

Bioinformatique Analyse de données issues de séquenceurs à très haut débit





Pierre Morisse Doctorat en Informatique 1° année

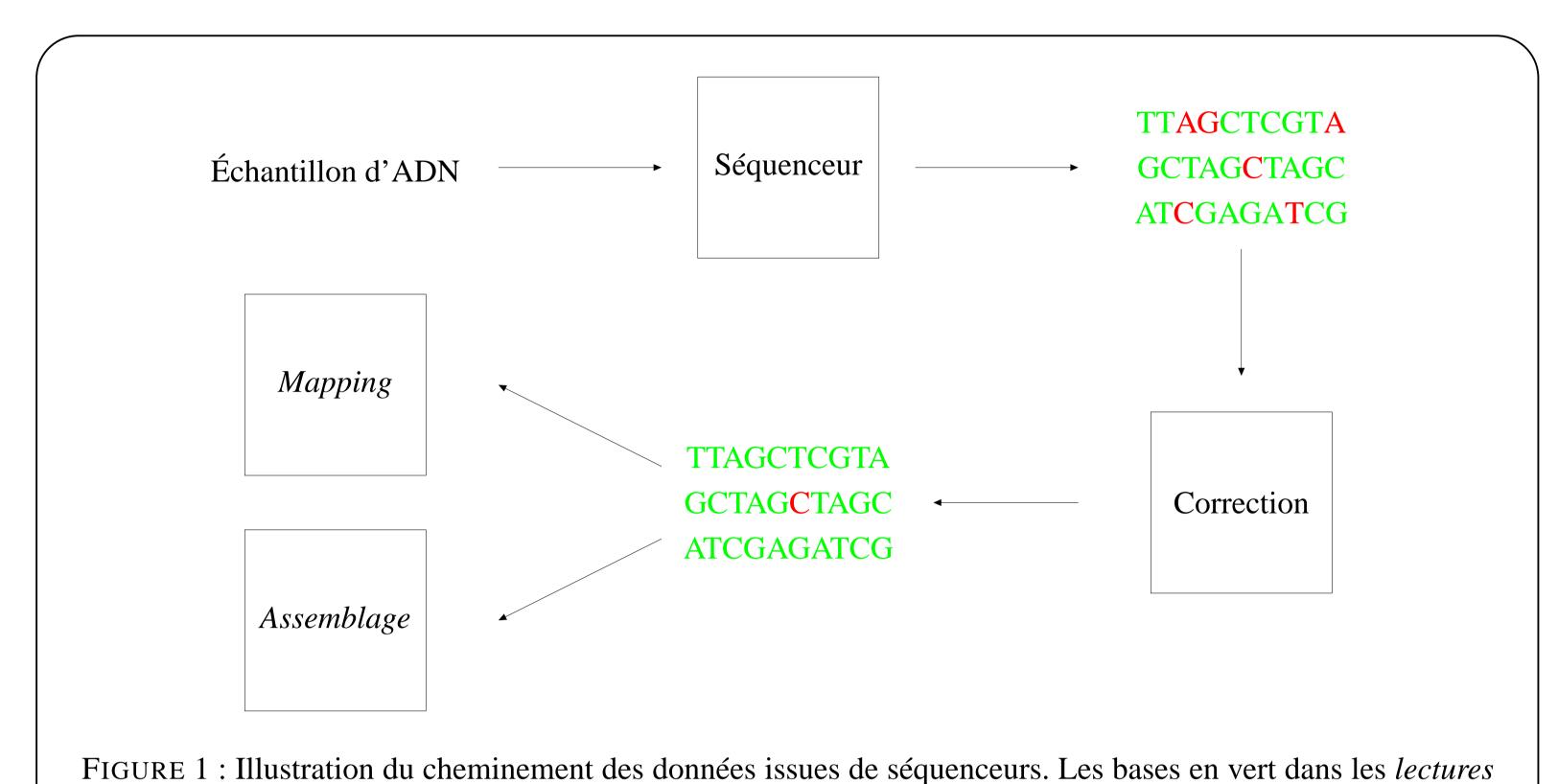


Introduction

Depuis le milieu des années 2000, les séquenceurs à très haut débit se développent et permettent de séquencer l'ADN d'un individu sous forme de courtes séquences appelées *lectures*, utilisées pour résoudre divers problèmes de génomique, notamment de *mapping* et d'*assemblage*.

Lors de du séquençage de l'ADN, les séquenceurs peuvent introduire des erreurs dans les *lectures* produites. Nous appelons ces erreurs des erreurs de séquençage, et il est donc souvent nécessaire de faire subir aux *lectures* une procédure de correction avant de les utiliser, afin d'améliorer leur précision.

Les séquenceurs produisant des millions de *lectures*, il est nécessaire de développer des outils informatiques adaptés au traitement de telles quantités de données, afin de permettre la résolution des différents problèmes.



correspondent aux bases correctes, et les bases en rouge aux erreurs de séquençage.

Évolution des séquenceurs

Les séquenceurs se développent et évoluent très rapidement. Ils deviennent moins imposants, moins coûteux, et visent ainsi à être plus accessibles au grand public. Ils produisent également des *lectures* de plus en plus longues, très utiles pour résoudre des problèmes complexes, bien que peu précises.



Principaux objectifs

Côté informatique :

- Développer des structures de données permettant de stocker et de traiter les grandes quantités de données formées par les *lectures*
- Développer des outils permettant aux biologistes de manipuler facilement les *lectures*

Côté biologie :

- Analyser les *lectures* afin détecter des mutations dans l'ADN d'un individu, et ainsi d'éventuelles pathologies
- Générer de nouveaux génomes de référence à partir des *lectures*, afin de les utiliser par la suite dans d'autres problèmes

3 principaux problèmes

1. Correction:

- Présence d'erreurs de séquençage dans les lectures, très nombreuses dans les lectures longues
- Nécessité de réduire le taux d'erreur afin d'améliorer la précision des lectures, et de faciliter leur utilisation
- Différentes approches (Comparaison des *lectures* entre elles, analyse des *facteurs* des *lectures*, etc)

2. Mapping:

- Aligner les *lectures* séquencées sur un génome de référence
- Comparer l'ADN d'un individu à l'ADN du génome de référence
- Détecter des mutations et d'éventuelles pathologies



FIGURE 2 : Illustration du *mapping* de *lectures* (en noir) sur un génome de référence (en bleu). A gauche, le *mapping* est réalisé avec des *lectures* courtes et met en évidence la présence de *gaps* (parties non couvertes du génome de référence), tandis qu'à droite, le *mapping* est réalisé avec des *lectures* longues, et permet de se débarrasser des *gaps* et de couvrir totalement le génome de référence.

3. Assemblage:

- Aligner les *lectures* entre elles afin de trouver des chevauchements
- Assembler les *lectures* se chevauchant afin de créer des *contigs*
- Reconstruire ainsi le génome dont les *lectures* sont originalement issues

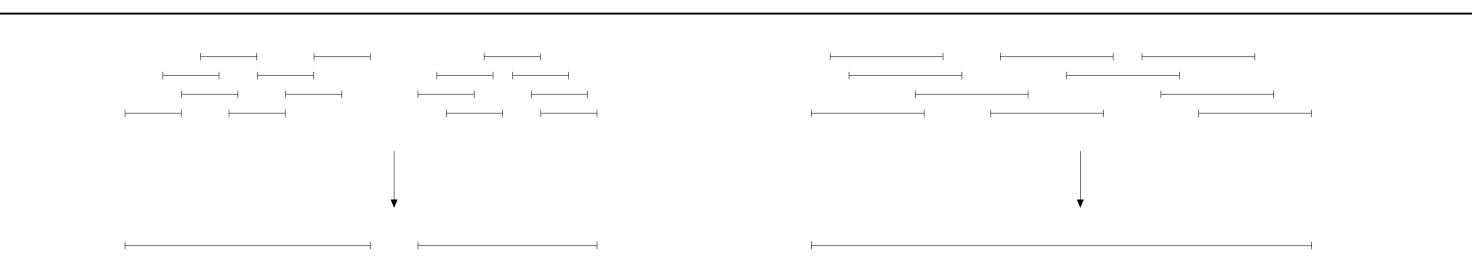


FIGURE 3 : Illustration de l'assemblage de *lectures*. À gauche, l'assemblage est réalisé avec des *lectures* courtes, et produit plusieurs *contigs*, tandis qu'à droite, l'assemblage est réalisé avec des *lectures* longues, et permet de ne produire qu'un unique *contig*.

Mon travail actuel

Mon travail actuel porte sur le développement d'une nouvelle méthode de correction de *lectures* longues. Plus précisément, de production de *lectures* longues dites *synthétiques*, car obtenues à partir d'un assemblage de *lectures* courtes. Pour cela, des *lectures* courtes sont alignées sur une *lecture* longue servant de *modèle*. Les *lectures* courtes totalement alignées servent alors de *graines*, et sont étendues à l'aide de *lectures* courtes partiellement alignées.

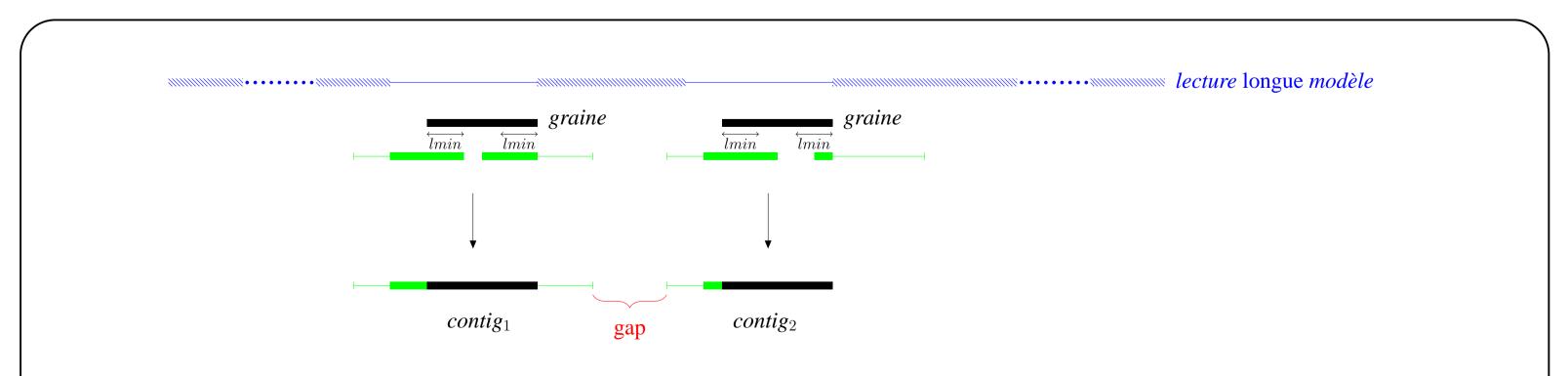


FIGURE 4 : Illustration de la méthode d'extension des *graines*. Les parties en gras des *lectures* courtes (en noir et en vert) correspondent aux parties de ces *lectures* correctement alignées sur le *modèle*. Les parties hachurées du *modèle* (en bleu) correspondent aux zones à fort taux d'erreurs de séquençage de celui-ci.

Les *graines* (en noir) sont étendues à l'aide de *lectures* courtes (en vert) dont seul un *préfixe*, ou seul un *suf-fixe*, s'est correctement aligné sur le modèle. Ces *lectures* partiellement alignées étendent alors effectivement la *graine* si elles la chevauchent sur un longueur au moins lmin, et un *contig* est alors obtenu.

Une fois ces *contigs* obtenus, le problème restant est alors le remplissage des *gaps*. Il existe pour cela une première méthode nécessitant d'aligner toutes les *lectures* courtes entre elles afin de les comparer et de les assembler et ainsi produire de nouveaux *contigs*. Cette méthode se montre cependant très coûteuse en terme de temps, et au vu du grand nombre de *lectures* longues à corriger, nous cherchons actuellement une autre solution permettant de combler ces *gaps* plus rapidement, sans nécessiter l'alignement de toutes les *lectures* courtes entre elles.