

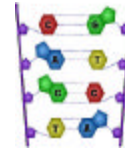
DNA Sequencing



Lecture 8, Thursday April 24, 2003

New topic: DNA sequencing

How we obtain the sequence of nucleotides of a species



```

...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAATATT...
    
```

Lecture 8, Thursday April 24, 2003

Which representative of the species?

Which human?



Answer one:

Answer two: it doesn't matter

Polymorphism rate: number of letter changes between two different members of a species

Humans: $\sim 1/1,000 - 1/10,000$

Other organisms have much higher polymorphism rates

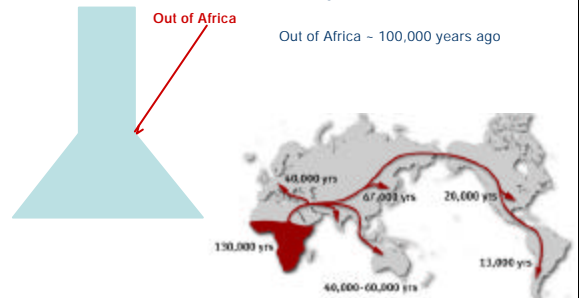


Lecture 8, Thursday April 24, 2003

Why humans are so similar

A small population that interbred reduced the genetic variation

Out of Africa - 100,000 years ago



Lecture 8, Thursday April 24, 2003

Migration of human variation

Early *Homo sapiens sapiens*
in Africa

150,000 to 100,000 BP



<http://info.med.yale.edu/genetics/kkidd/point.html>

Lecture 8, Thursday April 24, 2003

Migration of human variation

Homo sapiens sapiens
colonizing south west Asia
 $\sim 100,000$ BP



<http://info.med.yale.edu/genetics/kkidd/point.html>

Lecture 8, Thursday April 24, 2003

Migration of human variation



Lecture 8, Thursday April 24, 2003

DNA Sequencing

Goal:

Find the complete sequence of A, C, G, T's in DNA

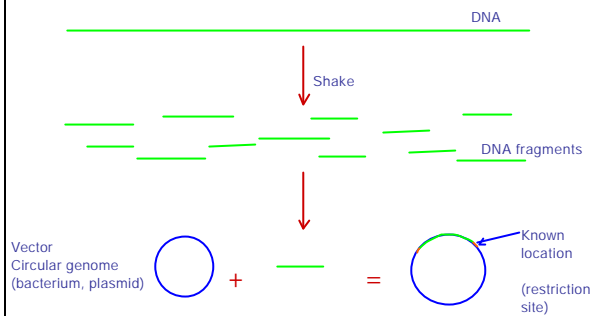
Challenge:

There is no machine that takes long DNA as an input, and gives the complete sequence as output

Can only sequence ~500 letters at a time

Lecture 8, Thursday April 24, 2003

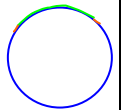
DNA sequencing – vectors



Lecture 8, Thursday April 24, 2003

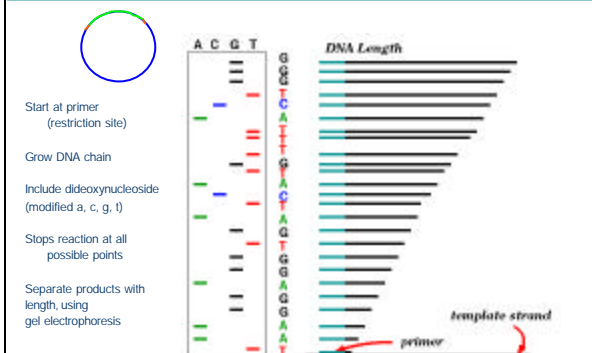
Different types of vectors

VECTOR	Size of insert
Plasmid	2,000-10,000 Can control the size
Cosmid	40,000
BAC (Bacterial Artificial Chromosome)	70,000-300,000
YAC (Yeast Artificial Chromosome)	> 300,000 Not used much recently



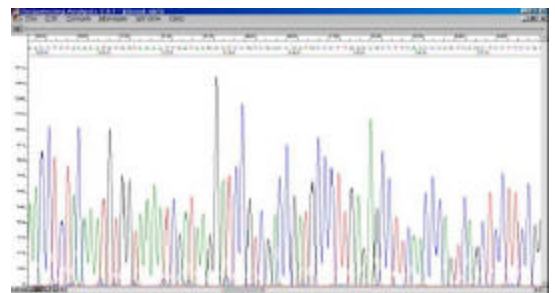
Lecture 8, Thursday April 24, 2003

DNA sequencing – gel electrophoresis



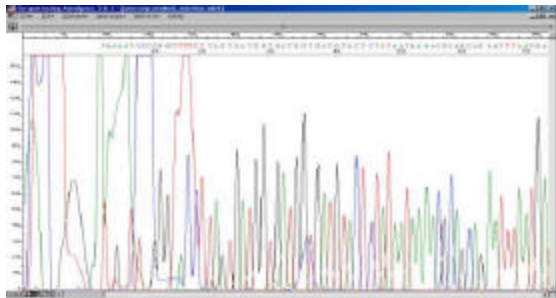
Lecture 8, Thursday April 24, 2003

Electrophoresis diagrams



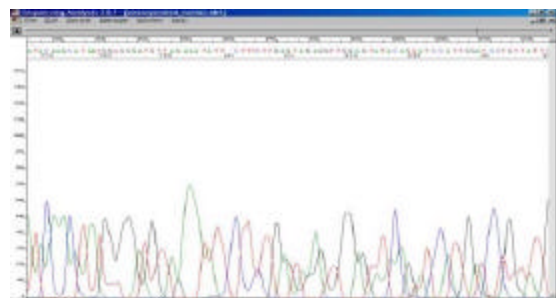
Lecture 8, Thursday April 24, 2003

Challenging to read answer



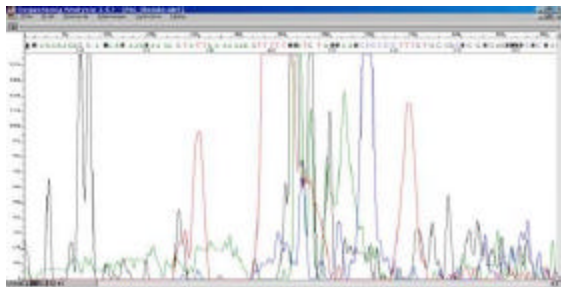
Lecture 8, Thursday April 24, 2003

Challenging to read answer



Lecture 8, Thursday April 24, 2003

Challenging to read answer



Lecture 8, Thursday April 24, 2003

Reading an electropherogram

1. Filtering
2. Smoothing
3. Correction for length compressions
4. A method for calling the letters – PHRED

PHRED – PHil's read editor (by Phil Green)
Based on dynamic programming

Several better methods exist, but labs are reluctant to change

Lecture 8, Thursday April 24, 2003

Output of PHRAP: a read

500-700 nucleotides

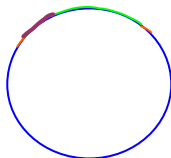
A C G A A T C A G A
16 18 21 23 25 15 28 30 32 21

Quality scores: $-10 \log_{10} \text{Prob}(\text{Error})$

Reads can be obtained from leftmost,
rightmost ends of the insert

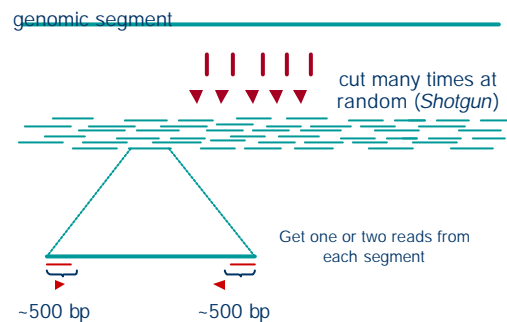
Double-barreled sequencing:

Both leftmost & rightmost ends are
sequenced



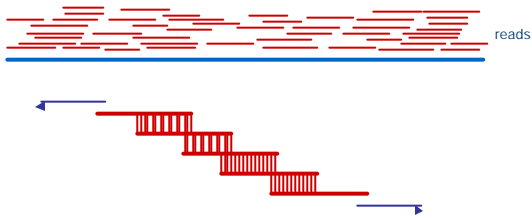
Lecture 8, Thursday April 24, 2003

Method to sequence segments longer than 500



Lecture 8, Thursday April 24, 2003

Reconstructing the Sequence (Fragment Assembly)



Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

Lecture 8, Thursday April 24, 2003

Definition of Coverage



Length of genomic segment: L
 Number of reads: n
 Length of each read: l

Coverage $C = nl/L$

How much coverage is enough?

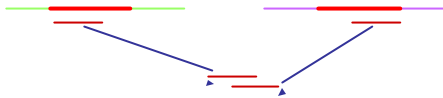
(Lander-Waterman model):
 Assuming uniform distribution of reads, $C=10$ results in 1 gapped region /1,000,000 nucleotides

Lecture 8, Thursday April 24, 2003

Challenges with Fragment Assembly

- Sequencing errors
~1-2% of bases are wrong

- Repeats



- Computation: ~ $O(N^2)$ where $N = \# \text{ reads}$

Lecture 8, Thursday April 24, 2003

Repeats

Bacterial genomes: 5%
 Mammals: 50%

Repeat types:

Low-Complexity DNA (e.g. ATATATATACATA...)
 Microsatellite repeats: $(a_1 \dots a_k)^n$ where $k \sim 3-6$
 (e.g. CAGCAGTAGCAGCACCAG)

Common Repeat Families
 SINE (Short Interspersed Nuclear Elements)
 (e.g. ALU: ~300-long, 10^6 copies)
 LINE (Long Interspersed Nuclear Elements)
 ~500-5,000-long, 200,000 copies

MIR
 LTR/Retroviral

Other
 -Genes that are duplicated & then diverge (paralogs)
 -Recent duplications, ~100,000-long, very similar copies

Lecture 8, Thursday April 24, 2003

Strategies for sequencing a whole genome

- Hierarchical - Clone-by-clone
 - Break genome into many long pieces
 - Map each long piece onto the genome
 - Sequence each piece with shotgun

Example: Yeast, Worm, Human, Rat

- Online version of (1) - Walking
 - Break genome into many long pieces
 - Start sequencing each piece with shotgun
 - Construct map as you go

Example: Rice genome

- Whole genome shotgun

One large shotgun pass on the whole genome

Example: Drosophila, Human (Celera), Neurospora, Mouse, Rat, Fugu

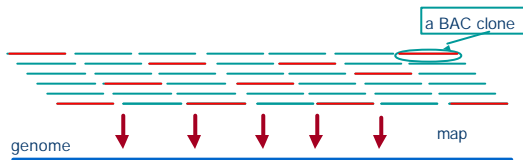
Lecture 8, Thursday April 24, 2003

Hierarchical Sequencing



Lecture 8, Thursday April 24, 2003

Hierarchical Sequencing Strategy



1. Obtain a large collection of BAC clones
2. Map them onto the genome (Physical Mapping)
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

Lecture 8, Thursday April 24, 2003

Methods of physical mapping



Goal:

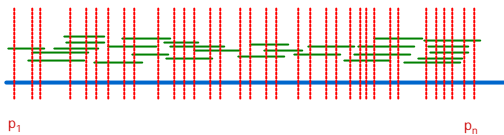
Make a map of the locations of each clone relative to one another
Use the map to select a minimal set of clones to sequence

Methods:

- Hybridization
- Digestion

Lecture 8, Thursday April 24, 2003

1. Hybridization



Short words, the *probes*, attach to complementary words

1. Construct many probes
2. Treat each BAC with all probes
3. Record which ones attach to it
4. Same words attaching to BACS X, Y \Rightarrow overlap

Lecture 8, Thursday April 24, 2003

Hybridization – Computational Challenge

Matrix:

m probes \times n clones

(i, j): 1, if p_i hybridizes to C_j
0, otherwise

Definition:

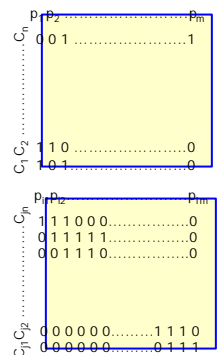
Consecutive ones matrix

A matrix 1s are consecutive

Computational problem:

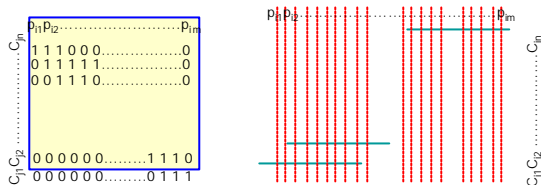
Reorder the probes so that matrix is in consecutive-ones form

Can be solved in $O(m^3)$ time ($m \gg n$)



Lecture 8, Thursday April 24, 2003

Hybridization – Computational Challenge



If we put the matrix in consecutive-ones form,
then we can deduce the order of the clones
& which pairs of clones overlap

Lecture 8, Thursday April 24, 2003

Hybridization – Computational Challenge

Additional challenge:

A probe (short word) can hybridize in many places in the genome

Computational Problem:

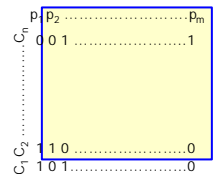
Find the order of probes that implies the minimal probe repetition

Equivalent: find the shortest string of probes such that each clone appears as a substring

APX-hard

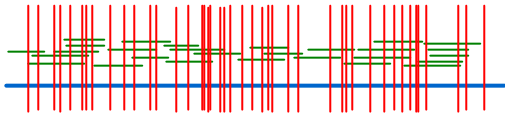
Solutions:

Greedy, Probabilistic, etc.
Lots of manual curation



Lecture 8, Thursday April 24, 2003

2. Digestion



Restriction enzymes cut DNA where specific words appear

1. Cut each clone separately with an enzyme
2. Run fragments on a gel and measure length
3. Clones C_a, C_b have fragments of length $\{l_p, l_j, l_k\} \Rightarrow$ overlap

Double digestion:

Cut with enzyme A, enzyme B, then enzymes A + B

Lecture 8, Thursday April 24, 2003

Online Clone-by-clone The Walking Method



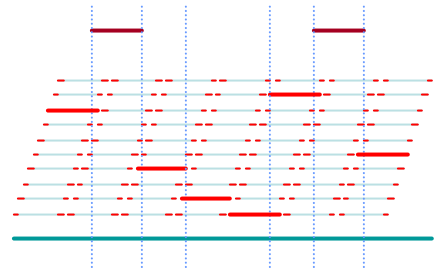
Lecture 8, Thursday April 24, 2003

The Walking Method

1. Build a very redundant library of BACs with sequenced clone-ends (cheap to build)
2. Sequence some "seed" clones
3. "Walk" from seeds using clone-ends to pick library clones that extend left & right

Lecture 8, Thursday April 24, 2003

Walking: An Example



Lecture 8, Thursday April 24, 2003

Advantages & Disadvantages of Hierarchical Sequencing

Hierarchical Sequencing

- ADV. Easy assembly
- DIS. Build library & physical map; redundant sequencing

Whole Genome Shotgun (WGS)

- ADV. No mapping, no redundant sequencing
- DIS. Difficult to assemble and resolve repeats

The Walking method - motivation

Sequence the genome clone-by-clone without a physical map

The only costs involved are:

- Library of end-sequenced clones (CHEAP)
- Sequencing

Lecture 8, Thursday April 24, 2003

Walking off a Single Seed



- Low redundant sequencing
- Too many sequential steps

Lecture 8, Thursday April 24, 2003

Walking off Several Seeds in Parallel



- Few sequential steps
- Additional redundant sequencing

In general, can sequence a genome in ~5 walking steps, with <20% redundant sequencing

Lecture 8, Thursday April 24, 2003

Next Lecture

Whole-genome shotgun sequencing

- Currently, the most popular method for sequencing a genome

Computational assembly of a genome

- Putting a large puzzle together

Lecture 8, Thursday April 24, 2003