UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

INTRODUCTION TO *DE NOVO* ASSEMBLY OF GENOMES AND TRANSCRIPTOMES

Laurent Falquet
MA/MER @ UniFr
Group Leader @ SIB

SIB
Swiss Institute of Bioinformatics

Overview

Part 1

What is *de novo* assembly?

Sequencing machines and data

Limitations of the techniques

Limitations of the sequences

Quality Control and Filtering

Algorithms for assembling

- **Greedy**
- **OLC**
- **DeBruijn**

Assembly quality measurements

Part 2

Scaffolding & closing

- **Long reads for assembling**
  - Long reads only
  - Hybrid & polishing (short + long)
- **HiC**
- **Optical maps**

Assembly assessment

Visualization

Transcriptome assembly specificity

Question you should ask yourself when you plan to sequence a full genome (or more than one)…

- **Why?**
- **Size of the genome? (eukaryote *vs* prokaryote)**
- **%GC of the genome?**
- **Are there any known or expected repeats?**
- **Is there a known genome available to serve as a reference?**
- **How far do you want to go?**
  - **Fully assembled genome**
  - **Comparative genomics (SNPs and/or Indels or SVs)**
  - **Partial assembly (draft genome)**
  - **Only contigs searching for specific coding regions (genes)**
  - **Functional annotation and Pathway classification**

Why answering those questions first?

The answers will drive the choice of the technique…
…the amount of time required
…and the costs!

Yes, the sequencing costs decrease every year…
…but don't underestimate the costs of the bioinformatics part:

- **Storage**
- **Calculation**
- **Expertise**

contact your bioinformatician before sequencing!

What is Next Generation Sequencing?

Figure 1. Original genomic DNA is broken into a collection of overlapping fragments

Figure 2. The ends of each fragment (drawn in green) are sequenced

Figure 3.

This is also called "Whole Genome Shotgun Sequencing"

Next Generation Sequencing and Genome Assembly:
a Simple Jigsaw Puzzle?

Yes, but you must deal with

- **Millions of pieces**
- **Lots of malformed pieces**
- **Often missing pieces**
- **Pieces mixed from another puzzle**
- **Lots of identical blue sky pieces…**

If *de novo* you…

Genome assembly, deep blue…

…don't even know the final picture…

Reminder of the sequencing methods

1986 - 2001 | 2005 | 2006 | 2010 | 2011 | 2014 | 2018

ABI Sanger | Roche 454 | Illumina | Life SOLiD | Life Ion Torrent | Pacific Biosciences | Oxford Nanopore

First Generation

Second Generation

Third Generation

REMINDER OF THE SEQUENCING METHODS

Limitations of the NGS techniques

All methods

- **sequencing errors (substitutions or insertions/deletions)**
- **missing data (sampling/coverage bias)**

Roche454, IonTorrent

- **long (>12) mononucleotide repeats**
- **short reads (200bp) for Ion Torrent**

Illumina, MGI

- **short reads (<300bp)**

Pacific Bioscience, Oxford Nanopore

- **high error rate (~15% indels) for long reads (raw data)**
- **Good news: PacBio HiFi reads have much better quality.**

Limitations of the sequences

Repeats

transposases, IS-elements, retroviruses, duplications, etc.

Polymorphisms

SNPs, CNV, multiploid, sample mixture, etc.

Sequence bias

%GC

Repeats are a major issue for all assemblers

Figure top. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

Figure bottom. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs.

HËNJ GENOME IKEA ASSEMBLY INSTRUCTIONS

U

58

Overview of a sequencing project

Overview of a sequencing project

Select DNA and sequencing method

Quality Control

if necessary Trimming/Filtering

De novo assembly with cleaned data

1st step: Assembly

consists mainly in building contigs from short reads

2nd step: Scaffolding (optional)

where contigs are ordered and oriented

3rd step: Finishing or Closing (optional)

where gaps are closed

Genome annotation (draft or final)

Submission to databases & publication (not discussed here)

Overview of a sequencing project

Select DNA and sequencing method

Quality Control

if necessary Trimming/Filtering **(done yesterday)**

De novo assembly with cleaned data

1st step: Assembly

consists mainly in building contigs from short reads

2nd step: Scaffolding (optional)

where contigs are ordered and oriented

3rd step: Finishing or Closing (optional)

where gaps are closed

Genome annotation (draft or final)

Submission to databases & publication (not discussed here)

UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG | Biochemistry Unit | Falquet Laurent

Genome assembly software

3 main algorithm classes

Greedy and other (mixture)

MIRA3, Ray, MaSuRCA, HGAP, Canu…

Overlap Layout Consensus (OLC) assemblers

Newbler, Edena, WGS, CABOG, miniasm, hifiasm, Flye…

deBruijn graph assemblers

ABySS, ALLPATH-LG, SOAP-denovo, (meta)SPAdes, MEGAHIT

...

Greedy

Greedy assemblers - The first assembly programs followed a simple but effective strategy in which the assembler greedily joins together the reads that are most similar to each other.
An example is shown below, where the assembler joins, in order, reads 1 and 2 (overlap = 200 bp), then reads 3 and 4 (overlap = 150 bp), then reads 2 and 3 (overlap = 50 bp) thereby creating a single contig from the four reads provided in the input. One disadvantage of the simple greedy approach is that because local information is considered at each step, the assembler can be easily confused by complex repeats, leading to mis-assemblies.

Leonhard Euler
1707 - 1783

Swiss mathematician
Euler's identity, the most famous formula!

$$e^{i\pi} + 1 = 0$$

Graph theory

In 1736, Euler solved the problem known as the Seven Bridges of Königsberg. The city of Königsberg, Prussia was set on the Pregel River, and included two large islands which were connected to each other and the mainland by seven bridges.

The problem is to decide whether it is possible to follow a path that crosses each bridge exactly once and returns to the starting point. It is not: there is no Eulerian circuit.

This solution is considered to be the first theorem of graph theory, specifically of planar graph theory.

Graph theory

A graph refers to a collection of **vertices** (or '**nodes**') and a collection of **edges** (or '**vectors**') that connect pairs of vertices.

A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another (digraph).

https://en.wikipedia.org/wiki/Graph_theory

Overlap-layout-consensus

Overlap-layout-consensus - The relationships between the reads provided to an assembler can be represented as a graph, where the nodes represent each of the reads and an edge connects two nodes if the corresponding reads overlap. The assembly problem thus becomes the problem of identifying a path through the graph that contains all the nodes - **a Hamiltonian path (Figure below).** This formulation allows researchers to use techniques developed in the field of graph theory in order to solve the assembly problem.
An assembler following this paradigm starts with an **overlap stage** during which all overlaps between the reads are computed and the graph structure is computed. In a **layout stage**, the graph is simplified by removing redundant information. Graph algorithms are then used to determine a layout (relative placement) of the reads along the genome. In a final **consensus stage**, the assembler builds an alignment of all the reads covering the genome and infers, as a consensus of the aligned reads, the original sequence of the genome being assembled.

Overlap graph for a bacterial genome. The thick edges in the picture on the left (a Hamiltonian cycle) correspond to the correct layout of the reads along the genome (figure on the right). The remaining edges represent false overlaps induced by repeats (exemplified by the red lines)

EXERCISE

Here is a set of reads:

```
TACAGT
  CAGTC
   AGTCA
      CAGA
```

1. How many *k*-mers are in these reads (including duplicates), for $k = 3$?
2. How many *distinct* *k*-mers are in these reads?
   - (i) for $k = 2$
   - (ii) for $k = 3$
   - (iii) for $k = 5$
3. Pretend these reads come from the genome TACAGTCAGA. What is the largest k such that the set of distinct *k*-mers in the genome is exactly the set of distinct *k*-mers in the reads above?
4. For any value of k, is there a mathematical relation between N, the number of *k*-mers (incl. duplicates) in a sequence, and L, the length of that sequence?

Eulerian *vs* Hamiltonian path/cycle ?

De Bruijn graph combines both

What happens in case of repeats?

No difference in the graph!

EXERCISE

In this exercise, for simplicity, ignore reverse complements.
Reads:

```
TACAGT
  CAGTC
   AGTCAG
     TCAGA
```

1. Construct the de Bruijn graph for $k = 3$.

   Reminder: edges are k-mers and nodes correspond to k-1 overlaps

2. How many contigs can be created?
3. At which value of k is there a single contig?
4. (optional) Find a mathematical relationship between k_a, the smallest k value with which a genome can be assembled into a single contig (using a de Bruijn graph), and ℓ_r, the length of the longest exactly repeated region in that genome.

Comparison of assemblies with different kmers viewed with Bandage

51-mer assembly
This k-mer size is too small, resulting in a complex and tangled graph with 4618 nodes and 6070 edges.

61-mer assembly
This graph is better than the 51-mer graph – it is much less complex (1357 nodes and 1768 edges) but still has very few dead ends.

Comparison of assemblies with different kmers viewed with Bandage

71-mer assembly
While the complexity of the graph has improved (it has 611 nodes and 765 edges), it now shows many more dead ends.

81-mer assembly
As compared to the 71-mer graph, the complexity has slightly improved (it has 490 nodes and 512 edges), but it has broken into many disconnected parts.

UNIVERSITÉ DE FRIBOURG / UNIVERSITÄT FREIBURG | Biochemistry Unit | Falquet Laurent

Comparison of assemblies with different kmers viewed with Bandage

91-mer assembly
This graph has 2386 nodes and 304 edges and mostly consists of disconnected nodes.
This k-mer size is definitely too large.

Assemblers become faster and more accurate

By adapting the graph to coverage

By applying error correction

By performing local assembly

Use of multiple kmers

Originally proposed by IDBA (Peng et al, 2010), Minia, MEGAHIT and (meta)SPAdes take advantage of combining multiple kmers iteratively.

Assembly quality measurements

Number of contigs

Ideally 1 for a bacterial genome…, but the lower the better

Contig sizes

The larger the better (up to the size of the genome), usually given in maximum, minimum and average lengths.

Correctness

Difficult to assess for a new genome

N50 **The most used quality value for *de novo* assembly**

The N50 is the size of the smallest of all the large contigs covering 50% of the genome

Kmer counting analysis

Allows for error and ploidy detection

N50 what's that?

1) Sort the contigs by size
2) Sum them starting with the largest until you reach 50% of the estimated genome size

Last contig added = N50

How to choose the Kmer parameter for deBruijn graphs?

K =	21	23	25	27	29	31
Velvet N50	12182	43427	67361	66898	66306	**107440**
Abyss N50	15928	25693	29334	30241	**31596**	29797
SOAPdenovo N50	98956	**99286**	82319	82910	84517	52098
SPAdes N50 (combined kmers)	44589	44589	44589	44589	44589	44589

Best scores compared

K =	Velvet31	ABySS29	SOAPdenovo23	SPAdes
Nr contigs	252	1339	213	**182**
Consensus size bp	2936521	3078819	**2825424**	2851989
N50	**107440**	31596	99286	44589
Max	**369778**	132996	252985	173042

However longest contig is not always the best…

Velvet: 369’778bp

ABySS: 88’077bp and 132’996bp

SPAdes: 173’042bp

Kmer counting

Originally proposed via Jellyfish to evaluate the error rate in reads. Reused by KAT (and Merqury) for counting kmers in both reads and genome assemblies can reveal discrepancies and help calculate the ploidy.

Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770 (first published online January 7, 2011) doi:10.1093/bioinformatics/btr011
Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. published online April 22, 2014 doi:10.1101/gr.170720.113

KAT (kmer analysis toolkit)

Visualise the copy number spectra of WGS data compared against an assembly

1st assembly contains most (but not all) the heterozygous content, and introduces more duplications on homozygous content (purple peak)

2nd assembly is more collapsed, including mostly a single copy of the homozygous content and heterozygous content removed (black peak)

Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, Bernardo J Clavijo, KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies, Bioinformatics, Volume 33, Issue 4, 15 February 2017, Pages 574–576, https://doi.org/10.1093/bioinformatics/btw663

Rhie, A., Walenz, B.P., Koren, S. et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245 (2020). https://doi.org/10.1186/s13059-020-02134-9