Stage de Maîtrise IGIS ITA

K-mers et k-mers espacés

Gabriel Toublanc

26 janvier 2017

Université de Rouen, U.F.R des Sciences et Techniques de Saint-Etienne-du-Rouvray, Equipe LITIS TIBS

Introduction

Introduction

Séquençage ADN :

• Lectures courtes (2005)

Introduction

Séquençage ADN:

- Lectures courtes (2005)
- Lectures longues (2010)

Comptage des k-mers

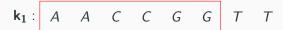
Les k-mers sont des facteurs de séquences ADN.

Les k-mers sont des facteurs de séquences ADN.

Ex: Avec la séquence x = AACCGGTT, on obtient les k-mers de taille 6 (6-mers) suivants :

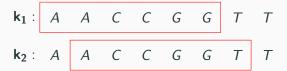
Les k-mers sont des facteurs de séquences ADN.

Ex: Avec la séquence x = AACCGGTT, on obtient les k-mers de taille 6 (6-mers) suivants :



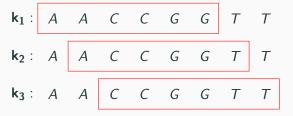
Les k-mers sont des facteurs de séquences ADN.

Ex: Avec la séquence x = AACCGGTT, on obtient les k-mers de taille 6 (6-mers) suivants :



Les k-mers sont des facteurs de séquences ADN.

Ex: Avec la séquence x = AACCGGTT, on obtient les k-mers de taille 6 (6-mers) suivants :



2

Utilisé afin de simuler les erreurs d'insertions sur les lectures longues.

Utilisé afin de simuler les erreurs d'insertions sur les lectures longues.

Ex : Avec la séquence x = AACCGGTT et le motif m = 111011, on obtient les 5-mers espacés suivants :

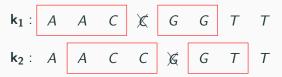
Utilisé afin de simuler les erreurs d'insertions sur les lectures longues.

Ex : Avec la séquence x = AACCGGTT et le motif m = 111011, on obtient les 5-mers espacés suivants :

k₁: A A C X G G T T

Utilisé afin de simuler les erreurs d'insertions sur les lectures longues.

Ex : Avec la séquence x = AACCGGTT et le motif m = 111011, on obtient les 5-mers espacés suivants :



Utilisé afin de simuler les erreurs d'insertions sur les lectures longues.

Ex : Avec la séquence x = AACCGGTT et le motif m = 111011, on obtient les 5-mers espacés suivants :



Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(\mathbf{L} - \mathbf{k} + \mathbf{1}) * \mathbf{4}^t$ k-mers espacés à insertion possibles.

Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(\mathbf{L} - \mathbf{k} + \mathbf{1}) * \mathbf{4}^{\mathbf{t}}$ k-mers espacés à insertion possibles.

Ex : Avec la séquence $\mathbf{x} = \mathbf{AACCGGTT}$ et le motif $\mathbf{m} = \mathbf{111011}$, on obtient les 6-mers espacés suivants :

Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(\mathbf{L} - \mathbf{k} + \mathbf{1}) * \mathbf{4}^{\mathbf{t}}$ k-mers espacés à insertion possibles.

Ex : Avec la séquence $\mathbf{x} = \mathbf{AACCGGTT}$ et le motif $\mathbf{m} = \mathbf{111011}$, on obtient les 6-mers espacés suivants :

 $\mathbf{k_1}: A \quad A \quad C \xrightarrow{A, C, G, T} C \quad G \quad G \quad T \quad T$

Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(\mathbf{L} - \mathbf{k} + \mathbf{1}) * \mathbf{4}^{\mathbf{t}}$ k-mers espacés à insertion possibles.

 $\mbox{\bf Ex}$: Avec la séquence $\mbox{\bf x} = \mbox{\bf AACCGGTT}$ et le motif $\mbox{\bf m} = 111011$, on obtient les 6-mers espacés suivants :

$$\mathbf{k_1}: A \quad A \quad C \xrightarrow{A, C, G, T} C \quad G \quad G \quad T \quad T$$
 $\mathbf{k_2}: A \quad A \quad C \quad C \xrightarrow{A, C, G, T} G \quad G \quad T \quad T$

Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(\mathbf{L} - \mathbf{k} + \mathbf{1}) * \mathbf{4}^{\mathbf{t}}$ k-mers espacés à insertion possibles.

 $\mbox{\bf Ex}$: Avec la séquence $\mbox{\bf x} = \mbox{\bf AACCGGTT}$ et le motif m=111011, on obtient les 6-mers espacés suivants :



Utilisé afin de simuler les erreurs de délétion sur les lectures longues.

 $(L-k+1)*4^t$ k-mers espacés à insertion possibles.

Ex : Avec la séquence $\mathbf{x} = \mathbf{AACCGGTT}$ et le motif $\mathbf{m} = \mathbf{111011}$, on obtient les 6-mers espacés suivants :

| k ₁ : | Α | Α | C A | , C, G, T | С | G | G | T | T |
|-------------------------|---|---|-----|-----------|---|-----|------------|---|---|
| | | | С | | | | | | |
| | | | С | | | | | | |
| k ₄ : | Α | Α | С | С | G | G E | 1, C, G, T | T | Т |

 $\textbf{Entr\'ees}: \texttt{table_hachage} \ \textit{table}, \ \texttt{cha\^{i}nes} \ \textit{lectures}, \ \texttt{cha\^{i}ne} \ \textit{motif}, \ \texttt{entier} \ \textit{k}$

Entrées : table_hachage table, chaînes lectures, chaîne motif, entier k
pour chaque lecture de lectures faire

Entrées : table_hachage table, chaînes lectures, chaîne motif, entier k pour chaque lecture de lectures faire

```
\begin{array}{ll} \mathbf{pour} \ i = 0; \ i + k \leq |\mathit{lecture}|; \ i + = 1 \ \mathbf{faire} \\ & \mathit{kmerEntier} = 0; \\ & \mathit{kmer} = ""; \end{array}
```

 $\begin{table}{ll} \bf Entrées: table_hachage \ \it table, \ chaînes \ \it lectures, \ chaîne \ \it motif, \ entier \ \it k \ \it pour \ chaque \ \it lectures \ \it faire \ \it lectures \ \it lectu$

```
\begin{array}{l} \mathbf{pour} \ i = 0; \ i+k \leq |\mathit{lecture}|; \ i+=1 \ \mathbf{faire} \\ \\ k\mathit{merEntier} = 0; \\ k\mathit{mer} = ""; \\ \\ \mathbf{pour} \ j = 0; \ j < k; \ j+=1 \ \mathbf{faire} \\ \\ \\ | \ \mathbf{si} \ \mathit{motif[j]} \neq 0 \ \mathbf{alors} \ \mathit{kmer} = \mathit{kmer} + \mathit{lecture[i+j]}; \\ \\ \mathbf{fin} \end{array}
```

```
Entrées: table_hachage table, chaînes lectures, chaîne motif, entier k
pour chaque lecture de lectures faire
    pour i = 0; i + k < |lecture|; i + = 1 faire
         kmerEntier = 0:
         kmer = "";
         pour j = 0; j < k; j+ = 1 faire
              si motif[j] \neq 0 alors kmer = kmer + lecture[i + j];
         fin
         pour chaque nucleotide de kmer faire
              kmerEntier* = 4;
              suivant valeur de nucleotide faire
                  cas où A faire:
                  cas où C faire kmerEntier + = 1;
                  cas où G faire kmerEntier + = 2;
                  cas où T faire kmerEntier+=3;
              fin
         fin
```

```
Entrées: table_hachage table, chaînes lectures, chaîne motif, entier k
pour chaque lecture de lectures faire
    pour i = 0; i + k < |lecture|; i + = 1 faire
         kmerEntier = 0:
         kmer = "";
         pour j = 0; j < k; j+ = 1 faire
              si motif[j] \neq 0 alors kmer = kmer + lecture[i + j];
         fin
         pour chaque nucleotide de kmer faire
              kmerEntier* = 4;
              suivant valeur de nucleotide faire
                  cas où A faire:
                  cas où C faire kmerEntier + = 1;
                  cas où G faire kmerEntier + = 2;
                  cas où T faire kmerEntier+=3;
              fin
         fin
         table[kmerEntier] + = 1;
    fin
fin
```

kmerExpand

kmerExpand

 $\mbox{\bf Entr\'ees}$: chaînes $\it lectures$, chaîne $\it motif$, entier $\it k$

kmerExpand

Entrées : chaînes lectures, chaîne motif, entier k pour chaque lecture de lectures faire

kmerExpand

```
\label{eq:charge_entropy} \begin{split} &\textbf{Entr\'ees}: \text{cha\^nes } \textit{lectures}, \text{ cha\^ne } \textit{motif}, \text{ entier } \textit{k} \\ &\textbf{pour chaque } \textit{lecture } \textit{de lectures faire} \\ & | & \textbf{pour } i = 0; \ i + k \leq |\textit{lecture}|; \ i + = 1 \ \textbf{faire} \\ & | & \textit{kmer} = \textit{lecture}[i:k] \ \textit{kmerExpandRec(kmer, motif, 0)} \\ & & \textbf{fin} \end{split}
```

kmerExpand

```
\label{eq:charge_entropy} \begin{split} &\textbf{Entr\'ees}: \text{cha\^{n}es } \textit{lectures}, \text{ cha\^{n}e } \textit{motif}, \text{ entier } \textit{k} \\ &\textbf{pour chaque } \textit{lecture } \textit{de lectures faire} \\ & | & \textbf{pour } i = 0; \ i + k \leq |\textit{lecture}|; \ i + = 1 \ \textbf{faire} \\ & | & \textit{kmer} = \textit{lecture}[i:k] \ \textit{kmerExpandRec(kmer, motif, 0)} \\ & & \textbf{fin} \end{split}
```

11111

kmerExpandRec

Entrées : chaîne kmer, chaîne nvKmer, chaîne motif, entier posMotif, entier posKmer

kmerExpand

```
Entrées : chaînes lectures, chaîne motif, entier k
pour chaque lecture de lectures faire
    pour i = 0; i + k \le |lecture|; i + = 1 faire
         kmer = lecture[i : k] kmerExpandRec(kmer, motif, 0)
    fin
fin
```

kmerExpandRec

Entrées : chaîne kmer, chaîne nvKmer, chaîne motif, entier posMotif, entier posKmer si posSeed == |motif| alors affiche(nvKmer);

kmerExpand

kmerExpandRec

```
\label{eq:continuous_section} \begin{split} &\textbf{Entr\'ees}: \text{cha\^ine } \textit{kmer}, \text{ cha\^ine } \textit{nvKmer}, \text{ cha\^ine } \textit{motif}, \text{ entier } \textit{posMotif}, \text{ entier } \textit{posKmer} \\ &\textbf{si } \textit{posSeed} == |\textit{motif}| \text{ alors } \textit{affiche(nvKmer)}; \\ &\textbf{sinon } \textbf{si } \textit{motif[posMotif]} \neq 0 \text{ alors } \\ &\textit{kmersExpandRec(kmer, nvKmer + kmer[posKmer], posMotif + 1, posKmer + 1)}; \end{split}
```

kmerExpand

kmerExpandRec

fin

```
\begin{split} \textbf{Entr\'ees}: & \text{cha\^ine } \textit{kmer}, \text{ cha\^ine } \textit{nvKmer}, \text{ cha\^ine } \textit{motif}, \text{ entier } \textit{posMotif}, \text{ entier } \textit{posKmer} \\ \textbf{si } \textit{posSeed} &== |\textit{motif}| \text{ alors } \textit{affiche}(\textit{nvKmer}) \, ; \\ \textbf{sinon } \textbf{si } \textit{motif}[\textit{posMotif}] \neq 0 \text{ alors } \\ \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + \textit{kmer}[\textit{posKmer}], \textit{posMotif} + 1, \textit{posKmer} + 1) \, ; \\ \textbf{sinon} \\ & \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + A, \textit{posMotif} + 1, \textit{posKmer}) \, ; \\ \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + C, \textit{posMotif} + 1, \textit{posKmer}) \, ; \\ \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + G, \textit{posMotif} + 1, \textit{posKmer}) \, ; \\ \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + T, \textit{posMotif} + 1, \textit{posKmer}) \, ; \\ \textit{kmersExpandRec}(\textit{kmer}, \textit{nvKmer} + T, \textit{posMotif} + 1, \textit{posKmer}) \, ; \\ \end{cases}
```

Mots Minimaux absents

Mots Minimaux absents

Plus long sous-mot commun

Plus long sous-mot commun

Résultats obtenus

Résultats : KmersDel et kmersExpand

Résultats : Mots Minimaux absents

Résultats : Plus long sous-mot commun

