CISC 636 Computational Biology & Bioinformatics (Fall 2016)

Whole genome sequencing

Mapping & Assembly

Some early completed genomes (source: TIGR CMR)

[A] Bacteria, 1.6 Mb,
~1600 genes
Science 269: 496

[B] 1997
Eukaryote, 13 Mb,
~6K genes
Nature 387: 1

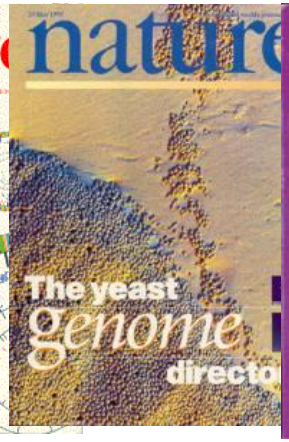
[C] 1998
Animal, ~100 Mb,
~20K genes
Science 282: 1945

[D] 2001
Agrobacteria,
5.67 Mb,
~5419 genes
Science 294:2317

[E,F] 2001
Human, ~3 Gb,
~35K genes
Science 291: 1304
Nature 409: 860

[G] 2005
Chimpanzee

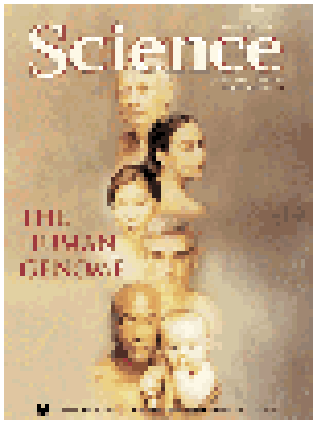
(A)



(B)

(C)

(D)



(E)

(F)

(G)

Sequencing Technologies and Strategies

- Sanger method (gel-based sequencing)
  - Top-Down or "BAC-to-BAC"
  - Whole Genome Shotgun
- Sequencing by Hybridization
- Pyrosequencing ("sequencing by synthesis") (*Science*, 1998)
- Polony sequencing (George Church et al , *Science*, *August* 2005)
- The 4-5-4 sequencing (*Nature*, *July* 2005)
- SBS (www.solexa.com)

Physical Mapping by using Sequence-Tagged-Sites

STSs are unique markers

Exercise: What is the difference between STS and EST, expressed sequence tags?

What is the chance a fragment of 20 bps to be unique in a genome of 3 billion bps?

(Visit http://www.ncbi.nlm.nih.gov/dbEST/ to learn more about EST as an alternative to whole genome sequencing.)

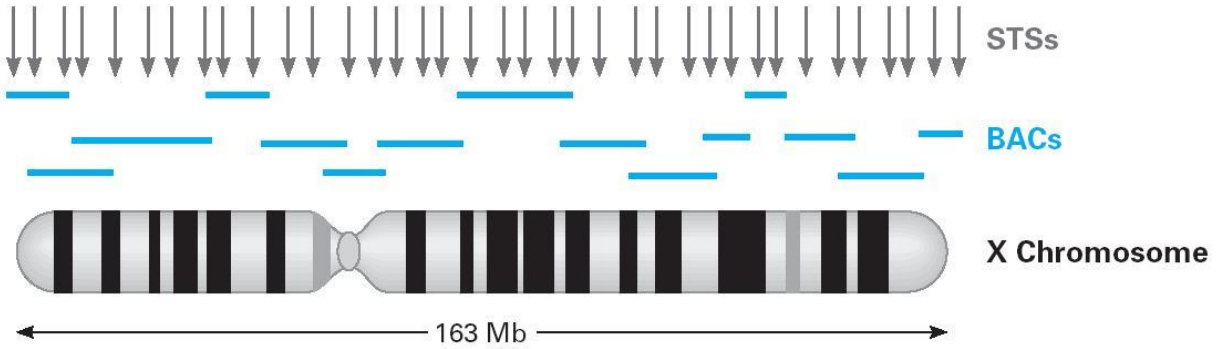
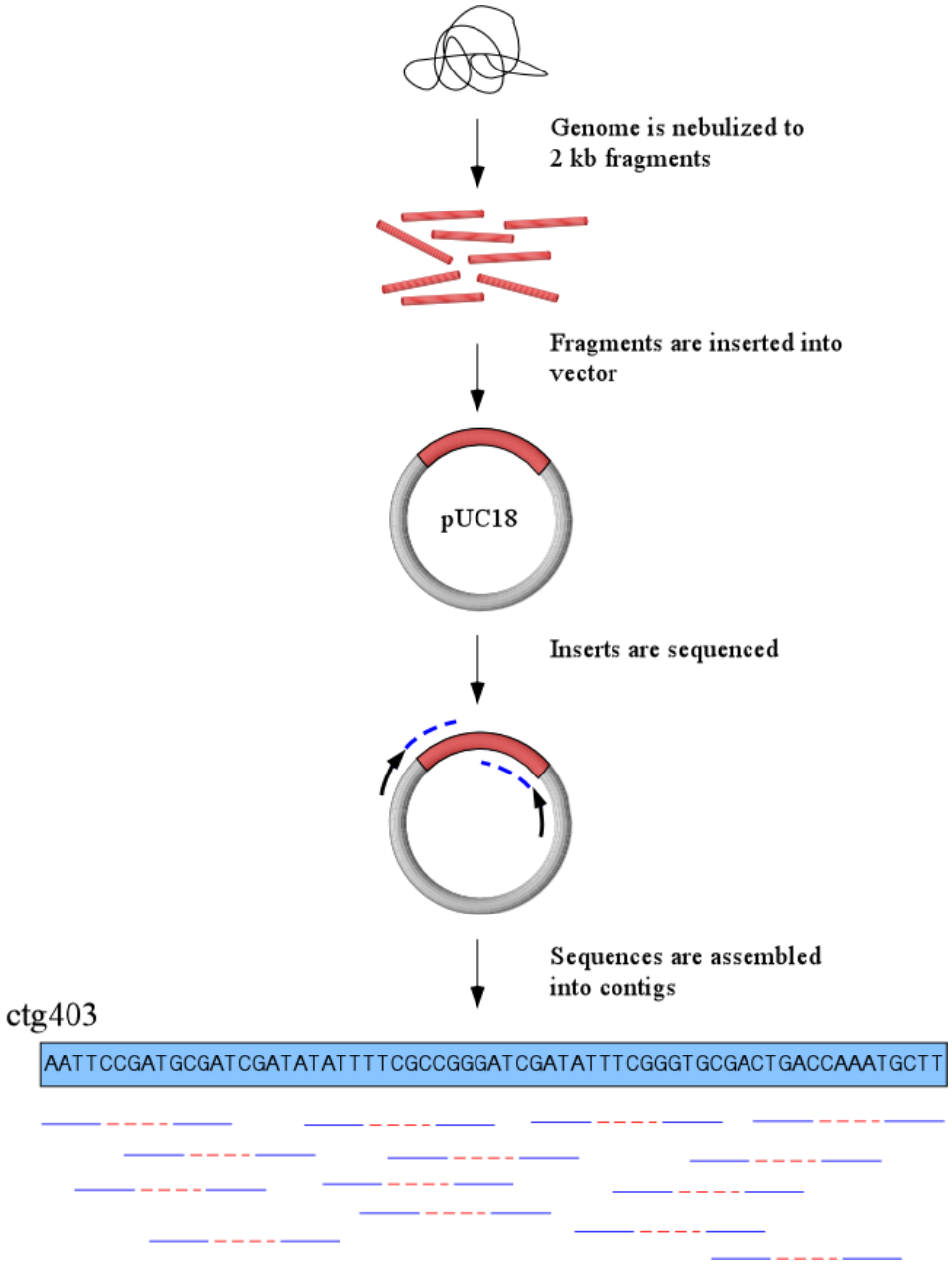


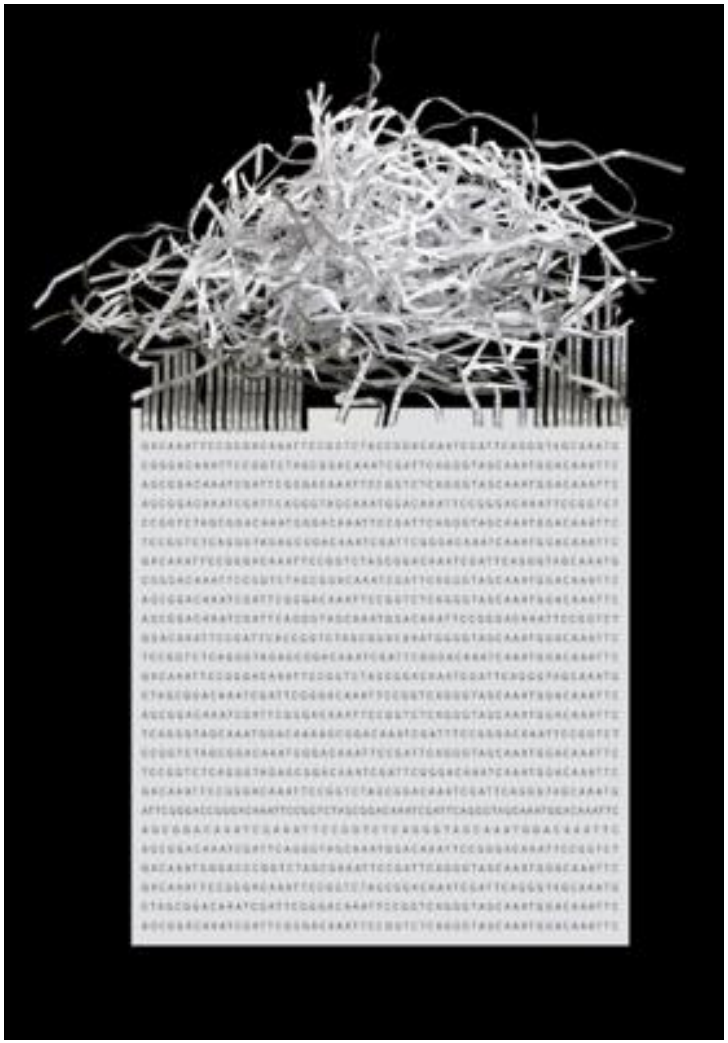
FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout the chromosome and on overlapping BACs.

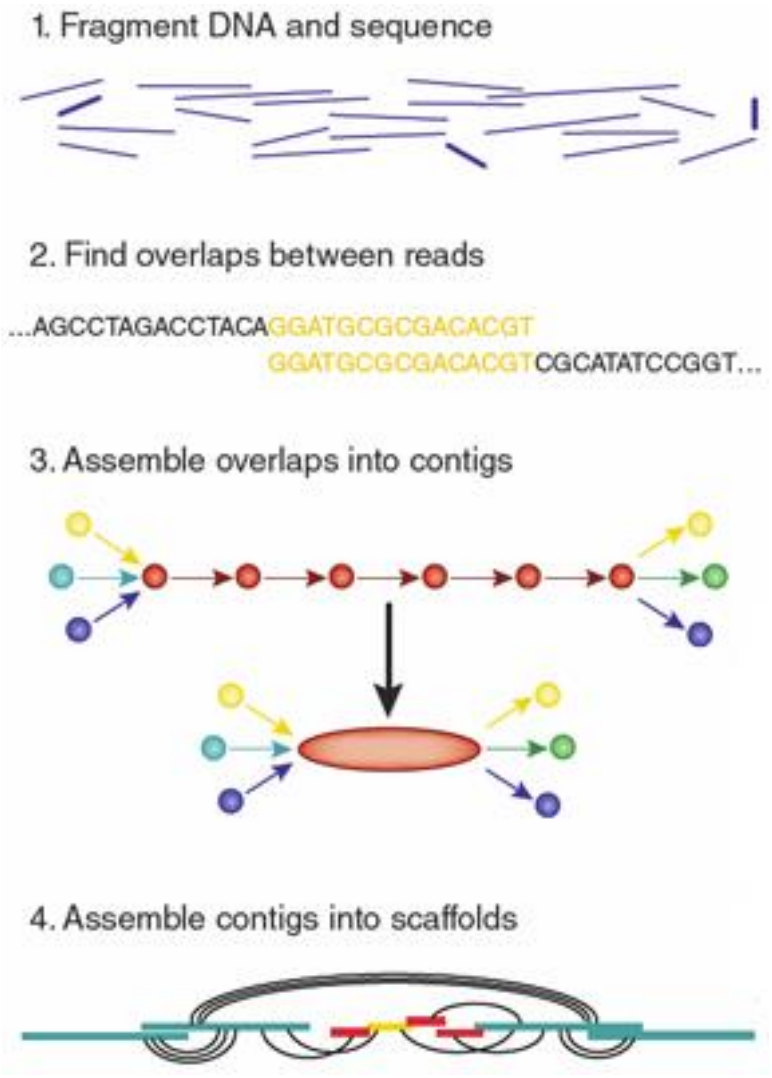
Courtesy of Discovering genomics, proteomics, & bioinformatics by Campbell & Heyer.

Terms

- BAC
- YAC
- Cosmid
- Mapping
- Tiling path
- Read
- Gap
- Contig
- Shotgun





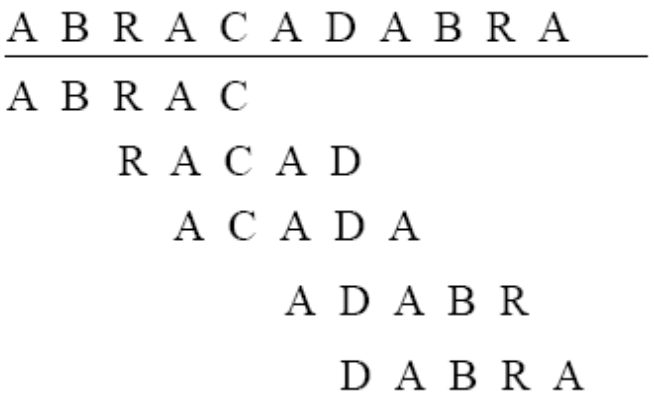


Credit: Nature Methods 9, 333–337 (2012)

Why is Assembly Difficult?

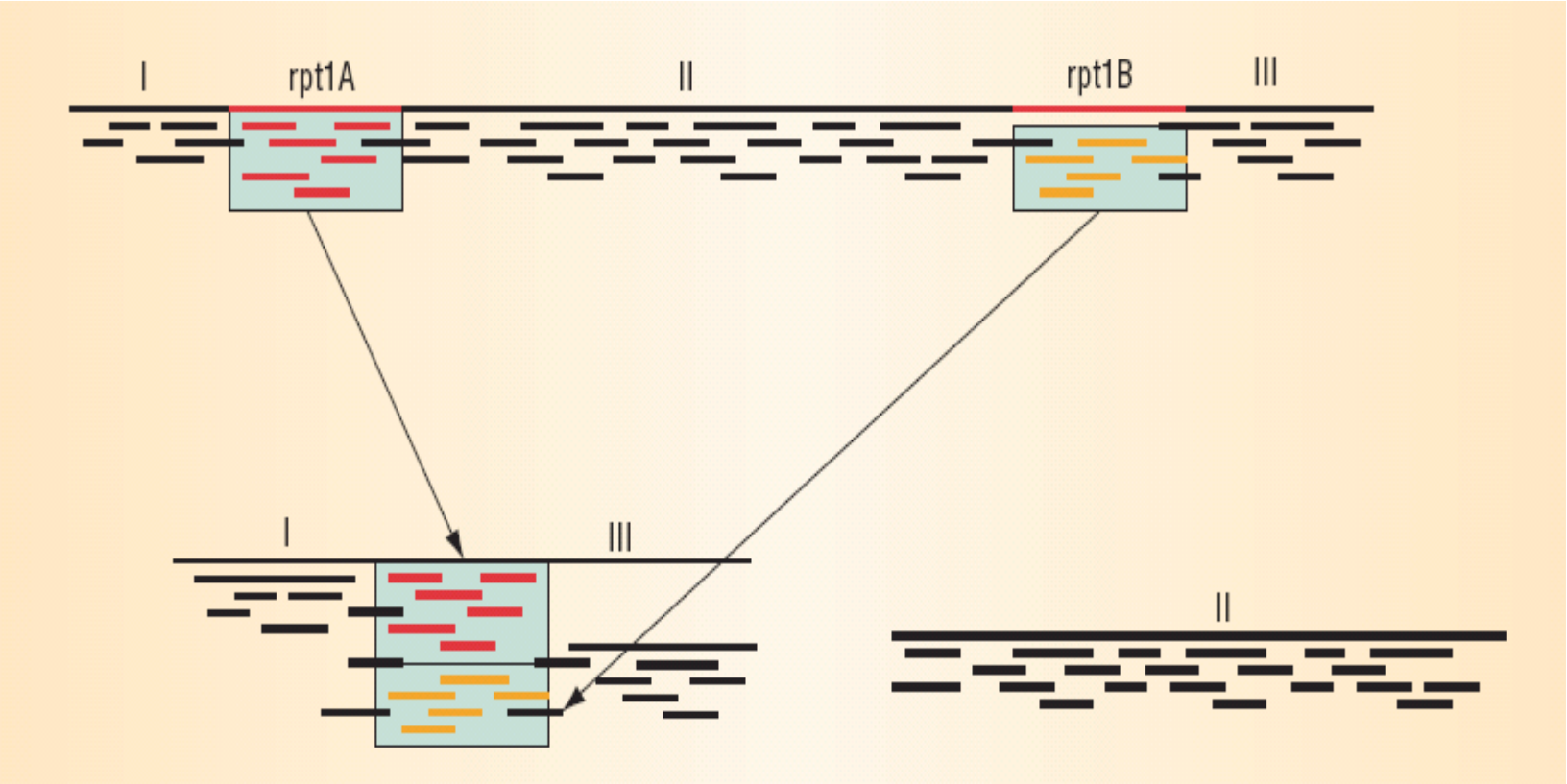
The most natural notion of assembly is to order the fragments so as to form the shortest string containing all of them.

A B R A C
A C A D A
A D A B R
D A B R A
R A C A D



However, the problem of finding the shortest common superstring of a set of strings is NP-complete.

Assembly is complicated by repeats

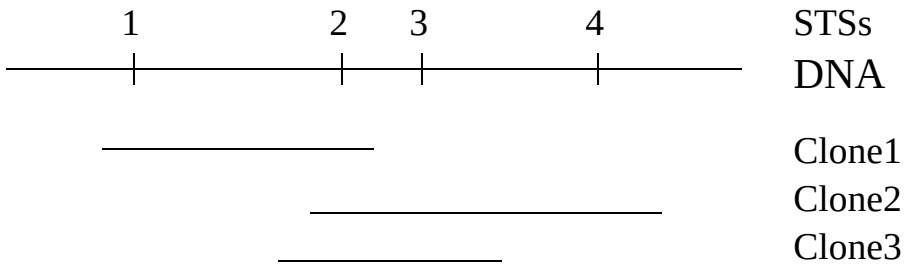


Informatics tasks

- Shotgun coverage (Lander-Waterman)
- Base-calling (Phred)
- Assembly (Phrap, www.phrap.com)
- Visualization (Consed– contig editor for phred and phrap)
- Post-assembly analyses
  - Sun Kim, Li Liao, Jean-Francois Tomb, "A Probabilistic Approach to Sequence Assembly Validation“
  - Alvaro Gonzalez, Li Liao, BMC Bioinformatics, **9**:102, 2008.
- …

STS-content mapping

a. Actual ordering that we want to infer:



b. Hybridization data:

Clone \ STS	2	4	1	3
1	1	0	1	0
2	1	1	0	1
3	1	0	0	1

What we do not know:

either the relative location of STSs in the genome, or

the relative location of clones in the genome.

c. Permutation of columns to have consecutive ones in rows

Clone \ STS	1	2	3	4
1	1	1	0	0
2	0	1	1	1
3	0	1	1	0

Linear time algorithm by Booth & Lueker (1976)

Sequence coverage (Lander-Waterman, 1988):

- Length of genome: G
- Length of fragment: L
- # of fragments: N
- Coverage: a = NL/G.

Fragments are taken randomly from the original full length genome.

Q: What is the probability that a base is not covered by any fragment?

Assumption: fragments are independently taken from the genome, in other words, the left-hand end of any fragment is "uniformly" distributed in (0,G).

Then, the probability for the LHE of a fragment to fall within an interval (x, x+L) is L/G.

Since there are N fragments in total, on average, any point in the genome is going to be covered by NL/G fragments.

Poisson distribution:

- The rate for an event A to occur is r.
  what is the rate to see a left-hand end of a fragment?

- Probability to see an event in time interval (t, t + dt) is
  $P(A|r) = r\,dt$

- h(t) = probability no event in (0,t)
  This is called exponential distribution

- Due to the independence of different time intervals, we have
  $h(t + dt) = h(t)\,[1 - r\,dt]$
  $\partial h/\partial t + r\,h(t) = 0 \Rightarrow$

$$h(t) = \exp(-rt).$$

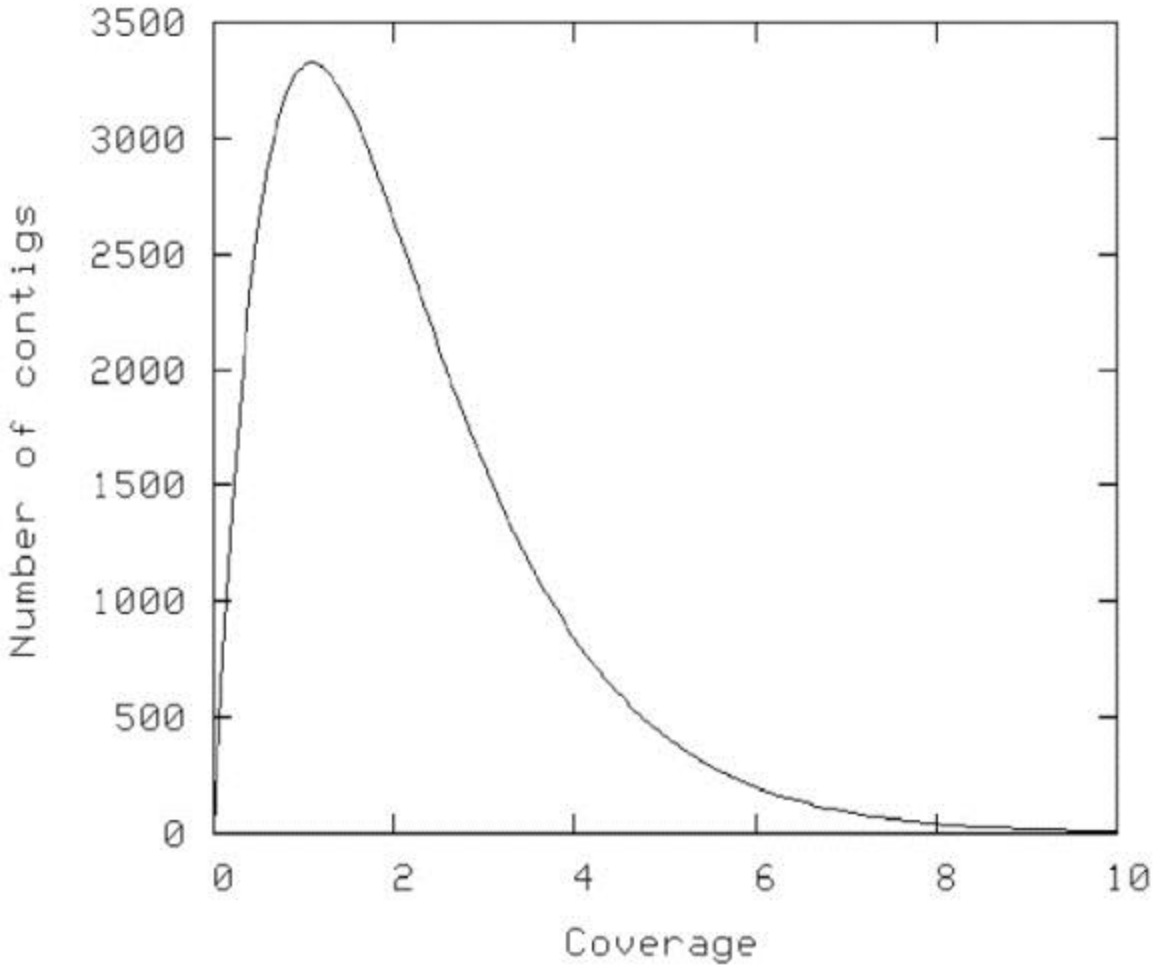
- Probability to have n events in (0, t)
  $P(n|r) = \exp(-rt)\,[(rt)^n / n!]$.
  This is called Poisson distribution.

What is the mean proportion of the genome covered by one or more fragments?

- Randomly pick a point, the probability that to its left, within L, where there are at least on fragment, is

$$1 - \exp(-NL/G)$$

- Example: to have the genome 99% covered, the coverage NL/G shall be 4.6; and 99.9% covered if NL/G is 6.9.
- Is it enough to have 99.9% covered? Human genome has 3 x 10^9 bps. A 6.9 x coverage will leave ~3,000,000 bps uncovered, which cause physical gaps in sequencing the human genome.
- Then, what is the number of possible gaps?


3500
3000
2500
2000
1500
1000
500
0
Number of contigs
0
2
4
6
8
10
Coverage

What is the mean # of contigs?

- N exp(-NL/G)
- For G = 100,000 bps, and L = 500

NL/G	1.0	1.5	2.0	3.0	4.0	5.0	6.0	7.0
Mean # of contigs	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

Assembly programs

- Phrap
- Cap
- TIGR assembler
- Celera assembler
- CAP3
- ARACHNE
- EULER
- AMASS

A genome sequence assembly primer is available at
http://www.cbcb.umd.edu/research/assembly_primer.shtml

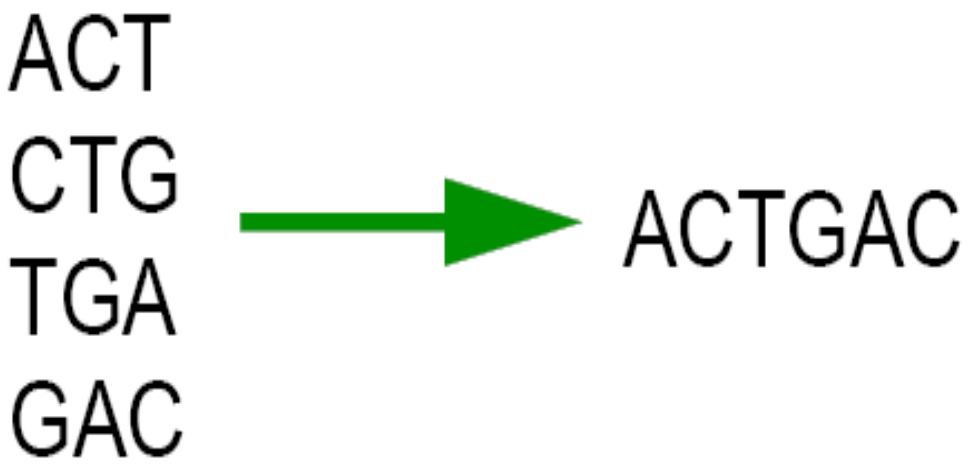
Sequencing by Hybridization

Hybridize target to array containing a spot for each possible k-mer.

TGT ACTGAC TGT	TGG ACTGAC ACTGAC TGG	TGA ACTGAC TGA
CTT ACTGAC CTT ACTGAC	ACTGAC CTG CTG ACTGAC	ACTGAC CTA ACTGAC CTA
GAA GAA ACTGAC	GAT GAT ACTGAC	GAC ACTGAC GAC ACTGAC

The spectrum of a sequence: multi-set of all its k-long substrings (k-mers).

Goal: reconstruct the sequence from its spectrum.



Pevzner 89: reconstruction is polynomial.

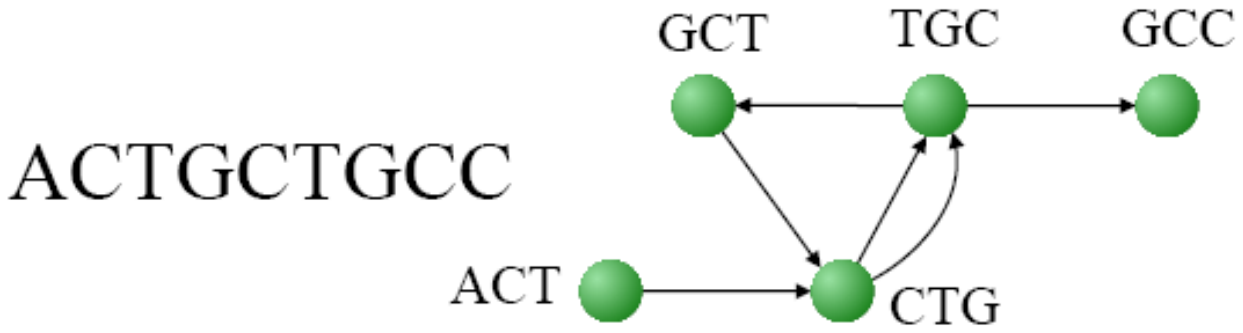
Sequence reconstruction and Eulerian Path Problem (Pavzner '89)

$A = a_1 \cdots a_{n+k-1}$: the sequence.

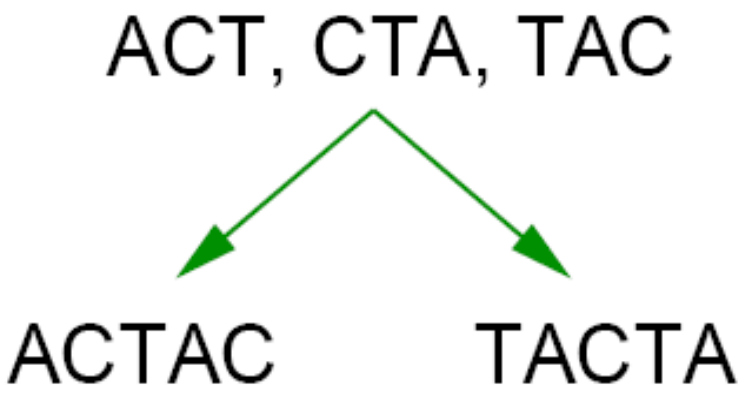
A_i : the $(k-1)$-mer $a_i a_{i+1} \cdots a_{i+k-2}$.

The de-Bruijn graph of A : $G_A = (V, E)$ where

- $V = \{A_i : i = 1, \ldots, n+1\}$
- $E = \{e_i : i = 1, \ldots, n\}$, $e_i = (A_i, A_{i+1})$



Different sequences can have the same spectrum:



Shorty: Assembly from Short Paired Reads

1. Clean input read-pairs to correct base-sequencing errors, using read frequency analysis and consensus read correction.
2. Construct the de Bruijn subgraph on "left" reads so as to group associated the "right" reads.
3. Construct the de Bruijn graph on the "right" reads of each left group.
4. Select contigs of sufficient size to pass through a shotgun assembler.
5. Post-assembly contig extension.

Chen & Skiena 2005 (http://www.cs.sunysb.edu/~skiena)