

MAPPI

5 tâches,

- ▶ *Tâche 1 : Nouvelles structures d'index pour la recherche de motifs approchés*
- ▶ *Tâche 2 : Mapping pour la métagenomique et la métatranscriptomique*
- ▶ *Tâche 3 : Outils d'assemblage pour les NGS*
- ▶ *Tâche 4 : Assemblage guidé de données de métatranscriptomique*
- ▶ *Tâche 5 : Pipeline bioinformatique*

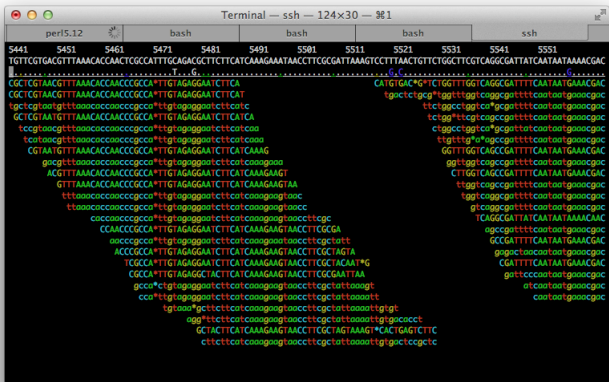
MAPPI

5 tâches, celles que je vais décrire dans le contexte Lillois

- ▶ *Tâche 1 : Nouvelles structures d'index pour la recherche de motifs approchés*
- ▶ *Tâche 2 : Mapping pour la métagenomique et la métatranscriptomique*
- ▶ *Tâche 3 : Outils d'assemblage pour les NGS*
- ▶ *Tâche 4 : Assemblage guidé de données de métatranscriptomique*
- ▶ *Tâche 5 : Pipeline bioinformatique*

Tâche 1 : Nouvelles structures d'index pour la recherche de motifs approchés

Contexte : *Read Mapping*



```
Terminal — ssh — 124x30 — 381
perl5.12  bash  bash  bash  ssh
5441 5451 5461 5471 5481 5491 5501 5511 5521 5531 5541 5551
TGTTCTGTGACGTTTAAACACCACTCCGCATTTCAGACGCTTCTTCATCAAGAATAAACCTCCGGATTAAGTCCCTTAACTGTCTGGCTCTGTCAAGCGATTACAATAAAMACGAC
.....T.....G.....C.....G.....
CGCTCGTAACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCACATGTGAC*G*TCGGTTGGTCAGCGGATTTCAATAATGAACGAC
CGCTCGTAACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATtgaactcgg*tgatttagtcagcgattttcaataatgaacgac
tgcctgtaagtgtttaaacacccaaccccgcca*ttgttagggaaacttcatcctctcagccctgac*gcgattttcaataatgaacgac
GCCTGTAATGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATCActgga*ttcagccctgac*gcgattttcaataatgaacgac
tccgtaacggttttaaacacccaaccccgcca*ttgttagggaaacttcatcctcctcagccctgac*gcgattttcaataatgaacgac
tctatcagcgttttaaacacccaaccccgcca*ttgttagggaaacttcatcctcctcagccctgac*ttgittg*agccgattttcaataatgaacgac
CGTAATGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATCAAGGGTTTGGTCAGCGGATTTCAATAATGAACGAC
gpcgttttaaacacccaaccccgcca*ttgttagggaaacttcatcctcctcaggaagggattagtcagccgattttcaataatgaacgac
ACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATCAAGAAGTCTTTGGTCAGCGGATTTCAATAATGAACGAC
GTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATCAAGAAGTAAttgatcagccgattttcaataatgaacgac
ttaaaccacccaaccccgcca*ttgttagggaaacttcatcctcctcaggaagtaacgtgagcagccgattttcaataatgaacgac
ttaaaccacccaaccccgcca*ttgttagggaaacttcatcctcctcaggaagtaaccctcagccagccgattttcaataatgaacgac
caccaccccgcca*ttgttagggaaacttcatcctcctcaggaagtaaccctcagccagccagccgattttcaataatgaacgac
CCAACCCCGCCA*TTGTAGAGGAATCTTCATCAAGAAGTAACTTCGGAAgcccattttcaataatgaacgac
aaccccgcca*ttgttagggaaacttcatcctcctcaggaagtaaccctcagctattGCCGATTTTCAATAATGAACGAC
ACCCCGCCA*TTGTAGAGGAATCTTCATCAAGAAGTAACTTCCTAGTAgagcacaataatgaacgac
TCGCCA*TTGTAGAGGAATCTTCATCAAGAAGTAACTTCCTACAAT*GCGATTTCAATAATGAACGAC
CGCCA*TTGTAGAGGCTACTTCATCAAGAAGTAACTTCGGAAATAAgattcccaataatgaacgac
gcca*ctgttagggaaacttcatcctcctcaggaagtaaccctcagctattaaagttatcagccgattttcaataatgaacgac
cca*ttgttagggaaacttcatcctcctcaggaagtaaccctcagctattaaagttatcagccgattttcaataatgaacgac
tgtaaagcttctcctcctcaggaagtaaccctcagctattaaagttgtgattcaataatgaacgac
agg*ttctcctcctcctcaggaagtaaccctcagctattaaagttgacacctcaataatgaacgac
GCTACTTCATCAAGAAGTAACTTCCTAGTAAAGT*CACTGAGTCTTC
cttctcctcctcctcaggaagtaaccctcagctattaaagttgactcagctc
```

Tâche 1 : Nouvelles structures d'index pour la recherche de motifs approchés

Contexte : *Read Mapping*

The image shows a terminal window titled 'Terminal — ssh — 124x30 — 381'. The terminal displays a read mapping visualization. At the top, several genomic coordinates are listed: 5441, 5451, 5461, 5471, 5481, 5491, 5501, 5511, 5521, 5531, 5541, and 5551. Below these, a reference sequence is shown: TGTTCTGTGACGTTTAAACACCACTTTCAGACGCTTCTTCATCAAGAAATAAACCCTTCGGATTAAAGTCTTTAACTGTCTGGCTCTGTCCAGGCATTATCAATTAATAAACGAC. The visualization consists of multiple lines of colored dots and dashes representing mismatches and matches between the read and the reference. The colors used are green, red, and black. The dots are arranged in a grid-like pattern, with some lines showing more mismatches than others. The terminal also shows the prompt 'perl5.12' and several 'bash' shells.

Tâche 1 : Nouvelles structures d'index pour la recherche de motifs approchés

Contexte : *Read Mapping*

Réalisé : 1. Portage de l'algorithme de Wu-Mamber sur GPU

[Bit-Parallel Multiple Pattern Matching. T. T. Tran, M. Giraud, J.-S. Varré PPAM / PBC 2011.]

2. Indexation des voisinages des k -mers

But : profiter de l'efficacité du cache GPU/Processeur

Deux méthodes d'indexation envisagées :

- ▶ indexation directe (tri des mots → recherche dichotomique)
- ▶ hachage parfait

+ non encore publié mais des résultats :

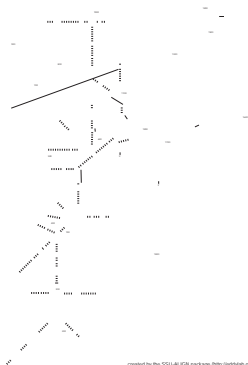
- ▶ mise en oeuvre en OpenCL (fonctionnelle sur CPU et GPU)
- ▶ gain en performance entre x10 et x60
- ▶ prototype de readmapper en cours

Tâche 4 : Assemblage guidé de données de métatranscriptomique

Contexte : identification d'ARN ribosomiques (16S/18S,23S/28S...)

- Buts :
- ▶ élimination
 - ▶ classification

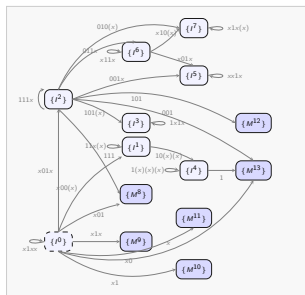
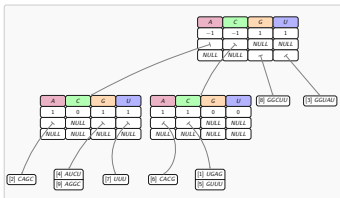
? : Problème nouveau sur données de *métatranscriptomique*



Tâche 4 : Assemblage guidé de données de métatranscriptomique

Contexte : identification d'ARN ribosomiques

- Réalisé :
- ▶ conception d'un filtre efficace pour la sélection des familles d'ARNr (SortMeRNA)
 - ▶ travail basé sur le *Burst Trie* et *l'automate de Levenstein*



Tâche 4 : Assemblage guidé de données de métatranscriptomique

Contexte : identification d'ARN ribosomiques

- En cours :
- ▶ communication aux *London Stringology Days*
 - ▶ publication en cours de soumission
 - ▶ séjour prévue au Génomoscope pour la transition
fin Tâche 4 / début Tâche 2, 10-13 avril