CISC 889 Bioinformatics (Spring 2004)
Lecture 3

Genome Sequencing

Li Liao
Computer and Information Sciences
University of Delaware

Administrative

- Have you visited The NCBI website?
- Have you read Hunter's Tutorial?
- Can you remember 20 amino acids now?
- Can you remember some genetic code?

- Do you have any questions?

Today's topic

- How are whole genomes sequenced?
- What are the challenges (to bioinformatics)?
- Is it a done deal now?

As of Feb. 2004, 159 completed microbial genomes

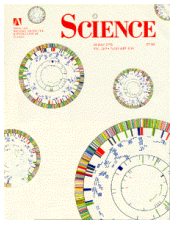
[A] Bacteria, 1.6 Mb, ~1600 genes Science 269: 496

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

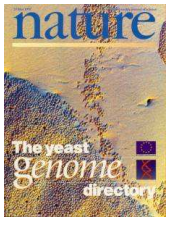
[C] 1998 Animal, ~100 Mb, ~20K genes Science 282: 1945

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

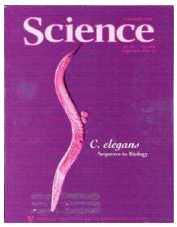
[F] 2001 Agrobacteria, 5.67 Mb, ~5419 genes Science 294:2317



(A)



(B)



(C)

(D)

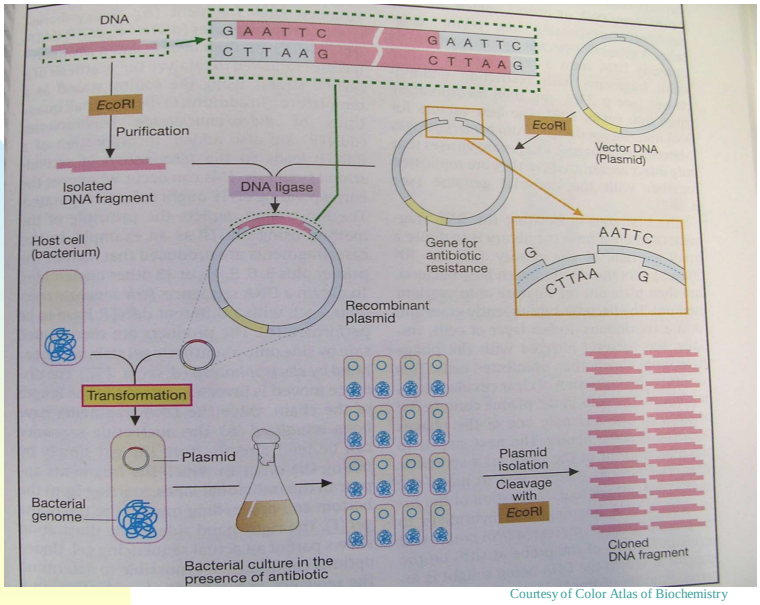
(E)

(F)

Terminologies

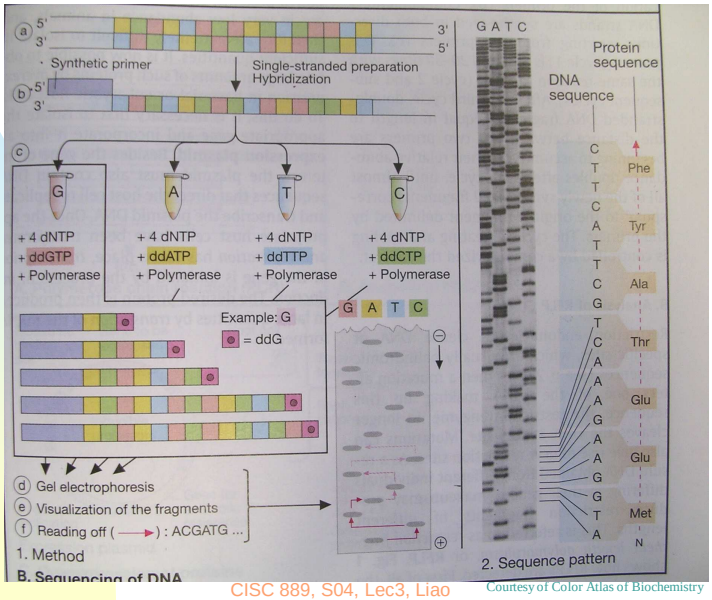
- Oligomer
- Primer
- dNTP and ddNTP
- Electrophoresis
- Gel
- Chromatogram
- Base-calling
- Phrap, Phred, Consed
- Strategies:
  - Shotgun
  - Transposon-based

DNA Cloning



Courtesy of Color Atlas of Biochemistry

Sequencing DNA (Sanger's method)

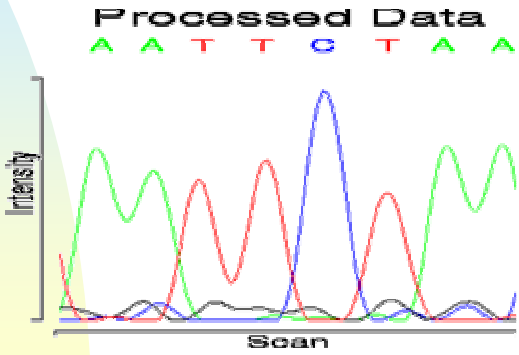


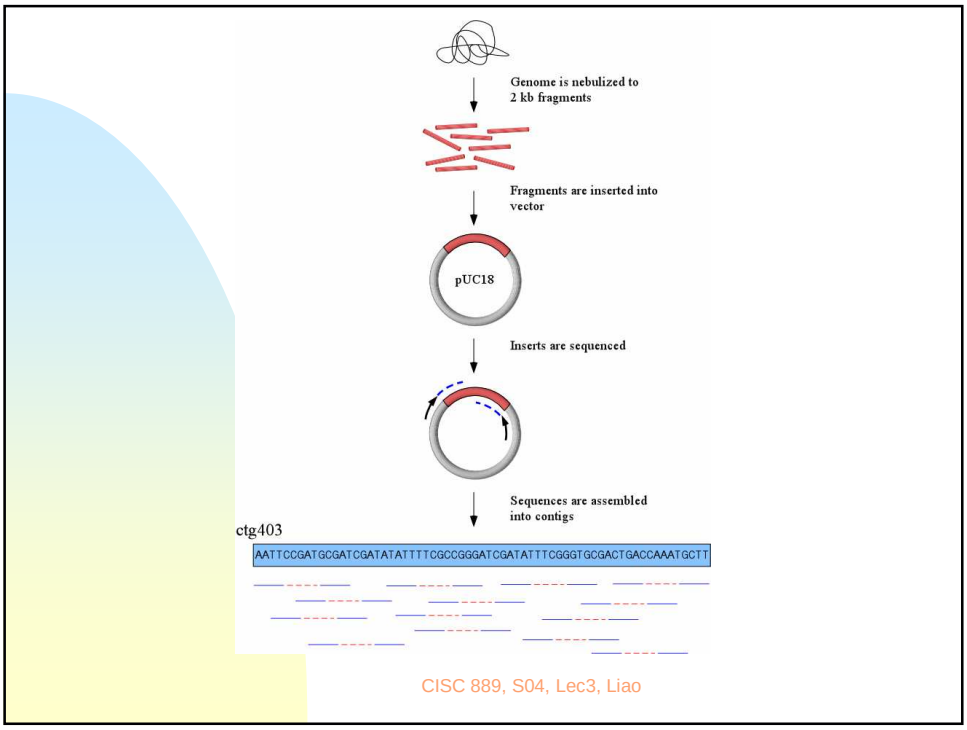
Courtesy of Color Atlas of Biochemistry

Improvement by Leory Hood:

4 color fluorescent dyes in single lane -> Chromatograms

Base-calling





STS: sequence-tagged-sites, served as 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.)

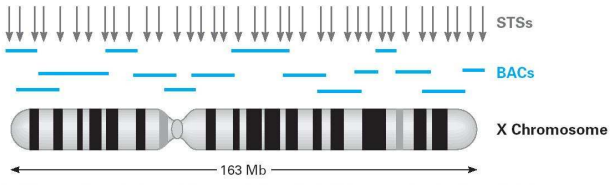
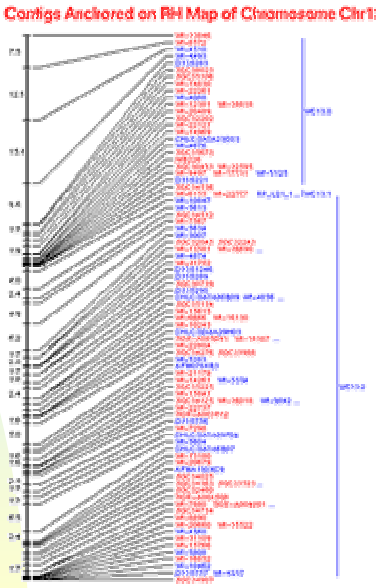


FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout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

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



Courtesy of http://www.broad.mit.edu/cgi-bin/contig/phys_map

CISC 889, S04, Lec3, Liao

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
- …

CISC 889, S04, Lec3, Liao

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

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

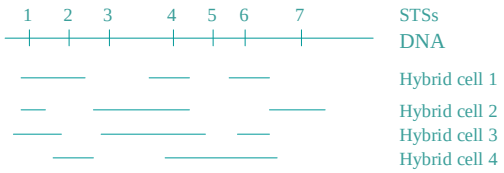
Linear time algorithm by Booth & Lueker (1976)

CISC 889, S04, Lec3, Liao

Radiation-hybrid mapping

Any single hamster cell contains ~ 5 to 10 disconnected, nonoverlapping fragments.

a. Actual ordering that we want to infer:



b. Hybridization data:

cells \ STS	2	4	5	3	1	7	6
1	1	1	0	0	1	0	1
2	0	1	0	1	1	1	0
3	0	1	0	1	1	0	1
4	1	1	1	0	0	0	1

Computational problem: to deduce the correct order of the STSs.

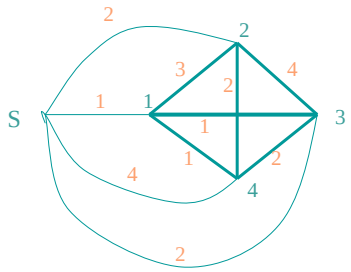
Answer: No exact solution.

CISC 889, S04, Lec3, Liao

Two transformed problems: Find a permutation of the columns that

1. minimize the total # of blocks of consecutive ones, or
2. minimize the maximum # of blocks that appear in any single row.

cells \ STS	2	4	3	1
1	1	1	0	1
2	0	1	1	1
3	0	1	1	1
4	1	1	0	0



Assign weight:

(S, v) = # of 1s in column v

(u,v) = Hamming distance between columns u and v.

Version 1 problem can then be solved as a traveling salesman problem, which is itself known to be NP-complete.

CISC 889, S04, Lec3, Liao

Sequence coverage:

- 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?

Assumptions:

1. left-hand end of any fragment is uniformly distributed in (0,G),
2. probability to be in an interval (x, x+L) is h/G.
3. on average, at any point are covered by NL/G fragments.
4. independence of different fragments (or time intervals)

Poisson distribution: (as a limiting "low counting rate" approximation to the binomial distribution)

- rate r for an event A to occur in time interval (t, t + dt) is
  $P(A|r) = r\,dt$
- h(t) = probability no event in (0,t)
- By independence of different time intervals
  $h(t + dt) = h(t)\,[1 - r\,dt]$
  $\partial h/\partial t + r\,h(t) = 0 \Rightarrow h(t) = \exp(-rt)$. This is the answer to our question.

- n events in (0, t)
  $P(n|r) = \exp(-rt)\,[(rt)^n / n!]$

CISC 889, S04, Lec3, Liao

More questions

- 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.
- 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

CISC 889, S04, Lec3, Liao