TP BIO-INFORMATIQUE

Plusieurs outils pour travailler en biologie moléculaire

Nous utiliserons les acronyme Français pour les acides nucléiques soit ARN / ADN / et pour parler des deux AxN

Par choix : nous travaillerons dans un organisme PROCARYOTE

Les captures d'écrans et les liens sont testés et validés au 26/01/2021
Si ces liens devaient ne plus fonctionner à l'avenir, merci de contacter jean-pascal.dufour@ac-creteil.fr

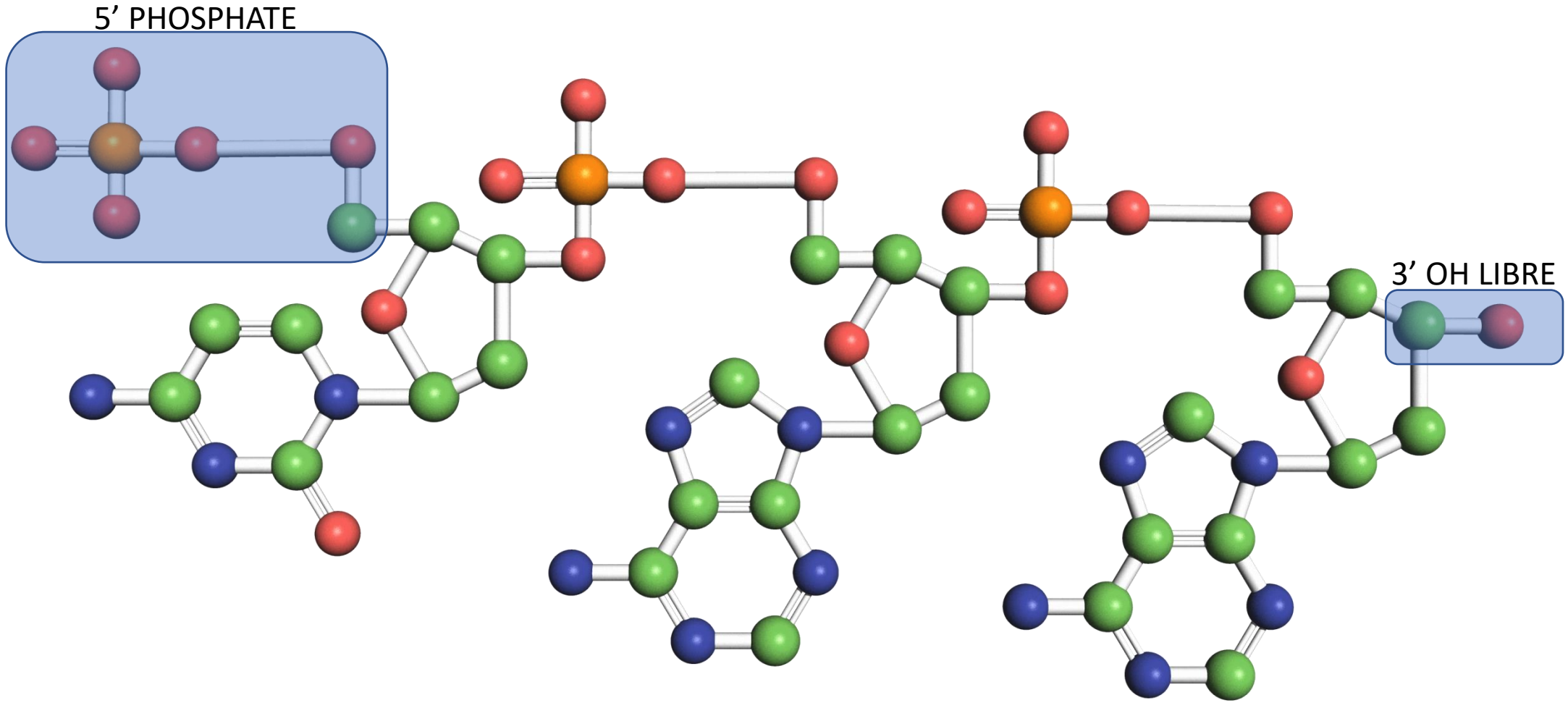
Petite introduction :

- Parler de Biologie moléculaire c'est quoi : parler ADN / ARN et parler PROTEINES (plus tous les liens qui se font entre ces éléments).
- Nous allons exploiter plusieurs outils (un maximum d'outils hors ligne), pour visualiser, annoter, digérer, amplifier des séquences.
- J'ai opté pour des logiciels gratuits (ou pour les versions gratuites de logiciels), nous allons de facto perdre certaines fonctionnalités et devoir combiner plusieurs outils pour arriver à nos fins.
- Pour faciliter l'usage des outils (souvent anglophones soyons réalistes), un petit rappel des fondamentaux « in english » vous est proposé en préambule de l'activité.

Les fondamentaux où il est question de s'orienter :

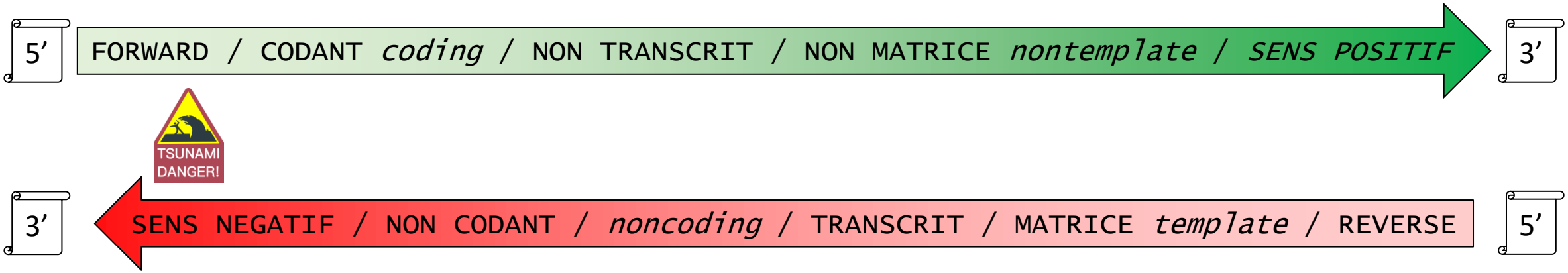
- Une des principales difficulté d'exploitation des séquences et des outils réside dans les termes employés pour parler du SENS de lecture des AxN, plusieurs termes sont employés pour définir la même chose et cela est source de confusions.
- Tout brin d'AxN qui se respecte est ORIENTE : 5' Phosphate → 3' OH Libre.
- La croissance d'un brin se fait dans la nature par ajout d'un nTP à l'extrémité 3' OH libre du brin en cours d'élongation.
- Là où ca coince en général : Les diverses enzymes parcourent l'ADN dans le sens 3'→5' affin de produire par complémentarité un NEOBRIN 5'→3'

Les fondamentaux où il est question de s’orienter :



Les fondamentaux où il est question de s'orienter :

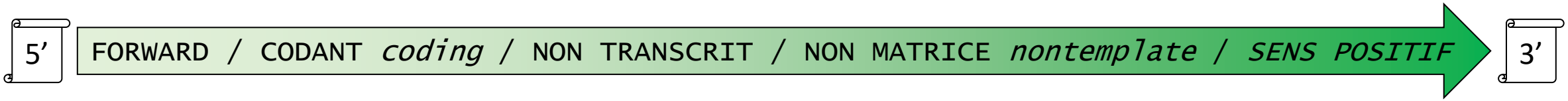
- Par convention, c'est le brin orienté 5'➔3' qui est écrit soit seul, soit sur la ligne supérieure d'un outil de bio-informatique.
- Nous travaillons et donc nous exprimerons toujours dans une logique de Transcription.



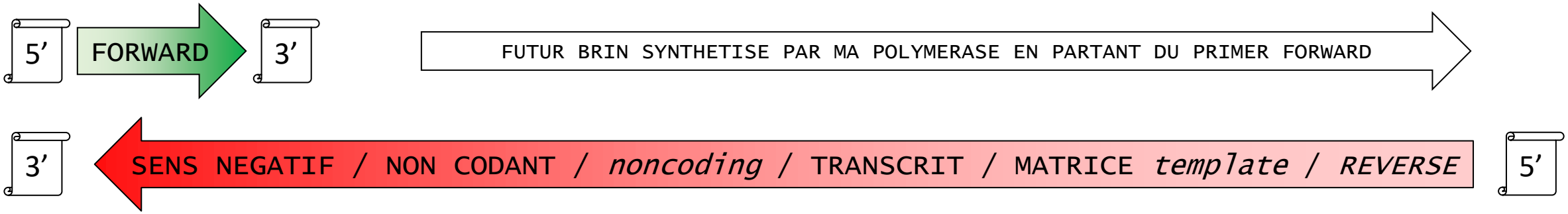
Risque MAJEUR de confusion : Pour la PCR, il est nécessaire de fournir 2 amorces, chacune devant être le début d'un des deux brins d'ADN à amplifier :
Pour synthétiser un nouveau BRIN POSITIF, je dois m'appuyer sur le brin NEGATIF existant. Mon amorce DEVRA DONC ETRE COMPLEMENTAIRE DU BRIN NEGATIF, cette amorce sera donc selon la logique de sens POSITIF donc appelée PRIMER FORWARD ☺

Les fondamentaux où il est question de s'orienter :

Pour synthétiser une copie du brin POSITIF (donc le brin FORWARD)



Je dois fournir le début de cette copie à la POLYMERASE, donc je dois fournir un petit morceau du FORWARD, il est donc d'usage (et logique) de nommer ce PRIMER le PRIMER FORWARD ☺



En conséquence, pour produire une COPIE du brin NEGATIF, je vais devoir fournir à la POLYMERASE un frangent de ce brin, nous nommerons donc ce second PRIMER, le PRIMER REVERSE.

Les fondamentaux où l'on révise son ETLV ☺

- Petit lexique à l'usage des honnêtes profs :

TERME ANGLAIS	EN FRANCAIS	COURTE EXPLICATION :
ORF : Open Reading Frame	Cadre ouvert de lecture	Séquence commençant par un codon START et terminant par un STOP. Permet de rechercher les éventuelles séquences pouvant êtres des CDS
PRIMER	Amorce	Amorce pour débuter le travail d'un POLYMERASE dans les PCR
PROMOTER	Promoteur	Séquence présente sur l'ADN favorisant l'accroche de l'ARN POLYMERASE (Via Facteur σ) en amont du gène à transcrire.
TERMINATOR	Terminateur	Séquence présente sur l'ADN induisant un arrêt de la transcription (Rho dépendant : séquence de fixation de la prot Rho) ou (Rho indépendant : formation d'une structure TIGE/BOUCLE par l'ARN)
FEATURE	Fonctionnalité	En Français nous avons plus tendance à parler d'Annotation
CDS : Coding Sequence	Séquence Codante ☺	Séquence présente sur l'ADN et se traduisant concrètement par une séquence peptidique
Tm : melting Temperature	Température de Fusion	Température pour laquelle, la stabilité d'un double brin d'AxN est rompue, induisant une séparation de ces derniers.
HAIRPIN	Epingle à cheveux	Formation d'une structure TIGE / BOUCLE dans l'ARN (ici utilisée pour sa capacité à stopper la transcription chez les PROCARYOTES) https://fr.wikipedia.org/wiki/Tige-boucle
FOLDING	Pliage / pliant	Processus d'obtention d'une structure 2ndaire des peptides et aa (et 3aire aussi selon)

PASSONS A LA PRATIQUE :

- Une séquence ADN bactérienne vous est proposée.
- Vos missions :
  - Etudier et annoter la séquence : Ori / MCS /Promoter / ORF / CDS / Terminator.
  - Une fois les CDS supposées identifiées : identifier si possible les protéines correspondantes.
  - Effectuer une mise à jour des annotations suite à vos découvertes.
  - Une protéine doit résister à vos tentatives d'identifications, mais nous allons néanmoins persévérer et entreprendre un travail plus approfondi :
    - Préparer des primers pour amplifier le gène codant cette protéine, disposer de tels primers sera utile si nous envisageons de vérifier la présence de ce plasmide dans une bactérie en culture.
    - S'essayer à la prédiction IN SILICO : conformation, éventuelles fonctions biologiques, liaisons avec des ligans ….

Cet entrainement est destiné à des enseignants STL BGB, disposant des prérequis de biologie moléculaire. La séquence de travail pourrait être grandement simplifiée (une seule CDS, codant une protéine connue) par exemple

La séquence :

Organisme Hôte : *E. coli* Longueur : 2730 pb Séquence : circulaire

Sans précision supplémentaire : vous considérez que le brin présenté est le brin POSITIF donc orienté 5'➔3'

ATTENTION : séquence d'entrainement, vous devez trouver 3 ORF (mais 2 sont très courts 10 aa)

TTGACAATTAATCATCGGCTCGTATAATGTCCTGCATAAGGAGGTTCCCTTACTATGGCACAGGAAGAAGAAGCAGAACAGAATCTGAGCGAACTGAGCGGTCCGTGGCGTACC
GTTTATATTGGTAGCACCAATCCGGAAAAAATTCAGGAAAATGGTCCGTTTCGTACCTATTTTCGTGAACTGGTTTTTGATGATGAAAAAGGTACCGTTGATTTTTATTTTAGCGT
TAAACGTGATGGTAAATGGAAAAATGTTCATGTTAAAGCAACCAAACAGGATGATGGTACCTATGTTGCAGATTATGAAGGTCAGAATGTTTTTAAAATTGTTAGCCTGAGCCGT
ACCCATCTGGTTGCACATAATATTAATGTTGATAAACATGGTCAGACCACCGAACTGACCGGTCTGTTTGTTAAACTGAATGTTGAAGATGAAGATCTGGAAAAATTTTGGAAAC
TGACCGAAGATAAAGGTATTGATAAAAAAAATGTTGTTAATTTTCTGGAAAATGAAGATCATCCGCATCCGGAACCGCGTCAGACCGAAATTAATGAAGATGAAACCGAAAGCA
CCGATGAACTGGCATTTCAGCGTATGGCAACCATTCAGAATTGTCGTGAAACCGAAATTCTGGGTCTGAGCCCGGAAGAAATGGCAATGCTGAAACATGGTGGTTAATACTAAG
GAGGTTCACCTCCAATGAATCTGTATATTCAGTGGCTGAAAGATGGTGGTCCGAGCAGCGGTCGTCCGCCGCCGAGCTAATCCTAAGGAGGTTCACCTCAAATGGATGCAGAAT
TTCGTCATGATAGCGGTTATGAAGTTCATCATCAGAAACTGGTTTTTTTTGCAGAAGATGTTGGTAGCAATAAATAAGCTAGCGCTACCGGACTCAGATCTCGAGCTCAAGCTTC
GAATTCTGCAGTCGACGGTACCGCGGGCCCGGGATCCACTAGTGCAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGT
CCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGAAAACTAAAGCGCCACAAGGGCGCTTTAGTTTGTTTTGATTTTCAGCCTGATACAGATTAAATCAGAACG
CAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAATCAGACCAAGTTTACTCATATATACTTTTGACAATT
AATCATCGGCTCGTATAATGTCCTGCATAGATTGATTTACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGC
CCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGC
ACCTCGACCCCAAAAAACTTGATTTGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT
GTTCCAAACTGGAACAACACTCAACCCTATCTCGGGCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACG
CGAATTTTAACAAAATATTAACGTTTACAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGAC
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAG
CTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTA
CATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCT
GAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCATTGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAG
GCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGA
CTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACACGTT
CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAG
CGGAAGAGCGCCTGACGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTA

Vos outils à disposition pour la partie génomique :

- Du plus basique aux plus modernes :

  - Un simple traitement de texte : c'est très proche de la méthode « à l'ancienne », travail hautement respectable et courageux, où l'on recherche des séquences consensus à la main ou en utilisant la fonction rechercher. Vous pouvez essayer car cette séquence synthétique est construite à partir de ce type de briques.

  - Logiciel Serial Clonner : ancien (il reste très pratique car stable, il nécessite en revanche une petite mise à jour manuelle de ses bases de données de FEATURES pour réussir correctement à annoter la séquence de test. Pour la PCR il offre des fonctions simples (vous devez trouver vos amorces, c'est en un sens plus pédagogique).

  - **Le logiciel SnapGene Viewer : plus moderne, soyons franc, la séquence de test est conçue pour obtenir une annotation automatique via les bases de données de cet outil ☺. Pour la PCR vous devez là aussi choisir vos primers.**

  - Le logiciel Genome Compiler : actuel, sa fonction d'annotation automatique ne donne pas de résultats avec la séquence de test, il sera en revanche utilisable avantageusement pour la partie PCR, il offre la possibilité de déterminer automatiquement les meilleurs PRIMERS via « primer 3 », outil efficace de prédiction des Tm , des CG% , des Δ Tm, possible de lui autoriser des Mismatchs pour anticiper des amplifications non désirées.

- Un seul site : n'y voyez pas une expression de la paresse de l'auteur, mais simplement le choix d'un site de référence
  - https://blast.ncbi.nlm.nih.gov/Blast.cgi

Vos outils à disposition pour la partie protéomique :

- Le choix des armes logicielles :
  - Pymol : pour visualiser les structures protéiques déjà déterminées que l'on trouvera dans la PDB ou que l'on fera calculer IN SILICO. Ses avantages sont tellement nombreux que je vous invite à consulter le tutoriel qui lui est consacré.

- Le choix des armes sur le net :
  - La Protein Data Bank : source inextinguible de structures 3D et +
  - RPBS : Ressources Parisiennes en Bio-informatique Structurale : pas exactement un site, mais un portail, ouvrant sur une multitudes d'outils. Nous nous concentrerons uniquement sur l'outil PEP-FOLD
  - ZhangLab : Site de l'université du Michigan : portail donnant accès à une multitude d'outils, nous nous concentrerons uniquement sur I-TASSER
  - OPM : Site de l'université du Michigan (encore) : Orientations of Proteins in Membranes, le titre est suffisamment évocateur normalement ☺
  - NCBI Blast : pour identifier des séquences protéiques (ou nucléiques)

Mémo de la bonne PCR :

• Longueur des amorces :	18 à 24 bases
• Température d'hybridation des amorces Ta en °C:	Ta = Tm – 5
• Température de « décrochage » des amorces Tm en °C :	Tm = 2(A+T) + 4(G+C)
• Spécificité des amorces :	Séquences uniques si possible (d'où le 18 à 24 bases, au-delà hybridation moins efficace et impact sur la quantité d'amplicons obtenus)
• Complémentarité intra et inter amorces :	Formation de tiges / boucles si intra et risque de formation de dimère d'amorces si inter
• Teneur en G/C 45 à 55% et suites poly T ou poly A à éviter :	Risque de « glissement » dans l'hybridation
• Séquence à l'extrémité 3' :	L'accroche G/C est recommandée pour éviter la respiration de l'extrémité (G/C 3 liaisons H contre 2 pour A/T, cela accroche donc mieux.

Piste pour ETLV ?

Sources :

- https://rnbio.upmc.fr/bio-mol_pcr2
- https://www.jove.com/video/3998/polymerase-chain-reaction-protocole-basic-plus-stratgies-de-dpannage?language=French

Le résultat final attendu :

Cette diapositive combine les résultats obtenus en utilisant :

- Snapgene Viewer : pour la carte du plasmide et l'identification des ORF
- NCBI Blast : pour identifier les protéines codées par les ORF

- Genomecompiler : pour les amorces et la simulation PCR
- Pymol (et la PDB) pour la représentation 3D des PROT
- La structure de la protéine de la formation est prédite via ZhangLab (24 Heures !!!!),