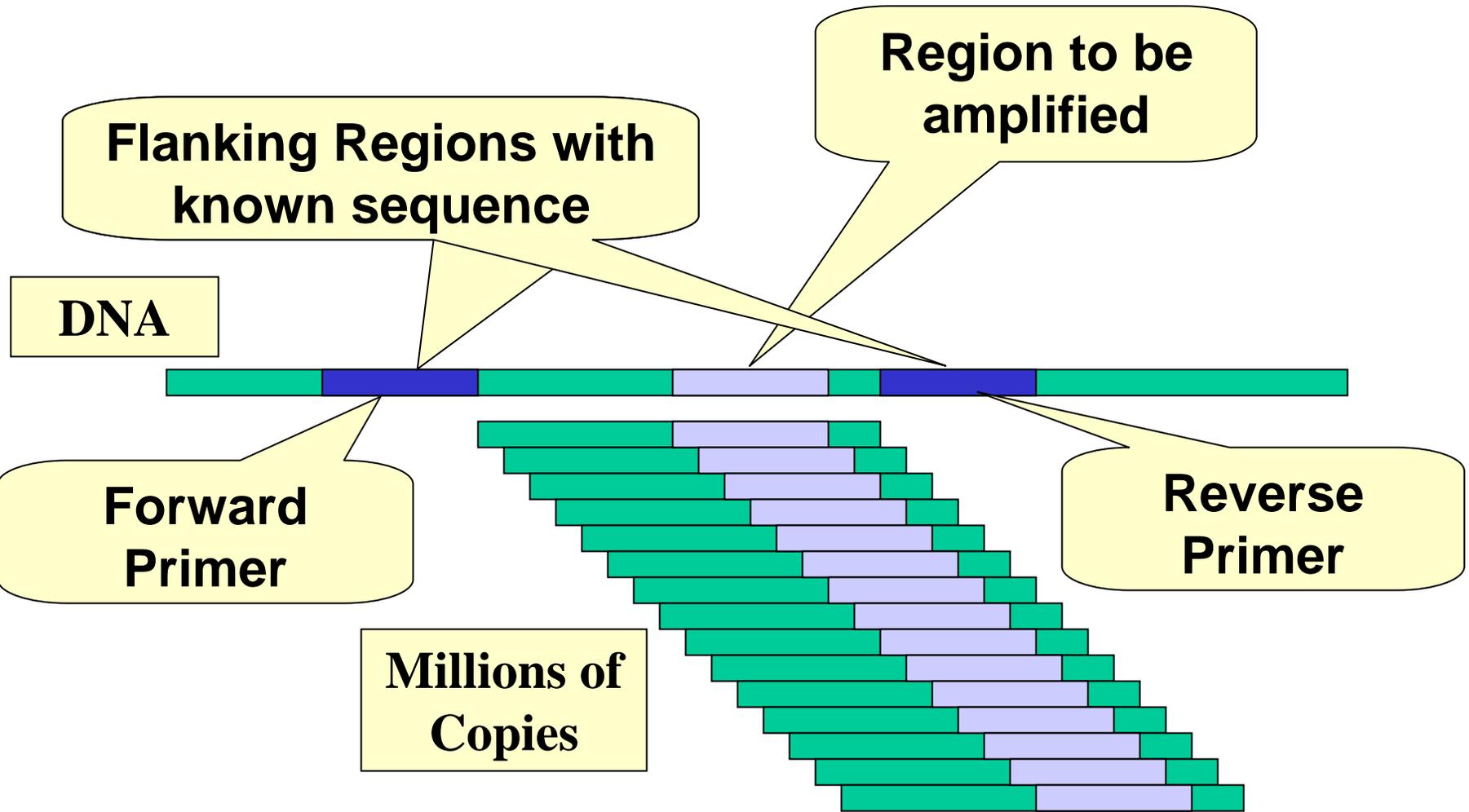


Polymerase Chain Reaction (PCR)

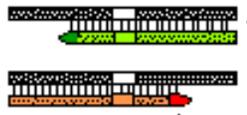
- For testing, large amount of DNA is needed
 - Identifying individuals for forensic purposes
 - (0.1 μL of saliva contains enough epithelial cells)
 - Identifying pathogens (viruses and/or bacteria)
- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences

PCR



POLYMERASE CHAIN REACTION

DNA region of interest.



primer

1. DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.



2. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

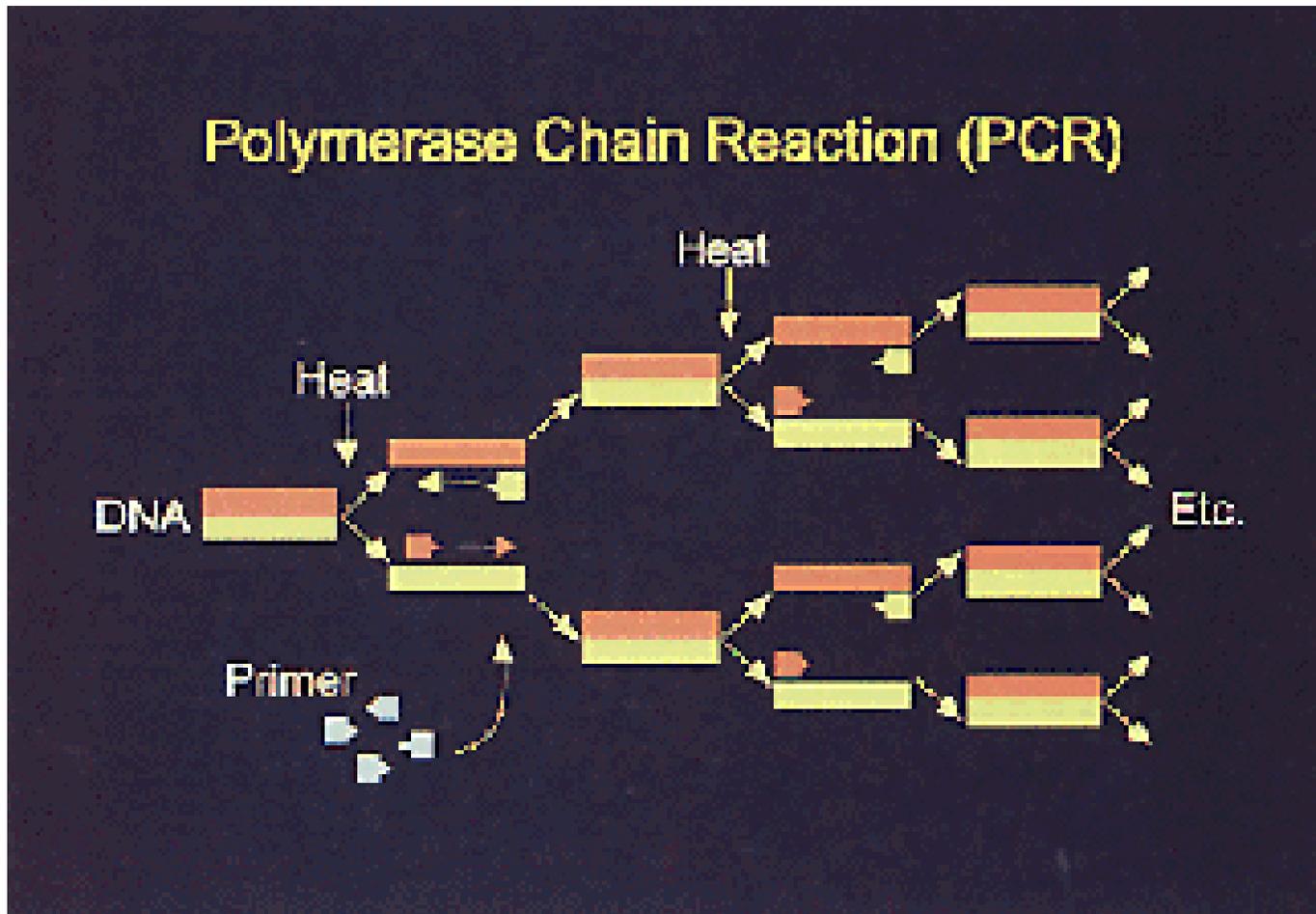
3. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

4. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

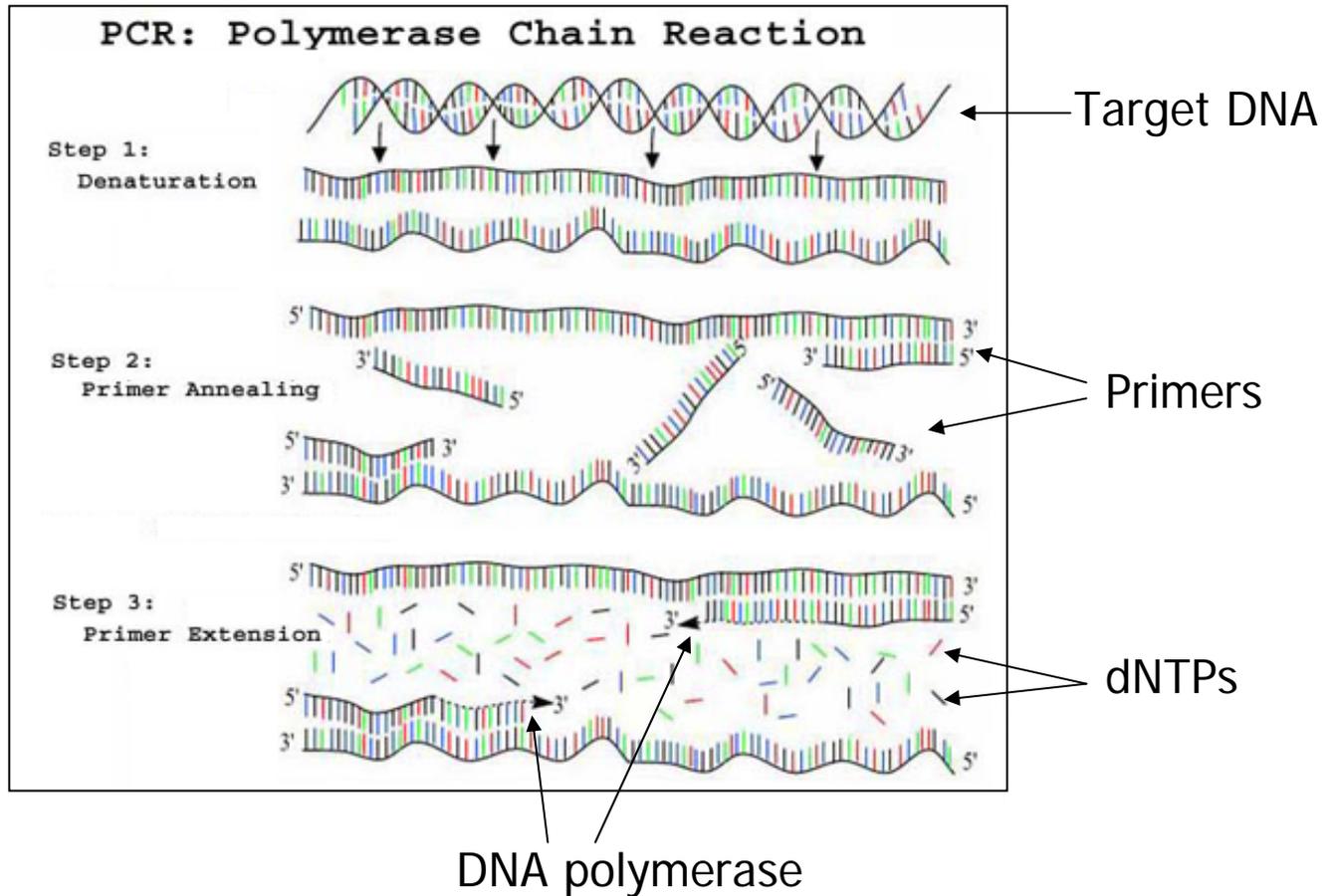
5. Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.



PCR



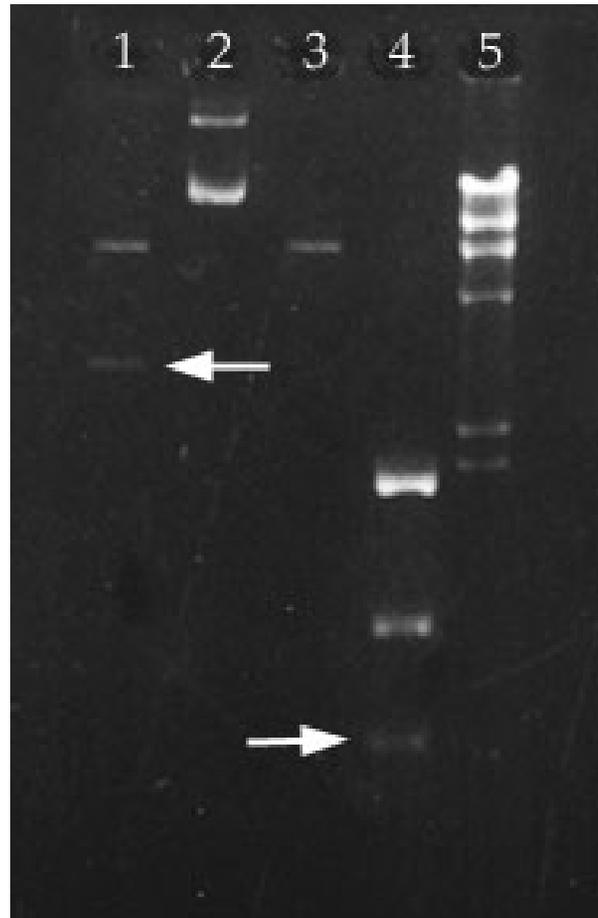
Schematic outline of a typical PCR cycle



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).

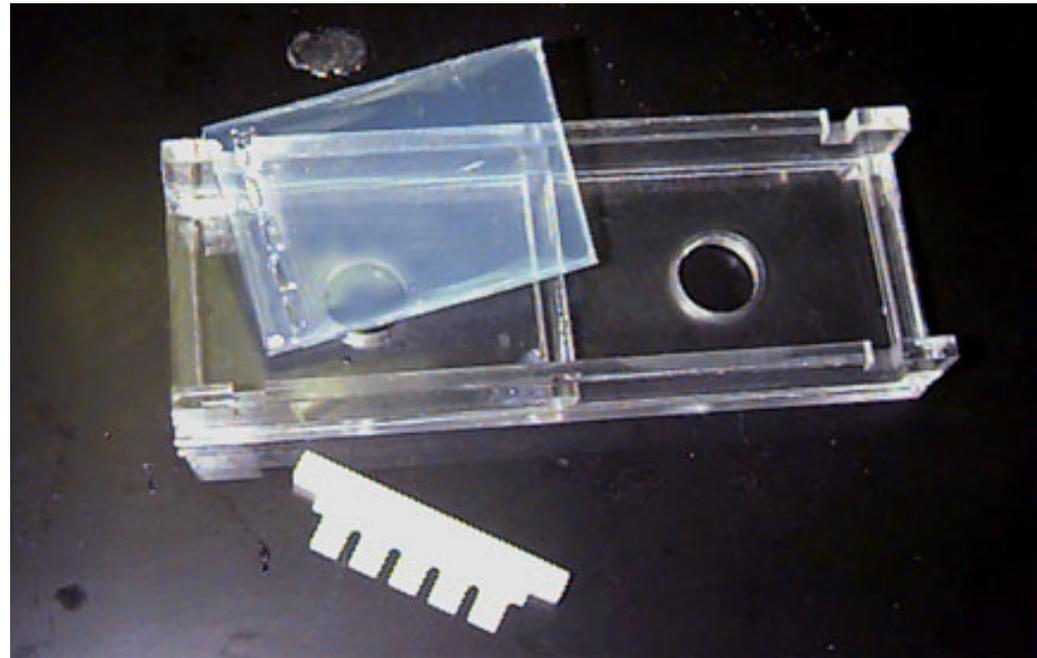
Gel Pictures



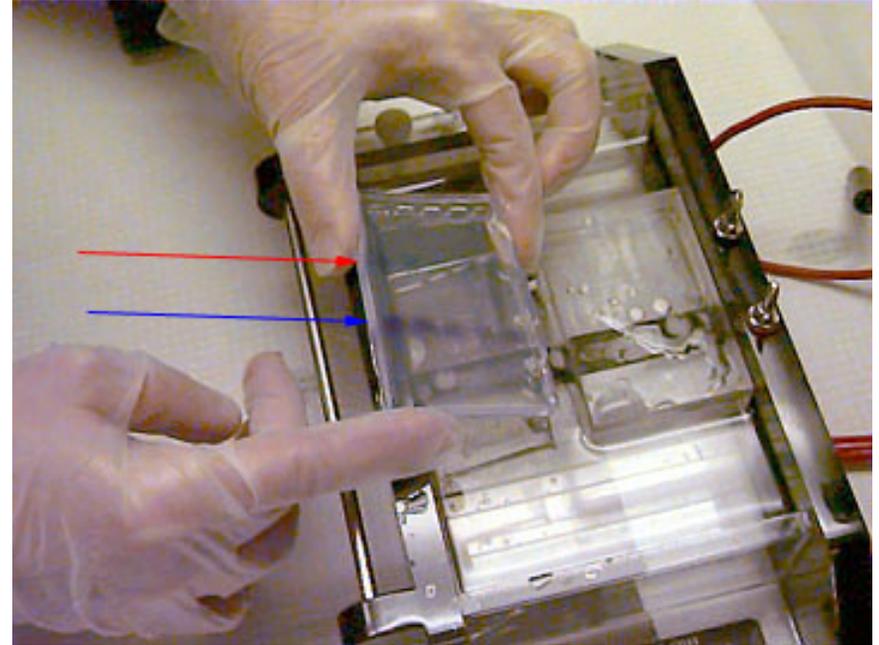
Gel Electrophoresis: Measure sizes of fragments

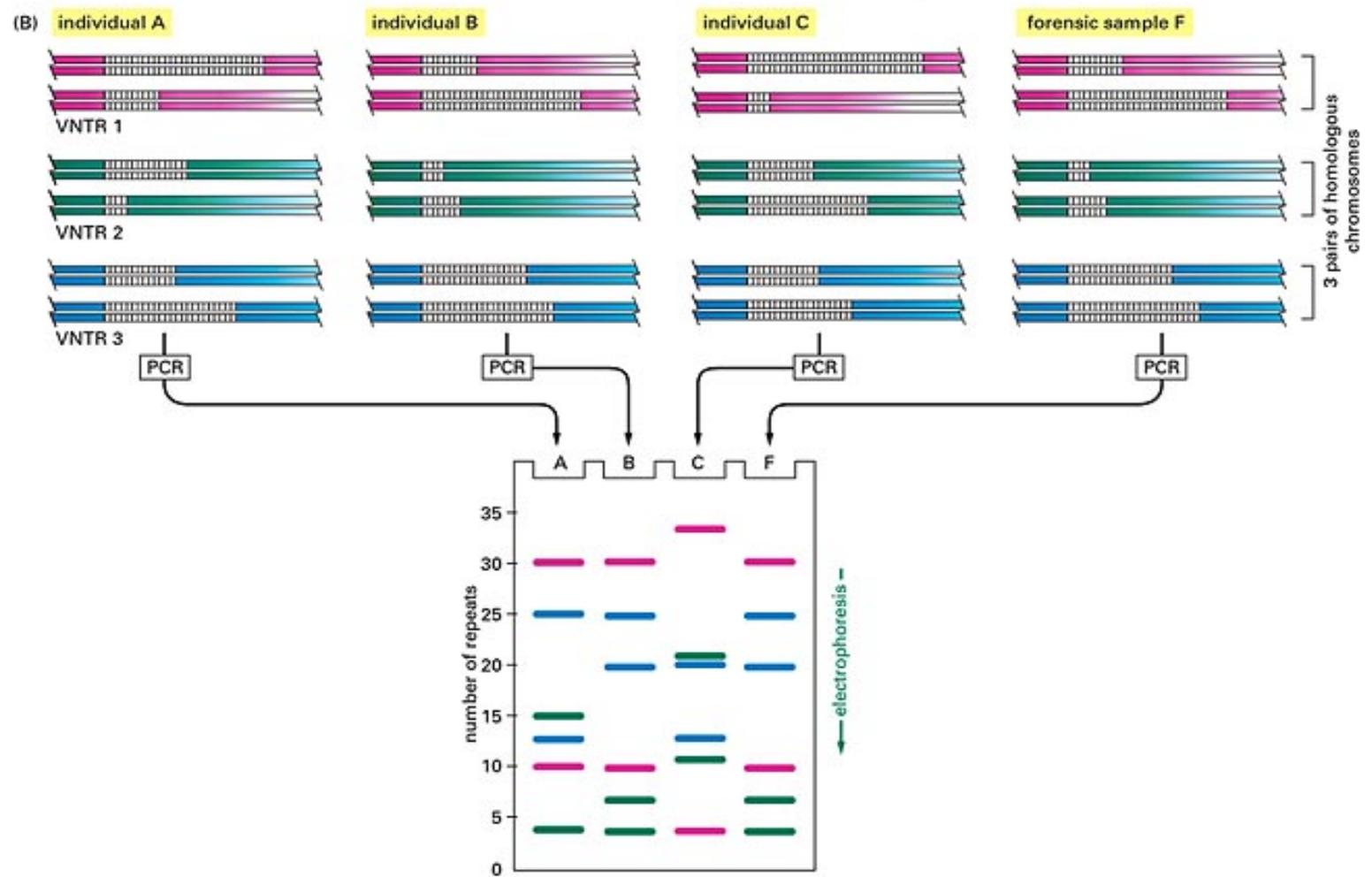
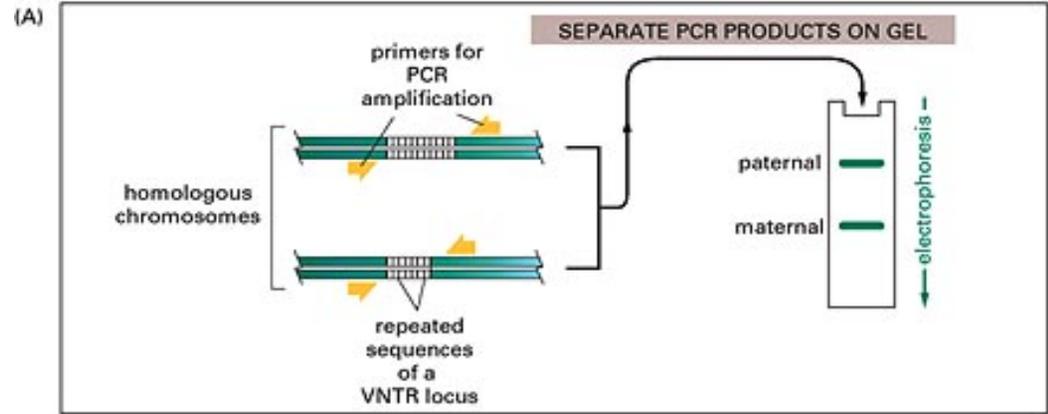
- The phosphate backbone makes DNA a highly negatively charged molecule. Thus DNA can be fractionated according to its size.
- **Gel:** allow hot 1 % solution of purified agarose to cool and solidify/polymerize (like Jello).
- DNA sample added to wells at the top of a gel and voltage is applied. Larger 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

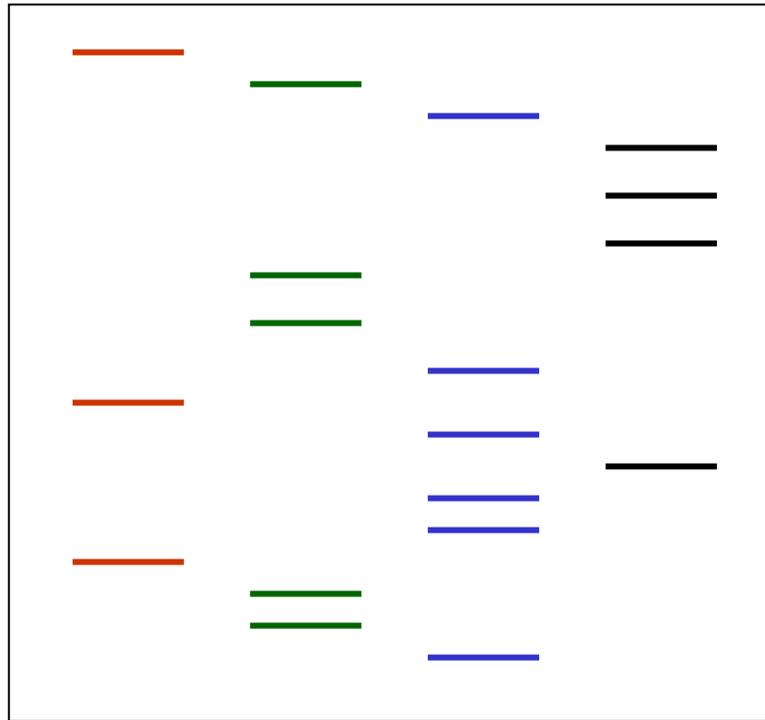




Sequencing a Fragment Using Gels

- 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

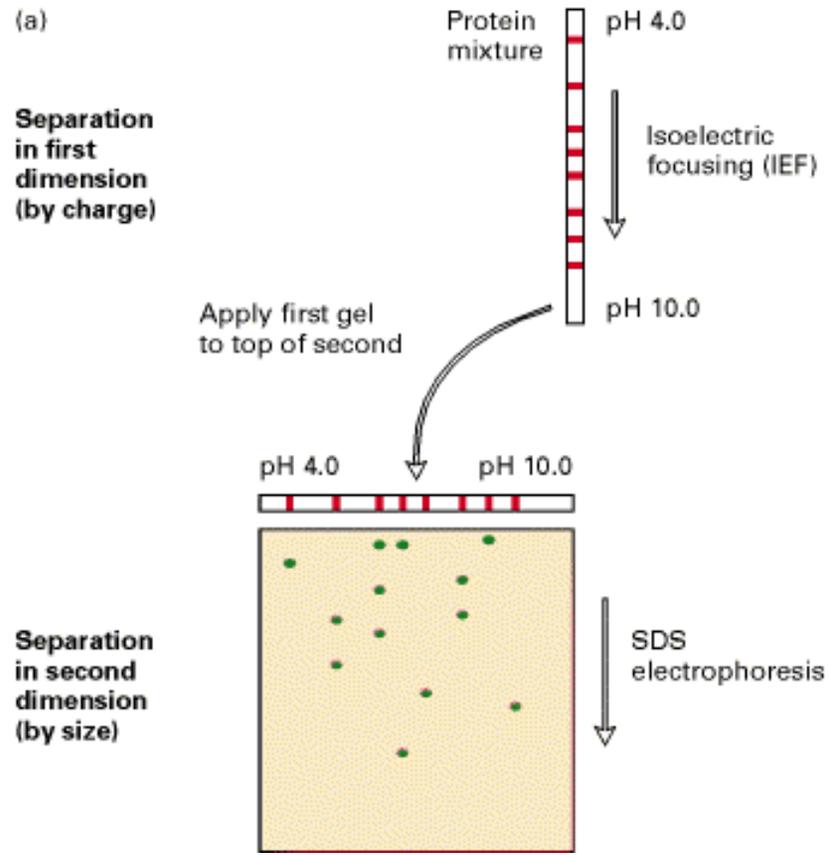
A

C

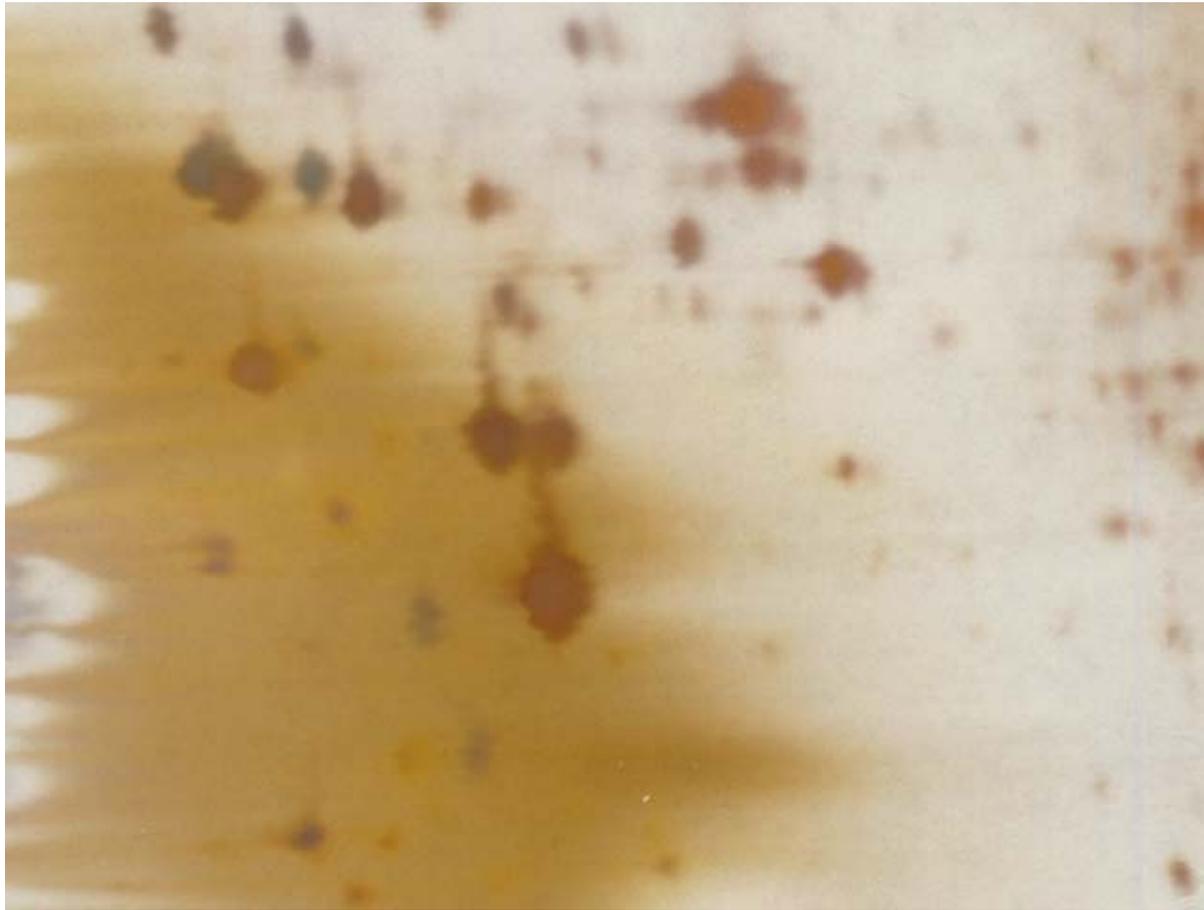
G

T

2D-Gels



2D Gel Electrophoresis



2D-Gels

First Dimension Methodology of a 2D Gel:

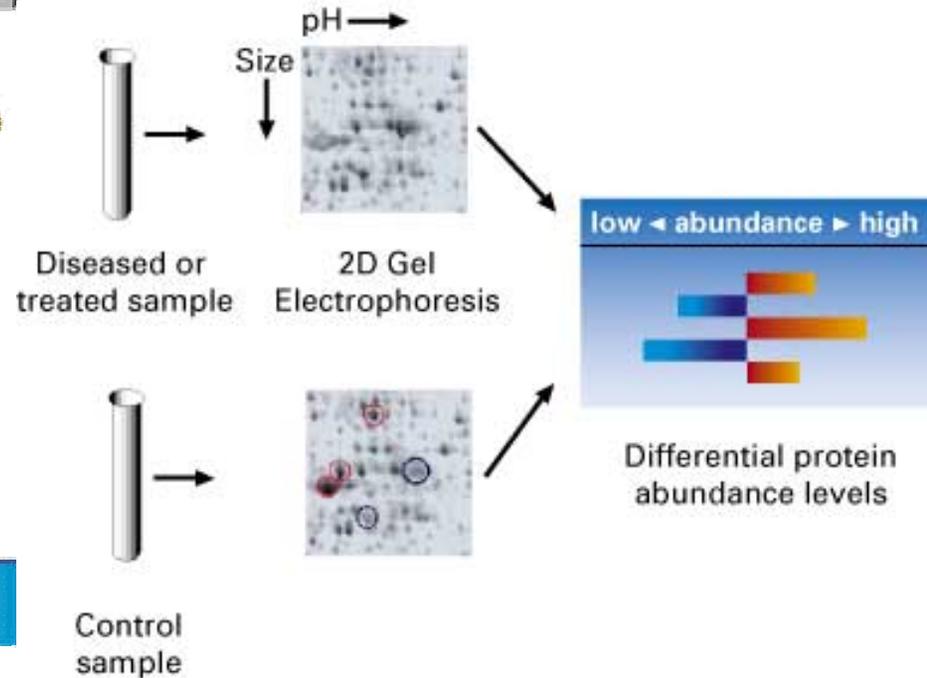
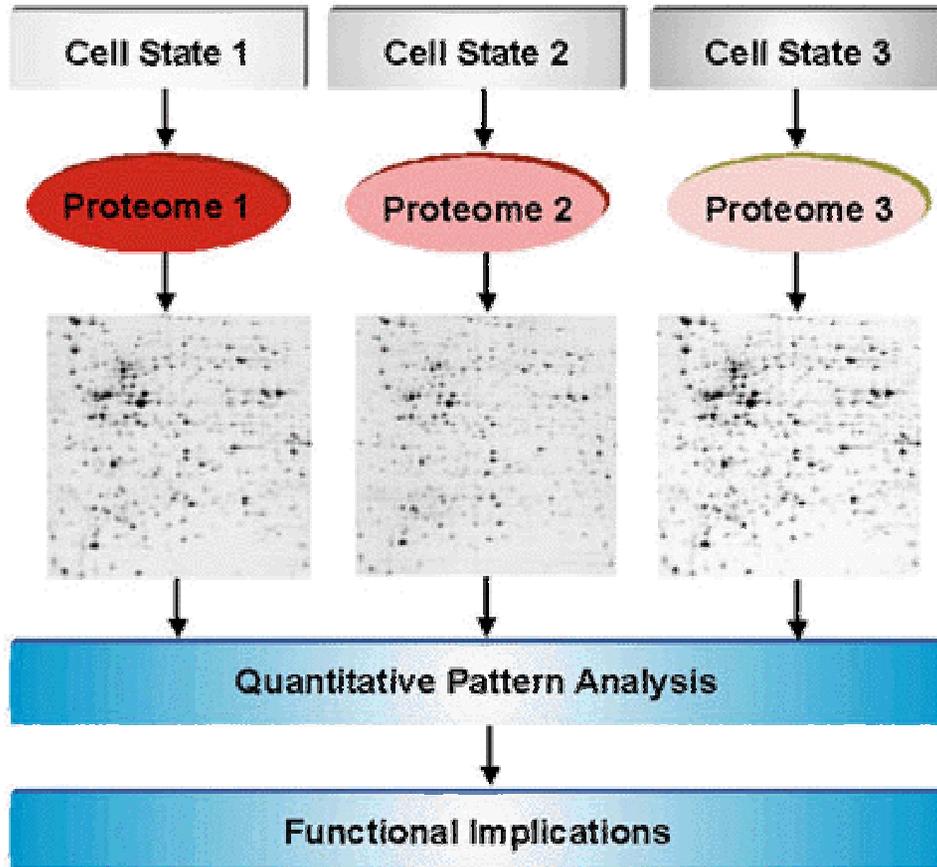
Denatured cell extract layered on a glass tube filled with polyacrylamide saturated with solution of ampholytes, a mixture of polyanionic[(-) charged] and polycationic [(+) charged] molecules. When placed in an electric field, the ampholytes separate and form continuous gradient based on net charge. 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.

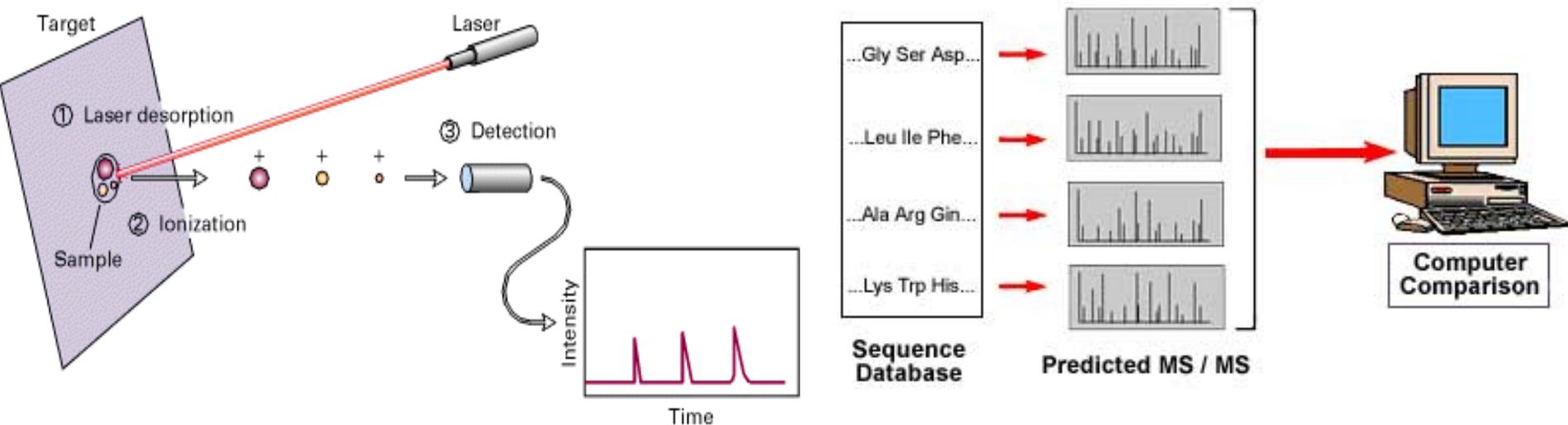
2D-gels



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.

Genomics

- Study of all genes in a genome, or comparison of whole genomes.
 - Whole genome sequencing
 - Whole genome annotation & Functional genomics
 - Whole genome comparison
 - **PipMaker**: uses BLASTZ to compare very long sequences (> 2Mb); <http://www.cse.psu.edu/pipmaker/>
 - **Mummer**: used for comparing long microbial sequences (uses Suffix trees!)

Genomics

- Study of all genes in a genome
 - Gene Expression
 - Microarray experiments & analysis
 - Probe design (**CODEHOP**)
 - Array image analysis (**CrazyQuant**)
 - Identifying genes with significant changes (**SAM**)
 - Clustering

Comparative Genomics

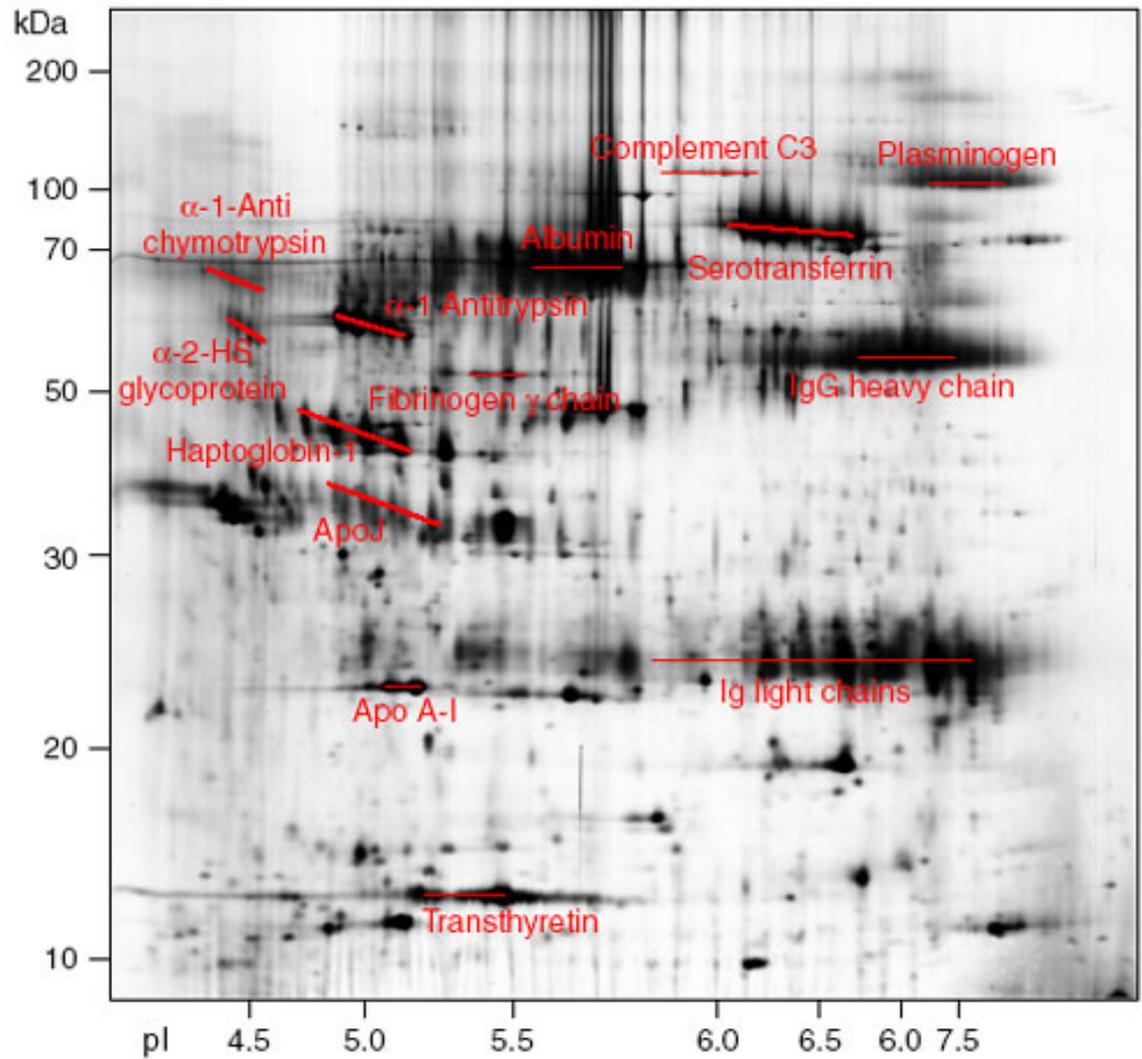
- Comparison of whole genomes.
 - Whole genome sequencing
 - Whole genome annotation & Functional genomics
 - Whole genome comparison
 - **PipMaker, MultiPipMaker, EnteriX**: PipMaker uses BLASTZ to compare very long sequences (> 2Mb);
<http://www.cse.psu.edu/pipmaker/>
 - **Mummer**: used for comparing long microbial sequences (uses Suffix trees!)
 - Many more!

Databases for Comparative Genomics

- PEDANT useful resource for standard questions in comparative genomics. For e.g., *how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.*
- COGs Clusters of orthologous groups of proteins.
- MBGD Microbial genome database searches for homologs in all microbial genomes

Proteomics

- Study of all **proteins** in a genome, or comparison of whole genomes.
 - Whole genome annotation & Functional proteomics
 - Whole genome comparison
 - Protein Expression: **2D Gel Electrophoresis**



TRENDS in Biotechnology

Other Proteomics Tools

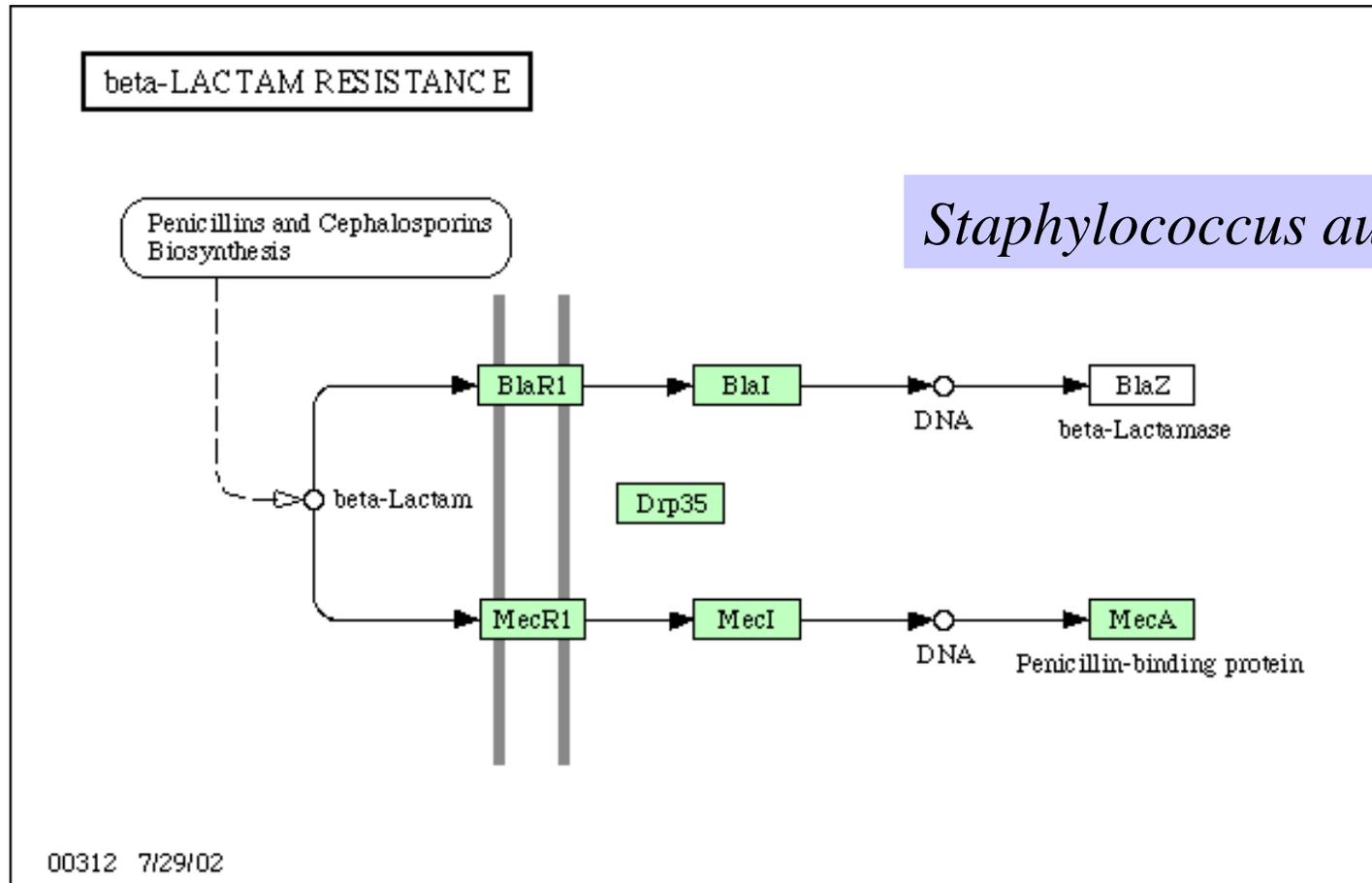
From ExPASy/SWISS-PROT:

- **AACompIdent** identify proteins from aa composition
[Input: aa composition, isoelectric point, mol wt., etc. Output: proteins from DB]
- **AACompSim** compares proteins aa composition with other proteins
- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins
- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins
- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.
- **PeptideMass** theoretical mass fingerprint for a given protein.
- **GlycoMod** predicts oligosaccharide modifications from mass difference
- **TGREASE** calculates hydrophobicity of protein along its length

Gene Networks & Pathways

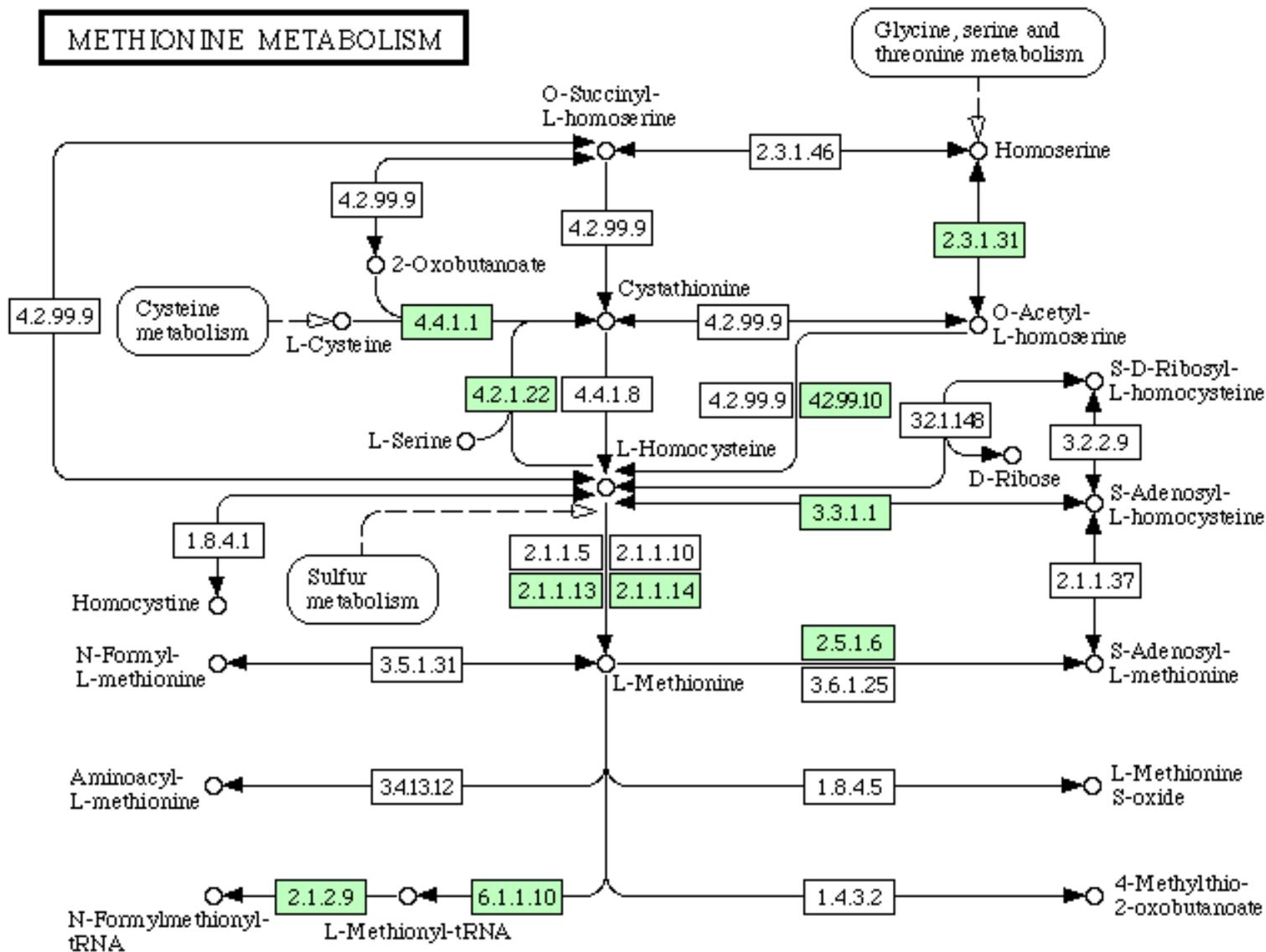
- Genes & Proteins act in concert and therefore form a complex network of dependencies.

Pathway Example from KEGG



Pseudomonas aeruginosa

METHIONINE METABOLISM



STSs and ESTs

- **Sequence-Tagged Site**: short, unique sequence
- **Expressed Sequence Tag**: short, unique sequence from a coding region
 - 1991: 609 ESTs [Adams et al.]
 - June 2000: 4.6 million in **dbEST**
 - Genome sequencing center at St. Louis produce 20,000 ESTs per week.

What Are ESTs and How Are They Made?

- Small pieces of DNA sequence (usually 200 - 500 nucleotides) of low quality.
- Extract mRNA from cells, tissues, or organs and sequence either end. Reverse transcribe to get cDNA (5' EST and 3'EST) and deposit in EST library.
- Used as "**tags**" or markers for that gene.
- Can be used to identify similar genes from other organisms (Complications: variations among organisms, variations in genome size, presence or absence of **introns**).
- 5' ESTs tend to be more useful (cross-species conservation), 3' EST often in UTR.

DNA Markers

- Uniquely identifiable DNA segments.
- Short, <500 nucleotides.
- Layout of these markers give a **map** of genome.
- Markers may be **polymorphic** (variations among individuals). Polymorphism gives rise to **alleles**.
- Found by PCR assays.

Polymorphisms

- Length polymorphisms
 - Variable # of tandem repeats (VNTR)
 - Microsatellites or short tandem repeats
 - Restriction fragment length polymorphism (RFLP) caused by changes in restriction sites.
- Single nucleotide polymorphism (SNP)
 - Average once every ~100 bases in humans
 - Usually biallelic
 - [dbSNP](#) database of SNPs (over 100,000 SNPs)
 - ESTs are a good source of SNPs

SNPs

- SNPs often act as “disease markers”, and provide “genetic predisposition”.
- SNPs may explain differences in drug response of individuals.
- **Association study**: study SNP patterns in diseased individuals and compare against SNP patterns in normal individuals.
- Many diseases associated with SNP profile.

Motifs in Protein Sequences

Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function**. They are also called **super-secondary structures**.

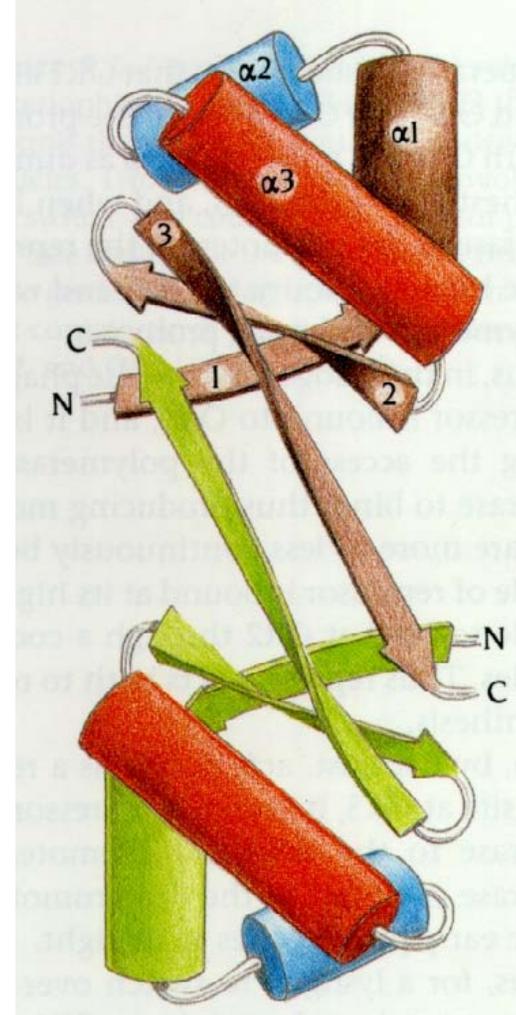
Examples: Helix-Turn-Helix, Zinc-finger, Homeobox domain, Hairpin-beta motif, Calcium-binding motif, Beta-alpha-beta motif, Coiled-coil motifs.

Several motifs may combine to form **domains**.

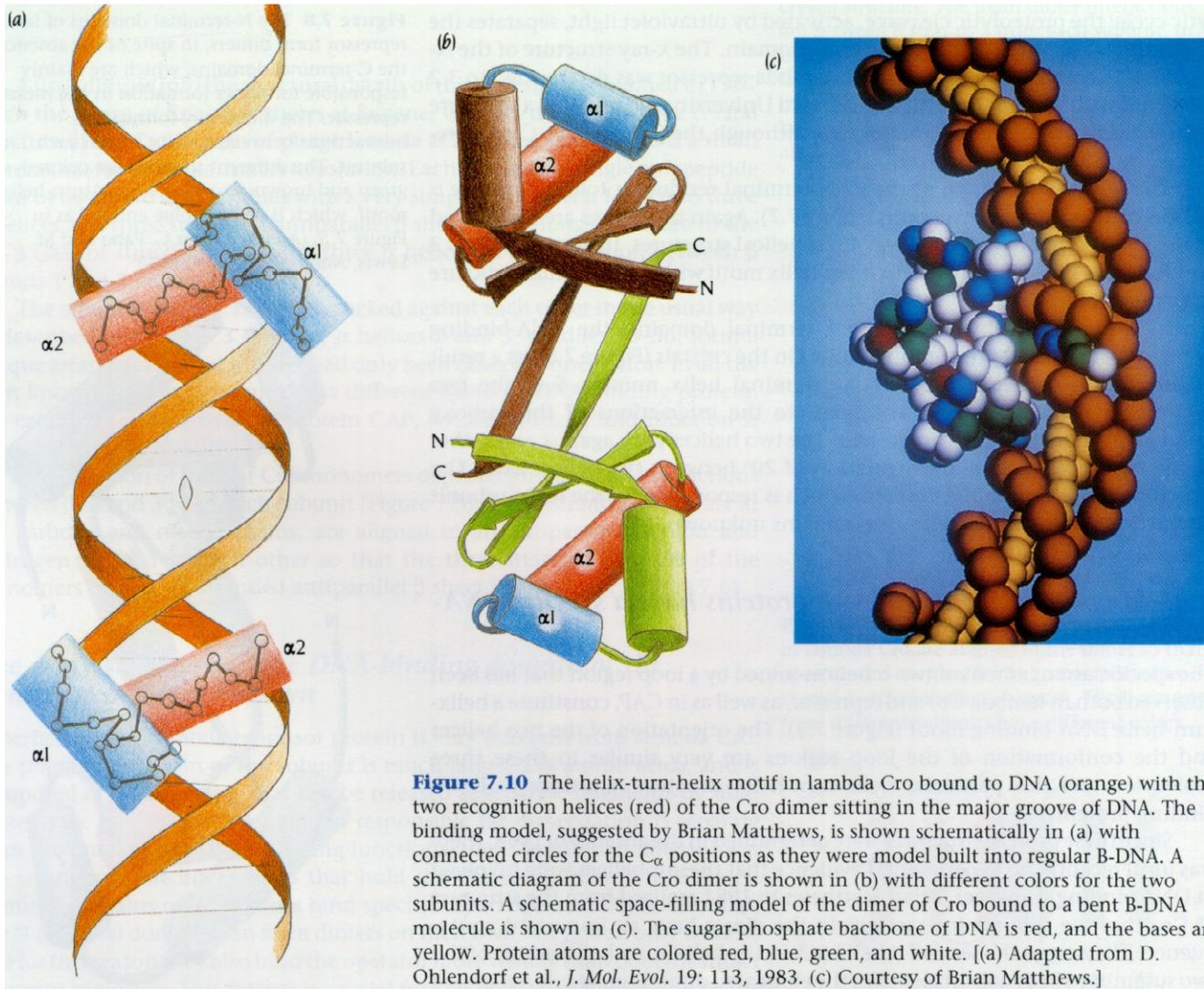
- Serine proteinase domain, Kringle domain, calcium-binding domain, homeobox domain.

Helix-Turn-Helix Motifs

- Structure
 - 3-helix complex
 - Length: 22 amino acids
 - Turn angle
- Function
 - Gene regulation by binding to DNA



DNA Binding at HTH Motif



HTH Motifs: Examples

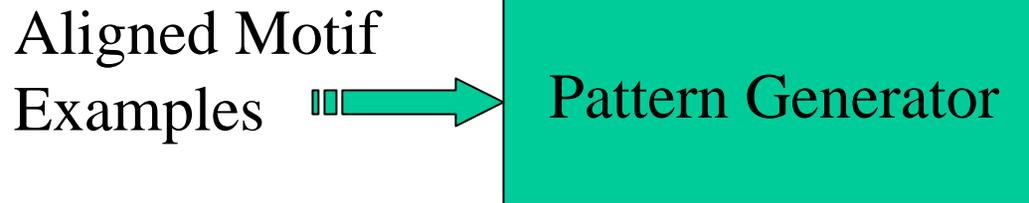
<i>Loc</i>	<i>Protein Name</i>	<i>Helix 2</i>									<i>Turn</i>				<i>Helix 3</i>								
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	Cro	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	434 Cro	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	P22 Cro	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	Rep	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	434 Rep	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	P22 Rep	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	CII	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	LacR	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	CAP	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	TrpR	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	BlaA Pv	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	TrpI Ps	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

Basis for New Algorithm

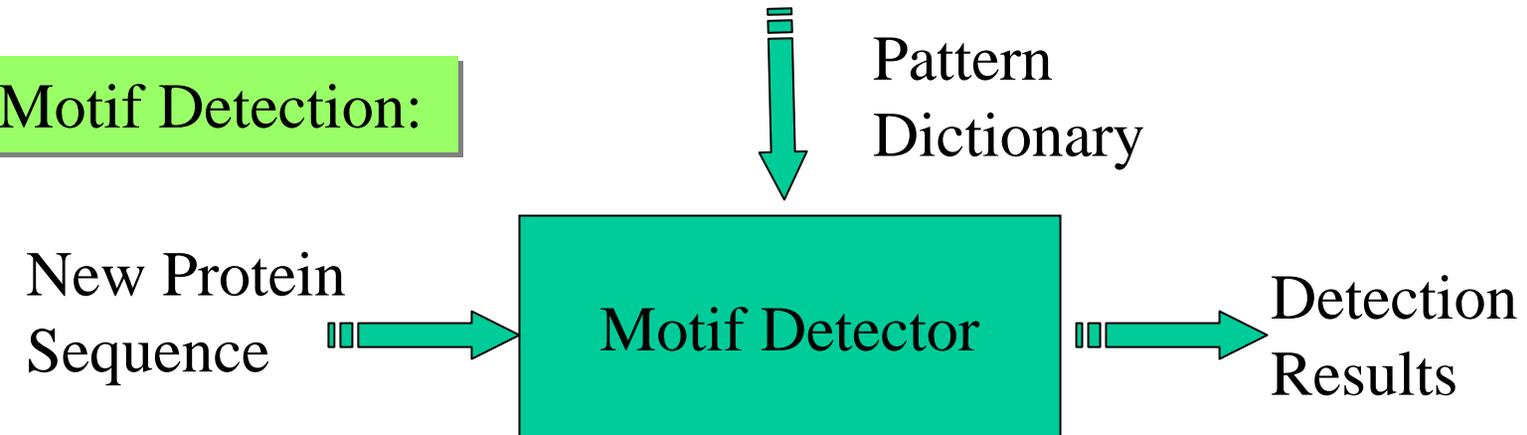
- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some **reinforcing** combinations are relatively rare.

New Motif Detection Algorithm

Pattern Generation:



Motif Detection:



Patterns

Loc	Protein Name	Helix 2									Turn				Helix 3								
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	Cro	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	434 Cro	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	P22 Cro	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	Rep	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	434 Rep	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	P22 Rep	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	CII	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	LacR	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	CAP	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	TrpR	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	BlaA Pv	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	TrpI Ps	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

- Q1 G9 N20
- A5 G9 V10 I15

Pattern Mining Algorithm

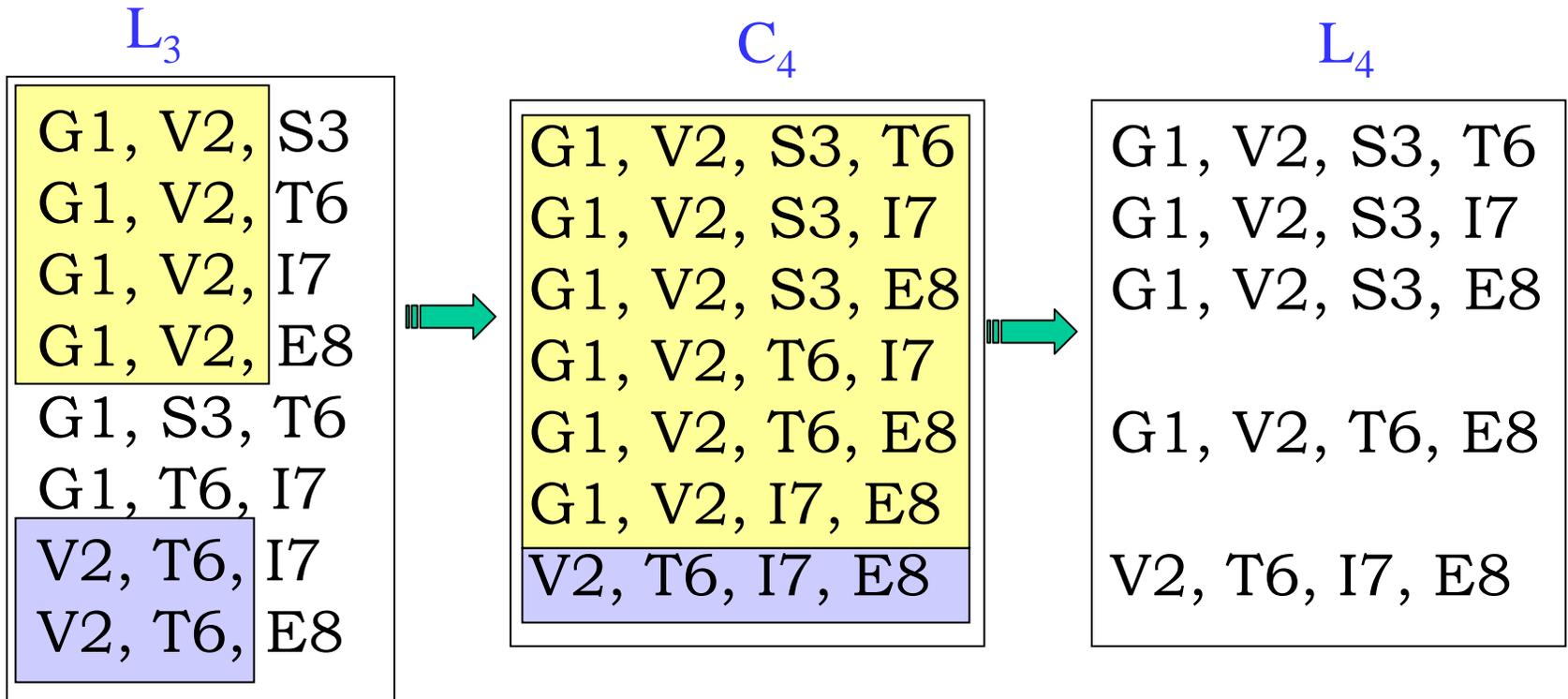
Algorithm **Pattern-Mining**

Input: Motif length **m**, support threshold **T**,
list of aligned motifs **M**.

Output: Dictionary **L** of frequent patterns.

1. $L_1 :=$ All frequent patterns of length 1
2. **for** $i = 2$ **to** **m** **do**
3. $C_i :=$ **Candidates**(L_{i-1})
4. $L_i :=$ Frequent candidates from C_i
5. **if** ($|L_i| \leq 1$) **then**
6. **return** **L** as the union of all L_j , $j \leq i$.

Candidates Function



Motif Detection Algorithm

Algorithm **Motif-Detection**

Input : Motif length **m**, threshold score **T**, pattern dictionary **L**, and input protein sequence **P**[1..n].

Output : Information about motif(s) detected.

1. **for** each location **i do**
2. **S** := **MatchScore**(**P**[**i**..**i+m-1**], **L**).
3. **if** (**S** > **T**) **then**
4. Report it as a possible motif

Experimental Results: GYM 2.0

<i>Motif</i>	<i>Protein Family</i>	<i>Number Tested</i>	<i>GYM = DE Agree</i>	<i>Number Annotated</i>	<i>GYM = Annot.</i>
<i>HTH Motif (22)</i>	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

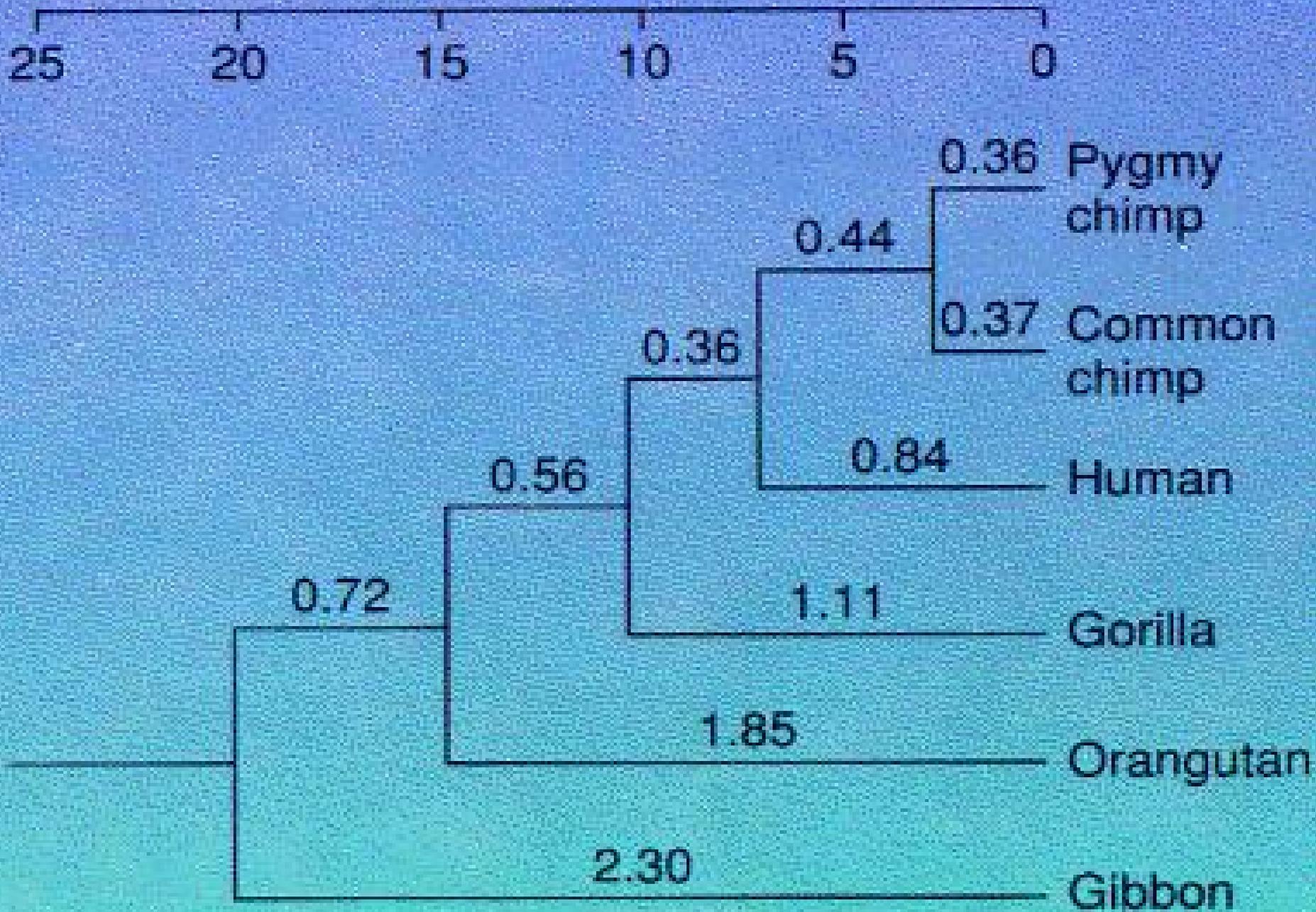
Experiments

- Basic Implementation (Y. Gao)
- Improved implementation & comprehensive testing (K. Mathee, GN).
- Implementation for homeobox domain detection (X. Wang).
- Statistical methods to determine thresholds (C. Bu).
- Use of substitution matrix (C. Bu).
- Study of patterns causing errors (N. Xu).
- Negative training set (N. Xu).
- NN implementation & testing (J. Liu & X. He).
- HMM implementation & testing (J. Liu & X. He).

Theory of Evolution

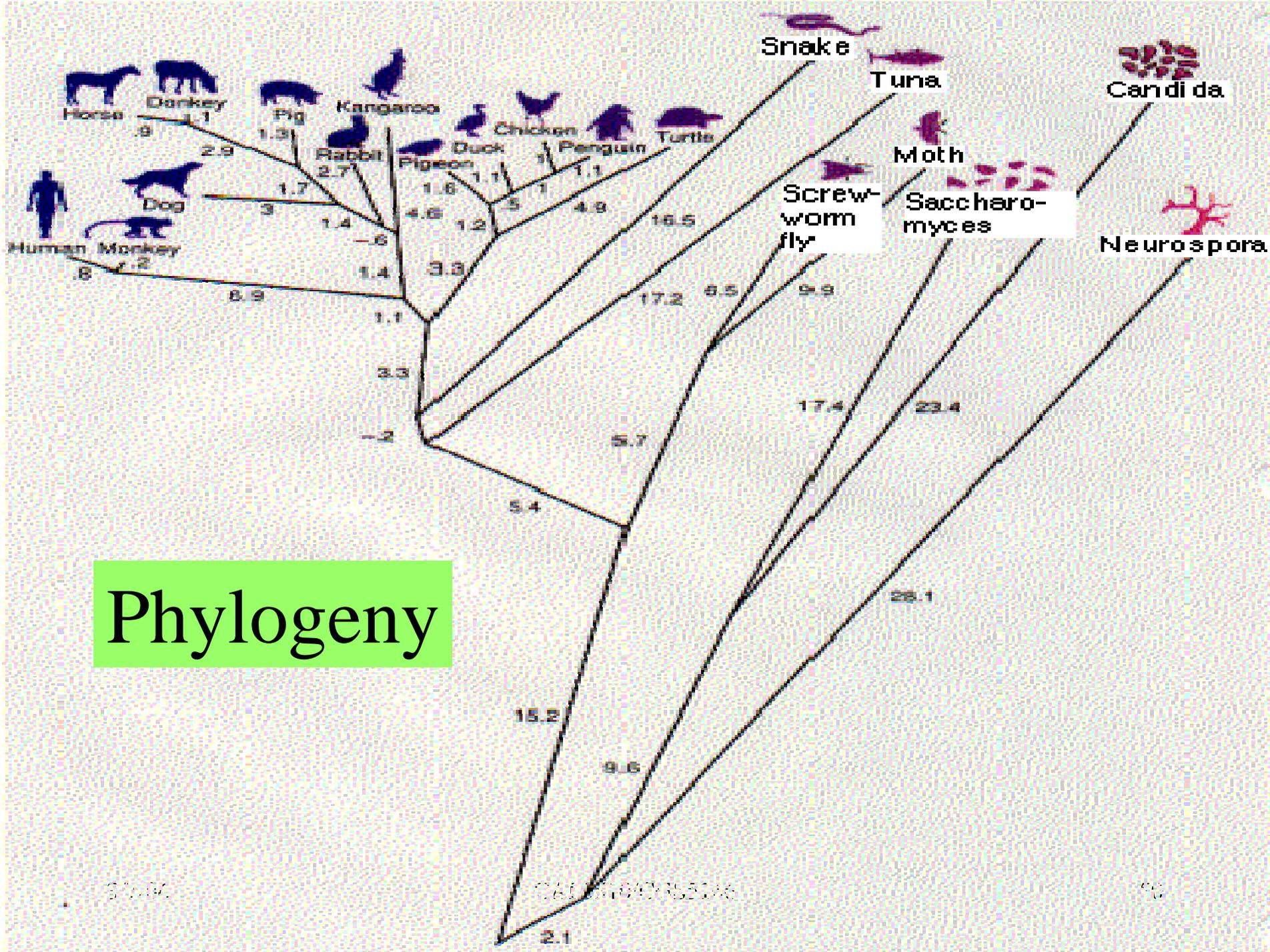
- Charles Darwin
 - **1858-59**: *Origin of Species*
 - 5 year voyage of H.M.S. Beagle (1831-36)
 - Populations have variations.
 - Natural Selection & Survival of the fittest: *nature selects best adapted varieties to survive and to reproduce.*
 - Speciation arises by splitting of one population into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

Millions of years



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**.



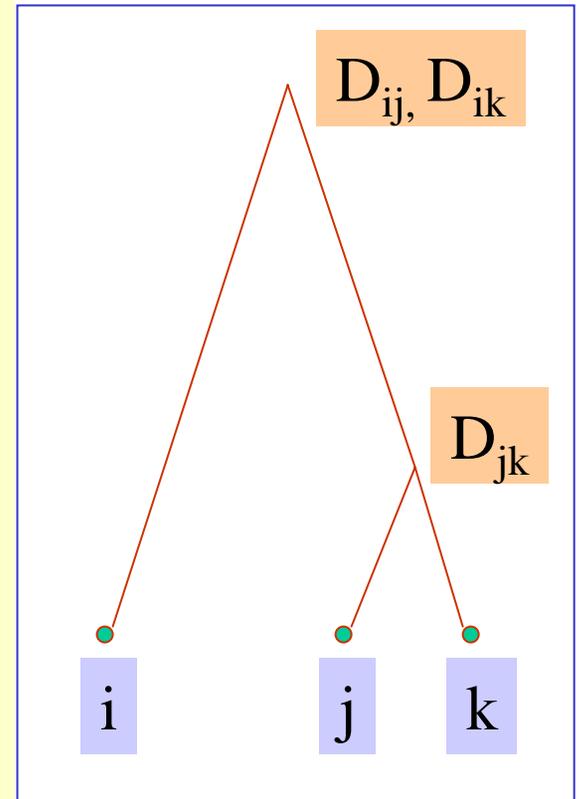
Phylogeny

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

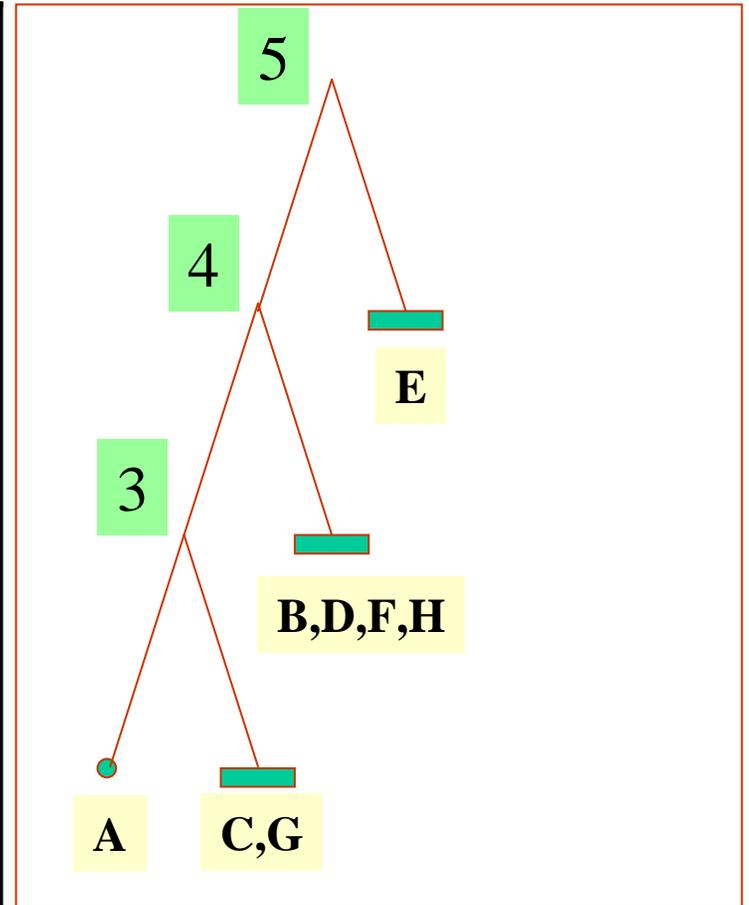
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: Accepted point mutations in amino acid sequence of a protein occurs at a **constant** rate.
 - Varies from protein to protein
 - Varies from one part of a protein to another

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**

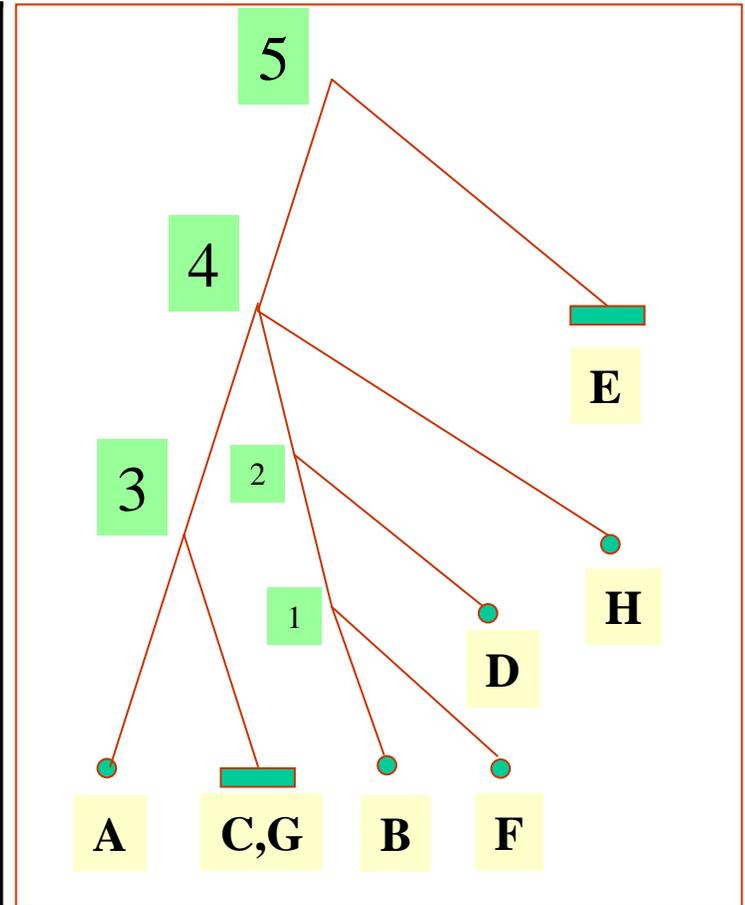
Ultrametric: Example

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



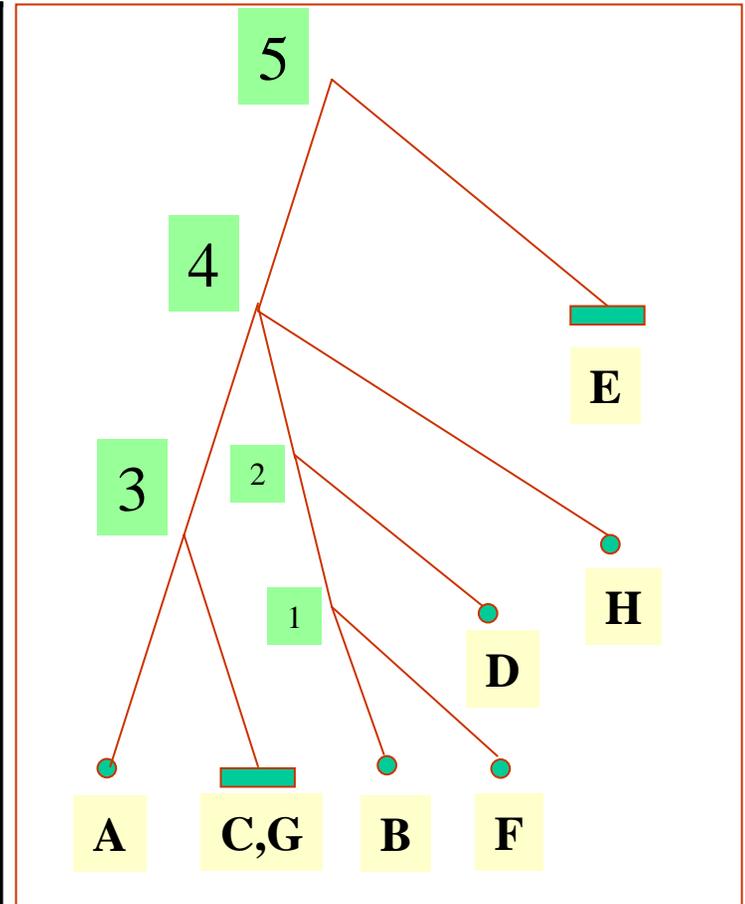
Ultrametric: Example

	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

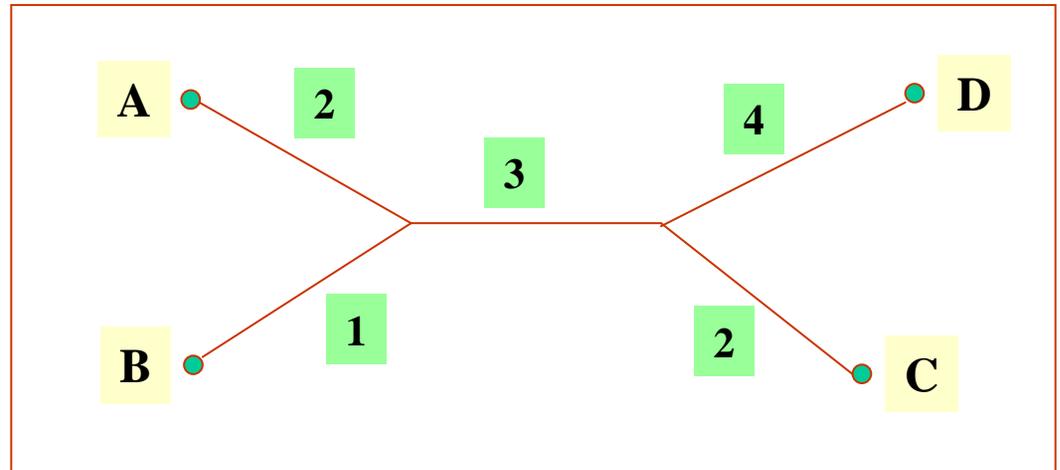
	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



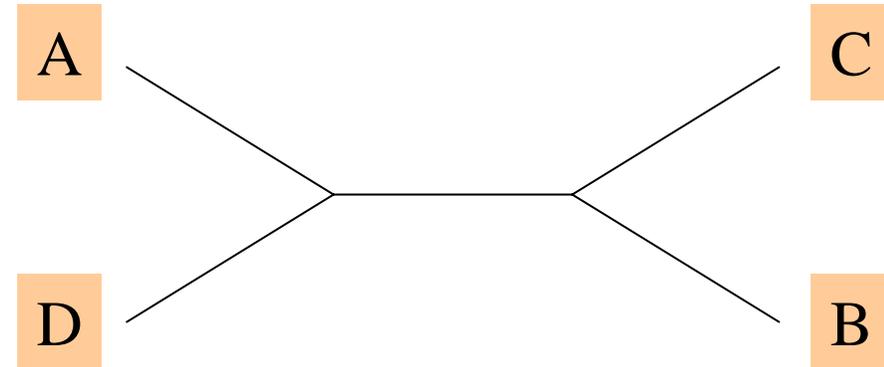
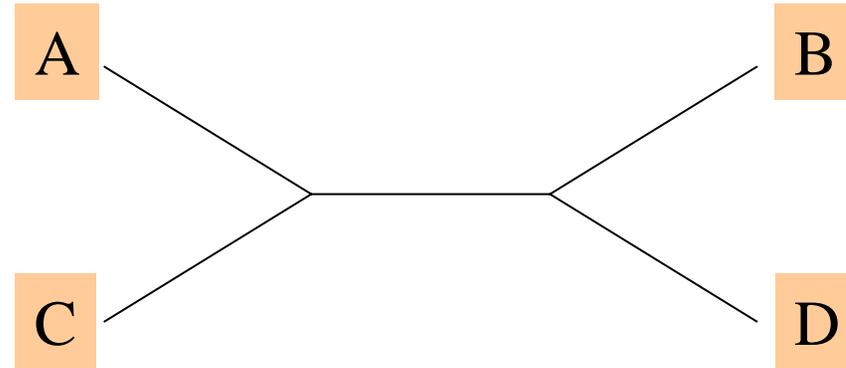
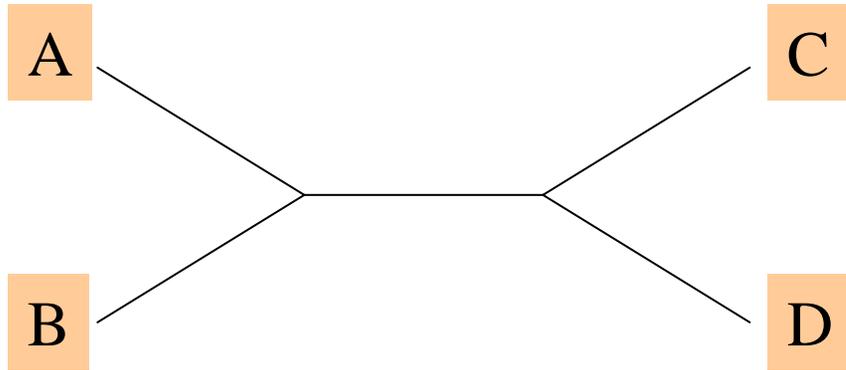
Additive-Distance Trees

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

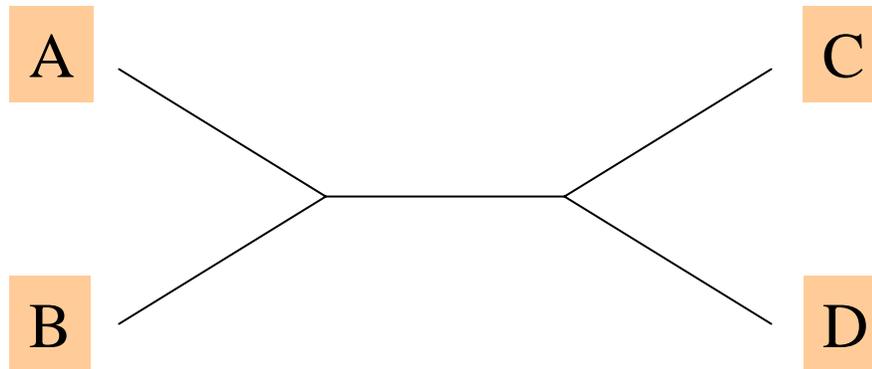


Four-Point Condition

- If the true tree is as shown below, then

1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and

2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

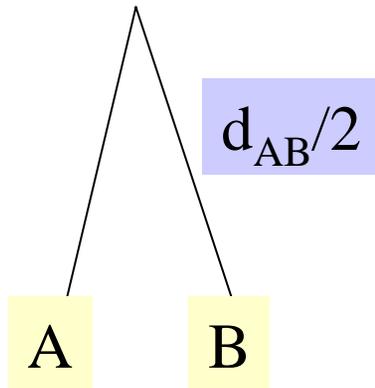


Unweighted pair-group method with arithmetic means (UPGMA)

	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

$$D_{ij}' = \frac{D_{ij} - D_{io} - D_{jo}}{2} + \left(\frac{\sum_{k=1}^n D_{ko}}{n} \right)$$

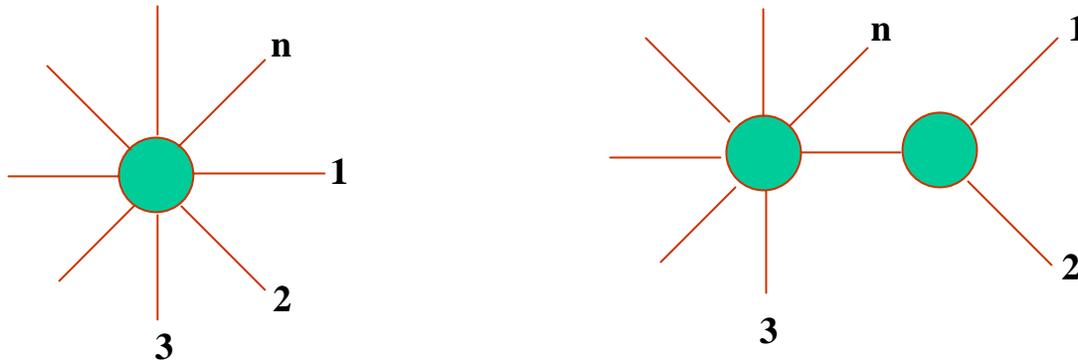
- Now apply UPGMA

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.

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

Neighbor-Joining



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

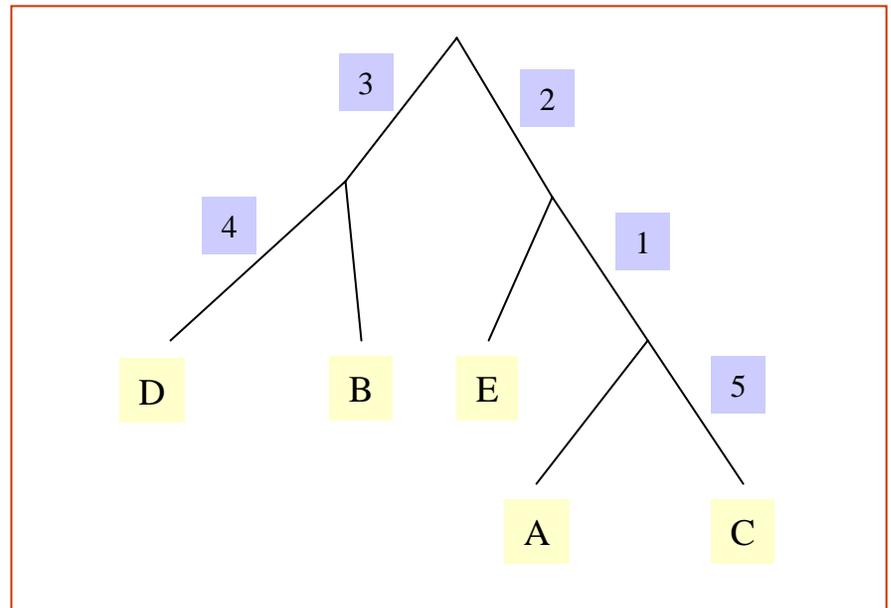
Constructing Evolutionary/Phylogenetic Trees

- 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 \Leftrightarrow 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

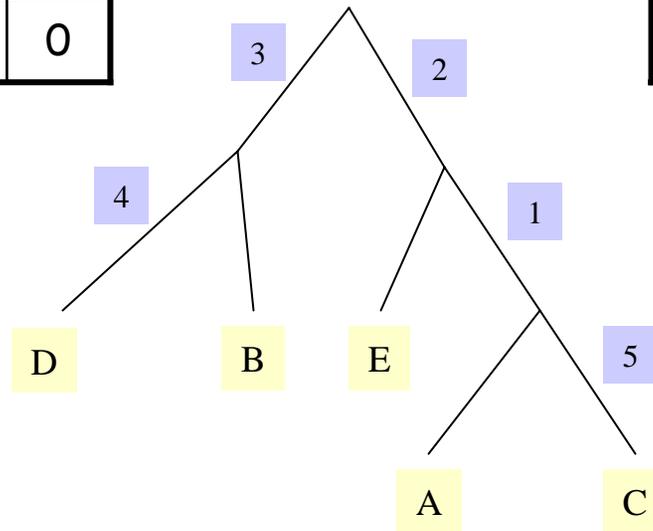


Example

- Perfect phylogeny does not always exist.

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

	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

Examples of Character Data

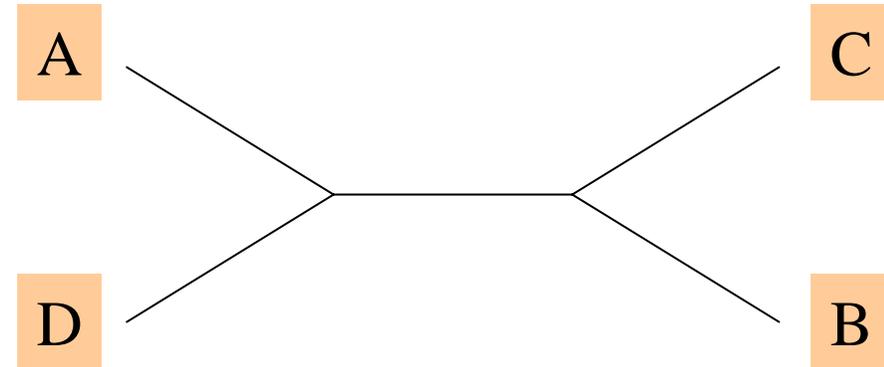
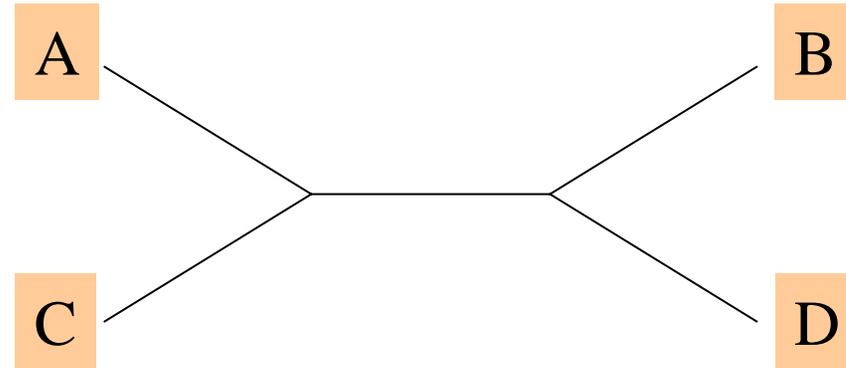
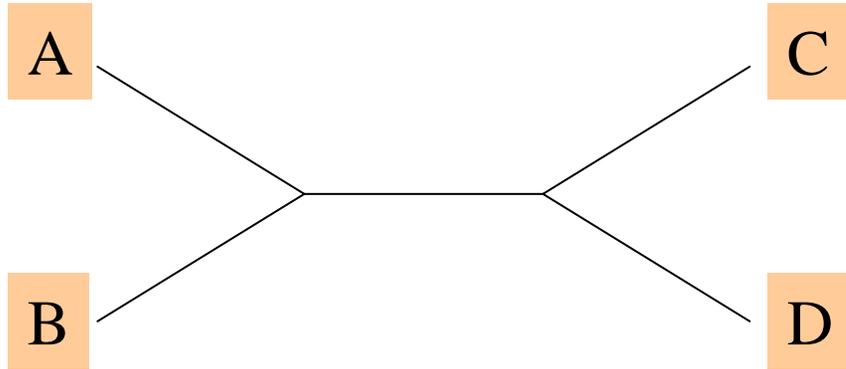
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Maximum Parsimony Method: Example

	Characters/Sites								
Sequence s	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Unrooted Trees on 4 Taxa



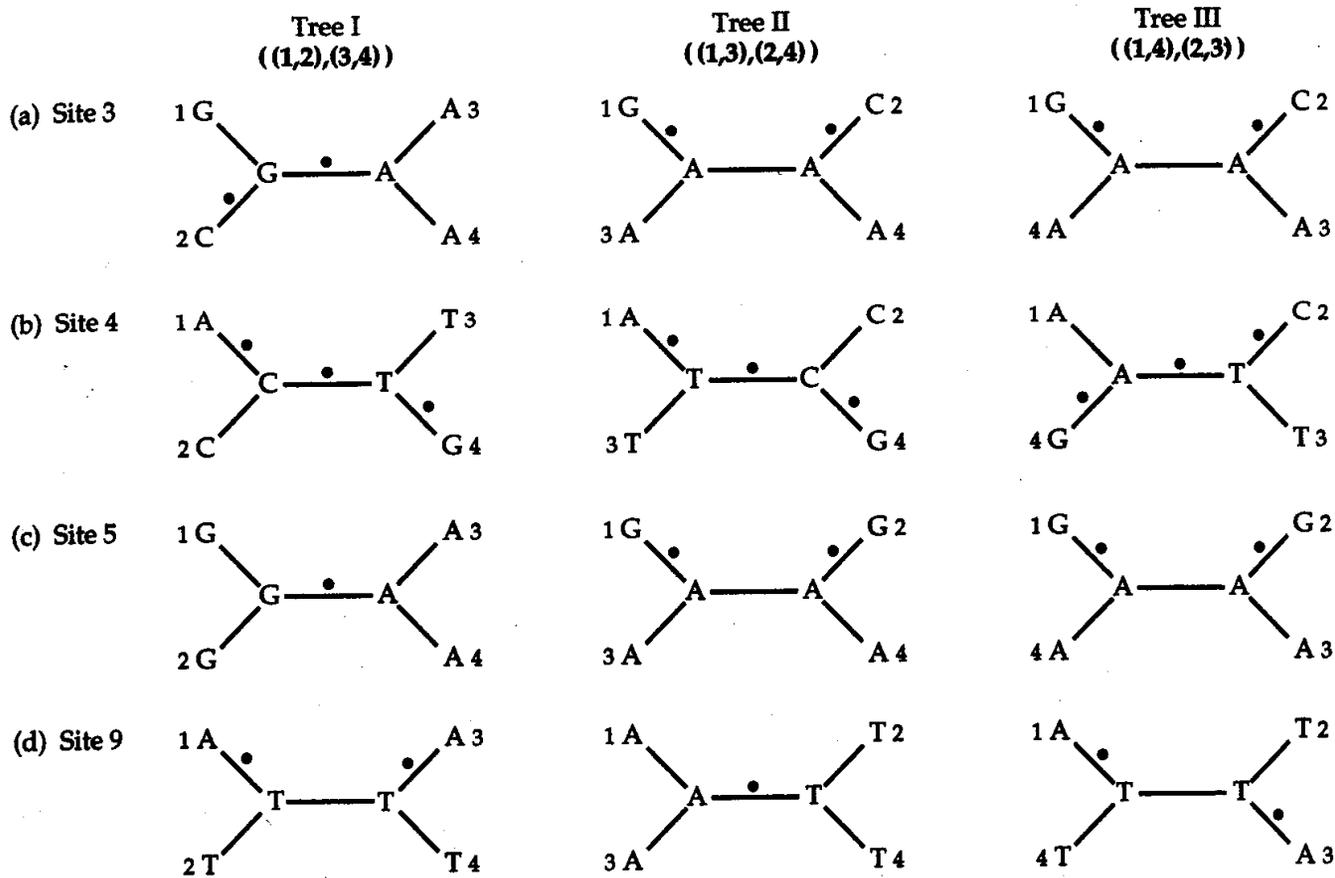


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III(d) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Inferring nucleotides on internal nodes

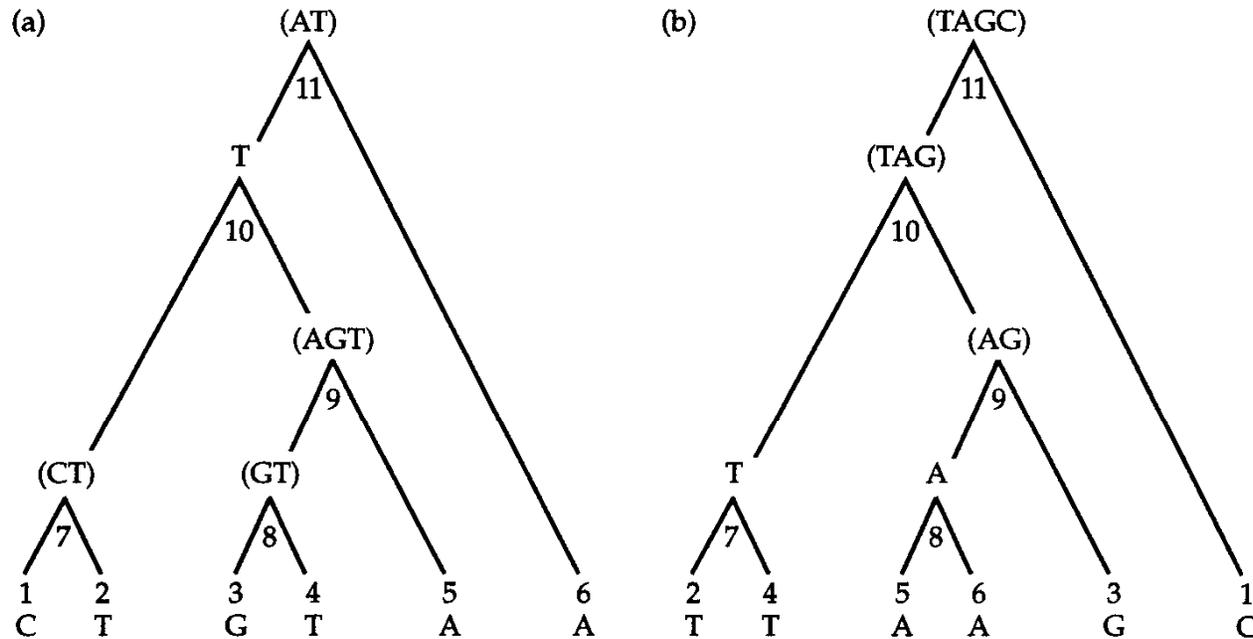


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).

Searching for the Maximum Parsimony Tree: Exhaustive Search

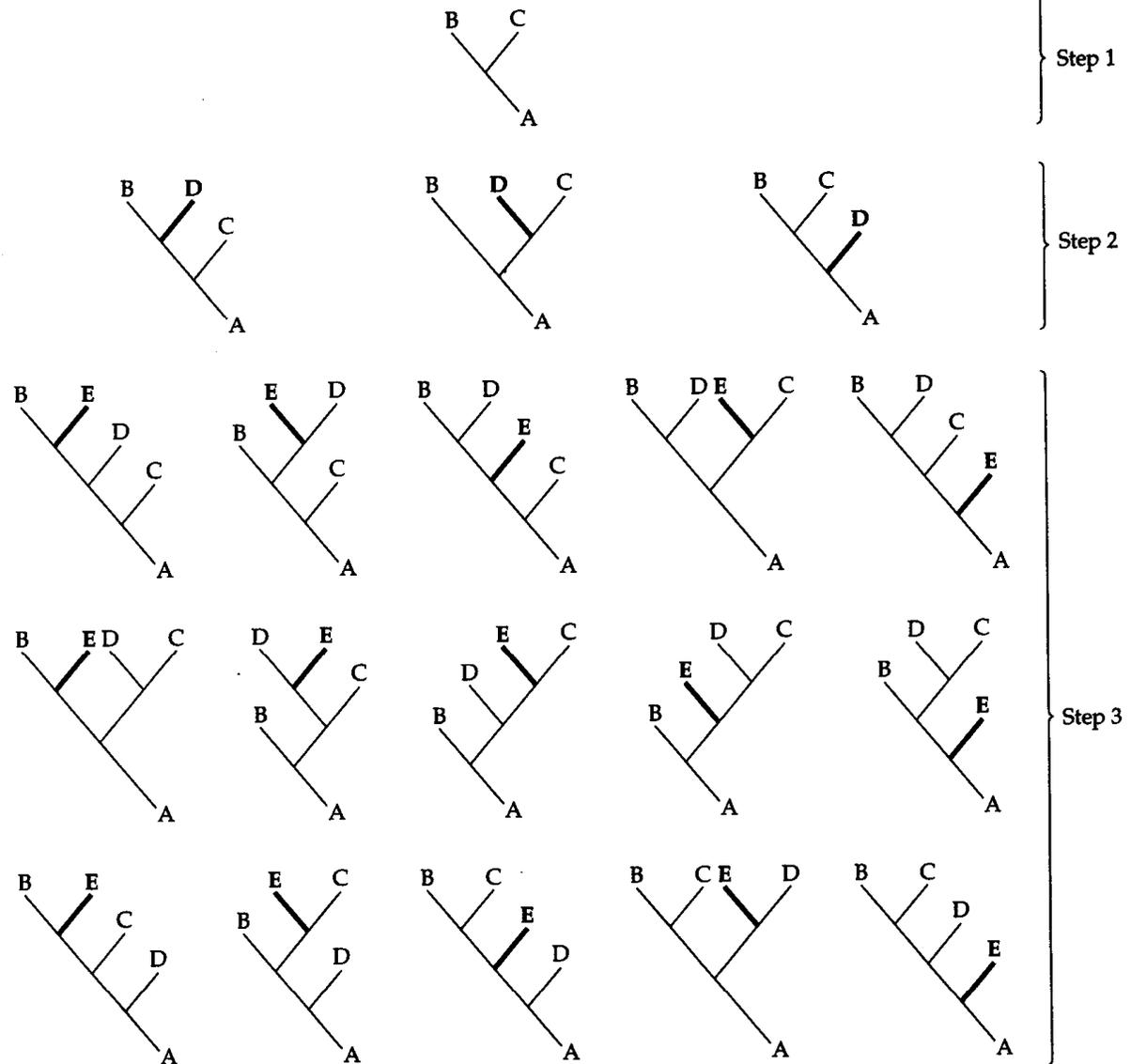
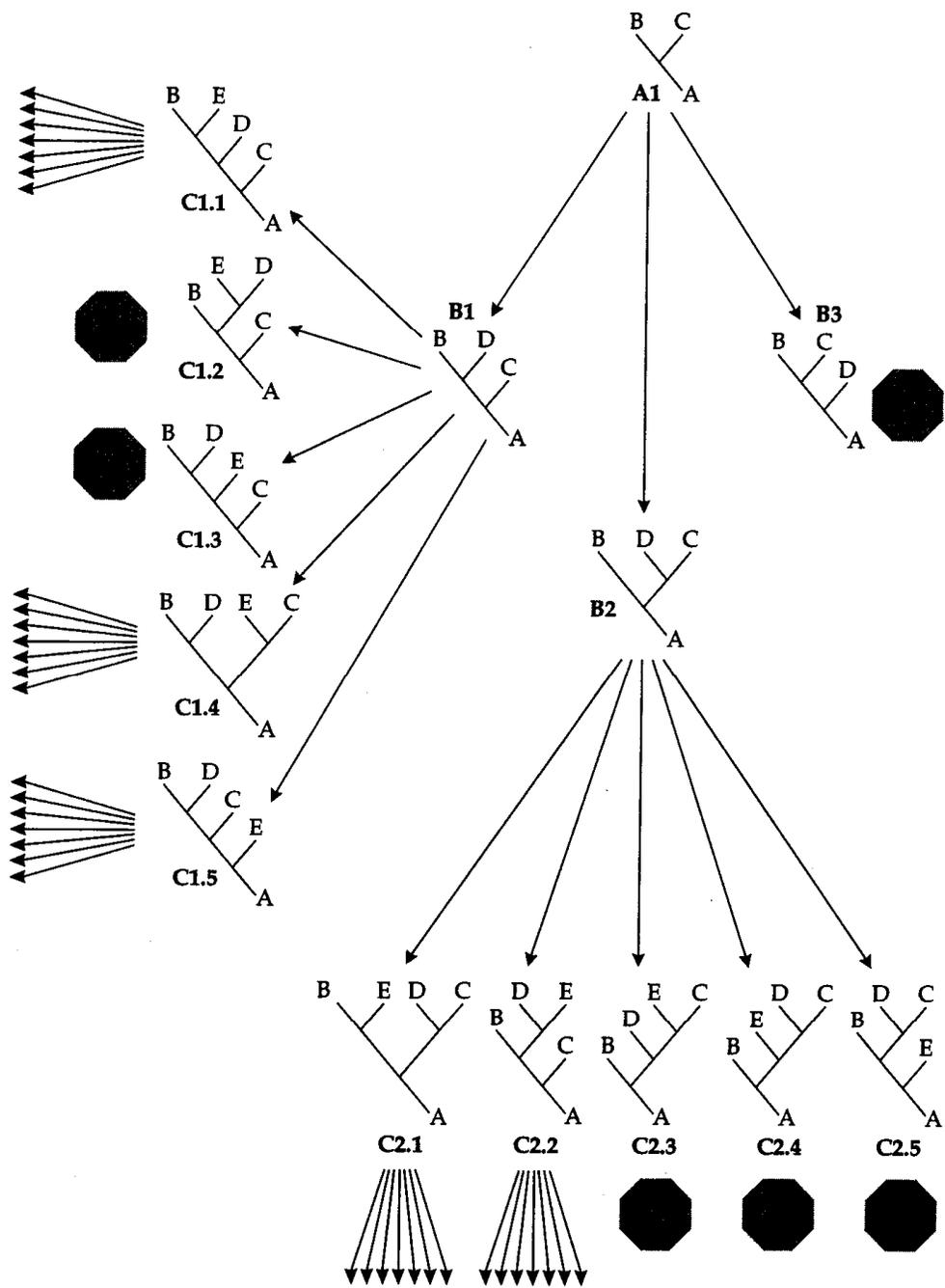


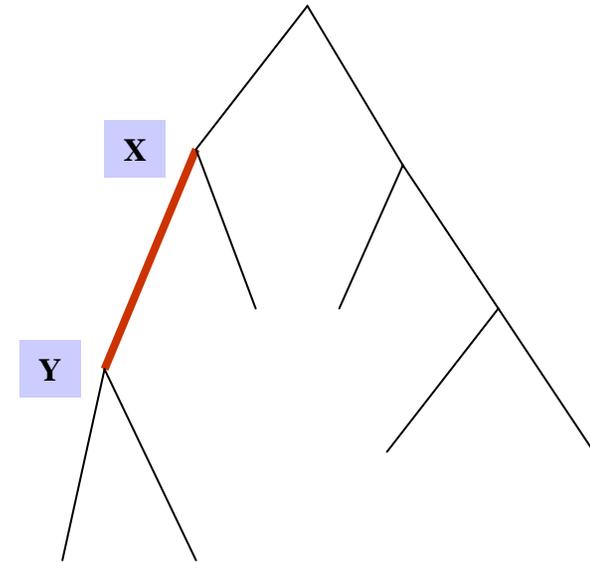
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four OTUs. In step 3, we add OTU E to each of the five branches of the three trees in step 2, thereby generating 15 unrooted trees. Additions of OTUs are shown as heavier lines. Modified from Swofford et al. (1996).



Searching for the
Maximum
Parsimony Tree:
Branch-&-Bound

Probabilistic Models of Evolution

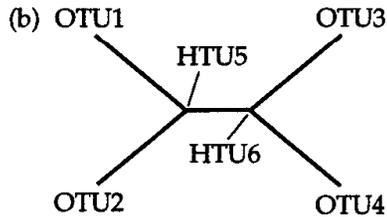
- Assuming a **model of substitution**,
 - $\Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$,
- Using this formula it is possible to compute the likelihood that data D is generated by a given phylogenetic tree T under a model of substitution. Now find the tree with the maximum likelihood.



- Time elapsed? Δ
- Prob of change along edge?
 $\Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$
- Prob of data? **Product of prob for all edges**

FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length n . (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all n sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).

(a)	1	2	3	4	5	6	7	8	9	...	n
OTU1	A	A	G	A	C	T	T	C	A	...	N
OTU2	A	G	C	C	C	T	T	C	T	...	N
OTU3	A	G	A	T	A	T	C	C	A	...	N
OTU4	A	G	A	G	G	T	C	C	T	...	N



(c)

$$\begin{aligned}
 L_{(5)} = & \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \begin{array}{c} A \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \begin{array}{c} T \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \begin{array}{c} G \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} C \\ \diagup \end{array} \begin{array}{c} A \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} C \\ \diagup \end{array} \begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} C \\ \diagup \end{array} \begin{array}{c} T \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} C \\ \diagup \end{array} \begin{array}{c} G \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} T \\ \diagup \end{array} \begin{array}{c} A \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} T \\ \diagup \end{array} \begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} T \\ \diagup \end{array} \begin{array}{c} T \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} T \\ \diagup \end{array} \begin{array}{c} G \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \begin{array}{c} A \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \begin{array}{c} T \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagdown \end{array} \begin{array}{c} G \\ \diagup \end{array} \begin{array}{c} G \\ \diagdown \end{array} \begin{array}{c} A \\ \diagup \end{array} \right)
 \end{aligned}$$

(d) $L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \dots \times L_{(n)} = \prod_{i=1}^n L_{(i)}$

(e) $\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \dots + \ln L_{(n)} = \sum_{i=1}^n \ln L_{(i)}$

Computing Maximum Likelihood Tree