LE SÉQUENÇAGE ET L'ASSEMBLAGE DES GÉNOMES

Comment obtenir la séquence du génome d'un individu ?

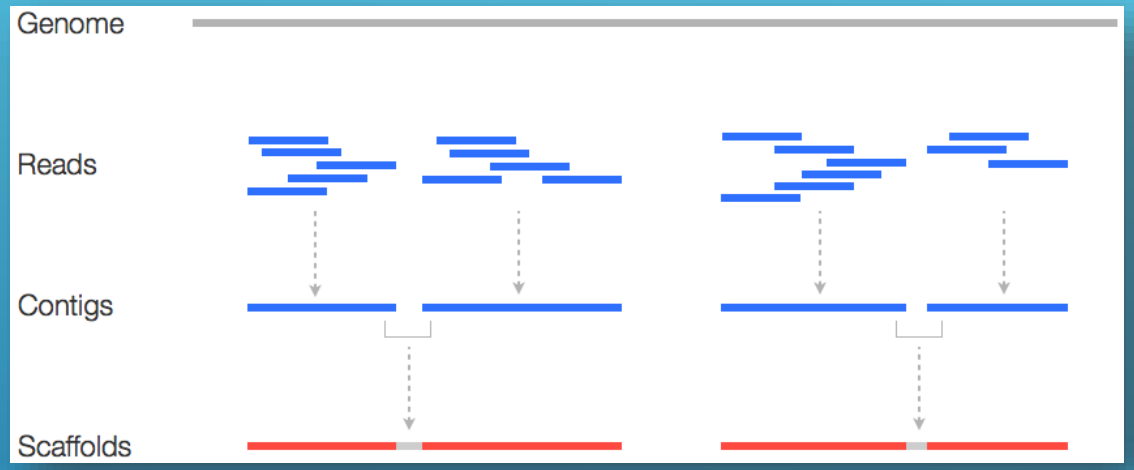
BOUILLÉ Pauline
CORDIER Mathis

- SÉQUENÇAGE DES GÉNOMES
  - Histoire
  - Machines
  - Aujourd'hui

- ASSEMBLAGE DES SÉQUENCES
  - Assemblage *Shotgun*
  - Assemblage par graphe de chevauchement
  - Assemblage par graphe de *De Bruijjn*

SÉQUENÇAGE DES GÉNOMES

INTRODUCTION

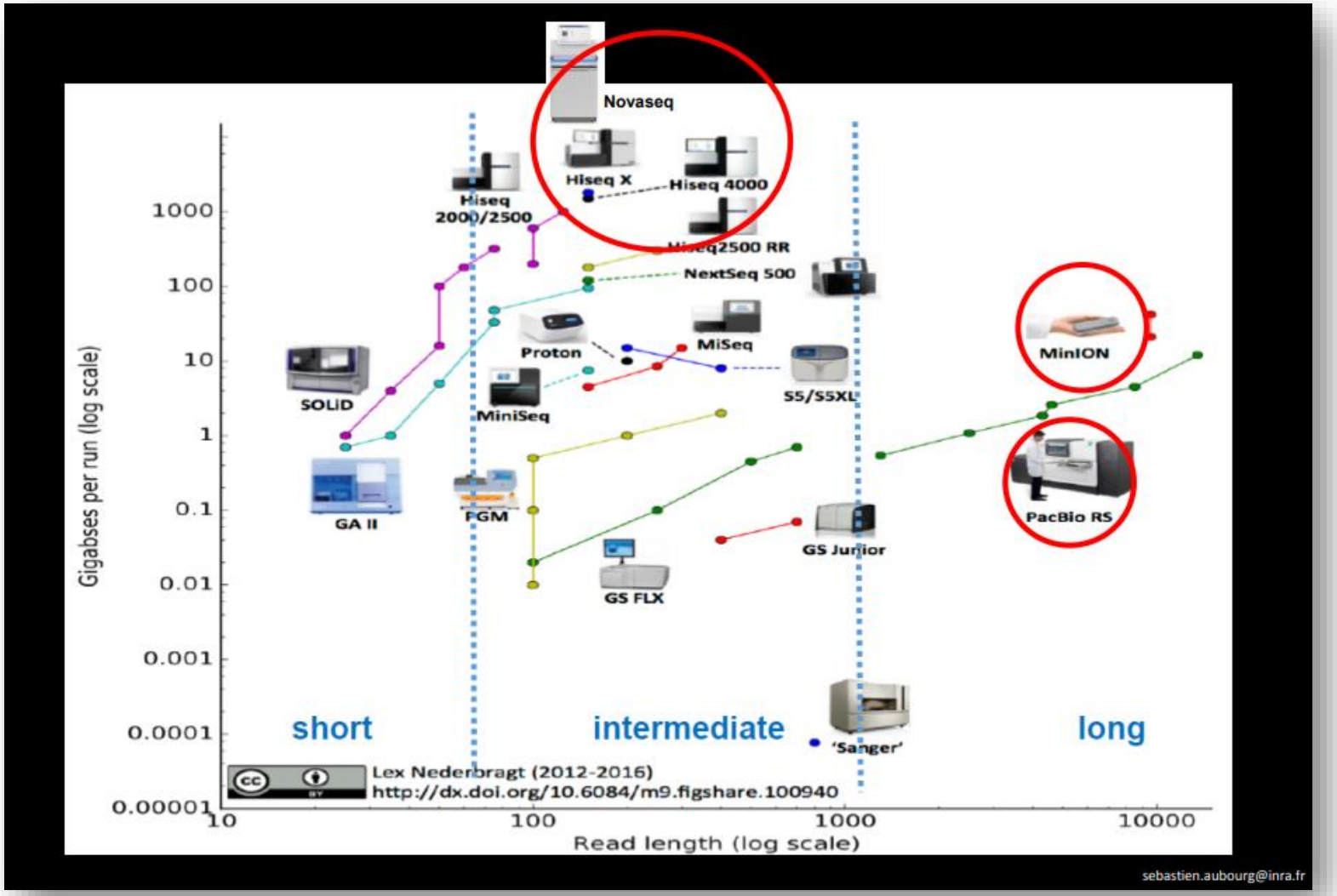


- Objectif du séquençage : Obtenir la séquence ATCG pour comprendre l'expression des gènes
- Obtention d'un grand nombre de sous-séquences de la séquence du génome

Histoire

1977	Frederick Sanger Biomimétisme : inspiré de la manière dont les enzymes reproduisent les brins d'ADN Utilisation de produits radioactifs et résultats sur plaque de gel 160 nucléotides par jour
1990	Produits fluorescents à la place de radioactifs et résultats sur capillaires en verre Début du projet Génome Humain
2000	Automatisation : 500.000 nucléotides par jour
2003	Projet Génome Humain : Première séquence du génome humain 20 institutions engagées dans le monde entier Séquence composée de 3 Mrd de nucléotides Coût estimé : 3 Mrd €
2006	Nanotechnologies : Séquençage à Haut Débit Tout est automatisé et résultats sur puce électronique 1 Mrd de nucléotides peuvent être obtenus en quelques heures

Machines


Novaseq
Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
SOLiD
MiniSeq
S5/S5XL
MinION
GA II
FGM
GS Junior
PacBio RS
GS FLX
'Sanger'
short
intermediate
long
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
sebastien.aubourg@inra.fr

Aujourd'hui

Méthode	Longueur de la lecture	Précision	Lectures par expérience	Temps d'expérience	Coût par million de bases
Ion semiconductor (Séquençage Ion Torrent)	Jusqu'à 400 Mb	98 %	Jusqu'à 80 millions	2 heures	1 $
Pyroséquençage (454)	700 Mb	99,9 %	1 million	1 jour	10 $
Séquençage par synthèse (Illumina)	50 à 300 Mb	99,9 %	Jusqu'à 6 milliards	1 à 11 jours	0.05/0.15 $

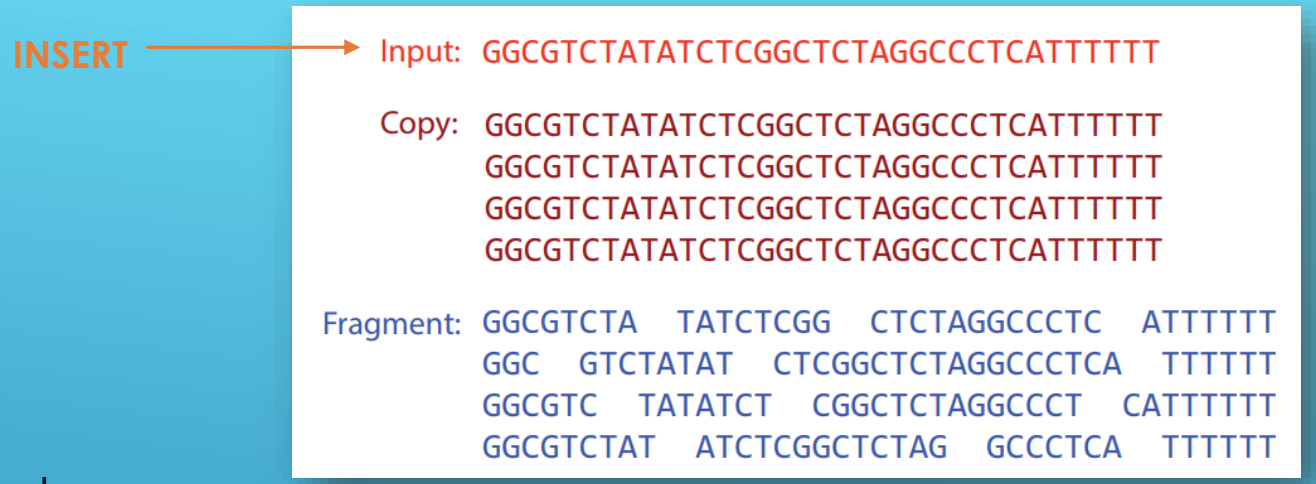
Aujourd'hui

Nom	Nombre de machines
Illumina HiSeq 2000	5490
Illumina Genome Analyser 2x	411
Roche 454	382
ABI SOLiD	326
Ion Torrent	301
Illumina MiSeq	299
Ion Proton	104
Pacific Biosciences	50
Oxford Nanopore MinION	14
Illumina NextSeq	3

ASSEMBLAGE DES SÉQUENCES

INTRODUCTION

- Assemblage de *Novo* : Aucune référence pour le génome étudié
- Comment assembler les séquences recueillies ?



- Séquençage aléatoire dans un insert
- On séquence de nombreux fragments de l'insert
- Comment assembler ces fragments ?

ASSEMBLAGE *SHOTGUN*

- Soit *L* la taille de l'ADN étudié.
  Soit *N* le nombre de nucléotides total des lectures.
  Soit *n* la taille de chaque lecture.
- On a donc la profondeur de lecture : $P = \frac{N}{L}$
- On peut considérer que l'évènement Ω suivant:
  Ω^x = {Une base de la séquence cible est représentée dans x lectures} suit une loi de Poisson de paramètre P, i.e.

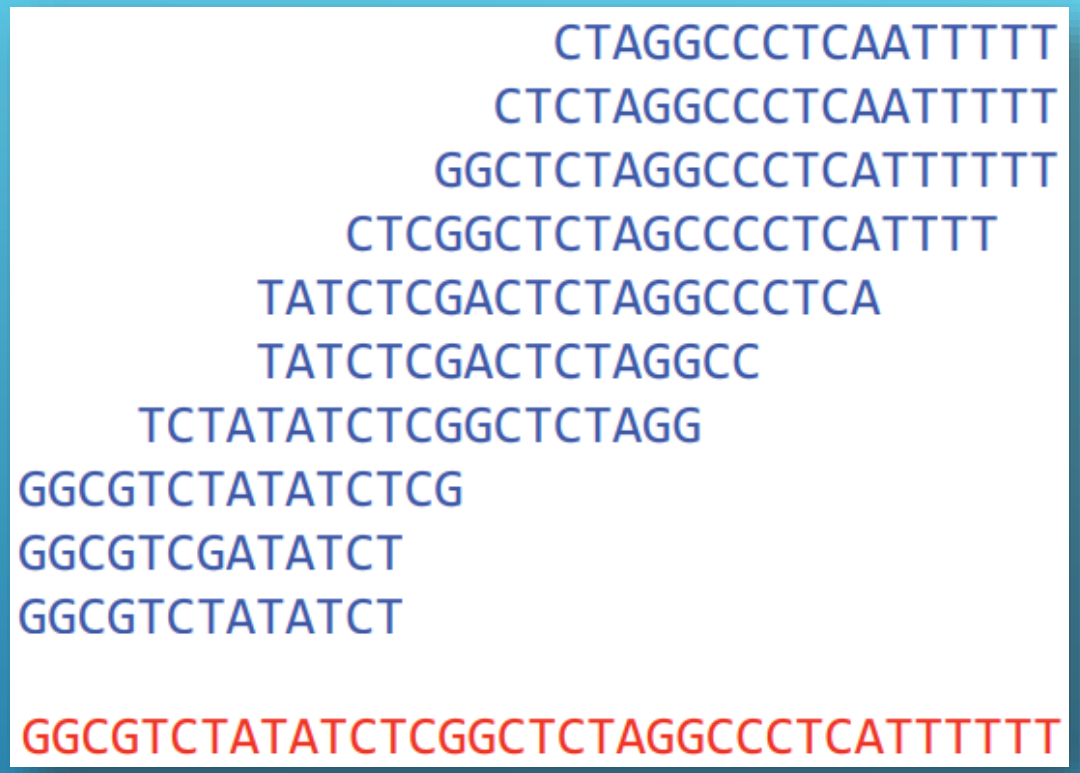
$$\mathbb{P}(\Omega^x) = \frac{P^x}{x!} e^{-P}$$

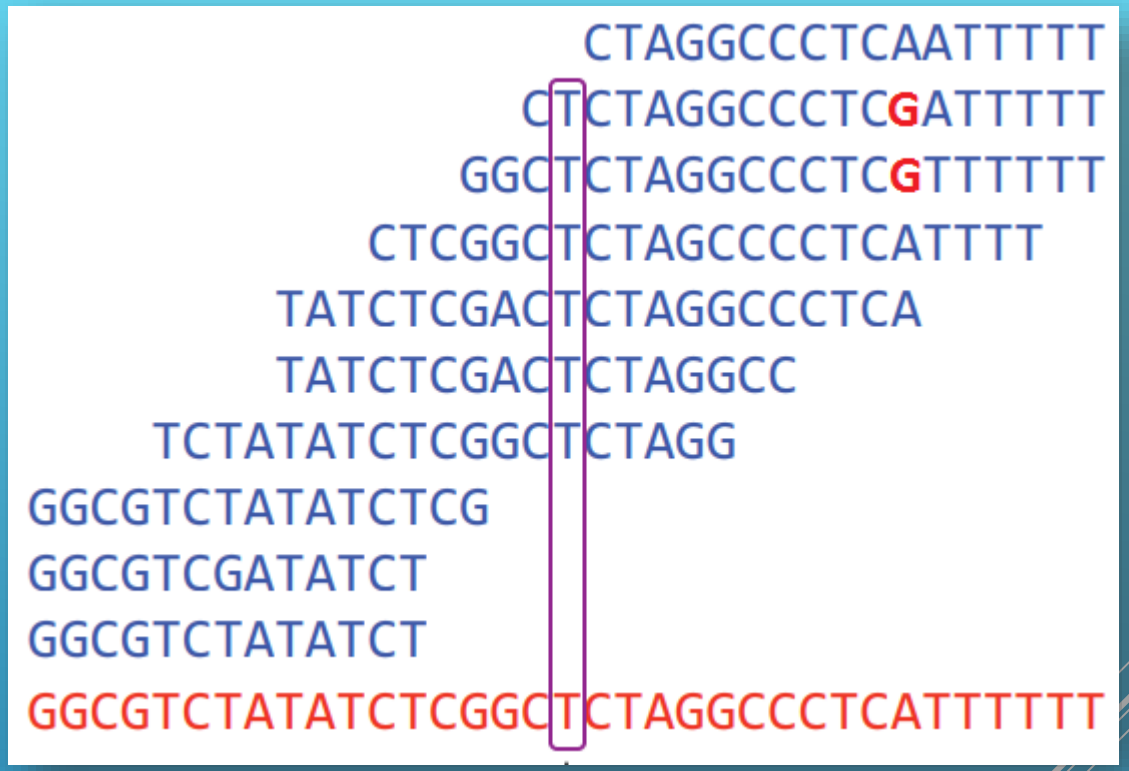
- *Alors on a*

$$taux_{adn\_lu} = 1 - \mathbb{P}(\Omega^0) = 1 - e^{-P}$$

$$N_{trous} = N_{lecture} \cdot \mathbb{P}(\Omega^0) = \frac{N}{n} \cdot e^{-P}$$

$$Taille_{trous} = L/N_{lecture} = \frac{n}{P}$$





10 Fragments : 177 nucléotides
1 Insert : 35 nucléotides

Profondeur : $\mathbf{P = \frac{177}{35} \approx 5\,X}$

- Il peut y avoir des différences entre les fragments
- On ne connait pas l'ordre des séquences

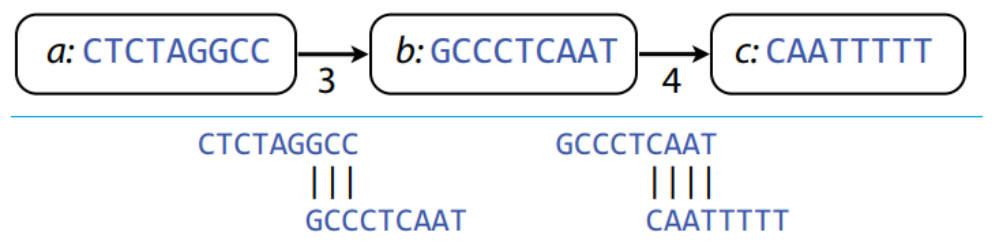
DEUX APPROCHES D'ASSEMBLAGE

- ASSEMBLAGE PAR GRAPHE DE CHEVAUCHEMENT
  Construction d'un graphe de chevauchement directement depuis les reads
  Simplification du graphe
  Assemblage par le graphe

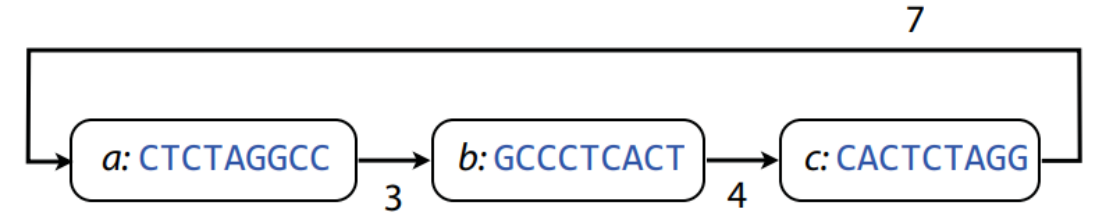
- ASSEMBLAGE PAR GRAPHE DE *DE BRUIJN*
  Construction d'un graphe des *k-mer*
  Élimination des reads originaux
  Assemblage par le graphe

ASSEMBLAGE PAR CHEVAUCHEMENT (OLC)

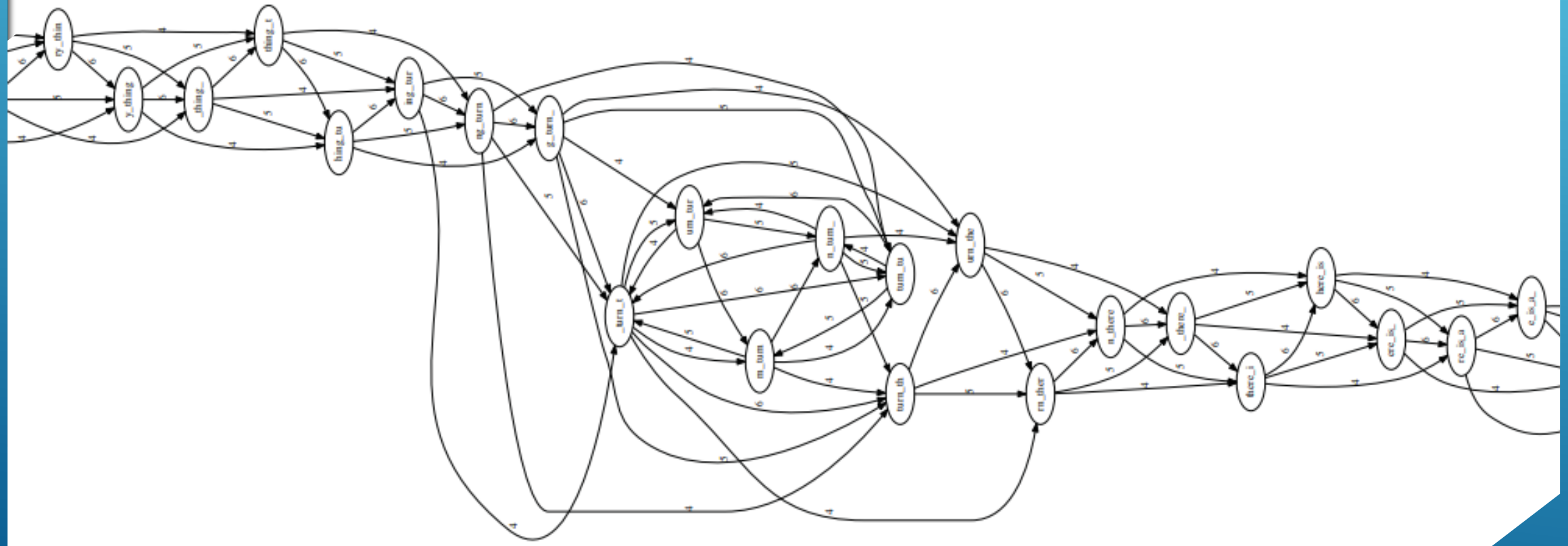
- Chaque sommet du graphe correspond à un fragment
- Les arêtes représentent les chevauchements entre les fragments
- Le nombre de bases qui se chevauchent entre les deux fragments est indiqué au niveau de l'arête



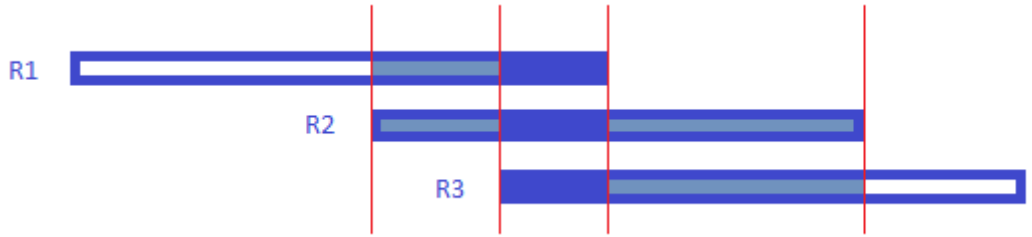
- Le graphe peut contenir des cycles
  La séquence du génome elle-même peut être circulaire



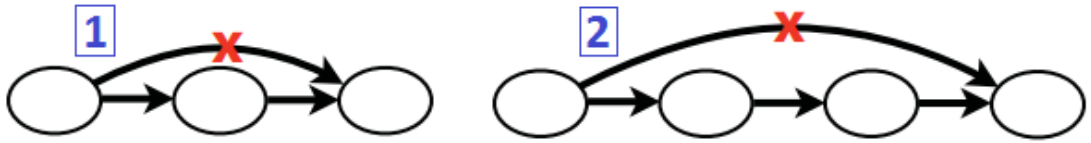
Séquence à retrouver	to_every_thing_turn_turn_turn_there_is_a_season
Taille des reads	7
Taille minimale des chevauchements entre 2 reads	3



On supprime alors tous les reads qui n'apportent pas d'information supplémentaire

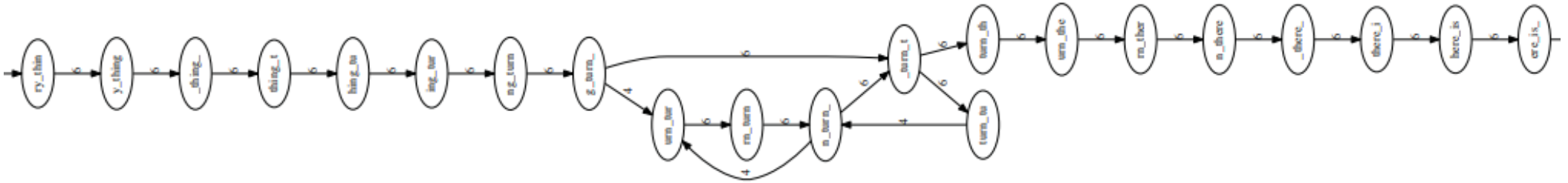


Nous les supprimons en commençant par ceux qui traversent le moins de nœuds possibles du graphe

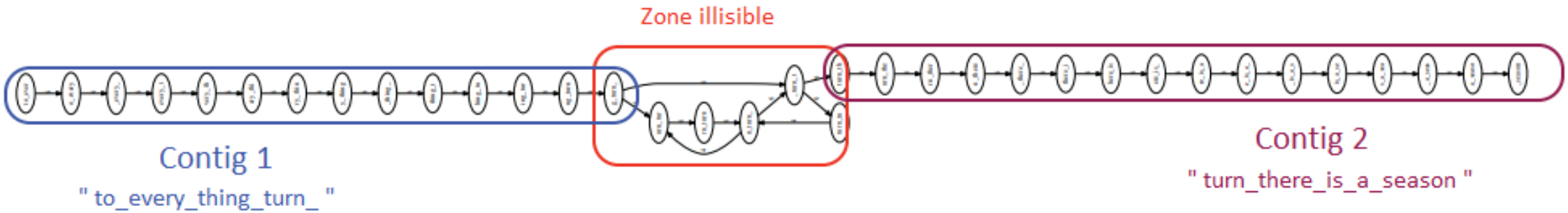


Avec R1,R2,R3, nous aurions

En supprimant les arêtes concernées nous obtenons le graphe suivant :

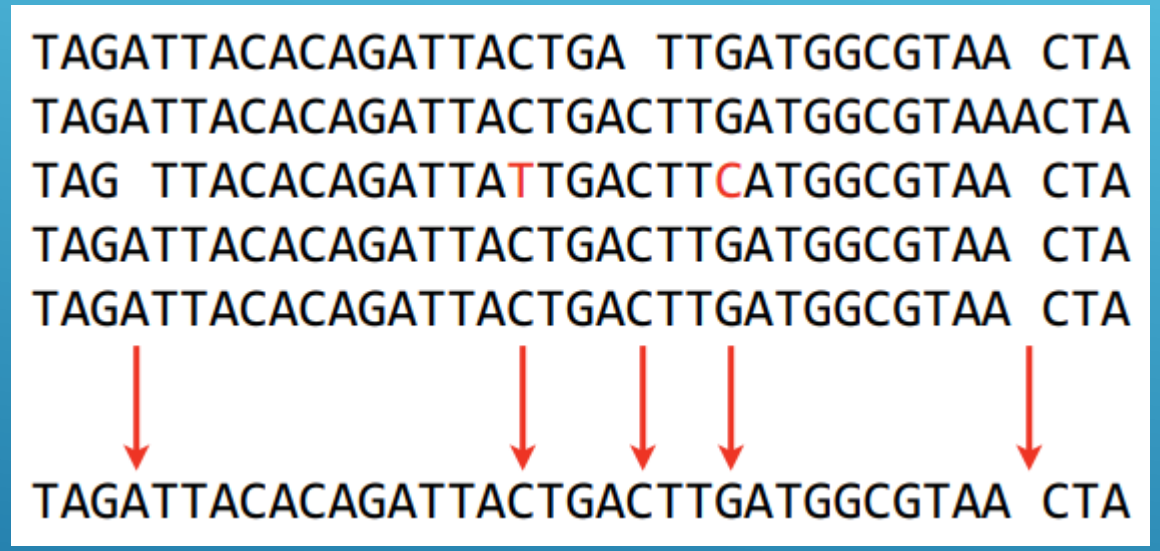


Nous avons alors obtenu 2 contigs et une zone que nous ne pouvons pas traiter



La zone illisible se situe sur l'endroit correspondant à la série de « .._turn_.. »
Les répétitions sont un gros facteurs d'erreurs ou d'incertitudes

CONSENSUS



Il s'agit de s'accorder sur la séquence qui semble être la vraie,
c'est-à-dire qu'il faut conserver le nucléotide le plus fréquent sur les bases où il y a un doute

DÉFAUTS DE LA MÉTHODE

- Cette méthode peut-être lente
  Ajouté à cela le très grand nombre de bases, elle devient très lente !

  En effet,
  Soit N le nombre de reads
  Nombre de calculs :

$$O(N^2)$$

ASSEMBLAGE PAR GRAPHE DE *DE BRUIJN* (DBG)

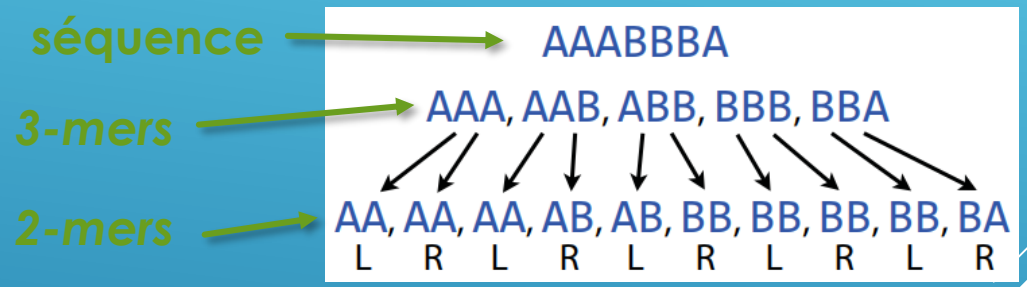
K-MER

- Sous-partie de la séquence de taille K

séquence → GGCGATTCATCG

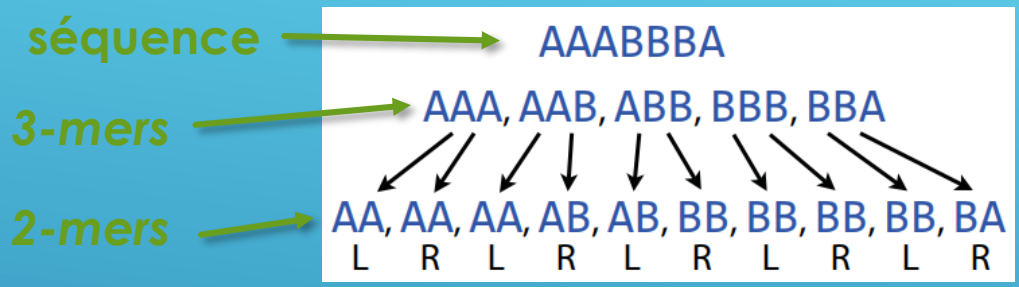
Tous les 3-mers associés

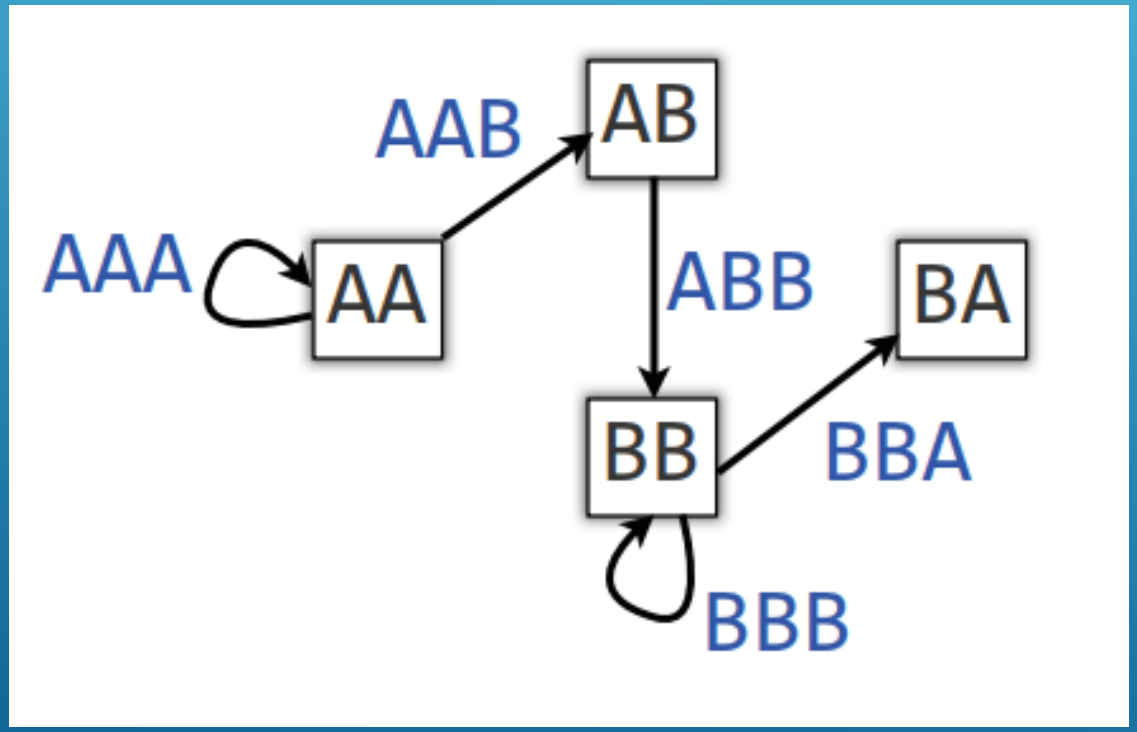
GGC
GCG
CGA
GAT
ATT
TTC
TCA
CAT
ATC
TCG



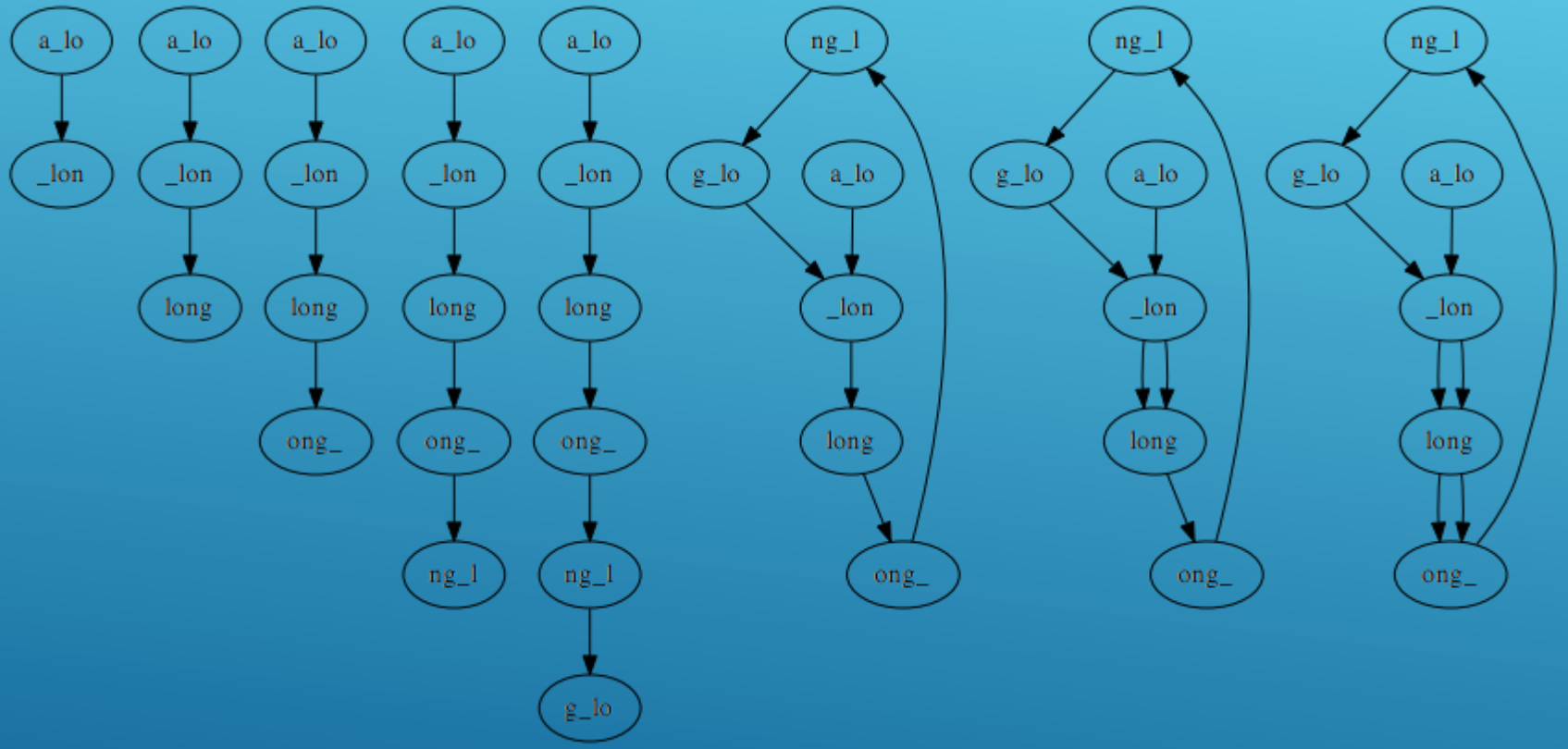
La méthode de *De Bruijn* utilise les *K-1*-mers

GRAPHE DE *DE BRUIJN*

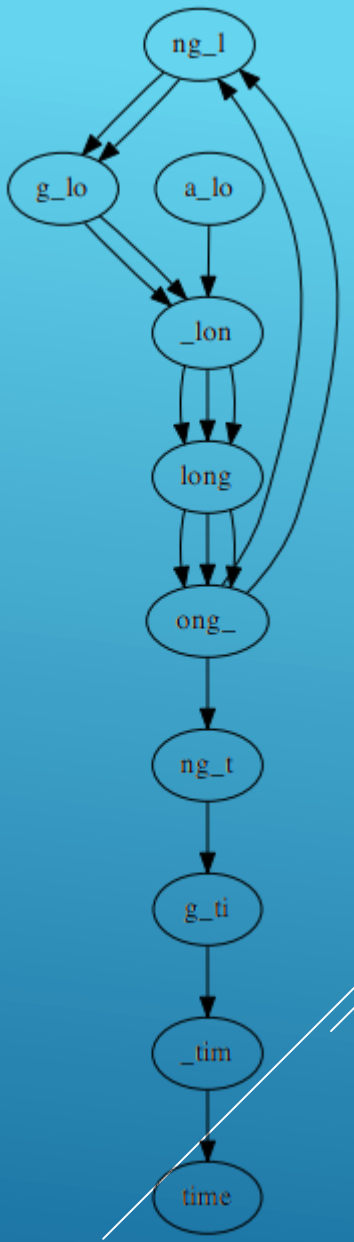




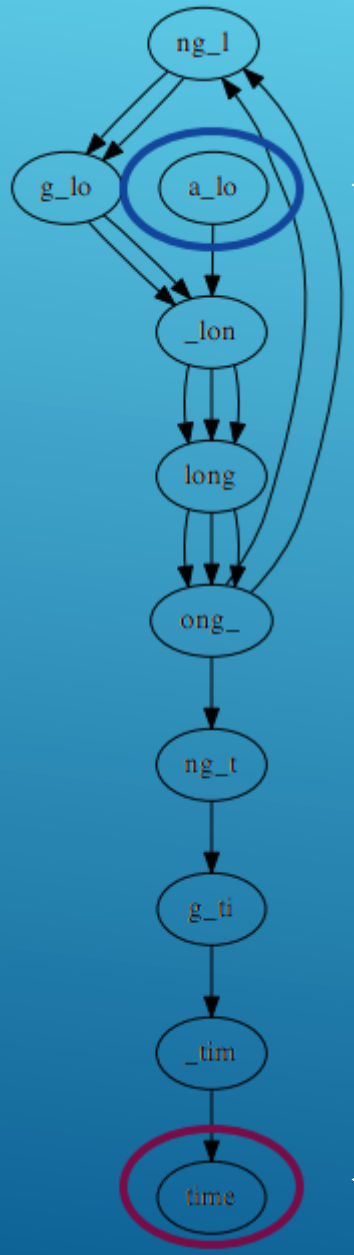
A_LONG_LONG_LONG_TIME



Les 8 premiers 5-mers associés à la séquence



Les 4-mers associés à la séquence



Premier *4-mer*

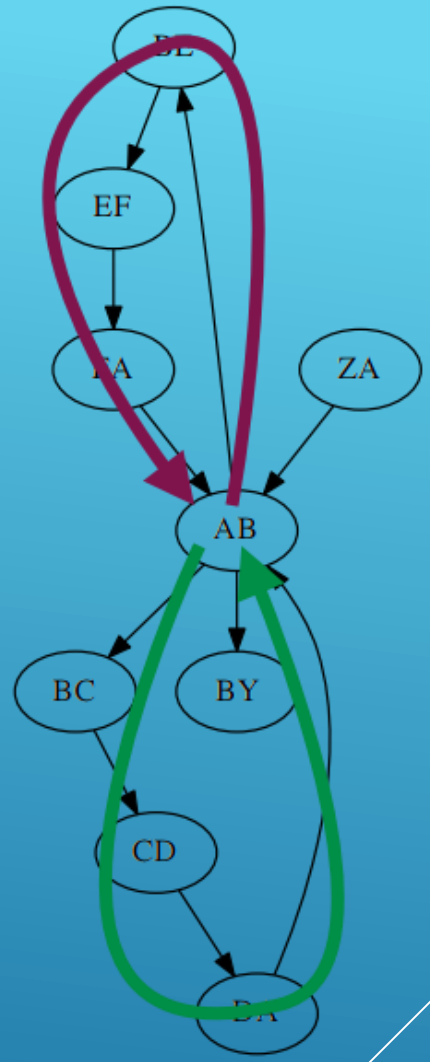
Dernier *4-mer*

PROBLÈME DE MULTIPLES CHEMINS

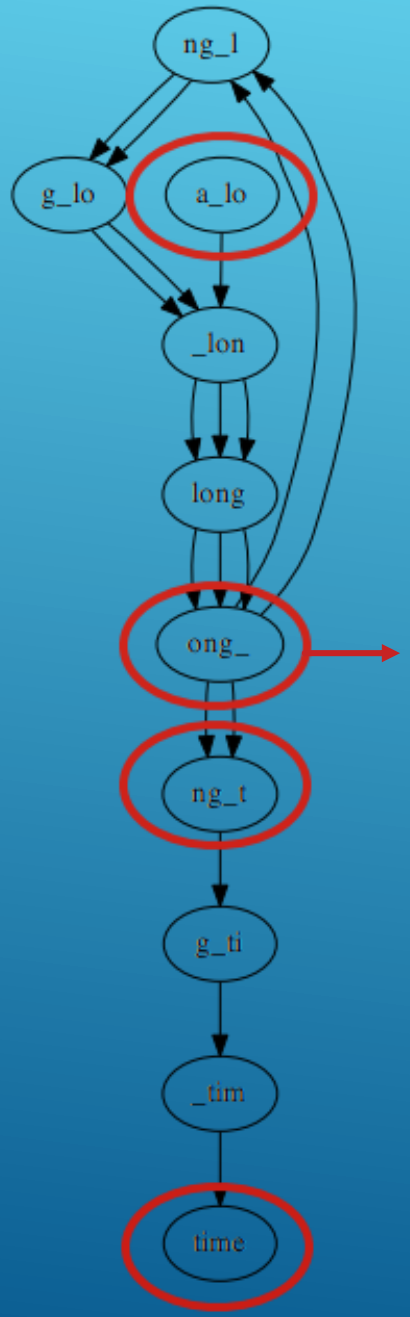
Deux lectures du graphe sont possibles

ZA → AB → BE → EF → FA → AB → BC → CD → DA → AB → BY

ZA → AB → BC → CD → DA → AB → BE → EF → FA → AB → BY

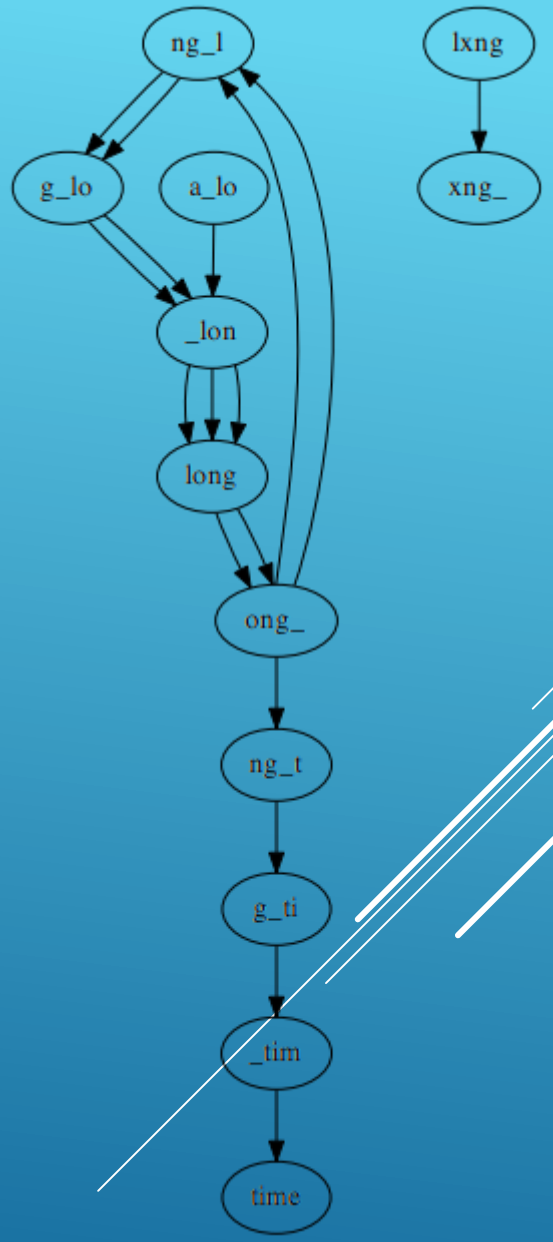


NOEUDS DÉSÉQUILIBRÉS



3 entrants
4 sortants

DIFFÉRENCES DE SÉQUENÇAGE



TEMPS DE CALCUL

- Pour chaque *K-mer*, on ajoute 1 arête et 2 nœuds
  Donc cela équivaut à du *O(1)*

- Soit N la taille de la séquence
  Il y a N-K *K-mers*
  Donc cela équivaut à du *O(N)*

- Nous avons alors
  Nombre de calculs :

$$\boldsymbol{O(N)}$$

	OLC t >= 75	De Bruijn k = 61	De Bruijn k = 67	De Bruijn k = 59

Table 1. Assembly statistics for *C. elegans* data set

	SGA	Velvet	ABySS	SOAPdenovo
Scaffold N50 size	26.3 kbp	31.3 kbp	23.8 kbp	31.1 kbp
Aligned contig N50 size	16.8 kbp	13.6 kbp	18.4 kbp	16.0 kbp
Mean aligned contig size	4.9 kbp	5.3 kbp	6.0 kbp	5.6 kbp
Sum aligned contig size	96.8 Mbp	95.2 Mbp	98.3 Mbp	95.4 Mbp
Reference bases covered	96.2 Mbp	94.8 Mbp	95.9 Mbp	95.1 Mbp
Reference bases covered by contigs ≥1 kb	93.0 Mbp	92.1 Mbp	93.9 Mbp	92.3 Mbp
Mismatch rate at all assembled bases	1 per 21,545 bp	1 per 8786 bp	1 per 5577 bp	1 per 26,585 bp
Mismatch rate at bases covered by all assemblies	1 per 82,573 bp	1 per 18,012 bp	1 per 8209 bp	1 per 81,025 bp
Contigs with split/bad alignment (sum size)	458 (4.4 Mbp)	787 (7.2 Mbp)	638 (9.1 Mbp)	483 (4.4 Mbp)
Total CPU time	41 h	2 h	5 h	13 h
Max memory usage	4.5 GB	23.0 GB	14.1 GB	38.8 GB

CONCLUSION

BIBLIOGRAPHIE

- MIT
- INSERM
- TED
- Wikipédia
- Biorigami
- SNJ Jussieu – Sorbonne
- IRO Université de Montreal
- France-Génomique