Lecture 1

Introduction:
The Molecular Basis of Life

An Historical Perspective

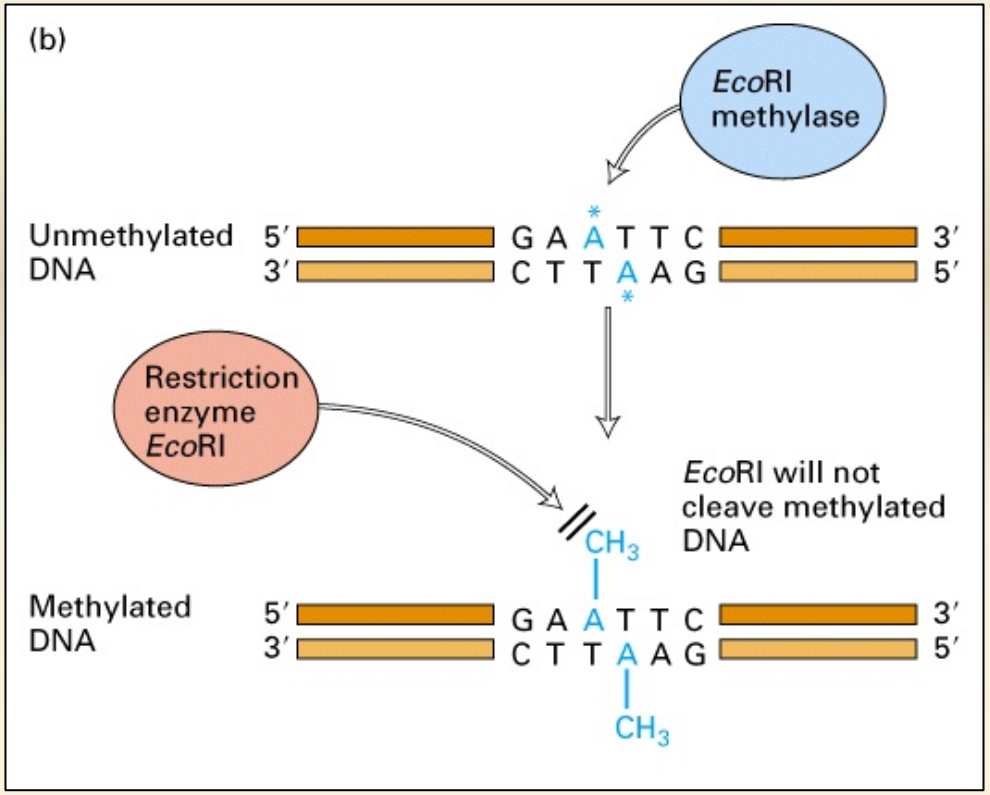
- … – 1900 Pre-Mendelian period
- 1900 – 1940 Pre-DNA period
- 1940 – 1990 DNA period
- ➢ 1990 – 2003 Genomic period
- 2003 – … Post-genomic era

Modern Biology

- Mechanism
- Cell theory
- Evolution

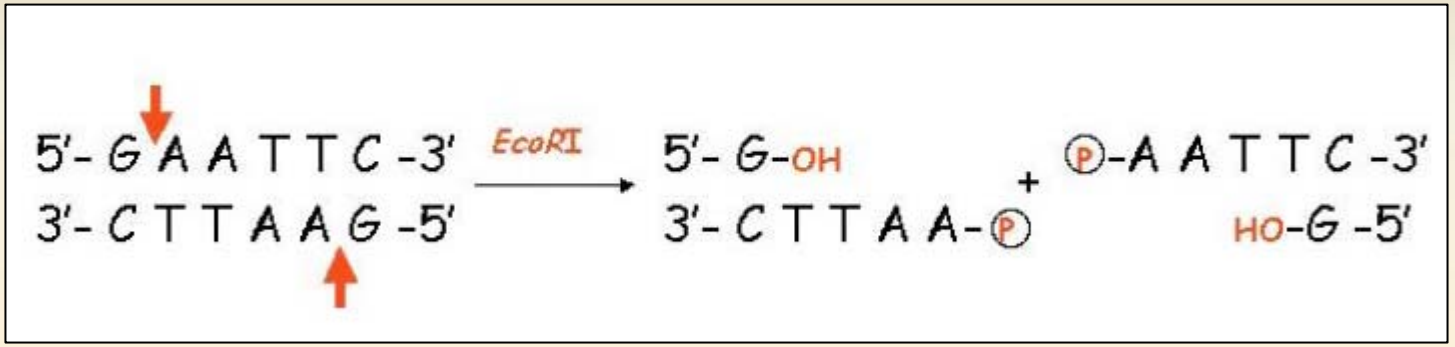
Manipulating DNA

- Restriction enzymes

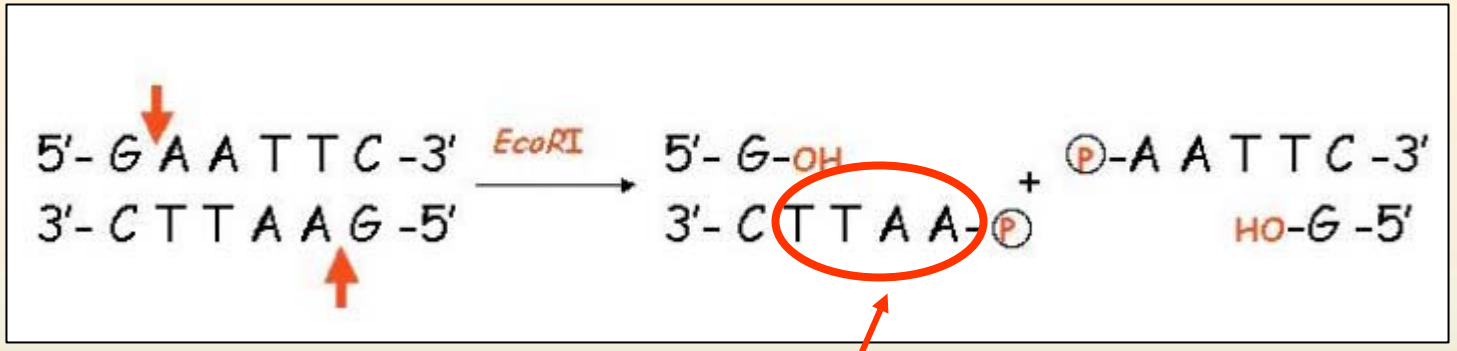


Manipulating DNA

- Restriction enzymes
  - Can cut DNA duplex at specific sites (palindrome sequence).
  - Do not discriminate between DNA from different organisms
  - A natural part of the bacterial defense system
  - High specificity for their recognition site means that DNA will be cut reproducibly into defined fragments



Manipulating DNA



- Restriction enzymes
  - Produce *sticky ends* of a single-stranded DNA which can base-pair (anneal) with any complementary single-stranded DNA sequence

Manipulating DNA

- Restriction enzymes
- Cloning vectors – replicating systems in addition to chromosomes:
  - Plasmids and *BACs* in Prokaryotes
  - Artificial chromosomes in Yeasts (Eukaryotes), *YACs*
  - Detailed restriction map of cloning vector
  - Marker – antibiotic resistance

Manipulating DNA

- Restriction enzymes
- Cloning vectors
- Reverse transcriptase
  - makes transcription from RNA to DNA (retroviruses – HIV)
  - we can take a mRNA (unstable) of any expressed gene and transcribe it into the DNA sequence (stable, double-stranded)
  - this DNA is called *cDNA*

Manipulating DNA

- Restriction enzymes
- Cloning vectors
- Reverse transcriptase
- Recombinant DNA
  - Self-replicating system containing artificially introduced gene
  - Example: production of the insulin
  - Future: production of spider silk, biodegradation of waste

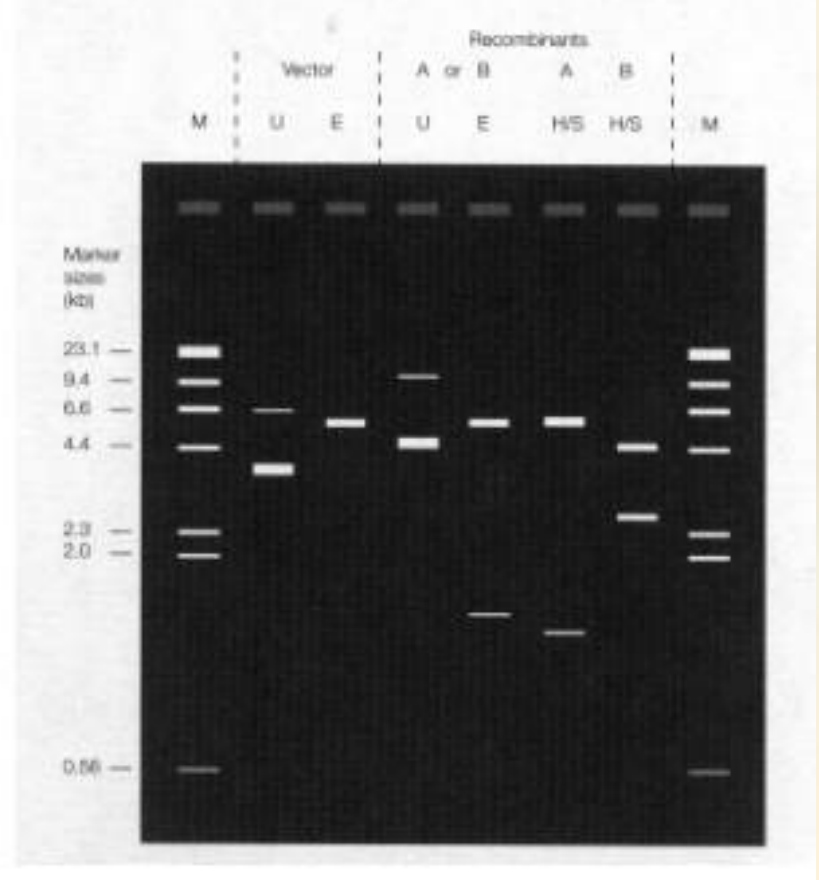
cDNA libraries

- Produce cDNA of a gene
- Clone this DNA in BAC, YAC or plasmid
- The amount of DNA sequence can be increased using Polymerase Chain Reaction (PCR)

Sequencing

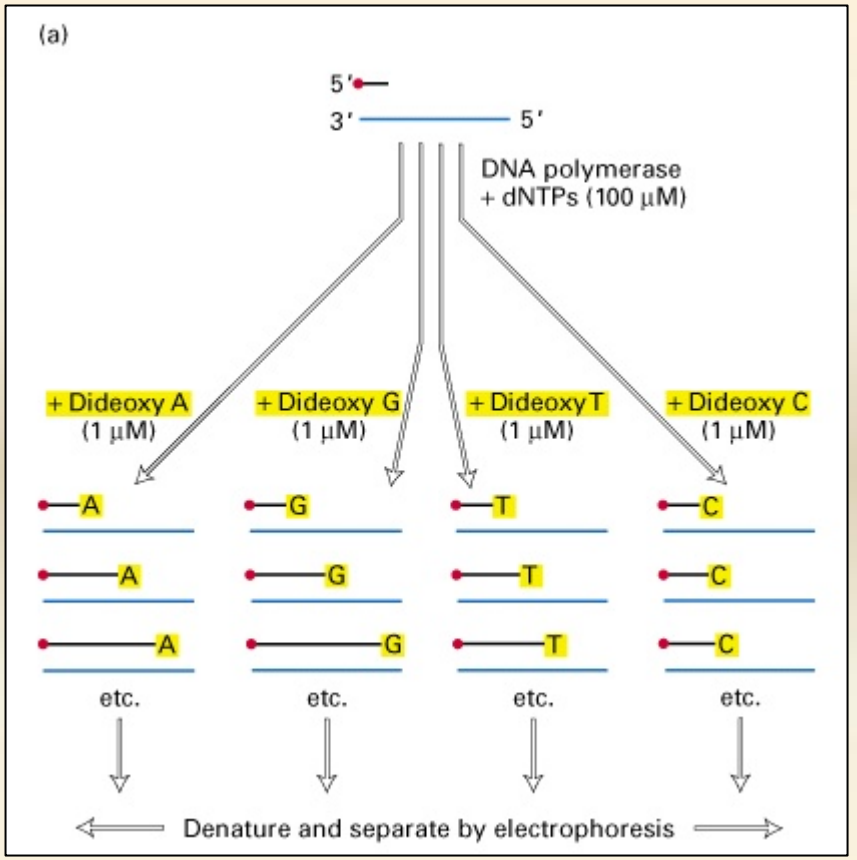
- Gel electrophoresis – determine length of DNA fragments

The length of DNA molecules is *decreasing* – smaller molecules run faster in a porous gel



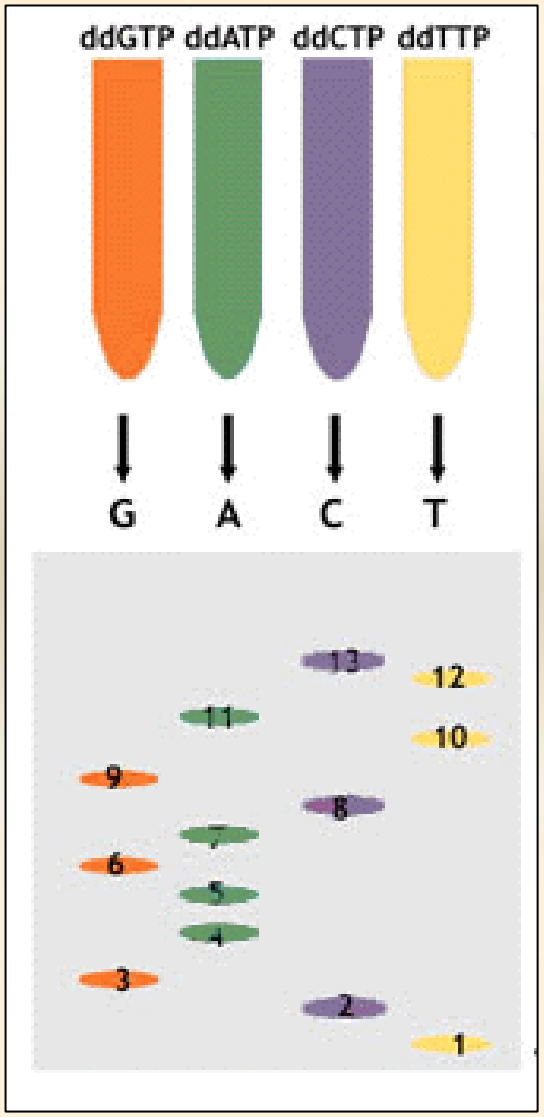
Sequencing

- Enzymic chain termination method
  - 4 different reaction tubes
  - Primer – sequence complementary to the start of the sequenced DNA
  - Mix of A,C,G,T radioactively labeled nucleotides
  - Small amount of dideoxynucleotides – when incorporated, no further chain growth



Sequencing

- The resulting DNAs from 4 tubes are loaded into 4 adjacent lines of the gel
- We can read the sequence from gel
- The sequence is read bottom up – from shorter to longer
- Process is automated
- Only up to 1000 nucleotides can be sequenced at a time
- We can determine sequence of all cDNAs in a cloning library



Sequencing genomes

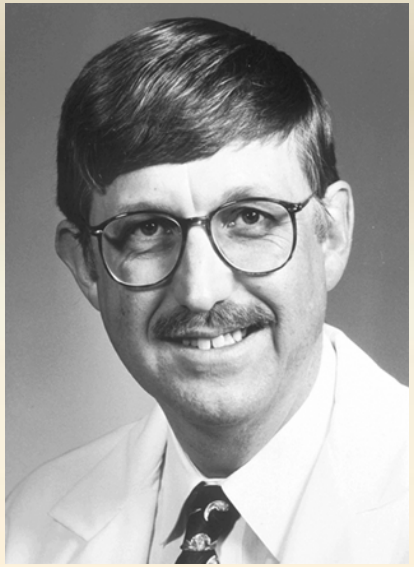
- 1985 – proposal to sequence entire Human genome. Financed by US Department of Energy (DOE), lead by Watson, at first, then by Francis Collins
  - "The fear is not *big* science so much as *bad* science," said Botstein, “the DOE's proposal is a scheme for unemployed bombmakers."
- First, model organisms were sequenced
  - E. coli (bacteria)
  - Drosophila (fruit fly)
  - C. elegans (round worm)

Human Genome Project – 1986-2003

- The scientific value seemed dubious. Although many biologists agreed that maps of the chromosomes would be useful for finding genes, what good would come from deciphering every A, T, G, and C, especially since most of them were "junk" that did not code for genes.
- Read more at:

Controversial From the Start

Why sequence junk?



Francis Collins

Craig Venter

Human Genome Project

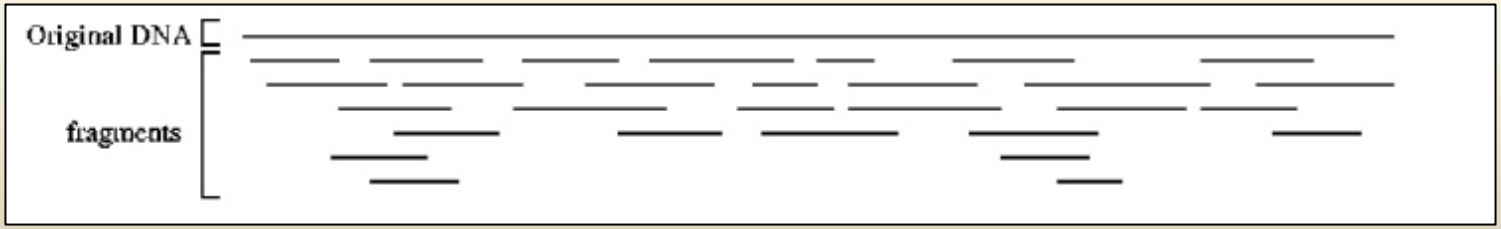
- 1985-the project initiated by Charles DeLisi, head of the department of energy (DoE) in the USA
- 1990-launched with the intention to be completed within 15 years and with a 3 billion dollar budget
- 1996-”Bermuda principles” – formalized the release of sequence data into public databases
- 1998-Craig Venter forms *Celera* company and promises to finish sequencing in 3 year with an ambitious “whole genome shotgun” approach
- 1999-the public project responds to Venter’s challenge and changes their target completion time
- December 1999-the first human chromosome sequence (22) published
- June 2000 – working draft announced
- February 2001 – the first draft published in nature and Science magazines

The Human Genome Sequence

- $3*10^9$ basepairs (30 times larger than fruit fly and round worm – both around 10^8 basepairs), 250 times larger than Yeast genome
- Coding regions not more than 3%
- Around 46% of the remaining DNA – repeating sequences
- The rest contains promoters and other regulatory sequences

Computational problem 1 – genome sequence assembly

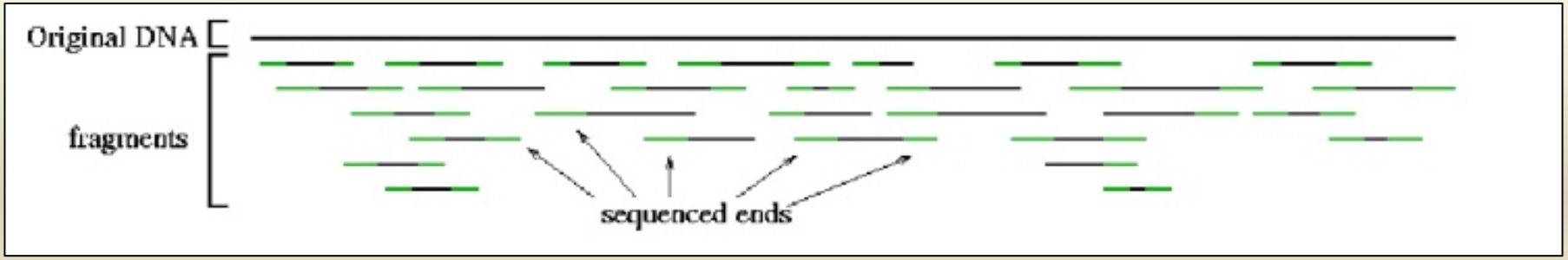
- Whole genome shotgun sequencing



Original DNA is broken into a collection of fragments

Computational problem 1 – genome sequence assembly

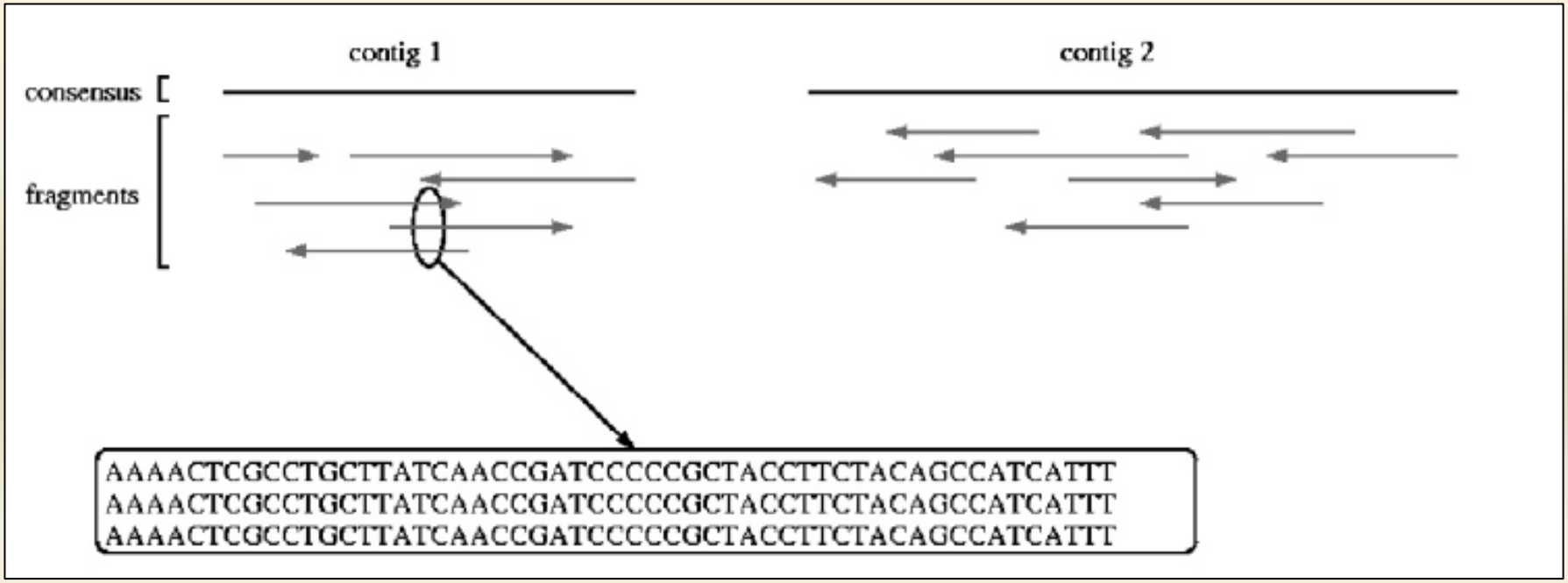
- Whole genome shotgun sequencing



The ends of each fragment (drawn in green) are sequenced

Computational problem 1 – genome sequence assembly

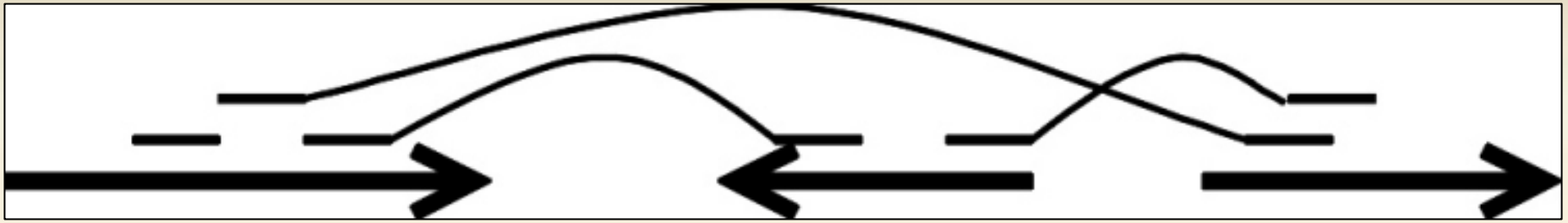
- Whole genome shotgun sequencing



The sequence reads are assembled together based on sequence similarity. The overlapping substrings are called *contigs*

Additional constrains

- Within the assembly the paired end reads must be placed at a distance consistent with the size of the library from which they originate and must be oriented towards each other.
- The constraints provided by mate pairs lead to constraints on the relative order and orientation of the contigs.

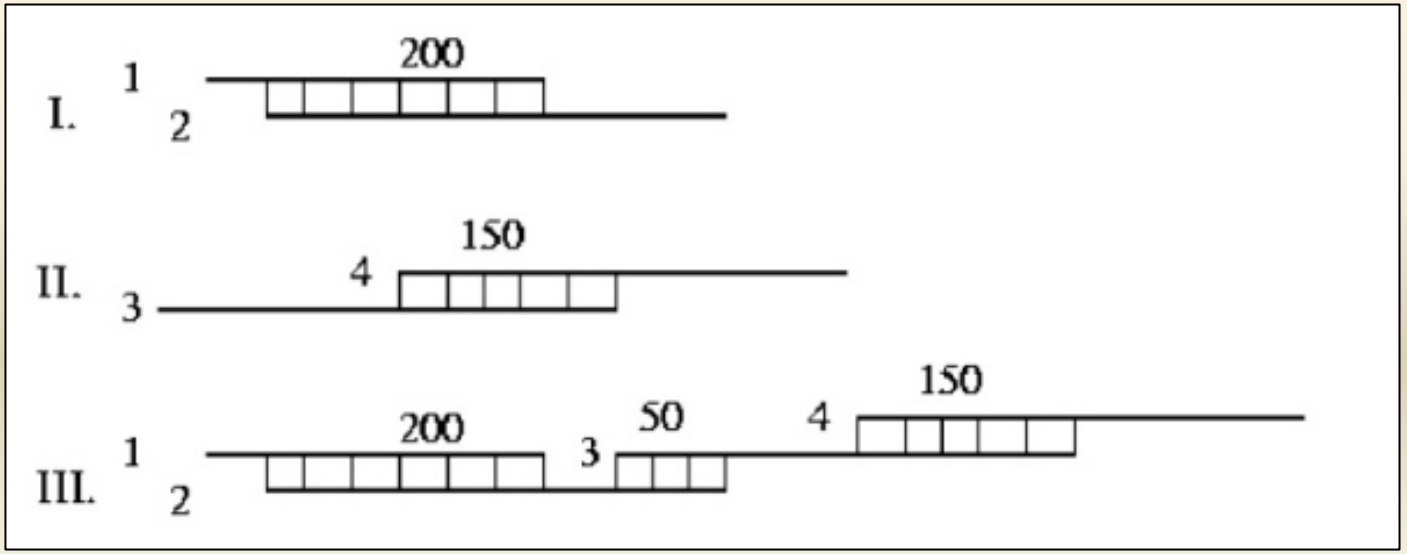


Assembly challenges

- Non-random fragments – not all pieces can be grown in *E. coli*, since their products are toxic to bacteria
- Repeats – lead to incorrectly computed overlaps

Assembly algorithms

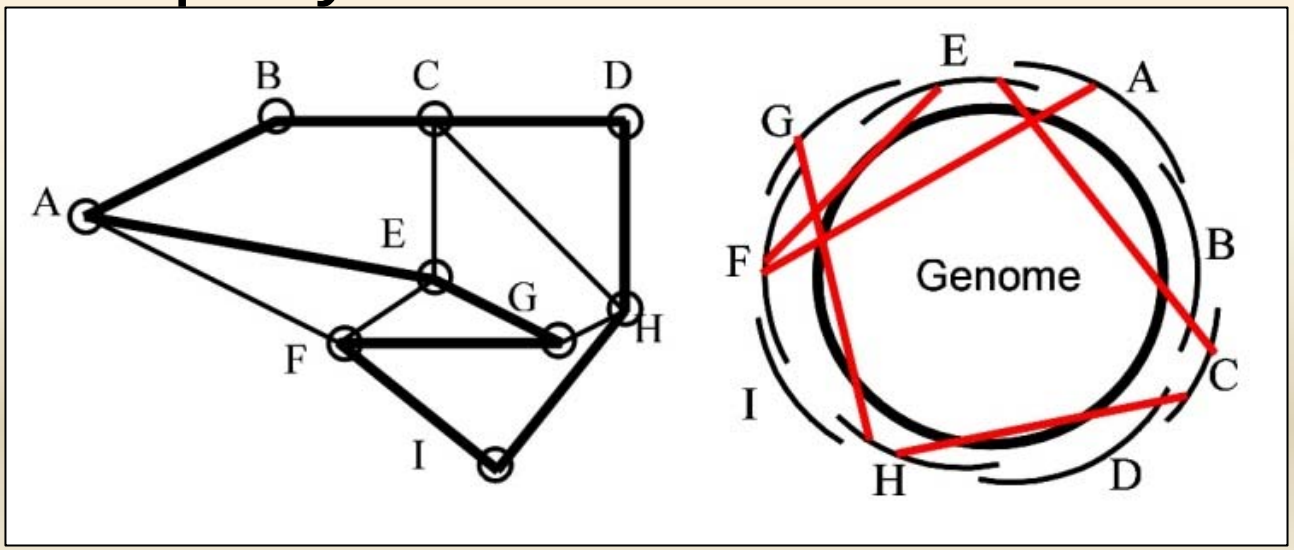
- Greedy assemblers



The assembler greedily joins together the reads that are most similar to each other.

Assembly algorithms

- Overlap-layout consensus

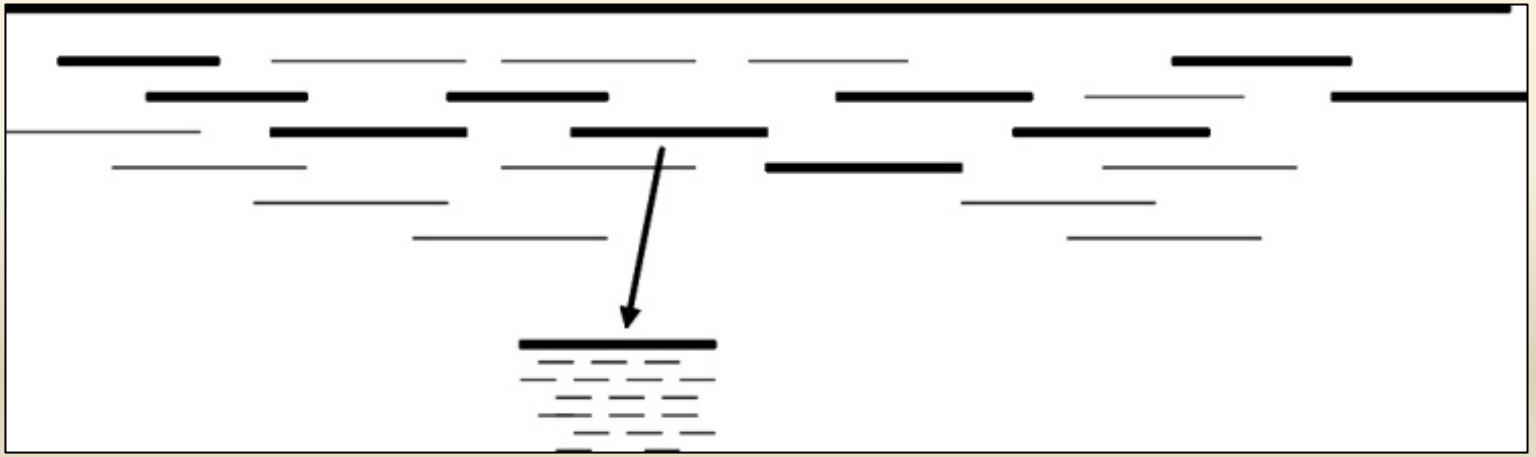


The relationships between the reads can be represented as a graph, where the nodes represent the reads and an edge connects two nodes if the corresponding reads overlap.

The problem of identifying a path through the graph that contains all the nodes - a **Hamiltonian path**

Computational problem 1 – genome sequence assembly

- **BAC-by-BAC (hierarchical) sequencing**



150 MB multiple random cuts are inserted into BACs

A *minimal tiling path* of BACs is chosen such that each base in the genome is covered by at least one BAC, and the overlap between BACs is minimized.

Each BAC fragment is sequenced separately

Chromosomal location of each BAC sequence is known, fewer random pieces to assemble

The computational biology

- The *bioinformatics* is born
  - the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of large-scale biological data.

Historical Perspective

- … – 1900 Pre-Mendelian period
- 1900 – 1940 Pre-DNA period
- 1940 – 1990 DNA period
- 1990 – 2003 Genomic period
- 2003 – … Post-genomic era

Future

- Systems biology
  - Complete set of all molecules of an organism
  - Complete set of interactions between these parts
  - Modeling of life
- Synthetic biology
  - *Mycoplasma laboratorium* is a minimal genome organism obtained by removal 100 genes from 482 genes of the smallest organism grown in culture, *M.*
- Evolution

Practical examples

- Gene therapy with no side effects
- Synthetic biology – engineering new products
  - Since natural biological systems are so complicated, we would be better off re-building the natural systems that we care about, from the ground up, in order to provide engineered surrogates that are easier to understand and interact with.
  - Biofuel in a minimal genome – Mycoplasma laboratorium
- Medicine and agriculture

The post-genomic era

- Let me now comment on the question "what next". Up to now we are working on the descriptive phase of molecular biology. ... But the real challenge will start when we enter the synthetic biology phase of research in our field. We will then devise new control elements and add these new modules to the existing genomes or build up wholly new genomes. This would be a field with the unlimited expansion potential and hardly any limitations to building "new better control circuits" and ..... finally other "synthetic" organisms, like a "new better mouse". ... I am not concerned that we will run out exciting and novel ideas…Waclaw Szybalski

Perspectives

- Computational tools instead of a microscope
- Very long period …