Sequence Assembly

Hammad Naveed

hammad.naveed@nu.edu.pk

The sequencing problem

- We want to determine the identity of the base pairs (ACTG) that make up:
  - A single large molecule of DNA
  - The genome of a single cell
  - The genome of an individual organism
  - The genome of a species

- But we can’t (currently) “read” off the sequence of an entire molecule all at once

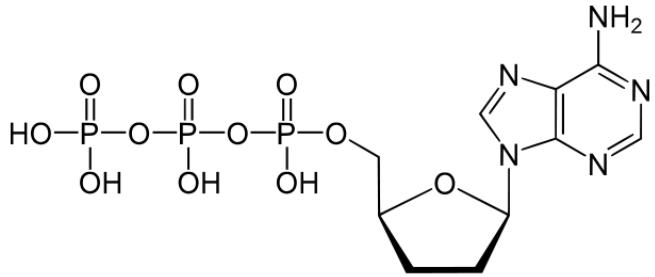
The strategy: substrings

- We *do* have the ability to read or detect *short* pieces (substrings) of DNA
  - Sanger sequencing: 500-700 bp/read
  - Hybridization arrays: 8-30bp/probe
  - Latest technologies:
    - 454 Genome Sequencer FLX: 250-600 bp/read
    - Illumina Genome Analyzer: 35-300 bp/read
    - Pacific Biosciences: ~10,000 bp/read

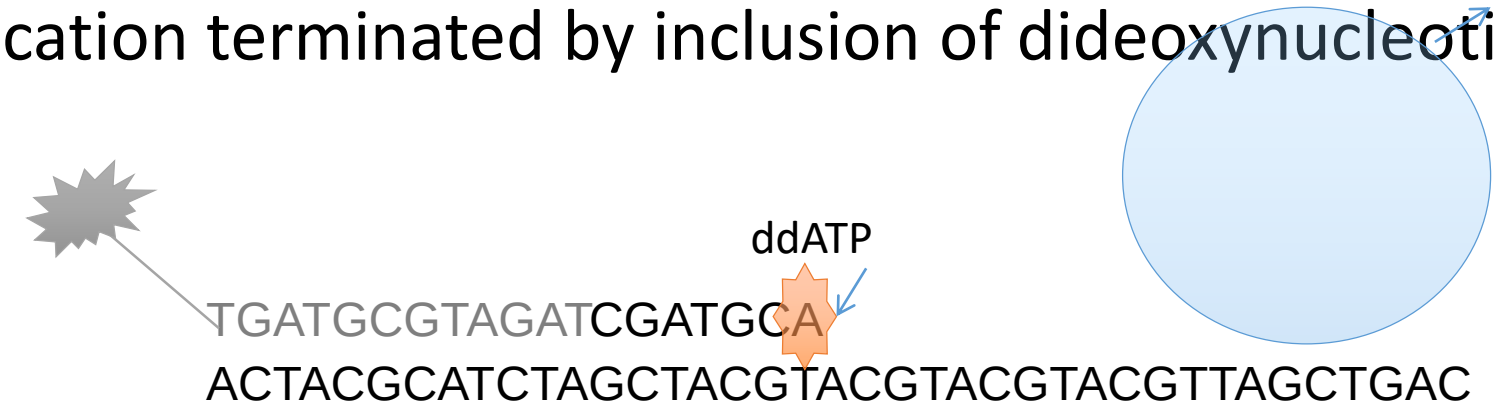
- bp = base pair = letter

Sanger sequencing

- Classic sequencing technique: "Chain-termination method"

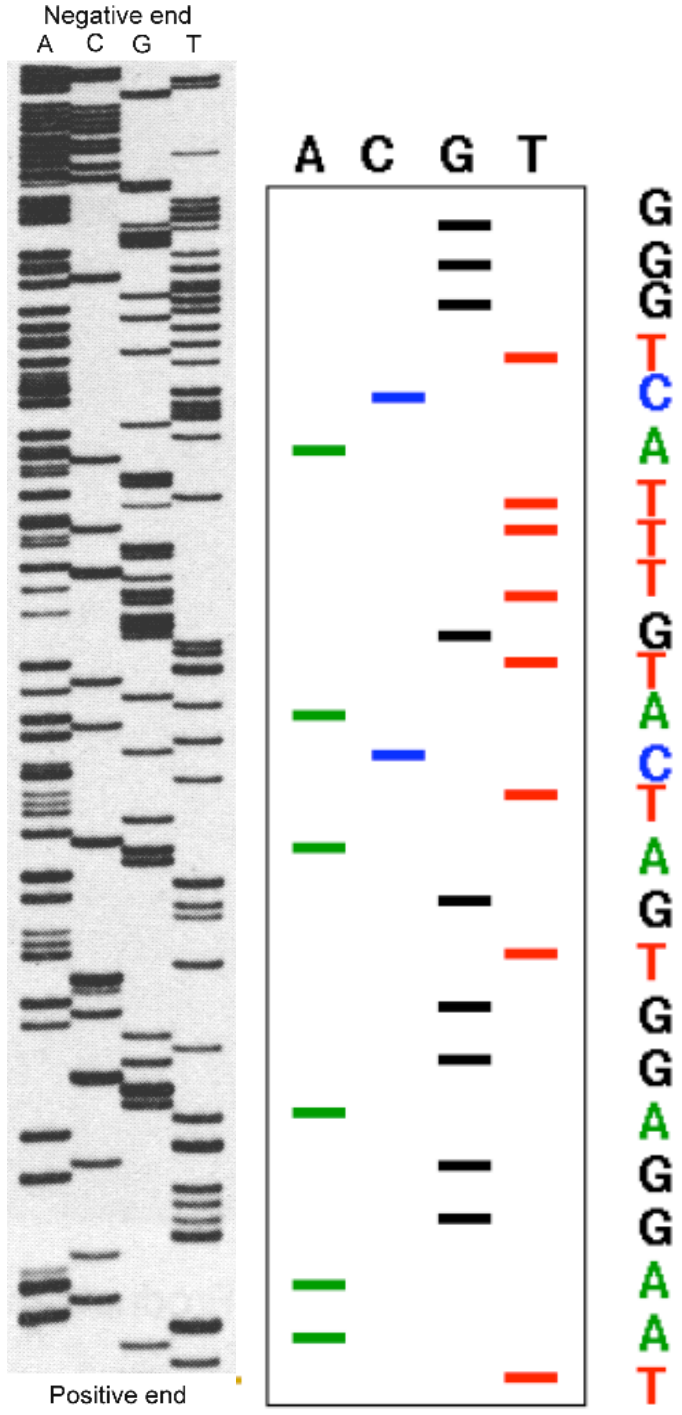


- Replication terminated by inclusion of dideoxynucleotide (ddNTP)



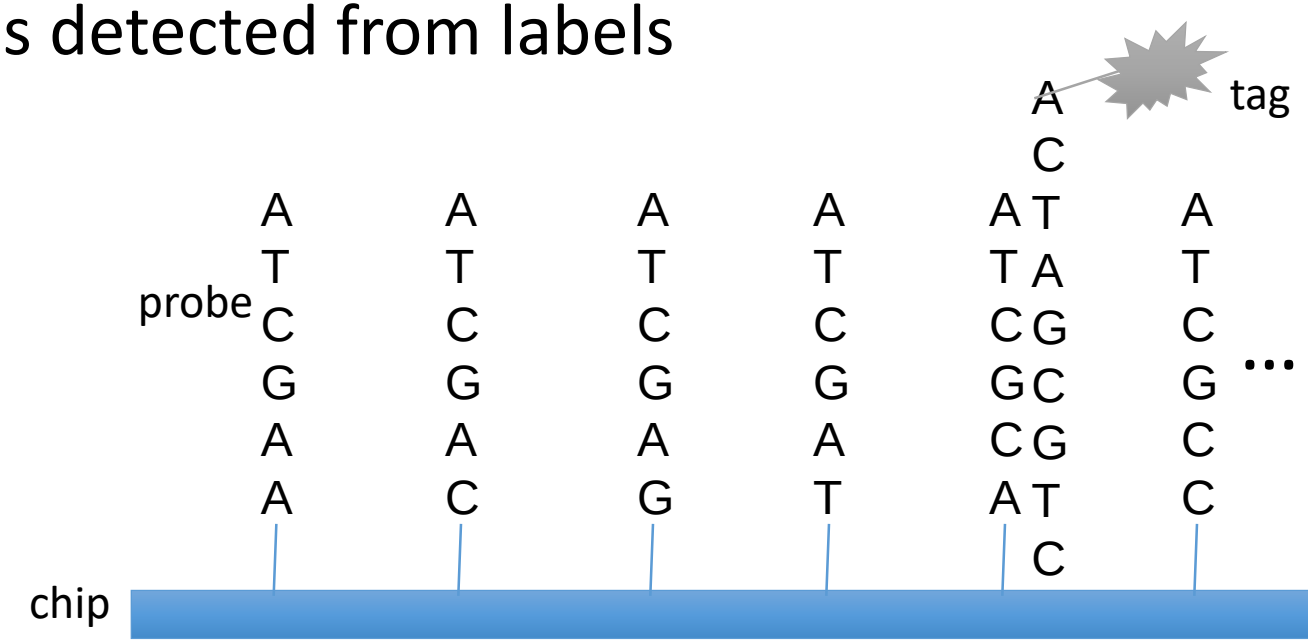
Sequencing gels

- Run replication in four separate test tubes
  - Each with one of some concentration of either ddATP, ddTTP, ddGTP, or ddCTP
- Depending on when ddNTP is included, different length fragments are synthesized
- Fragments separated by length with electrophoresis gel
- Sequence can be read from bands on gel

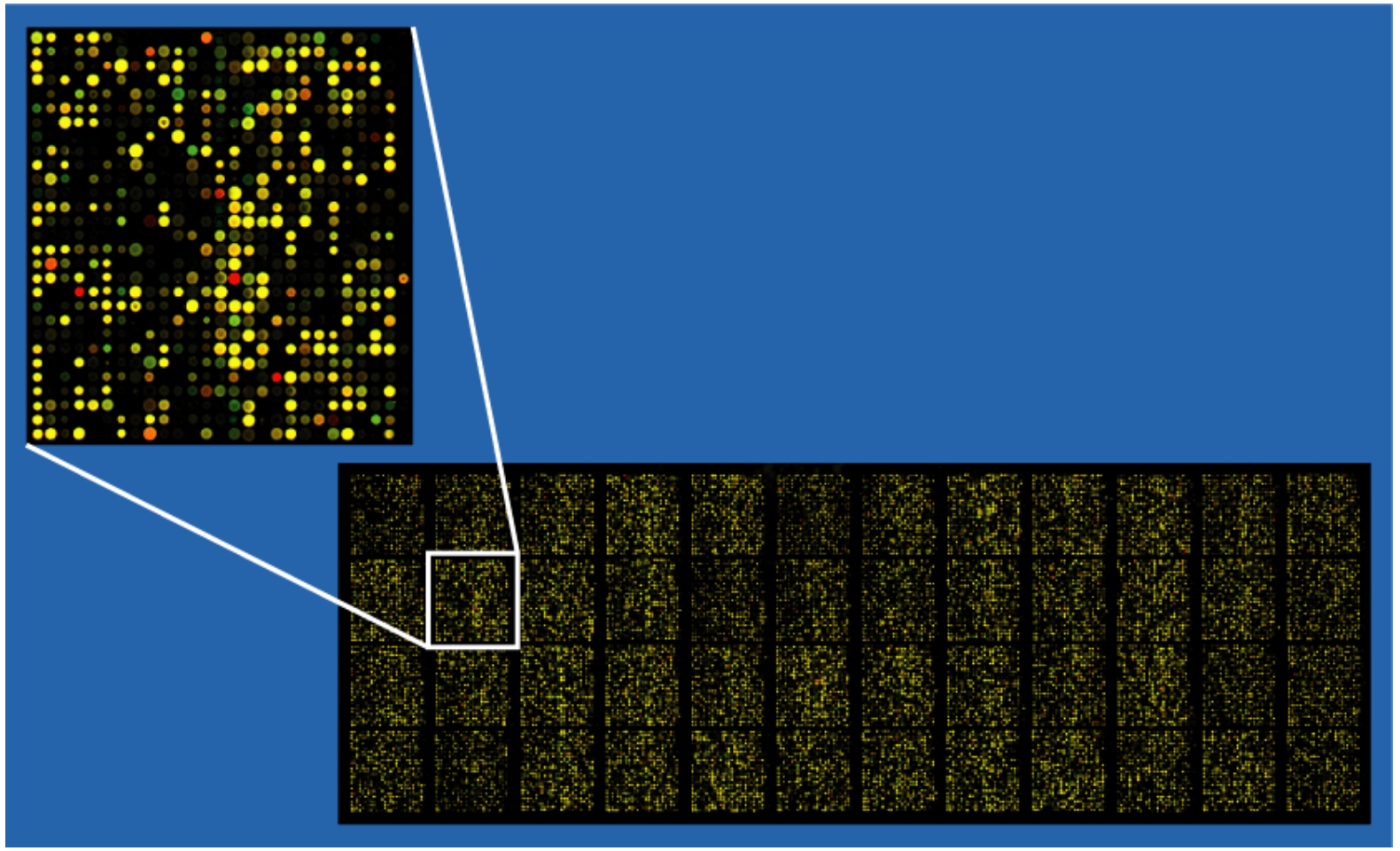


Universal DNA arrays

- Array with all possible oligonucleotides (short DNA sequence) of a certain length as probes
- Sample is labeled and then washed over array
- Hybridization is detected from labels



Reading a DNA array



Latest technologies

- 454
  - “Sequencing by synthesis”
  - Light emitted and detected on addition of a nucleotide by polymerase
  - 400-600 Mb / 10 hour run

- Illumina
  - Also “sequencing by synthesis”
  - ~100 Gb/day on one machine
  - Uses fluorescently-labeled reversible nucleotide terminators
  - Like Sanger, but detects added nucleotides with laser after each step

Latest technologies

- Pacific Biosciences:
  - “Sequencing by synthesis”
  - Single molecule sequencing
  - Detects addition of single fluorescently-labeled nucleotides by an immobilized DNA polymerase
  - Real-time: reads bases at the rate of DNA polymerase
  - 4 hours for sequencing with reads up to 60kb long
  - Video

Latest technologies

- Oxford Nanopore
  - Emerging technology
  - Pocket-sized
  - Higher error rate

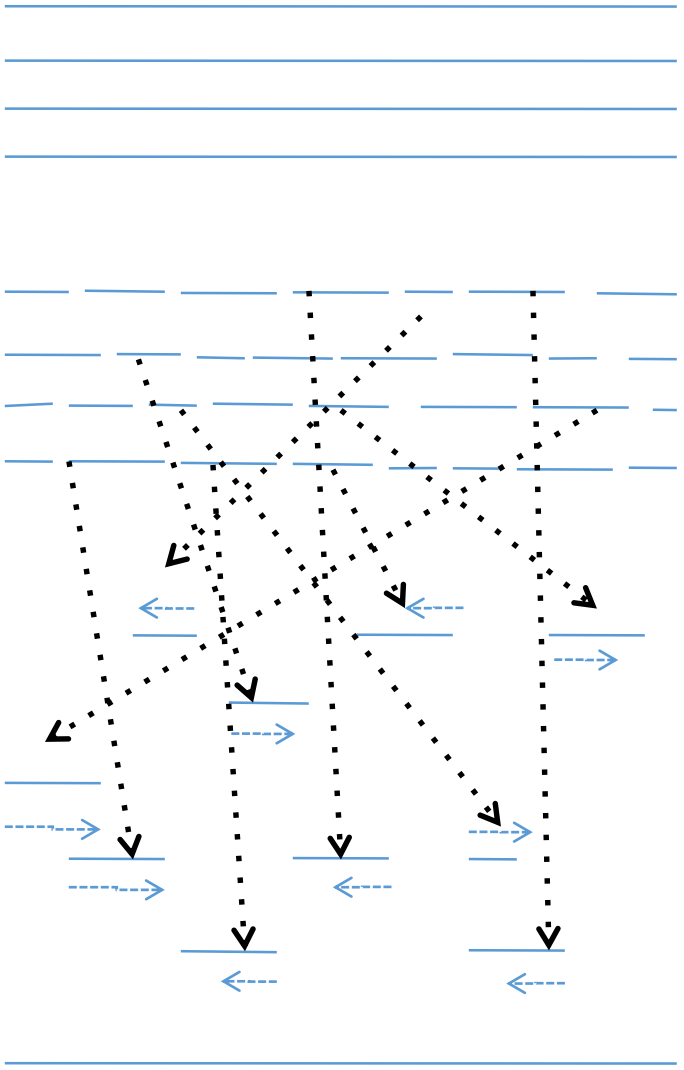
Shotgun Sequencing Fragment Assembly

Multiple copies of sample DNA

Randomly fragment DNA

Sequence sample of fragments

Assemble reads



The fragment assembly problem

- Given: A set of reads (strings) $\{s_1, s_2, \ldots, s_n\}$
- Do: Determine a large string *s* that "best explains" the reads

- What do we mean by "*best explains*"?
- What *assumptions* might we require?

Shortest superstring problem

- Objective: Find a string *s* such that
  - all reads s_1, s_2, … , s_n are substrings of *s*
  - *s* is as short as possible

- Assumptions:
  - Reads are 100% accurate
  - Identical reads must come from the same location on the genome
  - “best” = “simplest” (Ockham's razor)

Shortest superstring example

- Reads:

  {ACG, CGA, CGC, CGT, GAC, GCG, GTA, TCG}

- Shortest superstring (length 10)

TCGACGCGTA

TCG
 CGA
  GAC
   ACG
    CGC
     GCG
      CGT
       GTA

Algorithms for shortest substring problem

- This problem turns out to be *NP*-complete
- Simple *greedy* strategy:

  while # strings > 1 do

  merge two strings with maximum overlap

  loop
- Conjectured to give string with

  length ≤ 2 × minimum length
- “2-approximation”
- Other algorithms will require *graph theory…*