Sam Meyer - 3BIM-BISTA 2015-2016

Mini-projet: réponse globale des gènes bactériens à la relaxation de l'ADN et rôle de la séquence promotrice

1 Présentation du problème

Les bactéries régulent l'expression de leurs gènes à l'échelle globale en modifiant l'état physique de leur chromosome. Un des mécanismes majeurs consiste à réguler la topologie de l'ADN à l'aide d'enzymes (topoisomérases). Dans un environnement riche, où la bactérie se multiplie rapidement, la concentration d'ATP est forte dans la cellule, ce qui engendre une forte activité de la gyrase qui induit une torsion négative dans l'ADN. Dans un milieu pauvre, à l'inverse, la torsion est plus faible. Ce mécanisme est vital : plusieurs antibiotiques courants agissent en inhibant la gyrase, induisant une "relaxation" de l'ADN. On étudie l'impact de cette relaxation sur l'expression des gènes, dans une expérience où on a mesuré cette expression sur micro-puces à ADN, soit sur un groupe de cellules contrôle (fichiers `sans_1.pair` et `sans_2.pair`), soit dans un groupe de cellules dont l'ADN a subi une relaxation par la novobiocine (fichiers `relax_1.pair` et `relax_2.pair`). Il y a deux fichiers à chaque fois car l'expérience a été réalisés avec deux réplicats.

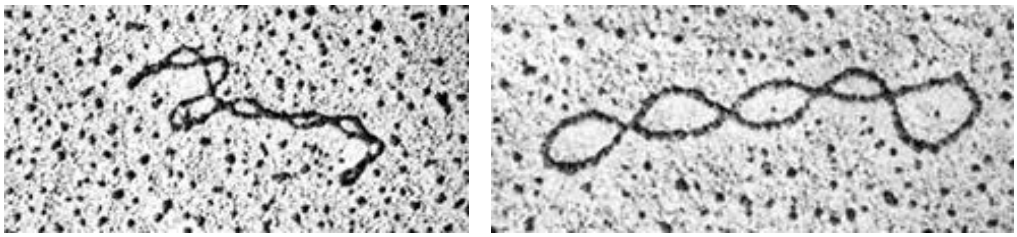

FIGURE 1 – Exemple d'un plasmide (molécule d'ADN circulaire de quelques kilobases) soumis à une torsion forte (à gauche) et relaxé (à droite)

La technique consiste à fixer des "sondes" sur une micro-puce (DNA microarray). Chaque sonde est une séquence d'environ 60 paires de bases, choisie dans la partie codante d'un gène. On mesure la quantité d'ARN messager du gène en question dans une population cellulaire, en rétro-transcrivant chaque ARN en ADN simple-brin complémentaire (cDNA), taggé avec un fluorophore. En hybridant ensuite tous ces cDNA sur la puce, ils vont venir s'hybrider sur les sondes. La fluorescence mesurée sur chaque sonde donne donc une mesure de la quantité d'ARN messager pour le gène en question. Une puce permet de fixer et localiser plusieurs dizaines de milliers de sondes à la fois : on a donc des informations pour l'ensemble des gènes.

Ici, on a donc réalisé chaque expérience en deux réplicats. De plus, on a construit cinq sondes différentes pour chaque gène, et chaque sonde a été placée en triple exemplaire sur chaque puce. Ainsi, au total, on a trente mesures de l'expression de chaque gène. Dans chacun des fichiers de mesure (résultats bruts, non analysés), le nom du gène est donné en deuxième colonne, le nom de la sonde en troisième colonne, et la fluorescence mesurée en avant-dernière colonne. Les lignes avec un nom de gène ne commençant pas par "ECH3837" sont des contrôles techniques et peuvent être ignorées.

On suppose qu'une mesure individuelle de fluorescence f peut être décrite par le modèle suivant :

$$\log(f) = \alpha + \log(n) + h_{sonde} + b_{puce} + \epsilon$$

— α caractérise les propriétés du fluorophore utilisé
— n est le nombre d'ARNm en solution, qui mesure le niveau d'expression du gène
— h_{sonde} est un paramètre qui décrit la probabilité d'hybridation à la sonde considérée. Selon la séquence, cette probabilité est en effet très différente, d'où les valeurs très différentes obtenues pour un même gène avec différentes sondes. Remarquez que les sondes utilisées sont les mêmes dans les 4 expériences.

— b_{puce} décrit des variations d'une puce à l'autre. Une partie du travail d'analyse consiste à évaluer ce paramètre en comparant p. ex. l'hybridation de séquences standard. Ici, pour simplifier on va supposer que ce paramètre est nul.
— ϵ est une variable aléatoire décrivant les variations résiduelles

2 Détection des gènes différentiellement exprimés

Dans un premier temps, on souhaite détecter les gènes qui répondent de façon significative (activation ou répression) à la relaxation de l'ADN.

1. Compte tenu de l'influence de la sonde sur la fluorescence, pouvez-vous directement comparer les différentes mesures de fluorescence obtenues pour chaque gène dans les deux conditions biologiques ?
2. L'effet de la sonde est le même dans les différentes expériences : suggérez une méthode pour vous en affranchir et obtenir, pour chaque gène, un échantillon représentatif de l'effet de la relaxation. A quelle valeur devez-vous comparer cet échantillon pour conclure sur cet effet ?
3. Suggérez deux tests que vous pourriez utiliser pour identifier l'effet de la relaxation sur chaque gène. Lequel choisissez-vous (justifiez votre réponse).
4. Ecrivez les hypothèses nulle et alternative pour chacun des tests que vous faites. Calculez la p-value de votre test pour chaque gène.
5. Nous souhaitons par la suite étudier spécialement les 10% des gènes qui répondent le plus significativement à la relaxation. Sélectionnez ces gènes, et exportez un tableau contenant le nom de ces gènes, la p-value associée, et une colonne indiquant si l'effet est une activation ou une répression. Combien y en a-t-il dans chaque catégorie ?
6. Quelle est la valeur de la p-value minimale de votre liste ? Compte tenu de cette valeur, estimez le risque qu'au moins un de vos gènes soit en fait un faux positif.
7. L'énoncé précise qu'on suppose un effet nul de la puce. Testez cette hypothèse en détaillant votre démarche. Si vous rejetez l'hypothèse nulle, modifiez votre stratégie de test !

3 Etude d'un lien statistique entre séquence promotrice et réponse à la relaxation

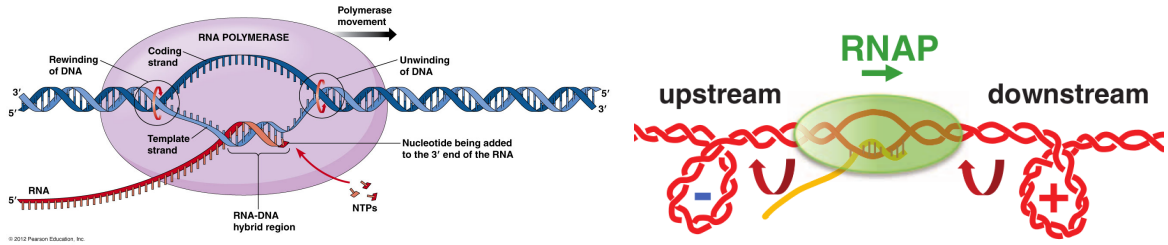

FIGURE 2 – Illustration du lien entre transcription et contraintes de torsion sur l'ADN : pour initier la transcription, la polymérase doit former une "bulle" de dénaturation de l'ADN (gauche) et lors de son avancée (élongation, à droite), elle doit désenchevêtrer les deux brins, ce qui engendre des contraintes de torsion considérables.

Nous travaillons maintenant sur les 10% des gènes qui répondent le plus significativement (positivement ou négativement) à la relaxation. Un des effets de la relaxation est de modifier les propriétés physiques de l'ADN lors de l'initiation ou l'élongation de la transcription : l'ARN polymérase doit ouvrir la double-hélice (initiation) puis avancer en séparant les deux brins enchevêtrés (élongation), et cela peut être *beaucoup* facilité si l'ADN était moins enroulé que dans son état de repos (torsion négative). Cette propriété est vraie quelle que soit la séquence, mais pour expliquer la réponse différentielle des gènes à la relaxation, on peut tester l'hypothèse que les gènes activés ou réprimés diffèrent par leurs propriétés mécaniques. Une signature naïve de ces propriétés est la proportion de bases A-T / G-C. On veut donc tester une différence significative de cette proportion, dans la région promotrice ou codante, entre les deux groupes de gènes.

Vous disposez du fichier contenant la séquence de référence de l'organisme (`sequence.fasta`) et des positions et orientations des gènes (`annotation_dickeya.xlsx`). Attention, dans ce dernier les colonnes indiquent la position gauche et droite : selon l'orientation du gène (brin de référence ou complémentaire), le début (codon start de traduction) peut être l'un ou l'autre ! Remarquez que le nom des gènes est le même que précédemment, sauf qu'il n'y a pas le préfixe indiquant l'organisme en question (''`ECH3937_v6b_`'') : à vous de faire la correspondance.

On souhaite comparer la proportion de GC sur une zone entre -300 et +500 paires de bases par rapport au début de traduction, qui correspond grossièrement à la zone promotrice et codante. Pour le calculer, on va diviser cette zone de taille 800 en 40 fenêtres de taille 20 (disjointes), et calculer la proportion de bases GC dans chaque fenêtre. On pourra ainsi comparer cette proportion dans les deux groupes, à différents endroits le long de la région choisie, pour identifier lesquels jouent potentiellement un rôle biologique.

8. Implémentez une fonction qui prend en entrée une séquence, et retourne la proportion de GC
9. Implémentez des fonctions, afin d'importer la liste des gènes différentiellement exprimés (activés ou réprimés), et pour chacun d'eux, calculer la liste des 40 proportions de bases GC dans l'ordre du brin codant (attention lorsqu'il ne s'agit pas du brin de référence !). Le résultat sera donc un dataframe numérique avec 40 colonnes, chaque ligne correspondant à un gène.
10. Tracez un graphe représentant le niveau de GC moyen et l'erreur-type sur ce niveau (intervalle de confiance de largeur σ, représenté par une barre d'erreur), le long de la zone étudiée (axe x), et pour chacun des groupes. Sur quelle hypothèse repose le calcul de l'erreur-type ? D'après ce graphe, y a-t-il une différence notable ? Si oui, où ? On se focalise dorénavant sur la fenêtre où on a trouvé la différence la plus forte à la question 3, car c'est elle qui joue un rôle potentiel.
11. On souhaite comparer le niveau de GC entre les deux groupes. Choisissez un type de test approprié, et justifiez son utilisation.
12. Quels sont les hypothèses nulle et alternative du test que vous effectuez ? Choisissez un risque α, calculez la p-value. Trouve-t-on une relation significative entre le niveau de GC dans la région choisie et la réponse du gène ?
13. Estimez le risque que l'effet en question soit en fait un faux positif. Pour éviter ce problème, comment devez-vous choisir le risque α à la question précédente au regard de notre protocole d'analyse ?