Expression stochastique des gènes chez Bacillus subtilis

Alexandre Deloupy

Sorbonne Université

Ecole doctorale : Physique en Île de France (ED 564)

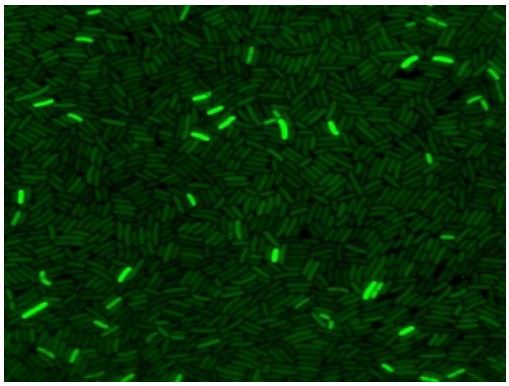
Laboratoire Jean Perrin

Expression stochastique des gènes chez *Bacillus subtilis*

Par Alexandre Deloupy

Thèse de doctorat de Physique

Dirigée par Jérôme Robert, Stéphane Aymerich et Lydia Robert.



Soutenue le 14 Décembre 2018

Devant le jury composé de :

Pr. Eric Clément	Président
Dr. Philippe Nghe	Rapporteur
Dr. Gregory Batt	Rapporteur
Dr. Stéphane Aymerich	Directeur de thèse
Dr. Sylvia De Monte	Examinatrice
Dr. Jerôme Robert	Invité
Dr. Lydia Robert	Invitée

« Any trouble, boy?

- *No, old man. Thought I was having trouble with my adding. It's all right now. »*

Table des matières

REMERCIEMENTS

Les remerciements constituent la dernière partie rédigée de cette thèse, et contrairement à ce que je pensais c'est aussi pénible à rédiger que le reste bien qu'on puisse être un peu plus libre contrairement au reste de la thèse.

En premier lieu je souhaite avant tout remercier ma famille, à savoir mes parents, pour leur amour inconditionnel, leur protection dans les moments difficiles, pour s'être assurer que je ne manquais de rien et pour leur (immense) potentiel comique. Je les remercie pour les week-ends bordelais passés ensemble qui ont été essentiels dans ce marathon qu'est la thèse. Comme le dit Tuco : « ça réconforte de savoir que, qu'il pleuve ou qu'il vente, il y a toujours quelque part une bonne soupe qui vous attend. ». A mon frère qui est responsable de la plupart des fous rires que j'ai pu avoir, et qui m'a permis d'avancer dans les moments un peu compliqués de ces trois années de thèse. Enfin, je remercie mes grands-parents pour avoir toujours cru en moi, en particulier ma grand-mère et mon grand-père de Mont de marsan, pour leur mot doux à chaque grande étape de ma vie d'étudiant. Une pensée pour papy Jo qui est parti au cours de ma première année de thèse et à Jean-Marie dont on ne peut sous-estimer l'absence. Enfin, une pensée pour Villegailhenc qui traverse une période difficile suite aux inondations du mois d'Octobre, inondations intervenues lors de la fin de la rédaction de ce manuscrit.

Je remercie mes directeurs de thèse, Lydia Robert, Jérôme Robert et Stéphane Aymerich pour m'avoir fait confiance sur cette thèse et de m'avoir proposé un sujet à la hauteur de mon talent pour me permettre d'obtenir le titre de docteur. Titre qui me revient d'ailleurs de droit divin (parce que je pense qu'on peut vraiment parler de droit divin). Je les remercie pour leur gentillesse à toute épreuve, notamment Lydia qui m'a beaucoup apporté et m'a redonné confiance dans les moments de doute. Je tiens aussi à remercier Matthieu Jules pour les discussions à Jouy et Vincent Sauveplane qui m'a beaucoup apporté durant cette thèse en me formant aux techniques de biologies moléculaires pour la construction des souches de *B. subtilis* et pour être venue m'ouvrir tous les matins en pleine crise du RER C. Je les remercie également pour leur accueil à l'institut Micalis à Jouy et pour m'avoir permis de manger gratuitement grâce à « la carte Micalis du 440 » et de découvrir les meilleurs yaourts qu'il m'ait été possible de déguster.

Cette thèse s'est essentiellement déroulée au laboratoire Jean Perrin. Je tiens donc à remercier l'ensemble des personnes du laboratoire. Une mention particulière pour l'équipe Biophysique des micro-organismes, à savoir Jérôme,

Lydia, Marina, Nelly, Maxime et Jean. Une mention particulière pour Philippe qui est parti pour Nice au cours de ma thèse et qui nous a fait entrer moi et mes acolytes dans la « NFL-zone ». Je remercie chaleureusement Didier Chatenay pour garantir une ambiance de travail amicale et privilégiée (cette dernière phrase est issue des remerciements de Florence). Je remercie également l'ensemble de l'équipe SYBER de l'institut Micalis pour leur gentillesse et leur très bon accueil, à savoir Matthieu, Vincent, Etienne, Anne-Gaëlle, Olivier, Stephen, Magali, Teddy, Rahma, Marwa et Gabriella.

Cette thèse aurait pu être comme un film d'Olivier Marchal (qui sont particulièrement mauvais), c'est à dire rugueux et ne laissant pas beaucoup de place à la fantaisie et l'humour. Heureusement, cette thèse a été l'occasion de croiser la route d'autres doctorants. Je tiens à remercier l'ensemble des doctorants que j'ai côtoyé en premier ceux du bureau 414 bien évidemment : Florence, avec qui j'ai passé l'intégralité de ces trois années de thèse. Si la thèse est un marathon alors les chronos réalisés sont franchement pas mal (contrairement à d'autres). Merci pour avoir supporté mes blagues souvent limites qui étaient la plupart du temps faites pour t'énerver je l'avoue. Il est tard (1h27 du matin) quand j'écris ces remerciements donc je vais pas trop m'étendre mais je suis très heureux et fier d'avoir partagé ces trois années de thèse avec toi. Je remercie également Anis pour l'animation politique des repas et des pauses, pour m'avoir fait découvrir Tutotal et pour tout un tas d'autre chose que je ne saurais pas vraiment dire. Je remercie ensuite Manon et ce malgré ton empreinte carbone déplorable (contrairement à la mienne qui est exemplaire). Tu as été une source de motivation par rapport à l'énergie qui t'habites et ton abnégation. Je remercie Jean-Baptiste un toulousain tout blond aux yeux bleus. Parce que c'est bien d'être toulousain, et c'est un bordelais qui le dit. Pour finir avec le bureau 414 je remercie mon acolyte de microbiologie Amaury, quand j'écris ces remerciements ton équipe mène 299 à 254. Et pour info il te reste 1 WR et moi 2 WR ma DEF et 1 DL donc c'est pas fini. Merci pour avoir pester avec moi sur cette p….. de hotte stérile et pour rentrer dans les gens comme tu sais si bien le faire et apprécier mes blagues. Je remercie également les doctorants « poissons » : Geoffrey, que j'ai aperçu en M1 notamment en rattrapage de mécanique quantique, te fréquenter m'a confirmer tout le bien que j'imaginais. Je remercie également Guillaume pour les sourires dans les couloirs et enfin la franco-germano-russe Sophia, parce que c'est bien d'être russe (c'est quand même le pays de Tatu).

Je remercie également les membres de mon jury de thèse, Eric Clément, pour avoir accepté d'être président du jury de thèse, Philippe Nghe et Gregory Bratt pour avoir accepté d'être les rapporteurs de ce manuscrit et enfin Silvia De Monte pour avoir accepté d'être examinatrice du jury.

Enfin je tiens à remercier Loïc Marrec (ou Loïkoum le gros loukoum) qui est en réalité la vrai découverte de cette thèse. Pour ces fois où nous sommes parties en *Leonie*, pour Romane, Adeline et bien sûr ce cher Fitzpatrick. Pour avoir été là du début de la rédaction jusqu'à la fin. Toi seul connaît le vrai sens de cette thèse qui fut un véritable combat. En tout cas Alexandre ne serait pas aller bien loin sans Loïc. *Goeie dag ! Aangename kennis ! Ons vir jou Suid Afrika. Hier kom die bokke. N'kosi Sikelel'iAfrika.* Pour finir, je remercie Olivier Noël pour les soirs à décompresser et les balades nocturnes autour de la tour Montparnasse et pour m'avoir inciter à courir. On a chacun écrit quelque chose et j'espère sincèrement que ton œuvre sera lu par (beaucoup) plus de monde que cette thèse.

SUMMARY

Stochastic gene expression in *Bacillus subtilis*:

A population of genetically identical individuals sharing the same environment exhibits some residual phenotypic variability. Such heterogeneity arises from the stochastic, or random, nature of gene expression also referred as noise. This stochasticity results on the one hand from the random encounter of chemical species during both transcription and translation (intrinsic noise), and on the other hand from the fluctuations in the concentration of these chemicals (extrinsic noise). A stochastic model involving only intrinsic noise predicts that phenotypic noise strength varies linearly with translational efficiency but does not depend on transcriptional one. This prediction was shown to be compatible with data on a limited number of strains and conditions but has never been fully tested on a large collection of strains with different transcription and translation efficiencies.

We aim to go further in the test of this prediction by using a collection of ~40 strains of the bacterium Bacillus subtilis where GFP is expressed under the control of different Promoters, TSS and RBS. For each strain, cell-to-cell heterogeneity is investigated by quantifying fluorescence signal at the single cell level, based on flow cytometry techniques and epifluorescence microscopy. Our results show that, contrary to expectations, phenotypic noise strength shows a strong positive correlation with transcriptional efficiency. We demonstrated that over a wide range of expression covering most of the proteome of *B. subtilis*, the expression noise is dominated by external noise sources. Therefore, stochastic models of gene expression are not suitable for quantifying the effects of translation and transcription on gene expression noise.

RESUME

Expression stochastique des gènes chez *Bacillus subtilis* :

Une population d'individus génétiquement identiques partageant le même environnement présente une certaine variabilité phénotypique résiduelle. Cette hétérogénéité découle de la nature stochastique, ou aléatoire, de l'expression des gènes, également appelée bruit. Cette stochasticité résulte d'une part de la rencontre aléatoire d'espèces chimiques pendant la transcription et la traduction (bruit intrinsèque), et d'autre part des fluctuations dans la concentration de ces substances chimiques (bruit extrinsèque). Un modèle stochastique ne faisant intervenir que le bruit intrinsèque prédit que la force du bruit phénotypique varie linéairement avec l'efficacité de la traduction, mais qu'il ne dépend pas du taux de transcription. Cette prédiction s'est révélée compatible avec des données portant sur un nombre limité de souches et de conditions, mais n'a jamais été entièrement testée sur un grand nombre de souches ayant différentes efficacités de transcription et de traduction.

Notre objectif est d'aller plus loin dans le test de cette prédiction en utilisant une collection d'une quarantaine de souches de la bactérie *Bacillus subtilis* où la protéine GFP est exprimée sous le contrôle de différents promoteurs, TSS et RBS. Pour chaque souche, l'hétérogénéité entre cellules est étudiée en quantifiant le signal de fluorescence au niveau de la cellule unique, à l'aide de techniques de cytométrie en flux et de microscopie en épifluorescence. Nos résultats montrent que, contrairement aux attentes, la force du bruit phénotypique montre une forte corrélation positive avec l'efficacité transcriptionnelle. Nous avons démontré que sur une large gamme d'expression couvrant la majeure partie du protéome de *B. subtilis*, le bruit d'expression est dominé par les sources de bruit extrinsèques. Par conséquent, les modèles stochastiques d'expression génique ne conviennent pas pour quantifier les effets de la traduction et de la transcription sur le bruit d'expression génétique.

INTRODUCTION :

I Expression des gènes :

La synthèse protéique au sein d'une cellule implique deux étapes biochimiques majeures (Figure 1-A) :

- Transcription : synthèse de l'ARN messager (ARNm) à partir du gène présent sur une molécule d'ADN.
- Traduction : synthèse de la protéine à partir d'une molécule d'ARNm.

A côté ce ces deux étapes du dogme central de la biologie moléculaire, il faut également ajouter les étapes de dégradations des ARN messagers et des protéines respectivement par les RNases et les protéases. La dégradation des ARN messagers limite fortement leur durée de vie dans la cellule, qui est en général de l'ordre de quelques minutes (Taniguchi et al, 2010). Les protéines sont plus stables et peuvent avoir un temps de demi-vie de plusieurs heures (Taniguchi et al, 2010).

I.1 Transcription :

L'initiation de la transcription nécessite la rencontre d'une enzyme, l'ARN polymérase, avec une région particulière de l'ADN en amont du gène appelée promoteur. Lors de l'initiation, la polymérase s'attache de façon réversible au promoteur pour former un complexe « fermé ». La polymérase déroule ensuite quelques bases de la double hélice pour permettre la lecture du brin matrice par son site catalytique, formant ainsi un complexe ouvert. Finalement, la polymérase quitte le promoteur et se déplace le long du gène pour synthétiser l'ARN complémentaire au brin matrice : c'est la phase d'élongation. Une fois le début de la phase d'élongation, le promoteur est libre pour accueillir une nouvelle ARN polymérase et initier la transcription à nouveau. L'affinité de la polymérase pour un promoteur donné dépend de la séquence de ce dernier. La polymérase est composée de 5 sous unités et une protéine spécifique, le facteur σ, va se rajouter à ce noyau enzymatique. Le rôle du facteur σ est de diriger l'ARN polymérase vers la séquence promotrice, en se fixant spécifiquement sur deux séquences de quelques nucléotides situées autour des positions -35 et -10. Ces séquences ont donc une grande influence sur le taux de transcription du gène en aval. Le site +1 indique en général la position du premier nucléotide transcrit. Cependant, l'ARN

polymérase peut parfois démarrer la phase d'élongation de l'ARN messager au niveau des positions +2 ou +3, phénomène observé à la fois chez *E. coli* et *B. subtilis*. (Guiziou et al, 2016). La probabilité de débuter la transcription à tel ou tel site dépend de la nature du nucléotide présent (A, T, G, C) et du niveau intracellulaire de nucléotides triphosphates (NTPs) affiliés. Ainsi, le taux de transcription dépend des séquences -35 et -10 mais aussi des quelques premiers nucléotides en aval du site « +1 ». L'efficacité de transcription peut également être contrôlée par certaines protéines régulatrices appelées facteurs de transcriptions. Cette régulation transcriptionnelle permet aux cellules d'exprimer certains gènes et d'autre pas à un instant donné, ou encore de moduler l'activité de certains gènes.

I.2 Traduction :

Le démarrage (ou initiation) de la traduction chez les bactéries débute par la formation d'un complexe entre la petite sous-unité 30S du ribosome et la molécule d'ARN messager. Cette association est assurée par l'appariement de l'ARN ribosomal 16S avec une séquence appelée « Ribosome Binding Site » située à l'extrémité 5' de l'ARN messager en amont du codon d'initiation. L'efficacité d'initiation de la traduction dépend de la séquence TIR (*Translational Initiation Region*), i.e. la région comprise entre le promoteur et le codon d'initiation (souvent AUG). Ce domaine comprend la séquence RBS et d'autres séquences (enhancers) permettant d'améliorer le processus de traduction (Vimberg et al, 2007). L'efficacité de la traduction peut être médiée par la capacité de la séquence RBS à recruter fréquemment les ribosomes. Cette séquence de quelques nucléotides est séparée du codon d'initiation par une région d'environ 10 nucléotides. La taille et la nature de cette séquence peuvent influencer le positionnement du ribosome sur le codon d'initiation et donc également stimuler ou altérer l'efficacité de traduction de l'ARN messager. Éventuellement d'autres éléments stimulateurs se trouvant en amont de la séquence RBS peuvent intervenir dans l'efficacité de la traduction.

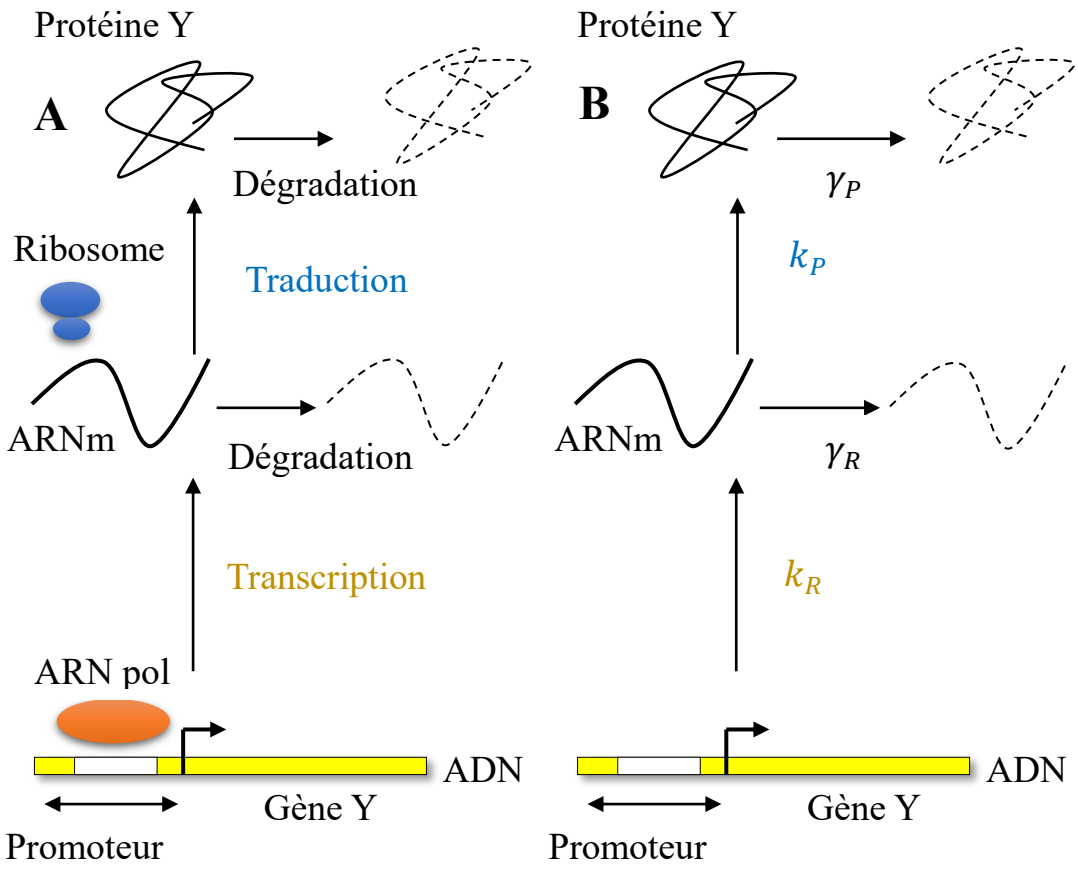


Figure 1 : Expression génique dans le cas d'une transcription non régulée. A) Un promoteur constitutif recrute une ARN polymérase et initie la transcription. L'ARN messager qui en résulte est par la suite traduit par les ribosomes pour former les protéines jusqu'à dégradation du transcrit. Les protéines sont par la suite dégradées. **B)** Interprétation mathématique du dogme central servant de base au modèle déterministe et stochastique à deux niveaux. Toutes les réactions sont supposées d'ordre un avec des taux de productions d'ARN messagers et de protéines k_R, k_P et des taux de dégradation pour ces molécules γ_R, γ_P.

II Stochasticité de l'expression génique :

L'expression d'un gène est contrôlée par les concentrations, activités et localisations de nombreuses molécules comme les polymérases, les facteurs de transcriptions ou encore les ribosomes. Ces espèces chimiques ainsi que l'ADN porteur de l'information génétique évoluent dans le cas des cellules procaryotes dans le cytoplasme et sont soumis à l'agitation thermique. Leur localisation résulte donc d'une marche aléatoire en 3D. Ainsi la rencontre entre ces différentes espèces chimiques se fait de manière aléatoire ou stochastique. Considérons deux cellules bactériennes d'une population monoclonale, donc génétiquement identiques. A un instant donné, si nous nous intéressons à un gène en particulier, il est possible que pour une des cellules, le promoteur de ce gène soit occupé par une polymérase, permettant l'initiation de la transcription, tandis que pour l'autre

cellule le promoteur soit libre. Ainsi le nombre de protéines dans les deux cellules sera différent. La nature probabiliste des rencontres entre espèces chimiques conduit également à des fluctuations du nombre de protéines au cours du temps au sein d'une même cellule. En particulier, les concentrations en polymérases et ribosomes peuvent fluctuer au sein d'une même cellule. Ces fluctuations peuvent se propager et ainsi entraîner des fluctuations dans le nombre de protéines entre cellules génétiquement identiques. L'expression des gènes est donc fondamentalement stochastique. La physique statistique réconcilie le caractère discret et stochastique des échelles microscopiques au caractère continu et déterministe des échelles macroscopiques (Erwin Schrödinger, *Qu'est-ce que la vie ?*). On peut citer par exemple le paramagnétisme où à l'échelle microscopique chaque molécule s'oriente aléatoirement à cause de l'agitation thermique mais du fait de leur grand nombre, elles produisent une orientation prépondérante dans la direction du champ appliqué. Pour la bactérie *Escherichia coli*, les nombres d'ARN polymérases et de ribosomes sont assez élevés, respectivement de l'ordre de 5000 et 27000 par cellules (Bremer & Dennis, 1996). Mais dans le cas de l'expression génétique, le facteur limitant reste la très faible quantité d'ADN dans une cellule, qui est présent en nombre de copies de l'ordre de l'unité. Le nombre de transcrits dépend beaucoup du gène considéré et particulièrement de l'efficacité du promoteur à recruter les ARN polymérases : de 1500 transcrits par génération à 1 transcrit par génération (Leroy, 2010). Ainsi, le caractère fondamentalement aléatoire de l'expression des gènes conduit à une variabilité facilement détectable entre cellules génétiquement identiques au vu des niveaux de fluctuations relatives dans la quantité de ces acteurs principaux. Une population d'individus génétiquement identiques et partageant le même environnement manifeste donc une diversité résiduelle appelée variabilité phénotypique ayant pour origine le caractère aléatoire de l'expression des gènes. Un autre terme employé pour désigner cette variabilité est celui de bruit.

II.1 Bruit intrinsèque et bruit extrinsèque :

Dans leurs travaux pionniers, Elowitz et al. et Swain et al. proposent de diviser les sources de bruit dans l'expression des gènes en deux catégories : le bruit intrinsèque et le bruit extrinsèque (Elowitz et al, 2002 ; Swain et al, 2002). Comme nous l'avons vu précédemment, l'expression de chaque gène est contrôlée par la concentration, l'activité et la localisation de nombreuses molécules, comme les polymérases, les ribosomes ou d'éventuelles protéines régulatrices. Ainsi, les fluctuations dans les concentrations de ces molécules vont entraîner des fluctuations dans l'expression génique. Elles représentent donc une source de bruit qui est globale à une cellule donnée et est catégorisée comme extrinsèque.

Si on suppose deux copies du même gène dans une cellule donnée, ces fluctuations extrinsèques à ces gènes devraient affecter de la même manière l'expression des deux copies tandis qu'elles affecteraient de manière différentes deux gènes identiques mais dans deux cellules différentes (caractérisés par des variables extrinsèques différentes). D'une manière générale, on regroupe dans le bruit extrinsèque tous les phénomènes indépendants du gène considéré et global à une cellule donnée, dont les fluctuations peuvent se propager à l'expression de notre gène. Maintenant, considérons deux cellules ne montrant aucune différence dans leurs variables extrinsèques. Nous avons déjà énoncé le fait que les rencontres entre polymérases et promoteurs ou encore entre ribosomes et RBS se font de manière aléatoire du fait de l'agitation thermique. Ainsi les initiations de la transcription et de la traduction n'auront pas lieu au même moment dans les deux cellules. Cette stochasticité résiduelle, inhérente au processus d'expression est catégorisée comme intrinsèque.

La distinction entre bruit intrinsèque et bruit extrinsèque a été formalisée mathématiquement par Swain et al (Swain et al, 2002). Dans cette étude, Swain et al base cette distinction sur la formule de la variance totale. Pour deux variables aléatoires Y et X, la variance de Y peut s'écrire selon :

$$Var(Y) = E\big(Var(Y|X)\big) + Var\big(E(Y|X)\big)$$

Où $E(\,)$ représente l'espérance mathématique. Swain et al applique cette formule lorsque la variable aléatoire Y est égale au nombre de protéines à un instant donné et X est égale à un vecteur des variables extrinsèques qui spécifie l'environnement cellulaire (état du cycle cellulaire, nombres d'ARN polymérases, ribosomes…) à un instant donné. Ainsi, en notant $P(I,E)$ le nombre de protéines comme une fonction du vecteur I des variables intrinsèques et du vecteur E des variables extrinsèques, on peut écrire :

$$Var\big(P(I,E)\big) = E\big(Var(P(I,E)|E)\big) + Var\big(E(P(I,E)|E)\big)$$

En définissant le bruit intrinsèque comme

$$\eta_{int}^2 = E\big(Var(P(I,E)|E)\big)/E(P(I,E))^2$$

Et le bruit extrinsèque comme

$$\eta_{ext}^2 = Var\big(E(P(I,E)|E)\big)/E(P(I,E))^2$$

On obtient la partition du bruit total en une composante intrinsèque et une composante extrinsèque :

$$\eta_{tot}^2 = \eta_{int}^2 + \eta_{ext}^2.$$

En parallèle de cette approche mathématique, Elowitz et al. ont mesuré les contributions intrinsèques et extrinsèques dans le bruit d'expression génétique chez *E. coli* (Elowitz et al 2002). Pour ce faire, ils ont introduit dans le chromosome d'*E. coli* les gènes de deux protéines fluorescentes CFP et YFP, respectivement de couleur bleue et jaune, contrôlés par le même promoteur et localisés de façon symétrique par rapport à l'origine de réplication du chromosome. Cette localisation permet d'éviter les différences d'expression causées par les fluctuations du nombre de copies de gènes liées au cycle cellulaire. Ainsi, une corrélation entre les nombres de protéines YFP et CFP par cellule indique la présence de bruit extrinsèque, tandis que le bruit intrinsèque tend à décorréler les deux signaux. Grâce à cette technique, Elowitz et al. ont montré que les deux sources de stochasticité, intrinsèque et extrinsèque contribuent de façon substantielle au bruit total. Dans leurs expériences, le niveau d'expression des gènes CFP et YFP pouvait être modulé grâce à l'usage d'un promoteur inductible. En modulant ainsi le niveau d'expression, ils ont montré que le bruit intrinsèque décroit avec le niveau d'expression tandis que le bruit extrinsèque évolue de façon non monotone et est maximal pour des niveaux d'induction intermédiaire. Ce comportement peut s'expliquer par les fluctuations dans la concentration d'inducteur, qui sont plus importantes pour des niveaux d'induction intermédiaires. A la suite de cette étude, une approche similaire a été utilisée pour évaluer l'importance relative du bruit extrinsèque et intrinsèque dans l'expression des gènes chez les eucaryotes (Raser and O'Shea, 2004). Raser et al. ont construit des souches de levure *Saccharomyces cerevisiae* exprimant la protéine CFP et la protéine YFP, à partir de gènes localisés sur le même locus sur des chromosomes homologues et contrôlés par le même promoteur. Comme pour les travaux de Elowitz et al., la corrélation entre les signaux de YFP et CFP reflète l'importance relative des sources extrinsèques et intrinsèques de bruit. Raser et al. ont utilisé plusieurs promoteurs et ont montré que le bruit était dominé par sa composante extrinsèque.

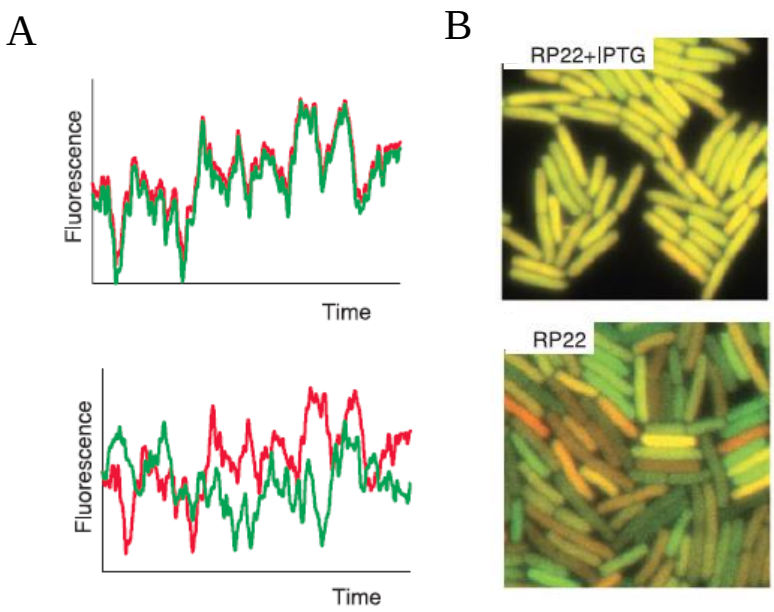


Figure 2 : Bruit intrinsèque et extrinsèque dans l'expression des gènes chez *E. coli.* A) En l'absence de bruit intrinsèque, le bruit extrinsèque cause des fluctuations corrélées pour 2 copies du même gène dans une cellule (rouge et vert, haut) ; la corrélation diminue en présence de bruit intrinsèque (bas) (Elowitz et al, 2002). **B)** Superposition d'images de fluorescence de YFP et CFP (ici rouge et vert) pour une souche portant les gènes CFP et YFP sous le même promoteur inductible, en condition d'induction totale (haut), ou de répression totale (bas) (Elowitz et al, 2002).

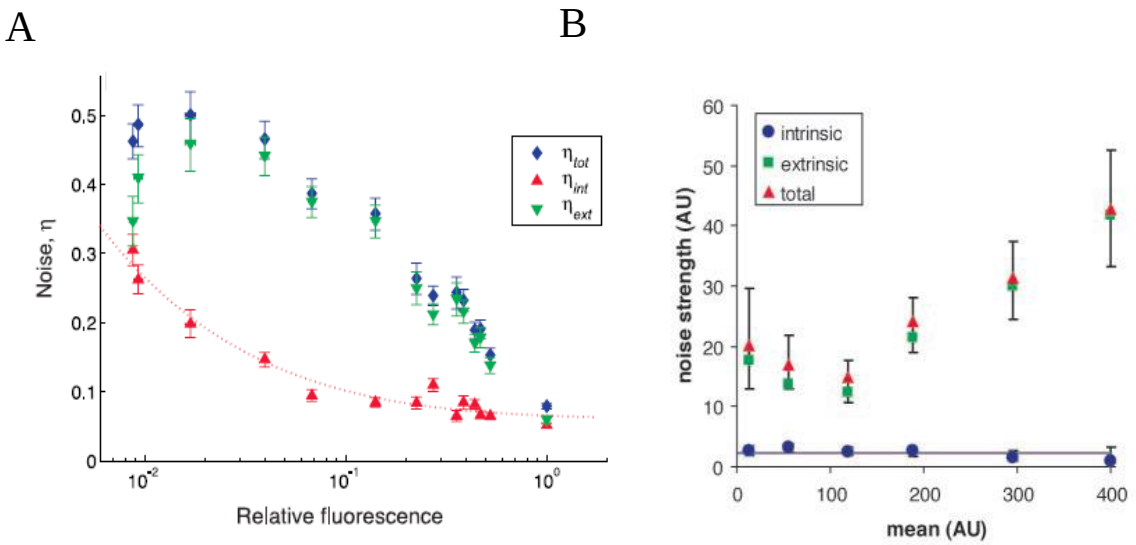


Figure 3 : Bruit intrinsèque et extrinsèque des gènes chez les procaryotes et les eucaryotes. A) Bruit extrinsèque (vert), intrinsèque(rouge) et total pour une souche d'*E. coli* portant les gènes CFP et YFP sous le contrôle d'un promoteur inductible (voir figure 2). Le niveau moyen de fluorescence (axe des x) est modulé en changeant le niveau d'induction du promoteur (Elowitz et al, 2002). **B)** Bruit extrinsèque (vert), intrinsèque (bleu) et total (rouge) pour une souche de *S. cerevisiae* exprimant les gènes YFP et CFP insérés à des loci homologues sous le contrôle du même promoteur. Le niveau moyen de fluorescence (axe des x) est modulé en changeant le niveau d'induction du promoteur (Raser & O'Shea, 2004).

II.2 Conséquences biologiques du bruit dans l'expression génétique :

L'utilisation du terme « bruit » est assez révélatrice de la vision négative associée à cette stochasticité, notamment en biologie synthétique. Le but de la biologie synthétique est d'étendre ou de modifier le comportement des organismes et de les concevoir pour accomplir de nouvelles tâches (E. Andrianantoandro, 2006), par exemple la synthèse de nouvelles molécules d'intérêt. Dans ce contexte, différents gènes sont mis en réseaux, une protéine synthétisée par un de ces gènes jouant un rôle dans l'expression d'une autre. Ainsi, le bruit généré au niveau d'une protéine donnée aura des répercussions sur la synthèse d'autres protéines avec lesquelles elle interagit. Cette propagation du bruit ne peut donc que nuire à la performance du processus de synthèse d'une protéine d'intérêt puisqu'on veut sa production prévisible, fiable et uniforme. Aussi, loin de ce contexte industriel, le bon fonctionnement d'une cellule nécessite des réactions biochimiques avec une stœchiométrie précise. Ainsi le bruit dans l'activité ou la quantité de ces réactants semble à première vue préjudiciable et force la cellule à développer des stratégies pour le combattre. Cependant, il a été montré récemment que le bruit pouvait également être exploité par les organismes vivants pour permettre à une population monoclonale de se diversifier phénotypiquement. Bien que la génération de phénotypes non parfaitement optimaux puisse être préjudiciable, certains de ces phénotypes peuvent être plus adaptés dans d'autres environnements et la maintenance d'une certaine diversité phénotypique peut permettre ainsi à la population de survivre dans des environnements fluctuants (stratégie dite de « *bet hedging* », par analogie aux stratégies de réduction des risques sur les marchés financiers). Dès lors, si le caractère probabiliste de l'expression génétique peut s'avérer être dangereux pour la cellule dans certains cas et bénéfique dans d'autres, il convient alors à celle-ci de pouvoir le contrôler : le minimiser pour s'en protéger ou l'amplifier pour en tirer profit.

II.3 Mesurer et quantifier la stochasticité dans l'expression génétique :

L'un des principaux objectifs des récentes recherches sur la stochasticité de l'expression génétique a été d'identifier et de quantifier ses différentes sources. Ceci a été rendu possible d'une part par les progrès en biologie moléculaire et d'autre part par l'accessibilité grandissante des outils quantitatifs. Ainsi, presque toutes les expériences visant à caractériser le bruit d'expression génétique débutent par l'insertion d'un gène rapporteur codant pour une protéine

fluorescente dans le génome de l'organisme étudié. Le niveau d'expression de ce gène pouvant être modulé suivant la force du promoteur ou du RBS situés en amont de la partie codante. Finalement, les signaux de fluorescence de nombreuses cellules isogéniques sont quantifiés par microscopie de fluorescence (Elowitz et al, 2002 ; Golding et al, 2005 ; Cai et al, 2006 ; Yu et al, 2006 ; Choi et al, 2008 ; Taniguchi et al, 2010) ou cytométrie en flux (Ozbudak et al, 2002 ; Blake et al, 2003 ; Raser & O'Shea, 2004 ; Newman et al, 2006 ; Bar-Even et al, 2006). Une fois mesurés les signaux de fluorescence, une mesure de l'hétérogénéité des niveaux de fluorescence doit être définie. L'écart-type de la distribution σ_P , défini par : $\sigma_P = \sqrt{\langle P^2 \rangle - \langle P \rangle^2}$, où P représente le signal de fluorescence (rapportant le nombre de protéines fluorescentes) et $\langle\ \rangle$ indique la valeur moyenne, reflète l'étendue des valeurs possibles des signaux de fluorescence autour de la valeur moyenne. Il semble donc être un bon candidat à première vue pour quantifier le bruit d'expression. Cependant il est important de noter que les signaux de fluorescence, bien que proportionnels aux quantités de protéines présentes dans une cellule, sont donnés en unité de fluorescence arbitraire. Ils ne seront donc pas nécessairement égaux d'une expérience à une autre si elles sont faites sur le même dispositif expérimental (les signaux de sortie dépendant grandement de l'électronique de la chaîne d'acquisition), et seront dans tous les cas très différents si le dispositif utilisé est différent entre les deux expériences, chaque système ayant sa propre unité de mesure. Ainsi l'écart type suivra la même logique : si une variable aléatoire (ici la fluorescence d'une cellule) est multipliée par une constante, l'écart-type sera multiplié par la même constante. Ainsi l'écart-type ne peut être un bon candidat pour la mesure de variabilité, puisque différentes unités de mesure donneraient une hétérogénéité différente pour la même variable. Par conséquent, une quantité souvent utilisée pour mesurer les fluctuations est le coefficient de variation, défini par :

$$\eta_P = \frac{\sigma_P}{\langle P \rangle}$$

Cette mesure est souvent appelée « bruit » dans la littérature. Cette quantité a l'avantage d'être sans dimension et n'est donc pas soumis à l'effet d'échelle mentionné précédemment. Cependant, cette quantité est en général soumis à un effet de taille (Ozbudak et al, 2002) : plus la moyenne sera faible, plus les fluctuations relatives seront importantes et inversement, plus la moyenne sera importante, moins la variabilité le sera. Par exemple, pour des fluctuations Poissonniennes,

$$\eta_P = \frac{\sigma_P}{\langle P \rangle} = \frac{1}{\sqrt{\langle P \rangle}}$$

Cet effet trivial pourrait parfois masquer les effets de certaines sources de stochasticité sur la variabilité phénotypique. Ainsi, une autre quantité est parfois proposée pour quantifier le bruit d'expression, il s'agit de la force du bruit (« Noise-strength » en anglais) ou facteur Fano défini par :

$$Noise - strength = \frac{\sigma_P^2}{\langle P \rangle} = \eta_P^2 \langle P \rangle$$

Cette quantité présente le désavantage d'être de la même dimension que le signal de fluorescence mais présente l'avantage d'être toujours égale à 1 pour des fluctuations Poissonniennes. Le facteur Fano peut donc permettre de contourner l'effet trivial d'une diminution du bruit avec une augmentation de la moyenne (Thattai & van Oudenaarden, 2001 ; Ozbudak et al, 2002) et permet de comparer la variabilité à une simple variabilité Poissonnienne. Si nous imaginons que nous soyons capables de quantifier les signaux de fluorescence en termes de nombres de molécules, alors si ces signaux sont distribués suivant une loi de Poisson, nous aurons un « Noise-strength » égal à 1. Ainsi, un facteur Fano supérieur à 1 implique une distribution qui est plus large qu'une distribution de Poisson avec la même moyenne, et à l'inverse, un facteur Fano inférieur à un une distribution plus étroite qu'une distribution de Poisson avec la même moyenne. A noter que dans le cas où nous mesurons les signaux en unité arbitraire, la force du bruit serait égale au coefficient de proportionnalité entre la mesure du signal et le nombre de protéines présent dans une cellule.

III Modèles stochastiques de l'expression génique :

Pour interpréter les mesures de variabilité, il est souvent fait appel à la modélisation. Nous présentons dans cette partie deux modèles développés par plusieurs auteurs pour essayer de relier la variabilité phénotypique à des paramètres biophysiques intervenant dans différentes étapes de l'expression génétique et ainsi identifier les différentes sources de stochasticité et quantifier l'importance relative de ces différentes sources. Ces modèles ont été à l'origine de nombreuses expériences qui ont alors pour but de tester leurs prédictions et seront présentés dans une seconde partie.

Modéliser l'expression génique revient à interpréter mathématiquement les différentes étapes biochimiques de la synthèse protéique et d'en déduire les quantités d'ARN messagers et de protéines à un instant donné. Deux modèles « standards » de l'expression génique ont été proposés : (i) un modèle pour les promoteurs constitutifs, c'est-à-dire non régulés et qui sont toujours aptes à recruter les ARN polymérases et ainsi à initier la transcription (Thattai & van Oudenaarden, 2001 ; Paulsson, 2005 ; Shahrezaei & Swain, 2008). Ce modèle est

appelé modèle à deux niveaux de l'expression génique (Shahrezaei & Swain, 2008). (ii) un modèle décrivant des promoteurs régulés (Paulsson, 2005 ; Shahrezaei & Swain, 2008). Cette régulation se fait au moyen de facteurs de transcription, qui peuvent augmenter l'affinité de la polymérase pour le promoteur ou bien au contraire en interdire l'accès. Ce modèle est appelé modèle à trois niveaux de l'expression génique (Shahrezaei & Swain, 2008). Il est important de noter que ces deux modèles ne prennent pas en compte les étapes du cycle cellulaire comme la réplication de l'ADN qui double le nombre de copies du gène au sein de la cellule, ainsi que la ségrégation des différentes molécules au cours de la division cellulaire. De plus, ces modèles décrivent la composante intrinsèque du bruit et ne prennent pas en compte les fluctuations extrinsèques, en particulier les fluctuations de concentration des machineries de transcription et traduction.

III.1 Modèle stochastique à deux niveaux de l'expression génique :

III.1.1 Présentation du modèle :

Le modèle à deux niveaux permet de décrire la dynamique de production des ARN messagers et des protéines dans le cas d'un promoteur constitutif. Ce modèle a été introduit au début des années 2000 par plusieurs auteurs (Paulsson, 2005 ; Thattai & van Oudenaarden, 2001 ; Swain et al, 2002 ; Shahrezaei & Swain, 2008). L'idée du modèle est de regrouper les différentes étapes de la transcription en une seule avec un taux de production d'ARN messagers k_R. On procède de la même manière pour la traduction avec un taux de production de protéines k_P par ARN messager, et pour la dégradation des messagers et des protéines avec des taux de dégradation respectifs $\gamma_R$et γ_P (Figure 1-B). Autrement dit si on note $r(t)$ et $p(t)$ respectivement les nombres d'ARN messagers et de protéines à un instant t, alors nous avons les réactions :

Transcription :	$r \xrightarrow{k_R} r+1$
Dégradation ARNm :	$r \xrightarrow{r\gamma_R} r-1$
Traduction :	$p \xrightarrow{rk_P} p+1$
Protéolyse :	$p \xrightarrow{p\gamma_P} p-1$

Ce modèle est illustré en figure 4. Les paramètres biophysiques k_R, k_P, γ_R et γ_P rendent compte d'une certaine réalité moléculaire. Par exemple dans ce modèle, k_R et k_P déterminent respectivement les efficacités de transcription et de traduction. Plus k_R est important, plus on pourra initier de transcription. D'un

point de vue moléculaire, un k_R important pourrait correspondre à un promoteur fort, c'est-à-dire capable de recruter fréquemment les ARN polymérases, et donc d'initier de nombreuses transcriptions successivement ; tandis qu'un k_R faible serait plutôt adapté à un promoteur faible. De même, un k_P important correspond à une séquence RBS capable de recruter fréquemment un ribosome et donc d'initier de nombreuses traductions successives, tandis qu'un k_P faible correspond à un ARN messager faiblement traduit. Les quantités d'ARN polymérases, ribosomes ou protéines du dégradosome assurant les dégradations des ARN messagers et des protéines se retrouvent dans les taux de réactions. On a par exemple : $k_R = \widetilde{k_R}[ARNpol]$ et $k_P = \widetilde{k_P}[Ribosome]$. On peut écrire une version déterministe de ce modèle, qui est développée en annexe (annexe 1). Dans un cadre déterministe, à l'état stationnaire toutes les cellules ont la même quantité d'ARNm $r^{st} = \frac{k_R}{\gamma_R}$ et la même quantité de protéines $p^{st} = \frac{k_R k_P}{\gamma_R \gamma_P}$. Cependant, comme le nombre d'ARN messagers est petit, de l'ordre de dix, ce modèle déterministe à variables continues ne peut représenter les réelles quantités de transcrits et de protéines au sein d'une cellule. De plus il ne peut expliquer les différences de niveau de protéines entre cellules observées expérimentalement. Nous devons donc pour cela considérer un modèle dans lequel $r(t)$ et $p(t)$ sont des processus stochastiques, pour rendre compte de la variabilité phénotypique observée. C'est ce modèle que nous présentons ici.

Les parties (i) à (iii) de l'annexe 2 présentent une analyse mathématique détaillée de chacune des étapes de ce modèle (transcription, traduction, dégradation).

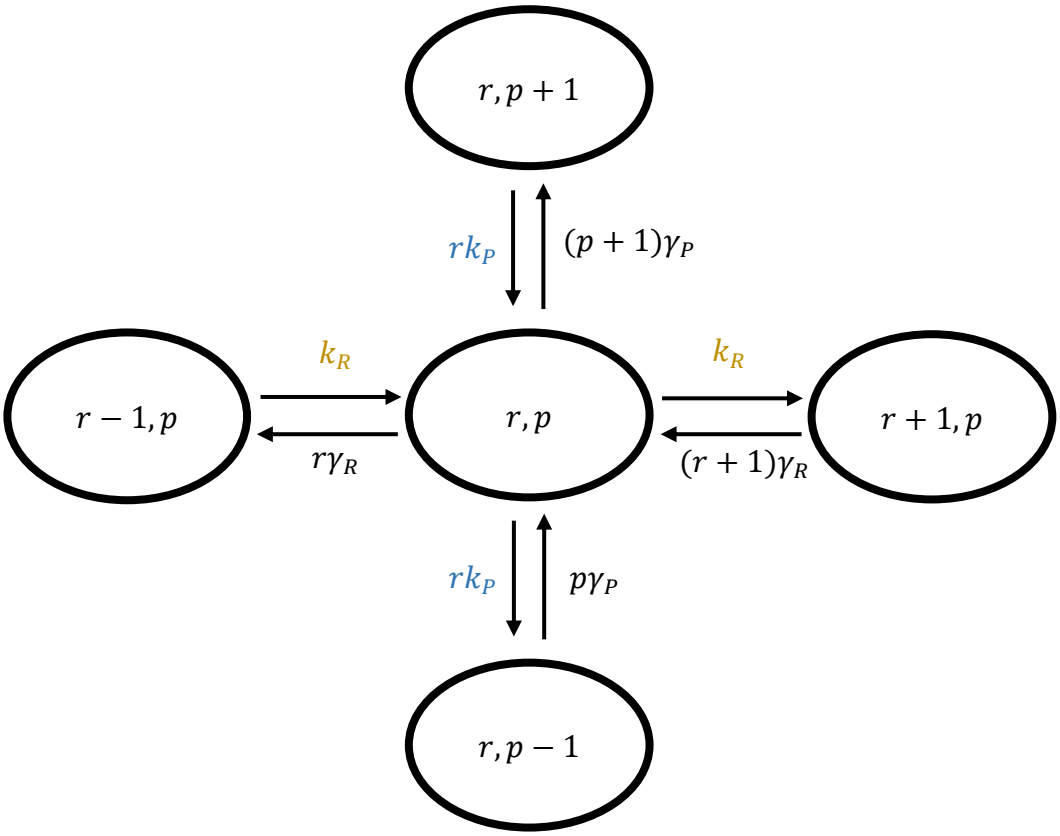


Figure 4 : Description de la stochasticité de l'expression génique par deux processus de vie et de mort Markoviens. Un état du système est caractérisé par la donnée du nombre d'ARN messagers r et du nombre de protéines p qui sont deux variables stochastiques. Ce schéma réactionnel permet notamment de déterminer l'équation maîtresse régissant l'évolution de la loi de probabilité jointe $P(r, p, t)$. En orange est indiquée la phase de transcription qui permet d'augmenter d'une unité le nombre de transcrits, en bleu la phase de traduction qui permet d'augmenter d'une unité le nombre de protéines et en noir les phases de dégradation des ARN messagers et des protéines qui diminuent leur nombre d'une unité.

Dans ce modèle, les nombres de copies d'ARN messagers et de protéines au cours du temps sont des processus stochastiques qui augmentent ou diminuent selon des temps d'attente distribués exponentiellement. Le nombre d'ARN messagers augmente d'une unité selon une loi exponentielle de paramètre k_R, et le nombre d'ARN messagers r synthétisés à un instant t donné satisfait un processus de Poisson. Le nombre d'ARN messager diminue également d'une unité selon une loi exponentielle de paramètre $r\gamma_R$; et le temps de vie pour un ARN messager est également distribué exponentiellement. De la même manière, le nombre de protéines augmente d'une unité selon une loi exponentielle de paramètre rk_P, et diminue d'une unité selon une loi exponentielle de paramètre $p\gamma_P$. Mis ensemble, ces quatre processus définissent deux processus markoviens

de naissance et de mort : le premier concernant la synthèse des ARN messagers (transcription) suivi de leur dégradation ; et le second la synthèse des protéines à partir des ARN messagers (traduction) et leur dégradation. Dans ce modèle, les fluctuations aléatoires du nombre d'ARN messager créent un environnement fluctuant pour la synthèse des protéines. On doit donc s'attendre à ce que les fluctuations de la quantité d'ARN se retrouvent au niveau des caractéristiques de la distribution du nombre de protéines.

III.1.2 Distribution stationnaire des nombres d'ARN messagers et de protéines :

Dans le cadre de ce modèle, les distributions du nombre d'ARNm et de protéines par cellule à l'état stationnaire peuvent être caractérisées. On peut en particulier démontrer (voir annexe 2, partie v) que le nombre de transcrits par cellule à un instant donné suit une loi de Poisson de paramètre $\lambda = k_R/\gamma_R$:

$$P(r) = \exp(-\lambda)\frac{(\lambda)^r}{r!}$$

La distribution du nombre de protéines est beaucoup plus difficile à calculer analytiquement. Différentes approches (voir annexe 2, partie vi) et approximations, notamment fondée sur un temps de vie des ARNm très inférieur au temps de vie des protéines, ont néanmoins permis des résolutions analytiques (Shahrezaei & Swain, 2008 ; Friedman et al, 2006) et montré que le nombre de protéines se distribue selon une loi Gamma de paramètre $a = k_R/\gamma_P$ et $b = k_P/\gamma_R$:

$$P(p) \cong \frac{p^{a-1}e^{-p/b}}{b^a\Gamma(a)}$$

De même, il est possible de calculer analytiquement le niveau d'expression moyen, le bruit d'expression (carré du coefficient de variation) et le facteur Fano (annexe 2 partie iv). On est alors capable de relier ces différentes quantités avec les différents paramètres biophysiques du modèle et de quantifier les effets relatifs de chaque étape, notamment la transcription et la traduction, sur les fluctuations du niveau de protéines.

III.1.3 Prédiction du modèle à deux niveaux sur l'effet de la transcription et de la traduction sur le bruit d'expression génique :

Les réactions biochimiques de transcription et de traduction sont toutes les deux sources de bruit. Cependant, afin de mieux comprendre comment la variabilité phénotypique est façonnée, il est important de comprendre l'importance relative de ces deux processus sur la génération du bruit.

Transcription et traduction sont des évènements séquentiels, le produit de l'un servant de support à l'autre. Le nombre d'ARN messagers étant en faible quantité par rapport aux nombres de protéines correspondantes, les fluctuations du nombre de messagers sont attendues pour être responsable d'une part importante des fluctuations au niveau protéique. Nous allons voir ce que prédit le modèle sur ces fluctuations.

- *Importance des fluctuations d'ARN messagers sur le bruit du niveau d'expression* :

Le bruit d'expression génique s'écrit d'après le modèle à deux niveaux selon :

$$\eta_p^2 = \frac{1}{\langle p \rangle} + \frac{1}{\langle r \rangle} \frac{\gamma_P}{\gamma_R + \gamma_P}$$

Dans cette équation, le premier terme représente des fluctuations Poissonniennes du nombre de protéines, qui sont dû aux processus de synthèse/dégradation des protéines (Paulsson,2005). En effet, si la synthèse des protéines se faisait à taux constant, alors le bruit d'expression serait donné par (Kauffman & van Oudenaarden, 2007) :

$$\eta_p^2 = \frac{1}{\langle p \rangle}$$

Le second terme quant à lui tient compte des fluctuations stochastiques du taux de synthèse des protéines (mk_P) générés par les fluctuations stochastiques du nombre d'ARN messagers. Ce terme se présente comme le produit des fluctuations Poissonniennes du nombre d'ARN messagers, soit η_m^2, par un terme de « lissage temporel » qui indique le fait que le niveau de protéines ne peut pas s'ajuster immédiatement aux changements dans le taux de synthèse (Paulsson, 2005 ; Pedraza & Paulsson, 2008). Autrement dit, ce terme quantifie les différences d'échelles de temps entre les fluctuations des deux processus, et ainsi leur capacité à se propager de l'un à l'autre. Par exemple, avec $\gamma_R \gg \gamma_P$, les fluctuations du nombre d'ARN messagers sont « rapides » par rapport au temps de vie des protéines, et l'effet de ces fluctuations en sont d'autant plus amorties puisque $\frac{\gamma_P}{\gamma_R+\gamma_P} \cong \frac{\gamma_P}{\gamma_R} \ll 1$ (Coulon, 2010). D'après cette équation, les fluctuations du niveau de protéine sont plus importantes que celles engendrées par une variable aléatoire ayant la même moyenne et suivant une loi de Poisson. A nombre de protéines constant, plus le nombre d'ARN messager est important, moins les fluctuations du niveau de ces molécules η_m^2 seront importantes, et ainsi une variabilité phénotypique η_p^2 moindre. A l'inverse, plus le nombre d'ARN messager est faible, plus la variabilité phénotypique η_p^2 est importante. Autrement

dit, pour un niveau d'expression donné, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit. Nous pouvons réécrire l'équation précédente pour montrer explicitement la dépendance de la variabilité phénotypique avec les taux de transcription et de traduction :

$$\eta_p^2 \cong \frac{\gamma_P \gamma_R}{k_R k_P} + \frac{\gamma_P}{k_R}$$

Ainsi, un taux de traduction élevé ne réduit que le premier terme qui correspond au bruit Poissonnien du nombre de protéines, tandis qu'un taux de transcription élevé réduit aussi celle générée par les fluctuations Poissonniennes du niveau d'ARN messager (Figure 5-C).

- *Facteur Fano et « burst » traductionnel :*

Les mêmes conclusions auraient également pu être faites à partir du facteur Fano. D'après l'équation précédente, l'augmentation du taux de transcription comme celle du taux de traduction réduit le bruit (Figure 5-C). Le facteur Fano s'écrit quant à lui :

$$\frac{\sigma_p^2}{\langle p \rangle} = 1 + \frac{k_P}{(\gamma_R + \gamma_P)} \cong 1 + \frac{k_P}{\gamma_R}$$

Il présente donc des dépendances qualitativement différentes selon la façon dont l'abondance de protéines exprimées varie : il augmente linéairement avec l'abondance moyenne des protéines lorsque l'efficacité de traduction est augmentée, mais reste constant lorsque le taux de transcription est augmenté (Figure 5-D). L'intérêt du facteur Fano est ainsi expérimental puisqu'il relie une quantité mesurable, le facteur Fano, à un paramètre biologique, l'efficacité de traduction, qui peut être modulé en introduisant des mutations dans la séquence TIR (Translation Initiation Region, contenant le RBS). L'indépendance de cette quantité avec l'efficacité de transcription peut également être testée en jouant sur la force du promoteur d'un gène d'intérêt (Ozbudak et al, 2002 ; Blake et al, 2003 ; Raser & O'Shea, 2004). Le facteur Fano permet donc de tester le modèle stochastique à deux niveaux sur ces conclusions. Nous discuterons de ces travaux expérimentaux dans une prochaine partie. Il est important de noter que l'augmentation du facteur Fano n'implique pas que la variabilité relative (dont la mesure est donnée par le coefficient de variation) ait également augmentée. Pour deux gènes caractérisés par des facteurs Fano faible et élevé, on peut conclure que le gène ayant le facteur Fano plus élevé présente une variabilité relative accrue seulement lorsque les deux gènes sont exprimés à des abondances similaires. Le temps de vie des ARN messagers étant relativement court par rapport à celui des

protéines, Thattai et van Oudenaarden proposent l'hypothèse que la traduction se produit de manière rare et brève. On a donc ainsi une courte période d'intense activité traductionnelle qui conduit à l'injection d'un « paquet » ou « burst » de protéines dans le cytoplasme (Figure 6). On peut montrer (annexe 2 partie vi) que le nombre de protéines contenues dans une salve est distribué selon une loi géométrique de paramètre $b/(1+b)$ de valeur moyenne b, avec $b = k_p/\gamma_R$. Ce mécanisme est appelé « *burst traductionnel* ». Le facteur Fano se réécrit :

$$\frac{\sigma_p^2}{\langle p \rangle} \cong 1 + b$$

Il est important de noter que le terme de burst disparaît complètement si la synthèse et la dégradation des transcrits est déterministe et reste inchangé si la traduction est déterministe (Paulsson, 2004). Ainsi, contrairement à ce que laisse entendre le terme de « burst traductionnel », ce dernier ne provient pas de la nature stochastique de la traduction mais des évènements stochastiques de synthèse/dégradation des transcrits. Ainsi, plus la traduction sera efficace, au point d'avoir $b \gg 1$, soit $k_P \gg \gamma_R$, plus on s'écartera du comportement Poissonnien de synthèse/dégradation des protéines et la variabilité observée sera majoritairement due aux événements de synthèse/dégradation des transcrits. On peut estimer qu'à partir de deux protéines par transcrits, les fluctuations du nombre d'ARN messagers dominent le bruit intrinsèque.

Pour conclure sur ces résultats, nous avons deux quantités qui permettent de caractériser les sources de la variabilité phénotypique observée :

- Le coefficient de variation quantifie la variabilité phénotypique et montre que les fluctuations du nombre d'ARN messagers ont un rôle important dans le bruit d'expression génique. D'autre part, le taux de transcription a un impact plus fort sur la génération du bruit d'expression que le taux de traduction.

- Le facteur Fano ne représente pas directement la variabilité phénotypique mais permet de quantifier par le terme de « *burst traductionnel* » les contributions relatives des effets des évènements de synthèse et dégradation des ARN messagers et des fluctuations dues à la synthèse/dégradation au niveau protéique sur la variabilité phénotypique.

Pour conclure, le modèle à deux niveaux prédit que pour un niveau d'expression donné, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit.

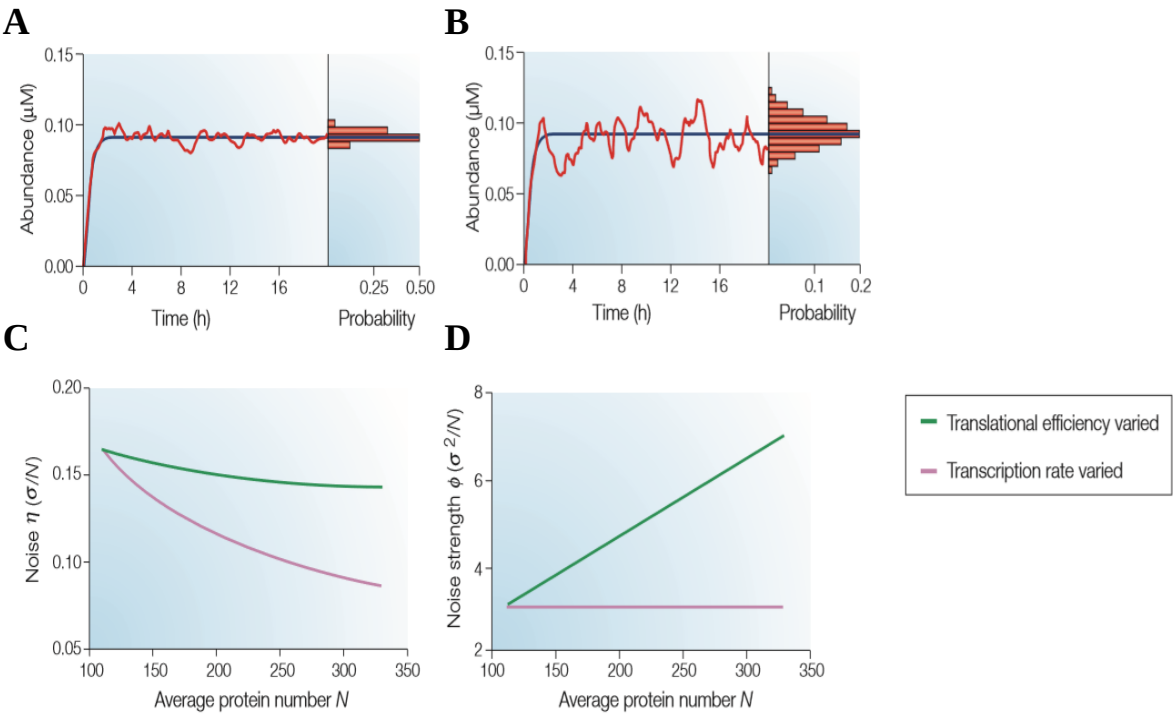


Figure 5 : Influence de la transcription et de la traduction sur le bruit d'expression génique prédit par le modèle à deux niveaux (d'après Kaern et al, 2005). **A-B)** Evolution temporelle de la concentration de protéines résultant de simulations déterministes (courbes bleues) et stochastiques (courbes rouges) du modèle à deux niveaux de l'expression génique. Les histogrammes représentent les distributions de la concentration en protéine. **A)** Fluctuations du nombre de protéines dans le cas où l'on a un nombre élevé de copies d'ARN messagers et de protéines (~3000 et ~10000 respectivement). Les taux de transcriptions k_R et de traductions k_P ont pour valeurs respectives 5 molécules par min et 0.2 molécules par minute. **B)** Fluctuations du nombre de protéines dans le cas où le taux de transcription est diminué de 100 fois, et le taux de traduction est augmenté de 100 fois pour conserver le nombre moyen de protéines constant : k_R=0.05 unités par min et k_P=20 unités par min ; auxquels correspondent un faible nombre moyen d'ARN messagers et un nombre élevé de protéines, respectivement ~30 et ~10000. Bien qu'au niveau protéique les systèmes **A)** et **B)** aient la même taille, les fluctuations sont plus importantes dans le cas **B)** où l'efficacité de transcription a été diminué et l'efficacité de traduction augmentée. Ainsi, pour un niveau d'expression donné, un gène faiblement transcrit et fortement traduit (cas **A**) génèrera plus de bruit qu'un gène fortement transcrit et faiblement traduit (cas **B**). **C-D)** Prédictions théoriques de l'influence des efficacités de transcription et de traduction sur le bruit (ou coefficient de variation) **C**) et le facteur Fano **D)**. L'efficacité de transcription est augmentée avec un taux constant de traduction, ce qui conduit à une augmentation du nombre moyen de protéines. De la même manière, l'efficacité de traduction est augmentée en laissant le taux de transcription inchangé.

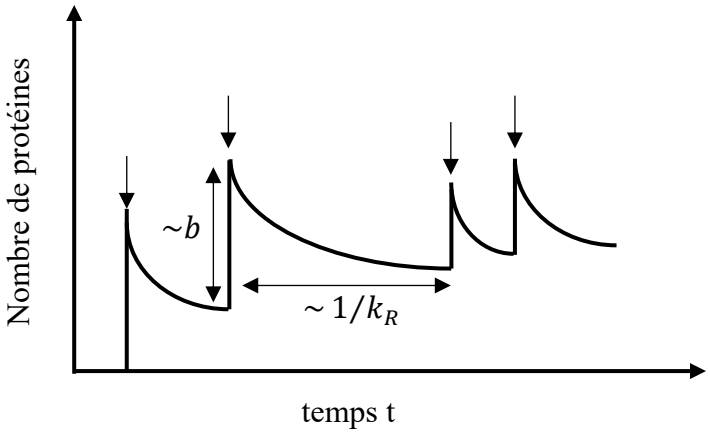


Figure 6 : Représentation schématique du burst traductionnel. Chaque flèche représente la synthèse d'un transcrit. Comme les ARN messagers sont beaucoup plus instable que les protéines, sur une échelle de temps calée sur la durée de vie moyenne d'une protéine, tout se passe comme si à chaque fois qu'un ARN messager est synthétisé ce dernier injecte dans la cellule un « burst » de protéines dans le cytoplasme avant sa dégradation qui survient relativement vite. La taille de chaque burst est distribuée géométriquement et a pour valeur moyenne $b = k_P/\gamma_R$. La durée entre chaque burst est distribuée exponentiellement et est égale au temps d'attente entre deux évènements de synthèse d'un ARN messager. La durée moyenne entre chaque burst est donnée par $1/k_R$. Entre chaque période de synthèse protéique, le nombre de protéines décroît du fait de leur dégradation. D'après Thattai et van Oudenaarden, 2001.

III.2 Modèle stochastique à trois niveaux de l'expression génique :

III.2.1 Présentation du modèle :

Historiquement, le modèle à deux niveaux a été proposé pour caractériser le bruit intrinsèque de l'expression génique chez les procaryotes (Thattai & van Oudenaarden, 2001 ; Swain et al, 2002). A côté de celui-ci, plusieurs modèles ont été développé pour caractériser le bruit intrinsèque de l'expression génique dans les cellules eucaryotes (Peccoud & Ycart, 1995 ; Paulsson, 2005 ; Blake et al, 2003 ; Raser & O'Shea, 2004). La motivation de développer d'autres modèles provient essentiellement de la plus grande complexité de l'expression génique chez les eucaryotes. Ces modèles se basent sur le même principe que le modèle à deux niveaux, à savoir que chaque réaction individuelle est un processus stochastique caractérisé par une variable aléatoire qui incrémente ou décrémente d'une unité avec des temps d'attente distribués exponentiellement. Les modèles eucaryotes se démarquent du modèle à deux niveaux par l'ajout de plusieurs étapes dans la phase de transcription, et l'ajout d'étapes de conformations du gène. Par exemple, Blake et al. introduisent des étapes de remodelage de la chromatine, d'assemblage séquentiel de la machinerie de transcription et de production en

rafales d'ARN messager due à la réinitiation de la transcription (Blake et al, 2003). La réinitiation de la transcription est un processus endémique aux cellules eucaryotes qui consiste au maintien des protéines de transcription à un gène activement transcrit. Ce mécanisme contribue au maintien de l'état transcriptionnel actif et augmente le taux des cycles de transcription suivant par rapport au cycle initial (Dieci et al, 2013) De la même manière, Raser et O'Shea proposent d'utiliser un modèle qui rajoute uniquement une étape d'activation/inactivation du promoteur d'un gène d'intérêt avant la phase de synthèse des ARN messagers (Raser & O'Shea, 2004). C'est ce modèle que nous présenterons par la suite et qui se nomme modèle à trois niveaux (Shahrezaei & Swain, 2008) et qui fut décrit analytiquement pour la première fois par Peccoud et Ycart puis par Paulsson (Peccoud & Ycart, 1995 ; Paulsson, 2005). L'intérêt de ce modèle est qu'il permet d'obtenir un modèle standard à la fois applicable pour les cellules eucaryotes et les cellules procaryotes. En effet, chez les eucaryotes, les transitions entre les états actifs et inactifs du promoteur peuvent s'expliquer comme des changements de structure de la chromatine, des attachements et détachement de protéines nécessaires à la transcription comme dans le cas de la réinitiation de la transcription. D'autre part, ce modèle permet de modéliser un promoteur régulé au moyen de répresseurs ou d'activateurs qui sont des mécanismes que l'on retrouve à la fois chez les cellules eucaryotes mais aussi chez les cellules procaryotes. Les évènements stochastiques d'attachement et de détachement du facteur de transcription (répresseur ou activateur) deviennent des sources de bruit intrinsèque supplémentaire tandis que les fluctuations dans leur nombre représentent une source de bruit extrinsèque. Comme pour le modèle à deux niveaux, le modèle à trois niveaux ne cherche qu'à quantifier les sources de bruit intrinsèque, et nous considèrerons donc par la suite que la concentration en facteur de transcription est constante. Dans le modèle à deux niveaux, le caractère aléatoire du nombre d'ARN messagers crée un environnement stochastique pour la synthèse des protéines. Dans le modèle à trois niveaux, du fait du caractère stochastique de l'état du promoteur, à savoir occupé par un répresseur/activateur ou non, la synthèse des messagers se fera également dans un environnement stochastique. Ce modèle introduit ainsi la notion de « *burst transcriptionnel* ». Ce modèle se différencie du modèle à deux niveaux par l'ajout en amont de la transcription d'une étape d'activation/inactivation du promoteur (Figure 7). On note k_+ et k_- respectivement le taux de transition entre l'état inactif et actif et le taux de transition entre l'état actif et inactif.

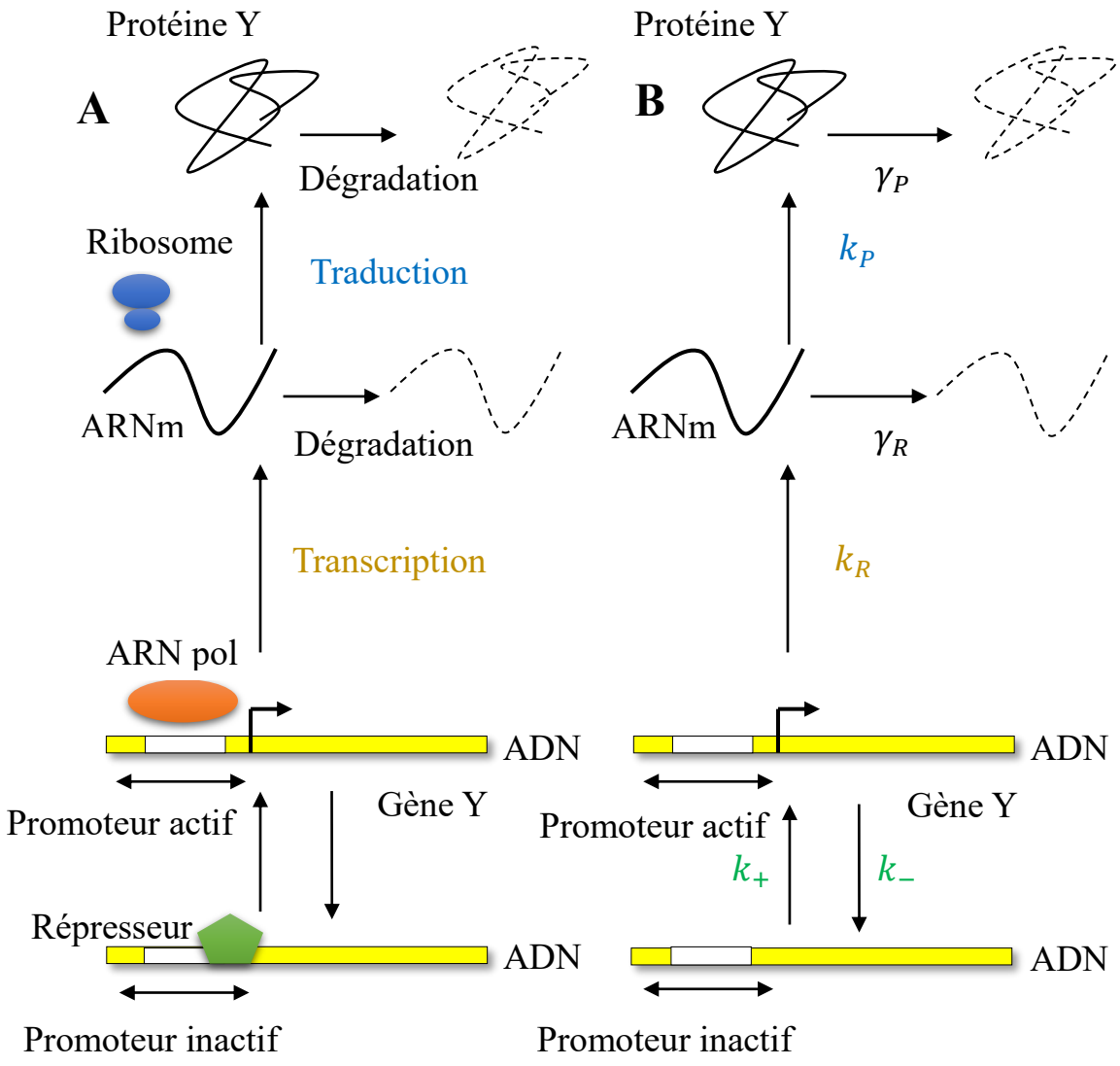


Figure 7 : Représentations du modèle à trois niveaux de l'expression génique. **A)** Mécanismes biologiques de l'expression génique pris en compte dans le modèle : régulation génétique, transcription, traduction et dégradations des messagers et des protéines. Les médiateurs de ces étapes biochimiques sont les répresseurs, les ARN polymérases et les ribosomes. Les dégradations des messagers et des protéines sont assurées respectivement par les ribonucléases et les protéases. **B)** Interprétation mathématiques des différentes étapes biochimiques présentées en A). Chaque étape est un processus stochastique qui incrémente ou décrémente des temps d'attente distribués exponentiellement.

III.2.2 Expression analytique du nombre moyen de protéines et du bruit d'expression génique d'après le modèle à trois niveaux :

On s'intéresse tout d'abord à la dynamique des ARN messagers, c'est-à-dire au processus stochastique $\{g, r\}$ dont une représentation est donnée dans l'annexe 3.

On peut déduire le niveau moyen d'ARN messager $\langle r \rangle$ ainsi que le bruit associé η_r^2 à l'état stationnaire (Paulsson, 2005) à partir du bruit au niveau des promoteurs $\eta_g^2 = \frac{1}{g^{tot}} \frac{k_-}{k_+}$ où g^{tot} est le nombre de copies du gène d'intérêt :

$$\langle r \rangle = \frac{k_R}{\gamma_R} \langle g \rangle = \frac{k_R}{\gamma_R} g^{tot} P_+$$

$$\eta_r^2 = \frac{1}{\langle r \rangle} + \eta_g^2 \frac{\tau_g}{\tau_R + \tau_g}$$

$$\frac{\sigma_r^2}{\langle r \rangle} = 1 + (1 - P_+) \frac{k_R \tau_R}{\tau_g^{-1} \tau_R + 1}$$

Avec $\tau_g = (k_+ + k_-)^{-1}$ une échelle de temps caractéristique des changements dans l'activité des promoteurs, $\tau_R = \gamma_R^{-1}$ le temps de vie moyen d'une molécule d'ARN messager, $P_+ = \frac{k_+}{k_+ + k_-}$ la probabilité pour que le système se trouve dans l'état actif et $\eta_g^2 = \frac{1 - P_+}{\langle g \rangle}$. On caractérise la « rapidité » de la dynamique des promoteurs en comparant τ_g avec τ_R. Le bruit se décompose en deux parties et s'interprète de la même manière que le bruit au niveau protéique dans le modèle à deux niveaux. Le premier terme correspond aux fluctuations Poissonniennes engendrées par le processus de synthèse/dégradation des transcrits, tandis que le second terme représente les fluctuations au niveau de l'activité des gènes qui ont été transmises avec un terme de « lissage temporel » (Coulon, 2010). Cette décomposition permet de déterminer la contribution relative des fluctuations d'activité des promoteurs dans le bruit d'expression (Annexe 3). Ainsi, pour un niveau d'ARN messager donné, un gène faiblement transcrit et souvent actif sera moins bruité qu'un gène fortement transcrit et rarement actif (Figure 8). Aussi, pour un niveau d'ARN messager donné, plus la dynamique d'activité d'un gène est lente plus le niveau d'ARN messager sera bruité.

Ce modèle permet de décrire un phénomène de « *burst transcriptionnel* » à travers le second terme du facteur Fano. Ce terme de burst transcriptionnel permet d'expliquer des écarts au comportement Poissonnien. Ce sera notamment le cas pour des gènes ayant une dynamique lente de l'activité des promoteurs $\tau_g \gg \tau_R$ avec des périodes d'activité sensiblement plus courtes que les périodes d'inactivité ($P_+ \ll 1$), et par un taux de transcription élevé (voir annexe 3).

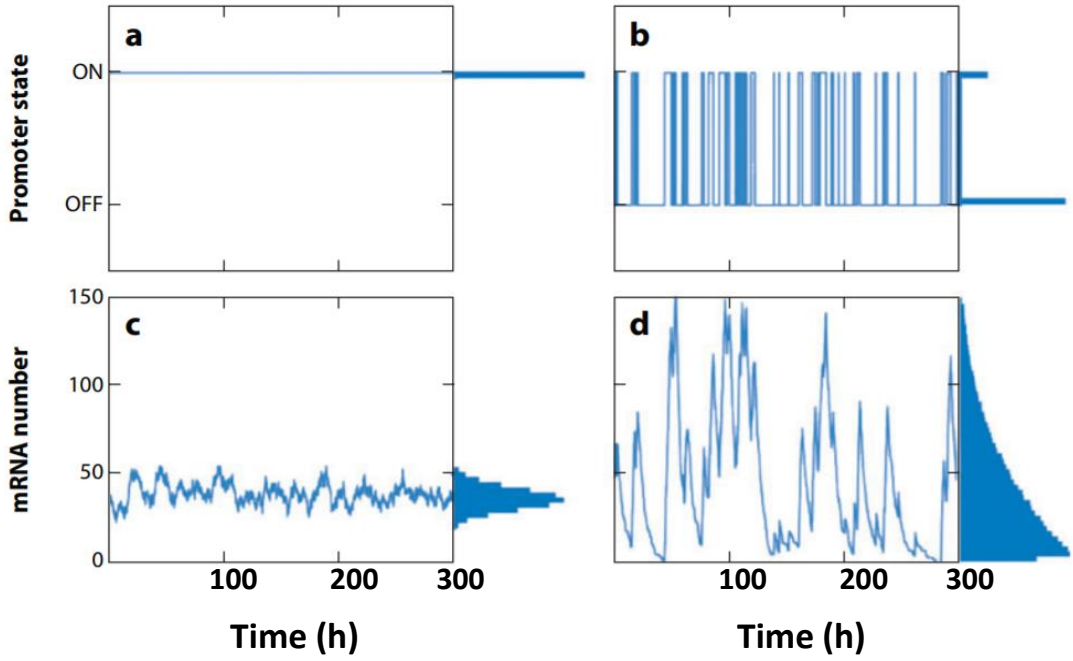


Figure 8 : Influence de l'activité d'un promoteur sur la variabilité observée au niveau des ARN messagers (Raj & van Oudenaarden, 2009). **a)** Dynamique d'un promoteur qui est toujours dans l'état actif. **b)** Dynamique d'un promoteur qui alterne des périodes d'activité et d'inactivité. Cette dynamique peut être rapide par rapport aux temps caractéristiques des fluctuations du nombre d'ARN messager ou lente. Dans tous les cas, le temps passé dans l'état actif est sensiblement plus court que le temps passé dans l'état inactif. **c)** Dynamique du nombre d'ARN messager dans le cas où le promoteur est toujours actif. Dans ce cas la distribution du nombre de transcrit est Poissonienne (équation 33 avec $P_+ = 1$). **d)** Dynamique du nombre d'ARN messagers dans le cas où le promoteur est libre de transiter entre les états actifs et inactifs. Parce que le temps de séjour dans l'état actif est beaucoup plus court que celui passé dans l'état inactif, les fluctuations engendrées au niveau de l'activité du promoteur ont une contribution importante sur la variabilité observée au niveau des transcrits. Cette dernière s'écarte d'un comportement Poissonien et les fluctuations sont plus importantes que dans **c)** malgré un nombre moyen d'ARN messager équivalent. Pour assurer l'égalité des quantités moyennes d'ARN messager, le taux de transcription a été augmenté en **d)** par rapport à **c)**. Les histogrammes à droite des dynamiques temporelles en **c)** et **d)** représentent l'allure des distributions du nombre d'ARN messagers. Ces distributions montrent que pour une quantité d'ARN messager donné, un gène souvent actif et faiblement transcrit sera moins bruité qu'un gène fortement transcrit et rarement actif.

On s'intéresse maintenant à la dynamique des protéines, et donc au processus stochastique Markovien de synthèse protéique dans sa globalité $\{g, r, p\}$ dont une représentation est donnée par la figure 9. On peut montrer que le niveau de bruit d'expression et le facteur Fano s'écrivent :

$$\eta_p^2 = \frac{1}{\langle p \rangle}\left(1 + k_P\tau_P\frac{\tau_R}{\tau_R+\tau_P} + k_P k_R \tau_R \tau_P (1-P_+)\frac{\tau_R}{\tau_R+\tau_P}\frac{\tau_g}{\tau_g+\tau_P}\frac{\tau_g+\tau_P+\tau_g\tau_P/\tau_R}{\tau_g+\tau_R}\right)$$

$$\frac{\sigma_P^2}{\langle p \rangle} = 1 + k_P \tau_P \frac{\tau_R}{\tau_R + \tau_P} + k_P k_R \tau_R \tau_P (1 - P_+) \frac{\tau_R}{\tau_R + \tau_P} \frac{\tau_g}{\tau_g + \tau_P} \frac{\tau_g + \tau_P + \tau_g \tau_P / \tau_R}{\tau_g + \tau_R}$$

Le premier terme du bruit correspond aux fluctuations Poissonniennes dues aux évènements stochastiques de synthèse/dégradation des protéines. Les deux termes suivants correspondent aux fluctuations du nombre d'ARN messagers qui ont été transmises au niveau protéique. Le premier de ces termes reprend les fluctuations Poissoniennes des évènements spontanés de naissance et de mort des transcrits, tandis que le second terme reprend les fluctuations dans l'activité des gènes. Le facteur Fano reprend les trois sources de bruit. D'une manière générale, pour que les fluctuations d'un niveau donné aient une contribution majoritaire au(x) niveau(x) en aval il faut que ses fluctuations soient importantes et/ou que ses fluctuations soient efficacement transmises. Il faut également que les contributions des autres niveaux lui soient inférieurs. Pour un gène ayant le troisième prédominant sur les deux autres, le bruit d'expression est essentiellement piloté par les évènements d'activation et d'inactivation des gènes. On parle de « *burst transcriptionnel* », la synthèse protéique se produit alors en « rafales » dont l'origine est au niveau de la dynamique des promoteurs. Ce sera notamment le cas si $P_+ \ll 1$, donc pour des promoteurs essentiellement dans l'état inactif et pour lesquels $\tau_g \gg \tau_P$. Pour un gène ayant le second terme prédominant, le bruit d'expression génique est essentiellement dû aux évènements de synthèse et de dégradation des ARN messagers. On se retrouve dans une situation similaire à celle du modèle à deux niveaux. On parle de « *burst traductionnel* », la synthèse protéique se produit par salves dont l'origine se situe au niveau des évènements stochastiques de synthèse et de dégradation des ARN messagers. Ce sera notamment le cas si $P_+ \sim 1$, (gène presque toujours actif) ou $\tau_g \ll \tau_P$ et avec une traduction efficace comparativement à la dégradation des ARN messagers. L'ensemble des contributions des sources de bruit intrinsèque sur le bruit d'expression est résumé sur la figure 10.

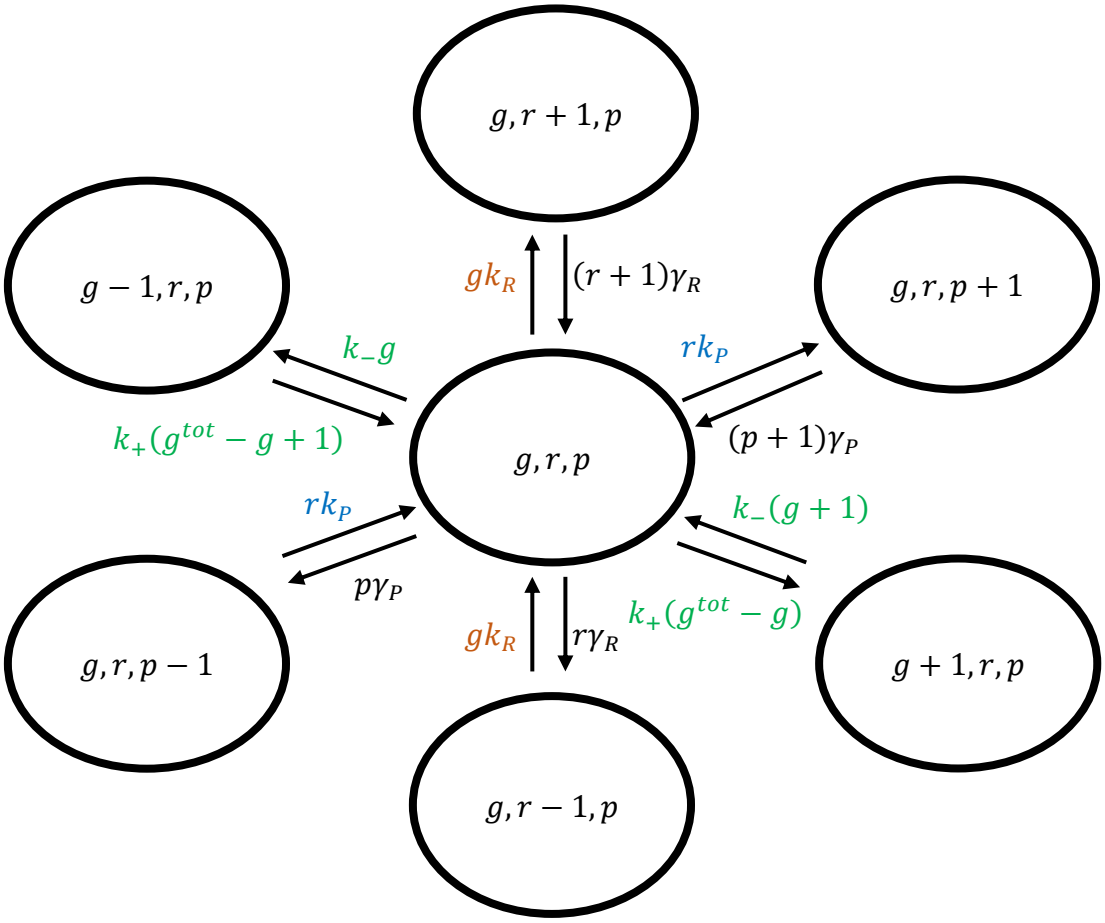
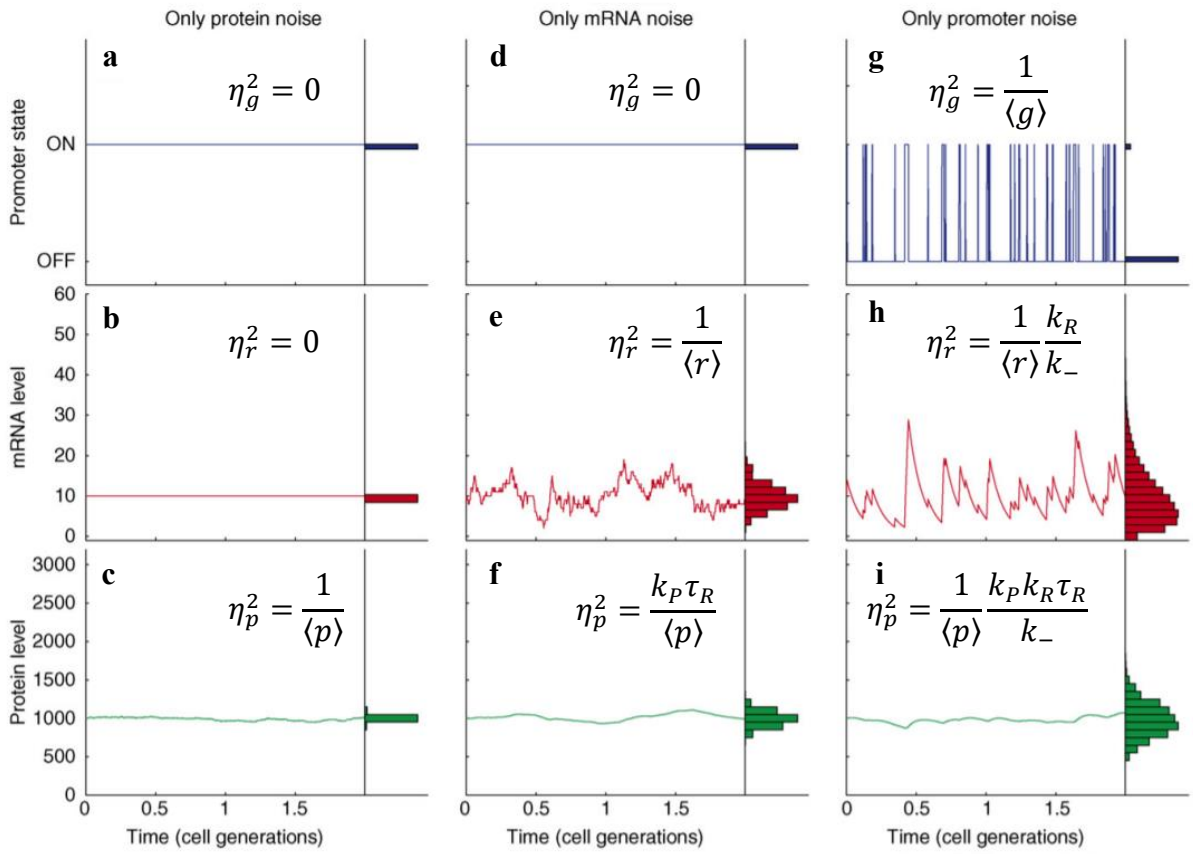


Figure 9 : Illustration du processus Markovien $\{\boldsymbol{g}, \boldsymbol{r}, \boldsymbol{p}\}$. En vert sont indiquées les transitions entre états actifs et inactifs des promoteurs, en orange la synthèse des ARN messagers (transcription) et en bleu la synthèse des protéines (traduction). En noir sont indiqués les étapes de dégradations des transcrits et des protéines.



<u>Figure 10 :</u> Comparaison des trois sources possibles de bruit (résultats de simulation numérique). D'après Kaufmann & van Oudenaarden, 2007. La première ligne représente la dynamique du promoteur (**a,d,g** bleu), la seconde ligne la dynamique du nombre d'ARN messagers (**b,e,h** rouge) et la troisième ligne la dynamique du nombre de protéines (**c,f,i** vert). La première colonne (**a,b,c**) décrit la situation dans laquelle seul les évènements de synthèse/dégradation des protéines sont stochastiques, les autres étapes étant déterministes. Dans ce cas, on obtient des fluctuations Poissoniennes du nombre de protéines. La seconde colonne (**d,e,f**) correspond au cas où seul les évènements de synthèse/dégradation des ARN messagers sont stochastiques (burst de traduction). Le processus stochastique de vie et de mort des transcrits donne des fluctuations Poissoniennes au niveau des messagers. Ces fluctuations se propagent au niveau protéique par le mécanisme de burst traductionnel et confèrent une variabilité phénotypique qui est inversement proportionnelle au nombre moyen de protéines. La constante de proportionnalité est donnée par la taille moyenne d'un burst de protéines. La troisième colonne (**g,h,i**) décrit la situation dans laquelle seul la dynamique d'activation/inactivation du promoteur est stochastique (burst de transcription). Dans ce cas le promoteur n'est plus toujours actif mais oscille aléatoirement entre des périodes brèves d'activité et des périodes plus longues d'inactivité, soit $k_- \gg k_+$. Cette stochasticité se transmet au niveau des messagers par le mécanisme de burst transcriptionnel (on a ici $\tau_g = k_-^{-1} \ll \tau_R$). Les courtes périodes d'activité du promoteur conduisent par la suite à un burst de protéines. Là encore, la variabilité phénotypique est inversement proportionnelle au nombre moyen de protéine, et la constante de proportionnalité est alors égale au nombre moyen de protéines produites sur une période active du promoteur. Pour chaque simulation, le nombre d'ARN messager est constant (10 molécules d'ARN messagers) ainsi que le nombre de protéines (1000 molécules). Les notations sont les mêmes que dans le texte principal.

III.2.3 Distribution analytique des nombres d'ARN messagers et de protéines :

Comme pour le modèle à deux niveaux, Shahrezaei & Swain (2008) proposent une méthode analytique pour déterminer les distributions à l'état stationnaire du nombre de protéines. Cette méthode se base sur le fait que les ARN messagers sont beaucoup plus instables que les protéines. A la différence du modèle développé par Paulsson (2005) leur modèle considère un seul promoteur qui peut commuter entre deux états (actif et inactif). La distribution trouvée pour le nombre de protéine est bien plus compliquée que la distribution binomiale négative ou Gamma obtenue dans le modèle à deux niveaux et n'est valable que dans le cas où $\frac{\gamma_R}{\gamma_P} \gg 1$ (Figure 11). Cette distribution peut générer une distribution bimodale avec un pic correspondant à zéro nombre de protéines (le promoteur est dans l'état inactif) et un autre pour un nombre différent de zéro dans le cas où les périodes d'activation/inactivation sont très longues (Figure 11C).

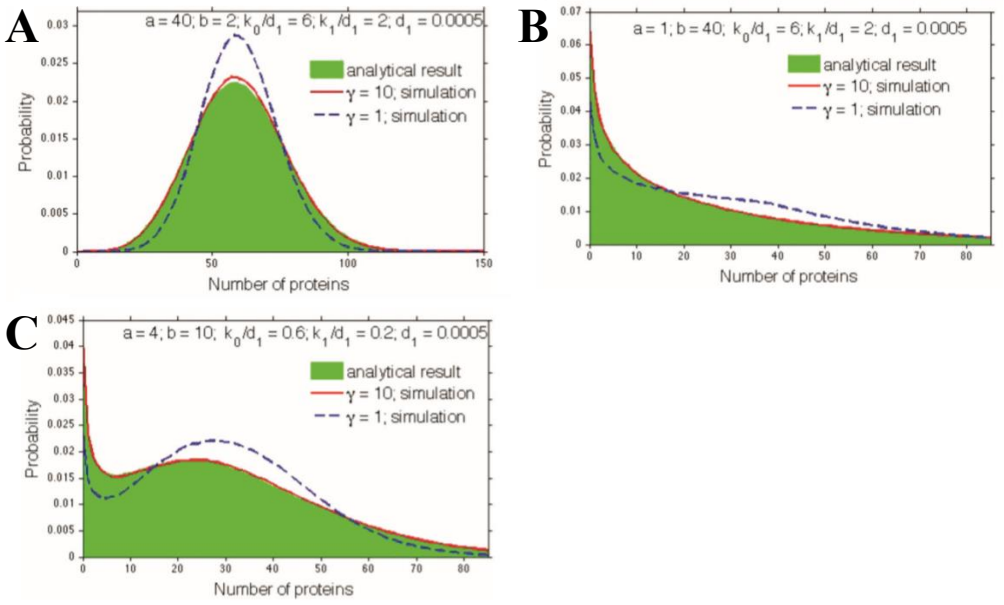


Figure 11 : Prédictions théoriques et simulations du modèle à trois niveaux de l'expression génique (Shahrezaei & Swain, 2008) sur la distribution du nombre de protéines. La courbe en verte représente le résultat analytique. Les autres courbes représentent les simulations stochastiques pour $\gamma = \gamma_R/\gamma_P \gg 1$ (courbes rouges) et $\gamma = 1$ (courbes bleues). Les simulations et la distribution analytique s'ajustent d'autant plus que le rapport des temps de vie γ est grand. $a = k_R/\gamma_P$; $b = k_P/\gamma_R$; $d_1 = \gamma_P$; $k_0 = k_+$et $k_1 = k_-$. **A)** Les paramètres de la simulation sont : $a = 40, b = 2$, $k_0 = 6d_1$, $k_1 = 2d_1$, $d_1 = 0.0005s^{-1}$. Le nombre moyen de protéines correspondant est $\langle p \rangle = 60$. **B)** $a = 1, b = 40$, $k_0 = 6d_1$, $k_1 = 2d_1$, $d_1 = 0.0005s^{-1}$, $\langle p \rangle = 30$. **C)** $a = 4, b = 10$, $k_0 = 0{,}6d_1$, $k_1 = 0{,}2d_1$, $d_1 = 0.0005s^{-1}$, $\langle p \rangle = 30$. **B-C)** Le fait de ralentir la dynamique des promoteurs fait apparaître un comportement bimodal de la distribution de protéines.

III.3 Limites des deux modèles de l'expression génique :

Une limitation importante des modèles stochastiques de l'expression génique que nous venons d'aborder est l'absence des sources de bruit extrinsèque. De nombreux mécanismes sont responsables de la production de bruit extrinsèque, notamment la réplication du chromosome au cours du cycle cellulaire, qui au sein d'une cellule peut induire des fluctuations périodiques du nombre de protéines et expliquer une part de l'hétérogénéité mesurée entre différentes cellules d'une même population qui ne seraient pas toutes au même stade du cycle cellulaire (Keren et al, 2015). De même, les fluctuations dans les quantités d'ARN polymérases, de ribosomes, de ribonucléases et de protéases sont attendues pour être une source de bruit extrinsèque importante (Yang, 2014). Nous allons présenter ici comment il est possible d'introduire ces sources dans les modèles stochastiques présentés précédemment. Nous verrons également l'impact de la division cellulaire, qui entraine une répartition stochastique des molécules entre les deux cellules filles. Pour finir, les modèles que nous avons présentés sont basés sur une hypothèse mathématique forte qui consiste à dire que les évènements pris en compte se produisent avec des intervalles de temps qui sont distribués exponentiellement. Nous aborderons également la pertinence de cette hypothèse. Pour finir nous aborderons l'effet de l'autorégulation sur le bruit d'expression

III.3.1 Prédiction des effets du bruit extrinsèque sur la variabilité phénotypique :

Les modèles que nous avons abordés ici partagent les mêmes hypothèses. Une des hypothèses consiste à considérer les différents taux intervenant dans les évènements de synthèse (k_R, k_P), de dégradations (γ_R, γ_P) et d'activation/inactivation des promoteurs (k_-, k_+) comme des constantes. Cela revient à considérer les quantités d'ARN polymérases, ribosomes, ribonucléases, protéases, et répresseurs comme constantes mais également à considérer l'environnement cellulaire comme invariable. Les fluctuations de ces composés et le cycle cellulaire ne sont donc pas pris en compte par ces modèles. Rosenfeld et al ont montré expérimentalement que le bruit extrinsèque engendre des fluctuations « lentes », c'est-à-dire qu'elles vont persister sur des échelles de temps comparable aux cycles cellulaires, tandis que les fluctuations intrinsèques vont être beaucoup plus rapide (Rosenfeld et al, 2005). Ainsi, pour une cellule donnée, l'écart entre le niveau de protéines et la valeur moyenne engendré par les sources de bruit extrinsèques va demeurer sur une échelle de temps proche du cycle cellulaire et ainsi conférer à la cellule une « mémoire ». Si on considère différentes cellules, les variables extrinsèques de chaque cellule seront différentes

et leurs lentes fluctuations maintiendront ces hétérogénéités. A partir de ce résultat, Taniguchi et ces collègues proposent d'injecter les sources de bruit extrinsèque dans le modèle à deux niveaux en permettant aux paramètres biophysiques k_R, k_P, γ_R et γ_P de fluctuer au cours du temps. Ces paramètres biophysiques qui deviennent alors des variables aléatoires sont regroupés dans les variables $a = k_R/\gamma_P$ et $b = k_P/\gamma_R$ qui représentent respectivement le nombre moyen d'ARN messagers produits durant le temps de vie d'une protéine et le nombre moyens de protéines synthétisé à partir d'un transcrit. Ces quantités varient donc lentement dans une cellule et sont différentes d'une cellule à une autre, les écarts pouvant être considérés comme quasi-statique. En supposant de plus que les variables a et b sont indépendantes, Taniguchi et al. obtiennent l'expression suivante pour le bruit total (Taniguchi et al, 2010) :

$$\eta_p^2 = \frac{1 + \langle b \rangle}{\langle p \rangle} + \eta_b^2 \frac{\langle b \rangle}{\langle p \rangle} + \eta_a^2 + \eta_a^2 \eta_b^2 + \eta_b^2$$

Où $\eta_a^2 = \sigma_a^2/\langle a \rangle^2$ et $\eta_b^2 = \sigma_b^2/\langle b \rangle^2$ sont les bruits des variables a et b Cette équation du bruit peut être séparée en deux parties. Une première partie qui représente la contribution du bruit intrinsèque tel qu'elle est prédite par le modèle à deux niveaux et qui présente une dépendance en $\langle p \rangle^{-1}$:

$$\eta_{int}^2 = \frac{1 + \langle b \rangle}{\langle p \rangle}$$

Et une seconde partie qui dépend des termes η_a^2 et η_b^2 et qui représente la contribution des sources de bruit extrinsèque :

$$\eta_{ext}^2 = \eta_b^2 \frac{\langle b \rangle}{\langle p \rangle} + \eta_a^2 + \eta_a^2 \eta_b^2 + \eta_b^2$$

Ainsi, le bruit extrinsèque ajoute des termes qui ne dépendent pas de l'abondance moyenne de protéines. Pour des gènes dont le niveau d'expression est élevé, à tel point que les termes en $\langle p \rangle^{-1}$ disparaissent, la variabilité phénotypique est alors très largement dominée par les trois derniers termes de l'expression du bruit extrinsèque. Ces termes génèrent donc un bruit global inévitable qui est indépendant de l'abondance moyenne des protéines et qui se manifeste à haut niveau d'expression. A l'inverse, le bruit intrinsèque pourra avoir une contribution importante pour des gènes faiblement exprimés. La variabilité phénotypique présentera dans ce cas une dépendance avec l'abondance moyenne des protéines $\langle p \rangle$. La limite inférieure du bruit intrinsèque représente des fluctuations Poissoniennes du nombre de protéines. Ainsi, dans le plan $(\langle p \rangle, \eta_p^2)$, à faible niveau d'expression, la variabilité phénotypique ne pourra pas se situer en deçà

de la courbe d'équation $\eta_p^2 = 1/\langle p \rangle$, et à fort niveau d'expression en deçà de la courbe d'équation $\eta_p^2 = \eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2$ (Figure 12). Taniguchi et al proposent donc une manière d'introduire les sources de bruit extrinsèque dans le modèle stochastique à deux niveaux. Leur méthode se base sur des données d'expériences qui montrent que les fluctuations engendrées par le bruit extrinsèque sont lentes devant les fluctuations d'origine intrinsèque. Cependant, les auteurs ne donne pas la nature exacte des sources de stochasticité responsable du « plateau » de bruit se manifestant à haut niveau d'expression. Son origine pourrait provenir des fluctuations dans le nombre d'ARN polymérases et/ou de ribosomes mais aussi des différentes étapes du cycle cellulaire comme la réplication du chromosome.

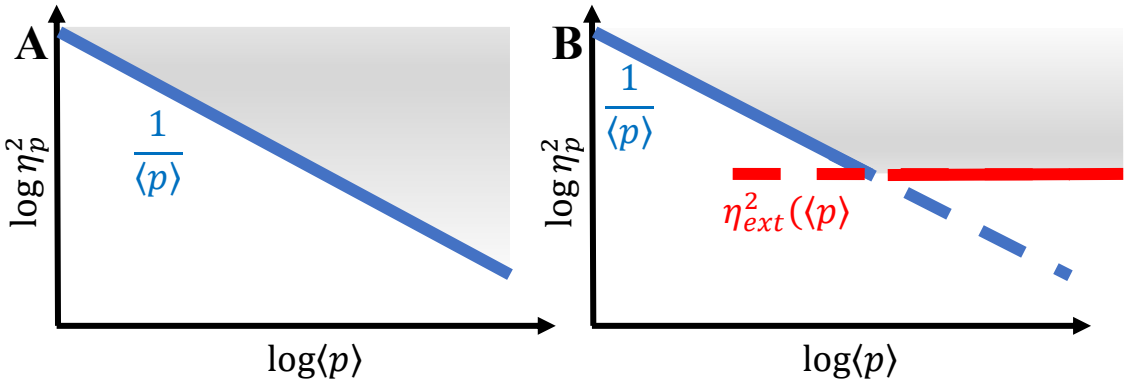


Figure 12 : Comportement asymptotique de la variabilité phénotypique dans le plan $(\langle p \rangle, \eta_p^2)$ et domaine du plan accessible aux différents gènes (zones grisées). **A)** Cas où il n'y a que le bruit intrinsèque. Pour un niveau d'expression donné, les fluctuations minimales du nombre de protéines sont de types Poissoniennes (courbe bleue). Ainsi, la variabilité phénotypique pour l'ensemble des niveaux niveau d'expression du protéome est contrainte à la zone grisée. **B)** Dans le cas où le bruit extrinsèque est pris en compte dans le modèle stochastique à deux niveaux (Taniguchi et al, 2010), les fluctuations Poissoniennes intrinsèques restent la limite inférieure pour les gènes s'exprimant faiblement, tandis que les sources de stochasticité extrinsèque qui sont globales à la cellule représentent la limite inférieure pour les gènes s'exprimant fortement (courbe rouge). Cette seconde limite qui se manifeste à haut niveau d'expression ne dépend pas de l'abondance moyenne des protéines $\langle p \rangle$.

Par ailleurs ce modèle ne prend pas en compte la variabilité introduite par la répartition des molécules lors de la division cellulaire.

III.3.2 Effets de la division cellulaire sur la variabilité phénotypique :

Nous allons ici détailler comment la division cellulaire introduit de la variabilité phénotypique supplémentaire. La plupart des bactéries se reproduisent par fission binaire : la cellule bactérienne croît et se divise en deux cellules de taille identique. La division cellulaire est assurée par la polymérisation de la protéine FtsZ qui crée un anneau constricteur de la paroi cellulaire lors de la division (appelé Z ring). Le contenu cellulaire en termes de protéines et d'ARN messagers va se retrouver partitionné en deux : on parle de ségrégation des

molécules lors de la division cellulaire. Ce partitionnement des molécules lors de la division cellulaire introduit des différences entre cellules sœurs. En effet, pour des organismes qui se divisent de façon symétrique comme *E. coli*, le nombre de molécules d'une espèce chimique donnée dont va hériter une cellule fille suit une loi binomiale de paramètre 0.5, chaque molécule ayant la même probabilité de se retrouver dans l'une ou l'autre des cellules filles. Pour être précis, la probabilité d'obtenir N_1 molécules dans la cellule fille 1 à partir de N molécules dans la cellule mères'écrit :

$$P(N_1, N) = \frac{N!}{N_1!\,(N - N_1)!} 1/2^N$$

On note $N_2 = N - N_1$ le nombre de molécules dans la cellule fille 2. On en déduit les propriétés suivantes : $\langle N_1 \rangle = N/2 = \langle N_2 \rangle$ et $\sigma_{N_1}^2 = N/4 = \sigma_{N_2}^2$. A partir de ces égalités, il est possible alors de quantifier la variabilité induite par la répartition binomiale des protéines :

$$\frac{\sigma_{N_1}}{\langle N_1 \rangle} = \frac{1}{\sqrt{N}}$$

Ainsi, la partition aléatoire introduit une variabilité d'autant plus importante que le nombre de molécules impliqué est faible. Cette répartition binomiale est notamment attendue pour les protéines et les ARN messagers et a été directement démontré dans la bactérie modèle *E. coli*. pour les ARN messagers (Golding et al, 2005) et pour les protéines (Rosenfeld et al, 2005) (Figure 13). Dans leurs travaux, Golding et al rendent fluorescentes les molécules d'ARN messagers issues de la transcription d'un gène d'intérêt. Il est alors possible de compter directement le nombre de ces molécules dans la cellule mère avant la division (N) et comment elles se répartissent dans chacune des deux cellules filles (N_1 et N_2). Rosenfeld et al s'intéressent quant à eux à la répartition d'une protéine fusionnée avec une protéine fluorescente YFP. Là encore, la fluorescence détectée permet de quantifier la répartition des protéines entre les deux cellules filles. La quantification du nombre ne se fait cependant pas directement, mais au moyen du signal de fluorescence de chaque cellule qui est proportionnel au nombre de molécules fluorescentes présentes dans chaque bactérie : $I = \alpha N, I_1 = \alpha N_1,$ et $I_2 = \alpha N_2$. Dans ces deux études, la mesure de l'écart $|\Delta N| = |N_1 - N_2|$ entre les deux cellules filles en fonction du nombre N de molécules dans la cellule mère est reportée. Pour chaque nombre de messagers ou de protéines initial N, la moyenne quadratique de $|\Delta N|$ est calculée et comparée à ce que donnerait une loi binomiale de paramètre 0.5. Cette distribution binomiale donne une moyenne quadratique de $|\Delta N|$ égale à : $\sqrt{\langle (N_1 - N_2)^2 \rangle} = \sqrt{N}$, et correspond bien aux comportements observés expérimentalement pour la répartition des ARN

messagers (Figure 13B) et des protéines (Figure 13C). Le caractère stochastique de la partition des molécules à la division cellulaire impacte également le nombre de protéines impliquées dans l'expression génique, à savoir les ARN polymérases, les ribosomes et autres protéines du dégradosome. La variabilité phénotypique induite par la partition peut être dans certain cas plus importante que celle donnée par une distribution binomiale de paramètre 0.5, notamment pour des protéines qui se regrouperaient ou s'agrégerait en vésicules et en organites. Au contraire, on pourrait imaginer que dans certains cas les cellules parviennent à une ségrégation plus ordonnée et ainsi réduire la variabilité phénotypique engendrée par le mécanisme de ségrégation du contenu cellulaire lors de la division cellulaire. Huh et Paulsson ont montré mathématiquement comment le partitionnement stochastique contribue à la variabilité phénotypique observée sur une population de cellules identiques (Huh & Paulsson, 2011). Pour cela, les auteurs considèrent un modèle dans lequel les ARN messagers et les protéines sont synthétisées et dégradées de façon déterministes au cours du cycle cellulaire avant d'être répartis aléatoirement lors de la division cellulaire. La distribution obtenue s'ajuste alors parfaitement avec la distribution binomiale négative prédite par le modèle stochastique à deux niveaux. Ainsi, les erreurs de partitionnement "imitent" le bruit d'expression génique, et l'hypothèse peut être émise que la variabilité phénotypique observée dans les différentes expériences cherchant à valider les modèles stochastiques de l'expression génique provient plutôt de la ségrégation aléatoire à la division cellulaire plutôt que de la stochasticité inhérente à l'expression génique (Huh & Paulsson, 2011). La division cellulaire n'intervient qu'au bout d'un certain temps T_d nommé temps de division qui dépend grandement du type cellulaire et des ressources disponibles. En milieu riche, le temps de division peut être d'une vingtaine de minute par exemple pour les organismes modèles *E. coli* et *B. subtilis* et atteindre des temps supérieurs à l'heure dans des milieux pauvres. Au cours de ce temps, la cellule croît de manière exponentielle caractérisé par le taux d'élongation $\mu = ln2/T_d$. Les modèles stochastiques ne s'intéressent qu'aux évènements de synthèse et dégradation des différents composés de l'expression génique mais ne prennent pas en compte le volume dans lequel ces réactions se produisent. Les modèles s'effectuent donc à volume constant. Considérons dans un premier temps que le volume en question représente le volume de la cellule. Ainsi, les nombres de protéines et de messagers prédits par le modèle représentent dans ce cas les quantités totales de ces molécules dans la cellule. Du fait de la division cellulaire, la moitié des transcrits et des protéines sont amenés à un instant donné (T_d) à « quitter » la cellule et à ne jamais y revenir. On parle alors de dilution. Cet effet sera d'autant plus important si ces molécules survivent jusqu'à l'instant T_d, c'est-à-dire pour $\tau_R, \tau_P \gg T_d$, où nous avons défini les temps de vie moyen τ_R et τ_P à partir des réactions de dégradation dues aux ribonucléases et protéases. Ainsi, il existe deux mécanismes entraînant une décroissance de la quantité d'ARN messagers et de protéines : la

dégradation et la dilution. La prépondérance de l'un ou l'autre sur la décroissance se mesure en comparant les temps de vie moyen et le temps de division T_d :

- Les ARN messagers sont très instable et leur temps de vie moyen est estimé à environ 4 minutes (Taniguchi et al, 2010). Ce qui est nettement inférieur au temps de division des bactéries. Par exemple, dans les expériences de Taniguchi et al., le temps de division est estimé à 150 minutes. Ainsi, le processus de décroissance du nombre d'ARN messagers est dominé par la dégradation.

- La plupart des protéines, comme nous l'avons vu précédemment, ont un temps de vie bien supérieur à celui des ARN messagers et peuvent même survivre à plusieurs cycles cellulaires (Koch et Levy, 1955). Ainsi, le processus de décroissance du nombre de protéines est dominé par la dilution.

Ainsi, il est possible d'injecter dans le modèle la division cellulaire (Taniguchi, 2010) en remplaçant le taux de dégradation des protéines par le taux de croissance de la cellule. Cependant, il est important de noter que le phénomène de dilution due à la division cellulaire est traité par les modèles stochastiques comme un processus dont les différents instants d'occurrence sont distribués exponentiellement, à l'instar des autres réactions du modèle. En réalité, la dilution intervient à intervalles réguliers correspondant aux temps de division. Enfin, une autre manière d'introduire la croissance cellulaire mais aussi la réplication des chromosomes est de raisonner non pas en nombre de composants, mais en termes de concentration, mais il faut pour cela modifier les modèles. En effet, l'expansion du volume de la cellule contribue à une diminution des concentrations de chaque composant (gène, ARN messagers et protéines) par dilution. Thattai et van Oudenaarden supposent que l'ADN est répliqué avec un taux correspondant exactement à la croissance cellulaire (Thattai & van Oudenaarden, 2001). Ainsi la concentration en ADN est constante au cours du cycle cellulaire. L'expansion du volume ainsi que les réactions de dégradation jouent de concert pour diminuer la concentration du nombre de copies d'ARN messagers et de protéines. Ainsi, Thattai & van Oudenaarden proposent d'ajouter au taux de dégradation des ARN messagers et des protéines le taux de croissance de la cellule : $\mu_i + \mu, i = R, P$ (Thattai & van Oudenaarden, 2001). Raisonner en termes de concentration revient à considérer une unité de volume entourant un gène d'intérêt et à calculer les fluctuations du nombre de copies d'ARN messager et de protéines à l'intérieur de ce volume. Ces fluctuations proviennent des évènements de synthèse/dégradation de ces composants et également de leur dilution : lorsqu'un composé (une protéine ou un ARN messager) quitte le volume, on suppose qu'il ne revient jamais à l'intérieur. Les instants auxquels les évènements de sortie se produisent sont distribués exponentiellement.

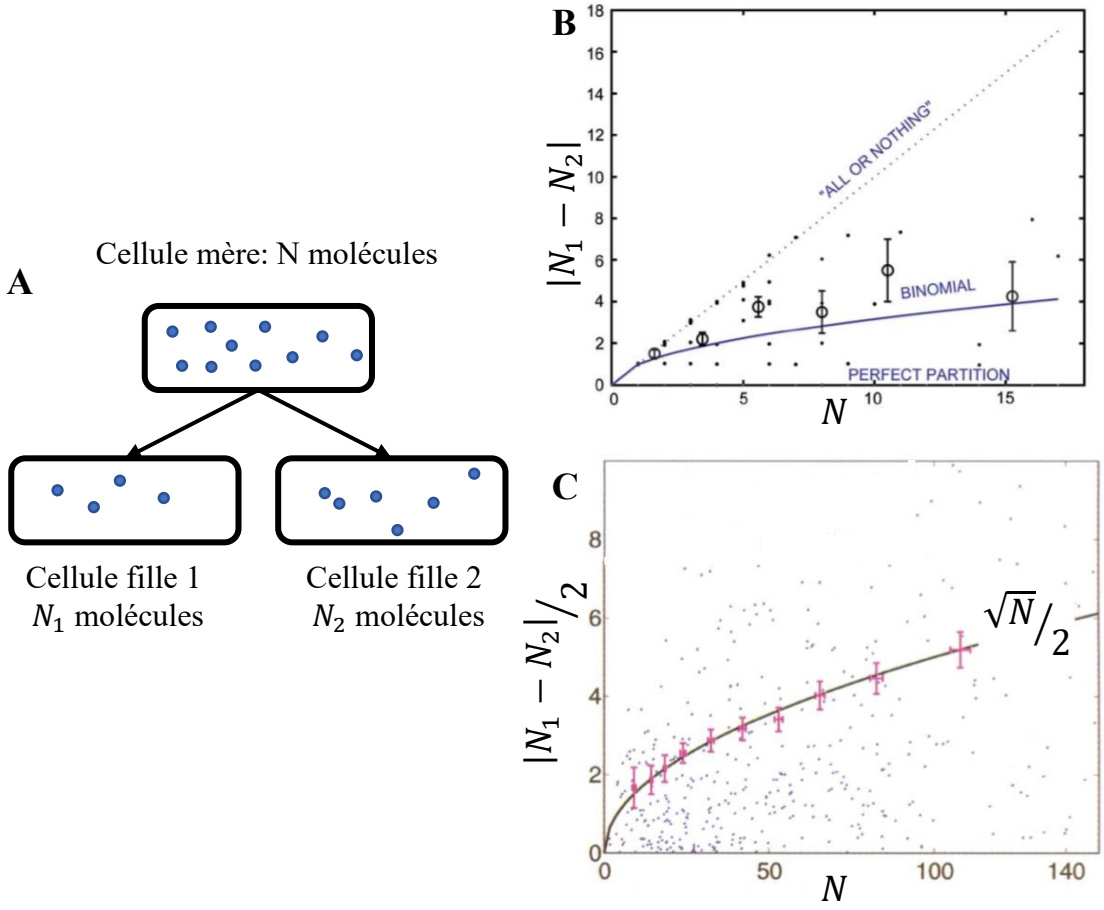


Figure 13 : Répartition binomiale des ARN messagers et des protéines durant la division cellulaire dans la bactérie *E. coli*. A) La cellule mère commence avec N protéines ou ARN messagers (cercles bleus). Au moment de la division cellulaire, ces molécules sont partitionnées dans les deux cellules filles suivant une loi binomiale de paramètre 0,5. **B)** Différence $|N_1 - N_2|$ du nombre d'ARN messagers entre les deux cellules filles sachant que la cellule mère en possédait N. Les courbes montrent trois mécanismes possibles de la répartition des messagers représentée ici par la moyenne quadratique. Le premier cas du « all or nothing » correspond au cas où une des cellules récupère l'ensemble des messagers. Le cas « perfect partition » décrit la situation dans laquelle chaque cellule fille reçoit exactement la moitié du nombre de messagers présent dans la cellule mère. Le dernier cas correspond à la répartition binomiale du nombre de messagers dans la cellule fille pour lequel $\sqrt{\langle (N_1 - N_2)^2 \rangle} = \sqrt{N}$. Ce dernier cas semble être le comportement le plus proche des données expérimentales. (Adapté de Golding et al, 2005). **C)** Différence $|N_1 - N_2|$ du nombre de protéines entre les deux cellules filles en fonction du nombre N de protéines dans la cellule mère. Là encore, le modèle de répartition binomiale (courbe noire) décrit bien la moyenne quadratique calculée à partir des données expérimentales (croix rouges). (Adapté de Rosenfeld et al, 2005).

III.3.3 Distribution exponentielle des temps d'attente :

Une autre hypothèse du modèle consiste à dire que tous les différents évènements intervenant dans l'expression génique (synthèse/dégradations des protéines et des messagers, activation/inactivation) se réalisent avec des intervalles de temps qui sont distribués exponentiellement. Les taux de ces différentes réactions sont soit des constantes, soit dépendent de l'état actuel du système. Cette hypothèse fait que les processus stochastiques sont Markoviens : les modèles sont « sans mémoire », c'est-à-dire que le futur du système ne dépend que de l'état présent mais pas des instants antérieurs à l'état actuel. La distribution exponentielle des temps d'attente est en réalité très bien adaptée pour modéliser la collision de deux espèces chimiques toutes deux soumises à des fluctuations Browniennes dues à l'agitation thermique. Ainsi, le temps d'attente entre deux rencontres entre une ARN polymérase et un promoteur peut être correctement modélisé par une distribution exponentielle. Les modèles stochastiques de l'expression génique considèrent quant à eux que ce sont les temps séparant deux évènements d'incrémentation ou de décrémentation d'une unité du nombre de promoteurs actifs, d'ARN messagers et de protéines qui sont distribués exponentiellement. Or, ces temps comprennent les temps de rencontre entre molécules et les temps nécessaires pour les réactions de synthèse et/ou dégradation. Ces réactions requièrent plusieurs étapes biochimiques. Dans les cellules procaryotes par exemple, pour la transcription, suite à la fixation de l'ARN polymérase sur le promoteur, la synthèse d'un ARN messager nécessite l'ouverture du complexe ARN polymérase-séquence promotrice, le détachement de l'ARN polymérase du promoteur qui n'intervient qu'après un certain nombre de cycles abortifs, et pour finir l'élongation durant laquelle l'ARN polymérase va parcourir la partie codante du gène durant laquelle l'enzyme va alterner des phases à vitesse constante et des temps de pause. Toutes ces étapes peuvent donc conduire à une distribution du temps d'attente entre deux synthèses d'ARN messager non exponentielle De plus, il est nécessaire pour certains gènes d'ajouter des étapes supplémentaires pour pouvoir décrire correctement leur expression et prendre en compte les éventuelles sources de stochasticité supplémentaire inhérentes à ces niveaux. Par exemple, de nombreuses protéines nécessitent une étape supplémentaire dite de maturation dans laquelle interviennent des protéines appelées chaperons dont la fonction est d'assister les protéines dans leur repliement pour les placer dans leur conformation fonctionnelle. Les protéines sont alors capables d'assurer leur fonction au sein de la cellule. Cette étape introduit de nouvelles sources de stochasticité intrinsèque et extrinsèque. Le repliement protéique est caractérisé par un temps moyen τ_m qui dépend de la nature intrinsèque de la protéine et de l'organisme dans lequel cette protéine a été synthétisé. Un autre phénomène pouvant conduire à des écarts avec le comportement exponentiel est celui de la « recherche » de la séquence promotrice par la polymérase. Des expériences menées à l'échelle de la molécule unique (Elf

et al, 2007) ont révélé que cette recherche se faisait en deux phases qui alternent aléatoirement du fait des chocs désordonnés avec les molécules du milieu environnant : une phase de diffusion à 1D le long du brin d'ADN (recherche locale), et une phase de diffusion à 3D dans le cytoplasme (recherche locale). On parle alors de « recherches intermittentes ». Cette stratégie permet notamment de réduire le temps de recherche et donc le temps entre deux transcriptions par rapport à une stratégie de type marche aléatoire simple (Coppey et al, 2004). De même, les réactions de dégradation ne se font pas de manière instantanée puisqu'elles résultent aussi de nombreuses étapes biochimiques (Kushner, 2002 ; Deutscher, 2006). Par exemple, la dégradation des transcrits chez *E. coli* nécessite dans un premier temps le recrutement du complexe multiprotéique constituant le dégradosome ; la protéines RNaseE se fixant à l'extrémité triphosphate 5', et la protéine PNPase au niveau de la queue polyadénylée en 3'. Ensuite, la RNaseE doit procéder au clivage endonucléolytique du messager en plusieurs régions riches en Adénine et Uracile. Et pour finir, les fragments générés doivent être par la suite digérés de leur extrémité 3' jusqu'en 5' par les exoribonucléases PNPase, RNaseII et RNaseR. Là encore, les intervalles de temps entre deux évènements de dégradation ne sont pas nécessairement distribués exponentiellement. Pedraza et Paulsson ont pris en compte ces distributions non-exponentielles des délais entre chaque réaction de synthèse ou de dégradation (Pedraza & Paulsson, 2008). Dans leur étude les auteurs ont notamment introduit des bursts de transcription séparés par des intervalles de temps aléatoires, de distribution non exponentielle. Ces temps entre chaque burst définissent des périodes gestatoires. De même, les étapes de dégradation des messagers et des protéines sont décomposées en plusieurs sous étapes, introduisant la notion de sénescence, c'est-à-dire une dégradation progressive des transcrits et des protéines. Les auteurs concluent que les périodes de gestation et de sénescence conduisent à une réduction de la variabilité du niveau d'expression du gène comparativement à ce que prédisent les modèles stochastiques « standards ». D'autres études sur le sujet ont été mené pour évaluer l'importance de ces aspects sur la variabilité phénotypique (Fromion et al, 2013 ; Leoncini et al, 2013).

III.3.4 Autorégulation

Certains gènes ont la capacité de s'autoréguler, c'est-à-dire que la protéine va influencer sa propre production à travers une boucle rétroactive ou « feedback ». Ce mécanisme intervient au niveau transcriptionnel, la protéine influençant la capacité de l'ARN polymérase à se lier au promoteur. Deux cas peuvent alors se présenter : (i) soit le gène est réprimé par ses propres protéines et on parle d'autorégulation négative ou « negative feedback » ; (ii) soit les protéines activent leur propre production en améliorant la force du promoteur. On parle dans ce cas d'autorégulation positive ou « positive feedback ». Ces deux « motifs » de régulation ont été exploré à la fois théoriquement (Thattai & van Oudenaarden,

2001) et expérimentalement chez les bactéries *E. coli* et *B. subtilis* (Becskei & Serrano, 2000 ; Becskei et al, 2001 ; Ozbudak et al, 2004 ; Maamar et al, 2007) pour déterminer l'effet de ces boucles de rétroaction sur la variabilité phénotypique. Par exemple, l'autorégulation négative peut être modélisée en considérant le taux d'initiation de la transcription comme une fonction décroissante non linéaire du nombre (ou de la concentration) de protéines, telle que la fonction de répression de Hill : $k_R = k_R^{max}/(1 + (p/K)^n)$, où K est appelée constante de dissociation et est égale au nombre ou à la concentration de protéine pour laquelle le taux de transcription est égale la moitié du taux de transcription maximal (c'est-à-dire en l'absence de boucle rétroactive) et n le coefficient de Hill (Thattai & van Oudenaarden, 2001). Une rétroaction négative conduit à une réduction du niveau de bruit intrinsèque, tandis qu'une rétroaction positive conduit à une amplification du bruit et peut même engendrer une distribution multimodale (Thattai & van Oudenaarden, 2001)). Ces deux « motifs » sont très importants puisqu'ils se retrouvent dans de nombreux réseaux de gènes naturels et synthétiques.

IV Mise en évidence expérimentale de la stochasticité dans l'expression génique :

On pourrait s'attendre à ce que les modèles stochastiques à deux et trois niveaux présentés ici soient difficiles à valider ou invalider notamment du fait des nombreux mécanismes cellulaires générant de la variabilité phénotypique autre que les sources de bruit intrinsèque. Cependant, comme nous allons le voir ces modèles ont été validés par des expériences ayant recours à des outils quantitatifs de plus en plus précis, menées à l'échelle de la cellule unique, ainsi que par des techniques de biologie moléculaire toujours plus pointues et relativement simples à mettre en place. Comme il l'a été dit précédemment, pour mesurer le bruit d'expression d'un gène en particulier, il faut dans un premier temps rendre mesurable l'abondance des protéines et éventuellement des ARN messagers. Pour ce faire, la protéine du gène d'intérêt peut être fusionnée avec une protéine fluorescente (GFP, mKate, mCherry…) appelée également fluorophore, ou bien le gène d'intérêt peut directement coder pour une protéine fluorescente. Ce gène peut être soit porté par le chromosome de la cellule, soit par un plasmide (dans le cas des bactéries) suivant les effets que l'on souhaite mesurer. La quantification et la caractérisation du niveau de bruit associé à l'expression du gène d'intérêt se fait au moyen du coefficient de variation et du facteur Fano. On doit donc préciser comment on calcule l'abondance moyenne du nombre de protéines (ou d'ARN messagers) ainsi que l'écart type de la distribution du aux fluctuations. On peut envisager pour cela deux procédures. (i) Suivre l'évolution temporelle du niveau de fluorescence d'une cellule, la fluorescence étant proportionnelle à la quantité

de fluorophores présents dans la cellule, et calculer les grandeurs statistiques sur un intervalle de temps. Cet intervalle de temps doit être suffisamment grand pour que le nombre de protéines ait parcouru l'ensemble des valeurs qui lui sont permis de sorte que la distribution de leur nombre soit la plus correcte possible et permette une bonne estimation de la valeur moyenne ainsi que de l'écart-type. Cette procédure est celle qui est sous entendue dans les modèles d'expression génique où les calculs ont été fait dans une cellule. (ii) Une autre approche consiste à prendre une collection d'un grand nombre de répliques de la cellule précédente, soit une population de cellules génétiquement identiques, chacune de ces cellules ayant un nombre de protéines différent. Ainsi, en supposant un nombre de cellules important, suffisamment pour couvrir l'ensemble du nombre de protéines accessible du fait des fluctuations, on peut caractériser la distribution du nombre de protéines et calculer le niveau d'expression moyen ainsi que l'écart type de la distribution. Une question se pose alors : la statistique obtenue à partir de la première procédure, c'est à dire à partir du comportement d'une seule cellule est-elle équivalente à la statistique obtenue à partir d'une population de cette même cellule ? En physique statistique, on réconcilie les deux approches à l'aide de l'hypothèse ergodique selon laquelle les statistiques obtenues sur une réalisation donnée d'un processus sont identiques à celles obtenues sur un ensemble de réalisations indépendantes observées à un instant donné (Coulon, 2010). Cette hypothèse, bien que rarement examinée, est supposée vérifiée et est à l'origine de nombreuses expériences cherchant à mesurer le bruit d'expression génique. Si on considère les deux procédés comme identique d'un point de vue statistique, la première approche, c'est-à-dire suivre l'évolution de la quantité d'une protéine dans une cellule apporte une information supplémentaire sur la dynamique du bruit tandis que la seconde approche donne seulement une information sur la variabilité à un instant donné. Observer le bruit d'expression génique nécessite d'avoir recours à des outils permettant de visualiser les cellules et permettant de plus de quantifier les niveaux de fluorescence. Les méthodes utilisées sont la cytométrie en flux et la microscopie de fluorescence. Nous allons présenter dans ce qui suit les travaux expérimentaux qui ont pour objectif d'identifier les sources du bruit d'expression en confrontant résultats expérimentaux et modèles de l'expression génique. Aussi, en étant capable de modifier et de contrôler les différents paramètres biologiques pris en compte dans les modèles, il est alors possible de vérifier leurs prédictions. Ces différentes expériences ont été menées sur des organismes dits « modèles ». Nous présenterons ici des expériences menées sur des unicellulaires qui ont été étudiés depuis longtemps par les biologistes, si bien que leur génome a été entièrement séquencé et qu'on sait parfaitement les cultiver dans différents milieux de culture référencés. Pour l'ensemble des espèces bactériennes, les organismes modèles les plus utilisés sont *Escherichia coli* (*E. coli*) pour les bactéries à Gram négatif et *Bacillus subtilis* (*B. subtilis)* pour les bactéries à Gram positif. Pour les cellules eucaryotes, un organisme modèle classique est la levure *Saccharomyces cerevisiae (S.*

cerevisiae). Les premiers travaux expérimentaux effectués pour caractériser le bruit d'expression génique se concentrent sur un gène rapporteur en particulier et observent à l'aide de constructions génétiques comment varient le bruit d'expression génique de ce gène au niveau des ARN messagers et des protéines lorsqu'est modifié le niveau d'expression moyen par action sur la transcription ou la traduction (Elowitz et al, 2002 ; Ozbudak et al, 2002 ; Blake et al, 2003 ; Raser et O'Shea, 2004 ; Golding et al, 2005). Enfin, certains travaux ont adopté une approche « génomique » en rapportant le bruit d'expression en fonction de l'abondance moyenne des protéines pour un ensemble de gènes dont la taille est comparable au génome (Newman et al, 2006 ; Bar-Even et al, 2006 ; Taniguchi et al, 2010) chez la levure *S. cerevisiae* et la bactérie *E. coli*. Ces études ont permis d'investiguer les propriétés du bruit sur l'ensemble du génome.

IV.1 Effets de la transcription et de la traduction sur le bruit d'expression : Etude sur un gène unique :

Le travail pionnier conduit par Elowitz et al. (Elowitz et al, 2002) a déjà été présenté lors de la présentation des sources de bruits intrinsèques et extrinsèques. Pour rappel, ces auteurs présentent une méthode astucieuse (méthode « deux couleurs ») pour séparer les contributions du bruit extrinsèque et du bruit intrinsèque sur la variabilité phénotypique totale dans la bactérie *E. coli*. Au moyen d'un promoteur inductible, Elowitz et al. réalisent également une première analyse du comportement des bruits intrinsèque et extrinsèque en fonction du niveau d'expression moyen. Ils ont établi que le bruit intrinsèque évoluait comme l'inverse du niveau d'expression moyen tandis que le bruit extrinsèque évoluait de façon non monotone, atteignant un maximum pour des niveaux de fluorescence intermédiaire. En comparant leur donnée avec un modèle développé par Swain et al. (Swain et al, 2002), ils interprètent la tendance observée pour le bruit intrinsèque comme provenant des évènements de synthèse/dégradation des ARN messagers. Une seconde étude pionnière très importante a été réalisé par Ozbudak et ses collaborateurs (Ozbudak et al, 2002) cette fois ci sur la bactérie *B. subtilis*. Les auteurs cherchent à quantifier l'influence des taux de transcription et de traduction sur la stochasticité de l'expression génique et permettent ainsi de tester le modèle à deux niveaux introduit par Thattai et van Oudenaarden (Thattai & van Oudenaarden, 2001). Ces deux investigations expérimentales sur la stochasticité de l'expression génique ont été mené chez des cellules procaryotes et quantifient le bruit d'expression au niveau protéique. Ces deux études « pionnières » été rapidement suivies par des expériences du même type chez les cellules eucaryotes en utilisant la levure *S. cerevisiae* (Blake et al, 2003 ; Raser & O'Shea, 2004). Enfin, une étude menée par l'équipe d'Ido Golding a quant à elle observée la

production d'ARN messagers et quantifiés le bruit d'expression génique au niveau des ARN messagers chez la bactérie *E. coli*.

IV.1.1 Prédiction des modèles stochastiques sur l'effet de la transcription et de la traduction sur le bruit d'expression génique :

Nous avons vu précédemment que la force du bruit dans le cadre du modèle à deux niveaux s'écrivait comme une fonction du taux de traduction k_P mais était indépendant du taux de transcription k_R :

$$\frac{\sigma_p^2}{\langle p \rangle} = 1 + \frac{k_P}{(\gamma_R + \gamma_P)} \cong 1 + \frac{k_P}{\gamma_R}$$

Thattai & van Oudenaarden introduisent l'hypothèse de burst traductionnel, qui permet de décrire la dynamique de la traduction. Selon cette hypothèse, un faible nombre d'ARN messager (potentiellement unique) est efficacement traduit ($k_P > \gamma_R$), de telle sorte que le nombre de protéines dépend du faible nombre d'ARN messager, ce qui permet de montrer l'importance des fluctuations au niveau des ARN messagers sur le bruit d'expression mesuré au niveau protéique. Le terme de burst traductionnel est utilisé pour caractériser la contribution des évènements stochastiques de synthèse/dégradation des ARN messagers sur le bruit d'expression génique. Le nombre moyen de protéines synthétisées par ARN messagers (donc par burst) est donné par la quantité k_P/γ_R. D'après ce modèle, pour un niveau d'expression donné, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit. C'est cette hypothèse que Ozbudak et al. cherchent à vérifier en quantifiant la force du bruit pour différentes constructions génétiques permettant à la fois de moduler de manière indépendante les taux de traduction et de transcription.

Pour le modèle à trois niveaux, la force du bruit présente un terme supplémentaire dû aux évènements stochastiques d'activation/inactivation du (ou des) promoteur(s) :

$$\frac{\sigma_P^2}{\langle p \rangle} = 1 + k_P \tau_R + k_P k_R \tau_R (1 - P_+) \frac{\tau_P}{1 + \tau_g^{-1} \tau_P}$$

Ce modèle permet d'introduire le concept de burst transcriptionnel selon lequel l'ADN est efficacement transcrit sur de courtes périodes d'activité des promoteurs, ce qui conduit à une synthétise d'ARN messagers en « rafales ». Ce phénomène de burst transcriptionnel aura un effet similaire sur la synthèse protéique que le phénomène de burst traductionnel, à savoir une synthèse protéique en rafale, mais leurs origines seront différentes. Le phénomène de burst

transcriptionnel est représenté par le troisième terme du facteur Fano, tandis que le second terme représente le phénomène de burst traductionnel. Suivant lequel de ces deux termes est dominant, le bruit d'expression tiendra son origine soit du burst de transcription, soit du burst de traduction. Expérimentalement, moduler le niveau d'expression moyen par action sur la traduction ou la transcription et mesurer la force du bruit correspondant permet de distinguer lequel des deux mécanises est majeure dans la génération du bruit.

IV.1.2 Données expérimentales sur l'effet de la transcription et de la traduction sur le bruit d'expression génique :

Le modèle stochastique à deux niveaux permet de prédire la dépendance du bruit avec les taux de traduction et transcription. Après avoir formulé ces prédictions, l'équipe de van Oudenaarden (Thattai & van Oudenaarden 2001) les a testées expérimentalement (Ozbudak et al, 2002), en variant indépendamment les taux de transcription et traduction d'un gène codant pour la Green Fluorescent Protein (GFP) introduit dans le chromosome de *B. subtilis*. Plus spécifiquement, Ozbudak et al. ont construit 4 souches de *B. subtilis* contenant le gène de la GFP traduit à différents taux, grâce à différentes séquences de RBS et différents codons d'initiation de la traduction. De plus, pour chaque souche le taux de transcription a été contrôlé grâce à l'utilisation d'un promoteur inductible à l'IPTG (promoteur Pspac). La force du bruit (facteur Fano) d'expression de la GFP a ainsi été mesurée pour les différentes souches dans les différentes conditions d'induction, sur la base de mesure de fluorescence de cellules uniques par cytométrie en flux. Pour réduire les effets de « taille » entre cellules, et s'affranchir d'une partie des sources de bruit extrinsèques, ces auteurs ont sélectionné les cellules selon une taille donnée. Le résultat est en accord avec les prédictions du modèle, la force du bruit étant fortement positivement corrélée avec le taux de traduction et seulement modérément corrélée avec le taux de transcription (Figure 14). Ce travail semble donc confirmer la conclusion théorique indiquant que pour le même niveau d'expression moyen, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit, et que le modèle stochastique à deux niveaux est bien adapté pour expliquer les données obtenues avec des cellules procaryotes.

Par la suite, le même type d'étude a été menée sur la levure *S. cerevisiae* (Blake et al, 2003). Là encore des promoteurs inductibles ont été utilisés pour moduler le taux de transcription, et différentes souches exprimant des variantes de protéines fluorescentes avec différents taux de traduction ont été utilisées. Comme dans le cas des bactéries, les données sur *S. cerevisiae* indiquent que la force du bruit augmente avec le taux de traduction. Mais au contraire des prédictions du modèle et des résultats obtenus sur *B. subtilis*, Blake et al. ont

montré que la force du bruit variait aussi avec l'efficacité transcriptionnelle, de façon non monotone, avec un maximum pour des taux de transcription intermédiaire. Si ces prédictions ne sont pas en accord avec le modèle stochastique à deux niveaux, ils sont en accord avec le modèle stochastique à trois niveaux et peuvent s'expliquer par le terme de burst transcriptionnel correspondant d'après ces auteurs à un processus lent de remodelage de la chromatine, ainsi qu'un processus d'assemblage séquentiels de la machinerie de transcription.

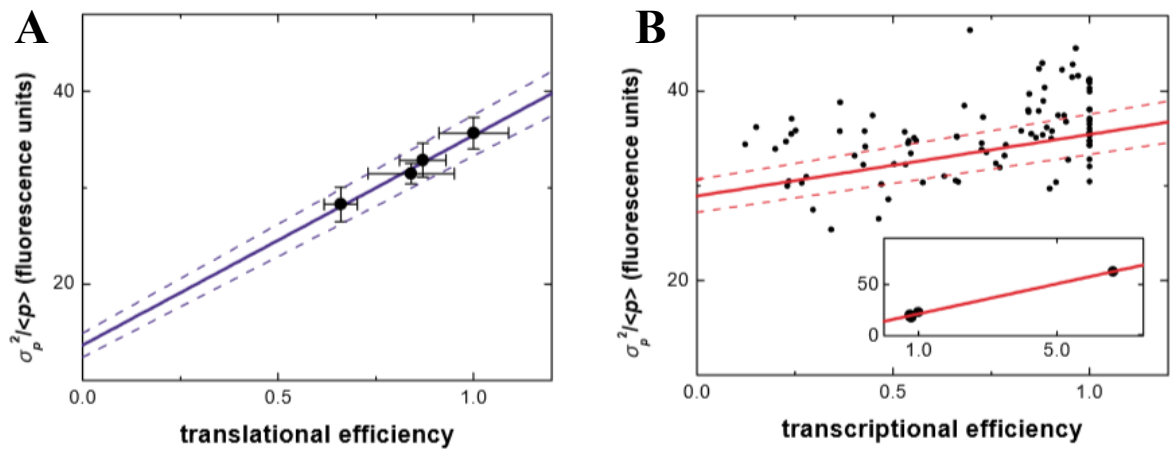


Figure 14 : Evolution du facteur Fano avec les efficacités de traduction et de transcription. A) Evolution du facteur Fano avec l'efficacité de traduction. L'efficacité de traduction est modifiée au moyen de différentes séquences RBS et codons d'initiation. **B)** Evolution du facteur Fano avec l'efficacité de transcription. La force de la transcription est modifiée au moyen d'un promoteur inductible à l'IPTG. Le facteur Fano présente d'après ces résultats une corrélation forte avec l'efficacité de traduction et une corrélation plus faible avec l'efficacité de transcription, ce qui confirme la conclusion du modèle à deux niveaux : le bruit d'expression tient son origine des « *bursts de traduction* » et même niveau d'expression moyen, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit. D'après Ozbudak et al (Ozbudak et al, 2002).

Il est important de constater que dans ces deux études sur *B. subtilis* et *S. cerevisiae*, le bruit mesuré comprend a priori une composante intrinsèque ainsi qu'une composante extrinsèque, cette dernière n'étant pas prise en compte dans les modèles. Peu de temps après, Raser & O'Shea ont effectué une autre étude sur *S. cerevisiae*, en utilisant la méthode développée par Elowitz et al. pour décomposer le bruit en composante extrinsèque et intrinsèque (Raser & O'Shea, 2004). Raser & O'Shea ont tout d'abord montré que le bruit intrinsèque représentait moins de 3% du bruit total dans leurs conditions. Ils ont ensuite testé la dépendance de la force du bruit intrinsèque avec l'efficacité de transcription sur plusieurs promoteurs inductibles, et ont obtenu des résultats différents pour les 3 promoteurs utilisés. Pour deux d'entre eux la force du bruit était constante quand le taux de transcription variait, tandis qu'une corrélation négative était obtenue pour le troisième promoteur. Les données collectées pour le bruit intrinsèque sont de nouveau en accord avec le modèle stochastique à trois niveaux et s'expliquent

par le terme de burst transcriptionnel. Les différents comportements du facteur Fano avec le niveau d'expression moyen correspondent à différentes cinétiques de l'étape d'activation/inactivation des promoteurs.

Ainsi, ces expériences tendent à montrer que le modèle à deux niveaux semble être suffisant pour décrire le bruit d'expression génique pour les cellules procaryotes. Le bruit d'expression tenant alors son origine du mécanisme de burst traductionnel. Ce modèle est cependant insuffisant pour décrire les données obtenues chez les eucaryotes. Pour ces derniers, le phénomène de burst transcriptionnel semble être la source principale du bruit intrinsèque et le modèle à trois niveaux le plus adapté pour décrire leurs résultats.

Cependant, les résultats obtenus par l'équipe de van Oudenaarden peuvent être sujet à discussion. Comme nous venons de le mentionner, les résultats obtenus s'appuient sur une comparaison entre des données de bruit total, sans partition entre le bruit intrinsèque et extrinsèque, et le modèle stochastique à deux niveaux qui ne prend en compte que les sources de bruit intrinsèque. Seul une sélection par la taille des cellules a été effectuée, ce qui devrait *a priori* seulement réduire les sources de bruits extrinsèques liés à certains aspects du cycle cellulaire, puisque les cellules collectées sont dans des états cellulaires équivalents. D'autre part, la transcription a été modulée au moyen d'un promoteur inductible. Le promoteur n'est donc pas constitutif mais régulé au moyen de répresseurs et de molécules d'inducteur. Ainsi, le modèle à trois niveaux pourrait être mieux adapté pour décrire correctement les données. C'est dans ce contexte que Golding et al (Golding et al, 2005) ont cherché à observer le bruit d'expression au niveau des ARN messagers en utilisant un promoteur inductible, donc dans conditions de transcription très proche de celles de Ozbudak et al. D'après les conclusions de l'étude d'Ozbudak et al, si on quantifie le nombre d'ARN messager entre cellules isogéniques, on devrait s'attendre à obtenir un facteur Fano égale à 1. Pour pouvoir observer les molécules d'ARN messagers, Golding et al utilisent la méthode de détection des ARN messagers liés à la protéine MS2. Ces auteurs conçoivent pour cela un gène pouvant transcrire un ARN messager présentant 96 copies d'une structure tige-boucle dans sa région non traduite. Ce gène est contrôlé par un promoteur inductible réprimé par LacI et activé par la protéine AraC. Chacune de ces structures secondaires se lient étroitement à la protéine formant la capside du bactériophage MS2. Cette protéine a été préalablement fusionné avec le fluorophore GFP. Lorsque 96 protéines MS2-GFP se lient à une molécule d'ARN messager, le signal de fluorescence est assez important pour pouvoir être détecté par microscopie de fluorescence et il est ainsi possible de compter le nombre d'ARN messagers présent dans une cellule. En mesurant l'hétérogénéité inter-cellulaire en termes de nombre d'ARN messagers, Golding et al mesurent un facteur Fano égal à 4 ce qui est une preuve solide du mécanisme de burst de transcription. Ces auteurs ont également mesuré l'activité

transcriptionnelle en temps réel et ont constaté que la synthèse d'ARN messagers se produit par « rafales », avec le promoteur qui permute aléatoirement entre les états actifs et inactifs, et montrent que le temps d'attente entre deux salves de synthèses de messagers est distribué exponentiellement. Enfin, le facteur Fano mesuré expérimentalement (~4,0) est en très bon accord avec les prédictions du modèle à trois niveaux (~4,6). Ainsi Golding et al montrent que le mécanisme de burst transcriptionnel n'est pas qu'endémique aux cellules eucaryotes mais peut être également être observé chez les cellules procaryotes dans le cas où le gène d'intérêt est soumis à des mécanismes de régulations. C'est effectivement le cas dans les études menées par Elowitz et al (Elowitz et al, 2002) et Ozbudak et al (Ozbudak et al, 2002) où les promoteurs utilisés sont inductibles.

IV.2 Etude sur une gamme de gènes comparable à la taille du génome :

IV.2.1 Chez la levure *S. cerevisiae* :

Si les études précédentes ont été menées sur un gène uniquement, Newman et al proposent une étude à l'échelle du génome de *S. cerevisiae* (Newman et al, 2006). Pour cela, ils constituent une banque de souches de *S. cerevisiae* dans laquelle la protéine fluorescente GFP est fusionnée à plus de 2500 protéines naturelles. Avec un système de mesures « haut débit » assuré par la cytométrie en flux, Newman et al mesurent l'abondance de ces 2500 protéines naturelles de *S. cerevisiae* fusionnées avec la protéine fluorescente GFP. Contrairement aux études précédentes, comme chaque gène ne partage pas nécessairement les mêmes promoteurs et RBS, il n'est pas possible avec cette collection de levure de quantifier les effets de la transcription et de la traduction sur la génération du bruit par exemple. Cependant, ils permettent de rendre compte de la dépendance de la variabilité phénotypique avec le niveau d'abondance moyenne de protéines naturellement présentes chez *S. cerevisiae*. En utilisant la stratégie « deux-couleurs » mise au point par Elowitz sur une partie de la collection les auteurs permettent de discerner la contribution des deux grandes classes de sources de bruit, à savoir extrinsèques et intrinsèques sur la variabilité phénotypique. La cytométrie en flux permet d'avoir pour chaque cellule des informations sur leur taille ainsi que leur granularité. A partir d'une population de levures exhibant une large gamme de tailles et d'états du cycle cellulaire, Newman et ses collègues sont alors capables de quantifier précisément l'effet de ces hétérogénéités sur le bruit extrinsèque. Ils constatent alors qu'en filtrant les cellules selon leur taille et leur granularité de sorte à ne garder que des cellules dans le même état cellulaire on réduit de manière significative le bruit extrinsèque alors que le bruit intrinsèque lui ne varie pas. Il est intéressant de noter que ce premier résultat valide la démarche d'Ozbudak et al pour réduire la contribution des sources de bruit extrinsèque. Ainsi, en sélectionnant des cellules dans le même état cellulaire, Newman et al observent que pour des gènes faiblement et moyennement

exprimés, le niveau de bruit intrinsèque est inversement proportionnel à l'abondance moyenne des protéines ; ce qui est compatible avec les modèles de l'expression génique (figure 15-A). Cependant, à fort niveau d'expression, cette tendance n'est plus vraie, mais elle est rétablie si on écarte les sources de bruit extrinsèque grâce à l'approche « deux couleurs ». Ainsi, les sources de bruit extrinsèque entraîneraient un bruit « plancher » indépendant du niveau d'expression moyen. Une autre équipe va réaliser un travail similaire mais contrairement à l'étude qui vient d'être mentionnée, celle-ci va confronter les résultats expérimentaux avec les prédictions théoriques. S'agissant de cellules eucaryotes, ces auteurs confrontent leur donnée avec le modèle à trois niveaux de l'expression génique (Bar-Even et al, 2006). Bar-Even et al. conçoivent une banque de souches de la levure *S. cerevisiae* dans laquelle 43 protéines naturelles ont été fusionnées avec la protéine GFP. Chaque cellule de cette banque est cultivée dans 11 conditions environnementales différentes, et la quantité de protéines pour chaque condition est déterminée par cytométrie en flux. Ces auteurs observent également une dépendance linéaire entre le bruit et l'inverse de l'abondance moyenne des protéines pour des souches et conditions expérimentales pour lesquels le niveau d'expression moyen est intermédiaire (Figure 15-B) La constante de proportionnalité entre le bruit intrinsèque et l'inverse de l'abondance moyenne des protéines est estimée à ~1200 :

$$\frac{\sigma_P^2}{\langle p \rangle^2} = \frac{C}{\langle p \rangle} = \frac{1200}{\langle p \rangle}$$

D'après le modèle stochastique à trois niveaux, une constante *C* nettement supérieur à 1 indique que le bruit d'expression est dû aux niveaux en amont de la phase de traduction. Cependant, Bar-Even et al. ne peuvent conclure quant à l'origine de ces fluctuations. Ces fluctuations peuvent en effet soit provenir des évènements de synthèse/dégradation des messagers, soit provenir des évènements d'activation/inactivation du (des) promoteur(s). Pour des hauts niveaux d'expression, la tendance précédente n'est plus vraie et atteint un niveau de bruit plateau en dessous duquel le bruit ne peut s'aventurer. Là encore, en employant l'approche "deux-couleurs" mise en place par Elowitz et al., les auteurs montrent expérimentalement que ce plateau est dû aux sources de bruits extrinsèques dont les effets ne peuvent être réduits par une sélection de la taille des cellules.

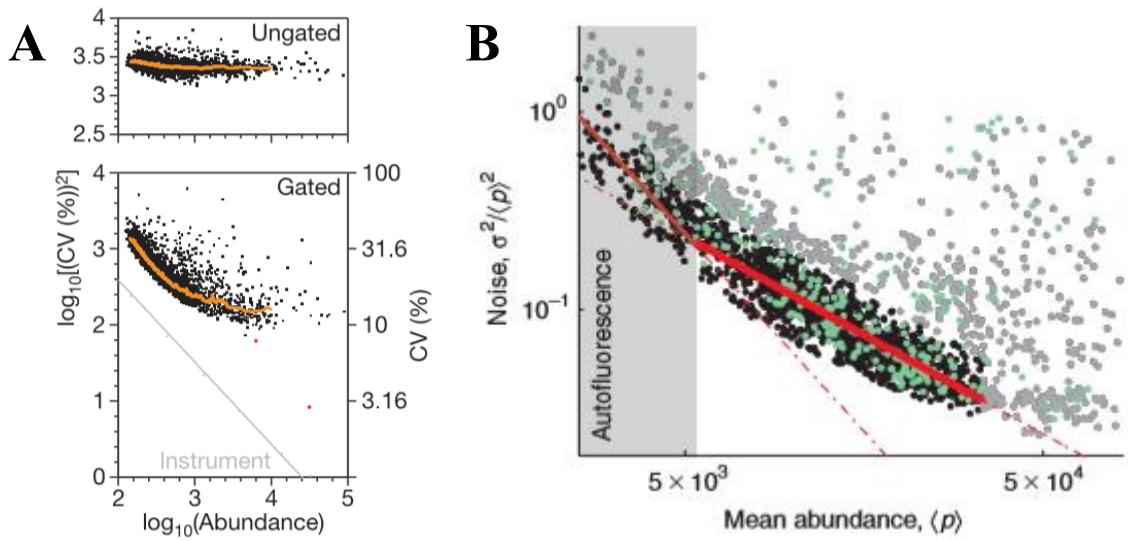


Figure 15 : Evolution du bruit d'expression en fonction du niveau moyen d'expression pour une collection de gènes naturellement présent chez *S. cerevisiae*. A-B) Les deux études menées par Newman et al (**A**) et Bar-Even et al (**B**) conduisent aux mêmes conclusions, à savoir une évolution inversement proportionnelle du bruit d'expression avec le niveau moyen d'expression pour des niveaux d'expression intermédiaire. Puis à fort niveau d'expression, les sources de bruit extrinsèque dominent et impose un bruit plancher indépendant du niveau moyen d'expression. Adaptée de Newman et al (Newman et al, 2006) (**A**) et Bar-Even et al (Bar-Even et al, 2006) (**B**).

IV.2.2 Chez la bactérie *E. coli* :

Taniguchi et al. ont réalisé une étude génomique comparable à ce qui a été fait sur la levure avec la bactérie *E. coli* en mesurant la variabilité d'expression génique sur un large nombre de protéines naturelles (1018) au moyen de la microscopie de fluorescence. La microscopie de fluorescence présente l'avantage d'avoir une meilleure résolution et une meilleure sensibilité, ce qui permet de quantifier des niveaux de fluorescence faible contrairement à la cytométrie en flux. Pour rendre leur système « haut débit » à l'instar de la cytométrie en flux employée précédemment pour les levures, Taniguchi et al ont développé un circuit micro-fluidique où chaque canal contient des cellules qui rapportent l'expression d'une protéine différente. Pour sonder la variabilité entre cellules dans le nombre de protéines, un gène rapporteur YFP a été insérée dans le chromosome des bactéries de telle sorte que la protéine fluorescente fusionne avec la protéine d'intérêt, si bien que le niveau de fluorescence détecté renseigne directement sur la quantité de cette protéine. Les signaux de fluorescence ont été de plus approximé en nombre de protéines, et les auteurs montrent qu'une normalisation de la fluorescence totale détectée par la surface des cellules permet d'éliminer les effets de taille sur la variabilité observée. Les résultats obtenus montrent que la variabilité phénotypique pour l'ensemble des protéines visitées peut être décrite par une loi Gamma de paramètre a et b , respectivement égale à l'inverse du carré

du coefficient de variation et au facteur Fano (ou force du bruit), ce qui est cohérent avec le modèle à deux niveaux (Figure 16 A-B).

Ces travaux font aussi apparaître une dépendance du niveau de bruit η^2 avec le niveau moyen d'expression des protéines $\langle p \rangle$. Ce niveau de bruit est tracé en fonction de l'abondance moyenne des protéines sur la figure 16-C où chaque point correspond à une protéine caractérisée par son niveau de bruit et son niveau d'expression moyen.

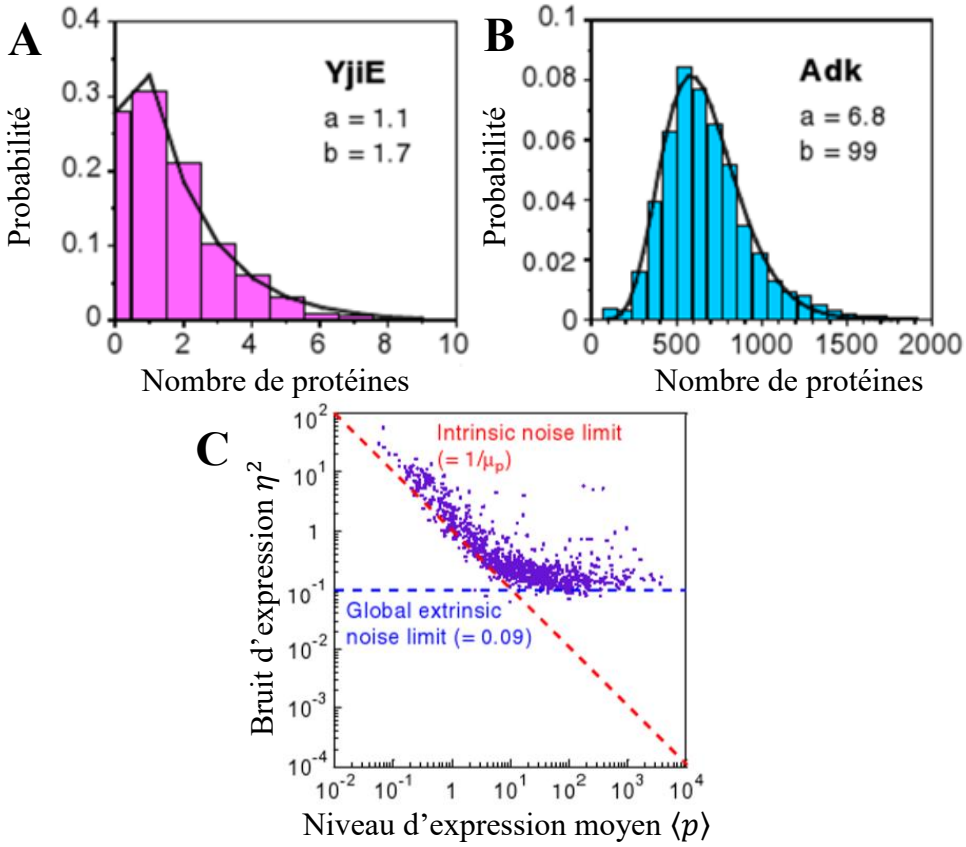


Figure 16 : Résultats de l'étude génomique sur *E. coli.* A)-B) Allure des distributions pour les protéines YjiE (**A**) et Adk (**B**). Ces deux distributions sont ajustées par des lois Gamma de paramètre a et b. Le fait que la variabilité inter-cellulaire se distribue selon une loi Gamma semble indiquer que pour cette étude génomique le modèle à deux niveaux est bien approprié pour décrire les données collectées **C**) Evolution du bruit d'expression avec le niveau moyen d'expression pour la collection de gènes visitée. Pour des niveaux d'expression inférieur à 10 protéines par cellule, les sources de bruit intrinsèques dominent (courbe rouge). Pour des niveaux d'expression supérieur à 10 protéines par cellule, les sources de bruit extrinsèque dominent et imposent un bruit « plancher » indépendant du niveau d'expression moyen en-dessous duquel on ne peut pas s'aventurer. Pour la protéine faiblement exprimée, YjiE les paramètres a et b peuvent être reliés aux paramètres biophysiques du modèle stochastique à deux niveaux tandis que pour la protéine fortement exprimée Adk, si la loi Gamma décrit bien la variabilité observée, les paramètres a et b ne peuvent plus être reliés aux paramètres du modèle de bruit intrinsèque. D'après Taniguchi et al, 2010

Les auteurs constatent que pour des niveaux d'expressions moyens inférieurs à environ 10 protéines par cellule, le bruit d'expression génique est inversement proportionnel au niveau d'expression moyen (courbe rouge) tandis que pour des protéines dont le niveau d'expression est supérieur à 10 protéines par cellule, le niveau de bruit atteint un plateau et devient indépendant du niveau d'expression moyen des protéines (courbe bleu). La dépendance inverse s'explique par la composante intrinsèque du bruit puisqu'en accord avec les modèles stochastiques de l'expression génique. D'autre part, puisque les distributions Gamma s'ajustent bien avec les distributions du nombre de protéines entre cellules, le modèle à deux niveaux semble pour ces niveaux d'expression le modèle le mieux adapté pour décrire les données expérimentales. Ainsi, le bruit d'expression génique pour des niveaux d'expression inférieur à 10 protéines par cellule est dû aux caractères stochastiques de production et de dégradation des ARN messagers au travers du mécanisme de burst traductionnel. Lorsque le nombre de protéines devient trop important, les auteurs montrent grâce à une approche « deux-couleurs » que le bruit généré par les sources de bruit intrinsèque devient faible et est dominé par les sources de bruits extrinsèques. Ces sources globales de bruit affectent les gènes de la même manière, ce qui se traduit par un plateau (représenté en bleu), au-dessous duquel le niveau de bruit ne peut pas descendre. Ceci suggère que chaque protéine a au moins 30% de variation dans son niveau d'expression. Taniguchi et al. montrent que les distributions Gamma sont toujours adaptés pour décrire leur donnée. Cependant, les paramètres a et b ne peuvent plus être interprété comme respectivement la fréquence des bursts traductionnels et la taille de ces bursts.

Ainsi, deux régimes distincts du bruit en fonction de l'abondance moyenne des protéines sont mesurés à la fois chez *S. cerevisiae* et *E. coli.* Un régime dans lequel le bruit d'expression est inversement proportionnel avec l'abondance moyenne des protéines et dominé par les sources de bruit intrinsèque, et un régime à plus fort niveau d'expression où les sources de bruit extrinsèque dominent. Chez *E. coli* cependant, le plateau est atteint beaucoup plus rapidement, dès que l'on dépasse 10 protéines par cellule, alors que les effets du bruit intrinsèque chez la levure se font encore ressentir pour un nombre de protéines d'environ 1000. De plus, l'allure des distributions obtenues chez *E. coli* indique que pour des gènes naturellement présents dans le génome de *E. coli* et dont le niveau d'expression moyen est faible, le modèle à deux niveaux semble être adapté pour décrire le bruit d'expression. Enfin, si ce dernier résultat semble conforter les résultats d'Ozbudak et al quant à la validité du modèle à deux niveaux, les niveaux d'expression qu'ils visitent dans leur expérience sont nécessairement au-delà de la limite des 10 protéines par cellule. En effet, compte tenu de la résolution et de la sensibilité de la cytométrie en flux, il est difficile de quantifier les signaux en deçà de 10 protéines par cellule. Ainsi, dans la plage de niveaux d'expressions visitée dans leur expérience, la contribution des sources de bruit extrinsèque est

dominante et il est alors difficile de confronter les résultats expérimentaux avec les modèles de bruit intrinsèque.

QUESTION POSEE ET STRATEGIE ADOPTEE

Comme nous venons de le voir dans l'introduction, une modélisation simple du processus stochastique d'expression des gènes (modèle à deux niveaux) prédit que la transcription et la traduction ont une influence différente sur le bruit d'expression. Un gène fortement transcrit et faiblement traduit serait moins bruité qu'un gène faiblement transcrit et fortement traduit. Cette prédiction a des conséquences importantes à la fois en terme évolutifs et pour la biologie synthétique. Comme nous l'avons évoqué en introduction, le bruit d'expression peut avoir des effets néfastes sur le fonctionnement de la cellule, en l'éloignant de sa stœchiométrie optimale, ou au contraire des effets bénéfiques en permettant une diversification phénotypique. Les taux de transcription et de traduction étant modulables par la séquence des promoteurs et des TIR, ils peuvent être optimisés par sélection naturelle au cours de l'évolution. Ainsi, s'ils ont un effet différent sur le niveau de bruit d'expression, ce dernier peut également évoluer par sélection. Pour un niveau d'expression moyen fixé, la cellule pourrait donc évoluer vers un taux de transcription fort et une traduction faible pour minimiser le bruit, ou l'inverse dans le cas où la variabilité est bénéfique. De la même façon, en termes de biologie synthétique un contrôle du bruit d'expression pourrait être envisagé simplement par modification des séquences des promoteurs et des TIR. En particulier, dans un système de production d'un composé d'intérêt par un micro-organisme, une telle stratégie pourrait permettre de minimiser assez simplement le bruit d'expression pour optimiser le rendement. Comme nous l'avons vu, le modèle à deux niveaux présente plusieurs limitations importantes et une validation expérimentale de cette prédiction est donc essentielle. En terme expérimental, si le travail effectué chez *B. subtilis* par Ozbudak et al. semble confirmer la prédiction du modèle, ce n'est pas le cas des résultats chez la levure (Blake et al, 2002 Raser & O'Shea, 2004). De plus, Ozbudak et al. n'ont utilisé que 4 souches différentes et les promoteurs utilisés n'étaient pas constitutifs, alors que l'utilisation d'un régulateur peut avoir un impact important sur le bruit (Elowitz et al 2002). Enfin, il est important de souligner également que les sources de bruits extrinsèques sont dominantes pour une majeure partie du protéome de *E. coli*, organisme procaryote comme *B. subtilis*.

Dans ce travail de thèse nous cherchons à quantifier l'effet des taux de transcription et de traduction sur le bruit d'expression, chez la bactérie *B. subtilis*, en utilisant une banque d'environ 40 souches dans lesquelles le gène de la protéine fluorescente GFPmut3 a été inséré dans le chromosome et exprimé selon une combinatoire des taux de transcription et de traduction. Cette banque a été conçue et construite dans l'équipe de Matthieu Jules et Stéphane Aymerich par Vincent Sauveplane (Institut Micalis, INRA). Le taux de transcription est modulé au moyen de 8 promoteurs constitutifs naturellement présents chez *B. subtilis* et le

taux de traduction est contrôlé par 5 séquences TIR à la fois naturelles et synthétiques. La quantité de protéines fluorescentes est estimée sur cellules uniques à la fois par microscopie à épifluorescence et par cytométrie en flux. L'utilisation de ces deux techniques nous permet de contrôler les éventuels biais expérimentaux qui pourrait être introduits par chacune des deux techniques. A partir des signaux de fluorescence obtenus, le bruit est mesuré par le facteur Fano. Pour chaque TIR, les effets de la transcription sur le bruit d'expression génique peuvent être investigués en visitant les différents promoteurs tandis que pour chaque promoteur, les effets de la traduction sur la variabilité phénotypique sont observés en parcourant les différents modules TIR.

MATERIELS ET METHODES

Nous détaillons dans cette partie les éléments nécessaires à la mise en place des expériences qui nous permettront de mesurer le bruit d'expression génique. Nous voulons quantifier les effets de la traduction et de la transcription sur la stochasticité de l'expression génique. Pour cela, nous devons disposer d'une banque de la bactérie *B. subtilis* nous permettant d'isoler chacun des deux mécanismes comme l'ont fait précédemment Ozbudak et al (Ozbudak et al, 2002). Nous présenterons dans un premier temps la banque de mutants utilisée. Nous décrirons ensuite comment nous quantifions le signal de fluorescence à la fois par une approche de cytométrie en flux et de microscopie de fluorescence.

I Organisme d'étude : *Bacillus subtilis* :

Bacillus subtilis est une bactérie du sol appartenant au phylum Firmicute et constitue l'organisme modèle d'étude des bactéries à gram positif. Cette bactérie a une forme de bâtonnet cylindrique, ou bacille de diamètre 0,87 µm et de longueur moyenne 4,7 µm dans des conditions de croissance rapide (Weart et al, 2007). Comme pour la bactérie *E. coli*, le temps de doublement des cellules en milieu riche est de l'ordre de 20 minutes. L'ADN génomique code pour environ 4100 gènes (Kobayashi et al, 2003). La séquence complète du génome, ainsi que les réponses du transcriptome et du protéome à diverses conditions environnementales ont été déterminées (Nicolas et al, 2012 ; Guiziou et al, 2016 ; Borkowski et al, 2016). *B. subtilis* est utilisé comme organisme châssis pour la synthèse d'enzyme d'intérêt (Van Dijl & Hecker, 2013). En effet, d'un point de vue ingénierie, *B. subtilis* présente de nombreux avantages comme sa compétence naturelle (chez *E. coli* les cellules sont rendues compétentes artificiellement), et donc une intégration chromosomique d'ADN exogène facilitée. Sa capacité de sporulation facilite grandement les conditions de stockage : les spores peuvent être placés dans des aliquots ou conservés sur boîte de pétri scellé et conservé à température ambiante pour des temps très long. D'autre part, *B. subtilis* est une bactérie non pathogène, classifiée par l'administration américaine des denrées alimentaires (FDA) comme " généralement considéré comme sûr " et ayant obtenue le statut de "présomption d'innocuité reconnue" par l'Autorité européenne de sécurité des aliments (EFSA). Enfin, la destinée cellulaire de *B. subtilis* est d'autre part très sensible aux fluctuations stochastiques de l'expression génique : compétence, sporulation, ou encore cannibalisme (Losick & Desplan, 2008). *B. subtilis* constitue donc un organisme de choix utile (non pathogénicité, transfert de matériel génétique facilité, facilité de stockage) et nécessaire (biotechnologie et aspects fondamentaux) pour l'étude du bruit dans l'expression génique.

II Banque de mutant de *B. subtilis* :

Nous présentons ici la banque de souches que nous avons utilisée. Dans une première partie nous décrirons la banque initialement construite par Vincent Sauveplane (Institut Micalis, INRA). Puis nous détaillerons la construction de souches supplémentaires que nous avons effectuée pour compléter cette banque.

II.1 Banque originale de *B. subtilis* :

La banque est constituée d'environ 46 souches, dans lesquelles le gène de la protéine gfpmut3, une variante de la GFP avec un temps de repliement plus rapide (de l'ordre de quelques minutes), est inséré sur le chromosome de *B. subtilis* à un locus proche de l'origine de réplication du chromosome. Les promoteurs et séquences TIR utilisés pour contrôler l'expression sont détaillées ci-dessous.

- **Promoteurs :**

Comme nous l'avons vu en introduction, le taux de transcription dépend des séquences -35 et -10 mais aussi des quelques premiers nucléotides en aval du site "+1". Les promoteurs ont donc été définis ici comme les séquences comprenant les 50 premiers nucléotides en amont du site "+1", ainsi que les 8 premiers nucléotides transcrits. La zone du promoteur correspondant aux 8 premiers nucléotides transcrits est également nommé TSS (Transcription Start Site). Pour rappel, le site +1 indique la position du premier nucléotide transcrit. Cependant, pour certains promoteurs, l'ARN polymérase peut démarrer la phase d'élongation de l'ARN messager au niveau des positions +2 ou +3, phénomène observé à la fois chez *E. coli* et *B. subtilis*. (Guiziou et al, 2016). Il existe donc une incertitude sur la localisation du site d'initiation de la transcription, ce dernier se situant soit en position +1, +2 ou +3. Ainsi, pour rendre compte des effets du démarrage de la phase d'élongation dans la transcription, il a été décidé de "prolonger" les promoteurs des 8 premiers nucléotides transcrits. Au total 8 promoteurs constitutifs naturels ont été sélectionnés, avec différentes activités et qui ont été caractérisé dans des études précédentes (Klumpp et al, 2009 ; Nicolas et al, 2012 ; Borkowski et al, 2016). En particulier, leur efficacité a été estimée dans différentes conditions de croissance. Le tableau 1 présente les 8 promoteurs et leur dépendance avec le taux de croissance μ et le tableau 2 indique leurs séquences.

Nom du promoteur	Dépendance avec μ	Gène natif code pour :
fbaA	Constant	Fructose-1, 6-biphosphate aldolase
ykpA	Constant	Transporteur ABC (protéine de liaison avec l'ATP)
zwf	Constant	Glucose 6-phosphate déshydrogénase
yufK	Positif	Protéine hypothétique membranaire
rrnJP2	Positif	16S ARNr
ylxM	Positif	Protéine hypothétique
yqzM	Négatif	Protéine hypothétique
ykwB	Négatif	N-acétyltransférase

Tableau 1 : Liste des promoteurs constitutifs sélectionnés dans la construction de la banque de souches de *B. subtilis*. La première colonne correspond au nom du promoteur qui correspond au nom du gène natif auquel il est rattaché chez *B. subtilis*. La deuxième colonne indique comment varie la force du promoteur, et donc le nombre d'ARN messagers lorsque le taux de croissance μ augmente. Par exemple, s'il est indiqué « positif », cela signifie que l'abondance du nombre d'ARN messagers augmente avec le taux de croissance. La troisième colonne indique la protéine qui est codée par le gène auquel est rattaché le promoteur dans son contexte naturel.

Nom	Séquence
fbaA	AATCATGTCATTATGTTGCCGATTTGTCGAAAAGTTGGTATCCTAGTTAT **GGAGAAAA**
ykpA	TTTATCAGGGAATCATTTCTCTTGCCCTGCATTCATGGTATACTTTTATT **GATGATAG**
zwf	AAAAGGGCTTAAATGTTTGCTTTCGTTGAATTTTAGATTTAAAATGAAGG **AAATATAC**
yufK	TGGAATAATTTATCTTGTCAATGTGTTTTAAGTCCTCCATAATAAATGAG **GTAGATTA**
rrnJP2	CTTCAAAAAAAGTTATTGACTTCACTGAGTCAACGAGTTATAATAATAAA **GTCGCTTG**
ylxM	ACAAGATAAAAACTTGACAGTGTCATTAAAACCGTGTAAACTAAGTTATC **GTAAAGGG**
yqzM	AAAACGTTACCTTAGCTTCTAGATTTTTCAAAACAATTCCATTATACTTA **AGTCAGTC**
ykwB	GAAAATCGGAATATATTTACTGAAAATTCAATCTTCGTTATAATGAAACA **AGTCAGTC**

Tableau 2 : Séquences des différents promoteurs utilisés dans la banque de souches. En gras sont indiqués les 8 premiers nucléotides transcrits.

- **Régions d'initiations de la traduction (TIR) :**

Pour moduler le taux de traduction, on utilise 5 différents TIR qui se différencient entre eux essentiellement par la séquence du RBS et la longueur du TIR (Tableau 3). Ces différents TIR interviennent soit naturellement dans certains gènes de *B. subtilis* (fbaA, gltX, tufA) soit dérivent du TIR du gène fbaA (fbaAshort, fbaAhs). Le gène gltX code pour la protéine glutamyl-ARNt synthétase et le gène tufA pour la protéine EF-Tu, un facteur d'élongation qui joue un rôle essentiel dans la traduction. Ces TIR ont été caractérisés dans deux précédentes études (Guiziou et al, 2016 ; Borkowsky et al, 2016), montrant notamment que leur efficacité de traduction dépend du taux de croissance des

bactéries : plus le taux de croissance est important, plus l'efficacité de traduction sera faible.

TIR	Séquence
fbaA	AGCGATCTGAGTATTTACATATGACAGCAATATATGGGTCATGCTAGGG TGGAAAGCTTTTTTCGCTAGAAGACAATCAGGCTACAGGTGGGAAGGAG GACATTCGAC
fbaAshort	AGCGATCTGAGTATTTACATATGACAGCAATATATGGGTCATGCTAGGG TGGAAAGCTTTTTTCGCTAGAAGACAATCAGGCTACA**GGTGGGAAGGA GGAACTACT**
fbaAhs	AGCGGATAACAATTGGTGGGAAGGAGGACATTCGAC
gltX	TTGCCTGAGGCCATACATGACATGAAAGGAAGTATTTGAAA
tufA	AGGAAGTGAAAGCTTTCTTTCACTTCCTATCACTCTATACATTACTAATT AAAAGCTCTTAAGGAGGATTTTAGA

Tableau 3 : Les 5 différents modules TIR avec différentes efficacités de traduction utilisés dans la banque de *B. subtilis*. Pour chaque séquence (colonne de gauche) sont indiquées en noir la séquence nucléotidique qui se situe en amont du RBS, en bleu la séquence du RBS et en vert la séquence comprise entre le RBS et le codon d'initiation qui est AUG pour l'ensemble des constructions (non représenté). Les différences dans les séquences conduisent à des différences de taux de traduction. Par exemple, pour les TIR fbaA et fbaAshort, la seule différence consiste dans la taille et la nature de la séquence comprise entre le RBS et le codon d'initiation, le RBS et les bases en amont étant identique. A l'inverse, le TIR fbaAshort est différent du TIR fbaA au niveau de la zone en amont du RBS.

La combinaison de ces différents modules génétiques (8 promoteurs et 5 modules TIR) donne théoriquement un nombre de 40 souches, auxquels s'ajoutent 7 autres souches pour lesquelles le module TIR est constant et couplé à 7 séquences promotrices qui ne diffèrent que sur les 7 derniers nucléotides, soit la partie TSS. Pour ces dernières constructions, la séquence des 50 premiers nucléotides est celle du promoteur rrnJP2, les séquences TSS correspondent aux séquences TSS des autres promoteurs autres que rrnJP2, et le module TIR est fbaAshort. En pratique, la banque contient exactement 44 souches, les combinaisons {ylxM-tufA}, {yqzM-fbaAshort} et {rrnJP2-ykpA-fbaAshort} n'ayant pas réussi à être assemblés ou à être intégré correctement dans le génome de *B. subtilis*. La liste complète des souches est donnée dans les tableaux 4 et 5.

Nom de la souche	Assemblage Promoteur-TIR	Nom de la souche	Assemblage Promoteur-TIR
PL1S01	fbaA-fbaA	PL1S21	ylxM-fbaA
PL1S02	fbaA-fbaAhs	PL1S22	ylxM-fbaAhs
PL1S03	fbaA-fbaAshort	PL1S23	ylxM-fbaAshort
PL1S04	fbaA-gtlX	PL1S24	ylxM-gtlX
PL1S05	fbaA-tufA	PL1S25	ylxM-tufA
PL1S06	rrnJP2-fbaA	PL1S26	yqzM-fbaA

PL1S07	rrnJP2-fbaAhs	PL1S27	yqzM-fbaAhs
PL1S08	rrnJP2-fbaAshort	PL1S28	yqzM-fbaAshort
PL1S09	rrnJP2-gtlX	PL1S29	yqzM-gtlX
PL1S10	rrnJP2-tufA	PL1S30	yqzM-tufA
PL1S11	ykpA-fbaA	PL1S31	yufK-fbaA
PL1S12	ykpA-fbaAhs	PL1S32	yufK-fbaAhs
PL1S13	ykpA-fbaAshort	PL1S33	yufK-fbaAshort
PL1S14	ykpA-gtlX	PL1S34	yufK-gtlX
PL1S15	ykpA-tufA	PL1S35	yufK-tufA
PL1S16	ykwB-fbaA	PL1S36	zwf-fbaA
PL1S17	ykwB-fbaAhs	PL1S37	zwf-fbaAhs
PL1S18	ykwB-fbaAshort	PL1S38	zwf-fbaAshort
PL1S19	ykwB-gtlX	PL1S39	zwf-gtlX
PL1S20	ykwB-tufA	PL1S40	zwf-tufA

Tableau 4 : Ensemble des souches constituant la banque originale de *B. subtilis*. Le nom de chaque souche est suivi de l'assemblage {Promoteur-TIR} qui lui correspond qui a été placé en amont de la séquence codante de la protéine GFPmut3. Les souches indiquées en rouge n'ont pas été construite.

Nom de la souche	Assemblage Promoteur[-50,-1]-TSS-TIR	Nom de la souche	Assemblage Promoteur[-50,-1]-TSS-TIR
PL1B01	rrnJP2-fbaA-fbaAshort	PL1B02	rrnJP2-ykpA-fbaAshort
PL1B03	rrnJP2-ykwB-fbaAshort	PL1B04	rrnJP2-ylxM-fbaAshort
PL1B05	rrnJP2-yqzM-fbaAshort	PL1B06	rrnJP2-yufK-fbaAshort
PL1B07	rrnJP2-zwf-fbaAshort		

Tableau 5 : Souches constituant la banque originale de *B. subtilis* pour lesquelles seul change la séquence TSS. Le nom de chaque souche est suivi de l'assemblage {Promoteur[-50,-1]-TSS-TIR} qui lui correspond qui a été placé en amont de la séquence codante de la protéine GFPmut3. La souche indiquée en rouge n'a pas été construite.

Nous devrons compléter cette banque de souches avec de nouvelles constructions. Une première série (1) de construction nous servira de contrôle et chaque souche de cette série présentera une séquence TSS unique (TSS du promoteur fbaA) et un module TIR unique (fbaAshort) assemblés avec les différentes séquences des 50 premiers nucléotides en amont du site +1 des promoteurs de la banque originale excepté le promoteur fbaA qui est redondant

avec la banque originale. Nous aurons donc au total 7 nouvelles constructions que nous nommerons Alex[01-07] (Tableau 6). La deuxième série de construction (2) consistera à intégrer au génome de B. subtilis un deuxième fluorophore : la protéine mKate2 dans un locus voisin de celui de la GFPmut3. Les modules synthétiques en amont de la partie codante pour le fluorophore mKate2 devront être les mêmes que pour le fluorophore GFPmut3. Ces modules seront ceux construits pour la série (1), à savoir différentes séquences [-50,-1] du promoteur assemblé avec un TSS unique et un module TIR unique (respectivement fbaA et fbaAshort). Ces nouvelles constructions seront nommées Alex[01-07]-mKate. En parallèle de ces constructions, nous devrons vérifier que les séquences des modules présents dans la banque originale ne sont pas mutées. Nous présentons les techniques de biologie moléculaire utilisées pour la réalisation des nouvelles constructions et pour le séquençage des constructions originales dans la partie suivante.

Nom de la souche	Assemblage Promoteur [-50,-1]-TSS-TIR
Alex01	rrnJP2-fbaA-fbaAshort
Alex02	ykpA-fbaA-fbaAshort
Alex03	ykwB-fbaA-fbaAshort
Alex04	ylxM-fbaA-fbaAshort
Alex05	yqzM-fbaA-fbaAshort
Alex06	yufK-fbaA-fbaAshort
Alex07	zwf-fbaA-fbaAshort

Tableau 6 : Assemblages génétiques réalisées pour la construction des nouvelles souches (Alex [01-07]).

II.2 Constructions de nouvelles souches :

Les inserts sont synthétisés dans un premier temps sous forme de plasmide à l'aide de réactions de PCR où on génère plusieurs fragments qui sont ensuite assemblées par un assemblage Gibson. Les plasmides formés sont utilisés pour transformer dans un premier temps *E. coli* qui nous sert d'organisme châssis pour amplifier le nombre de plasmides. Ces plasmides sont ensuite extraits et purifiés puis utilisés pour transformer *B. subtilis*. Les plasmides présentent une région d'homologie avec le chromosome de *B. subtilis* permettant son intégration dans un locus spécifique du chromosome, ainsi que des gènes de résistance à l'ampicilline et à la spectinomycine. La démarche expérimentale est la suivante :

- Réactions de PCR
- Visualisation des produits de PCR sur gel d'électrophorèse

- Purifications de fragments
- Assemblage Gibson
- Transformation d'*E. coli*
- Extraction/purification de plasmides
- Séquençage des plasmides
- Transformation de *B. subtilis*

Pour le séquençage des anciennes constructions, la zone d'intérêt {Promoteur-TIR-GFP} est sélectionnée et amplifiée grâce à une PCR sur colonie. Les produits de PCR sont visualisés sur gel d'électrophorèse puis purifiés et séquencés.

II.2.1 Réactions de polymérisation en chaîne (PCR) :

La technique de réaction de polymérisation en chaîne ou PCR ("Polymerisation Chain Reaction") consiste à réaliser un nombre très élevé de copies d'une séquence d'ADN d'intérêt. Une ADN polymérase est utilisée pour catalyser la réaction de polymérisation à partir d'amorces complémentaire de chaque brin d'ADN à amplifier (une amorce sens ou *forward* et une amorce anti-sens ou *reverse*). L'enzyme allonge les amorces dans le sens 5'→3'. A partir de chaque amorce un brin antiparallèle se forme qui présentera à son extrémité opposée une séquence complémentaire de l'autre amorce. Après plusieurs cycles de PCR, la séquence comprise entre les amorces initiales se trouve ainsi amplifiée un grand nombre de fois : le nombre de fragments amplifiés après n cycles est égale à 2^n. Les réactions de PCR ont lieu dans un thermocycleur qui permet un contrôle de la température du milieu réactionnel pour chaque étape (Figure 17). Plusieurs types de PCR vont être réalisées pour réaliser les constructions et pour le séquençage. Le *design* des différentes amorces (ou oligonucléotides) utilisées a été réalisé grâce au logiciel *Geneious*. La synthèse de ces amorces a été effectuée par Eurofins Genomics. Les amorces se trouvent sous forme de poudre que l'on réhydrate avec de l'eau stérile de telle sorte que la concentration en amorce soit de 100 pmol/µL. Les différentes ADN polymérases utilisées sont stockées à -20°C. Nous présentons tour à tour les différentes PCR utilisées.

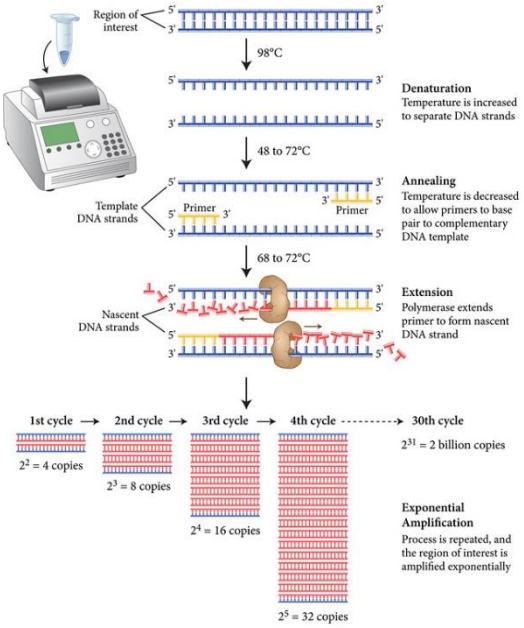


Figure 17 : Schéma récapitulatif du Principe de la PCR (Neb.com). Il y est notamment indiqué les différentes étapes thermiques contenues dans un cycle de PCR.

- **PCR sur colonie :**

La PCR sur colonie est une méthode à haut débit qui permet de déterminer la présence ou l'absence d'ADN d'insertion dans les constructions génomiques, ou encore d'isoler un fragment si on est sûre de sa présence. Les cellules d'un clone isolé de *E. coli* ou *B. subtilis* sur une boîte de pétri sont sélectionnées et placées dans le milieu réactionnel de PCR. Les cellules sont lysées pendant l'étape initiale de dénaturation à 98°C. Cette première étape de chauffage provoque la libération de l'ADN génomique, de sorte qu'il peut servir de brin matrice pour la réaction d'amplification. Des amorces conçues pour cibler spécifiquement l'ADN de l'insert peuvent être utilisées pour déterminer si la construction contient le fragment d'ADN d'intérêt. Ce type de PCR sera effectué pour isoler et amplifier les séquences {Promoteur-TIR-GFPmut3} de chaque souche de la banque originale de *B. subtilis* afin d'être séquencées. Nous utiliserons pour cela les amorces VS19 (*forward*) et VS20 (*reverse*). L'amorce VS19 s'hybride en amont du gène GFPmut3 et en aval de la zone d'homologie tandis que l'amorce VS20 s'hybride en aval du gène GFPmut3. Ce type de PCR sera également effectué sur les cellules transformées de *E. coli* et *B. subtilis* avec les nouvelles constructions pour nous assurer de leur insertion dans les plasmides obtenus par assemblage Gibson et dans le chromosome de *B. subtilis* respectivement avec un choix judicieux d'amorces. Pour la première série de construction (souches Alex[01-07]), les PCR de vérifications peuvent aussi être réalisées avec les couples d'amorces VS19 et VS20, mais également avec les couples d' amorces VS[58-

64]/gfp-middle-rev. Chaque amorce *forward* VS[58-64] s'hybride spécifiquement dans la séquence promotrice tandis que l'amorce *reverse* gfp-middle-rev s'hybrident dans la partie codante de la GFPmut3. Par exemple, l'amorce VS58 s'hybrident dans le promoteur rrnJP2. Ainsi, si on obtient un produit de PCR en utilisant ces couples d'amorce, cela signifiera que l'insert est bien présent dans le génome de *B. subtilis* ou dans le plasmide recombinant obtenu par assemblage Gibson. Pour la seconde série de constructions (souches Alex[01-07]-mKate), on utilise les couples d'amorces VS24/gfp-middle-rev et/ou VS19/gfp-middle-rev pour contrôler la présence de la cassette d'expression GFPmut3. Le premier couple permet de s'assurer que la zone {Promoteur-TIR-GFPmut3} est bien intégré au chromosome et qu'il se situe au locus désiré (l'oligonucléotide VS24 s'hybride en amont de la zone d'homologie). Le deuxième couple sera utilisé pour vérifier la présence de la zone {Promoteur-TIR-GFPmut3} sur les constructions plasmidiques. Pour cette même série de construction, on utilise le couple d'amorce VS85/VS86 permettant d'amplifier la zone {Promoteur-TIR-mKate2} pour nous assurer que cette séquence et donc le fluorophore mKate2 est bien intégré aux constructions plasmidiques et au chromosome de *B. subtilis*. Pour toutes les PCR sur colonie, nous utilisons l'ADN polymérase Taq (DreamTaq Green PCR Master Mix (2X), Thermo Scientific), isolée à partir de la bactérie *Thermus aquaticus* qui possède un taux d'erreur relativement élevé. Les fragments générés lors de ces PCR n'étant pas attendus pour être long, les risques de mutations sont relativement rares et on privilégiera cette polymérase. La solution DreamTaq Green PCR Master Mix (2X) de chez Thermo Scientific contient les dNTPs (désoxyribonucléotides), les polymérases, un tampon, et un tampon de charge pour la visualisation des produits de PCR sur gel d'électrophorèse. Pour 50µL de volume réactionnel, le mélange se compose de :

- 0,25 µL amorce *forward*
- 0,25 µL amorce *reverse*
- 25 µL DreamTaq 2X Master Mix
- Cellules transformées
- 24,5 µL eau

Les différentes amorces utilisées pour les PCR sur colonies sont résumées dans le tableau 7.

Nom	Séquence
GFPmiddle-Rv	CTTCTTTAAAATCAATACCTTTTAACTCG
VS19	CCCGCTCATTAGGCGGGCTG
VS20	GCGGCAACCGAGCGTTCTGA
VS24	GCGGTCGGCGCAGGTATAGG
VS85	CCTTGCATAGGGGGATCTCG
VS86	AAAGTCTGGAATCCCCTGCG
VS58	TGACTTCACTGAGTCAACGAGT
VS59	TCTTGCCCTGCATTCATGGT
VS60	ACTGAAAATTCAATCTTCGTT

VS61	TGACAGTGTCATTAAAACCGTGT
VS62	ACGTTACCTTAGCTTCTAGATTTTTCA
VS63	TGTCAATGTGTTTTAAGTCCTCCA
VS64	AGGGCTTAAATGTTTGCTTTCGT

Tableau 7 : Séquences des amorces utilisées pour les PCR sur colonie

- **PCR effectuée pour la construction de la série (1):**

On procède à une amplification séparée de la séquence {Promoteur[-50,-1]-TSS-TIR} et du corps du plasmide avec le plasmide PL1S03 comme matrice qui présente le TSS du promoteur fbaA et le TIR fbaAshort (Figure 18).

- Amplification N°1 : On amplifie le corps du plasmide par le couple d'amorce VS57/P-PS-AM avec l'ADN polymérase Invitrogen Platinum SuperFi (Invitrogen Platinum SuperFi DNA Polymerase, ThermoFisher) dont le niveau de fidélité est environ 100 fois supérieur à la polymérase Taq. Cette polymérase est contenue dans un mélange réactionnel contenant un tampon spécifique ainsi que les différents dNTPs. Pour 50µL de volume réactionnel, le mélange réactionnel se compose de :
  - 0,25 µL amorce *forward*
  - 0,25 µL amorce *reverse*
  - Quantité variable de matrice PL1S03 (selon la concentration)
  - 25 µL Invitrogen Platinum SuperFi DNA Polymerase 2X MasterMix
  - Eau jusqu'à 50 µL

- Amplification N°2 : On amplifie la cassette d'expression GFPmut3, c'est à dire l'ensemble {Promoteur[-50,-1]-fbaA-fbaAshort} avec les couples de primers VS49-55/VS56 et avec l'ADN polymérase hautement fidèle Q5 (Q5® High-Fidelity DNA Polymerase, NEB). Ces amorces s'hybrident au niveau de la séquence TSS fbaA du plasmide PL1S03 matrice et en amont de cette séquence d'hybridation se trouve les différentes séquences promotrices [-50,-1] ainsi qu'une région d'homologie nécessaire pour l'assemblage Gibson. Pour 50µL de volume réactionnel, le mélange réactionnel se compose de :
  - 10 µL 5XQ5 Reaction Buffer
  - 1 µL 10 mM dNTPs
  - 0,25 µL amorce *forward*
  - 0,25 µL amorce *reverse*
  - Quantité variable de matrice PL1S03 (selon la concentration)
  - 0,5 µL polymérase Q5
  - Eau jusqu'à 50µL

Les deux fragments linéaires amplifiés à partir de ces PCR présentent des zones d'homologie sur leur extrémité permettant de les assembler par la suite grâce à un

assemblage Gibson. Ces zones d'homologie font environ une vingtaine de paires de bases.

Les différentes amorces utilisées pour réaliser ces PCR à partir du plasmide matrice sont résumées dans le tableau 8.

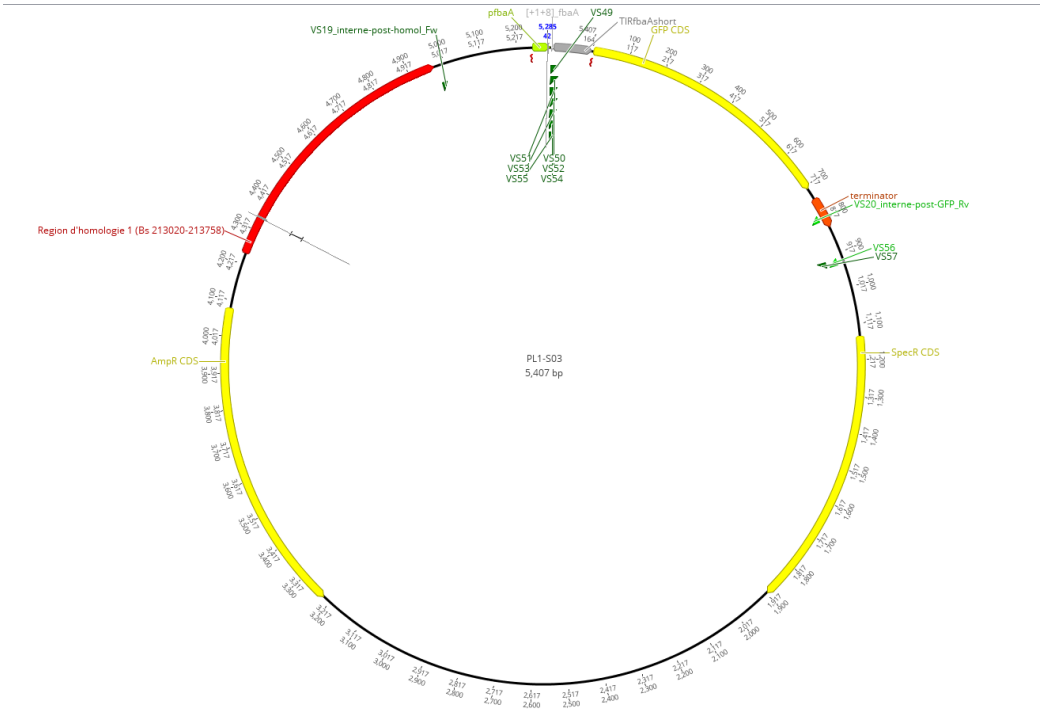


Figure 18 : Carte du plasmide PL1S03. Les différentes amorces pour la première série de construction y sont indiquées. Y sont indiquée également la zone d'homologie pour l'intégration dans le chromosome de *B. subtilis* par simple cross-over, la cassette d'expression GFPmut3 avec l'assemblage {fbaA-fbaAshort} ainsi que les cassettes d'expression de résistance aux antibiotiques ampicilline et spectinomycine.

	Zone amplifiée	Matrice	Séquence
VS49	{rrnJP2-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCCTTCAAAAAAAGTTATTGACTTCACTGAGTCAACGAGTTATAATAATAAAGGAGAAAAAGCGATCTGAGTATTTAC
VS50	{ykpA-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCTTTATCAGGGAATCATTTCTCTTGCCCTGCATTCATGGTATACTTTTATTGGAGAAAAAGCGATCTGAGTATTTAC
VS51	{ykwB-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCGAAAATCGGAATATATTTACTGAAAATTCAATCTTCGTTATAATGAAACAGGAGAAAAAGCGATCTGAGTATTTAC
VS52	{ylxM-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCACAAGATAAAAACTTGACAGTGTCATTAAAACCGTGTAAACTAAGTTATCGGAGAAAAAGCGATCTGAGTATTTAC
VS53	{yqzM-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCAAAACGTTACCTTAGCTTCTAGATTTTTCAAAACAATTCCATTATACTTAGGAGAAAAAGCGATCTGAGTATTTAC
VS54	{yufK-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCTGGAATAATTTATCTTGTCAATGTGTTTTAAGTCCTCCATAATAAATGAGGGAGAAAAAGCGATCTGAGTATTTAC
VS55	{zwf-fbaA-fbaAshort-GFPmut3} *Forward*	PL1S03	TACCGCGGGCTTTCCCAGCAAAAGGGCTTAAATGTTTGCTTTCGTTGAATTTTAGATTTAAAATGAAGGGGAGAAAAAGCGATCTGAGTATTTAC
VS56	{Promoteur-fbaAshort-GFPmut3} *reverse* (universel)	PL1S03	ATCTCTTGCCAGTCACGTTACG

VS57	Corps de plamide *forward*	PL1S03	GTAACGTGACTGGCAAGAG
P-PS-AM	Corps de plamide *reverse*	PL1S03	GCTGGGAAAGCCCGCGGTAAAAGTCGACG

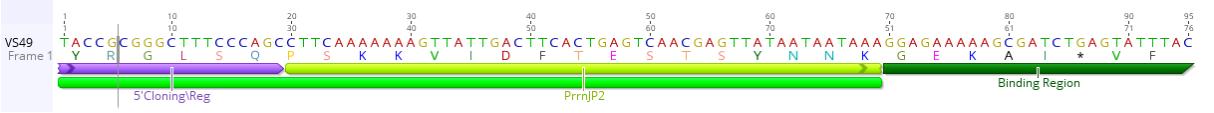


Tableau 8 : Amorces utilisées pour la première série de construction. Les séquences en violettes correspondent à la zone d'homologie permettant une recircularisation par assemblage Gibson, en vert la séquence des 50 nucléotides du promoteur en amont du site +1, et en noir la région d'hybridation sur le plasmide matrice PL1S03 correspondant à la séquence du TSS fbaA.

- **PCR effectuée pour la construction de la série (2):**

Les plasmides obtenus lors de la construction de la série (1) sont utilisés pour la construction des plasmides présentant les deux fluorophores mKate2 et GFPmut3 pilotés par les mêmes séquences {Promoteur[-50,-1]-TSS-TIR}. On procède ici à l'amplification séparée de 3 fragments.

- Amplification N°1 : Amplification des "corps de plasmide" à partir des plasmides matrices Alex[01-07] et PL1S03 avec les couples d'amorces VS72/VS73. Les corps de plasmide contiennent l'ensemble {Promoteur[-50,-1]-fbaA-fbaAshort-GFPmut3} ainsi que la zone d'homologie et les cassettes de résistance à l'ampicilline et à la spectinomycine. On utilise l'ADN polymérase hautement fidèle Q5 (Q5® High-Fidelity DNA Polymerase, NEB).

- Amplification N°2 : Amplification de la région {Promoteur[-50,-1]-fbaA-fbaAshort} sur les plasmides matrices Alex[01-07] et PL1S03 par les amorces VS[76-83]/VS84. Le fragment généré présente une extrémité homologue avec le corps du plasmide issu de l'amplification N°1.

- Amplification N°3 : Amplification de la séquence codante de la protéine mKate2 sur plasmide matrice pSG15 avec le couple d'amorces VS74/VS75 et l'ADN polymérase hautement fidèle Q5 (Q5® High-Fidelity DNA Polymerase, NEB). Le fragment généré présente une extrémité homologue avec le corps de plasmide de l'amplification N°1 et une autre extrémité homologue avec le TIR fbaAshort de l'amplification N°2.

Les différentes amorces utilisées pour réaliser ces PCR à partir du plasmide matrice sont résumées dans le tableau 9

	Zone amplifiée	Matrice	Séquence
VS72	Corps de plasmide {Promoteur-fbaAshort-GFPmut3} *reverse*	PL1S03 Alex[01-07]	CATGATTACGCCAAGCTCGC
VS73	Corps de plasmide {Promoteur-fbaAshort-GFPmut3} *forward*	PL1S03 Alex[01-07]	GTCATAGCTGTTTCCTGTGTG
VS74	{mKate2} *forward*	pSG15	GGTGGGAAGGAGGAACTACTATGTCAGAACTAATCAAAGAGAATATGCAC
VS75	{ mKate2} *reverse*	pSG15	CACACAGGAAACAGCTATGACTATAAACGCAGAAAGGCCCACC
VS76	{rrnJP2-fbaA-fbaAshort } *forward*	Alex01	GCGAGCTTGGCGTAATCATGCTTCAAAAAAAGTTATTGACTTCACTGAGTC
VS77	{ykpA-fbaA-fbaAshort } *forward*	Alex02	GCGAGCTTGGCGTAATCATGTTTATCAGGGAATCATTTCTCTTGCCC
VS78	{ykwB-fbaA-fbaAshort } *forward*	Alex03	GCGAGCTTGGCGTAATCATGGAAAATCGGAATATATTTACTGAAAATTCAATCTTC
VS79	{ylxM-fbaA-fbaAshort } *forward*	Alex04	GCGAGCTTGGCGTAATCATGACAAGATAAAAACTTGACAGTGTCATTA
VS80	{yqzM-fbaA-fbaAshort } *forward*	Alex05	GCGAGCTTGGCGTAATCATGAAAACGTTACCTTAGCTTCTAGATTTTTC
VS81	{yufK-fbaA-fbaAshort } *forward*	Alex06	GCGAGCTTGGCGTAATCATGTGGAATAATTTATCTTGTCAATGTGTTTTAAG
VS82	{zwf-fbaA-fbaAshort } *forward*	Alex07	GCGAGCTTGGCGTAATCATGAAAAGGGCTTAAATGTTTGCTTTCG
VS83	{fbaA-fbaA-fbaAshort } *forward*	PL1S03	GCGAGCTTGGCGTAATCATGAATCATGTCATTATGTTGCCGATTTGT
VS84	{Promoteur-fbaAshort } *reverse* (universel)	PL1S03 Alex[01-07]	AGTAGTTCCTCCTTCCCACC

Tableau 9 : Amorces utilisées pour la deuxième série de construction.

- **Programmation du thermocycleur:**

Le mélange réactionnel est d'abord chauffé à 96°C pendant 5 minutes afin de dénaturer les brins d'ADN et de lyser les cellules dans le cas des PCR sur colonie, puis soumis à trente cycles d'amplification. Un cycle d'amplification comprend 15 s de dénaturation à 96°C, 30s d'hybridation des amorces à 54°C-60°C (selon la température optimale d'hybridation des amorces utilisées), et un temps d'élongation à 72°C qui dépend de la taille de l'amplicon à obtenir et de la vitesse de travail de l'ADN polymérase (l'ADN polymérase Q5 a une vitesse d'élongation de 20-30s par kilobase (kb) tandis que l'ADN polymérase DreamTaq travaille à 1 min par kb). La réaction est achevée par une élongation finale de 5 min à 72°C

Les amplicons obtenus au cours des PCR sont contrôlés en fonction de leur taille par électrophorèse sur gel d'agarose. La taille attendue des fragments est donnée par le logiciel *Geneious.*

II.2.2 Electrophorèse sur gel d'agarose :

Cette technique permet de séparer les fragments d'ADN en fonction de leur taille. Du tampon de charge est additionnée si besoin à la solution de produits de PCR. Pour les PCR réalisées avec la polymérase DreamTaq, le tampon de charge est déjà incorporé dans la solution contenant la polymérase et les dNTPs, ce qui n'est pas le cas des PCR réalisées avec les polymérases Platinum SuperFi et Q5. Pour ces dernières, il faudra donc ajouter le tampon de charge à la solution

d'ADN. Le tampon de charge ajoute de la couleur à l'échantillon, ce qui simplifie le processus de chargement sur gel d'agarose et permet de contrôler la migration lors de l'application du champ électrique. On prépare un gel d'agarose à 0,7% (w/v) (UltraPure Agarose, invitrogen) (le solvant est du tampon TAE) dans lequel on incorpore du Midori Green (intercalant de l'ADN) (1µL de Midori Green pour 10 mL de gel). On laisse le mélange se solidifier pendant ~30 min à température ambiante. Une fois solidifié, on charge le gel avec la solution de produit de PCR. Pour 50 µL de produit de PCR, on charge environ 5µL dans les puits du gel. On charge également 8 µL d'une solution "étalon" contenant des fragments d'ADN de longueurs connues (SmartLadder MW-1700-10, Eurogentec) permettant d'estimer la taille de nos produits de PCR. L'électrophorèse est effectuée dans du tampon TAE sous une tension de 125 V pour une durée de ~30 minutes (la durée et le voltage dépendent de la taille attendue des amplicons). L'ADN, coloré par le Midori Green intercalé entre les plateaux de bases, est visualisé sous ultraviolets. La taille des fragments d'ADN est déterminée par comparaison avec la migration de la solution "étalon".

II.2.3 Digestion par l'enzyme de restriction DpnI :

L'enzyme DpnI coupe spécifiquement les ADN méthylés, ce qui est le cas des plasmides utilisés dans les réactions de PCR puisque provenant de transformations d'*E. coli*, mais pas des produits de PCR. Un traitement des produits de PCR par l'enzyme DpnI permet donc d'éliminer les plasmides matrice de PCR. Les plasmides matrices étant transformants, alors lors de la transformation d'*E. coli* par nos nouvelles constructions, nous pourrions obtenir des "faux positifs" provenant de la transformation d'*E. coli* par ces plasmides. Ils doivent donc être éliminé. La digestion par une enzyme de restriction se fait selon les indications du fournisseur de l'enzyme. Nous utilisons l'enzyme DpnI de chez NEB (stockage à -20°C). Pour chaque 50 µL de produits de PCR, on ajoute 0,5 µL de la solution de DpnI. Le mélange est placé à 37°C sans agitation pendant une nuit.

II.2.4 Purification des produits de PCR traité à la DpnI :

Après traitement à la DpnI, les produits de PCR sont purifiés afin d'éliminer les amorces et les dNTPs notamment ainsi que les débris cellulaires pour les PCR sur colonie. Cette étape est importante pour optimiser la réaction de séquençage, mais aussi les transformations à suivre. Pour la purification, on utilise le kit de purification des produits de PCR QIAquick® PCR Purification Kit de chez QIAGEN en suivant les instructions du fabricant.

II.2.5 Quantification de la concentration en ADN :

Une fois les produits de PCR purifiés, on procède à la quantification de la concentration en ADN en utilisant un spectrophotomètre Nanodrop de chez Thermo Scientific qui permet de quantifier rapidement la concentration en ADN à partir d'un échantillon de 1,5 µL de solution à titrer. Pour les acides nucléiques, le spectrophotomètre mesure l'absorbance à la longueur d'onde 260 nm et en déduit la concentration. La pureté de la solution en acide nucléique est estimée à l'aide des rapports de l'absorbance mesurée à 260 nm, avec l'absorbance mesurée à 280 nm et 230 nm. D'après les données du constructeur, une solution en ADN sera considérée comme pure si le premier rapport est aux alentours de 1,80 et le second autour de 2,0. Les concentrations obtenues sont importantes pour l'assemblage Gibson et également pour la préparation des échantillons pour le séquençage. Les concentrations sont données en ng/µL.

II.2.6 Clonage des inserts dans les corps de plasmide par assemblage Gibson :

L'assemblage Gibson est une méthode d'assemblage de multiples fragments d'ADN ayant des extrémités homologues qui ne nécessite pas d'enzyme de restriction (Gibson et al, 2009). L'assemblage se fait à température constante. L'assemblage est réalisé par trois enzymes différentes : des 5'exonucléases, des ADN ligases et des ADN polymérases. Les fragments d'ADN sont conçus pour avoir des séquences de 15-25 pb à leurs extrémités identiques avec les extrémités des fragments auxquels ils doivent être assemblés (Figure 19-A). Dans un premier temps, la 5'exonucléase digère les extrémités 5' de deux fragments à ligaturer et crée ainsi deux extrémités cohésives. Ces deux extrémités cohésives débordantes s'associent ensuite grâce aux liaisons hydrogènes entre séquences complémentaires. L'ADN polymérase incorpore les nucléotides pour combler les vides et étend ainsi les extrémités 3'. Enfin les ADN ligases lie de manière covalente les fragments d'ADN en créant des liaisons phosphodiester 3'-5' au niveau des jonctions. Cette procédure crée un ADN double brin. C'est cette méthode que nous allons utiliser pour assembler les plusieurs fragments obtenus par PCR. Pour la première série de construction, nous avons 2 fragments à assembler et pour la deuxième série 3 fragments à assembler. Pour réaliser ces assemblages Gibson, on utilise le kit Hifi DNA Assembly Master Mix (2X) de chez NEB. Le kit est stocké à -20°C. D'après les instructions du fabricant, lorsqu'on veut assembler 2 ou 3 fragments, les rapports entre les corps de plasmide (vecteurs de clonage) et les fragments à y cloner (inserts) est de 1 pour 2. Pour chaque assemblage Gibson, on travaille dans un volume total de 10µL. La quantité totale de fragments dans le volume réactionnel doit être de 0,1 pmol, ce qui représente un volume X (en µL). On y ajoute 5 µL de Hifi DNA Assembly Master Mix (2X). On complète avec un volume 5-X µL d'eau stérile. Pour calculer les quantités optimales de fragments, on convertit les concentrations de

chaque fragment obtenu avec le dispositif Nanodrop en ng/µL en pmol/µL selon la formule (N représente le nombre de paires de bases du fragment) :

$$\text{Concentration en pmol/µL} = (\text{Concentration en ng/µL}) \times 1000 / (N \times 650 \text{ daltons})$$

Le mélange des différents constituants est réalisé dans des aliquots (un aliquot pour chaque assemblage) placés dans la glace. Une fois les mélanges réalisés,

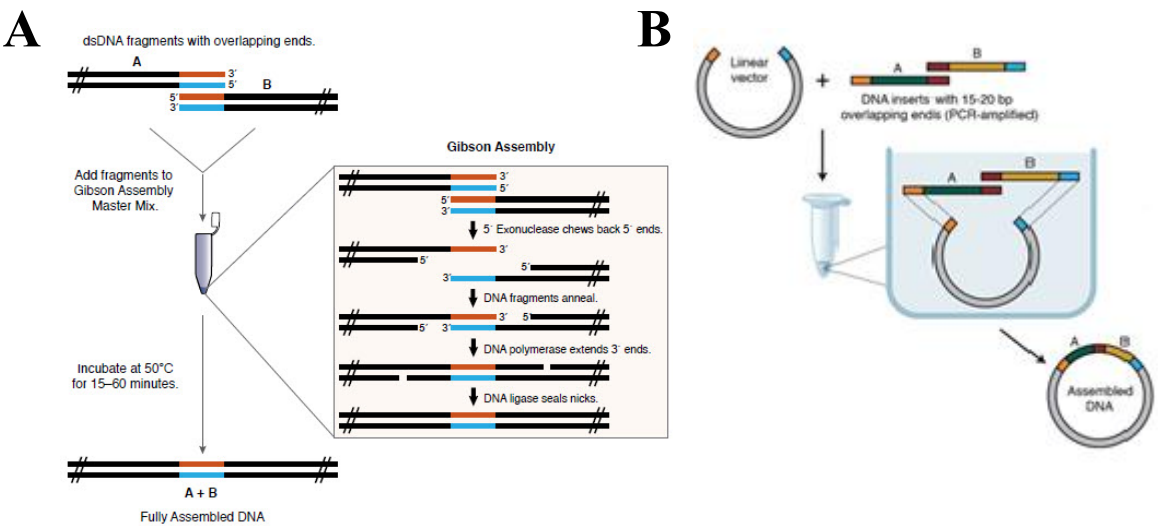


Figure 19 : Assemblage Gibson (neb.com). **A)** Schéma descriptif des différentes étapes biochimiques de l'assemblage Gibson. **B)** Illustration du clonage de deux fragments linéaires dans un vecteur de clonage linéaire. Chaque fragment possède des zones d'homologies qui permettent une recircularisation et un assemblage par assemblage Gibson. Cette dernière illustration représente la construction d'un plasmide tel que nous l'avons fait pour la deuxième série de construction. Le vecteur linéaire représente le fragment appelé dans le texte principal corps de plasmide. Les fragments A et B sont respectivement la séquence codante de la protéine mKate2 et la zone {Promoteur-TIR}. Après la réaction, on doit obtenir un plasmide circulaire qui est représenté en figure 20.

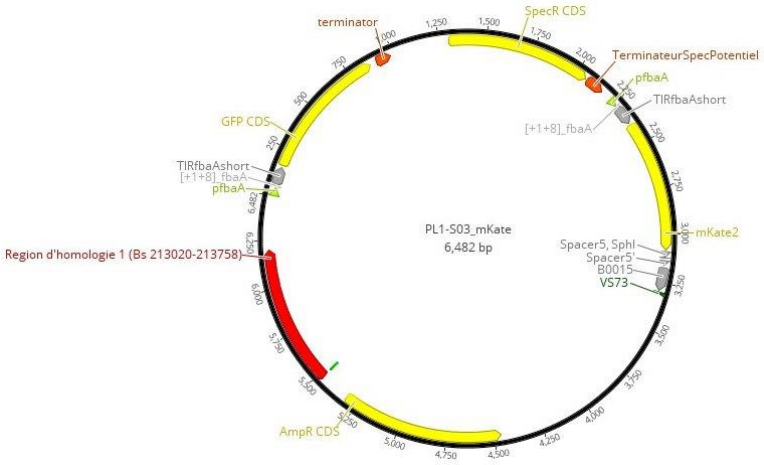


Figure 20 : Plasmide PL1S03-mKate attendu après recircularisation par assemblage Gibson. Les fluorophores GFPmut3 et mKate2 sont contrôlés par les mêmes séquences promotrices et modules TIR.

l'ensemble est placé 30 min dans un bain marie à 50°C. Cette opération nous permet d'obtenir à partir de fragments linéaires (produits de PCR) des plasmides circulaires pouvant être transformé (ADN recombinant) chez *E. coli* (Figure 19).

II.2.7 Transformation d'*E. coli* Mach1 :

Cette étape est utilisée pour amplifier les différents plasmides synthétisés par assemblage Gibson. La transformation est un processus qui apparaît lorsqu'une cellule ingère un ADN étranger contenue dans son environnement. Ces cellules bactériennes capable de capturer de l'ADN depuis leur environnement sont dites compétentes. La transformation se repose dans notre cas sur une désorganisation de la paroi bactérienne par traitement chimique (développement de la compétence) suivie d'un choc thermique (introduction de l'ADN recombinant dans les cellules). Ainsi, on transforme des cellules chimio-compétentes de la souche Mach1 de *E. coli* par choc thermique avec l'ADN plasmidique recombinant synthétisé par assemblage Gibson. Les cellules compétentes sont stockées à -80°C dans des aliquots de 1.5mL contenant 200µL de suspension de cellules. Du fait de leur compétence, ces cellules sont extrêmement fragiles et les aliquots sont placés dans un bac de glace pour éviter tout choc thermique lors de la sortie du congélateur à -80°C. Pour chaque plasmide construit, on mélange dans un aliquot de 1,5 mL, 5µL de produit d'assemblage Gibson (contenant le plasmide recombinant) avec 90µL de cellules compétentes. Ce mélange se fait dans la glace. On laisse le mélange pendant 30 min dans la glace. On effectue ensuite un choc thermique à 42°C dans un bain marie pendant 1 min. On transfère ensuite les aliquots de nouveau dans la glace pendant 5 min. Pour chaque aliquot de transformation, on ajoute 900µL de milieu LB (Lysogeny Broth) froid pour un volume total ~1mL. Les mélanges sont ensuite placés dans une étuve à 37°C pendant 40 minutes sans agitation. Cette dernière étape permet aux cellules transformées d'exprimer le phénotype (notamment la résistance aux antibiotiques spectinomycine et ampicilline) correspondant à l'ADN introduit. On procède ensuite à un étalement sur boîte de pétri contenant du milieu LB gélosé supplémenté de l'antibiotique ampicilline à la concentration de 50µg/mL. L'ajout de l'antibiotique permet d'ajouter une pression de sélection pour ne sélectionner que les cellules transformées. Pour les plasmides contenant les deux fluorophores, on ajoute une pression sélective supplémentaire en ajoutant l'antibiotique spectinomycine (50µg/mL). Pour chaque transformation (~1mL de culture), on réalise des étalements à différentes concentrations cellulaires avec un volume constant de ~100µL :

- étalement au 1/100 : 10 µL de culture supplémentée de 100 µL de LB.
- étalement au 1/10 : 100 µL de culture.

- étalement au 9/10 : on centrifuge le volume de culture restant (~900 µL) pendant 1 min à 6000 g, puis on élimine ~800µL de surnageant et on resuspend le culot cellulaire dans les 100µL restant.

Une fois les 100 µL de suspensions déposées, on étale à l'aide de billes préalablement stérilisées. L'étalement par bille permet d'avoir un étalement homogène et d'isoler les cellules afin d'obtenir des clones uniques. Les boîtes de transformations sont ensuite placées à 37°C pendant une nuit. Les solutions mères d'antibiotique sont conservées à -20°C. Les concentrations des solutions mères d'ampicilline et de spectinomycine sont respectivement 50 mg/mL et 100 mg/mL. Le solvant est de l'eau stérile.

II.2.8 Extraction et purification des plasmides :

Après incubation, pour chaque boîte de transformation correspondant à nos différentes constructions, on sélectionne plusieurs clones. Ces clones sont striés sur de nouvelles boîtes de pétri (LB + antibiotique(s)) fraîches et placées à 37°C pendant une nuit. Le lendemain, on effectue un criblage des différents clones par PCR sur colonie pour nous assurer que les inserts {Promoteur-TIR-GFPmut3} et/ou {Promoteur-TIR-mKate2} sont bien présents dans les plasmides assemblés et recircularisés par assemblage Gibson avec un choix d'amorces judicieux (voire partie précédente). On sélectionne les boîtes de repiquage ayant donné des bons résultats par PCR sur colonie. On incube pendant une nuit à 37°C des cultures LB liquide supplémentées d'antibiotique(s) de ~4 mL et ensemencés par les cellules issues des boîtes de pétri retenues. Le lendemain, on récupère les cultures précédentes à l'état stationnaire, et on procède à partir de ces cultures à l'extraction et à la purification de l'ADN plasmidique en utilisant le kit QIAprep Spin Miniprep Kit de chez QIAGEN. Pour cela, les bactéries issues des cultures stationnaires sont récoltées par centrifugation et utilisées pour l'extraction d'ADN plasmidique en suivant les instructions données par le fournisseur. Une fois effectuée l'extraction et la purification, on quantifie la concentration en ADN plasmidique (spectrophotomètre Nanodrop). Les solutions d'ADN plasmidique sont stockées à -20°C. Avant de transformer *B. subtilis* avec ces plasmides, on doit séquencer sur ces plasmides les gènes permettant d'exprimer les fluorophores afin de s'assurer qu'aucune mutation n'ait été introduite durant les cycles de PCR notamment.

II.2.9 Séquençage automatique d'ADN plasmidique et génomique :

Si les PCR sur colonie permettent de vérifier la présence d'un insert, cette méthode ne permet pas d'avoir accès directement à la séquence nucléotidique de l'insert. Pour cela, on doit procéder à un séquençage. Plusieurs séquençages sont effectués. On séquence dans un premier temps une partie de l'ADN génomique

des souches de la banque originale de *B. subtilis* contenant le gène rapporteur GFPmut3, c'est à dire l'ensemble {Promoteur-TIR-GFPmut3}. Dans un deuxième temps, on séquence les ensembles {Promoteur-TIR-GFPmut3} et {Promoteur-TIR-mKate2} sur les nouveaux plasmides construits par assemblage Gibson et qui ont été amplifiés dans *E. coli*. Le séquençage est réalisé par l'entreprise GATC BIOTECH (LightRun Tube Sequencing). La technique utilisée dérive de la méthode Sanger qui met en œuvre une succession d'étapes comportant la séparation des deux brins d'ADN de la région à séquencer, l'hybridation d'une amorce, et une polymérisation par une ADN polymérase en présence de didésoxyribonucléosides triphosphate (ddNTP). L'incorporation aléatoire des quatre désoxyribonucléotides au cours de la synthèse du brin d'ADN entraîne l'arrêt de l'élongation de la chaîne d'ADN. Chacun des didésoxyribonucléotides (ddATP, ddTTP, ddGTP, et ddCTP) est couplé à une fluorescéine ayant une longueur d'excitation qui lui est propre permettant son identification. La préparation des échantillons se fait selon les instructions de l'entreprise GATC BIOTECH. Pour chaque séquençage, on doit préparer une solution de 10 µL dans un aliquot de 1,5 mL contenant :

- 5 µL d'amorce à la concentration de 5 µM (5 pmol/µL). Nos amorces étant à la concentration de 100 µM, on doit dans un premier temps procéder à une dilution par 20. Pour le séquençage de la région {Promoteur-TIR-GFPmut3}, que ce soit pour les ADN génomiques ou plasmidiques, on utilise comme amorce l'oligonucléotide VS19. Ce dernier se situe à ~400 pb de la région {Promoteur-TIR-GFPmut3}, ce qui permet d'optimiser la fiabilité du résultat du séquençage dans cette région d'intérêt. De même, pour séquencer la région {Promoteur-TIR-mKate2}, nous utilisons l'amorce VS86.

- X µL d'ADN plasmidique ou génomique purifié à séquencer. Le volume X est calculé à partir de la concentration (en ng/µL) en ADN mesurée au Nanodrop de telle sorte que la quantité d'ADN présente dans l'échantillon soit de ~450 ng.

- On complète avec de l'eau stérile jusqu'à 10 µL.

II.2.10 Transformation de *B. subtilis* :

- **Préparation des cellules compétentes :**

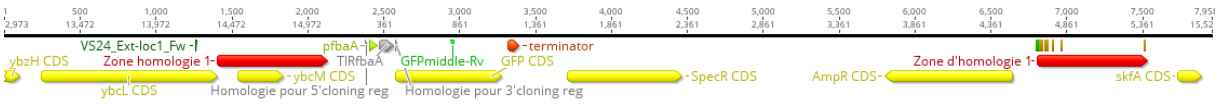
La transformation de *B. subtilis* utilise sa compétence naturelle en début de phase stationnaire. *B. subtilis* est rendue compétente lors d'une croissance en milieu minimum enrichi MG1 suivie d'une dilution dans un milieu carencé MG2. On doit au préalable préparer les milieux de culture suivant: milieu MM 10X

($(NH_4)_2S0_4$: 20 g/L, citrate de Na tri-sodique: 10 g/L, $K_2HPO4,3H_2O$: 140 g/L, KH_2PO4: 60 g/L); milieu MG0 (pour 100mL: 10 mL MM10X, 1 mL glucose 50%, 160 µL $MgSO_4$ 1M, 90 mL H_2O); milieu MG1 (pour 50 mL: 50 mL MG0, 250 µL Hydrolysat caséine 5%, 1 mL yeast extract 5%); milieu MG2 (pour 50 mL: 50 mL MG0, 125 µL Hydrolysat caséine 5%, 250 µL yeast extract 5%, 125 µL $MgSO_4$ 1M, 250 µL $Ca(NO_3)_2$ 0,5M filtré). Le jour avant le lancement des transformations, on prépare une pré-culture stationnaire de la souche de *B. subtilis* BSB168 (souche *trp*+) en milieu LB. Pour la préparation des cellules compétentes, on suit le protocole suivant :

- On ensemence du milieu MG1 au 1/100ème à partir de la pré-culture LB, soit 200 µL de préculture pour 20 mL milieu MG1.
- On incube à 37°C sous agitation et on attend que la préculture MG1 atteigne une DO à 600 nm de ~0,4.
- On ensemence le milieu MG2 au 1/10 avec la pré-culture MG1, soit 2 mL de préculture MG1 pour 18 mL de milieu MG2.
- On laisse incuber 1h30 à 37°C sous agitation. Les bactéries deviennent compétentes
- On centrifuge les cellules à température ambiante (2500 g pendant 5 min).
- On re-suspend les cellules dans 1/10 de volume initial en MG2. Pour cela, on retire 18 mL et on resuspend dans 2 mL.

- **<u>Transformation de *B. subtilis*</u>** :

Dans un aliquot de 1,5 mL, on place 100 µL de culture de cellules compétentes en contact avec 1-5 µL de plasmides transformants. On incube pendant 1 h au bain marie à 37°C ou sous agitation à 37°C pour favoriser l'introduction des plasmides dans *B. subtilis*. Au cours de la transformation, l'ADN plasmidique s'associe aux cellules compétentes et passe à travers la membrane plasmique. L'ADN transformant est circulaire et contient une région d'homologie avec le chromosome de *B. subtilis* qui spécifie le locus d'intégration. Lors de la recombinaison homologue par simple cross-over, l'ADN transformant contenant le(s) gène(s) rapporteur(s) ainsi qu'un gène de résistance à la spectinomycine est intégré dans le chromosome de *B. subtilis* (Figure 21). On procède ensuite à un étalement des 100 µL du mélange de transformation par billes sur boîte de pétri contenant du milieu LB gélosé supplémenté de l'antibiotique spectinomycine à la



<u>Figure 21</u> : Insert de l'ADN porté initialement par le plasmide PL1S01 dans le chromosome de *B. subtilis*. Cette insertion s'étant effectuée par simple cross-over, l'insert est encadré de part et d'autre par la séquence d'homologie (en rouge). Entre ces deux séquences homologues, on retrouve les gènes de résistance aux antibiotiques ainsi que le gène codant pour la protéine GFPmut3, soit l'ensemble {Promoteur:fbaA-TIR:fbaA-GFPmut3}.

concentration de 50µg/mL. L'ajout de l'antibiotique permet d'ajouter une pression de sélection pour ne conserver que les cellules ayant intégré l'ADN transformant dans leur chromosome. Les cultures sont placées à l'étuve à 37°C pendant une nuit. Le lendemain, on effectue un criblage sur plusieurs clones par transformation et on vérifie par PCR sur colonie la présence des gènes rapporteurs.

III Quantification des signaux de fluorescence par cytométrie en flux :

III.1 Préparation des cultures :

Pour obtenir une culture de *B. subtilis* en état de croissance "équilibrée" (balanced growth), permettant ainsi d'avoir un état physiologique bien contrôlé et des résultats reproductibles, nous avons établi un protocole spécifique de préparation des cultures. Toutes les étapes d'incubation sont effectuées à 37°C sous agitation. Les cultures sont effectuées à partir d'un stock de bactéries conservées dans du glycérol à -80°C, inoculé dans du LB supplémenté avec de la spectinomycine et incubé une nuit à 37°C sous agitation. Cette culture est diluée par 100 dans du LB, incubée pendant 2 heures, diluée 50 fois dans du milieu S (milieu minimum Spizizen (Spizizen, 1958)) (pour 100 mL : 0.2 g $(NH_4)_2SO_4$, 1.4 g K_2HPO_4, 0.6 g KH_2PO_4, 0.1 g Sodium citrate, 80 µL $MgSO_4$ à 1M, 6.6 µL $MnSO_4$ à 1M, 1 mL Glucose 50%, 100 µL de solution G ($FeCl_3$ et acide citrique) et 99 mL H20) et incubée 2 heures. La culture est alors diluée par 8 dans du milieu S, incubée 3 heures, puis diluée une dernière fois par 70000 dans du milieu S et incubée une nuit jusqu'à atteindre une densité optique à 600 nm (DO) de 0.2 le lendemain. Le lendemain, la culture incubée toute la nuit est alors diluée par 40 dans un tampon spécifique appelé FACS Flow filtré et peut alors être analysée par cytométrie en flux.

III.2 Cytométrie en flux :

La cytométrie en flux est une technique haut-débit permettant de mesurer simultanément plusieurs caractéristiques physiques de cellules uniques, typiquement le volume et la fluorescence, sur un grand nombre de cellules. Les cellules sont en suspension dans un milieu tampon spécifique (FACS Flow) qui s'écoule le long d'un canal jusqu'à atteindre un faisceau laser. La cellule est alors illuminée et la lumière diffractée dans l'axe du faisceau laser est utilisée comme une mesure indicatrice du volume de la cellule (FSC). Quand la cellule contient un fluorophore, le signal émis par ce dernier est également collecté,

perpendiculairement au faisceau laser. Ce signal permet d'estimer la quantité de fluorophores présents dans la cellule. Nous utilisons un cytomètre BD FACSCalibur, équipé d'un laser émettant à 488 nm, qui permet d'exciter la protéine GFPmut3 exprimée par les souches de B. subtilis que nous utilisons. Le cytomètre est contrôlé par le logiciel Cell *Quest*. Comme nous disposons que d'un seul laser, le niveau d'expression du fluorophore mKate2 ne pourra pas être caractérisé par cytométrie en flux. Pour toutes les souches de la banque, nous pouvons donc estimer la taille des cellules et la quantité de protéines GFPmut3 par cellule, sur un échantillon de 10^4 à 10^5 cellules. Le débit de l'écoulement et la densité de cellules dans l'échantillon sont choisis de telle sorte que les signaux mesurés correspondent bien à une cellule unique et non à des amas de cellules. Toutes les souches de la banque ont été caractérisées avec les mêmes paramètres d'acquisition, à savoir la même puissance de laser et le même voltage aux bornes des détecteurs pour la détection des signaux FSC et de fluorescence. On utilise une photodiode pour le signal FSC, et un photomultiplicateur pour la fluorescence. Ces paramètres ont été choisis de telle sorte que les souches dont les niveaux de fluorescence sont les plus forts ne saturent pas les détecteurs. Le signal de fluorescence provenant d'une cellule issue d'une des souches de la banque est en réalité la somme (i) des signaux émanant des fluorophores, (ii) des signaux provenant de l'autofluorescence de la cellule, et (iii), des signaux générés par l'électronique d'acquisition et qui représente le bruit de fond du dispositif :

$$I_{\,mesurée} = I_{GFP} + I_{\,Auto-fluorescence} + I_{Bruit\ de\ fond}$$

La souche BSB168 de *B. subtilis* non fluorescente est caractérisée par cytométrie en flux pour mesurer le niveau moyen de l'autofluorescence et comment celui-ci se distribue autour de cette valeur. Pour chaque expérience de cytométrie en flux, on estime le bruit de fond généré par l'électronique d'acquisition sur les signaux FSC et de fluorescence en passant dans le cytomètre un échantillon de Facs Flow sans cellules. Cela permet également de définir un seuil sur le signal FSC permettant de discriminer les événements de détection cellulaire et les événements associés au bruit de fond. Ainsi, ne seront pris en compte que les événements appartenant à une fenêtre d'acquisition correspondant uniquement aux cellules. Les signaux de fluorescence originaires du bruit de fond se situant majoritairement en deçà du seuil de détection du cytomètre (~65%), le niveau de bruit sera d'une part faible, et d'autre part difficile à estimer. Nous les négligerons donc par la suite, et les données acquises par cytométrie en flux seront uniquement corrigées de l'autofluorescence des cellules. On estime également la dérive et les fluctuations de la puissance du laser et de l'électronique d'acquisition en mesurant pour chaque jour d'expérience les signaux FSC et de fluorescence de billes fluorescentes de 5,0 µm excité à 488 nm et émettant dans la gamme 515–660 nm (flow cytometry alignment beads, Thermo Fisher). Les données du

cytomètre en flux indique pour chaque cellule une information sur la taille (signal FSC) et une autre sur le nombre de protéines total (signal de fluorescence). Ces données sont issues des signaux recueillis par les détecteurs qui convertissent les photons détectés en un signal électrique puis leur assignent un numéro de canal (valeur numérique codé sur 8 bits). Ces numéros de canaux vont de 0 à 1023. C'est dans cette échelle que sont exprimées les données "brutes" en sortie du cytomètre. Cependant, ces valeurs numériques ne représentent pas une échelle linéaire, dans le sens où un signal deux fois supérieur à un autre signal ne sera pas dans cette échelle égale au double de ce dernier. Pour rétablir la linéarité, les données recueillies du cytomètre doivent être transformées selon la formule suivante :

$$valeur\ linéaire\ =\ 10^{(numéro\ de\ canal/facteur\ d'échelle)}$$

Le facteur d'échelle est choisi de telle sorte que l'échelle des valeurs linéaires s'étende de 1 à 10000, soit 4 décades parcourues. Avec la résolution de notre cytomètre (10 bits), le facteur d'échelle est égal à 256. Avant interprétation des signaux collectés par cytométrie en flux, nous devons transformer l'ensemble des données par l'équation précédente.

IV Quantification des signaux de fluorescence par microscopie de fluorescence :

IV.1 Préparation des échantillons pour la microscopie :

IV.1.1 Préparation des cultures de *B. subtilis* :

La préparation des cultures en phase exponentielle est exactement la même que celle suivie pour la cytométrie en flux. Seulement, au lendemain de la culture en milieu S en phase exponentielle, on dilue cette culture à un rapport 1:300 dans un milieu S préchauffé à 37°C. On laisse incuber pendant une heure avant de charger les cellules sur une fine tranche de gel d'agarose (appelé pad d'agar dans ce qui suit) placé sur une lame de verre. Une dilution au 1:300 suivie d'une culture de 1h conduit à une densité en cellules permettant un espacement correct entre cellule unique sur le pad d'agar pour la microscopie. On pourrait directement placer les cellules sur le pad d'agar juste après la dilution. Cependant, nous avons remarqué que l'enchaînement de la dilution et de la mise en place sur le gel d'agarose conduisait à un fort taux de mort cellulaire. Une étape d'adaptation de 1h dans un milieu S frais permettrait aux cellules de "survivre" au changement d'environnement entre culture liquide et substrat solide.

IV.1.2 Préparation des lames de microscope contenant le gel d'agarose :

Cette étape est faite avant la dilution au 1:300 expliquée précédemment et donc avant que la culture en milieu S n'atteigne une DO de 0,2. On nettoie dans un premier temps deux lames de microscopes en verres (Knittel Glass 76x26 mm) avec une solution d'éthanol à 70% dans l'eau afin d'éliminer les poussières et prévenir d'éventuelles contaminations. On utilise ensuite un cadre appelé *gene frame* (Thermo scientific; 125 µL, 1.7x2.8 cm) et présentant deux faces collantes. Le cadre permet d'accueillir le gel d'agarose. On colle donc une des faces du *gene frame* sur une des lames en laissant l'autre face protégée. On prépare ensuite le milieu de culture (milieu S) supplémenté avec l'agarose. Pour cela, on dissout dans un premier temps 245 mg (2,45%) d'agarose (high-resolution low-gelling temperature; Sigma) dans 100 mL d'eau Milli-Q dans un autoclave à 121°C pendant 15 minutes. L'utilisation de l'autoclave permet de dissoudre et également de stériliser le mélange. Après refroidissement à 50°C dans un bain marie, environ 12 mL de la solution précédente sont supplémentés avec les différents composants du milieu S, qui peuvent être thermosensible, sous un environnement stérile (étape effectuée sous PSM) de sorte que la concentration en agarose finale soit de 1,5% et que les concentrations des composants du milieu S soient respectées. La quantité restante du mélange eau-agarose est laissée à température ambiante et se solidifie. Il sera rendu liquide pour les expériences suivantes en utilisant un micro-onde. On place ensuite 1 mL du mélange S-agarose liquide à 50°C dans le cadre du *gene frame*. On place par-dessus la seconde lame de verre pour homogénéiser le niveau et éviter que le gel ne sèche en refroidissant. Pour accélérer le refroidissement et la solidification du gel, on place la préparation à 4°C pendant une heure. Après solidification, on découpe dans le gel 3 bandelettes d'environ 5 mm d'épaisseur et espacées d'environ 4 mm (de Jong et al, 2011). L'espace entre les bandelettes permet de fournir l'oxygène essentiel à la croissance de *B. subtilis* qui est strictement aérobie et les trois bandes d'agarose permettent d'analyser trois cultures en parallèle. On dépose ensuite sur chaque lamelle 2,5 µL des cultures issues de la dilution 1:300. On laisse sécher l'agarose jusqu'à évaporation/absorption des gouttes de culture déposées. On colle ensuite une lamelle de verre (Knittel Glass Cover Slips 24x60mm) préalablement nettoyée à l'éthanol 70% sur la face libre du *gene frame.* L'échantillon est ensuite incubé à 37°C pendant une heure. Cette dernière étape est importante pour permettre au gel d'agarose de s'équilibrer à 37°C et ainsi éviter les pertes de focus dues à la dilatation des *pads* d'agar lorsqu'on cherchera à contrôler la croissance des bactéries et acquérir le signal de fluorescence.

IV.2 Le microscope :

Le système de vidéo-microscopie se compose de :

- Un microscope *DeltaVision* (GE Healthcare) entièrement motorisé.
- Une caméra *PCO Edge sCMOS Camera*.
- Une chambre thermostatée à 37°C.

Le microscope et la caméra sont contrôlés par le logiciel d'acquisition *Softworx* (GE Healthcare). Le microscope utilisé permet de travailler en contraste de phase et en fluorescence à champs large (Widefield microscopy). L'éclairage pour le contraste de phase est assuré par une diode électroluminescente (LED) blanche. La fluorescence est assurée par un module composé de plusieurs diodes électroluminescentes et comportant plusieurs filtres d'excitation (Solid State Illumination) permettant d'exciter l'échantillon à la longueur d'onde souhaitée. Un miroir dichroïque est placé sur le trajet optique pour permettre de séparer la longueur d'onde d'excitation issue de la source et qui doit se diriger vers l'échantillon, et la longueur d'onde d'émission issue de l'échantillon et qui doit se diriger vers la caméra. Sur ce dernier trajet se trouve un filtre d'émission permettant de sélectionner une gamme étroite de longueur d'onde autour de la longueur d'onde d'émission du fluorophore. Pour le fluorophore GFPmut3(excitation:488nm/émission:510nm), nous utilisons un filtre d'excitation autour de 475 nm et un filtre d'émission autour de 525 nm et pour le fluorophore mKate2(excitation:588nm/émission:633nm) un filtre d'excitation autour de 575 nm et un filtre d'émission autour de 632 nm. L'objectif utilisé est un objectif à immersion à huile x100 d'ouverture numérique 1,40. Nous utilisons le même objectif pour les images en contraste de phase et en fluorescence. L'huile utilisée a pour indice 1,518 à 37°C. Le microscope est couplé à une platine motorisée permettant de se déplacer dans l'échantillon dans le plan (x,y). Durant les acquisitions, le focus est maintenu à l'aide du système "Ultimatefocus" utilisant un laser infrarouge et piloté par le logiciel *Softworx*. Les intensités de la lumière issues des différentes LED, que ce soient pour la technique de contraste de phase ou pour la fluorescence, peuvent être modulées en jouant sur la tension électrique aux bornes de ces modules. L'intensité émise par chaque source est exprimée en termes de pourcentage. Un pourcentage de 100% indique que l'intensité de la lumière qui atteint l'échantillon est maximale. Toutes les images acquises avec la caméra *PCO Edge sCMOS Camera* ont une taille de 1024x1024 et les niveaux de gris des images obtenues obtenus sont codés sur 15 bits. Les niveaux d'intensité se répartissent alors sur une gamme dynamique de 0 à 32767.

IV.3 Suivi de la croissance par imagerie en contraste de phase :

Dans un microscope classique en champs clair la lumière diffusée par l'échantillon, composé d'objet transparent comme les cellules, est déphasée par rapport aux rayons qui n'ont pas traversé l'échantillon. Cependant l'intensité lumineuse est peu affectée si bien qu'il est difficile d'identifier les cellules. La

technique de microscopie en contraste de phase (Zernike, 1953) permet de convertir les variations de phase introduites par l'échantillon en variation d'intensité, ce qui permet d'obtenir des images très contrastés des cellules. Pour cela, un anneau de phase est introduit au niveau de la source de lumière, et une pastille transparente dite pastille de phase est placée dans le plan focal image de l'objectif. La technique de contraste de phase permet donc d'identifier facilement les cellules sur les *pads* d'agar et de pouvoir les détecter sur les images acquises lors de la segmentation. On sélectionne ~30 cellules uniques sur chaque bandelette, soit pour chaque souche. Ainsi, pour chaque lame de microscopie, on enregistre ~90 positions dont les coordonnées sont enregistrées dans le programme *Softworx*. A partir de ce même logiciel, on lance une acquisition d'environ 4 heures dans laquelle chaque position est imagée toutes les 10 minutes par contraste de phase. Une durée de 4 heures assure que les cellules sont en phase exponentielle de croissance, que les colonies formées sont en monocouches, et que le nombre total de cellules dont on collectera le signal de fluorescence est suffisant. Le système de focalisation par laser infrarouge "Ultimate focus" permet d'actualiser la position de l'objectif pour compenser la dérive thermique et conserver les bactéries dans le plan focal. On est alors capable de suivre l'évolution temporelle de chaque micro-colonie formée à partir d'un seul individu. Pour ces acquisitions, on utilise une intensité d'illumination de 32% et un temps d'exposition de 50 ms. Les séries d'images pour chaque position sont analysées dans l'environnement Matlab.

IV.4 Acquisition du signal de fluorescence :

Après 4h de croissance à 37°C, on capture une image en contraste de phase (illumination:32%, exposition:50ms) et en fluorescence, avec des paramètres d'illumination et d'acquisition précisés ultérieurement, de chaque micro-colonie dont on a suivi la croissance.

- **Configuration du microscope :**

La méthode d'imagerie de fluorescence utilisée est celle dite à champ large. Il existe deux types d'illuminations : l'illumination de Köhler et l'illumination critique. L'illumination de Köhler consiste à focaliser la lumière d'excitation dans le plan focal arrière de l'objectif de façon à ce que l'échantillon soit illuminé de façon homogène par un faisceau parallèle de grande étendue spatiale. L'illumination critique consiste quant à elle à focaliser la lumière d'excitation au niveau de l'échantillon. Dans notre système la lumière est émise par une diode et guidée par une fibre optique et l'illumination critique offre une meilleure homogénéité d'éclairement. C'est donc ce mode d'illumination qui a été adopté

pour l'acquisition du signal de fluorescence. Avec notre stratégie d'imagerie à champs large, tout l'échantillon contenu dans la zone d'observation est excité de façon homogène et l'ensemble des fluorophores présents dans les cellules émettent simultanément de la lumière qui est recueillie sur la caméra.

- **Intensité du faisceau d'excitation et temps d'exposition de la caméra :**

Les différentes souches de la banque de *B. subtilis* ne possédant pas les mêmes niveaux d'expression de la protéine GFPmut3, nous devons adapter l'intensité du faisceau d'excitation et/ou le temps d'exposition de la caméra pour chacune d'entre elles. Pour les souches avec une expression faible, ces paramètres ont été sélectionnés de telle sorte que la fluorescence puisse être détectable au-dessus du bruit de fond. En effet, comme tout l'échantillon est illuminé, le détecteur reçoit non seulement de la lumière de la partie de l'échantillon dans le plan focal, mais aussi des parties situées hors focus et qui correspondent essentiellement à l'autofluorescence de l'agarose. Ces dernières vont alors contribuer à ajouter un signal à celui émanant des fluorophores, et nuire à la sensibilité de la mesure. Pour les autres souches, on choisit les paramètres afin d'éviter la saturation de la caméra et également de réduire les effets du *photobleaching* qui sera présenté dans une prochaine partie. Plusieurs stratégies ont ainsi été utilisées. Pour la banque originale, la stratégie employée est de faire varier l'intensité de la lumière d'excitation tout en gardant le temps d'exposition de la caméra constant. Pour les souches venues compléter la banque, la stratégie inverse a été utilisée, à savoir garder l'intensité du faisceau d'excitation constante tandis qu'on modifie le temps d'exposition de la caméra. Pour la première stratégie, les souches de la banque d'origine sont divisées en trois catégories selon leur niveau d'expression, toutes les souches appartenant à une de ces catégories étant illuminées avec la même intensité :

- Souches avec un niveau d'expression faible et souche BSB168 pour l'autofluorescence : illumination :100% ; temps d'exposition :1s.
- Souches avec un niveau d'expression intermédiaire : illumination :32% ; temps d'exposition :1s.
- Souches avec un niveau d'expression élevé : illumination :2% ; temps d'exposition : 1s.

Pour la deuxième stratégie, l'intensité d'excitation est de 50%, et les différents temps d'exposition sont 1s ; 0,5s et 0,2s. Pour quantifier le signal de la protéine mKate2, l'intensité d'excitation est de 100%, et les différents temps d'exposition sont : 0,5s et 1s.

IV.5 Traitement du signal de fluorescence des images de microscopie :

Pour extraire le niveau de fluorescence de chaque cellule imagée, nous utilisons le programme “Schnitzcells” développé par l’équipe de M.Elowitz (California Institute of Technology). Les images de fluorescence doivent au préalable être corrigées du bruit de fond et des éventuels défauts d’homogénéité de l’éclairage d’excitation.

IV.5.1 Correction des images de fluorescence :

On cherche ici à convertir les images de fluorescence en une information sur la concentration et/ou le nombre total de protéines présent dans la cellule. L’image de fluorescence “brute” doit être traitée pour éliminer deux sortes d’artefacts introduits par le montage expérimental : (i) les signaux correspondant à l’autofluorescence de l’agarose qui est éclairé et qui se situe hors-focus, et à l’électronique d’acquisition. (ii) inhomogénéité de l’intensité d’excitation sur l’échantillon.

- **Inhomogénéité de l’intensité d’excitation :**

On corrige ce défaut en réalisant un “fond blanc”. Pour cela, au lieu de déposer une culture de bactérie sur une lamelle d’agarose, on y dépose 2,5 µL d’une solution de fluorescéine. On image plusieurs (~40) champs de fluorescéine dont on fait une image moyenne $b(x,y)$ et qui correspond au fond blanc. Ainsi, si on note $I_0(x,y)$ une image “brute”, alors l’image corrigée du défaut d’illumination $I_1(x,y)$ s’écrit :

$$I_1(x,y) = I_0(x,y)\,\frac{\langle b(x,y)\rangle}{b(x,y)}$$

Où $\langle b(x,y)\rangle$ est l’intensité moyenne de l’image calculée sur tous les pixels. La quantité $\langle b(x,y)\rangle / b(x,y)$ ne dépend pas des paramètres de l’acquisition (intensité d’excitation et temps d’exposition), et le défaut d’illumination est le même pour chaque image acquise au cours d’une même expérience, et entre différentes expériences menées à des jours différents sans qu’il n’y ait eu d’intervention sur le chemin optique de la lumière d’excitation.

- **Correction du bruit de fond :**

Une fois obtenue les images corrigées du défaut d’illumination, il reste à estimer et supprimer le bruit de fond. Pour cela, grâce à un programme similaire à celui utilisé pour détecter les colonies sur les images en contraste de phase, on identifie pour chaque image de fluorescence la colonie entière, et on définit le bruit de fond (autofluorescence de l’agarose et bruits électroniques) comme l’intensité de chaque pixel ne faisant pas partie de la micro-colonie détectée. On calcule alors à partir de ces pixels l’intensité moyenne du bruit de fond $B(x,y)$

pour les 30 images acquises par souche. L'image corrigée du bruit de fond $I_2(x,y)$ s'écrit alors :

$$I_2(x,y) = I_1(x,y) - \langle B(x,y) \rangle$$

C'est à partir de ces images corrigées que nous allons quantifier les niveaux d'expression de la GFPmut3 et de la protéine mKate2.

IV.5.2 Segmentation des cellules et quantification des signaux de fluorescence :

Pour quantifier le signal de fluorescence des cellules individuelles il faut tout d'abord segmenter les cellules dans chaque micro-colonie. Ceci est rendu possible grâce au programme "Schnitzcells" (Elowitz et al). A chaque image de fluorescence correspond une image prise au même moment en contraste de phase. C'est cette dernière qui sert à l'étape de segmentation, ce qui permet de rendre indépendantes la quantification du signal de fluorescence d'une cellule et la détection de ses contours. La segmentation est effectuée de manière automatique par "Schnitzcells" mais il est nécessaire de faire une vérification manuelle et de corriger les résultats. En particulier certaines cellules sont mal séparées et les cellules qui se chevauchent dans la colonie doivent être supprimées. Le masque finalement obtenu est superposé à l'image de fluorescence correspondante et permet de mesurer pour chaque cellule sa surface (en pixels), sa fluorescence totale, c'est à dire intégrée sur tous les pixels constituant la cellule, et enfin sa fluorescence moyenne (moyennée sur l'ensemble des pixels constituant la cellule). On utilise dans ce travail la grandeur "fluorescence moyenne" afin d'éliminer la variation du nombre de protéines due au cycle cellulaire. En analysant l'ensemble des 30 colonies imagées pour une souche, on est capable d'obtenir les informations précédentes pour ~1000-2000 individus et calculer les différentes grandeurs statistiques caractérisant le bruit d'expression (Figure 22).

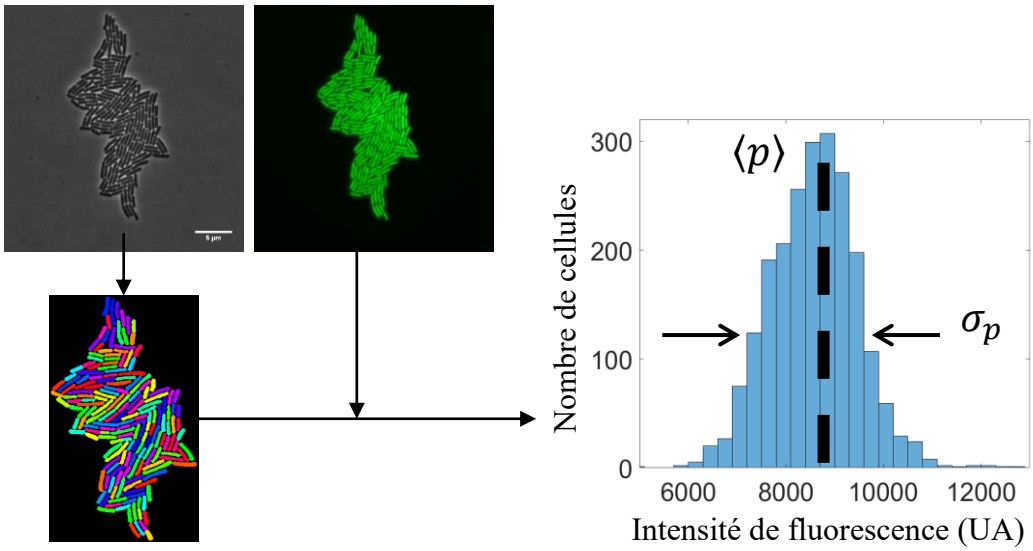


Figure 22 : Mesure du bruit d'expression génique par microscopie de fluorescence. A partir de l'image de contraste de phase, on procède à une segmentation qui permet d'identifier les cellules de la micro-colonie. Ce masque binaire est ensuite superposé à l'image de fluorescence, et il est alors possible de quantifier le signal de fluorescence moyen de chaque cellule de la micro-colonie. En répétant cette procédure pour plusieurs micro-colonies, on peut estimer le niveau moyen d'expression, le bruit d'expression génique et le facteur Fano.

Cependant, les différentes souches ne sont pas illuminées par les mêmes paramètres d'illumination, ce qui nécessitent de contrôler au préalable la manière de comparer les souches entre elles en termes de fluorescence moyenne et de bruit d'expression. Par exemple, les pertes de fluorescence des fluorophores entre différentes souches ne sont pas identiques suivant l'intensité du faisceau d'excitation et le temps d'exposition de la caméra utilisée. Pour chaque paramètre d'illumination utilisés, nous devons donc quantifier la photo-stabilité des fluorophores utilisés. De plus, il nous faut également contrôler la manière de normaliser les signaux. Par exemple, pour comparer les niveaux moyens et de bruit d'expression entre une souche illuminée à 100% et à 32%, on doit déterminer la loi d'évolution de l'intensité d'émission en fonction de l'intensité d'excitation et de même évaluer l'évolution de l'intensité d'émission en fonction du temps d'exposition. Les images acquises lors de ces contrôles sont corrigées de la même manière que les images nous permettant d'évaluer le niveau de bruit d'expression génique.

IV.6 Expériences de quantification du *Photobleaching* :

Les fluorophores soumis à l'éclairage de la lampe fluorescente sont excités et se relaxent vers leur niveau d'énergie fondamentale en émettant un rayonnement de fluorescence. Ce cycle pourrait continuer tant que le fluorophore est excité mais au bout d'un certain temps, au lieu de se désexciter vers l'état

fondamental, la protéine transite vers un état “sombre” et n’est plus capable de fluorescer. Ce phénomène est nommé *photobleaching* et correspond à des dommages chimiques et des modifications des liaisons covalentes induits par les photons et peut dépendre grandement de l’environnement. Pour un fluorophore, le nombre de cycles d’excitation et d’émission qui se produisent avant le *photobleaching* est une variable stochastique, et le “temps de vie” d’un fluorophore (d’un point de vue sa capacité à émettre de la fluorescence) est distribué selon une loi exponentielle caractérisé par un taux de “dégradation” k_{PB}. Pour les illuminations 100% et 32%, on caractérise la photo-stabilité de la GFPmut3 en suivant l’évolution de la fluorescence moyenne de chaque cellule d’une micro-colonie d’une centaine d’individus afin de contrôler les pertes de fluorescence. Pour cela, la micro-colonie est illuminée et imagée à fréquence élevée (mode d’acquisition fast acquisition) à la longueur d’onde d’excitation du fluorophore avec un temps d’exposition constant (300 ms). Dans le mode fast acquisition, le temps entre deux acquisitions est donné par le temps d’exposition augmenté de 33 ms, soit ici 333 ms. Contrairement aux expériences menées pour quantifier les niveaux moyens et de bruits d’expression des fluorophores, on ne prend pas d’image en contraste de phase et on n’effectue donc aucune opération d’autofocus. Cependant, le laser permettant de maintenir le focus est tout de même maintenu pour rester dans les mêmes conditions d’illumination. Le mode fast acquisition permet de déplacer les électrons plus rapidement hors de la puce CCD de la caméra. Cette technique d’imagerie à grande vitesse nous permet d’obtenir une centaine d’image de fluorescence en 1 minute. Ce temps étant inférieur au temps de doublement de nos cellules (~1h), la croissance des bactéries dans cette plage temporelle est négligeable et on segmente uniquement la première image de fluorescence acquise. A partir de la segmentation, on peut déterminer la dépendance temporelle de la fluorescence moyenne de chaque cellule formant la micro-colonie.

IV.7 Normalisation des signaux de fluorescence :

Pour pouvoir comparer les souches entre elles, et donc quantifier les effets de la séquence promotrice et de la séquence TIR sur le bruit d’expression, nous devons déterminer quelle normalisation il convient d’appliquer. Pour cela, nous devons caractériser l’évolution de l’intensité d’émission lorsque varie (i) l’intensité d’excitation et (ii) le temps d’exposition :

(i) On choisit une souche de la banque possédant un niveau d’expression du fluorophore GFPmut3 important et un temps d’exposition de la caméra qui nous permettent d’obtenir un signal de fluorescence quantifiable pour l’ensemble des intensités d’excitation. On choisit pour cela la souche PL1S05 et on prend un temps d’exposition de 150 ms. Pour une intensité du faisceau d’excitation donnée,

on image 5 micro-colonies (en contraste de phase et en fluorescence) et on recueille après segmentation la fluorescence moyennes de chaque cellule de ces micro-colonies. En répétant ceci pour l'ensemble des intensités d'illuminations possible, on est capable d'obtenir la dépendance de l'intensité de fluorescence émise en fonction de l'intensité d'excitation. On image également des billes fluorescentes de 5,0 µm excité également à 488 nm et émettant dans la gamme 515–660 nm (flow cytometry alignment beads, Thermo Fisher) pour contrôler la tendance observée avec les cellules bactériennes.

(ii) A l'aide de la même souche et en fixant cette fois-ci l'intensité de l'illumination (%T=10%), on parcourt plusieurs temps d'exposition en considérant pour chacun de ces temps 5 micro-colonies dont on quantifie le niveau de fluorescence moyen de chaque cellule. L'évolution de la fluorescence moyenne en fonction du temps d'exposition est de nouveau contrôlée avec des billes fluorescentes.

RESULTATS

I Présentation et interprétation des signaux de cytométrie en flux et de microscopie :

I.1 Données issues de la microscopie de fluorescence :

I.1.1 Raisonnement en termes de fluorescence moyenne :

A partir des images de fluorescence, nous avons vu qu'il était possible de recueillir plusieurs informations, une information sur la surface de la cellule (en nombre de pixels), la fluorescence totale (intégrée sur tous les pixels de la cellule) et la fluorescence moyenne (moyennée sur tous les pixels de la cellule). Les différentes cellules imagées ne se trouvent pas au même stade du cycle cellulaire, ce qui se manifeste en particulier par des tailles différentes. Ces différences contribuent à générer de la variabilité sur le nombre de protéines par cellule et représentent une source de bruit extrinsèque. Cette source de bruit se retrouvera dans notre analyse si on mesure l'hétérogénéité des signaux de fluorescence en termes de fluorescence totale. Pour s'affranchir de cette contribution, plusieurs auteurs (Elowitz et al, 2002 ; Golding et al, 2005 ; Taniguchi et al, 2010) suggèrent de normaliser le signal de fluorescence total par la taille des cellules, ce qui revient à raisonner en termes de fluorescence moyenne c'est à dire en termes de concentration de fluorophores. Nous montrons sur la figure 23-A l'évolution de la fluorescence totale, qui est une indication du nombre de fluorophores contenus dans la cellule, en fonction de la surface de la bactérie. On observe une corrélation forte (coefficient de corrélation linéaire de Pearson de 0,9) entre ces deux quantités. Une régression linéaire imposant le passage par zéro rend bien compte de l'évolution de la fluorescence totale avec la surface des cellules. Après normalisation du signal de fluorescence totale par la surface des cellules, la corrélation linéaire entre les deux quantités chute pour atteindre -0.097. Ainsi, la normalisation du signal de fluorescence total par la surface des bactéries permet ainsi de supprimer la dépendance du signal de fluorescence avec la taille.

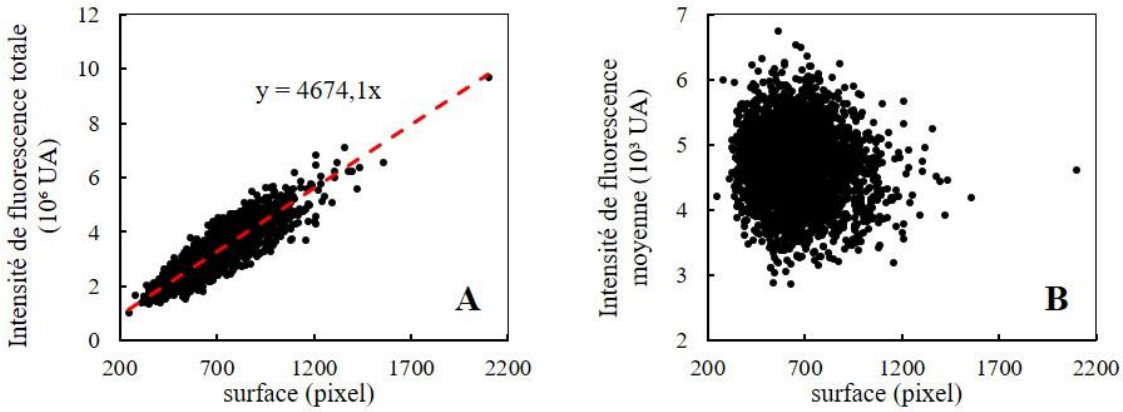


Figure 23 : Signaux de fluorescence en fonction de la taille des cellules. A) Signaux de fluorescence total en fonction de la surface des cellules. On observe une forte corrélation linéaire entre les deux quantités. **B)** Après normalisation de l'intensité de fluorescence total avec la surface des bactéries, on perd la corrélation précédente. Pour une meilleure quantification du bruit d'expression génique, on décide donc de normaliser par la taille des cellules, ce qui *a priori* permet de réduire les effets du cycle cellulaire sur la variabilité phénotypique. La variabilité quantifiée à partir des intensités moyennes de chaque cellule peut ainsi être plus sensible aux paramètres génétiques de nos différentes souches.

I.1.2 Photobleaching :

Dans les expériences visant à caractériser le phénomène du *photobleaching*, nous avons recueillis les données de fluorescence moyenne de plusieurs cellules appartenant à une même micro-colonie à différents instants. Pour un fluorophore donné, le temps d'attente avant la perte de sa capacité à fluorescer est distribué exponentiellement et caractérisé par un taux de *photobleaching* constant k_{PB}. Ainsi, pour une population N_0 de fluorophores émettant un signal I_0à un instant $t = 0$, l'intensité du signal émis par cette même population après un temps t d'illumination peut être décrit par une fonction exponentielle décroissante :

$$I(t) = I_0 \exp(-k_{PB} t)$$

Dans nos expériences, pour chaque instant où une image a été acquise, la fluorescence mesurée pour une cellule est l'intégrale de l'intensité lumineuse $I(t)$ sur un temps Δt correspondant au temps d'acquisition de la caméra. Par exemple, pour la première image acquise, on détecte pour une cellule donnée une fluorescence moyenne F_0 qui s'écrit selon :

$$F_0 = \int_0^{\Delta t} I_0 \exp(-k_{PB} t) dt$$

Ce qui s'intègre pour donner :

$$F_0 = \frac{I_0}{k_{PB}}[1 - \exp(-k_{PB}\Delta t)]$$

Pour la deuxième image, le signal mesuré s'écrit :

$$F_1 = \int_{\Delta t}^{2\Delta t} I_0 \exp(-k_{PB} t)dt$$

Soit :

$$F_1 = \frac{I_0}{k_{PB}}[\exp(-k_{PB}\Delta t) - \exp(-k_{PB}2\Delta t)]$$
$$= \frac{I_0}{k_{PB}}\exp(-k_{PB}\Delta t)\,[1 - \exp(-k_{PB}\Delta t)] = F_0 \exp(-k_{PB}\Delta t)$$

On montre par récurrence que la fluorescence moyenne d'une cellule F_n au bout d'un temps cumulatifs d'illumination de $n\Delta t$ s'écrit :

$$F_n = F_0 \exp(-k_{PB} n\Delta t)$$

En introduisant $t_{cum} = n\Delta t$, la fluorescence moyenne d'une cellule au bout d'un temps d'illumination t_{cum} s'écrit :

$$F(t_{cum}) = F_0 \exp(-k_{PB} t_{cum})$$

Ainsi, pour chaque cellule, la répartition des points montrant la dépendance de la fluorescence $F(t_{cum})$ en fonction du temps t_{cum} est ajustée numériquement par la fonction précédente. On en déduit alors pour chaque cellule le paramètre k_{PB} dont on déduit un k_{PB} moyen pour chaque intensité d'illumination. On est alors capable d'estimer les pertes de fluorescence.

D'après les résultats obtenus, la décroissance des niveaux moyens de fluorescence de chaque cellule est bien exponentielle. En effet, si on trace le logarithme népérien de la fluorescence moyenne en fonction du temps on constate qu'une droite d'équation $y = ax + b$ ajuste parfaitement la répartition des données (Figure 24-A-B). La pente de la droite nous permet donc, d'après ce qui précède, de déduire le taux k_{PB}.

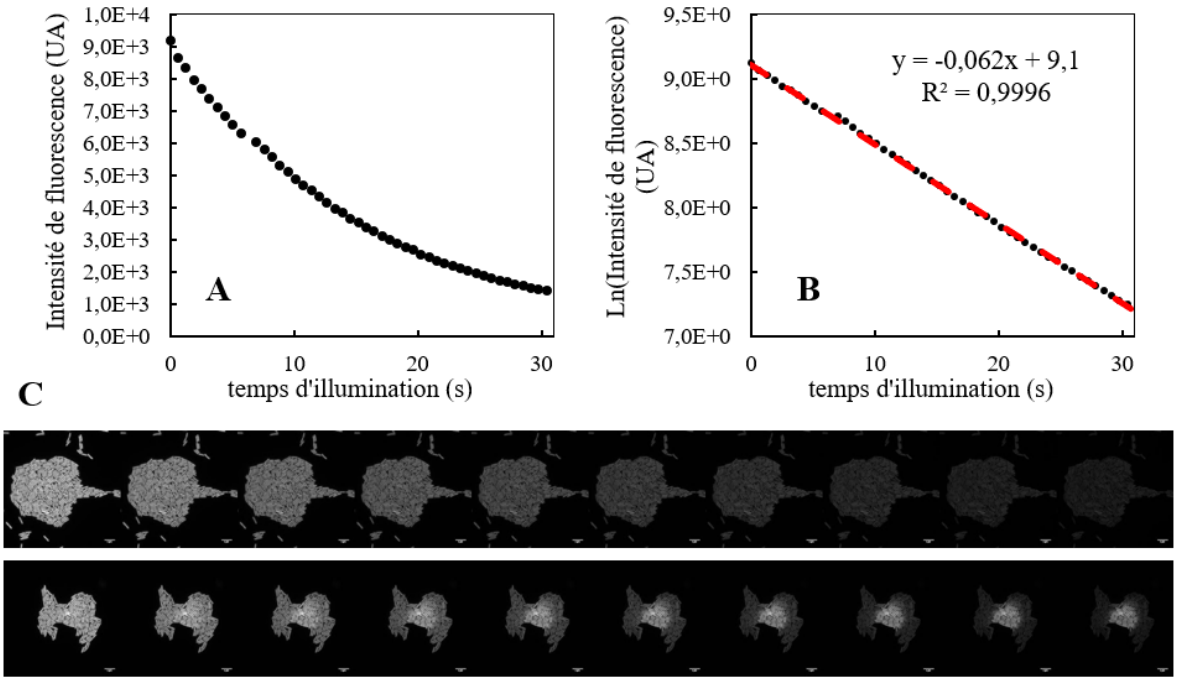


Figure 24: Phénomène de *photobleaching* du fluorophore GFPmut3. A-B) Décroissance exponentielle du signal de fluorescence d'une cellule en fonction du temps d'illumination des sondes fluorescentes contenues dans cette cellule. Un ajustement de cette courbe par une exponentielle décroissante permet de quantifier le coefficient k_{PB}. **C)** Suivi temporel de la fluorescence d'une micro-colonie en fonction du temps. L'intervalle de temps entre deux images consécutives est de 3s. L'image de haut correspond au cas où le laser d'auto-focus n'est pas activé. On observe alors une décroissance homogène du signal de fluorescence sur l'ensemble de la micro-colonie. L'image du bas correspond au cas où le laser d'auto-focus est activé. Dans ce cas, la décroissance du signal de fluorescence n'est plus homogène. les cellules contenues au centre de l'image sont « blanchies » moins rapidement que les cellules excentrées.

Deuxièmement, on constate que ces taux varient considérablement d'une cellule à une autre, ce qui peut être vue qualitativement sur la figure 24-C. Il apparaît clairement que les cellules contenues dans un disque autour du centre de l'image ont un signal très peu dégradé. Ce résultat est d'autant plus surprenant que, bien que *a priori* corrigé, le défaut d'illumination du dispositif fait que les bactéries se situant au centre de l'image devrait être plus éclairé et donc avoir un temps caractéristique de dégradation du signal de fluorescence plus faible que les cellules excentrées. Plus surprenant encore, cet effet disparaît si on effectue la même expérience avec des billes fluorescentes de 5 µm mais réapparaît si on image des cellules d'*E. coli* exprimant le fluorophore mCherry. Ce phénomène semble donc propre aux cellules et ne dépend pas du fluorophore utilisé. Cependant, en éliminant le laser permettant de maintenir le focus, on observe que cet effet disparaît. Dans ce cas, on constate une nette diminution de l'écart type de la constante k_{PB} autour de sa valeur moyenne : de $0{,}0182\ s^{-1}$ avec le laser activé à $0{,}0048\ s^{-1}$ quand le laser est désactivé (ces valeurs sont obtenues avec une illumination de 32%, un temps d'exposition de 300 ms et avec la souche PL1S24). Le taux moyen $\langle k_{PB} \rangle$, quant à lui, augmente lorsque on éteint le laser

(de 0,0499 s^{-1} à 0,0643 s^{-1}). Ce phénomène semble donc résulter d'un processus biologique (les billes ne sont pas affectées) couplé avec un processus photochimique, le rayonnement du laser jouant un rôle majeur dans ce processus.

Enfin, pour estimer l'importance du *photobleaching* sur nos expériences, on analyse les données collectées sur la souche PL1S05 avec l'illumination maximale (100%) qui devrait correspondre *a priori* au cas le plus affecté. Compte tenu de la présence du laser d'auto-focus dans les expériences, on considère les taux k_{PB} maximal et minimal et on évalue pour chacun la perte de fluorescence au bout d'une seconde au lieu d'estimer le taux k_{PB} moyen :

$$k_{PBmax} \sim 0{,}1\ s^{-1} \implies F(1s) = 0{,}9F(t=0)$$
$$k_{PBmin} \sim 0{,}02\ s^{-1} \implies F(1s) = 0{,}98F(t=0)$$

On a donc au maximum pour les souches excentrées une diminution de ~10% contre 2% au centre de l'image. Ces différences ajoutent de la variabilité supplémentaire qui n'est pas d'origine « biologique » et doit être corrigé. Cependant, cette correction peut s'avérer fastidieuse. Pour contourner cette difficulté, nous tâcherons pour chaque expérience de décentrer les colonies dans la mesure du possible avant de réaliser les acquisitions de fluorescence, en particulier pour les souches ayant un faible niveau d'expression et nécessitant une intensité d'excitation maximale. Les pertes de fluorescence dans ce cas seront homogènes sur l'ensemble de la population imagée et de l'ordre de 10%, ce que nous considèrerons par la suite comme négligeable. Pour les autres intensités du faisceau d'excitation utilisées, (50%, 32% et 2%), les pertes de signal par *photobleaching* sont négligeables.

I.1.3 Normalisation des signaux de fluorescence :

On détermine l'évolution de l'intensité d'émission du fluorophore GFPmut3 en fonction de l'intensité d'excitation. Pour cela, on utilise la souche PL1S05 que l'on image avec les différentes intensités d'illumination (exprimée en pourcentage de lumière transmise) que nous permet d'utiliser le dispositif expérimental (%T : 2%, 5%, 10%, 15%, 20%, 25%, 32%, 50% et 100%) avec un temps d'exposition fixe (150 ms). Pour rappel, pour chaque intensité d'excitation, on image 5 micro-colonies. Pour chaque micro-colonies, on mesure la fluorescence moyenne de chaque cellule puis on en déduit une moyenne sur l'ensemble des cellules de la micro-colonie. On moyenne de nouveau la fluorescence moyenne de chaque micro-colonie (5 au total). On en déduit un niveau de fluorescence moyen pour une intensité d'excitation donnée (Figure 25-A). On procède de même avec le temps d'exposition (dans ce cas on image 1 micro-colonie de la souche PL1S05 par temps d'exposition) en fixant l'intensité

d'excitation à 10%. Les différents temps d'exposition visitées vont de 5 ms à 1s. les résultats sont reportés sur la figure 25-B.

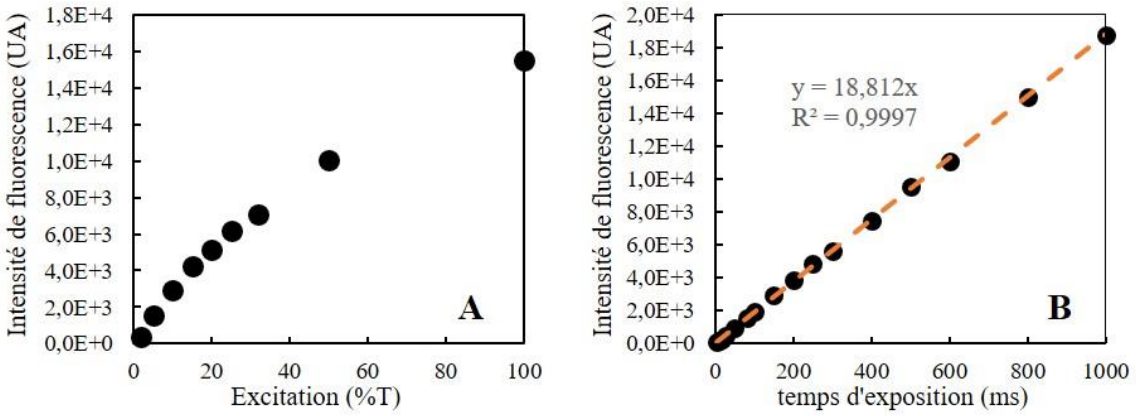


Figure 25 : Evolution de l'intensité de fluorescence moyen en fonction de l'intensité de la lumière d'excitation (A) et du temps d'exposition de la caméra (B).

On constate que lorsque varie l'intensité d'excitation, l'intensité de fluorescence n'évolue pas linéairement avec l'intensité d'excitation, si bien que pour normaliser les signaux de fluorescence entre souches imagées avec des intensités d'excitation différentes, une simple multiplication par le rapport des illuminations ne suffit pas et il faut considérer chaque illumination au cas par cas et normaliser en multipliant par le rapport des signaux de fluorescence mesuré ici. Toutes les souches sont normalisées par rapport à l'intensité d'excitation maximale (100%). Les constantes multiplicatives sont répertoriées dans le tableau 10. Par exemple, pour comparer une souche imagée avec une illumination de 100% et une autre imagée avec une intensité d'excitation de 2%, multiplier le niveau de fluorescence moyen ainsi que la variance du signal de fluorescence de cette dernière souche respectivement par 50 et 2500 entraînera une surestimation de ces quantités. Au lieu de ça, on doit multiplier la moyenne par 39,7 et la variance par le carré de cette valeur. En effectuant la même expérience avec des billes fluorescentes de 5 µm, on obtient le même comportement et les constantes de normalisation à appliquer sont identiques, ce qui montre que ce comportement est dû au système d'imagerie. En revanche, l'intensité de fluorescence est proportionnelle au temps d'exposition de la caméra. Dans ce cas, une simple multiplication par le rapport des temps d'exposition permet de correctement comparer des souches imagées avec des temps d'exposition différents. Tous les résultats présentés dans ce qui suit ont été normalisés par les procédures expliquées ici. Ceci revient à considérer que toutes les souches ont été acquises avec la même illumination (100%) et le même temps d'exposition (1 s).

Illumination initiale (%T)	Illumination finale (%T)	Constante de normalisation
2	100	**39,7**
5	100	10
10	100	5,35
15	100	3,68
20	100	3,01
25	100	2,53
32	100	**2,19**
50	100	**1,54**
100	100	**1**

Tableau 10 : Constante de normalisation pour comparer des souches qui n'ont pas été caractérisées avec les mêmes paramètres d'illumination. Sont indiqués en gras les intensités d'illuminations utilisées dans cette étude ainsi que les constantes de normalisation à utiliser.

I.2 Données issues de la cytométrie en flux :

Dans un plan où est reporté le signal FSC en abscisse et le signal de fluorescence en ordonné (Figure 26-A), on obtient un nuage de points représentatif de la variabilité de la population en termes de taille et de fluorescence, dans lequel chaque point représente une cellule. Le signal de fluorescence collecté reflète ici le nombre total de fluorophores présents dans chaque cellule. Quantifier l'hétérogénéité entre cellules de ces signaux de fluorescence inclut nécessairement une part importante de bruit extrinsèque notamment due dans notre cas au cycle cellulaire. Pour éliminer ces effets du cycle cellulaire, deux approches peuvent être envisagées : (i) on sélectionne les cellules ayant une taille donnée (Newman et al, 2006), ou (ii) comme pour les données de microscopie, on normalise par rapport à la taille des cellules. Nous présentons les deux approches dans ce qui suit :

I.2.1 Sélection des cellules selon leur taille :

Newman et al ont montré qu'en sélectionnant les cellules selon une taille donnée, on était capable de réduire considérablement la contribution du bruit extrinsèque due au cycle cellulaire (Newman et al, 2006). En adoptant cette stratégie, on est donc capable de détecter avec plus de sensibilité la contribution des sources de bruits intrinsèques. Pour sélectionner une taille, on détermine pour l'ensemble des souches le mode de la distribution du signal FSC (c'est à dire la valeur du signal FSC pour laquelle la densité de probabilité est maximale). Ce mode est identique pour toutes les souches, et on collecte uniquement les données correspondant à des cellules dont le signal FSC est compris dans une fenêtre

encadrant cette valeur. L'inconvénient de cette méthode est qu'on réduit considérablement le nombre d'individus, passant de ~10^4-10^5 à ~10^3 cellules, ce qui réduit la précision d'estimation des grandeurs statistiques telles que la moyenne et l'écart type du niveau de fluorescence. Une autre stratégie est de se baser sur celle adoptée en microscopie, à savoir normaliser le signal de fluorescence par la taille des cellules.

I.2.2 Normalisation du signal de fluorescence par le signal FSC :

Contrairement à la microscopie où nous pouvons directement mesurer la dimension des bactéries, en cytométrie en flux, la relation liant le signal FSC à la taille des bactéries n'est pas bien connue et fait toujours débat. Si certains l'ont supposé proportionnel au volume des cellules, d'autres ont déterminé expérimentalement que la diffraction de la lumière par la cellule est en réalité une fonction complexe de la taille de cette dernière (Julià et al, 2000; Troussellier et al, 1999) et dépend également de la forme des objets diffusants, de leur structure, de leur composition chimique et également de l'angle de détection (Vives-Rego et al, 2000). Cependant, en regroupant les cellules de la population en fonction du signal FSC (binning), et en moyennant pour chaque valeur de ces signaux les différents niveaux de fluorescences, on constate que les niveaux de fluorescence $\langle p \rangle(FSC)$ évoluent linéairement avec le signal FSC :

$$\langle p \rangle(FSC) = aFSC + b$$

Les paramètres a et b dépendent de la souche (Figure 26-B). Comme nous avons vu en microscopie que la fluorescence totale des cellules était directement proportionnelle à leur surface/volume (au bruit d'expression près), on peut envisager une dépendance linéaire du signal FSC avec le volume des cellules. Pour supprimer la corrélation entre les dimensions d'une cellule ($FSC = v$) et le signal de fluorescence détecté (F), on normalise la fluorescence F par la fluorescence donnée par la régression linéaire du signal FSC, à savoir $av + b$. En traitant les données de cette manière, toutes les souches de la banque auront un niveau moyen d'expression voisin de 1, et ne rendra donc pas compte de la force des différents promoteurs et TIR. Pour corriger cet effet, on multiplie les données traitées par la fluorescence moyenne de fluorescence pris sur l'ensemble de la population avant normalisation.

Les deux approches, i.e. sélection d'une sous-population de cellules ou bien normalisation par le signal FSC conduisant aux mêmes résultats quantitatifs, nous présenterons ici les résultats obtenus avec la normalisation, qui présente l'avantage de conserver l'intégralité des données. La sensibilité du cytomètre est très inférieure à celle du microscope. Si on est capable d'extraire du bruit de fond les souches pour lesquelles les fluorophores sont les moins exprimés en

microscopie, il est très difficile d'extraire ces mêmes souches du bruit en cytométrie en flux. En effet, pour de nombreuses souches, une partie de leur distribution vers les faibles valeurs de fluorescence se retrouvent en deçà du seuil de détection et se retrouve donc tronquée. Ainsi, leur moyenne et écart type ne pourront pas être estimés. Nous sommes donc contraints d'éliminer ces souches pour l'analyse des données de cytométrie en flux. De façon générale, la microscopie offre des mesures plus précises et nous utiliserons les données de cytométrie essentiellement comme contrôle.

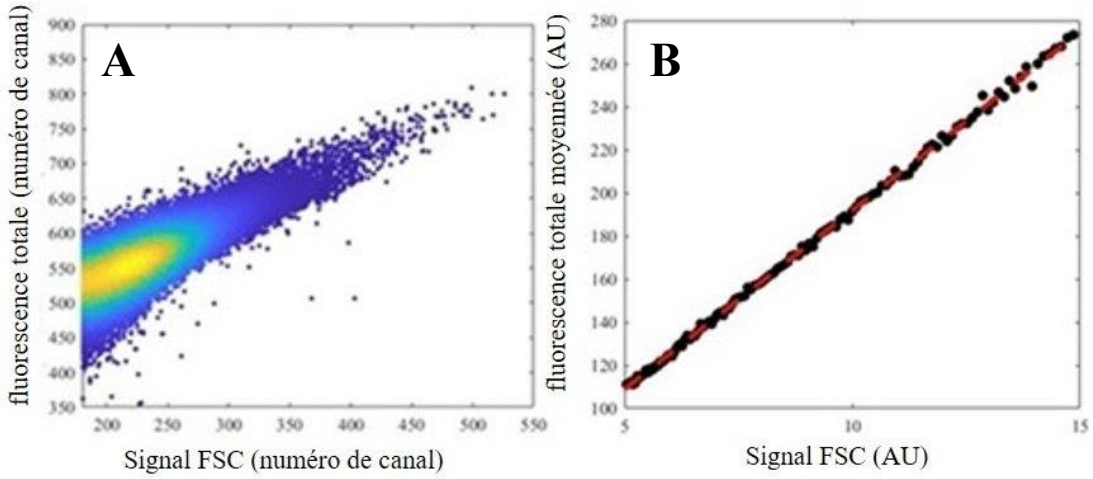


Figure 26 : Données de fluorescence totale et de taille (FSC) des bactéries collectées en cytométrie en flux (souche Alex01). On reporte les données collectées dans un plan où figure en ordonné le signal de fluorescence et en abscisse le signal FSC (**A**). On observe comme en microscopie une forte corrélation linéaire entre la fluorescence totale et la taille des cellules. Si pour chaque taille, on calcule la fluorescence moyenne, on obtient une répartition de points parfaitement ajustable par une régression linéaire (**B**). On a ainsi : $\langle p \rangle(FSC) = aFSC + b$.

I.3 Correction de l'autofluorescence :

Certains composants naturels de la cellule sont excités à la même longueur d'onde que les fluorophores, et la relaxation radiative peut libérer des photons de même énergie que les protéines rapportrices. Ce signal supplémentaire est nommé autofluorescence de la cellule. Les signaux collectés pour la souche non fluorescente BSB168 permettent de quantifier son niveau moyen ainsi que sa variabilité entre cellules. Les signaux émanant de cette souche sont traités de la même manière que les souches de la banque, que ce soit en cytométrie et en microscopie de fluorescence. En cytométrie en flux cependant, une partie des cellules de la souche BSB168 ont un niveau d'autofluorescence plus faible que le seuil de détection. Pour remédier à ceci, on s'appuie sur la distribution d'autofluorescence obtenue en microscopie afin d'inférer la partie tronquée de la distribution et d'estimer la moyenne et la variance de l'autofluorescence. Afin de quantifier uniquement le niveau d'expression moyen des fluorophores ainsi que leur niveau de bruit, on considère que les signaux dus à l'autofluorescence et aux fluorophores sont indépendants, de sorte que :

$$\langle I_{mesurée} \rangle = \langle I_{fluorophore} \rangle + \langle autofluorescence \rangle$$
$$Var(I_{mesurée}) = Var\left(I_{fluorophore}\right) + Var(autofluorescence)$$

On soustrait donc au niveau d'expression moyen mesuré pour une souche de la banque, l'intensité moyenne obtenue pour l'autofluorescence, et on procède de même pour la variance.

I.4 Suivi de la croissance par imagerie en contraste de phase :

Pour pouvoir évaluer le taux de croissance de chaque micro-colonie, on doit détecter la micro-colonie sur chaque image. Le programme Matlab que nous avons écrit utilise la fonction "*edge*" qui permet de détecter les contours de la colonie. Après remplissage, on obtient une image binaire où une valeur de 1 a été assignée aux pixels contenus dans la micro-colonie et 0 en dehors, ce qui nous donne la surface de la micro-colonie (en pixels). En effectuant ceci à chaque instant, on peut reporter l'évolution de la surface de la micro-colonie au cours du temps et en déduire son taux de croissance (Figure 27). Pour chaque souche, on est capable de mesurer le taux de croissance de 30 micro-colonies. Le taux de croissance ayant un effet sur l'efficacité de transcription et de traduction, il est important de vérifier qu'il ne varie pas trop d'une expérience à une autre, afin de ne pas introduire de biais dans la comparaison des souches.

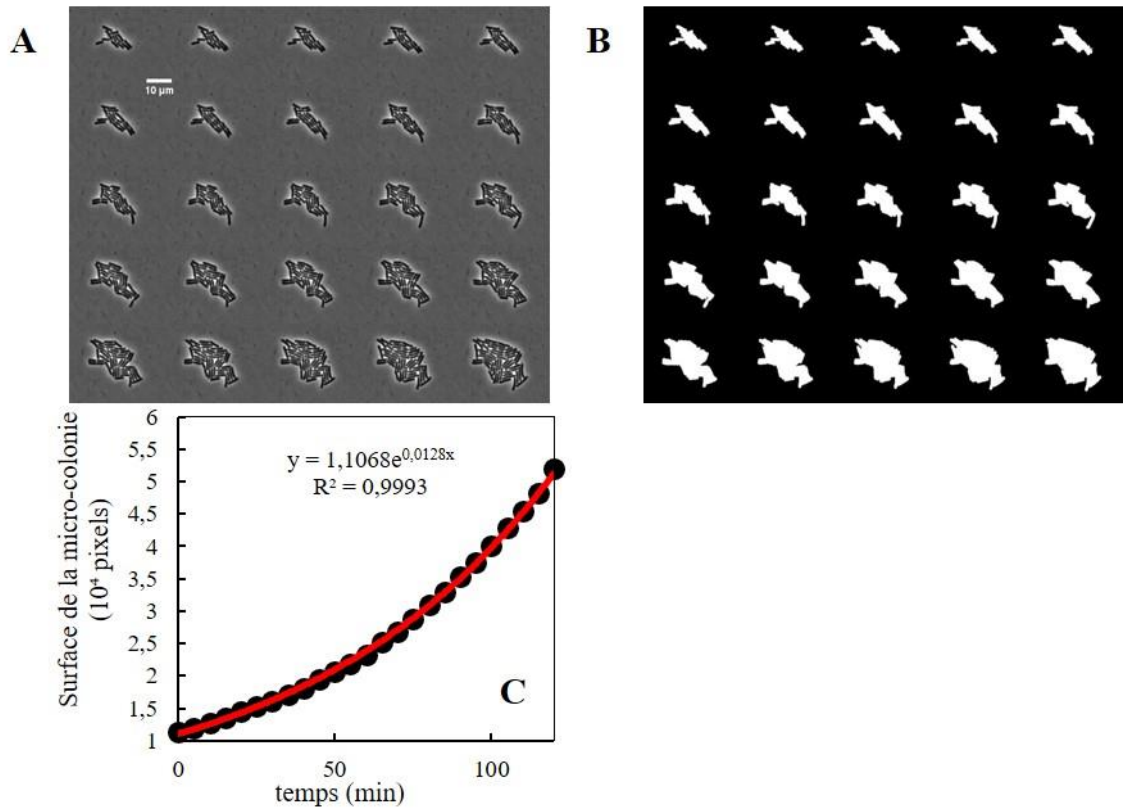


Figure 27 : Croissance d'une micro-colonie sur lamelle d'agarose. A) Suivi de la croissance d'une micro-colonie sur lamelle d'agarose en microscopie de contraste de phase à 37°C. Deux images successives sont espacées de 15 min. L'échelle indique 10µm **B)** Sur chaque image en contraste de phase, on détecte la micro-colonie et on est alors capable de calculer la surface de la micro-colonie à chaque instant. **C**) Surface de la micro-colonie en fonction du temps. La ligne rouge est un ajustement des données par une simple exponentielle, ce qui permet de déterminer le taux de croissance de la micro-colonie et un temps caractéristique de la division cellulaire. Ici, on obtient un temps de doublement de la taille de la population de 54 min.

I.5 Reproductibilité des expériences et correspondance entre les expériences de cytométrie en flux et de microscopie :

Les données de fluorescence obtenues pour les différentes souches sont dans un premier temps analysées à la lumière du taux de croissance moyen obtenu sur les 30 micro-colonies. Puisque les différentes constructions génétiques montrent des dépendances fortes avec la croissance des cellules, nous nous sommes efforcés de quantifier le temps de doublement des micro-colonies avant chaque acquisition des niveaux de fluorescence. Si le taux de croissance est reproductible en culture liquide avec un temps de doublement d'environ une heure, les cultures sur *pad d'agar* (substrat solide) peuvent présenter une variabilité des temps de doublement relativement importante entre différentes séries d'expériences menées sur des souches identiques (de ~40min à ~1h). Cette variabilité pourrait s'expliquer par l'impossibilité de contrôler quantitativement l'évaporation de l'eau contenue dans le *pad d'agar* lors de l'étape d'évaporation/absorption suite au dépôt de la culture cellulaire, potentiellement combinée à la consommation par les bactéries de l'agarose en plus du glucose. Nous avons vérifié que la variabilité des mesures de signaux de fluorescence entre réplicats n'était pas due à des différences de croissance. Ainsi, la majorité des données collectées en microscopie ont été traitées et insérées dans les résultats de notre étude. Nous vérifions la reproductibilité sur le niveau d'expression moyen de chaque souche ainsi que sur l'écart type de la distribution du niveau de fluorescence. Pour cela, nous comparons les résultats obtenus sur deux séries d'expériences menées sur l'ensemble de la collection de souches pour l'intensité de fluorescence moyenne ainsi que pour l'écart type de la distribution du signal de fluorescence (Figure 28-A-B-C-D). On voit que les moyennes et écarts type sont assez reproductibles d'une expérience à une autre. Enfin, en croisant les données obtenues en cytométrie en flux et en microscopie de fluorescence on s'assure que les niveaux d'expression moyen et l'écart type des distributions mesurées sont cohérents (Figure 28-E-F). Cependant, bien que les écarts types et les moyennes soient cohérents entre les deux appareils, les coefficients de variation sont assez différents, les valeurs de bruits sont notamment plus élevées dans le cas des expériences menées en cytométrie en flux. Ceci pourrait s'expliquer par la moins bonne résolution de la cytométrie en flux ou encore par

la différence de l'état des cellules lorsqu'elles sont dans le milieu tampon Facs Flow où les cellules sont stoppées dans leur croissance tandis que dans les expériences de microscopie les cellules peuvent croître.

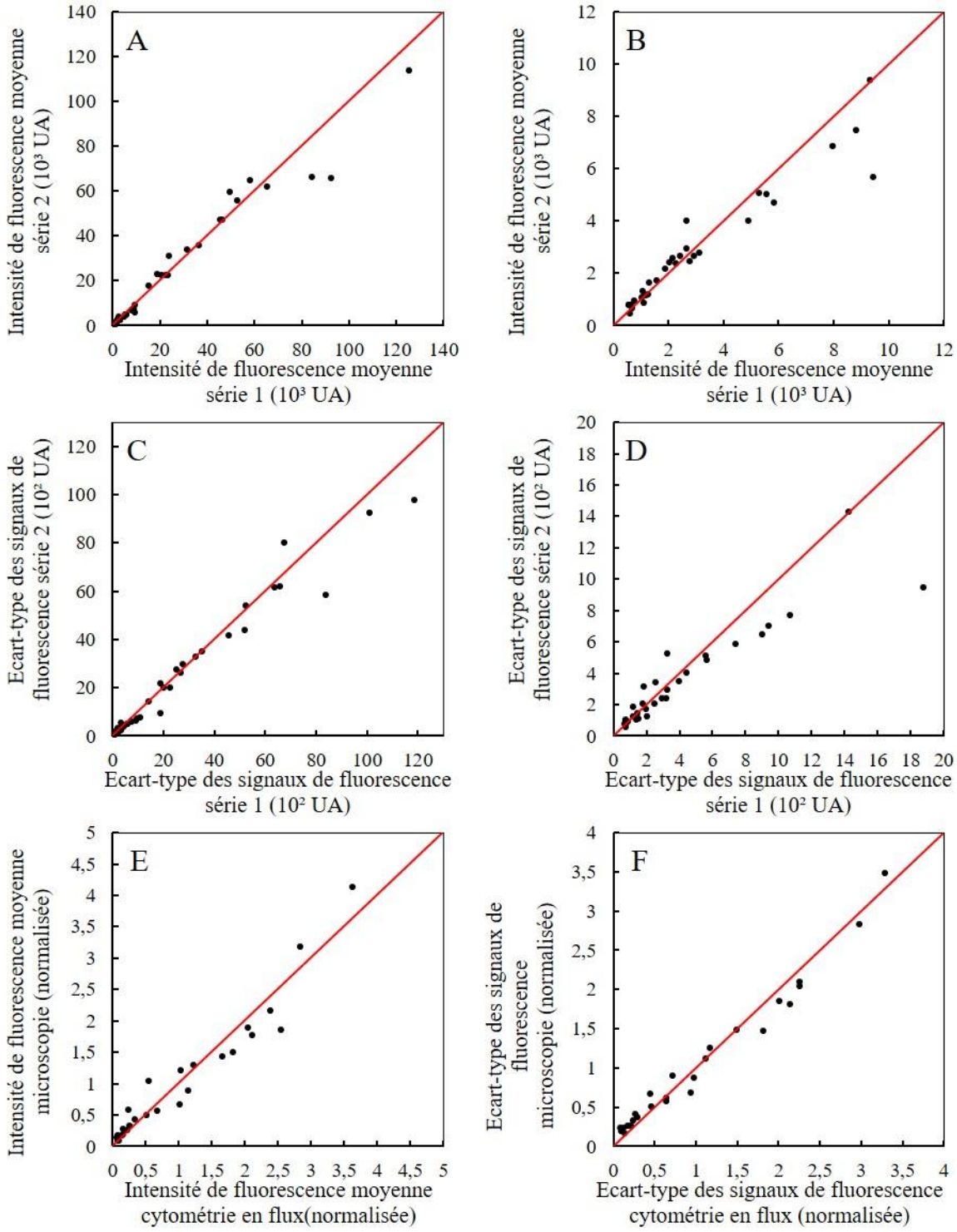


Figure 28 : Reproductibilité des niveaux moyens de fluorescence (**A-B**) et des écart-types des signaux de fluorescence (**C-D**) entre deux séries d'expériences en microscopie de fluorescence. **E-F**) Correspondance entre les expériences effectuées en cytométrie en flux et en microscopie.

II Caractérisation de la banque de souches :

La cytométrie en flux ayant une sensibilité moins importante que la microscopie, nous nous concentrerons sur les données de microscopie de fluorescence pour caractériser le niveau d'expression moyen ainsi que le bruit d'expression de chacune des souches de la banque originale de *B. subtilis*. D'autre part, d'après le séquençage de ces souches, la souche PL1S10 {rrnJP2-tufA} est mutée au niveau de la séquence promotrice de la cassette d'expression GFPmut3. Cette souche est donc exclue de notre analyse.

II.1 Niveau d'expression moyen des souches :

Comme on peut le voir sur la figure 29, la banque présente une large gamme de niveaux d'expression, qui est bien représentative de la gamme d'expression naturelle du protéome de *B. subtilis*. La caractérisation du niveau d'expression moyen de la protéine GFPmut3 des souches de la banque nous permet dans un premier temps de classer les différents promoteurs et TIR selon leur force. De cette caractérisation ressort le classement suivant de la force des promoteurs :

ykwB<yufK<yqzM<zwf<ykpA<fbaA<rrnJP2<ylxM,

Et le classement suivant de la force des modules TIR :

fbaAhs<fbaA<fbaAshort<gtlX<<tufA.

Toutes les souches suivent ce classement à l'exception de PL1S04 et PL1S07. Il est probable que l'expression dans ces deux souches soit perturbée par une interaction imprévue entre le promoteur et le TIR. On peut par exemple imaginer un repliement particulier de l'ARNm induit par la combinaison d'un TIR avec les derniers nucléotides du promoteur, qui correspondent aux quelques premiers nucléotides de l'ARNm. C'est pourquoi ces deux souches seront par la suite exclues de l'analyse.

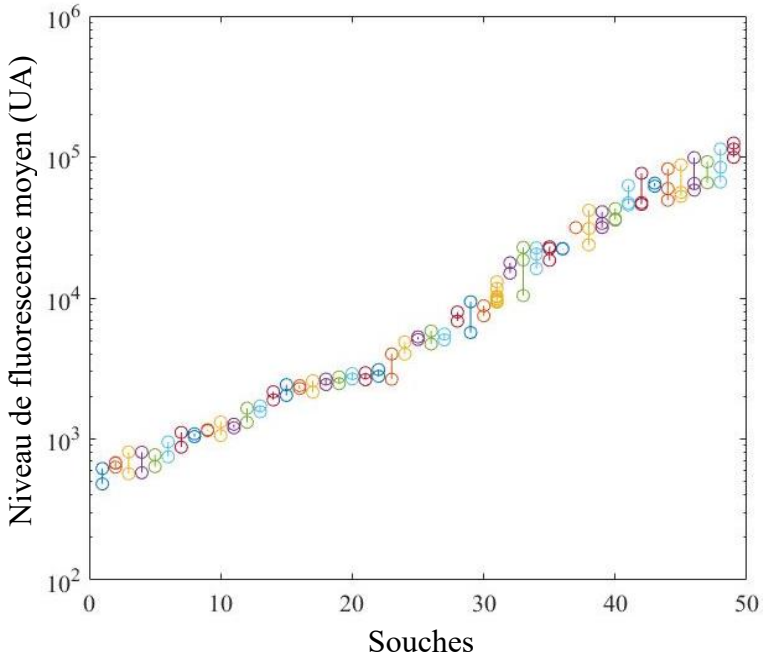


Figure 29 : Champs d'expression visité par la banque de *B. subtilis.* Suivant l'axe des abscisses est ordonné l'ensemble des souches dans l'ordre croissant de leur niveau d'expression moyen représenté sur l'axe des ordonnées. La plage d'expression s'étend sur ~10^3 décades, ce qui est représentatif d'une grande partie du protéome de *B. subtilis*. Pour chaque souche, les niveaux de fluorescence obtenue pour différentes séries d'expériences ont été reportés

II.2 Niveau du bruit d'expression de l'ensemble des souches :

On estime pour toutes les souches le bruit d'expression génique, défini ici comme le carré coefficient de variation. La dépendance du niveau de bruit avec le niveau d'expression moyen peut apporter des informations importantes sur les mécanismes responsables de la génération du bruit (Bar-Even et al, 2006, (Newman et al, 2006). On trace donc ce bruit d'expression en fonction du niveau moyen d'expression du fluorophore pour chaque souche (Figure 30). On constate que si le bruit est légèrement plus élevé pour les souches à très faible niveau d'expression, sur l'ensemble des souches le niveau de bruit est à peu près constant et indépendant du niveau moyen. Ce résultat est sensiblement différent de ce qui a été obtenu dans les études génomiques effectuées chez la levure *S. cerevisiae* menées par Bar-Even et al. (Bar-Even et al, 2006) et Newman et al. (Newman et al, 2006) (Figure 15). Pour rappel, ces deux études montrent que le bruit d'expression présentait deux régimes selon le niveau moyen d'expression du gène: une évolution du bruit inversement proportionnelle à la moyenne pour des gènes exprimées à des niveaux faibles et intermédiaires, signature des sources de bruits intrinsèques modélisées par les modèles stochastiques de l'expression génique; et pour les hauts niveaux d'expression, la tendance précédente n'est plus

vraie et le bruit d'expression atteint un niveau de bruit plateau en dessous duquel le bruit ne peut s'aventurer. Les différentes valeurs de bruit obtenues dans notre étude chez *B. subtilis* révèle que ce "plateau" est beaucoup plus vite atteint. En effet, si chez la levure la dépendance du bruit en $1/\langle p \rangle$ est encore vraie pour des niveaux moyens de 10^3 à 10^4 protéines, nous observons un plateau pour lequel le niveau de bruit est totalement décorrélé pour l'ensemble de la gamme d'expression visitée, sachant que le niveau d'expression le plus faible correspond à environ une cinquantaine de fluorophores. Cependant, si ce comportement est différent de celui constaté chez *S. cerevisiae*, il est cohérent avec celui observé chez la bactérie modèle *E.coli* dans l'étude menée par Taniguchi et al. Ces derniers observent en effet chez cette bactérie les deux mêmes régimes du niveau de bruit en fonction du niveau d'expression moyen, mais dans le cas de *E. coli*, le plateau est atteint beaucoup plus rapidement, c'est à dire dès que le niveau moyen excède 10 protéines. Ainsi, l'allure de la répartition des valeurs de bruit en fonction de la concentration moyenne de protéines observée dans notre banque chez *B. subtilis* est cohérent avec ce qui est constaté sur une majeure partie du génome chez *E. coli*.

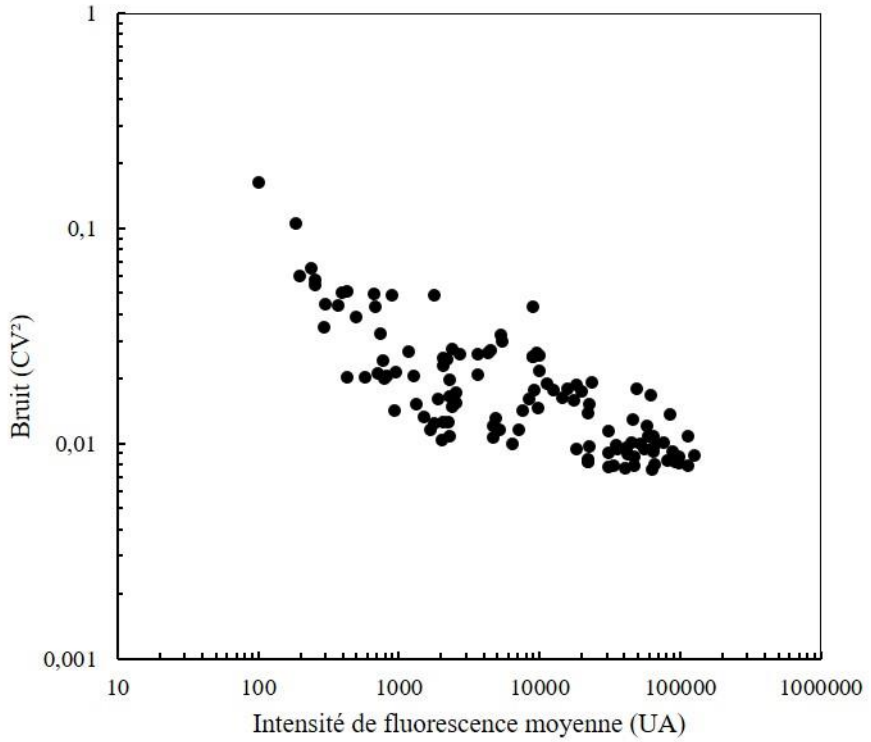


Figure 30 : Bruit d'expression génique en fonction du niveau moyen d'expression. Pour la majeure partie des niveaux d'expression visités, le niveau de bruit semble ne pas dépendre du niveau d'expression moyen et semble atteindre une limite à fort niveau d'expression.

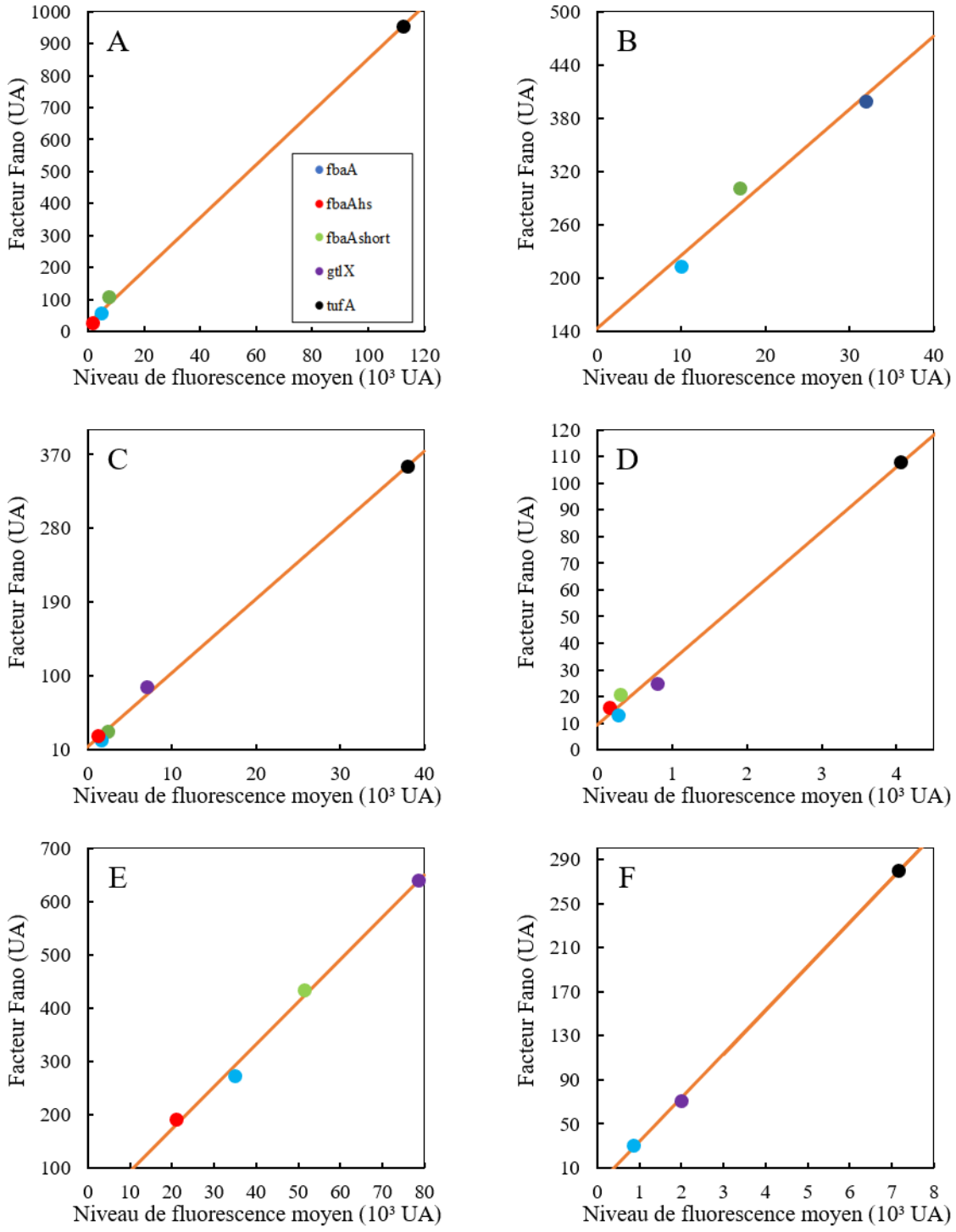
III Effets de la traduction et de la transcription sur le bruit d'expression génique :

Pour rappel, le modèle stochastique classique à deux niveaux prédit que pour un niveau d'expression donné, un gène fortement transcrit et faiblement traduit sera moins bruité qu'un gène faiblement transcrit et fortement traduit. Ozbudak et al (Ozbudak et al, 2002) ont cherché à vérifier expérimentalement cette prédiction chez *B. subtilis* en utilisant un rapporteur fluorescent exprimé sous le contrôle d'un promoteur inductible, dans 4 souches présentant des taux de traduction différents. Les résultats obtenus mettent en évidence une forte corrélation linéaire entre le facteur Fano (FF) et le taux de traduction tandis que le FF est indépendant du taux de transcription. Nous revisitons ce phénomène en mesurant pour chacune de nos souches le FF. La combinatoire de la banque de *B. subtilis* nous permet de tester les effets individuels de la traduction et de la transcription. La banque disposant de 8 différents promoteurs et 5 modules TIR, nous pouvons tester les effets de la traduction sur le bruit d'expression dans 8 conditions différentes en visitant pour chacune 5 différentes efficacités de traduction. De même, les effets de la transcription sur le bruit d'expression peuvent être évalués dans 5 conditions différentes en parcourant pour chaque conditions 8 différentes efficacités de transcription.

III.1 Effets de la traduction sur le bruit d'expression génique :

Pour chaque module promoteur, on parcourt l'ensemble des différents modules TIR et on mesure le niveau d'expression moyen et le facteur Fano du signal de fluorescence. On reporte ces données sur un graphe où apparaît en abscisse le niveau moyen d'expression et en ordonnée le facteur Fano pour chaque promoteur (Figure 31). Sur chaque figure, les différents TIR visités sont représentés par des couleurs différentes pour pouvoir les identifier d'un graphe à un autre : bleu pour fbaA, rouge pour fbaAhs, vert pour fbaAshort, magenta pour gtlX et noir pour tufA. On constate alors une évolution linéaire du facteur Fano avec le niveau d'expression moyen de chaque souche. Puisque le module promoteur est gardé constant, les différents niveaux moyens d'expressions reflètent différents taux de traduction. Ces résultats sont cohérents avec ce qui a été observé par Ozbudak et al. Sur chaque graphe, on effectue une régression linéaire de la répartition des points. Les pentes et ordonnées à l'origine sont résumées dans le tableau 11. Ces résultats sont vérifiés par cytométrie en flux. En effet, lorsque nous quantifions le bruit d'expression en microscopie de fluorescence, nous collectons des données émanant de cellules issues de plusieurs micro-colonies. Dans chaque micro-colonie, les cellules imagées à un instant donné appartiennent à la même lignée cellulaire, si bien qu'on pourrait imaginer que les quantités de protéines présentes dans les cellules d'une même colonie ne soient pas indépendantes, et que ces

corrélations introduites par notre stratégie de quantification du bruit viennent gêner nos conclusions. Pour contrôler cet éventuel “biais statistique”, nous utilisons les données de cytométrie en flux qui sont issues d’un tirage aléatoire des cellules contenues dans le milieu de culture. Les quantités de protéines mesurées dans chaque cellule sont dans ce cas *a priori* indépendantes. En ordonnant les moyennes et les facteurs Fano des signaux de fluorescence obtenus pour chaque souche en fonction des différents promoteurs, on obtient la même corrélation linéaire entre le facteur Fano et le taux de transcription.



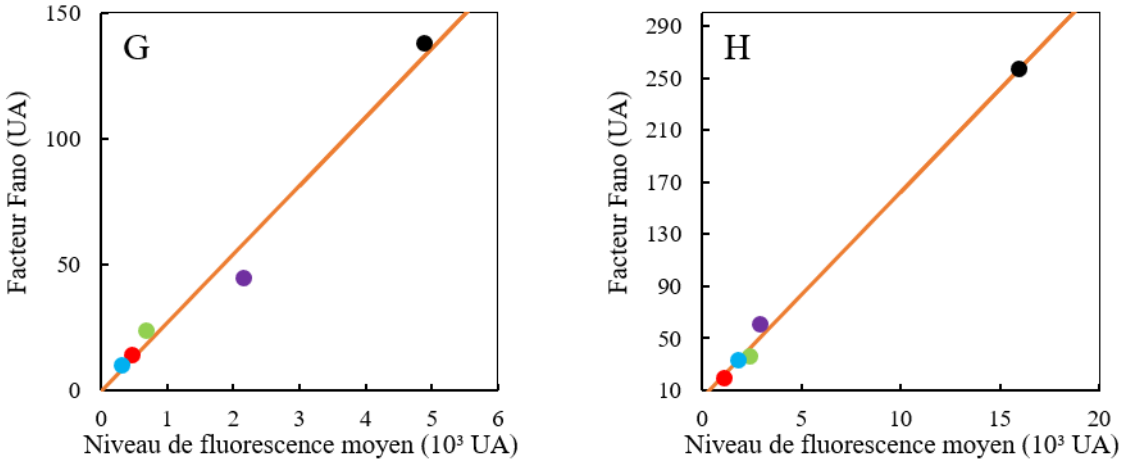


Figure 31 : Evolution du facteur Fano en fonction de la force de la traduction pour différents promoteurs. La force de traduction est modulée par différents module TIR (représenté par différentes couleurs). Chaque graphe correspond à un promoteur : **A)** fbaA **B)** rrnJP2 **C)** ykpA **D)** ykwB **E)** ylxM **F)** yqzM **G)** yufK **H)** zwf. Les courbes en rouge sont des régressions linéaires du type $y = ax + b$. Les résultats obtenus ici sont en adéquation avec le modèle à deux niveaux.

Promoteurs + TSS	Pente	Ordonnée à l'origine (AU)
ykwB	0,024	9,2
yufK	0,027	-0,25
yqzM	0,04	-5,9
zwf	0,016	4,5
ykpA	0,009	13
fbaA	0,0083	24
rrnJP2	0,0082	140
ylxM	0,008	13,1

Tableau 11 : Pentes et ordonnées à l'origine des régressions linéaires réalisées sur les données expérimentales (Figure 31). Les promoteurs sont classés de haut en bas du plus faible (ykwB) au plus fort (ylxM).

On en conclut que dans la gamme d'expression visitée, *une augmentation du taux de traduction s'accompagne d'une augmentation linéaire du facteur Fano*, ce qui semble confirmer le modèle stochastique à deux niveaux de l'expression génique.

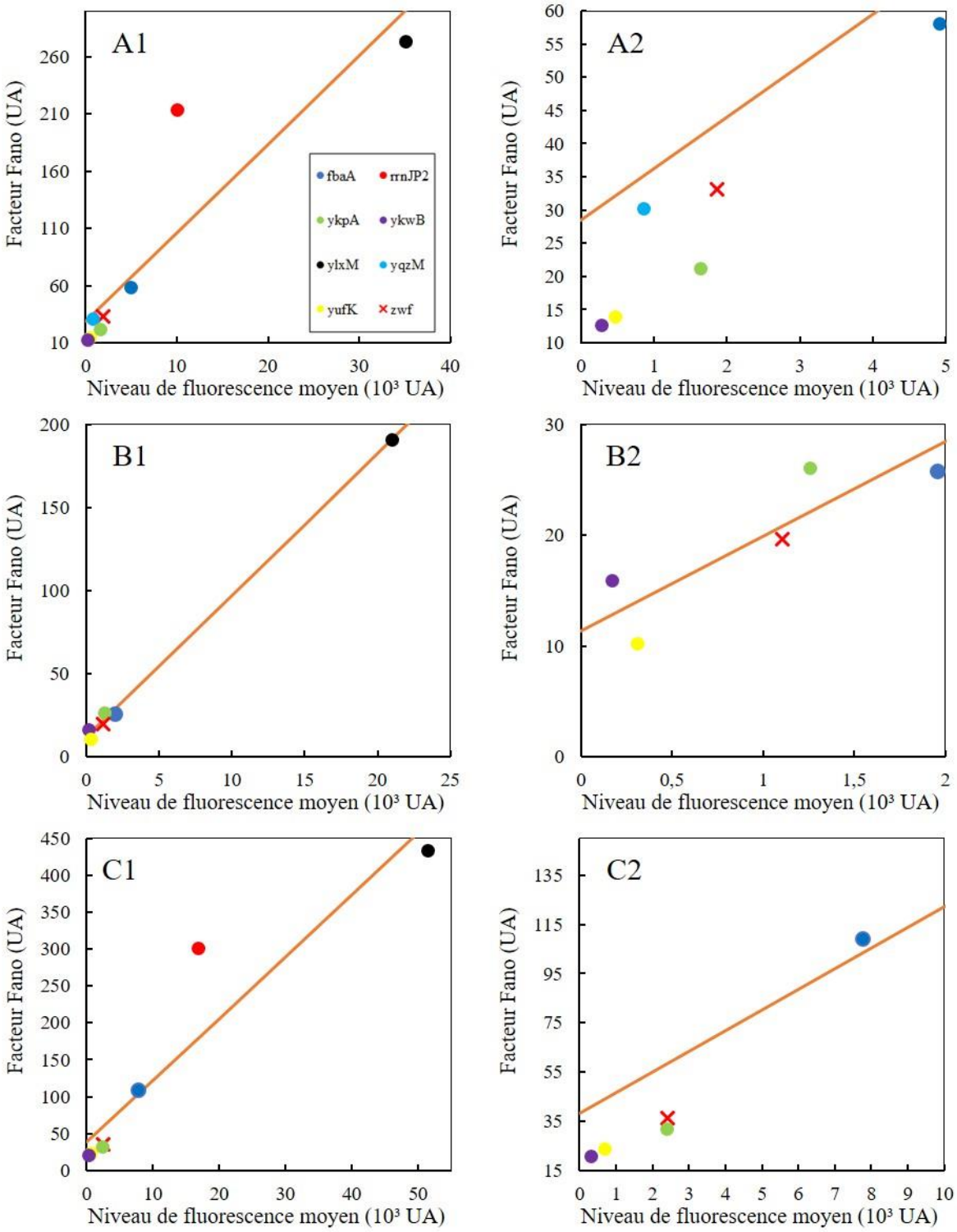
III.2 Effets de la transcription sur le bruit d'expression génique :

On procède de même pour la transcription. Pour chaque modules TIR, on parcourt les différents promoteurs, et on calcule pour chacune des souches correspondantes, à partir de la distribution des signaux de fluorescence, le niveau moyen d'expression et le facteur Fano. Les résultats obtenus pour les 5 différents TIR sont reportés sur la figure (Figure 32). Sur chaque graphe, les différents promoteurs sont indiqués par différentes couleurs et/ou différents symboles pour pouvoir les identifier d'un graphe à un autre. En bleu est indiqué le promoteur fbaA, en rouge le promoteur rrnJP2, en vert le promoteur ykpA, en magenta le promoteur ykwB, en noir le promoteur ylxM, en cyan le promoteur yqzM, en jaune le promoteur yufK et enfin la croix rouge correspond au promoteur zwf. On constate comme dans le cas précédent une évolution linéaire du facteur Fano avec le niveau d'expression moyen lorsqu'est modifié le promoteur. Une régression linéaire est effectuée pour chaque graphe. Les pentes et ordonnées à l'origine obtenues sont données dans le tableau 12.

On peut alors observer que les pentes obtenues avec certains modules TIR lorsque sont modifié les promoteurs sont comparables à ce que l'on obtient avec certains promoteurs lorsque nous faisons varier la force des modules TIR. Ainsi, la force du promoteur, c'est à dire sa capacité à recruter de nombreuses ARN polymérases (région en amont du +1 de transcription) et/ou à rapidement débuter la phase d'élongation de l'ARN messager (premiers nucléotides de l'ARNm) une fois la polymérase fixée à l'ADN, contribue également à l'établissement du niveau de bruit au sein d'une population monoclonale, et affecterait le facteur fano de la même manière que le taux de traduction; ce qui est en désaccord avec les résultats obtenus par Ozbudak et al et le modèle stochastique à deux niveaux de l'expression génique. Là encore, ce comportement inattendu est confirmé par les données recueillies par cytométrie en flux (Figure 32-F).

Il est cependant important de noter que les promoteurs tels que nous les avons définis contiennent les 8 premiers nucléotides transcrits (région appelée TSS, pour Transcription Start Site). Le repliement de l'ARN messager intervenant rapidement dès la transcription des premiers nucléotides en sortie de la polymérase, la séquence TSS y joue un rôle important, si bien que différentes séquences du TSS peuvent conduire à différents repliements, certains pouvant conduire par exemple à une séquestration du RBS et ainsi empêcher la sous-unité 30S du ribosome de se lier à l'ARN messager. De plus, le repliement de l'ARN messager peut également protéger l'extrémité 5' de l'action des ribonucléases et ainsi jouer sur la stabilité de l'ARN messager. La séquence du TSS pourrait donc avoir un effet sur le taux de traduction et/ou sur le taux de dégradation de l'ARN messager. La corrélation linéaire obtenue entre le facteur Fano et le niveau d'expression moyen résultant de différents promoteurs pourrait ainsi être due à un

effet traductionnel du TSS ou encore à différents temps de vie de l'ARN messager. Les résultats précédents sur l'étude des effets de la transcription sur le bruit d'expression génique doivent donc être contrôlés par la construction de nouvelles souches pour lesquelles les différents promoteurs seront suivis d'un TSS identique, de sorte que les ARN messagers soient similaires dans chacune de ces souches. La quantification du bruit d'expression associé à ces nouvelles souches devrait nous permettre de conclure quant aux effets de la transcription sur le bruit d'expression génique.



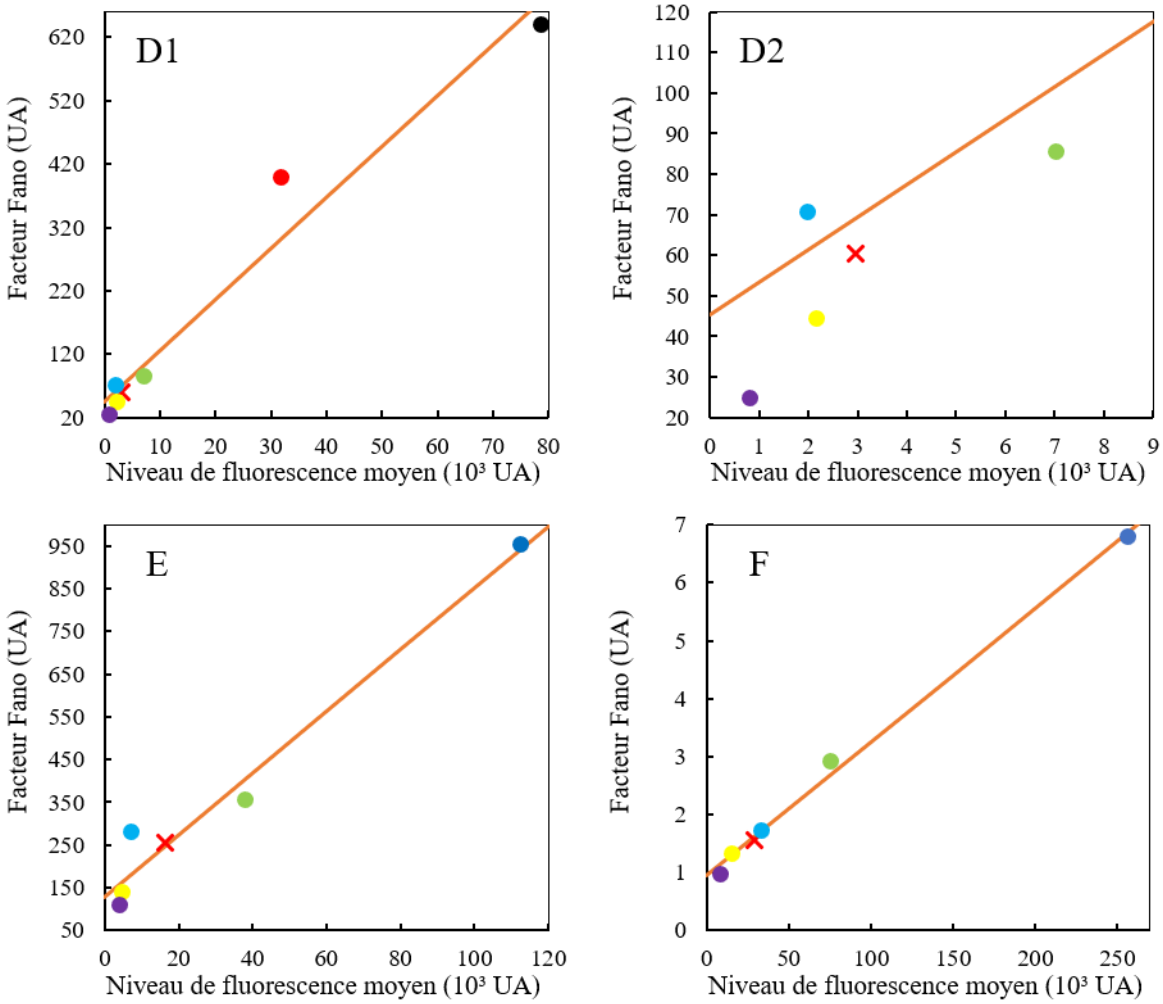


Figure 32 : Evolution du facteur Fano en fonction de la force de la transcription pour différents TIR. La force de transcription est modulée par différents promoteurs (représenté par différentes couleurs et marqueurs). Chaque paire de graphes correspond à un TIR : **A1-A2)** fbaA **B1-B2)** fbaAhs **C1-C2)** fbaAshort **D1-D2)** gtlX **E-F)** tufA. Le graphe (**F**) correspond aux données de cytométrie en flux. Les droites sont des régressions linéaires du type $y = ax + b$. Contrairement aux conclusions du modèle à deux niveaux, le facteur Fano dépend linéairement de la force des différents promoteurs.

TIR	Pente	Ordonnée à l'origine (AU)
fbaAhs	0.0086	11
fbaA	0.0077	28
fbaAshort	0.0084	38
gtlX	0.008	45
tufA	0.007	130

Tableau 12 : Pentes et ordonnées à l'origine des régressions linéaires réalisées sur les données expérimentales (Figure 32). Les TIR sont classés de haut en bas du plus faible (fbaAhs) au plus fort (tufA).

III.3 Effets de la transcription sur le bruit d'expression génique: Etude sur des souches avec TSS "identiques" :

Comme indiqué précédemment, pour étudier l'effet de la transcription nous devons maintenant séparer la région promotrice en amont du +1 de transcription de la région TSS, pour éliminer les éventuels effets de cette dernière sur la traduction et sur la dégradation des ARN messagers. Ceci nous permet de vérifier si la corrélation linéaire entre le facteur Fano et le niveau d'expression obtenue précédemment lorsqu'on varie l'ensemble {Promoteurs-TSS} est bien imputable au taux de transcription. Nous appellerons dorénavant promoteur uniquement la séquence en amont du +1 de transcription. Nous présentons dans un premier temps la construction de nouvelles souches avec les différents promoteurs et des TSS identiques. Puis nous présentons les résultats obtenus avec ces souches.

III.3.1 Construction des nouvelles souches avec TSS "identiques/unique" :

Avant d'être intégré au chromosome de *B. subtilis*, les différents assemblages Promoteur-TSS-TIR-GFP constituant la banque originale de *B. subtilis* ont été fabriqués sous forme de plasmide par assemblage Gibson. Nous utilisons la même stratégie ici pour la construction de nos nouvelles souches. On choisit comme TSS celui du gène fbaA et comme module TIR, le TIR fbaAshort. On procède dans un premier temps par PCR à une amplification séparée de la zone {Promoteur-TSS-TIR-GFP} et du corps de plasmide. Ces deux types d'amplification utilisent le plasmide utilisé pour construire la souche PL1S03 qui contient le TSS fbaA et le TIR fbaAshort. Pour la première amplification, nous utilisons 7 amorces sens (*forward*) qui s'hybrident sur la région du plasmide comprenant le TSS et les premiers nucléotides du TIR (leurs séquences ont été données dans la partie Matériel et Méthodes). En amont de la région d'hybridation se trouve pour chaque amorce un promoteur et une zone nécessaire à la recircularisation par assemblage Gibson. Les amorces anti-sens (*reverse*) s'hybrident après la partie codante de la GFP. La deuxième amplification sert à amplifier le corps de plasmide qui sera le même pour toutes les constructions. Ce corps de plasmide possède notamment les cassettes de résistance à l'ampicilline et à la spectinomycine ainsi qu'une zone d'homologie nécessaire à l'intégration dans le chromosome de *B. subtilis*. Les tailles des différents produits de PCR sont vérifiées par électrophorèse sur gel. Une fois validés, les produits de PCR sont mis en contact avec l'enzyme de restriction DpnI pour éliminer le plasmide matrice PL1S03, puis purifiés. Chaque assemblage {Promoteur-TSS-TIR} et le corps de plasmide (tous deux linéaires) présentent des extrémités homologues qui permettent l'assemblage final et donc la recircularisation par assemblage Gibson. Ces plasmides néosynthétisés sont utilisés pour transformer des cellules d'*E. coli* compétentes. Pour la transformation d'*E. coli*, les cellules compétentes sont placées en contact avec les différents plasmides. Après choc thermique, les

cellules sont étalées sur boîte de pétri contenant un milieu nutritif sélectif (milieu LB supplémenté d'ampicilline). Les cultures sont placées à 37°C pendant une nuit (*Over Night*). Le lendemain, plusieurs clones (ou colonie unique) sont sélectionnés sur chaque boîte de transformation, on en extrait les plasmides puis ces derniers sont purifiés et séquencés (séquençage selon la méthode de Sanger). Les plasmides correspondant bien aux assemblages désirés sont sélectionnés pour la transformation de *B. subtilis*. La souche de *B. subtilis* que nous utilisons est BSB168 (trp+). Après les avoir rendus compétentes, les cellules de *B. subtilis* sont mises en contact avec les plasmides transformants puis étalées sur milieu nutritif et sélectif (milieu LB supplémenté de spectinomycine) et placées à 37°C pour une culture *ON*. Au cours du processus de transformation, l'ADN plasmidique est intégré au chromosome de la cellule par recombinaison homologue (cross over) dans un locus spécifique. Ceci est rendu possible par la présence d'une zone homologue entre le plasmide transformant et le chromosome. Le lendemain, plusieurs colonies sont sélectionnées sur les boîtes de transformation et on procède alors à des PCR de vérifications sur ces colonies (criblage des clones). Une première vérification sur la présence de l'insert (assemblage Promoteur-TSS-TIR-GFP) dans le chromosome s'effectue en sélectionnant des amorces sens et anti-sens qui s'hybrident sur des zones encadrant ce dernier. Une deuxième vérification sur la présence des promoteurs consiste à choisir des amorces sens qui s'hybrident directement dans le promoteur de l'insert et une amorce anti-sens qui s'hybrident dans la partie codante de la GFP. Les clones donnant des signaux positifs lors de la migration des produits de PCR sur gel d'électrophorèse indiquent que l'insert est bien intégré et que le promoteur désiré s'y situe bien. Les clones positifs sont intégrés dans la banque originale de souches. Les nouvelles souches et les plasmides utilisées sont nommés de Alex01 jusqu'à Alex07, chaque souche exprimant un promoteur différent.

III.3.2 Evolution du facteur Fano suivant la force du promoteur :

Ces nouvelles souches possèdent toutes le même TSS et le même module TIR. A ses souches doit être ajoutée la souche PL1S03 de la banque originale qui possède également le TSS fbaA et le TIR fbaAshort. Ces 8 souches diffèrent uniquement par leur promoteur, si bien que des niveaux moyens d'expression différents reflètent des taux de transcription différents. Comme précédemment, on mesure pour chaque souche le niveau moyen et le facteur Fano des signaux de fluorescence. On reporte les résultats sur le graphe de la figure 33. On constate alors une augmentation linéaire du facteur Fano avec la force du promoteur. Une régression linéaire donne alors une pente comparable à ce que nous avons obtenu avec ce même module TIR mais lorsque la séquence promotrice s'étendait jusqu'au 8 premiers nucléotides transcrits, soit ~0,012. On en conclut alors que les corrélations linéaires entre le facteur Fano et le niveau moyen d'expression des fluorophores lorsqu'est modifié le module {Promoteur-TSS} s'explique

essentiellement par une modification du taux de transcription, et non par un effet sur la traduction ou sur la stabilité des ARN messagers. Cette tendance est encore une fois confirmée par les données de cytométrie en flux. Ainsi, *une augmentation du taux de transcription s'accompagne d'une augmentation du facteur Fano*, ce qui est en désaccord avec le modèle stochastique à deux niveaux de l'expression génique et avec les résultats expérimentaux d'Ozbudak et al.

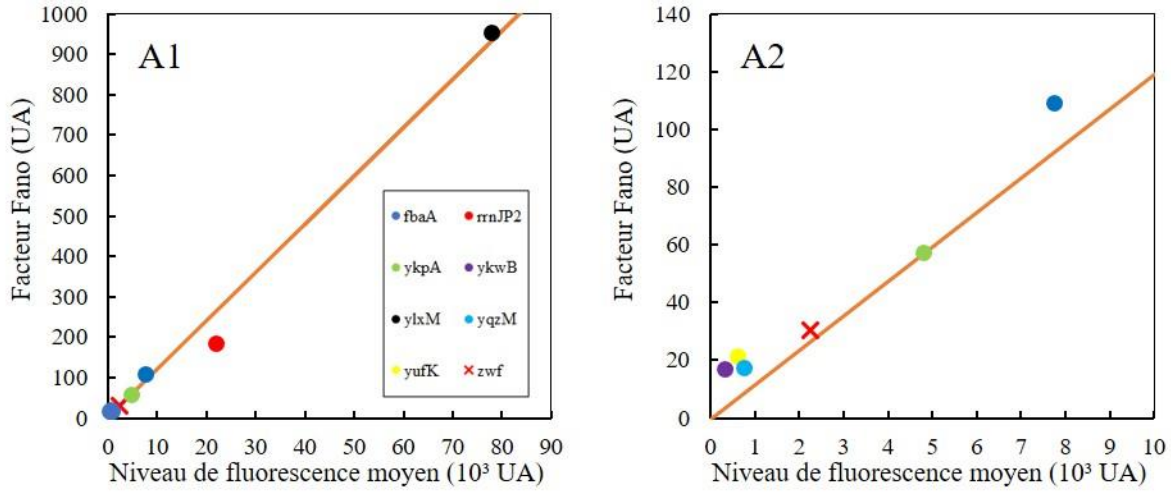


Figure 33 : Evolution du facteur Fano en fonction de la force de transcription lorsque seul change les 50 nucléotides en amont de la séquence TSS. La séquence TSS est ici gardée constante à travers les différentes souches tout comme le TIR. On obtient un comportement très similaire à celui obtenu sur les graphes de la figure 32, ce qui tend à montrer que l'évolution du facteur Fano avec le niveau moyen d'expression est uniquement due à une modification de la transcription.

III.4 Comparaison des différentes stratégies de modulation de l'expression génique (promoteurs, TSS, {promoteur-TSS}, modules TIR) :

- **Comparaison des effets sur le bruit d'expression génique de la modulation du TSS et de la modulation du promoteur :**

La même analyse menée sur les souches pour lesquelles seule change la séquence TSS (souches PL1B01 à PL1B07) permet de rendre compte de l'effet individuel de cette séquence sur la moyenne et le facteur Fano des niveaux d'expression du fluorophore. On peut alors comparer comment varie le facteur Fano avec le niveau d'expression moyen lorsqu'est modifié la séquence TSS d'une part et la séquence promotrice d'autre part. En effet, à partir de la souche contenant l'assemblage {rrnJP2-fbaA-fbaAshort} (Alex01 ou PL1B01), il est possible, grâce aux séries PL1B01-07 (+PL1S08) et Alex01-07 (+PL1S03), de respectivement modifier le TSS ou la séquence promotrice et de voir si ces deux stratégies conduisent à un comportement similaire ou bien si chacune emprunte un chemin différent dans le plan $(\langle p \rangle, FF)$. Les résultats sont reportés sur la figure 34-A. On constate alors que du point de vue du bruit d'expression génique, il est

identique de modifier la séquence du TSS (points bleus) ou la séquence promotrice (carrés rouges). Le point d'intersection des deux stratégies correspond aux souches PL1B01 et Alex01 qui présentent toutes les deux le même assemblage {Promoteur-TSS-TIR} en amont de la partie codante de la protéine GFPmut3. Ce résultat suggère alors que les effets sur le bruit d'expression génique du caractère transcriptionnel, stabilisateur de l'ARN messager, ou éventuellement traductionnel associé au TSS sont équivalents à ceux générés par les effets purement transcriptionnels associés au promoteur. Ceci peut également être observé dans le graphe de la figure 34-B, où nous avons placé dans le même plan $(\langle p \rangle, FF)$ l'ensemble des souches possédant le module TIR fbaAshort, c'est à dire les séries Alex01-07 (marqueurs rouge), PL1B01-07 (marqueurs bleus), et PL1S03-08-13-18-23-33-38 (croix noires). Ce graphe permet de résumer l'ensemble des stratégies adoptées ici ayant une action sur l'efficacité de transcription du gène codant pour la GFPmut3, à savoir une modification (i) du promoteur, (ii) du TSS ou (iii) de l'ensemble {Promoteurs-TSS}.

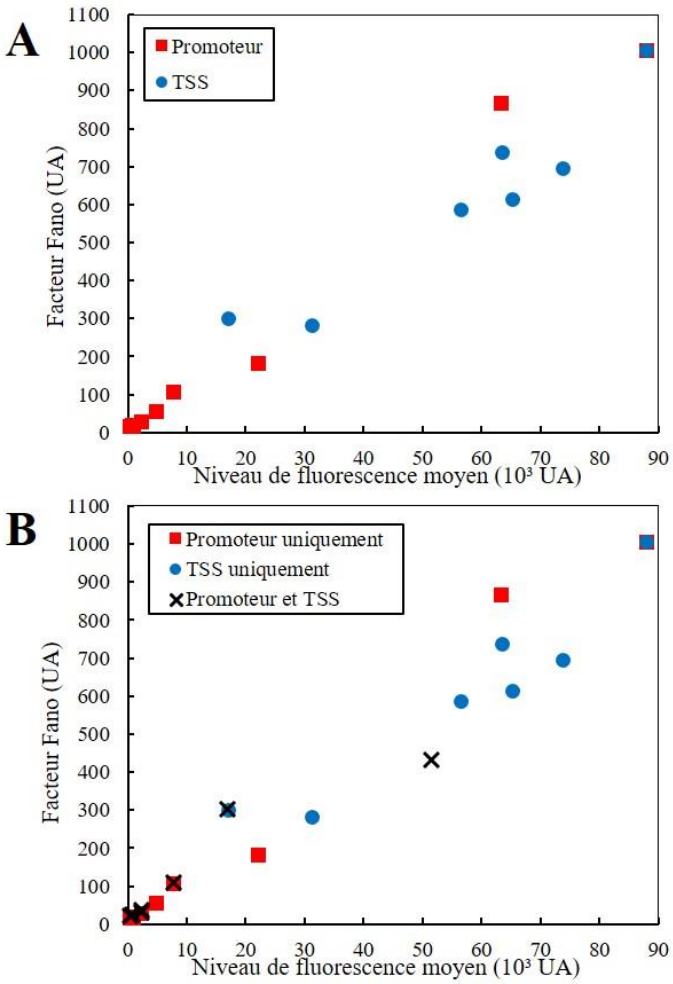
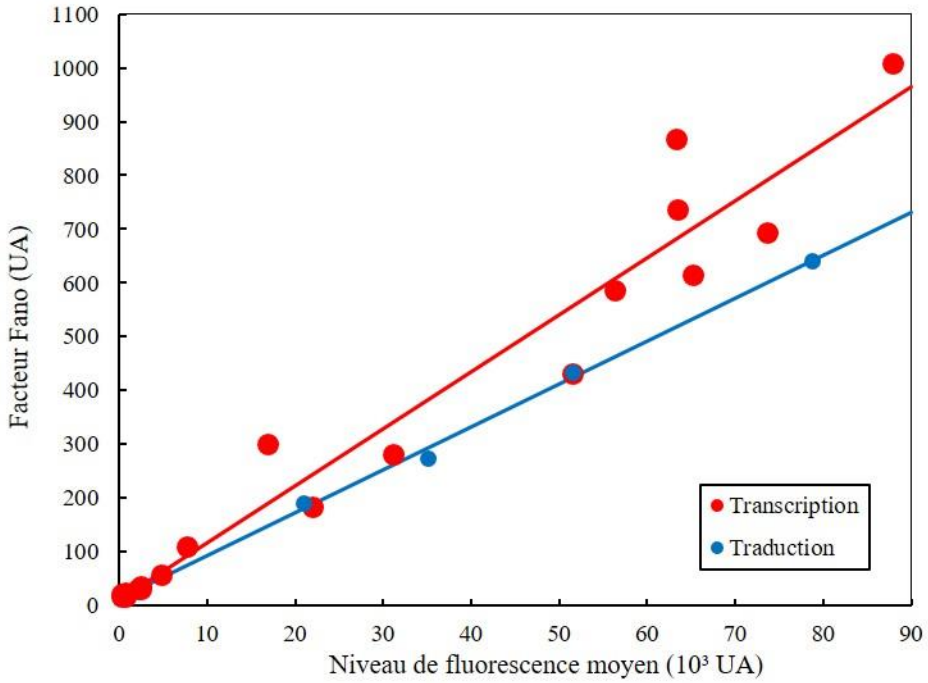


Figure 34 : Transcription et bruit d'expression. A) Une modification des 50 premiers nucléotides en amont du TSS (promoteur) ou du TSS donnent des résultats similaires sur le bruit d'expression génique. **B)** Une modification de l'ensemble {Promoteur-TSS} donne là encore des résultats similaires sur l'évolution du bruit d'expression génique.

Ces trois stratégies donnent des résultats qualitativement très similaire en ce qui concerne le bruit d'expression**.**

- **<u>Comparaison des effets sur le bruit d'expression génique entre la modulation du TIR et la modulation des ensembles jouant un rôle dans la transcription :</u>**

Nous savons que la région promotrice et le TSS jouent un rôle important sur le taux de transcription. Nous venons de voir que ces deux entités ont un effet similaire sur le bruit d'expression génique. L'évolution du facteur Fano évolue de manière similaire avec le niveau moyen d'expression selon que l'on modifie le TSS, le promoteur ou même les deux à la fois. On cherche maintenant à comparer cette tendance avec celle que donne les changements de TIR. Contrairement aux séquences génétiques précédentes, celle-ci ne joue qu'au niveau traductionnel. La figure 35 permet de comparer l'effet d'un changement de TIR avec un changement des différents ensembles jouant un rôle dans la transcription. A partir de la souche PL1S23 ayant pour promoteur et TSS ylxM et TIR fbaAshort (cette souche apparaît comme un point bleu et rouge superposés), on varie le module TIR (en bleu) ou les différents modules impactant la transcription (en rouge). On observe alors des effets quantitativement très similaire. Nous voyons ainsi que dans certaines conditions, il n'y a pas de distinction claire entre transcription et traduction par rapport au bruit d'expression génique. Ceci peut également être vu en comparant les pentes des régressions linéaires dans les tableaux 11 et 12.



<u>Figure 35 :</u> Evolution du facteur Fano selon que l'on modifie de manière indépendante la transcription (en rouge) ou la traduction (en bleu). On constate alors que dans certaines conditions, la traduction et la transcription peuvent avoir un effet similaire sur le bruit d'expression génique.

On remarque alors que selon les conditions de transcription ou de traduction, les pentes des courbes peuvent être très comparable.

Ainsi, dans le cas de *B. subtilis*, et dans la gamme de niveaux d'expression visitée, *pour un niveau d'expression moyen donné, un gène fortement transcrit et faiblement traduit ne sera pas nécessairement moins bruité qu'un gène faiblement transcrit et fortement traduit*. Nous présentons dans la partie qui suit une possible explication de ce comportement. Pour cela, nous nous basons sur les paramètres issus des régressions linéaires (pente et ordonnée à l'origine) sur les courbes donnant l'évolution du facteur Fano avec le niveau moyen d'expression obtenu lorsque varie soit le taux de transcription, soit le taux de traduction.

IV Hypothèse d'explication des comportements obtenus : présence du bruit extrinsèque :

A partir des pentes et ordonnées à l'origine des régressions linéaires effectuées sur les résultats de nos mesures, on est capable de dégager plusieurs tendances suivant que la traduction ou la transcription est modifiée. Ainsi, lorsqu'est modifié la traduction, on constate que plus le promoteur est fort, plus la pente de la régression linéaire est faible. On a un coefficient de corrélation de rang entre la pente et la force du promoteur de -0,9, ce qui montre une forte dépendance de la pente avec la force des promoteurs. Pour le promoteur le plus faible de notre collection à savoir le promoteur ykwB, la pente est de 0,024 tandis que pour les promoteurs les plus forts, rrnJP2 et ylxM, les pentes sont ~0,008. Les ordonnées à l'origine ont quant à elles une corrélation de rang moins forte avec la force du promoteur (0,76) (Tableau 11). De même, lors de l'étude des effets de la transcription sur le bruit d'expression, on peut constater que les pentes des régressions linéaires sont constantes, de l'ordre de ~0,008, quel que soit le TIR, (coefficient de corrélation de rang de -0,7) mais les ordonnées à l'origine quant à elle semble croître avec la force du module TIR (coefficient de corrélation de rang de 1) (Tableau 12). D'une valeur de 11 AU pour le TIR le plus faible (fbaAhs) jusqu'à 130 AU pour le TIR le plus fort (tufA). Si ces résultats ne peuvent pas être expliqués par le modèle stochastique à deux niveaux ne prenant en compte que les sources de stochasticité intrinsèques, ils peuvent s'expliquer si on considère les sources de bruit extrinsèques. Taniguchi et al ont introduit les sources de bruit extrinsèques en permettant aux paramètres biophysiques du modèle intrinsèque de fluctuer selon des distributions stationnaires. En regroupant les paramètres dans les variables a $(= k_R/\gamma_P)$ et b $(= k_P/\gamma_R)$ représentant respectivement le nombre moyen d'ARN messagers produits durant le temps de vie d'une protéine et le nombre moyens de protéines synthétisé à partir d'un transcrit, le bruit d'expression génique s'écrit :

$$\eta_p^2 = \frac{1+\langle b\rangle}{\langle p\rangle} + \eta_b^2\frac{\langle b\rangle}{\langle p\rangle} + \eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2$$

Cette équation peut être séparée en deux parties, une représentant les sources de bruits intrinsèques et une autre qui représente la contribution des sources de bruits extrinsèques :

$$\eta_{int}^2 = \frac{1+\langle b\rangle}{\langle p\rangle}$$

$$\eta_{ext}^2 = \eta_b^2\frac{\langle b\rangle}{\langle p\rangle} + \eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2$$

Le facteur Fano se déduit facilement du bruit d'expression :

$$\frac{\sigma_p^2}{\langle p\rangle} = 1 + \langle b\rangle + \eta_b^2\langle b\rangle + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\langle p\rangle$$

Où $\eta_a^2 = \sigma_a^2/\langle a\rangle^2$, $\eta_b^2 = \sigma_b^2/\langle b\rangle^2$ et $\langle p\rangle = \langle a\rangle\langle b\rangle$. Si on suppose qu'uniquement le taux de traduction est modifiée, soit la quantité $\langle b\rangle$ par modulation de k_P, alors on élimine la quantité $\langle b\rangle$ en la remplaçant par $\langle p\rangle/\langle a\rangle$ dans l'expression du facteur Fano de sorte que les modulations de $\langle b\rangle$ soit contenue dans la variable $\langle p\rangle$:

$$\frac{\sigma_p^2}{\langle p\rangle} = 1 + \frac{\langle p\rangle}{\langle a\rangle} + \eta_b^2\frac{\langle p\rangle}{\langle a\rangle} + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\langle p\rangle$$

$$= 1 + \left(\frac{1}{\langle a\rangle}(1+\eta_b^2) + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\right)\langle p\rangle \qquad (1)$$

Si maintenant on suppose que seule la transcription est modifiée, soit $\langle a\rangle$, on doit considérer l'équation :

$$\frac{\sigma_p^2}{\langle p\rangle} = 1 + (1+\eta_b^2)\langle b\rangle + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\langle p\rangle \qquad (2)$$

Pour l'ensemble des souches de *B. subtilis,* nous supposons que les quantités η_a^2 et η_b^2 sont constantes. Cette hypothèse semble raisonnable pour deux raisons : (i) les variations des quantités a et b sont par exemple dû aux fluctuations du nombre d'ARN polymérases et de ribosomes et sont identiques dans les différentes souches ; (ii) les nombres de polymérases et de ribosomes affectent multiplicativement a et b *à priori* : $k_R = \widetilde{k_R}[ARNpol]$, où $\widetilde{k_R}$ représente la force du promoteur et $[ARNpol]$ la concentration d'ARN polymérases dans la cellule.

On note par la suite $C_1 = \eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2$ et $C_2 = 1 + \eta_b^2$. Les équations (1) et (2) s'écrivent donc respectivement :

$$\frac{\sigma_p^2}{\langle p \rangle} = 1 + \left(\frac{C_2}{\langle a \rangle} + C_1\right)\langle p \rangle \qquad (3)$$

$$\frac{\sigma_p^2}{\langle p \rangle} = 1 + C_2\langle b \rangle + C_1\langle p \rangle \qquad (4)$$

On peut tirer à partir de ces équations plusieurs prédictions. Premièrement, ces équations sont des fonctions linéaires de l'abondance moyenne des protéines $\langle p \rangle$, que ce soit lorsque c'est la transcription qui est modifiée ou la traduction. Deuxièmement, d'après l'équation (3), quand on modifie le taux de traduction dans différentes conditions de transcription, la pente dépend de la force du promoteur ($\langle a \rangle$) , et décroît avec cette dernière, tandis que l'ordonnée à l'origine est constante. Cette prédiction est en accord avec nos résultats expérimentaux où nous avons vu que plus le promoteur était fort, plus la pente de la régression linéaire était faible tandis que l'ordonnée à l'origine, bien que non constante, ne semble pas être corrélé avec la force des promoteurs utilisés. La pente semble d'autre part converger vers ~0,008. Cette valeur de pente est atteinte dès le promoteur ykpA. Cette convergence semble indiquer d'après l'équation (3) que :

$$C_1 = \eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2 \sim 0{,}008$$

Troisièmement, d'après l'équation (4), quand on modifie le taux de transcription, les pentes des régressions linéaires obtenues dans différentes conditions de traduction doivent être constante tandis que les ordonnées à l'origine doivent augmenter avec la force de la traduction. De plus, les pentes des régressions linéaires doivent être égale à la constante C_1, c'est-à-dire la pente que l'on obtient lorsqu'on fait varier la traduction en maintenant un promoteur fort. Là encore, les données expérimentales confirment ces dernières prédictions. En effet, les pentes sont constantes et se répartissent autour de 0,008, ce qui correspond aux pentes obtenues lorsque nous avons modifié les modules TIR dans des contextes de transcription forte, et les ordonnées à l'origine augmentent bien avec la force des modules TIR. Ainsi, les prédictions du modèle de Taniguchi et al combinant sources de bruits intrinsèques et extrinsèques sont vérifiées par les différentes régressions linéaires effectuées sur nos résultats expérimentaux, qui imposent une forme du facteur Fano (en unité arbitraire de fluorescence) du type :

$$\frac{\sigma_p^2}{\langle p \rangle} = C_3 + C_2\langle b \rangle + C_1\langle p \rangle$$

Le modèle de Taniguchi et al permet d'interpréter les constantes C_1 et C_2 en termes de bruits extrinsèques. Nous devons maintenant vérifier que les sources de bruits extrinsèques jouent bien un rôle important dans la gamme d'expression visitée pour qu'elles puissent expliquer les comportements obtenus. Pour cela, nous adoptons la stratégie « deux couleurs » mise en place par Elowitz et al. Comme précédemment lorsque nous avons contrôlé la possibilité d'effets sur la traduction ou sur la stabilité des ARN messagers de la séquence TSS, nous devons construire de nouvelles souches exprimant deux fluorophores, et imager ces nouvelles souches de la même manière que les précédentes.

V Décomposition du bruit d'expression total en bruit extrinsèque et intrinsèque :

V.1 Construction des souches exprimant les deux fluorophores :

Nous ajoutons le gène codant pour un nouveau fluorophore, la protéine mKate2, dans le chromosome de *B. subtilis* dans un locus voisin de celui exprimant la protéine GFPmut3. Dans chaque construction les séquences codantes pour ces protéines sont contrôlées par le même promoteur, le même TSS et le même TIR. L'augmentation linéaire du facteur Fano avec l'abondance moyenne des protéines est attendue lorsque varie le taux de traduction mais pas lorsque varie le taux de transcription. Ainsi, nous décidons d'ajouter le second fluorophore dans le chromosome de souches permettant de visiter l'ensemble des promoteurs mais avec des séquences TSS et des modules TIR identiques ; c'est à dire les souches Alex[01-07] et PL1S03. Nous avons besoin pour cela des plasmides synthétisés dans la partie précédente (Alex[01-07]) et du plasmide PL1S03, et d'un plasmide supplémentaire dans lequel se trouve la séquence codant pour la protéine mKate2. Il s'agit donc de construire plusieurs inserts contenant pour chacun un module promoteur différent, un module TSS fbaA et un module TIR fbaAshort placés en amont d'une séquence codant pour la protéine mKate2. Ces différentes cassettes d'expression sont assemblées et clonées dans les plasmides Alex[01-07] et PL1S03 en plusieurs étapes: (i) On amplifie l'ensemble des plasmides Alex[01-07] et PL1S03 par PCR. Les fragments constitués à partir de chaque plasmide constituent les corps de plasmide et contiennent notamment la cassette d'expression de la GFPmut3, les cassettes de résistance aux antibiotiques ampicilline et spectinomycine ainsi que la zone d'homologie nécessaire à la recombinaison homologue avec le chromosome de *B. subtilis*. (ii) On isole et on amplifie par PCR sur les plasmides matrices Alex[01-07] et PL1S03 les différents assemblages {Promoteur-TSS-TIR}. Les amorces sens spécifiques à chaque promoteur permettent d'ajouter une zone de liaison avec le corps du plasmide sur une des extrémités. (iii) On isole et on amplifie par PCR à partir du plasmide

matrice pSG15 la séquence codant pour la protéine mKate2. On y ajoute grâce aux séquences amorces une zone permettant la liaison avec l'ensemble {Promoteur-TSS-TIR} sur une des extrémités du produit de PCR, et sur l'autre extrémité une zone permettant la liaison avec le corps de plasmide. (iv) Les trois fragments linéaires obtenus par PCR, à savoir les corps de plasmides contenant la séquence GFP, les différents ensembles {Promoteur-TSS-TIR} et la partie codante pour la protéine mKate2 sont lysés par DpnI pour digérer les plasmides matrices méthylés utilisés dans les différentes PCR, puis purifiés. Après digestion, les trois fragments constitués à chaque étape précédente sont assemblés par assemblage Gibson à l'aide des différentes zones de recouvrement entre fragments. Les 8 plasmides ainsi obtenus se nomment respectivement Alex[01-07]-mKate et PL1S03-mKate. Comme précédemment, on transforme des souches de *E. coli* compétentes à partir de ces plasmides. Pour la transformation d'*E. coli*, les cellules compétentes sont placées en contact avec les différents plasmides. Après plusieurs chocs thermiques, les cellules sont étalées sur boîte de pétri contenant un milieu nutritif sélectif (milieu LB supplémenté d'ampicilline et de spectinomycine). Les cultures sont placées à 37°C pendant une nuit. Le lendemain, plusieurs clones (ou colonie unique) sont sélectionnés sur chaque boîte de transformation, on en extrait les plasmides, on les purifie et on procède au séquençage (séquençage selon la méthode de Sanger). Les plasmides correspondant bien aux assemblages désirés (ne présentant pas de mutations dans les différentes zones d'intérêt) sont sélectionnés pour la transformation de *B. subtilis* BSB 168. Après préparation des cellules compétentes, on procède à la transformation de *B. subtilis*. Pour cela, les différents plasmides sont mis en contact avec les cellules compétentes à 37°C pendant 1h sous agitation puis étaler sur milieu nutritif sélectif (milieu LB supplémenté avec de la spectinomycine) et placé à 37°C pour une nuit. Le lendemain, on collecte différents clones par boîtes de transformation que l'on soumet à des PCR de vérifications. Pour les constructions "deux couleurs", on effectue deux types de PCR sur colonie : une première PCR pour vérifier l'intégration au bon locus de la zone {Promoteur-TIR-GFPmut3} d'une part et une seconde PCR pour vérifier la présence du groupement {Promoteur-TIR-mKate2} d'autre part, à l'aide d'amorces s'hybridant dans différentes régions du chromosome. Les clones donnant des résultats positifs après migration sur gel des produits de PCR sont ensemencés sur boîtes de pétri fraîches (LB + spectinomycine). On obtient ainsi plusieurs clones par constructions génétiques. Ces constructions étant sensiblement plus compliquées que les précédentes, on mesure les niveaux de fluorescence moyens de chaque fluorophore pour l'ensemble des clones par microscopie de fluorescence (on image 5 colonies pour chaque clone retenue). Ceci permet de vérifier si (i) les signaux du fluorophore GFPmut3 dans ces nouvelles souches "deux-couleurs" sont identiques avec les signaux de ce même fluorophore dans les souches "mono-couleurs" (Alex[01-07] + PL1S03). Ceci permet de vérifier que les deux gènes rapporteurs sont indépendants statistiquement. Enfin, on

vérifie également si (ii) les signaux moyens de fluorescence de la protéine GFPmut3 et de la protéine mKate2 sont corrélées à travers les différentes souches (les rapports entre les fluorescences moyennes de la GFPmut3 et de la mKate2 doivent être comparable d'un clone à un autre et d'une souche à une autre). On retient pour chaque souche un clone ayant validé le contrôle précédent. Ces souches se nomment respectivement Alex[01-07]-mKate et PL1S03-mKate. Seul la construction Alex04-mKate ne répond pas aux critères, et est donc exclue de l'analyse. Ces 7 souches supplémentaires sont imagées de manière à pouvoir recueillir les signaux de fluorescence de la GFPmut3 (verte) et de la mKate2 (rouge) (Figure 36).

V.2 Quantification de la contribution des sources de bruit extrinsèque et intrinsèque :

Les souches constituées précédemment possèdent dans leur chromosome deux gènes rapporteurs (GFPmut3 et mKate2) situés à des loci voisins dans le chromosome. Pour chaque souche, les deux gènes sont contrôlés par des séquences promotrices, TSS et TIR similaires. A partir des signaux GFPmut3 et mKate2 collectés, on peut observer et quantifier le degré de corrélation de ces signaux d'une part et quantifier les bruits intrinsèques et extrinsèques d'autre part (Figure 36). L'observation des corrélations nous permet de rendre compte qualitativement de l'importance de l'une ou l'autre des deux types de bruits et d'observer son évolution avec le niveau d'expression moyen et dans notre cas avec la force du promoteur (avec nos constructions, seule la séquence promotrice change d'une souche à une autre). Pour cela, nous plaçons dans le plan donnant en ordonnée le signal de fluorescence de la protéine mKate2 (p_m) et en abscisse le signal de fluorescence de la protéine GFPmut3 (p_g), l'ensemble des données de fluorescence recueillies pour les deux fluorophores de toutes les cellules imagées. Les fluctuations corrélées des niveaux de GFPmut3 et de mKate2 entre cellules sont la signature des sources de bruits extrinsèques, tandis que les fluctuations entre cellules non corrélées sont la signature des sources de bruit intrinsèque (Figure 37). On observe alors que pour l'ensemble de nos souches, et donc pour l'ensemble des promoteurs visités, les coefficients de corrélations linéaires sont forts (de ~0,5 à 0,86) et le sont d'autant plus que le niveau moyen d'expression (ou la force du promoteur) est élevé (coefficient de corrélation de rang (Spearman) de ~0,94). On en déduit donc que les sources de bruit extrinsèques ont un effet majeur sur le bruit d'expression génique total.

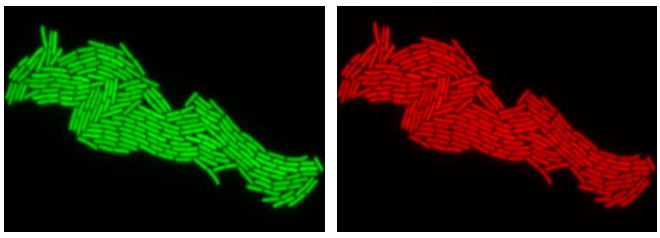


Figure 36 : Images de fluorescence GFP (droite) et mKate2 (gauche) d'une micro-colonie de *B. subtilis* (ici Alex01-mKate2). Le niveau de corrélation entre les deux signaux permet d'estimer l'importance relative des sources de bruit intrinsèque et extrinsèque sur le bruit d'expression total.

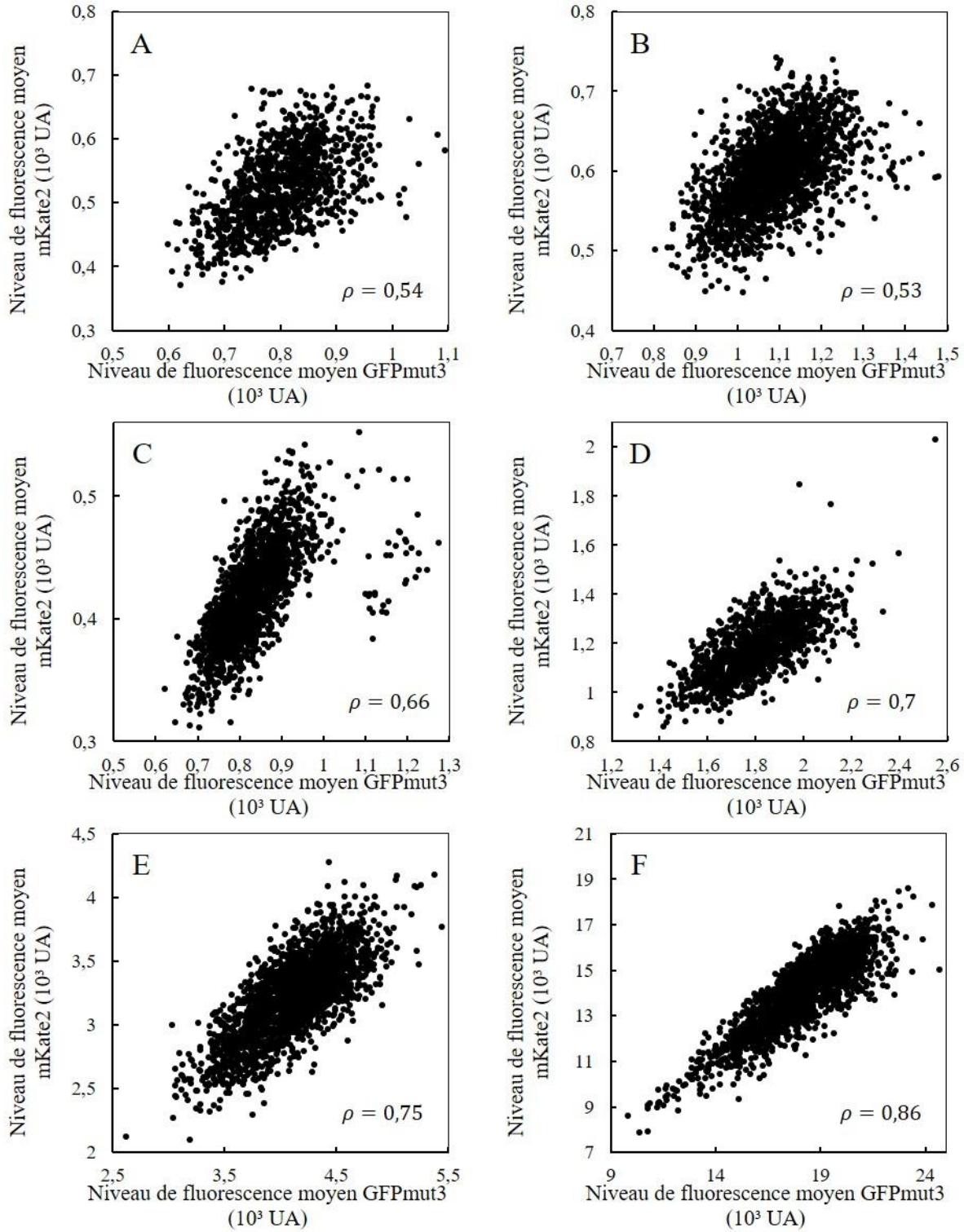


Figure 37 : Corrélation entre les signaux de fluorescence mKate2 et GFPmut3. Pour chaque graphe, on place en ordonnée le signal de fluorescence issue du fluorophore mKate2 et en abscisse le signal de fluorescence du fluorophore GFPmut3. Chaque point représente une cellule unique. On indique également sur chaque graphe le coefficient de corrélation linéaire (Pearson). Un graphe correspond à une souche et ont été ordonné selon le niveau d'expression moyen de la souche (du plus faible au plus fort). On observe des niveaux de corrélation relativement élevé qui augmente suivant la force des promoteurs, ce qui montre une domination plus forte des sources de bruit extrinsèques sur les sources de bruit intrinsèques au fur et à mesure qu'augmente la force de transcription.

On décompose maintenant le bruit d'expression total mesuré. En présence de bruit intrinsèque et extrinsèque, la variance totale d'une quantité stochastique *x* peut être décomposée selon (Swain et al, 2002 ; Schultheiβ Araùjo et al, 2018) : $\sigma^2 = \sigma_{int}^2 + \sigma_{ext}^2$, avec :

$$\sigma_{int}^2 = \langle\langle x^2|z\rangle\rangle_e - \langle\langle x|z\rangle^2\rangle_e$$

$$\sigma_{ext}^2 = \langle\langle x|z\rangle^2\rangle_e - \langle\langle x|z\rangle\rangle_e^2$$

Où *z* dénote l'ensemble des variables ou processus stochastiques extrinsèques, les crochets $\langle\ \rangle$ représentent la moyenne sur une sous population de cellules ayant une valeur donnée de *z.* Cette moyenne est donc calculée pour toutes les valeurs des variables stochastiques représentant les sources de bruit intrinsèque. Les crochets $\langle\ \rangle_e$ indique une moyenne pour toutes les valeurs possibles de *z*, donc sur l'ensemble de la population (si elle présente un assez grand nombre d'individus, on peut supposer que l'ensemble des valeurs possibles de *z* sont représentées dans la population). Dans la méthode « deux-couleurs », où on caractérise le bruit d'expression génique à l'aide des protéines GFPmut3 et mKate2. On note les variables stochastiques *g* et *m* représentant respectivement le niveau d'abondance de GFPmut3 et de mKate2 dans une cellule. On écrit :

$$\langle\langle(g-m)^2|z\rangle\rangle_e = \langle\langle g^2|z\rangle\rangle_e + \langle\langle m^2|z\rangle\rangle_e - 2\langle\langle g|z\rangle\langle m|z\rangle\rangle_e$$

$$\langle\langle gm|z\rangle\rangle_e - \langle\langle g|z\rangle\rangle_e\langle\langle m|z\rangle\rangle_e = \langle\langle g|z\rangle\langle m|z\rangle\rangle_e - \langle\langle g|z\rangle\rangle_e\langle\langle m|z\rangle\rangle_e$$

La dernière égalité est uniquement vraie si les deux rapporteurs sont statistiquement indépendants, ce qui est le cas avec les souches construites. De plus, les deux rapporteurs étant contrôlés par les mêmes séquences promotrices, TSS et TIR, ils sont identiques et leur abondance moyenne sont égales, si bien que :

$$\langle\langle(g-m)^2|z\rangle\rangle_e = 2\langle\langle m^2|z\rangle\rangle_e - 2\langle\langle m|z\rangle^2\rangle_e = 2\sigma_{int}^2$$
$$\langle\langle gm|z\rangle\rangle_e - \langle\langle g|z\rangle\rangle_e\langle\langle m|z\rangle\rangle_e = \langle\langle m|z\rangle^2\rangle_e - \langle\langle m|z\rangle\rangle_e^2 = \sigma_{ext}^2$$

D'après les définitions des bruits intrinsèques et extrinsèques

$$\eta_{int}^2 = \frac{\sigma_{int}^2}{\langle m\rangle^2}; \qquad \eta_{ext}^2 = \frac{\sigma_{ext}^2}{\langle m\rangle^2}$$

Et des facteurs Fano (*FF*) associés :

$$FF_{int} = \frac{\sigma_{int}^2}{\langle m\rangle}; \qquad FF_{ext} = \frac{\sigma_{ext}^2}{\langle m\rangle}$$

On arrive aux expressions :

$$\eta_{int}^2 = \frac{\langle (g-m)^2 \rangle}{2\langle m \rangle \langle g \rangle}; \qquad \eta_{ext}^2 = \frac{\langle gm \rangle - \langle g \rangle \langle m \rangle}{\langle g \rangle \langle m \rangle}$$

$$FF_{int} = \frac{\langle (g-m)^2 \rangle}{2\sqrt{\langle m \rangle \langle g \rangle}}; \qquad FF_{ext} = \frac{\langle gm \rangle - \langle g \rangle \langle m \rangle}{\sqrt{\langle g \rangle \langle m \rangle}}$$

On utilise la notation $\langle\ \rangle$ pour condenser les moyennes effectuées pour l'ensemble des variables stochastiques intrinsèques et extrinsèques. Ces moyennes sont celles effectuées sur l'ensemble de la population. Dans les expériences que nous avons réalisées, nous n'avons pas accès directement aux nombres de molécules de fluorophores, mais uniquement aux signaux de fluorescence qu'ils émettent une fois excités. Le signal de fluorescence que l'on recueille par cellule est directement proportionnel à l'abondance du fluorophore dans une cellule :

$$p_g = k_g g;\ p_m = k_m m$$

Les constantes de proportionnalité des deux fluorophores ne sont pas égales. La constante de proportionnalité dépend en effet notamment des propriétés de chaque fluorophore comme la brillance ou le rendement quantique, mais aussi des paramètres d'illuminations avec lesquelles on excite chaque fluorophore. Ces différences rendent compliquées l'estimation du bruit intrinsèque à partir des données de fluorescence :

$$\frac{\langle (p_g - p_m)^2 \rangle}{2\langle p_g \rangle \langle p_m \rangle} = \frac{k_g}{k_m} \frac{\langle \left(g - \frac{k_m}{k_g} m\right)^2 \rangle}{2\langle g \rangle \langle m \rangle} \neq \frac{\langle (g-m)^2 \rangle}{2\langle m \rangle \langle g \rangle}$$

On normalise toutes les données recueillies par leur valeur moyenne, de tel sorte à pouvoir estimer le bruit intrinsèque :

$$\frac{\langle (p_g/\langle p_g \rangle - p_m/\langle p_m \rangle)^2 \rangle}{2} = \frac{\langle (g/\langle g \rangle - m/\langle m \rangle)^2 \rangle}{2} = \frac{\langle (g-m)^2 \rangle}{2\langle m \rangle \langle g \rangle}$$

La dernière égalité n'est vraie que si $\langle m \rangle = \langle g \rangle$, ce qui est supposé être le cas dans les expériences « deux couleurs ». Cependant, il est important de remarquer que les valeurs moyennes $\langle m \rangle$ et $\langle g \rangle$ présentes dans l'équation précédente représente les quantités moyennes de fluorophores capables de fluorescer, c'est-à-dire les fluorophores dans un état fonctionnel et donc repliés. Ainsi, estimer les bruits intrinsèques et extrinsèques à partir des données de fluorescence n'est possible que si les temps de repliement des fluorophores sont identiques. Cette condition

est notamment supposée vrai dans les travaux d'Elowitz et al mais n'a jamais été vérifié. Nous considérons dans un premier temps que les temps de repliement des protéines mKate2 et GFPmut3 sont comparables, ce qui nous permet d'estimer le bruit intrinsèque du bruit extrinsèque, mais ceci devra être contrôlé (voire partie discussion). Des différences de temps de repliement biaisent les estimations des composantes intrinsèque et extrinsèque du bruit dans le sens d'une sous-évaluation de la composante extrinsèque.

V.3 Evolution du bruit intrinsèque et extrinsèque en fonction de la force des promoteurs :

A partir des données de fluorescence émanant des deux fluorophores, nous calculons le bruit extrinsèque et intrinsèque à partir des formules énoncées précédemment. Pour le bruit intrinsèque, nous traçons η_{int}^2 en fonction de $(1/\langle p \rangle)$ avec $\langle p \rangle$ le niveau d'expression moyen du fluorophore GFPmut3 (Figure 38-A). On observe alors une évolution linéaire du bruit intrinsèque avec l'inverse du niveau d'expression moyen, ce qui est conforme aux observations d'Elowitz et al. obtenues sur la bactérie *E. coli* en faisant varier le niveau d'expression du gène d'intérêt au moyen d'un promoteur (PLlac01) inductible à l'IPTG (Elowitz et al, 2002) ; ainsi qu'à celles faites par Newman et al et Bar-Even et al sur une partie du génome de la levure *S. cerevisiae* (Newman et al, 2006 ; Bar-Even et al, 2006). Cette dépendance est en accord avec le modèle à deux niveaux de l'expression génique qui prévoient une évolution du bruit d'expression génique en $1/\langle p \rangle$, où la constante de proportionnalité ne dépend que de l'efficacité de traduction, qui est inchangée à travers les souches exprimant les deux fluorophores. En revanche, le bruit extrinsèque, et donc le bruit total, ne suit pas cette loi. La décomposition du bruit total en bruit intrinsèque et extrinsèque est représentée sur la figure 38-B. On représente en noir le bruit d'expression total, en rouge le bruit extrinsèque et en bleu le bruit intrinsèque. La courbe bleue est la régression linéaire réalisée sur la figure 38-A. Cette décomposition montre que le bruit extrinsèque est toujours dominant, et l'est d'autant plus que le promoteur est fort, si bien que dès que le promoteur n'est pas trop faible, le bruit extrinsèque est très fortement dominant devant le bruit d'expression généré par les sources de bruit intrinsèque. Le bruit d'expression total décroît jusqu'à un certain seuil qui semble fixé par le niveau de bruit extrinsèque, tandis que le niveau de bruit intrinsèque diminue de manière monotone.

V.4 Evolution du facteur Fano intrinsèque et extrinsèque en fonction de la force des promoteurs :

A partir des données de fluorescence émanant des deux fluorophores, nous calculons maintenant les facteurs Fano extrinsèque, intrinsèque et total. Les résultats sont reportés sur la figure 38-B. Les marqueurs bleus représentent le

facteur Fano associé aux sources de bruits intrinsèques, les marqueurs rouges celui associé aux sources de bruit extrinsèque et les marqueurs noirs le facteur Fano total. On constate alors que les facteurs Fano extrinsèque et intrinsèque évoluent linéairement avec la force des promoteurs. Les régressions linéaires de la répartition des données extrinsèques et intrinsèques donnent des pentes ~ 0,011 et ~ 0,0014 respectivement. La pente obtenue pour le facteur Fano total est ~ 0,012. Ainsi, la pente obtenue pour la partie extrinsèque du facteur Fano est près de 10 fois supérieure à celle obtenue pour la partie intrinsèque, et cette pente est de ce fait similaire à la pente du facteur Fano total. Ainsi, l'augmentation inattendue du facteur Fano avec la force du promoteur semble en grande partie être due au comportement du bruit extrinsèque avec le taux de transcription. Si on reprend l'analyse des pentes des régressions linéaires des données obtenues avec les souches "mono-couleur" et le modèle développé par Taniguchi et al, quand on modifie uniquement la transcription, les pentes obtenues sont constantes (~0,008) et ne dépendent pas des conditions de traduction. D'après les données "deux-couleurs", la valeur obtenue correspond à un terme de bruit extrinsèque. Quand on modifie uniquement la traduction à l'aide des différents TIR de la collection, si le promoteur (ainsi que son TSS naturel) n'est pas trop faible, le terme de bruit extrinsèque est dominant dans la pente de la régression et la valeur de la pente est également autour de 0,008. Dans le cas où les conditions de transcription sont faibles, il faut ajouter à cette valeur de pente un terme inversement proportionnelle à la force du promoteur, ce qui conduit à une augmentation de la pente. En considérant la hiérarchisation des promoteurs suivante : ykwB<yufK<yqzM<zwf<ykpA<fbaA<rrnJP2<ylxM, on constate que la contribution des sources de bruit intrinsèque devient négligeable devant le bruit d'expression génique pour les souches possédant des promoteurs au moins aussi forts que le promoteur ykpA.

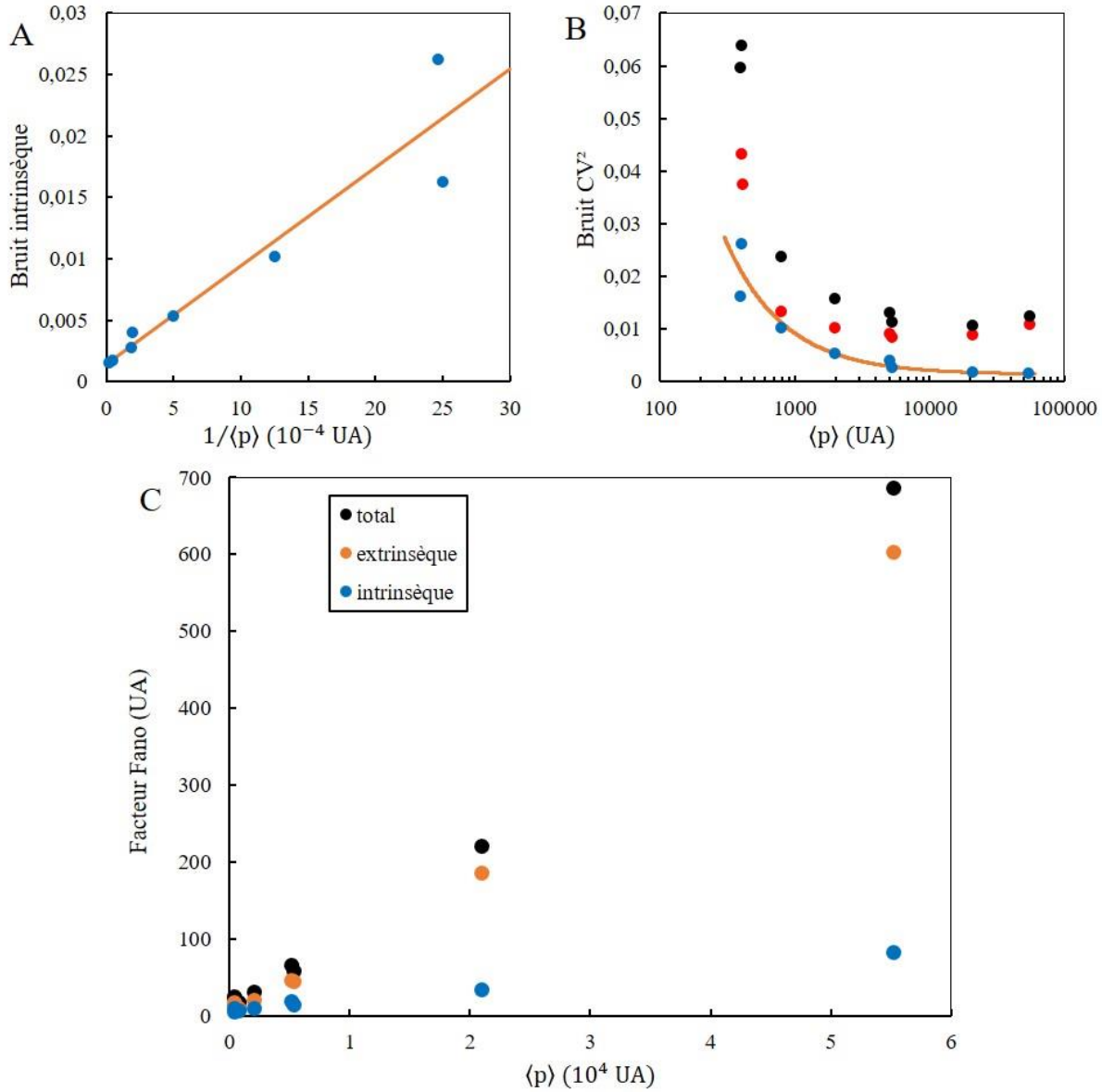


Figure 38 : Partitionnement du bruit total en bruit intrinsèque et extrinsèque. A) Evolution du bruit intrinsèque avec l'inverse du niveau moyen d'expression ($\langle p \rangle$). Une régression linéaire du type $y = ax + b$ donne une droite qui s'ajuste bien avec les données expérimentales. Ce résultat est en accord avec le modèle à deux niveaux de l'expression génique, puisqu'avec les souches utilisées, la traduction n'est pas modifiée. **B)** Bruit extrinsèque (rouge), intrinsèque (bleu) et total (noir) en fonction du niveau moyen d'expression. Le niveau moyen de fluorescence est modulé en utilisant les différents promoteurs (50 nucléotides en amont du TSS) de la banque. Ces résultats sont similaires à ceux obtenus par Elowitz et al (Elowitz et al, 2002). **C)** Facteur Fano extrinsèque (rouge), intrinsèque (bleu) et total (noir) en fonction du niveau moyen d'expression. Les graphes **B)** et **C)** montrent que les sources de bruit extrinsèques sont dominantes par rapport aux sources de bruit intrinsèque et expliquent les résultats obtenus à partir des souches « mono-couleurs ».

DISCUSSION

I Résultats de notre étude :

D'après nos résultats, chez la bactérie modèle Gram positive *B. subtilis* et dans la gamme d'expression visitée, la composante extrinsèque du bruit d'expression génique domine la composante intrinsèque. Cette forte composante extrinsèque du bruit fait que pour la plupart des souches possédant des promoteurs pas trop faibles et d'un point de vue du bruit d'expression génique, pour obtenir un niveau moyen d'expression, il est équivalent de modifier la traduction ou la transcription. Autrement dit, pour un niveau d'expression moyen, et dès que les promoteurs ne sont pas trop faibles, un gène avec une traduction forte et une transcription faible ne sera pas nécessairement plus bruité qu'un gène avec une transcription forte et une traduction faible (Figure 35). Pour des promoteurs très faibles cependant, une contribution du bruit intrinsèque supplémentaire fait qu'il vaut mieux dans ce cas augmenter la force du promoteur qu'augmenter la force du RBS si on veut augmenter le niveau moyen d'expression en minimisant le bruit.

Les résultats obtenus suggèrent de plus que les sources de bruit extrinsèque peuvent être une raison possible au "plateau" du bruit d'expression observé précédemment, ce qui serait en accord avec les conclusions de Taniguchi et al dans leur étude génomique sur *E. coli* (Taniguchi et al, 2010). Ces auteurs démontrent en effet expérimentalement que les sources de bruit extrinsèques deviennent majoritaires sur le bruit d'expression total dès que le niveau d'expression moyen dépasse ~10 protéines par cellule. Les souches utilisées dans notre étude dépassant toutes cette limite pour les fluorophores utilisés, les résultats obtenus chez *B. subtilis* semblent cohérents avec ce qui a été conclu pour la bactérie *E. coli* : pour la plus grande majorité des niveaux d'expression présents naturellement chez ces bactéries, le bruit d'expression est dominé par les sources de bruit extrinsèque. Il a également été montré que les sources de bruits extrinsèques sont à l'origine d'un "bruit plancher" chez la levure qui intervient cependant pour des niveaux d'expression moyens beaucoup plus élevé que chez *E. coli* et *B. subtilis*.

Sur la figure 38-B, on note que si le comportement du bruit intrinsèque est similaire à celui observé dans les expériences d'Elowitz et al, le bruit extrinsèque ne suit pas la même tendance. Dans notre cas, le bruit extrinsèque est une fonction décroissante du niveau d'expression moyen tandis que dans l'étude d'Elowitz, le bruit extrinsèque n'est pas une fonction monotone et atteint un maximum pour des niveaux d'expression intermédiaire. Cette différence provient probablement des différences de stratégie pour modifier le niveau d'expression moyen, et le

maximum obtenu avec un promoteur inductible reflète probablement l'influence des fluctuations de concentration du répresseur Lac.

II Bruit extrinsèque et validité du modèle à deux niveaux :

Nos résultats sur le bruit d'expression total sont en désaccord avec les prédictions du modèle stochastique à deux niveaux qui ne prend en compte que les sources de bruit intrinsèque (Thattai & van Oudenaarden, 2001). Ce modèle prévoit une évolution linéaire du facteur Fano avec l'abondance moyenne des protéines lorsque celle-ci est modifiée par modification du taux de traduction tandis qu'une modification du taux de transcription ne doit pas modifier le facteur Fano : $FF = 1 + b$. Nos données, en revanche, semblent en accord avec le modèle développé par Taniguchi et al qui prévoit une évolution du facteur Fano selon :

$$FF = 1 + (1 + \eta_b^2)\langle b\rangle + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\langle a\rangle\langle b\rangle$$

Ou

$$FF = 1 + (1 + \eta_b^2)(\langle p\rangle/\langle a\rangle) + (\eta_a^2 + \eta_a^2\eta_b^2 + \eta_b^2)\langle p\rangle$$

Où le dernier terme qui s'explique par la contribution de sources de bruits extrinsèque permet d'observer la dépendance linéaire du facteur Fano avec le niveau moyen d'expression lorsqu'est modifié le taux de transcription. Les données concernant le bruit d'expression total ne sont donc pas bien décrites par le modèle à deux niveaux qui ne prend en compte que les sources de bruits intrinsèques. Mais peut-on conclure quant à la validité du modèle à deux niveaux en ce qui concerne les données de bruit intrinsèque ? Pour rappel, les souches exprimant les deux fluorophores nous ont permis d'estimer le niveau de bruit intrinsèque et le facteur Fano associé et d'observer son évolution en fonction de la force des promoteurs utilisés dans ces souches. Nous observions alors que la pente obtenue après régression des données intrinsèques était près de 10 fois inférieure à la pente obtenue pour les données de bruit total (Figure 38-B). Cependant, cette évolution linéaire du facteur Fano intrinsèque avec la force des promoteurs peut être discutée et remise en question. En effet, la méthode "deux couleurs" suppose que (i) les deux rapporteurs ont des temps de maturation identiques, et repose (ii) sur l'interprétation de chaque terme de la loi de variance totale comme le bruit intrinsèque d'une part et le bruit extrinsèque d'autre part.

II.1 Temps de maturation des fluorophores :

Des temps de maturation différents entre les deux fluorophores entraînent une perte de corrélation entre les signaux de fluorescence qui n'est pas due aux sources de bruit intrinsèque. Ainsi, le degré de corrélation entre les niveaux

d'expression des deux fluorophores se retrouve diminué, et donc la contribution du bruit extrinsèque sur le bruit total est sous-évaluée. A l'inverse, la contribution des sources de bruit intrinsèque se retrouve surévaluée. Par exemple, le temps de repliement des fluorophores utilisés dans l'étude d'Elowitz et al (Elowitz et al, 2002), à savoir EYFP et ECFP, ont été mesuré *in vitro* par Iizuka et al (Iizuka et al, 2011). Ces auteurs obtiennent pour EYFP et ECFP respectivement ~16 min et ~10 min. Dans les conditions de leur expérience, le système de double rapporteur d'Elowitz et al surévalue donc a priori légèrement le bruit intrinsèque et sous-évalue le bruit extrinsèque. Nous devons contrôler ceci pour notre étude. D'après la littérature, le temps de maturation de la protéine GFPmut3 est estimé à environ 6,5 min chez *E. coli* (Megerle et al, 2008), tandis que le temps de repliement du fluorophore mKate2 a été estimé *in vitro* comme étant inférieur à 20 min (Shcherbo et al, 2009). Cependant le temps de repliement d'une protéine dépend fortement de l'organisme dans lequel se situe la molécule en question. Il faut donc comparer des temps de repliement obtenus dans des conditions similaires. Nous devons donc mesurer l'écart entre temps de repliement des fluorophores GFPmut3 et mKate2 dans la bactérie *B. subtilis*. Pour cela plusieurs méthodes peuvent être utilisées. Une première technique couramment utilisée consiste à mesurer le temps de repliement des deux protéines à l'échelle de la cellule unique (Gordon et al, 2007). La technique consiste alors à laisser croître les cellules exprimant nos deux fluorophores dans un milieu de culture, puis à exposer ces dernières à une forte concentration d'un antibiotique inhibiteur de la traduction comme le chloramphénicol. Une telle concentration permet de bloquer la production de protéine ainsi que la croissance. Ainsi, on peut négliger la dégradation/dilution des protéines, et l'augmentation de la fluorescence totale de chaque cellule au cours du temps est uniquement due à la maturation des protéines fluorescentes déjà traduite avant l'ajout de l'antibiotique. On peut montrer à partir d'un modèle cinétique du premier ordre que cette évolution de la fluorescence s'écrit :

$$M(t) = M_0 + I_0(1 - \exp(-t/\tau_{mat}))$$

Où $M(t)$ est la quantité de protéines fluorescentes ayant adopté leur forme fonctionnelle à l'instant t après ajout de l'antibiotique, M_0 et I_0 sont respectivement les quantités de protéines matures et immatures au moment de l'ajout de l'antibiotique, et τ_{mat} est le temps de maturation. Ainsi, en collectant les données de fluorescence de plusieurs cellules à différents instants par microscopie de fluorescence, on est capable pour chaque cellule de tracer une courbe de maturation. Un ajustement numérique des courbes par l'équation précédente permet de déterminer un temps de maturation pour chaque cellule, et donc un temps de maturation moyen pour un fluorophore donné dans l'organisme *B. subtilis*. Cette méthode nécessite d'introduire en cours d'expérience un antibiotique dans le milieu de culture. Ceci n'est pas possible avec notre dispositif instrumental qui est scellé et dans lequel le milieu est un substrat solide sur lequel

les cellules sont fixées. Il faudrait alors utiliser un dispositif de canaux microfluidiques permettant l'injection de l'antibiotique et le suivi de cellules individuelles. Pour remédier à ces difficultés expérimentales, une approche par cytométrie en flux peut également être envisagée. Dans ce cas, l'ajout d'antibiotique peur se faire aisément dans une culture liquide, puis un échantillon de cette culture est prélevé à différents instants pour pouvoir recueillir les données de fluorescence. Bien que ne permettant pas un suivi individuel, cette méthode permet d'observer l'évolution de la fluorescence moyenne. Par un ajustement adéquat des données, on peut remonter aux temps de repliement des fluorophores. L'inconvénient de cette méthode est le manque de sensibilité du cytomètre qui pourrait nous empêcher de détecter l'évolution de la fluorescence suite à la maturation des protéines. D'autre part, le cytomètre que nous utilisons ne disposant que d'un laser, nous ne pouvons pas quantifier le signal de fluorescence de la protéine mKate2 par cytométrie en flux. Une alternative à la cytométrie en flux est de mesurer la fluorescence émanant des deux fluorophores à l'échelle de la population "*en bulk*" grâce à un lecteur de microplaques. Enfin, une autre méthode consiste à utiliser comme "outil de quantification" un lecteur de plaques mais contrairement aux méthodes évoquées précédemment on ne cherche pas à quantifier les temps de repliement des deux fluorophores mais à quantifier leur différence. Pour ce faire, nous sommes en train de construire deux souches de *B. subtilis* exprimant chacune un fluorophore sous le contrôle d'un promoteur inductible. Dans notre cas le promoteur hyperspank (Phs). Après induction avec une forte concentration d'IPTG, l'évolution de la fluorescence des populations de chaque souche peut être suivie. Les deux évolutions devraient avoir des allures de sigmoïdes, et le décalage temporel entre les deux courbes devraient refléter l'écart entre temps de maturation des deux fluorophores.

II.2 Interprétation de la formule de la variance totale :

Le principe théorique sur lequel se repose la méthode des deux rapporteurs pour estimer les bruits intrinsèques et extrinsèques est la formule de la variance totale (Swain et al, 2002 ; Hilfinger & Paulsson, 2011). En utilisant les notations de l'article de Hilfiger & Paulsson, si X représente le nombre de protéines d'une cellule donnée à un instant donné et si $\boldsymbol{Z}$ représente l'ensemble des variables spécifiant l'état de la cellule (nombre d'ARN polymérase, ribosomes, volume de la cellule, états du cycle cellulaire…), alors il est théoriquement possible de décomposer la variance de X en une partie expliquée par $\boldsymbol{Z}$ et une autre partie qui n'est pas expliquée par Z :

$$\mathrm{Var}(X) = \underbrace{\mathrm{E}\big(\mathrm{Var}(X|Z)\big)}_{\text{Inexpliquée par Z}} + \underbrace{\mathrm{Var}\big(\mathrm{E}(X|Z)\big)}_{\text{Expliquée par Z}}$$

Pour rappel, $Var(X|\boldsymbol{Z})$ est la variance de X dans une sous-population de cellules réalisant la valeur $\boldsymbol{Z}$, l'espérance mathématique de cette quantité conduisant au premier terme est calculée sur l'ensemble de sous-populations caractérisées par des valeurs différentes de $\boldsymbol{Z}$. Le second terme représente la variance de l'espérance conditionnelle de X sachant $\boldsymbol{Z}$. Dans la méthode des deux rapporteurs et pour une cellule donnée, chaque niveau d'expression des deux gènes rapporteurs peut être interprété comme une réalisation de X dans un environnement commun spécifié par $\boldsymbol{Z}$. Les approches deux couleurs permettent d'estimer cette décomposition et interprète la partie inexpliquée par $\boldsymbol{Z}$ comme le bruit intrinsèque, et la partie expliquée par $\boldsymbol{Z}$ comme le bruit extrinsèque. Ces interprétations constituent les définitions des bruits intrinsèques et extrinsèques (Swain et al, 2002). Cependant, Hilfinger & Paulsson ont montré mathématiquement que cette décomposition basée sur l'état actuel de l'environnement ne permettait pas de séparer les contributions des sources de bruit intrinsèque et extrinsèque. Ils proposent alors une nouvelle version de la procédure de la décomposition qui ne prend pas en compte uniquement l'état actuel de l'environnement mais aussi son histoire. Ils montrent alors qu'avec ce type de partitionnement, la méthode des deux rapporteurs permet d'estimer de manière séparée les bruits intrinsèques et extrinsèques. Cependant, ce niveau de description des bruits intrinsèques et extrinsèques est phénotypique et ne tient pas compte des mécanismes moléculaires sous-jacents. Ces auteurs se demandent alors si les termes de bruit intrinsèque et extrinsèque peuvent être comparés à des modèles stochastiques de l'expression génique qui ne tiennent compte respectivement que des sources de bruit intrinsèque et extrinsèque. Hilfinger & Paulsson montrent mathématiquement que la partie du bruit d'expression génique imputable aux sources de bruit extrinsèque déduite de la décomposition suivant l'histoire de l'environnement cellulaire peut être analysé et interprété à l'aide de modèles de l'expression génique ne prenant en compte que les sources de bruits extrinsèques, c'est à dire avec notamment des taux de traduction et de transcription variant dans le temps, et ignorant les sources de bruit intrinsèque. A l'inverse, le bruit intrinsèque déduit de cette même décomposition ne correspond pas au bruit intrinsèque obtenu avec des modèles stochastiques de l'expression génique excluant totalement les sources de bruits extrinsèques, c'est à dire avec des taux de transcription et de traduction fixes par exemple. Ainsi, faire appel à la méthode des deux rapporteurs pour éliminer les influences inconnues ou complexes de l'environnement cellulaire sur le bruit d'expression génique et ne tenir compte alors que du bruit intrinsèque ne permet pas de conclure quant à la validité du modèle à deux niveaux.

La différence des temps de maturation des protéines rapportrices et l'interprétation faite de la formule de la variance totale ne gênent en rien les conclusions de notre étude quant à la dominance du bruit extrinsèque. Par contre, ces considérations empêchent une validation rigoureuse des modèles stochastique

de l'expression génique qui ne tiennent compte que des sources de bruits intrinsèques.

III Origine du bruit extrinsèque :

Nous avons vu que le bruit extrinsèque pouvait correctement être interprété à partir de la décomposition du bruit suivant l'histoire de l'environnement cellulaire comme la partie du bruit d'expression expliquée par l'environnement cellulaire et son histoire. De nombreux mécanismes, déterministes ou stochastiques peuvent être proposés comme sources de bruit extrinsèque : partition des constituants de la cellule lors de la division cellulaire, réplication des gènes, fluctuations des nombres d'ARN polymérases et de ribosomes. Cependant, l'importance relative de chacun de ces termes reste obscure. En effet, si de nombreuses études menées rendent compte des effets du bruit extrinsèque à la fois chez la levure *S. cerevisiae* et la bactérie *E. coli* (Bar-Even et al, 2006 ; Newman et al, 2006 ; Taniguchi et al, 2010), les mécanismes impliqués et leur contribution relative au niveau de bruit extrinsèque restent indéterminés. Néanmoins, deux études, une première menée chez *S. cerevisiae* (Keren et al, 2015) et une seconde chez *E. coli* (Yang et al, 2014) ont respectivement permis de rendre compte des effets du cycle cellulaire et des fluctuations de la concentration d'ARN polymérases sur le niveau de bruit extrinsèque. Keren et al ont réalisé une étude sur *S. cerevisiae* comparable à celle menée par Newman et al (Keren et al, 2015). Les auteurs observent l'évolution du bruit d'expression et du niveau d'expression moyen lorsqu'est modifié le taux de croissance des cellules et montrent expérimentalement que la variabilité phénotypique est une fonction de l'environnement, ce qui fait qu'en cas de faible croissance, une population de cellules isogéniques sera généralement plus variable que la même population dans un milieu plus riche. Autrement dit, pour un niveau d'expression moyen identique, le bruit d'expression génique sera plus important pour une population avec un taux de croissance faible que pour une population dans un milieu riche. Leurs résultats sont en accord avec un modèle de bruit extrinsèque qui considère une population de cellules asynchrones, dans laquelle les cellules sont à différentes étapes du cycle cellulaire. Chaque cellule peut être soit dans un état G1 (avant réplication du gène), soit dans un état G2 (après réplication du gène). A travers la dépendance du niveau bruit d'expression avec le taux de croissance de la population des cellules, les auteurs concluent que le cycle cellulaire a une contribution importante sur le niveau de bruit extrinsèque. Yang et al ont cherché de leur côté à quantifier la contribution des fluctuations du nombre d'ARN polymérases sur le niveau de bruit chez *E. coli* (Yang et al, 2014). Ces auteurs développent un système permettant de quantifier à la fois les variations de concentration en ARN polymérase et également les bruits extrinsèques et

intrinsèques d'une protéine en aval par une approche deux couleurs. Pour cela, des ARN polymérases T7 sont fusionnées avec le fluorophore YFP. Le niveau moyen d'expression et de bruit des polymérases est contrôlé au moyen d'un promoteur inductible. Ensuite, deux gènes exprimant deux fluorophores différents (CFP et mCherry) sont intégrés symétriquement dans le chromosome d'*E. coli* afin d'estimer les bruits intrinsèque et extrinsèque. Ces gènes sont contrôlés par un promoteur uniquement sensible aux ARN polymérases T7, ce qui permet d'observer la propagation du bruit en amont au niveau des ARN polymérases T7 sur les niveaux de bruit extrinsèque et intrinsèque en aval. D'après leur résultat, les variations dans la concentration d'ARN polymérase se propagent en effet vers le niveau de bruit des protéines en aval, mais n'affectent que le bruit extrinsèque, et ce dernier évolue linéairement lorsque varie le bruit d'expression des polymérases. De plus, quand le bruit au niveau des ARN polymérases est considérablement réduit, le bruit plancher dû au bruit extrinsèque est réduit de 30%. Yang et al concluent alors que le bruit dû aux fluctuations de la concentration d'ARN polymérases est une source majeure du bruit extrinsèque observé à haut niveau d'expression.

Dans notre étude, les données collectées par cytométrie en flux ont été traité de deux manières, la première méthode consistant à sélectionner les cellules selon une taille donnée, et la seconde en effectuant une normalisation par la taille. Cette dernière opération étant effectué sur l'ensemble de la population, on pourrait s'attendre à ce que les effets du cycle cellulaire se fassent ressentir. Etant donné que les résultats que nous avons obtenus sont similaires suivant les deux approches, nous pouvons en déduire que les effets du cycle cellulaire ne devraient pas être *a priori* une contribution majeure au bruit extrinsèque. De même, les données de microscopie de fluorescence ont été normalisées par la taille (ou surface) des cellules, ce qui d'après Taniguchi et al devrait réduire les effets du cycle cellulaire sur le bruit extrinsèque. Ainsi, on pourrait s'attendre à ce que le bruit extrinsèque soit dû aux fluctuations du nombre d'ARN polymérases et de ribosomes par exemple. Un premier contrôle à effectuer serait de quantifier le taux de croissance individuel de chaque cellule et d'observer les corrélations entre le niveau de fluorescence de chaque cellule et leur taux de croissance individuel. Une corrélation forte indiquerait que le bruit d'expression mesuré reflète des différences de taux de croissance entre cellules. De plus, il a été établi expérimentalement que les nombres de polymérases et de ribosomes augmentent fortement quand le taux de croissance augmente (Klumpp & Hwa, 2008). Ainsi, des cellules possédant des taux de croissances différentes contiennent différentes quantités d'ARN polymérases et de ribosomes ; si bien que si dans nos données, une corrélation est observée entre bruit d'expression et fluctuations du taux de croissance, ceci pourrait *a priori* indiquer un effet des variations du nombre de polymérase et/ou de ribosomes. Pour aller plus loin dans l'identification des mécanismes responsables du niveau de bruit extrinsèque, il serait également

intéressant de quantifier directement les fluctuations de polymérases et de ribosomes en fusionnant ces complexes enzymatiques avec des protéines fluorescentes (Yang et al, 2014), ou encore utiliser différents milieux de culture (Keren et al, 2015) et comparer l'évolution du niveau de bruit dans ces différents milieux selon le taux de croissance de la population.

IV Pertinence biologique de la variabilité phénotypique et conséquences des résultats obtenus :

Le modèle à deux niveaux prédit que pour un niveau d'expression donné, un gène fortement transcrit et faiblement traduit montrera moins de variabilité qu'un gène faiblement transcrit et fortement traduit. Les taux de transcription et de traduction sont des caractères codés dans le génome, qui dépendent notamment des séquences de la région promotrice et du RBS. Par conséquent, ils peuvent évoluer en fonction de pression sélectives, ce qui fait de la variabilité de l'expression génique un caractère évolutif. L'idée que le bruit d'expression génique est soumis à des forces évolutives est soutenue par le fait que des groupes de gènes de fonctions connexes présentent des niveaux différents de bruit d'expression. En particulier, il a été montré chez la levure *S. cerevisiae* que le bruit d'expression des gènes "essentiels" est particulièrement faible, contrairement aux gènes impliqués dans la réponse à des changements de milieu qui ont un bruit d'expression élevé (réponse à des situations de stress) (Fraser & Kaern, 2009 ; Newman et al, 2006 ; Bar-Even et al, 2006). Ces résultats suggèrent alors que le bruit peut être (i) nuisible dans certaines circonstances et la cellule doit trouver des stratégies pour le combattre, tandis qu'il peut être (ii) avantageux dans d'autres cas. Nous discutons des deux cas dans ce qui suit où nous confrontons les conclusions établies à la lumière du modèle à deux niveaux et ce que nos résultats impliquent sur ces conclusions.

Le correct fonctionnement d'une cellule nécessite un équilibre biochimique précis des quantités de nombreuses molécules données par la stœchiométrie des réactions impliquant ces molécules. Ainsi, le bruit d'expression des gènes codant pour ces molécules peut être dans ce cas nuisible pour la cellule. On peut donc s'attendre à ce que dans de nombreux cas, le caractère stochastique de l'expression génique soit soumis à des pressions de sélection conduisant à une minimisation du bruit. Cette question a été explorée chez la levure *S. cerevisiae* par Fraser et al (Fraser et al, 2004). Cette étude se base sur les prédictions du modèle à deux niveaux vérifiés expérimentalement par Ozbudak et al chez *B. subtilis* (Ozbudak et al, 2002). Les auteurs ont cherché à savoir si les gènes dont la variabilité d'expression devait être particulièrement délétère présentaient un niveau de bruit réduit. Ils ont estimé que pour deux groupes de gènes, les gènes

essentiels (dont la délétion est létale pour l'organisme) et ceux codant pour des protéines qui sont des sous-unités de complexe multiprotéique et dont la concentration est donc essentielle à l'assemblage correct du complexe, le bruit d'expression devrait être particulièrement nuisible. Afin de savoir si la sélection naturelle permet effectivement de minimiser les fluctuations aléatoires d'expression de ces gènes, ils ont estimé le bruit d'expression de presque tous les gènes de *S. cerevisiae*. Ils observent alors que pour un niveau d'expression donné, les gènes essentiels et ceux codant pour des sous-unités de complexes ont en moyenne un taux de transcription plus élevé et un taux de traduction plus faible, une stratégie qui d'après le modèle stochastique à deux niveaux conduit à un bruit d'expression génique plus faible. D'après les auteurs, ceci suggère que pour un grand nombre de gènes, le bruit d'expression est en effet nuisible et minimisé par la sélection naturelle. D'après nos résultats, nous avons vu que pour des promoteurs faibles, il était avantageux d'un point de vue du bruit d'expression d'avoir une traduction faible et une transcription forte, mais que pour des promoteurs plus forts, ceci n'est plus nécessairement vrai du fait du bruit extrinsèque. Ainsi, nos résultats suggèrent l'existence d'autres cibles sur lesquelles s'exercent une pression sélective. Notamment, les forces de traduction et de transcription évoluent selon le taux de croissance des cellules. Pour des promoteurs constitutifs, la dépendance du taux de transcription avec le taux de croissance se manifeste par une augmentation pour des taux de croissance faible, puis une saturation pour les forts taux de croissance (Klumpp et al, 2009). Pour les éléments TSS, suivant la nature de la première base (soit G soit A), la tendance précédente des promoteurs peut être modifiée. Ainsi, dans le génome de *B. subtilis*, la force de transcription peut avoir des comportements très différent selon le gène lorsque varie le taux de croissance. Par exemple, l'efficacité de transcription du gène fbaA est constante à travers différents taux de croissance tandis que l'efficacité de transcription du gène rrnJP2 augmente avec le taux de croissance (Nicolas et al, 2012). Pour ce qui est de la traduction, il a été montré qu'une augmentation du taux de croissance entraîne une chute du nombre de ribosomes libres, ce qui conduit à son tour à une décroissance de l'efficacité de traduction des modules TIR (Borkowski et al, 2016). Cependant, la sensibilité de la réponse à la variation du nombre de ribosomes disponibles n'est pas identique pour tous les TIR et dépend de la séquence de ces derniers. Ainsi, la combinaison de différents promoteurs, TSS et TIR conduit à des comportements différents quand le taux de croissance varie. Ainsi, au lieu d'être modulé par pression sélective au niveau du bruit d'expression, il se pourrait que les séquences contrôlant la transcription et la traduction de gènes essentiels ait été modelées afin d'assurer un niveau d'expression moyen permettant aux cellules d'assurer certaines fonctions essentielles à leur survie à travers différents milieux.

A côté de cette stratégie évolutive de minimisation du bruit pour les gènes essentiels, il a été montré expérimentalement que les gènes liés aux situations de

stress présentaient des niveaux de bruit d'expression particulièrement élevés (Bar-Even et al, 2006 ; Newman et al, 2006). Ces niveaux élevés de bruit peuvent être vus comme une stratégie de diversification des phénotypes appelée stratégie de "*bet-hedging*" (Blake et al, 2006). La stratégie de *bet-hedging* est une stratégie évolutive qui consiste à privilégier la réduction du risque à long terme au détriment de la "*fitness*" à court terme grâce à la diversification de phénotypes. En d'autres termes, une population peut être composée de sous-populations de phénotypes différents qui sont adaptées de façon optimale à des environnements différents. Par conséquent, dans n'importe quel environnement, une partie de la population n'est pas adaptée de façon optimale, mais si l'environnement fluctue, la population pourrait être plus apte à faire face aux changements. Il a été montré que le bruit d'expression génique était à l'origine de stratégie de *bet-hedging*, notamment chez *B. subtilis* avec les phénomènes de sporulation et de compétence (Figure 39). Ces deux phénomènes interviennent lorsqu'une population isogénique de *B. subtilis* entre en phase stationnaire dans laquelle la population subit une pénurie de nutriments. Une fraction des cellules de la population entre dans un processus de sporulation qui aboutit à la formation d'un type de cellule dormante appelée spore. Les spores sont plus résistants à de nombreux stress environnementaux que les cellules non sporulées. Cependant, ce processus est énergétiquement coûteux et si la limitation en nutriments est passagère, les cellules qui ne sont pas entrées dans le processus de sporulation peuvent reprendre leur croissance plus facilement (Dubnau & Losick, 2006). Une autre sous partie de la population va quant à elle développer la capacité d'absorber l'ADN présent dans l'environnement, un phénomène appelé compétence. Les cellules compétentes peuvent donc incorporer du nouveau matériel génétique, ce qui peut permettre l'acquisition de nouvelles fonctions, mais elles cessent temporairement de croître. Ainsi, une population mixte composée de spores, de cellules compétentes et de cellules n'ayant développé aucun des deux phénotypes précédents peut permettre à la population de faire face à l'incertitude quant à la disponibilité future des nutriments. Il a été montré que le processus de sporulation s'appuie sur l'expression bistable des gènes de sporulation générées par une boucle complexe d'autostimulation contrôlant l'activité du régulateur Spo0A (Veening et al, 2005 ; Veening et al, 2008). De même, il a été démontré expérimentalement que la décision de chaque cellule de devenir compétente est basée sur le bruit dans le niveau de la protéine régulatrice comK, qui intervient dans l'expression de tous les gènes de compétence et est sujette à une autorégulation positive (Maamar et al, 2007). Pour montrer que l'entrée dans la phase de compétence est basée sur le bruit d'expression de la protéine comK, Maamar et al se sont appuyés sur le modèle stochastique à deux niveaux : en augmentant le taux de transcription du gène codant pour la protéine comK tout en réduisant le taux de traduction, afin d'assurer un niveau d'expression moyen de la protéine constant, on devrait observer une réduction du bruit et une diminution du nombre de cellules compétentes. C'est en effet ce qui est observé par les auteurs,

à savoir une réduction du bruit conduisant à une diminution du nombre de cellules compétentes dans la population. Ainsi, le modèle stochastique à deux niveaux semble validé dans ces expériences. Le gène comK possède naturellement une transcription très faible, si bien que les observations effectuées par Maamar et al sont cohérentes avec nos résultats : pour des promoteurs très faibles, une augmentation de la force de transcription et une réduction de la force de traduction s'accompagne bien d'une légère réduction du bruit d'expression.

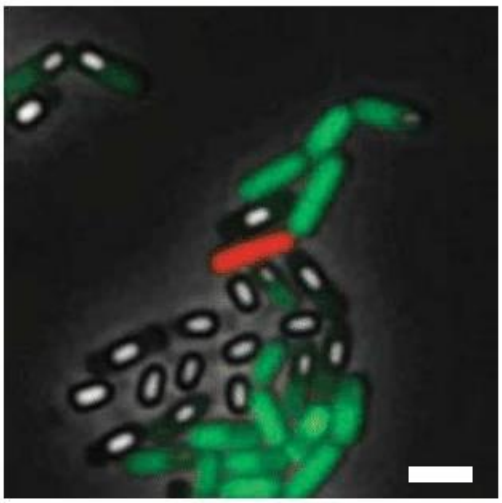


Figure 39 : Stratégie de « *bet-hedging* » chez *B. subtilis*. Quand des cellules d'un population isogénique de *B. subtilis* rencontre une limitation de nutriments, le bruit d'expression génique conduit ces cellules à des destins différents. Les cellules en vertes croissent végétativement, les objets blancs sont des spores résultant du processus de sporulation et la cellule rouge est quant à elle dans un état de compétence. D'après Losick & Desplan (2008). La barre d'échelle en blanc indique 2µm.

ANNEXE 1

- Modèle à deux niveaux, cas déterministe :

Pour mieux comprendre ce modèle, considérons en premier lieu qu'il n'y ait pas de fluctuations. On peut écrire l'évolution de ces quantités par des équations cinétiques chimiques du premier ordre en ayant au préalable défini des taux de production et de dégradation des ARN messagers et des protéines. Considérons la phase de transcription par exemple qui avec un certain niveau de détail suit un certain nombre d'étapes biochimiques. Les étapes les plus importantes sont l'attachement réversible de la polymérase (P) sur la région promotrice d'un gène (D) pour former un complexe « fermé » (PDc). La polymérase déroule ensuite

$$P + D \underset{k_-}{\overset{k_+}{\rightleftarrows}} PD_C \xrightarrow{k_{ouverture}} PD_o \xrightarrow{k_{echappement}} elongation$$

quelques bases de la double hélice pour permettre la lecture du brin matrice par son site catalytique, formant ainsi un complexe ouvert (PDo). Finalement, la polymérase quitte le promoteur et se déplace le long du gène pour synthétiser l'ARN complémentaire au brin matrice : c'est la phase d'élongation. Une fois le début de la phase d'élongation, le promoteur est libre pour accueillir une nouvelle ARN polymérase et initier le début de la transcription. On peut représenter ces trois réactions par le schéma cinétique suivant (Phillips, 2012), où $k_+, k_-, k_{ouverture}, k_{echappement}$ représentent respectivement les taux d'attachement, de détachement, d'ouverture de l'ADN, et d'échappement de la polymérase de la région promotrice. L'idée du modèle est de regrouper ces différentes étapes de la transcription en une seule avec un taux de production d'ARN messagers k_R. On procède de la même manière pour la dégradation des messagers en définissant un taux de dégradation γ_R, la traduction avec un taux de production de protéines k_P par ARN messagers, et la dégradation des protéines avec un taux de dégradation γ_P (Figure 1). Autrement dit si on note $r(t)$ et $p(t)$ respectivement les nombres d'ARN messagers et de protéines à un instant t, alors nous avons les équations :

$$\frac{\partial r}{\partial t} = k_R - r\gamma_R \quad (1.1)$$

$$\frac{\partial p}{\partial t} = rk_P - p\gamma_P \quad (1.2)$$

Les paramètres biophysiques k_R, k_P, γ_R et γ_P rendent compte d'une certaine réalité moléculaire. Par exemple, dans ce modèle, k_R et k_P déterminent respectivement les efficacités de transcription et de traduction. Plus k_R est important, plus on pourra initier de transcription. D'un point de vue moléculaire,

un k_R important pourrait correspondre à un promoteur fort, c'est-à-dire capable de recruter fréquemment les ARN polymérases, et donc d'initier de nombreuses transcriptions successivement ; tandis qu'un k_R faible serait plutôt adapté à un promoteur faible. De même, un k_P important correspond à une séquence RBS capable de recruter fréquemment un ribosome et donc d'initier de nombreuses traductions successives, tandis qu'un k_P faible correspond à un ARN messager faiblement traduit. Les quantités d'ARN polymérases, ribosomes ou autres protéines du dégradosome assurant les dégradations des ARN messagers et des protéines se retrouvent dans les taux de réactions. On a par exemple : $k_R = \widetilde{k_R}[ARNpol]$ et $k_P = \widetilde{k_P}[Ribosome]$.

Partant de conditions initiales nulles en nombre d'ARN messagers et protéines, la résolution du système d'équations couplées (1) et (2) donne les évolutions temporelles de $r(t)$ et $p(t)$:

$$r(t) = \frac{k_R}{\gamma_R}(1 - e^{-\gamma_R t})$$

$$p(t) = \frac{k_R k_P}{\gamma_R \gamma_P}(1 - e^{-\gamma_P t}) - \frac{k_R k_P}{\gamma_R(\gamma_R - \gamma_P)}(e^{-\gamma_R t} - e^{-\gamma_P t})$$

A l'état stationnaire, c'est-à-dire pour $t \to +\infty$, lorsque le nombre de molécules créées dans la cellule à chaque instant compense le nombre de molécules dégradées, les quantités d'ARN messagers et de protéines sont données par :

$$r^{st} = \frac{k_R}{\gamma_R};\ p^{st} = \frac{k_R k_P}{\gamma_R \gamma_P}$$

Cependant, comme le nombre d'ARN messagers est de l'ordre de dix, ce modèle déterministe ne peut représenter les réelles quantités de transcrits et de protéines au sein d'une cellule et ne peut expliquer les différences de niveau de protéines observées expérimentalement. Nous devons donc pour cela considérer un modèle stochastique pour rendre compte de la variabilité phénotypique observée.

ANNEXE 2

- Modèle à deux niveaux, Cas stochastique :

Au lieu d'être des quantités déterministes, $r(t)$ et $p(t)$ doivent être considérés comme des variables aléatoires. Le modèle stochastique adopte les mêmes caractéristiques que le modèle déterministe, c'est-à-dire que les taux de productions et de dégradations deviennent les probabilités de transitions par unité de temps entre différents états. Ici un état est défini par le doublet $\{r, p\}$. La probabilité de trouver le système dans cet état est spécifiée par la distribution jointe de probabilité $P_{r,p}(t)$, qui évolue suivant le schéma suivant (Thattai & van Oudenaarden, 2001) :

Transcription : $P_{r,p} \xrightarrow{k_R} P_{r+1,p}$

Dégradation ARNm : $P_{r,p} \xrightarrow{r\gamma_R} P_{r-1,p}$

Translation : $P_{r,p} \xrightarrow{rk_P} P_{r,p+1}$

Protéolyse : $P_{r,p} \xrightarrow{p\gamma_P} P_{r,p-1}$

Chacune de ces étapes individuelles est un processus stochastique que nous allons détailler dans la prochaine partie.

i) Transcription :

Considérons uniquement la première réaction de transcription. D'après le modèle, la probabilité que le nombre d'ARN messagers augmente de r à $r+1$ sur un intervalle de temps Δt est égale à $k_R \Delta t$. On appelle r_t le nombre d'évènement « synthèse d'un ARN messager » entre l'instant initial $t = 0$ et l'instant t. Cette variable aléatoire caractérise le processus de transcription. La quantité $-r_t$ représente le nombre d'évènements entre t et $t + t'$ aussi appelé accroissement du processus. On note t_i l'instant du $i^{ème}$ évènement qui est également une variable aléatoire. Le temps séparant le $(i-1)^{ème}$ évènement du $i^{ème}$ évènement est notée T_i. Ce temps représente donc le temps de séjour dans un état (caractérisée par le nombre de messagers) et est également une variable aléatoire. On a notamment la relation : $t_i = \sum_{j=1}^{i} T_j$. D'après le modèle adopté, le processus caractérisé par la donnée r_t est un processus de comptage, c'est-à-dire un processus croissant : pour deux instants s et t tel que $s \leq t$, on a $r_s \leq r_t$. De plus, le processus de transcription tel qu'il est défini ici est un processus à accroissement indépendant, c'est-à-dire que pour tous les instants $t_1, t_2, \dots, t_i$ tels

que $t_1 < t_2 < \cdots < t_i$, les accroissements $r_{t_1} - r_{t_0}, r_{t_2} - r_{t_1}, \ldots, r_{t_i} - r_{t_{i-1}}$ sont des variables aléatoires indépendantes. Ainsi, le nombre de messagers à un instant t_i , r_{t_i}, dépend du nombre de messagers à l'instant t_{i-1} , $r_{t_{i-1}}$, mais ne dépend pas des instants ultérieurs. Ce processus est donc un processus de Markov, ou processus sans mémoire. Enfin, ce processus est stationnaire puisque les accroissements suivent tous la même logique. Si on considère différentes réalisations $r_{i,t}$ du processus (Figure 2-1), chaque réalisation définissant une trajectoire possible, on peut calculer la distribution de probabilité du nombre d'ARN messagers à un instant donné.

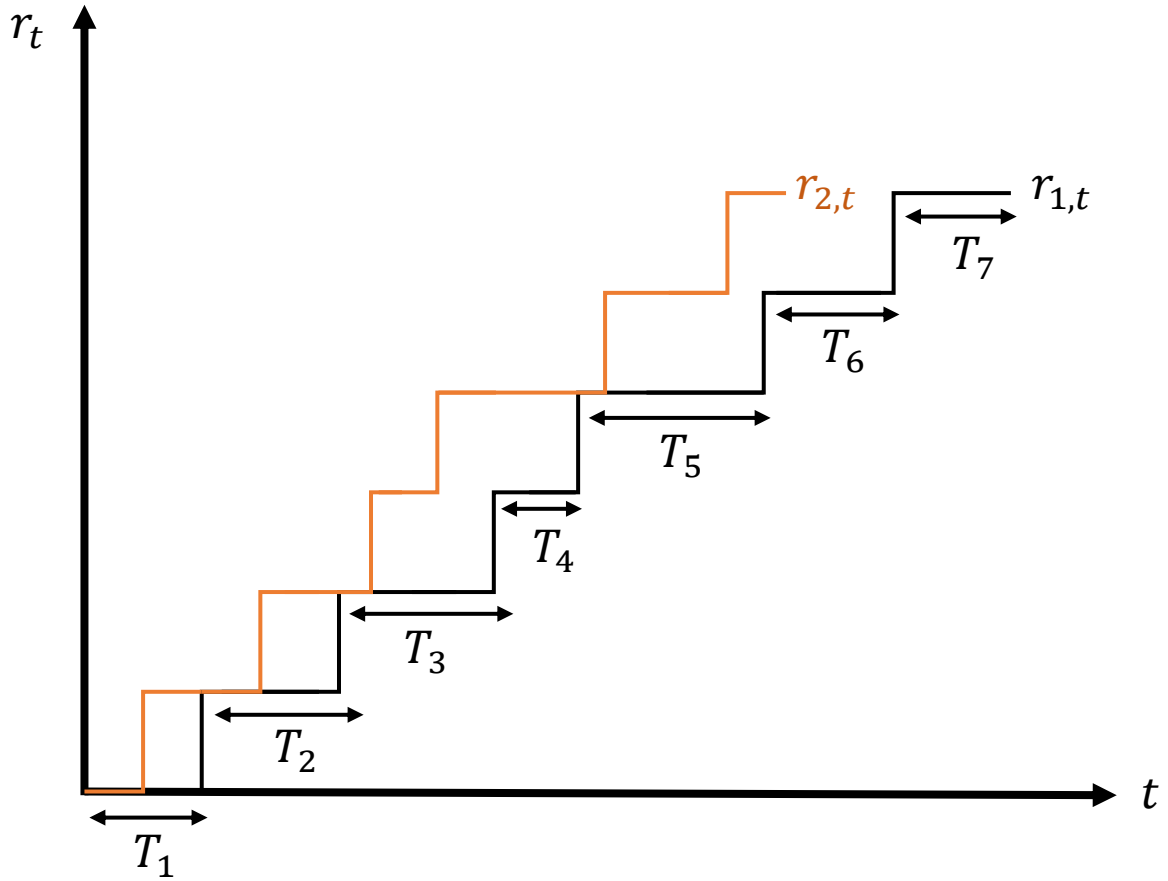


Figure 2-1 : Deux réalisations $r_{1,t}$ et $r_{2,t}$ possibles du processus stochastique de comptage représentant la synthèse des ARN messagers. Le temps séparant le $(i-1)^{ème}$ évènement du $i^{ème}$ évènement est notée T_i. Si on considère plusieurs trajectoires, on peut alors construire les distributions de probabilités à chaque instant du nombre d'ARN messagers. On peut également construire les distributions de temps de séjours dans chaque état sur une même trajectoire.

La probabilité que le nombre d'ARN messagers augmente de r à $r+1$ dans un intervalle de temps Δt est égale à $k_R \Delta t$:

$$P(r \rightarrow r+1 \; pendant \; \Delta t) = \; k_R \Delta t$$

Ainsi, la probabilité d'avoir m messagers à l'instant $t + \Delta t$ s'écrit :

$$P(r, t + \Delta t) = k_R \Delta t P(r-1, t) + (1 - k_R \Delta t) P(r, t)$$

Où le premier terme représente la probabilité d'augmenter d'une unité le nombre d'ARN messagers et qu'il y en avait $r-1$ à l'instant t. Le deuxième terme représente la probabilité que le nombre de messagers n'évoluent pas durant Δt et qu'il y en avait déjà r à l'instant t. On en déduit :

$$\frac{P(r, t + \Delta t) - P(r, t)}{\Delta t} = k_R P(r-1, t) - k_R P(r, t)$$

Pour des temps $\Delta t \to 0$, l'équation précédente s'écrit :

$$\frac{dP(r, t)}{dt} = k_R [P(r-1, t) - P(r, t)] \quad (*)$$

Pour résoudre cette équation, on part de conditions initiales nulles :

$$P(r, 0) = \delta_{r,0} \text{ et } P(-1, t) = 0.$$

On introduit la fonction génératrice : $G(z, t) = \sum_{r=0}^{+\infty} z^r P(r, t)$.

En multipliant l'équation (*) par z^r, et en sommant sur m, on obtient :

$$\frac{dG(z, t)}{dt} = k_R \sum_{r=0}^{+\infty} z^r P(r-1, t) - k_R \sum_{r=0}^{+\infty} z^r P(r, t)$$

En remarquant que :

$$\sum_{r=0}^{+\infty} z^r P(r-1, t)$$

$$= \sum_{n=-1}^{+\infty} z^{n+1} P(n, t) = \sum_{n=-1}^{+\infty} z^{n+1} P(n, t) = z \sum_{r=0}^{+\infty} z^r P(r, t) = zG$$

L'équation (4) devient :

$$\frac{dG(z, t)}{dt} = k_R (z-1) G$$

Soit : $$G(z, t) = G(z, 0) \exp(k_R (z-1) t)$$

D'après la condition initiale : $P(r, 0) = \delta_{r,0}$, on en déduit :

$$G(z, 0) = \sum_{r=0}^{+\infty} z^r P(r, 0) = \sum_{r=0}^{+\infty} z^r \delta_{r,0} = 1, \text{ d'où :}$$

$$G(z, t) = \exp(k_R z t) \exp(-k_R t)$$

Ce qui s'écrit aussi, en utilisant le développement en série entière de la fonction exponentielle :

$$G(z,t) = \exp(k_R zt)\exp(-k_R t) = \exp(-k_R t)\sum_{r=0}^{+\infty}\frac{1}{r!}(k_R t)^r z^r$$

En identifiant avec la définition de la fonction génératrice, on obtient :

$$P(r,t) = \frac{(k_R t)^r}{r!}\exp(-k_R t)$$

C'est-à-dire une loi de Poisson de moyenne $k_R t$ et de variance égale à la moyenne, et donc un facteur Fano égal à un ;1, ce qui correspond à une variabilité importante. Ainsi, d'après ce modèle, la transcription correspond à un processus de Poisson. On peut également calculer la distribution de densité de probabilité des temps de séjour dans chaque état. Pour cela, on considère de nouveaux différentes réalisations du processus de transcription. Comme nous avons un processus stationnaire, nous pouvons nous intéresser à la distribution des instants séparant l'instant initial t_0 où il n'y a pas encore d'ARN messagers et l'instant t_1 où on a la synthèse du premier messager, soit la distribution des instants T_1. Pour calculer cette densité de probabilité, on s'intéresse donc à l'évènement : « Synthèse du premier messager entre les instants T_1 et $T_1 + \Delta t$ » dont la probabilité s'écrit : $P(T_1)\Delta t$. On discrétise le temps T_1 en N intervalles : $T_1 = N\Delta t$. Ainsi, la probabilité de l'évènement considéré s'écrit :

$$P(T_1)\Delta t = (1 - k_R\Delta t)^N k_R\Delta t = (1 - k_R\frac{T_1}{N})^N k_R\Delta t$$

D'autre part, $(1 - k_R\frac{T_1}{N})^N = \exp[\ln\left(1 - k_R\frac{T_1}{N}\right)^N] = \exp[Nln\left(1 - k_R\frac{T_1}{N}\right)]$

Dans la limite où $N \to +\infty$, c'est-à-dire dans limite des temps continus :

$$ln\left(1 - k_R\frac{T_1}{N}\right) \approx -k_R\frac{T_1}{N}$$

Si bien que : $\exp\left[Nln\left(1 - k_R\frac{T_1}{N}\right)\right] = \exp(-k_R T_1)$, ce qui nous conduit à la loi de probabilité :

$$P(T_1)dt = k_R\exp(-k_R T_1)dt$$

On en déduit donc une distribution exponentielle des temps de séjour (ou temps d'attente entre deux évènements). Le temps de séjour moyen τ est donné par :

$$\tau = \frac{1}{k_R}$$

ii) Dégradation des ARN messagers :

Dès les premiers ARN messagers synthétisés commence une compétition entre les ribosomes et les RNases pour soit initier la traduction, soit la dégradation des ribosomes. D'après notre modèle, La probabilité que le nombre d'ARN messagers décroît de r à $r-1$ dans un intervalle de temps Δt est égale à $r\gamma_R \Delta t$:

$$P(r \to r-1 \; pendant \; \Delta t) = \; r\gamma_R \Delta t$$

Ce processus stochastique est également markovien et stationnaire comme le processus de transcription et est caractérisé par le nombre d'ARN messager r_t à un instant t. Etablissons la loi de probabilité du nombre d'ARN messager à un instant t, sans prendre en compte la phase de transcription et sachant que l'on a M ARN messagers à l'instant initial. Le calcul complet sera présenté dans la prochaine partie. Considérons tout d'abord une molécule d'ARN messager. Soit $P_1(t+\Delta t)$ la probabilité de survie de cet ARN messager au temps $t+\Delta t$. On a :

$$P_1(t+\Delta t) = \; (1-\gamma_R \Delta t) P_1(t)$$

Ce qui donne :

$$\frac{dP_1}{dt} = \; \gamma_R \Delta t P_1$$

Soit :

$$P_1(t) = \exp(\gamma_R t)$$

On en déduit directement la probabilité $P_0(t)$ que le messager soit dégradé au bout d'un temps t s'écrit :

$$P_0(t) = 1 - P_1(t) = 1 - \exp(\gamma_R t)$$

Ainsi, dans une population de M ARN messagers, chaque messager a une probabilité $1-\exp(\gamma_R t)$ d'être dégradé au bout d'un temps t et une probabilité $\exp(\gamma_R t)$ de survivre au bout du temps t. La probabilité d'avoir r messagers à l'instant t, $P(r,t)$, revient donc à considérer la probabilité d'avoir $M-r$ molécules dégradées et r molécules qui ont survies. Cette quantité doit être pondérée par le nombre de façons d'obtenir les r ARN messagers non dégradés parmi les M messagers de la population initiale. Soit :

$$P(r,t) = \binom{M}{r} \exp(-r\gamma_R t)(1-\exp(-\gamma_R t))^{M-r}$$

Comme pour le cas de la transcription, on peut montrer que le temps de séjour dans un état donné, ou le temps d'attente entre deux évènements de dégradation est distribué exponentiellement : si on note T le temps d'attente du premier évènement, on a :

$$P(T)dt = r\gamma_R \exp(-r\gamma_R T)\, dt$$

Si on considère une molécule d'ARN messager, le temps de séjour ou temps d'attente représente également son temps de vie. On peut montrer que les temps de vie d'un ARN messager sont distribués exponentiellement et le temps de vie moyen τ_{ARNm} d'un ARN messager est donné par :

$$\tau_{ARNm} = \langle T \rangle = \frac{1}{\gamma_R}$$

iii) Traduction et protéolyse :

Pris individuellement, c'est-à-dire sans prendre en compte les étapes de synthèse et de dégradation des ARN messagers, La traduction suit la même logique que la transcription, à savoir que, partant de d'une seule molécule d'ARN messager, la probabilité de synthétiser une protéine et donc d'augmenter le nombre de protéines d'une unité (on passe de p à $p+1$) en un temps Δt est donnée par $k_P \Delta t$:

$$P(p \to p+1\ pendant\ \Delta t) = k_P \Delta t$$

Dans le cas où on a r ARN messagers dans le cytoplasme, cette probabilité devient :

$$P(p \to p+1\ pendant\ \Delta t) = r k_P \Delta t$$

Ainsi, ce processus est un processus de comptage, stationnaire et markovien, et le nombre de protéines p synthétisées par un ARN messager durant son temps de vie t suit un processus de Poisson :

$$P(p,t) = \frac{(k_P t)^p}{p!} e^{-k_P t}$$

Dans le cas où on a r ARN messagers, cette loi se réécrit :

$$P(p,t) = \frac{(rt)^p}{p!} e^{-r k_P t}$$

Les temps d'attente T entre deux évènements sont quant à eux distribués selon une loi exponentielle de paramètre $r k_P$.

La réaction de dégradation des protéines est identique à la réaction de dégradation des messagers et les temps de vie T d'une protéine sont distribués exponentiellement :

$$P(T)dt = p\gamma_P \exp(-p\gamma_P T)\, dt$$

Et le temps de vie moyen d'une protéine est donnée par : $\tau_{Proteine} = \langle T \rangle = \frac{1}{\gamma_P}$

iv) Calcul du bruit dans le nombre d'ARNm et de protéines :

Les variables stochastiques à prendre en compte dans ce modèle stochastique à deux niveaux sont le nombre d'ARN messagers $r(t)$ et de protéines $p(t)$. Un état du système est donc spécifié par la donnée du doublon $\{r(t), p(t)\}$. La probabilité d'observer le système dans cet état s'écrit : $P(r, p, t)$. L'équation qui régit l'évolution de cette probabilité se nomme équation maîtresse et se détermine facilement à partir des probabilités de transitions par unité de temps (Figure 4) :

$$\begin{aligned}\partial_t P(r,p,t) = {} & k_R P(r-1,p,t) + \gamma_R (r+1) P(r+1,p,t) + k_P r P(r,p-1,t) \\ & + \gamma_P (p+1) P(r,p+1,t) - (k_R + \gamma_R r + k_P r + \gamma_P p) P(r,p,t)\end{aligned}$$

On introduit la fonction génératrice :

$$G(z,z',t) = \sum_{r,p=0}^{+\infty} z^r z'^p \, P(r,p,t)$$

Ainsi, en multipliant l'équation maîtresse par $z^r z'^p$ et en sommant sur r et p, on obtient :

$$\begin{aligned}\partial_t G(z,z',t) = {} & k_R z G + \gamma_R \partial_z G + k_P z z' \partial_z G + \gamma_P \partial_{z'} G - k_R G - \gamma_R z \partial_z G \\ & - k_P z \partial_z G - \gamma_P z' \partial_{z'} G \quad (2.1)\end{aligned}$$

La fonction génératrice permet de calculer simplement le moment d'ordre n des variables aléatoires $r(t)$ et $p(t)$ définies par les espérances mathématiques de r^n et p^n, n étant un entier naturel : $\mu_{r,n} = \langle r^n \rangle$ et $\mu_{p,n} = \langle p^n \rangle$. La donnée de tous les moments (n allant de 1 à $+\infty$) caractérise presque totalement la distribution de la variable aléatoire considérée. Nous nous limitons ici aux moments d'ordre un et

deux à partir desquelles nous pouvons calculer le coefficient de variation et le facteur Fano. On a ainsi, compte tenu de la définition de la fonction génératrice :

$$\partial_z G(z,z',t)|_{z=z'=1} = \langle r \rangle$$

$$\partial_{z'} G(z,z',t)|_{z=z'=1} = \langle p \rangle$$

$$\partial_z \partial_{z'} G(z,z',t)|_{z=z'=1} = \langle pr \rangle$$

$$\partial_z \partial_z G(z,z',t)|_{z=z'=1} = \langle r^2 \rangle - \langle r \rangle$$

$$\partial_{z'} \partial_{z'} G(z,z',t)|_{z=z'=1} = \langle p^2 \rangle - \langle p \rangle$$

Plaçons-nous à l'état stationnaire, c'est-à-dire que la distribution jointe, et donc la fonction génératrice, ne dépend plus explicitement du temps : $\partial_t G(z,z',t) = 0$. On évalue alors l'équation (2.1) en $z' = 1$:

$$0 = (k_R z G + \gamma_R \partial_z G - k_R G - \gamma_R z \partial_z G)|_{z'=1} \quad (2.2)$$

$$\Longrightarrow 0 = (1-z)(-k_R G + \gamma_R \partial_z G)|_{z'=1}$$

Cette dernière relation devant être valable quel que soit z, on en déduit :

$$(-k_R G + \gamma_R \partial_z G)|_{z'=1} = 0 \quad (2.3)$$

Pour $z = 1$, on en déduit, sachant que $G(1,1) = \sum_{r,p=0}^{+\infty} P(r,p) = 1$:

$$-k_R G + \gamma_R \langle r \rangle = 0 \Longrightarrow \langle r \rangle = \frac{k_R}{\gamma_R}$$

Pour le moment d'ordre deux du nombre d'ARN messager, on dérive l'équation (2.3) par rapport à z et on prend par la suite $z = 1$:

$$-k_R \langle r \rangle + \gamma_R (\langle r^2 \rangle - \langle r \rangle) = 0 \Longrightarrow \langle r^2 \rangle = \left(\frac{k_R}{\gamma_R} + 1\right) \langle r \rangle = \langle r \rangle^2 + \langle r \rangle$$

On en déduit donc la variance :

$$\sigma_R^2 = \langle r^2 \rangle - \langle r \rangle^2 = \langle r \rangle$$

On en déduit donc un coefficient de variation et un facteur Fano :

$$\eta_R = \frac{\sigma_R}{\langle r \rangle} = \frac{1}{\sqrt{\langle r \rangle}}$$

$$\frac{\sigma_R^2}{\langle r \rangle} = 1$$

Ainsi, le modèle à deux niveaux donne un facteur Fano égale à un en ce qui concerne le nombre d'ARN messager ; ce qui signifie qu'une fois l'état

stationnaire atteint, les niveaux de fluctuations sont du même niveau qu'une variable aléatoire ayant pour distribution de probabilité une loi de Poisson et ayant la même moyenne. La loi de probabilité du nombre d'ARN messager sera déterminée dans une prochaine section. Regardons maintenant les niveaux de fluctuations du nombre de protéines. Pour cela, on évalue cette fois l'équation (2.1) en $z = 1$:

$$0 = k_P z' \partial_z G + \gamma_P \partial_{z'} G - k_P \partial_z G - \gamma_P z' \partial_{z'} G$$

$$\Longrightarrow 0 = (1 - z')(\gamma_P \partial_{z'} G - k_P \partial_z G)|_{z=1}$$

Cette relation devant être valable quel que soit z', on en déduit :

$$(\gamma_P \partial_{z'} G - k_P \partial_z G)|_{z=1} = 0 \quad (2.4)$$

On en déduit le nombre moyen de protéines à l'état stationnaire avec $z' = 1$:

$$\gamma_P \langle p \rangle - k_P \langle r \rangle = 0 \Longrightarrow \langle p \rangle = \frac{k_R k_P}{\gamma_R \gamma_P}$$

Pour le moment d'ordre un, on dérive l'équation (2.4) par rapport à z' et on prend par la suite $z' = 1$:

$$\gamma_P(\langle p^2 \rangle - \langle p \rangle) - k_P \langle rp \rangle = 0 \Longrightarrow \langle p^2 \rangle = \frac{k_P}{\gamma_P} \langle rp \rangle + \langle p \rangle \quad (2.5)$$

Pour calculer le moment d'ordre un de la variable stochastique rp, on peut dériver l'équation (2.1) prise dans l'état stationnaire par la variable z:

$$\begin{aligned} 0 = {} & k_R G + k_R z \partial_z G + \gamma_R \partial_z \partial_z G + k_P z' \partial_z G + k_P z z' \partial_z \partial_z G + \gamma_P \partial_z \partial_{z'} G \\ & - k_R \partial_z G - \gamma_R \partial_z G - \gamma_R z \partial_z \partial_z G - k_P \partial_z G - k_P z \partial_z \partial_z G \\ & - \gamma_P z' \partial_z \partial_{z'} G \end{aligned}$$

Puis $z = 1$:

$$\begin{aligned} 0 = {} & k_R G + k_P z' \partial_z G + k_P z' \partial_z \partial_z G + \gamma_P \partial_z \partial_{z'} G - \gamma_R \partial_z G - k_P \partial_z G - k_P \partial_z \partial_z G \\ & - \gamma_P z' \partial_z \partial_{z'} G|_{z=1} \end{aligned}$$

On dérive maintenant par rapport à z' :

$$\begin{aligned} 0 = {} & k_R \partial_{z'} G + k_P \partial_z G + k_P z' \partial_{z'} \partial_z G + k_P \partial_z \partial_z G + k_P z' \partial_{z'} \partial_z \partial_z G \\ & + \gamma_P \partial_{z'} \partial_z \partial_{z'} G - \gamma_R \partial_{z'} \partial_z G - k_P \partial_{z'} \partial_z G - k_P \partial_{z'} \partial_z \partial_z G \\ & - \gamma_P \partial_z \partial_{z'} G - \gamma_P z' \partial_{z'} \partial_z \partial_{z'} G|_{z=1} \end{aligned}$$

Puis $z' = 1$:

$$0 = k_R \partial_{z'} G + k_P \partial_z G + k_P \partial_z \partial_z G - \gamma_R \partial_{z'} \partial_z G - \gamma_P \partial_z \partial_{z'} G|_{z=z'=1}$$

Ce qui donne au final, avec $\langle r^2 \rangle - \langle r \rangle = \langle r \rangle^2$:

$$0 = k_R\langle p \rangle + k_P\langle r \rangle + k_P(\langle r^2 \rangle - \langle r \rangle) - \gamma_R\langle pr \rangle - \gamma_P\langle pr \rangle$$

$$\Longrightarrow \langle pr \rangle = \frac{k_R\langle p \rangle + k_P\langle r \rangle(\langle r \rangle + 1)}{\gamma_R + \gamma_P} \quad (2.6)$$

Nous pouvons revenir sur le calcul du moment d'ordre deux du nombre de protéines en combinant (2.5) et (2.6) :

$$\langle p^2 \rangle = \frac{k_P}{\gamma_P}\left(\frac{k_R\langle p \rangle + k_P\langle r \rangle(\langle r \rangle + 1)}{\gamma_R + \gamma_P}\right) + \langle p \rangle$$

Soit :

$$\langle p^2 \rangle = \frac{k_P}{(\gamma_R + \gamma_P)}\left(\frac{k_R\langle p \rangle + k_P\langle r \rangle(\langle r \rangle + 1)}{\gamma_P}\right) + \langle p \rangle$$

$$\Longrightarrow \langle p^2 \rangle = \frac{k_P}{(\gamma_R + \gamma_P)}\left(\frac{k_R}{\gamma_P}\langle p \rangle + \frac{k_R}{\gamma_R}\langle p \rangle + \langle p \rangle\right) + \langle p \rangle$$

On en déduit la variance :

$$\sigma_p^2 = \langle p^2 \rangle - \langle p \rangle^2 = \frac{k_P}{(\gamma_R + \gamma_P)}\left(\frac{k_R}{\gamma_P}\langle p \rangle + \frac{k_R}{\gamma_R}\langle p \rangle + \langle p \rangle\right) + \langle p \rangle - \langle p \rangle^2$$

Le facteur Fano s'écrit donc :

$$\frac{\sigma_p^2}{\langle p \rangle} = \frac{k_P}{(\gamma_R + \gamma_P)}\left(\frac{k_R}{\gamma_P} + \frac{k_R}{\gamma_R} + 1\right) + 1 - \frac{k_R k_P}{\gamma_R \gamma_P}$$

$$\Longrightarrow \frac{\sigma_p^2}{\langle p \rangle} = 1 + \frac{k_P}{(\gamma_R + \gamma_P)} = 1 + \frac{k_P/\gamma_R}{(1 + \gamma_P/\gamma_R)}$$

L'approximation :

$$\tau_{proteines} \gg \tau_{ARNm} \Longrightarrow \gamma_R \gg \gamma_P$$

Nous conduits à :

$$\frac{\sigma_p^2}{\langle p \rangle} \cong 1 + \frac{k_P}{\gamma_R} \quad (2.7)$$

Le coefficient de variation se déduit facilement du facteur Fano (sans l'approximation sur les temps de vie) :

$$\eta_p^2 = \frac{1}{\langle p \rangle} + \frac{\gamma_P}{k_R}\frac{\gamma_R}{(\gamma_R + \gamma_P)} = \frac{1}{\langle p \rangle} + \frac{1}{\langle r \rangle}\frac{\gamma_P}{\gamma_R + \gamma_P} \quad (2.8)$$

Avec l'approximation sur les temps de vie, on obtient :

$$\eta_p^2 \cong \frac{1}{\langle p \rangle} + \frac{\gamma_P}{k_R} \quad (2.9)$$

Ainsi, pour résumer ce qui se passe au niveau protéique, le niveau d'expression moyen d'un gène (représenté par le nombre moyen de protéines) s'exprime à partir des probabilités de transitions entre les différents états selon la relation :

$$\langle p \rangle = \frac{k_R k_P}{\gamma_R \gamma_P}$$

Ce résultat indique notamment que le nombre de protéines peut être modulable suivant les efficacités de transcription et de traduction et les temps de vie des ARN messagers et des protéines. Si on ne tient pas compte des dégradations, alors pour obtenir un niveau d'expression $\langle p \rangle$ donné, nous pouvons jouer sur deux « leviers » : la transcription et la traduction. On peut en effet avoir pour un nombre moyen de protéines donnée, un gène fortement transcrit et faiblement traduit, ou à l'inverse un gène faiblement transcrit mais fortement traduit. Regardons maintenant ce qu'il en est du bruit d'expression à travers le coefficient de variation (ou bruit) ainsi que le facteur Fano (ou force du bruit). Si nous avons vu que les fluctuations du nombre d'ARN messager étaient comparable à celles que l'on aurait obtenue avec une distribution de Poisson ayant la même moyenne, le coefficient de variation ainsi que le facteur Fano du nombre de protéines conduisent à des conclusions différentes.

v) <u>Distribution du nombre d'ARN messager :</u>

On s'intéresse ici uniquement aux processus de synthèse/dégradation des ARN messagers. Un état du système est caractérisé par la donnée du nombre de messagers à un instant donné $r(t)$. On note $P(r,t)$ la probabilité d'observer la valeur r à un instant t. Nous allons chercher ici à caractériser la dynamique de la distribution d'ARN messager. Pour cela, nous allons comme précédemment écrire l'équation maîtresse qui régit l'évolution de $P(r,t)$ (Figure 2-2) :

$$\partial_t P(r,t) = k_R P(r-1,t) + \gamma_R (r+1) P(r+1,t) - (k_R + \gamma_R r) P(r,t)$$

Avec $r \geq 0$ et $P(-1,t) \equiv 0$.

On a de plus comme condition initiale : $P(r,0) = \delta_{r,0}$, soit une probabilité nulle de trouver $r \neq 0$ à l'instant initial.

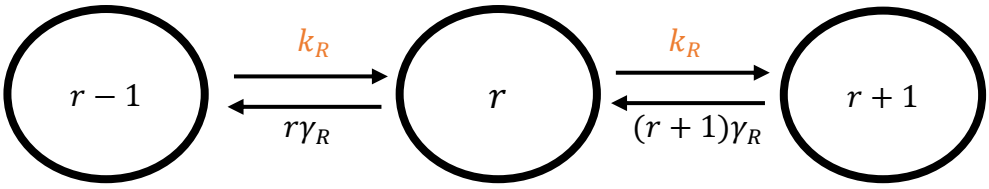


Figure 2-2 : Illustration du processus de vie et de mort modélisant la stochasticité du nombre d'ARN messager. La probabilité par unité de temps k_R permet d'incrémenter d'une unité le processus stochastique (synthèse représentée en orange) et la probabilité par unité de temps $r\gamma_R$ de décrémenter d'une unité sachant que l'état initial est r (dégradation représentée en noir).

On introduit la fonction génératrice :

$$G(z,t) = \sum_{r=0}^{+\infty} z^r \, P(r,t)$$

En multipliant l'équation par z^r et en sommant sur r, on obtient :

$$\partial_t G(z,t) = k_R z G(z,t) + \gamma_R \partial_z G(z,t) - k_R G(z,t) - \gamma_R z \partial_z G$$

Ce qui se réécrit :

$$\partial_t G + \gamma_R (z-1) \partial_z G = k_R (z-1) G \qquad (2.10)$$

Il s'agit d'une équation aux dérivées partielles du premier ordre avec des coefficients non constants, c'est-à-dire de la forme :

$$a(z,t)\frac{\partial G}{\partial t} + b(z,t)\frac{\partial G}{\partial z} = F\big(z,t,G(z,t)\big)$$

Une méthode standard pour résoudre une telle équation est la méthode des caractéristiques. L'idée de cette technique est de chercher dans le plan (z,t) des courbes $(z(s),t(s))$ paramétrées par s telle que sur ces courbes l'équation précédente s'écrit :

$$\frac{dG}{ds} = F\left(z(s),t(s),G\big(z(s),t(s)\big)\right) \Rightarrow \frac{dG}{ds} = \frac{\partial G}{\partial s} + \frac{\partial z}{\partial s}\frac{\partial G}{\partial z} + \frac{\partial t}{\partial s}\frac{\partial G}{\partial t} = F$$

On en déduit par identification avec l'équation (2.10) :

$$\begin{cases} \dfrac{\partial z}{\partial s} = \gamma_R (z(s)-1) & (2.11) \\ \dfrac{\partial t}{\partial s} = 1 & (2.12) \\ \dfrac{dG}{ds} = k_R (z(s)-1) G & (2.13) \end{cases}$$

L'équation (2.12) donne $t = s$, si bien que :

$$(2.11) \implies \frac{\partial z}{\partial t} = \gamma_R (z-1)$$

$$(2.13) \implies \frac{dG}{dt} = k_R (z-1) G$$

La résolution de l'équation sur $z(t)$ donne :

$$z(t) = 1 + a e^{\gamma_R t}$$

Où a paramétrise les conditions initiales. Il s'en suit que :

$$\frac{dG}{dt} = k_R a e^{\gamma_R t} G$$

Ce qui s'intègre en :

$$G(t) = F(a) \exp\left(\frac{k_R a e^{\gamma_R t}}{\gamma_R}\right)$$

Où F est une fonction du paramètre a que nous déterminerons par les conditions initiales.

On peut d'autre part réécrire le paramètre a selon : $a = (z-1)e^{-\gamma_R t}$ de sorte que :

$$G(z,t) = F((z-1)e^{-\gamma_R t}) \exp\left(\frac{k_R (z-1)}{\gamma_R}\right)$$

D'autre part, d'après la condition initiale $P(r,0) = \delta_{r,0}$ conduit à :

$$G(z,0) = \sum_{r=0}^{+\infty} z^r \, P(r,0) = \sum_{r=0}^{+\infty} z^r \, \delta_{r,0} = 1$$

ce qui donne :

$$F(z-1) \exp\left(\frac{k_R (z-1)}{\gamma_R}\right) = 1 \implies F(z) = \exp\left(\frac{-k_R z}{\gamma_R}\right)$$

Finalement :

$$G(z,t) = \exp\left(\frac{k_R ((1-z)e^{-\gamma_R t})}{\gamma_R}\right) \exp\left(\frac{k_R (z-1)}{\gamma_R}\right)$$

Que l'on réécrit :

$$G(z,t) = \exp\left(\frac{k_R z (1 - e^{-\gamma_R t}))}{\gamma_R}\right) \exp\left(\frac{k_R (e^{-\gamma_R t} - 1)}{\gamma_R}\right)$$

On pose : $\lambda(t) = \frac{k_R}{\gamma_R}(1 - e^{-\gamma_R t})$. On développe ensuite en série entière la première exponentielle suivant la variable z :

$$\exp\big(z\lambda(t)\big) = \sum_{r=0}^{+\infty} \frac{(z\lambda(t))^r}{r!}$$

$$\Longrightarrow G(z,t) = \exp\big(-\lambda(t)\big) \sum_{r=0}^{+\infty} \frac{\big(z\lambda(t)\big)^r}{r!} \stackrel{\text{def}}{=} \sum_{r=0}^{+\infty} z^r P(r,t)$$

On en déduit donc par identification la loi de probabilité du processus stochastique de synthèse/dégradation des ARN messagers :

$$P(r,t) = \exp\big(-\lambda(t)\big) \frac{\big(\lambda(t)\big)^r}{r!} \qquad (2.14)$$

C'est-à-dire une loi de Poisson de paramètre $\lambda(t) = \frac{k_R}{\gamma_R}(1 - e^{-\gamma_R t})$. A un instant t donné, le nombre moyen d'ARN messagers, ainsi que la variance sont donnés par :

$$\langle r(t) \rangle = \lambda(t) = \frac{k_R}{\gamma_R}(1 - e^{-\gamma_R t})$$

$$\sigma^2_{r(t)} = \lambda(t) = \frac{k_R}{\gamma_R}(1 - e^{-\gamma_R t})$$

On retrouve les résultats de l'étude précédente pour l'état stationnaire atteint pour $t \to +\infty$:

$$\langle r \rangle^{st} = \sigma_r^{2\,st} = \frac{k_R}{\gamma_R}$$

Nous avons donc démontré que le processus stochastique de synthèse/dégradation des ARN messagers implique que le nombre de transcrits à un instant donné suit une loi de Poisson selon le modèle à deux niveaux de l'expression génique. Ce modèle permet de décrire, on le rappelle, un promoteur non régulé ou constitutif, qui est à chaque instant disponible pour recruter une polymérase et initier la transcription. Nous allons maintenant passer à la dynamique de la distribution du nombre de protéines.

vi) <u>Distribution du nombre de protéines :</u>

Nous nous intéresserons en premier lieu au « burst traductionnel » pour déterminer la distribution du nombre de protéines synthétisées à partir d'une molécule d'ARN messager unique au cours de sa vie. Puis, nous déterminerons la distribution stationnaire du nombre de protéines lorsque l'on a r ARN messagers dans la cellule.

Considérons une molécule d'ARN messager. Nous allons calculer la probabilité que cet ARN messager produise p protéines au cours de sa « vie ». Le calcul que nous reprenons ici a notamment été réalisé par plusieurs auteurs pour démontrer l'hypothèse de burst traductionnel (Philipps et al, 2012 ; Bresslof, 2014). Nous noterons par la suite $P(p)$ cette probabilité. Dans un intervalle de temps Δt, deux évènements peuvent se produire pour notre ARN : (i) le messager peut recruter un ribosome et initier la traduction, avec le modèle adopté, cela correspond à un incrément d'une unité du processus stochastique représenté par le nombre n de protéines. La probabilité d'un tel évènement est $k_P \Delta t$. (ii) le messager peut être dégradé par une ribonucléase, la probabilité qu'un tel évènement se produise pendant l'intervalle de temps considéré s'écrit : $\gamma_R \Delta t$. Ici, puisqu'on regarde ce qui se passe au niveau d'un ARN messager, alors l'état du système est caractérisé par la donnée de $\{r, p\}$, avec p qui peut prendre n'importe quelle valeur dans l'ensemble des entiers naturels, et $r = \{0,1\}$, suivant que l'ARN messager est dégradé ($r = 0$) ou non ($r = 1$). Ainsi, on note $P(r = 1, p, t)$ la probabilité qu'à l'instant t on ait n protéines et l'ARN messager n'est pas dégradé. On note $P(r = 0, p, t)$ la probabilité qu'au bout du temps t on ait p protéines et le messager est dégradé. Bien évidemment, une fois que l'ARN est dégradé, il n'est plus possible de synthétiser des protéines supplémentaires. D'autre part, puisque le temps de vie d'un ARN messager est très petit devant le temps de vie d'une protéine, on peut faire l'approximation qu'au cours de sa vie, un ARN messager n'a pas le temps de « voir » les protéines se dégrader. Ainsi, on négligera ici la dégradation des protéines. Les équations maîtresses qui régissent les évolutions temporelles des deux équations s'écrivent :

$$\frac{\partial P(r = 1, p, t)}{\partial t} = k_P P(r = 1, p - 1, t) - k_P P(r = 1, p, t) - \gamma_R P(r = 1, p, t) \quad (2.15)$$

$$\frac{\partial P(r = 0, p, t)}{\partial t} = \gamma_R P(r = 1, p, t) \quad (2.16)$$

Comme précédemment, on doit imposer $P(-1, t) \equiv 0$ pour le cas où $p = 0$. La probabilité qui nous intéresse est $P(p)$, soit la probabilité d'avoir p protéines au bout du temps de vie d'une protéine. Cette probabilité doit donc être liée à $P(r =$

$0, p, t)$. Pour déterminer $P(p)$, il nous faudrait mesurer pour un nombre important de différents ARN messagers le nombre de protéines que ces derniers ont permis de synthétiser avant leur dégradation. Cependant, comme les temps de vie des ARN messagers sont une variable stochastique distribué selon notre modèle par une distribution exponentielle, tous les ARN messagers ne se dégraderont pas en même temps, et une mesure à un instant fixe ne nous permettra pas de conclure sur la distribution $P(p)$. Il faut donc ajouter une condition à la probabilité $P(r = 0, p, t)$ pour pouvoir en déduire $P(p)$. D'autre part, on sait qu'au bout d'un temps long comparativement au temps de vie moyen d'un ARN messager (γ_R^{-1}), on est certain que ce dernier sera dégradé, si bien que $\lim_{t \to +\infty} P(r = 1, p, t) = 0$. Ainsi, en mesurant la quantité de protéines synthétisée par chaque ARN au bout d'un temps long, on pourra en déduire $P(p)$:

$$P(p) = \lim_{t \to +\infty} P(r = 0, p, t)$$

Puisque cette probabilité se calcule sur des temps longs, cela revient à considérer comme atteint l'état stationnaire. La distribution $P(p)$ est donc stationnaire et ne présente aucune dépendance temporelle.

D'après l'équation (21),

$$P(r = 0, p, t) = \gamma_R \int_0^t P(r = 1, p, t')dt'$$

On en déduit donc :

$$P(p) = \gamma_R \int_0^{+\infty} P(r = 1, p, t)dt \quad (2.17)$$

On intègre maintenant l'équation (2.15) :

$$\int_0^{+\infty} \frac{\partial P(r = 1, p, t)}{\partial t} dt = \int_0^{+\infty} k_P P(r = 1, p - 1, t)dt - \int_0^{+\infty} k_P P(r = 1, p, t)dt - \int_0^{+\infty} \gamma_R P(r = 1, p, t)dt$$

En utilisant l'équation (2.17), on obtient :

$$\lim_{t \to +\infty} P(r = 1, p, t) - P(r = 1, p, 0) = -P(p) + \frac{k_P}{\gamma_R}(P(p - 1) - P(p))$$

De plus, $\lim_{t\to+\infty} P(r=1,p,t)=0$ et en utilisant comme condition initiale qu'au temps $t=0$ on a une molécule d'ARN messager et aucune protéine synthétisée correspondante, soit $P(r=1,p,0)=\delta_{p,0}$. L'équation établie précédemment se réécrit selon :

$$-\delta_{p,0} = -P(p) + \frac{k_P}{\gamma_R}\big(P(p-1) - P(p)\big) \quad (2.18)$$

Ainsi, pour $p=0$:

$$P(0) = \frac{\gamma_R}{k_P+\gamma_R} \quad (2.19)$$

Et pour $p \geq 1$, nous avons la relation de récurrence :

$$P(p) = \frac{k_P}{k_P+\gamma_R} P(p-1)$$

Avec un raisonnement par récurrence à partir de l'égalité (2.19), on obtient :

$$P(p) = \frac{\gamma_R}{k_P+\gamma_R}\left(\frac{k_P}{k_P+\gamma_R}\right)^p \quad (2.20)$$

Qui peut se réécrire, en posant $b=\frac{k_P}{\gamma_R}$:

$$P(p) = \left(1-\frac{b}{1+b}\right)\left(\frac{b}{1+b}\right)^p \quad (2.21)$$

Ainsi, le nombre de protéines synthétisé par ARN messagers est distribué selon une loi géométrique de paramètre $b/(1+b)$. On peut également calculer le nombre moyen de protéines synthétisées par ARN messager en utilisant la fonction génératrice :

$$\langle p\rangle = \partial_z G(z)|_{z=1} \; ; \; G(z) = \sum_{p=0}^{+\infty} z^p P(p)$$

Pour déterminer $G(z)$, on part de l'équation (2.18) en multipliant chaque membre par z^n et on somme sur n, on obtient alors :

$$G(z) = \frac{-1}{b(z-1)-1}$$

En dérivant par rapport à z et en évaluant cette dérivée en $z=1$, nous en déduisons le nombre moyen de protéines produit par un ARN messager :

$$\langle p \rangle = \partial_z G(z)|_{z=1} = b = \frac{k_P}{\gamma_R}$$

Nous retrouvons le terme de « burst traductionnel ». L'idée de burst traductionnel repose sur le fait que les protéines ont un temps de vie beaucoup plus long que les ARN messagers. Ainsi, chaque protéine synthétisée à partir d'un ARNm donné ne sera pas dégradé (en moyenne) avant que le messager ne le soit. Sur l'échelle de temps des protéines, tout se passe comme si les protéines étaient synthétisées simultanément à partir d'un messager (Figure 2.3). La distribution (2.21) aurait également pu être déterminée plus intuitivement (Yu et al, 2006 ; Shahrezaei & Swain, 2008). En effet, une fois l'ARN messager synthétisé et dans un intervalle de temps Δt, différents évènements peuvent se produire : le messager peut soit se lier avec un ribosome, soit rencontrer une ribonucléase, ou ne se fixer à aucune de ces molécules. Si on considère qu'un évènement de rencontre avec un ribosome ou une RNase ait bien lieu, alors la probabilité d'initier la traduction ou de déclencher la dégradation sachant l'évènement de rencontre, s'écrit désormais comme une probabilité conditionnelle, les deux évènements en compétition s'excluant mutuellement. La probabilité d'initier la traduction s'écrit donc : $\rho = {k_P}/{k_P + \gamma_R}$, et la probabilité de rencontrer une ribonucléase s'écrit ${\gamma_R}/{k_P + \gamma_R} = 1 - \rho$. Ainsi, pour observer un « burst » de protéines de taille p, il faut pour cela que p ribosomes se soient fixés au niveau du RBS, suivi de la fixation de la ribonucléase. Ainsi, on obtient :

$$P(p) = \frac{\gamma_R}{k_P + \gamma_R}\left(\frac{k_P}{k_P + \gamma_R}\right)^p = \left(1 - \frac{b}{1+b}\right)\left(\frac{b}{1+b}\right)^p$$

Une autre manière de procéder encore consiste à raisonner sur le temps de vie τ d'un ARN messager (Shahrezaei & Swain, 2008). En effet, ces derniers sont distribués exponentiellement :

$$P(\tau) = \gamma_R \exp(-\gamma_R \tau)$$

Nous avons vu que la synthèse protéique à partir d'une molécule d'ARN messager est un processus de Poisson. Ainsi, la probabilité d'avoir p protéines au bout d'un temps de vie τ de la molécule d'ARN messager s'écrit : $\frac{(k_P\tau)^p}{p!}\exp(-k_P\tau)$. Cette dernière probabilité est une probabilité conditionnelle. Ainsi, la probabilité d'observer p protéines et un temps de vie du messager τ s'écrit :

$$P(\tau)\frac{(k_P\tau)^p}{p!}\exp(-k_P\tau) = \gamma_R \exp(-\gamma_R\tau)\frac{(k_P\tau)^p}{p!}\exp(-k_P\tau)$$

Le nombre de protéines p peut être obtenu pour différents temps de vie du messager. On intègre donc la probabilité précédente par rapport à τ pour obtenir la probabilité recherchée :

$$P(p) = \int_0^{+\infty} d\tau \gamma_R \exp(-\gamma_R \tau) \frac{(k_P \tau)^p}{p!} \exp(-k_P \tau)$$

En utilisant $\int_0^{+\infty} dx x^n e^{-ax} = \frac{n!}{a^{n+1}}$, on obtient bien l'équation (2.20).

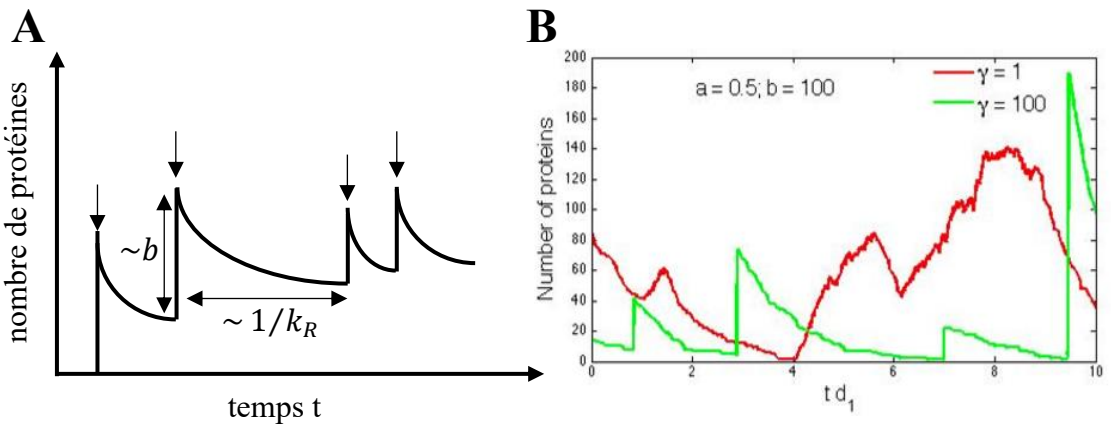


Figure 2-3 : Burst traductionnel. **A)** Représentation schématique du phénomène de burst traductionnel. Chaque flèche représente la synthèse d'un transcrit. Comme les ARN messagers sont beaucoup plus instable que les protéines, alors sur une échelle de temps calée sur la durée de vie moyenne d'une protéine, tout se passe comme si à chaque fois qu'un ARN messager est synthétisé ce dernier injecte dans la cellule un « burst » de protéines dans le cytoplasme avant sa dégradation qui survient relativement vite. La taille de chaque burst est distribuée géométriquement et a pour valeur moyenne $b = k_P/\gamma_R$. La durée entre chaque burst est distribuée exponentiellement et est égale au temps d'attente entre deux évènements de synthèse d'un ARN messager. La durée moyenne entre chaque burst est donnée par $1/k_R$. Entre chaque période de synthèse protéique, le nombre de protéines décroît du fait de leur dégradation. D'après Thattai & van Oudenaarden, 2001 **B)** Evolutions temporelles du nombre de protéines résultant de simulations stochastiques du modèle à deux niveaux de l'expression génique pour deux valeurs différentes du rapport $\gamma = \gamma_R/\gamma_P$. Quand on augmente γ de 1 (courbe rouge) à 100 (courbe verte), on voit apparaître des bursts abrupts de synthèse protéique : la faible durée de vie des transcrits conduit à une activité de synthèse protéique courte mais très intense. Un rapport $\gamma \gg 1$ implique une synthèse protéique par « rafales » successives. $a = k_R/\gamma_P$; $b = k_P/\gamma_R$ et $d_1 = \gamma_P$ est le taux de dégradation des protéines. Les paramètres de la simulation sont : $a = 0.5$; $b = 100$ et $d_1 = 0.0005s^{-1}$. D'après Shahrezaei & Swain, 2008.

Nous avons ici calculé la distribution du nombre de protéines générées par burst à partir d'une molécule d'ARN messager unique dans l'état stationnaire. Cependant, le nombre total de protéines est le résultat de la traduction de nombreuses molécules d'ARN messagers. Pour calculer la distribution du nombre de protéines dans ce cas plus complexe, une des approches possibles est d'écrire l'équation maîtresse régissant l'évolution de la probabilité jointe $P(r, p, t)$:

$$\partial_t P(r,p,t) = k_R P(r-1,p,t) + \gamma_R (r+1) P(r+1,p,t) + k_P r P(r,p-1,t) + \gamma_P (p+1) P(r,p+1,t) - (k_R + \gamma_R r + k_P r + \gamma_P p) P(r,p,t)$$

Puis de sommer sur l'ensemble des valeurs possible de r pour déterminer la distribution du nombre de protéines $P(p)$ (Shahrezaei & Swain, 2008).

Une autre approche consiste à écrire une équation maîtresse effective en se basant sur le concept de burst traductionnel : puisque la synthèse d'un transcrit et le burst de protéines de tailles j générées par ce dernier se déroulent simultanément, on regroupe dans une même étape les processus de synthèse/dégradation des ARN messagers ainsi que la traduction. La seule variable aléatoire à prendre en compte est donc le nombre de protéines $p(t)$ qui caractérise le processus. Cependant, cette équation maîtresse tient compte de manière implicite des fluctuations du nombre d'ARN messagers (Shahrezaei & Swain, 2008). Le nombre de protéines $p(t)$ évoluent suivant le schéma cinétique suivant :

$$p \xrightarrow{k_R P(j)} p + j$$

$$p \xrightarrow{p\gamma_P} p - 1$$

La première réaction correspond à un incrément du nombre de protéines de la quantité j apporté par la transcription et le burst traductionnel de taille j. La probabilité que dans un temps Δt un tel évènement se produise est $k_R P(j)\Delta t$. La seconde réaction correspond à la réaction de protéolyse. Dans un temps Δt, la probabilité d'un tel évènement s'écrit $p\gamma_P \Delta t$. On en déduit l'équation maîtresse donnant l'évolution de $P(p,t)$:

$$\frac{\partial P(p,t)}{\partial t} = k_R \left(\sum_{j=1}^{p} P(j)\, P(p-j,t) - \sum_{j=1}^{+\infty} P(j)\, P(p,t) \right) + \gamma_R (p+1) P(p+1,t) - \gamma_R p P(p,t)$$

Comme $\sum_{j=1}^{+\infty} P(j) = 1$, on en déduit :

$$\frac{\partial P(p,t)}{\partial t} = k_R \left(\sum_{j=1}^{p} P(j)\, P(p-j,t) - P(p,t) \right) + \gamma_R (p+1) P(p+1,t) - \gamma_R p P(p,t) \qquad (22)$$

Au lieu de résoudre directement cette équation, nous allons faire l'hypothèse qu'il y a suffisamment de ribosomes et autres molécules impliquées dans la traduction de telles sortes que la traduction de chaque ARN messager est indépendante (Philipps et al, 2013). Supposons que nous avons $r = 2$ ARN messagers. Du fait de l'hypothèse précédente, la probabilité de produire p protéines du aux traductions indépendantes de chaque ARNm s'écrit :

$$P_2(p) = \sum_{j=0}^{p} P(j)P(p-j)$$

Cette dernière égalité représente le produit de convolution de la probabilité $P(p)$ avec elle-même (Philipps et al, 2013).

Pour le cas où $r = 3$, on obtient :

$$P_3(p) = \sum_{j=0}^{p} P(j)P_2(p-j) = \sum_{j=0}^{p} P(j) \sum_{j'=0}^{p-j} P(j')P(p-j-j')$$

$$= \sum_{j=0}^{p} \sum_{j'=0}^{p-j} P(j)P(j')P(p-j-j')$$

…

$$P_r(p) = \int_0^p P(j)P_{r-1}(p-j)dj$$

Ainsi, avec cette hypothèse de travail, Philipps et al concluent que calculer la distribution du nombre total de protéines synthétisées par r ARNm revient à calculer de multiples produits de convolutions de la probabilité individuelle $P(p)$ avec elle-même. En considérant le nombre de protéines relativement grand de tel sorte que l'on puisse remplacer les sommes par des intégrales, on obtient :

$$P_r(p) = \int_0^p P(j)P_{r-1}(p-j)dj$$

En utilisant la transformation de Laplace défini par : $\tilde{P}(s) = \int_0^{+\infty} P(t)e^{-st}dt$, on montre que la transformée de Laplace du produit de convolution $C(t) = \int_0^t f(t')g(t-t')$ s'écrit :

$$\tilde{C}(s) = \tilde{f}(s)\tilde{g}(s)$$

La transformée de Laplace de $P_r(p)$ s'écrit donc :

$$\tilde{P}_r(s) = [\tilde{P}(s)]^r$$

Pour calculer la probabilité d'avoir p protéines à partir de r ARN messagers, on doit dans un premier temps calculer l'intégrale $\tilde{P}(s)$ avec :

$$P(p) = \left(1 - \frac{b}{1+b}\right)\left(\frac{b}{1+b}\right)^p = \left(\frac{b}{1+b}\right)^{p+1}$$

$$\tilde{P}(s) = \int_0^{+\infty} P(p)e^{-sp}dp$$

Ainsi, on obtient :

$$\tilde{P}(s) = \int_0^{+\infty} \left(\frac{b}{1+b}\right)^{p+1} e^{-sp}dp = \frac{b}{1+b}\int_0^{+\infty} \left(\frac{b}{1+b}e^{-s}\right)^p dp$$

Avec :

$\left(\frac{b}{1+b}e^{-s}\right)^p = \exp\left(ln\left(\left(\frac{b}{1+b}e^{-s}\right)^p\right)\right) = \exp\left(p\, ln\left(\left(\frac{b}{1+b}e^{-s}\right)\right)\right)$, on en déduit que :

$$\tilde{P}(s) = \frac{b}{1+b}\frac{\left(\frac{b}{1+b}e^{-s}\right)^p}{\ln\left(b/1+b\right) - s}\Bigg|_0^{+\infty} = -\left[(1+b)\left(\ln\left(b/1+b\right) - s\right)\right]^{-1}$$

D'où l'on déduit :

$$\tilde{P}_r(s) = [\tilde{P}(s)]^r = \left[-(1+b)\left(\ln\left(b/1+b\right) - s\right)\right]^{-r}$$

La transformation inverse a une forme analytique et s'écrit (Philipps et al, 2013) :

$$P_r(p) = \left(\frac{b}{1+b}\right)^p \left(\frac{1}{1+b}\right)^r \frac{p^{r-1}}{\Gamma(r)}$$

La distribution précédente est une première approximation de la loi binomiale négative dans le cas où $p \gg 1$. La distribution du nombre de protéines s'écrit en fait (Shahrezaei & Swain, 2008) :

$$P_r(p) = \left(\frac{b}{1+b}\right)^p \left(\frac{1}{1+b}\right)^r \frac{\Gamma(r+p)}{\Gamma(p+1)\Gamma(r)}$$

Où $\Gamma(r) = (r-1)!$,avec r un entier est la fonction Gamma. Dans le cas où $b, p \gg 1$, cette distribution peut également se réécrire :

$$P_r(p) \cong \frac{p^{r-1}e^{-p/b}}{b^r \Gamma(r)} \qquad (2.23)$$

Cette distribution se nomme loi Gamma (Figure 2-4). Il s'agit d'une loi continue qui se justifie par l'approximation $b, p \gg 1$. Elle représente la probabilité d'avoir p protéines à partir de r ARN messagers. Ici, r n'est pas une variable stochastique mais un paramètre que nous avons fixé pour les besoins du calcul. Ce paramètre est par la suite imposé pour permettre à la distribution précédente d'être solution de l'équation maîtresse (2.22). D'autre part, d'après les propriétés de la loi Gamma :

$$\langle p \rangle = rb, \qquad \sigma_p^2 = rb^2, \qquad \frac{\sigma_p^2}{\langle p \rangle} = b, \qquad \frac{\sigma_p^2}{\langle p \rangle^2} = \frac{1}{r}$$

Le fait que le facteur Fano soit égal à la taille moyenne d'un burst dévie du comportement observé lorsque nous avons calculé ce facteur (équation (2.7)) à partir du modèle discret correspondant à l'équation maîtresse sur la probabilité jointe $P(r, p, t)$. Cette différence provient du caractère continu de la distribution $P_r(p)$ (Friedman et al, 2006). La déviation est cependant faible dans le cas où $b > 1$, et justifie *a posteriori* l'approximation continue et la loi Gamma qui en découle. En comparant la moyenne donnée par la distribution et celle calculée précédemment, on peut alors en déduit la valeur du paramètre r :

$$\langle p \rangle = rb = r\frac{k_P}{\gamma_R} = \frac{k_R k_P}{\gamma_R \gamma_P} \Longrightarrow r = \frac{k_R}{\gamma_P}$$

Cette valeur du paramètre r permet également à la distribution (2.23) d'être solution de l'équation maîtresse (2.22). k_R représentant le taux de synthèse d'ARN messagers par unité de temps et $1/\gamma_P$ le temps de vie moyen d'une protéine, on peut interpréter le paramètre r comme le nombre moyen d'ARN messagers synthétisés sur le temps de vie d'une protéine. Comme chaque transcrit injecte quasi-simultanément un burst de protéines dans le cytoplasme, on en déduit que ce terme représente le nombre de bursts que « voit » une protéine au cours de sa vie. La distribution (2.23) est donc totalement caractérisée par la donnée des paramètres dynamiques k_R/γ_P et k_P/γ_R.

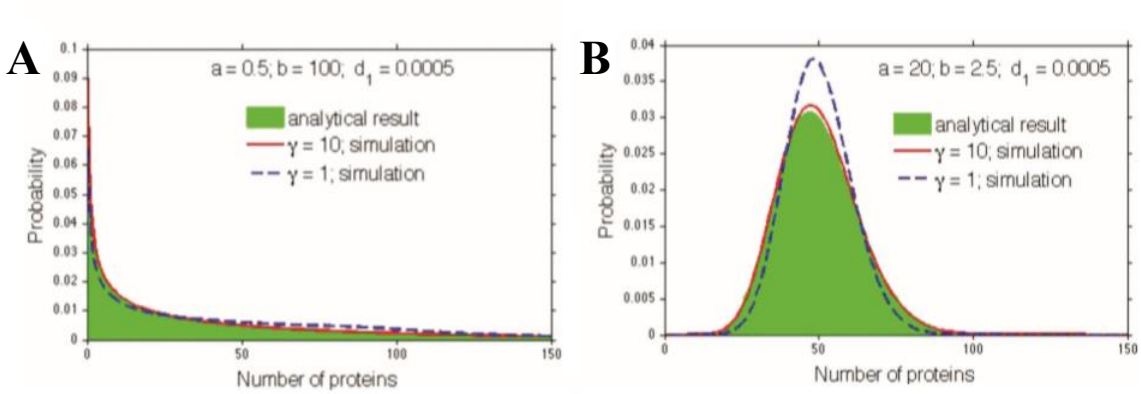


Figure 2-4 : Prédictions théoriques et simulations du modèle à deux niveaux de l'expression génique (Shahrezaei & Swain, 2008). La courbe en verte représente le résultat analytique donné par l'équation (28). Les autres courbes représentent les simulations stochastiques pour $\gamma = \gamma_R/\gamma_P \gg 1$ (courbes rouges) et $\gamma = 1$ (courbes bleues). L'exactitude de la loi Gamma avec les simulations est d'autant plus importante que le rapport des temps de vie γ est grand. $a = k_R/\gamma_P$; $b = k_P/\gamma_R$ et $d_1 = \gamma_P$ est le taux de dégradation des protéines. **A)** Les paramètres de la simulation sont : $a = 0.5;\ b = 100$ et $d_1 = 0.0005s^{-1}$. **B)** $a = 0.5;\ b = 100$ et $d_1 = 0.0005s^{-1}$.

ANNEXE 3

- Modèle stochastique à trois niveaux :

Si les phases de synthèse et de dégradations des messagers et des protéines sont identiques au cas du modèle à deux niveaux, voyons plus en détail l'étape d'activation/inactivation du promoteur. Pour cela, nous allons caractériser l'état du promoteur par la variable aléatoire g qui peut prendre deux valeurs : $g = 0$ pour le cas où le promoteur est inactif, et $g = 1$ dans le cas contraire. Imaginons le cas d'un contrôle transcriptionnel par un répresseur. Le répresseur qui inactive le gène se lie à la région promotrice avec une probabilité par unité de temps k_-. Le répresseur quitte le promoteur après un temps d'attente distribué selon une loi exponentielle de paramètre k_+, permettant au promoteur de passer dans l'état actif. Le temps d'attente avant qu'un nouveau répresseur ne vienne s'attacher à la région promotrice est distribué selon une loi exponentielle de paramètre k_-. Les probabilités de transition k_- et k_+ entre ces deux états contiennent implicitement l'information sur le nombre de répresseurs supposé ici constant. Leur valeur sont donc déterministes et ne prennent en compte que les sources de bruit intrinsèque. Nous avons considéré ici le cas simple où nous n'avons qu'une seule copie du gène. Cependant, du fait de la réplication du gène et suivant la localisation chromosomique du gène d'intérêt (aussi appelé locus), il est possible que plusieurs copies d'un même gène soit présent dans la cellule (Paulsson, 2005). Soit g^{tot} le nombre de copies de ce gène et supposons ce nombre constant (ce qui est évidemment faux en toute rigueur). Si on caractérise par $g(t)$ le nombre de promoteurs actifs à un instant t donné, alors on a $g^{tot} - g(t)$ promoteurs inactifs. Ainsi la probabilité en un temps Δt d'incrémenter g d'une unité s'écrit : $k_+(g^{tot} - g)\Delta t$. De même, la probabilité de décrémenter g d'une unité en un temps Δt s'écrit : $k_- g \Delta t$. Un état du système est décrit par la donnée de $\{g, r, p\}$ qui est un processus stochastique Markovien dont les différentes étapes s'écrivent (Paulsson, 2005) :

Activation du promoteur :	$g \xrightarrow{k_+(g^{tot}-g)} g+1$
Inactivation du promoteur :	$g \xrightarrow{k_- g} g-1$
Transcription :	$r \xrightarrow{k_R g} r+1$
Dégradation ARNm :	$r \xrightarrow{\gamma_R r} r-1$
Traduction :	$p \xrightarrow{r k_P} p+1$
Protéolyse :	$p \xrightarrow{p\gamma_P} p-1$

La probabilité d'observer le système dans l'état $\{g, r, p\}$ est donnée par la distribution jointe de probabilité $P(g, r, p, t)$.

i) Expression analytique du nombre moyen de protéines et du bruit d'expression génique d'après le modèle à trois niveaux :

Pour caractériser le bruit d'expression génique à l'aide de ce modèle, nous devons comme nous l'avons fait précédemment calculer les moments d'ordre un et d'ordre deux de chaque variable aléatoire qui caractérise le processus stochastique, à savoir g, r et p. Pour mieux interpréter l'expression analytique du bruit d'expression génique, nous allons commencer par voir les expressions analytiques des fluctuations au niveau de l'activité des gènes $\left(\eta_g^2\right)$, puis au niveau des ARN messagers (η_m^2), et enfin au niveau protéique $\left(\eta_p^2\right)$.

i.i) Fluctuations au niveau de l'activité des promoteurs :

A partir de la figure 3-1, on peut établir l'équation maîtresse sur la probabilité $P(g, t)$:

$$\frac{dP(g,t)}{dt} = k_+(g^{tot} - g + 1)P(g-1,t) - k_+(g^{tot} - g)P(g,t) - k_-gP(g,t) + k_-(g+1)P(g+1,t)$$

Avec les conditions : $P(-1) = 0$ et $P(g > g^{tot}) = 0$.

Considérons en premier lieu le cas où $g = 1$ et $g^{tot} = 1$. C'est-à-dire le cas où on a une seule copie du gène d'intérêt et ce dernier est actif.

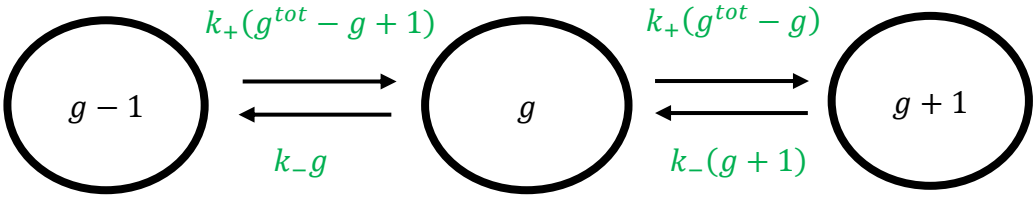


Figure 3-1 : Illustration du processus stochastique de vie et de mort Markovien g. En vert sont indiquées les transitions entre états actifs et inactifs des promoteurs.

L'équation maîtresse précédente se réécrit dans ce cas particulier selon :

$$\frac{dP(g=1)}{dt} = k_+P(g=0) - k_-P(g=1)$$
$$P(g=1) + P(g=0) = 1$$

A l'état stationnaire, la probabilité de trouver le promoteur dans l'état actif ou inactif s'écrit respectivement :

$$P(g = 1) = P_+ = \frac{k_+}{k_+ + k_-}$$
$$P(g = 0) = P_- = \frac{k_-}{k_+ + k_-}$$

Comme chaque copie du gène est indépendante, ces probabilités stationnaires représentent également la probabilité de trouver un gène tiré au hasard parmi les g^{tot} copies soit dans l'état actif, soit dans l'état inactif. Si on revient au cas où l'on a g^{tot} copies du même, on détermine les quantités $\langle g \rangle$ et η_g^2 à l'état stationnaire en introduisant la fonction génératrice $G(z,t) = \sum_{g=0}^{g^{tot}} z^g P(g,t)$ dans l'équation maîtresse stationnaire. Après calcul, on obtient (Paulsson, 2005 ; Bar-Even et al, 2006) :

$$\langle g \rangle = g^{tot} \frac{k_+}{k_+ + k_-} = g^{tot} P_+ \qquad (3.1)$$

$$\eta_g^2 = \frac{1}{g^{tot}} \frac{k_-}{k_+} = \frac{1 - P_+}{\langle g \rangle} \qquad (3.2)$$

Ainsi, d'après l'équation (3-2), pour un niveau moyen d'activité donné, le bruit (quantifié par η_g^2) est plus faible que dans le cas de fluctuations Poissoniennes. Ceci est notamment dû à la limite du nombre total de gènes g^{tot} (Paulsson, 2005). En effet, à partir de l'équation maîtresse, il est possible de montrer que la distribution stationnaire du nombre de gènes actifs est binomiale de paramètre g^{tot} (qui pourrait représenter le nombre de fois que l'on jette un dé ou une pièce) et P_+ (qui représente la probabilité d'un succès). A partir de ces équations, on peut considérer deux cas de figures : si le temps durant lequel un gène est activé est très petit devant le temps où il est inactif, alors $k_+ \ll k_-$. D'après les équations (3-1) et (3-2), $\langle g \rangle$ sera alors faible et les fluctuations autour de cette valeur seront importantes et proche d'un comportement Poissonien. Dans le cas contraire, si le gène est actif sur un temps beaucoup plus long que celui pendant lequel il est inactif alors $\langle g \rangle$ tend vers la valeur g^{tot}, et les fluctuations tendent vers zéro. En d'autres termes, le dernier cas correspond à la situation où le modèle à trois niveaux rejoint celui à deux niveaux.

i.ii) Fluctuations au niveau du nombre d'ARN messager et « burst transcriptionnel » :

On s'intéresse maintenant au processus stochastique $\{g, r\}$ dont une représentation est donnée par la figure 3-2. L'équation maîtresse régissant la probabilité $P(g, r, t)$ s'écrit :

$\frac{dP(g,r,t)}{dt} = k_+(g^{tot} - g + 1)P(g-1,r,t) + k_-(g+1)P(g+1,r,t) + (r+1)\gamma_R P(g,r+1,t) + gk_R P(g,r-1,t) - (k_- g + k_+(g^{tot} - g) + gk_R + r\gamma_R)P(g,r,t)$

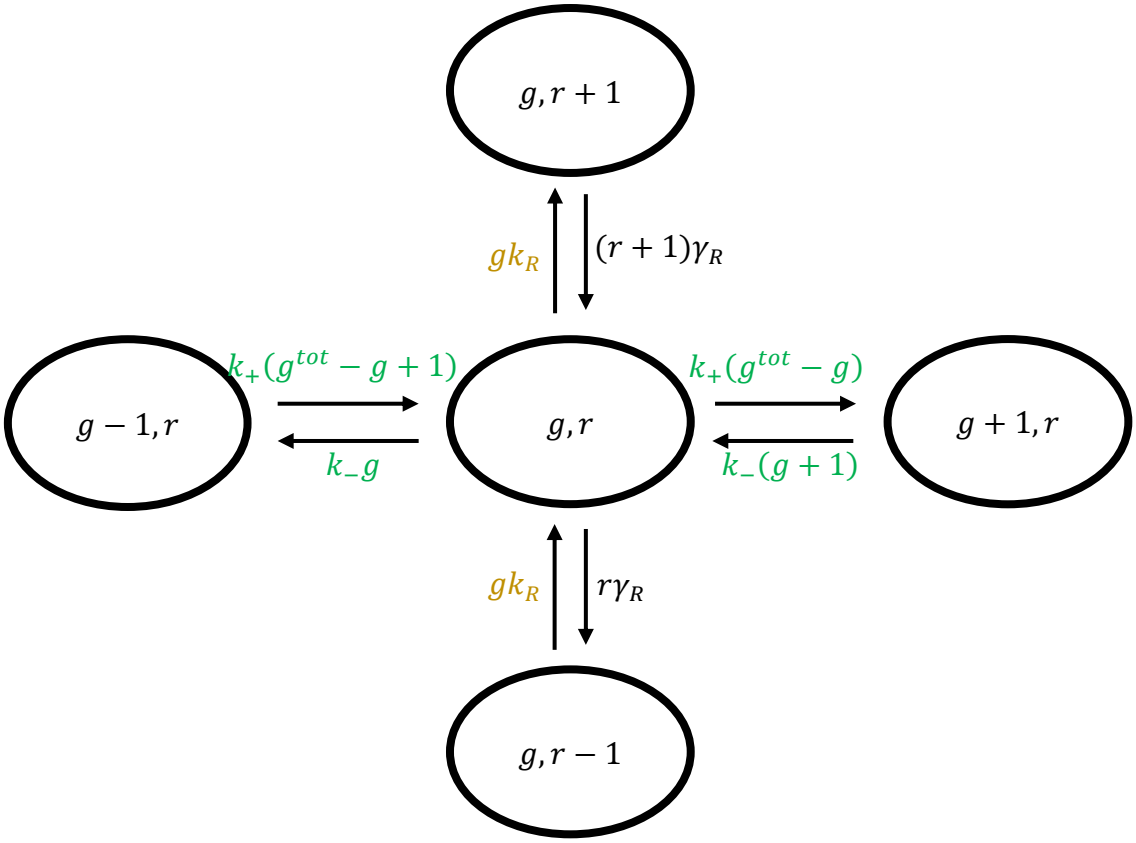


Figure 3-2 : Illustration du processus Markovien $\{g, r\}$. En vert sont indiquées les transitions entre états actifs et inactifs des promoteurs, en orange la synthèse des ARN messagers (transcription) et en noir la dégradation des messagers.

Comme dans le cas précédent, en introduisant la fonction génératrice :

$$G(z, z', t) = \sum_{\substack{0 \le g \le g^{tot} \\ 0 \le r < +\infty}} z^g z'^r P(g, r, t)$$

On en déduit le niveau moyen d'ARN messager $\langle r \rangle$ ainsi que le bruit associé η_r^2 à l'état stationnaire (Paulsson, 2005) :

$$\langle r \rangle = \frac{k_R}{\gamma_R}\langle g \rangle = \frac{k_R}{\gamma_R} g^{tot} P_+ \quad (3.3)$$

$$\eta_r^2 = \frac{1}{\langle r \rangle} + \eta_g^2 \frac{\tau_g}{\tau_R + \tau_g} \quad (3.4)$$

Avec $\tau_g = (k_+ + k_-)^{-1}$ une échelle de temps caractéristique des changements dans l'activité des promoteurs et $\tau_R = \gamma_R^{-1}$ le temps de vie moyen d'une molécule d'ARN messager. On peut expliciter la dépendance du bruit associé au nombre d'ARN messager avec certains paramètres biophysique du système :

$$\eta_r^2 = \frac{\gamma_R}{k_R\langle g\rangle} + \frac{1-P_+}{\langle g\rangle}\frac{\tau_g}{\tau_R+\tau_g};\ \langle g\rangle = g^{tot}P_+ \qquad (3.5)$$

On en déduit que le bruit associé à l'activité des promoteurs aura une contribution importante par rapport aux évènements de synthèse/dégradation des transcrits sur le niveau de bruit des messagers si :

- $\frac{\tau_g}{\tau_R+\tau_g}$ se rapproche de l'unité, c'est-à-dire si $\tau_g \gg \tau_R$, autrement dit pour une dynamique lente de l'activité des promoteurs par rapport au temps caractéristique des fluctuations τ_R du nombre de messagers.

- le bruit η_g^2 tend vers sa valeur maximale, obtenue lorsque $P_+ \ll 1$, c'est-à-dire pour $k_+ \ll k_-$, autrement dit un temps d'activité des promoteurs très court devant le temps où ces derniers sont inactifs.

- le taux de transcription k_R est élevé, ce qui permet de réduire les fluctuations Poissoniennes des évènements spontanés de synthèse et de dégradation des transcrits.

Ainsi, des gènes dont les paramètres biophysiques remplissent les conditions précédentes auront un niveau de bruit du nombre d'ARN messager ayant une forte contribution du bruit généré en amont. Si la dynamique de l'activité des promoteurs est rapide devant le temps des fluctuations du nombre de messagers, soit $\tau_R \gg \tau_g$, alors le bruit η_g^2 se transmettra difficilement au niveau supérieur puisque $\frac{\tau_g}{\tau_R+\tau_g} \approx \frac{\tau_g}{\tau_R} \ll 1$. De même, si $k_- \ll k_+$, c'est-à-dire un promoteur essentiellement actif, alors η_g^2 tend vers sa valeur minimale et n'aura qu'une faible contribution sur le niveau de bruit η_r^2.

Nous pouvons à partir de l'équation (3-4) déterminer le facteur Fano associé au nombre d'ARN messager à l'état stationnaire :

$$\eta_r^2 = \frac{1}{\langle r\rangle} + \eta_g^2\frac{\tau_g}{\tau_R+\tau_g} = \frac{1}{\langle r\rangle}\left(1 + \frac{k_R\tau_R\langle g\rangle(1-P_+)}{\langle g\rangle}\frac{\tau_g}{\tau_R+\tau_g}\right)$$
$$= \frac{1}{\langle r\rangle}\left(1 + (1-P_+)\frac{k_R\tau_R}{\tau_g^{-1}\tau_R+1}\right)$$

Soit :

$$\frac{\sigma_r^2}{\langle r \rangle} = 1 + (1 - P_+) \frac{k_R \tau_R}{\tau_g^{-1} \tau_R + 1} \qquad (3.6)$$

On retrouve alors les conclusions établies précédemment avec le coefficient de variation η_r^2. Si le facteur Fano est proche de 1, alors la variabilité observée au niveau des transcrits est essentiellement due à la stochasticité des évènements de synthèse/dégradation des messagers tandis qu'un facteur Fano très supérieur à 1 (ou de manière plus exacte dès que le facteur Fano devient supérieur à 2), le bruit associé au nombre d'ARN messager est majoritairement dû aux caractères aléatoires des transitions entre les états actif et inactif des promoteurs. Ce sera notamment le cas lorsque $\tau_g \gg \tau_R$, $P_+ \ll 1$ soit $k_+ \ll k_-$ et un taux de transcription k_R élevé. Nous venons de voire quelles étaient les conditions pour que le bruit au niveau de la dynamique du gène ait un effet majeur sur le bruit observé au niveau du nombre d'ARN messager. Maintenant, ce modèle permet de décrire un phénomène de « *burst transcriptionnel* ». De manière rigoureuse, et de la manière dont il a été introduit dans le cas de la synthèse protéique dans le modèle à deux niveaux, le phénomène de « burst transcriptionnel » doit correspondre à de très brèves et intenses périodes de synthèse d'ARN messagers sans que la dégradation des transcrits n'ait lieu. Ainsi, il faut que la condition $k_+ \ll k_-$ soit vérifiée, mais aussi que le temps passé dans l'état actif (k_-^{-1}) soit nettement inférieur au temps de vie des ARN messagers (τ_R), soit : $\tau_g \ll \tau_R$, ce qui nous place dans le contexte d'une dynamique rapide de l'activité des gènes. Enfin, puisque les périodes d'activité sont très courtes, le taux de transcription doit être élevé pour pouvoir convertir chaque période d'activité du gène en une salve d'ARN messagers ($k_R > k_-$). Dans ces conditions, le facteur Fano de la du nombre d'ARN messager s'écrit :

$$\frac{\sigma_r^2}{\langle r \rangle} = 1 + (1 - P_+) \frac{k_R \tau_R}{\tau_g^{-1} \tau_R + 1} \approx 1 + \frac{k_R}{\tau_g^{-1}} = 1 + \frac{k_R}{k_-}$$

Où on retrouve un terme de « burst » correspondant au nombre moyen d'ARN messager synthétisé pendant une période active du promoteur ($\frac{k_R}{k_-}$). La fréquence des bursts est estimé par le temps moyen d'inactivité d'un gène entre deux transitions, soit k_+^{-1}.

i.iii) <u>Fluctuations au niveau du nombre de protéines</u> :

On s'intéresse maintenant au processus stochastique Markovien de synthèse protéique dans sa globalité $\{g, r, p\}$ dont une représentation est donnée par la figure 3-3. L'équation maîtresse régissant la probabilité $P(g, r, p, t)$ s'écrit :

$$\frac{\partial P(g,r,p,t)}{\partial t}$$
$$= k_+(g^{tot}-g+1)P(g-1,r,p,t) + k_-(g+1)P(g+1,r,p,t)$$
$$+ gk_R P(g,r-1,p,t) + (r+1)\gamma_R P(g,r+1,p,t)$$
$$+ rk_P P(g,r,p-1,t) + (p+1)\gamma_P P(g,r,p+1,t)$$
$$- [k_- g + k_+(g^{tot}-g) + r\gamma_R + gk_R + p\gamma_P + rk_P]P(g,r,p,t)$$

A partir de la fonction génératrice :

$$G(z,z',z'',t) = \sum_{\substack{0\le g\le g^{tot}\\ 0<r<+\infty\\ 0<p<+\infty}} z^g z'^r z''^p P(g,r,p,t)$$

Et en évaluant l'équation maîtresse à l'état stationnaire où $(\partial./\partial t = 0)$, on en déduit le niveau moyen d'expression ainsi que la variabilité phénotypique à l'état stationnaire :

$$\langle p\rangle = \frac{k_P}{\gamma_P}\langle m\rangle = \frac{k_P}{\gamma_P}\frac{k_R}{\gamma_R}\langle g\rangle = \frac{k_P}{\gamma_P}\frac{k_R}{\gamma_R}g^{tot}P_+ \qquad (3.7)$$

$$\eta_p^2 = \frac{1}{\langle p\rangle} + \frac{1}{\langle m\rangle}\frac{\tau_R}{\tau_R+\tau_P} + \frac{1-P_+}{\langle g\rangle}\frac{\tau_R}{\tau_R+\tau_P}\frac{\tau_g}{\tau_g+\tau_P}\frac{\tau_g+\tau_P+\tau_g\tau_P/\tau_R}{\tau_g+\tau_R} \qquad (3.8)$$

Où $\tau_P = 1/\gamma_P$ est la durée de vie moyen d'une protéine ou temps caractéristique des fluctuations. En combinant les équations (3.7) et (3.8), on en déduit :

$$\eta_p^2 = \frac{1}{\langle p\rangle}\left(1 + k_P\tau_P\frac{\tau_R}{\tau_R+\tau_P} + k_P k_R\tau_R\tau_P(1-P_+)\frac{\tau_R}{\tau_R+\tau_P}\frac{\tau_g}{\tau_g+\tau_P}\frac{\tau_g+\tau_P+\tau_g\tau_P/\tau_R}{\tau_g+\tau_R}\right)$$

D'où le facteur Fano du nombre de protéines :

$$\frac{\sigma_P^2}{\langle p\rangle} = 1 + k_P\tau_P\frac{\tau_R}{\tau_R+\tau_P} + k_P k_R\tau_R\tau_P(1-P_+)\frac{\tau_R}{\tau_R+\tau_P}\frac{\tau_g}{\tau_g+\tau_P}\frac{\tau_g+\tau_P+\tau_g\tau_P/\tau_R}{\tau_g+\tau_R} \qquad (3.9)$$

Dans l'équation (3.8), le premier terme correspond aux fluctuations Poissoniennes dues aux évènements stochastiques de synthèse/dégradation des protéines. Les deux termes suivants correspondent aux fluctuations du nombre d'ARN messagers qui ont été transmises au niveau protéique. Le premier de ces termes reprend les fluctuations Poissoniennes des évènements spontanés de naissance et de mort des transcrits, tandis que le second terme reprend les fluctuations dans l'activité des gènes. Le facteur Fano (équation 3.9) reprend les trois sources de bruit en éliminant les effets de taille et est plus sensible aux différentes sources biophysique de stochasticité. Si les fluctuations proviennent majoritairement du processus stochastique de création/destruction des transcrits, alors :

$$\frac{\sigma_P^2}{\langle p \rangle} \approx k_P \tau_P \frac{\tau_R}{\tau_R + \tau_P}$$

L'approximation souvent valide que les ARN messagers ont des temps de vie beaucoup plus court que les protéines soit $\tau_R \ll \tau_P$, conduit à :

$$\frac{\sigma_P^2}{\langle p \rangle} \approx k_P \tau_R = \frac{k_P}{\gamma_R}$$

On retrouve ici le terme de burst traductionnel, qui est la signature de la stochasticité des évènements de synthèse/dégradation des transcrits. Pour observer un tel comportement, à savoir une contribution importante de ces évènements sur le niveau de variabilité phénotypique, il faut un taux de traduction efficace pour que les fluctuations Poissoniennes au niveau protéique soit négligeable comme dans le modèle à deux niveaux, mais il faut aussi que les fluctuations au niveau de l'activité des gènes le soient également. Nous avons vu que ces fluctuations n'affectent que très peu la dynamique du nombre d'ARN messagers lorsque le temps caractéristique des transitions (caractérisé par τ_g) entre les états actifs et inactifs étaient très rapide devant le temps caractéristique des fluctuations au niveau des ARN messagers (caractérisé par τ_R). Soit $\tau_R \gg \tau_g$. Puisque $\tau_R \ll \tau_P$, il vient également : $\tau_R, \tau_P \gg \tau_g$. Cette stochasticité générée au niveau de l'activité des promoteurs est d'autant moins ressenti au niveau des messagers et donc des protéines que le taux de transcription est faible (afin d'éviter le phénomène de burst transcriptionnel), ou que le gène est très souvent dans l'état actif comparativement au temps passé dans l'état inactif, soit $k_+ \gg k_-$ ou encore $P_+ \approx 1$. Dans ce cas, la variabilité phénotypique s'écrit comme l'inverse de l'abondance moyenne des protéines avec un terme multiplicatif égal à la taille moyenne d'un burst de traduction :

$$\eta_p^2 = \frac{1}{\langle p \rangle} \frac{k_P}{\gamma_R} \qquad (3.10)$$

Si maintenant la stochasticité au niveau des protéines est due aux évènements aléatoires d'activation et inactivation des promoteurs, alors le facteur Fano s'écrit :

$$\frac{\sigma_P^2}{\langle p \rangle} = k_P k_R \tau_R \tau_P (1 - P_+) \frac{\tau_R}{\tau_R + \tau_P} \frac{\tau_g}{\tau_g + \tau_P} \frac{\tau_g + \tau_P + \tau_g \tau_P / \tau_R}{\tau_g + \tau_R}$$

L'approximation $\tau_R \ll \tau_P$ conduit à :

$$\frac{\sigma_P^2}{\langle p \rangle} \approx k_P k_R \tau_R^2 (1 - P_+) \frac{1}{1 + \tau_g^{-1} \tau_P} \frac{\tau_g^{-1} \tau_P + \tau_P / \tau_R}{1 + \tau_g^{-1} \tau_R}$$
$$\approx k_P k_R \tau_R (1 - P_+) \frac{\tau_P}{1 + \tau_g^{-1} \tau_P}$$

Pour que la stochasticité générée au niveau de la dynamique des promoteurs soit responsable de la majeure partie des fluctuations observées au niveau protéique, il est nécessaire qu'elle ait également une contribution élevée sur la distribution du nombre d'ARN messager. Pour que ce soit le cas, nous avons vu dans la partie précédente qu'une des conditions est que le temps passé dans l'état inactif est beaucoup plus long que le temps passé dans l'état actif, soit : $k_- \gg k_+$, ou encore $P_+ \ll 1$. Dans ces conditions :

$$\frac{\sigma_P^2}{\langle p \rangle} \approx k_P k_R \tau_R \frac{\tau_P}{1 + \tau_g^{-1} \tau_P}$$

Ou encore, avec $\tau_g = (k_+ + k_-)^{-1} \approx k_-^{-1}$:

$$\frac{\sigma_P^2}{\langle p \rangle} \approx k_P k_R \tau_R \frac{\tau_P}{1 + k_- \tau_P} \qquad (3.11)$$

Nous avons vue qu'une condition supplémentaire est que les gènes doivent être activés/inactivés sur des échelles de temps beaucoup plus long que le temps de vie des transcrits, soit $\tau_g \gg \tau_R$, et que le taux de transcription doit être important. Deux situations peuvent alors se présenter, soit $\tau_g \gg \tau_P$, soit $\tau_g \ll \tau_P$, en d'autres termes, soit les transitions se font sur des échelles de temps beaucoup plus long que le temps de vie d'une protéine, soit sur des temps beaucoup plus court. Examinons chacun des deux cas.

- Cas 1 : $\tau_g \gg \tau_P$: dynamique lente de l'activité des promoteurs par rapport au temps caractéristique des fluctuations du nombre de protéines τ_P :

Le facteur Fano s'écrit dans ce cas

$$\frac{\sigma_p^2}{\langle p \rangle} \approx \frac{k_P k_R}{\gamma_R \gamma_P} \Longrightarrow \eta_p^2 \approx \frac{1}{\langle p \rangle} \frac{k_P k_R}{\gamma_R \gamma_P} = \frac{1}{\langle g \rangle} = \eta_g^2 \qquad (3.12)$$

Le facteur Fano est alors égale au nombre de protéines moyens donné par le modèle à deux niveaux. Cette quantité représente le nombre moyen de protéines que produirait chaque promoteur dans l'état actif. En effet, le fait que $\tau_g \gg \tau_P, \tau_R$ permet alors d'atteindre un état stationnaire pour chaque période d'activité des promoteurs. La dynamique de transition est tellement lente que le nombre de protéines et d'ARN messagers suivent l'état des promoteurs, et la variabilité phénotypique est égale à la variabilité au niveau de l'activité des gènes. Ainsi, si on considère un seul gène, les protéines (et les ARN messagers) atteignent un niveau stationnaire donné par le modèle à deux niveaux pendant la longue période d'activité, puis un niveau proche de zéro pendant la période d'inactivité. Cette dynamique du promoteur conduit à une distribution bimodale du nombre de protéines (Figure 3-4-C). On peut noter que ce genre de comportement ne nécessite pas la condition $k_- \gg k_+$. La condition $\tau_g \gg \tau_P, \tau_R$ assure la transmission des fluctuations des promoteurs aux niveaux situés en aval, si bien que dans le cas où $k_- = k_+$, on peut encore négliger les autres contributions sur la variabilité phénotypique et écrire :

$$\frac{\sigma_p^2}{\langle p \rangle} \approx (1 - P_+) \frac{k_P k_R}{\gamma_R \gamma_P} \Longrightarrow \eta_p^2 \approx \frac{(1 - P_+)}{\langle p \rangle} \frac{k_P k_R}{\gamma_R \gamma_P} = \frac{(1 - P_+)}{\langle g \rangle} = \eta_g^2$$

- Cas 2 : $\tau_g \ll \tau_P$: dynamique rapide des promoteurs par rapport au temps caractéristique des fluctuations du nombre de protéines τ_P

Dans ces conditions :

$$\frac{\sigma_p^2}{\langle p \rangle} \approx \frac{k_P k_R}{\gamma_R k_-} \Longrightarrow \eta_p^2 = \frac{1}{\langle p \rangle} \frac{k_P k_R}{\gamma_R k_-} \qquad (3.13)$$

La quantité k_R / k_- représente le nombre moyen d'ARN messagers synthétisé durant la période d'activité du promoteur et k_P / γ_R le nombre moyen de protéines produit par une molécule d'ARN messager au cours de sa vie. Ainsi, le facteur Fano représente ici le nombre moyen de protéines produit pendant la période d'activité du promoteur. Nous avons vu qu'une dynamique rapide des promoteurs ($\tau_g \ll \tau_R$) avec des temps de séjour moyens passés dans l'état actif beaucoup plus court que le temps moyen passé dans l'état inactif, et avec un taux de transcription élevé conduisait à un phénomène de burst transcriptionnel. Ce phénomène rendant alors la contribution des fluctuations au niveau des promoteurs majoritaires dans

la variabilité du nombre d'ARN messagers avec $k_R > k_-$. Dans ce cas, nous avons nécessairement $\tau_g \ll \tau_P$, et les fluctuations dans le niveau des protéines engendrées par le mécanisme de burst transcriptionnel s'écrivent selon l'équation (3.13). Ainsi les conditions $\tau_g \gg \tau_R$, $\tau_g \ll \tau_P$ et $k_- \gg k_+$ font que sur une échelle de temps comparable au temps de vies des protéines, la synthèse protéique se fait de manière intensive sur les courtes périodes d'activité des gènes. La taille moyenne de chaque salve de protéines est alors donnée par : $k_R k_P / \gamma_R k_-$. Une synthèse de protéines en burst peut donc provenir de deux mécanismes : soit (i) d'un burst transcriptionnel qui provient des courtes périodes d'activation des gènes couplée avec une transcription efficace ($k_R > k_-$) suivie d'une traduction efficace ($k_P > \gamma_R$) soit (ii), les gènes sont toujours actifs et la soudaine production de protéines provient seulement d'une traduction efficace des ARN messagers durant leur temps de vie ($k_P > \gamma_R$), ce qui correspond au cas de burst traductionnel. La prépondérance des fluctuations de la dynamique des promoteurs sur la variabilité phénotypique est souvent caractérisée par le terme de burst transcriptionnel, en opposition au concept de burst traductionnel qui caractériserait une prépondérance des évènements de synthèse/dégradation des messagers sur le bruit d'expression génique. Cependant, le terme de burst transcriptionnel correspond en toute rigueur au cas où la synthèse des messagers se fait sur les courtes périodes d'activité des promoteurs sans que les messagers néosynthétisés est le temps de se dégrader. Dans le texte principal et dans ce qui suit, nous parlerons de burst transcriptionnel pour caractériser la contribution de la dynamique des promoteurs sur le bruit d'expression génique.

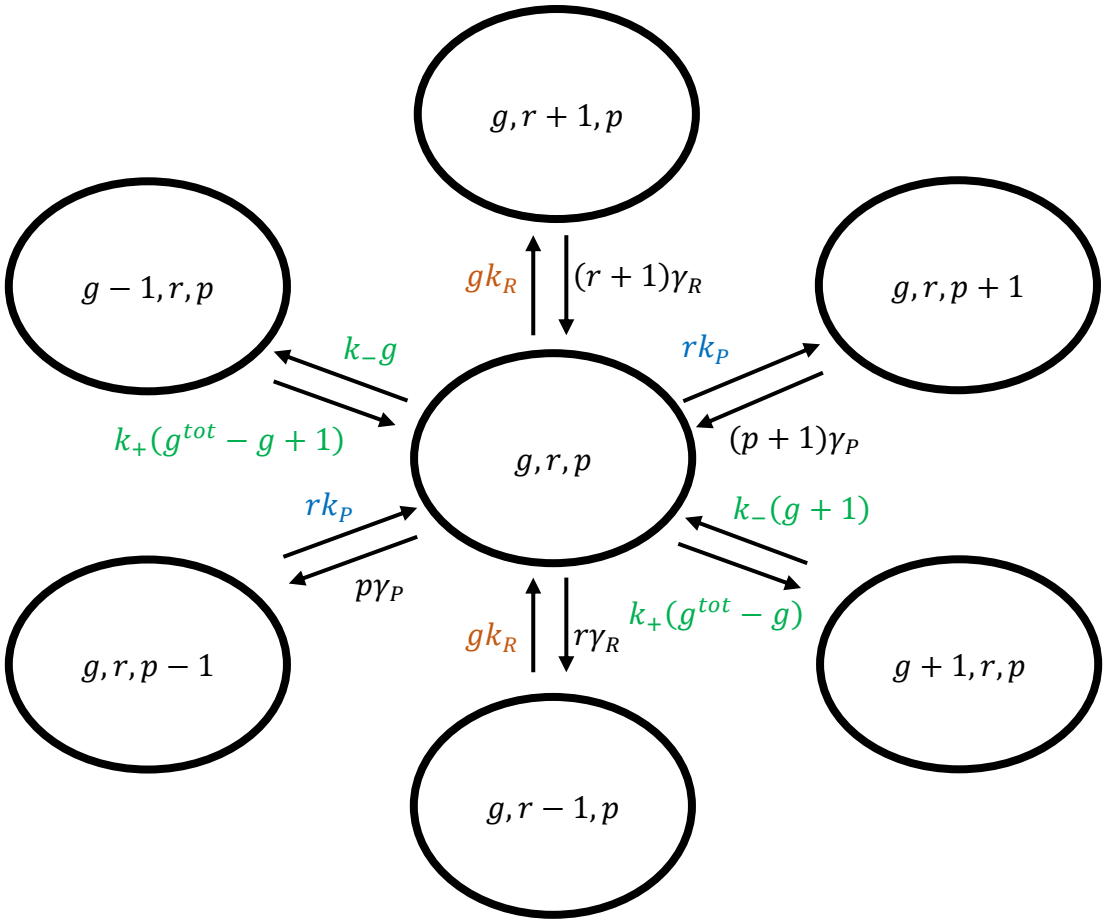


Figure 3-3 : Illustration du processus Markovien $\{g, r, p\}$. En vert sont indiquées les transitions entre états actifs et inactifs des promoteurs, en orange la synthèse des ARN messagers (transcription) et en bleu la synthèse des protéines (traduction). En noir sont indiqués les étapes de dégradations des transcrits et des protéines.

Pour conclure, nous rappelons l'expression de la variabilité phénotypique prédite par le modèle à trois niveaux :

$$\eta_p^2 = \frac{1}{\langle p \rangle} + \frac{1}{\langle m \rangle}\frac{\tau_R}{\tau_R + \tau_P} + \frac{1 - P_+}{\langle g \rangle}\frac{\tau_R}{\tau_R + \tau_P}\frac{\tau_g}{\tau_g + \tau_P}\frac{\tau_g + \tau_P + \tau_g\tau_P/\tau_R}{\tau_g + \tau_R}$$

D'après cette égalité, le premier terme correspond à un comportement Poissonnien des fluctuations du nombre de protéines et représente la borne inférieure des variations stochastiques de la quantité de protéines. Les autres termes correspondent aux fluctuations provenant des niveaux en amont qui sont transmises au niveau protéique. L'«efficacité» de la transmission étant quantifiée par les termes de lissage temporels D'une manière générale, pour que les fluctuations d'un niveau donné aient une contribution majoritaire au(x) niveau(x) en aval il faut que ses fluctuations soient importantes et/ou que ses fluctuations soient efficacement transmises. Il faut également que les contributions des autres niveaux lui soient inférieur. Le meilleur moyen d'identifier les conditions pour lesquelles les fluctuations d'un niveau donné sont majoritaires par rapport au(x)

niveau(x) en aval sur la variabilité phénotypique se fait au moyen du facteur Fano :

$$\frac{\sigma_P^2}{\langle p \rangle} = 1 + k_P \tau_R + k_P k_R \tau_R (1 - P_+) \frac{\tau_P}{1 + \tau_g^{-1} \tau_P}$$

Où on a fait l'approximation $\tau_P \gg \tau_R$. Nous résumons les conditions « optimales » déduites du facteur Fano pour que chacun des deux trois niveaux soient majoritaires sur la variabilité phénotypique (Figure 3-4) :

- Source principale de bruit : activation/inactivation des promoteurs (burst traductionnel)

Le gène d'intérêt doit avoir une dynamique lente des promoteurs, à savoir :$\tau_g \gg \tau_R, \tau_P$, et un temps de séjour moyen dans l'état actif beaucoup plus court que le temps passé dans l'état inactif, soit : $k_- \gg k_+$. Une transcription et une traduction efficace contribueront également à rendre cette source majoritaire sur le bruit intrinsèque. Dans ce cas le bruit s'écrit :

$$\frac{\sigma_P^2}{\langle p \rangle^2} \approx \frac{k_P k_R \tau_R \tau_P}{\langle p \rangle} = \frac{1}{\langle g \rangle} \approx \eta_g^2$$

Dans la figure 3-4 est présentée le cas du burst transcriptionnel obtenue avec $k_- \gg k_+$ et une dynamique rapide du promoteur (Figure 3-4-g-h-i)) .

- Source principale de bruit : évènements de synthèse/dégradation des ARN messagers :

Le gène d'intérêt doit avoir une dynamique rapide des promoteurs, à savoir :$\tau_g \ll \tau_R, \tau_P$, et un temps de séjour moyen dans l'état actif beaucoup plus long que le temps passé dans l'état inactif, soit : $k_+ \gg k_-$. Une transcription faible et une traduction forte contribueront également à rendre cette source majoritaire sur le bruit intrinsèque (Figure 3-4-d-e-f)). Dans ce cas le bruit s'écrit en termes de burst traductionnel :

$$\frac{\sigma_P^2}{\langle p \rangle^2} \approx \frac{k_P \tau_R}{\langle p \rangle}$$

- Source principale de bruit : évènements de synthèse/dégradation des protéines :

Cette source sera majoritaire pour des gènes faiblement transcrit et faiblement traduit, et pour lesquelles le promoteur adopte une dynamique rapide des transitions entre états actifs et inactifs, et que ces derniers se trouvent en majorité dans l'état actif (Figure 3-4-a-b-c)). Le bruit s'écrit alors :

$$\frac{\sigma_P^2}{\langle p \rangle^2} \approx \frac{1}{\langle p \rangle}$$

Nous observons alors que quel que soit la source de stochasticité qui génère la variabilité phénotypique, le bruit intrinsèque présente toujours une dépendance avec l'inverse de l'abondance moyenne des protéines :

$$\eta_P^2 = \frac{C}{\langle p \rangle}$$

C dépendant des paramètres biophysiques pilotant l'expression du gène d'intérêt.

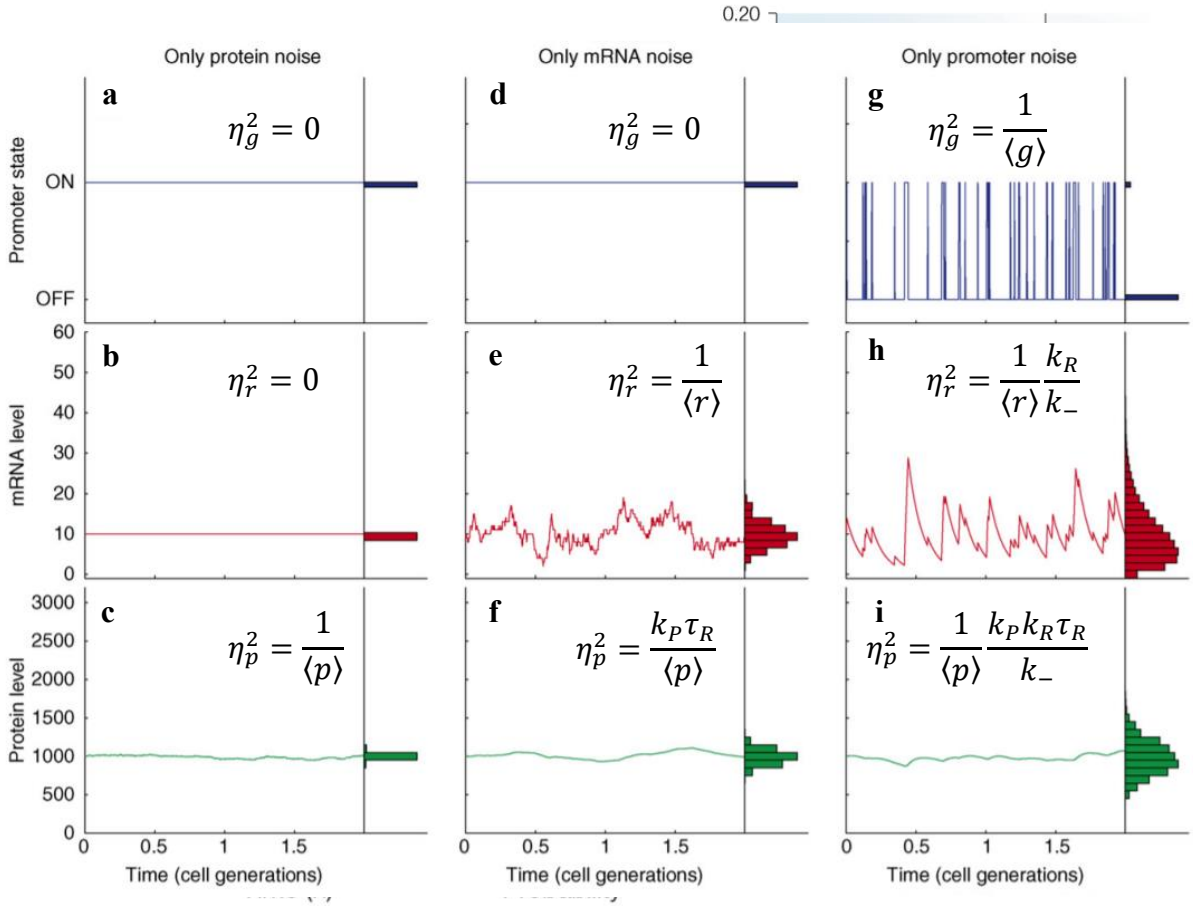


Figure 3-4 : Comparaison des trois sources possibles de bruit (résultats de simulation numérique). D'après Kaufmann & van Oudenaarden, 2007. La première ligne représente la dynamique du promoteur (**a,d,g** bleu), la seconde ligne la dynamique du nombre d'ARN messagers (**b,e,h** rouge) et la troisième ligne la dynamique du nombre de protéines (**c,f,i** vert). La première colonne (**a,b,c**) décrit la situation dans laquelle seul les évènements de synthèse/dégradation des protéines sont stochastiques, les autres étapes étant déterministes. Dans ce cas, on obtient des fluctuations Poissoniennes du nombre de protéines. La seconde colonne (**d,e,f**) correspond au cas où seul les évènements de synthèse/dégradation des ARN messagers sont stochastiques. Le processus stochastique de vie et de mort des transcrits donne des fluctuations Poissoniennes au niveau des messagers. Ces fluctuations se propagent au niveau protéique par le mécanisme de burst traductionnel et confèrent une variabilité phénotypique qui est inversement proportionnelle au nombre moyen de protéines. La constante de proportionnalité est donnée par la taille moyenne d'un burst de protéines. La troisième colonne (**g,h,i**) décrit la situation dans laquelle seul la dynamique d'activation/inactivation du promoteur est stochastique. Dans ce cas le promoteur n'est plus toujours actif mais oscille aléatoirement entre des périodes brèves d'activité et des périodes plus longues d'inactivité, soit $k_- \gg k_+$. Cette stochasticité se transmet au niveau des messagers par le mécanisme de burst transcriptionnel (on a ici $\tau_g = k_-^{-1} \ll \tau_R$). Les courtes périodes d'activité du promoteur conduisent par la suite à un burst de protéines. Là encore, la variabilité phénotypique est inversement proportionnelle au nombre moyen de protéine, et la constante de proportionnalité est alors égale au nombre moyen de protéines produites sur une période active du promoteur. Pour chaque simulation, le nombre d'ARN messager est constant (10 molécules d'ARN messagers) ainsi que le nombre de protéines (1000 molécules). Les notations sont les mêmes que dans le texte principal.

ii) Distribution analytique des nombres d'ARN messagers et de protéines :

Comme pour le modèle à deux niveaux, Shahrezaei & Swain (2008) proposent une méthode analytique pour déterminer les distributions à l'état stationnaire du nombre de protéines. Cette méthode se base sur le fait que les ARN messagers sont beaucoup plus instables que les protéines. A la différence du modèle développé par Paulsson (2005) leur modèle considère un seul promoteur qui peut commuter entre deux états (actif et inactif). Dans leur calcul, les auteurs introduisent $P_{r,p}^{(0)}$ la probabilité d'avoir r ARN messagers et p protéines quand le promoteur est inactif, et $P_{r,p}^{(1)}$ la probabilité d'avoir r ARN messagers et p protéines quand le promoteur est actif. En écrivant les équations maîtresses pour chacune des deux probabilités, on obtient un système de deux équations couplées. Shahrezaei & Swain proposent alors de résoudre ces équations à l'état stationnaire en introduisant les fonctions génératrices $G^{(0)}(z,z') = \sum_{r,p} z^r\, z'^p P_{r,p}^{(0)}$ et $G^{(1)}(z,z') = \sum_{r,p} z^r\, z'^p P_{r,p}^{(1)}$ définies pour chaque état du promoteur. Par la suite, la probabilité d'observer p protéines est déterminée à partir de la fonction génératrice de la variable aléatoire p, à savoir : $G(z') = \sum_p z'^p P_p = G^{(0)}(z') + G^{(1)}(z')$. La probabilité recherchée s'écrit alors :

$$P(p) = P_p = \frac{\Gamma(\alpha+p)\Gamma(\beta+p)\Gamma(k_-/\gamma_P + k_+/\gamma_P)}{\Gamma(p+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(k_-/\gamma_P + k_+/\gamma_P + p)} \times \left(\frac{b}{1+b}\right)^p \left(1-\frac{b}{1+b}\right)^\alpha \times {}_2F_1\left(\alpha+p, k_-/\gamma_P + k_+/\gamma_P - \beta, k_-/\gamma_P + k_+/\gamma_P + p; \frac{b}{1+b}\right)$$

Où

$$\alpha = \frac{1}{2}(a + k_-/\gamma_P + k_+/\gamma_P + \varphi)$$
$$\beta = \frac{1}{2}(a + k_-/\gamma_P + k_+/\gamma_P - \varphi)$$
$$\varphi^2 = (a + k_-/\gamma_P + k_+/\gamma_P)^2 - 4ak_+$$
$$a = {}^{k_R}/_{\gamma_P}\,; b = {}^{k_P}/_{\gamma_R}$$

La fonction ${}_2F_1(a,b,c;z)$ est la fonction hypergéométrique de Gauss. Cette distribution est bien plus compliquée que la distribution binomiale négative (ou loi Gamma) donné par le modèle à deux niveaux et est valable dans le cas où $\frac{\gamma_R}{\gamma_P} \gg 1$ (Figure 3-5). Cette distribution peut générer une distribution bimodale avec un

pic correspondant à zéro nombre de protéines (le promoteur est dans l'état inactif) et un autre pour un nombre différent de zéro dans le cas où les périodes d'activation/inactivation sont très longues (Figure 3-5-C). On retrouve la distribution binomiale négative lorsque $k_- \to 0$, c'est-à-dire lorsque le gène est toujours actif à l'état stationnaire, ou lorsque $\frac{(k_+ + k_-)}{\gamma_P} = \frac{\tau_P}{\tau_g} \gg 1$, soit pour une dynamique rapide du promoteur (Shahrezaei & Swain, 2008).
Shahrezaei & Swain (2008) déterminent également la forme analytique de la distribution du nombre d'ARN messagers à l'état stationnaire :

$$P(r) = \frac{r_s^r e^{-r_s}}{r!} \frac{\Gamma(\zeta_+ + r)\Gamma(\zeta_+ + \zeta_-)}{\Gamma(\zeta_+ + \zeta_- + r)\Gamma(\zeta_+)} {}_1F_1(\zeta_-, \zeta_+ + \zeta_- + r; r_s)$$

Avec : $r_s = {k_R}/{\gamma_R}$, $\zeta_+ = {k_+}/{\gamma_R}$, $\zeta_- = {k_-}/{\gamma_R}$, et ${}_1F_1(a, b; z)$ est la fonction hypergéométrique confluente. Comme pour la distribution de protéines, cette dernière peut être bimodale, ou tendre vers une distribution binomiale négative lorsque $\zeta_- \gg 1$, puisque les ARN messagers sont alors produits par rafales ou en burst. Enfin, on retrouve la distribution Poissonnienne du modèle à deux niveaux dans le cas où le promoteur est toujours actif soit $\zeta_- \to 0$.

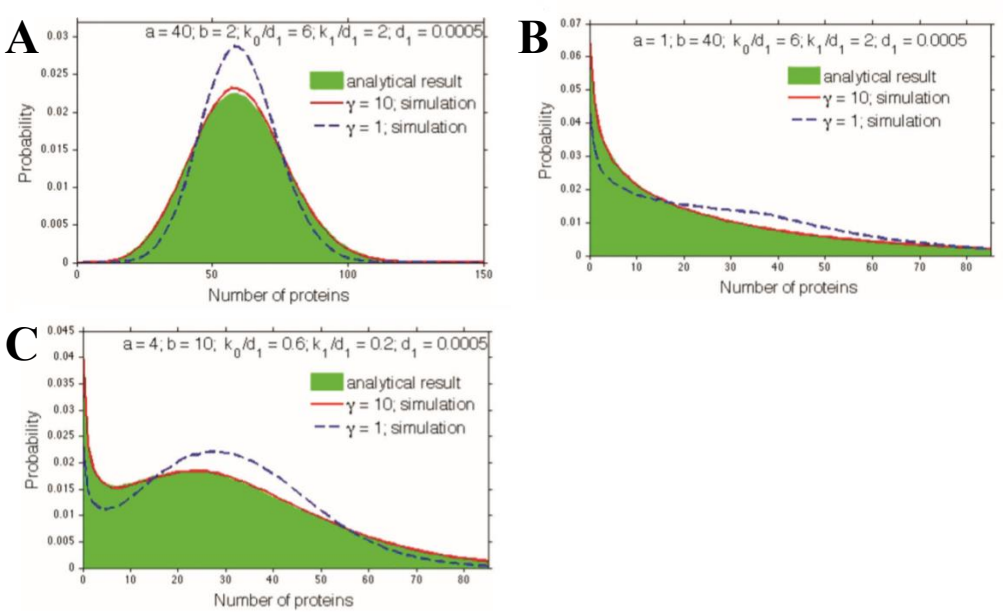


Figure 3-5 : Prédictions théoriques et simulations du modèle à trois niveaux de l'expression génique (Shahrezaei & Swain, 2008) sur la distribution du nombre de protéines. La courbe en verte représente le résultat analytique. Les autres courbes représentent les simulations stochastiques pour $\gamma = \gamma_R/\gamma_P \gg 1$ (courbes rouges) et $\gamma = 1$ (courbes bleues). Les simulations et la distribution analytique s'ajustent d'autant plus que le rapport des temps de vie γ est grand. $a = k_R/\gamma_P$; $b = k_P/\gamma_R$; $d_1 = \gamma_P$; $k_0 = k_+$et $k_1 = k_-$. **A)** Les paramètres de la simulation sont : $a = 40, b = 2$, $k_0 = 6d_1$, $k_1 = 2d_1$, $d_1 = 0.0005s^{-1}$. Le nombre moyen de protéines correspondant est $\langle p \rangle = 60$. **B)** $a = 1, b = 40$, $k_0 = 6d_1$, $k_1 = 2d_1$, $d_1 = 0.0005s^{-1}$, $\langle p \rangle = 30$. **C)** $a = 4, b = 10$, $k_0 = 0{,}6d_1$, $k_1 = 0{,}2d_1$, $d_1 = 0.0005s^{-1}$, $\langle p \rangle = 30$. **B-C)** Le fait de ralentir la dynamique des promoteurs fait apparaître un comportement bimodal de la distribution de protéines.

BIBLIOGRAPHIE

Andrianantoandro, E., Basu, S., Karig, D.K., and Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol *2*, 2006.0028.

Araújo, I.S., Pietsch, J.M., Keizer, E.M., Greese, B., Balkunde, R., Fleck, C., and Hülskamp, M. (2017). Stochastic gene expression in Arabidopsis thaliana. Nature Communications *8*, 2132.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. Nat. Genet. *38*, 636–643.

Becskei, A., and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. Nature *405*, 590–593.

Becskei, A., Séraphin, B., and Serrano, L. (2001). Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. EMBO J. *20*, 2528–2535.

Blake, W.J., KAErn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633–637.

Blake, W.J., Balázsi, G., Kohanski, M.A., Isaacs, F.J., Murphy, K.F., Kuang, Y., Cantor, C.R., Walt, D.R., and Collins, J.J. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. Mol. Cell *24*, 853–865.

Borkowski, O., Goelzer, A., Schaffer, M., Calabre, M., Mäder, U., Aymerich, S., Jules, M., and Fromion, V. (2016). Translation elicits a growth rate-dependent, genome-wide, differential protein production in Bacillus subtilis. Mol. Syst. Biol. *12*, 870.

Bremer, H., and Dennis, P.P. (2008). Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. EcoSal Plus *3*.

Bressloff, P.C. (2014). Stochastic Processes in Cell Biology (Springer International Publishing).

Cai, L., Friedman, N., and Xie, X.S. (2006). Stochastic protein expression in individual cells at the single molecule level. Nature *440*, 358–362.

Choi, P.J., Cai, L., Frieda, K., and Xie, X.S. (2008). A stochastic single-molecule event triggers phenotype switching of a bacterial cell. Science *322*, 442–446.

Coppey, M., Bénichou, O., Voituriez, R., and Moreau, M. (2004). Kinetics of target site localization of a protein on DNA: a stochastic approach. Biophys. J. *87*, 1640–1649.

Coulon, A. (2010). Stochasticité de l'expression génique et régulation transcriptionnelle : Modélisation de la dynamique spatiale et temporelle des structures multiprotéiques. thesis. Lyon, INSA.

Dessalles, R. (2017). Stochastic models for protein production : the impact of autoregulation, cell cycle and protein production interactions on gene expression. thesis. Paris Saclay.

Deutscher, M.P. (2006). Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic Acids Res. *34*, 659–666.

Dieci, G., Bosio, M.C., Fermi, B., and Ferrari, R. (2013). Transcription reinitiation by RNA polymerase III. Biochim. Biophys. Acta *1829*, 331–341.

van Dijl, J.M., and Hecker, M. (2013). Bacillus subtilis: from soil bacterium to super-secreting cell factory. Microb. Cell Fact. *12*, 3.

Dubnau, D., and Losick, R. (2006). Bistability in bacteria. Mol. Microbiol. *61*, 564–572.

Elf, J., Li, G.-W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Fraser, D., and Kærn, M. (2009). A chance at survival: gene expression noise and phenotypic diversification strategies. Molecular Microbiology *71*, 1333–1340.

Fraser, H.B., Hirsh, A.E., Giaever, G., Kumm, J., and Eisen, M.B. (2004). Noise Minimization in Eukaryotic Gene Expression. PLoS Biol *2*.

Friedman, N., Cai, L., and Xie, X.S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Phys. Rev. Lett. *97*, 168302.

Fromion, V., Leoncini, E., and Robert, P. (2012). Stochastic Gene Expression in Cells: A Point Process Approach. ArXiv:1206.0362 [Math, q-Bio].

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison Iii, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods *6*, 343–345.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell *123*, 1025–1036.

Gordon, A., Colman-Lerner, A., Chin, T.E., Benjamin, K.R., Yu, R.C., and Brent, R. (2007). Single-cell quantification of molecules and rates using open-source microscope-based cytometry. Nat. Methods *4*, 175–181.

Guiziou, S., Sauveplane, V., Chang, H.-J., Clerté, C., Declerck, N., Jules, M., and Bonnet, J. (2016). A part toolbox to tune genetic expression in Bacillus subtilis. Nucleic Acids Res. *44*, 7495–7508.

Hilfinger, A., and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc. Natl. Acad. Sci. U.S.A. *108*, 12167–12172.

Huh, D., and Paulsson, J. (2011). Non-genetic heterogeneity from stochastic partitioning at cell division. Nat. Genet. *43*, 95–100.

Iizuka, R., Yamagishi-Shirasaki, M., and Funatsu, T. (2011). Kinetic study of de novo chromophore maturation of fluorescent proteins. Anal. Biochem. *414*, 173–178.

de Jong, I.G., Beilharz, K., Kuipers, O.P., and Veening, J.-W. (2011). Live Cell Imaging of Bacillus subtilis and Streptococcus pneumoniae using Automated Time-lapse Microscopy. J Vis Exp.

Kaern, M., Elston, T.C., Blake, W.J., and Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. *6*, 451–464.

Kaufmann, B.B., and van Oudenaarden, A. (2007). Stochastic gene expression: from single molecules to the proteome. Curr. Opin. Genet. Dev. *17*, 107–112.

Keren, L., van Dijk, D., Weingarten-Gabbay, S., Davidi, D., Jona, G., Weinberger, A., Milo, R., and Segal, E. (2015). Noise in gene expression is coupled to growth rate. Genome Res. *25*, 1893–1902.

Klumpp, S., Zhang, Z., and Hwa, T. (2009). Growth rate-dependent global effects on gene expression in bacteria. Cell *139*, 1366–1375.

Koch, A.L., and Levy, H.R. (1955). Protein turnover in growing cultures of Escherichia coli. J. Biol. Chem. *217*, 947–957.

Kushner, S.R. (2004). mRNA decay in prokaryotes and eukaryotes: different approaches to a similar problem. IUBMB Life *56*, 585–594.

Leoncini, E. (2013). Towards a global and systemic understanding of protein production in prokaryotes. thesis. Palaiseau, Ecole polytechnique.

Leroy, P. (2005). Erreurs de processivité lors de la synthèse protéique chez Escherichia coli. thesis. Paris 6.

Losick, R., and Desplan, C. (2008). Stochasticity and cell fate. Science *320*, 65–68.

Maamar, H., Raj, A., and Dubnau, D. (2007). Noise in gene expression determines cell fate in Bacillus subtilis. Science *317*, 526–529.

Megerle, J.A., Fritz, G., Gerland, U., Jung, K., and Rädler, J.O. (2008). Timing and dynamics of single cell gene expression in the arabinose utilization system. Biophys. J. *95*, 2103–2115.

Mertz, J. (2010). Introduction to optical microscopy (Roberts).

Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., et al. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science *335*, 1103–1106.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. *31*, 69–73.

Ozbudak, E.M., Thattai, M., Lim, H.N., Shraiman, B.I., and Van Oudenaarden, A. (2004). Multistability in the lactose utilization network of Escherichia coli. Nature *427*, 737–740.

Paulsson, J. (2004). Summing up the noise in gene networks. Nature *427*, 415–418.

Paulsson, J. (2005). Models of stochastic gene expression. Physics of Life Reviews *2*, 157–175.

Peccoud, J., and Ycart, B. (1995). Markovian Modeling of Gene-Product Synthesis. Theoretical Population Biology *48*, 222–234.

Pedraza, J.M., and van Oudenaarden, A. (2005). Noise propagation in gene networks. Science *307*, 1965–1969.

Pedraza, J.M., and Paulsson, J. (2008). Effects of molecular memory and bursting on fluctuations in gene expression. Science *319*, 339–343.

Phillips, R., Theriot, J., Kondev, J., and Garcia, H. (2012). Physical biology of the cell (Garland Science).

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Raj, A., and van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. Annu Rev Biophys *38*, 255–270.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene regulation at the single-cell level. Science *307*, 1962–1965.

Shahrezaei, V., and Swain, P.S. (2008). Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. U.S.A. *105*, 17256–17261.

Shcherbo, D., Murphy, C.S., Ermakova, G.V., Solovieva, E.A., Chepurnykh, T.V., Shcheglov, A.S., Verkhusha, V.V., Pletnev, V.Z., HazelWood, K.L., Roche, P.M., et al. (2009). Far-red fluorescent tags for protein imaging in living tissues. Biochem J *418*, 567–574.

Spizizen, J. (1958). Transformation of biochemically deficient strains of bacillus subtilis by deoxyribonucleate. Proc Natl Acad Sci U S A *44*, 1072–1078.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. U.S.A. *99*, 12795–12800.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Thattai, M., and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. Proc. Natl. Acad. Sci. U.S.A. *98*, 8614–8619.

Veening, J.-W., Hamoen, L.W., and Kuipers, O.P. (2005). Phosphatases modulate the bistable sporulation gene expression pattern in Bacillus subtilis. Mol. Microbiol. *56*, 1481–1494.

Veening, J.-W., Stewart, E.J., Berngruber, T.W., Taddei, F., Kuipers, O.P., and Hamoen, L.W. (2008). Bet-hedging and epigenetic inheritance in bacterial cell development. Proc. Natl. Acad. Sci. U.S.A. *105*, 4393–4398.

Vimberg, V., Tats, A., Remm, M., and Tenson, T. (2007). Translation initiation region sequence preferences in Escherichia coli. BMC Mol. Biol. *8*, 100.

Walker, N.E. (2016). A single-cell study on stochasticity growth and gene expression. Delft University of Technology.

Yang, S., Kim, S., Rim Lim, Y., Kim, C., An, H.J., Kim, J.-H., Sung, J., and Lee, N.K. (2014). Contribution of RNA polymerase concentration variation to protein expression noise. Nat Commun *5*, 4761.

Young, J.W., Locke, J.C.W., Altinok, A., Rosenfeld, N., Bacarian, T., Swain, P.S., Mjolsness, E., and Elowitz, M.B. (2012). Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. Nature Protocols *7*, 80–88.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X.S. (2006). Probing gene expression in live cells, one protein molecule at a time. Science *311*, 1600–1603.