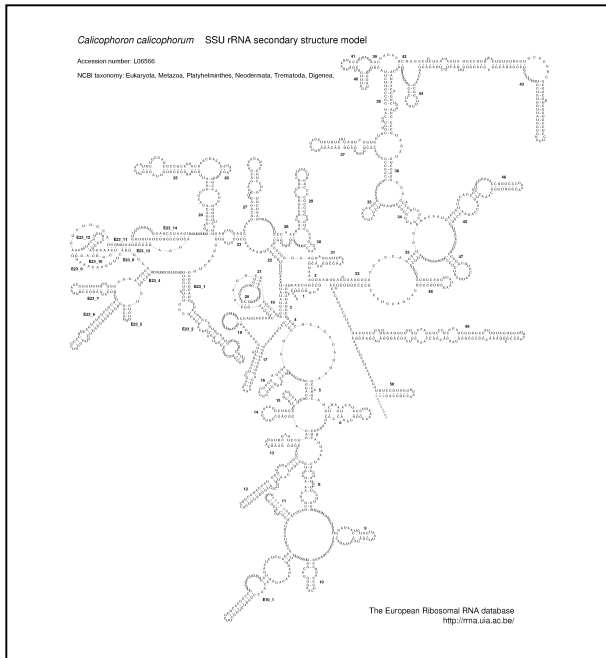
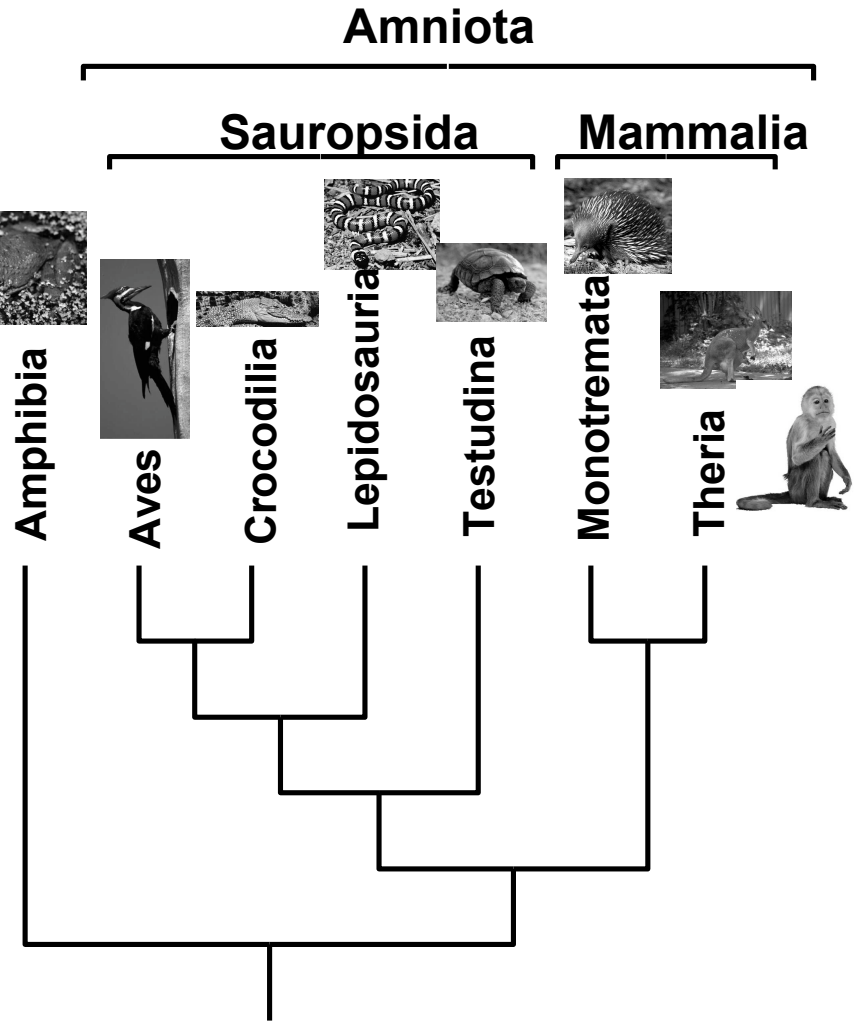


# SEQUENCE ALIGNMENT AND PHYLOGENETIC ANALYSIS

1 AGTACAGGGAGAGCTACGA  
 2 AGTACAGGGAGAGCTACGA  
 3 AGTACAAGGAGTGCTTCGA  
 4 AGTACAAGGAGTGCTTCGA  
 5 AGTACAAGGA - - - CTACGA



# Multiple Sequence Alignment

Inference of positional homology (nucleotides, amino acids)

Homology- similarity attributable to common ancestry

1	<b>AGTACAGGGAGAGCTACGA</b>
2	<b>AGTACAGGGAGAGCTACGA</b>
3	<b>AGTACAAGGAGTGCTTCGA</b>
4	<b>AGTACAAGGAGTGCTTCGA</b>
5	<b>AGTACAAGGA - - - CTACGA</b>

- Prerequisite for phylogeny analysis
- Divergence estimates
- Comparative structural evaluation (e.g., inference of gene secondary structure)

<b>GCGGCCCA</b>	<b>TCAGGTAGTT</b>	<b>GGTGG</b>
<b>GCGGCCCA</b>	<b>TCAGGTAGTT</b>	<b>GGTGG</b>
<b>GCGTTCCA</b>	<b>TCAGCTGGTT</b>	<b>GGTGG</b>
<b>GCGTCCCA</b>	<b>TCAGCTAGTT</b>	<b>GGTGG</b>
<b>GCGGCGCA</b>	<b>TTAGCTAGTT</b>	<b>GGTGA</b>
<b>*****</b>	<b>*****</b>	<b>*****</b>

<b>TTGACATG</b>	<b>CCGGGG---A</b>	<b>AACCG</b>
<b>TTGACATG</b>	<b>CCGGTG--GT</b>	<b>AAGCC</b>
<b>TTGACATG</b>	<b>-CTAGG---A</b>	<b>ACGCG</b>
<b>TTGACATG</b>	<b>-CTAGGGAAC</b>	<b>ACGCG</b>
<b>TTGACATC</b>	<b>-CTCTG---A</b>	<b>ACGCG</b>
<b>*****</b>	<b>??????????</b>	<b>*****</b>

# SUBSTITUTIONS

ACTGGACGTAAC ← Correct alignment  
ACTGAACGCAAC

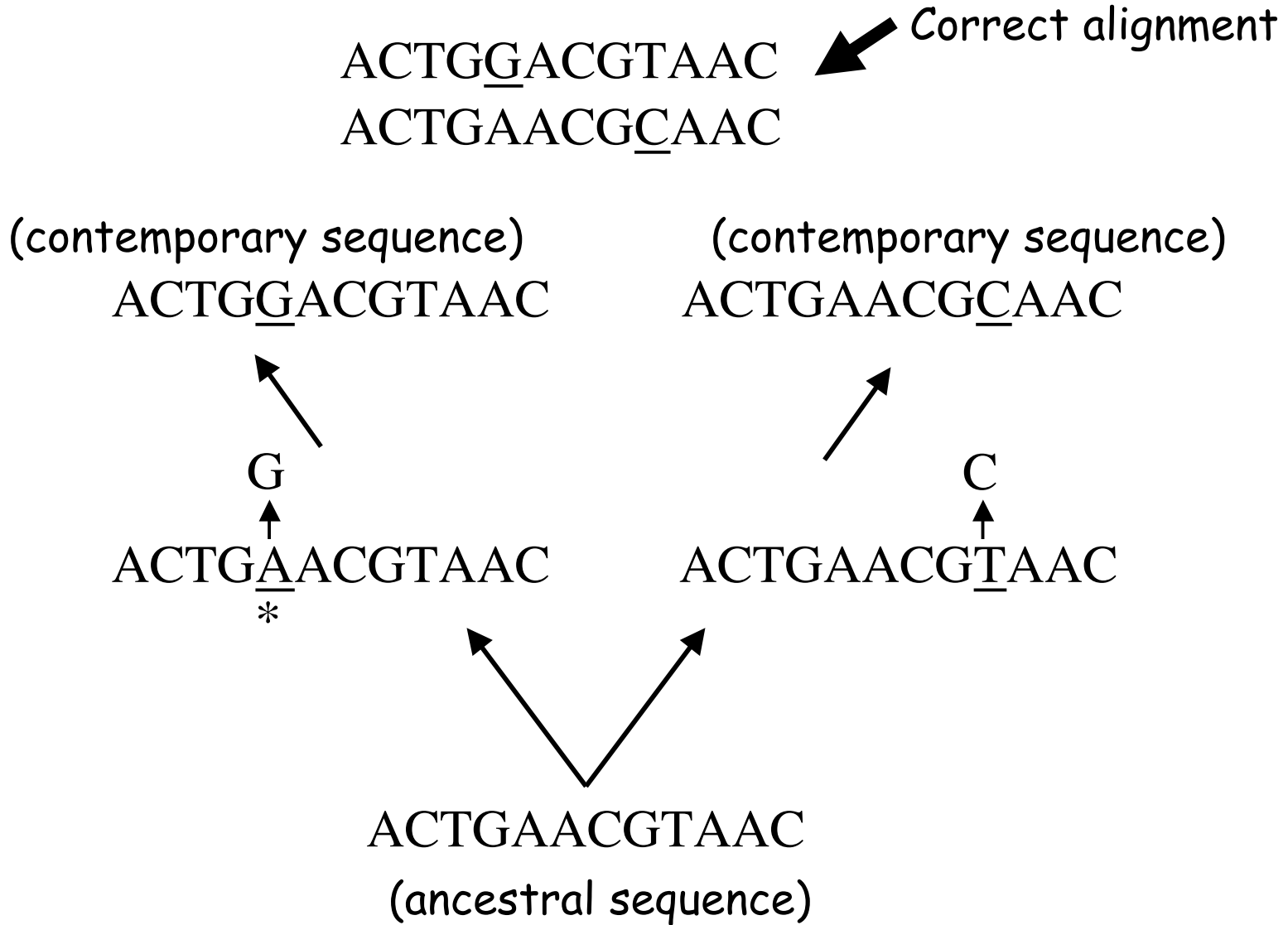
(contemporary sequence)  
ACTGGACGTAAC

(contemporary sequence)  
ACTGAACGCAAC

G  
↑  
ACTGAACGTAAC  
\*

C  
↑  
ACTGAACGTAAC

ACTGAACGTAAC  
(ancestral sequence)





## Substitutions vs Gaps

	Sub. Cost	Gap cost
1 ACTTCCGAATTG- GCT 	0(1)	5(1) = 5
2 ACT- - CGA - TT- GC - CT	0(1)	5(2) = 10

VS

1 ACTTCCGAATTGGCT       * *         *	3(1)	2(1) = 5
2 ACTC - - - GATT - GCCT	3(1)	2(2) = 7

Maximize similarity while minimizing gaps

After Siddall

Amino Acid sequence alignments:  
similar considerations

# Multiple Alignment Methods

- Manual
- Automatic

Manual- minimal length variation, refinement of automatic alignment

# Multiple alignment programs

## Progressive alignments

E.g., Clustal X, Pileup

- 1-pairwise similarities between sequences calculated
- 2-guide tree constructed (UPGMA, Neighbor-Joining, etc.)
- 3-sequences (pairs) aligned (N-W) in order of decreasing similarity as determined by guide tree

Principal user options:

- gap insertion weights (gap opening penalty; *GOP*)
- gap length weights (gap extension penalty; *GEP*)

Arbitrary!

## PAIRWISE DISTANCE MATRIX

**A** CGAAGTCATGCTAAAGTA

**B** CGATATCAGGTAAAGAT

    \*\*      \*  \*         \*\*

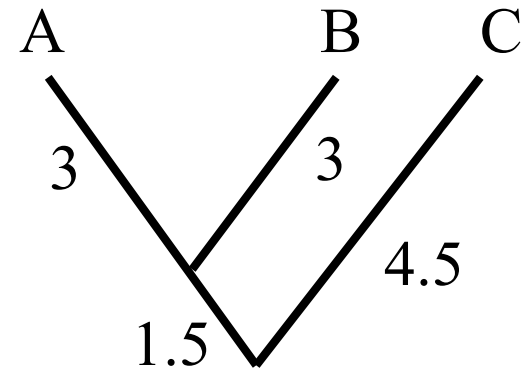
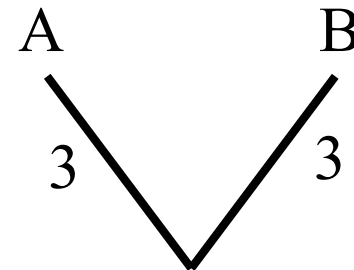
	A	B	C	D
A	-	6	8	12
B		-	10	14
C			-	16
D				-

# GUIDE TREE CONSTRUCTION ( e.g., UPGMA, etc.)

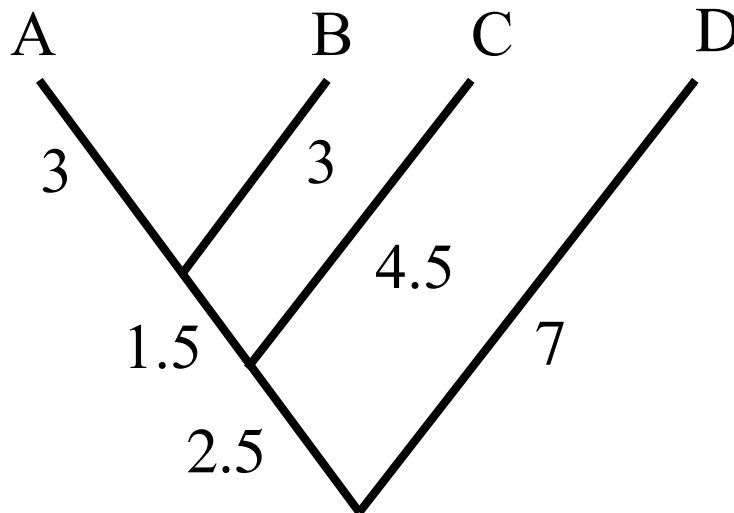
	A	B	C	D
A	-	6	8	12
B		-	10	14
C			-	16
D				-

A-C = 8  
 B-C = 10  
 Average =  $18/2 = 9$

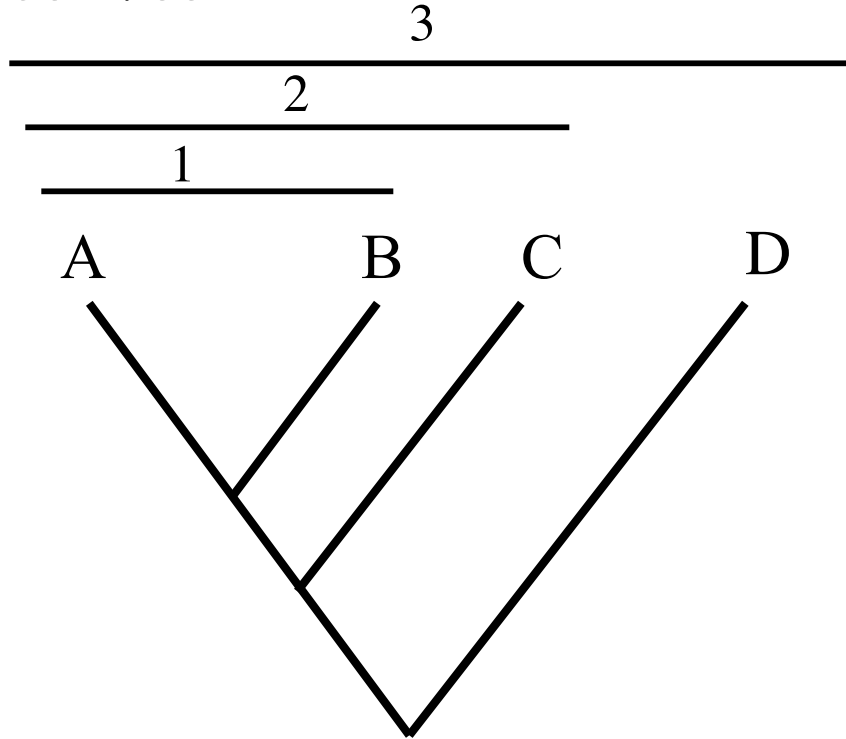
A-D = 12  
 B-D = 14  
 Average =  $26/2 = 13$



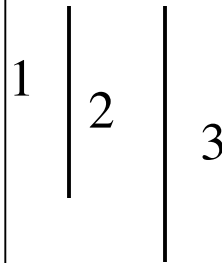
A-D = 12  
 B-D = 14  
 C-D = 16  
 $42/3 = 14$

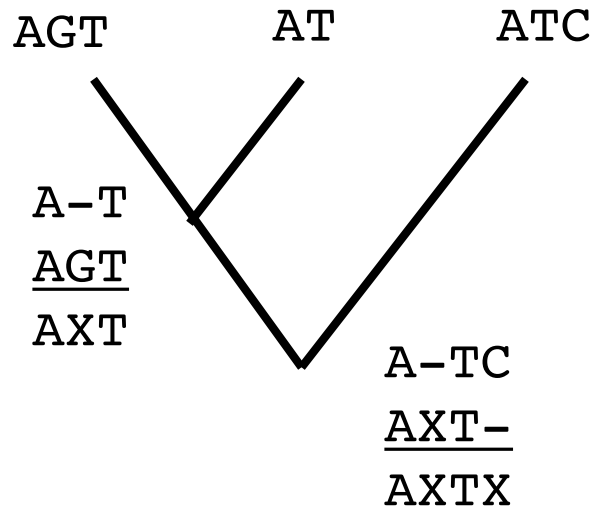


# Guide Tree



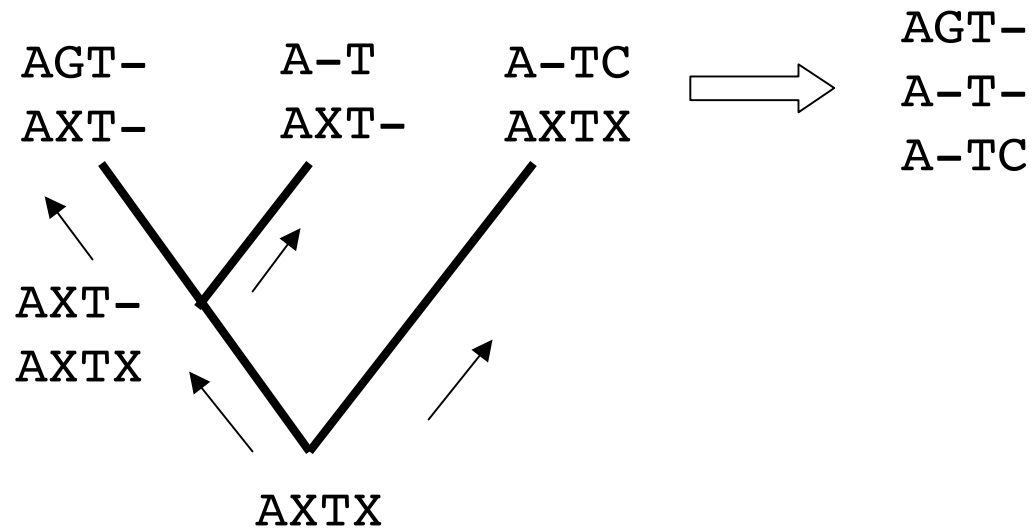
**AGTACAAGGAGTGCTTCGA**  
**AGTACAAGGAGTGCTTCGA**  
**AGTACAAGGA- - - CTACGA**  
**AGTACAGGGAGAGCTACGA**





Alignment fixed between pairs  
 "Once a gap, always a gap."  
 Feng and Doolittle, 1987.

Aloysius et al. 2000



# Refinement

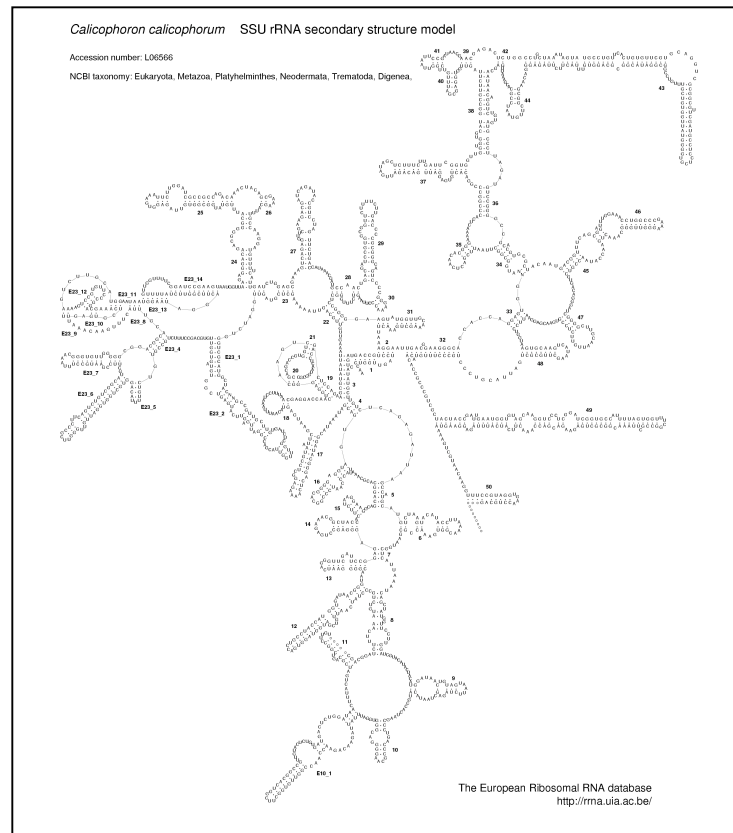
How?

1	<b>GTACAGGGAGAGCTACGA</b>
2	<b>GTACAGGGAGAGCTACGA</b>
3	<b>GTACAAGGAGTGCTTCGA</b>
4	<b>GTACAAGGAGTGCTTCGA</b>
5	<b>GTACAAGGA - - - CTACGA</b>

# Secondary Structure Determination

```

<----- (-----HELIX 19-----)
<----- (2222222-000000-111111-00000-111111-0000-2222222
Thermus ruber      UCCGAUGC-UAAAGA-CCGAAG=CUCAA=CUUCGG=GGGU=GCGUUGGA
Th. thermophilus  UCCCAUGU-GAAAGA-CCACGG=CUCAA=CCGUGG=GGGA=GCGUGGGA
E.coli            UCAGAUGU-GAAAU-CCCGGG=CUCAA=CCUGGG=AAU=GCAUCUGA
Ancystr. nidulans UCUGUUGU-CAAAGC-GUGGGG=CUCAA=CCUCAU=ACAG=GCAAUGGA
B.subtilis       UCUGAUGU-GAAAGC-CCCCGG=CUCAA=CCGGGG=AGGG=UCAUUGGA
Chl.aurantiacus  UCGGCGCU-GAAAGC-GCCCCG=CUUAA=CGGGGC=GAGG=CGCGCCGA
match            **          ***          *  *  *  *          **
  
```



# Divergence Estimates

Characters converted to evolutionary distances.

1 AAGTCATGCT  
2 AAATCAGGCT  
3 CAGACAGTCA  
4 CACTCCCA

1 AAGTCATGCT 2/10 = 0.2  
2 AAATCAGGCT  
\* \*

	1	2	3	4
1	-			
2	0.2	-		
3	0.5	0.5	-	
4	0.7	0.6	0.3	-

# PHYLOGENETIC ANALYSIS

## PAUP\*

### Phylogenetic Analysis Using Parsimony and Other Methods

Reconstruction of evolutionary relationships (Phylogeny)

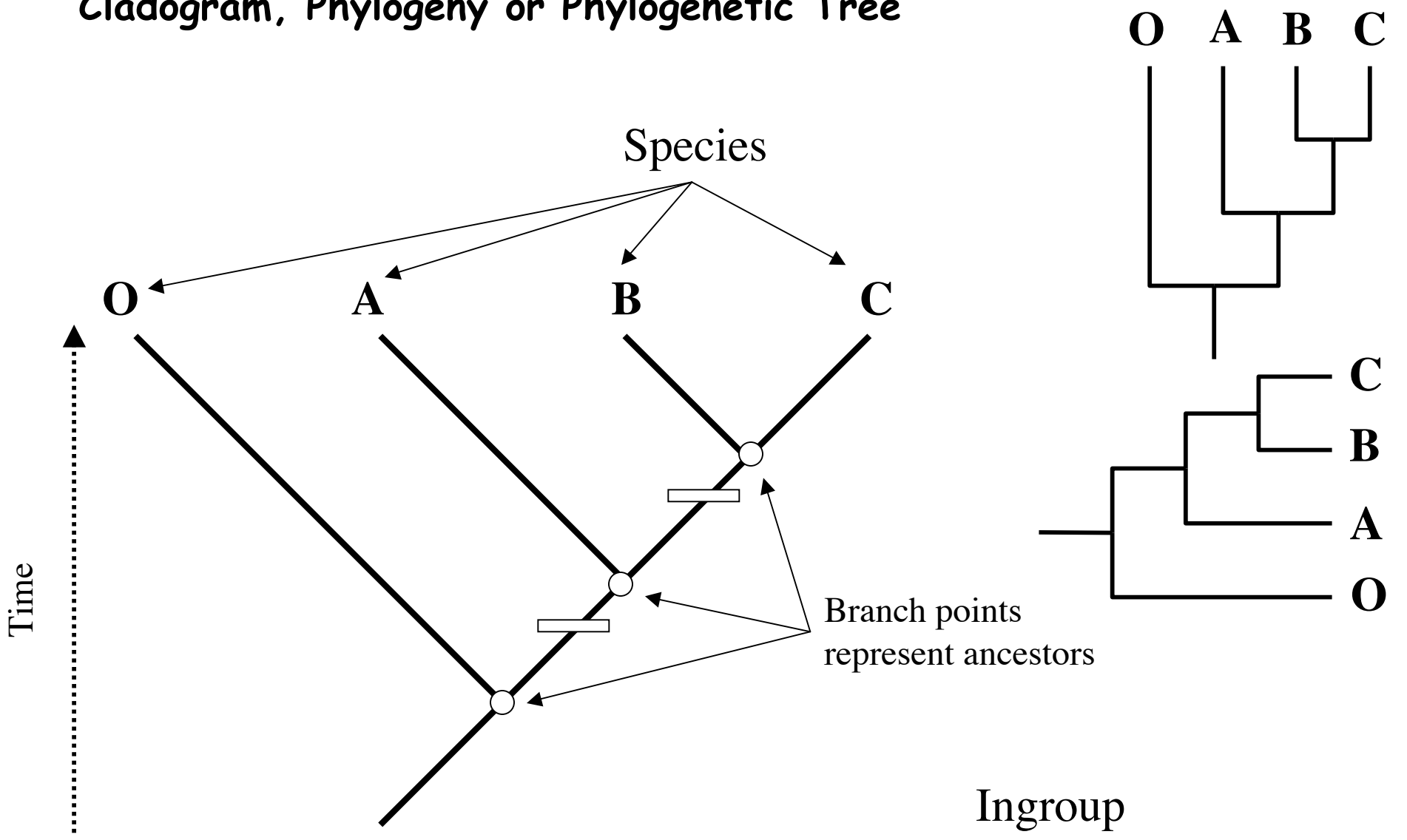
- Character evolution
- Classification

Trace origin of diseases

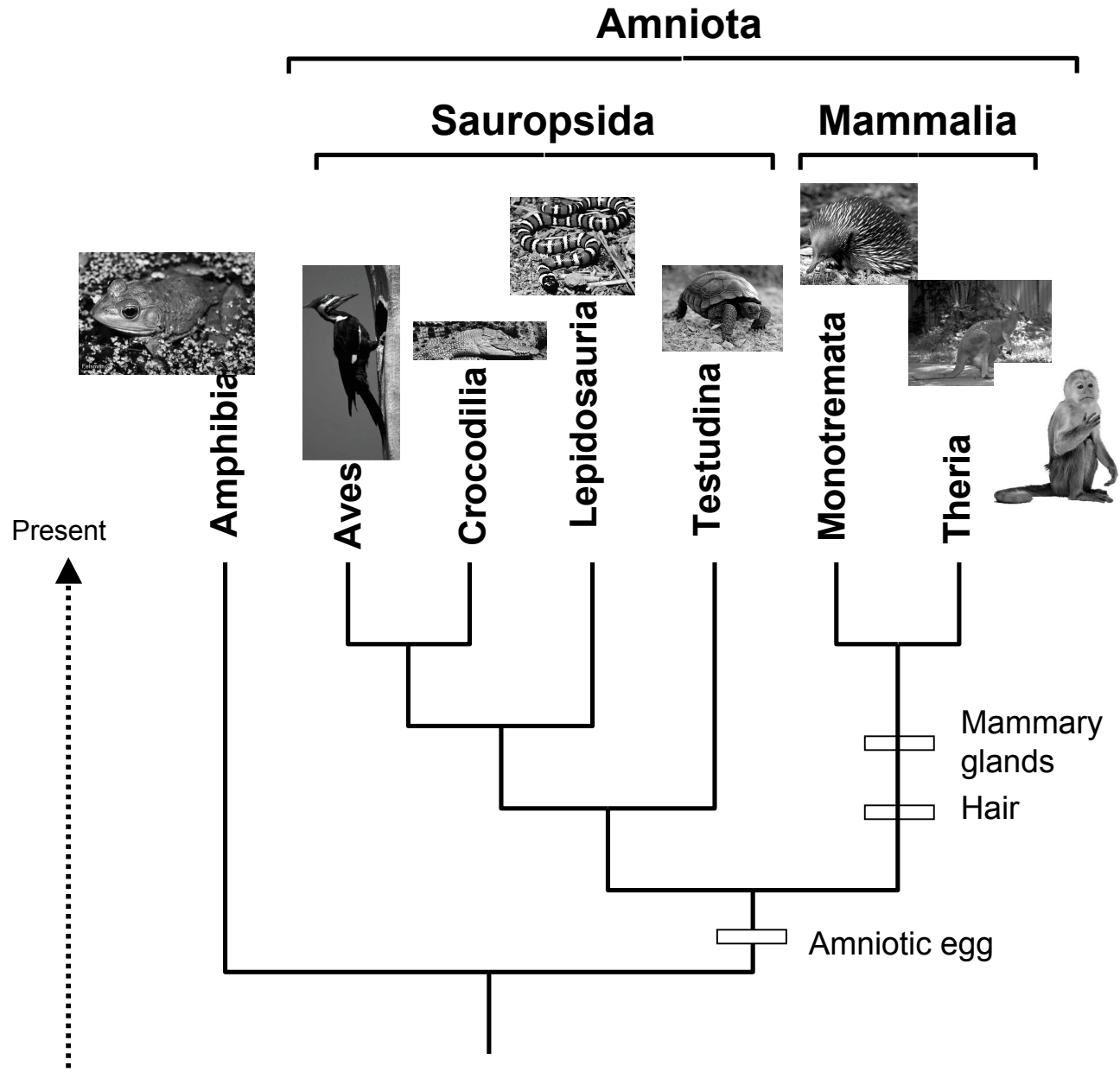
Forensic applications

- Character Methods
- Distance Methods

# Cladogram, Phylogeny or Phylogenetic Tree

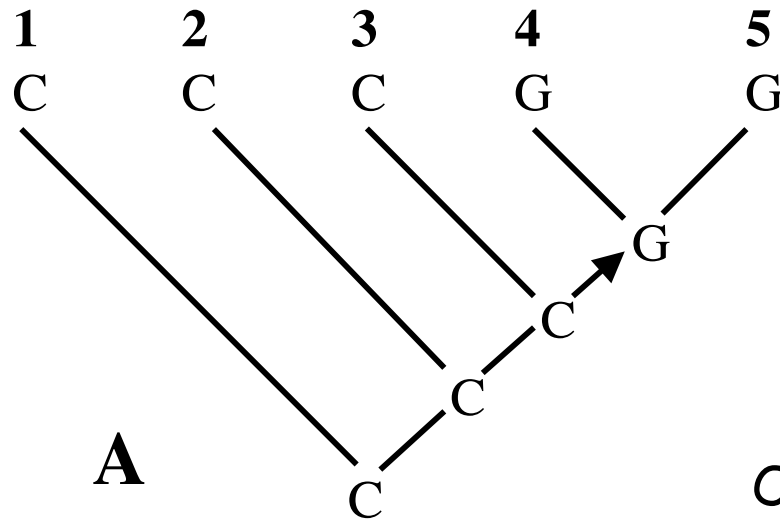


An hypothesis of evolutionary relationships!

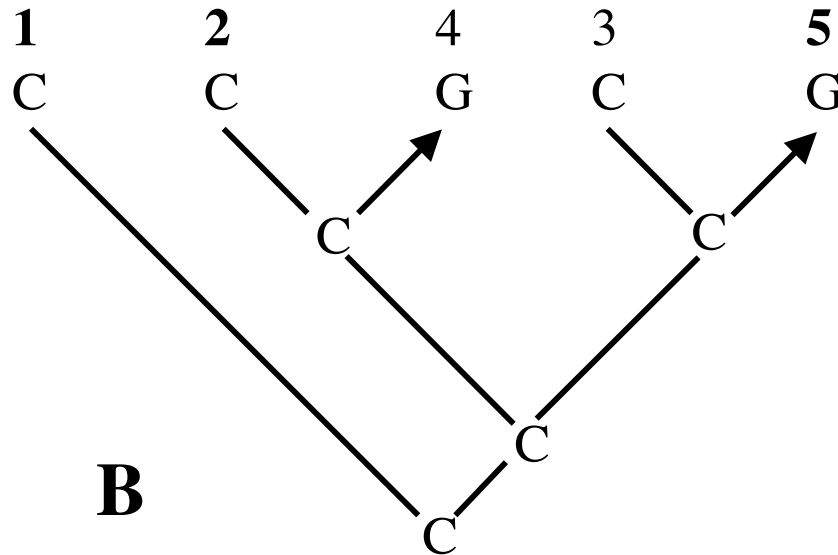


# Parsimony

<b>1</b>	A	C	C	C	T
<b>2</b>	A	A	C	C	T
<b>3</b>	A	C	C	C	T
<b>4</b>	A	T	G	C	A
<b>5</b>	A	G	G	C	A



Optimal tree?



# TREE SEARCHING

Exact searches:

Exhaustive

Branch and Bound

Approximate searches:

Heuristic search

Trees:

4 taxa = 3 trees

5 taxa = 15 trees

10 taxa = 2,027,025

# Branch and Bound Search

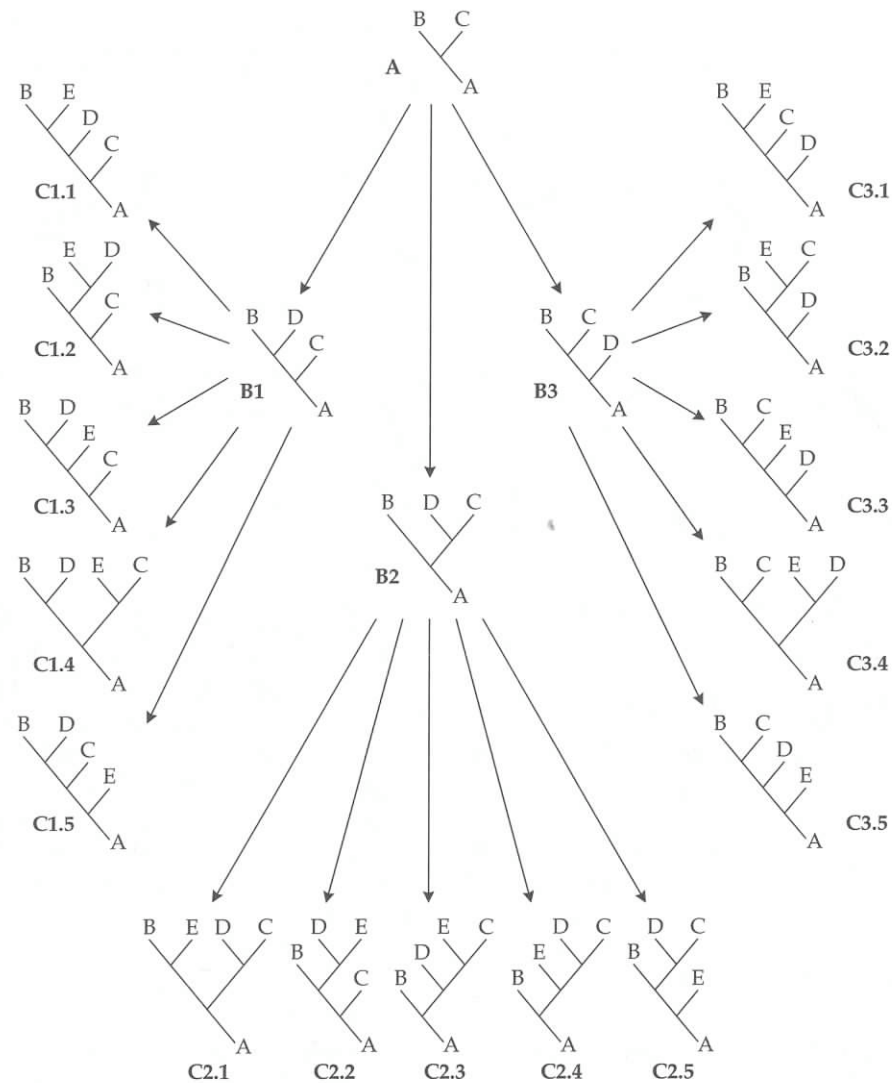


Figure 24 Search tree for branch-and-bound algorithm (see text).

From Swofford et al. 1996

# Heuristic Search

Approximate

1) Generate starting trees:  
Multiple Options

2) Improving on starting trees

Branch swapping

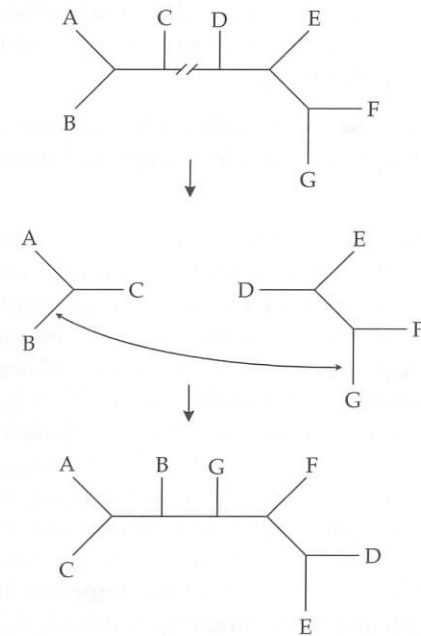


Figure 28 Branch swapping by tree bisection and reconnection. The tree is bisected along a branch, yielding two disjoint subtrees. The subtrees are then reconnected by joining a pair of branches, one from each subtree. All possible bisections and pairwise reconnections are evaluated.

From Swofford et al. 1996