TP 2 Implémentation de PLAST

Pierre Emery 20278920 Mehdi Qostali 20260322

Automne 2025

Question 1

Voici l'output de notre programme pour chacune des séquences du fichier `unknown.fasta` qui contient des séquences d'ARNt dont la nature est inconnue. On utilise les paramètres par défaut (`-ss = 0.001`, `-E = 4`, `seed = '11111111111'`).

Pour interpréter les sorties, il faut comprendre la signification de chaque bloc de résultat :

— Chaque bloc commence par une ligne d'en-tête de la forme `>AcideAminé`Anticodon|Espèce|. C'est l'identifiant de la séquence de la banque qui aligne le mieux avec la séquence inconnue.

— **Score brut** : score d'alignement S calculé avec notre matrice de scores (+5 match, -4 mismatch) après extension et fusion des HSP. Il dépend de la longueur et de la qualité de l'alignement.

— **Bitscore** : version normalisée du score brut, ce qui permet de comparer des HSP de longueurs différentes. Un bitscore plus élevé indique un alignement plus pertinent.

— **E-value** : nombre attendu, par hasard, d'alignements avec un bitscore au moins aussi élevé. Plus la e-value est petite, plus le hit est significatif. Seuls les hits avec `e-value` $\leq$ `-ss` (ici 10^{-3}) sont reportés.

— Les lignes `Query : ...` et `Sbjct : ...` montrent l'alignement local entre un segment de la séquence inconnue (`Query`) et la séquence de la banque (`Sbjct`), avec les positions de début et de fin.

— Pour une séquence inconnue donnée, les différents hits sont triés par pertinence (e-value croissante). Le *meilleur hit* est donc celui avec la plus petite e-value (et, en général, le plus grand bitscore).

Résultats

Pour chaque séquence inconnue, nous reportons ci-dessous l'output retourné par PLAST :

Séquence unknown 1

```
    X|???|Malus_domestica
AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT
```

Output 1

```
>M|cat|Carica_papaya
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Oryza_sativa_Japonica_Group
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
```

```
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Vitis_vinifera_2
# Best HSP score:251.00, bitscore:72.00, evalue: 4.91e-16
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCCCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Arabidopsis_thaliana
# Best HSP score:232.00, bitscore:67.00, evalue: 1.57e-14
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGC 72
22 ACTCATCAGGCTCATGACCTGAAGATTACAGGTTCGAATCCTGTCCCCGC 72
----------------------------------------
>M|cat|Bracteacoccus_minor_2
# Best HSP score:109.00, bitscore:33.00, evalue: 2.70e-04
12 GAATTGGTTTACTCATCAGGCTCATGACC 41
11 GTAGTGGTTAACTCATCGGGCTCATGACC 40
----------------------------------------
Total : 5
```

Séquence unknown 2

```
    X|???|Nephroselmis_olivacea
ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATGTA
```

Output 2

```
    >R|tcg|Marchantia_polymorpha
# Best HSP score:301.00, bitscore:86.00, evalue: 3.00e-20
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATG 72
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATG 72
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:188.00, bitscore:55.00, evalue: 6.43e-11
0 ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTG 43
0 ACATTCTTAGCTCAGTTGGATAGAGCAACGGCCTTCTAAGCTG 43
----------------------------------------
>R|tct|Marchantia_polymorpha
# Best HSP score:127.00, bitscore:38.00, evalue: 8.43e-06
1 CATCCTTAGCTCAGTAGGATAGAGCAACA 30
1 CATTCTTAGCTCAGTTGGATAGAGCAACA 30
----------------------------------------
Total : 3
```

Séquence unknown 3

```
    X|???|Phoenix_dactylifera
CGCGGAGTAGAGCAGTTTGGTAGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCATCCCTA
```

Output 3

```
    >P|tgg|Oryza_sativa_Japonica_Group
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
```

```
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
---------------------------------------
>P|tgg|Sorghum_bicolor
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
---------------------------------------
>P|tgg|Vitis_vinifera
# Best HSP score:132.00, bitscore:39.00, evalue: 4.22e-06
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCCAATCCTGTCATCCCTA 74
---------------------------------------
Total : 3
```

Séquence unknown 4

```
    X|???|Chara_vulgaris
GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGCT
```

Output 4

```
    >R|tct|Marchantia_polymorpha
# Best HSP score:347.00, bitscore:99.00, evalue: 3.66e-24
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGAATGC 73
---------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:338.00, bitscore:96.00, evalue: 2.93e-23
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATGC 73
---------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:167.00, bitscore:49.00, evalue: 4.12e-09
31 CCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAG 68
31 CCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAG 68
---------------------------------------
Total : 3
```

Question 2 – Déduction de la nature des séquences

Énoncé : En déduire, si possible, la nature (acide aminé et anticodon) de chacune des séquences. L'identifiant des séquences de la banque est de la forme `AcideAminé|Anticodon|Espèce`. Vous pouvez au besoin ajuster le seuil de significativité `-ss`.

Pour chaque séquence du fichier `unknown.fasta`, nous utilisons le meilleur hit retourné par PLAST (plus petit e-value, plus grand bitscore) pour déduire l'acide aminé et l'anticodon. L'identifiant des séquences de la banque a la forme `AcideAminé|Anticodon|Espèce`.

Séquence unknown 1 (`X|???|Malus_domestica`)

Les meilleurs hits sont :

```
>M|cat|Carica_papaya      (bitscore 75, e-value ~ 6.1e-17)
>M|cat|Oryza_sativa_...   (bitscore 75, e-value ~ 6.1e-17)
...
```

Ils partagent tous le même identifiant `M|cat|...`. On en conclut que :

— Acide aminé : M = méthionine

— Anticodon : `cat` (en ADN), soit CAU en ARN

La séquence unknown 1 correspond donc à un ARNt$^{\text{Met}}$ avec anticodon CAU.

Séquence unknown 2 (`X|???|Nephroselmis_olivacea`)

Le meilleur hit est :

```
>R|tcg|Marchantia_polymorpha   (bitscore 86, e-value ~ 3.0e-20)
```

Deux autres hits significatifs existent (`R|tct|...`) mais avec des scores et des e-values moins bons.

Nous en déduisons que :

— Acide aminé : R = arginine

— Anticodon le plus probable : `tcg` (ADN), soit UCG en ARN

La séquence unknown 2 est donc très probablement un ARNt$^{\text{Arg}}$ anticodon UCG.

Séquence unknown 3 (`X|???|Phoenix_dactylifera`)

Les meilleurs hits sont :

```
>P|tgg|Oryza_sativa_...   (bitscore 42, e-value ~ 5.3e-07)
>P|tgg|Sorghum_bicolor    (bitscore 42, e-value ~ 5.3e-07)
...
```

L'identifiant commence par `P|tgg|...`. On en déduit :

— Acide aminé : P = proline

— Anticodon : `tgg` (ADN), soit UGG en ARN

La séquence unknown 3 correspond à un ARNt$^{\text{Pro}}$ avec anticodon UGG.

Séquence unknown 4 (`X|???|Chara_vulgaris`)

Les deux meilleurs hits sont :

```
>R|tct|Marchantia_polymorpha   (bitscore 99, e-value ~ 3.7e-24)
>R|tcg|Marchantia_polymorpha   (bitscore 96, e-value ~ 2.9e-23)
```

Ils correspondent tous les deux à des ARNt$^{\text{Arg}}$, mais avec deux anticodons différents. Le meilleur hit (`R|tct|...`) a le score le plus élevé et la plus petite e-value.

Nous concluons donc :

— Acide aminé : R = arginine

— Anticodon le plus probable : `tct` (ADN), soit UCU en ARN

La séquence unknown 4 est donc très probablement un ARNt$^{\text{Arg}}$ anticodon UCU.

Question 3 – Comparaison avec BLASTN

Énoncé : Vérifiez vos résultats en vous servant du véritable outil BLASTN (NCBI). On vous demande de comparer les deux résultats.

unknown 1

Nous avons soumis la première séquence (`X|???|Malus_domestica`) à BLASTN (NCBI). Les résultats montrent plusieurs dizaines de hits ayant 100 % d'identité sur toute la longueur alignée, avec des e-values de l'ordre de 10^{-28}. Ces hits correspondent à des génomes mitochondriaux de plantes, en particulier de nombreuses entrées annotées *Malus domestica*, ce qui s'aligne bien avec l'entête d'unknown 1. En revanche, les espèces retrouvées par BLASTN (génomes mitochondriaux complets) ne sont pas les mêmes que celles de notre petite banque de tRNAs utilisée par PLAST.

unknown 2

Pour la deuxième séquence (`X|???|Nephroselmis_olivacea`), BLASTN retourne un hit principal ayant 100 % d'identité sur toute la longueur alignée avec une séquence de *Nephroselmis olivacea*, ce qui correspond directement à l'entête d'unknown 2. On observe également une série de hits chez *Marchantia polymorpha* avec environ 91,55 % d'identité et des e-values de l'ordre de 6×10^{-17}. Ces résultats sont cohérents avec notre outil PLAST, dont le meilleur hit significatif se trouvait justement dans des tRNAs de *Marchantia polymorpha*.

unknown 3

Pour la troisième séquence (`X|???|Phoenix_dactylifera`), BLASTN retourne plusieurs hits ayant 100 % d'identité sur toute la longueur alignée, tous annotés comme provenant de *Phoenix dactylifera*. Cela confirme l'entête d'unknown 3 et la nature « tRNA de plante » suggérée par PLAST. Comme pour unknown 1, les espèces exactes dans BLASTN ne coïncident pas avec celles de notre petite banque PLAST (par exemple *Oryza sativa*, *Sorghum bicolor*, *Vitis vinifera*), mais les alignements restent parfaitement cohérents avec l'idée d'un tRNA très conservé entre différentes espèces végétales.

unknown 4

Pour la dernière séquence (`X|???|Chara_vulgaris`), BLASTN retourne plusieurs hits ayant 100 % d'identité sur toute la longueur alignée, dont certains annotés comme *Chara vulgaris*. Cela correspond bien à l'entête d'unknown 4. Comme pour les autres séquences, les espèces de tRNAs trouvées par PLAST (basées sur notre banque restreinte) ne sont pas exactement les mêmes que celles listées par BLASTN, mais les niveaux d'identité et les e-values très faibles montrent que les deux outils pointent vers le même type de molécule (tRNA) dans des organismes phylogénétiquement cohérents.

Commentaires

Nous voulions finalement commenter les résultats de la question 3, qui peuvent paraître un peu décevants au premier abord. Les limites de notre implémentation viennent surtout du fait que notre banque de référence est très restreinte (un petit ensemble de tRNAs) alors que BLASTN interroge une base de données nucléotidique gigantesque.

BLASTN n'affiche que les 100 meilleurs alignements, triés par score et e-value. Ainsi, même si notre outil PLAST trouve un HSP très significatif (bitscore élevé, e-value très faible) vers une certaine espèce de notre petite banque, rien ne garantit que cette même espèce apparaisse dans le « top 100 » des hits BLASTN.

Le cas d'unknown 2 illustre bien ce point : notre PLAST trouve comme meilleur hit un tRNA de *Marchantia polymorpha*, et BLASTN retourne justement des alignements significatifs vers *Marchantia polymorpha* (en plus du hit parfait vers *Nephroselmis olivacea*, qui correspond à l'en-tête de la séquence inconnue). À l'inverse, pour unknown 4, PLAST donne un alignement

encore plus significatif (bitscore plus élevé, e-value plus faible) vers des tRNAs de *Marchantia polymorpha*, mais dans BLASTN les 100 meilleurs hits sont occupés par des génomes de *Chara vulgaris* et d'espèces très proches. On ne voit donc pas forcément *Marchantia* dans la liste, non pas parce que PLAST est « faux », mais parce que les bases de données et les jeux de références ne sont pas les mêmes.

Question 4 (Bonus) – Impact de la longueur de la graine

Énoncé : Qu'arrive-t-il lorsque vous utilisez des graines plus longues ou plus courtes (impact sur vitesse, précision, sensibilité) ?

Expériences réalisées

Pour analyser l'impact de la taille de la graine, nous avons relancé PLAST sur les quatre séquences « unknown » en faisant varier le paramètre `-seed`. Nous avons utilisé les valeurs suivantes :

- `-seed '11'`
- `-seed '11111111111'`
- `-seed '1111111111111111111'`

Observations

De manière qualitative, nous avons constaté les comportements suivants :

- Pour certains inconnus (par ex. `unknown 1` et `unknown 2`), les graines plus longues ($k = 11$ ou $k = 20$) permettent de retrouver des HSPs plus longs et mieux scorés (bitscores plus élevés, e-values plus faibles). Avec la graine très courte ($k = 2$), certains de ces meilleurs alignements globaux ne sont pas retrouvés.
- Pour `unknown 4`, la graine courte ($k = 2$) génère davantage de HSPs, dont plusieurs alignements secondaires avec des e-values proches du seuil de significativité. Lorsque la graine est plus longue, ces hits « bruités » disparaissent et seuls les alignements les plus pertinents restent.
- Pour `unknown 3`, les résultats sont pratiquement identiques pour les trois tailles de graine : la zone alignée est suffisamment conservée pour que toutes les graines testées mènent au même meilleur HSP.
- Du point de vue du temps d'exécution, la graine courte `'11'` prend environ 1,0–1,15 secondes par requête, alors que les graines plus longues `'11111111111'` et `'1111111111111111111'` tournent autour de 0,04–0,08 secondes. Les graines plus longues sont donc nettement plus rapides sur notre petite banque.

Analyse

- **Vitesse** : nos mesures montrent que la graine très courte (`'11'`) est d'un ordre de grandeur plus lente (environ 1 s) que les graines plus longues (environ 0,05 s). Cela confirme qu'une graine plus longue réduit fortement le nombre de hits initiaux à étendre.
- **Précision** : les graines plus longues favorisent des hits très spécifiques. On obtient alors moins de HSP, mais ceux-ci sont en général de meilleure qualité. Par contre, des alignements avec plus de mismatches répartis risquent de ne jamais être trouvés si aucun k-mer long exact ne se forme, ce qui réduit la sensibilité à des séquences très éloignées.

Question 5 (Bonus) – Graines de PatternHunter

Énoncé : Adapter l'algorithme aux graines espacées (avec positions "don't care") comme `111010010100110111` et comparer les performances.

Adaptation du code

Nous avons généralisé les fonctions `get_kmers` et `find_seed_hits` pour gérer des graines espacées de type PatternHunter

Expérience réalisée

Pour comparer l'algorithme original et la version PatternHunter, nous avons lancé PLAST sur les quatre séquences `unknown 1-4` avec les mêmes paramètres d'extension (`-E 4`) et de significativité (`-ss 1e-3`), en ne changeant que la graine :

— graine normale : `"11111111111"` ;

— graine espacée PatternHunter : `"1101001101110111"`.

Pour chaque configuration, nous avons observé la liste des hits, leurs scores bruts, bitscores, e-values, les alignements produits puis le temps d'exécution.

Observations

Nous avons constaté :

— **unknown 1** : les deux graines retrouvent les mêmes espèces (par exemple *Carica papaya*, *Oryza sativa*, *Vitis vinifera*, *Arabidopsis thaliana*), mais la graine espacée déclenche un HSP qui s'étend sur pratiquement toute la séquence et donne un score brut et un bitscore plus élevés, avec une e-value plus faible. On voit également apparaître quelques hits supplémentaires de score intermédiaire.

— **unknown 2** : avec la graine contiguë, le meilleur hit est *Marchantia polymorpha*, suivi de *Mesostigma viride*. Avec la graine espacée, le classement peut s'inverser : le HSP sur *Mesostigma* devient légèrement meilleur, ce qui montre que la graine espacée est plus sensible à certaines variantes de l'alignement.

— **unknown 3** : les résultats sont identiques pour les deux graines (mêmes espèces, mêmes scores). Quand l'alignement est déjà très conservé, le type de graine ne change pratiquement rien.

— **unknown 4** : les deux variantes retrouvent les mêmes HSP forts sur *Marchantia polymorpha* et *Mesostigma viride*, mais la graine espacée détecte en plus quelques HSP de plus faible score sur d'autres tRNA apparentés.

— Les temps d'exécution restent du même ordre pour les deux graines (environ 0,05 secondes par requête), sans différence nette de performance en temps sur notre petite banque de tRNA.

Analyse

En résumé, l'utilisation de graines espacées de type PatternHunter a les effets suivants :

— **Impact sur le nombre de hits** : le nombre total de HSP a tendance à augmenter légèrement. On retrouve toujours les hits de très bon score, mais on voit apparaître davantage de HSP intermédiaires ou faibles, notamment pour des séquences apparentées (même espèce ou espèces proches).

— **Impact sur la sensibilité** : les graines espacées sont plus sensibles, car elles permettent de déclencher un hit même si les identités sont réparties sur une région plus longue avec quelques différences. Elles détectent donc des similarités plus “dégénérées”, au prix d’un peu plus de bruit et d’un classement des hits qui peut légèrement changer.

— **Temps de calcul** : d’après nos mesures, les temps d’exécution pour la graine contiguë et la graine espacée sont très proches (tous autour de 0,05 s), ce qui confirme que, sur une petite banque de tRNA, l’impact des graines espacées sur le temps de calcul est négligeable par rapport à l’impact sur la sensibilité.

Annexe A – Sorties détaillées pour la question 4

Dans cette annexe, nous regroupons les sorties de PLAST utilisées pour analyser l’impact de la longueur de la graine (question 4). Mêmes si les sorties pour la graine par défaut `'11111111111'` sont déjà présentées dans l’énoncé de la question 1 nous les avons remis dans cette annexe puisque nous avons ajouté le temps d’exécution à la sortie. Nous ajoutons en plus les résultats pour les graines `'11'` et `'1111111111111111111'`.

Graine '11'

Séquence unknown 1 (X|???|Malus_domestica)

```
$ python plast.py -i AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCC

>M|cat|Vitis_vinifera_2
# Best HSP score:251.00, bitscore:72.00, evalue: 4.91e-16
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCCCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Arabidopsis_thaliana
# Best HSP score:232.00, bitscore:67.00, evalue: 1.57e-14
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGC 72
22 ACTCATCAGGCTCATGACCTGAAGATTACAGGTTCGAATCCTGTCCCCGC 72
----------------------------------------
Total : 2
Temps d'exécution : 1.1390 secondes
```

Séquence unknown 2 (X|???|Nephroselmis_olivacea)

```
$ python plast.py -i ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Mesostigma_viride
# Best HSP score:188.00, bitscore:55.00, evalue: 6.43e-11
0 ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTG 43
0 ACATTCTTAGCTCAGTTGGATAGAGCAACGGCCTTCTAAGCTG 43
----------------------------------------
>R|tct|Marchantia_polymorpha
# Best HSP score:174.00, bitscore:51.00, evalue: 1.03e-09
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTG 43
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTG 43
----------------------------------------
>R|acg|Mesostigma_viride
```

```
# Best HSP score:117.00, bitscore:35.00, evalue: 6.75e-05
7 TAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTG 43
7 TAGTTCAATAGGATAGAGCATCAGACTACGAATCTG 43
----------------------------------------
Total : 3
Temps d'exécution : 1.1528 secondes
```

Séquence unknown 3 (X|???|Phoenix_dactylifera)

```
$ python plast.py -i CGCGGAGTAGAGCAGTTTGGTAGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCATCC

>P|tgg|Oryza_sativa_Japonica_Group
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Sorghum_bicolor
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Vitis_vinifera
# Best HSP score:132.00, bitscore:39.00, evalue: 4.22e-06
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCCAATCCTGTCATCCCTA 74
----------------------------------------
Total : 3
Temps d'exécution : 1.0511 secondes
```

Séquence unknown 4 (X|???|Chara_vulgaris)

```
$ python plast.py -i GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Marchantia_polymorpha
# Best HSP score:343.00, bitscore:98.00, evalue: 7.31e-24
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGCT 74
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGAATGCG 74
----------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:334.00, bitscore:95.00, evalue: 5.85e-23
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGCT 74
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATGCG 74
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:178.00, bitscore:52.00, evalue: 5.15e-10
31 CCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATG 72
31 CCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAGAATG 72
----------------------------------------
>K|ttt|Marchantia_polymorpha
# Best HSP score:120.00, bitscore:36.00, evalue: 3.37e-05
```

```
32 CTTCTAAGTTGAAGGTCACAGGTTCAAATCCTG 65
31 CTTTTAACTTAAAGGTCGCAGGTTCAAGTCCTG 64
----------------------------------------
>K|ttt|Bracteacoccus_minor
# Best HSP score:106.00, bitscore:32.00, evalue: 5.40e-04
16 TGGATAGAGCAACAACCTTCTAAGTTGAAGGT 48
15 TCGGTCGAGCAACAAGCTTTTAACTTGAAGGT 47
----------------------------------------
Total : 5
Temps d'exécution : 0.9747 secondes
```

Graine '11111111111'

Séquence unknown 1 (X|???|Malus_domestica)

```
$ python plast.py -i AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCC

>M|cat|Carica_papaya
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Oryza_sativa_Japonica_Group
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Vitis_vinifera_2
# Best HSP score:251.00, bitscore:72.00, evalue: 4.91e-16
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCCCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Arabidopsis_thaliana
# Best HSP score:232.00, bitscore:67.00, evalue: 1.57e-14
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGC 72
22 ACTCATCAGGCTCATGACCTGAAGATTACAGGTTCGAATCCTGTCCCCGC 72
----------------------------------------
>M|cat|Bracteacoccus_minor_2
# Best HSP score:109.00, bitscore:33.00, evalue: 2.70e-04
12 GAATTGGTTTACTCATCAGGCTCATGACC 41
11 GTAGTGGTTAACTCATCGGGCTCATGACC 40
----------------------------------------
Total : 5
Temps d'exécution : 0.0625 secondes
```

Séquence unknown 2 (X|???|Nephroselmis_olivacea)

```
$ python plast.py -i ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGAT

>R|tcg|Marchantia_polymorpha
# Best HSP score:301.00, bitscore:86.00, evalue: 3.00e-20
```

```
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATG 72
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATG 72
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:188.00, bitscore:55.00, evalue: 6.43e-11
0 ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTG 43
0 ACATTCTTAGCTCAGTTGGATAGAGCAACGGCCTTCTAAGCTG 43
----------------------------------------
>R|tct|Marchantia_polymorpha
# Best HSP score:127.00, bitscore:38.00, evalue: 8.43e-06
1 CATCCTTAGCTCAGTAGGATAGAGCAACA 30
1 CATTCTTAGCTCAGTTGGATAGAGCAACA 30
----------------------------------------
Total : 3
Temps d'exécution : 0.0450 secondes
```

Séquence unknown 3 (`X|???|Phoenix_dactylifera`)

```
$ python plast.py -i CGCGGAGTAGAGCAGTTTGGTAGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCATCC

>P|tgg|Oryza_sativa_Japonica_Group
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Sorghum_bicolor
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Vitis_vinifera
# Best HSP score:132.00, bitscore:39.00, evalue: 4.22e-06
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCCAATCCTGTCATCCCTA 74
----------------------------------------
Total : 3
Temps d'exécution : 0.0425 secondes
```

Séquence unknown 4 (`X|???|Chara_vulgaris`)

```
$ python plast.py -i GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Marchantia_polymorpha
# Best HSP score:347.00, bitscore:99.00, evalue: 3.66e-24
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGAATGC 73
----------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:338.00, bitscore:96.00, evalue: 2.93e-23
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATGC 73
```

```
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:167.00, bitscore:49.00, evalue: 4.12e-09
31 CCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAG 68
31 CCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAG 68
----------------------------------------
Total : 3
Temps d'exécution : 0.0513 secondes
```

Graine '11111111111111111111'

Séquence unknown 1 (X|???|Malus_domestica)

```
$ python plast.py -i AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCC

>M|cat|Carica_papaya
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Oryza_sativa_Japonica_Group
# Best HSP score:260.00, bitscore:75.00, evalue: 6.14e-17
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Vitis_vinifera_2
# Best HSP score:251.00, bitscore:72.00, evalue: 4.91e-16
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
22 ACTCATCAGGCCCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Arabidopsis_thaliana
# Best HSP score:232.00, bitscore:67.00, evalue: 1.57e-14
22 ACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGC 72
22 ACTCATCAGGCTCATGACCTGAAGATTACAGGTTCGAATCCTGTCCCCGC 72
----------------------------------------
Total : 4
Temps d'exécution : 0.0761 secondes
```

Séquence unknown 2 (X|???|Nephroselmis_olivacea)

```
$ python plast.py -i ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGAT

>R|tcg|Marchantia_polymorpha
# Best HSP score:301.00, bitscore:86.00, evalue: 3.00e-20
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATG 72
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATG 72
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:115.00, bitscore:34.00, evalue: 1.35e-04
45 GGTCACAGGTTCAAATCCTGTAG 68
45 GGTCACAGGTTCAAATCCTGTAG 68
----------------------------------------
```

```
>R|tct|Marchantia_polymorpha
# Best HSP score:115.00, bitscore:34.00, evalue: 1.35e-04
45 GGTCACAGGTTCAAATCCTGTAG 68
45 GGTCACAGGTTCAAATCCTGTAG 68
----------------------------------------
Total : 3
Temps d'exécution : 0.0541 secondes
```

Séquence unknown 3 (X|???|Phoenix_dactylifera)

```
$ python plast.py -i CGCGGAGTAGAGCAGTTTGGTAGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCATCC

>P|tgg|Oryza_sativa_Japonica_Group
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Sorghum_bicolor
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
Total : 2
Temps d'exécution : 0.0546 secondes
```

Séquence unknown 4 (X|???|Chara_vulgaris)

```
$ python plast.py -i GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Marchantia_polymorpha
# Best HSP score:347.00, bitscore:99.00, evalue: 3.66e-24
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGAATGC 73
----------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:338.00, bitscore:96.00, evalue: 2.93e-23
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATGC 73
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:167.00, bitscore:49.00, evalue: 4.12e-09
31 CCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAG 68
31 CCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAG 68
----------------------------------------
Total : 3
Temps d'exécution : 0.0632 secondes
```

Annexe B – Sorties détaillées pour la question 5 (PatternHunter)

Dans cette annexe, nous présentons les sorties de PLAST obtenues avec la graine espacée de type PatternHunter `'1101001101110111'` utilisée pour la question 5.

Séquence unknown 1 (X|???|Malus_domestica)

```
$ python plast.py -i AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCC

>M|cat|Carica_papaya
# Best HSP score:352.00, bitscore:100.00, evalue: 1.83e-24
0 AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
0 AGCGGGGTAGAGGAATTGGTCGACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Oryza_sativa_Japonica_Group
# Best HSP score:352.00, bitscore:100.00, evalue: 1.83e-24
0 AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
0 AGCGGGGTAGAGGAATTGGTCAACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Vitis_vinifera_2
# Best HSP score:343.00, bitscore:98.00, evalue: 7.31e-24
0 AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
0 AGCGGGGTAGAGGAATTGGTCGACTCATCAGGCCCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
----------------------------------------
>M|cat|Arabidopsis_thaliana
# Best HSP score:325.00, bitscore:93.00, evalue: 2.34e-22
0 AGCGGGGTAGAGGAATTGGTTTACTCATCAGGCTCATGACCTGAAGACTGCAGGTTCGAATCCTGTCCCCGCCT 74
0 AGCGGGGTAGAGGAATTGGTCAACTCATCAGGCTCATGACCTGAAGATTACAGGTTCGAATCCTGTCCCCGCAT 74
----------------------------------------
>M|cat|Bracteacoccus_minor_2
# Best HSP score:109.00, bitscore:33.00, evalue: 2.70e-04
12 GAATTGGTTTACTCATCAGGCTCATGACC 41
11 GTAGTGGTTAACTCATCGGGCTCATGACC 40
----------------------------------------
>M|cat|Neochloris_aquatica
# Best HSP score:105.00, bitscore:32.00, evalue: 5.40e-04
7 TAGAGGAATTGGTTTACTCATCAGGCTCAT 37
7 TAGAGCAATTGGTTAGCTTATCGGGCTCAT 37
----------------------------------------
Total : 6
Temps d'exécution : 0.0491 secondes
```

Séquence unknown 2 (X|???|Nephroselmis_olivacea)

```
$ python plast.py -i ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Mesostigma_viride
# Best HSP score:316.00, bitscore:90.00, evalue: 1.87e-21
0 ACATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATGTA 74
0 ACATTCTTAGCTCAGTTGGATAGAGCAACGGCCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAGAATGTA 74
----------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:301.00, bitscore:86.00, evalue: 3.00e-20
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAGCTGGTGGTCACAGGTTCAAATCCTGTAGGATG 72
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATG 72
----------------------------------------
>R|tct|Marchantia_polymorpha
```

```
# Best HSP score:168.00, bitscore:49.00, evalue: 4.12e-09
1 CATCCTTAGCTCAGTAGGATAGAGCAACAGCCTTCTAAG 40
1 CATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAG 40
----------------------------------------
Total : 3
Temps d'exécution : 0.0460 secondes
```

Séquence unknown 3 (X|???|Phoenix_dactylifera)

```
$ python plast.py -i CGCGGAGTAGAGCAGTTTGGTAGCTCGCAAGGCTCATAACCTTGAGGTCACGGGTTCAAATCCTGTCATCC

>P|tgg|Oryza_sativa_Japonica_Group
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Sorghum_bicolor
# Best HSP score:141.00, bitscore:42.00, evalue: 5.27e-07
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCAAATCCTGTCATCCCTA 74
----------------------------------------
>P|tgg|Vitis_vinifera
# Best HSP score:132.00, bitscore:39.00, evalue: 4.22e-06
44 AGGTCACGGGTTCAAATCCTGTCATCCCTA 74
44 ATGTCACGGGTTCCAATCCTGTCATCCCTA 74
----------------------------------------
Total : 3
Temps d'exécution : 0.0514 secondes
```

Séquence unknown 4 (X|???|Chara_vulgaris)

```
$ python plast.py -i GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGAT

>R|tct|Marchantia_polymorpha
# Best HSP score:347.00, bitscore:99.00, evalue: 3.66e-24
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGAATGC 73
----------------------------------------
>R|tcg|Marchantia_polymorpha
# Best HSP score:338.00, bitscore:96.00, evalue: 2.93e-23
0 GCATTCTTAGCTCAGCTGGATAGAGCAACAACCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATGC 73
0 GCATTCTTAGCTCAGTTGGATAGAGCAACAACCTTCGAAGTTGATGGTCACAGGTTCAAATCCTGTAGGATGC 73
----------------------------------------
>R|tct|Mesostigma_viride
# Best HSP score:178.00, bitscore:52.00, evalue: 5.15e-10
31 CCTTCTAAGTTGAAGGTCACAGGTTCAAATCCTGTAGGATG 72
31 CCTTCTAAGCTGTAGGTCACAGGTTCAAATCCTGTAGAATG 72
----------------------------------------
>K|ttt|Marchantia_polymorpha
# Best HSP score:120.00, bitscore:36.00, evalue: 3.37e-05
32 CTTCTAAGTTGAAGGTCACAGGTTCAAATCCTG 65
31 CTTTTAACTTAAAGGTCGCAGGTTCAAGTCCTG 64
```

```
----------------------------------------
>K|ttt|Bracteacoccus_minor
# Best HSP score:106.00, bitscore:32.00, evalue: 5.40e-04
16 TGGATAGAGCAACAACCTTCTAAGTTGAAGGT 48
15 TCGGTCGAGCAACAAGCTTTTAACTTGAAGGT 47
----------------------------------------
Total : 5
Temps d'exécution : 0.0575 secondes
```