

DNA Sequencing



Lecture 9, Tuesday April 29, 2003

Reading

Basic:

ARACHNE: A Whole-Genome Shotgun Assembler

Euler: A shotgun assembler based on finding Eulerian paths

Optional:

Transposons: Genome Sizes:

ARACHNE 2: Assembly of the mouse genome

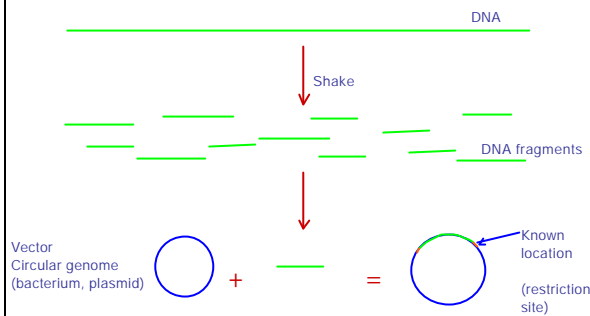
Skim through following 2 free *Nature* issues:

Mouse (December 2002);

50 year anniversary (last week!)

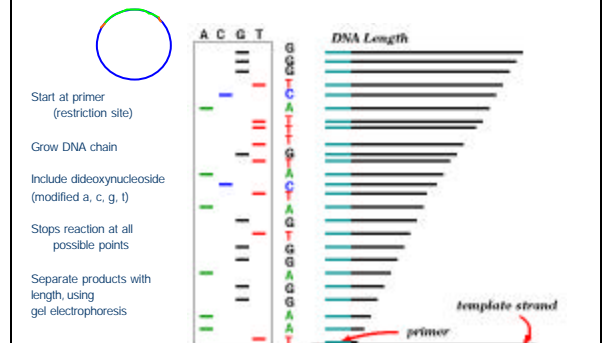
Lecture 9, Tuesday April 29, 2003

DNA sequencing – vectors



Lecture 9, Tuesday April 29, 2003

DNA sequencing – gel electrophoresis



Lecture 9, Tuesday April 29, 2003

Output of PHRAP: a read

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

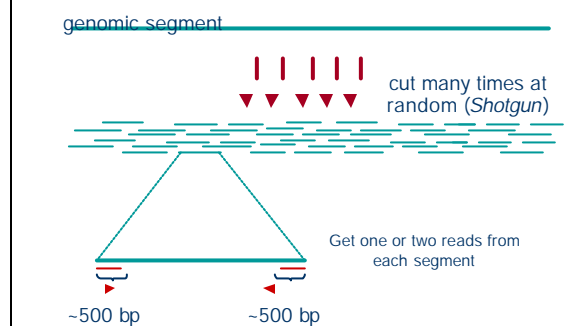
Reads can be obtained from leftmost, rightmost ends of the insert

Double-barreled sequencing:

Both leftmost & rightmost ends are sequenced

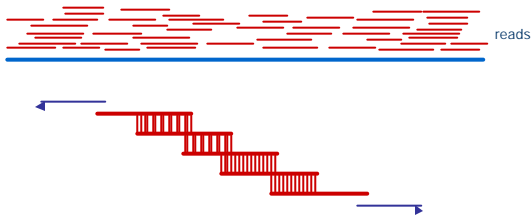
Lecture 9, Tuesday April 29, 2003

Method to sequence segments longer than 500



Lecture 9, Tuesday April 29, 2003

Reconstructing the Sequence (Fragment Assembly)



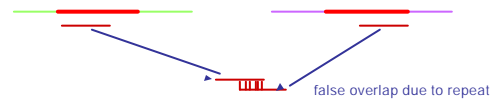
Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region

Lecture 9, Tuesday April 29, 2003

Challenges with Fragment Assembly

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



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

Lecture 9, Tuesday April 29, 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

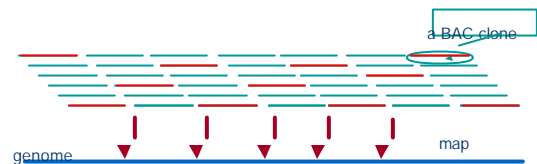
- Whole genome shotgun

One large shotgun pass on the whole genome

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

Lecture 9, Tuesday April 29, 2003

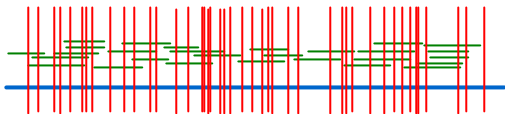
Hierarchical Sequencing Strategy



- Obtain a large collection of BAC clones
- Map them onto the genome (Physical Mapping)
- Select a minimum tiling path
- Sequence each clone in the path with shotgun
- Assemble
- Put everything together

Lecture 9, Tuesday April 29, 2003

2. Digestion



Restriction enzymes cut DNA where specific words appear

- Cut each clone separately with an enzyme
- Run fragments on a gel and measure length
- Clones C_A, C_B have fragments of length $\{l_i, l_j, l_k\} \Rightarrow$ overlap

Double digestion:

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

Lecture 9, Tuesday April 29, 2003

Online Clone-by-clone The Walking Method



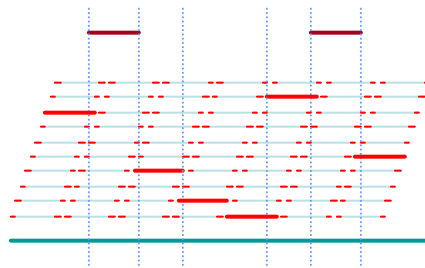
Lecture 9, Tuesday April 29, 2003

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 9, Tuesday April 29, 2003

Walking: An Example



Lecture 9, Tuesday April 29, 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 9, Tuesday April 29, 2003

Walking off a Single Seed



- Low redundant sequencing
- Many sequential steps

Lecture 9, Tuesday April 29, 2003

Walking off a single clone is impractical

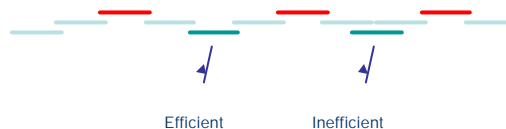
Cycle time to process one clone: 1-2 months

1. Grow clone
2. Prepare & Shear DNA
3. Prepare shotgun library & perform shotgun
4. Assemble in a computer
5. Close remaining gaps

A mammalian genome would need 15,000 walking steps !

Lecture 9, Tuesday April 29, 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 9, Tuesday April 29, 2003

Using Two Libraries

Most inefficiency comes from closing a small ocean with a much larger clone

Solution: Use a second library of small clones

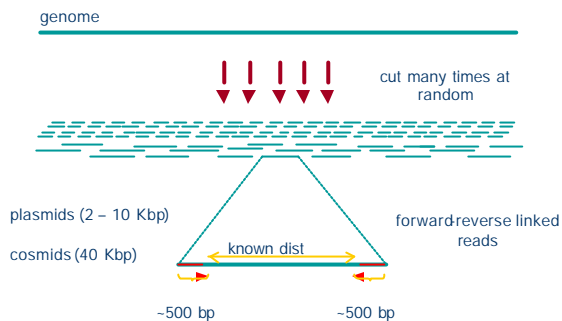
Lecture 9, Tuesday April 29, 2003

Whole-Genome Shotgun Sequencing



Lecture 9, Tuesday April 29, 2003

Whole Genome Shotgun Sequencing



Lecture 9, Tuesday April 29, 2003

ARACHNE: Steps to Assemble a Genome

1. Find overlapping reads

2. Merge good pairs of reads into longer contigs

3. Link contigs to form supercontigs

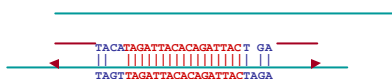
4. Derive consensus sequence

..ACGATTACAATAGGTT..

Lecture 9, Tuesday April 29, 2003

1. Find Overlapping Reads

- Sort all k-mers in reads ($k = 24$)
- Find pairs of reads sharing a k-mer
- Extend to full alignment – throw away if not >95% similar



Lecture 9, Tuesday April 29, 2003

1. Find Overlapping Reads

One caveat: repeats

A k-mer that appears N times, initiates N^2 comparisons

ALU: 1,000,000 times

Solution:

Discard all k-mers that appear more than $c \times \text{Coverage}$, ($c = 10$)

Lecture 9, Tuesday April 29, 2003

1. Find Overlapping Reads

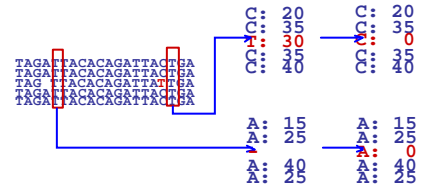
Create local multiple alignments from the overlapping reads



Lecture 9, Tuesday April 29, 2003

1. Find Overlapping Reads (cont'd)

- Correct errors using multiple alignment



- Score alignments
- Accept alignments with good scores

Lecture 9, Tuesday April 29, 2003

Basic principle of assembly

Repeats confuse us

Ability to merge two reads is related to our ability to detect repeats

We can dismiss as repeat any overlap of $< 1\%$ similarity

Role of error correction:

Discards ~90% of single-letter sequencing errors

⇒ Threshold 1% increases

Lecture 9, Tuesday April 29, 2003

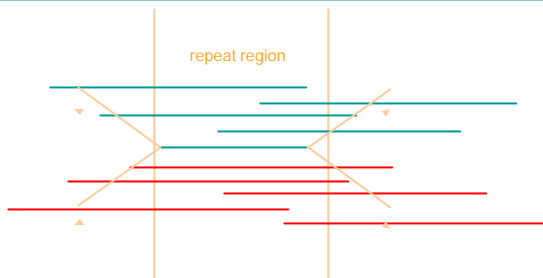
2. Merge Reads into Contigs (cont'd)



Merge reads up to potential repeat boundaries
(Myers, 1995)

Lecture 9, Tuesday April 29, 2003

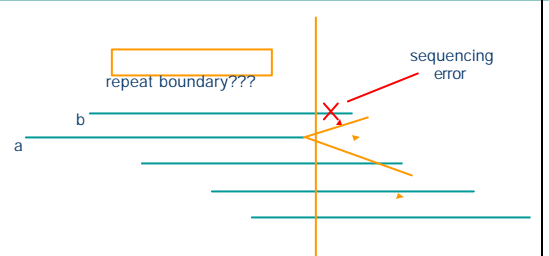
2. Merge Reads into Contigs (cont'd)



- Ignore non-maximal reads
- Merge only maximal reads into contigs

Lecture 9, Tuesday April 29, 2003

2. Merge Reads into Contigs (cont'd)



- Ignore "hanging" reads, when detecting repeat boundaries

Lecture 9, Tuesday April 29, 2003

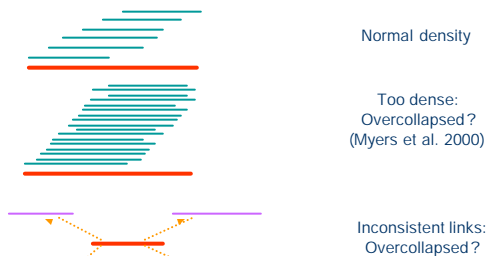
2. Merge Reads into Contigs (cont'd)



- Insert non-maximal reads whenever unambiguous

Lecture 9, Tuesday April 29, 2003

3. Link Contigs into Supercontigs



Lecture 9, Tuesday April 29, 2003

3. Link Contigs into Supercontigs (cont'd)

Find all links between unique contigs

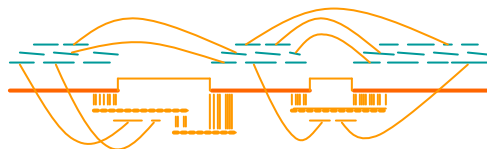
Connect contigs incrementally, if ≥ 2 links



Lecture 9, Tuesday April 29, 2003

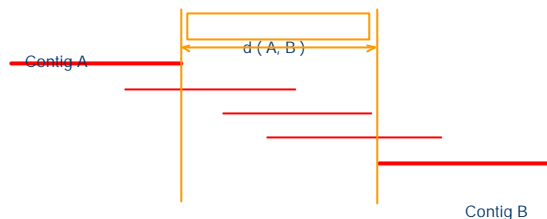
3. Link Contigs into Supercontigs (cont'd)

Fill gaps in supercontigs with paths of overcollapsed contigs



Lecture 9, Tuesday April 29, 2003

3. Link Contigs into Supercontigs (cont'd)

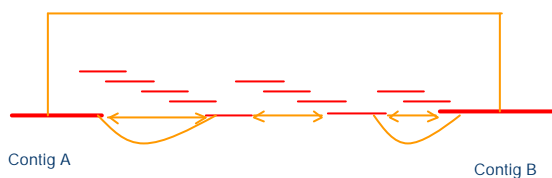


Define $G = (V, E)$
 V := contigs
 E := (A, B) such that $d(A, B) < C$

Reason to do so: Efficiency; full shortest paths cannot be computed

Lecture 9, Tuesday April 29, 2003

3. Link Contigs into Supercontigs (cont'd)



Contig A Contig B

Define T : contigs linked to either A or B

Fill gap between A and B if there is a path in G passing only from contigs in T

Lecture 9, Tuesday April 29, 2003

4. Derive Consensus Sequence

```

TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTATGGCGTAAACTA
TAG TTACACAGATTATGACTT CATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGTATGGCGTAA CTA
    
```

↓ ↓ ↓ ↓

TAGATTACACAGATTACTGACTTGTATGGCGTAA CTA

Derive **multiple alignment** from pairwise read alignments

Derive each consensus base by weighted voting

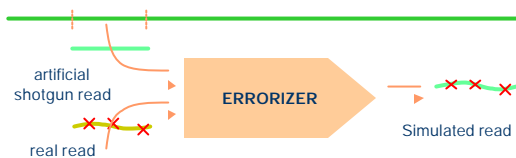
Lecture 9, Tuesday April 29, 2003

Simulated Whole Genome Shotgun

- Known genomes
Flu, yeast, fly, Human chromosomes 21, 22
- Make "realistic" shotgun reads
- Run ARACHNE
- Align output with genome and compare

Lecture 9, Tuesday April 29, 2003

Making a Simulated Read



Simulated reads have error patterns taken from random real reads

Lecture 9, Tuesday April 29, 2003

Human 22, Results of Simulations

Plasmid/ Cosmid	5 X / 0.5 X	1 X / 0.5 X	3 X / 0.5 X
	959 Kb	15 Kb	2.7 Kb
	142 Kb	10.6 Kb	2.0 Kb
	3 Mb	3 Mb	4.1 Kb
	41	32	26
	97.3	91.1	67

Lecture 9, Tuesday April 29, 2003

Neurospora crassa Genome (Real Data)

- 40 Mb genome, shotgun sequencing complete (W-CGR)
- Evaluated assembly using 1.5Mb of finished BACs
- 1% uncovered (of finished BACs)

Coverage:
1705 contigs
368 supercontigs

Efficiency:
Time: 20 hr
Memory: 9 Gb

Accuracy:
< 3 misassemblies
compared with 1 Gb of
finished sequence
Errors/10⁸ letters:
Subst. 260

10/20/01, 1/2/02



Mouse Genome

Improved version of ARACHNE assembled the mouse genome

Several heuristics of iteratively:
Breaking supercontigs that are suspicious
Rejoining supercontigs

Size of problem: 32,000,000 reads

Time: 15 days, 1 processor
Memory: 28 Gb

N50 Contigsize: 16.3 Kb → 24.8 Kb
N50 Supercontig size: .265 Mb → 16.9 Mb

Lecture 9, Tuesday April 29, 2003