Correction hybride : Les reads longs synthétiques

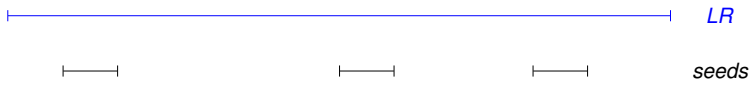
Pierre Morisse

20 janvier 2017

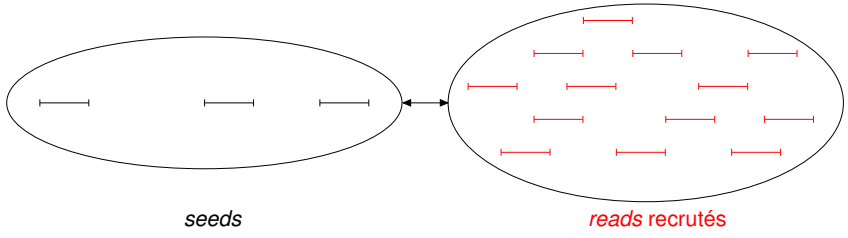
NaS

Idée : Création de reads longs synthétiques via une approche hybride, en assemblant des reads courts à l'aide d'un LR template

1. Alignement des SR sur le LR, afin de trouver des seeds



2. Recrutement de nouveaux SR, similaires aux seeds, en alignant les SR entre eux

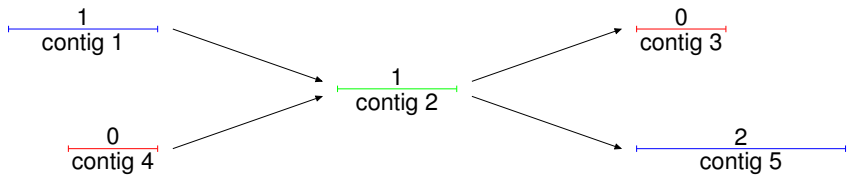


3. Assemblage de l'ensemble de SR obtenu, et obtention d'un contig

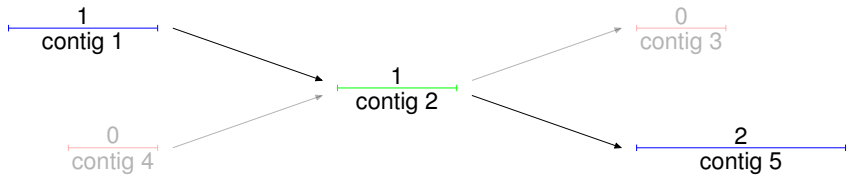
NaS

En général un unique contig est produit, mais de mauvais SR peuvent être recrutés et produire des contigs erronés

1. Construction du graphe des contigs



2. Sélection du chemin optimal



3. Vérification du contig obtenu, par alignement des SR

Notre méthode

Idée : Comme NaS, créer des LR synthétiques à partir d'un assemblage de SR et d'un LR template, mais en s'affranchissant de l'étape d'alignement des SR entre eux, qui est l'étape la plus coûteuse en temps de NaS

Méthode divisée en 5 étapes

Notre méthode

1. Correction des SR (avec Quorum)

2. Alignement des SR sur le LR, afin de trouver des seeds (avec BLAT)

3. Fusion des seeds se chevauchant sur une longueur assez importante

4. Relier les seeds en les étendant à l'aide de chevauchements parfaits avec les k-mers des SR

5. Extension du LR sythétique obtenu, à gauche (resp. à droite) du seed le plus à gauche (resp. à droite)

Outil utilisé

PgSA (Pseudogenome Suffix Array) permet d'indexer un ensemble de reads, et de répondre aux 7 requêtes suivantes, pour une chaîne *f* donnée :

1. Dans quels *reads f* apparaît ?
2. Dans combien de *reads f* apparaît ?
3. Quelles sont les occurrences de *f* ?
4. Quel est le nombre d'occurrences de *f* ?
5. Dans quels *reads f* n'apparaît qu'une fois ?
6. Dans combien de *reads f* n'apparaît qu'une fois ?
7. Quelles sont les occurrences de *f* dans les *reads* où *f* n'apparaît qu'une fois ?

Parmi ces requêtes, la 3ème va nous permettre de trouver des chevauchements parfaits entre les k-merrs

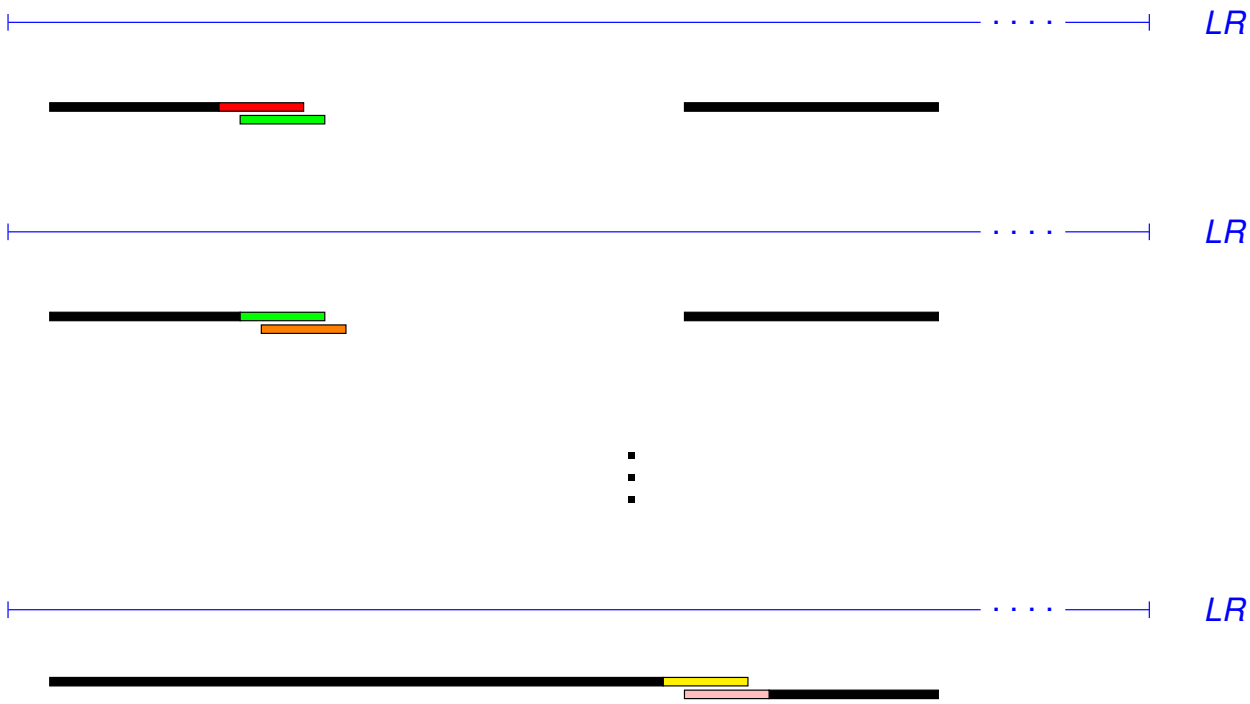
Outil utilisé

D'autres structures (Gk-Arrays, Compressed Gk-Arrays) permettent le traitement des ces requêtes, mais la longueur k de f doit être fixée à la compilation, alors que PgSA permet de traiter les requêtes pour des valeurs de k variables.

=> Permet de chercher des chevauchements de longueur k-2 si aucun chevauchement de longueur k-1 n'a été trouvé, sans avoir besoin de recalculer l'index

Étape 4

Indexation de l'ensemble des k-mers des SR avec PgSA, et boucle sur la requête 3 afin de trouver des chevauchements parfaits de k-mers permettant de lier les seeds entre eux



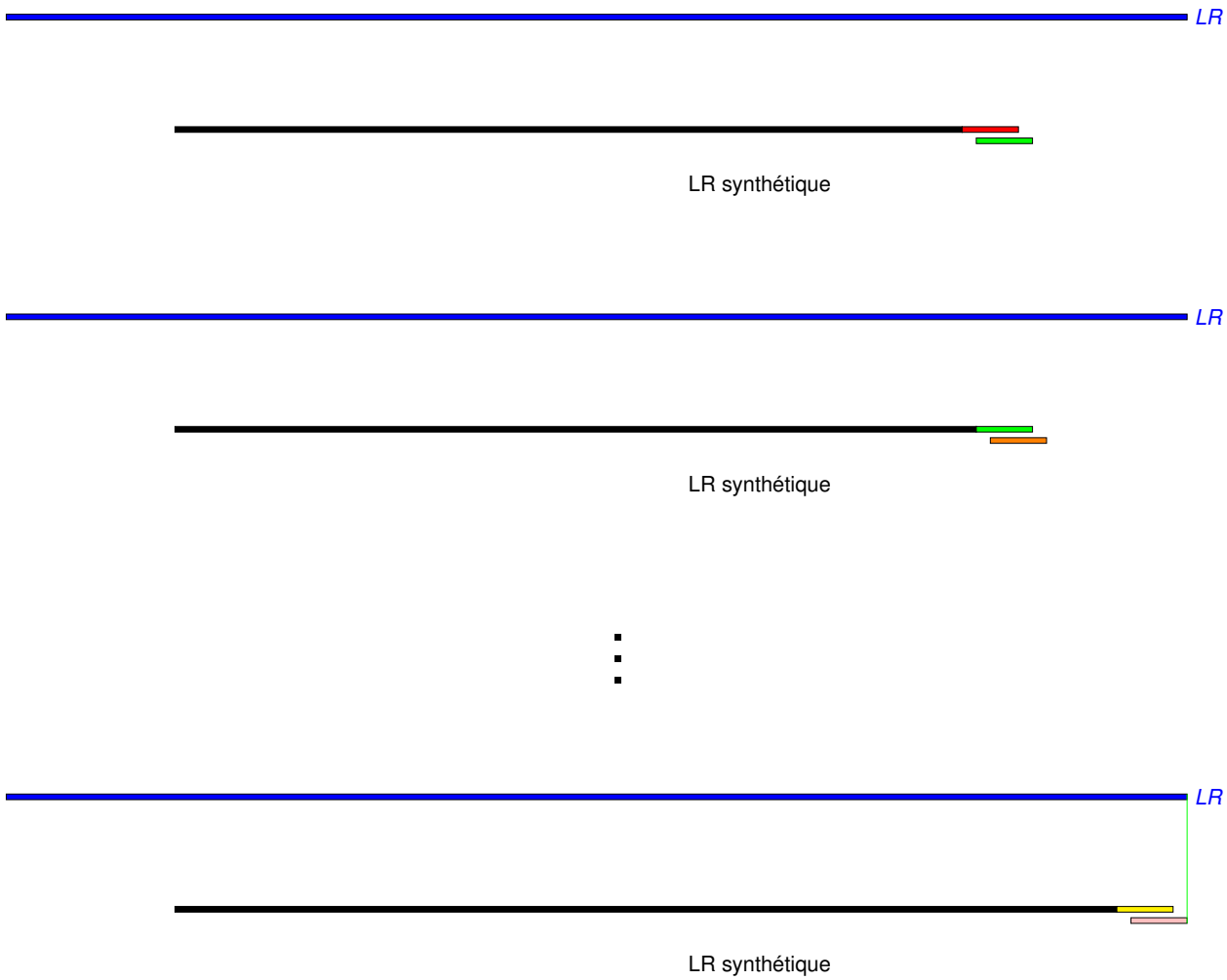
Remarque : Il est possible qu'un k-mer chevauche parfaitement plusieurs k-mers $\Rightarrow$ Exploration de toutes les extensions possibles avec du backtracking

Étape 5

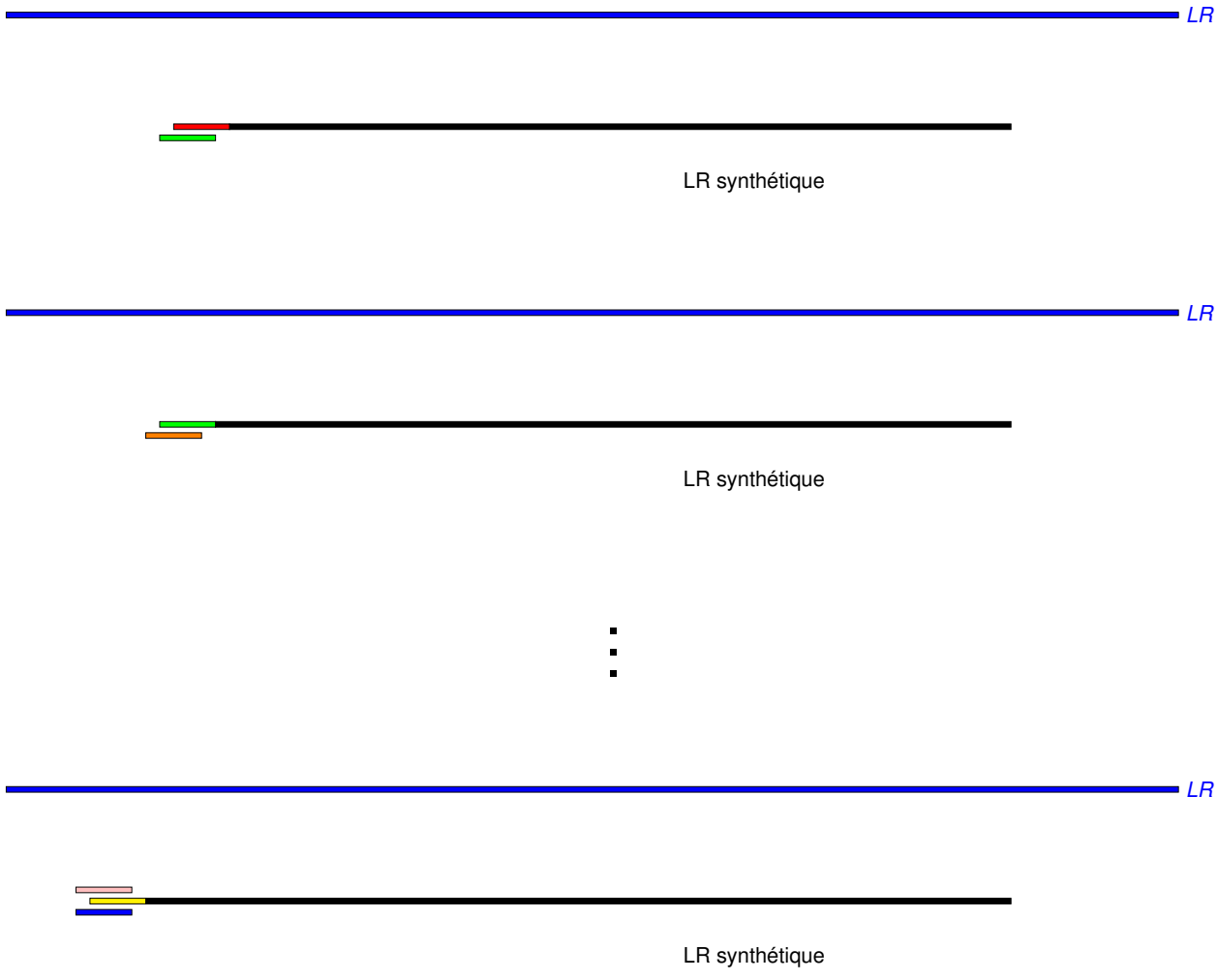
Le seed le plus à gauche ne s'aligne pas toujours en position 0 sur le LR, et de même, le seed le plus à droite n'atteint pas toujours l'extrémité droite du LR.

=> Une fois tous les seeds reliés et le LR synthétique produit, on étend, à l'aide de chevauchements parfaits de k-mers, son extrémité gauche et son extrémité droite, jusqu'à atteindre les extrémités du LR initial, ou une ambiguïté (extension possible à l'aide de plus d'un k-mer)

Étape 5, Cas 1 : Pas d'ambiguïté



Étape 5, Cas 2 : Présence d'une ambiguïté



Remarques

- Lors de l'étape 4, certains seeds peuvent être impossibles à relier $\Rightarrow$ Production d'un LR synthétique fragmenté en plusieurs parties

- Lorsqu'un LR ne possède qu'un seed, on se contente alors d'étenre celui-ci au maximum à gauche et à droite, jusqu'à atteindre les extrémités du LR ou une ambiguïté

Résultats et comparaison avec NaS

Sur un jeu de LR 1D de ADP1 :

	Nombre de reads	**Longueur moyenne**	**Taille totale**	**Identité moyenne**	s **Temps**
LR bruts	10 567	1 873	19 788 858	3,68 %	N.A.
NaS	784	3 764	2 951 256	99,76 %	11h20
Nous	786 LR (dont 59 fragmentés)	9 095	7 757 758	99,52 %	21 min

Sur un jeu de LR 2D de ADP1 :

	Nombre de reads	**Longueur moyenne**	**Taille totale**	**Identité moyenne**	**Temps**
LR bruts	602	5 197	3 128 834	8,83 %	N.A.
NaS	445	5 798	2 580 256	99,83 %	5h49
Nous	443 (dont 22 fragmentés)	5 627	2 633 407	99,91 %	8 min

À faire

- Tester sur d'autres jeux de données
- Trouver une solution pour éviter les LR synthétiques fragmentés (tuning des paramètres de BLAT ?)
- Utiliser un autre aligneur pour mimer les modes fast et sensitive de NaS
- Paralléliser la correction