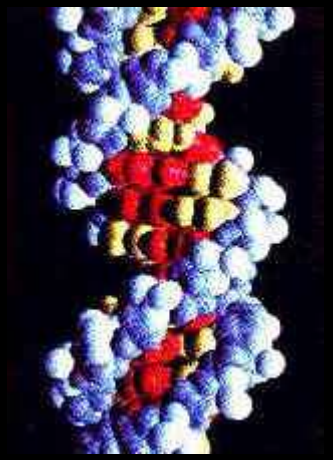
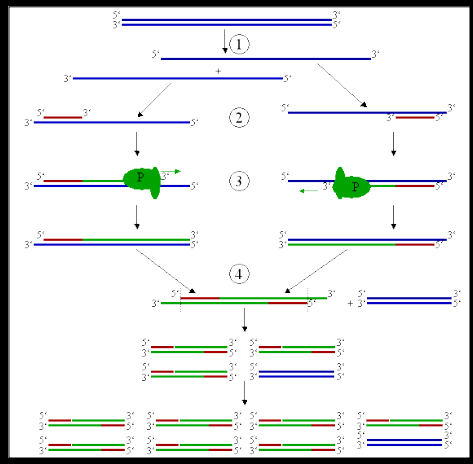
A4G

Sequencing genomes

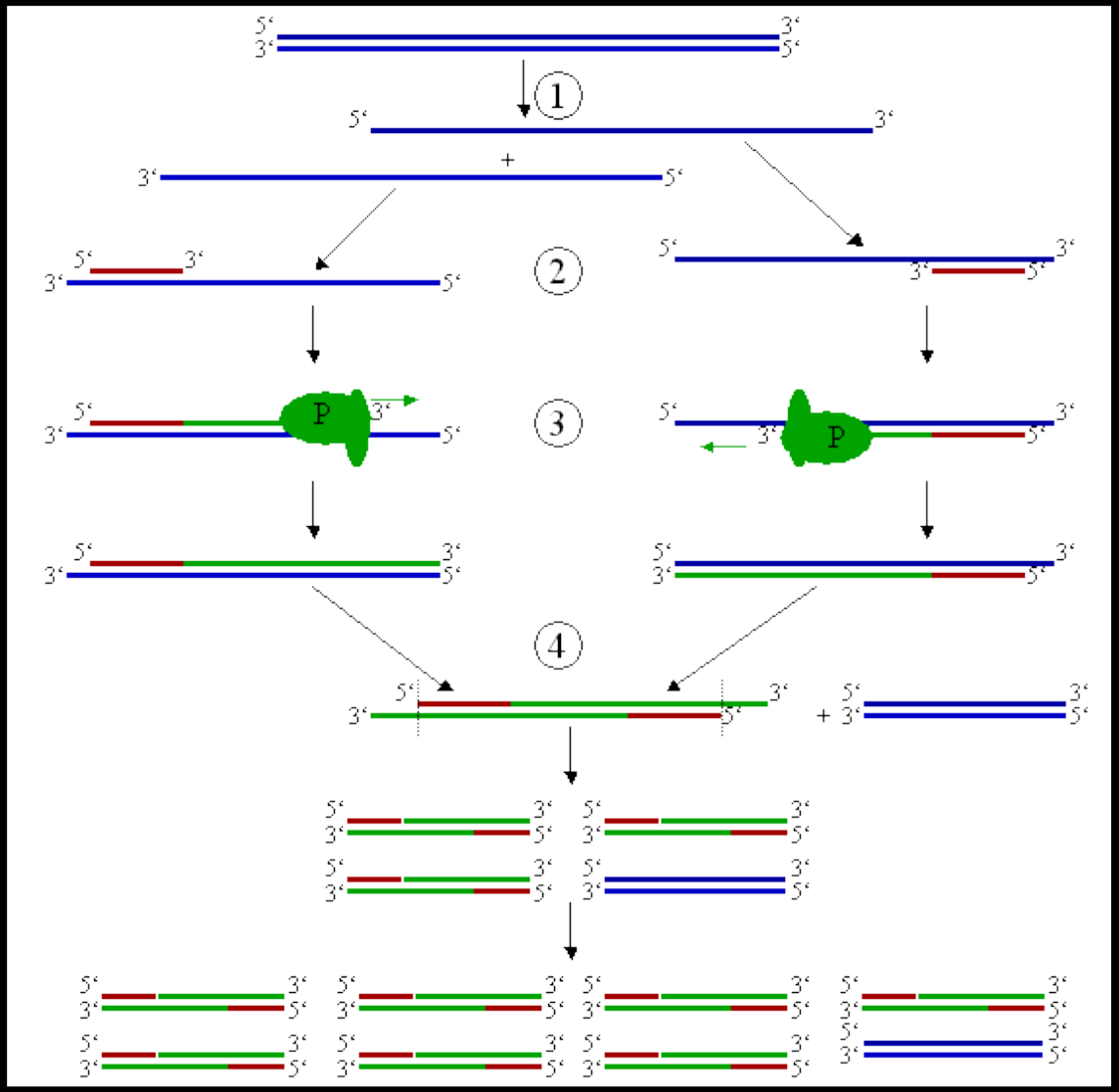


Polymerase chain reaction

- PCR is a molecular biological technique for creating large amount of DNA
- We need:
  - DNA template
  - Two primers
  - DNA-Polymerase
  - Nucleotides
  - Buffer - a suitable chemical environment



Polymerase chain reaction

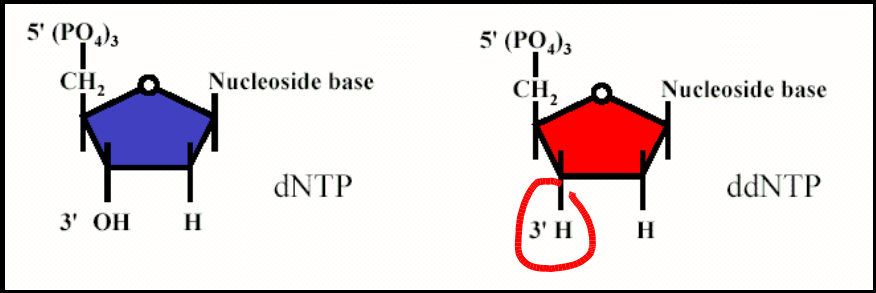


http://en.wikipedia.org/wiki/Polymerase_chain_reaction

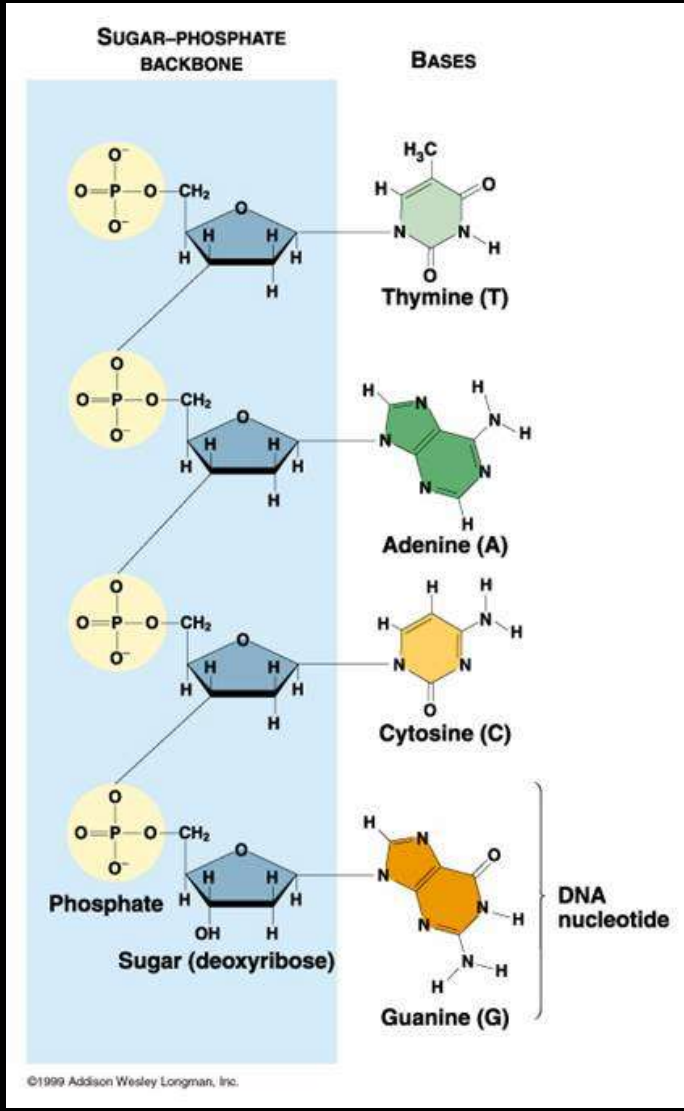
DNA Sequencing

- Chain Termination Method
  - Sanger, 1977
  - single stranded DNA, 500-700b
  - Method:
    - Electrophoresis can separate DNA molecules differing 1bp in length
    - Dideoxynucleotide (*ddNTP*) are used - which stop replication

ddNucleotides

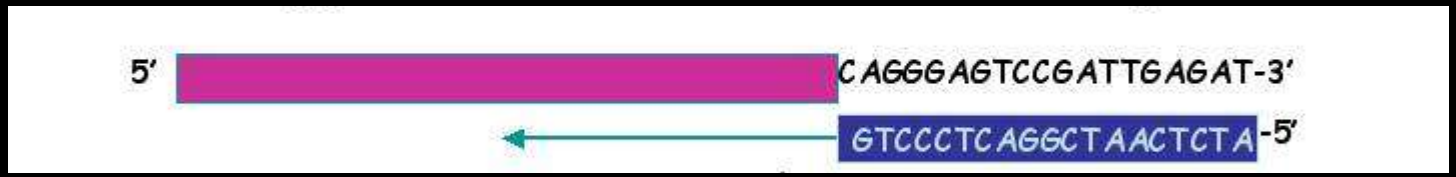


- ddA, ddT, ddC, ddG
- Each type marked with fluorescent dye
- When incorporated into DNA chain – stops replication



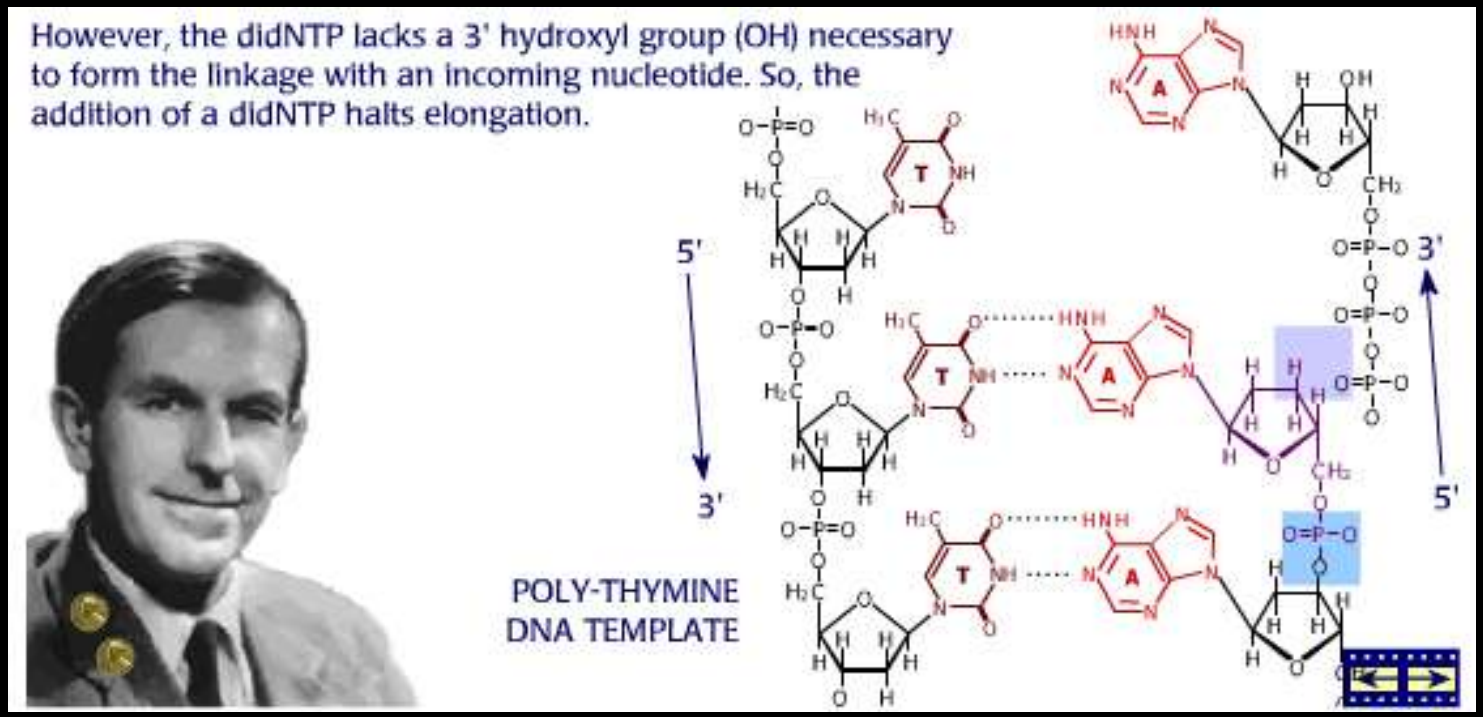
Chain Termination Method, An Outline

- Start four separate replications reactions
  - first obtain single stranded DNA
  - add a (universal) primer
- Start each replications in a soup of A,T,C,G

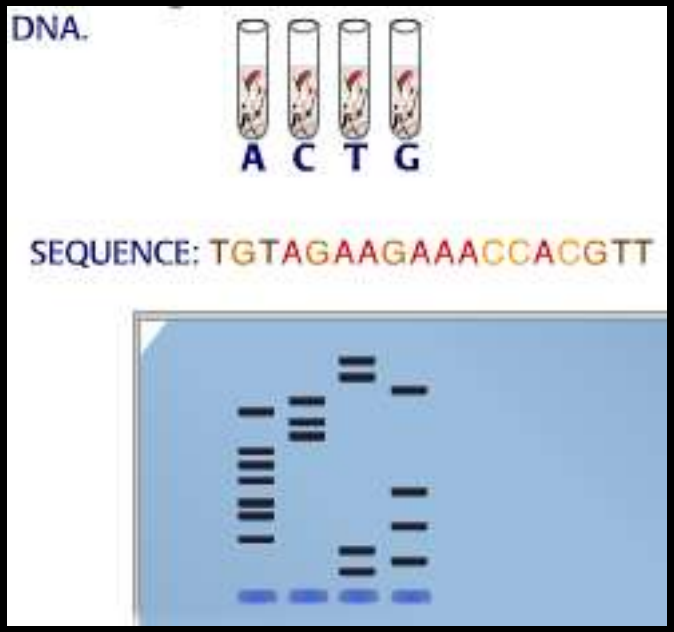


Chain Termination Method, An Outline

- add tiny amounts of
  - ddA to the first reaction,
  - ddT to second, ddC 3rd, ddG 4th

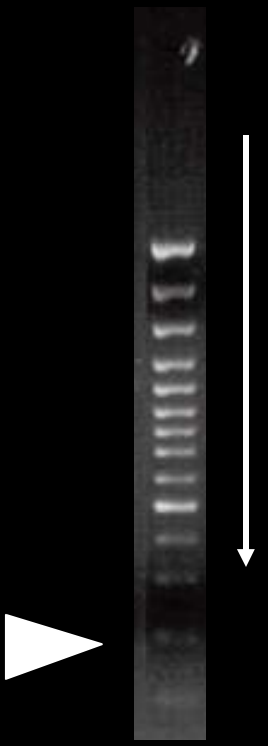


Chain Termination Method A read



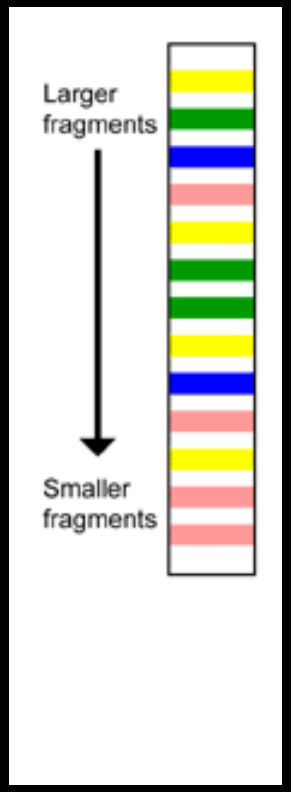
Chain Termination Method, Reading the Sequence

- Recent improvements:
  - one reaction and Four types of ddNTP have four different fluorescent labels
  - automated reading



See: www.dnai.org/timeline/index.html -> 70s -> DNA sequencing

Chain Termination Method, Results



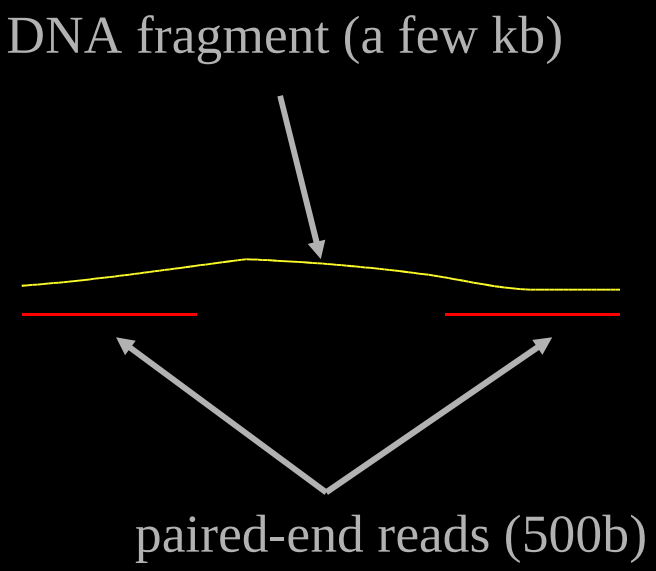
Signal

time

fragment size

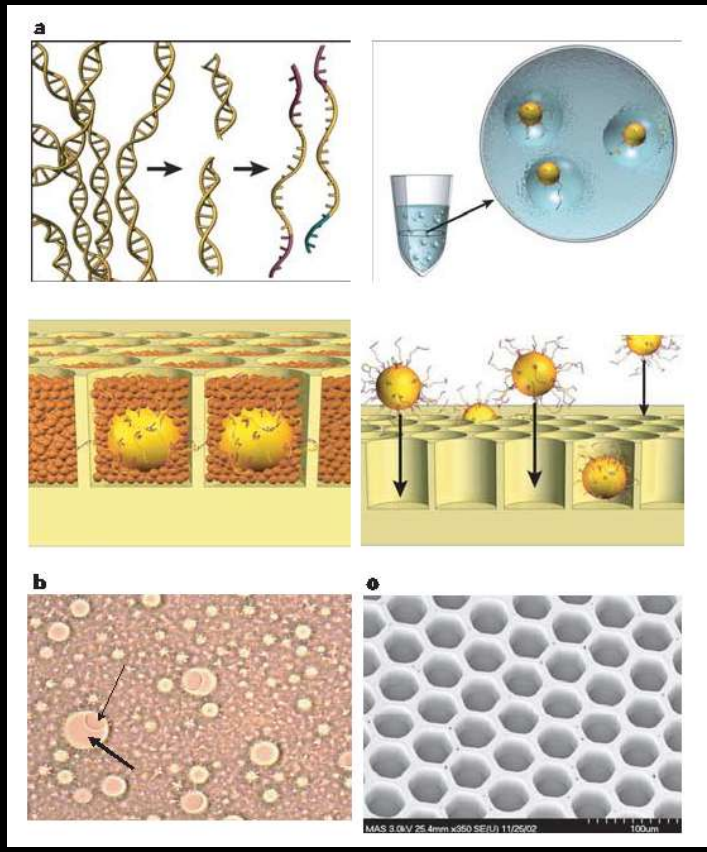
www.newscientist.com

Paired–end reads



Massively parallel picolitre–scale sequencing: 454

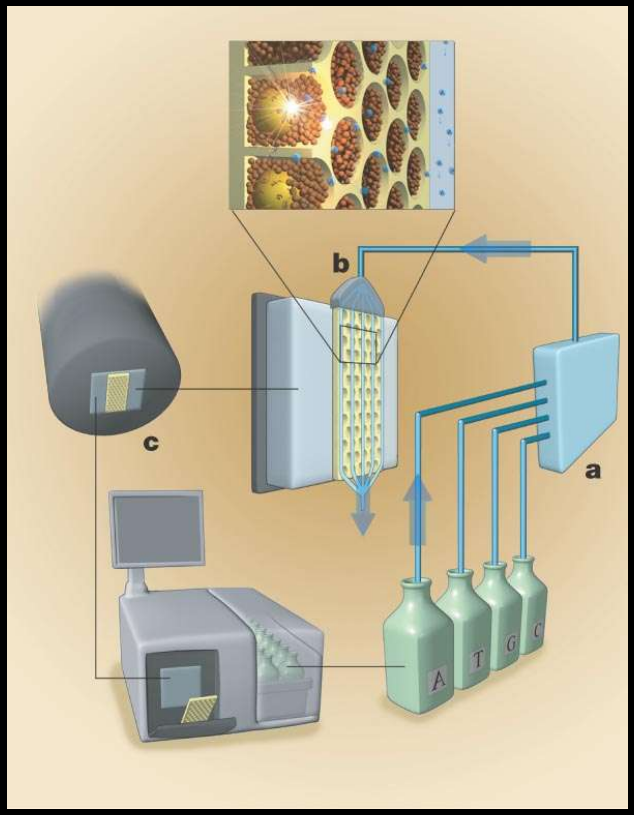
- fragment single strand DNA (ssDNA)
- fragments bound to beads (1 f/bead)
- replication in oil droplets
  - 1 bead/droplet
  - 10mln copies/bead
- beads are deposited in 1.6mln microscopic wells



Margulies *et al.*, Nature Vol 437, 15 September 2005, doi:10.1038/nature03959

Massively parallel picolitre–scale sequencing: 454

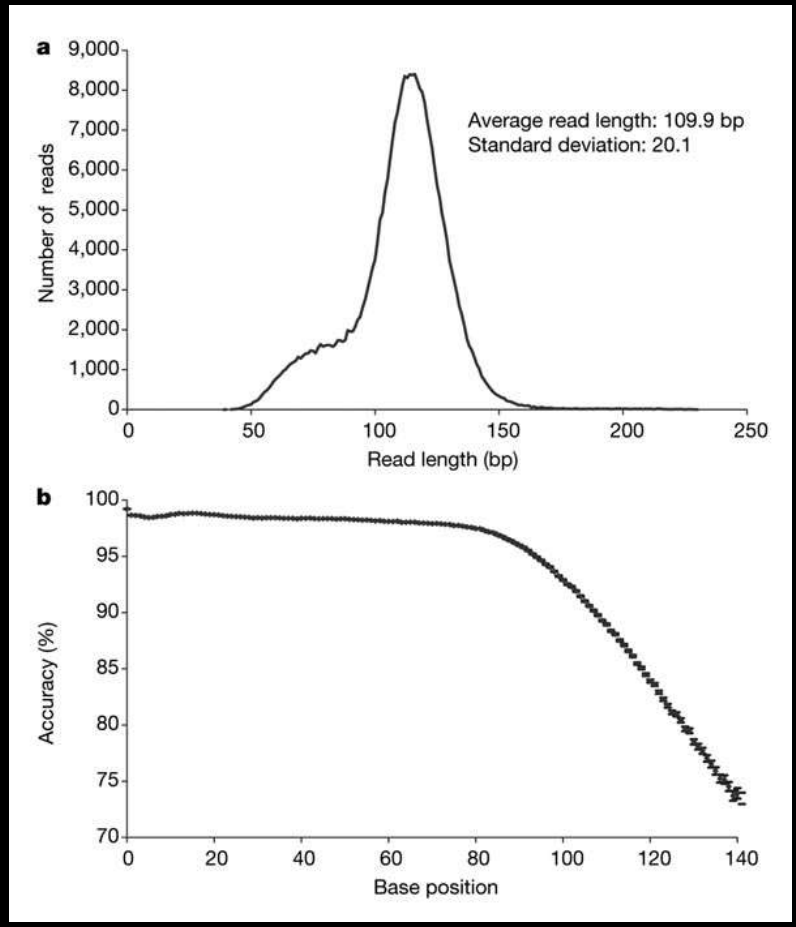
- ssDNA (ready to make a complement) in each well
- sequencing-by-synthesis
  - wash the plate with special nucleotides
  - emits light when DNA grows
  - record on the camera



Margulies *et al.*, Nature Vol 437, 15 September 2005, doi:10.1038/nature03959

454 – results

- advantages
  - 100x faster (25mln nucleotides/h)
  - 1 operator
- disadvantages
  - short reads
  - accuracy



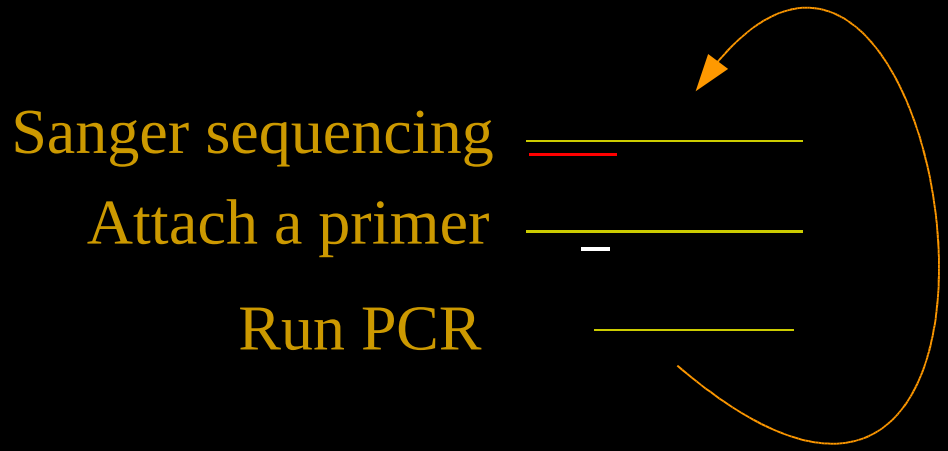
Margulies *et al.*, Nature Vol 437, 15 September 2005, doi:10.1038/nature03959

Sequencing methods

- Directed
- Top-down (hierarchical)
- Bottom-up (shotgun)

Directed sequencing 1

- *Primer walking* using PCR



Directed sequencing 2

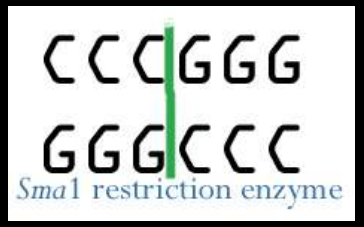
- *Nested deletion*
  - cut DNA with exonuclease
  - “eats up” DNA from an end
    - one bp at a time
    - either 3' or 5'

Directed sequencing summary

- sequential
- gets stuck
- used for short seqs (~tens kb)

Restriction enzymes

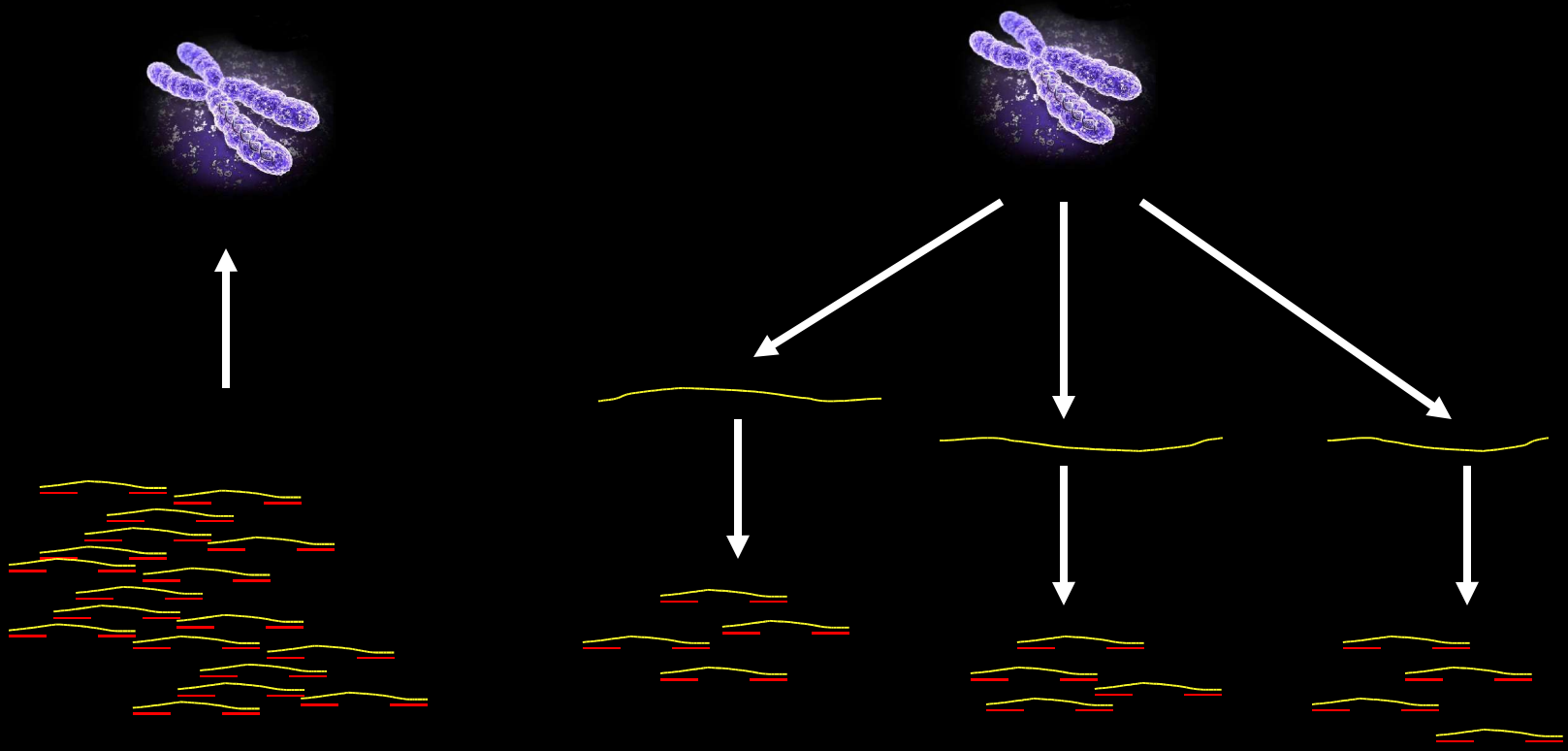
- proteins
- cut DNA
- at a specific pattern



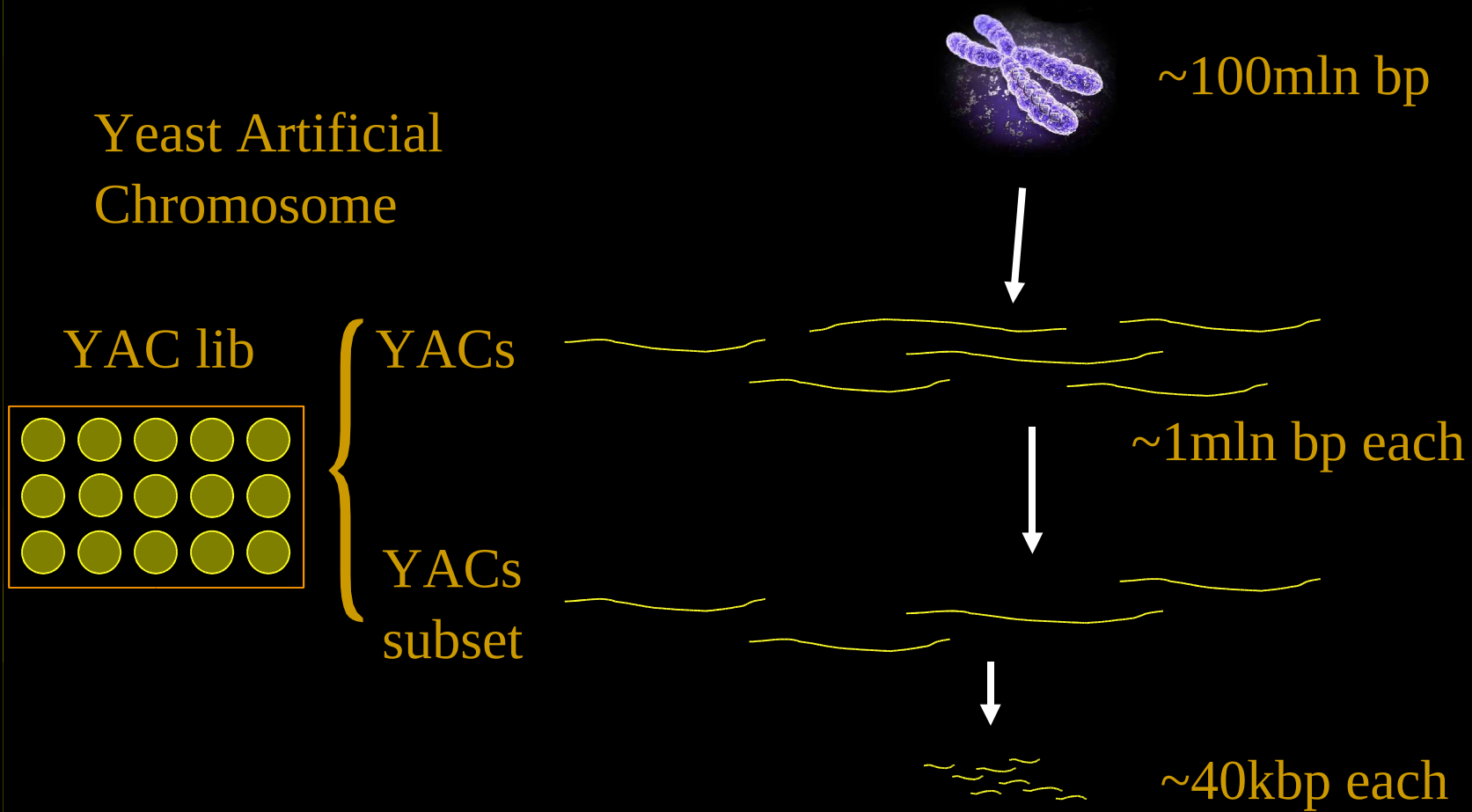
http://en.wikipedia.org/wiki/Restriction_enzymes

Shotgun vs. Hierarchical Method

- Shotgun
  - bottom-up
- Hierarchical
  - top-down

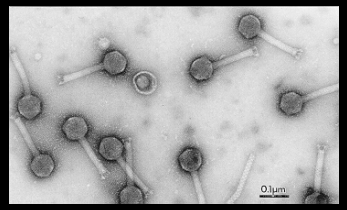


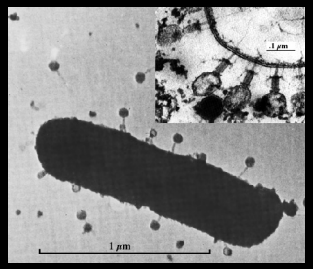
Hierarchical sequencing



Hierarchical sequencing

~40kbp each

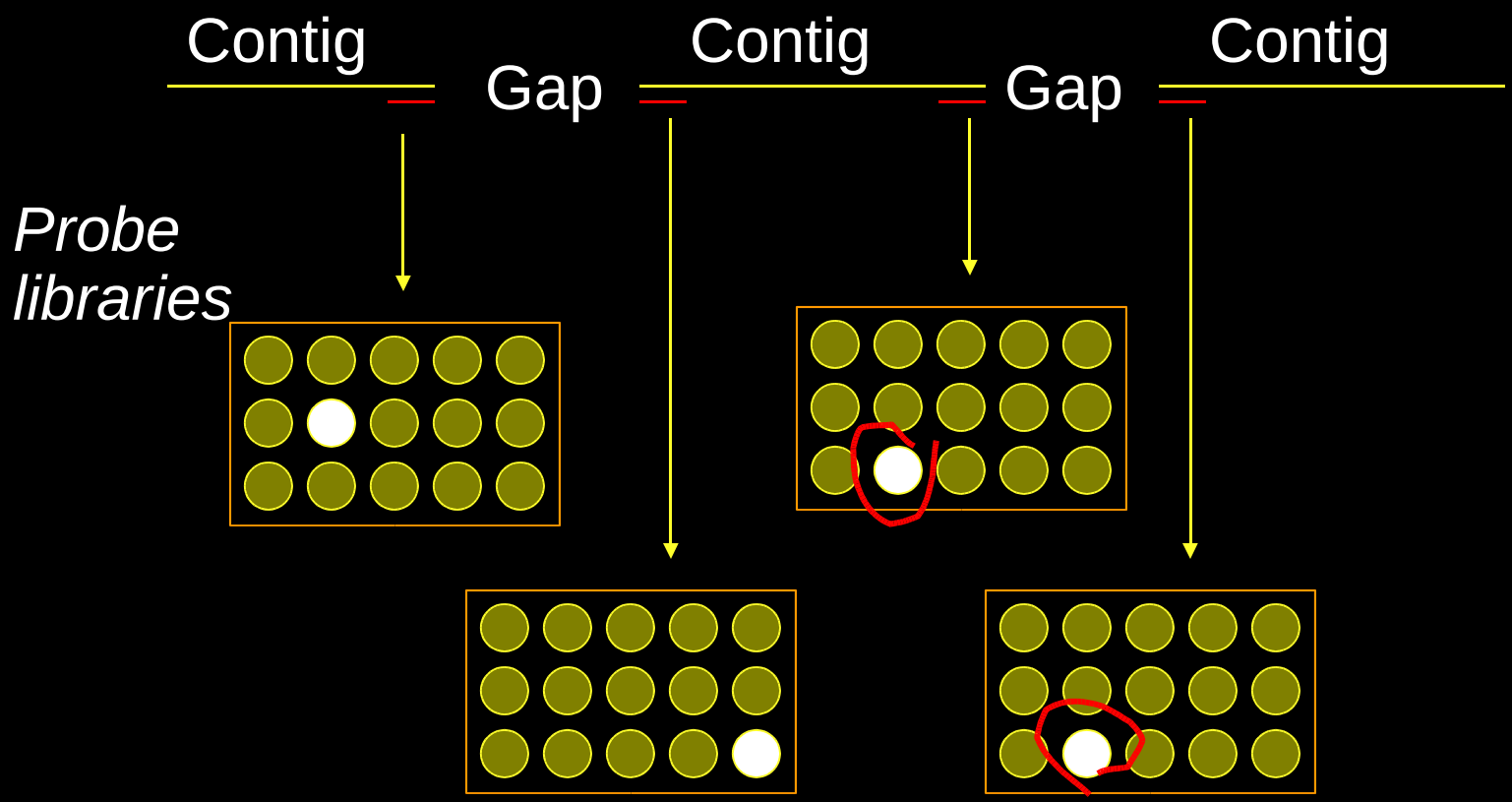




BACs

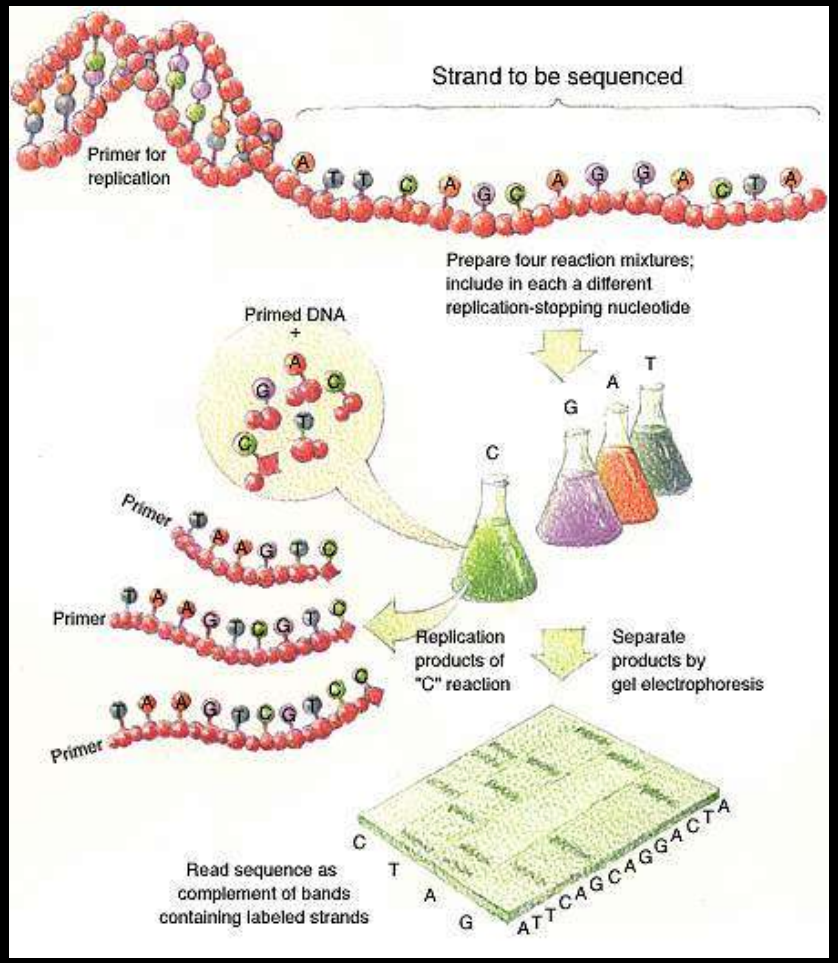
sequencing (easy, short sequence)

Filling in gaps

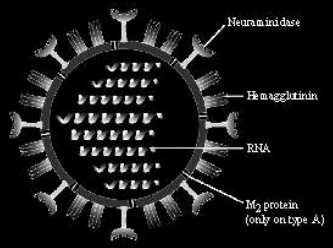


Shotgun DNA Sequencing

- Shear DNA into millions of small fragments
- Read 500 – 700 nucleotides at a time from the small fragments (Sanger method)

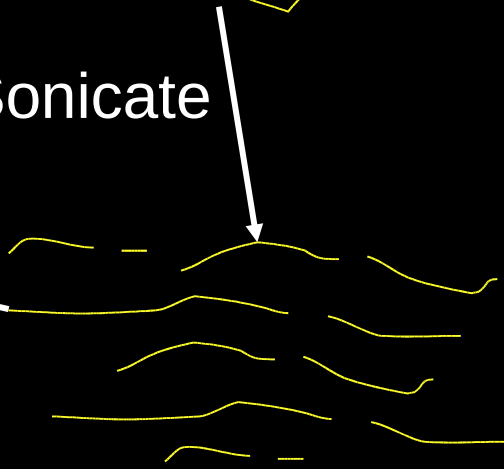


Shotgun Method – Haemophilus Influenzae Sequencing

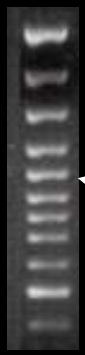


Extract DNA

Sonicate

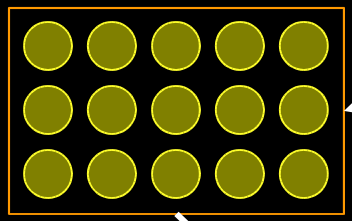


Electrophoresis

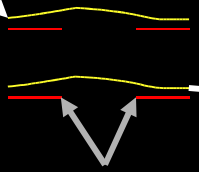


1.5-2kb

DNA library



Sequence

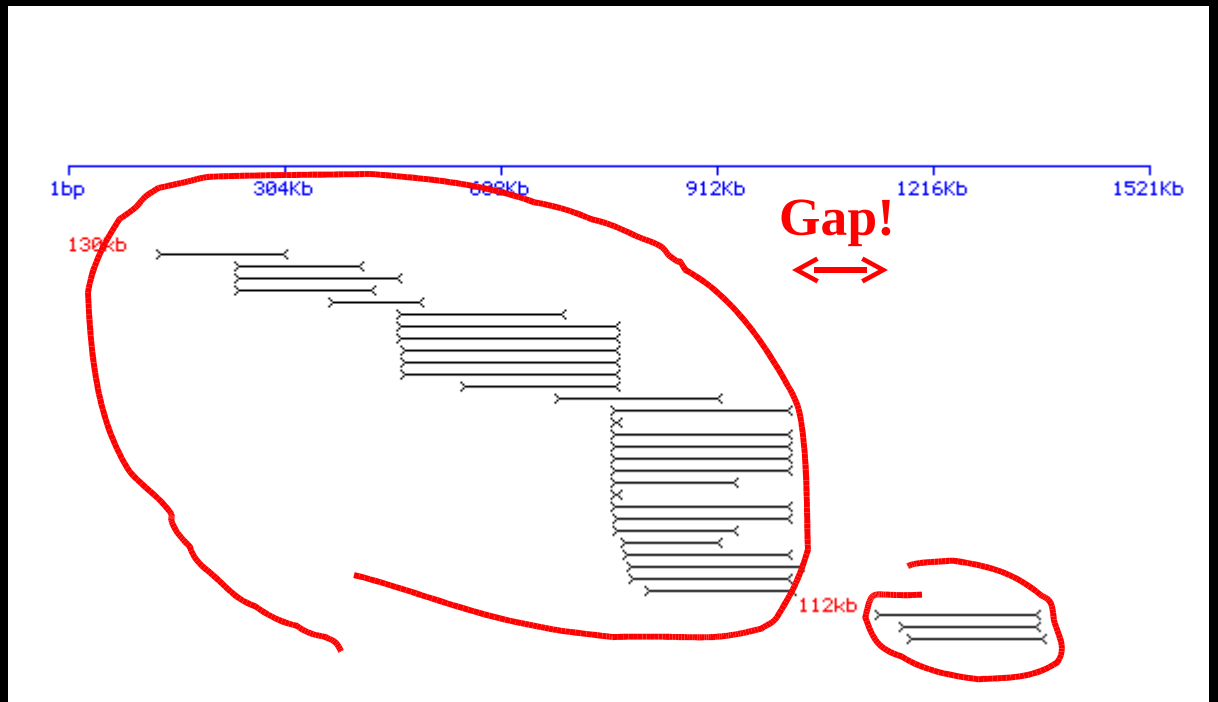


paired-end reads

Construct seq

A contig

- **Contig** – a continuous set of overlapping sequences



Read Coverage

Length of genomic segment: ***L***

Number of reads: ***n*** Coverage ***C = n l / L***

Length of each read: ***l***

How much coverage is enough?

Lander-Waterman model:

Assuming uniform distribution of reads, *C*=10 results in 1 gapped region per 1,000,000 nucleotides

Shotgun Method – Pros and Cons

- Pros
  - Human labour reduced to minimum
- Cons
  - Computationally demanding – $O(n^2)$ comparisons
  - High error rate in contig construction
    - Repeats as the main problem

Shotgun vs. Hierarchical Method

- Celera vs. Human Genome Project
- Hierarchical (top-down) assembly:
  - The genome is carefully mapped
  - "Shotgun" into large chunks of 150kb
    - Exact location of each chunk is known
  - Each piece is again "shotgun" into 2kb and sequenced

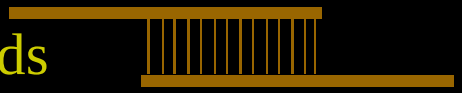
Assembling the genome

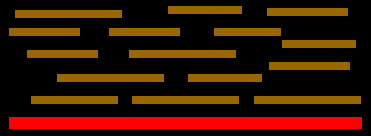
- Given a set of (short) fragments from shotgun sequencing...
  - find overlap between all pairs
  - → find the order of reads in DNA
  - determine a consensus sequence

Assembling the genome: Overlap–Layout–Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

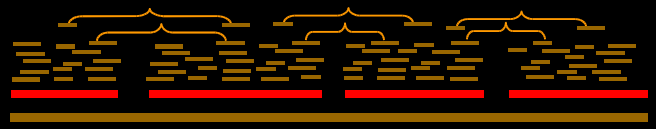
Overlap: find potentially overlapping reads





Layout: merge reads into contigs and

contigs into supercontigs



Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

Fragment Assembly

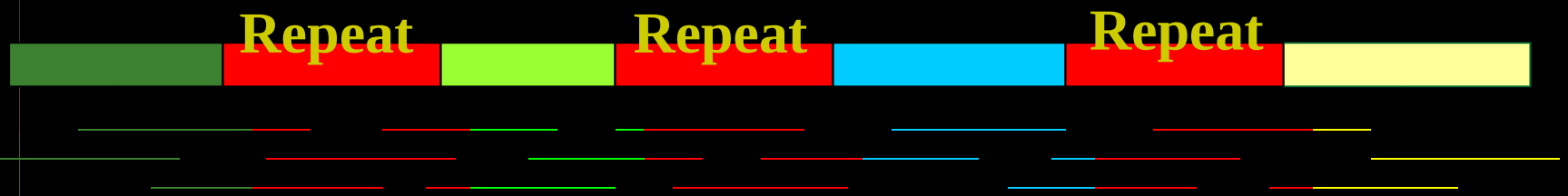
- **Computational Challenge:** assemble individual short fragments (reads) into a single genomic sequence (“contig”)
- Until late 1990s the shotgun fragment assembly of human genome was viewed as intractable problem

Challenges in Fragment Assembly

- Repeats: A **major** problem for fragment assembly
- > 50% of human genome are repeats:

  - over 1 million *Alu* repeats (about 300 bp)

  - about 200,000 LINE repeats (1000 bp and longer)



Green and blue fragments are interchangeable when assembling repetitive DNA

Repeat Types

- *Low-Complexity DNA* (e.g. ATATATATACATA…)
- *Microsatellite repeats* $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- *Transposons/retrotransposons*
  - **SINE** Short Interspersed Nuclear Elements
    (e.g., *Alu*: ~300 bp long, 10^6 copies)
  - **LINE** Long Interspersed Nuclear Elements
    ~500 - 5,000 bp long, 200,000 copies
  - **LTR retroposons** Long Terminal Repeats
    (~700 bp) at each end
- *Gene Families* genes duplicate & then diverge
- *Segmental duplications* ~very long, very similar copies

Paired-end reads help to resolve repeat order

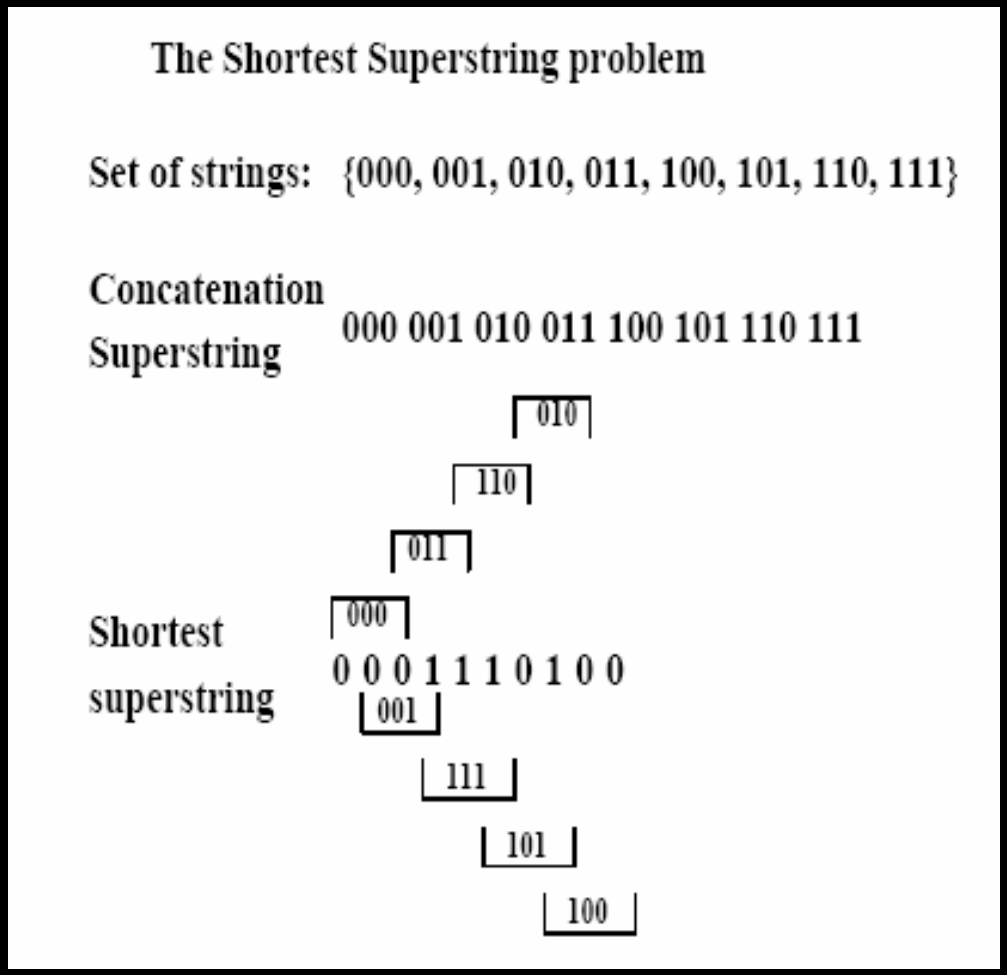
Shortest Superstring Problem

- Problem: Given a set of strings, find a shortest string that contains all of them
- Input: Strings $s_1, s_2, \ldots, s_n$
- Output: A string s that contains all strings

  $s_1, s_2, \ldots, s_n$ as substrings, such that the length of s is minimized

- **Complexity:** NP – hard
- **Note:** this formulation does not take into account sequencing errors

Shortest Superstring Problem: Example



Reducing SSP to TSP

- Traveling Salesman Problem
- Define *overlap* (s_i, s_j) as the length of the longest prefix of s_j that matches a suffix of s_i.

aaaggcatcaaatctaaaggcatcaaa

aaaggcatcaaatctaaa

What is** overlap (s_i, s_j) **for these strings?

Reducing SSP to TSP

- Define *overlap* (s_i, s_j) as the length of the longest prefix of s_j that matches a suffix of s_i.

aaaggcatcaaatctaaaggcatcaaa

aaaggcatcaaatctaaa

aaaggcatcaaatctaaa

overlap=12

Reducing SSP to TSP

- Define *overlap* (s_i, s_j) as the length of the longest prefix of s_j that matches a suffix of s_i.

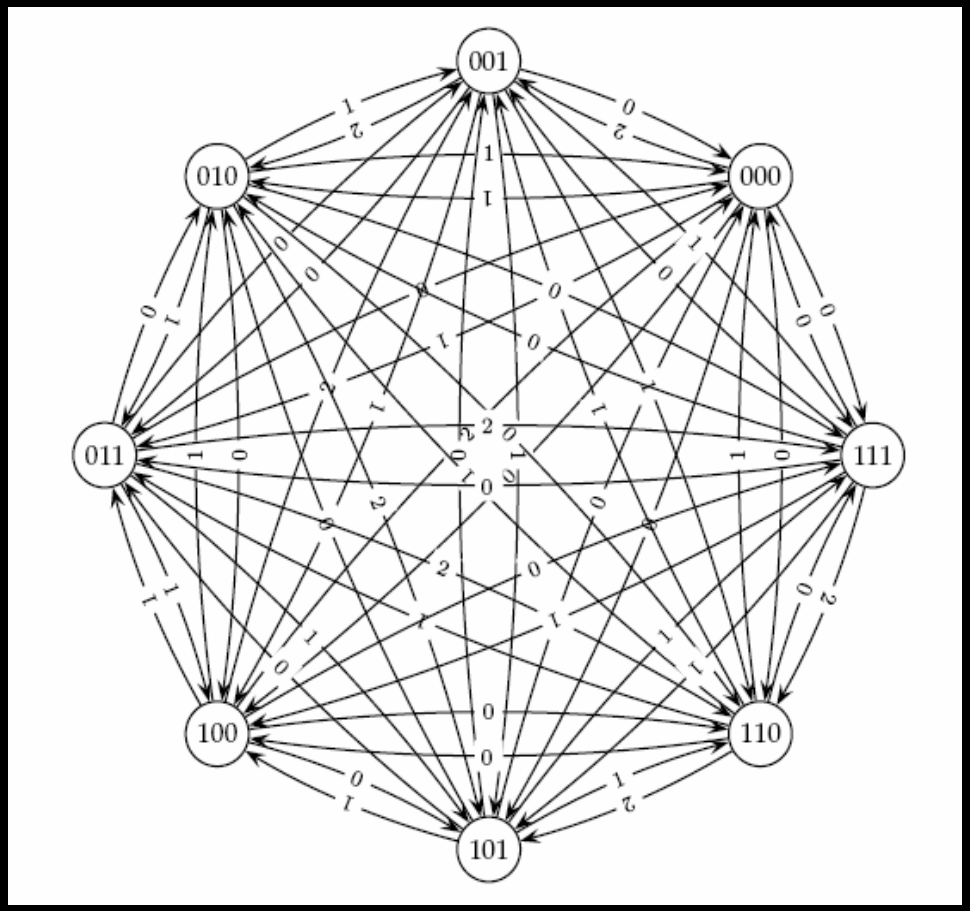
  aaaggcatcaaatctaaaggcatcaaa

  aaaggcatcaaatctaaa

  aaaggcatcaaatctaaa

- Construct a graph with n vertices representing the n strings $s_1, s_2, \ldots, s_n$.
- Insert edges of length *overlap* (s_i, s_j) between vertices s_i and s_j.
- Find the shortest path which visits every vertex exactly once. This is the **Traveling Salesman Problem** (TSP), which is also NP – complete.

Reducing SSP to TSP (cont'd)

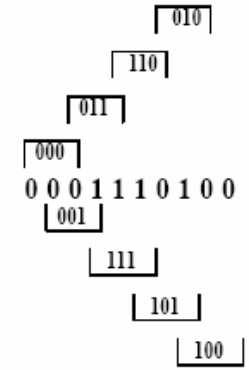


The Shortest Superstring problem

Set of strings: {000, 001, 010, 011, 100, 101, 110, 111}

Concatenation Superstring: 000 001 010 011 100 101 110 111

Shortest superstring: 0 0 0 1 1 1 0 1 0 0

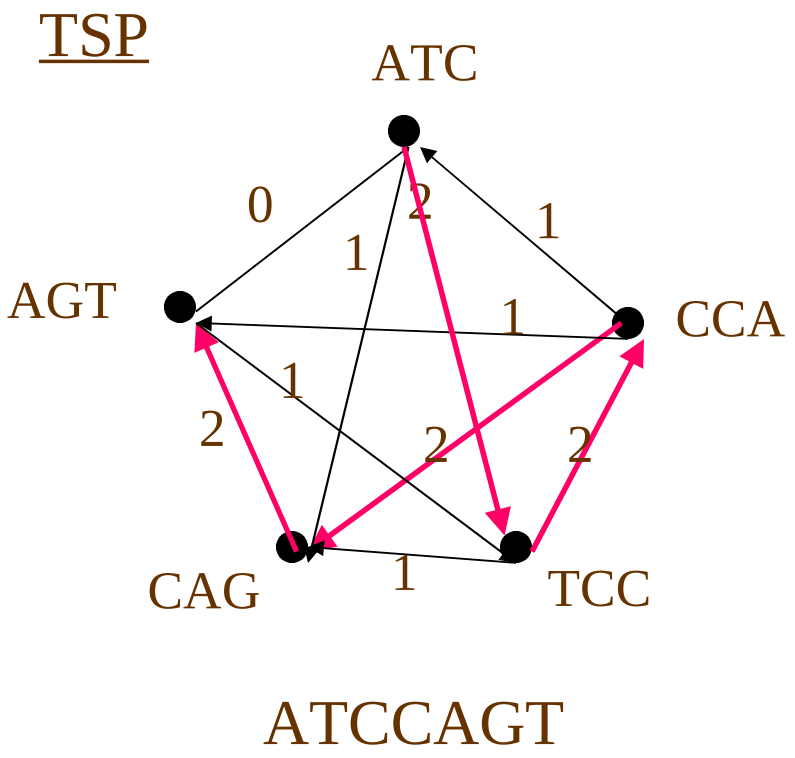


SSP to TSP: An Example

S = { ATC, CCA, CAG, TCC, AGT }

SSP

AGT
CCA
ATC
ATCCAGT
TCC
CAG

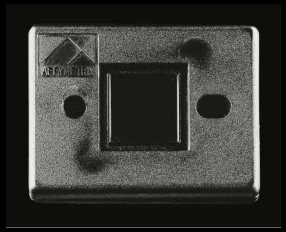


Sequencing by Hybridization (SBH): History

1988: SBH suggested as an an alternative sequencing method. Nobody believed it will ever work

First microarray prototype ***(1989)***

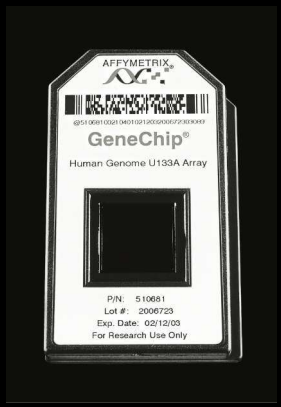
First commercial DNA microarray prototype w/16,000 features ***(1994)***



1991: Light directed polymer synthesis developed by Steve Fodor and colleagues.

1994: Affymetrix develops first 64-kb DNA microarray

500,000 features per chip ***(2002)***



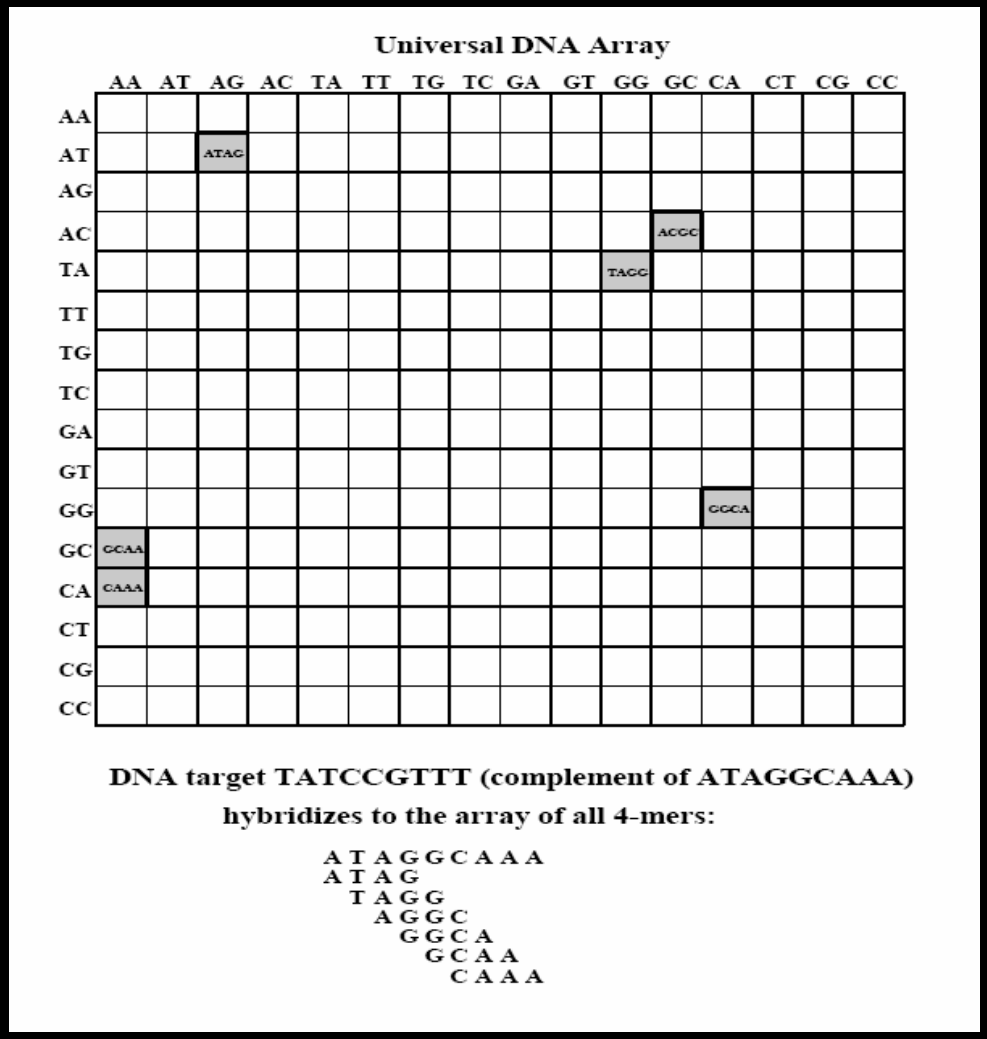
DNA microarray

- a chip which contains short probes
  - ssDNA sequences, millions of them
- make DNA for sequncing fluorescent
- wash it over the chip
- DNA hybridizes to its complementary strand
- cells light up

Universal DNA microarrray

- A DNA microarray which contains **all** seqs of length *l* (*l-mers*)
- therefore, we can determine *l*-mer composition

Hybridization on DNA Array



l–mer composition

- ***Spectrum (s, l)***
  - a set of all possible *l*-mers
- *Spectrum (TATGGTGC, 3):*

  {ATG, GGT, GTG, TAT, TGC, TGG}

Different sequences – the same spectrum

- Different sequences may have the same spectrum:

  Spectrum(GTATCT,2)=

  Spectrum(GTCTAT,2)=

  {AT, CT, GT, TA, TC}

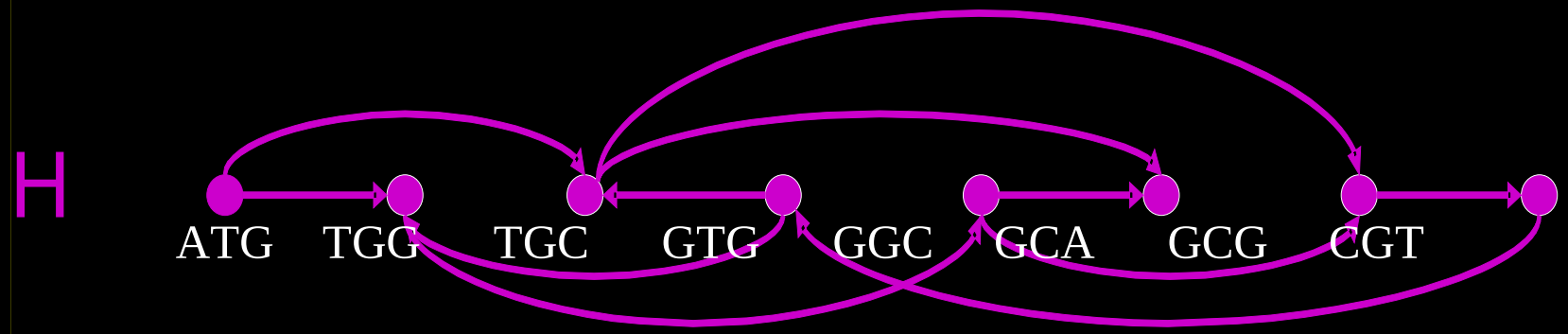
The SBH Problem

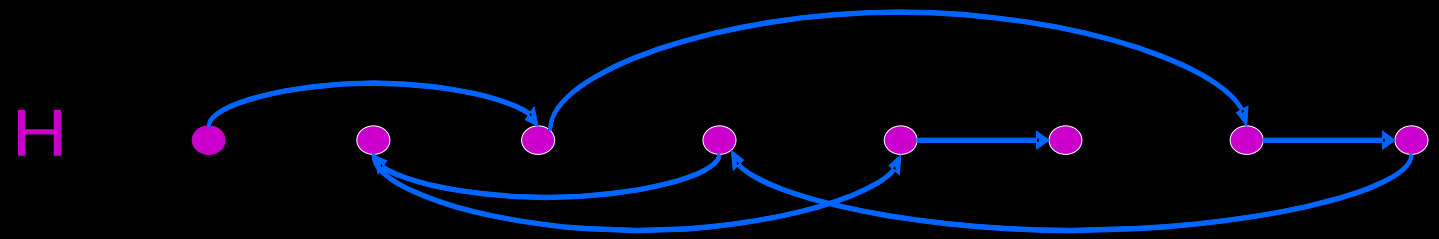
- Goal: Reconstruct a string from its *l*-mer composition
- Input: A set *S*, representing all *l*-mers from an (unknown) string *s*
- Output: String *s* such that *Spectrum (s,l) = S*
- This is a *special case* of SSP

SBH: Hamiltonian Path Approach

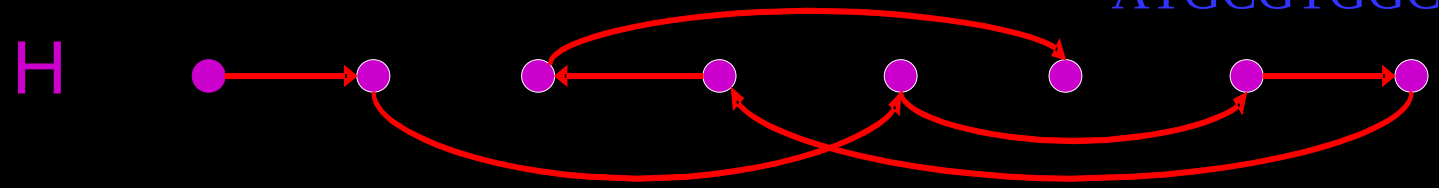
A graph:

S = { ATG TGG TGC GTG GGC GCA GCG CGT }





ATGCGTGGCA



ATGGCGTGCA

SBH: Eulerian Path Approach

S = { ATG, TGC, GTG, GGC, GCA, GCG, CGT }

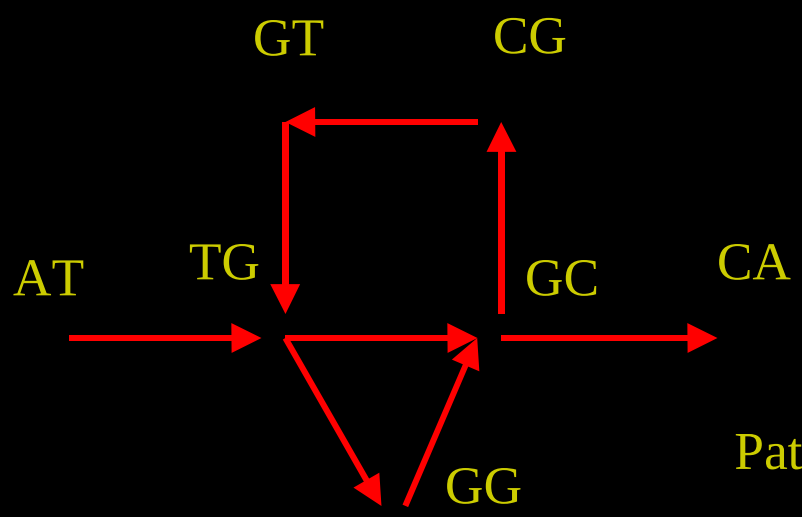
- Vertices correspond to all (*l* – 1) – mers :

  { AT, TG, GC, GG, GT, CA, CG }

- There's an edge S_1->S_2 *iff* there's a substring in the spectrum for which the first l-1 nucleotides correspond to S_1, and the last l-1 nucleotides correspond to S_2

SBH: Eulerian Path Approach

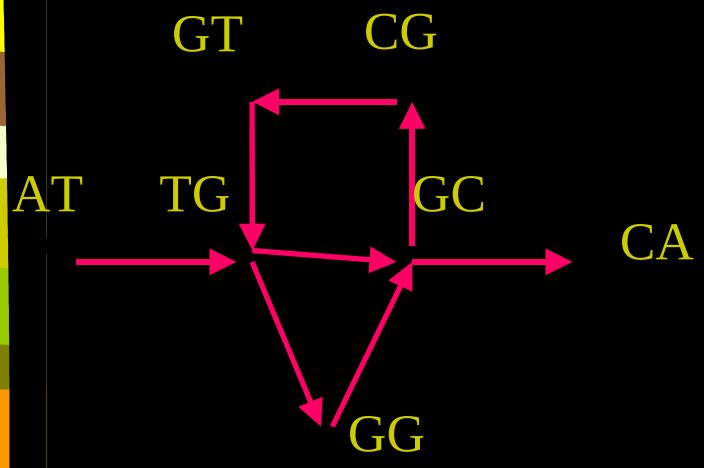
S = { ATG, TGC, GTG, GGC, GCA, GCG, CGT }



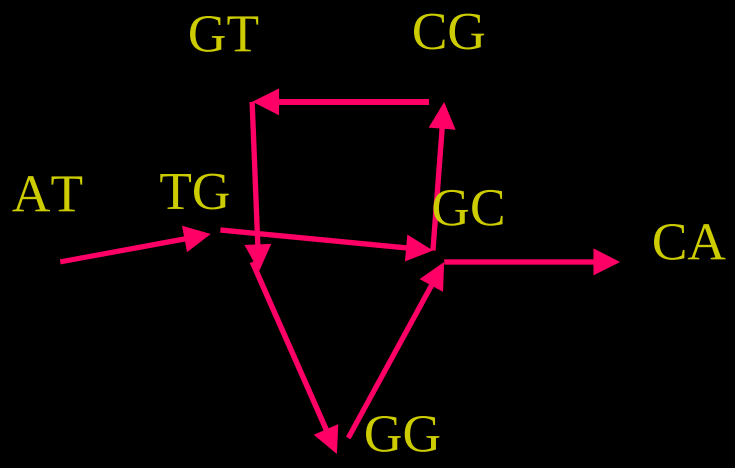
Path visited every EDGE once

SBH: Eulerian Path Approach

S = { AT, TG, GC, GG, GT, CA, CG } corresponds to two different paths:



ATGGCGTGCA



ATGCGTGGCA

Euler Theorem

- A graph is balanced if

  in(v)=out(v) for every v

- **Theorem**: *A connected graph is Eulerian if and only if each of its vertices is balanced.*

Euler Theorem: Proof

- Eulerian → balanced

  for every edge entering *v* (incoming edge) there exists an edge leaving *v* (outgoing edge). Therefore

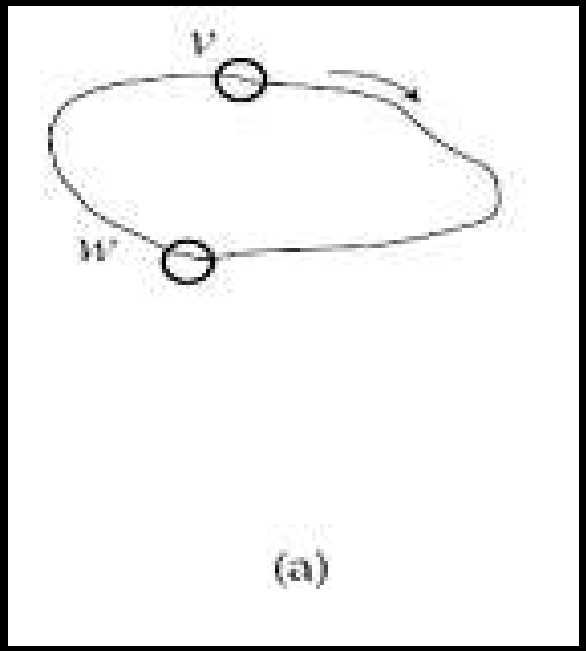
  $$in(v)=out(v)$$

- Balanced → Eulerian

  ???

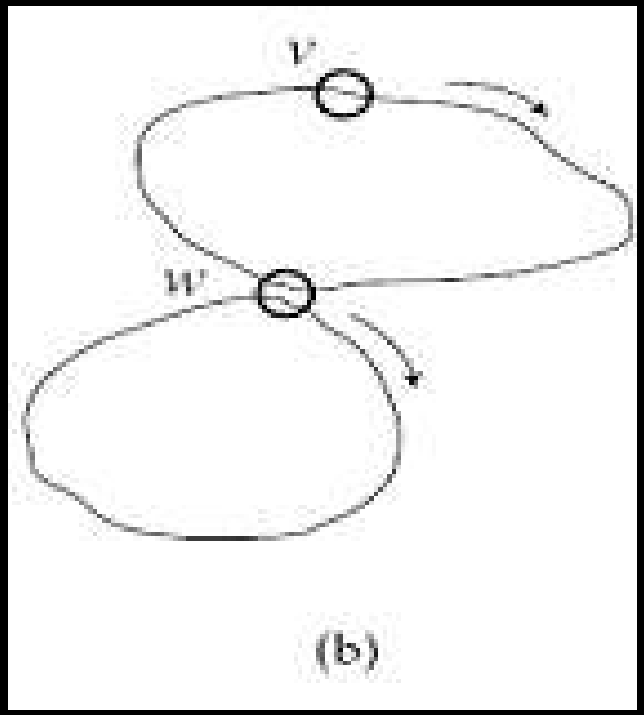
Algorithm for Constructing an Eulerian Cycle

a. Start with an arbitrary vertex *v* and form an arbitrary cycle with unused edges until a dead end is reached. Since the graph is Eulerian this dead end is necessarily the starting point, i.e., vertex *v*.



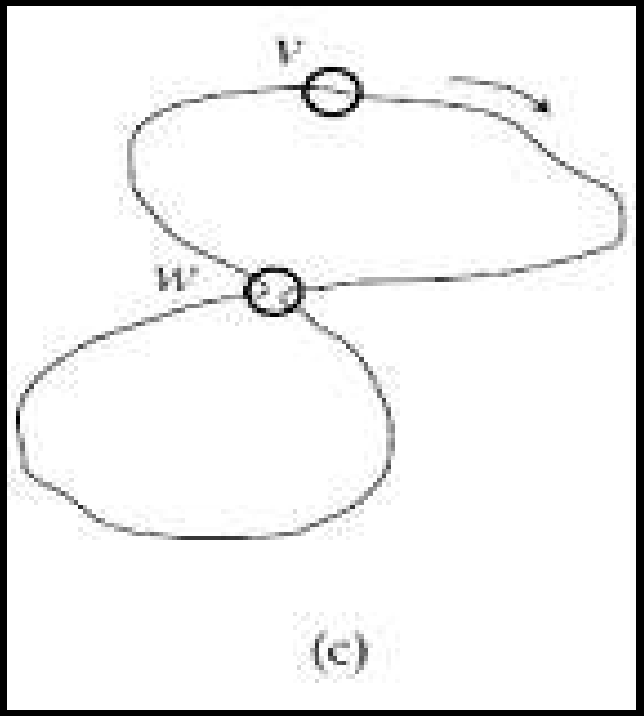
Algorithm for Constructing an Eulerian Cycle (cont'd)

b. If cycle from (a) above is not an Eulerian cycle, it must contain a vertex *w*, which has untraversed edges. Perform step (a) again, using vertex *w* as the starting point. Once again, we will end up in the starting vertex *w*.



Algorithm for Constructing an Eulerian Cycle (cont'd)

c. Combine the cycles from (a) and (b) into a single cycle and iterate step (b).



Euler Theorem: Extension

- **Theorem**: *A connected graph has an Eulerian path if and only if it contains at most two semi-balanced vertices and all other vertices are balanced.*

Some Difficulties with SBH

- **Fidelity of Hybridization:** difficult to detect differences between probes hybridized with perfect matches and 1 or 2 mismatches
- **Array Size:** Effect of low fidelity can be decreased with longer *l*-mers, but array size increases exponentially in *l*. Array size is limited with current technology.
- **Instead microarrays are used for**:
  - gene expression analysis
  - SNP analysis techniques (longer probes in both cases)

References

- www.bioalgorithms.info
- Simons, Robert W. *Advanced Molecular Genetics Course*, UCLA (2002). http://www.mimg.ucla.edu/bobs/C159/Presentations/Be
- Batzoglou, S. *Computational Genomics Course*, Stanford University (2004). http://www.stanford.edu/class/cs262/handouts.html