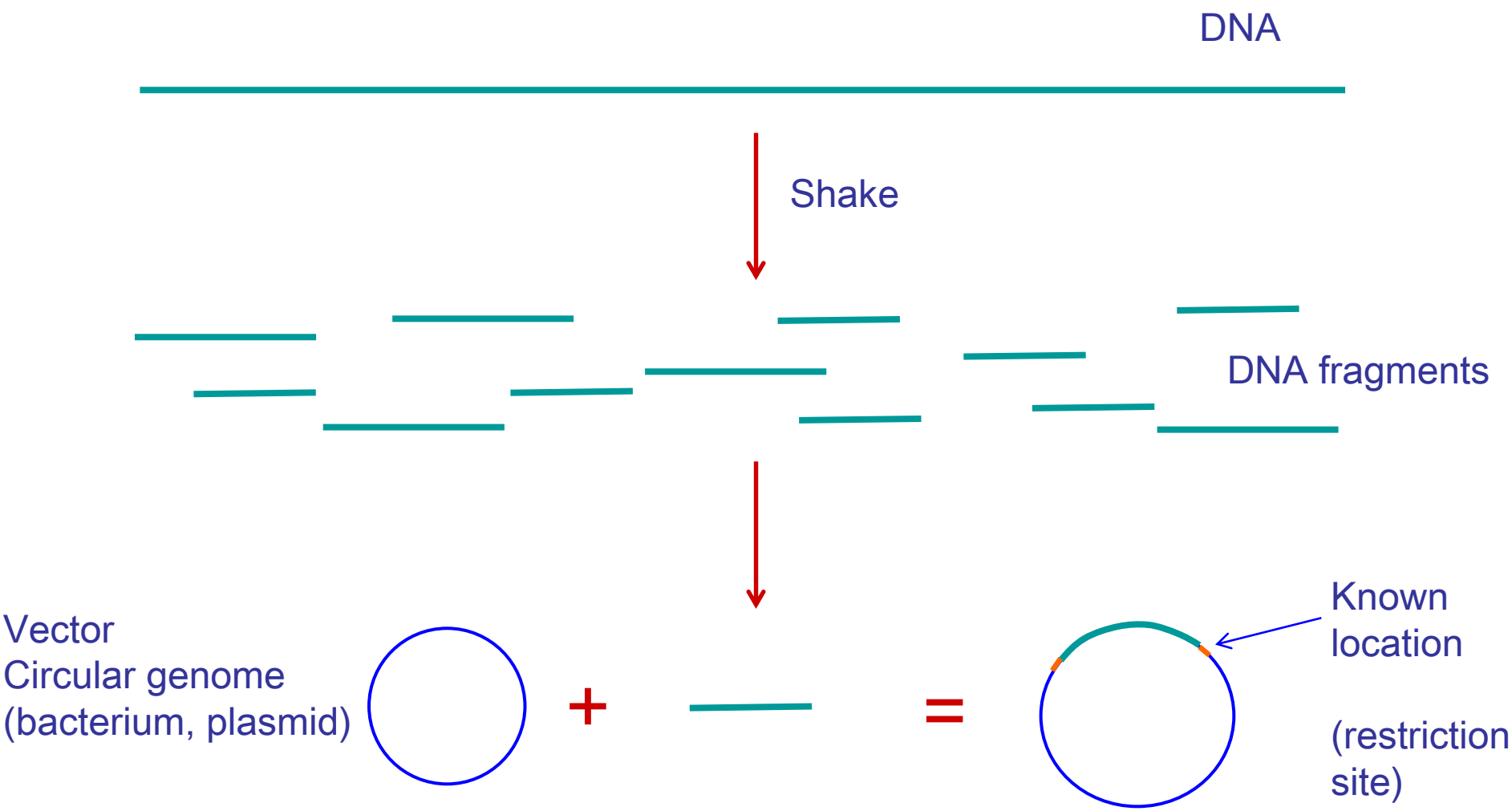
Creating Artificial Datasets

Michael Panitz
Mathias Ganter

Computational Biology CSE 527
University of Washington

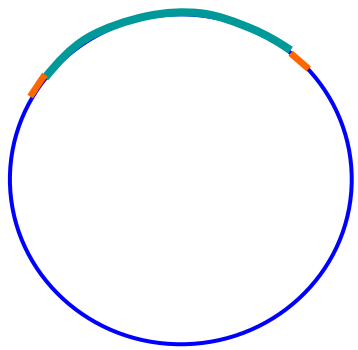
13h December 2004

DNA sequencing – vectors

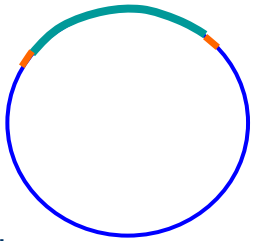


Different types of vectors

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



DNA sequencing – gel electrophoresis



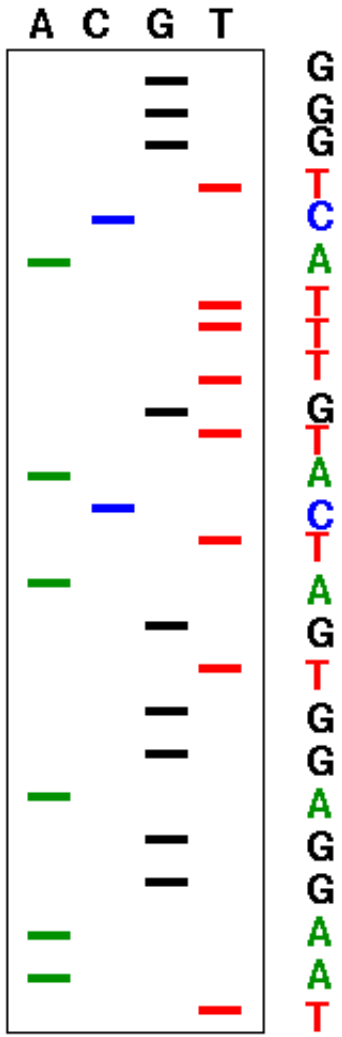
Start at primer
(restriction site)

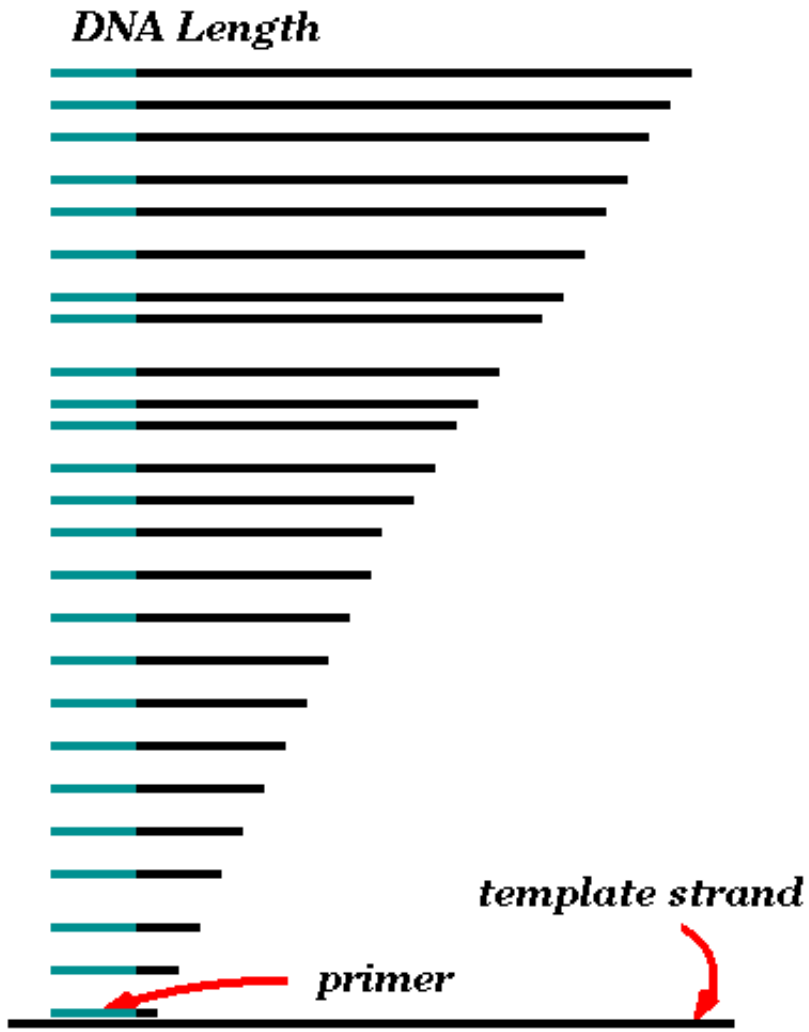
Grow DNA chain

Include dideoxynucleoside
(modified a, c, g, t)

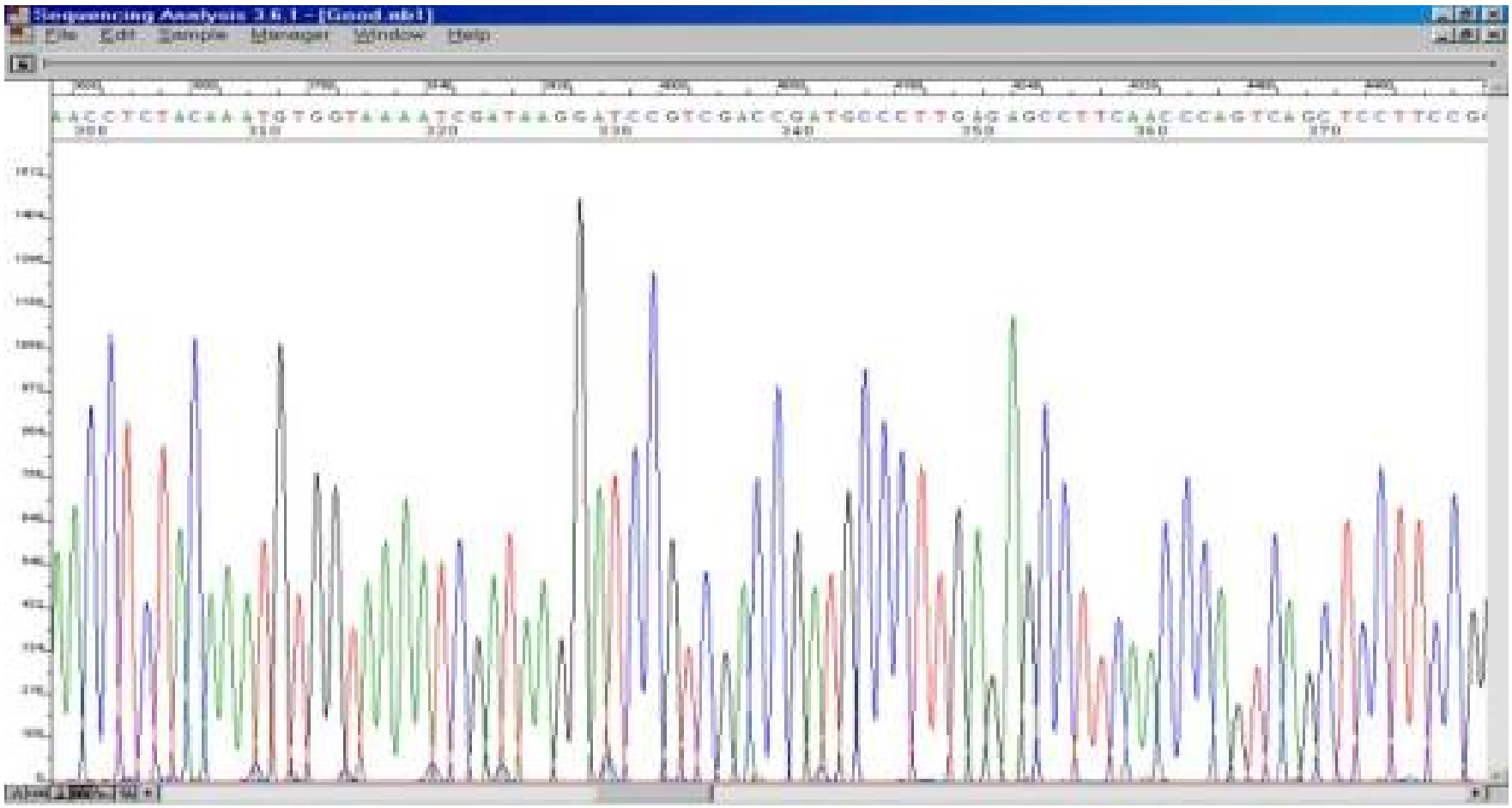
Stops reaction at all
possible points

Separate products with
length, using
gel electrophoresis





Electrophoresis diagrams



Output of gel electrophoresis: a *read*

A read: 500-700 nucleotides

A C G A A T C A G …. A

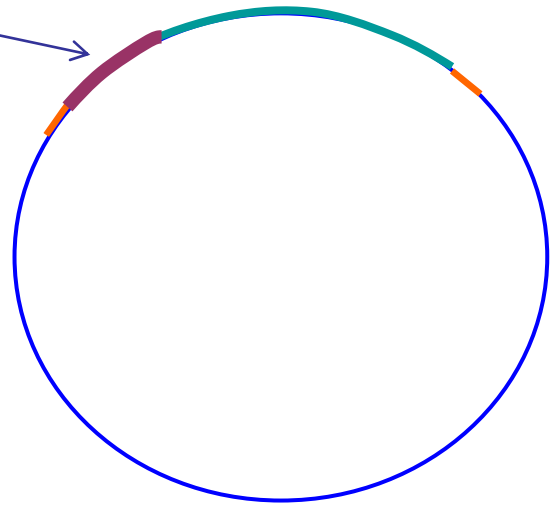
16 18 21 23 25 15 28 30 32 21

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

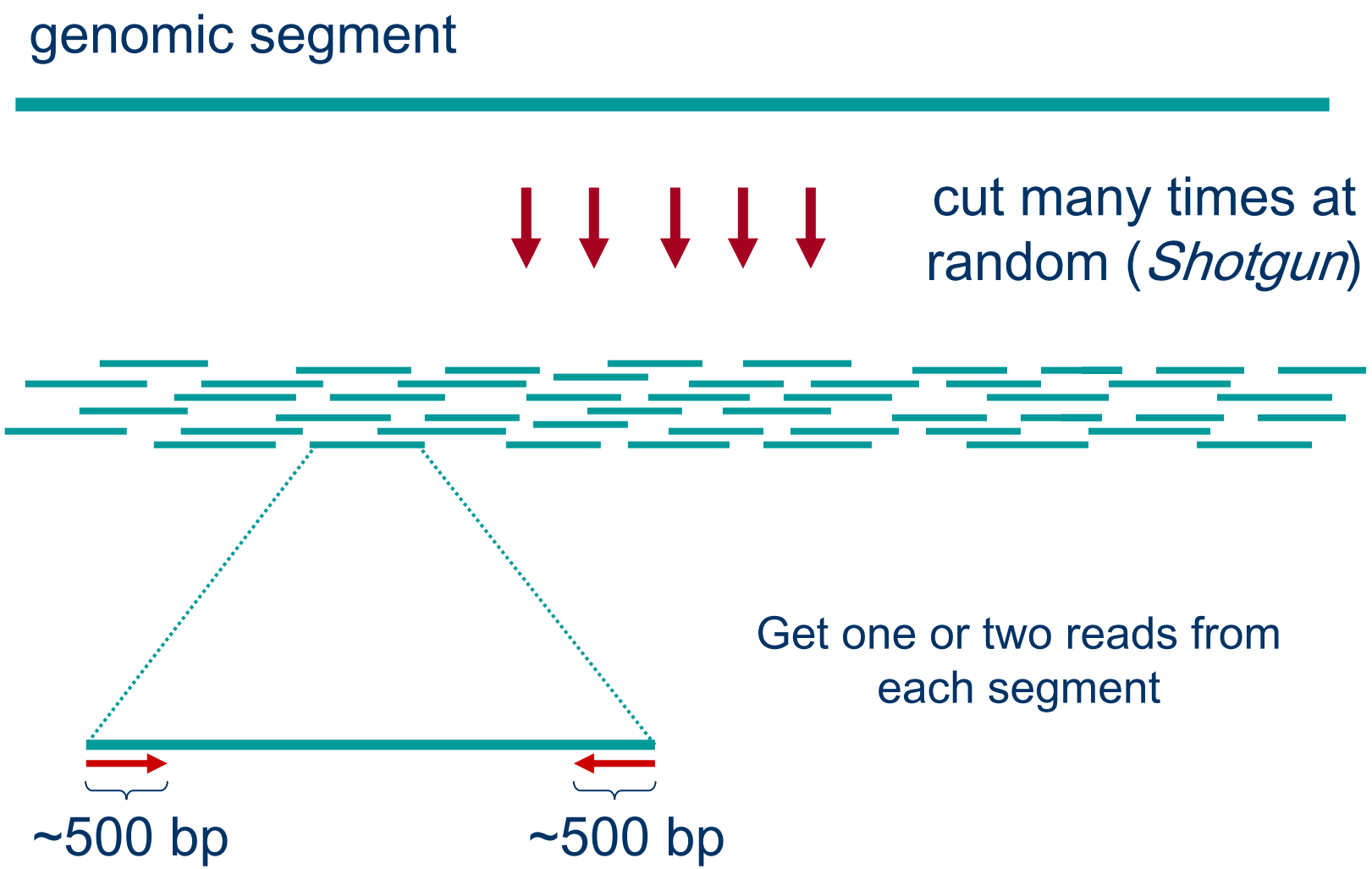
Reads can be obtained from leftmost, rightmost ends of the insert

Double-barreled sequencing:

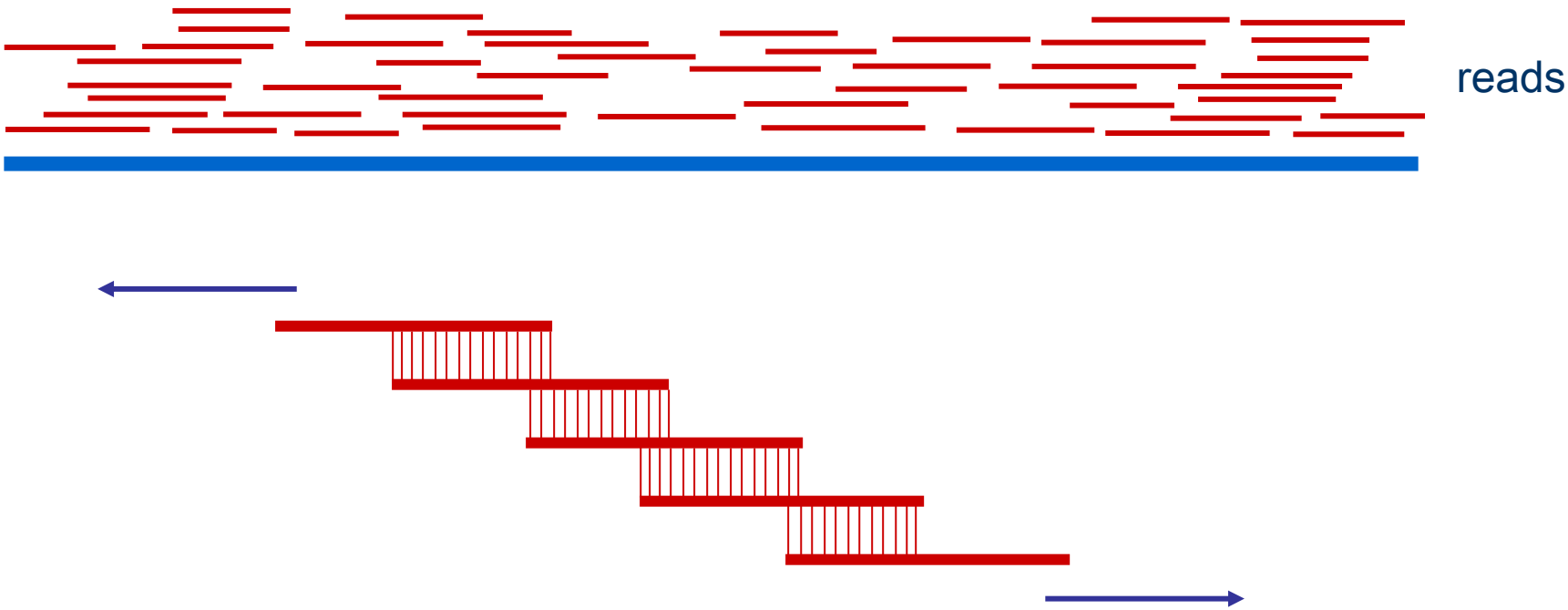
Both leftmost & rightmost ends are sequenced



Method to sequence segments longer than 500



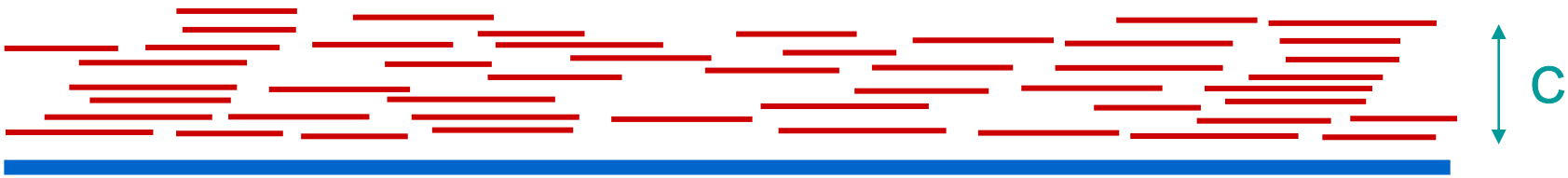
Reconstructing the Sequence (Fragment Assembly)



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

Overlap reads and extend to reconstruct the original genomic region

Definition of Coverage



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

Definition: Coverage C = nl/L

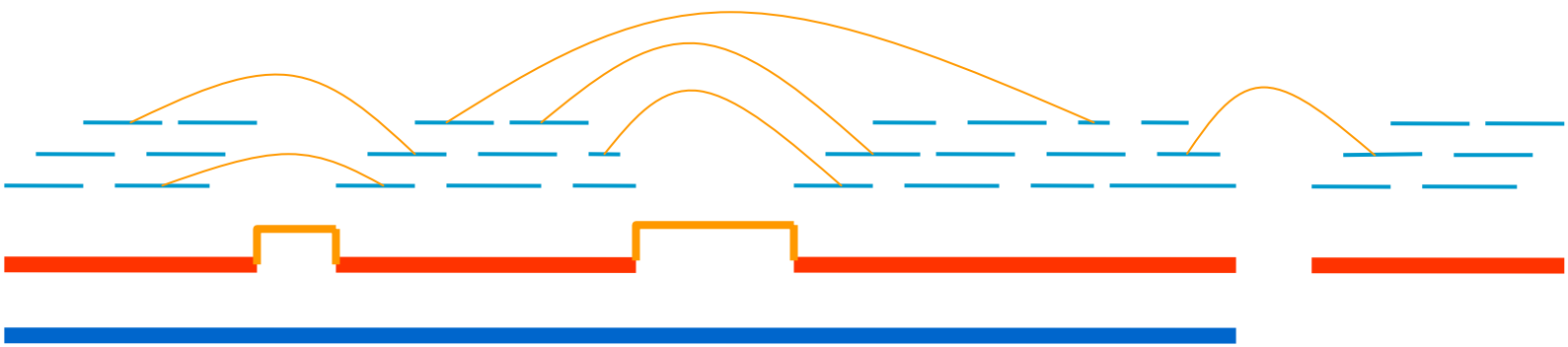
How much coverage is enough?

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

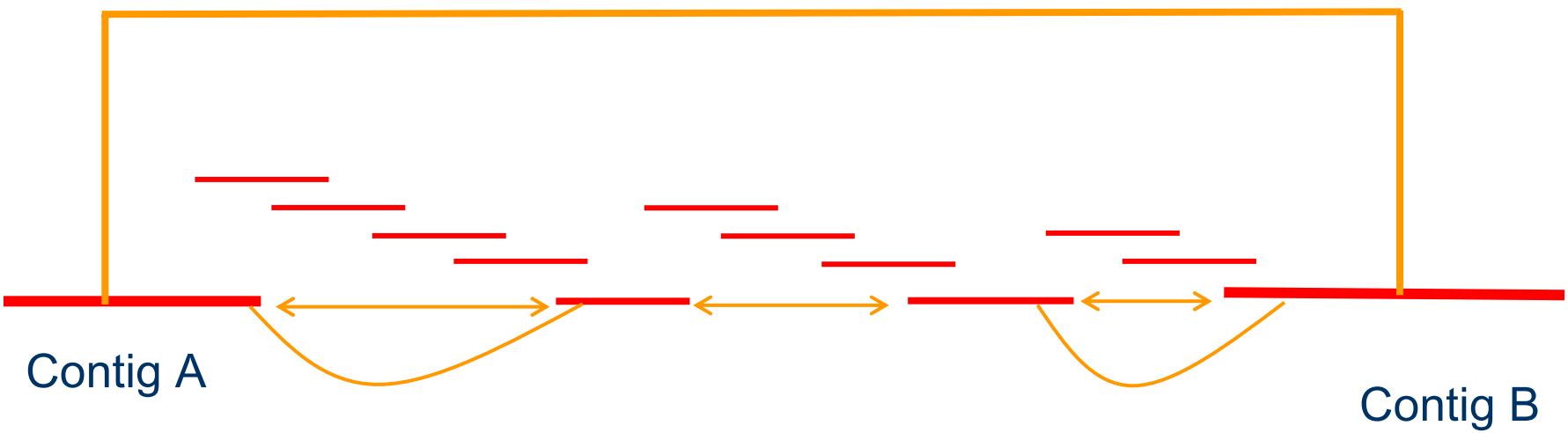
3. Link Contigs into Supercontigs (cont'd)

Find all links between unique contigs

Connect contigs incrementally, if ≥ 2 links



3. Link Contigs into Supercontigs



Define T: contigs linked to either A or B

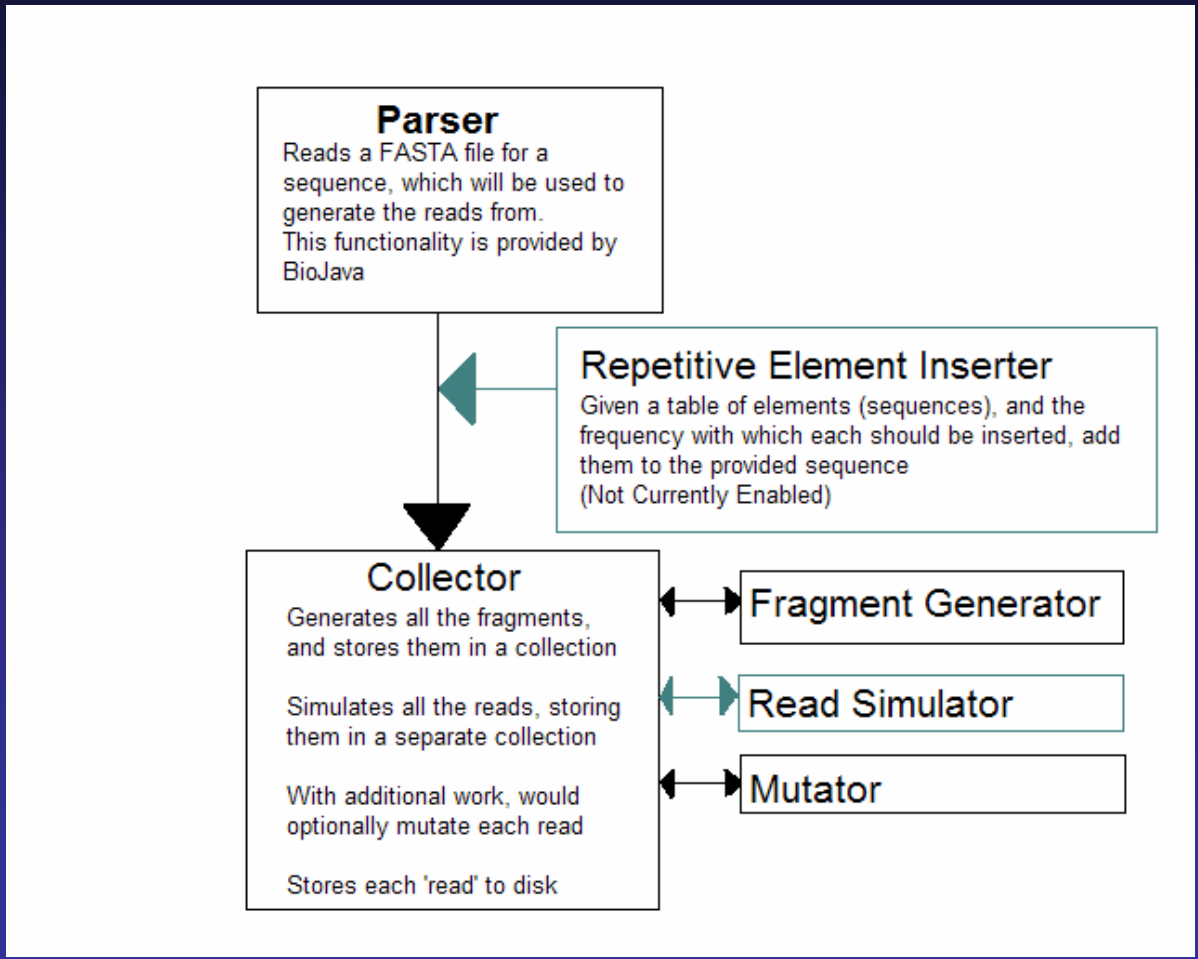
Fill gap between A and B if there is a path in G passing only from contigs in T

4. Derive Consensus Sequence

Derive multiple alignment from pairwise read alignments

Derive each consensus base by weighted voting

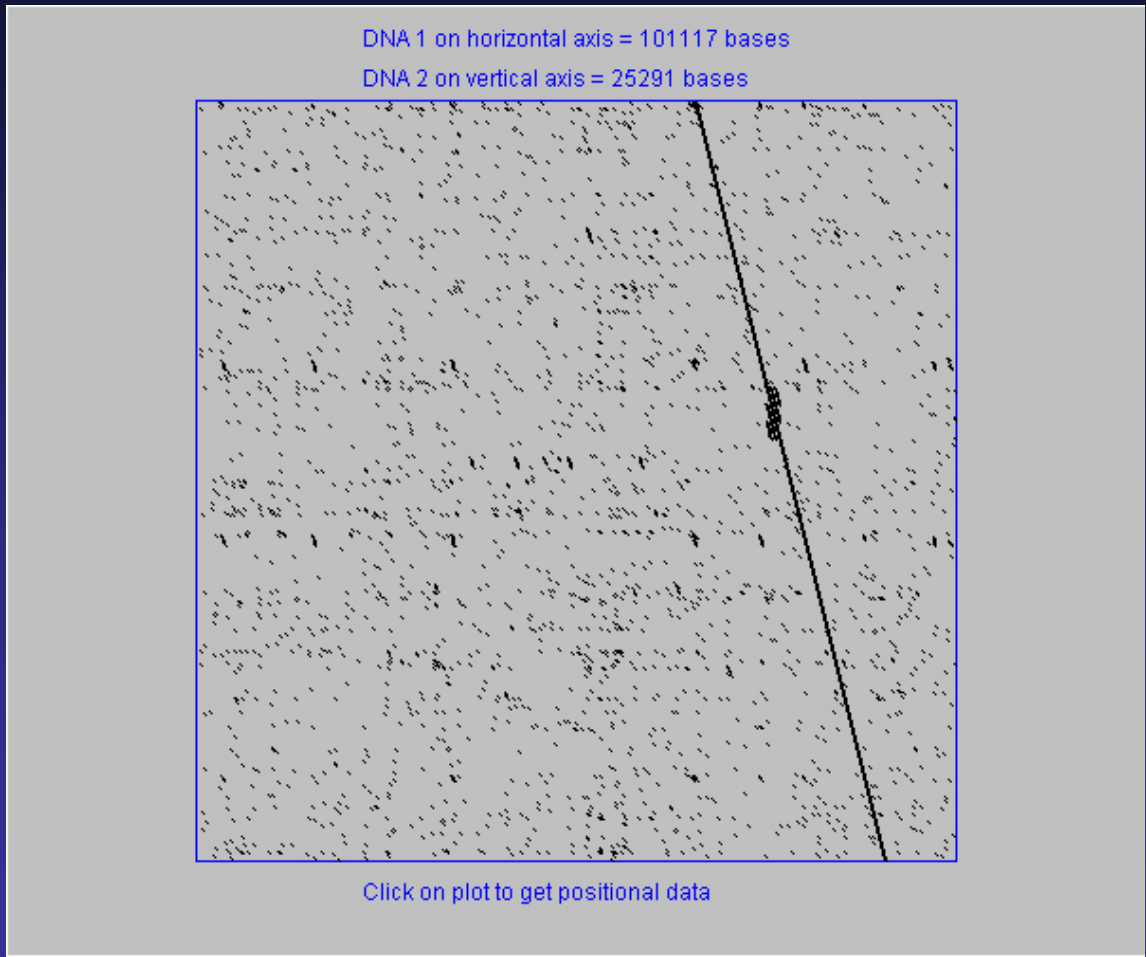
Implementation Details



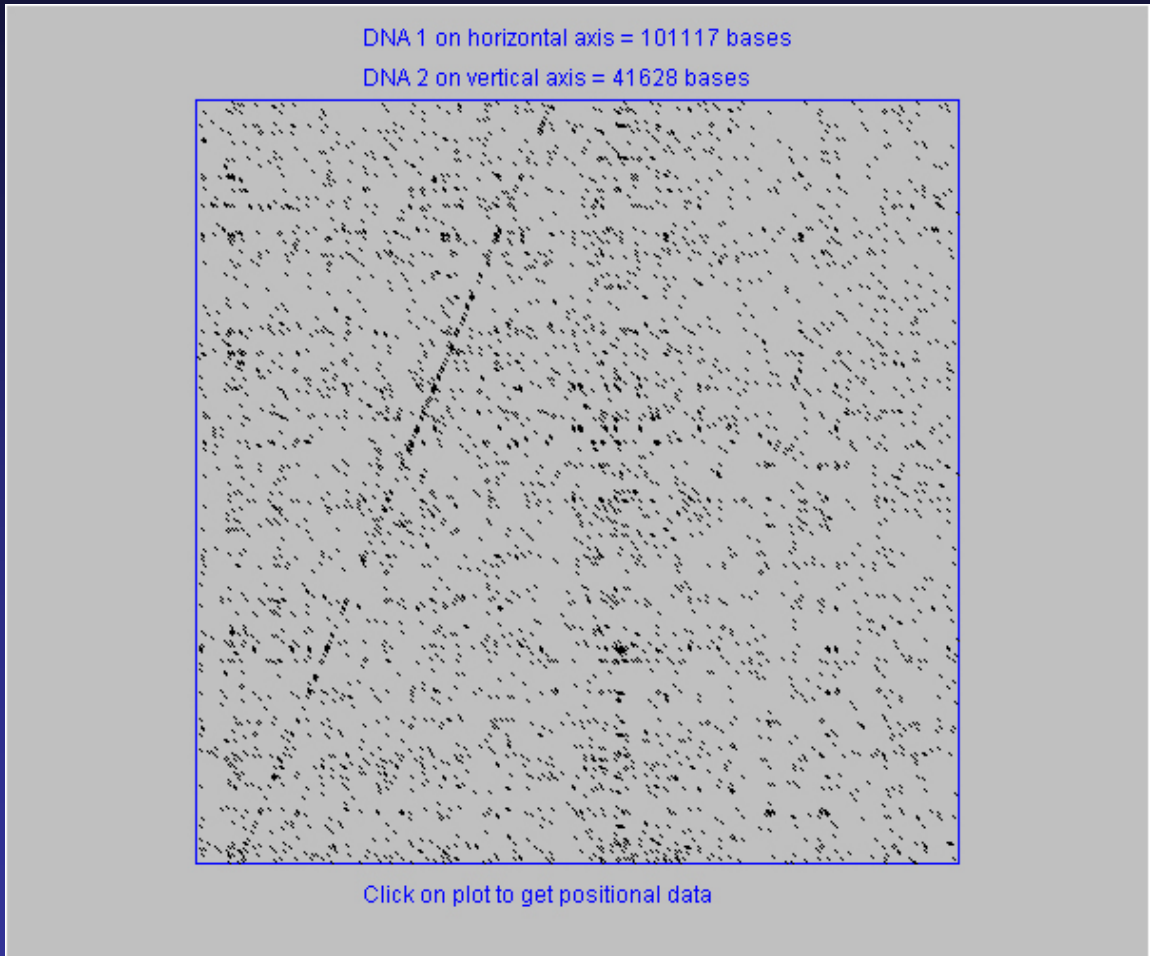
Analyzing Methods – Dot Plots

- developed in the early 1980s
- similarity matrix and a visual representation between two sequences
- provide an easy and powerful means of sequence analysis, useful for searching out regions of similarity in two sequences and repeats within a single sequence.
- The principle
  - A matrix comparison of two sequences (or one with itself) is prepared by "sliding" a window of user-defined size (called window size) along both sequences.
  - If the two sequences within that window match with a precision set by the mismatch limit, a dot is placed in the middle of the window signifying a match. Variations in both the size of the sliding window and the stringenthy factor can be used to separate more significant data from less important data.

Explanation and Results



Explanation and Results (continued)



Analyzing Methods

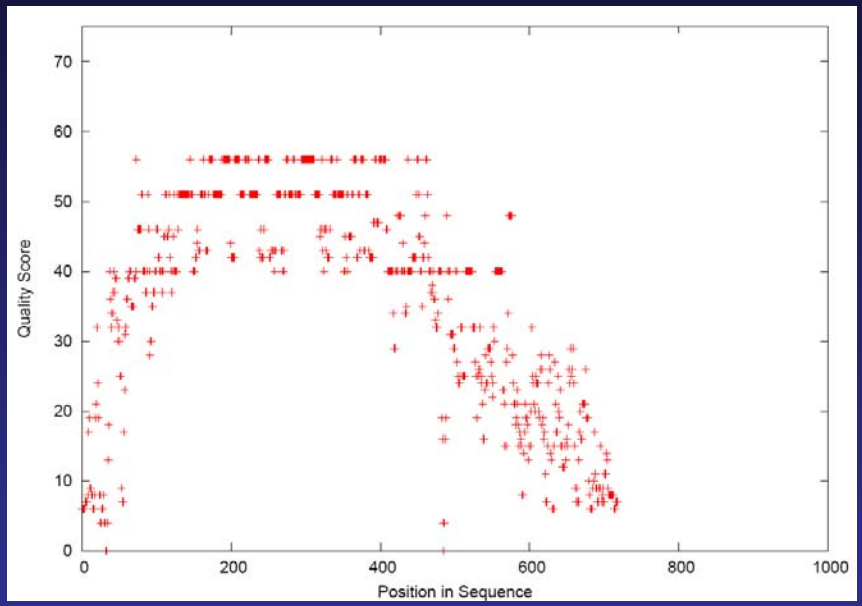
- There are 3 key issues:
  - What kind of algorithms to use to find an alignment (Needleman-Wunsch, Smith-Waterman, FSA-model, HMM, multiple sequence alignments)
  - What kind of scoring systems to use to rank alignments
  - What kind of statistical methods to use to evaluate the significance of an alignment score

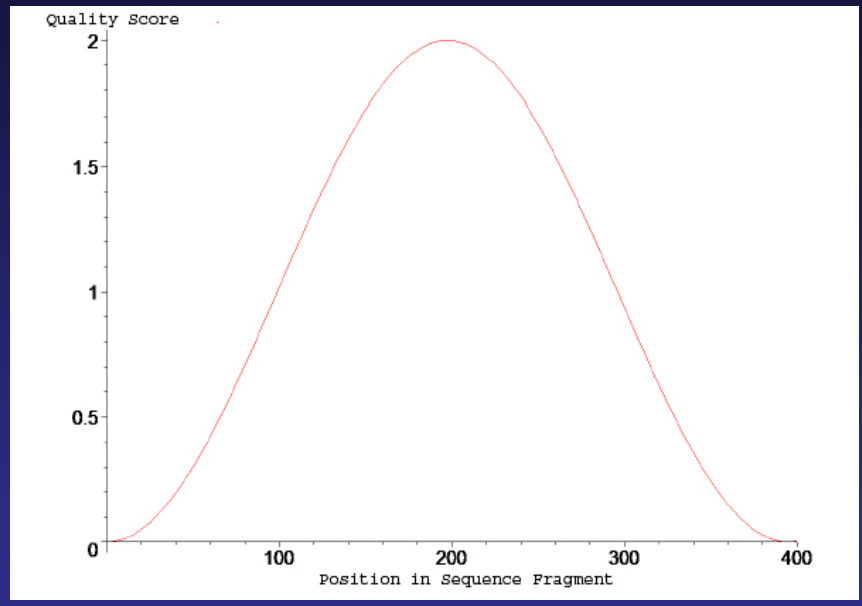
- A different approach we though of is using phylogenetic analyzing methods to explore the relationship between various generated sequences and the initial one.
  - This can be done because we used one initial sequence and built all the other sequences out of this one. One could say that they all diverged from one common ancestor by a simulated process of mutation and selection. This can be interpreted as the relative likelihood that the sequences are related, compared to being unrelated.

Areas of Future Work

- Better quality score function
- Synthetic data sets
  - Generating data sets from scratch, given user parameters
    - Repetitive element insertion
- Mutations
  - Realistic rates of substitution, indel
- Numeric metric for accuracy of reassembled contigs
- ‘Pipelining’ – generate one read at a time to save memory

Quality function





$$score = A \cdot \cos(w \cdot x + q) + z$$