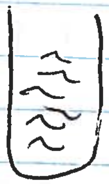
DNA sequencing + Genome sequencing

Goal :



→ File :

```
>seq1
 ACTAGGTA   ] read1
>seq
 TGACTAG. . ] read2
   ⋮
```

Illumina Sequencing : See video.

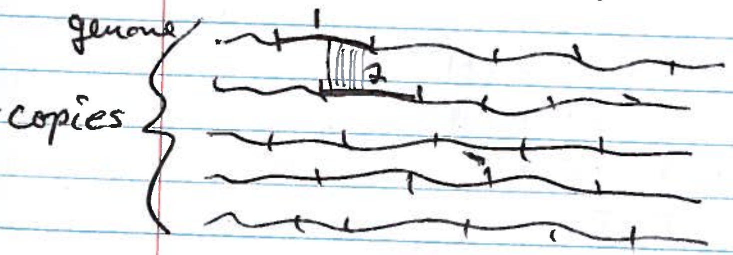
Limitation : - Length of a read is limited (Illumina ≤300bp)

Genome sequencing + Assembly

Goal: Get entire DNA seq. of a long genome

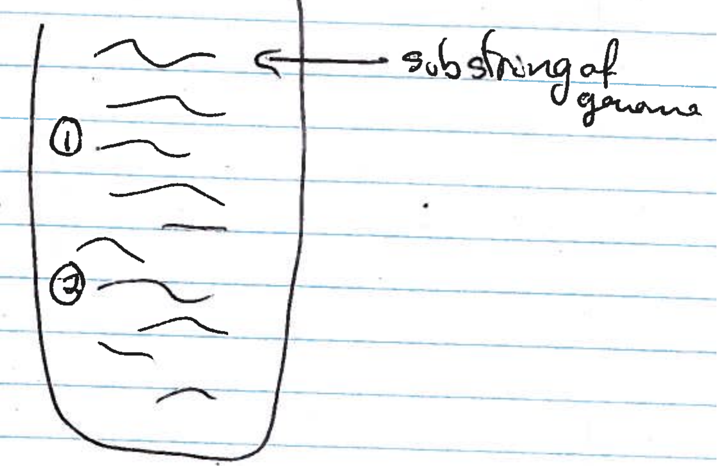
Problem: Seq. machines produce reads that short

Shotgun sequencing



① Generate many copies of genome

② Cut seq. randomly into small pieces
↳ sonication, re~~striction enzyme~~



③ Sequence DNA fragments
ACTAGG
TAGCC...
⋮

④ Assemble reads into genome

True genome : ACTAGCTTCTTAGCCTT
(unknown)

Read1	CTTCTT
Read2	ACTAGC
Read3	AGCCTT
Read4	CTTAGC
Read5	TAGCTT
Read6	TTCTTA

Problem: Shortest Superstring Problem

Given: Set of reads $R_1 \ldots R_n$, of length L = six

Find: Sequence G such that

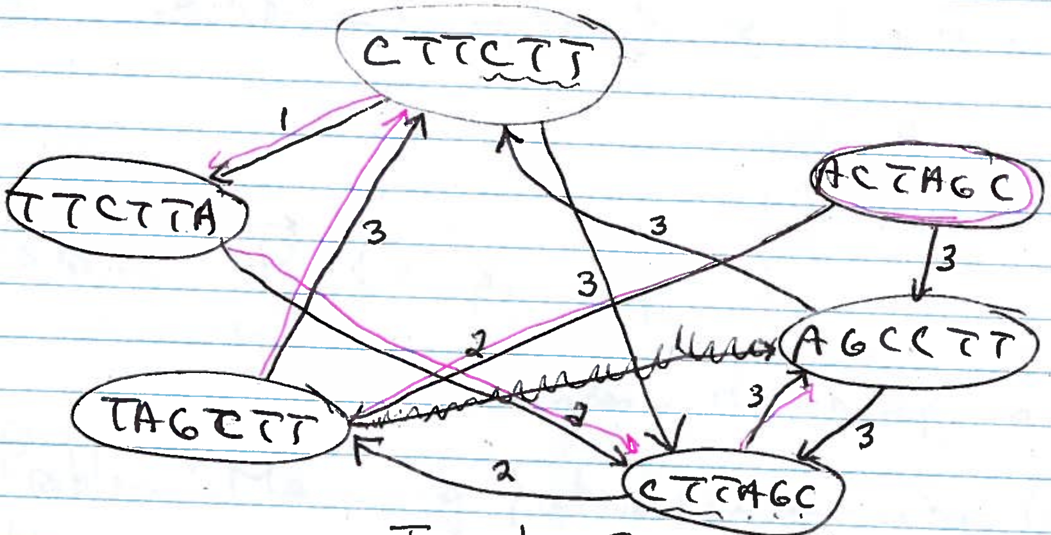
1) Each R_i is a substring of G
2) G is a short as possible

Assumption: No sequencing error.

Overlap-Layout-Consensus Approach

Build Graph: V: set of reads

E: overlaps btw reads of at least $\overset{=3}{k}$ bases



Traveling Salesperson Problem

Goal: Find shortest Hamiltonian Path in G → NP-Complete Problems

shortest → Smallest total weight

Hamiltonian Path → Path that visits each vertex exactly once

```
ACTAGC
  TAGCTT
     CTTCTT
      TTCTTA
        CTTAGC
           AGCCTT
```

weight: 2+3+1+2+3
= 11

Predicted Genome ACTAGCTTCTTAGCCTT