurochem *75*, 872-881. [10899966]

Eggertsson, G., and Soll, D. (1988). Transfer ribonucleic acid-mediated suppression of termination codons in Escherichia coli. Microbiol Rev *52*, 354-374. [3054467]

Egli, M., Minasov, G., Su, L., and Rich, A. (2002). Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. Proc Natl Acad Sci U S A *99*, 4302-4307. [11904368]

Fagegaltier, D., Hubert, N., Yamada, K., Mizutani, T., Carbon, P., and Krol, A. (2000). Characterization of mSelB, a novel mammalian elongation factor for selenoprotein translation. Embo J *19*, 4796-4805. [10970870]

Fahrenkrug, S. C., Joshi, B., Hackett, P. B., Jr., and Jagus, R. (2000). Alternative transcriptional initiation and splicing define the translational efficiencies of zebrafish mRNAs encoding eukaryotic initiation factor 4E. Differentiation *66*, 15-22. [10997588]

Farabaugh, P. J. (1996). Programmed translational frameshifting. Annu Rev Genet *30*, 507-528. [8982463]

Farabaugh, P. J., Zhao, H., and Vimaladithan, A. (1993). A novel programed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. Cell *74*, 93-103. [8267715]

Fletcher, C. M., Pestova, T. V., Hellen, C. U., and Wagner, G. (1999). Structure and interactions of the translation initiation factor eIF1. Embo J *18*, 2631-2637. [10228174]

Flower, A. M., and McHenry, C. S. (1990). The gamma subunit of DNA polymerase III holoenzyme of Escherichia coli is produced by ribosomal frameshifting. Proc Natl Acad Sci U S A *87*, 3713-3717. [2187190]

Forchhammer, K., Leinfelder, W., and Bock, A. (1989). Identification of a novel translation factor necessary for the incorporation of selenocysteine into protein. Nature *342*, 453-456. [2531290]

Forster, C., Ott, G., Forchhammer, K., and Sprinzl, M. (1990). Interaction of a selenocysteine-incorporating tRNA with elongation factor Tu from E.coli. Nucleic Acids Res *18*, 487-491. [2408012]

Frank, J. (2001). Ribosomal dynamics explored by cryo-electron microscopy. Methods *25*, 309-315. [11860285]

Frank, J., and Agrawal, R. K. (2000). A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature *406*, 318-322. [10917535]

Frolova, L. Y., Tsivkovskii, R. Y., Sivolobova, G. F., Oparina, N. Y., Serpinsky, O. I., Blinov, V. M., Tatkov, S. I., and Kisselev, L. L. (1999). Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. Rna *5*, 1014-1020. [10445876]

Fujiwara, T., Busch, K., Gross, H. J., and Mizutani, T. (1999). A SECIS binding protein (SBP) is distinct from selenocysteyl-tRNA protecting factor (SePF). Biochimie *81*, 213-218. [10385002]

Gao, D., and McHenry, C. S. (2001). tau binds and organizes Escherichia coli replication through distinct domains. Partial proteolysis of terminally tagged tau

to determine candidate domains and to assign domain V as the alpha binding domain. J Biol Chem *276*, 4433-4440. [11078743]

Gesteland, R. F., and Atkins, J. F. (1996). Recoding: dynamic reprogramming of translation. Annu Rev Biochem *65*, 741-768. [8811194]

Gesteland, R. F., Weiss, R. B., and Atkins, J. F. (1992). Recoding: reprogrammed genetic decoding. Science *257*, 1640-1641. [1529352]

Giedroc, D. P., Cornish, P. V., and Hennig, M. (2003). Detection of scalar couplings involving 2'-hydroxyl protons across hydrogen bonds in a frameshifting mRNA pseudoknot. J Am Chem Soc *125*, 4676-4677. [12696863]

Gil, J., Esteban, M., and Roth, D. (2000). In vivo regulation of protein synthesis by phosphorylation of the alpha subunit of wheat eukaryotic initiation factor 2. Biochemistry *39*, 7521-7530. [10858301]

Goyer, C., Altmann, M., Trachsel, H., and Sonenberg, N. (1989). Identification and characterization of cap-binding proteins from yeast. J Biol Chem *264*, 7603-7610. [2651444]

Groft, C. M., Beckmann, R., Sali, A., and Burley, S. K. (2000). Crystal structures of ribosome anti-association factor IF6. Nat Struct Biol *7*, 1156-1164. [11101899]

Gross, J. D., Moerke, N. J., von der Haar, T., Lugovskoy, A. A., Sachs, A. B., McCarthy, J. E., and Wagner, G. (2003). Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. Cell *115*, 739-750. [14675538]

Grundner-Culemann, E., Martin, G. W., 3rd, Tujebajeva, R., Harney, J. W., and Berry, M. J. (2001). Interplay between termination and translation machinery in eukaryotic selenoprotein synthesis. J Mol Biol *310*, 699-707. [11453681]

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999). Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res *9*, 417-427. [10330121]

Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., and Chan, M. K. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. Science *296*, 1462-1466. [12029132]

Harding, H. P., Novoa, I., Zhang, Y., Zeng, H., Wek, R., Schapira, M., and Ron, D. (2000). Regulated translation initiation controls stress-induced gene expression in mammalian cells. Mol Cell *6*, 1099-1108. [11106749]

Harrell, L., Melcher, U., and Atkins, J. F. (2002). Predominance of six different hexanucleotide recoding signals 3' of read-through stop codons. Nucleic Acids Res *30*, 2011-2017. [11972340]

Hatfield, D. L., Smith, D. W., Lee, B. J., Worland, P. J., and Oroszlan, S. (1990). Structure and function of suppressor tRNAs in higher eukaryotes. Crit Rev Biochem Mol Biol *25*, 71-96. [2183969]

Heider, J., Baron, C., and Bock, A. (1992). Coding from a distance: dissection of the mRNA determinants required for the incorporation of selenocysteine into protein. Embo J *11*, 3759-3766. [1396569]

Hellen, C. U., and Sarnow, P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev *15*, 1593-1612. [11445534]

Herr, A. J., Atkins, J. F., and Gesteland, R. F. (1999). Mutations which alter the elbow region of tRNA2Gly reduce T4 gene 60 translational bypassing efficiency. Embo J *18*, 2886-2896. [10329634]

Herr, A. J., Gesteland, R. F., and Atkins, J. F. (2000). One protein from two open reading frames: mechanism of a 50 nt translational bypass. Embo J *19*, 2671-2680. [10835364]

Herr, A. J., Nelson, C. C., Wills, N. M., Gesteland, R. F., and Atkins, J. F. (2001). Analysis of the roles of tRNA structure, ribosomal protein L9, and the bacteriophage T4 gene 60 bypassing signals during ribosome slippage on mRNA. J Mol Biol *309*, 1029-1048. [11399077]

Herr, A. J., Wills, N. M., Nelson, C. C., Gesteland, R. F., and Atkins, J. F. (2001). Drop-off during ribosome hopping. J Mol Biol *311*, 445-452. [11492998]

Hill, K. E., Lloyd, R. S., Yang, J. G., Read, R., and Burk, R. F. (1991). The cDNA for rat selenoprotein P contains 10 TGA codons in the open reading frame. J Biol Chem *266*, 10050-10053. [2037562]

Himeno, S., Chittum, H. S., and Burk, R. F. (1996). Isoforms of selenoprotein P in rat plasma. Evidence for a full-length form and another form that terminates at the second UGA in the open reading frame. J Biol Chem *271*, 15769-15775. [8663023]

Ho, J. H., Kallstrom, G., and Johnson, A. W. (2000). Nmd3p is a Crm1p-dependent adapter protein for nuclear export of the large ribosomal subunit. J Cell Biol *151*, 1057-1066. [11086007]

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc Natl Acad Sci U S A *71*, 4135-4139. [4530290]

Howard, M. T., Shirts, B. H., Zhou, J., Carlson, C. L., Matsufuji, S., Gesteland, R. F., Weeks, R. S., and Atkins, J. F. (2001). Cell culture analysis of the regulatory frameshift event required for the expression of mammalian antizymes. Genes Cells *6*, 931-941. [11733031]

Huang, W. M., Ao, S. Z., Casjens, S., Orlandi, R., Zeikus, R., Weiss, R., Winge, D., and Fang, M. (1988). A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science *239*, 1005-1012. [2830666]

Hubert, N., Walczak, R., Carbon, P., and Krol, A. (1996). A protein binds the selenocysteine insertion element in the 3'-UTR of mammalian selenoprotein mRNAs. Nucleic Acids Res *24*, 464-469. [8602359]

Hung, M., Patel, P., Davis, S., and Green, S. R. (1998). Importance of ribosomal frameshifting for human immunodeficiency virus type 1 particle assembly and replication. J Virol *72*, 4819-4824. [9573247]

Ivanov, I. P., Gesteland, R. F., and Atkins, J. F. (2000). Antizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit. Nucleic Acids Res *28*, 3185-3196. [10954585]

Ivanov, I. P., Matsufuji, S., Murakami, Y., Gesteland, R. F., and Atkins, J. F. (2000). Conservation of polyamine regulation by translational frameshifting from yeast to mammals. Embo J *19*, 1907-1917. [10775274]

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447-458. [2846182]

Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature *331*, 280-283. [2447506]

Jacks, T., Townsley, K., Varmus, H. E., and Majors, J. (1987). Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins. Proc Natl Acad Sci U S A *84*, 4298-4302. [3035577]

Karow, M. L., Rogers, E. J., Lovett, P. S., and Piggot, P. J. (1998). Suppression of TGA mutations in the Bacillus subtilis spoIIR gene by prfB mutations. J Bacteriol *180*, 4166-4170. [9696765]

Kawakami, K., Inada, T., and Nakamura, Y. (1988). Conditionally lethal and recessive UGA-suppressor mutations in the prfB gene encoding peptide chain release factor 2 of Escherichia coli. J Bacteriol *170*, 5378-5381. [3053663]

Kawakami, K., Pande, S., Faiola, B., Moore, D. P., Boeke, J. D., Farabaugh, P. J., Strathern, J. N., Nakamura, Y., and Garfinkel, D. J. (1993). A rare tRNA-Arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in Saccharomyces cerevisiae. Genetics *135*, 309-320. [8243996]

Keiper, B. D., and Rhoads, R. E. (1999). Translational recruitment of Xenopus maternal mRNAs in response to poly(A) elongation requires initiation factor eIF4G-1. Dev Biol *206*, 1-14. [9918691]

Kim, Y. G., Maas, S., and Rich, A. (2001). Comparative mutational analysis of cis-acting RNA signals for translational frameshifting in HIV-1 and HTLV-2. Nucleic Acids Res *29*, 1125-1131. [11222762]

Klein, D. J., Schmeing, T. M., Moore, P. B., and Steitz, T. A. (2001). The kink-turn: a new RNA secondary structure motif. Embo J *20*, 4214-4221. [11483524]

Kohrl, J., Brigelius-Flohe, R., Bock, A., Gartner, R., Meyer, O., and Flohe, L. (2000). Selenium in biology: facts and medical perspectives. Biol Chem *381*, 849-864. [11076017]

Kollmus, H., Hentze, M. W., and Hauser, H. (1996). Regulated ribosomal frameshifting by an RNA-protein interaction. Rna *2*, 316-323. [8634912]

Kollmus, H., Honigman, A., Panet, A., and Hauser, H. (1994). The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II in vivo. J Virol *68*, 6087-6091. [8057488]

Kozak, M. (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. Mol Cell Biol *7*, 3438-3445. [3683388]

Kozak, M. (1997). Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. Embo J *16*, 2482-2492. [9171361]

Kromayer, M., Wilting, R., Tormay, P., and Bock, A. (1996). Domain structure of the prokaryotic selenocysteine-specific elongation factor SelB. J Mol Biol *262*, 413-420. [8893853]

Kryukov, G. V., and Gladyshev, V. N. (2000). Selenium metabolism in zebrafish: multiplicity of selenoprotein genes and expression of a protein containing 17 selenocysteine residues. Genes Cells *5*, 1049-1060. [11168591]

Kryukov, G. V., Kryukov, V. M., and Gladyshev, V. N. (1999). New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. J Biol Chem *274*, 33888-33897. [10567350]

Kurland, C. G. (1992). Translational accuracy and the fitness of bacteria. Annu Rev Genet *26*, 29-50. [1482115]

Larsen, B., Gesteland, R. F., and Atkins, J. F. (1997). Structural probing and mutagenic analysis of the stem-loop required for Escherichia coli dnaX ribosomal frameshifting: programmed efficiency of 50%. J Mol Biol *271*, 47-60. [9300054]

Larsen, B., Peden, J., Matsufuji, S., Matsufuji, T., Brady, K., Maldonado, R., Wills, N. M., Fayet, O., Atkins, J. F., and Gesteland, R. F. (1995). Upstream stimulators for recoding. Biochem Cell Biol *73*, 1123-1129. [8722029]

Larsen, B., Wills, N. M., Gesteland, R. F., and Atkins, J. F. (1994). rRNA-mRNA base pairing stimulates a programmed -1 ribosomal frameshift. J Bacteriol *176*, 6842-6851. [7961443]

Leger, M., Sidani, S., and Brakier-Gingras, L. (2004). A reassessment of the response of the bacterial ribosome to the frameshift stimulatory signal of the human immunodeficiency virus type 1. Rna *10*, 1225-1235. [15247429]

Lescure, A., Gautheret, D., Carbon, P., and Krol, A. (1999). Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. J Biol Chem *274*, 38147-38154. [10608886]

Li, Z., Stahl, G., and Farabaugh, P. J. (2001). Programmed +1 frameshifting stimulated by complementarity between a downstream mRNA sequence and an error-correcting region of rRNA. Rna *7*, 275-284. [11233984]

Liphardt, J. (1999) The mechanism of -1 ribosomal frameshifting: experimental and theoretical analysis, Ph.D., Churchill College, Cambridge.

Liu, Z., Reches, M., and Engelberg-Kulka, H. (1999). A sequence in the Escherichia coli fdhF "selenocysteine insertion Sequence" (SECIS) operates in the absence of selenium. J Mol Biol *294*, 1073-1086. [10600367]

Ma, S., Hill, K. E., Caprioli, R. M., and Burk, R. F. (2002). Mass spectrometric characterization of full-length rat selenoprotein P and three isoforms shortened at the C terminus. Evidence that three UGA codons in the mRNA open reading frame have alternative functions of specifying selenocysteine insertion or translation termination. J Biol Chem *277*, 12749-12754. [11821412]

Maden, B. E. (1990). The numerous modified nucleotides in eukaryotic ribosomal RNA. Prog Nucleic Acid Res Mol Biol *39*, 241-303. [2247610]

Major, L. L., Poole, E. S., Dalphin, M. E., Mannering, S. A., and Tate, W. P. (1996). Is the in-frame termination signal of the Escherichia coli release factor-2 frameshift site weakened by a particularly poor context? Nucleic Acids Res *24*, 2673-2678. [8758994]

Maldonado, R., and Herr, A. J. (1998). Efficiency of T4 gene 60 translational bypassing. J Bacteriol *180*, 1822-1830. [9537381]

Mao, X., Schwer, B., and Shuman, S. (1995). Yeast mRNA cap methyltransferase is a 50-kilodalton protein encoded by an essential gene. Mol Cell Biol *15*, 4167-4174. [7623811]

Marquez, V., Wilson, D. N., and Nierhaus, K. H. (2002). Functions and interplay of the tRNA-binding sites of the ribosome. Biochem Soc Trans *30*, 133-140. [12023840]

Martin-Romero, F. J., Kryukov, G. V., Lobanov, A. V., Carlson, B. A., Lee, B. J., Gladyshev, V. N., and Hatfield, D. L. (2001). Selenium metabolism in Drosophila: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. J Biol Chem *276*, 29798-29804. [11389138]

Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J. F., Gesteland, R. F., and Hayashi, S. (1995). Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell *80*, 51-60. [7813017]

Matsufuji, S., Matsufuji, T., Wills, N. M., Gesteland, R. F., and Atkins, J. F. (1996). Reading two bases twice: mammalian antizyme frameshifting in yeast. Embo J *15*, 1360-1370. [8635469]

Matsuo, H., Li, H., McGuire, A. M., Fletcher, C. M., Gingras, A. C., Sonenberg, N., and Wagner, G. (1997). Structure of translation factor eIF4E bound to m7GDP and interaction with 4E-binding protein. Nat Struct Biol *4*, 717-724. [9302999]

McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J., and Tate, W. P. (1995). Translational termination efficiency in mammals is influenced by the base following the stop codon. Proc Natl Acad Sci U S A *92*, 5431-5435. [7777525]

Miyazaki, Y., Matsufuji, S., and Hayashi, S. (1992). Cloning and characterization of a rat gene encoding ornithine decarboxylase antizyme. Gene *113*, 191-197. [1572540]

Moazed, D., and Noller, H. F. (1989). Intermediate states in the movement of transfer RNA in the ribosome. Nature *342*, 142-148. [2682263]

Morikawa, S., and Bishop, D. H. (1992). Identification and analysis of the gag-pol ribosomal frameshift site of feline immunodeficiency virus. Virology *186*, 389-397. [1310175]

Morris, D. K., and Lundblad, V. (1997). Programmed translational frameshifting in a gene required for yeast telomere replication. Curr Biol *7*, 969-976. [9382847]

Mottagui-Tabar, S., Bjornsson, A., and Isaksson, L. A. (1994). The second to last amino acid in the nascent peptide as a codon context determinant. Embo J *13*, 249-257. [8306967]

Mottagui-Tabar, S., and Isaksson, L. A. (1997). Only the last amino acids in the nascent peptide influence translation termination in Escherichia coli genes. FEBS Lett *414*, 165-170. [9305752]

Namy, O., Duchateau-Nguyen, G., and Rousset, J. P. (2002). Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. Mol Microbiol *43*, 641-652. [11929521]

Namy, O., Hatin, I., and Rousset, J. P. (2001). Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep *2*, 787-793. [11520858]

Namy, O., Rousset, J. P., Napthine, S., and Brierley, I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol Cell *13*, 157-168. [14759362]

Nasim, M. T., Jaenecke, S., Belduz, A., Kollmus, H., Flohe, L., and McCarthy, J. E. (2000). Eukaryotic selenocysteine incorporation follows a nonprocessive mechanism that competes with translational termination. J Biol Chem *275*, 14846-14852. [10809727]

Ninio, J. (1975). Kinetic amplification of enzyme discrimination. Biochimie *57*, 587-595. [1182215]

Noller, H. F., Yusupov, M. M., Yusupova, G. Z., Baucom, A., and Cate, J. H. (2002). Translocation of tRNA during protein synthesis. FEBS Lett *514*, 11-16. [11904173]

Oeschger, M. P., Oeschger, N. S., Wiprud, G. T., and Woods, S. L. (1980). High efficiency temperature-sensitive amber suppressor strains of Escherichia coli K12: isolation of strains with suppressor-enhancing mutations. Mol Gen Genet *177*, 545-552. [6991863]

Pande, S., Vimaladithan, A., Zhao, H., and Farabaugh, P. J. (1995). Pulling the ribosome out of frame by +1 at a programmed frameshift site by cognate binding of aminoacyl-tRNA. Mol Cell Biol *15*, 298-304. [7799937]

Pape, T., Wintermeyer, W., and Rodnina, M. (1999). Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. Embo J *18*, 3800-3807. [10393195]

Parker, J. (1989). Errors and alternatives in reading the universal genetic code. Microbiol Rev *53*, 273-298. [2677635]

Pestova, T. V., Borukhov, S. I., and Hellen, C. U. (1998). Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. Nature *394*, 854-859. [9732867]

Phan, L., Zhang, X., Asano, K., Anderson, J., Vornlocher, H. P., Greenberg, J. R., Qin, J., and Hinnebusch, A. G. (1998). Identification of a translation initiation factor 3 (eIF3) core complex, conserved in yeast and mammals, that interacts with eIF5. Mol Cell Biol *18*, 4935-4946. [9671501]

Plant, E. P., Jacobs, K. L., Harger, J. W., Meskauskas, A., Jacobs, J. L., Baxter, J. L., Petrov, A. N., and Dinman, J. D. (2003). The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. Rna *9*, 168-174. [12554858]

Polycarpo, C., Ambrogelly, A., Ruan, B., Tumbula-Hansen, D., Ataide, S. F., Ishitani, R., Yokoyama, S., Nureki, O., Ibba, M., and Soll, D. (2003). Activation of the pyrrolysine suppressor tRNA requires formation of a ternary complex with class I and class II lysyl-tRNA synthetases. Mol Cell *12*, 287-294. [14536069]

Poole, E. S., Major, L. L., Mannering, S. A., and Tate, W. P. (1998). Translational termination in Escherichia coli: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. Nucleic Acids Res *26*, 954-960. [9461453]

Potapov, A. P. (1982). A stereospecific mechanism for the aminoacyl-tRNA selection at the ribosome. FEBS Lett *146*, 5-8. [6923830]

Prescott, C. D., Kleuvers, B., and Goringer, H. U. (1991). A rRNA-mRNA base pairing model for UGA-dependent termination. Biochimie *73*, 1121-1129. [1742356]

Quiocho, F. A., Hu, G., and Gershon, P. D. (2000). Structural basis of mRNA cap recognition by proteins. Curr Opin Struct Biol *10*, 78-86. [10679461]

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. Cell *108*, 557-572. [11909526]

Raught, B., and Gingras, A. C. (1999). eIF4E activity is regulated at multiple levels. Int J Biochem Cell Biol *31*, 43-57. [10216943]

Ringquist, S., Schneider, D., Gibson, T., Baron, C., Bock, A., and Gold, L. (1994). Recognition of the mRNA selenocysteine insertion sequence by the specialized translational elongation factor SELB. Genes Dev *8*, 376-385. [8314089]

Rogers, G. W., Jr., Richter, N. J., Lima, W. F., and Merrick, W. C. (2001). Modulation of the helicase activity of eIF4A by eIF4B, eIF4H, and eIF4F. J Biol Chem *276*, 30914-30922. [11418588]

Rother, M., Resch, A., Gardner, W. L., Whitman, W. B., and Bock, A. (2001). Heterologous expression of archaeal selenoprotein genes directed by the SECIS element located in the 3' non-translated region. Mol Microbiol *40*, 900-908. [11401697]

Rother, M., Wilting, R., Commans, S., and Bock, A. (2000). Identification and characterisation of the selenocysteine-specific translation factor SelB from the archaeon Methanococcus jannaschii. J Mol Biol *299*, 351-358. [10860743]

Shabalina, S. A., Ogurtsov, A. Y., Rogozin, I. B., Koonin, E. V., and Lipman, D. J. (2004). Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res *32*, 1774-1782. [15031317]

Shehu-Xhilaga, M., Crowe, S. M., and Mak, J. (2001). Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity. J Virol *75*, 1834-1841. [11160682]

Shibagaki, Y., Itoh, N., Yamada, H., Nagata, S., and Mizumoto, K. (1992). mRNA capping enzyme. Isolation and characterization of the gene encoding mRNA guanylytransferase subunit from Saccharomyces cerevisiae. J Biol Chem *267*, 9521-9528. [1315757]

Skuzeski, J. M., Nichols, L. M., Gesteland, R. F., and Atkins, J. F. (1991). The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. J Mol Biol *218*, 365-373. [2010914]

Somogyi, P., Jenner, A. J., Brierley, I., and Inglis, S. C. (1993). Ribosomal pausing during translation of an RNA pseudoknot. Mol Cell Biol *13*, 6931-6940. [8413285]

Song, H., Mugnier, P., Das, A. K., Webb, H. M., Evans, D. R., Tuite, M. F., Hemmings, B. A., and Barford, D. (2000). The crystal structure of human eukaryotic release factor eRF1--mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. Cell *100*, 311-321. [10676813]

Spahn, C. M., Beckmann, R., Eswar, N., Penczek, P. A., Sali, A., Blobel, G., and Frank, J. (2001). Structure of the 80S ribosome from Saccharomyces cerevisiae--tRNA-ribosome and subunit-subunit interactions. Cell *107*, 373-386. [11701127]

Spahn, C. M., Gomez-Lorenzo, M. G., Grassucci, R. A., Jorgensen, R., Andersen, G. R., Beckmann, R., Penczek, P. A., Ballesta, J. P., and Frank, J. (2004). Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. Embo J *23*, 1008-1019. [14976550]

Spahn, C. M., and Nierhaus, K. H. (1998). Models of the elongation cycle: an evaluation. Biol Chem *379*, 753-772. [9705140]

Srinivasan, G., James, C. M., and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science *296*, 1459-1462. [12029131]

Stansfield, I., and Tuite, M. F. (1994). Polypeptide chain termination in Saccharomyces cerevisiae. Curr Genet *25*, 385-395. [8082183]

Stark, H., Rodnina, M. V., Wieden, H. J., van Heel, M., and Wintermeyer, W. (2000). Large-scale movement of elongation factor G and extensive conformational change of the ribosome during translocation. Cell *100*, 301-309. [10676812]

Steneberg, P., Englund, C., Kronhamn, J., Weaver, T. A., and Samakovlis, C. (1998). Translational readthrough in the hdc mRNA generates a novel branching inhibitor in the drosophila trachea. Genes Dev *12*, 956-967. [9531534]

Su, L., Chen, L., Egli, M., Berger, J. M., and Rich, A. (1999). Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. Nat Struct Biol *6*, 285-292. [10074948]

Sundararajan, A., Michaud, W. A., Qian, Q., Stahl, G., and Farabaugh, P. J. (1999). Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. Mol Cell *4*, 1005-1015. [10635325]

Sung, D., and Kang, H. (1998). Mutational analysis of the RNA pseudoknot involved in efficient ribosomal frameshifting in simian retrovirus-1. Nucleic Acids Res *26*, 1369-1372. [9490779]

Svitkin, Y. V., Gradi, A., Imataka, H., Morino, S., and Sonenberg, N. (1999). Eukaryotic initiation factor 4GII (eIF4GII), but not eIF4GI, cleavage correlates with inhibition of host cell protein synthesis after human rhinovirus infection. J Virol *73*, 3467-3472. [10074204]

Tang, C. K., and Draper, D. E. (1989). Unusual mRNA pseudoknot structure is recognized by a protein translational repressor. Cell *57*, 531-536. [2470510]

Tate, W. P., and Brown, C. M. (1992). Translational termination: "stop" for protein synthesis or "pause" for regulation of gene expression. Biochemistry *31*, 2443-2450. [1547227]

Tate, W. P., and Mannering, S. A. (1996). Three, four or more: the translational stop signal at length. Mol Microbiol *21*, 213-219. [8858577]

ten Dam, E., Brierley, I., Inglis, S., and Pleij, C. (1994). Identification and analysis of the pseudoknot-containing gag-pro ribosomal frameshift signal of simian retrovirus-1. Nucleic Acids Res *22*, 2304-2310. [8036158]

ten Dam, E. B., Pleij, C. W., and Bosch, L. (1990). RNA pseudoknots: translational frameshifting and readthrough on viral RNAs. Virus Genes *4*, 121-136. [2402881]

ten Dam, E. B., Verlaan, P. W., and Pleij, C. W. (1995). Analysis of the role of the pseudoknot component in the SRV-1 gag-pro ribosomal frameshift signal: loop lengths and stability of the stem regions. Rna *1*, 146-154. [7585244]

Tork, S., Hatin, I., Rousset, J. P., and Fabret, C. (2004). The major 5' determinant in stop codon read-through involves two adjacent adenines. Nucleic Acids Res *32*, 415-421. [14736996]

Tsuchihashi, Z. (1991). Translational frameshifting in the Escherichia coli dnaX gene in vitro. Nucleic Acids Res *19*, 2457-2462. [1710356]

Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc Natl Acad Sci U S A *87*, 2516-2520. [2181440]

Tsukamoto, T., Shibagaki, Y., Imajoh-Ohmi, S., Murakoshi, T., Suzuki, M., Nakamura, A., Gotoh, H., and Mizumoto, K. (1997). Isolation and characterization of the yeast mRNA capping enzyme beta subunit gene encoding RNA 5'-triphosphatase, which is essential for cell viability. Biochem Biophys Res Commun *239*, 116-122. [9345280]

Tu, C., Tzeng, T. H., and Bruenn, J. A. (1992). Ribosomal movement impeded at a pseudoknot required for frameshifting. Proc Natl Acad Sci U S A *89*, 8636-8640. [1528874]

Tujebajeva, R. M., Copeland, P. R., Xu, X. M., Carlson, B. A., Harney, J. W., Driscoll, D. M., Hatfield, D. L., and Berry, M. J. (2000). Decoding apparatus for eukaryotic selenocysteine insertion. EMBO Rep *1*, 158-163. [11265756]

Tujebajeva, R. M., Harney, J. W., and Berry, M. J. (2000). Selenoprotein P expression, purification, and immunochemical characterization. J Biol Chem *275*, 6288-6294. [10692426]

Tujebajeva, R. M., Ransom, D. G., Harney, J. W., and Berry, M. J. (2000). Expression and characterization of nonmammalian selenoprotein P in the zebrafish, Danio rerio. Genes Cells *5*, 897-903. [11122377]

Uritani, M., and Miyazaki, M. (1988). Role of yeast peptide elongation factor 3 (EF-3) at the AA-tRNA binding step. J Biochem (Tokyo) *104*, 118-126. [3065333]

Van Ryk, D. I., and Dahlberg, A. E. (1995). Structural changes in the 530 loop of Escherichia coli 16S rRNA in mutants with impaired translational fidelity. Nucleic Acids Res *23*, 3563-3570. [7567470]

von Der Haar, T., Ball, P. D., and McCarthy, J. E. (2000). Stabilization of eukaryotic initiation factor 4E binding to the mRNA 5'-Cap by domains of eIF4G. J Biol Chem *275*, 30551-30555. [10887196]

Walczak, R., Westhof, E., Carbon, P., and Krol, A. (1996). A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. Rna *2*, 367-379. [8634917]

Wang, Z., Day, N., Trifillis, P., and Kiledjian, M. (1999). An mRNA stability complex functions with poly(A)-binding protein to stabilize mRNA in vitro. Mol Cell Biol *19*, 4552-4560. [10373504]

Wang, Z., and Kiledjian, M. (2000). The poly(A)-binding protein and an mRNA stability protein jointly regulate an endoribonuclease activity. Mol Cell Biol *20*, 6334-6341. [10938110]

Wassarman, P. M., and Kinloch, R. A. (1992). Gene expression during oogenesis in mice. Mutat Res *296*, 3-15. [1279405]

Weiner, A. M., and Weber, K. (1971). Natural read-through at the UGA termination signal of Q-beta coat protein cistron. Nat New Biol *234*, 206-209. [5288807]

Weiss, R., and Gallant, J. (1983). Mechanism of ribosome frameshifting during translation of the genetic code. Nature *302*, 389-393. [6339944]

Weiss, R. B., Dunn, D. M., Dahlberg, A. E., Atkins, J. F., and Gesteland, R. F. (1988). Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli. Embo J *7*, 1503-1507. [2457498]

Weiss, R. B., Dunn, D. M., Shuh, M., Atkins, J. F., and Gesteland, R. F. (1989). E. coli ribosomes re-phase on retroviral frameshift signals at rates ranging from 2 to 50 percent. New Biol *1*, 159-169. [2562219]

Weiss, R. B., Huang, W. M., and Dunn, D. M. (1990). A nascent peptide is required for ribosomal bypass of the coding gap in bacteriophage T4 gene 60. Cell *62*, 117-126. [2163764]

Wells, S. E., Hillner, P. E., Vale, R. D., and Sachs, A. B. (1998). Circularization of mRNA by eukaryotic translation initiation factors. Mol Cell *2*, 135-140. [9702200]

Wills, N. M., Gesteland, R. F., and Atkins, J. F. (1994). Pseudoknot-dependent read-through of retroviral gag termination codons: importance of sequences in the spacer and loop 2. Embo J *13*, 4137-4144. [8076609]

Wilson, K. S., Ito, K., Noller, H. F., and Nakamura, Y. (2000). Functional sites of interaction between release factor RF1 and the ribosome. Nat Struct Biol *7*, 866-870. [11017194]

Wilting, R., Schorling, S., Persson, B. C., and Bock, A. (1997). Selenoprotein synthesis in archaea: identification of an mRNA element of Methanococcus jannaschii probably directing selenocysteine insertion. J Mol Biol *266*, 637-641. [9102456]

Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. Nature *407*, 327-339. [11014182]

Wolin, S. L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. Embo J *7*, 3559-3569. [2850168]

Xu, Z., Choi, J., Yen, T. S., Lu, W., Strohecker, A., Govindarajan, S., Chien, D., Selby, M. J., and Ou, J. (2001). Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. Embo J *20*, 3840-3848. [11447125]

Yelverton, E., Lindsley, D., Yamauchi, P., and Gallant, J. A. (1994). The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in Escherichia coli. Mol Microbiol *11*, 303-313. [8170392]

Yonath, A. (2002). High-resolution structures of large ribosomal subunits from mesophilic eubacteria and halophilic archaea at various functional States. Curr Protein Pept Sci *3*, 67-78. [12370012]

Yoon, H., and Donahue, T. F. (1992). Control of translation initiation in Saccharomyces cerevisiae. Mol Microbiol *6*, 1413-1419. [1625572]

Yun, D. F., Laz, T. M., Clements, J. M., and Sherman, F. (1996). mRNA sequences influencing translation and the selection of AUG initiator codons in the yeast Saccharomyces cerevisiae. Mol Microbiol *19*, 1225-1239. [8730865]

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H., and Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 A resolution. Science *292*, 883-896. [11283358]

Yusupova, G. Z., Yusupov, M. M., Cate, J. H., and Noller, H. F. (2001). The path of messenger RNA through the ribosome. Cell *106*, 233-241. [11511350]

Zhang, S., Ryden-Aulin, M., and Isaksson, L. A. (1998). Functional interaction between tRNA2Gly2 at the ribosomal P-site and RF1 during termination at UAG. J Mol Biol *284*, 1243-1246. [9878344]

Zinoni, F., Heider, J., and Bock, A. (1990). Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. Proc Natl Acad Sci U S A *87*, 4660-4664. [2141170]

Résultats

2 Résultats

2.1 Raffinement du modèle

Ce travail a été réalisé en collaboration avec les équipes de Bioinformatique des Génomes de l'Institut de Générique et Microbiologie et de Bioinformatique du Laboratoire de Recherche en Informatique (Université Paris-Sud)

2.1.1 Résumé

Les gènes présentant un événement de décalage de phase de lecture en -1 ont un message génétique qui, par définition, n'est pas en phase. Ils risquent donc d'être répertoriés comme séquences non-codantes ou pseudogènes. Il a été montré que le mécanisme est corrélé à la présence de signaux particuliers au sein de la séquence codante (Jacks *et al.*, 1988). Cependant, il apparaît que, d'un organisme à l'autre, les signaux présentent à la fois des propriétés communes et des dissemblances. Pour permettre la recherche systématique des gènes soumis à décalage dans les génomes séquencés, une modélisation fine de ces signaux doit être réalisée. Des travaux dans ce sens ont déjà été effectués (Hammell *et al.*, 1999; Liphardt, 1999). Cependant, les modèles proposés n'ont pas permis pour l'instant de trouver de nouveaux sites authentiques de décalage.

Le but de ce travail a donc été la conception d'un modèle plus élaboré, notamment en tenant compte de paramètres précédemment négligés et en considérant le décalage dans son aspect probabiliste. À la suite des travaux de Ian Brierley et d'Amy Hammell (Brierley *et al.*, 1989; Brierley *et al.*, 1991; Hammell *et al.*, 1999) une étude systématique de la structure secondaire a été réalisée, uniquement fondée sur une modélisation des interactions et pas sur la structure tridimensionnelle. Pour ce faire, nous avons décrit plus finement cette structure en y ajoutant des attributs supplémentaires, principalement d'ordre « linguistique » (fréquences des nucléotides dans les différentes parties de la structure). Nous avons fait appel à un outil d'apprentissage des concepts « il y a » vs. « il n'y a pas » décalage pour tenter d'inférer de nouvelles règles liant, dans un premier temps, description de la structure secondaire et taux de décalage.

Cette méthode d'étude a permis d'identifier les nouvelles caractéristiques impliquées dans le décalage de phase de lecture en -1. Il a ainsi été mis en évidence l'importance de la dissymétrie de répartition des appariements `C-G` et `G-C` dans la première tige du pseudonœud et l'importance de la composition nucléotidique de l'espaceur. L'utilisation de l'outil d'apprentissage GloBo (Torre, 2000) a mené à une ensemble de règles de décisions, dont les deux règles majeures ont pu être vérifiées expérimentalement. La pertinence de leurs attributs avec le mécanisme de décalage de phase de lecture a été examinée en créant les sites artificiels qui suivaient les règles inférées par l'approche bioinformatique. La capacité de décalage de phase de lecture a été évaluée *in vivo* chez la levure *S. cerevisiæ*. L'originalité de ces règles provient de leur forme conjonctive. Chaque règle fournit un ensemble de caractéristiques qui doit être respecté pour répondre au modèle. Nos résultats montrent ainsi que l'*espaceur* est impliqué dans l'efficacité du décalage de phase. Bien que l'on connaisse depuis longtemps l'importance de la longueur de cette région, ce n'est que récemment qu'il a été montré que la séquence elle-même pouvait également être importante dans le cas des décalages de phase bactériens (Bertrand *et al.*, 2002). Des séquences artificielles ont été générées et évaluées *in vivo* afin de valider ces règles. Les résultats présentés ici démontrent que la séquence est également importante chez les eucaryotes. Différents mécanismes peuvent expliquer cet effet : les nucléotides peuvent agir directement avec des composants de la machinerie de traduction (ARN ou protéines ribosomiques), ou indirectement, par interaction codon/anticodon, ou par la disponibilité des ARNt correspondants. La séquence de l'espaceur doit être considérée comme une structure stimulatrice modulant l'efficacité de décalage de phase de lecture. Un autre élément identifié au cours de cette étude est le rôle multiplicatif du triplet `XXX` dans l'efficacité de

déphasage. Ceci suggère que le signal pourrait en fait être un tetra-nucléotide et que le triplet XXX est utilisé pour en moduler l'efficacité. Il est établi que le mécanisme de décalage de phase bactérien peut impliquer seulement un tétramère de la forme Y-YYZ. Chez les eucaryotes, bien que le glissement en tandem des ARNt soit le mécanisme principal, le dérapage peut avoir lieu alors que seul le site P du ribosome est occupé (Baranov *et al.*, 2004; Jacks *et al.*, 1988; Yelverton *et al.*, 1994).

2.1.2 Article

Towards a computational model for -1 eukaryotic frameshifting sites
Bekaert M., Bidou L., Denise A, Duchateau-Nguyen G., Forest JP., Froidevaux C., Hatin I., Rousset JP. & Termier M.
Bioinformatics 2003 **19**(3):327-35.

- *Voir annexe 1* -

2.1.3 Recherche de structures secondaires

2.1.4.1 Méthodologie

Après avoir extrait des règles qui impliquent un décalage de phase élevé, j'ai construit des mutants à partir de sites avérés qui respectaient ces règles. Nous avons ensuite recherché de nouvelles séquences dans le génome de la levure. Les séquences identifiées ont été classées et évaluées *in vivo*.

L'objectif de cette approche est de concevoir une étape de filtrage capable de sélectionner des régions susceptibles de produire un décalage de phase de lecture dans les génomes complets. La méthodologie employée a été la suivante : d'abord des séquences glissantes (heptamères en phase avec un codon d'initiation) ont été recherchées dans le génome de la levure à l'aide d'un automate. Dans un second temps un pseudonœud a été recherché en aval de cette séquence à l'aide d'un programme spécifique établissant si la séquence en aval est susceptible de se replier comme un site avéré, à l'aide d'un algorithme de prédiction de structures secondaires (algorithme Orpheo ; J-P Forest, communication personnelle). Dans un second temps, ces régions identifiées ont été évaluées et classées à l'aide des règles précédemment identifiées (ce qui inclut les 2 règles principales exposées dans l'article précédent, mais aussi d'autres moins prédictives, mais importantes à cette étape).

L'utilisation de systèmes d'apprentissage supervisé tel que GloBo (algorithme stochastique) (Bekaert *et al.*, 2003; Torre, 2000), Naives-

Bayes, C4.5 (algorithmes symboliques) ou M5' (algorithme numérique) appliqués à un ensemble de candidats potentiels, destinés à être testés expérimentalement, ont permis d'ordonner ces candidats en fonction de leurs chances de succès. Cette approche a mené à des règles et des prédictions qui ont put être vérifiées expérimentalement. La pertinence de leurs attributs avec le mécanisme de décalage de phase de lecture a été examinée en créant les sites artificiels qui suivaient les règles inférées par l'approche bioinformatique. Cette capacité de décalage de phase de lecture a été évaluée *in vivo* chez la levure *S. cerevisiæ*. La particularité de ces règles provient de leur forme conjonctive. À savoir, chaque règle fournit un ensemble de caractéristiques qui doivent être respectées pour répondre au modèle. L'originalité de ce travail repose sur l'utilisation de matrice d'alignement asymétrique permettant de respecter la dissymétrie C-G / G-C déjà évoquée et la procédure de vote basée sur des règles précédemment extraites afin d'ordonner les candidats. Ce travail faisant l'objet de la thèse de Jean-Paul Forest, je n'entrerai pas plus dans les détails de cette approche.

2.1.4.2 Résultats préliminaires

Cette approche a fourni 185 régions candidates au décalage de phase de lecture en -1 chez la levure *S. ceresisiæ*. Les 6 séquences les mieux classées ont été évaluées *in vivo*. Afin d'avoir une estimation à la fois du nombre de séquences identifiables par hasard (faux positif) et de la qualité du modèle utilisé, nous avons engendré un génome aléatoire de levure a l'aide du logiciel GenRGenS (Denise *et al.*, 2003) selon un modèle markovien, dont les paramètres ont été calculés sur la fréquence en hexamères du génome de *S. cerevisiæ*. Dans un génome aléatoire de la même taille que le sauvage, 104 candidats ont été obtenus. Les 3 séquences les mieux classées ont été évaluées *in vivo*.

Source	Position	ID	Efficacité
S. cerevisiæ	VI. 240146.c	54	1%±0
S. cerevisiæ	XIII.381816.c	129	1%±0
S. cerevisiæ	XIV.392491.w	135	5%±1
S. cerevisiæ	IV.1016654.w	20	7%±1
S. cerevisiæ	VII.557876.w	67	>0,1%
S. cerevisiæ	VII.678122.w	186	13%±1
Aléatoire	-	86	5%±1
Aléatoire	-	16	>0,1%
Aléatoire	-	37	>0,1%

Tableau 1 : Résultats des efficacités de décalage des séquences identifiées. Les séquences issues du génome de la levure présentent 3 séquences actives. Le génome aléatoire en présente 1.

Nous avons découvert 3 nouveaux sites de décalage de phase de lecture en -1. Notre étude est la première qui ait détecté autant de sites simultanément (Tableau 1). De plus, d'autres candidats bien classés sont probablement également pertinents. Par ailleurs, le génome aléatoire a produit au moins une séquence active. Ce résultat valide le modèle utilisé, puisqu'à partir d'une séquence purement aléatoire il nous est possible de retrouver au moins une séquence fonctionnelle *in vivo*.

2.1.4.3 Le meilleur candidat

Le meilleur candidat (13% de décalage de phase de lecture) de cette approche est aussi retrouvé dans l'approche présentée ci-dessous (2.3). Une étude plus approfondie a été initiée afin de caractériser ce gène de fonction inconnue. L'ARNm a été isolé, et l'ADNc séquencé entièrement en 5' et 3'. La Figure 1 schématise ce messager.

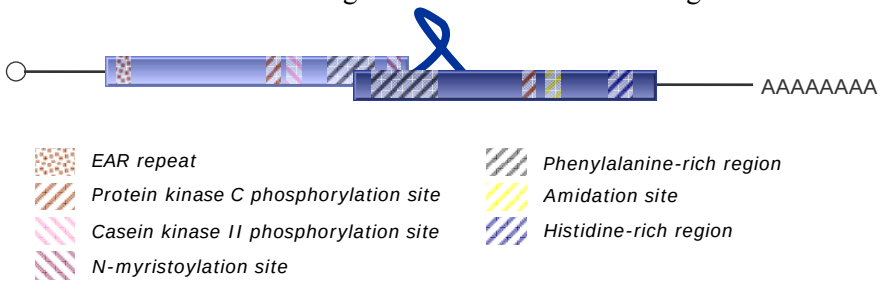


Figure 1 : ARNm de PRF13 (*Programmed ribosomal frameshift 13%*).

Aucune fonction n'a pu être attribuée à cette structure. Les motifs ne m'ont pas permis d'identifier une fonction probable. Aucune structure similaire n'a été identifiée dans d'autres génomes. J'ai donc initié une analyse fonctionnelle en commençant par réaliser la délétion complète du gène ; cette dernière est viable et ne présente pas de déficience de croissance. Une caractérisation plus complète de la protéine elle-même est nécessaire afin d'évaluer la fonction des formes courtes et étendues de cette dernière ainsi que du rôle joué par la régulation par décalage de phase de lecture en -1. A ce stade je ne peux pas encore exclure que cette région n'ait aucune fonction biologique réelle (pseudogène).

2.2 Caractérisation de sites viraux et de l'influence du site E

2.2.1 Résumé

Le but initial de ce travail était de caractériser de nouveaux sites viraux de décalage de phase de lecture en -1 afin d'enrichir notre base de données et de les inclure dans de futures modélisations. J'ai mesuré

l'efficacité de déphasage de 20 sites viraux identifiés uniquement sur la base de leur séquence et donc seulement présumés efficaces. À partir de ces résultats, j'ai développé une recherche linguistique fondée sur un modèle de Markov caché (HMM). Cet algorithme m'a permis de détecter 74 sites viraux de décalage de phase de lecture en -1 parmi tous les virus entièrement séquencés disponibles. L'alignement des séquences a mis en évidence un fort biais de composition juste en amont de l'heptamère glissant. Afin de déterminer le rôle de ces deux nucléotides en amont de l'heptamère dans le décalage de phase de lecture en -1, nous avons construit dans un vecteur rapporteur les 16 séquences possibles dans le contexte du coronavirus aviaire IBV (virus modèle dans l'étude du décalage de phase en -1, chez lequel un très grand nombre de mutagenèses a été réalisé ; Bekaert *et al.*, 2003; Brierley *et al.*, 1992; Brierley *et al.*, 1991). L'estimation du taux de décalage a montré l'existence de deux groupes de séquences, l'un induisant un décalage *fort* (~20%), l'autre un décalage *faible* (~10%). Cette région affecte donc l'efficacité de décalage de phase de lecture en -1 chez *S. cerevisiæ*. Ces résultats ont aussi révélé le rôle de la pseudouridine en position 39 de l'ARNt situé au site E du ribosome de la cellule hôte dans l'efficacité de décalage de phase. En utilisant des mutants Δ*PUS3* de la levure, incapables de modifier les positions 38/39 des ARNt, j'ai prouvé que l'efficacité de décalage était modulée par l'état de modification de l'ARNt au site E. Deux hypothèses peuvent expliquer le rôle de l'ARNt au site E dans le décalage : Soit le décalage pourrait être augmenté par l'absence d'ARNt au site E ; la modification Ψ39 déstabiliserait l'interaction ARNt/site E ; soit Ψ39 interfère, directement ou indirectement sur l'interaction de l'ARNt au site P avec le mRNA, diminuant la stabilité de l'appariement.

Mes résultats démontrent que la région glissante des signaux de décalage de phase de lecture en -1 est plus complexe que ce qui était précédemment établi. Les trois sites du ribosome semblent interférer avec le mécanisme du décalage. Comparé au modèle initial (Jacks *et al.*, 1988), les deux régions en 3' et 5' de l'heptamère sont maintenant impliquées dans l'efficacité de décalage de phase par des interactions entre l'ARNt, l'ARNm et le ribosome.

2.2.2 Article

An extended signal involved in eukaryotic -1 frameshifting operates through modification of the E-site tRNA

Bekaert M. & Rousset JP.

Mol. Cell 2005 **17**(1):61-68.

- *Voir annexe 2* -

2.2.3 Complément d'informations

Les 74 virus identifiés se répartissent en genres et familles distincts. La liste ci-dessous reprend les noms et les séquences glissantes de chaque virus. Deux virus sont non assignés. La structure de leur heptamère et les régions les bordant permettent cependant de classer le *Sugarcane yellow leaf virus* et l'*Acyrthosiphon pisum virus* comme respectivement Polerovirus et Sobemovirus.

Nom du virus	GenBank	Région glissante (5' vers 3')	Locus
dsRNA virus			
Totiviridae			
Giardiavirus			
Giardia lamblia virus	NC_003555	CCUGCGCCAU CCCUUUA UCCGAUCGUG	2841
Trichomonas vaginalis virus 3	NC_004034	GCGCGUAUCA GGGCCCU CGCUUGCAGG	2450
Trichomonas vaginalis virus II	NC_003873	CUGCCUACCA GGGCCCC AGCUUCGCGC	2382
Totivirus			
Saccharomyces cerevisiae virus L-A	NC_003745	GUACUCAGCA GGGUUUA GGAGUGGUAG	1964
Saccharomyces cerevisiae virus L-BC	NC_001641	CUGAGAAGUU GGAUUUU CGUGUAGCAG	1973
Helminthosporium victoriae virus 190S	NC_003607	CUGAUCGGGC CGAGGGA CAAUGAGUGA	2602
Retroid virus			
Retroviridae			
Alpharetrovirus			
Avian leukosis virus	NC_001408	UCCGCUUGAC AAAUUUA UAGGGAGGGC	2475
Rous sarcoma virus	NC_001407	UCCGCUUGAC AAAUUUA UAGGGAGGGC	2482
Betaretrovirus			
Enzootic nasal tumour virus of goats	NC_004994	CCCCGGUUUC GGGAAAC UGGGUGAGGG	1953*
Mason-Pfizer monkey virus	NC_001550	CACCCCAUCA GGGAAAC GGGAUGAGGG	2092*
Mouse mammary tumor virus	NC_001503	CUGAAAAUUC AAAAAAC UUGUAAAGGG	2082*
Ovine pulmonary adenocarcinoma virus	NC_001494	CCCCGGUUUC GGGAAAC UGGGUGAGGG	1960*
Simian SRV-1 type D retrovirus	NC_001551	CACCCCAUCA GGGAAAC GGACUGAGGG	2329*
Deltaretrovirus			
Bovine leukemia virus	NC_001414	CCCUCAAAUC AAAAAAC UAAUAGAGGG	1596*
Human T-lymphotropic virus 1	NC_001436	UCCCACACCC AAAAAAC UCCAUAGGGG	1718*
Human T-lymphotropic virus 2	NC_001488	CUACUGAGGA AAAAAAC UCCUUAAGGG	2087*
Primate T-lymphotropic virus 3	NC_003323	UCCGGGAGCA AAAAAAC UCCUCAGGGG	2001*
Simian T-lymphotropic virus 1	NC_000858	UCCCACACCC AAAAAAC UCCAUAGGGG	2067*
Simian T-lymphotropic virus 2	NC_001815	CCACCGAGGA AAAAAAC UCCCUGGGGG	2033*
Lentivirus			
Bovine immunodeficiency virus	NC_001413	ACUGCAGGUC AAAAAAC GGGAAGUGCU	1635
Caprine arthritis-encephalitis virus	NC_001463	GAAAACAGCA GGGAAAC GGGAGGAGGG	1810
Equine infectious anemia virus	NC_001450	GAAGUGUUCC AAAAAAC GGGAGCAAGG	1787
Feline immunodeficiency virus	NC_001482	GAAAGAAUUC GGGAAAC UGGAAGGCGG	1884
Human immunodeficiency virus 1	NC_001802	GACAGGCUAA UUUUUUA GGGAAGAUCU	1637
Human immunodeficiency virus 2	NC_001722	GACAGGCAGG UUUUUUA GGGUUGGGCC	2401
Jembrana disease virus	NC_001654	ACUGCAAAUC AAAAAAC GGGAGGCGCU	1441
Ovine lentivirus	NC_001511	CAUCACAGCA GGGAAAC AGCAGGAGGG	1812
Simian immunodeficiency virus	NC_001549	GACAGGCAAA UUUUUUA GGGUAUGGCC	2198
Simian immunodeficiency virus 2	NC_004455	AGAUGGUGAA UUUUUUA GGGAAUACCC	2056
Simian-Human immunodeficiency virus	NC_001870	GACAGGCGGG UUUUUUA GGCCUUGGUC	1840
Visna virus	NC_001452	AGAAACAGCA GGGAAAC AACAGGAGGG	1769
ssRNA+ virus			
Astroviridae			
Avastrovirus			
Avian nephritis virus	NC_003790	UUUGUAAGUC AAAAAAC UAAAUGACCC	3025
Turkey astrovirus	NC_002470	CUACGUGUUC AAAAAAC UAGAUAGUCA	3307
Mamastrovirus			

Human astrovirus	NC_001943	ACAAGGCCCC AAAAAAC UACAAAGGGC	2845
Mink astrovirus	NC_004579	AGCAGAAGCC AAAAAAC GGGAAGAGGG	2621
Ovine astrovirus	NC_002469	UCCAGCAGCC AAAAAAC UCCAAAGGGG	2535
Luteoviridae			
Enamovirus			
Pea enation mosaic virus-1	NC_003629	CCAGACGCUC GGGAAAC GGAUUAUUCC	2035
Luteovirus			
Barley yellow dwarf virus - GAV	NC_004666	UUGACUCUGU GGGUUUU UAGAGGGGCU	1159
Barley yellow dwarf virus - MAV	NC_003680	UUGACUCUGU GGGUUUU AGAGGGGCUC	1122
Barley yellow dwarf virus - PAS	NC_002160	UUGACUCUGU GGGUUUU UAGAGGGGCU	1158
Bean leafroll virus	NC_003369	UCACCUCAUC GGUUUUU UAGAGGGGCU	1282
Soybean dwarf virus	NC_003056	UAAACGCUGA GGGUUUU UAGAGGGGCU	1226
Polerovirus			
Beet chlorosis virus	NC_002766	CACAUCUGCC GGGAAAU GGACUGAGCG	1587
Beet mild yellowing virus	NC_003491	CCGGAACAAC CGGAAAC GCAAGCACCC	1591
Beet western yellows virus	NC_004756	CCAAGAGCUC GGGAAAC GGGAGAGCGG	1481
Cereal yellow dwarf virus - RPS	NC_002198	CCGGAAAGUC GGGAAAC GCCAAGGCGG	1602
Cereal yellow dwarf virus - RPV	NC_004751	AAGACGAGUC GGGAAAC GGGAAGGCGG	1699
Cucurbit aphid-borne yellows virus	NC_003688	AAUACGAGUC GGGAAAC GGGCAGGCGG	1488
Potato leafroll virus	NC_001747	CAAACAAGCC GGGAAAU GGGCAAGCGG	1774
Turnip yellows virus	NC_003743	AAGAUCUGUC GGGAAAC GGAGUGCGCG	1559
Unassigned Luteoviridae			
Sugarcane yellow leaf virus	NC_000874	CUCCAGACCA GGGAAAU GAGCCAAGUG	1759
Nidovirales / Arteriviridae			
Arterivirus			
Equine arteritis virus	NC_002532	CAGUGAAUCA GUUAAAC UGAGAGCGCC	5405
Lactate dehydrogenase-elevating virus	NC_002534	AGGCAUCGGC UUUAAAC UGCUAGCCAC	6836
Porcine reproductive and respiratory syndrom virus	NC_001961	AGGAGCAGUG UUUAAAC UGCUAGCCGC	7695
Simian hemorrhagic fever virus	NC_003092	CAUCUGAAGC UUUAAAC UGCUAACCGC	6521
Nidovirales / Coronaviridae			
Coronavirus			
Avian infectious bronchitis virus	NC_001451	AUAAGAAUUA UUUAAAC GGGUACGGGG	12354
Bovine coronavirus	NC_003045	AUACUAAUUU UUUAAAC GGGUUCGGGG	13341
Human coronavirus 229E	NC_002645	AUAACAGUUA UUUAAAC GAGUCCGGGG	12520
Human coronavirus OC43	NC_005147	ATACUAAUUU UUUAAAC GGGUUCGGGG	13341
Murine hepatitis virus	NC_001846	ACACGAACUU UUUAAAC GGGUUCGGGG	13601
Porcine epidemic diarrhea virus	NC_003436	AUAUGGCUUA UUUAAAC GAGUACGGGG	12620
SARS coronavirus	NC_004718	CAUCAACGUU UUUAAAC GGGUUUGCGG	13398
Transmissible gastroenteritis virus	NC_002306	AUCAAAGUUA UUUAAAC GAGUGCGGGG	12338
Tombusviridae			
Dianthovirus			
Carnation ringspot virus RNA 1	NC_003530	AAUCCCUCGA GGAUUUU UAAGUGCCCC	765
Red clover necrotic mosaic virus RNA 1	NC_003756	AAUCCCUUGA GGAUUUU UAGGCGGCCC	831
Sweet clover necrotic mosaic virus RNA 1	NC_003806	AAUCCCUUGA GGAUUUU UAGGCGGCGG	828
Sobemovirus			
Cocksfoot mottle virus	NC_002618	CAAUCCGGCC UUUAAAC UACCAGCGGG	1640
Subterranean clover mottle virus	NC_004346	UGCUCGAGCA UUUAAAC UGCCAGCGGG	1852
Turnip rosette virus	NC_004553	CAGUGAGCUC UUUAAAC UGCCAGCGGG	1757
Umbravirus			
Carrot mottle mimic virus	NC_001726	CACCCAUCCG GGAUUUU UACUAGGGGA	1031
Groundnut rosette virus	NC_003603	CCGGGGCACA AAAUUUU UAGUUGGGGA	867
Pea enation mosaic virus-2	NC_003853	GGCGCGCGGC GGAUUUU UGGUAGGGGC	923
Tobacco bushy top virus	NC_004366	GUGGGCCCAA GGAUUUU UGCUAGGGGA	952
Unassigned			
Acyrthosiphon pisum virus	NC_003780	AAGGCUAUUC UUUAAAC UUCUAGCCCC	8154

Tableau 2 : Liste des virus identifiés. La colonne locus indique la position du nucléotide lu deux fois par le ribosome (base Z de l'heptamère). L'étoile (*) correspond aux virus ayant deux sites de décalage de phase de lecture en -1 ; dans ce cas seule la jonction entre Gag et Pro a été considérée.

2.3 Recherche de sites de recodage par une approche sans *a priori*

Ce travail a été réalisé en collaboration avec le Laboratoire de Statistique et Génome (Université d'Evry).

2.3.1 Résumé

Dans cette étude, est présenté un système d'identification de gènes dont l'expression est contrôlée par décalage de phase de lecture en -1, sans connaissance *a priori* du mécanisme impliqué. Deux approches indépendantes ont été utilisées *in silico* sur le génome de *S. cerevisiæ*. La première est fondée sur l'identification de régions génomiques dont la traduction produirait des motifs protéiques identifiables qui peuvent être associés sur un même polypeptide par un événement de décalage de phase de lecture en -1. La seconde approche utilise une méthodologie linguistique fondée sur la fréquence d'utilisation des codons afin de discriminer les régions codantes des régions non codantes. A après avoir identifié des candidats potentiels à l'aide d'un modèle de Markov caché, ceux-ci sont classés par un calcul de vraisemblance. Cette stratégie ne se fonde sur aucun modèle de site de décalage de phase et est adaptée à la détection *de novo* d'événements de décalage de phase de lecture en -1.

En utilisant ce système, nous avons sélectionné 186 régions candidates. J'ai alors vérifié l'expression des ARNm et évalué l'efficacité de déphasage *in vivo*. Finalement 11 régions, sur les 55 testées, présentent effectivement un décalage de phase de lecture au moins 50 fois au dessus du bruit de fond. Plusieurs de ces candidats sont des gènes avec des fonctions connues, qui permettront d'analyser le rôle physiologique de l'événement de décalage de phase. De façon générale, ces résultats démontrent que le décalage de phase de lecture en -1 est un événement plus fréquent chez les eucaryotes qu'il n'était jusqu'alors proposé.

2.3.2 Article en préparation

- *Voir annexe 3* -

2.3.3 Complément d'informations

Cette approche a conduit à de nombreux tests *in vivo* afin d'identifier les candidats et de vérifier la validité des modèles utilisé. Le tableau ci-dessous reprend l'ensemble des analyses. Devant le nombre important de RT-PCR positives sur des ARNm des candidats les plus prometteurs, nous avons vérifié la pertinence de ce crible en recherchant les

ARNm chez les candidats les moins probables (« *Témoin RT-PCR négative* ») sur 10 tests 1 seul a été positif, validant bien l'approche.

fsORF	Source	Chr.	Position	ADNg	ARNm	ADNc	%	Notes
1*	Motif	I	192541-196178	+1 nt	-	-	**-**	Ré-annoté dans SGD
10	HMM	IV	205690-205988	oui	oui	oui	**0.1**	
11*	HMM	IV	384077-381986	oui	oui	oui	**11.0**	
12	HMM	IV	630075-630598	oui	intron	-	**-**	
13	HMM	V	183582-183327	oui	non	-	**-**	Temoin RT-PCR négative
14*	Motif	V	298948-301706	oui	non	-	**-**	
15*	HMM	VI	123462-129904	oui	?	?	**?**	
16	Motif/HMM	VI	15473-14309	oui	oui	oui	**9.0**	
17*	Motif	VII	1068995-1067213	oui	non	-	**-**	
18	HMM	VII	146543-146769	oui	non	-	**-**	Temoin RT-PCR négative
19*	Motif	VII	270340-267730	oui	oui	oui	**?**	
2	Motif	II	289386-290383	oui	oui	oui	**6.0**	
20	Motif	VII	425616-425971	oui	oui	oui	**?**	
21	Motif/2D	VII	677871-678301	oui	oui	oui	**13.0**	
22	HMM	VIII	262554-262197	oui	oui	oui	**0.1**	
23	HMM	VIII	35126-34916	oui	non	-	**-**	Temoin RT-PCR négative
24	HMM	VIII	499891-499585	oui	non	-	**-**	
25*	HMM	X	219713-217406	oui	oui	?	**?**	
26*	HMM	X	405173-406968	+ 1 nt	-	-	**-**	Ré-annoté dans SGD
27	HMM	X	732756-732555	oui	non	-	**-**	control
28	HMM	X	74021-74610	oui	intron	-	**-**	
29	Motif	XI	172169-171299	oui	oui	oui	**0.1**	
3	HMM	II	554266-553504	+ 1 nt	-	-	**-**	Ré-annoté dans SGD
30	HMM	XI	374144-374853	oui	oui	oui	**12.0**	
31*	Motif	XI	549085-551003	+ 1 nt	-	-	**-**	Ré-annoté dans SGD
32	HMM	XI	611160-611899	oui	oui	oui	**7.0**	
33	HMM	XI	639597-638535	+ 1 nt	-	-	**-**	Ré-annoté dans SGD
34*	Motif	XII	200413-200654	oui	non	-	**-**	
35	Motif	XII	203255-204786	oui	oui	oui	**?**	
36	HMM	XII	767116-766933	oui	non	-	**-**	Temoin RT-PCR négative
37*	Motif	XII	857539-861524	oui	non	-	**-**	
38*	HMM	XIII	263477-266754	oui	oui	?	**?**	
39	Motif	XIII	349605-348426	oui	oui	oui	**?**	
4	Motif	II	701799-700347	oui	non	-	**-**	
40*	Motif/HMM	XIII	436627-438788	oui	oui	oui	**5.0**	
41*	Motif	XIII	509318-507416	oui	oui	oui	**5.0**	
42	Motif	XIII	623212-622159	oui	oui	oui	**0.1**	
43	Motif	XIII	650035-651026	oui	oui	oui	**10.0**	
44	HMM	XIV	394359-394026	oui	oui	oui	**?**	
45	Motif	XIV	40618-42065	oui	non	-	**-**	
46	HMM	XIV	429214-428983	oui	non	-	**-**	Temoin RT-PCR négative
47	HMM	XIV	537790-538010	oui	oui	oui	**?**	
48	Motif	XV	742910-744210	oui	oui	oui	**5.0**	
49	Motif	XV	758330-759354	oui	non	-	**-**	
5*	Motif	III	200170-197617	oui	oui	oui	**0.1**	
50	Motif	XV	1026837-1028101	oui	oui	oui	**7.0**	
51	HMM	XV	782222-782003	oui	non	-	**-**	
52	HMM	XV	80639-81189	oui	intron	-	**-**	
53	Motif	XVI	117365-117062	oui	oui	oui	**0.1**	
54	Motif	XVI	138830-139449	oui	oui	oui	**3.0**	
55	HMM	XVI	935319-935028	oui	non	-	**-**	Temoin RT
6*	HMM	III	220178-218372	oui	non	-	**-**	Temoin RT
7	HMM	III	222829-223097	oui	oui	oui	**0.1**	Temoin RT
8	HMM	III	91686-91455	oui	non	-	**-**	Temoin RT
9*	Motif	IV	167806-164992	oui	oui	oui	**3.0**	

Tableau 3 : Liste des régions candidates testées - résultats préliminaires (1 seul dosage). Source : Approche utilisée ; Position : Numéro du chromosome et coordonnée du candidat ; ADNg : Séquence de la région sur le Génome (souche FY) ;ARNm : Détection d'un ARNm couvrant la totalité de la structure ;ADNc : Séquence de la l'ARNm ; % : Efficacité de décalage de phase.

Peu d'erreurs de séquence ont été observées, généralement la séquence téléchargée sur GenBank était correcte, les seuls cas d'erreurs de séquence étaient déjà ré-annotés sur le site du SGD. Trois introns « non annotés » ont aussi été identifiés. Le résultat important est l'identification de 11 séquences transcrites et porteuses de régions induisant un taux de déphasage supérieur à 5%.

Une extension en cours de ce travail est d'utiliser des approches similaire sur d'autres génomes mais aussi, au prix de quelques modifications du modèle d'étendre les recherches à d'autres types de recodage (décalage de phase de lecture en +1, translecture).

2.4 Références

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2004). P-site tRNA is a crucial initiator of ribosomal frameshifting. Rna 10, 221-230. [14730021]

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics 19, 327-335. [12584117]

Bertrand, C., Prere, M. F., Gesteland, R. F., Atkins, J. F., and Fayet, O. (2002). Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression. Rna 8, 16-28. [11871658]

Brierley, I., Digard, P., and Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell 57, 537-547. [2720781]

Brierley, I., Jenner, A. J., and Inglis, S. C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. J Mol Biol 227, 463-479. [1404364]

Brierley, I., Rolley, N. J., Jenner, A. J., and Inglis, S. C. (1991). Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. J Mol Biol 220, 889-902. [1880803]

Denise, A., Ponty, Y., and Termier, M. (2003). Random generation of structured genomic sequences, Paper presented at: Recomb'03 (Berlin).

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999). Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res 9, 417-427. [10330121]

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell 55, 447-458. [2846182]

Liphardt, J. (1999) The mechanism of -1 ribosomal frameshifting: experimental and theoretical analysis, Ph.D., Churchill College, Cambridge.

Torre, F. (2000) Intégration des biais de langage à l'algorithme générer-et-tester - Contributions à l'apprentissage disjonctif, Ph.D., Univeristé Paris-Sud, Orsay.

Yelverton, E., Lindsley, D., Yamauchi, P., and Gallant, J. A. (1994). The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in Escherichia coli. Mol Microbiol 11, 303-313. [8170392]

Implémentation

3 Implémentation

3.1 Préambule

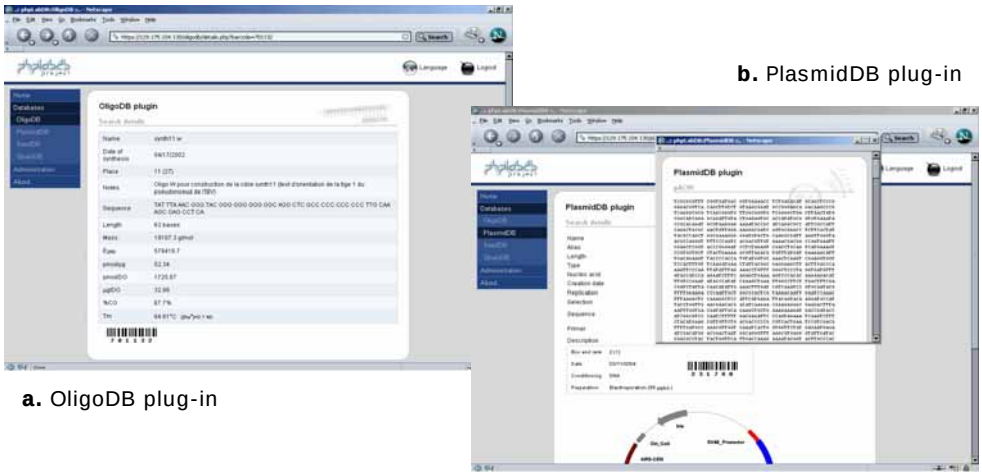
Afin de réaliser l'ensemble des travaux présentés, plusieurs systèmes logiciels ont été développés. Ils s'articulent autour d'une première suite d'utilitaires intégrés (phpLabDB) qui permet de traiter l'ensemble des bases de données du laboratoire et la recherche de sites de recodage, à l'aide d'une autre suite d'utilitaires (GenRecode).

- *phpLabDB* est un outil écrit en PHP/SQL prévu pour manipuler/visualiser les bases de données du laboratoire à travers le Web. Actuellement il gère l'ensemble des oligonucléotides de synthèse du laboratoire, les plasmides, les souches etc. Il permet de mettre en commun une importante masse de données immédiatement disponible. phpLabDB est un projet *open-source*.
- *GenRecode* est plus spécifique à mes travaux (voir la recherche de site de recodage sans *a priori*). Il permet l'acquisition, l'extraction et l'analyse des bases de données de séquences lors de la recherche de sites de recodage. C'est une suite de scripts perl/bioperl (Stajich *et al.*, 2002) dont le but est de rechercher des organisations génomiques compatibles avec le recodage. Associé avec Interpro (Mulder *et al.*, 2003), il permet de faire des recherches de motifs protéiques sur ces régions. L'ensemble est regroupé dans une base de données relationnelle. Une interface graphique est implémentée dans phpLabDB.

3.2 phpLabDB

3.2.1 Résumé

J'ai conçu et développé un système et une interface graphique qui regroupent différentes bases de données. Ceci permet de partager l'ensemble des ressources et connaissances du laboratoire. Ce système indépendant de la plateforme peut être consulté à distance. Le but du phpLabDB est que chaque membre de l'équipe puisse contrôler ses données rapidement et les rendre disponibles à la communauté. En raison du support sur l'Intranet et l'Extranet, les données peuvent être consultées à partir de n'importe quel ordinateur dans le laboratoire, ce qui accélère les recherches. phpLabDB est un ensemble de modules qui fonctionne sur un serveur Apache/PHP/SQL et est consultable avec un navigateur Web. L'utilisateur peut enrichir la base de données aussi bien qu'y effectuer des recherches.



a. OligoDB plug-in

b. PlasmidDB plug-in

Figure 1 : Captures d'écran de modules de phpLabDB. a) Détails d'un enregistrement d'OligoDB plug-in, incluant la séquence d'un oligonucléotide, un commentaire, le Tm et son identificateur code-barre. b) Détails d'un enregistrement de PlasmidDB plug-in. Carte du plasmide et une fenêtre *pop-up* de la séquence nucléotidique.

3.2.2 Article soumis

phpLabDB: a new gateway for private databases
Bekaert M. & Rousset JP.

Bioinformatics - soumis.

- *Voir annexe 3* -

3.2.2 Disponibilité

phpLabDB est librement téléchargeable sous licence artistique (*Artistic License*) aux adresses suivantes :

```
http://phplabdb.sourceforge.net/projects/
http://sourceforge.net/projects/phplabdb/
http://bioinformatics.org/project/?group_id=368/
```

3.3 GenRecode

Le but et l'application de GenRecode ont déjà été discutés précédemment (voir la recherche de sites de recodage sans *a priori*). Ici ne seront exposés que le principe, l'utilisation du système et la base de données relationnelle sur laquelle repose l'ensemble des programmes ainsi que les algorithmes exacts qui ont été utilisés.

3.3.1 Structuration des données

Afin de décrire les régions où l'organisation génomique est compatible avec le recodage, ce qui comprend la définition des régions, les informations déduites et leurs relations, j'ai utilisé le langage UML (Unified Modeling Langage), un langage graphique orienté objets physiques ou conceptuels ; les classes y sont caractérisées par leurs attributs et relations. J'ai donc élaboré, à partir du schéma conceptuel UML, la structure relationnelle de la base, en utilisant des règles de traduction des modèles conceptuels en modèles de type relationnel.

Les différentes bases de données développées au sein du laboratoire (voir phpLabDB) utilisent le système de gestion de bases de données relationnelles (SGBDR) PostgreSQL qui est un système performant et robuste. Ce type de recherche génère une grande quantité de données et leur gestion met en jeu des requêtes simultanées sur plusieurs tables, ce qui peut devenir très vite pénalisant en termes de temps de réponse.

Le modèle composé de trois parties a été établi : une phase extraction de séquences correspondant à l'extraction directe des séquences et de leurs annotations depuis les bases de données de séquences (GenBank ou EMBL ; Benson *et al.*, 2004; Kulikova *et al.*, 2004) ; une phase d'organisation des régions susceptibles de subir un recodage ; et une phase d'acquisition de données sur chacune de ces séquences. Une représentation simplifiée du modèle, réduit aux classes essentielles, est présentée Figure 2.

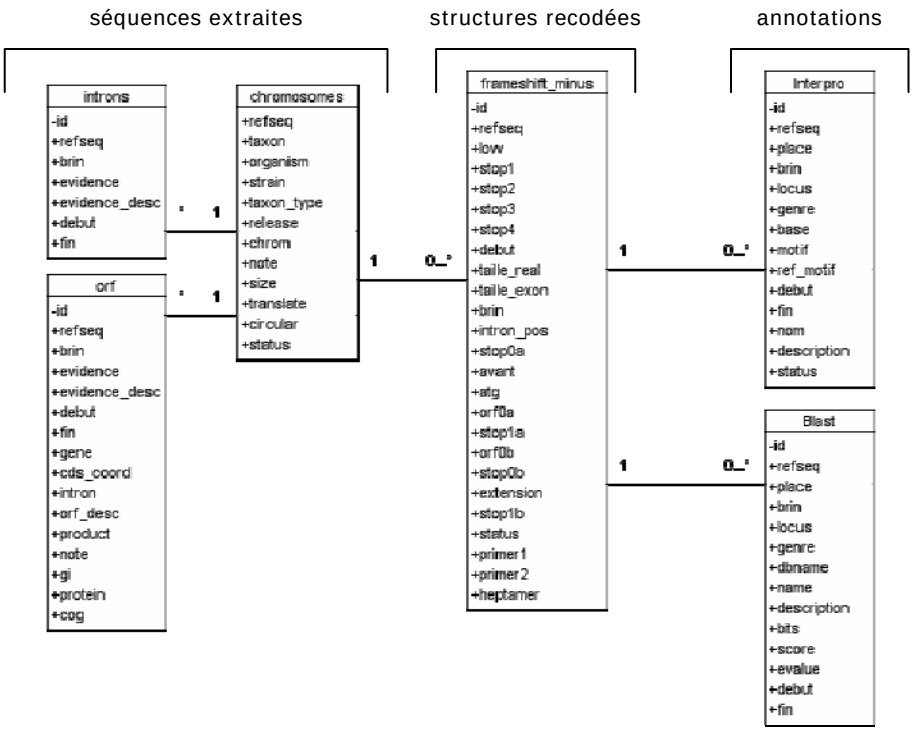


Figure 2 : Schéma conceptuel simplifié de la base de données db_recode. Les classes sont symbolisées par des tableaux dont le titre est le nom de la classe. Les attributs sont listés sous le nom de l'objet. Les classes sont reliées entre elles par des liens qui symbolisent les relations qui les unissent.

3.3.2 Présentation technique

La base de données de GenRecode, db_recode, est accessible aux différents systèmes de calcul afin de répartir les temps d'analyse très importants, et doit rester accessible aux différents clients graphiques du laboratoire pour l'affichage des résultats. J'ai développé une implémentation de GenRecode pour phpLabDB qui utilise les ressources du serveur Apache/PHP et PostgreSQL.

Le GenRecode lui-même est essentiellement une collection de modules en Perl avec l'implémentation Bioperl (Stajich *et al.*, 2002), facilitant le développement et l'analyse de vastes séquences génomiques. InterproScan (Zdobnov and Apweiler, 2001) a été implémenté manuellement pour permettre une analyse complète des données générées.

3.3.2 Extraction des séquences

L'extraction des séquences se fait à partir de trois scripts différents, ce qui permet d'effectuer plusieurs tâches simultanément ou de

les répartir sur plusieurs processeurs / ordinateurs (par exemple un cluster) :

- *add_chromosome.pl* : émet une requête auprès de GenBank (pour les génomes entiers issue de RefSeq) ou de l'EMBL pour les chromosomes isolés ;
- *analysis.pl* : parcourt la séquence et identifie les organisations génomiques compatibles avec un recodage.
  Le script stocke les données dans la base de données. L'algorithme d'identification de séquences est le suivant :

a. Décalage de phase de lecture en -1

```
/($stop)(((?!($debut))([ATCG]{3})(?!($stop)))*)
 ($debut)(((?!($stop))([ATCG]{3}))*[ATCG]{2})
 ($stop)([ATCG]((?!($stop))([ATCG]{3}))*)($stop)
 ([ATCG]{2}((?!($stop))([ATCG]{3}))*)($stop)/
```

b. Translecture

```
/($stop)(((?!($debut))([ATCG]{3})(?!($stop)))*)
 ($debut)(((?!($stop))([ATCG]{3}))*)($stop)
 (((?!($stop))([ATCG]{3}))*)($stop)/
```

c. Décalage de phase de lecture en +1

```
/($stop)(((?!($debut))([ATCG]{3})(?!($stop)))*)
 ($debut)(((?!($stop))([ATCG]{3}))*[ATCG])($stop)
 ([ATCG]{2}((?!($stop))([ATCG]{3}))*)($stop)
 ([ATCG]((?!($stop))([ATCG]{3}))*)($stop)/
```

Figure 3 : Expressions régulières de recherche de : a. décalage de phase de lecture en -1 ; b. translecture ; c. décalage de phase de lecture en +1. $stop représente l'ensemble des codons stop possibles pour l'organisme considéré. De même, $debut représente l'ensemble de codons initiateurs.

Ensuite, seulement, les limites de taille sont fixées. Ceci permet une plus grande souplesse dans le paramétrage des tailles au prix d'une perte de temps minime ;

- *extract.pl* : script utilisant une version dérivée de *mdust* du TIGR afin de filtrer les séquences de basses complexités pour les analyses HMM.

3.3.4 Analyse des séquences

L'analyse des séquences après l'extraction et la préparation des données se fait à partir de deux scripts différents :

- *pattern.pl* : prépare les données pour InterproScan (Zdobnov and Apweiler, 2001). Afin d'optimiser les temps de calcul d'IntroproScan et particulièrement de HMMER (Eddy, 1998), le script prépare une analyse sur la plus petite des ORF et ne soumet la seconde à IntroproScan que si la

première possède un motif. Cette étape a été optimisée pour fonctionner sur des systèmes de calculs parallélisés tels que des clusters ;

- *blasted.pl* : Effectue des BLASTs (Altschul *et al.*, 1990) sur les séquences sélectionnées.

3.3.5 Visualisation des résultats

La visualisation interprétable des résultats est pratiquement aussi importante que les résultats eux-mêmes. Il ne sert à rien d'accumuler des données si on ne peut pas les traiter. Un module a été développé afin d'avoir une vue synthétique (mais exhaustive) des informations de la base de données. L'utilisateur final n'a pas besoin de connaître la structure de la base de données. Objet, classe, attribut et relations sont modélisés de manière simplifiée au travers de « vues ». Le module permet d'une part d'obtenir des données statistiques sur la répartition des sites de recodage et d'autre part de détailler chacune de ces régions potentielles. Ceci inclut, la visualisation des régions ayant des motifs Interpro, les rapports BLAST, le design d'amorce de PCR, de RT-PCR ou de séquences et l'exportation vers d'autres logiciels tel qu'Artemis (Mural, 2000).

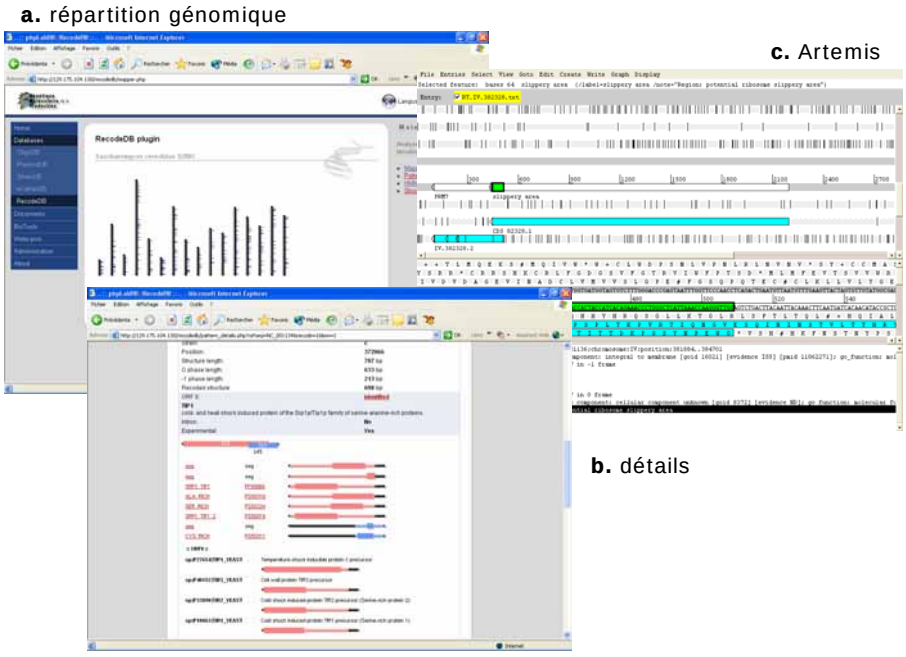


Figure 4 : Capture d'écrans de RecodeDB plug-in et d'Artemis. a. affichage de la répartition des régions de recodage identifiées sur les chromosomes de *S. cerevisiae* ; b. détail d'un site potentiel de décalage de phase de lecture (taille, positions, motifs et rapports BLAST) ; c. affichage *Artemis* d'une région de décalage de phase de lecture.

3.4 Références

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol 215, 403-410. [2231712]

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res 32, D23-26.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics 14, 755-763. [9918945]

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., et al. (2004). The EMBL Nucleotide Sequence Database. Nucleic Acids Res 32, D27-30.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., et al. (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31, 315-318. [12520011]

Mural, R. J. (2000). ARTEMIS: a tool for displaying and annotating DNA sequence. Brief Bioinform 1, 199-200. [11465031]

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12, 1611-1618. [12368254]

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848. [11590104]

Discussion et perspectives

4 Discussion et perspectives

4.1 Les sites de décalages

Les objectifs de ce travail étaient d'identifier des gènes soumis à un mécanisme de décalage de phase de lecture en -1. Les différentes approches conjointement élaborées ont permis d'identifier plusieurs candidats sérieux. Ces études systématiques ont été précisées par des analyses physiologiques et moléculaires complémentaires.

4.1.1 Nouveaux sites

Jusqu'à présent la grande partie des événements de décalage de phase de lecture en -1 chez les eucaryotes ont été identifiés chez les virus. Les deux seules recherches systématiques chez la levure, réalisées par Hammell et Liphardt, n'ont pas identifié de sites authentiques (Hammell *et al.*, 1999; Liphardt, 1999). Une troisième recherche centrée uniquement sur *Drosophila melanogaster* ne conclut pas sur des données biologiques avérées mais seulement sur des spéculations probabilistes (Sato *et al.*, 2003). Quant aux approches plus général d'identification de gènes chez la levure récentes, elles n'ont pas recherché ce type d'événements ou les ont classés comme pseudogènes (Harrison *et al.*, 2002; Kessler *et al.*, 2003; Oshiro *et al.*, 2002).

Le travail réalisé ici s'est appuyé sur les connaissances issues des exemples viraux et appliquées à la fois avec ou sans *a priori* sur divers génomes. La levure *S. cerevisiæ* restant cependant l'organisme modèle où

mes recherches se sont focalisées et où l'essentiel des évaluations biologiques ont été réalisées.

Au cours de ces travaux, des régions génomiques compatibles avec un décalage de phase de lecture ont été sélectionnées par trois méthodes complémentaires :

- La recherche de motifs protéiques fonctionnels déphasés. Cette méthode n'utilise que les données des banques de données de motifs et reste sans *a priori* sur le site de décalage lui-même ;
- L'identification d'une discontinuité dans la linguistique du codant (l'utilisation des codons) qui n'utilise que les probabilités de présence des codons dans les gènes déjà identifiés de l'organisme étudié. Cette méthode reste, elle aussi, sans *a priori* sur la zone de décalage ;
- La recherche de pseudonœuds caractéristique de sites de décalage. Cette approche utilise, au contraire des deux autres, des *a priori* forts sur le site de décalage et la présence d'une structure secondaire (qui généralement est un pseudonœud).

Les limites de chaque méthode ont déjà été discutées (voir résultats) : la première méthode nécessite que le décalage de phase de lecture fasse apparaître un motif protéique déjà identifié par ailleurs ; la seconde se restreint à l'observation de déphasages produisant une extension de la phase de lecture et la dernière nécessite la présence d'un pseudonœud proche du site de décalage. Cependant le principe de ce travail est d'identifier les sites réels de décalage de phase de lecture en -1 et d'être exigeant sur les critères utilisés afin de minimiser le nombre de faux positifs communs (au risque de ne pas être exhaustif) ; ensuite, l'utilisation des trois approches, en regroupant tous les sites distincts identifiés, permet au contraire d'éviter les biais inhérents à chacune des méthodes.

4.1.1.1 *Saccharomyces cerevisiæ*

La levure *S. cerevisiæ* possède 16 chromosomes pour 13 Mb de séquence. Treize gènes ont été identifiés par l'une de ces méthodes et biologiquement validés. Une analyse moléculaire des séquences a révélé que l'efficacité de déphasage était supérieure à 5% et qu'il existait un ARNm correspondant aux deux ORF déphasées *in vivo*. Ces résultats ne nous donnent évidemment pas une fonction, mais permettent de conclure sur l'existence du phénomène ; les motifs protéiques, quant à eux, permettent d'identifier quelques pistes. Sept gènes sont déjà partiellement caractérisés dans les banques de données, l'une des deux ORF qui les constituent étant déjà connue, mais cela n'informe pas forcément sur le rôle

du décalage de phase de lecture en -1, ni sur leur fonction, même si l'une des ORF a déjà été étudiée indépendamment de l'autre.

fsORF	%	Sage	ORF0	ORF-1	Notes
2	6%	Faible	SCO2		SCO2 *(involved in stability of Cox1p and Cox2p)*
11	11%	Faible	*YDL038C**	PRM7	PRM7 *(pheromone-regulated membrane protein)*
16	9%	-	AAD6	*AAD16**	AAD6 *(high similarity with the AAD of P. chrysosporium)*
21	13%	-			Intergénique / PRF13
30	12%	Moyen	YKL033W-A*		-
32	7%	Fort		SRL3	SRL3 *(Suppressor of Rad53 null Lethality)*
40	5%	-	*YMR084W**	*YMR085W**	putative glutamine--fructose-6-phosphate transaminase
41	5%	Faible	ADE17		ADE17 *(AICAR transformylase/IMP cyclohydrolase)*
43	10%	-	MRPL24		MRPL24 *(Mitochondrial ribosomal protein)*
48	5%	Faible	STE4		STE4 *(GTP-binding protein beta subunit of the pheromone pathway)*
50	7%	Faible		RAD17	RAD17 *(DNA damage checkpoint control protein)*
ID.135	5%	Faible			Intergénique
ID.20	7%				Intergénique

Tableau 1 : Liste des régions candidats actifs. * : ORF putative.

L'étude plus approfondie du gène PRF13 a été initiée, mais n'a pas encore assigné de fonction à ce dernier. Il reste donc encore une étape de caractérisation fine de chacun de ces gènes et du rôle du décalage de phase de lecture dans leur régulation.

4.1.1.2 Autres organismes

A terme l'objectif est de pourvoir parcourir l'ensemble des organismes séquencés à la recherche de sites authentiques de décalage de phase de lecture. Plusieurs difficultés restent à régler :

- L'identification des introns avec un maximum de robustesse. Dans les organismes complexes le nombre élevé d'introns reste le principal obstacle à ce type de recherche, puisqu'ils peuvent changer la phase ou posséder des snRNA très structurés avec des pseudonœuds, autant d'éléments qui perturbent l'analyse ;
- L'identification des erreurs de séquence et des pseudogènes ou du polymorphisme naturel.

4.1.2 Multicritères

Deux analyses ont été réalisées, l'une sur l'ensemble des virus entièrement séquencés à ce jour. 74 sites ont été identifiés dont 10 nouveaux. Pour ces virus la méthode la plus efficace n'a pas été l'identification de sites chevauchants comme précédemment présenté, mais la recherche d'un motif glissant simple avec un profil HMM. Cette recherche est à la fois plus rapide et plus pertinente dans les organismes aux séquences courtes et très compactes avec de nombreux gènes

chevauchant. Pour des organismes plus complexes et moins compacts, la recherche multicritère de sites chevauchants paraît plus adaptée.

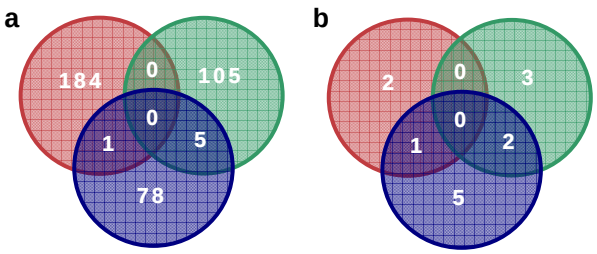


Figure 1 : Répartition des candidats identifiés par les trois méthodes présentées. a : Candidats identifiés *in silico* ; b : Séquences biologiquement actives. En rouge l'approche par recherche de structure secondaire ; en vert la recherche par HMM ; en bleu la recherche par motifs protéiques.

4.1.3 Conclusion

Ce travail suggère fortement que le décalage de phase de lecture en -1 n'est pas un phénomène limité aux gènes viraux, mais est aussi retrouvé dans des gènes cellulaires eucaryotes. Si jusqu'à présent les cas identifiés ne l'ont généralement été que dans des génomes viraux, cela reflète donc uniquement le fait que les virus possèdent de petits génomes, qui sont intensivement étudiés pour des raisons médicales ou agro-industrielles. Les critères utilisés sont plus restrictifs et produisent peu de faux positifs, mais accroissent de manière significative la qualité de la prédiction.

4.2 Dynamique du ribosome

4.2.1 Stimulateurs

Afin de mieux caractériser les facteurs *cis* et *trans* qui interviennent dans le déphasage des ribosomes, différents éléments ont été étudiés de manière plus approfondie.

L'alignement des séquences des sites viraux a initialement permis de mettre en évidence un fort biais de composition juste en amont de l'heptamère glissant. J'ai démontré que la séquence de cette région affectait réellement l'efficacité de décalage de phase le lecture en -1 chez *S. cerevisiae*. Ces résultats ont aussi mis en évidence le rôle de la pseudouridine en position 39 de l'ARNt situé dans le site E du ribosome de la cellule hôte, dans l'efficacité du décalage de phase probablement en déstabilisant le l'ARNt au site P ou en découplant l'ARNt du site E

prématurément. Un autre élément identifié est le rôle stimulateur du triplet XXX de l'heptamère dans l'efficacité de déphasage. Ceci suggère que le signal pourrait être la trace d'un tetra-nucléotide ancestral et que le triplet XXX est utilisé pour en moduler l'efficacité. En effet, chez les eucaryotes, bien que le glissement en tandem des ARNt soit le mécanisme principal, le dérapage peut avoir lieu alors que seul le site P du ribosome est occupé (Baranov *et al.*, 2004; Jacks *et al.*, 1988; Yelverton *et al.*, 1994). Enfin, ces approches m'ont permis de constater l'existence de nucléotides interdits ou préférentiels dans la séquence de l'espaceur, et de mettre en évidence que la modification de ces nucléotides module largement le niveau de décalage de phase de lecture en -1.

4.2.2 Vision intégrée

Le décalage de phase en -1 peut être vu comme une suite de mécanismes stimulateurs ingénieusement associés afin de permettre à chaque gène d'assurer une régulation optimale. En fait le ribosome en élongation « rencontre » ces éléments de manière séquentielle.

Le mouvement à grande échelle de la tige L1 a été récemment suggéré par des reconstructions des complexes de pre- et de post-translocation du ribosome. Ce mouvement est probablement associé au positionnement de l'ARNt au site E et à une dynamique de conformation du ribosome (Agrawal *et al.*, 1999; Valle *et al.*, 2003). Par ailleurs lors de la translocation, un pseudonœud peut induire une pause du ribosome (Lopinski *et al.*, 2000; Somogyi *et al.*, 1993; Wolin and Walter, 1988), fournissant une résistance au mouvement relatif du ribosome en se coinçant dans l'entrée ou dans le tunnel d'entrée de l'ARNm. Des expériences récentes de cryo-EM impliquent le pseudonœud à l'intérieur du ribosome en élongation (O. Namy, communication personnelle). Une résistance pourrait donc induire un état de « flottement du ribosome ». Cet état induirait une conformation étendue de la région de l'espaceur créant une tension locale de l'ARNm entre le site A et le pseudonœud. Cette tension pouvant être libérée soit par l'ouverture du pseudonœud, permettant à la région en aval d'avancer, soit par le glissement de la région proximale de l'ARNm d'une base en arrière (Kim *et al.*, 2001; Plant *et al.*, 2003). La propension à adopter une conformation étendue dépend de la composition nucléotidique de l'espaceur et peut-être de ses interactions avec d'autres ARN (Bekaert *et al.*, 2003; Bertrand *et al.*, 2002). Le résultat observé de ce mécanisme est un décalage d'une base en 5' : un décalage de phase de lecture en -1.

Un tel modèle, confère au triplet XXX de l'heptamère un rôle stimulateur critique (Bekaert *et al.*, 2003; Brierley *et al.*, 1992; Brierley *et*

al., 1997) et implique largement l'ARNt du site E (Leger *et al.*, 2004). Chez *E. coli* une étude récente impliquerait en fait que le décalage aurait lieu lorsque les dé-acétyl-ARNt et peptidyl-ARNt sont aux sites E et P du ribosome, plutôt que les peptidyl-ARNt et aminoacyl-ARNt aux sites P et A. Ceci expliquerait que les mutations de l'ARN 16S qui facilitent le logement de l'aminoacyl-ARNt au site A diminuent l'efficacité de décalage, ce dernier ayant alors lieu alors que l'aminoacyl-ARNt occupe le site A/T (Leger *et al.*, 2004).

4.3 Perspectives

Une des perspectives immédiates de ce travail, sera de mieux caractériser les candidats identifiés, afin de valider leur pertinence biologique en plus de valider le modèle utilisé. Une caractérisation des protéines elles-mêmes est nécessaire afin d'évaluer la fonction des formes courtes et étendues de chaque candidat, ainsi que le rôle joué par la régulation par décalage de phase de lecture en -1.

Une autre voie serait de poursuivre l'analyse du site de décalage et plus particulièrement de l'espaceur. Actuellement l'impact exact de la composition nucléotidique de l'espaceur et ses interactions avec d'autres ARN n'ont pas été étudiées de manière approfondie. Des expériences ont déjà été initiées dans ce but, mais elles restent encore insuffisantes pour permettre une conclusion

Un autre développement de ce travail, serait d'étendre ce genre d'analyse à d'autres génomes ou à d'autres événements de recodage (comme le décalage du cadre de lecture en +1 ou la translecture). Cependant la recherche de sites de recodage se heurte à une difficulté majeure, qui est la complexité des génomes analysés. Effectivement chez *S. cerevisiæ*, la densité de gènes est importante, les introns sont peu nombreux et assez bien caractérisés. Une autre difficulté reste aussi l'identification des erreurs de séquence et la caractérisation des pseudogènes ou du polymorphisme naturel.

Ce problème pose plus généralement la question de la définition d'une phase ouverte de lecture. Celles-ci sont annotées dans les banques de données entre un codon AUG et un codon stop. Or, en absence d'intron dans cette phase ouverte de lecture, il est admis qu'une seule protéine sera synthétisée. Nous venons de voir que, dans certains cas, ce postulat va s'avérer faux, puisqu'au moins grâce aux quatre recodages connus il sera possible d'obtenir plusieurs protéines à partir d'un même ARNm. Ces événements apparaissent encore actuellement comme des exceptions aux mécanismes généraux de traduction, mais nos résultats suggèrent que les mécanismes de recodages pourraient être plus fréquents qu'on ne le pense.

La recherche de cibles de recodage se heurte à un second problème, qui est l'erreur de séquence vraie ou supposée. Certains programmes informatiques ont été conçus pour identifier et corriger les erreurs de séquençage (Fukunishi and Hayashizaki, 2001). Mais comment se caractérise une erreur de séquençage ? Principalement de deux manières : soit par l'apparition d'un codon stop en phase, soit par la création d'un décalage du cadre de lecture. Ces erreurs de séquençage ressemblent donc fortement à des sites de recodage, et en l'absence d'une analyse plus fine des séquences, il n'est pas possible de les différencier. Ainsi, en voulant corriger automatiquement les séquences il y a un risque majeur de faire disparaître des banques de données une information importante à propos des sites de recodage. On aboutit alors au fait que le codon stop ou le décalage n'est plus annoté.

4.4 Références

Agrawal, R. K., Lata, R. K., and Frank, J. (1999). Conformational variability in Escherichia coli 70S ribosome as revealed by 3D cryo-electron microscopy. Int J Biochem Cell Biol *31*, 243-254. [10216957]

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2004). P-site tRNA is a crucial initiator of ribosomal frameshifting. Rna *10*, 221-230. [14730021]

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics *19*, 327-335. [12584117]

Bertrand, C., Prere, M. F., Gesteland, R. F., Atkins, J. F., and Fayet, O. (2002). Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression. Rna *8*, 16-28. [11871658]

Brierley, I., Jenner, A. J., and Inglis, S. C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. J Mol Biol *227*, 463-479. [1404364]

Brierley, I., Meredith, M. R., Bloys, A. J., and Hagervall, T. G. (1997). Expression of a coronavirus ribosomal frameshift signal in Escherichia coli: influence of tRNA anticodon modification on frameshifting. J Mol Biol *270*, 360-373. [9237903]

Fukunishi, Y., and Hayashizaki, Y. (2001). Amino acid translation program for full-length cDNA sequences with frameshift errors. Physiol Genomics *5*, 81-87. [11242592]

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999). Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res *9*, 417-427. [10330121]

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., and Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J Mol Biol *316*, 409-419. [11866506]

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447-458. [2846182]

Kessler, M. M., Zeng, Q., Hogan, S., Cook, R., Morales, A. J., and Cottarel, G. (2003). Systematic discovery of new genes in the Saccharomyces cerevisiae genome. Genome Res *13*, 264-271. [12566404]

Kim, Y. G., Maas, S., and Rich, A. (2001). Comparative mutational analysis of cis-acting RNA signals for translational frameshifting in HIV-1 and HTLV-2. Nucleic Acids Res *29*, 1125-1131. [11222762]

Leger, M., Sidani, S., and Brakier-Gingras, L. (2004). A reassessment of the response of the bacterial ribosome to the frameshift stimulatory signal of the human immunodeficiency virus type 1. Rna *10*, 1225-1235. [15247429]

Liphardt, J. (1999) The mechanism of -1 ribosomal frameshifting: experimental and theoretical analysis, Ph.D., Churchill College, Cambridge.

Lopinski, J. D., Dinman, J. D., and Bruenn, J. A. (2000). Kinetics of ribosomal pausing during programmed -1 translational frameshifting. Mol Cell Biol *20*, 1095-1103. [10648594]

Oshiro, G., Wodicka, L. M., Washburn, M. P., Yates, J. R., 3rd, Lockhart, D. J., and Winzeler, E. A. (2002). Parallel identification of new genes in Saccharomyces cerevisiae. Genome Res *12*, 1210-1220. [12176929]

Plant, E. P., Jacobs, K. L., Harger, J. W., Meskauskas, A., Jacobs, J. L., Baxter, J. L., Petrov, A. N., and Dinman, J. D. (2003). The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. Rna *9*, 168-174. [12554858]

Sato, M., Umeki, H., Saito, R., Kanai, A., and Tomita, M. (2003). Computational analysis of stop codon readthrough in D.melanogaster. Bioinformatics *19*, 1371-1380. [12874049]

Somogyi, P., Jenner, A. J., Brierley, I., and Inglis, S. C. (1993). Ribosomal pausing during translation of an RNA pseudoknot. Mol Cell Biol *13*, 6931-6940. [8413285]

Valle, M., Zavialov, A., Sengupta, J., Rawat, U., Ehrenberg, M., and Frank, J. (2003). Locking and unlocking of ribosomal motions. Cell *114*, 123-134. [12859903]

Wolin, S. L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. Embo J *7*, 3559-3569. [2850168]

Yelverton, E., Lindsley, D., Yamauchi, P., and Gallant, J. A. (1994). The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in Escherichia coli. Mol Microbiol *11*, 303-313. [8170392]

Annexes

Article 1

BIOINFORMATICS

Vol. 19 no. 3 2003, pages 327–335
DOI: 10.1093/bioinformatics/btf868

Towards a computational model for −1 eukaryotic frameshifting sites

Michaël Bekaert[1], Laure Bidou[1], Alain Denise[2,3], Guillemette Duchateau-Nguyen[2], Jean-Paul Forest[3], Christine Froidevaux[3,*], Isabelle Hatin[1], Jean-Pierre Rousset[1] and Michel Termier[2]

[1]Génétique Moléculaire de la Traduction, [2]Bioinformatique des Génomes, Institut de Génétique et Microbiologie (IGM), UMR CNRS 8621 and [3]Laboratoire de Recherche en Informatique (LRI), UMR CNRS 8623, Université Paris-Sud, 91405 Orsay Cedex, France

Received on February 1, 2002; revised on June 6, 2002; accepted on August 27, 2002

ABSTRACT

Motivation: Unconventional decoding events are now well acknowledged, but not yet well formalized. In this study, we present a bioinformatics analysis of eukaryotic −1 frameshifting, in order to model this event.

Results: A consensus model has already been established for −1 frameshifting sites. Our purpose here is to provide new constraints which make the model more precise. We show how a machine learning approach can be used to refine the current model. We identify new properties that may be involved in frameshifting. Each of the properties found was experimentally validated. Initially, we identify features of the overall model that are to be simultaneously satisfied. We then focus on the following two components: the spacer and the slippery sequence. As a main result, we point out that the identity of the primary structure of the so-called spacer is of great importance.

Availability: Sequences of the oligonucleotides in the functional tests are available at http://www.igmors.u-psud.fr/rousset/bioinformatics/

Contact: bekaert@igmors.u-psud.fr; jpforest@lri.fr; chris@lri.fr

INTRODUCTION

The universality of the genetic code is the initial step of automatic determination for hypothetical open reading frames (ORFs) using very simple methods, such as seeking long, terminator-less phases. However, translation machinery appears capable of decoding not only the classical genetic code but also several kinds of signalling patterns embedded in the mRNA (Gesteland *et al.*, 1992).

Three major forms of recoding have been identified: stop codon *readthrough* by the ribosome; ribosomal *frameshifting*, where the ribosome slips either forward or backward; and ribosome *hopping* where dozens of nucleotides on the message can be skipped by the decoding machinery (Gesteland and Atkins, 1996).

In the present work, we focus on −1 frameshifting. Most of these events are found in viruses and transposons, where they serve to produce the replicase domain needed for the life cycle. Enhancing or reducing the effectiveness of the mechanism can dramatically influence virus viability (Dinman *et al.*, 1998). Very few cellular genes using −1 frameshifting are presently known: the *dnaX* gene of *Escherichia coli* (Tsuchihashi and Kornberg, 1990), the cdd gene of *Bacillus subtilis* (Mejlhede *et al.*, 1999) and, more recently, the Edr gene in mice (Shigemoto *et al.*, 2001). To date, there is no general method to identify such genes.

The genetic information carried by genes expressed through a frameshifting event is, by definition, out of frame. Therefore, these genes could be annotated as non-coding (Medigue *et al.*, 1999). The development of molecular approaches has permitted the demonstration that −1 frameshifting is correlated with the presence of specific signals on the coding sequence, which in turn has led to the design of an initial model.

The goal of this research is to establish a program to identify new genes that use −1 frameshifting for their expression. Our present objective is to improve the known computational model of frameshifting sites in eukaryotic viruses. For this purpose, we use a combination of bioinformatics methods, computer science concepts and biological experimentation. This paper presents our approach and our initial results towards the conception of a more refined computational model.

BIOLOGICAL MODEL AND STATE OF THE ART

The current model for eukaryotic frameshift sites consists of two main components: a *slippery site*, which

*To whom correspondence should be addressed.

mechanically promotes frameshifting, and a *stimulatory structure* which probably acts by pausing the ribosome (Jacks and Varmus, 1985; Farabaugh, 1996; Tu *et al.*, 1992; Somogyi *et al.*, 1993; Lopinski *et al.*, 2000; Kontos *et al.*, 2001). These signals are carried by the mRNA and superimposed on the coding sequence. Although the presence of the slippery site is sufficient to induce a low but biologically significant level of frameshifting, that of the stimulatory structure is not (Kollmus *et al.*, 1996). The short sequence between these two components is called the *spacer* and is denoted SP (Figure 1).

1. The *slippery site*, denoted H (for Heptamer), is the place where the ribosome actually shifts. It is an heptameric sequence conforming to the motif X XXY YYZ (spaces display boundaries of codons in frame 0). The special arrangement of the nucleotides in the heptameric site allows the two tRNAs in both frames (0 and −1) to be paired: XXY and XXX for the tRNA in the P-site of the ribosome, YYZ and YYY for the tRNA in the A-site.
2. The *stimulatory structure*, denoted E (for Enhancer), increases the probability of 5′ ribosomal movement. It can be either a single stem-loop or a pseudoknot. The following subsequences can be distinguished:
   (a) **ES1.5**′, first stem, 5′-arm;
   (b) **EL1**, beginning of first loop;
   (c) **ES2.5**′, second stem, 5′-arm;
   (d) **EL1**′, end of first loop;
   (e) **ES1.3**′, first stem, 3′-arm;
   (f) **EL2**, second loop;
   (g) **ES2.3**′, second stem, 3′-arm.

Parts ES2.5′, EL2 and ES2.3′ are present only in the case of a pseudoknot.

For each subsequence of this model, the following features are of interest: (i) the length; (ii) the identity of the base at a given position; (iii) the number of occurrences of a given base; and (iv) the presence of a simple stem-loop or of a pseudoknot.

Experimental analyses have shown the influence of modifying some of these features on frameshifting efficiency. They have demonstrated that the following features are relevant:

(a) For the heptamer, the identity of Y and of Z: in the canonical model, Y is either an A or a U, Z cannot be a G and there is no constraint on X. Some variants are known. In particular, the three X can differ (we will write X1, X2 and X3 when necessary; Brierley *et al.*, 1992).

(b) For the first stem (ES1), the length and the G–C pair number (Brierley *et al.*, 1991).

(c) For the spacer, the length must be taken into account (Napthine *et al.*, 1999).

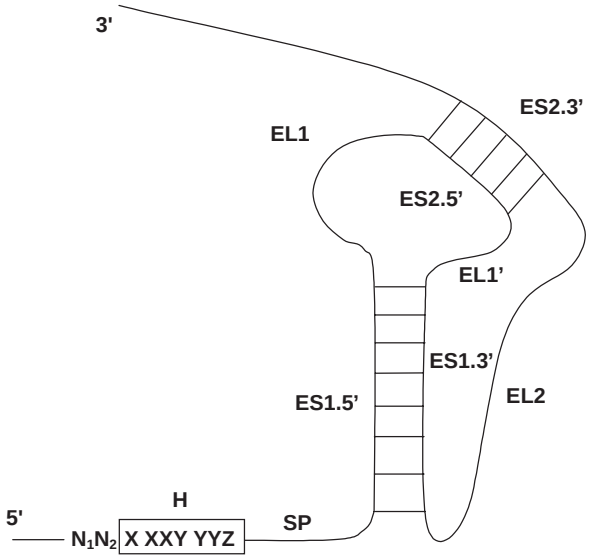


Fig. 1. Labels of subsequences collectively forming the frameshifting signal.

Moreover, pseudoknots are more efficient than stem-loops (Brierley, 1995).

Two attempts to model frameshifting sites were previously made. (Hammell *et al.*, 1999) proposed a model whose main parameters are the structure of H, constraints of lengths, and numbers of pairings in the stems. Liphardt investigated the possibility of using stochastic context-free grammars (SCFG) to model the stimulatory structure (Liphardt, 1999). Both approaches led to models which take into account the main known parameters of the phenomenon. However, the programs based on these models, when applied to entire genomes, found too many false positives. This may be due to two reasons. Firstly, the number of parameters to be considered in each site is large, and therefore difficult to handle 'by hand'. Secondly, the relatively small amount of data (known frameshift sites in wild viruses and mutants) is insufficient to lead to an accurate model of the phenomenon. Nonetheless, these studies constitute a significant step towards modelling frameshifting sites in order to design a prediction tool. In particular, they demonstrate that use of a stochastic model is rather promising, and that filtering constraints on the current model are necessary to increase the efficiency of any search program.

METHODOLOGY

In order to deal with the large number of possibly relevant parameters, we chose to adopt a strategy based on bioinformatics and computational methods. Moreover, since the articles of Hammell *et al.* (1999) and Liphardt,

more sites have been studied by biologists, therefore more data are available.

The general methodology is summarized in the following 'cyclic scheme':

(1) Take a set of sequences that induce frameshifting with a known efficiency level. It will be used as a training set to learn a proper description of the frameshifting event. We can consider this either as a binary event (it occurs or not) or as an event occurring with a given rate. The data representation of the sequences must take into account the consensus organization and the properties known or supposed to be relevant in the frameshifting process.

(2) Refine the model: For this purpose, we use machine learning approaches (supervized or not), associated with classical bioinformatics methods. The aim is to discover new properties shared by frameshifting sites that belong to the same class or have a pre-determined rate.

(3) Test the model: According to the new properties, design a set of sequences that conform to the model. This can be done in two ways: (i) *ab initio* designing new sequences which do not exist in any organism; (ii) using the model to find sequences in genomic databases. Predict their respective classes (or their effective rates) according to the model, and then biologically evaluate their functionality.

(4) Evaluate the model by comparing predictions and experimental results, and modify it if necessary. For instance, some attributes may be added. The cycle can be restarted, with the new sequences added to the learning set. When the model is considered to be sufficiently reliable, it will be used to construct an effective prediction tool.

MATERIALS AND METHODS

Sequences

The sequences under study come from the literature (Brierley *et al.*, 1992; Brierley, 1995; Marczinke *et al.*, 1998; Napthine *et al.*, 1999; ten Dam *et al.*, 1994, 1995; Kim *et al.*, 1999) as well as from electronic resources: Recode (Baranov *et al.*, 2001) and PseudoBase (van Batenburg *et al.*, 2000). A total of 27 wild-type frameshifting sites and 196 mutant sequences were used for computational work. Biological studies were performed on the avian coronavirus, *Infectious bronchitis virus* (IBV) with a minimal pseudoknot (Brierley *et al.*, 1992) (noted IBV.m) and gag/pro frameshifting site of wild type simian retrovirus 1 (ten Dam *et al.*, 1994) noted SRV.wt [†].

Computational methods and learning systems

Each frameshifting site is composed of several subsequences characterized by specific properties. These properties are formalized by attributes that measure some of their different aspects. We consider that an attribute can be of three types: (1) *numerical*: e.g. the G–C pair number in the first stem; (2) *Boolean*: e.g. Y = Z equality in H (possible values: true or false) or (3) *categorical*: e.g. Y identity in H (possible values: A, C, G or U).

We described the sequences with approximately 120 attributes. We then used a machine learning system to identify the relevant attributes and their values such that a given sequence induces efficient frameshifting. However, using so many attributes and relatively few sequences has two limitations: it is computationally expensive and it might decrease the learning performance. In fact, irrelevant attributes deteriorate learning (as expected) but so might redundant ones. To rectify this, we chose a simple decision tree algorithm to perform a selection. We used Weka's implementation (Witten and Frank, 2000) of C4.5 (Quinlan, 1993).

We were interested in learning the frameshifting concept (the target concept). Initially, we simply determined whether the phenomenon would occur or not. We thus considered the binary concept *efficient_frameshifting*. We did not take into account the actual frameshifting rate, but used it to sort the sequences in order to obtain examples (sequences inducing efficient frameshifting) and counter-examples (sequences with low frameshifting efficiency) of the target concept. Since we were interested in defining a 'frontier' between examples and counter-examples, the latter had to be as close as possible to the former.

We know that all sequences do not induce frameshifting in the same way (Giedroc *et al.*, 2000). For example, in some sequences, long stems (hence stable secondary structures) promote efficient frameshifting (Napthine *et al.*, 1999), while in some others bent stems do so (Chen and Tinoco, 1995; Kang *et al.*, 1996; Chen *et al.*, 1996). This can be described with several rules of the form:

R: '*if* condition C1 *and* condition C2 *and* ... *and* condition Ck *then* efficient_frameshifting', in which conditions specify relevant attributes and their values for the concept *efficient_frameshifting* to be satisfied. As a unique rule is not sufficient to cover all the cases, we look for a disjunction of rules of this form (disjunctive learning). Under such rules, efficient frameshifting occurs if at least one of them is satisfied. If a given sequence satisfies the conditions of one rule, it is a good candidate for an actual frameshifting site. Note that we will have only sufficient conditions, not necessary ones, and that a given sequence can satisfy several rules. Moreover, we do not attempt a description of the cases where frameshifting does not occur.

We chose GloBo (Torre, 2000) as a disjunctive learning tool because it performed well on a problem similar to ours (PTE Challenge, Srinivasan *et al.*, 1999). The intuition underlying the GloBo algorithm is as follows: each example is used as a seed to gather the largest possible subset 'around' it without encompassing any

[†] http://www.igmors.u-psud.fr/rousset/bioinformatics/

counter-examples (thus yielding a correct subsets). Once all the subsets have been built (there are as many subsets as examples), a collection of a few subsets is selected such that all examples are covered. We keep only a minimal collection of such sets in order to obtain few rules, which allows for more intelligibility of the target concept. Each subset is associated to a rule. The algorithm is stochastic, which means many correct subsets are tested before the algorithm computes the actual result. In practice, it seems important to cover each known example with at least one rule, even at the expense of covering a few counter-examples (called false positives). Of course, the rules obtained will also be used as predictions (see **Methodology**, step 4) and only experimentation is able to establish the veracity of the prediction.

In order to determine the precise level of frameshifting, we used classical tools that deal with quantitative prediction, such as regression trees (Breiman *et al.*, 1984). They combine both decision tree and regression techniques. Regression trees are like decision trees, except that each leaf is labelled with a number that is the average of the values of the data that reach the leaf and the splitting attributes are chosen to minimize the intrasubset variation in the class values down each branch (Witten and Frank, 2000). The label of each leaf represents the average value of the target concept for the data belonging to it.

Biological materials and methods

We chose the yeast *Saccharomyces cerevisiae* as an experimental system for the following reasons: (i) it is a simple eukaryotic model; (ii) its genome has been fully sequenced (Goffeau *et al.*, 1996); (iii) its translational machinery is able to perform -1 frameshifting (Dinman *et al.*, 1991); and (iv) it is well suited to reverse genetics experiments.

Strain and medium The *S. cerevisiae* FY1679-18B strain (his3-Δ200, trp1-Δ63, ura3-52, leu2-Δ1 MATα; Dujon *et al.*, 1994) was used in this study. Cultures were grown in YNB medium (0.67% Yeast Nitrogen Base, 2% glucose) under standard growth conditions.

Plasmids Constructs of each sequence tested were obtained by insertion of double stranded oligonucleotides into the *MscI* cloning site of the pAC99 plasmid, a centromeric vector carrying the LEU2 selective marker (Figure 2; Bidou *et al.*, 2000). The *MscI* cloning site is present at the junction of a *lacZ-luc* fusion gene.

Enzymatic activities and frameshifting frequency Reporter plasmids were transferred into yeast strains using the lithium acetate method. In each case at least three transformants cultivated in the same conditions were assayed as previously described. Frameshifting frequency expressed as percentage was calculated by dividing the luciferase/β-galactosidase ratio obtained from each test construct by the same ratio obtained with an in-frame control construct (Bidou *et al.*, 2000).

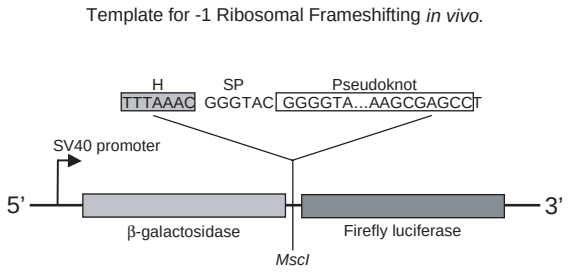


Fig. 2. Template constructed for frameshifting quantification. The tested sequence is localized at the junction between *lacZ/luc* fusion gene. The *in vivo* template is transcribed using SV40 promoter.

RESULTS

Formalizing complete frameshifting sites

We used only pseudo-knotted sites, as they have been more thoroughly studied experimentally than others.

The attributes we used were either new or already known to be relevant. The new attributes we considered are as follows (the others are given in the section **Biological model and state of the art**):

— For the slippery sequence H, the value of each base and the equalities X1 = X2, X1 = X3, X2 = X3, X3 = Y and Y = Z.

— For each other subsequence the nucleotidic composition (number and percentage of each base).

The decision tree method was used on the whole set of sequences. It identified a subset of attributes that are sufficient for correctly classifying almost all sequences. 139 examples and 57 counter-examples were used.

Let |M| be the length of the subsequence M, $|M|_B$ the number of B bases in M and $\%M_B$ the percentage of B in M. Those attributes are (see Figure 1):
Heptamer H: $|H|_A$, $|H|_C$, X1 base, Y base, Z base, X1 = X2, and X3 = Y equalities;
Spacer SP: $|SP|_C$, $|SP|_U$;
First loop EL1 part: |EL1|, $|EL1|_C$, $|EL1|_U$;
First loop EL1′ part: |EL1′|;
Second loop EL2: |EL2|, $|EL2|_A$, $|EL2|_C$, $|EL2|_U$;
First stem ES1: the G-C pair number denoted $|ES1|_{G-C}$, $|ES1.5'|_G$, $\%ES1.5'_G$, $|ES1.3'|_U$;
Second stem ES2: $|ES2.5'|_C$, $|ES2.5'|_G$, $|ES2.5'|_U$, $|ES2.3'|_A$.

This selection step allowed us to reduce the number of attributes from 120 to 25. We added to them a few attributes that seemed to us biologically relevant.

Note that even using the 120 attributes does not allow to correctly classify all the examples and counter-examples. Namely, some examples and counter-examples have the same values for those attributes and cannot be distinguished from one another. In the following, we left the problematic sequences aside. We thus worked with 135 examples and 51 counter-examples.

For our purposes, sequences having a frameshifting rate above 5% were considered to be examples and those having a frameshifting rate below 2% were considered to be counter-examples: sequences whose efficiency is between 2% and 5% were left aside to avoid an overly arbitrary frontier between examples and counter-examples (see **Methodology** step 1). Varying the boundaries did not significantly modify the results.

GloBo ran 25 times using only the subset of attributes selected. Before each run, the training set (70% of data) and the test set (30% remaining) were chosen randomly. Each execution gave between 8 and 13 rules. Some of these rules bore a strong similarity with one another and could even be identical in distinct runs.

We focused on the following two rules (see **Methodology** step 2) because they were almost invariant from one run to another and covered a large amount of examples:

R1: *if* $X1 \neq C$ *and* $X1 \neq U$ *and* $|H|_A \leqslant 5$
and $|ES1|_{G-C} \in [6, 9]$ *and* $|ES2.5'|_U \leqslant 1$
then efficient_frameshifting.

This rule covers about 44% of the examples (training and test set) and no counter-examples.

R2: *if* $Y \neq G$ *and* $Z \neq G$ *and* $|H|_A \leqslant 4$
and $|SP|_C \geqslant 1$
and $\%ES1.5'_G \leqslant 65$ *and* $|ES1|_{G-C} \geqslant 6$
then efficient_frameshifting.

This rule covers about 33% of the examples (training and test set) and no counter-examples.

Note that these rules are not mutually exclusive: some examples are covered by both.

The presence of the attributes dealing with ES1 and ES2 in the rules is not surprising, as they are linked to the stability of the stems and hence to the stability of the pseudoknot. Interestingly, R1 and R2 focus on a rich G–C pair content of ES1. Moreover R2 gives an upper limit on the G content of ES1.5′. The other attributes discriminate between different families of mutants: R1 covers most SRV.wt-like sequences whereas R2 covers most IBV.m-like ones. Together they cover 65% of the examples. We therefore clearly obtain two rules of a disjunctive description of the frameshifting process. Moreover, the conjunctive form of each rule implies that the constraint links the values of several attributes together. Previously the ranges (that is, the sets of relevant values) of the attributes were determined independently.

To test these rules (see **Methodology** step 3), we designed *ab initio* sequences that satisfy them and that remain in the same general context: R1 was thus tested on SRV mutants (see Figure 3a) whereas R2 was tested on IBV mutants (see Figure 3b). We chose to focus on the stems in both series of constructs because they are known to be critical for frameshifting. These two figures give only the elements that were changed in the wild type. ES1 was modified in length for SRV mutants and in composition for IBV mutants. ES2 was also modified in IBV mutants. Given a sequence satisfying a rule, we wanted to verify whether mutants of this sequence that only differ from it in attributes that do not occur in the rule still promote frameshifting (of course these mutants satisfy this rule too).

The results in Figure 3 show that the efficiency level for SRV mutants was relatively stable (between 11% and 18%) and that most mutants were more efficient than the wild type. Concerning IBV mutants, the results show that, although the efficiency level varied from 9% to 25%, all the tested constructs were able to drive a significant frameshifting. In comparison, frameshifting efficiency of defective mutants, IBV pKA9 and IBV pKA96 (Napthine *et al.*, 1999), was 1.2% in our experimental system, similar to that obtained *in vitro* by (Napthine *et al.*, 1999) ($\leqslant$2%). However, important variations were observed between constructs showing the same proportion of G–C versus A–U pairs and differing only by their repartition (compare for example constructs IBV.s1 and s3). This is probably related to the three-dimensional structure of the artificial pseudoknot (Farabaugh, 1996).

Overall, these results demonstrate that frameshifting indeed occurs on constructs that follow one rule, increasing our confidence in the rules (see **Methodology** step 4).

Spacer

A detailed study of the spacer of 27 sites from 24 different viruses (see **Methodology** step 1) has revealed a striking regularity, in spite of the variability of its length[‡].

Notably, it was clear that in efficient -1 frameshifting sites specific nucleotides are found at given positions in the spacer. We verified that similarities are not due to homologies by performing pairwise alignments of nucleic and protein sequences from the 24 different viruses.

We focused on the three first positions of the spacer, where the consensus is particularly clear: the first nucleotide is either a G or a U; the second is either a G or an A, with three known exceptions (including the case of RSV where the spacer is only one nucleotide long) and eventually, at the third position, one finds also a G or an A, with three exceptions. We measured the frameshifting efficiency of a number of spacer mutants, based on the

[‡] http://www.igmors.u-psud.fr/rousset/bioinformatics/

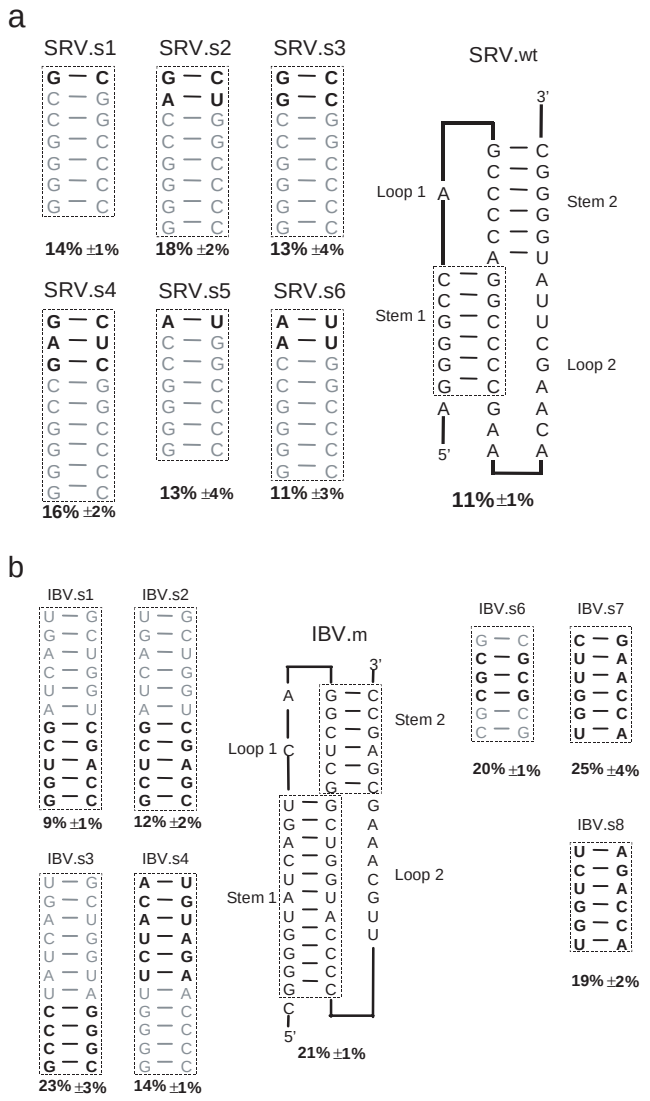


Fig. 3. Secondary-structure of the SRV (a) and IBV (b) pseudoknots. Dashed lines surround the region tested. Modified nucleotides are in bold and the frameshifting rate is indicated below. In the IBV mutants, stem 1 was modified only in IBV.s1 to IBV.s4 and stem 2 was modified only in IBV.s6 to IBV.s8.

wild-type IBV spacer. We systematically modified each of the first three nucleotides in turn. The results are in Table 1.

For each of the three positions, variation of a single base induces significant variations of the frameshifting level. An up to 4-fold difference was observed between constructs, but the extent of the effect seems to be more important for the first two nucleotides than for the third. Although the effect of the spacer sequence is important, the frameshifting efficiency never decreases below the 5% limit.

Table 1. Derived mutations of the three first nucleotides of the spacer (SP). Wild-type spacer is given by IBV.m (GGGUAC)

Construct	Spacer	FS −1 rate
IBV.m	**GGG**UAC	21% ± 1%
IBV.sp1	**A**GGUAC	15% ± 2%
IBV.sp2	**U**GGUAC	25% ± 4%
IBV.sp3	**C**GGUAC	6% ± 1%
IBV.sp4	G**A**GUAC	7% ± 2%
IBV.sp5	G**U**GUAC	19% ± 2%
IBV.sp6	G**C**GUAC	17% ± 1%
IBV.sp7	GG**A**UAC	15% ± 3%
IBV.sp8	GG**U**UAC	19% ± 2%
IBV.sp9	GG**C**UAC	11% ± 1%

Slippery sequence

Brierley *et al.* analyzed a large number of slippery heptamer sequences in the context of the IBV pseudoknot structure, in order to better understand the respective role of these two elements (Brierley *et al.*, 1992). Among the 64 possibilities for the X XXY YYZ heptamer, 44 mutations were created, the remaining 20 were thought to be non-functional.

We analyzed the frameshifting efficiency for the 64 mutants obtained from the wild-type IBV, X, Y and Z being successively replaced by A, C, G and U in the heptamer. We assign a level of 0 to the frameshifting level of the 20 sequences discarded (Brierley *et al.*, 1992). The frameshifting efficiency of the sequences measured below 1 were considered as 0.5. The attributes used were the same as above. First, we used a regression tree technique that (i) chooses relevant attributes to split the set of mutants into classes that share common properties of primary structure and have almost the same efficiency level and (ii) calculates the average efficiency level of each class. Note that the classes are not determined *a priori*. We got the following regression tree rules, where AL denotes the average level and s.d. the standard deviation.

If (Y = C *or* Y = G) *then* AL = 1.1 (s.d. 1.8)
If (Y = A *or* Y = U) *and* (Z = G) *then* AL = 2.68 (s.d. 2.8)
If (Y = A *or* Y = U) *and* (Z = A *or* Z = U *or* Z = C) *and* (X = C) *then* AL = 6.53. (s.d. 4.9)
If (Y = A *or* Y = U) *and* (Z = A *or* Z = U *or* Z = C) *and* (X = A *or* X = G *or* X = U) *then* AL = 22.7 (s.d. 9.7)

The order that the rules and their conditions are written in is important: it gives the attributes in decreasing order according to the minimization of the intrasubset variation. These rules confirm results already established by Brierley for IBV (Brierley *et al.*, 1992):

(1) only triplets YYY of A and of U are functional;

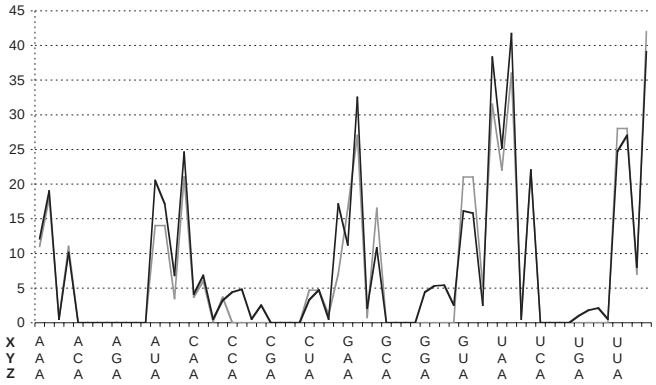


Fig. 4. Frameshifting efficiencies of slippery site mutants in IBV. Comparison between experimental results of Brierley *et al.* (1992) (in black) and the formula above (in grey). The sequences XYC, XYG and XYU lie between XYA and X′Y′A.

(2) the identity of Z in the heptamer is a critical determinant of frameshifting efficiency;

(3) triplets XXX of A, C, G and U are functional but C-triplets are the least slippery.

Moreover, we have a crude approximation of the expected efficiency in each case.

We then used *ad hoc* mathematical tools to show how precisely the frameshifting level depends on the identity of the nucleotides.

Figure 4 presents the frameshifting level for the 64 mutants of the heptamer (they are sorted in lexicographical order). It reveals a strong regularity in the frameshifting levels. Not only can one note a pseudoperiod in the values (as expected given the known rules about $Y = A$ or $Y = U$), but also the main peaks remarkably line up.

The regression equation given below (see the grey graphical on Figure 4) is an approximation function that expresses as closely as possible the 64 sequence levels. (In this equation, expressions in square brackets must be evaluated at 1 if the conditions are satisfied, at 0 otherwise).

$F(XXXYYYZ) = F1xF2xF3$, with

$F1 = (1/3)[X = C] + [X = A] + (3/2)[X = G] + 2[X = U]$

$F2 = [Y = A \text{ or } Y = U]$

$F3 = 11 + 3[Y = U] - 10.5[Z = G] + 7[Y = A][Z = C] + 7[Y = U][Z = U]$.

This equation reveals an unknown characteristic, namely the very interesting role of the XXX triplet expressed in F1. It acts as a multiplicative factor. Thus it appears that the identity of X influences the frameshifting efficiency following the order: $C < A < G < U$.

DISCUSSION

This study validates the methodological approach used to refine the current model. Specifically, our learning method allowed us to identify new features that may be involved in frameshifting and to specify the range of the values of the corresponding attributes. As we saw above (see **Results**), the set of attributes used was not sufficient to distinguish all the examples from the counter-examples. This will lead us to introduce more attributes. Using a disjunctive learning tool such as GloBo led us to two rules which were then verified experimentally. The relevance of their attributes to the mechanism of frameshifting was tested by creating artificial frameshifting sites that follow the rules identified by our bioinformatics approach. The capacity of these sites to induce frameshifting was then assessed *in vivo* in yeast cells. As these two rules do not cover all the examples, other rules should be investigated.

The originality of the obtained rules is derived from their conjunctive form. Namely, each rule provides a set of values that must be assigned to several attributes concerning possibly distinct components of the model.

Our results show that the spacer is involved in the efficiency of frameshifting. Although it has been known for many years that the length of the spacer region is crucial in frameshifting (Brierley, 1995), it has been recently shown that the sequence could also be important in bacterial frameshifting sites (Bertrand *et al.*, 2002). The results presented here demonstrate that this sequence is also important in eukaryotic frameshifting. Different mechanisms could account for this effect: the nucleotides may directly interact with components of the translational machinery (i.e. ribosomal RNA or protein), or indirectly, by codon–anticodon interaction, or through the availability of the corresponding tRNA. In this case, the spacer could not be seen as a sum of individual nucleotides but as a unit. If this is so, the activity of a sequence may not be defined by a single nucleotide but by a suite of nucleotides, and there may be different suites that work well. In order to shed some light on this point, the spacer should be analyzed systematically using a combinatorial approach, similar to a SELEX experiment, that selects the most efficient spacer sequences. Such a procedure, adapted for translational regulation, is available and in use in our laboratory to analyze programmed readthrough (Namy *et al.*, 2001).

Another interesting point is the multiplicative role of the XXX triplet in frameshifting efficiency. This suggests that the core frameshifting signal could in fact be a tetranucleotide and that the XXX triplet is used to modulate the efficiency. It is well known that the bacterial frameshifting mechanism often involves only a tetramer sequence of the form Y YYZ. In eukaryotes also, although simultaneous tandem tRNA slippage is the main mechanism, single slippage has a place, with a low but

significant frequency, at a given site (Jacks *et al.*, 1988; Yelverton *et al.*, 1994).

It is worth noting that this study was only conducted in the IBV context. However, the few experiments that have been done up to now on different kinds of viruses have shown that hierarchies of heptamer efficiency are generally conserved from one virus to another (Farabaugh, 1996). It would be of particular interest to complete the work of Brierley, by testing the remaining 20 heptamers in eukaryotic models and to test the frameshifting efficiency directed by the 64 possible heptamers and the 16 possible tetramers in a bacterial model. These 16 tetramer sequences should also be tested in eukaryotic cells, to assess whether some of them allow efficient frameshifting.

ACKNOWLEDGEMENTS

The authors are deeply grateful to Celine Fabret for critical reading of this manuscript and helpful suggestions. We also thank anonymous referees for pertinent suggestions.

This work was supported by a CNES grant on the Preparatory Program Mars Sample Analysis and by a CNRS INRA - INRIA - INSERM Bioinformatics grant.

REFERENCES

Baranov,P., Gurvich,O., Fayet,O., Prere,M., Miller,W., Gesteland,R., Atkins,J. and Giddings,M. (2001) RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Res.*, **29**, 264–267. http://recode.genetics.utah.edu.

Bertrand,C., Prere,M.-F., Gesteland,R., Atkins,J. and Fayet,O. (2002) Influence of the stacking potential of the base 3′ of tandem shift codons on −1 ribosomal frameshifting used for gene expression. *RNA*, **8**, 16–28.

Bidou,L., Stahl,G., Hatin,I., Namy,O., Rousset,J.-P. and Farabaugh,P. (2000) Nonsense-mediated decay mutants do not affect programmed −1 frameshifting. *RNA*, **6**, 952–961.

van Batenburg,F., Gultyaev,A., Pleij,C. and Olienhoek,J. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204. http://wwwbio.LeidenUniv.nl/∼Batenburg/PKB.html

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth, Monterey, CA.

Brierley,I. (1995) Ribosomal frameshifting on viral RNAs. *J. Gen. Virol.*, **76**, 1885–1892.

Brierley,I., Rolley,N.J., Jenner,A.J. and Inglis,S.C. (1991) Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.*, **220**, 889–902.

Brierley,I., Jenner,A.J. and Inglis,S.C. (1992) Mutational analysis of the 'slippery-sequence' component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol.*, **227**, 463–479.

Chen,L.X. and Tinoco,I. (1995) The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *J. Mol. Biol.*, **247**, 963–978.

Chen,X., Kang,H., Shen,L.X., Chamorro,M., Varmus,H.E. and Tinoco,I. (1996) A characteristic bent conformation of RNA pseudoknots promotes −1 frameshifting during translation of retroviral RNA. *J. Mol. Biol.*, **260**, 479–483.

ten Dam,E., Brierley,I., Inglis,S. and Pleij,C. (1994) Identification and analysis of the pseudo-knot containing *gag-pro* ribosomal frameshift signal of simian retrovirus −1. *Nucleic Acids Res.*, **22**, 2304–2310.

ten Dam,E.B., Verlaan,P.W. and Pleij,C.W. (1995) Analysis of the role of the pseudoknot component in the SRV-1 *gag-pro* ribosomal frameshift signal: loop lengths and stability of the stem regions. *RNA*, **1**, 146–154.

Dinman,J.D., Icho,T. and Wickner,R.B. (1991) A −1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion domain. *Proc. Natl Acad. Sci. USA*, **88**, 174–178.

Dinman,J., Ruiz-Echevarria,M. and Peltz,S. (1998) Translating old drugs into new treatments: ribosomal frameshifting as a target for antiviral agents. *Trends Biotechnol.*, **16**, 190–196.

Dujon,B., Alexandraki,D., Andre,B., Ansorge,W., Baladron,V., Ballesta,J., Banrevi,A., Bolle,P., Bolotin-Fukuhara,M. Bossier,P. *et al.* (1994) Complete DNA sequence of yeast chromosome XI. *Nature*, **369**, 371–378.

Farabaugh,P. (1996) Programmed translational frameshifting. *Microbiological review*, **60**, 103–134.

Gesteland,R. and Atkins,J. (1996) Recoding: dynamic reprogramming of translation. *Annu. Rev. Biochem.*, **65**, 741–768.

Gesteland,R., Weiss,R. and Atkins,J.F. (1992) Recoding: reprogrammed genetic decoding. *Science*, **257**, 1640–1641.

Giedroc,D.P., Theimer,C.A. and Nixon,P.L. (2000) Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.*, **298**, 167–185.

Goffeau,A., Barrell,B., Bussey,H., Davis,R., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J., Jacq,C. and Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.

Hammell,A.B., Taylor,R.C., Peltz,S.W. and Dinman,J.D. (1999) Identification of putative programmed −1 ribosomal frameshift signals in large DNA databases. *Genome Res.*, **9**, 417–427.

Jacks,T. and Varmus,H. (1985) Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. *Science*, **230**, 1237–1242.

Jacks,T., Power,M., Masiarz,F., Luciw,P., Barr,P. and Varmus,H. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, **331**, 280–283.

Kang,H., Hines,J.V. and Tinoco,I. (1996) Conformation of a non-frameshifting RNA pseudoknot from mouse mammary tumor virus. *J. Mol. Biol.*, **259**, 135–147.

Kim,Y.-G., Su,L., Maas,S., O'Neill,A. and Rich,A. (1999) Specific mutations in a viral RNA pseudoknot drastically change ribosomal frameshifting efficiency. *Proc. Natl Acad. Sci. USA*, **96**, 14234–14239.

Kollmus,H., Hentze,M. and Hauser,H. (1996) Regulated ribosomal frameshifting by an rna-protein interaction. *RNA*, **2**, 316–323.

Kontos,H., Napthine,S. and Brierley,I. (2001) Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Mol. Cell. Biol.*, **21**, 8657–8670.

Liphardt,J. (1999) *The mechanism of −1 ribosomal frameshifting: experimental and theoretical analysis*, PhD thesis, Churchill College, Cambridge.

Lopinski,J., Dinman,J. and Bruenn,J. (2000) Kinetics of ribosomal pausing during programmed −1 translational frameshifting. *Mol. Cell. Biol.*, **20**, 1095–1103.

Marczinke,B., Fisher,R., Vidakovic,M., Bloys,A. and Brierley,I. (1998) Secondary structure and mutational analysis of the ribosomal frameshift signal of Rous sarcoma virus. *J. Mol. Biol.*, **284**, 205–225.

Medigue,C., Rose,M., Viari,A. and Danchin,A. (1999) Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Res.*, **9**, 1116–1127.

Mejlhede,N., Atkins,J. and Neuhard,J. (1999) Ribosomal −1 frameshifting during decoding of *Bacillus subtilis* cdd occurs at the sequence CGA AAG. *J. Bacteriol.*, **181**, 2930–2937.

Namy,O., Hatin,I. and Rousset,J.P. (2001) Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep.*, **2**, 787–793.

Napthine,S., Liphardt,J., Bloys,A., Routledge,S. and Brierley,I. (1999) The role of RNA pseudoknot stem 1 length in the promotion of efficient −1 ribosomal frameshifting. *J. Mol. Biol.*, **288**, 305–320.

Quinlan,J. (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann, San Fransisco.

Shigemoto,K., Brennan,J., Walls,E., Watson,C., Stott,D., Rigby,P. and Reith,A. (2001) Identification and characterisation of a developmentally regulated mammalian gene that utilises −1 programmed ribosomal frameshifting. *Nucleic Acid Res.*, **29**, 4079–4088.

Somogyi,P., Jenner,A.J., Brierley,I. and Inglis,S.C. (1993) Ribosomal pausing during translation of an RNA pseudoknot. *Mol. cell. Biol.*, **13**, 6931–6940.

Srinivasan,A., King,R. and Bristol,D. (1999) An assessment of submissions made to the Predictive Toxicology Evaluation Challenge. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, Vol. 1.

Torre,F. (2000) *Intégration des biais de langage à l'algorithme générer-et-tester—Contributions à l'apprentissage disjonctif*, PhD thesis, LRI, Université Paris-Sud, Orsay.

Tsuchihashi,Z. and Kornberg,A. (1990) Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc. Natl Acad. Sci. USA*, **87**, 2516–2520.

Tu,C., Tzeng,T.-H. and Bruenn,J.A. (1992) Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc. Natl Acad. Sci. USA*, **89**, 8636–8640.

Witten,I.H. and Frank,E. (2000) *Data mining*. Morgan Kaufmann, San Fransisco, http://www.cs.waikato.ac.nz/ml/weka.

Yelverton,E., Lindsley,D., Yamauchi,P. and Gallant,J. (1994) The function of a ribosomal frameshifting signal from human immunodeficiency virus-1 in *Escherichia coli*. *Mol. Microbiol.*, **11**, 303–313.

Article 2

Molecular Cell, Vol. 17, 61–68, January 7, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.molcel.2004.12.009

An Extended Signal Involved in Eukaryotic −1 Frameshifting Operates through Modification of the E Site tRNA

Michaël Bekaert and Jean-Pierre Rousset*
Institut de Génétique et Microbiologie
Centre National de la Recherche Scientifique-Unité Mixte de Recherche 8621
Bâtiment 400
Université Paris-Sud
91405 Orsay Cedex
France

Summary

By using a sensitive search program based on hidden Markov models (HMM), we identified 74 viruses carrying frameshift sites among 1500 fully sequenced virus genomes. These viruses are clustered in specific families or genera. Sequence analysis of the frameshift sites identified here, along with previously characterized sites, identified a strong bias toward the two nucleotides 5′ of the shifty heptamer signal. Functional analysis in the yeast *Saccharomyces cerevisiae* demonstrated that high frameshifting efficiency is correlated with the presence of a Ψ39 modification in the tRNA present in the E site of the ribosome at the time of frameshifting. These results demonstrate that an extended signal is involved in eukaryotic frameshifting and suggest additional interactions between tRNAs and the ribosome during decoding.

Introduction

The universal rules of genetic translation have long been known. However, the present situation is more complex. Firstly, the genetic code is not universal. In several organelles and protists, all genes are decoded by a variant code in which a stop codon reads as a sense codon. Moreover, the genetic code can be locally extended by alteration of standard rules in an individually specific manner for given mRNAs. Such extensions of the genetic code are termed reprogrammed genetic decoding or recoding events. Recoding is determined by particular sequences that force the ribosome to escape standard translation. In vivo, recoding is often in competition with standard decoding and permits the synthesis of an elongated polypeptide. Only a defined proportion of the ribosomes translating a given recoded mRNA actually use reprogrammed genetic decoding. Recoding extends the possibilities of increased diversity in gene expression or regulation.

During programmed −1 frameshifting, ribosomes switch to an alternative frame at a specific shift site, and classical triplet decoding follows. Efficient −1 frameshifting necessitates specific signals on the mRNA. Basically, the model proposed by Jacks and Varmus identified two components of frameshift signals: a slippery heptamer sequence, X XXY YYZ (the frame of the initiator AUG is indicated by a space), and a downstream structural element, often a pseudoknot or a stem-loop (Jacks et al., 1988). The secondary structure induces a ribosomal pause at the slippery site, and the heptamer sequence allows the slippage of ribosome bound A and P site tRNAs by one nucleotide in the 5′ direction. The pause increases the probability of ribosomal movement in the 5′ direction, although pausing seems necessary, but not sufficient, for stimulation (Kontos et al., 2001). More recently, a new sequence element—the spacer region located between the heptamer and the secondary structure—has been shown to modulate frameshift efficiency both in prokaryotes and eukaryotes (Bekaert et al., 2003; Bertrand et al., 2002).

Most −1 frameshift events have been reported in viruses and transposons, where they serve in synthesis of replicase activities. This mode of expression allows both a very precise control of the Gag-Pol/Gag ratio and a means of incorporating the enzymes necessary for the replication cycle into the viral particle. Enhancement or reduction of the efficiency of this mechanism can influence virus viability (Dinman and Wickner, 1992; Shehu-Xhilaga et al., 2001). Very few cellular genes with −1 frameshifting are presently known: the *dnaX* gene of *Escherichia coli* (Tsuchihashi and Kornberg, 1990), the *cdd* gene of *Bacillus subtilis* (Mejlhede et al., 1999), and the *Edr* gene of mouse (Shigemoto et al., 2001). Even though the identification of novel genes regulated by −1 frameshifting constitutes one of the postgenomic challenges, there is presently no general method to do so (Namy et al., 2004). We recently began a computational approach to model frameshifting sites. The rationale was to observe more elements of the sequence in order to obtain a more precise description of characterized frameshift signals, identify pertinent characteristics, and develop appropriate algorithms. This approach previously allowed us to identify a strong bias in the spacer sequence of eukaryotic viral frameshift signals. This bias was shown to be functionally relevant for the frameshifting mechanism (Bekaert et al., 2003). However, only few frameshift signals have been functionally characterized, although several other sites are suspected to possess such signals from sequence analysis. The original goal of this study was to identify new −1 frameshift sites in viruses, so as to enrich our collection of sites in the modeling process. We first quantified frameshifting efficiency directed by 20 viruses carrying a putative or characterized frameshift site. From these results, we were able to develop a sensitive search based on HMM. This algorithm enabled us to detect −1 frameshift sites in 74 viruses among the available, fully sequenced viruses. Sequence alignment of these virus sites identified a strong composition bias just upstream of the shifty heptanucleotide site. We demonstrate that the sequence of this region actually affects frameshift efficiency in *Saccharomyces cerevisiae*. These results point to the possible involvement of pseudouridine at position 39 of the tRNA present in the E site of the host cell ribosome in modulating frameshifting efficiency. By using *PUS3* yeast mutants lacking Ψ38/39 modifications, we show that frameshifting efficiency is modu-

*Correspondence: jean-pierre.rousset@igmors.u-psud.fr

lated by the modification status of the E site tRNA. Overall, our results propose an extended model for −1 frameshift sites.

Results and Discussion

Characterization of Viral −1 Frameshift Sites

The RECODE database resource (http://recode.genetics.utah.edu, Baranov et al., 2001) describes 35 viruses suspected or demonstrated to carry a frameshift site. Few −1 frameshift signals are fully documented. The genomes of these 35 viruses are entirely sequenced, and only five sites are precisely characterized, including the structure of the stimulatory pseudoknot (BWYV, HIV-1, MMTV, PEMV-1, and SRV-1). For the remaining sites, 13 have been analyzed by extensive directed mutagenesis coupled with quantification of frameshifting efficiency, but the others are only partially characterized. Most of the sites are therefore putative; i.e., they carry the typical heptamer and secondary structure but have never been proven to be functional.

Initially, we functionally characterized a larger number of viruses containing a putative frameshifting site. To explore the widest viral diversity possible (order, family, and genus), we deduced a neighbor-joining tree from the 35 viruses, based on the multiple alignment of the polymerase protein sequence (Figure 1). From this tree, we selected a subset of 20 viruses representative of the global viral diversity. To assay the frameshift competence of each putative site, we cloned the entire viral −1 frameshift region of the 20 representative viruses in a dual-reporter vector and estimated in vivo frameshifting efficiency in yeast (see Experimental Procedures). Frameshift sites from different eukaryotic species have been shown to function in yeast (Bekaert et al., 2003; Stahl et al., 1995). The existence of a functional frameshift signal was demonstrated for all candidates (Table 1) except the *Sugarcane yellow leaf virus* (ScYLV). It is unlikely that the low level of expression of ScYLV is due to the use of a heterologous host cell, because ten frameshift sites from other plant viruses are functional in our assay. This site might be nonfunctional or carry polymorphic variations. The −1 frameshifting frequencies varied between 8% and 31%, compatible with those previously obtained with in vitro or in vivo assays (e.g., EAV 15%–20%, den Boon et al. [1991]; BWYV 5%, Kim et al. [1999]; HCV 20%–30%, Herold and Siddell [1993]; MMTV 20%, Chamorro et al. [1992]; and RSV 4%, Marczinke et al. [1998]). Although our assay does not provide exact frameshift rates in natural host cells (except for L-A and LB-C viruses, which are natural in *S. cerevisiae*), these results strongly suggest that most of the frameshift sites identified purely by sequence analysis are in fact functional.

We then aligned the newly characterized sites with sites already identified. Strikingly, we observed an important bias not only at the slippery heptamer but also in the spacer region and just upstream of the heptamer. The upstream bias was never before observed, and its detailed analysis is presented below.

Sensitive Search of Viral Frameshift Sites

We established a HMM profile of efficient viral −1 frameshift signals with the alignment of the slippery regions from the 20 viruses that we had functionally characterized (see Experimental Procedures) and used it to search the GenBank viral genome database (release 02/10/2004). 285 motifs were identified and subsequently manually inspected to eliminate false positives. We checked (1) that the first nucleotide of the heptamer is in frame with the AUG of the upstream coding region, (2) for a protein motif associated to the upstream and downstream coding regions, and (3) for the presence of a potential secondary structure downstream of the heptamer. By this procedure, we identified 74 frameshift sites in viral genomes. Most false positives exhibited no secondary structure after the shifty site and were found in the large and highly complex *herpesvirus*, *papillomavirus*, and *nucleopolyhedrovirus* genomes. We consider this assessment to be accurate, because it depends not only on in silico methods but also on the biological assay of the HMM learning set. It is noteworthy that this method is very efficient, even though we did not take into account the stimulatory secondary structures when we defined the profile. RNA folding algorithms are time consuming and cannot be restrained to a defined window in the vicinity of the heptamer. Moreover, the theoretical evaluation of thermodynamic stability of secondary structures is not accurate for pseudoknots (Walter et al., 1994).

With the HMM profile based on only 20 sites, we were able to find all known frameshifting viruses and 39 that are new or uncharacterized (Table 2). The list of the viruses with the position of their frameshift signals is in Supplemental Table S1 available online at http://www.molecule.org/cgi/content/full/17/1/61/DC1/. Ten putative frameshift sites were never previously annotated and are associated to an upstream and a downstream coding region. 12 frameshifting structures were already annotated as such in the RECODE database, and eight were only annotated in the sequence field of GenBank or in relation to a publication that did not mention any evidence of frameshift. For the remaining 44 sequences, a site was suspected, but it was not precisely localized between two coding regions. For those, we were able to propose a precise position for the frameshift event and in some cases, a more accurate annotation. For example, for the *Ovine astrovirus* (ssRNA$^+$, *Astroviridae* family), a putative −1 frameshift event sequence was previously reported but without data on the position of the frameshift site (Jonassen et al., 1998). By adding the newly identified virus to the initial set, we established an enhanced profile HMM of viral −1 frameshift sites (see Experimental Procedures), available as Supplemental Data.

Among 82 viral families, only seven are involved in −1 frameshifting: *Astroviridae*, *Arteriviridae*, *Coronaviridae*, *Luteoviridae*, *Retroviridae*, *Tombusviridae*, and *Totiviridae*. Within each family, only a few subfamilies/genera were capable of −1 frameshifting (see Supplemental Table S1 for details). However, in this latter case, all members of the genus submitted to HMM analyses appear capable of −1 frameshifting: they carry not only the HMM profile but also secondary structures as a canonical frameshift signal (Supplemental Table S1). For example, manual checking of the *Poleroviruses* found by using HMM successfully identifies a pseudoknot three to nine nucleotides downstream from the heptamer site.

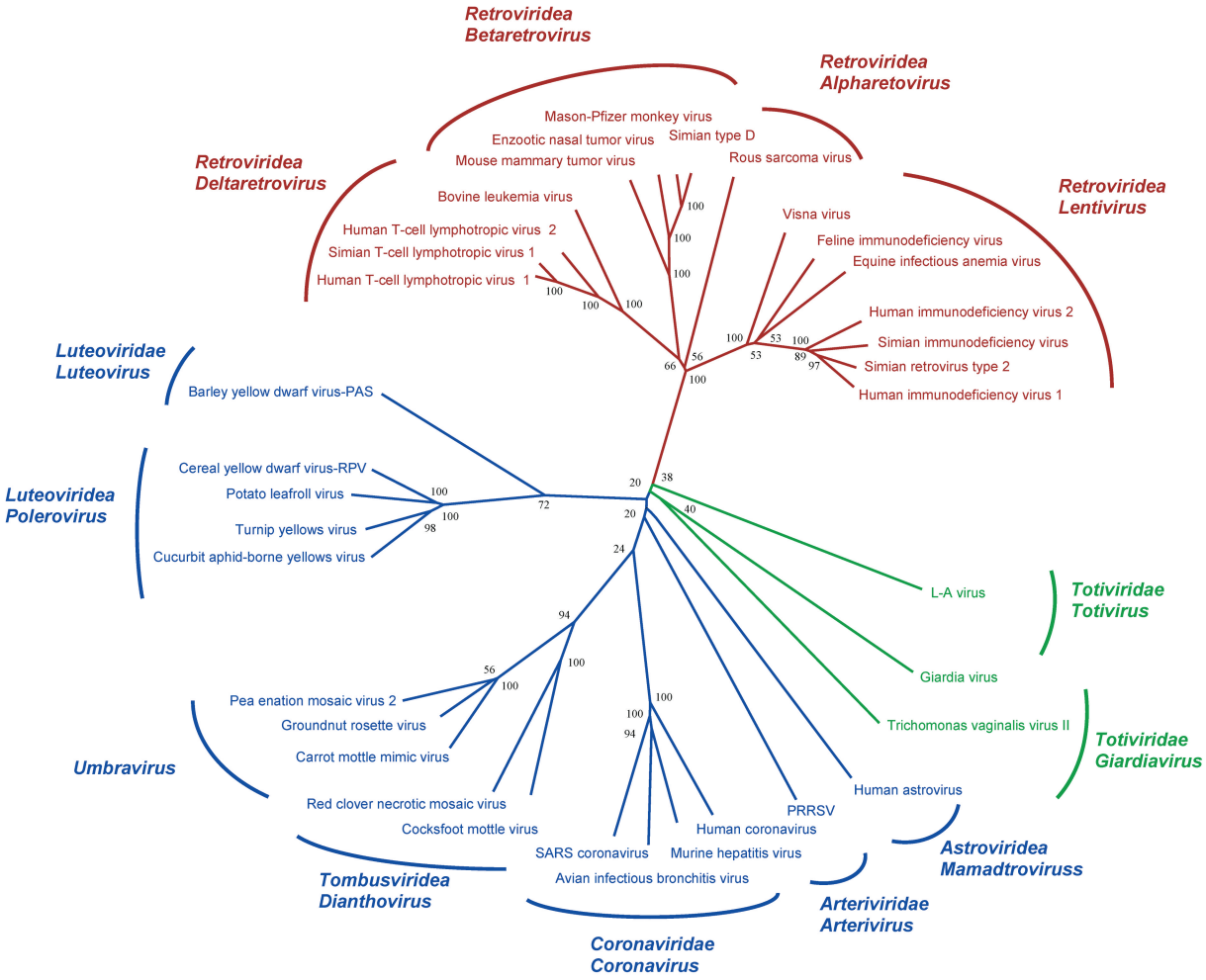


Figure 1. Unrooted Neighbor-Joining Phylogeny Based on the Amino Acid Sequence of the *Pol* Protein

Labels at branch tips represent taxa as provided by the Universal Virus Database, ICTVdB (http://phene.cpmc.columbia.edu; Büchen-Osmond, 2003). Numbers at branch nodes indicate the percentage of bootstrap support for that node based on 1000 replications. Color codes are according to the clade, with the following scheme: ssRNA-RT viruses (red), positive-sense ssRNA viruses (blue), and dsRNA viruses (green). Abbreviations: PRRSV, *Porcine reproductive and respiratory syndrome virus*.

In the *Totiviridae* family (dsRNA virus), only the *Totivirus* genus, represented by the L-A and LB-C viruses (two yeast viruses), has a −1 frameshift signal. Moreover, all viruses from the *Totivirus* and the *Giardiavirus* genera appear to exhibit such sites. The only exception is the *Ustilago maydis* virus H1 (*Totivirus* genus): it shows a perfectly conserved canonical slippery sequence and a strong pseudoknot, despite an in-frame *gag-pol* gene. This could be due to either a sequencing/annotation error or to an unusual configuration where canonical decoding would be responsible for the synthesis of the Gag-Pol protein, whereas the frameshift would lead to the production of only the Gag domain, reminiscent of the control of the *dnaX* gene expression in *E. coli* (Tsuchihashi and Kornberg, 1990). Another ambiguous case concerns the *Helminthosporium victiriae virus 190S*, where previous experimental investigation of gene expression concluded that translation of the second ORF is initiated on its own internal AUG codon (Huang and Ghabrial, 1996). However, this does not exclude the possibility that both mechanisms are at play to express different polypeptides involved in polymerase activity, as observed in some bacterial transposons (Fayet et al., 1990).

Upstream Bias

As mentioned above, the alignment of the initial set of 20 viruses −1 frameshift signal sequences revealed that the base composition around the slippery sequence follows a preferential use of nucleotides. Composition bias in the spacer has been previously reported for first nucleotides (Bekaert et al., 2003; Bertrand et al., 2002); the second part of the bias has been reported in relation with the first stem composition bias (ten Dam et al., 1990). Bias in upstream sequences of the slippery regions was accurately detected in the larger scale data derived from the 74 virus sequences identified through an order 1 HMM search where the probability of a given nucleotide is dependent on the identity of the previous nucleotide. Accordingly, Figure 2 shows the bias of dinucleotide distribution. The χ^2 score for the last dinucleotide position before the heptamer is 80 with 15 degrees

Table 1. Nucleic Acid Alignment of Heptameric Sequence

Virus Acronym	FS-1		Slippery Region		Genbank
BChV	15.8% ± 2%	cacaucugcC	**GGgAAAu**	gGacuGaGcG	NC_002766
BLV gag/pro	8.1% ± 1%	cccucaaaUC	**aaAAAAC**	UaauAGaGGG	NC_001414
BWYV	12.0% ± 1%	ccaagagcUC	**GGgAAAC**	gGGagaGcGG	NC_004756
BYDV	12.2% ± 1%	uugacucugu	**GGguuuu**	UaGagGGGcu	NC_002160
CABYV	17.5% ± 1%	aauacgagUC	**GGgAAAC**	gGGcAGGcGG	NC_003688
EIAV	7.0% ± 1%	gaaguguucC	**aaAAAAC**	gGGagcaaGG	NC_001450
FIV	9.0% ± 1%	gaaagaauUC	**GGgAAAC**	UGGaAGGcGG	NC_001482
HIV1	6.0% ± 1%	gacaggcuaa	**uuuuuua**	gGGaAGaucu	NC_001802
IBV	19.3% ± 1%	auaagaauUa	**uuuAAAC**	gGGuAcGGGG	NC_001451
L-A	10.0% ± 1%	guacucagca	**GGguuua**	gGaguGGuaG	NC_003745
L-BC	13.0% ± 2%	cugagaagUu	**GGauuuu**	cGuguaGcaG	NC_001641
LDV	13.1% ± 1%	aggcaucggC	**uuuAAAC**	UGcuAGccac	NC_002534
MMTV gag/pro	20.2% ± 2%	cugaaaauUC	**aaAAAAC**	UuGuAaaGGG	NC_001503
PEMV1	31.0% ± 2%	ccagacgcUC	**GGgAAAC**	gGauuauucc	NC_003629
PLRV	19.0% ± 1%	caaacaagcC	**GGgAAAu**	gGGcAaGcGG	NC_001747
PLRV-W	17.8% ± 2%	caaacaagcC	**uuuAAAu**	gGGcgaGcGG	Y07496
PRRSV	15.7% ± 1%	aggagcagUg	**uuuAAAC**	UGcuAGccGc	NC_001961
SARS	10.3% ± 1%	caucaacgUu	**uuuAAAC**	gGGuuuGcGG	NC_004718
ScYLV	0.7% ± 0%	cuccagacca	**GGgAAAu**	gaGccaaGuG	NC_000874
SRV1 gag/pro	13.0% ± 2%	caccccauca	**GGgAAAC**	gGacuGaGGG	NC_001551
Pseudoconsensus		xxxxxxxxUC	**GGAAAAC**	UGGxAGGGGG	

Nucleotides in agreement with the functional pseudoconsensus inferred from the HMM profile are in uppercase. Acronyms are as follows: BChV, *Beet chlorosis virus*; BLV, *Bovine leukemia virus* (gag/pro junction); BWYV, *Beet western yellows virus*; BYDV, *Barley yellow dwarf virus*; CABYV, *Cucurbit aphid-borne yellows virus*; EIAV, *Equine infectious anemia virus*; FIV, *Feline immunodeficiency virus*; HIV1, *Human immunodeficiency virus 1*; IBV, *Avian infectious bronchitis virus*; L-A, *Saccharomyces cerevisiae virus L-A*; L-BC, *Saccharomyces cerevisiae virus L-BC*; LDV, *Lactate dehydrogenase-elevating virus*; MMTV, *Mouse mammary tumor virus* (gag/pro junction); PEMV1, *Pea enation mosaic virus 1*; PLRV, *Potato leafroll virus*; PLRV-W, *Potato leafroll virus*, Germany strain (Wageningen); PRRSV, *Porcine reproductive and respiratory syndrome virus*; SARS, *SARS coronavirus*; ScYLV, *Sugarcane yellow leaf virus*; and SRV1, *Simian type D virus 1* (gag/pro junction).

of freedom, which makes it significant for a p value of 6.4×10^{-11}. The $-4/-5$ position also seems biased, but this has not been analyzed further.

To determine the role of this dinucleotide in frameshifting, we constructed dual-reporter vectors with the 16 possible sequences within the context of the wild-type (wt) frameshift signal of the *Avian infectious bronchitis virus* (IBV), because it has been extensively used as a model virus for -1 frameshifting studies (Brierley et al., 1991, 1992). Table 3 shows that a 3.3-fold variation was found between the frameshifting efficiencies directed by these 16 IBV variant sites. Compared to the wt sequence, the frameshifting level is significantly reduced (p value $< 10^{-4}$) in ten of the mutants.

Role of Base 39 of tRNA

The dinucleotide situated 5′ of the heptamer corresponds to the first two nucleotides of the preceding codon; its impact can thus be interpreted as an effect either of the amino acid, the codon, or the decoding tRNA. Because it was previously shown that tRNA modification can affect recoding efficiency (Lecointe et al., 2002; Licznar et al., 2003), we looked for a bias in modifications of tRNA involved in decoding high and low frameshifting constructs. A correlation between the presence of pseudouridine at position 39 (Ψ39) of the tRNA anticodon domain was observed (Table 3): all constructs that exhibited a high-frameshifting level use a cognate (or near-cognate) tRNA carrying the Ψ39 modification. Conversely, the sequences that do not involve a codon decoded by a tRNA with the Ψ39 modification direct low-frameshifting efficiency. This observation prompted us to investigate the effect of the mutation of

Table 2. Detailed Distributions of the HMM Profile Results from GenBank Compared to the Initial Data Set (RECODE database)

Status	GenBank (1500 virus)	RECODE (35 viruses)
New viruses	10	–
New annotations	32	15
Frameshift localized	12	8
Already annotated	20	12
	74	35

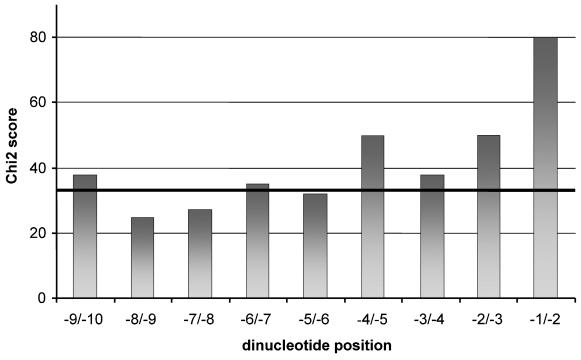


Figure 2. Estimation of Dinucleotide Usage Bias

χ^2 values were calculated at each position upstream of the slippery heptamer. Black line represents the threshold score for a p value of 1% with 15 degrees of freedom. Higher values are more indicative of a significant bias. The first dinucleotide is strongly significant (p value = $6.4\ 10^{-11}$).

Table 3. Effect of the Upstream Dinucleotide on −1 Frameshifting Efficiency

Plasmids	Modified Sequence	tRNA (Anticodon loop)	Frameshift
pAC.5.AA	aau AAu **uua aac**	ACU **GUU** t^6 AAΨ	18.0% ± 2%
pAC.5.AC	aau ACu **uua aac**	Am^5CU **IGU** t^6 AAΨ	22.0% ± 1%
pAC.5.UA	aau UAu **uua aac**	ACU **GΨA** i^6 AAΨ	19.3% ± 1%
pAC.5.UC	aau UCu **uua aac**	AΨU **IGA** i^6 AAΨ	22.1% ± 2%
pAC.5.UG	aau UGu **uua aac**	AΨU **GCA** i^6 AAΨ	19.0% ± 2%
pAC.5.UU	aau UUu **uua aac**	ACmU **GmAA** YAΨ	21.0% ± 1%
pAC.5.AG	aau AGu **uua aac**	CCU **GCU** AAG	9.3% ± 1%
pAC.5.AU	aau AUu **uua aac**	GCU **IAU** t^6 AAC	7.0% ± 1%
pAC.5.CA	aau CAu **uua aac**	GΨU **GUG** m^1 GCC	9.0% ± 1%
pAC.5.CC	aau CCu **uua aac**	CUU **AGG** GUG	9.0% ± 1%
pAC.5.CG	aau CGu **uua aac**	GCU **ICG** AAC	7.5% ± 1%
pAC.5.CU	aau CUu **uua aac**	GUC **GAG** GUC	10.0% ± 2%
pAC.5.GA	aau GAu **uua aac**	CΨU **GUC** m^1GCG	9.4% ± 1%
pAC.5.GC	aau GCu **uua aac**	CUU **IGC** m^1IΨG	8.0% ± 1%
pAC.5.GG	aau GGu **uua aac**	GΨU **GCC** AΨC	8.5% ± 1%
pAC.5.GU	aau GUu **uua aac**	CΨU **IAC** ACG	6.7% ± 1%

The FY strain was transformed with one of the plasmids harboring the test sequence as indicated (from 5′ to 3′). Frameshifting efficiencies were measured at 30°C, and the data are expressed as percentages. Codon including the dinucleotides are underlined and anticodon of tRNA anticodon loops are in bold (Lecointe, 2002); heptamers are in bold and dinucleotides in uppercase.

the *PUS3* gene, whose product is specifically responsible for the Ψ39 modification (Lecointe et al., 1998). If Ψ39 is actually involved, inactivation of *PUS3* should result in a lower frameshift efficiency.

Two low- and two high-frameshift rate constructs were tested in modification mutants (Table 4). With the low-frameshifting rate subset (frameshift efficiency lower than 10%), *pus3*Δ mutants show no significant effect (Table 4). In contrast, with the high-frameshifting rate subset (frameshift efficiency higher than 18%), which involves decoding by a Ψ39 modified tRNA, a reduced frameshifting frequency was observed in *pus3*Δ mutants. This frequency was similar to that directed by the low-frameshifting rate subset, indicating that most of the effect was reversed in the mutant. We verified that the effect is actually due to the modifying activity of Pus3p and not to a possible chaperone-like activity by using the *pus3*[D151A] mutant, which harbors a mutation in the active site of the PUS3 protein. In this mutant, the high-frameshifting constructs yield lower frameshifting efficiency, as in a *pus3*Δ mutant context (Table 4).

The effect of the dinucleotide upstream of the heptamer suggests that the three ribosomal site tRNAs are involved in the mechanism of −1 frameshifting. However, although the mechanism of frameshifting in eukaryotes is thought to involve mostly tandem slippage of the tRNAs occupying the A and P sites, single slippage at the P site has been reported to occur (Jacks et al., 1988). If this is the case in the experimental system used here, there is no tRNA in the A site at the time of slippage (Baranov et al., 2004; Leger et al., 2004). To test the occurrence of single slippage, we used a mutant site in which the UUUAAAC heptamer was mutated to UUU AUAC. In this case, tandem slippage should be inefficient due to the presence of two mismatches after repairing of the A site tRNA in the −1 frame, but single slippage would not be affected. The frameshifting efficiency obtained with this construct was <0.1%, similar to the background level. This result demonstrates that in these experiments, frameshifting actually occurred through a tandem tRNA slippage mechanism. This implies that the three sites are involved in ribosomal frameshifting (see below).

The Ψ39 modification is conserved over the tree of life; its role on −1 frameshifting could thus be similar in a broad spectrum of organisms. This is consistent with the fact that the bias at the two positions upstream of the heptamer was deduced from a wide variety of viruses of different origins. However, each host cell, like the yeast strains used here, carries a specific tRNA pool that differs from one organism to another. This could explain the different dinucleotide usage observed between viruses; however, not enough sequence data are available to assess this point. In any case, the existence of a bias indicates an important role of tRNA modification on −1 frameshifting in eukaryotes. A role of tRNA

Table 4. Effect of *PUS3* Gene Deletion on −1 Frameshifting Efficiency

Plasmids	Wt	*pus3*Δ	*pus3*Δ + pRS315	*pus3*Δ + *PUS3*	*pus3*Δ + *pus3*[D151A]
pAC.5.CG	5.3% ± 1%	5.9% ± 1% (1.1)	5.8% ± 0% (1.1)	5.5% ± 1% (1.0)	6.0% ± 1% (1.1)
pAC.5.GA	7.6% ± 1%	8.0% ± 1% (1.1)	8.1% ± 1% (1.1)	7.2% ± 0% (0.9)	7.3% ± 1% (1.0)
pAC.5.UA	21.7% ± 1%	12.8% ± 1% (0.6)	11.5% ± 1% (0.5)	19.3% ± 1% (0.9)	11.8% ± 1% (0.5)
pAC.5.UC	19.5% ± 1%	10.3% ± 1% (0.5)	10.2% ± 1% (0.5)	18.7% ± 2% (1.0)	12.7% ± 1% (0.7)

Wild-type (wt) and *pus3*Δ mutants of 74-D694 strains were transformed with the test plasmids. The 74-D694 *pus3*Δ strain was also transformed with empty pRS315 or the same plasmid containing the *PUS3* gene or the mutant *pus3*[D151A] gene, as indicated. −1 frameshifting efficiencies were measured at 30°C, and the data are expressed as percentages. Numbers in parentheses correspond to ratios of recoding efficiency in the wt strain over the recoding efficiency in the *pus3*Δ derivative strain. No significant difference can be expected by a Mann-Whitney statistical test, except between pAC.5.UA/UC in wt or *pus3*Δ + *PUS3* compared to other transformed strains (p value < 0.005).

modifications on +1 frameshifting has been previously described both in *E. coli* and in *S. cerevisiae* (Bjork et al., 1989; Lecointe et al., 2002; Urbonavicius et al., 2001). For −1 frameshifting, a few examples have been reported in *E. coli* (Brierley et al., 1997; Licznar et al., 2003), but not in eukaryotes. In these cases, the tRNAs involved were acting at the A or P site.

Overall, these results demonstrate that the effect of the upstream context of the heptamer is directed by the modification status of the tRNA decoding the −1 codon.

Conclusions

Viral Frameshifting Signals

It is striking that all members of a genus (or family, in some cases) use a frameshifting event to produce their Pol protein but that phylogenetic analyses of frameshift sequences give rise to patterns inconsistent with accepted trees (data not shown). Inconsistency of frameshifting patterns with accepted phylogenetic trees is not surprising taking into account the recombinant nature of many viruses; functional requirements probably account for both this complete conservation and the variability of the frameshifting site sequences. Indeed, in the *Retroviridae* family, the *Alpharetrovirus* genus is exceptional because some members exhibit frameshift signals but others do not. In fact, this genus is subdivided in two categories: replication-competent viruses, which possess the *pol* gene, and defective viruses, which do not. Logically, frameshift signals are found only in the latter category. It is even more interesting that despite their position among the *Totiviridae*, the *Leishmaniavirus* genus members do not carry −1 frameshift sites but, rather, use +1 frameshifting to express their polymerase domain. This suggests that strong biological constraints are at play in the selection of a recoding event in the life cycle of these viruses, possibly related to the incorporation of the polymerase as a fusion protein in the viral particle.

Role of E Site in Frameshifting

An interesting feature of the results presented here is the involvement of an extended nonanucleotide signal in ribosomal frameshifting. As demonstrated above, no single slippage is observed in the experimental system used here; this nonanucleotide-directed frameshifting thus involves classical tandem slippage where both A and P site tRNAs slip by one nucleotide upstream. This implies that the three ribosomal sites are involved in −1 frameshifting. Two hypotheses can be proposed to account for the role of the E site tRNA in frameshifting. Firstly, frameshifting might be enhanced by the absence of a tRNA in the E site. In this case, the Ψ39 modification would destabilize the tRNA:E site interaction. Secondly, Ψ39 might interfere directly or indirectly with the interaction of the P site tRNA with the mRNA, decreasing pairing stability.

The first hypothesis is supported by recent results in which premature release of the E site tRNA from the ribosome has been shown to be coupled with high-level +1 frameshifting at the *prfB* gene, encoding the prokaryotic termination factor RF2 (Marquez et al., 2004). Likewise, in eukaryotes, Ψ39 may induce an unusual E site conformation. If this is the case, one would predict that the Ψ39 modification induces a higher frequency of release of the tRNA from the E site. If E-tRNA normally helps prevent tRNA slippage in the P site, this could explain the different susceptibilities of a given heptamer to slippage. Probably the E-tRNA is released during the accommodation step of the A-tRNA and not during the preceding decoding reaction (Nierhaus, 1990; Noller et al., 2002). However, in the case of a −1 frameshift event, E-tRNA release at the decoding step would facilitate the slippery event of A and P site tRNAs, and this precisely might be the effect of Ψ39. Biochemical experiments will be required to clarify this point.

The second hypothesis is supported by structural data on the prokaryotic ribosome (Ramakrishnan and Moore, 2001; Yusupov et al., 2001) and inferred cryo-EM reconstruction of the yeast ribosome (Spahn et al., 2001) that strongly suggest that the E-tRNA interacts with several partners. The closest distance between the anticodon stem backbones of the P- and E-tRNAs is about 6 Å, which is closer than the distance separating the A- and P-tRNAs. The two tRNAs are not in direct contact but are linked by the 16S rRNA helices H24, H28, and H29, and loops 690 and 790, both of which they directly interact with through their anticodon loops (Yusupov et al., 2001). Another link between E and P sites is through the mRNA. A single possible contact was noted between the mRNA and E-tRNA in the crystal structure, but the latter was noncognate. Even this noncognate E site anticodon was close enough to the codon, such that cognate interaction would be structurally plausible; moreover, there is biochemical evidence for codon-anticodon specificity in the E site (Lill and Wintermeyer, 1987; Rheinberger et al., 1986). E site tRNA is thus sufficiently connected to the P site to suggest that it very likely plays a role in promoting the stability of P site codon-anticodon pairing. Ψ39 modification can be expected to improperly fill the E site during the slippage-prone state, probably resulting in an unstable P site codon-anticon interaction and enhanced −1 frameshifting. This is reminiscent of the role played by a particular context of a bacterial tmRNA resume codon. In this case, an unusual E site conformation destabilizes the P site codon-anticodon interaction and induces frameshifting (Trimble et al., 2004).

The results presented here demonstrate that the slippery component of −1 frameshift signals, at least in yeast, is more complex than previously anticipated. Compared to the initial model of Jacks et al. (1988), sequence elements of both the 3′ and 5′ heptamer elements are now shown to participate in frameshiting efficiency through interactions between tRNA, mRNA, and the ribosome. Similarly, downstream secondary structures can directly or indirectly influence frameshifting. A combinatorial use of upstream codons, heptamer sequences, downstream codons, and stimulatory secondary structures permit a given frameshifting efficiency for a given virus in a given host. Whether or not these different sequence elements act independently remains to be established.

Experimental Procedures

Polymerase Tree

A ClustalW 1.83 (Thompson et al., 1994) alignment of viral polymerase amino acid sequences retrieved from GenBank was used. It was

employed to deduce a neighbor-joining tree with 1000 bootstrap replications (Saitou and Nei, 1987) by using Mega 2.1 package (Kumar et al., 2001), which provides a graphical representation repartition of selected viruses. Pairwise distances were calculated as mean observed substitutions per site. The unrooted tree is shown in Figure 1 and is color coded to mark each clade.

Profile Construction

Frameshift sites—the heptamers surrounded by ten nucleotides on both sides—from the selected viruses were used to construct and calibrate an HMM by using the HMMER package 2.3.2 (Eddy, 1998). Each sequence was aligned on the shifty heptamer and the HMM established. Sequences from viruses not selected as representative subset but reported as frameshifting viruses were used to validate our profile. All sites were found (data not shown).

Searches with Profile

With this HMM profile, searches against the publicly available viral genome database (GenBank, downloaded February 10th, 2004) were carried out. All searches against nucleotide databases were performed with the HMMER 2.3.2 package. As threshold, we assigned a minimal e value of 0.5. Subsequently, an enhanced HMM profile was established with the HMMER package 2.3.2. The frameshift sites found—the heptamers surrounded by ten nucleotides on either side—were used to construct and calibrate the enhanced HMM profile.

Bias

Dinucleotide biases were estimated by counting each dinucleotide of the 74 sequences and comparing the distribution with an equiprobability model. Because we compared different viruses from different hosts, we should use the lesser bias model where the frequency of each dinucleotide is 1/16. This estimation used a χ^2 probability with 15 degrees of freedom.

Yeast Strains and Media

The *S. cerevisiae* strains used were FY1679-18B (Mat α *his3*-Δ200, *trp1*-Δ63, *ura3*-52, and *leu2*-Δ1), 74-D694 (Mat a, *ade1*-14, *trp1*-289, *his3*Δ200, *leu2*-3, 112, and *ura3*-52), and its derivative *pus3*Δ (Mat a, *ade1*-14, *trp1*-289, *his3*Δ200, *leu2*-3, 112, *ura3*-52, and *pus3*Δ::*KAN*). Strains were grown in minimal media (0.67% yeast nitrogen base, 2% glucose) supplemented with the appropriate amino acids to allow maintenance of the different plasmids under standard growth conditions. Yeast transformations were performed by the lithium acetate method (Ito et al., 1983).

Plasmids and Molecular Biology Methods

pAC99 derivatives were constructed by cloning the synthetic oligonucleotides of interest at the unique *MscI* site of pAC99 (Bidou et al., 2000). For viral frameshift sites, heptamer, spacer, and secondary structure surrounded by ten nucleotides on each side were inserted. For mutants of the upstream region of the heptamer, the IBV frameshift site was used and the wt sequence (UA) was changed to the 15 other possible sequences (see Supplemental Table S2). Plasmids containing the *PUS3* gene or the mutant *pus3*[D151A] gene were from Lecointe et al. (2002). All constructs were verified by sequencing the region of interest.

Quantification of −1 Frameshifting Efficiency

Luciferase and β-galactosidase activities were assayed in the same crude extract as previously described (Stahl et al., 1995). The assays were carried out at least five times by using two independent transformants grown in the same conditions. The luciferase β-galactosidase ratio obtained with test constructs was normalized to the ratio obtained with the in-frame control and expresses frameshift efficiency.

Acknowledgments

We would like to thank Dominique Fourmy, Henri Grosjean, and Olivier Namy for helpful discussions and suggestions and François Lecointe for providing us with the *pus3*Δ stains and plasmids. We thank members of the Génétique Moléculaire de la Traduction laboratory and the "frameshift team" for numerous stimulating discussions. We are especially grateful to Anne-Lise Haenni for critically reading the manuscript. This work was supported by the Association pour la Recherche sur le Cancer (contract 4699).

Received: July 20, 2004
Revised: September 15, 2004
Accepted: October 26, 2004
Published: January 6, 2005

References

Baranov, P.V., Gurvich, O.L., Fayet, O., Prere, M.F., Miller, W.A., Gesteland, R.F., Atkins, J.F., and Giddings, M.C. (2001). RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res. *29*, 264–267.

Baranov, P.V., Gesteland, R.F., and Atkins, J.F. (2004). P-site tRNA is a crucial initiator of ribosomal frameshifting. RNA *10*, 221–230.

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J.P., Froidevaux, C., Hatin, I., Rousset, J.P., and Termier, M. (2003). Towards a computational model for −1 eukaryotic frameshifting sites. Bioinformatics *19*, 327–335.

Bertrand, C., Prere, M.F., Gesteland, R.F., Atkins, J.F., and Fayet, O. (2002). Influence of the stacking potential of the base 3′ of tandem shift codons on −1 ribosomal frameshifting used for gene expression. RNA *8*, 16–28.

Bidou, L., Stahl, G., Hatin, I., Namy, O., Rousset, J.P., and Farabaugh, P.J. (2000). Nonsense-mediated decay mutants do not affect programmed −1 frameshifting. RNA *6*, 952–961.

Bjork, G.R., Wikstrom, P.M., and Bystrom, A.S. (1989). Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. Science *244*, 986–989.

Brierley, I., Rolley, N.J., Jenner, A.J., and Inglis, S.C. (1991). Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. J. Mol. Biol. *220*, 889–902.

Brierley, I., Jenner, A.J., and Inglis, S.C. (1992). Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. J. Mol. Biol. *227*, 463–479.

Brierley, I., Meredith, M.R., Bloys, A.J., and Hagervall, T.G. (1997). Expression of a coronavirus ribosomal frameshift signal in Escherichia coli: influence of tRNA anticodon modification on frameshifting. J. Mol. Biol. *270*, 360–373.

Büchen-Osmond, C. (2003). The universal virus database ICTVdB. Comput. Sci. Eng. *5*, 16–25.

Chamorro, M., Parkin, N., and Varmus, H.E. (1992). An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. Proc. Natl. Acad. Sci. USA *89*, 713–717.

den Boon, J.A., Snijder, E.J., Chirnside, E.D., de Vries, A.A., Horzinek, M.C., and Spaan, W.J. (1991). Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily. J. Virol. *65*, 2910–2920.

Dinman, J.D., and Wickner, R.B. (1992). Ribosomal frameshifting efficiency and gag/gag-pol ratio are critical for yeast M1 double-stranded RNA virus propagation. J. Virol. *66*, 3669–3676.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Fayet, O., Ramond, P., Polard, P., Prere, M.F., and Chandler, M. (1990). Functional similarities between retroviruses and the IS3 family of bacterial insertion sequences? Mol. Microbiol. *4*, 1771–1777.

Herold, J., and Siddell, S.G. (1993). An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. Nucleic Acids Res. *21*, 5838–5842.

Huang, S., and Ghabrial, S.A. (1996). Organization and expression of the double-stranded RNA genome of Helminthosporium victoriae 190S virus, a totivirus infecting a plant pathogenic filamentous fungus. Proc. Natl. Acad. Sci. USA *93*, 12541–12546.

Ito, H., Fukuda, Y., Murata, K., and Kimura, A. (1983). Transformation

of intact yeast cells treated with alkali cations. J. Bacteriol. *153*, 163–168.

Jacks, T., Madhani, H.D., Masiarz, F.R., and Varmus, H.E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447–458.

Jonassen, C.M., Jonassen, T.O., and Grinde, B. (1998). A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. J. Gen. Virol. *79*, 715–718.

Kim, Y.G., Su, L., Maas, S., O'Neill, A., and Rich, A. (1999). Specific mutations in a viral RNA pseudoknot drastically change ribosomal frameshifting efficiency. Proc. Natl. Acad. Sci. USA *96*, 14234–14239.

Kontos, H., Napthine, S., and Brierley, I. (2001). Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. Mol. Cell Biol. *21*, 8657–8670.

Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. Bioinformatics *17*, 1244–1245.

Lecointe, F. (2002) Etude d'Enzymes de Modification de Nucléotides des ARNt et Leurs Fonctions dans le Métabolisme Cellulaire Chez *Saccharomyces cerevisiae* (Orsay, France: Université Paris XI).

Lecointe, F., Simos, G., Sauer, A., Hurt, E.C., Motorin, Y., and Grosjean, H. (1998). Characterization of yeast protein Deg1 as pseudouridine synthase (Pus3) catalyzing the formation of psi 38 and psi 39 in tRNA anticodon loop. J. Biol. Chem. *273*, 1316–1323.

Lecointe, F., Namy, O., Hatin, I., Simos, G., Rousset, J.P., and Grosjean, H. (2002). Lack of pseudouridine 38/39 in the anticodon arm of yeast cytoplasmic tRNA decreases in vivo recoding efficiency. J. Biol. Chem. *277*, 30445–30453.

Leger, M., Sidani, S., and Brakier-Gingras, L. (2004). A reassessment of the response of the bacterial ribosome to the frameshift stimulatory signal of the human immunodeficiency virus type 1. RNA *10*, 1225–1235.

Licznar, P., Mejlhede, N., Prere, M.F., Wills, N., Gesteland, R.F., Atkins, J.F., and Fayet, O. (2003). Programmed translational −1 frameshifting on hexanucleotide motifs and the wobble properties of tRNAs. EMBO J. *22*, 4770–4778.

Lill, R., and Wintermeyer, W. (1987). Destabilization of codon-anticodon interaction in the ribosomal exit site. J. Mol. Biol. *196*, 137–148.

Marczinke, B., Fisher, R., Vidakovic, M., Bloys, A.J., and Brierley, I. (1998). Secondary structure and mutational analysis of the ribosomal frameshift signal of rous sarcoma virus. J. Mol. Biol. *284*, 205–225.

Marquez, V., Wilson, D.N., Tate, W.P., Triana-Alonso, F., and Nierhaus, K.H. (2004). Maintaining the ribosomal reading frame: the influence of the E site during translational regulation of release factor 2. Cell *118*, 45–55.

Mejlhede, N., Atkins, J.F., and Neuhard, J. (1999). Ribosomal −1 frameshifting during decoding of *Bacillus subtilis* cdd occurs at the sequence CGA AAG. J. Bacteriol. *181*, 2930–2937.

Namy, O., Rousset, J.P., Napthine, S., and Brierley, I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol. Cell *13*, 157–168.

Nierhaus, K.H. (1990). The allosteric three-site model for the ribosomal elongation cycle: features and future. Biochemistry *29*, 4997–5008.

Noller, H.F., Yusupov, M.M., Yusupova, G.Z., Baucom, A., and Cate, J.H. (2002). Translocation of tRNA during protein synthesis. FEBS Lett. *514*, 11–16.

Ramakrishnan, V., and Moore, P.B. (2001). Atomic structures at last: the ribosome in 2000. Curr. Opin. Struct. Biol. *11*, 144–154.

Rheinberger, H.J., Sternbach, H., and Nierhaus, K.H. (1986). Codon-anticodon interaction at the ribosomal E site. J. Biol. Chem. *261*, 9140–9143.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Shehu-Xhilaga, M., Crowe, S.M., and Mak, J. (2001). Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity. J. Virol. *75*, 1834–1841.

Shigemoto, K., Brennan, J., Walls, E., Watson, C.J., Stott, D., Rigby, P.W., and Reith, A.D. (2001). Identification and characterisation of a developmentally regulated mammalian gene that utilises −1 programmed ribosomal frameshifting. Nucleic Acids Res. *29*, 4079–4088.

Spahn, C.M., Beckmann, R., Eswar, N., Penczek, P.A., Sali, A., Blobel, G., and Frank, J. (2001). Structure of the 80S ribosome from *Saccharomyces cerevisiae*-tRNA-ribosome and subunit-subunit interactions. Cell *107*, 373–386.

Stahl, G., Bidou, L., Rousset, J.P., and Cassan, M. (1995). Versatile vectors to study recoding: conservation of rules between yeast and mammalian cells. Nucleic Acids Res. *23*, 1557–1560.

ten Dam, E.B., Pleij, C.W., and Bosch, L. (1990). RNA pseudoknots: translational frameshifting and readthrough on viral RNAs. Virus Genes *4*, 121–136.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Trimble, M.J., Minnicus, A., and Williams, K.P. (2004). tRNA slippage at the tmRNA resume codon. RNA *10*, 805–812.

Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc. Natl. Acad. Sci. USA *87*, 2516–2520.

Urbonavicius, J., Qian, Q., Durand, J.M., Hagervall, T.G., and Bjork, G.R. (2001). Improvement of reading frame maintenance is a common function for several tRNA modifications. EMBO J. *20*, 4863–4873.

Walter, A.E., Turner, D.H., Kim, J., Lyttle, M.H., Muller, P., Mathews, D.H., and Zuker, M. (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc. Natl. Acad. Sci. USA *91*, 9218–9222.

Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., and Cate, J.H. (2001). Crystal structure of the ribosome at 5.5 A resolution. Science *292*, 883–896.

Article en préparation

Identification of programmed translational -1 frameshifting sites in the genome of *Saccharomyces cerevisiae*

Michaël Bekaert (1), Hugues Richard (2),
Bernard Prum (2) & Jean-Pierre Rousset (1)*

(1) Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud, 91405 Orsay cedex, France
(2) Laboratoire Statistique et Génome, CNRS-INRA-Université d'Evry, 91000 Evry, France

*Corresponding author
Phone: 33 (0) 1 69 15 50 51
Fax: 33 (0) 1 69 15 46 29
E-mail : jean-pierre.rousset@igmors.u-psud.fr

Abstract

Frameshifting is a recoding event that allows the expression of two polypeptides from the same mRNA molecule. Most recoding events described so far are used by viruses and transposons to express their replicase activity. The few cellular genes known to be expressed following this control mode have been found by chance and not in systematic surveys for -1 ribosomal frameshifting sites. The goal of the present work was to set up such a systematic strategy, based on complementary bioinformatics approaches, without a priori knowledge of the involved mechanism. Two independent methods were devised. The first looks for genomic regions in witch two ORFs, each carrying a protein pattern, are in a frameshifted arrangement. The second uses Hidden Markov Models and likelihood in a two step approach. When this strategy was applied on the *Saccharomyces cerevisiae* genome, 189 candidate regions were found, of which 55 were further functionally investigated. Twenty eight of them expressed a full length mRNA covering the two ORFs and 11 showed a -1 frameshift efficiency 50-fold higher than background. Several of these candidates correspond to genes with known functions. These results strongly suggest that -1 frameshifting might be more widely spread than previously thought.

Introduction

Sequencing programs, along with various projects in the pharmaceutical, agricultural, aquacultural, and forestry industries, are creating an explosion of DNA sequence data. With this abundance of data, there is a growing need for more effective tools and methods to extract vital information from raw DNA sequences. Algorithms for identifying protein coding regions and predicting complete genes are of particular importance. Since the early 1990s, a number of computer programs for eukaryotic gene identification have been developed: GENMARK (Borodovsky and McIninch, 1993), FGENEH (Solovyev and Salamov, 1997; Solovyev et al., 1994), GeneParser (Snyder and Stormo, 1995), GeneWise (Birney et al., 1996), GenScan (Burge

and Karlin, 1997), and Procruts (Gelfand et al., 1996; Mironov et al., 1998). Most of these programs make use of sophisticated pattern recognition techniques, such as linear discriminant analyses, neural networks, or Hidden Markov models to identify coding regions. Some programs also make use of database sequences alignment methods, such as BLAST (Altschul et al., 1990), to further improve their predictions. Generally, these algorithms classify out of frame ORFs as either a sequencing error or a pseudogene signature (Harrison et al., 2002). Up to now only a few algorithms assign a frameshift as a possible regulatory process. However, frameshifting is involved in the expression of numerous genes, most of them being found in viruses or transposable elements (Baranov et al., 2002). Together with readthrough of stop codons and ribosome hopping, frameshifting is part of the reprogrammed genetic decoding ("recoding") events that allow expression of several polypeptides from the same mRNA (Gesteland et al., 1992). Although most of the recoding events described so far have been found in small autonomous genetic elements, a few cellular genes are known to be expressed by this mode of control (Namy et al., 2004). Up to now, these cellular recoded genes have been found serendipitously. Recently, a few *in silico* analyses have been described that allow performing a systematic search of recoding sites at the genomic scale. However, only a few authentic genes have yet been identified (Hammell et al., 1999; Harrison et al., 2002; Liphardt, 1999; Namy et al., 2003; Namy et al., 2002; Sato et al., 2003). The goal of the present work was to set up a comprehensive strategy, based on complementary bioinformatics approaches and functional *in vivo* analyses, to identify -1 ribosomal frameshifting sites in cellular genomes.

Jacks and Varmus described 20 years ago the first programmed -1 ribosomal frameshifting, from which they established the canonical model of eukaryotic -1 frameshifting site (Jacks et al., 1988; Jacks and Varmus, 1985). Today, several tens of viruses and one mouse nuclear gene (Shigemoto et al., 2001) have been identified as bearing such a -1 frameshifting site. A typical site contains a slippery heptamer in 5', where both A- and P-site tRNAs slip by one nucleotide upstream, followed by a stimulatory structure (stem loop, or pseudoknot) downstream (Brierley et al., 1989). The slippery heptamer is separated from the stimulatory structure by a short sequence, the so-called spacer. Based on this model, studies have been undertaken to identify frameshifting sites in the nuclear genome of the yeast *Saccharomyces cerevisiae* (Hammell et al., 1999; Liphardt, 1999). However, none of these allowed identifying with certainty authentic expressed genes controlled by -1 frameshifting. Two reasons might be proposed: first, the model might not

be precise enough, leading to the identification of too many false positive candidates (Bekaert et al., 2003); conversely the model might be too rigid, failing to identify true positive candidates. This would be the case, for example, if -1 frameshift could be directed by a more "degenerated" structure, or by mechanisms that rely on other types of signals.

In this report, we present a strategy to systematically extract genes in the yeast *S. cerevisiae*, whose expression are controlled by a -1 frameshifting, without a priori knowledge of the mechanism involved. We devised two independent methods to look for frameshifting sites *in silico*. The first is based on the search for genomic regions where two domains, each carrying a protein pattern, can be associated on a same polypeptide by a single -1 frameshifting. The second is performed by a two step selection with Hidden Markov Models, which, after discriminating on the potential candidates likely to possess a coding constrained region after their stop, rank them by likelihood ratio based on available biological knowledge. These two approaches do not rely on any model of frameshifting site and thus are well adapted for *de novo* detection of frameshift events.

We validated this strategy by analysing the genome of *S. cerevisiae*. A total of 189 candidate regions was found. We assessed the presence of a full-length mRNA and quantified -1 frameshift efficiency for a subset of the highest ranked candidates. Among the 55 characterized regions, 28 were analysed for their ability to induce -1 frameshifting in vivo; 11 showed a frameshift efficiency 50-fold higher than the background. Several of these candidates correspond to genes with known functions, which will allow further analysis of the physiological role of the frameshifting event. Overall, these results strongly suggest that -1 frameshift might be a more widely used way of controlling gene expression than previously thought.

Results

General strategy

Figure 1 shows the pipeline of our -1 frameshifting candidate identification strategy. The system first download and parse the nucleic acid sequences, the intron/exon data and position of specified chromosome from the GenBank database and stock them in local database for more

reliability. Our system seeks genomic configurations compatible with a -1 ribosomal frameshifting by using the following criteria: two open reading frames, one in 0 frame (ORF0), the other in -1 frame (ORF-1), that overlap along an intermediate shared region (figure 2). We fixed a length of at least 99 nucleotides for both ORF0 and ORF-1 areas, and of at least 149 base pairs for the whole structure (Step 1).

The second step was to filter undesirable low complexity sequences that may overload the next levels. From this step, remaining sequences were classified according to whether the 0 or/and -1 frames are already annotated as an ORF, in order to performed the subsequent HMM step. We define four class, "*left*" (ORF0 is annotated), "*right*" (ORF-1 is annotated), "*both*" (both ORFs are annotated) and "*none*" for all the others (Step 2).

The next step retained regions that have protein patterns in both ORF0 and ORF-1. For this purpose, we used InterPro database and InterProScan. (Step 3).

In parallel, Hidden Markov Models filtering and estimation was performed to predict coding regions that may continue in the -1 frame after the stop. This was followed by a ranking step where we compared the likelihood ratio of each selected candidate structure on the two following assumptions: "the sequence possesses a frameshift", and, "the sequence does not possess any frameshift", taking into account the class of the candidate defined in step 3. (Step 3' and Step 3'').

We then tested the candidate regions for expression *in vivo*, by looking for the presence of a full length polyadenylated mRNA, using oligo(dT) primed RT-PCR (Step 4). Finally, for the remaining candidates, -1 frameshifting efficiencies were determined *in vivo*, using a dual reporter system (Step 5).

Creating a dataset of potential -1 frameshift regions

The goal of this step was to identify structures exhibiting a genomic organization compatible with a translational -1 frameshift mode of expression. We chose to search first for overlapping ORFs. All searches were performed independently on four sets of data: the *S. cerevisiae* genome (12 Mbp), the genome of the yeast L-A virus (4,579 bp) which is known to bear an authentic -1 ribosomal frameshifting site, and artificial genomes which exhibits the same hexamer frequencies as the yeast and L-A genome respectively. The artificial genomes were generated using Markov chains (see Methods). All possible frameshifted structures are then automatically extracted.

Among all potential -1 frameshifts, some are DNA microsatellites (Hamada et al., 1984): tandem repetitions of the same triplet, which are read as repetitions of two different amino acids, depending on the reading frame. Such sequences were clean up by using mdust, which remove low-complexity sequence. From this analysis 22,445 regions were found in the yeast genome, 24,248 in the artificial genome, 10 in the yeast L-A virus genome and 8 regions in the artificial L-A genome.

Assessing functional frameshifting by InterproScan

The hit sequences were then subjected to protein motifs search. Each candidate sequence was kept only if it exhibits, in both frames, a pattern in InterPro database. InterPro database includes BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, ScanRegExp and SuperFamily. The default parameter settings were used for the search.
This approach was validated as far as only the actual frameshifting region was retrieved from the L-A virus genome. Moreover, we found 84 candidates in the yeast genome and 11 in the yeast artificial genome. Among these 84 regions, three categories can be defined: 69 exhibit domains that contain stretches of repeated amino-acids in each of the two frames. These are not low-complexity sequences that were already discarded at step 2. Noteworthy, no such candidates were found in the random genome. The second category is composed of regions where the two ORFs bear similar protein patterns, or two distinct but functionally compatible motifs (e.g. a sugar transporter and a sugar binding site). We found 6 such regions. The third category includes 8 regions which bear functional regions in one ORF and repetitions of amino acids in the other ORF. All the 11 candidate sequences from the random genome belong to this category.

Obtaining structure candidates by HMM

One of the more efficient methods to segment sequences in coding and non coding regions (allowing different phases and genes on both strands) is the Hidden Markov Model (HMM). It was introduced by Rabiner for speech recognition {ref}. This method is now commonly used in bioinformatics, from gene detection to prediction of protein domains {ref}.
For each step to be performed in a HMM framework, one has to completely specify a model, *i.e.* a probability law on the hidden states structure and a law for the emission of observed letters within each state.

One has to note that the aim pursued here is not to simply detect genes, but rather to select candidates for which the extension after the stop is similar to coding regions. As far as we know, existing software designed for gene detection does not offer such flexibility. Moreover, they are trained and tested on a sufficiently large set of labeled examples, which doesn't yet exist in the case of ribosomal frameshift. In the following, we detail the construction of the HMM and the strategy used for detection and ranking.

First, one needs to describe a model fitting on genes structure constraints. The simplest structure can be seen as the one summarized in figure 3, and corresponds to the one used by common gene detectors {ref}. A gene begins with a start, continues by stretches of three bases corresponding to the codons and ends on a stop. In addition, to take into account codons and thus amino acids heterogeneities within yeast genes, we allowed the model to alternate between up to three different laws for codons. All parameters of this model were first estimated on a similarity reduced set of 3,158 ORFs (see Methods for details).

Then, to adapt our model for the detection of frameshifted genes, we allowed coding regions to appear in -1 frame after the stop. For this purpose, we inserted a transition from the state corresponding to the last base of the stop to the second coding frame of each coding type. We only kept those sequences for witch the sum θ of the corresponding transition probabilities was more than 0.95 (the histogram in figure 4 serves us as a justification of this threshold).

As a positive control test, we tested this step of our approach on the L-A virus. This virus is selected with a probability θ of 1.0 (this is only due to approximation errors).

Using this criterion, a final set of 110 candidates was retrieved. Since experimental validation could not be performed on such a number of sequences, it was necessary to rank them in order to functionally test only the more likely candidates.

In order to incorporate for each selected candidate the known coding status of the two possible coding frames, we separately treated the sequences in the four classes defined above: *left, right, both* and *none.*

Then, in each class, we ranked the sequences accordingly to the likelihood ratio:

```
L_x = P(X | theta_fs, S) / P(X | theta_nofs, S)
```

Where `theta_fs` and `theta_nofs` stand respectively for the parameters of the model under the two following assumptions: “a frameshift appears” and, “no frameshift appears” conditionally on the status of the ORF. More details about the models used conditionally on the subset can be

found in the Methods section and in supplementary material.
Ranked candidates with their score are summarized in table 1. From these scores, we selected 20 candidates to be tested (5 from *none* class, 7 from *both* class, 4 from *left* class and 4 from *right* class).

Common candidates

Finally, we crossed the results obtained using the protein motifs search and the HMM search. 5 common candidates were identified by comparing the 84 regions obtained in the first approach with the 110 regions obtained in the second one. As the two methods are completely independent, these common candidates together with 28 candidates from the protein motifs approach and the 20 best ranked candidates from the HMM approach were selected for further biological investigation. We also selected the 10 worse candidates to serve as a control of the relevance of the ranking procedure.

Genomic sequence of the candidates

Since an authentic frameshifting is indistinguishable from a sequencing error, we first verified the sequence of the genomic region spanning the overlap between ORF0 and ORF-1. Among the 55 candidate sequences analyzed, 5 did not show the presence of the expected frameshift. Since the strain used here (FY1679-18B) is different from the strain (S288C) that has been used for the *S. cerevisiae* sequencing project, either a sequencing/annotation error or a gene polymorphism could explain this discrepancy.

Expression of candidate sequences

The next step was to test whether the candidate sequences correspond to expressed ORFs. Since most of these regions were previously considered as intergenic region, they have not been included in systematic expression analyses. However, for those which are constituted of at least one previously annotated ORF (*right*, *left* and *both* classes), at least partial information was available and is indicated in supplementary material table 1. However, even in the cases where the 2 ORFs were previously identified (*both* class), the presence of mRNA corresponding to each ORF was tested independently. To check whether an mRNA spanning the 2 ORFs is actually expressed, we examined the 50 remaining candidate sequences by RT-PCR. This was carried out

using first a reverse transcriptase step with an oligo(dT) primer that allows amplifying primarily polyadenylated messenger RNA. The second PCR step was performed with an upper primer located in 5' of first ORF (0 phase) and a lower primer located near beyond the stop codon in second ORF (-1 phase) to ensure that a full length message is actually present in the cell. For a few exceptionally long regions, random primer was used at the reverse transcriptase step and two couples of internal primers were used for secondary PCR instead (see Supplementary Material Table S1 for the list of oligonucleotides used). No signal was observed in the absence of reverse transcriptase and a unique specific amplification was obtained for 28 candidate sequences (figure 5 and table 1). These results demonstrate that the same molecule of mRNA cover both ORFs and that these mRNAs are polyadenylated.

The region of overlap of the cDNAs corresponding to all the bicistronic mRNAs were analyzed by gel electrophoresis and subsequently sequenced (data not shown). For three candidate regions, the presence of an unexpected intron was demonstrated (table 1). For the remaining candidates there was no evidence of length or sequence polymorphism, suggesting that no splicing or editing event is involved in the production of a frameshifted product (see Discussion).

Quantification of -1 frameshift efficiency

It cannot be predicted whether ribosomes can actually shift from ORF0 to ORF-1 for any of these 28 candidate expressed sequences, since none of them carry a canonical -1 frameshift signal with a heptamer followed by a secondary structure. To quantify -1 frameshift accuracy, each fragment (about 50 nt either side of the overlapping areas) was amplified by PCR (see Supplementary Material Table S1 for the list of oligonucleotides used) from genomic DNA of a wild-type yeast strain (FY1679-18B) and cloned into the pAC99 dual reporter vector (Bidou et al., 2000). In this reporter system, each translating ribosome gives rise to β-galactosidase activity while only those that frameshift in the overlapping region spanning ORF0 and ORF-1 would give rise to luciferase activity. Frameshifting efficiency is estimated by dividing the luciferase/β-galactosidase ratio obtained from the test construct by the corresponding ratio obtained from an in frame control construct (see Methods). Eleven fragments displayed a -1 frameshift efficiency ≥50-fold of the background (0.1%).

Discussion

Although most translational recoding events are found in viruses and transposons, a few cellular genes have been identified that use this mode of expression. These genes are involved in a variety of biological processes and are sometimes subject to an autoregulatory process. Recoding is also widely distributed between organisms; it is thus likely that numerous novel recoded cellular genes are still to be discovered. However, the prediction of recoding sites from genomic databases is currently a difficult task. Since most recoding events generate a premature in-frame stop codon, this is generally categorized as an error by computer programs, leading to improper gene annotation. A bioinformatics strategy has been developed to identify recoded genes, based on the knowledge of the recoding mechanism (model-based approach). In this case, genomic sequences are searched for regions exhibiting an already known recoding signal. Such analyses have already allowed the identification of several candidate recoded genes (Baranov et al., 2002; Hammell et al., 1999; Namy et al., 2003). This approach suffer major drawbacks: an imprecise model leading to a high number of false positive candidates and too rigid a model failing to identify truly positive candidates. For this reason, we and others have undertaken to develop bioinformatics approaches that do not depend on models of recoding sites and can be performed without a priori knowledge of the mechanism involved (Harrison et al., 2002; Sato et al., 2003). These approaches seek genomic configurations compatible with recoding, such as two ORFs overlapping or separated by a unique stop codon. The high number of candidates is then filtered by secondary constraints (length, presence of protein motifs, etc.). Several candidate recoded genes have already been identified in yeast (Harrison et al., 2002; Namy et al., 2003) and in Drosophila (Sato et al., 2003) by this way. However, except in a one study (Namy et al., 2003), no biological validation has been performed to assess whether the candidate regions actually induce recoding in vivo.

Here, we described a comprehensive analysis of the *S. cerevisiae* genome which attempted to identify cellular recoding events occurring during translational -1 frameshifting. We developed a genomic approach, seeking genes with an extended coding potential, without prior constraint from existing ideas on -1 frameshift mechanism.

In a first step, 22,445 genomic structures were extracted from the genome of *S. cerevisiae*. This value relies on two strong assumptions. First, we chosen to collect only extensions of polypeptide

but no premature ending, although biologically pertinent frameshifting events, like in the *E. coli DnaX*, could lead to the synthesis of a shortened product (Tsuchihashi and Kornberg, 1990). Second, we specified the minimal size of each ORF to 99 nucleotides (33 amino acids). Preliminary analysis (data not shown) had shown that decreasing this size by only 3 nucleotides (96) increased two fold the numbers of retrieved structures. Thus, this limit was chosen to keep the number of candidates compatible with the biological validation step.

[…]

Our approach identified 189 candidates in the yeast genome. Eleven sequences displayed a -1 frameshifting efficiency 50-fold higher than background. For each of these candidates, a unique mRNA covering both ORFs is present in the cell. Although it is only for those candidates showing the highest frameshifting levels that one could expect to detect an mRNA editing mechanism, we did sequence the RT-PCR products for each of them. No RNA post-transcriptional modification was identified (table 2). Moreover, from the amplification of the mRNA using a poly(dT) primer at the reverse transcription step, we concluded that these mRNA are polyadenylated and not rapidly degraded. Such candidates to -1 frameshift event could exhibit a degenerate frameshift site different from Jacks model. In fact Three candidates exhibit a shifty heptamer in the appropriate frame but no detectable secondary structure (see Table 2). Other might be a -1 event carrying a more degenerate site or even correspond to a mechanism completely different but ending to an apparent -1 frameshift event, such as ribosome hopping or alternative splicing. Some of these candidates might also turn out to be irrelevant with respect to frameshifting. In particular, some may correspond to pseudogenes or long 5' or 3' UTR. However, we think this explanation is unlikely since reading frame maintenance is a very robust mechanism. Several parameters converge to keep processivity errors like frameshifting at a low level in normal conditions. Indeed, during the last years we have tested several dozens of constructs for basal frameshifting efficiency and found systematically a background value between 10^{-4} and 10^{-3}. Finally, the finding of such putative frameshifting sites in the highly ranked candidates and not in the lowest ranked candidates is a strong argument in favor of their biological significance.

In conclusion, the combination of two simple approaches made it possible to identify several

candidate genes potentially controlled by a -1 frameshift mechanism. Our approach is promising and could be straightforwardly extended to other organisms, eukaryotic as well as prokaryotic (Bertrand et al., 2002) and to other recoding events. Finally, we hope that the identification of new cellular recoded genes will also tell us whether they share similar properties or play common physiological roles in the cell.

Methods

Data sources

The system used entire chromosome sequences from the GenBank/RefSeq database (Maglott et al., 2000) as inputs. *S. cerevisiae* chromosomes NC_001133 to NC_001148 (downloaded on March 5, 2003) and *S. cerevisiae* virus L-A, NC_003754 (downloaded on December 25, 2003).

Random sequences

To define random background to be compared with real genome analyses, searches were performed independently on artificial genomes which exhibit the same hexamer frequencies as the *S. cerevisiae* genome or the L-A virus genome. We used the GenRGenS software v1.0 (Denise et al., 2003) for random generation of genomic sequences, using Markov chains of order 5.

Implementation

The main system is implemented in Perl, Bioperl 1.1 (Stajich et al., 2002) and PostgreSQL.
To detect protein signatures in the sequences, the motif database InterPro release 7.0 (Mulder et al., 2003) was used along with the software InterProScan version 3.1 (Zdobnov and Apweiler, 2001).
In terms of family coverage, the protein signature databases are similar in size but differ in content. While all of the methods share a common interest in protein sequence classification, some focus on divergent domains (e.g., Pfam), some focus on functional sites (e.g., PROSITE), and others focus on families, specializing in hierarchical definitions from superfamily down to

subfamily levels in order to pin-point specific functions (e.g., PRINTS). TIGRFAMs focus on building HMMs for functionally equivalent proteins and PIR SuperFamilies always produce HMMs over the full length of a protein and have protein length restrictions to gather family members. SUPERFAMILY is based on structure using the SCOP superfamilies as a basis for building HMMs. ProDom uses PSI-BLAST to find homologous domains that are clustered in the same ProDom entry. The clustered resources are derived automatically from the UniProt databases.

Low complexity filtering

We used the mdust algorithm (available from TIGR) to mask nucleic acid low-complexity regions, in particular from microsatellite areas, which enhance background noise and false positive.

Hidden Markov Models specification and estimation.

Each estimation and computation on Hidden Markov Models was done using the software SHOW {ref}. For the estimation of the coding parameters (defined as coding state), we used the ORF list of 5,861 sequences available on the SGD website (http://yeastgenome.org). As *S. cerevisiae* is known to possess a large proportion of paralogous genes, we then wiped out proteins presenting more than 70% of full length similarity. All of these alignments were done using the FASTA program using a BLOSUM62 matrix. Proteins were then clustered using a *p*-value threshold of 10^{-3} leading to a set of 3,526 sequences. The estimation of the intergenic state (composed of one state of order 2) was done on the complementary of all of the annotated ORFs. For the filter step (3'), the added links starting from the stop add three degrees of freedom to the model (the probabilities of shifting to the three possible coding states). In addition, we added three other parameters which correspond to the law of the three lengths between the STOP0 and the STOP-1. Moreover, there is an important proportion of the 22,445 considered sequences where the composition of the intergenic segment may have an influence on the estimation of the parameters related to the length. More precisely some intergenic composition could be better fitted by a mixture of two or three coding regions than by the intergenic law. We thus estimated these 6 new parameters on the *left*, *right* and *both* subset. Probabilities of transition from the stop to the shifted coding regions were then deduced with a classical forward-backward algorithm on

the 22,445 candidate structures to achieve step 3'.

For the ranking step, we calculated the likelihood of filtered sequences under the two assumptions: "the sequence contains a frameshift" and "the sequence doesn't contain any frameshift".

Whereas the first assumption corresponds to the same model for all of the candidates, we designed different models for each of the classes *left*, *right, none*, *both* for the second one. These correspond to the following facts:

- none: "all the sequence is intergenic";
- left: "coding is followed by intergenic after $STOP_1$";
- right: "coding ending on $STOP_3$ is preceded by intergenic";
- both: "coding ends on $STOP_1$, followed by intergenic and coding ending on $STOP_3$".

Detailed models can be found in supplementary materials. The sequences were then ranked within each class on the log odd-ratio of the two concerned assumptions, rescaled by their length.

Yeast strains and media

The *Saccharomyce cerevisiae* strain used for this work was FY1679-18B (Mat α *his3*-Δ200, *trp1*-Δ63, *ura3*-52, *leu2*-Δ1). The strain was grown in minimal media (0.67% Yeast Nitrogen Base, 2% glucose) supplemented with the appropriate amino acids to allow maintenance of the different plasmids under standard growth conditions. Yeast transformations were performed by the lithium acetate method (Ito et al., 1983).

Plasmids

The pAC99 reporter plasmid has been previously described (Bidou et al., 2000). Constructs were obtained by inserting a PCR fragment containing the overlapping region into the *MscI* cloning site, between the *lacZ* and *luc* genes in the plasmid pAC99. For -1 frameshift measurements, an in-frame control was used which allowed the production of 100% fusion protein (β-galactosidase-luciferase). The region including the inserted fragment was sequenced in the newly constructed plasmids. Each construct was then sequenced to check that no error occurred during PCR amplification.

Enzymatic activities and -1 frameshift efficiency

The yeast strains were transformed with the reporter plasmids using the lithium acetate method (Ito et al., 1983). In each case, at least five independent assays were realized in the same conditions. Cells were broken using acid-washed glass beads; luciferase and β-galactosidase activities were assayed in the same crude extract, as previously described (Stahl et al., 1995). Efficiency of -1 frameshift is defined as the ratio of luciferase activity to β-galactosidase activity. To establish the relative activities of β-galactosidase and luciferase when expressed in equimolar amounts, the ratio of luciferase activity to β-galactosidase from an in-frame control plasmid was taken as a reference. Efficiency of -1 frameshift, expressed as percentage, was calculated by dividing the luciferase/β-galactosidase ratio obtained from each test construct by the same ratio obtained with the in-frame control construct (Bidou et al., 2000).

Molecular biology procedures and RT-PCR

Each overlapping fragment corresponding to the candidate sequences was amplified from FY1679-18B genomic DNA by PCR, using *Pfu* polymerase (Promega), and cloned into the pAC99 vector.

Total RNA was extracted from 5 ml of exponential yeast culture (Schmitt et al., 1990). Each RNA sample was subjected to digestion with 10 U of RNAse-free DNAse I (Boerhinger) at 37°C for 1 h. DNAse I was inactivated by heating at 90°C for 5 min, as recommended by the manufacturer. RNA was reverse-transcribed with oligo(dT) or random primer by Superscript II Kit (Invitrogen) for PCR amplification with Taq polymerase (Amersham) in a Primus thermocycler (MWG-Biotch). PCR fragments were visualized in a 1.5% agarose gel. The sequences of the primers used either in amplification or RT-PCR experiments are shown in Table S1, which is published as Supplementary Material.

Acknowlegments

We are very grateful to Christine Froidevaux, Michel Termier and members of the GMT laboratory for stimulating discussions.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002). Recoding: translational bifurcations in gene expression. Gene *286*, 187-201.

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002). Release factor 2 frameshifting sites in different bacteria. EMBO Rep *3*, 373-377.

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics *19*, 327-335.

Bertrand, C., Prere, M. F., Gesteland, R. F., Atkins, J. F., and Fayet, O. (2002). Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression. Rna *8*, 16-28.

Bidou, L., Stahl, G., Hatin, I., Namy, O., Rousset, J. P., and Farabaugh, P. J. (2000). Nonsense-mediated decay mutants do not affect programmed -1 frameshifting. Rna *6*, 952-961.

Birney, E., Thompson, J. D., and Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res *24*, 2730-2739.

Borodovsky, M., and McIninch, J. (1993). Recognition of genes in DNA sequence with ambiguities. Biosystems *30*, 161-171.

Brierley, I., Digard, P., and Inglis, S. C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell *57*, 537-547.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Denise, A., Ponty, Y., and Termier, M. (2003). Random generation of structured genomic sequences. Paper presented at: Recomb'03 (Berlin).

Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A *93*, 9061-9066.

Gesteland, R. F., Weiss, R. B., and Atkins, J. F. (1992). Recoding: reprogrammed genetic decoding. Science *257*, 1640-1641.

Hamada, H., Petrino, M. G., Kakunaga, T., Seidman, M., and Stollar, B. D. (1984). Characterization of genomic poly(dT-dG).poly(dC-dA) sequences: structure, organization, and conformation. Mol Cell Biol *4*, 2610-2621.

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999). Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res *9*, 417-427.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., and Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J Mol Biol *316*, 409-419.

Ito, H., Fukuda, Y., Murata, K., and Kimura, A. (1983). Transformation of intact yeast cells treated with alkali cations. J Bacteriol *153*, 163-168.

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447-458.

Jacks, T., and Varmus, H. E. (1985). Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science *230*, 1237-1242.

Liphardt, J. (1999) The mechanism of -1 ribosomal frameshifting: experimental and theoretical analysis, Churchill College, Cambridge.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. (2000). NCBI's LocusLink and RefSeq. Nucleic Acids Res *28*, 126-128.

Mironov, A. A., Roytberg, M. A., Pevzner, P. A., and Gelfand, M. S. (1998). Performance-guarantee gene predictions via spliced alignment. Genomics *51*, 332-339.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P.*, et al.* (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res *31*, 315-318.

Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M., and Rousset, J. P. (2003). Identification of stop codon readthrough genes in Saccharomyces cerevisiae. Nucleic Acids Res *31*, 2289-2296.

Namy, O., Duchateau-Nguyen, G., and Rousset, J. P. (2002). Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. Mol Microbiol *43*, 641-652.

Namy, O., Rousset, J. P., Napthine, S., and Brierley, I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol Cell *13*, 157-168.

Sato, M., Umeki, H., Saito, R., Kanai, A., and Tomita, M. (2003). Computational analysis of stop codon readthrough in D.melanogaster. Bioinformatics *19*, 1371-1380.

Schmitt, M. E., Brown, T. A., and Trumpower, B. L. (1990). A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. Nucleic Acids Res *18*, 3091-3092.

Shigemoto, K., Brennan, J., Walls, E., Watson, C. J., Stott, D., Rigby, P. W., and Reith, A. D. (2001). Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. Nucleic Acids Res *29*, 4079-4088.

Snyder, E. E., and Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. J Mol Biol *248*, 1-18.

Solovyev, V., and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol *5*, 294-302.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res *22*, 5156-5163.

Stahl, G., Bidou, L., Rousset, J. P., and Cassan, M. (1995). Versatile vectors to study recoding: conservation of rules between yeast and mammalian cells. Nucleic Acids Res *23*, 1557-1560.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., *et al.* (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome Res *12*, 1611-1618.

Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc Natl Acad Sci U S A *87*, 2516-2520.

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

Figure 1

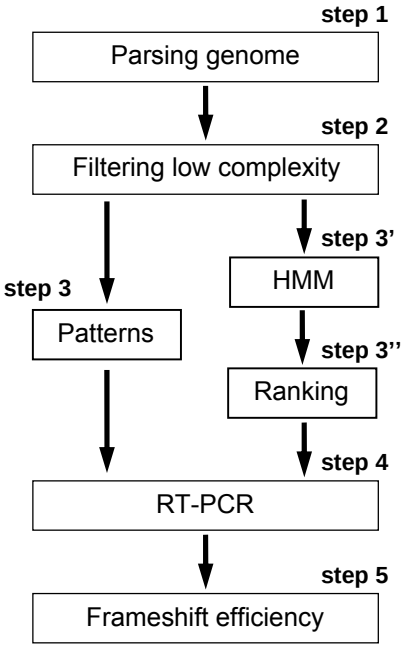

step 1
Parsing genome
step 2
Filtering low complexity
step 3'
HMM
step 3
Patterns
step 3''
Ranking
step 4
RT-PCR
step 5
Frameshift efficiency

Figure 2

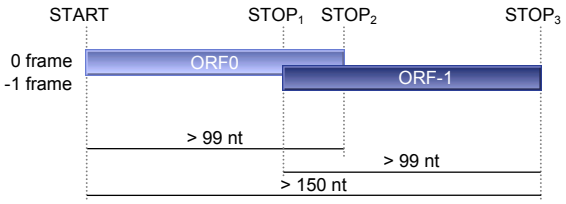

START
STOP1
STOP2
STOP3
0 frame
-1 frame
ORF0
ORF-1
> 99 nt
> 99 nt
> 150 nt

Figure 3

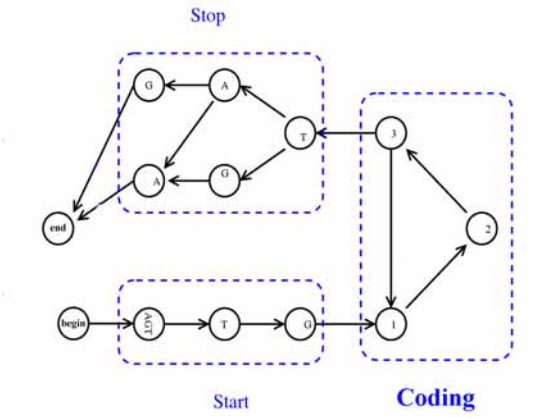


Figure 4

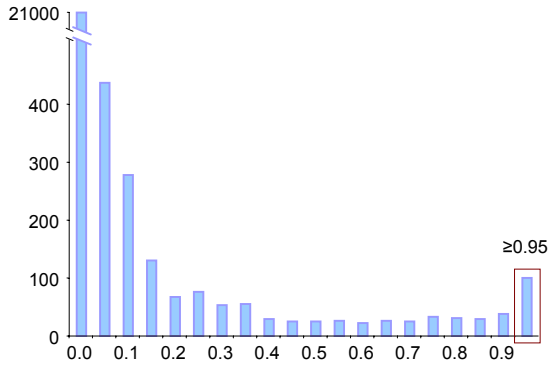
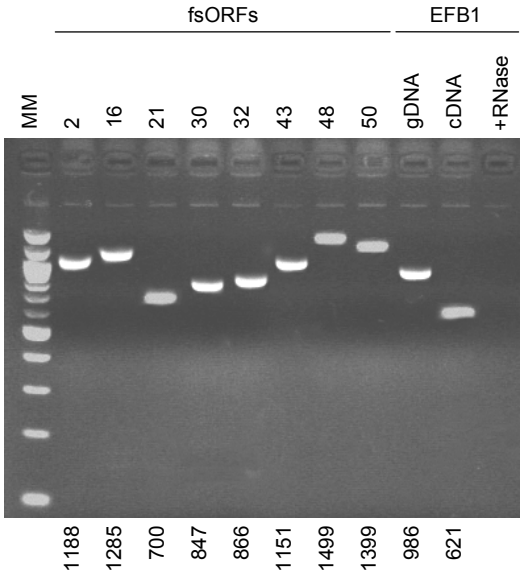


Figure 5


fsORFs
EFB1
MM
2
16
21
30
32
43
48
50
gDNA
cDNA
+RNase
1188
1285
700
847
866
1151
1499
1399
986
621

Figures legends

Figure 1. Pipeline of frameshifting candidate identification strategy

Figure 2. Schematic representation of the genomic configurations compatible with a -1 ribosomal frameshifting.

Figure 3. Illustration of a simple gene model. Begin and end stats are virtual stat figuring the begin and the end of the sequence. Furthermore they do not emit observation.

Figure 4. Distribution of the probability of transition from the 0 to the -1 frame for the candidate regions compatible with a -1 frameshifting event. As evidenced in this distribution, a clear pick was observed at the 0.95 limit. This threshold value was thus chosen as a cutoff for choosing the candidates to be ranked.

Figure 5. RT-PCRs. Total RNA was extracted as described in Methods, and treated by DNAse I. RT-PCR was carried out in two steps. First, a reverse transcription was done using a oligo(dT) primer, allowing only the reverse transcription of poly(A) mRNA. Then a standard PCR was performed on the mRNA after the reverse transcription. PCR products were visualized in a 2% agarose gel stained with ethidium bromide. A single amplification product was seen in all lanes, the expected size is indicated (in nucleotides) for each product at the bottom of the gel. Control use was *EFB1* mRNA witch includes an intron. Specific PCR on genomic DNA and cDNA exhibits two different products. Reverse transcription after RNase shows no DNA contamination during the process.

Table 1

Pattern results

fsORF	Chr.	Location	gDNA	mRNA	cDNA	FS	Notes	Class
1*	I	192541-196178	+1 nt	-	-	-	Re-annoted by SGD	both
2	II	289386-290383	yes	yes	yes	6.0		left
4	II	701799-700347	yes	no	-	-		left
5*	III	200170-197617	yes	yes	yes	0.1		both
9*	IV	167806-164992	yes	yes	yes	3.0		both
14*	V	298948-301706	yes	no	-	-		left
16	VI	15473-14309	yes	yes	yes	9.0		both
17*	VII	1068995-1067213	yes	no	-	-		left
19*	VII	270340-267730	yes	yes	yes	?		both
20	VII	425616-425971	yes	yes	yes	?		none
21	VII	677871-678301	yes	yes	yes	13.0		none
29	XI	172169-171299	yes	yes	yes	0.1		left
31*	XI	549085-551003	+ 1 nt	-	-	-	Re-annoted by SGD	left
34*	XII	200413-200654	yes	no	-	-		none
35	XII	203255-204786	yes	yes	yes	?		both
37*	XII	857539-861524	yes	no	-	-		both
39	XIII	349605-348426	yes	yes	yes	?		left
40*	XIII	436627-438788	yes	yes	yes	5.0		both
41*	XIII	509318-507416	yes	yes	yes	5.0		left
42	XIII	623212-622159	yes	yes	yes	0.1		left
43	XIII	650035-651026	yes	yes	yes	10.0		left
45	XIV	40618-42065	yes	no	-	-		left
48	XV	742910-744210	yes	yes	yes	5.0		left
49	XV	758330-759354	yes	no	-	-		left
50	XV	1026837-1028101	yes	yes	yes	7.0		left
53	XVI	117365-117062	yes	yes	yes	0.1		none
54	XVI	138830-139449	yes	yes	yes	3.0		left

HMM results

fsORF	Chr.	Location	gDNA	mRNA	cDNA	FS	Notes	Class	Rank
16	VI	15473-14309	yes	yes	yes	9.0		both	1
26*	X	405173-406968	+ 1 nt	-	-	-	Re-annoted by SGD	both	2
25*	X	219713-217406	yes	yes	?	?		both	3
11*	IV	384077-381986	yes	yes	yes	11.0		both	4
40*	XIII	436627-438788	yes	yes	yes	5.0		both	5
15*	VI	123462-129904	yes	?	?	?		both	6
38*	XIII	263477-266754	yes	yes	?	?		both	7
30	XI	374144-374853	yes	yes	yes	12.0		left	1
22	VIII	262554-262197	yes	yes	yes	0.1		left	2
3	II	554266-553504	+ 1 nt	-	-	-	Re-annoted by SGD	left	3
33	XI	639597-638535	+ 1 nt	-	-	-	Re-annoted by SGD	left	4
10	IV	205690-205988	yes	yes	yes	0.1		none	1
24	VIII	499891-499585	yes	no	-	-		none	2
47	XIV	537790-538010	yes	yes	yes	?		none	2
44	XIV	394359-394026	yes	yes	yes	?		none	4
51	XV	782222-782003	yes	no	-	-		none	5
28	X	74021-74610	yes	intron	-	-		right	1
52	XV	80639-81189	yes	intron	-	-		right	2
32	XI	611160-611899	yes	yes	yes	7.0		right	3
12	IV	630075-630598	yes	intron	-	-		right	4
6*	III	220178-218372	yes	no	-	-	control	right	
7	III	222829-223097	yes	yes	yes	0.1	control	none	
8	III	91686-91455	yes	no	-	-	control	none	
13	V	183582-183327	yes	no	-	-	control	none	
18	VII	146543-146769	yes	no	-	-	control	none	
23	VIII	35126-34916	yes	no	-	-	control	none	
27	X	732756-732555	yes	no	-	-	control	none	
36	XII	767116-766933	yes	no	-	-	control	none	
46	XIV	429214-428983	yes	no	-	-	control	none	
55	XVI	935319-935028	yes	no	-	-	control	none	

* RT-PCR was carried with two sets of primers

Table 2

fsORF	Level	Hetamer	Sage	Overlap	size (aa)	ORF0	ORF-1	Notes
2	6%	AAAAAAA	Low	34	332	SCO2		SCO2 (involved in stability of Cox1p and Cox2p)
11*	11%	CCCAAAG	Low	64	698	*YDL038C***	PRM7	PRM7 (pheromone-regulated membrane protein)
16	9%		-	145	389	AAD6	*AAD16***	AAD6 (high similarity with the AAD of P. chrysosporium)
21	13%	UUUUUUU	-	88	143			Intergenic
30	12%		Medium	40	236	*YKL033W-A***		-
32	7%		High	46	246		SRL3	SRL3 (Suppressor of Rad53 null Lethality)
40*	5%		-	43	720	*YMR084W***	*YMR085W***	putative glutamine--fructose-6-phosphate transaminase
41*	5%		Low	121	635	ADE17		ADE17 (AICAR transformylase/IMP cyclohydrolase)
43	10%		-	28	330	MRPL24		MRPL24 (Mitochondrial ribosomal protein)
48	5%		Low	199	433	STE4		STE4 (GTP-binding protein beta subunit of the pheromone pathway)
50	7%		Low	49	421		RAD17	RAD17 (DNA damage checkpoint control protein)

* RT-PCR was carried with two sets of primers

** Hypothetical ORF

Article soumis

phpLabDB: a new gateway for private databases

Bekaert M.* and Rousset J.-P.

Institut de Génétique et Microbiologie, UMR CNRS 8621, Bâtiment 400, Université Paris-Sud,

91405 Orsay cedex, France

Running head: A gateway for laboratory databases

* To whom correspondence should be addressed

Abstract

Summary: We have conceived and developed a system for an easy design and access to shared relational databases. Built around independent modules that interact with each others, phpLabDB manages a variety of day-to-day useful laboratory information. It is also a highly customisable open-source laboratory database project.
Availability: The computer system core, with related modules, is available at http://phplabdb.sourceforge.net/
Contact: bekaert@igmors.u-psud.fr

Introduction

Advances and new developments of biology have produced a wide variety of biological data. The quantity of information generated in laboratories is in perpetual growth, and just as in sequence data banks, like the exponentially growing GenBank (Benson *et al.*, 2004), the number of sequenced plasmids, strain genotypes or synthesised oligonucleotides also grows rapidly (at the laboratory scale). Starting from this observation, labworkers often develop home tools to stock and to retrieve all those data.
We have conceived and developed a laboratory information management system (LIMS) implemented by a graphic browser, that groups independent results in a relational format. These latter would otherwise remain stored in a format of their own, depending on preferences of whom generated them. This allows to share resources and knowledge. We designed a platform-independent system which can be accessed remotely.
The aim of phpLabDB is that each scientist in the laboratory can manage his data quickly and make them available to the community. Due to Intranet and Extranet support, data can be accessed from any connected computer in the laboratory, which accelerates information retrieval. Laboratory staff spends less time on recurrent and time-consuming operations and dedicates more

time to research activity.
phpLabDB is a whole of modules for research laboratories, specifically designed for molecular biology laboratories but it can be generalised to other fields. It runs on a server and is accessed through a Web browser. It is constructed around an administration core and a user area. The user area allows to enrich the database as well as to browse and search it. The administration platform, which is user-restricted, allows to manage the database.

Core Program overview

The phpLabDB core is the central resource that all phpLabDB applications have in common; it includes some coding standards, common code and inter-module communication. The shared code provides common ways of handling features such as preferences, permissions, browser detection and help to user. Modules can easily interact and retrieve valuable information from others with relational databases across shared code. As actually released, provided modules are designed to support experimental data from a molecular biology laboratory, but phpLabDB is not exclusively dedicated to such use: rather, it is a generic LIMS platform.

Modules

Each module is focused towards a particular type of data and is dedicated to the collection and the retrieval of information relative to this data, as simple and clear as possible. For example, all actual modules include information on the location of a particular sample in the laboratory. However, most features remain optional to allow a greater flexibility in sample handling. Accordingly, several modules have been developed.
Like the core program, the phpLabDB modules are also written in PHP and when needed, some of them use Scalable Vector Graphics (SVG) for graphical data display. Several phpLabDB modules are already available but other modules can easily be developed and added to

phpLabDB.

- OligoDB plug-in is an oligonucleotide database manager. This module stores information about synthetic oligonucleotides and PCR primers. It allows searching, browsing, manipulating, associating, inserting, deleting, viewing of records stored in SQL database (Figure 1A) and calculating Tm (Howley *et al.*, 1979). Search engine can also accept sequence queries.
- PlasmidDB plug-in is a manager for a database of plasmid constructs (for use in transformation assays or in sequencing). This module stores and allows the search for plasmid data. PlasmidDB records can also be associated to a sequence and to a map (Figure 1B). Search engine can accept sequence queries. One can also find matching primers for the plasmid sequence. Annealing search is done in both direct and reverse directions.
- SeedDB plug-in is a seed database manager. SeedDB stores many details about seed prospections (e.g. species, landscape nature or plant precocity). The module can also keep track of crosses and genealogy (e.g. F1, F2...Fn, selfing, androgenesis, backcross). Records include details like seed availability or cross type.
- StrainDB plug-in is a strain database manager. This module stocks and displays information about all bacterial or other organism strains owned by the laboratory. It includes an efficient search for the origin of a strain, PudMed links, and storage location.

Implementation

phpLabDB provides a useful programming platform that a LIMS developer can use to set up a common permission control system in a multi-lab, multi-user environment. phpLabDB core is written in PHP and HMTL/CSS/JavaScript (Bos *et al.*, 1998 ; Raggett *et al.*, 1999), and is distributed under the Artistic Licence. Using such script language for the majority of the code allows for easy distribution of updates and development of new features. PHP is especially suited for Web development and explicitly designed to be embedded into HTML. PHP can also be embedded directly into the Apache web server as a module, making it extremely fast and versatile. It has a very fast engine and it features server plug-ins for Apache, IIS or Netscape.

phpLabDB core can be used with postgreSQL or MySQL as SQL server. Modules use Scalable Vector Graphics 1.1 to display data (Ferraiolo *et al.*, 2003). Installation is simple. The program has been tested on Linux (Suse, RedHat, Mandrake), Mac OS X and Windows 2000/XP.

Finally, phpLabDB is a web-style solution that gives labtools to easily manage crucial data and information. Since it is designed as an Intranet or Internet solution, it fully uses network capabilities and is accessible over all computers in a laboratory. Other modules can be developed and added to the system. The following interactive features have been implemented.

- Support for popular database servers MySQL and PostgreSQL.
- Shared code for common feature handling (e.g. preferences, permissions).
- Secured access for administration and management of user rights.
- Multiple language interface available and easily extended.
- Easy to install.
- Developed for Platform Portability.
- Supports barcode Code-39 for quick and easy storage.

Acknowledgements

The authors are deeply grateful to Christine Froidevaux and Marie-Stanislas Remigereau for critical reading of this manuscript and helpful suggestions.

References

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, Database issue, D23-26.

Bos,B., Lie,H.W., Lilley,C., and Jacobs,I. (1998) Cascading Style Sheets, level 2. Available at: http://www.w3.org/TR/CSS2/.

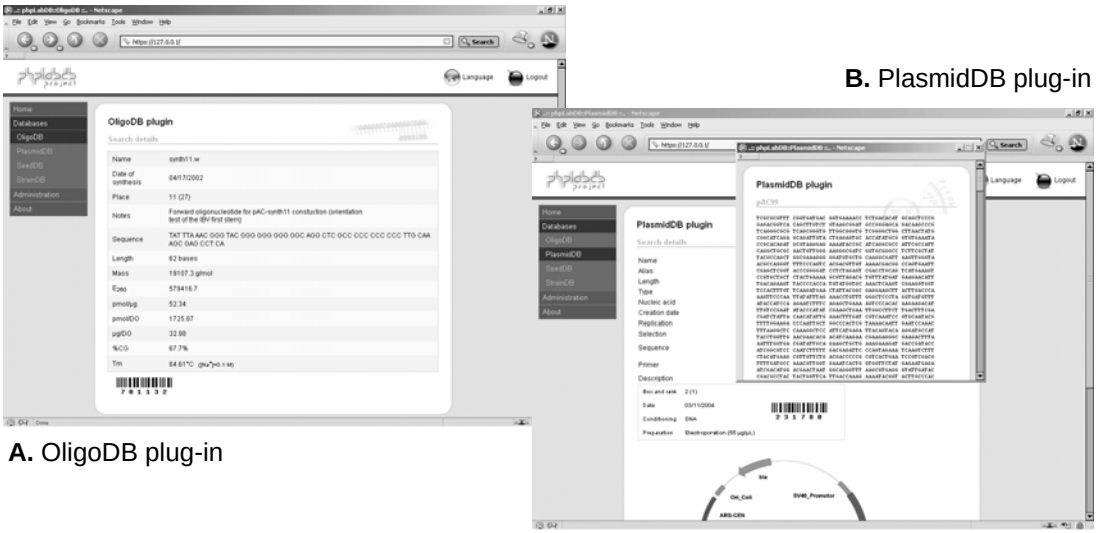
Ferraiolo,J., Fujisawa,J., and Jackson,D. (2003) Scalable Vector Graphics (SVG) 1.1 Specification. Available at: http://www.w3.org/TR/SVG11/.

Howley,P.M., Israel,M.A., Law,M.F., and Martin,M.A. (1979) A rapid method for detecting and mapping homology between heterologous DNAs. Evaluation of polyomavirus genomes. *J Biol Chem.*, **254**, 4876-4883.

Raggett,D., Le Hors,A., and Jacobs,I. (1999) HTML 4.01 Specification. Available at: http://www.w3.org/TR/html401/.

Figure

Figure 1



A. OligoDB plug-in

B. PlasmidDB plug-in

Fig. 1. Screenshot of phpLabDB modules. A) Details of records in the OligoDB plug-in, including oligonucleotide sequence, notes, Tm and barcode. B) Details of records in the PlasmidDB plug-in with SVG plasmid map and pop-up windows with plasmid sequence.

Étude du décalage de phase de lecture dans le génome de *Saccharomyces cerevisiæ*

Résumé

Le décalage de phase de lecture en -1 est un mécanisme de traduction non conventionnel des ARNm en protéines. Il contrôle la production de deux peptides différents à partir d'un messager unique. Bien que les exemples actuels de décalage de phase de lecture en -1 soient en grande partie limités aux génomes viraux et aux transposons, quelques événements bactériens et eucaryotes sont également documentés.

Mon travail de thèse a eu pour objet la recherche de gènes contrôlés par décalage de phase de lecture chez la levure *Saccharomyces cerevisiæ*, par des approches de biologie expérimentale et de bioinformatique. Une partie de cette étude a été réalisée en collaboration. Au cours de ces travaux, l'étude du mécanisme du décalage de phase de lecture eucaryote a aussi été abordée. Mes résultats ont permis de mettre en évidence l'effet de la modification des ARNt présents au site E du ribosome au moment du décalage sur l'efficacité de changement de cadre de lecture en -1.

Mots clés

Saccharomyces cerevisiæ ; décalage de phase de lecture ; HMM ; site E ; ribosome ; ARNt

Frameshift study in the genome of *Saccharomyces cerevisiæ*

Abstract

The ribosomal -1 frameshift is an unconventional mechanism of mRNA translation into protein. It controls the production of two different peptides from only one messenger. Although current examples of frameshifting are mainly limited to viral genomes and transposons, some bacterial and eukaryote events are also documented.

My thesis work aimed at the search for frameshift-controlled genes in *Saccharomyces cerevisiæ*, by experimental biology and bioinformatics approaches. Part of this study was collaborative. During this work, the study of -1 frameshifting mechanism was also tackled. Results enabled to highlight the effect of modifications within the tRNA present in the ribosomal E-site during the shift, on the frameshift efficiency.

Keywords

Saccharomyces cerevisiæ; frameshift; HMM; E-site; ribosome; tRNA