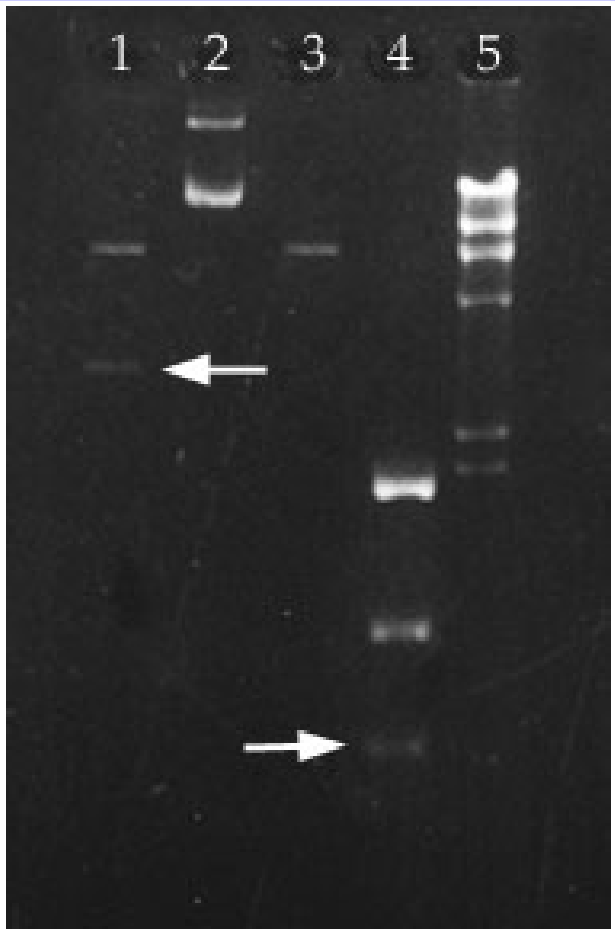
Gel Electrophoresis

- Used to measure the lengths of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

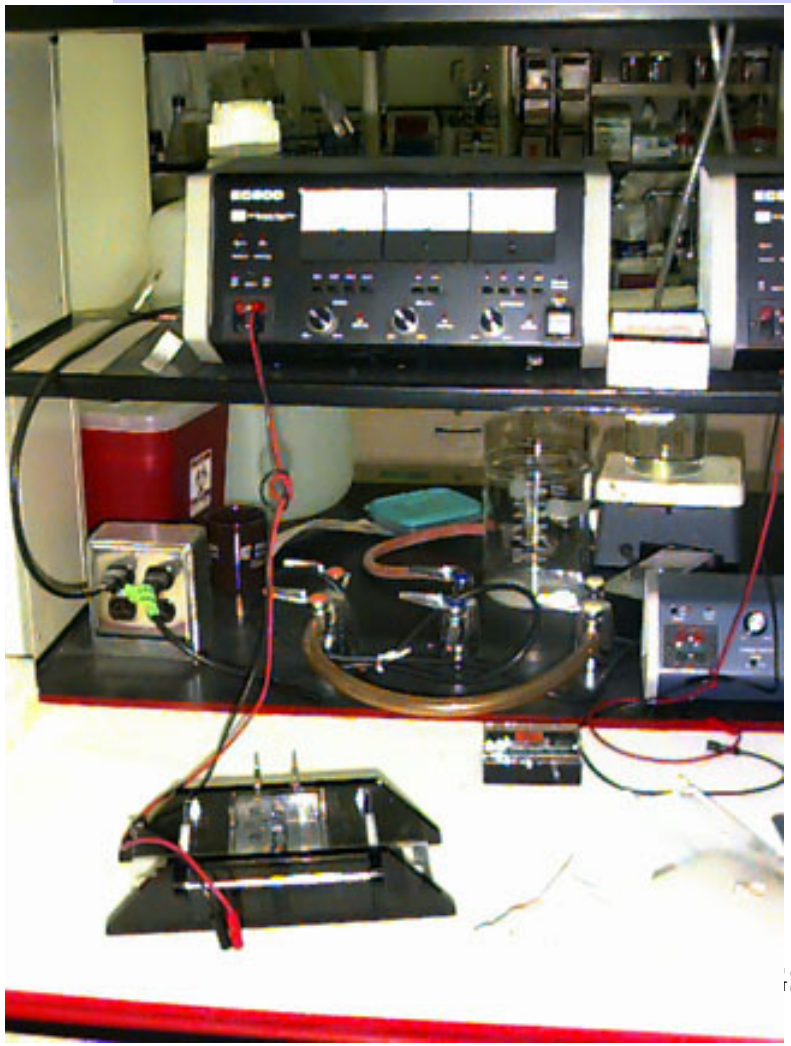
Gel Pictures

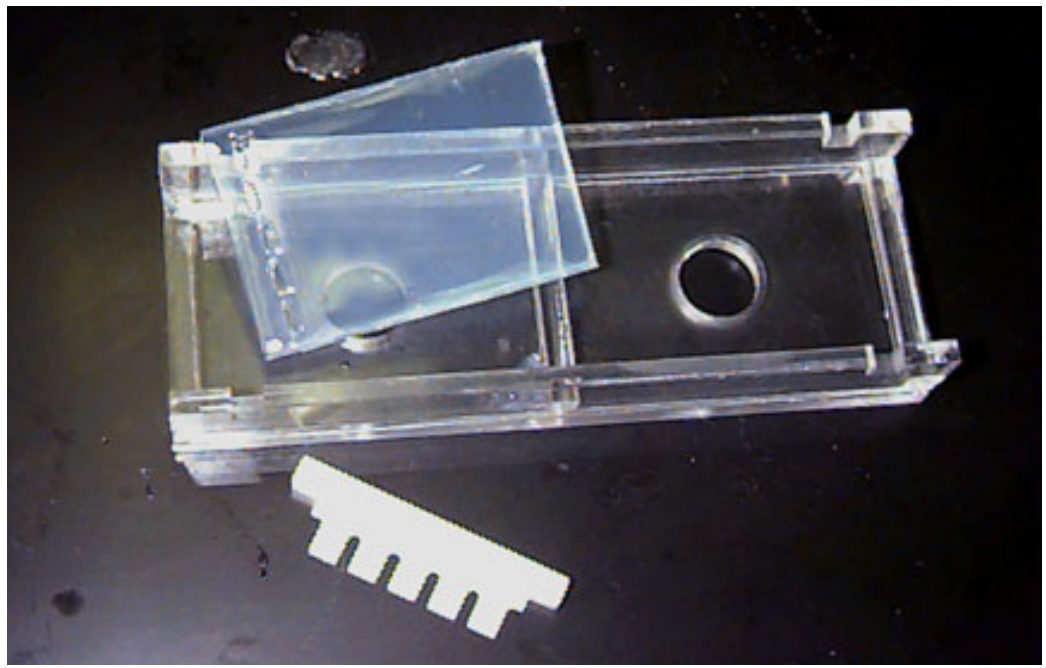


Gel Electrophoresis: Measure sizes of fragments

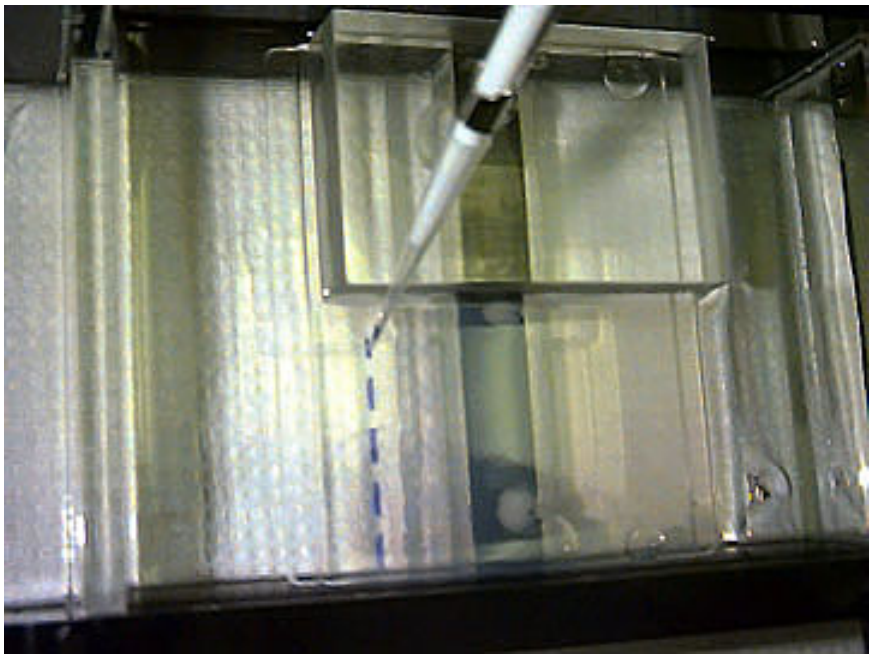
- The phosphate backbone makes DNA a highly negatively charged molecule. Thus DNA can be fractionated according to its size.
- Gel: allow hot 1 % solution of purifed agarose to cool and solidify/polymerize (like Jello).
- DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.
- Varying concentration of agarose makes different pore sizes & results.

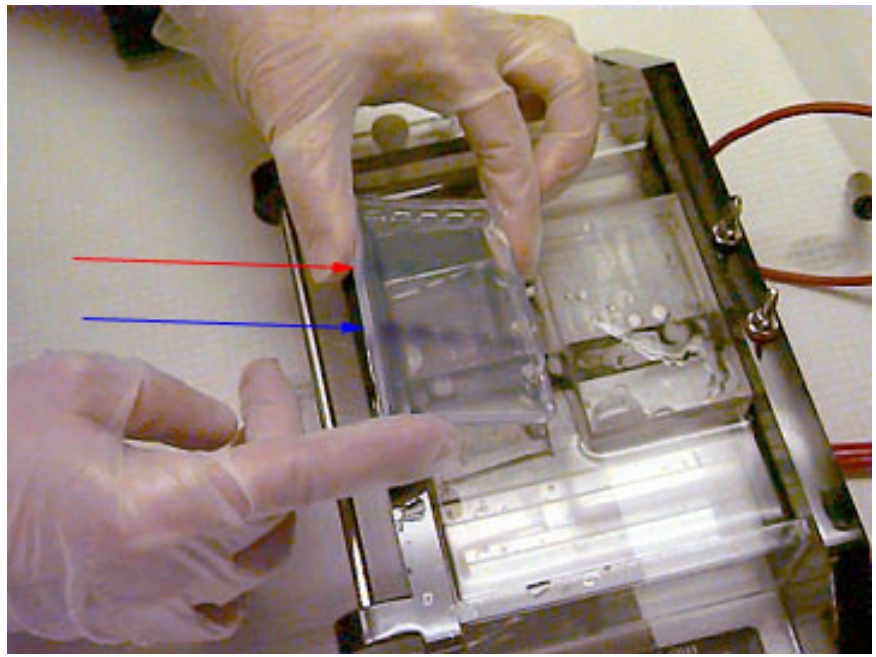
Gel Electrophoresis

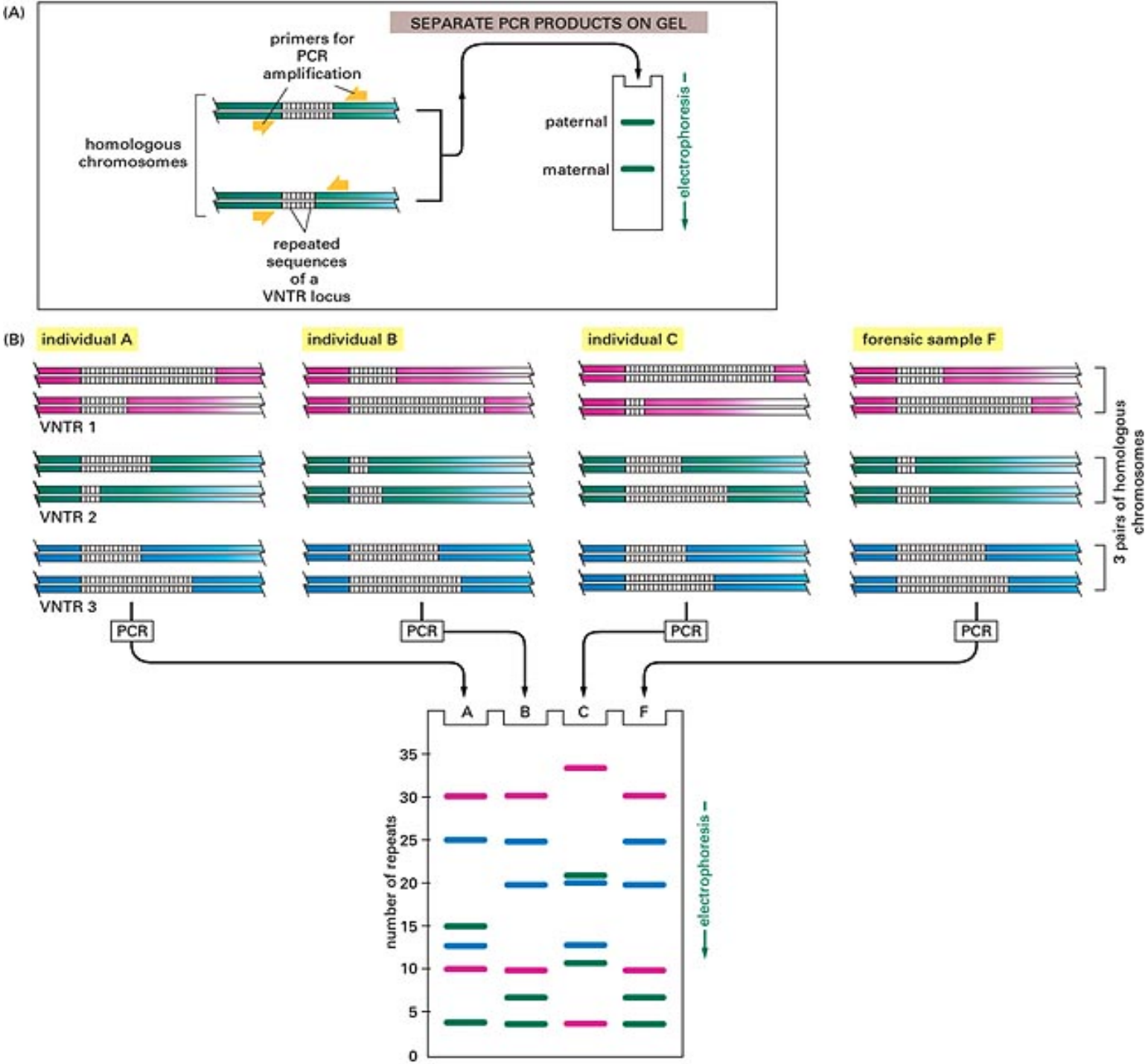




Gel Electrophoresis



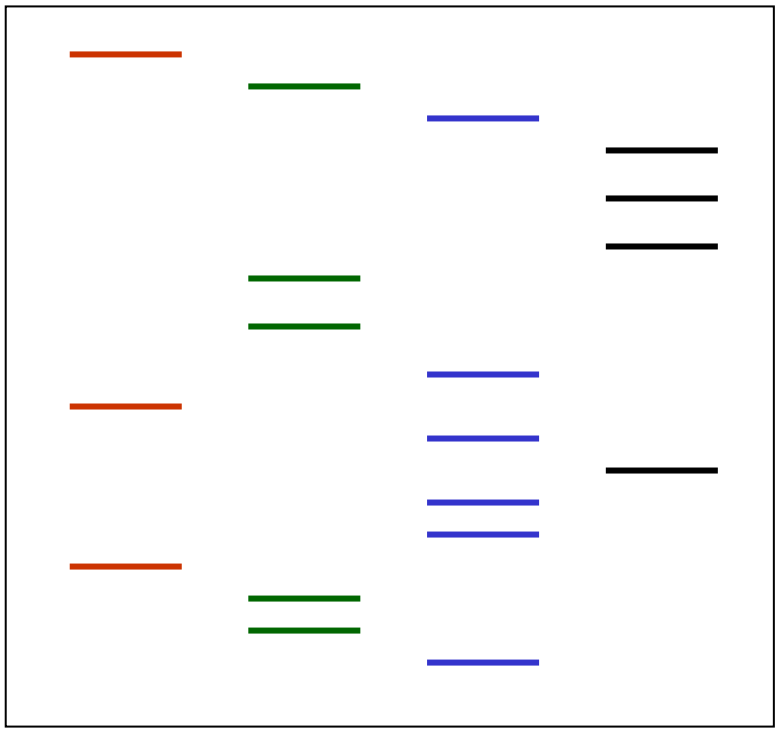



(A)
SEPARATE PCR PRODUCTS ON GEL
primers for PCR amplification
homologous chromosomes
repeated sequences of a VNTR locus
paternal
maternal
electrophoresis
(B)
individual A
individual B
individual C
forensic sample F
VNTR 1
VNTR 2
VNTR 3
3 pairs of homologous chromosomes
PCR
A
B
C
F
number of repeats
35
30
25
20
15
10
5
0
electrophoresis

Sequencing a Fragment Using Gels

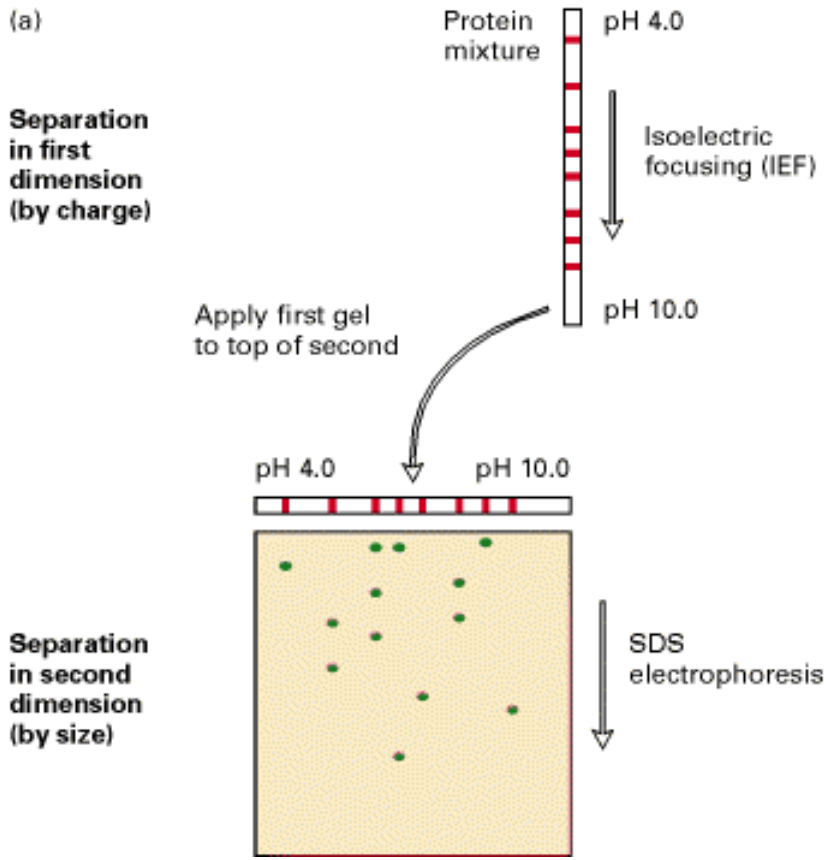
- Isolate the desired DNA fragment.
- Using the "starving method" obtain all fragments that end in A, C, G, T
- Run gel with 4 lanes and read the sequence

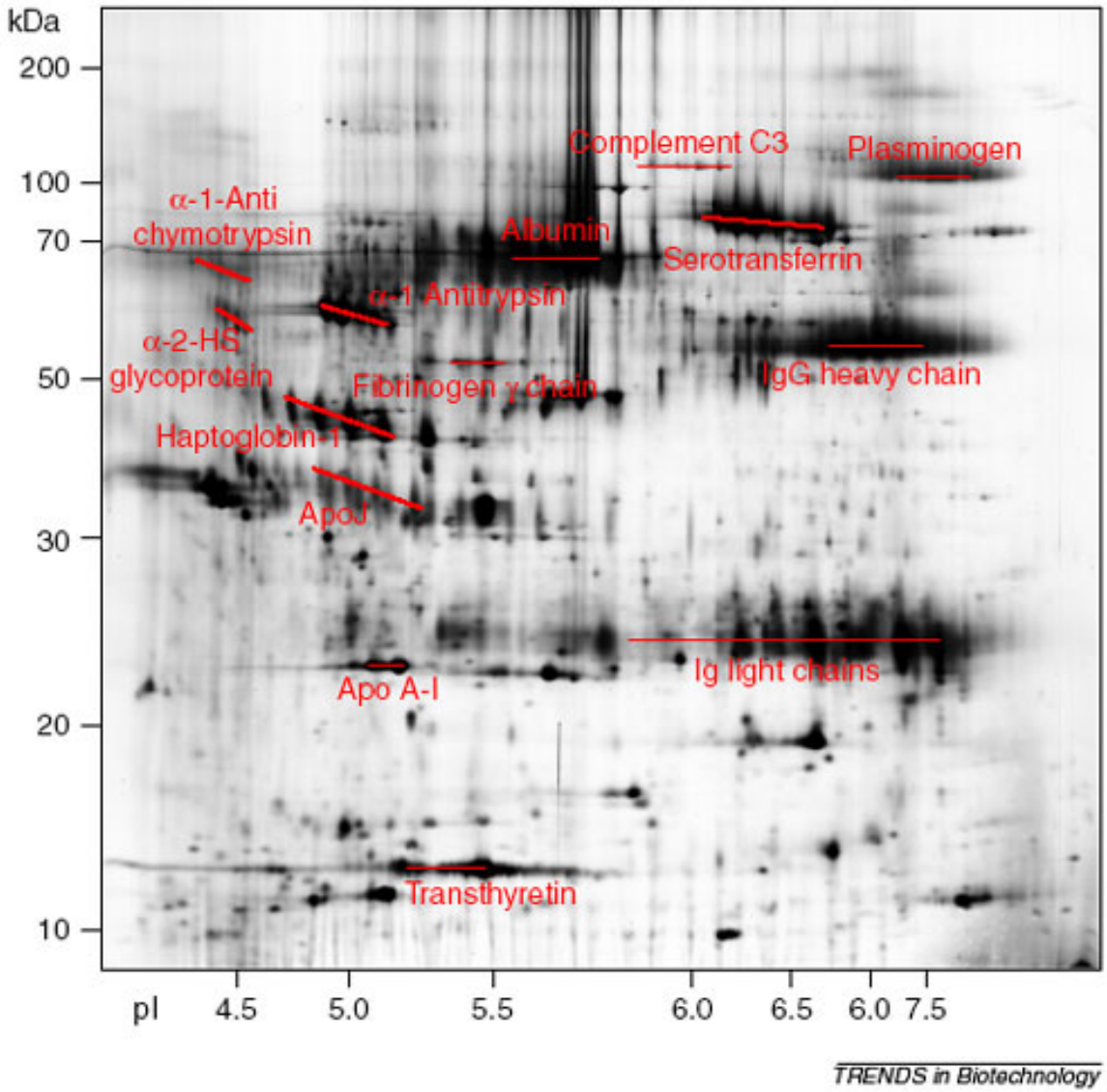
Application of Gels: Sequencing



GCCAGGTGAGCCTTTGCA

2D-Gels


(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis


kDa
200
100
70
50
30
20
10
pI
4.5
5.0
5.5
6.0
6.5
6.0
7.5
Complement C3
Plasminogen
α-1-Anti
chymotrypsin
Albumin
Serotransferrin
α-1 Antitrypsin
α-2-HS
glycoprotein
Fibrinogen γ chain
IgG heavy chain
Haptoglobin-1
ApoJ
Ig light chains
Apo A-I
Transthyretin
TRENDS in Biotechnology

2D-Gels

First Dimension Methodology of a 2D Gel:

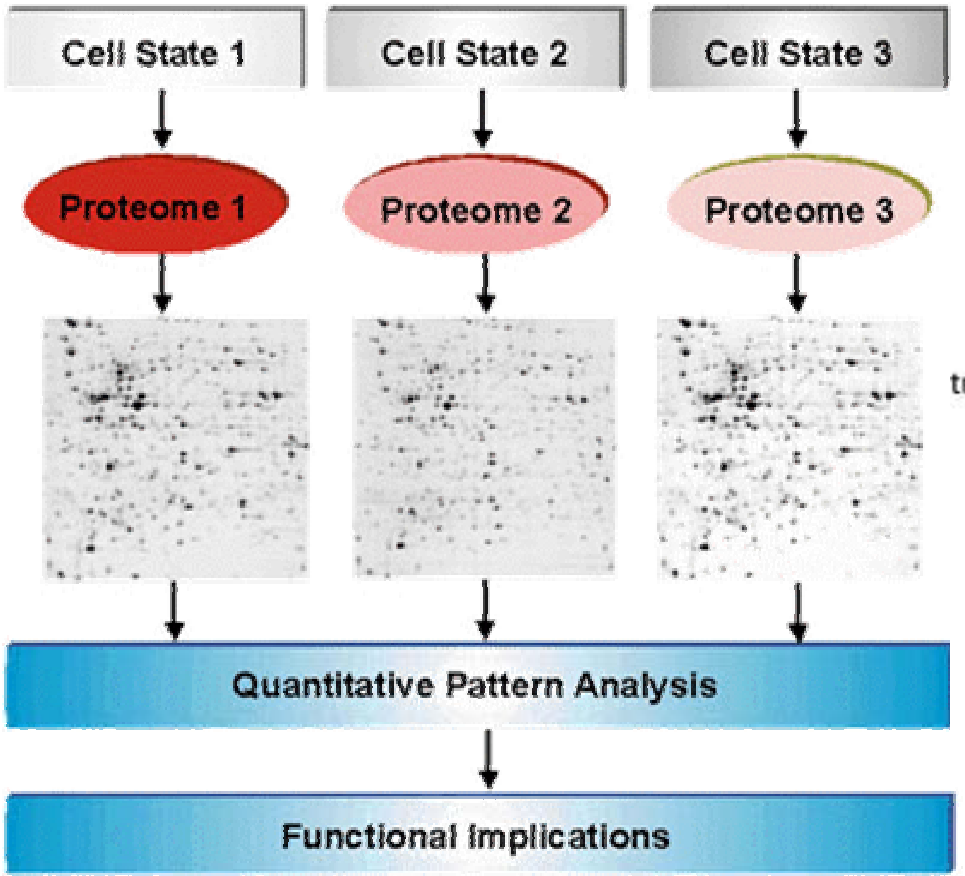
Denatured cell extract layered on a glass tube filled with polyacrylamide saturated with solution of ampholytes, a mixture of polyanionic[(-) charged] and polycationic [(+) charged] molecules. When placed in an electric field, the ampholytes separate and form continuous gradient based on net charge. Highly polyanionic ampholytes will collect at one end of tube, highly polycationic ampholytes will collect at other end. Gradient of ampholytes establishes pH gradient. Charged proteins migrate through gradient until they reach their pI, or isoelectric point, the pH at which the net charge of the protein is zero. This resolves proteins that differ by only one charge.

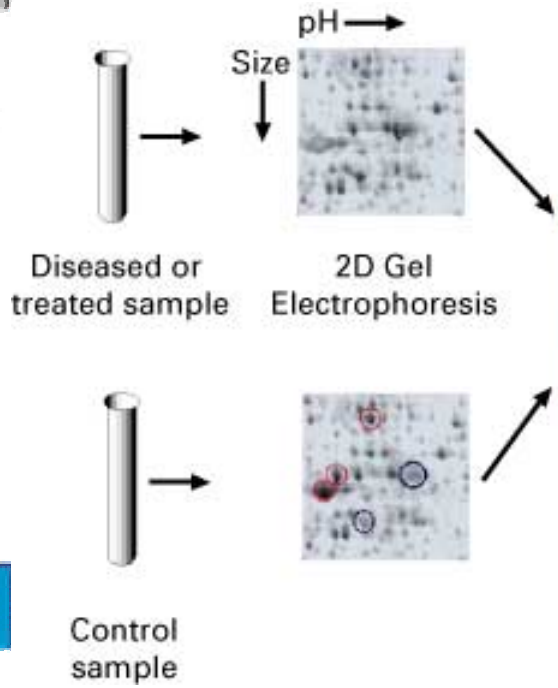
Entering the Second Dimension:

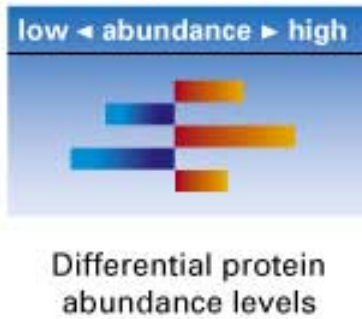
Proteins that were separated on IEF gel are next separated in the second dimension based on their molecular weights. The IEF gel is extruded from tube and placed lengthwise in alignment with second polyacrylamide gel slab saturated with SDS. When an electric field is imposed, the proteins migrate from IEF gel into SDS slab gel and then separate according to mass. Sequential resolution of proteins by their charge and mass can give excellent separation of cellular proteins. As many as 1000 proteins can be resolved simultaneously.

*Some information was taken from Lodish *et al.* Molecular Cell Biology.

2D-gels



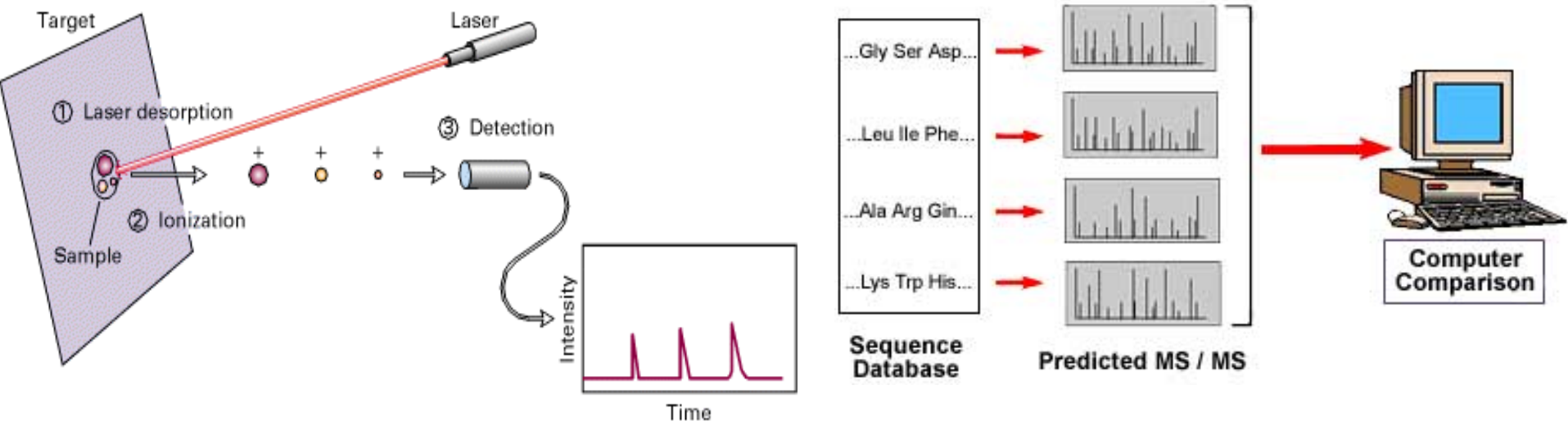




Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels

Mass Spectrometry

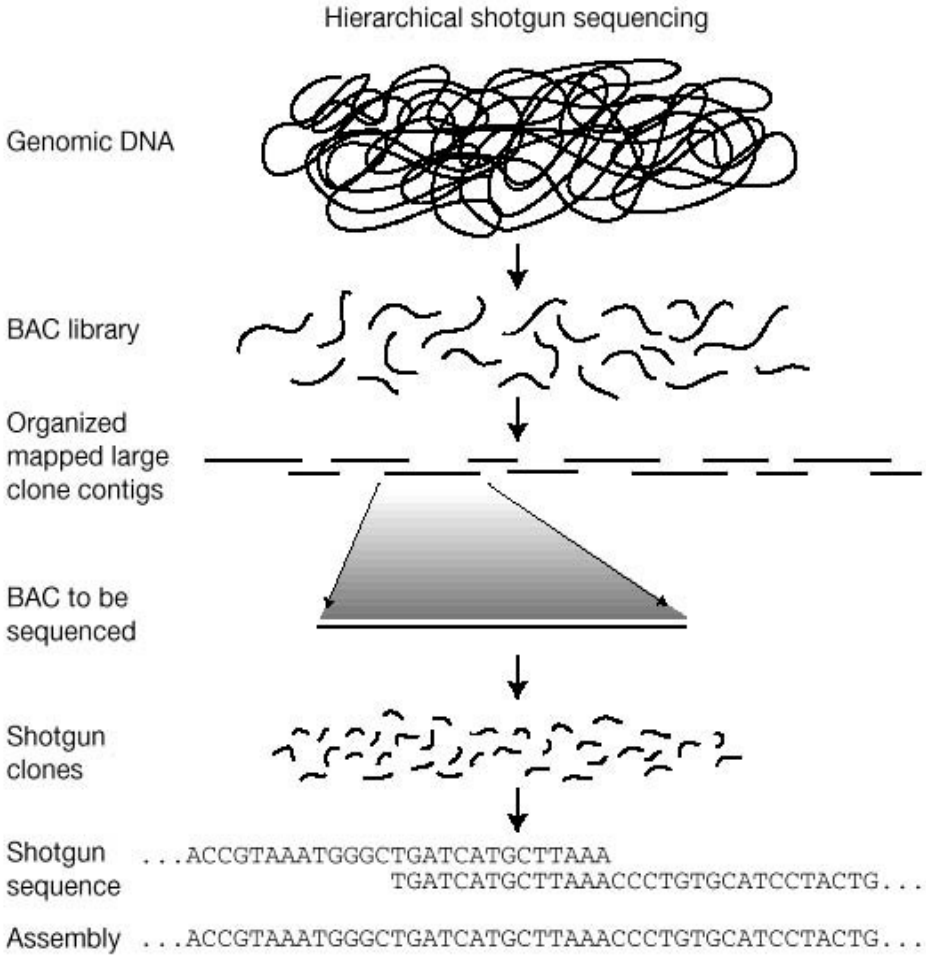


Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.
- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.

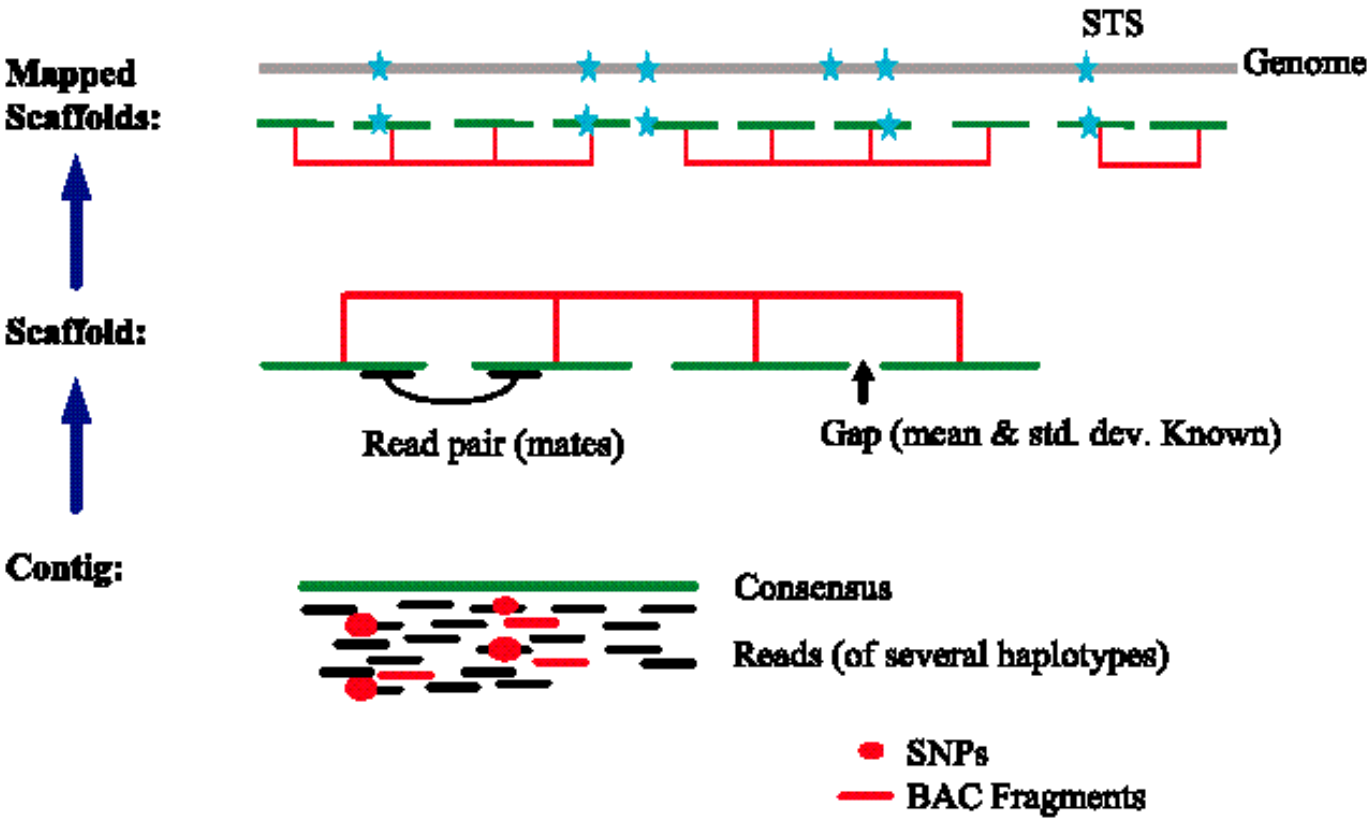
Sequencing

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
- Then run it on your PC.

Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-
  - **--ACCGT--**
  - **----CGTGC**
  - **TTAC-----**
  - **-TACCGT—**
  - **TTACCGTGC**

Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage

Assembly: Complications

w = AGTATTGGCAATC

z = AATCGATG

u = ATGCAAACCT

x = CCTTTTGG

y = TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
--------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

FIGURE 4.20

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

FIGURE 4.21

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.

Assembly: Complications

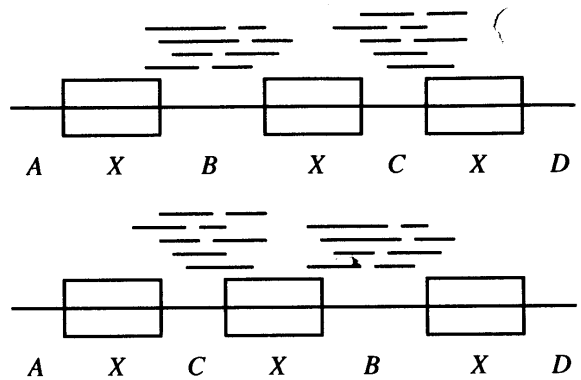


FIGURE 4.8

Target sequence leading to ambiguous assembly because of repeats of the form XXX.

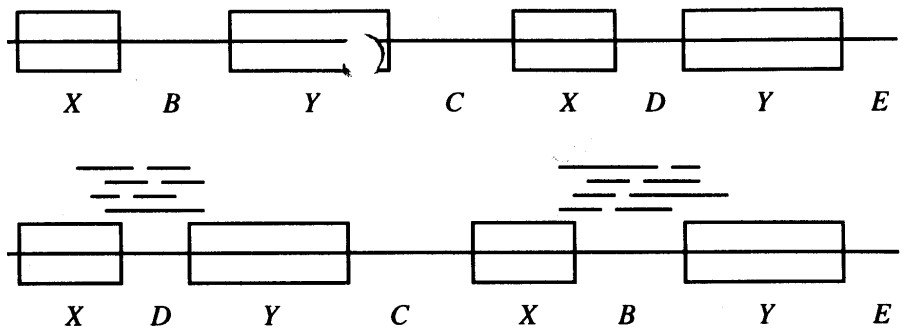


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

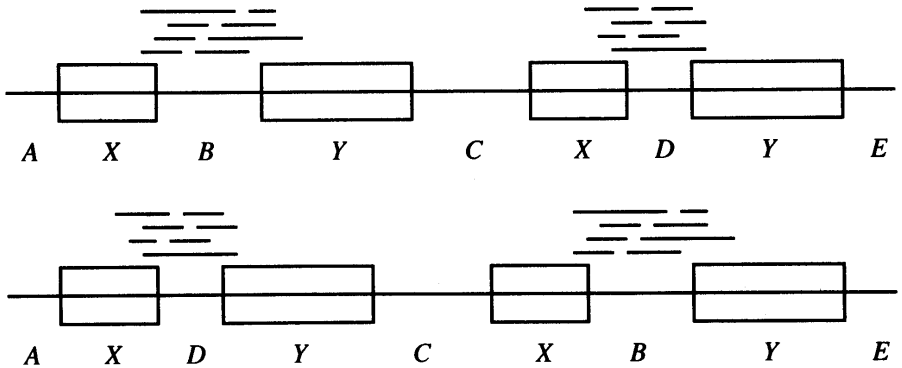


FIGURE 4.9

Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

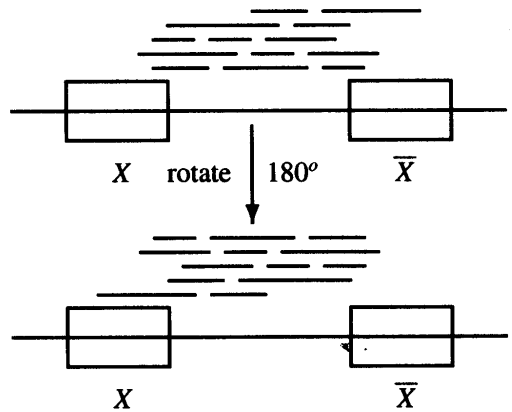


FIGURE 4.10

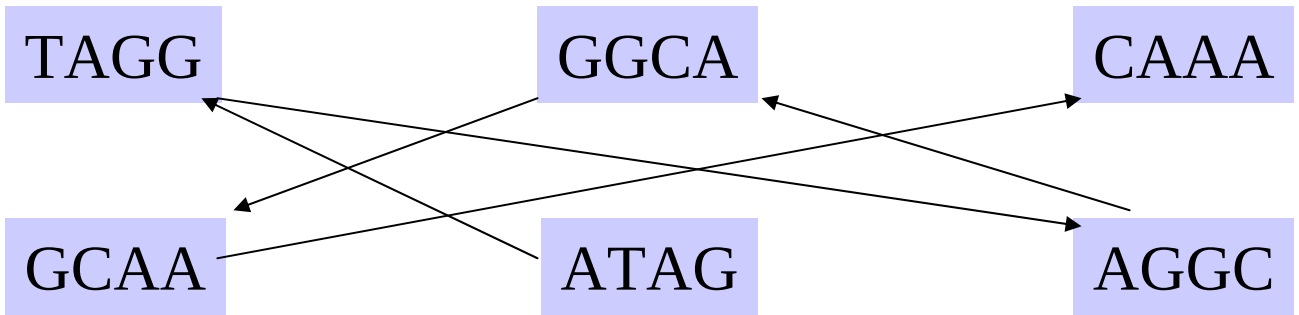
Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.

Miscellaneous

- Contig: A continuously covered region in the assembly.
- Other sequencing methods:
  - Sequencing by Hybridization (SBH)
  - Dual end sequencing
  - Chromosome Walking (see page 5-6 of Pevzner's text).

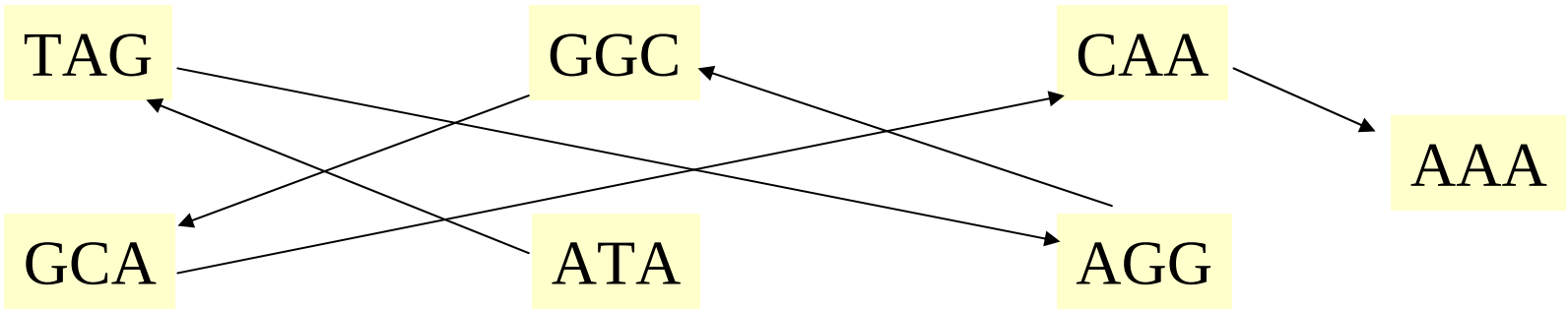
SBH

- Suppose that the only length 4 fragments that hybridize to S are: **TAGG, GGCA, CAAA, GCAA, ATAG, AGGC**. Then what is S, if it is of length ~9?



Hamiltonian Path Problem

SBH



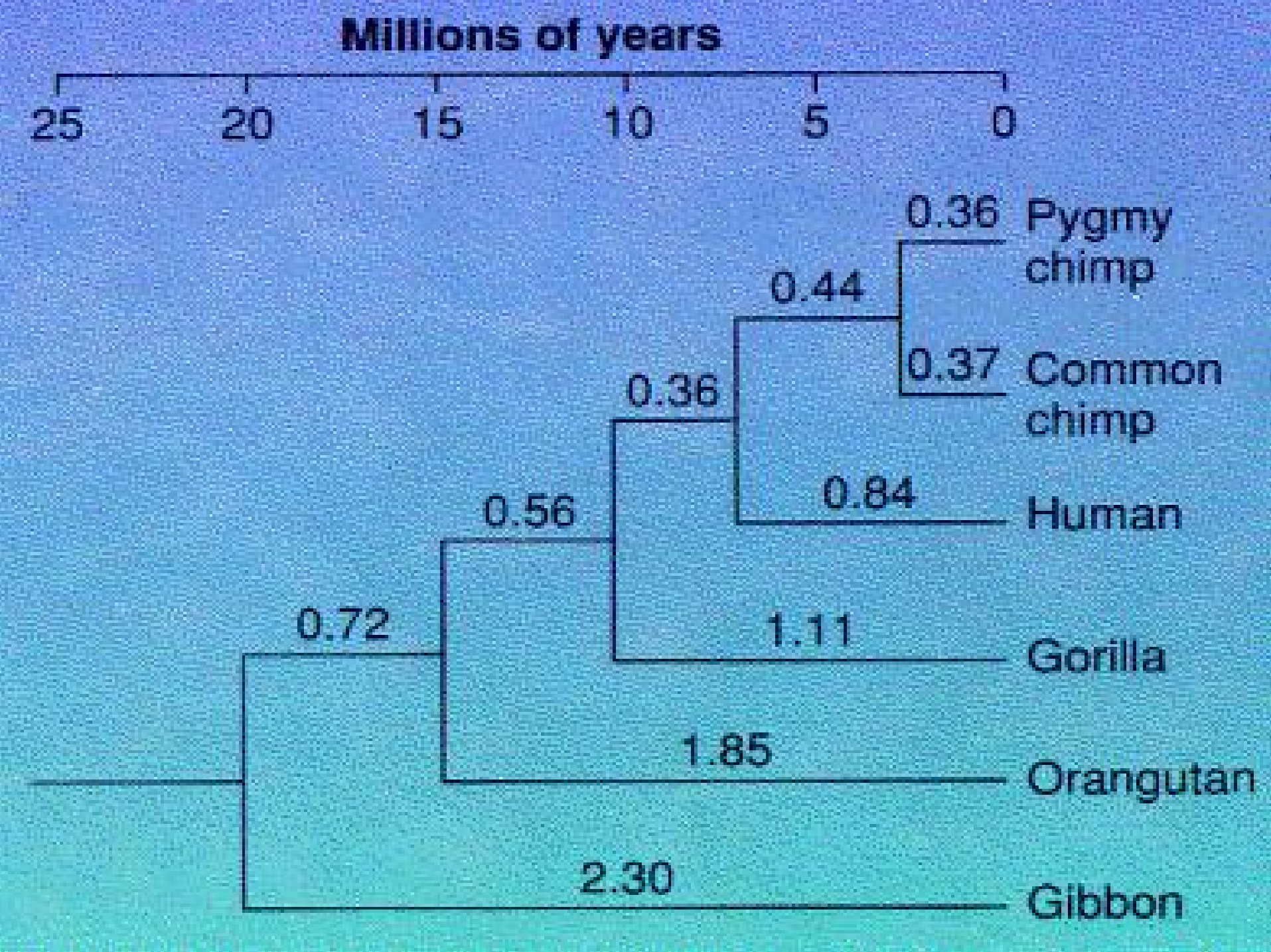
Eulerian Path Problem

Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]
- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).
- Phrap,(http://bozeman.genome.washington.edu/phrap.docs/phrap.html)
- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)
- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]
- GCG (http://www.gcg.com)
- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)
- SPACE, SPASS [Lawrence Berkeley Labs] (http://www-hgc.lbl.gov/inf/space.html)
- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)

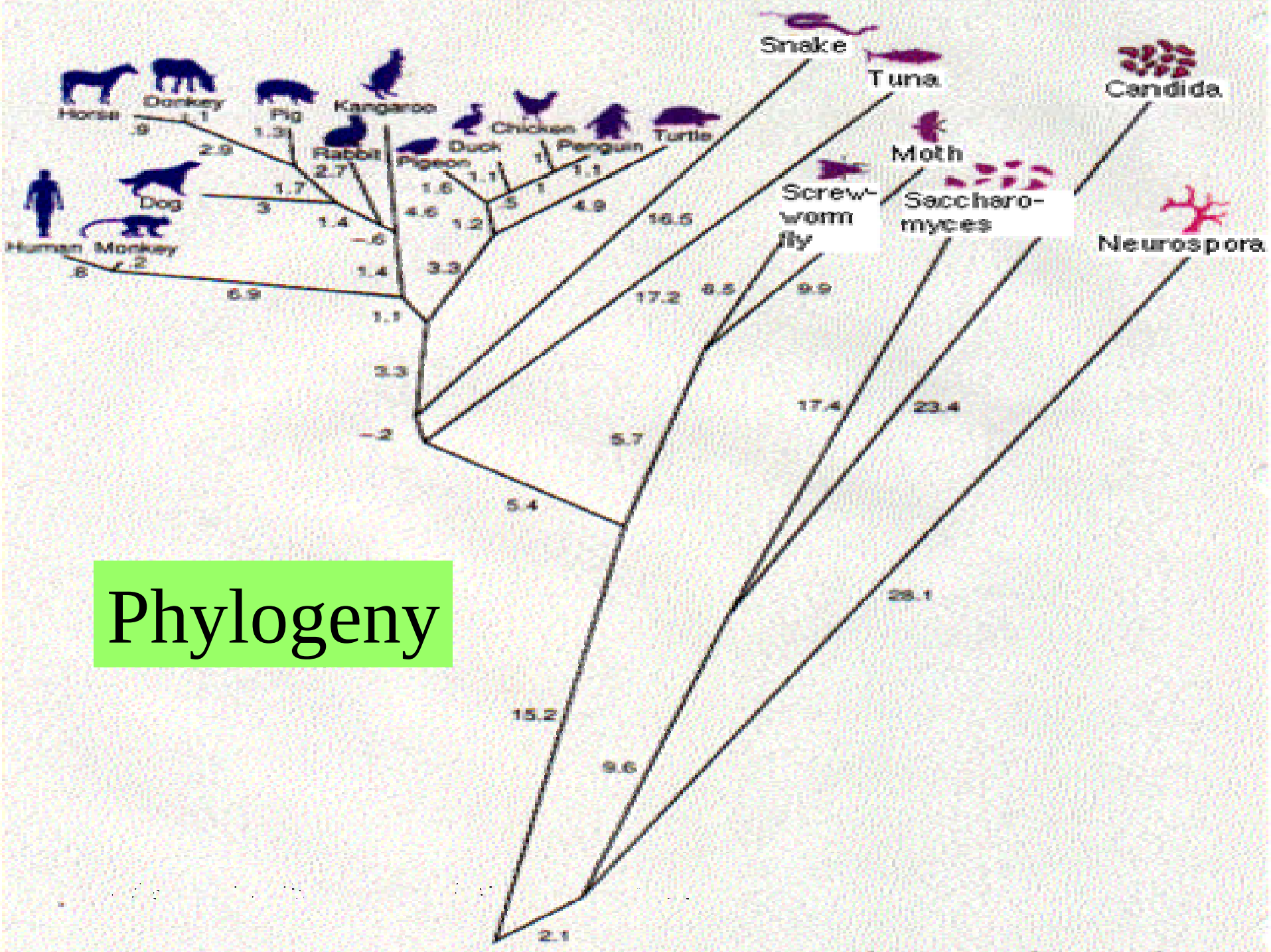
Theory of Evolution

- Charles Darwin
  - **1858-59:** *Origin of Species*
  - 5 year voyage of H.M.S. Beagle (1831-36)
  - Populations have variations.
  - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
  - Speciation arises by splitting of one population into subpopulations.
  - Gregor Mendel and his work (1856-63) on inheritance.


Millions of years
25
20
15
10
5
0
0.36
Pygmy chimp
0.44
0.37
Common chimp
0.36
0.84
Human
0.56
0.72
1.11
Gorilla
1.85
Orangutan
2.30
Gibbon

Dominant View of Evolution

- All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.
- Organization: **Directed Rooted Tree**; Existing species: Leaves; Common ancestor species (divergence event): Internal node; Length of an edge: Time.


Horse
Donkey
Pig
Kangaroo
Rabbit
Pigeon
Duck
Chicken
Penguin
Turtle
Snake
Tuna
Moth
Screw-worm fly
Saccharo-myces
Candida
Neurospora
Dog
Human
Monkey
.9
1.1
1.3
2.9
2.7
1.7
3
1.4
1.6
4.6
1.2
1.1
1
5
1
1.1
4.9
16.5
.6
1.4
3.3
6.9
.8
.2
1.1
3.3
17.2
8.5
9.9
17.4
23.4
5.7
5.4
28.1
15.2
9.6
2.1

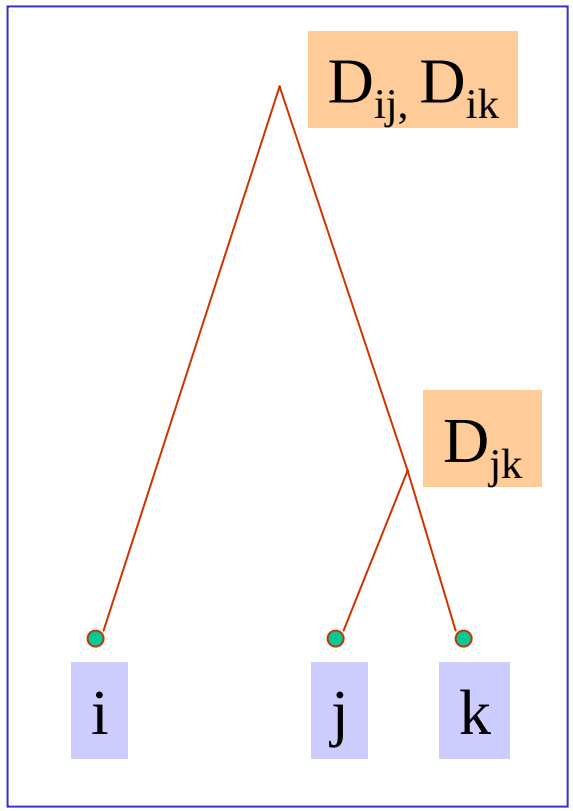
Phylogeny

Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Ultrametric

- An ultrametric tree:
  - decreasing internal node labels
  - distance between two nodes is label of least common ancestor.
- An ultrametric distance matrix:
  - Symmetric matrix such that for every i, j, k, there is tie for maximum of D(i,j), D(j,k), D(i,k)



Ultrametric: Assumptions

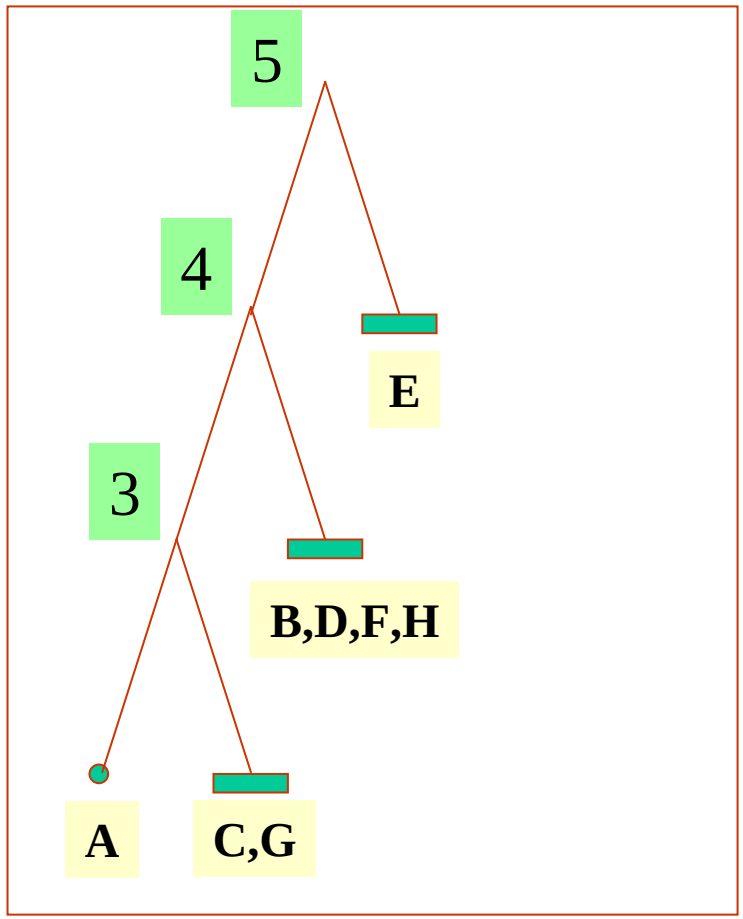
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

Ultrametric Data Sources

- Lab-based methods: hybridization
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: distance

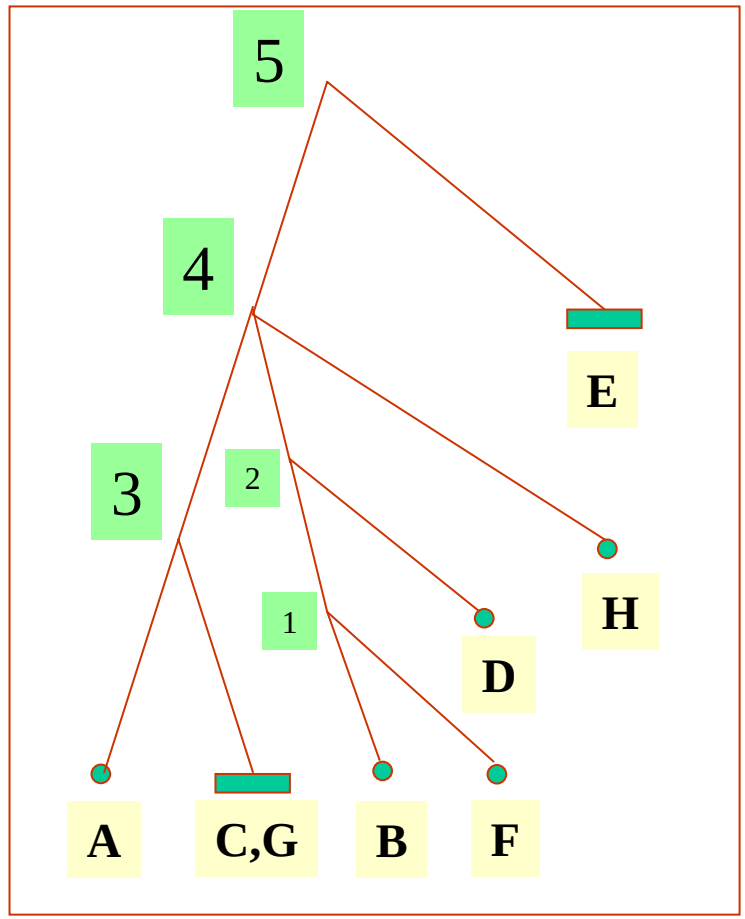
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B								
C								
D								
E								
F								
G								
H								



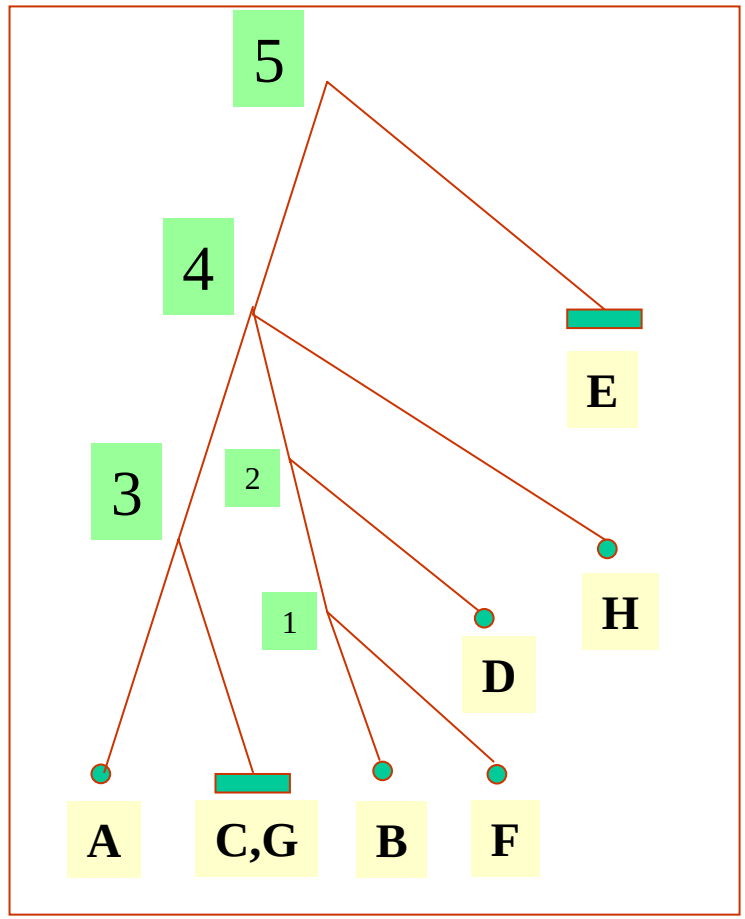
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C								
D								
E								
F								
G								
H								



Ultrametric: Distances Computed

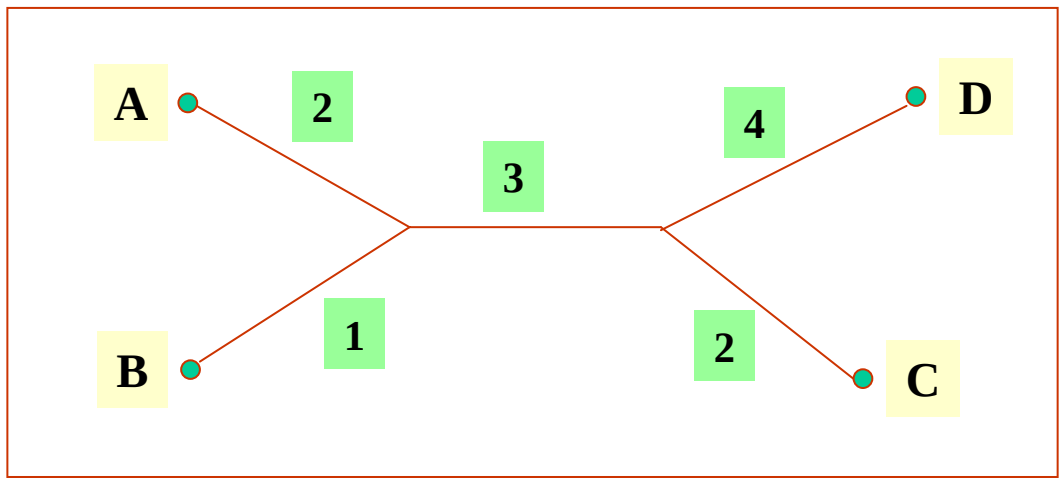
	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



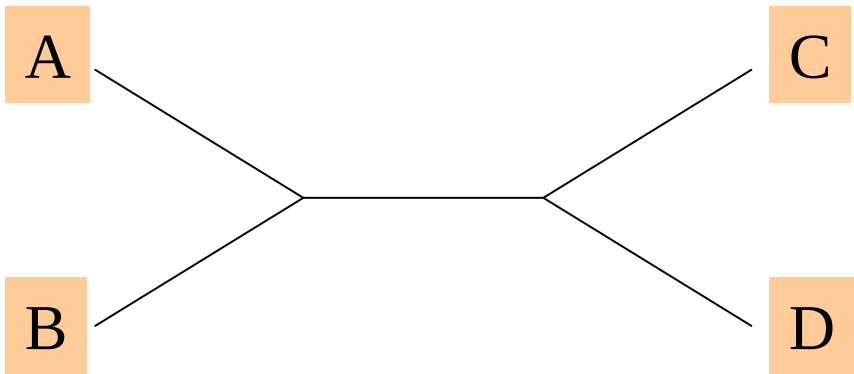
Additive-Distance Trees

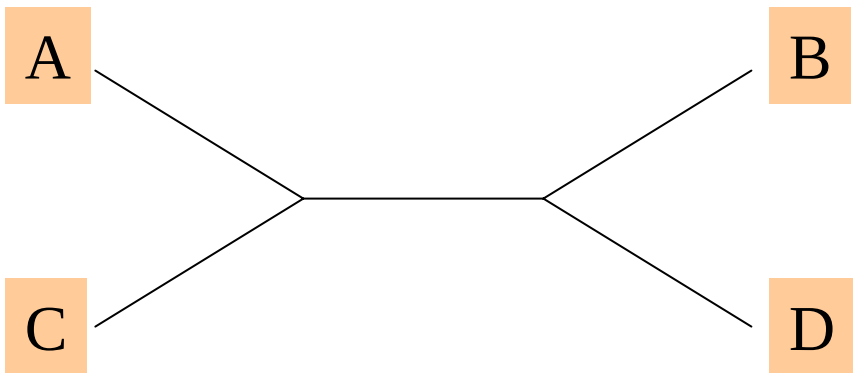
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

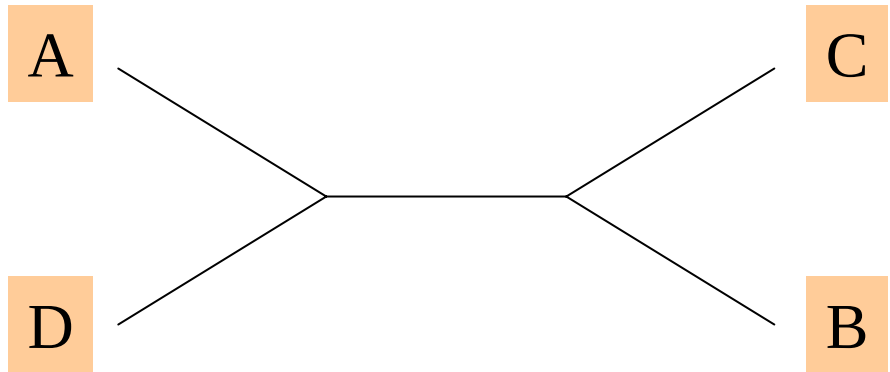
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

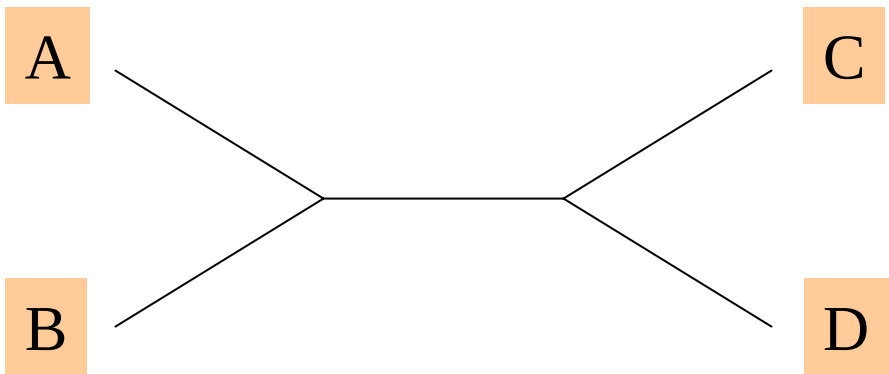






Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

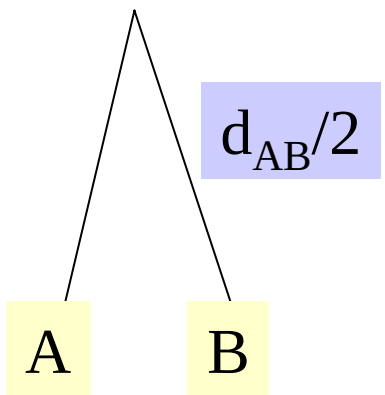


Unweighted pair-group method with arithmetic means (UPGMA)

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

	AB	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$$d_{(AB)C} = (d_{AC} + d_{BC}) /2$$



Transformed Distance Method

- UPGMA makes errors when rate constancy among lineages does not hold.
- Remedy: introduce an outgroup & make corrections

$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\sum_{k=1}^{n} D_{kO} \Big/ n \right)$$

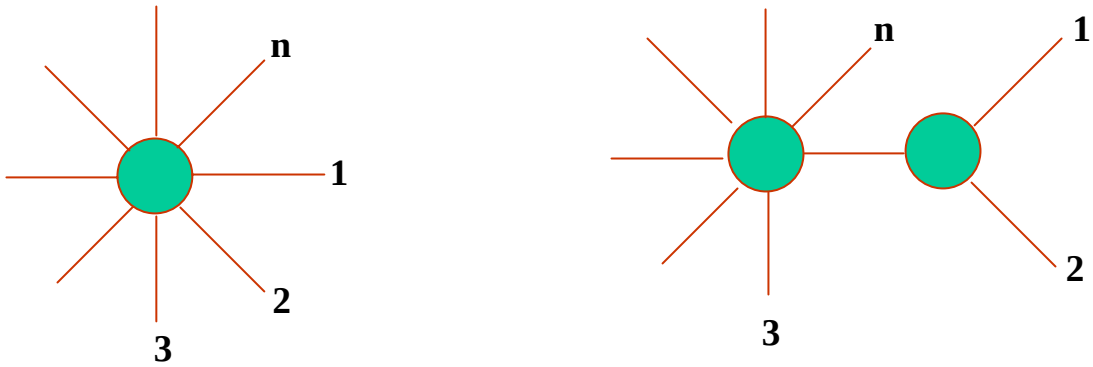
- Now apply UPGMA

Saitou & Nei: Neighbor-Joining Method

- Start with a star topology.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

Neighbor-Joining



$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

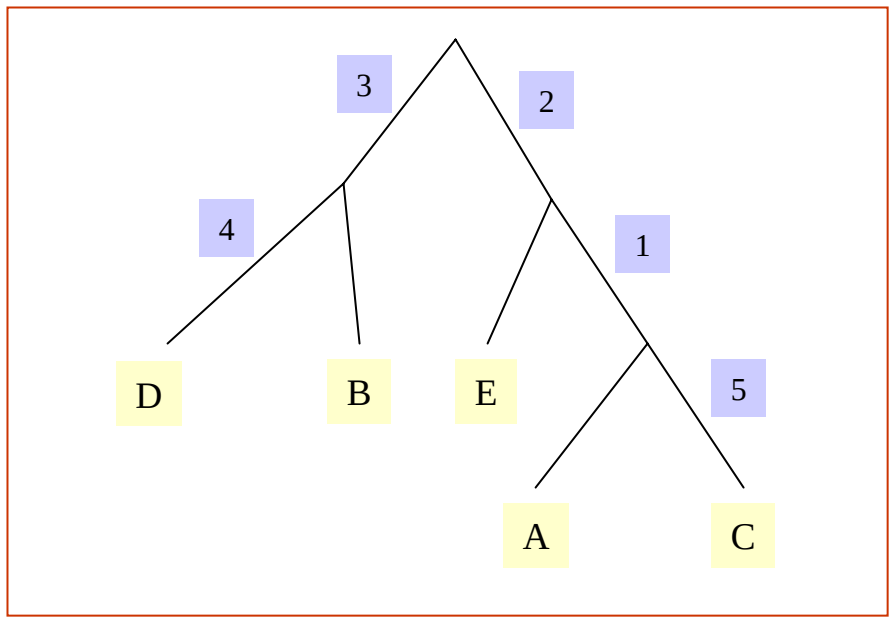
Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Character-based Methods

- Input: characters, morphological features, sequences, etc.
- Output: phylogenetic tree that provides the history of what features changed. [Perfect Phylogeny Problem]
- one leaf/object, 1 edge per character, path ⇔changed traits

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0



Example

- Perfect phylogeny does not always exist.

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

Maximum Parsimony

- Minimize the total number of mutations implied by the evolutionary history