Assemblage et Annotation de génome

David Couvin
(Institut Pasteur de la Guadeloupe)

Assemblage

- En bio-informatique, l'assemblage consiste à aligner et/ou fusionner des fragments d'ADN ou d'ARN issus d'une plus longue séquence afin de reconstruire la séquence originale. Il s'agit d'une étape d'analyse in silico qui succède au séquençage de l'ADN ou de l'ARN d'un organisme unique, d'une colonie de clones (bactériens par exemple), ou encore d'un mélange complexe d'organismes.

- Le problème de l'assemblage peut être comparé à celui de la reconstruction du texte d'un livre à partir de plusieurs copies de celui-ci, préalablement déchiquetées en petits morceaux.

Source : https://fr.wikipedia.org/wiki/Assemblage_(bio-informatique)

Assembleurs

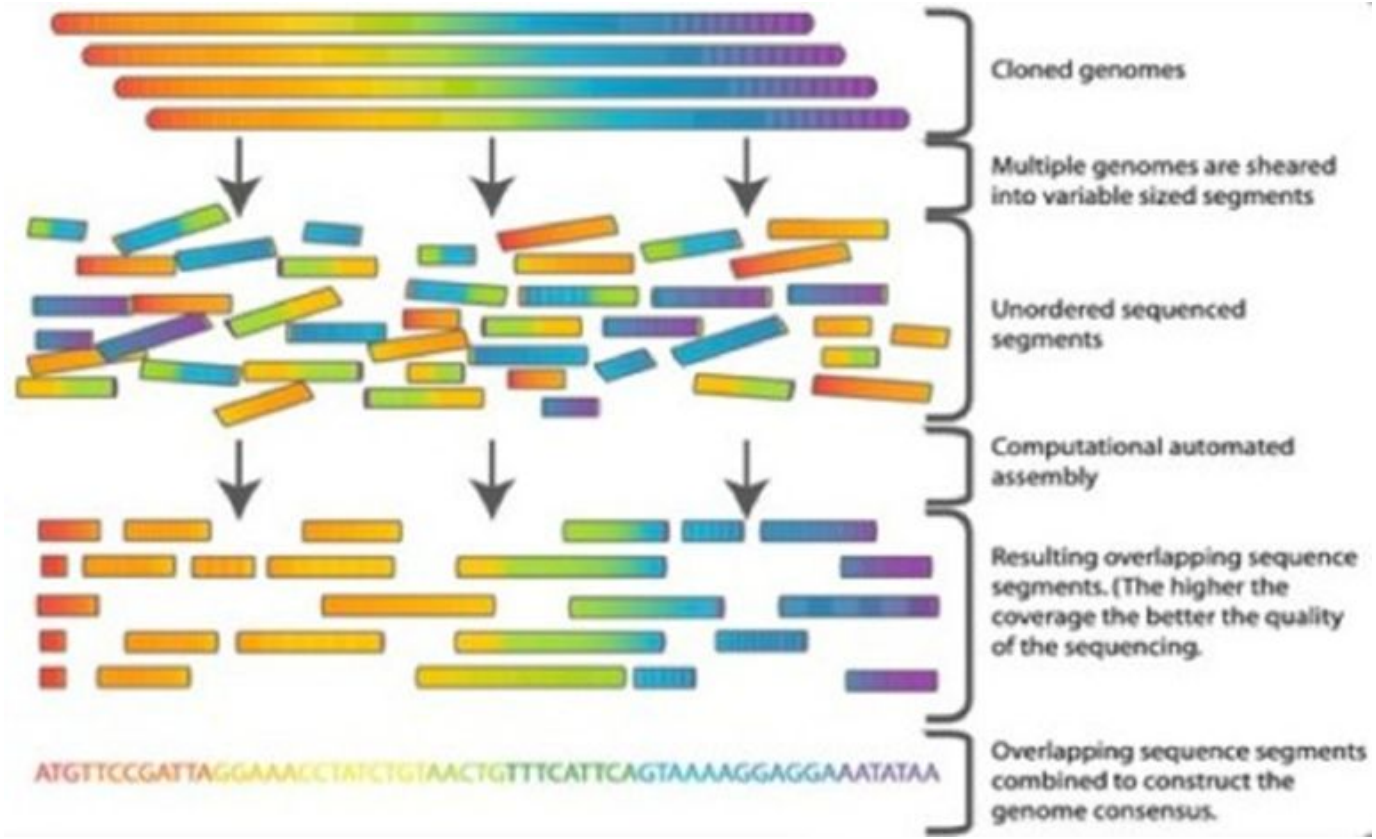
List of notable de-novo assemblers

Name	Type	Technologies	Author	Presented / Last updated	Licence*	Homepage
Velvet	(small) genomes	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007 / 2011	OS	link
SPAdes	(small) genomes, single-cell	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	Bankevich, A et al.	2012 / 2018	OS	link
Plass	Protein-level assembler: assembles six-frame-translated sequencing reads into protein sequences	Illumina	Steinegger, M et al.[7]	2018 / 2018	OS	link
Phrap	genomes	Sanger, 454, Solexa	Green, P.	1994 / 2008	C / NC-A	link
Newbler	genomes, ESTs	454, Sanger	454/Roche	2004/2012	C	link
MaSuRCA	Any size, haploid/diploid genomes	Illumina and PacBio/Oxford Nanopore data, legacy 454 and Sanger data	Zimin A, et al.	2011 / 2018	OS	link
Hinge	Small microbial genomes	PacBio/Oxford Nanopore reads	Kamath et al.[11]	2016 / 2018	OS	link
HGAP	Genomes up to 130 MB	PacBio reads	Chin et al.[8]	2011 / 2015	OS	link
Falcon	Diploid genomes	PacBio reads	Chin et al.[9]	2014 / 2017	OS	link
DNASTAR Lasergene Genomics Suite	(large) genomes, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	DNASTAR	2007 / 2016	C	link
DNA Baser Sequence Assembler	DNA sequence assembly with automatic end trimming & ambiguity correction. Includes a base caller.	Sanger, Illumina	Heracle BioSoft SRL	2018.09	C ($69)	www.DNABaser.com
Canu	Small and large, haploid/diploid genomes	PacBio/Oxford Nanopore reads	Koren et al.[10]	2001 / 2018	OS	link
AFEAP cloning Lasergene Genomics Suite	a precise and efficient method for large DNA sequence assembly	two rounds of PCRs followed by ligation of the sticky ends of DNA fragments	AFEAP cloning	2017 / 2018	C	link

*Licences: OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics

Source : https://en.wikipedia.org/wiki/Sequence_assembly

Reconstitution d'un génome à partir de reads

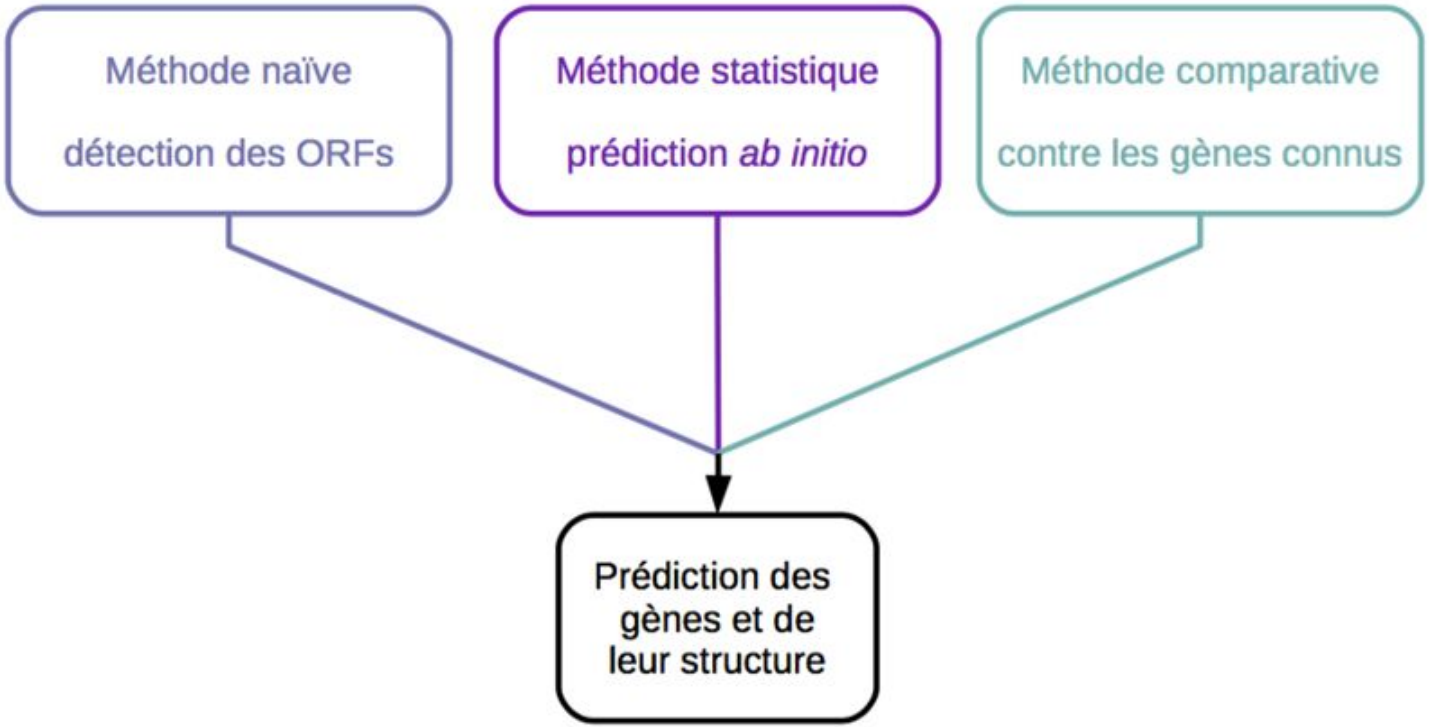


Annotation

- L’annotation d’un génome consiste à analyser la séquence nucléotidique qui constitue l’information brute pour en extraire l'information biologique. Cette analyse poursuit deux objectifs successifs, le premier est de localiser les gènes et les régions codantes et le second est, une fois ces gènes localisés, d'identifier ou de prédire leur fonction biologique. Ces deux étapes reposent initialement sur l'utilisation d'outils algorithmiques sophistiqués, dont le développement constitue l'un des champs de la bio-informatique.
- Pour prédire la fonction potentielle de ces gènes (leur attacher une étiquette, portant leur nom probable, leur fonction probable, leurs interactions probables), on utilise des programmes de recherche d'homologie de séquence.

Source : https://fr.wikipedia.org/wiki/G%C3%A9nome#Annotation_des_g%C3%A9nomes

Méthodes principales



Méthode naïve

- Principe : recherche les signaux des séquences codantes

- Débutent par un codon d'initiation ATG + autres

- Se terminent par un codon de terminaison TAA, TAG ou TGA

- Possèdent une taille multiple de 3

Cas des gènes sans intron

- Mise en œuvre: détecter les phases ouvertes de lecture

- ORFs = Open Reading Frames

- Phases (cadres) pouvant contenir un gène

- >50 nt entre un codon d'initiation et un codon de terminaison

- Traduction à l'aveugle dans les 6 phases de lecture à 3 phases par brin d'ADN

Les 6 phases de lectures d'une séquence nucléique (méthode naïve)

Méthode Naïve

- Avantages :

- méthode ab initio : sans connaissances préalables

- diminue la quantité de données à analyser pour la comparaison de séquences

- Limites :

- toutes les ORFs ne sont pas des gènes
- sensible aux erreurs de déquençage
- peu utile pour les gènes eucaryotes (présence d’introns)

Exemples d’outils d’annotation

- Prodigal,
- Prokka,
- Glimmer,
- GeneMark,
- Maker,
- …
- Utilisation de bases de données de séquences (GenBank, UniProt, …) pour la comparaison et la recherche de fonction.