

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

The terminal window shows a read mapping exercise. The reference genome is represented by a sequence of 'A', 'C', 'G', and 'T' characters. A read is shown below it, with some characters in red. The read contains several mismatches compared to the reference. The terminal title is 'Terminal — ssh — 124x30 — 381'.

```
perl5.12  bash  bash  bash  ssh
5441 5451 5461 5471 5481 5491 5501 5511 5521 5531 5541 5551
TGTTCGTGACGTTTAAACACCACTCCGCATTTCAGACGCTTCTTCATCAAGAATAAACCTCCGGATTAAGCTCTTAACTGTCTGGCTCTGTCAGGGCGATTACAATAAAMACGAC
.....T.....G.....C.....G.....
CGCTCGTAACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCACATGTGAC*G*TCGGTTGGTCAGGGCGATTTCAATAATGAACGAC
CGCTCGTAACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCAT      tgaactcgg*tgatttagtcagcgattttcaataatgaacgac
tgcctgtaagtttttaaacacccaaccccgcca*ttgttagggaaacttcatca      tctcagccctggtca*gcgattttcaataatgaacgac
GCCTGTAATGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCATCA      tctggt*ttcgtcagcgattttcaataatgaacgac
tccgtatcgttttaaacacccaaccccgcca*ttgttagggaaacttcatca      ctggcctggtca*gcgatttcaataatgaacgac
tctatcgttttaaacacccaaccccgcca*ttgttagggaaacttcatca      ttgttt*a*agccattttcaataatgaacgac
CGTAATGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCACAAG      GTTTTGGTCAGGGCGATTTCAATAATGAACGAC
ggccttttaaacacccaaccccgcca*ttgttagggaaacttcatcaagaa      ggttagtccagcgattttcaataatgaacgac
ACGTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCACAAGAAGT      CTGGTCAGGGCGATTTCAATAATGAACGAC
GTTTAAACACCAACCCCGCCA*TTGTAGAGGAATCTTCACAAGAAGTAA      ttggtcagcgattttcaataatgaacgac
tttaaacacccaaccccgcca*ttgttagggaaacttcatcaagaaagtaac      ttgtcagcgattttcaataatgaacgac
tttaaacacccaaccccgcca*ttgttagggaaacttcatcaagaaagtaacc      ctgagcgattttcaataatgaacgac
caaacacccaaccccgcca*ttgttagggaaacttcatcaagaaagtaacctcgc      TCAGGGGATTAACAATAAAMACGAC
CCAACCCCGCCA*TTGTAGAGGAATCTTCACAAGAAGTAACTTCGGGA      agccgattttcaataatgaacgac
aaaccccgcca*ttgttagggaaacttcatcaagaaagtaacctcgcatt      GCCGATTTCAATAATGAACGAC
ACCCCGCCA*TTGTAGAGGAATCTTCACAAGAAGTAACTTCCTAGTA      gtagcacaataatgaacgac
TCGCCA*TTGTAGAGGAATCTTCACAAGAAGTAACTTCCTACAAT*G      CGATTTCAATAATGAACGAC
CGCCA*TTGTAGAGGCTACTTCACAAGAAGTAACTTCGGGAATTA      gattcccaataatgaacgac
gcca*ctgttagggaaacttcatcaagaaagtaacctcgcattaaagt      ctcaataatgaacgac
cca*ttgttagggaaacttcatcaagaaagtaacctcgcattaaagatt      caataatgaacgac
tgtaaagcttctcatcaagaaagtaacctcgcattaaagattgtgt
agg*ttctctcatcaagaaagtaacctcgcattaaagattgacacct
GCTACTTCACAAGAAGTAACTTCCTAGTAAAGT*CACTGAGCTTC
cttctcatcaagaaagtaacctcgcattaaagattgactcgcctc
```


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 :

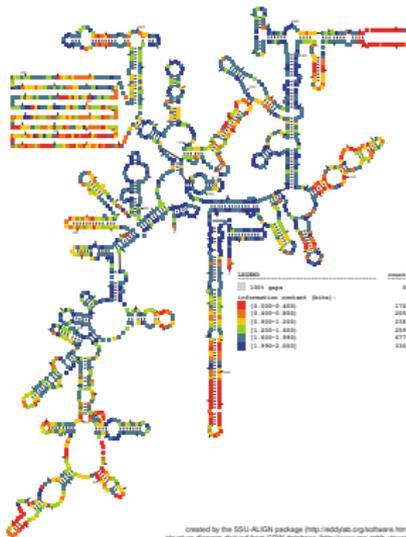
- ▶ mise en oeuvre en OpenCL (fonctionnelle sur CPU et GPU)
- ▶ gain en performance entre x10 et x60
- ▶ prototype de readmapper en cours

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

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