Exemple de scénario utilisant des outils bioinformatiques

Dans le cadre d'une toxi-infection incriminant des plats surgelés, on isole une protéine suspectée d'être responsable du pouvoir toxique. La protéine est retrouvée à une concentration significative, de plus on retrouve le germe producteur, une entérobactérie, mais que l'on n'arrive pas à identifier complètement par microbiologie traditionnelle.
On dispose de la séquence de la protéine afin d'enquêter sur son origine.

La séquence en acides aminés de la protéine est la suivante :

```
PSVNLIPSRKICLQNMINKDNVSVETIQSLLHSKQLPYFSDKRSFLLNLNCQVTDHSGRLIVCRHLASY
WIAQFNKSSGHVDYHHFAFPDEIKNYVSVSEEEKAINVPAIIYFVENGSWGDIIFYIFNEMIFHSEKSRA
LEISTSNHNMALGLKIKETKNGGDFVIQLYDPNHTATHLRAEFNKFNLAKIKKLTVDNFLDEKHQKCYGL
ISDGMSIFVDRHTPTSMSSIIRWPNNLLHPKVIYHAMRMGLTELIQKVTRVVQLSDLSDNTLELLLAAKN
DDGLSGLLLALQNGHSDTILAYGELLETSGLNLDKTVELLTAEGMGGRISGLSQALQNGHAETIKTYGRL
LKKRAINIEYNKLKNLLTAYYYDEVHRQIPGLMFALQNGHADAIRAYGELILSPPLLNSEDIVNLLASRR
YDNVPGLLLALNNGQADAILAYGDILNEAKLNLDKKAELLEAKDSNGLSGLFVALHNGCVETIIAYGKIL
HTADLTPHQASKLLAAEGPNGVSGLIIAFQNRNFEYMGIIKNENITPEEIAEHLDKKNGSDFLEIM
```

Etape 1 : Identification de la protéine par comparaison à une base de donnée

On cherche à savoir si cette protéine est connue et figure dans une base de données. On utilise donc BLAST-P afin d'aligner la séquence par rapport à la base de données.

Procédure :

-Sur le portail NCBI, dans le menu sélectionner **BLAST**
-Dans la page suivante, aller à **BASIC BLAST**, puis sélectionner **Protein Blast**
-Par copier-coller, insérer la séquence étudiée (au format FASTA) dans le logiciel
-Plus bas, dans les options de recherche sélectionner **« Non redundant protein sequences »** à la ligne database. Laisser les autres options par défaut.
-Cocher en bas de page **« Show results in new window »**
-Puis cliquer sur le bouton **BLAST.**

Attention : la requête prend plusieurs secondes.

-Les résultats sont rendus tout d'abord sous forme de graphiques montrant les alignements de séquences probables, chaque séquence étant décrite en bas de page. Les alignements les plus probables figurent en haut de classement (score et % d'identité élevés).

Dans un document texte (Word), coller les résultats obtenus :

-Cliquer sur lien de l'alignement le plus probable.

Remarque : Ligne Query = séquence analysée
Ligne centrale = séquence retenue pour l'alignement
Ligne Subject = séquence référencée dans la base

-Un motif fonctionnel connu a été repéré dans la séquence de la protéine. Comment se nomme t-il ? Chez quels organismes le retrouve t-on ? En quelques mots quel est son mode d'action ?
-Quel est le score ? Quel est le nombre de match, de mismatch et de gap ?
-Quel est le nombre d'acides aminés de cette protéine ?

Etape 2 : Traduction de la séquence protéique en séquence nucléotidique

tBLAST P ou N permettent de convertir une séquence protéique en séquence nucléotidique (et réciproquement) et de comparer la séquence obtenue à une base de données.

On souhaite déterminer si le gène correspondant à la protéine identifiée précédemment figure dans une base de données et si sa structure est connue. On utilise donc tBLAST-N (traduction protéines → nucléotides).

Procédure :

-Sur le portail NCBI, dans le menu sélectionner **BLAST**
-Dans la page suivante, aller à **BASIC BLAST**, puis sélectionner **tBlastN**
-Par copier-coller, insérer la séquence étudiée dans le logiciel
-Plus bas, dans les options de recherche sélectionner **« Nucleotide collections »** à la ligne database.
-Puisque l'on sait maintenant que l'on travaille sur *E.coli*, dans la ligne **Organism**, saisir **« escherichia coli »**. Faire « echap » pour ignorer la liste automatique.
-Cocher en bas de page **« Show results in new window »**
-Puis cliquer sur le bouton **BLAST.**

Dans un document Word, coller les résultats obtenus :

-Identifier le taxon le plus probable.
-Cliquer sur le lien de l'alignement le plus probable. Quel est le score ?
-Cliquer ensuite sur le lien correspondant au n° d'accession (à droite en haut). Quelle est la localisation du gène codant la protéine étudiée ?
-Quelle est la longueur en paires de bases de la séquence nucléotidique ?
-Regarder en bas de page, l'étendue de cette séquence nucléotidique.

-Transformer la page d'information au format FASTA afin de récupérer la séquence nucléotidique plus facilement. Il suffit de cliquer dans le menu supérieur (sous le titre de la séquence) sur « FASTA ». Coller la séquence obtenue dans un document Wordpad et enregistrer avec un nom explicite au format .txt.

Etape 3 : Recherche des amorces permettant l'amplification du plasmide

Vous avez montré précédemment que la protéine étudiée est codée par un plasmide porté par *E.coli*. On chercher maintenant à amplifier ce plasmide par PCR pour en obtenir des quantités importantes en vue d'une utilisation ultérieure (électrophorèse).
Vous disposez de tout le matériel nécessaire pour réaliser la PCR (plasmide isolé à partir de la souche, Taq polymérase, thermocycleur, etc.). Il vous reste à définir la séquence des amorces qu'il vous faudra commander pour réaliser la manipulation.

Le logiciel **Primer3** permet d'obtenir, pour une séquence nucléotidique, les amorces oligonucléotidiques (= primers) qu'il faudra utiliser pour amplifier la séquence par PCR.

Procédure :

-Dans un moteur de recherche, saisir **PRIMER3.** On recherche le logiciel Primer3 Input en version 0.4.0.
-Ouvrir la page, puis par copier-coller, insérer la séquence du plasmide à amplifier.
-Conserver les réglages par défaut, puis lancer la recherche d'amorce par « **Pick Primers** »

Dans un document Word, coller les résultats obtenus :

Donner les 2 amorces droite et gauche possibles pour amplifier le plasmide par PCR.

→ il ne reste plus qu'à commander l'amorce chez un fournisseur de matériels de biologie moléculaire, en donnant la séquence voulue.

Etape 4 : Recherche des sites de restriction

Vous avez amplifié par PCR le plasmide portant le gène codant la protéine d'intérêt. Vous disposez maintenant du plasmide en quantité importante. Au laboratoire, on réalise la digestion du plasmide à l'aide d'enzyme de restriction de façon à isoler le gène codant la protéine d'intérêt.

Après digestion, on réalisera une électrophorèse sur gel de façon à séparer les fragments de plasmide et isoler le gène d'intérêt.

Le logiciel **NEB Cutter** permet de rechercher sur une séquence nucléotidique, les zones de restriction, susceptibles d'être clivées par des enzymes de restriction.

Procédure :

-Dans un moteur de recherche, saisir **« nebcutter ».** On recherche le logiciel NEBcutter V2.0 ouvrir la page du logiciel.

-Au lieu de coller la séquence du plasmide comme on le faisait précédemment, saisir directement à la ligne **« Genbank number »** le n° identifiant du plasmide au format FASTA que vous précédemment placé dans un fichier texte, ou bien charger le fichier dans « local sequence file ».

-Dans les options de réglages, préciser que la séquence est **« circular »** (le plasmide est circulaire) et demander la recherche avec toutes enzymes en sélectionnant **« All specificities »**.
-Régler la valeur « Minimum ORF length to display » par rapport à la taille attendue de la protéine.

-Lancer la recherche en cliquant sur **« Submit »** et patienter !

-Le résultat montre la carte du plasmide avec de nombreuses informations :

-les ORF (Open Reading Frame = cadre de lecture ouvert) matérialisés par des flèches grises épaisses : chaque ORF correspond à un gène codant une protéine donnée.

-les sites de restriction, matérialisés par les points de coupure associés au nom des enzymes de restriction (BamHI, EcoRI, etc.)

-Repérer sur la carte la position de l'ORF correspondant à notre protéine d'intérêt : cliquer sur la flèche grise associée pour ouvrir une fenêtre dédiée à cette protéine. Problème : la représentation circulaire du plasmide ne permet pas de bien visualiser les petits ORF. Préférez dans ce cas une visualisation linéaire (Menu inférieur > Display > Linear)…une fois la protéine repérée, cliquez sur son nom.

-Repérer immédiatement le couple d'enzymes de restriction proposé permettant de couper le gène.

-En complément dans le menu supérieur, une fonction permet de rechercher les différentes enzymes de restriction coupant autour de ce gène **« Locate multiple cutters that excise this ORF »**. Cliquer sur ce lien et patientez pendant le « Running digest ».

De nombreuses enzymes de restriction sont possibles…au final, on utilisera le couple d'enzymes proposé par défaut. Calculer la taille du fragment digéré contenant notre gène d'intérêt.

… Ne fermez pas le logiciel, l'étape 5 est dans la suite logique !

Etape 5 : Prévision des conditions électrophorétiques

-Simuler le découpage du plasmide à l'aide de l'option **« New custom digest »** dans « Main options ». Cochez les enzymes de restriction retenues précédemment par défaut, puis cliquez sur « Digest ».

- Après digestion, vous pourrez alors réaliser l'électrophorèse de façon à isoler la bande correspondant au gène d'intérêt.

Nebcutter propose une option **« View gel »** permettant de simuler l'allure du gel d'électrophorèse après migration pour prévoir qu'elle sera la bande présentant le gène d'intérêt.
Réaliser cette simulation :

- Quel est le numéro de la bande correspondant à notre gène d'intérêt ?
- Quel type de gel utiliserez-vous pour faciliter la récupération de la bande ?