Sequencing and fragment assembly: The human genome

The sequencing the human genome was a tremendous scientific accomplishment, requiring large-scale collaboration between computational and life sciences.

However, the *intellectual* breakthrough that lead to successful sequencing came from computer scientists, not life scientists.

We will study the basic technology underlying all sequencing projects, and compare the somewhat different experimental strategies employed by the two groups.

Amazing progress in the scale of sequencing projects has been achieved, largely through automation:

Phage Phi-x174: 5kb, 1977.

Bacteriophage lambda: 50kb, 1982.

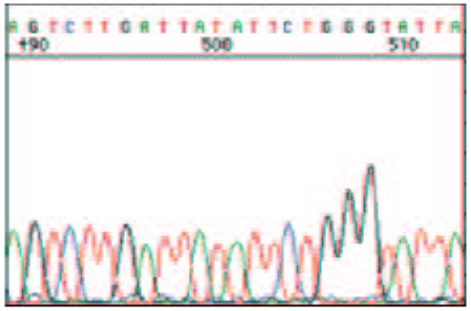
Haemophilus influenzae: 1.8Mb, July 1995.

Drosophila: 180Mb, March 2000.

Human: 3 billion bases, February 2001.

DNA sequencing machines

Sequencing machines today use the same basic principles as the original Gilbert-Sanger method. There has been tremendous progress in automating the procedure, however.

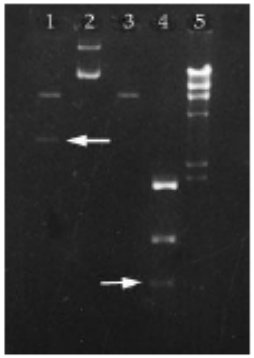


Read lengths have gotten only slightly longer with time, perhaps from 500 bp to 700 bp.

The sample to be sequenced is replicated in four distinct bins, using a distinct fluorescently labeled PCR primer. The bin associated with a given base x is given a mixture of functional and non-functional versions of x, where non-functional bases terminate transcription. This creates labeled fragments of all sizes ending in x.

Using gel electrophoresis, the fragments are separated by length. The presence or absence of a labeled band in each lane denotes whether the sequence has the given base in each position.

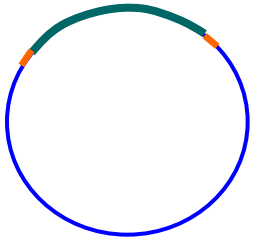
DNA sequencing machines



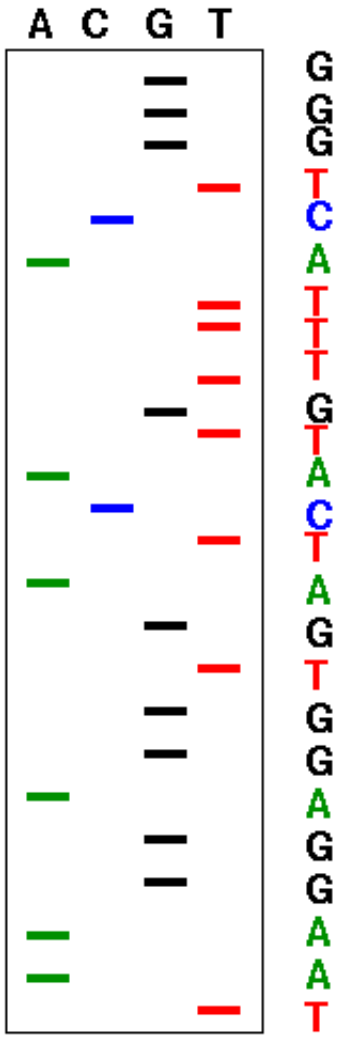
Modern capillary machines use smaller amounts of reagents and avoid problems with wandering lanes.

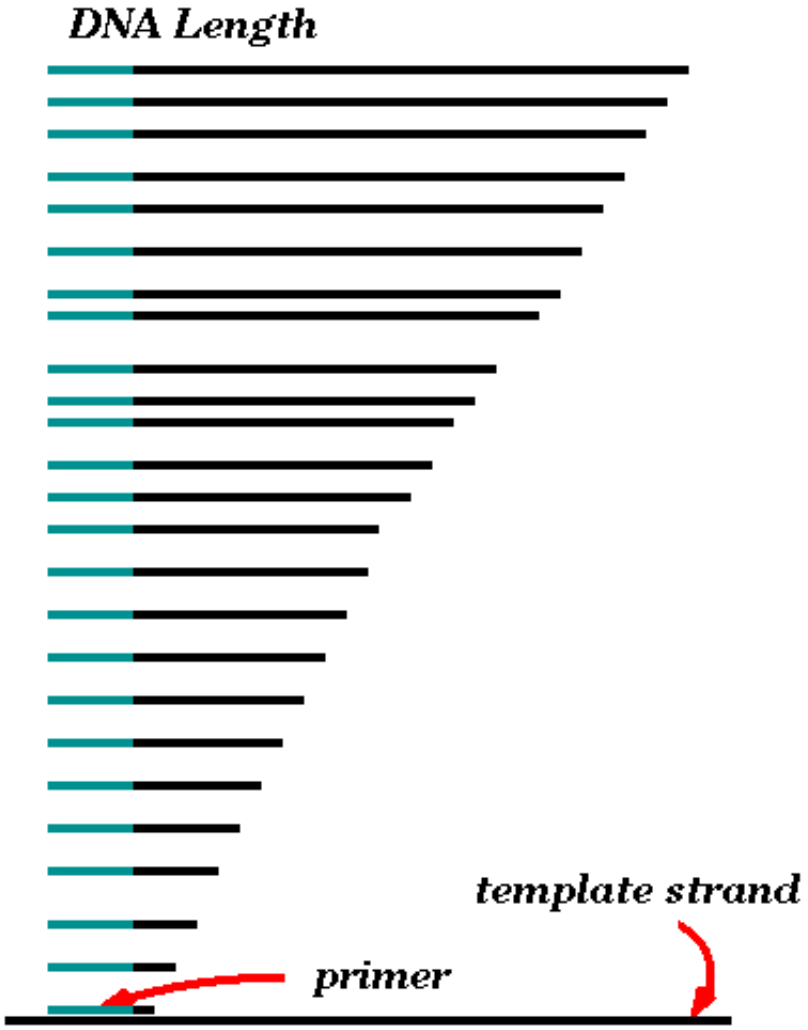
In the good regions of a read, the base error rate should be below 2%.

DNA Sequencing – gel electrophoresis



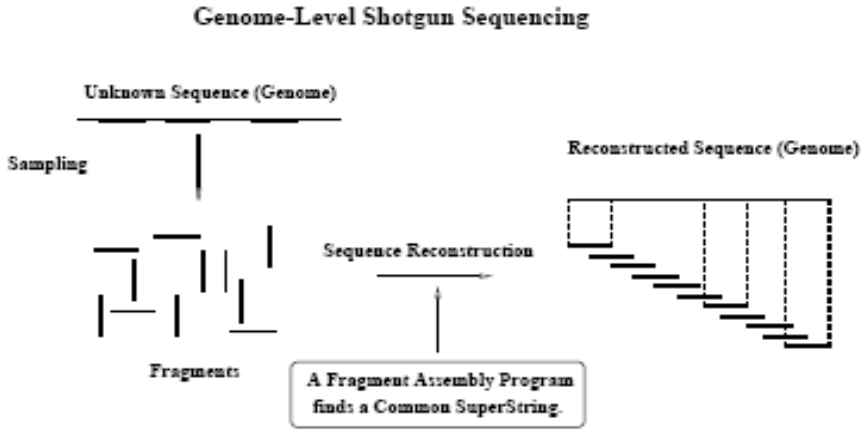
1. Start at primer (restriction site)
2. Grow DNA chain
3. Include dideoxynucleoside (modified a, c, g, t)
4. Stops reaction at all possible points
5. Separate products with length, using gel electrophoresis





DNA sequencing machines

In traditional *shotgun sequencing*, whole genomes are sequenced by making clones, breaking them into small pieces, and trying to put the pieces together again based on overlaps.



Note that the fragments are *randomly* sampled, and thus no positional information is available.

Gaps

Since we rely on fragment overlaps to identify their position, we must sample sufficient fragments to ensure enough overlaps.

Let T be the length of the target molecule being sequenced using n random fragments of length l, where we recognize all overlaps of length t or greater.

The *Lander-Waterman* equation gives the expected number of *gaps* g as

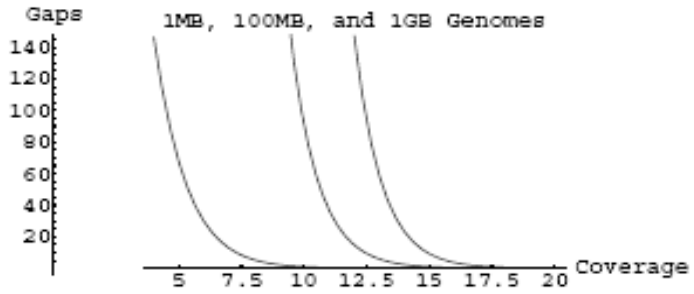
$$g = ne^{-n(l-t)/T}$$

Where does the e come from? Suppose we have as many fragments as bases, i.e. $T = n$ and each fragment is length 1. The probability p that base i is *not* sampled is

$$p = (\frac{n-1}{n})^n \rightarrow \frac{1}{e}$$

Coverage

The *coverage* of a sequencing project is the ratio of the total sequenced fragment length to the genome length, i.e. nl/T.



Gaps are very difficult and expensive to close in any sequencing strategy, meaning that very high coverage is necessary to use shotgun sequencing on a large genome.

Sequencing strategies

The effectiveness of a genome sequencing strategy depends upon the degree of *coverage*, the length of the inserts, and the auxiliary *mapping* information available to help assembly.

The DNA fragments or *clones* are replicated by inserting them into a living organism, the *cloning vector*.

Small fragments (40,000 bp) can be cut and pasted into a bacterial *cosmid*. Bigger fragments (up to 2,000,000 bp) can be replicated as a bacterial or yeast artificial chromosome, a *BAC* or *YAC*.

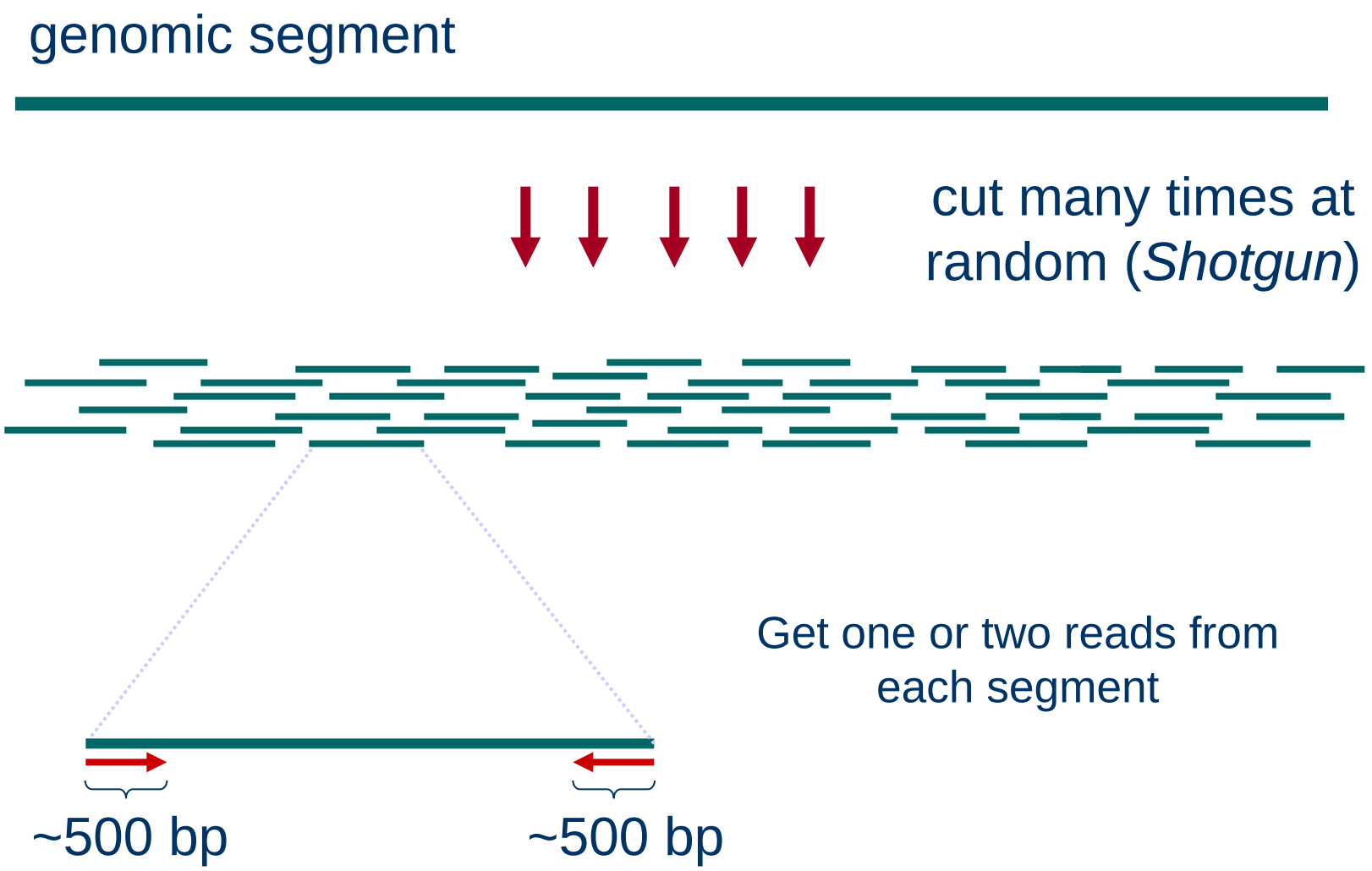
Sequencing strategies

After sequencing both ends of a given insert, we know roughly how far apart they should be in the final assembly.

Selecting the right mix of insert sizes can simplify assembly. Small inserts give tight assembly constraints, but big inserts help us build a scaffolding across the entire genome.

The internals of clones can be sequenced, but it is much more expensive than end sequencing. Thus it is done only in the closing gaps.

Method to sequence longer regions

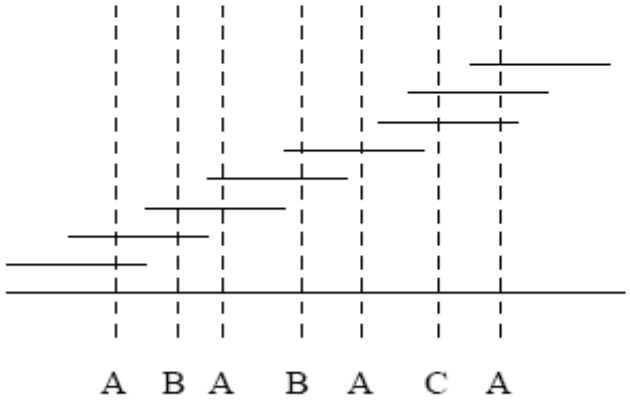


Mapping data

The high coverage necessary to sequence large genomes without gaps frightened most laboratories away from pure shotgun sequencing strategies.

A different approach is to construct a *map* showing where each clone lies on the human genome, and use this map to guide end sequencing and assembly.

Mapping data can be based on (1) using hybridization to detect the presence or absence of a given short sequence (*STS*) in a given clone, or (2) using *restriction enzymes* to cut each clone at a given pattern, and looking for similar fragment lengths.

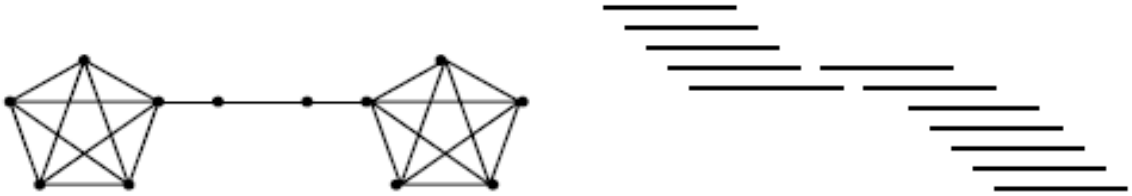


With a good enough map, the required coverage might go down to 2 or 3.

The algorithmics of mapping

Note that the correct ordering is a Hamiltonian path on the clones. Reconstructing clone order from mapping data tends to be an NP-complete.

However, the difficulty is due to errors and ambiguity in the mapping data since the problem of recognizing *interval graphs* can be done in linear time!



The public consortium used a sequencing strategy based on mapping the clones first.

Celera used hundreds of high-throughput sequencing machines to obtain enough coverage to shotgun sequence the human genome.

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.

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

```
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
```

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

Even worse, we have to deal with significant errors in the sequence fragments.

Even worse, genomes tend to have many *repeats* (approximate copies of the same sequence), which are very hard to identify and reconstruct.

Due to repeats, the shortest common superstring is typically *shorter* than the real sequence.

Overlap detection

Even worse, the size of the problem is very large. Celera's Human Genome sequencing project contained roughly 26.4 million fragments, each about 550 bases long.

To decide what overlaps what, we *could* compare each fragment against each other fragment via $O(n^2)$ dynamic programming, but faster methods are needed.

$$(26.4 \text{ million})^2 \times (550)^2 = 2.1 \times 10^{20} \text{ operations!}$$

Celera's assembly involved 500 million trillion base to base comparisons, requiring over 20,000 CPU (central processor unit) hours on their supercomputer.

Thus efficient overlap detection is critical, more critical than the NP-complete part of the problem!

Overlap detection must be tolerant of sequencing error, but even an error rate of 2% means one should be able to find fairly long ($\sim$ 25 bp) exact matches in a long overlap.