Le métabarcoding ou comment analyser la biodiversité d'un échantillon environnemental

Le principe consiste à extraire l'ADN d'un échantillon environnemental puis à amplifier par PCR un fragment cible (appelé code-barre) à l'aide d'un couple d'amorces prédéfini.
Le choix du code-barre est déterminant car celui-ci doit être suffisamment variable entre espèces pour pouvoir les discriminer mais très conservée au sein d'une même espèce. De plus, il doit présenter des zones conservées d'une espèce à l'autre pour permettre l'amplification du fragment par PCR chez l'ensemble des espèces étudiées. Les matrices étudiées contiennent généralement de faibles quantités d'ADN endogène ce qui nécessite l'utilisation privilégiée de fragments d'ADN mitochondriaux ou chloroplastiques car leur nombre de copies par cellule est 100 à 1 000 fois supérieur à celui de l'ADN nucléaire. Il est également utile que les code-barres ADN soient phylogénétiquement informatifs, c'est à dire que le niveau de divergence entre ces séquences de référence reflète le niveau de divergence entre les espèces qui les portent. Cette propriété permet d'assigner des espèces inconnues à un taxon d'ordre supérieur (genre, famille, etc.).

La première étape consiste en l'échantillonnage selon des normes précises afin d'éviter toute contamination par de l'ADN exogène. L'étape suivante consiste en l'extraction de l'ADN selon un protocole adapté au type d'échantillon à étudier. Les ADN extraits servent ensuite de matrice à une amplification par PCR avec les amorces correspondant au code-barre préalablement choisi. Après cette étape, les produits PCR obtenus correspondent à un mélange d'amplicons représentatif des ADN des espèces contenus dans l'échantillon de départ.
Pour pouvoir être séquencé, des adaptateurs spécifiques à la technologie de séquençage utilisée doivent être rajoutés ainsi que des *tags* (oligonucléotides uniques pour chaque échantillon) qui permettront d'attribuer chaque séquence à son échantillon de départ.

Nous testons dans le projet Onema-Méthodes deux types de constructions de librairies couramment utilisées en metabarcoding :

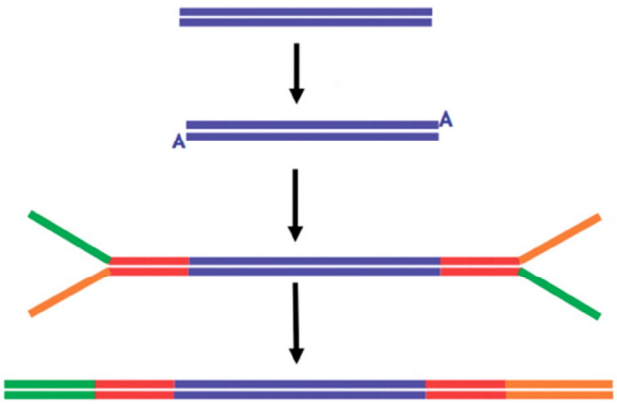
1/ **Le protocole Tailed-PCR** : Deux PCR successives vont permettre d'amplifier la région ciblée et d'y attacher de chaque côté les adaptateurs et tags spécifiques pour chaque échantillon

Amplification de la région cible avec amorces spécifiques + queue moléculaire

Amplification avec amorces contenant la séquence tag + adaptateur

Amplicons prêts à être séquencés

2/ **Le protocole Ligation** : Une seule PCR est réalisée permettant l'amplification de la région ciblée suivie d'une ligation des adaptateurs et du tag associé.



Amplification de la région cible

Ajout d'une adénine en 3' de chaque extrémité

Ligation des adaptateurs et tags à chacune des extrémités

Amplicons prêts à être séquencés

Les librairies amplicons obtenues par chaque protocole sont ensuite séquencées. Il existe différents types de séquenceurs de seconde génération proposés sur le marché avec des caractéristiques propres à chacun quant à la qualité, quantité, la longueur des séquences ainsi que sur le mode d'acquisition des données : optique pour Illumina et électrochimique pour Ion Torrent.

Après le séquençage, les séquences sont triées par échantillon grâce aux *tags*, puis assignées à des taxons par comparaison avec des séquences de référence. L'outil bio-informatique est indispensable pour trier les données, constituer les bases de référence, assigner les séquences aux taxons via ces bases et gérer les erreurs de séquençage.

Nous comparons dans le projet Onema-Méthodes, les données de séquences du code-barre RbcL (marqueur chloroplastique) sur les mêmes échantillons environnementaux avec un séquençage Illumina (MiSeq) ou Ion Torrent (PGM). Le séquençage des librairies sur MiSeq génère en sortie deux fragments par échantillon, dénommés R1 et R2, chacun représentant la séquence construite à partir d'une des amorces délimitant la région amplifiée. Le contigage (ou assemblage) de ces deux fragments R1 et R2 est ensuite réalisé pour produire la séquence entière de la région amplifiée. Il est possible d'ajouter un critère de qualité à ce contigage en jouant sur les paramètres de longueur de recouvrement de la séquence R1 et R2 et sur le nombre de mismatch autorisé dans cette zone de recouvrement.

D'autres étapes de nettoyage des données peuvent également être réalisées (élimination des séquences de mauvaise qualité, des séquences de petites tailles …)

Une fois ce travail réalisé, l'étape suivante consiste à relier chaque séquence à une séquence de référence. On dispose pour cela d'une base de référence pour la région amplifiée qui sert à l'étude, et qui contient des séquences d'organismes qui ont été identifiés de la façon la plus précise possible au niveau taxonomique, recouvrant idéalement l'ensemble de la diversité interspécifique. L'idée générale est de comparer une séquence inconnue (issue d'un échantillon environnemental) à chaque séquence de la base, et de lui affecter le nom attaché à une séquence de la base si la distance est inférieure à un seuil appelé « barcoding gap » (en général, un nombre de mismatch inférieur à 3 % de la longueur de la séquence). Il s'agit d'un inventaire par classification supervisée. On dispose donc :

- d'un fichier fasta de queries (Q)

- d'un fichier fasta des références (R) (ici, de la base R-Syst::diatoms) où chaque séquence est annotée taxonomiquement (ordre, famille, genre, espèce)

Le BLAST est l'outil standard pour de telles comparaisons entre des séquences dites query et des séquences dites de référence. Il existe cependant plusieurs outils qui permettent de construire un inventaire, comme Dada2, où des méthodes intégrées aux suites Mothur et QIIME. Nous développons ici une méthode qui ne fait appel à aucune heuristique, et donc devrait donner les résultats les plus précis, mais au prix d'un investissement plus fort dans le calcul.

Voici ci-dessous le détail de l'analyse réalisée dans le projet Onema-Méthodes :

Etape 1 : calculer toutes les distances entre les séquences de Q et les séquences de R.

Si on a par exemple 50 000 reads dans l'échantillon et 2000 séquences de référence, cela produit un tableau à 50 000 lignes et 2 000 colonnes (distance entre read query ligne et read référence colonne). Les distances sont calculées de façon « exacte » avec l'algorithme de Smith-Waterman. C'est la phase intensive du calcul, qui peut-être implémentée en boucle en utilisant le programme mpi-disseq, sur un cluster de calcul (il est utile de révoir plusieurs centaines de CPU). La sortie de cette étape est donc un grand tableau.

Étape 2 : repérer pour chaque query les reads de référence qui en sont proches. Il s'agit simplement de faire une liste pour chaque query des séquences de référence qui en sont proches, c.a.d. à une distance ≤ au barcoding gap (seuil habituellement de 3 %).

Etape 3 : annoter taxonomiquement ce voisinage. Cette étape consiste à tester si tous les reads de référence à distance ≤ barcoding gap sont de la même espèce. Si oui, le nom de l'espèce est transféré sur le read query, si non, il n'y a pas de résultat pour le read query.

Etape 4 : construire l'inventaire. Cette étape consiste essentiellement à lister les espèces de la base de référence qui ont été reconnues au moins une fois (un read query au moins y a été affecté) compter le nombre de read query qui proviennent de cette espèce.

L'analyse du jeu de données du projet Onema-Méthodes nous conduit à la conclusion que le choix du protocole de construction de librairies et la technologie de séquençage ont un impact significatif sur les résultats d'inventaires obtenus pour un même échantillon. Cette variabilité des résultats quantitatifs produits est cependant estompée lorsque l'utilisation des inventaires fait appel aux fréquences relatives plus qu'aux abondances (indices de diversité, indice de polluo-sensibilité).

Perspectives : L'utilisation d'une technologie émergente comme celle développée par la société d'Oxford Nanopore Technologies (cf https://nanoporetech.com/) permettrait, avec un appareil de la taille d'une grosse clé USB, le séquençage de longs fragments d'ADN (plus résolutif que les technologies de séquençage actuellement utilisées axé sur des fragments courts d'ADN) et l'identification taxonomique en temps réel.