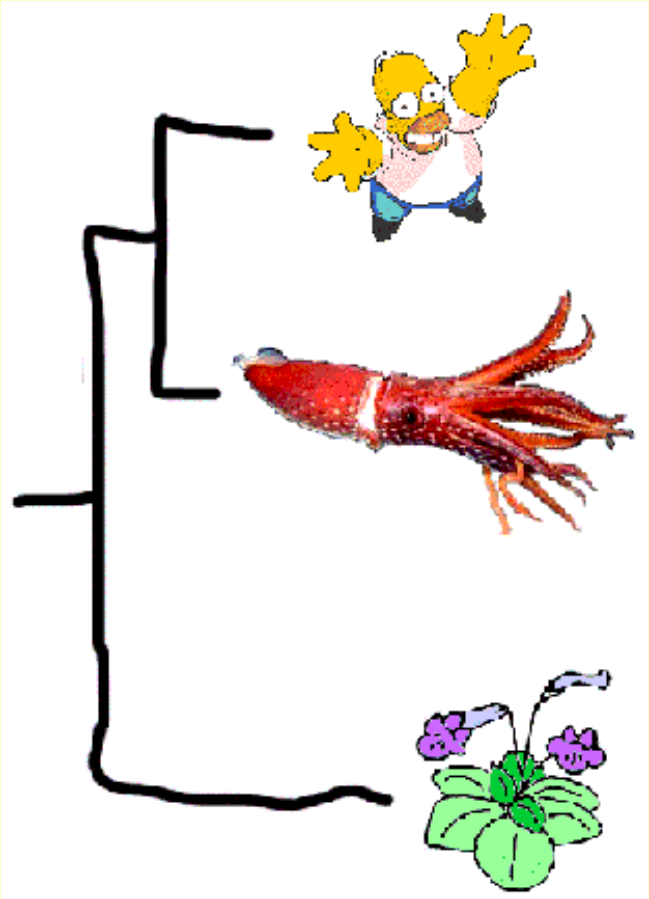
Computational Molecular Biology

Biochem 218 – BioMedical Informatics 231

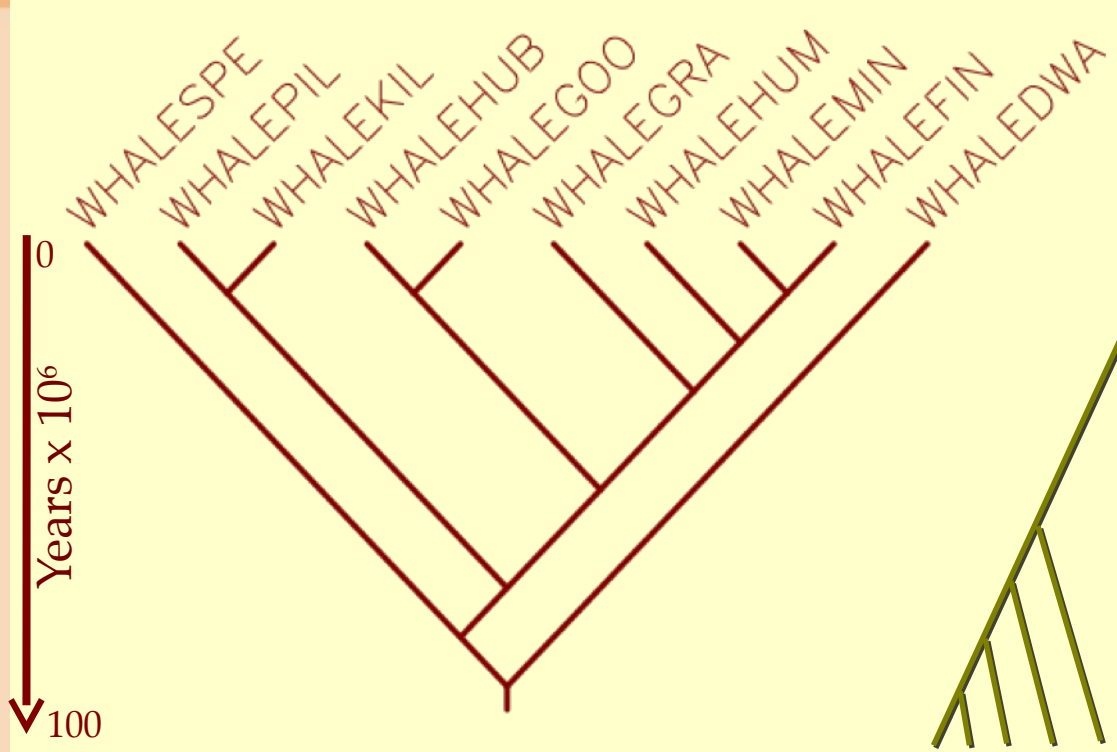
http://biochem218.stanford.edu/

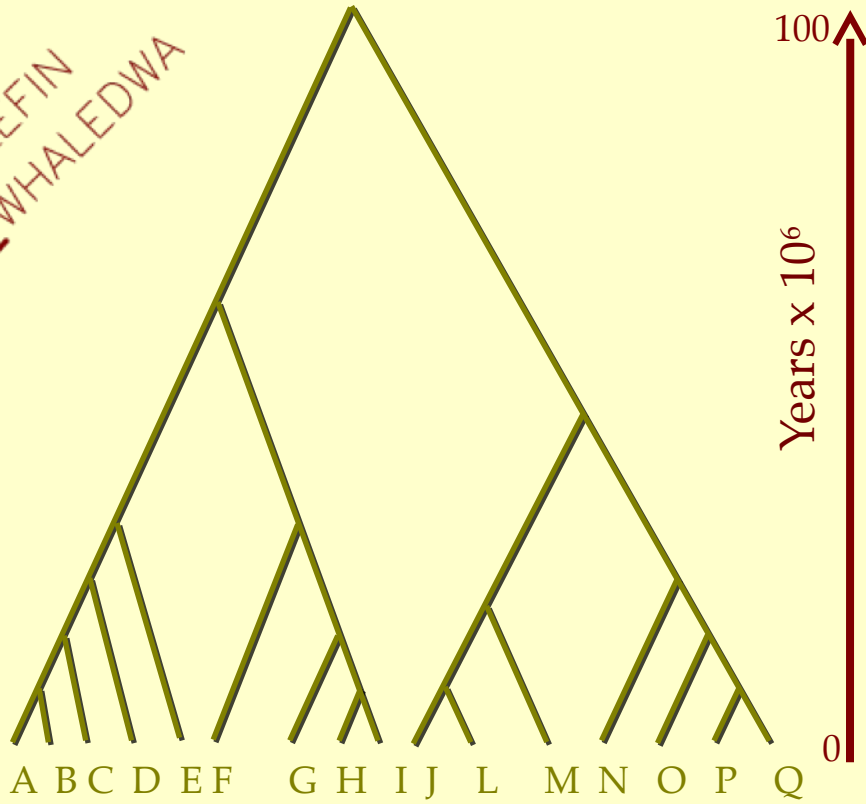
Phylogenies



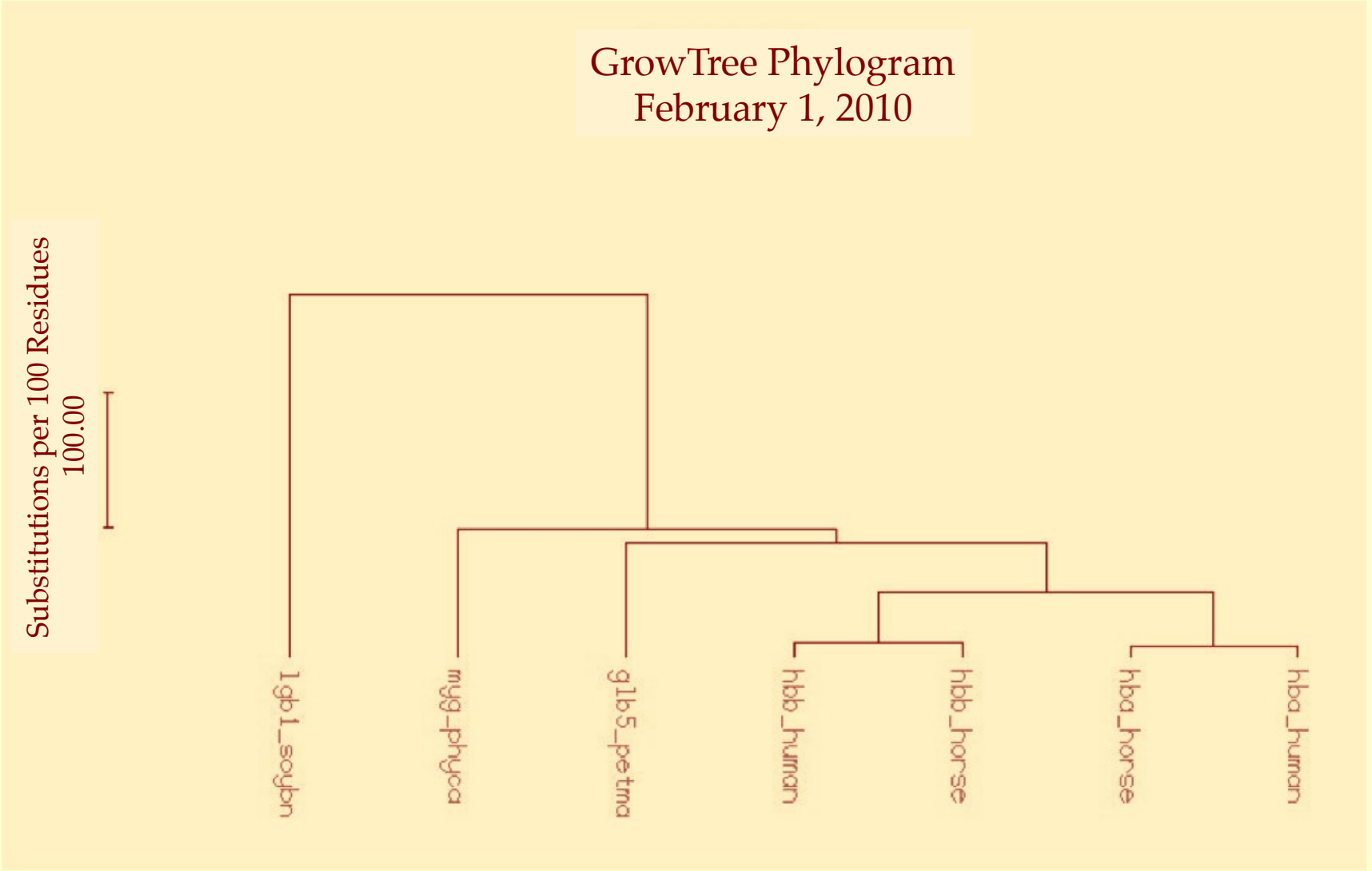
Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)

Cladogram Representation of Phylogenies

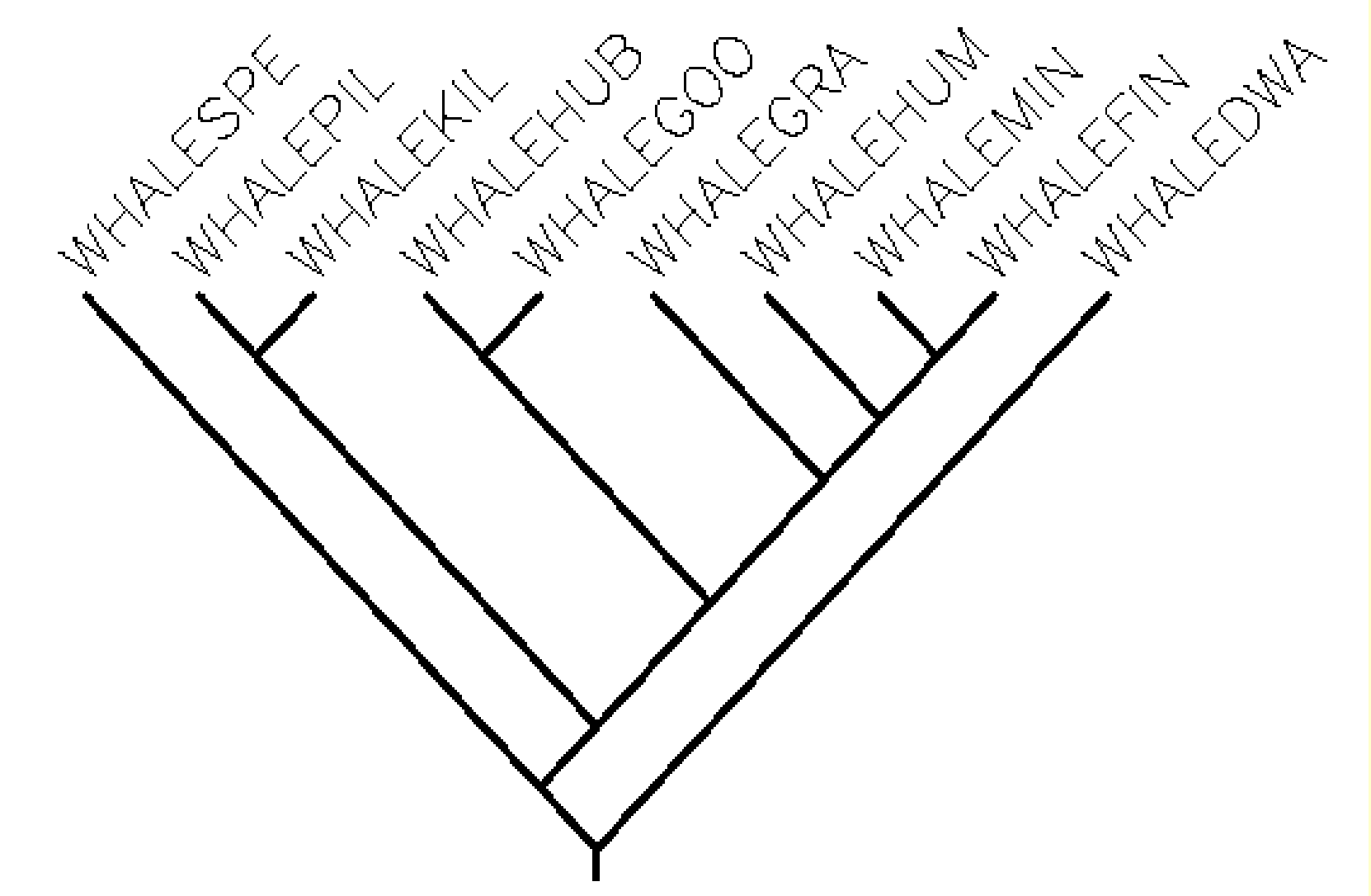




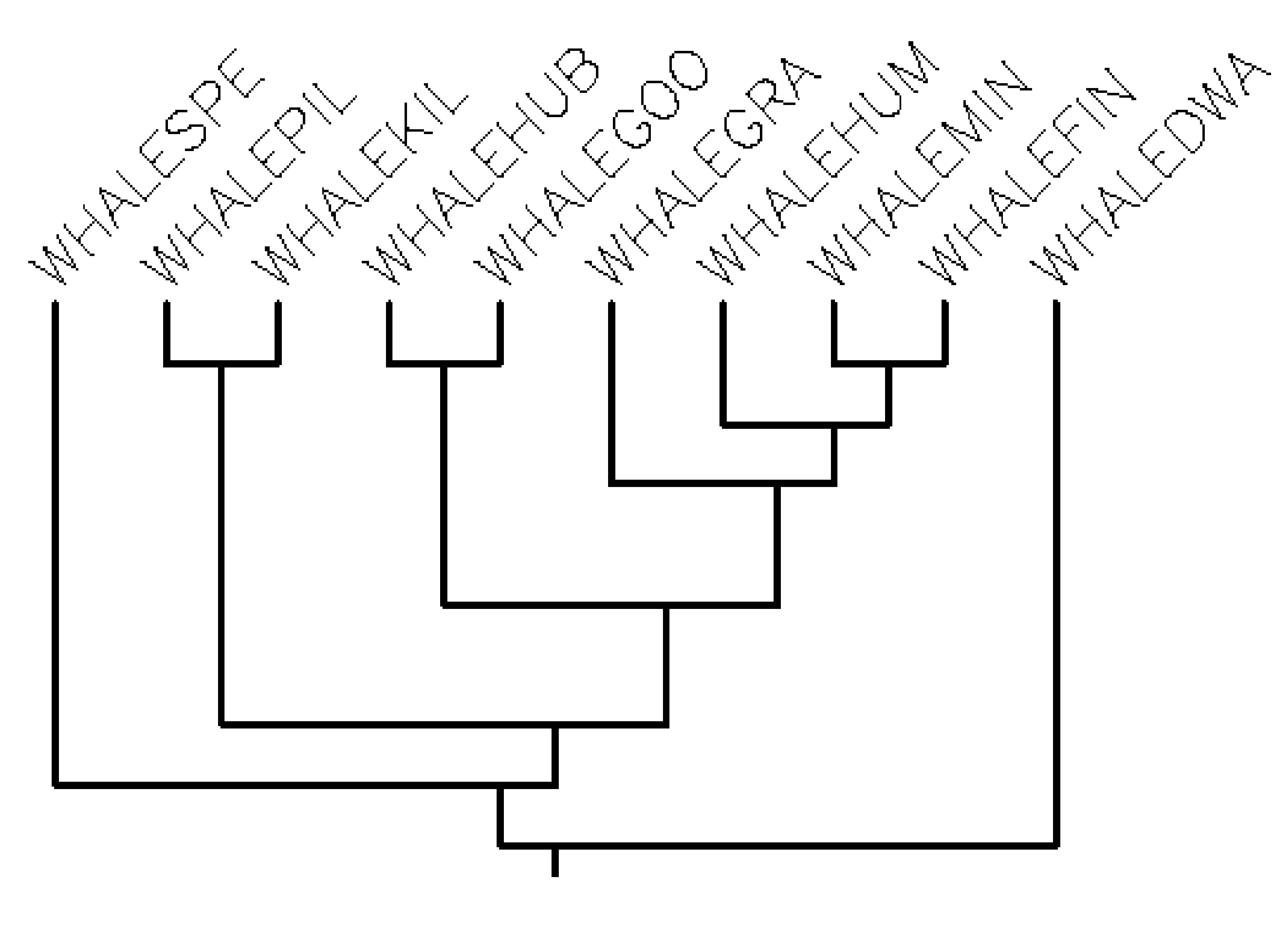
Dendrogram Representation of Phylogenies



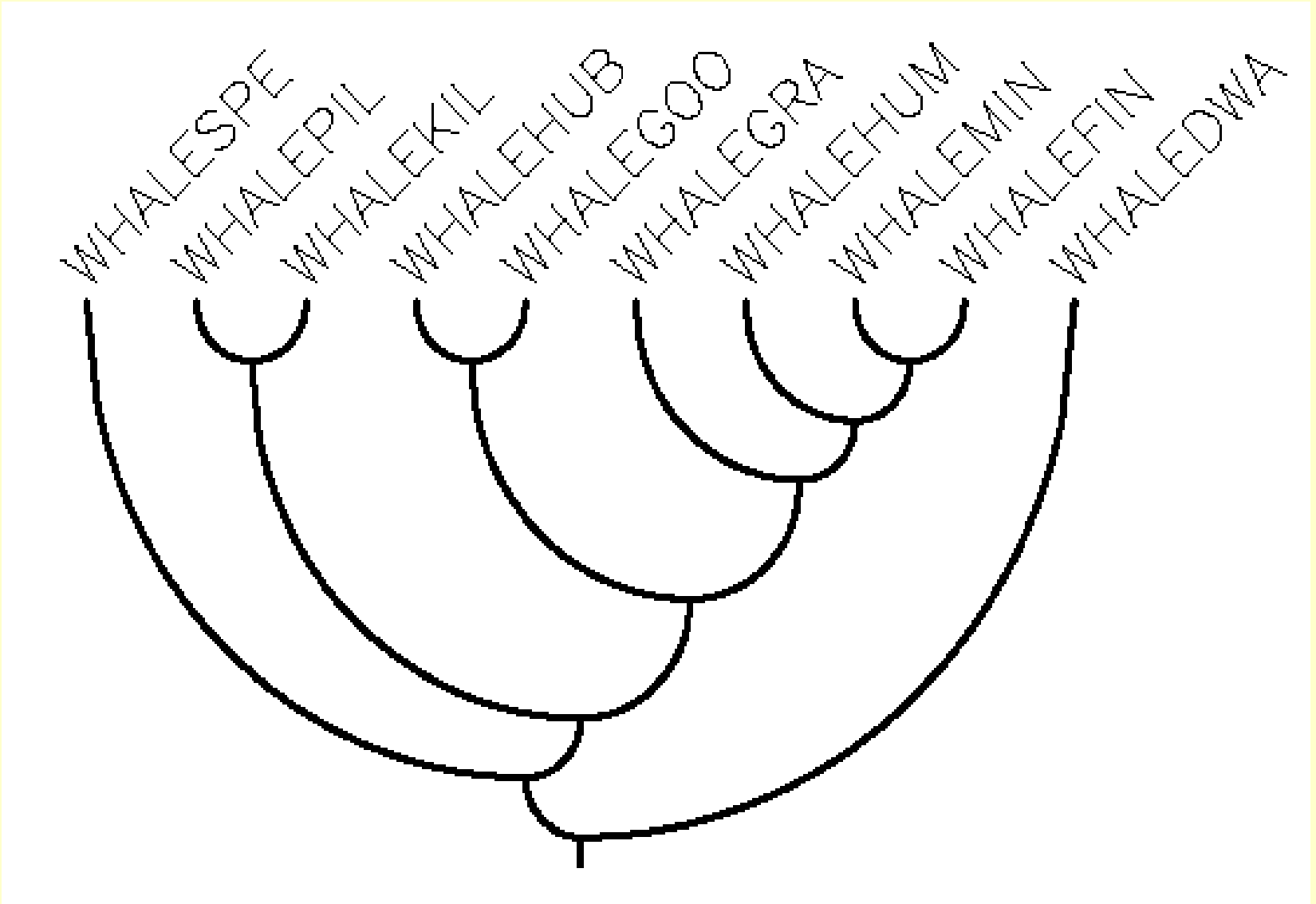
Cladogram



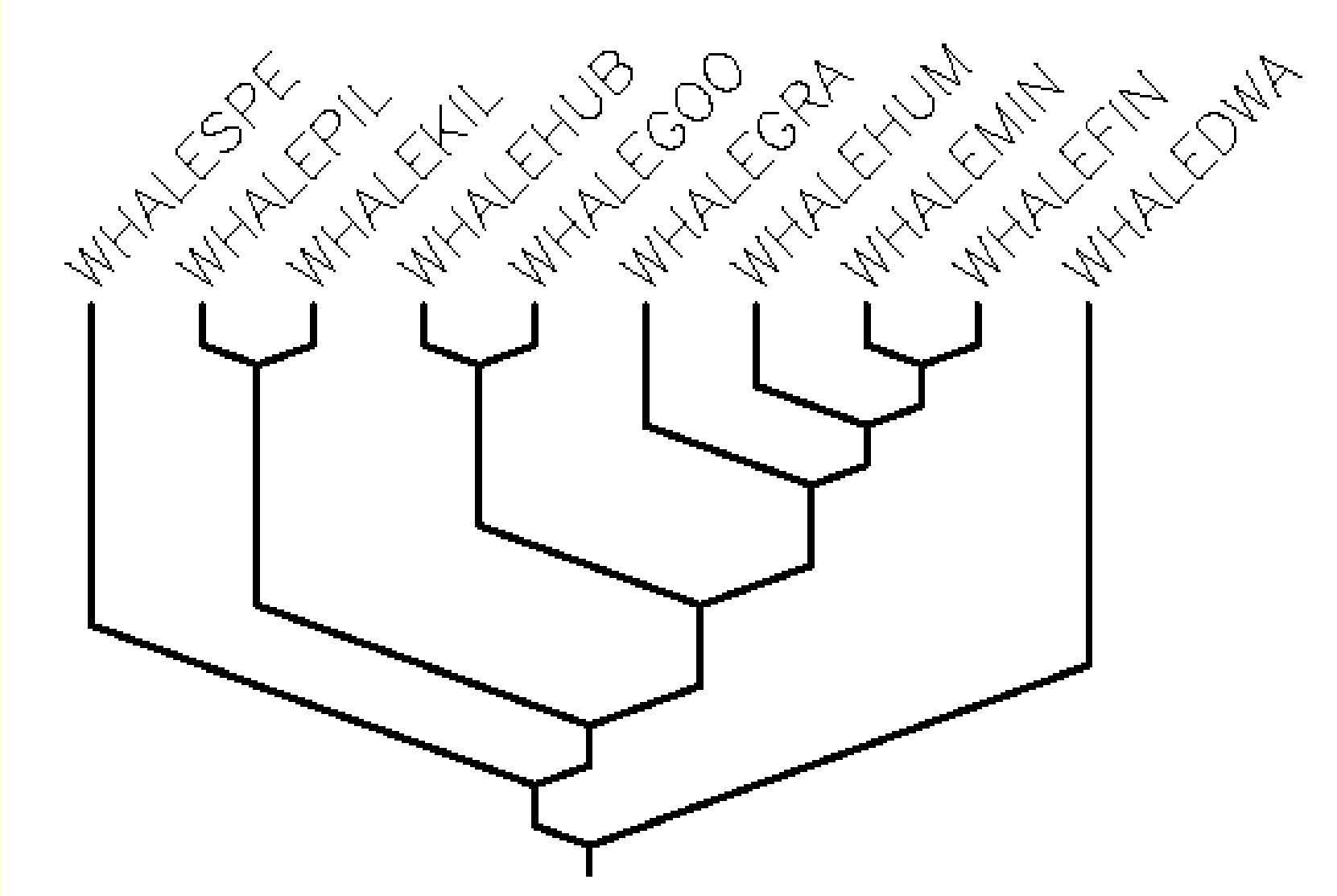
Phenogram



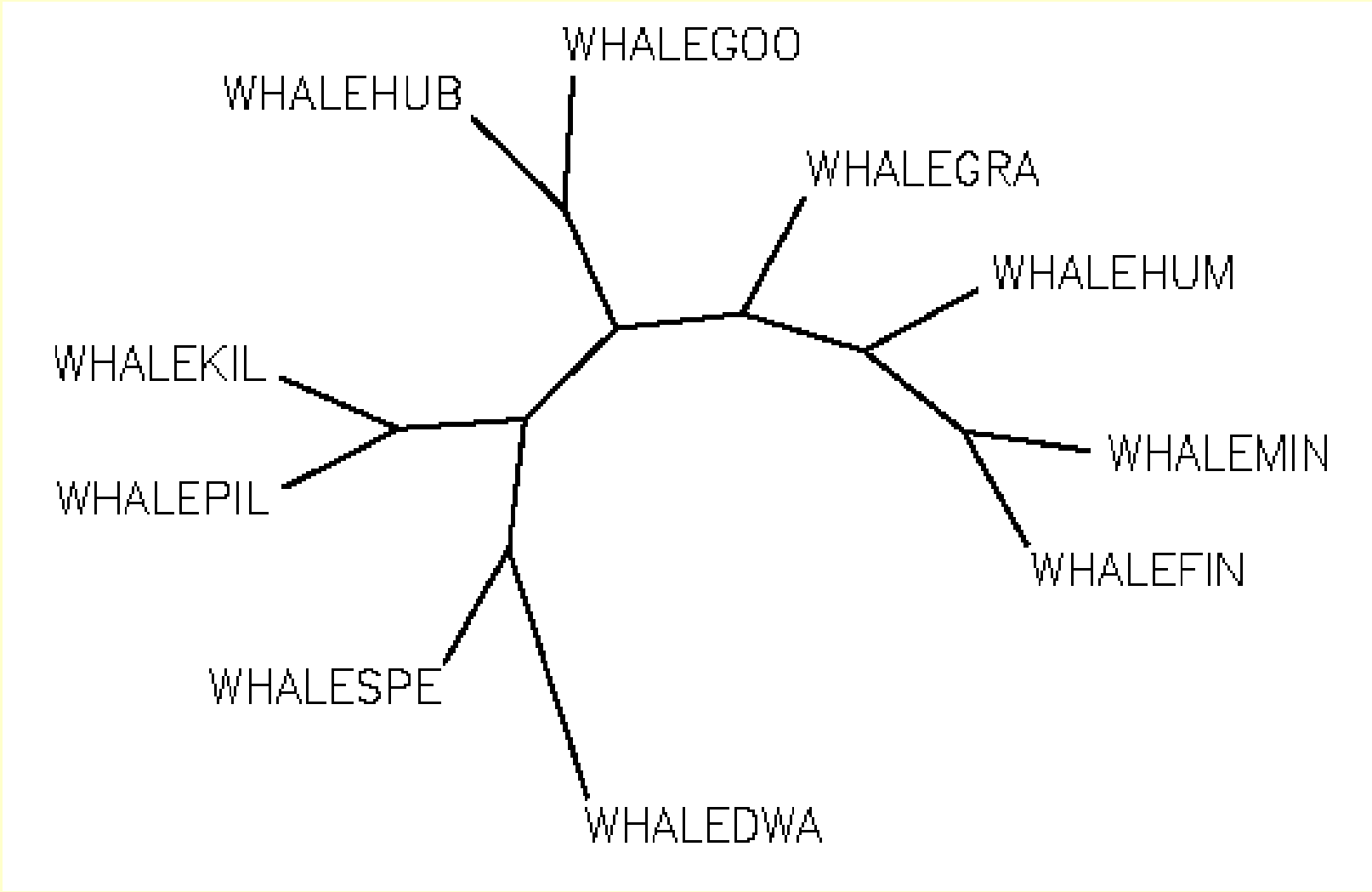
Curve-O-Gram



Eurogram

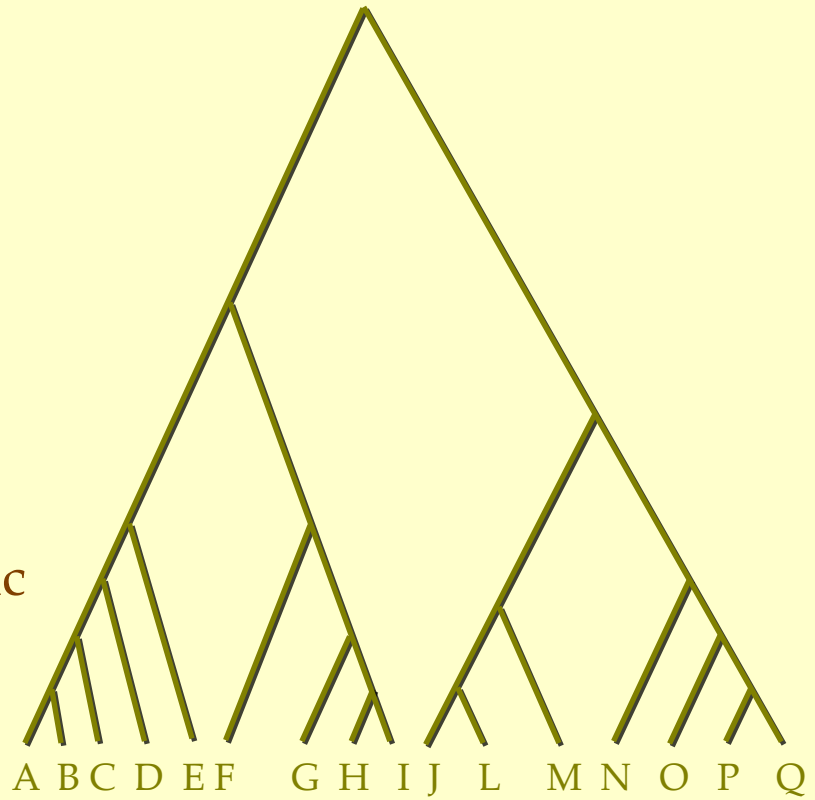


RadialGram



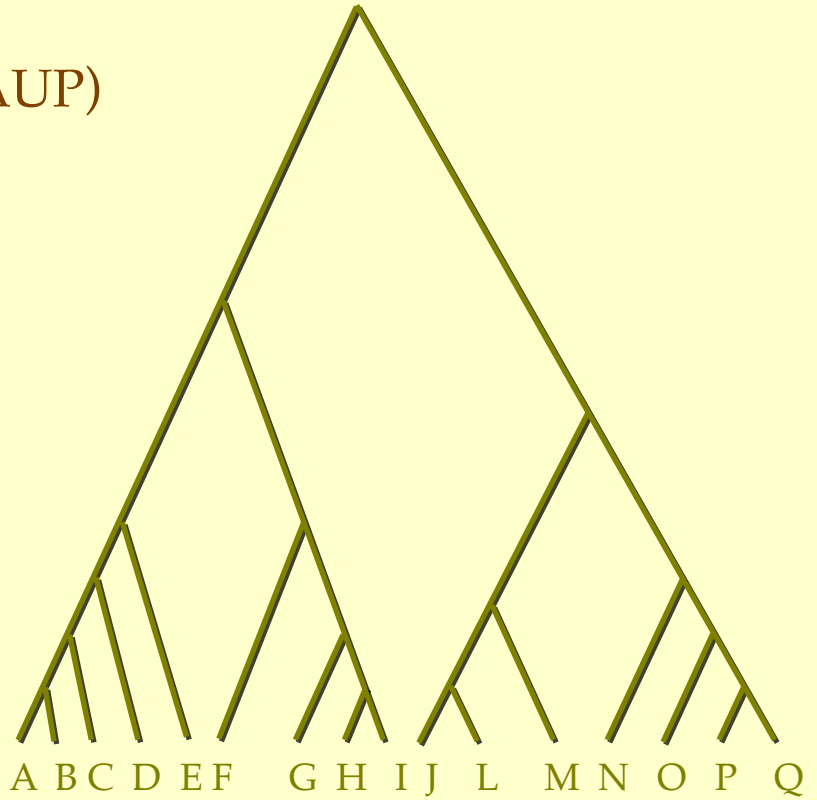
Methods for Determining Phylogenies

- Parsimony (character based)
  - Assigns mutations to branches
  - Minimize number of edits
  - Topology maximizes similarity of neighboring leaves
- Distance methods
  - Branch lengths = D(i,j)/2 for sequences i, j
  - Distances must be at least metric
  - Distances can reflect time or edits
  - Distance must be relatively constant per unit branch length



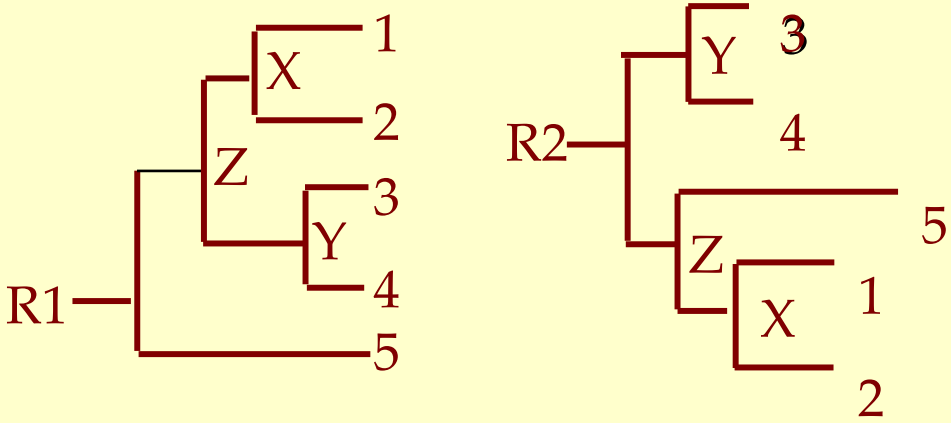
Methods for Determining Phylogenies

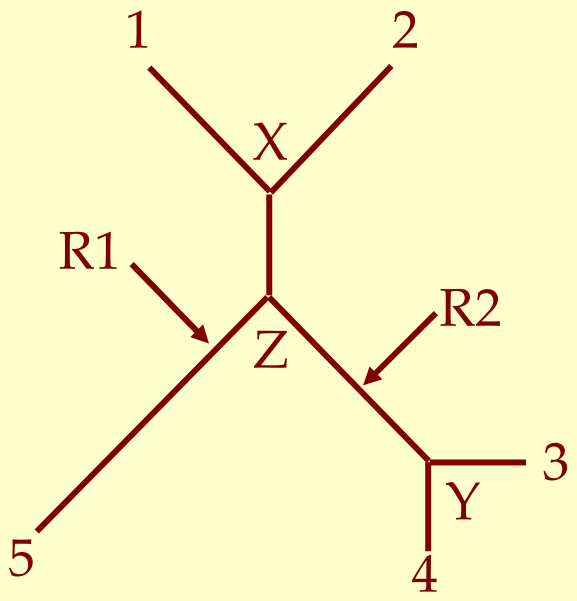
- Parsimony
  - Minimum mutation (Fitch, PAUP)
  - Minimal length encoding
- Probabilistic
  - Branch and Bound
  - Maximum likelihood
- Distance methods
  - Ultrametric Trees
  - Additive Trees
  - UPGMA
  - Neighbor Joining



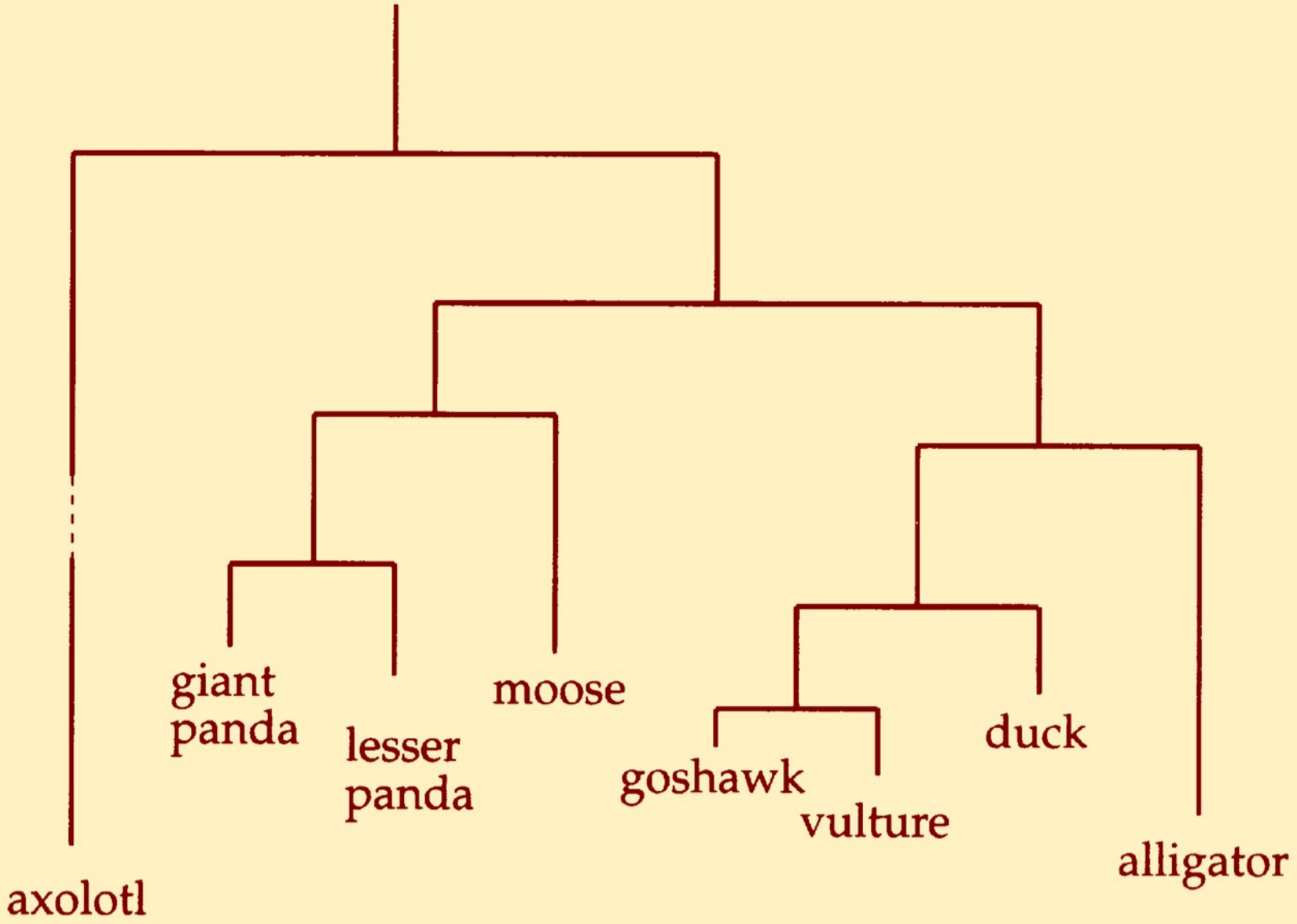
Properties of Trees

- Rooted or Unrooted
- Nodes and Branches
  - Internal Nodes
  - External Nodes - leaves
- Operational Taxonomic Units
- Outgroups
- Topology
- One path/pair
- Distances

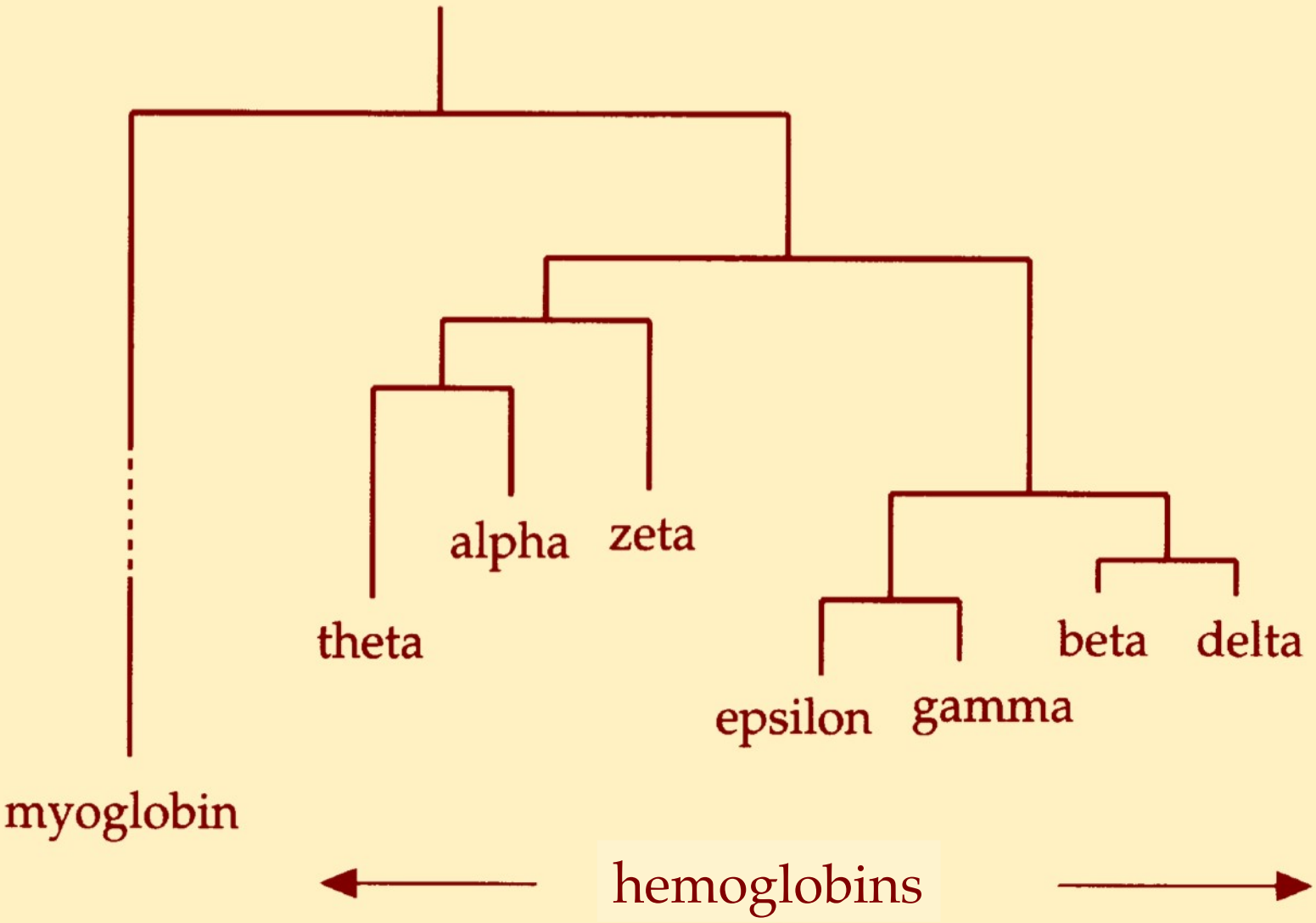




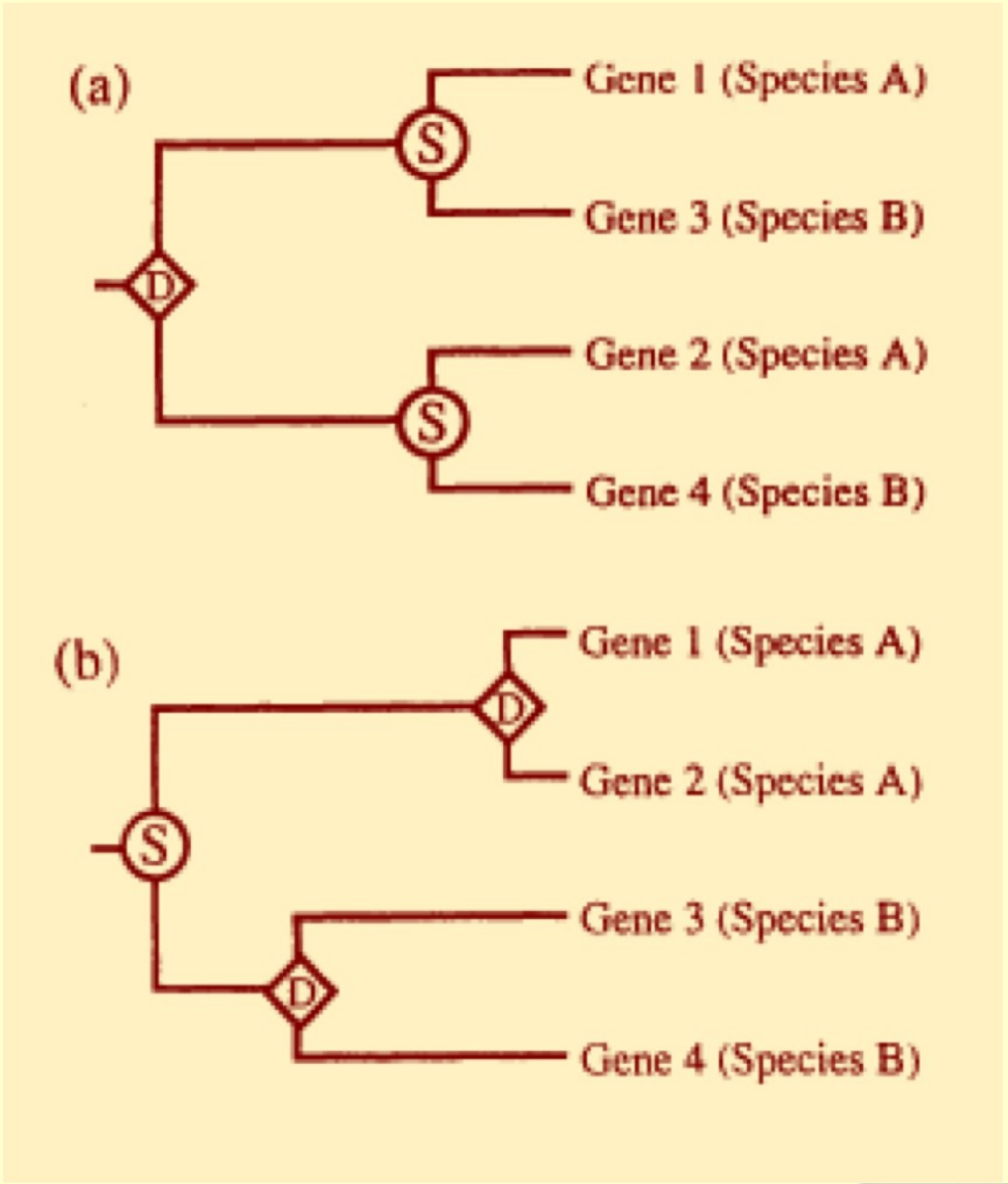
Orthologous Evolution



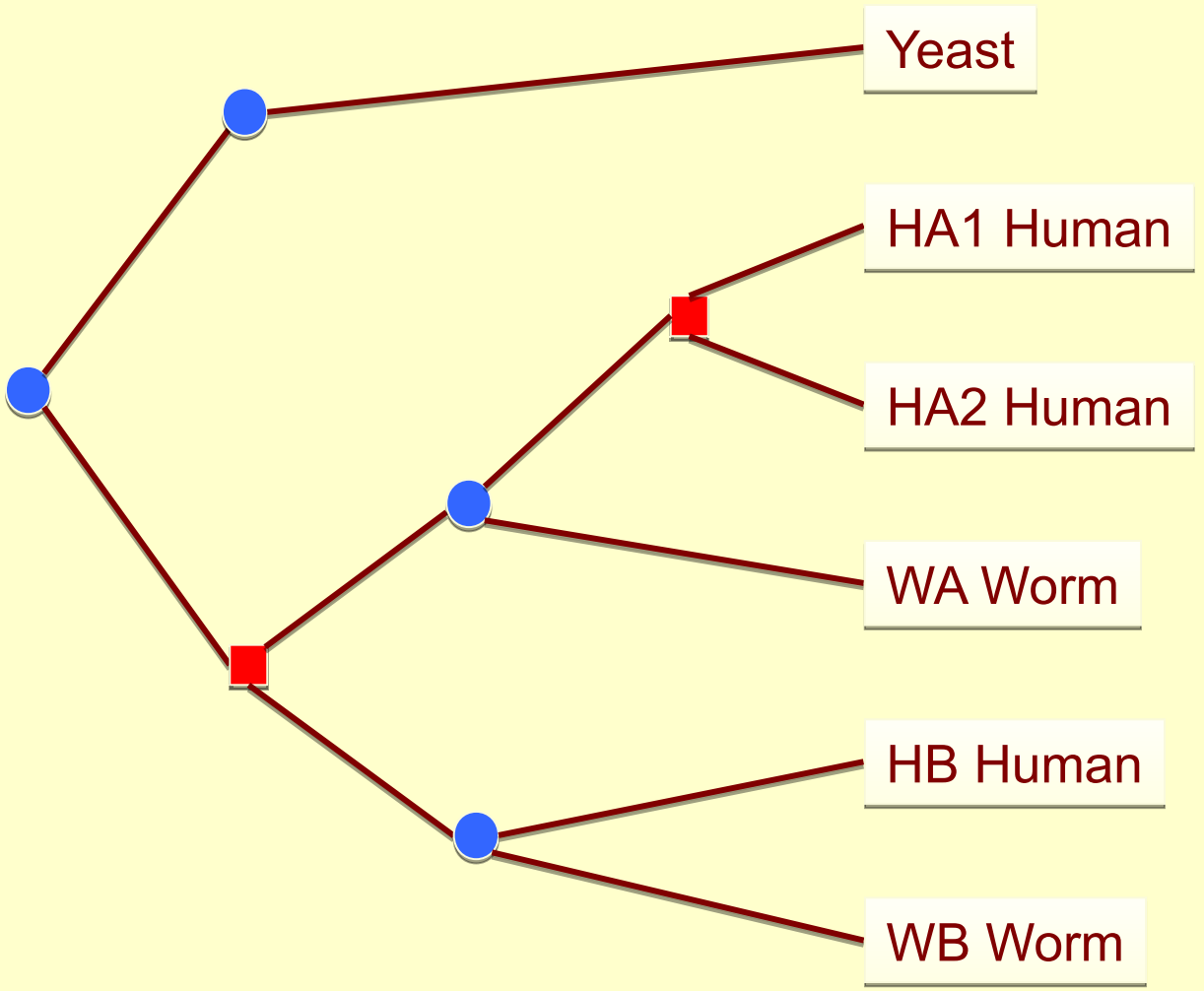
Paralogous Evolution



Challenges Making Trees: Gene Duplication versus Speciation



Orthology and Paralogy



Orthologs: *Derived by speciation*

Paralogs: *Gene Duplications*

Thanks to Seraphim Batzoglou

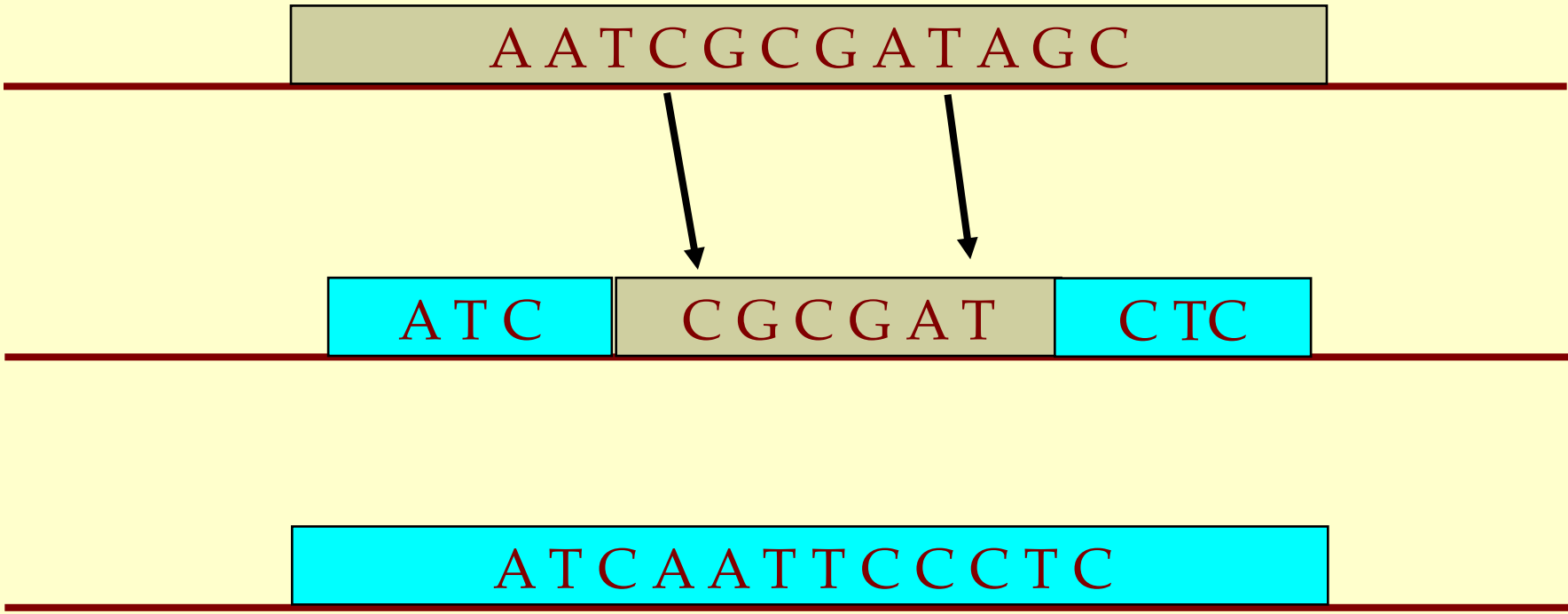
Gene Conversion

A A T C G C G A T A G C

A T C A A T T C C C T C

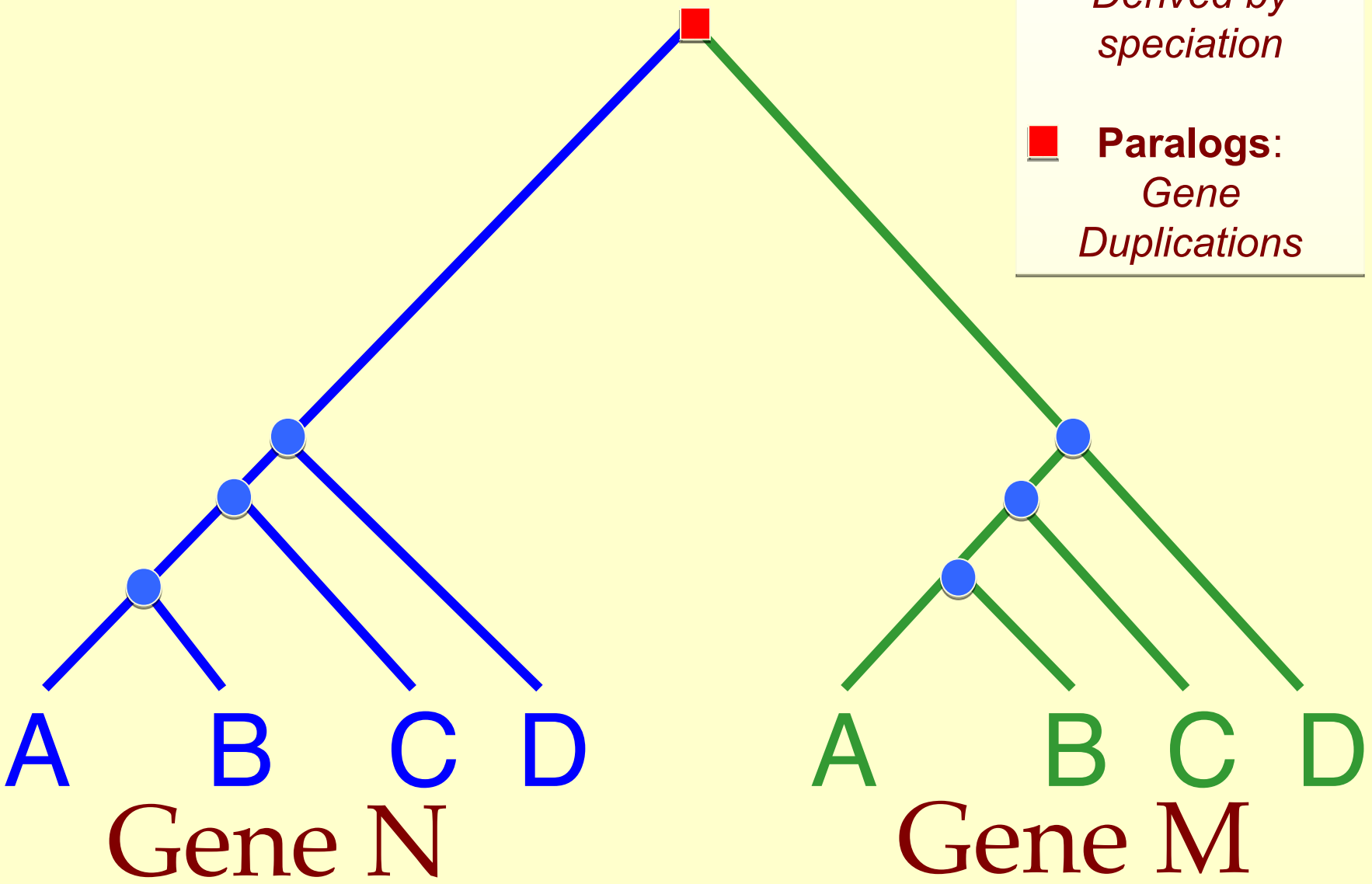
Thanks to Maryellen Ruvolo

Gene Conversion



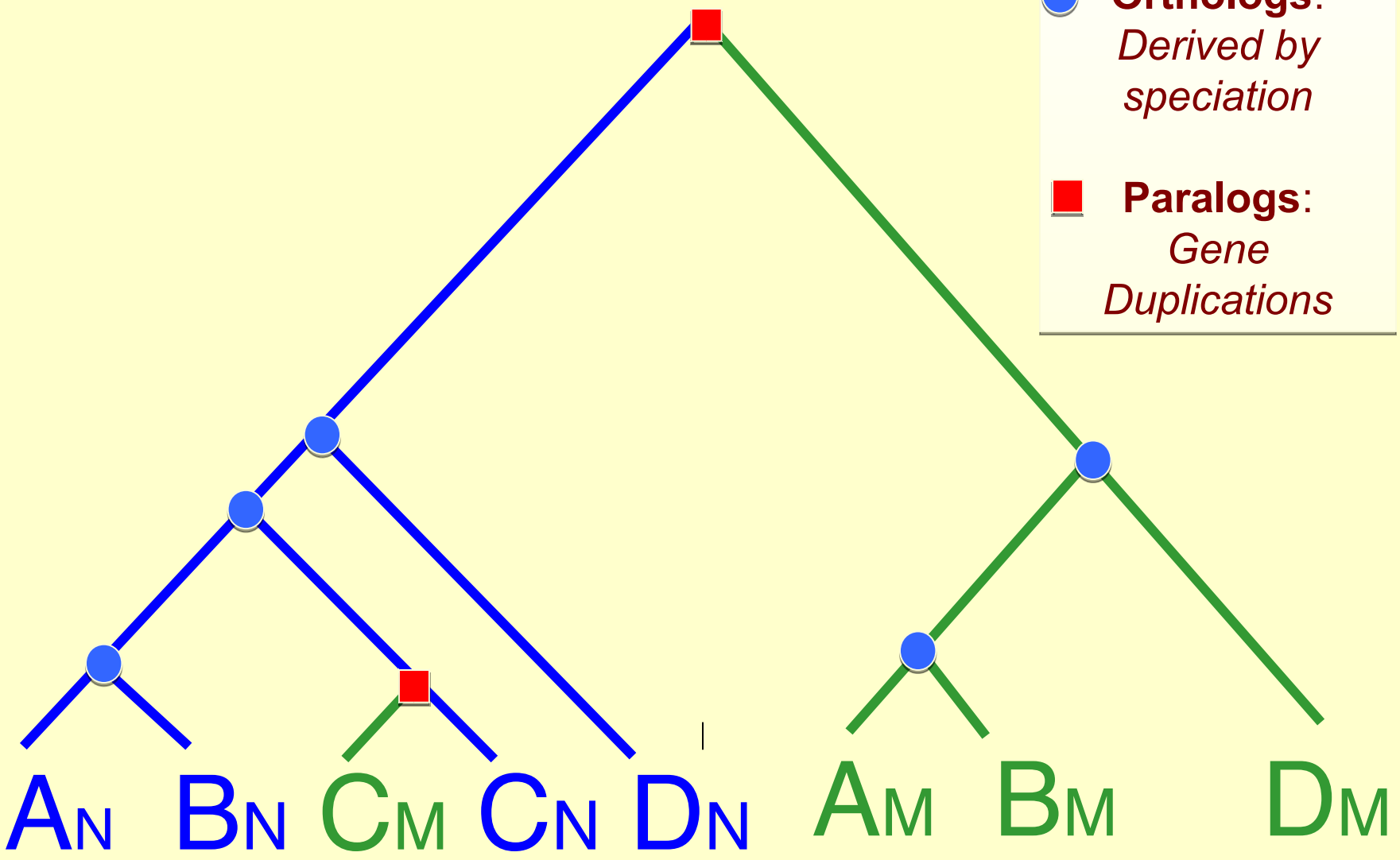
Thanks to Maryellen Ruvolo

Challenges Making Trees: Gene Conversion



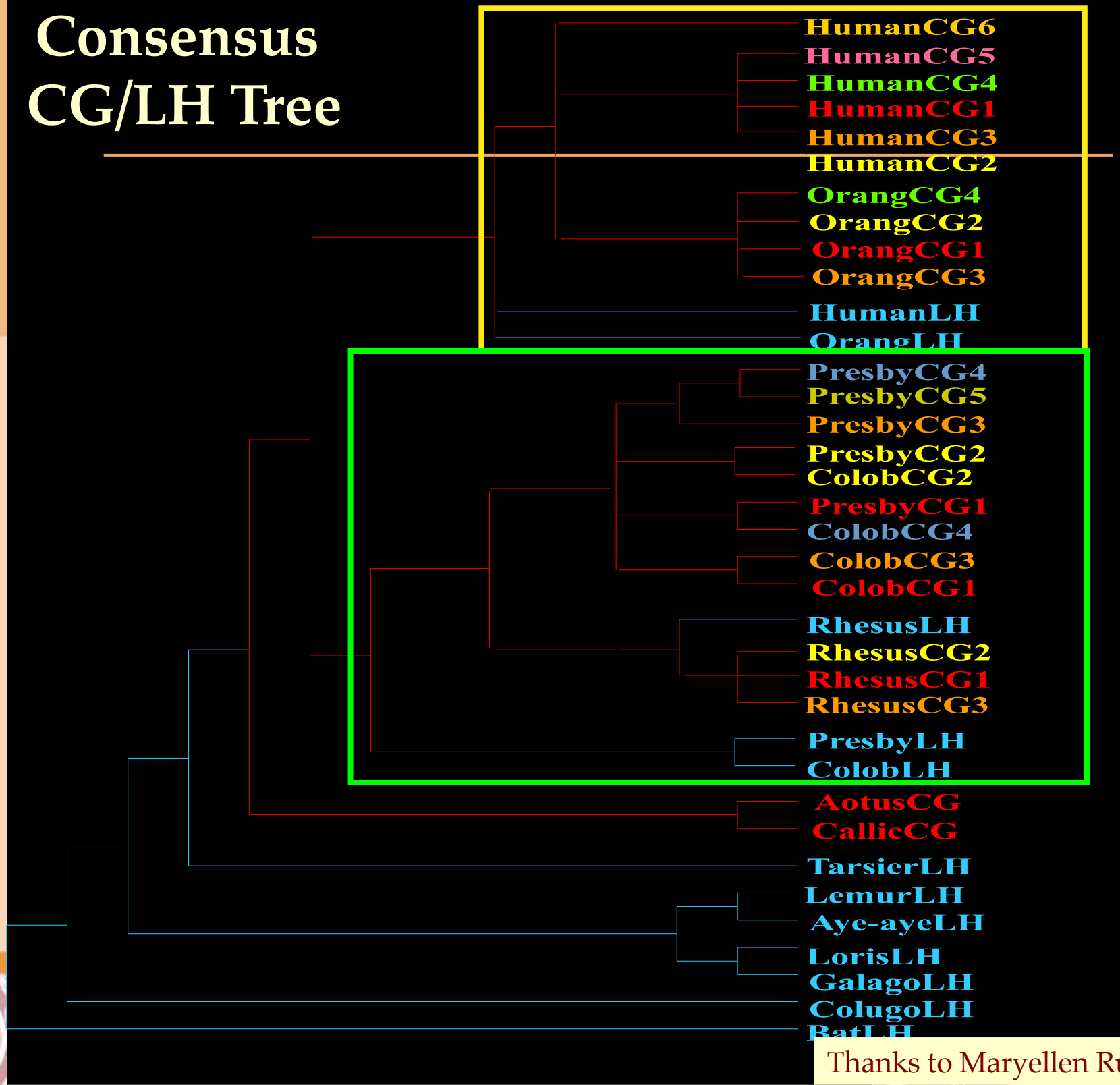
Thanks to Maryellen Ruvolo

Challenges Making Trees: C_M Has Been Converted from C_N



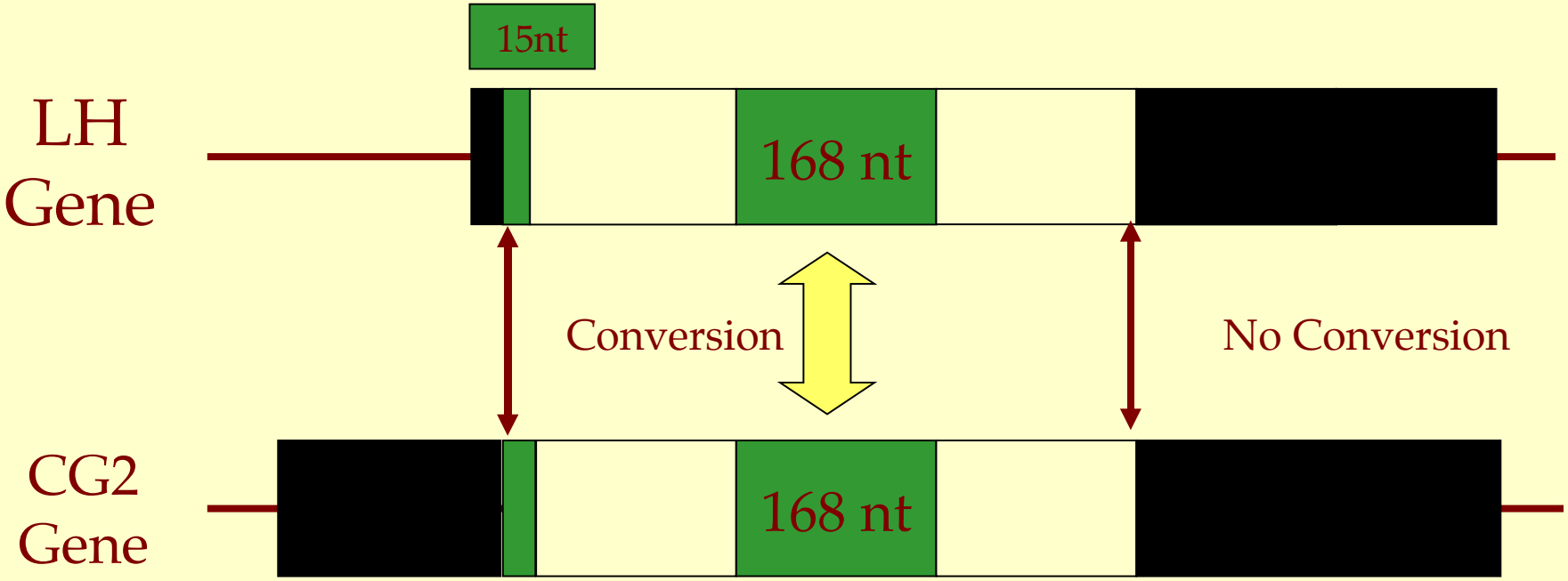
Thanks to Maryellen Ruvolo

Consensus CG/LH Tree



Thanks to Maryellen Ruvolo

Gene conversion between 1st & 2nd exons of LH, CG2 Genes



T
h

Thanks to Maryellen Ruvolo

Challenges Making Trees: Varying Rates of Mutation

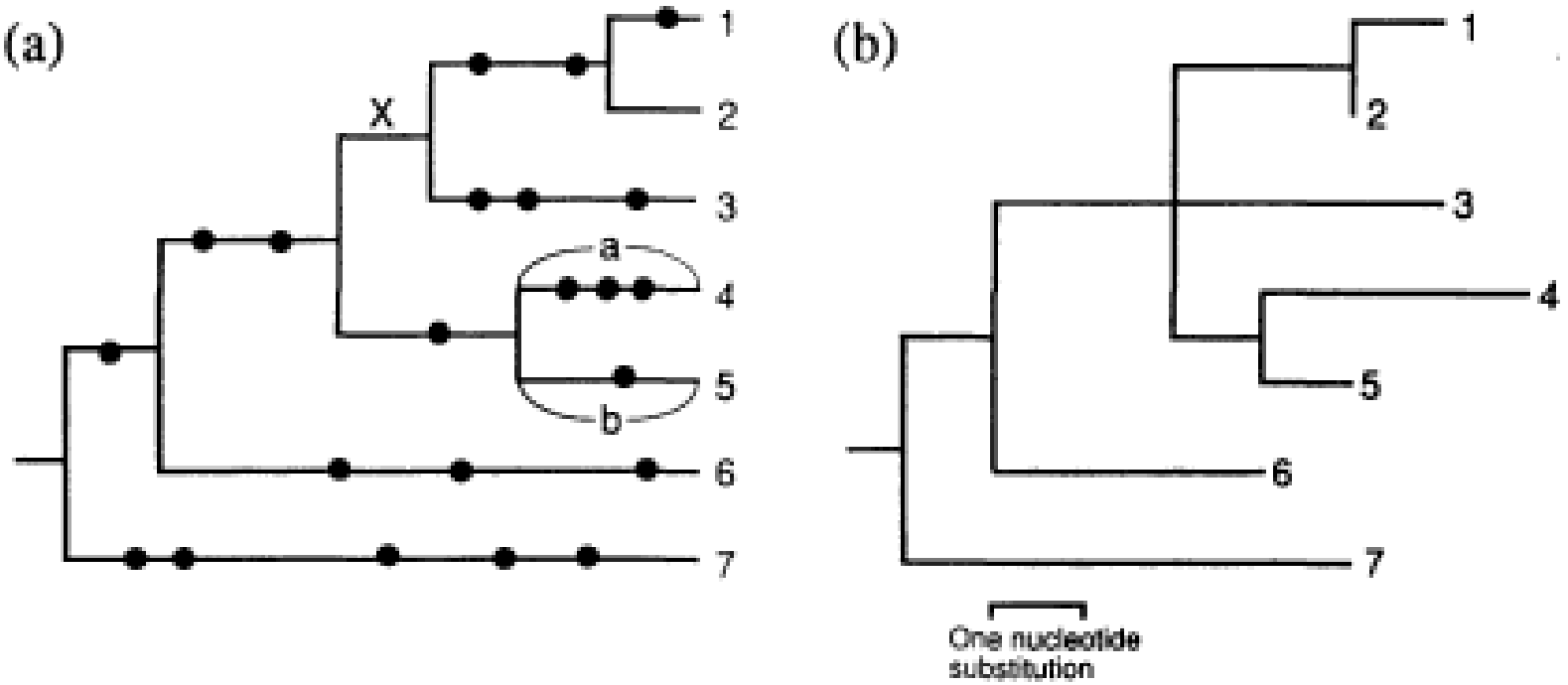
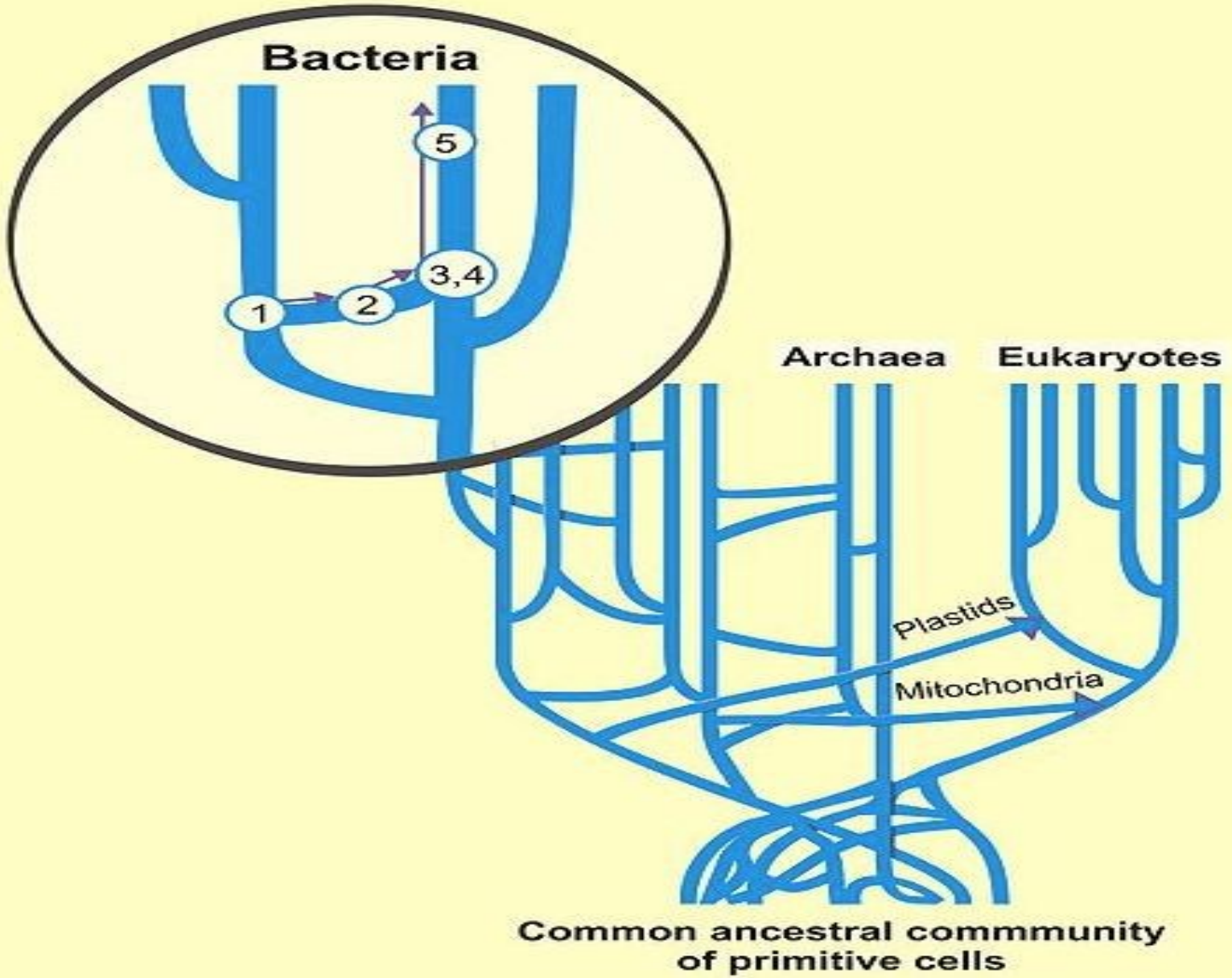


FIG. 5. Examples of the expected gene tree (a) and the corresponding realized gene trees (b). Filled circles on the expected gene tree denote nucleotide substitutions. Because no substitution occurred at branch X of the expected gene tree (a), the corresponding branch does not exist in the realized gene tree (b).

Challenges Making Trees: Horizontal Gene Transfer

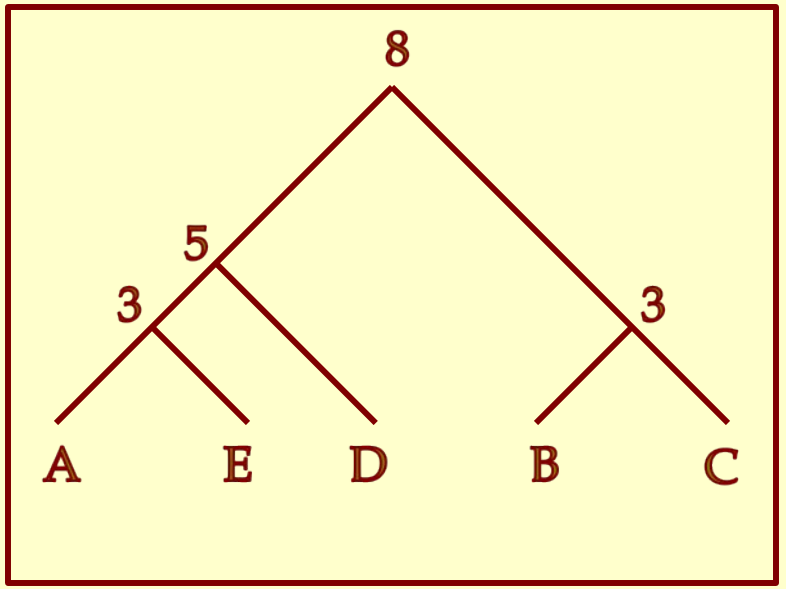


Maximum Ultrametric Distance Trees

Matrix D

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

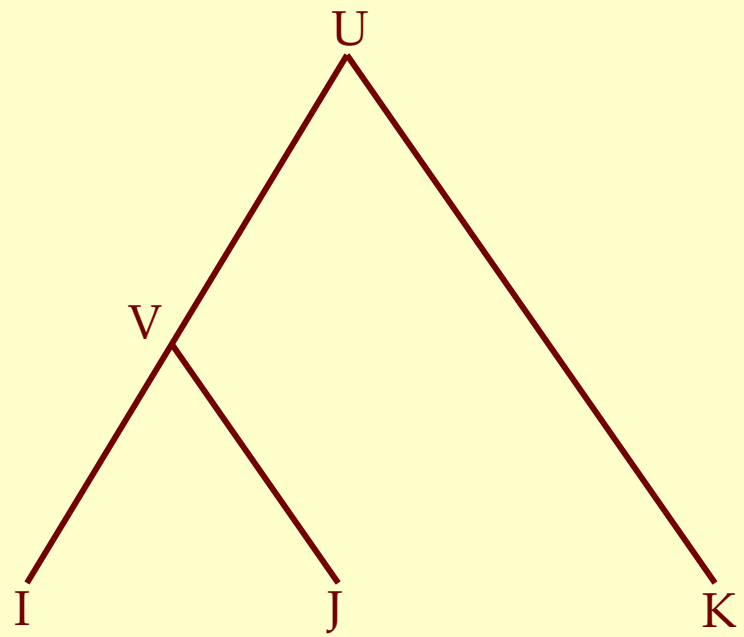
Tree T



- Matrix D is ultrametric for tree T if:
  - If D is a symmetric n by n matrix of distances
  - T contains n leaves, one from each row or column
  - Each node of T labeled by one entry from D
  - Numbers from root to leaves strictly decrease
  - For any two leaves i, j, D(i,j) labels nearest common ancestor of i and j in tree

Maximum Ultrametric Distance Trees

A symmetric matrix D is ultrametric if and only if for every three leaves i, j, and k, there is a tie for the maximum distance between D(i,j), D(i,k) and D(j,k).

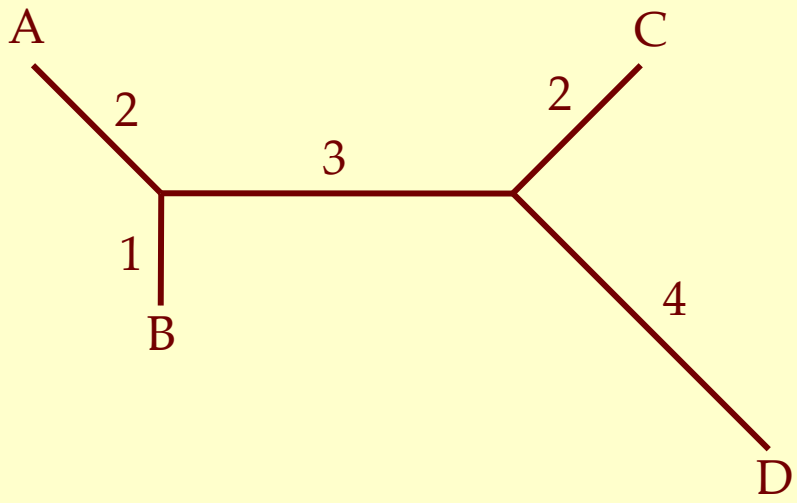


Additive Distance Trees

Matrix D

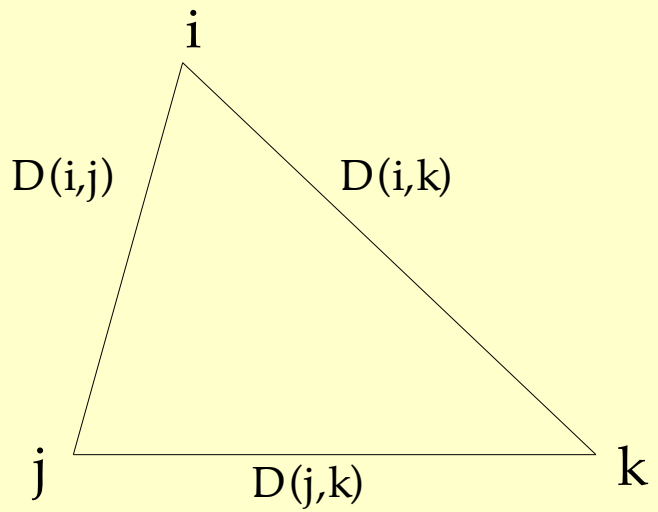
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0

Tree T

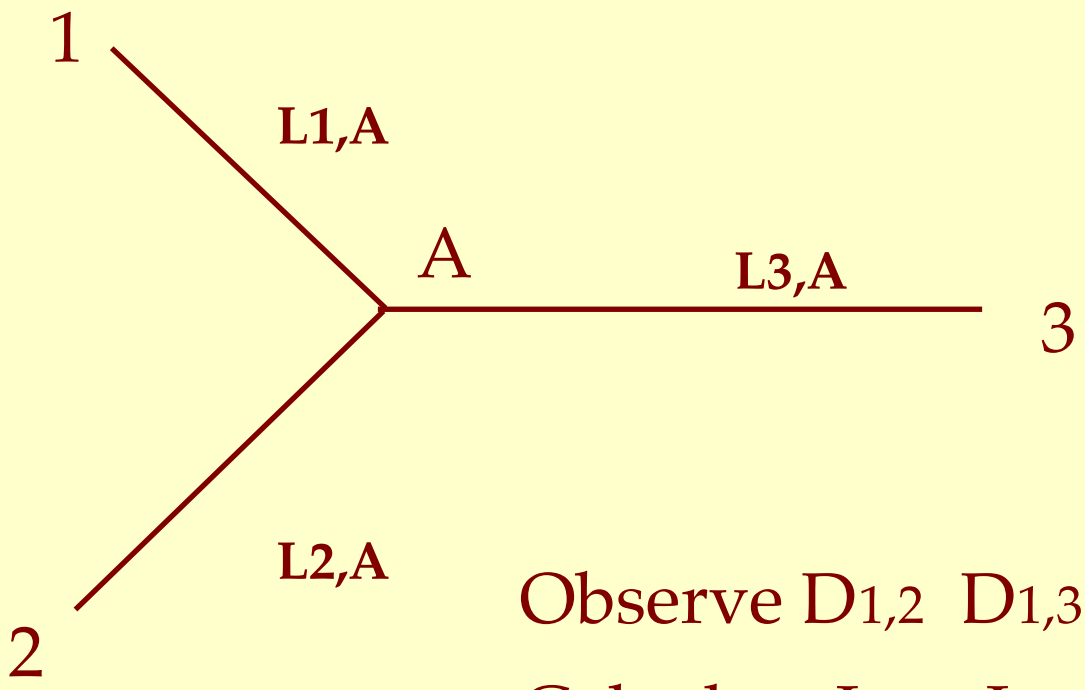


Distance Metrics Obey the Triangle Inequality

- $D(i,j) \leq D(i,k) + D(j,k)$ for all i, j, k
- (Max Score - Smith-Waterman Score) is a Metric if
  - If Gap-penalty $\geq$ 1+ Gap-size/(n-1)
  - Assuming match = 1 and mismatch = -1



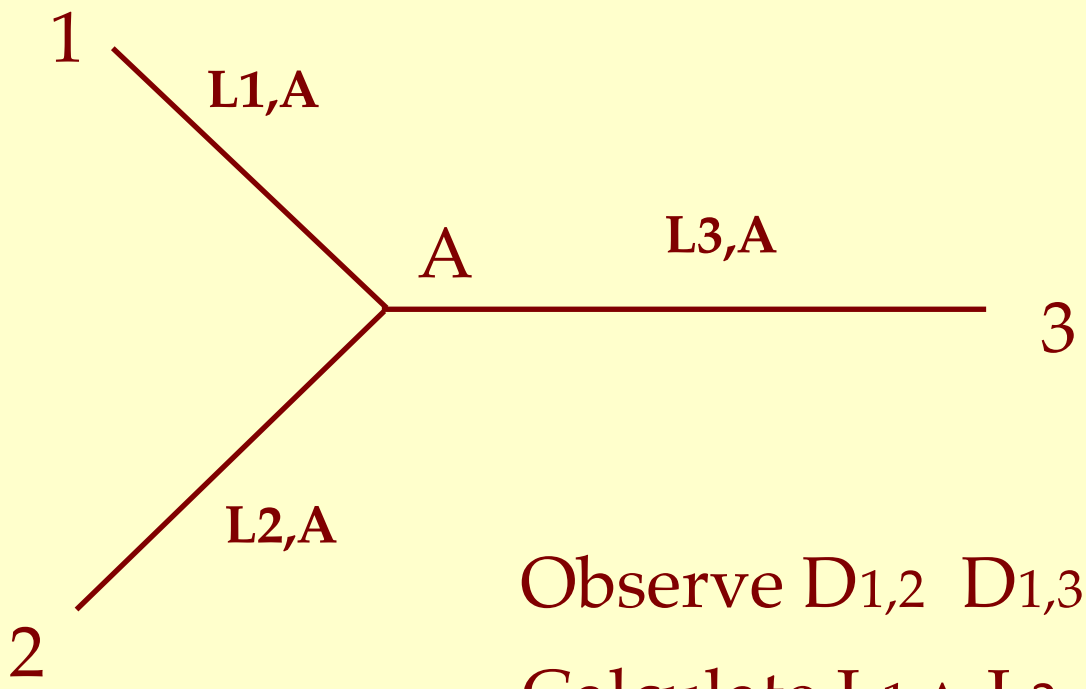
Three Leaf Tree



Observe $D_{1,2}$ $D_{1,3}$ $D_{2,3}$

Calculate $L_{1,A}$ $L_{2,A}$ $L_{3,A}$

Three Leaf Tree



Observe $D_{1,2}$ $D_{1,3}$ $D_{2,3}$

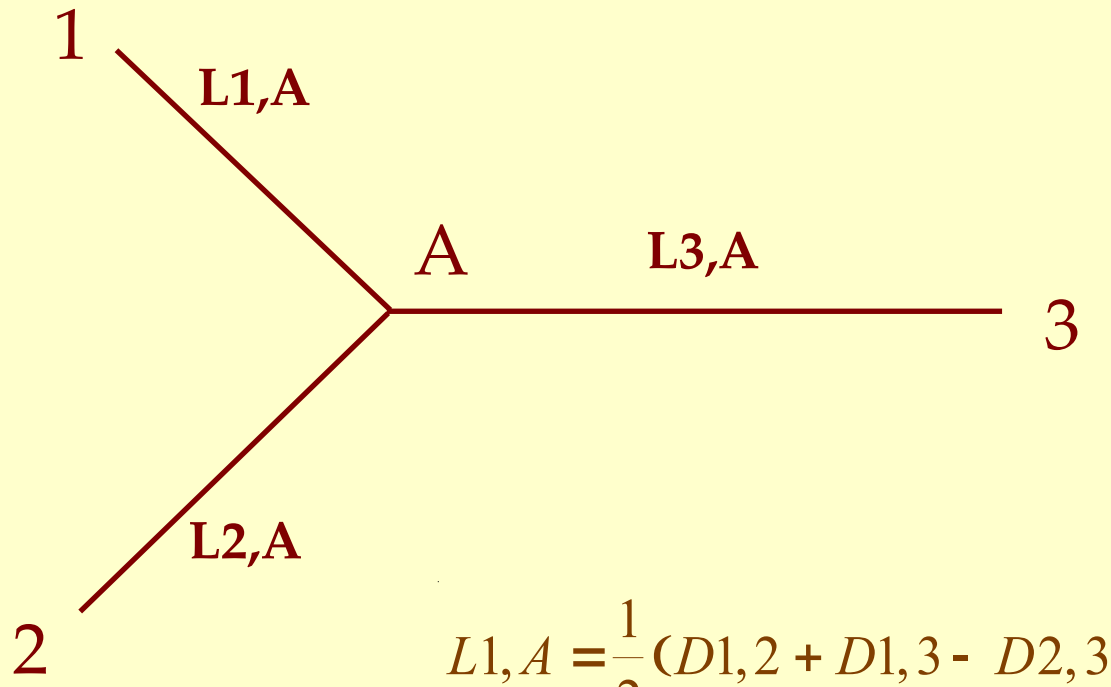
Calculate $L_{1,A}$ $L_{2,A}$ $L_{3,A}$

$D_{1,2}=L_{1,A}+L_{2,A}$

$D_{1,3}=L_{1,A}+L_{3,A}$

$D_{2,3}=L_{2,A}+L_{3,A}$

Solution to Three Species Tree

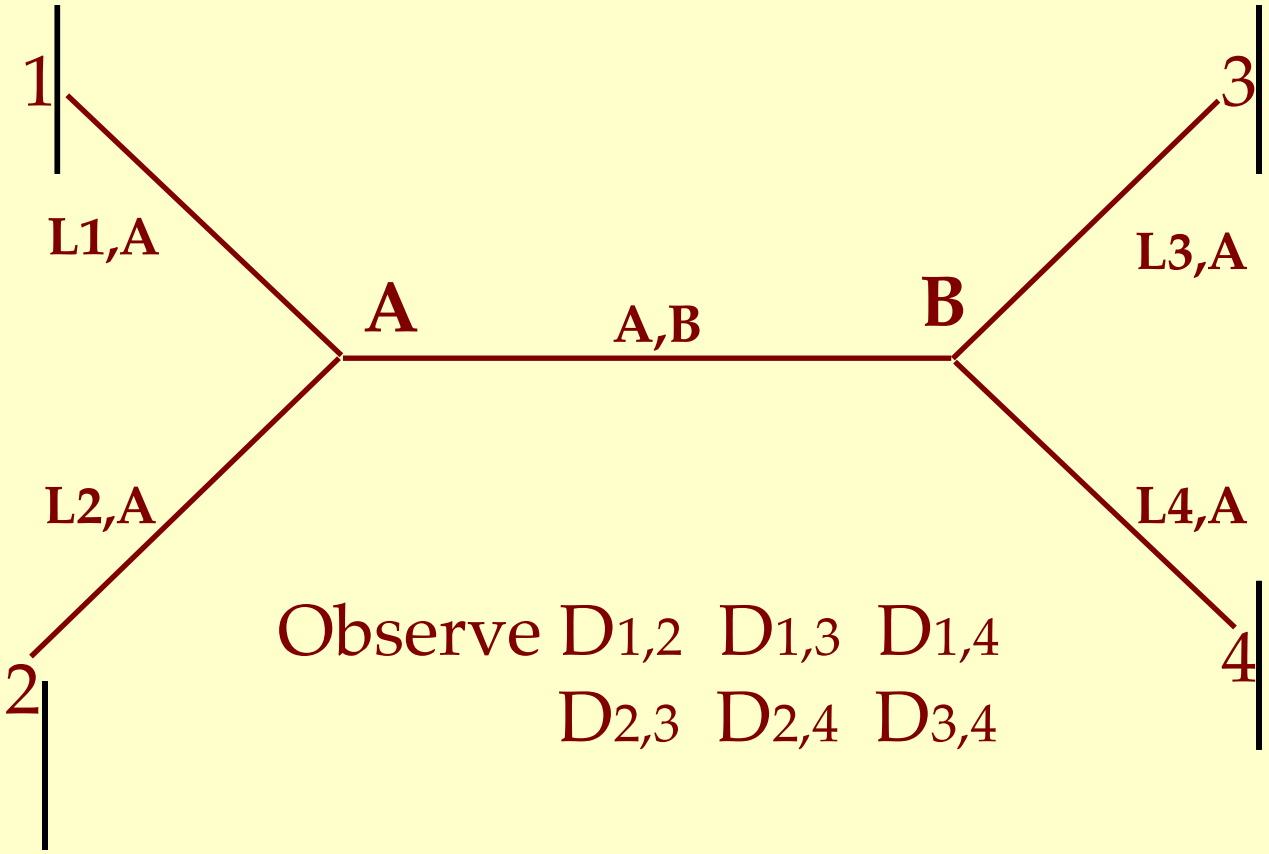


$$L1,A = \frac{1}{2}(D1,2 + D1,3 - D2,3)$$

$$L2,A = \frac{1}{2}(D1,2 + D2,3 - D1,3)$$

$$L3,A = \frac{1}{2}(D1,3 + D2,3 - D1,2)$$

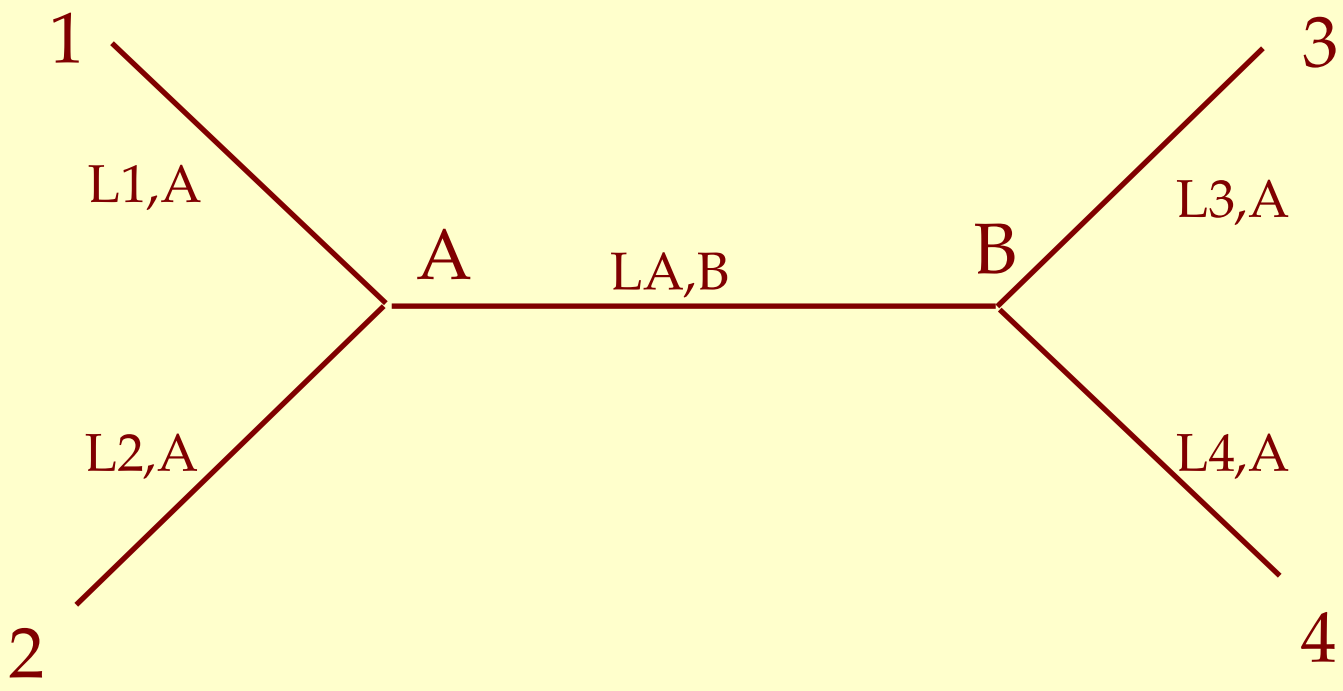
Four Species Tree



Observe $D_{1,2}$ $D_{1,3}$ $D_{1,4}$
$D_{2,3}$ $D_{2,4}$ $D_{3,4}$

Calculate $L_{1,A}$ $L_{2,A}$ $L_{3,B}$ $L_{4,B}$, $L_{A,B}$

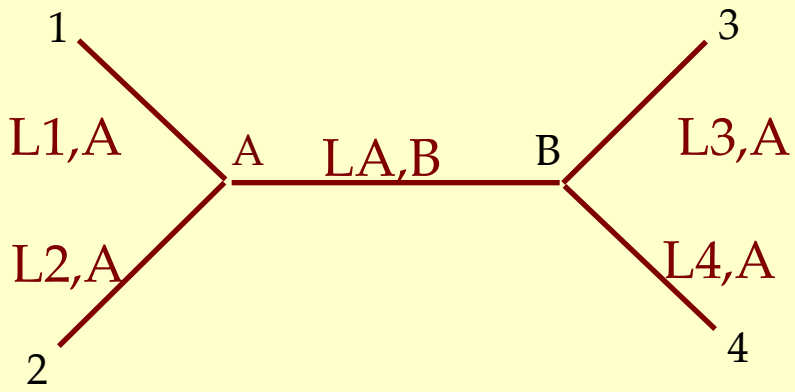
Four Species Topology



Label species 1, 2, 3, and 4 so that:

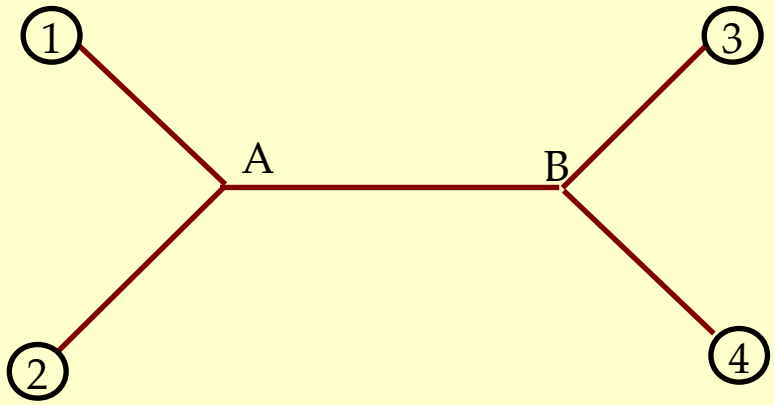
$$D(1,2) + D(3,4) \leq D(1,3) + D(2,4) = D(1,4) + D(2,3)$$

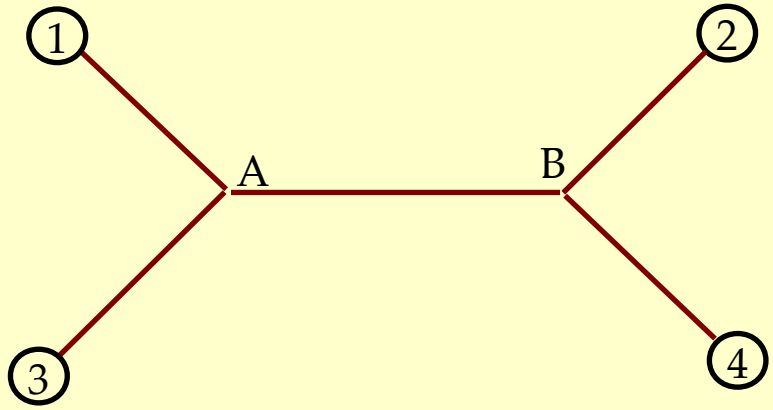
Solution for Four Species

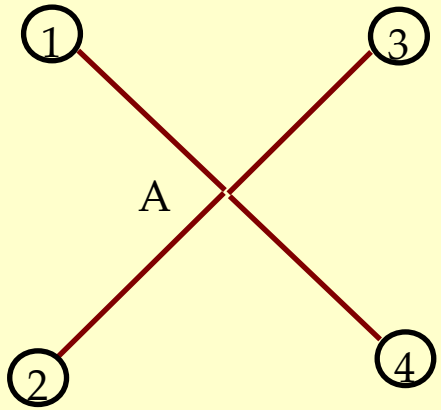


$$L1,A = 1/4*(D1,3 + D1,4 - D2,3 - D2,4) + 1/2*D1,2$$
$$L2,A = 1/4*(D2,3 + D2,4 - D1,3 - D1,4) + 1/2*D1,2$$
$$LB,3 = 1/4*(D1,3 + D2,3 - D1,4 - D2,4) + 1/2*D3,4$$
$$LB,4 = 1/4*(D1,4 + D2,4 - D1,3 - D2,3) + 1/2*D3,4$$
$$LA,B = 1/4*(D1,3 + D1,4 + D2,3 + D2,4) - 1/2*(D1,2 + D3,4)$$

Four Species =>Three Topologies







Species, Distances, Branches & Topologies

Number of Species	Number of Distances	Number of Branches	Number of Topologies
2	1	1	1
3	3	3	1
4	6	5	3
5	10	7	15
6	15	9	105

Species, Distances, Branches & Topologies

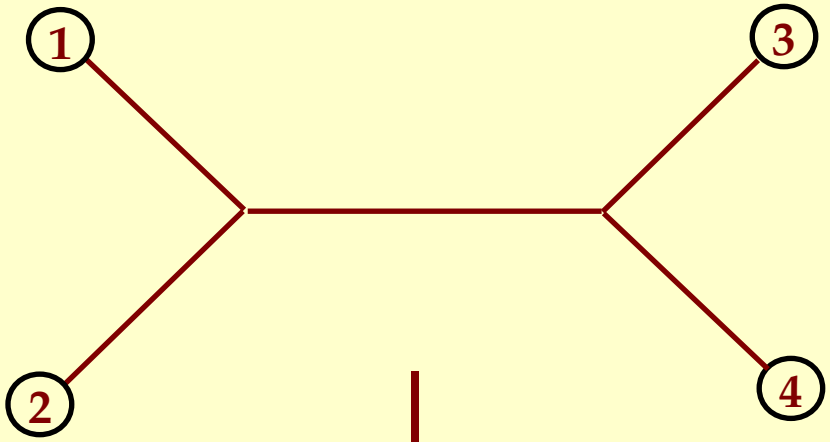
Number of Species	Number of Distances	Number of Branches	Number of Topologies
n	Dn	Ln	Tn
n+1	Dn+n	Ln+2	Ln*Tn
•	•	•	•
n	$\binom{n}{2} = \frac{n!}{2!(n-2)!}$	$(2n-3)$	$\prod_{i=1}^{n-2}(2n-1)$

Number of Topologies for n Species

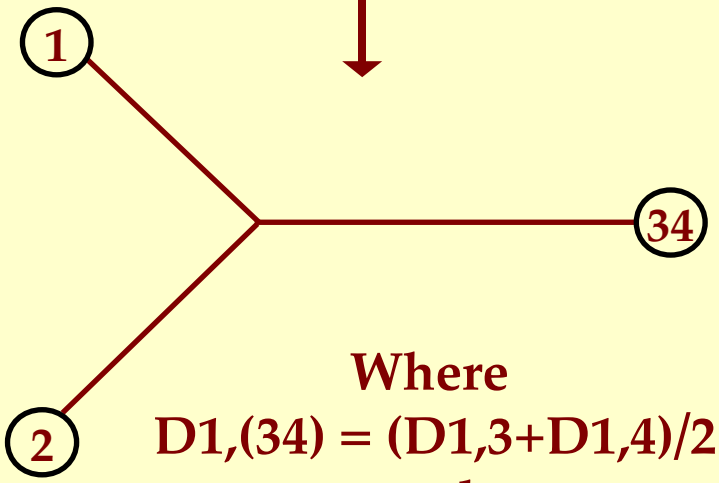
n	Tn
3	1
4	3
5	15
6	105
7	945
8	10,395
9	1.35×10^5
10	2.03×10^6
15	2.13×10^{14}
20	8.20×10^{21}

UPGMA: Unweighted Pair Group Method with Arithmetic Average

OTU	1	2	3
2	D1,2		
3	D1,3	D2,3	
4	D1,4	D2,4	D3,4

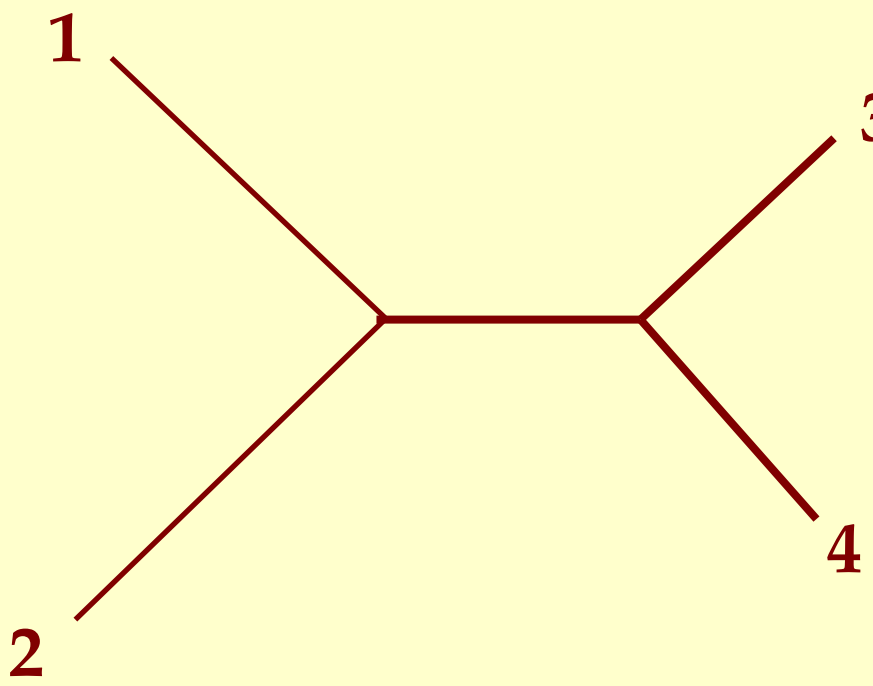


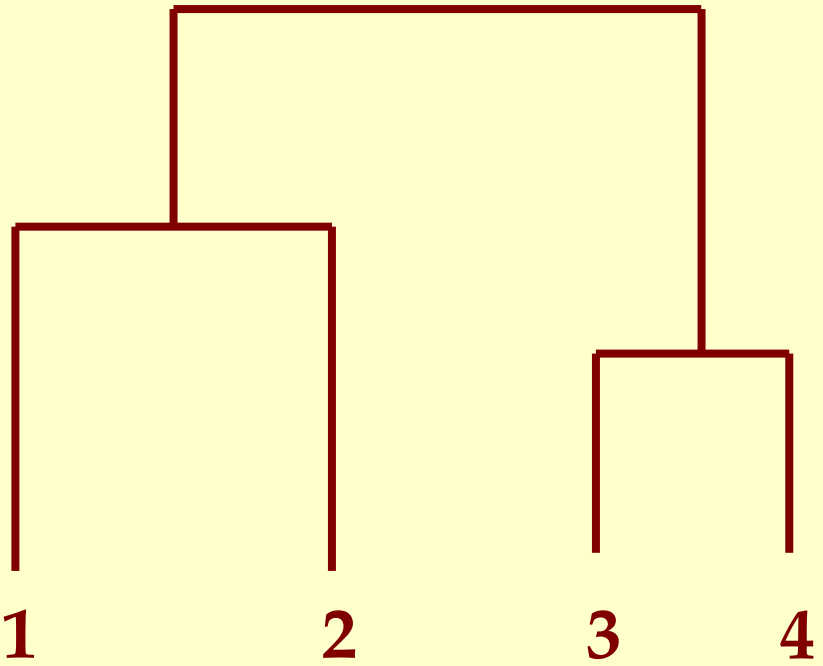
OTU	1	2
2	D1,2	
(34)	D1,(34)	D2,(34)



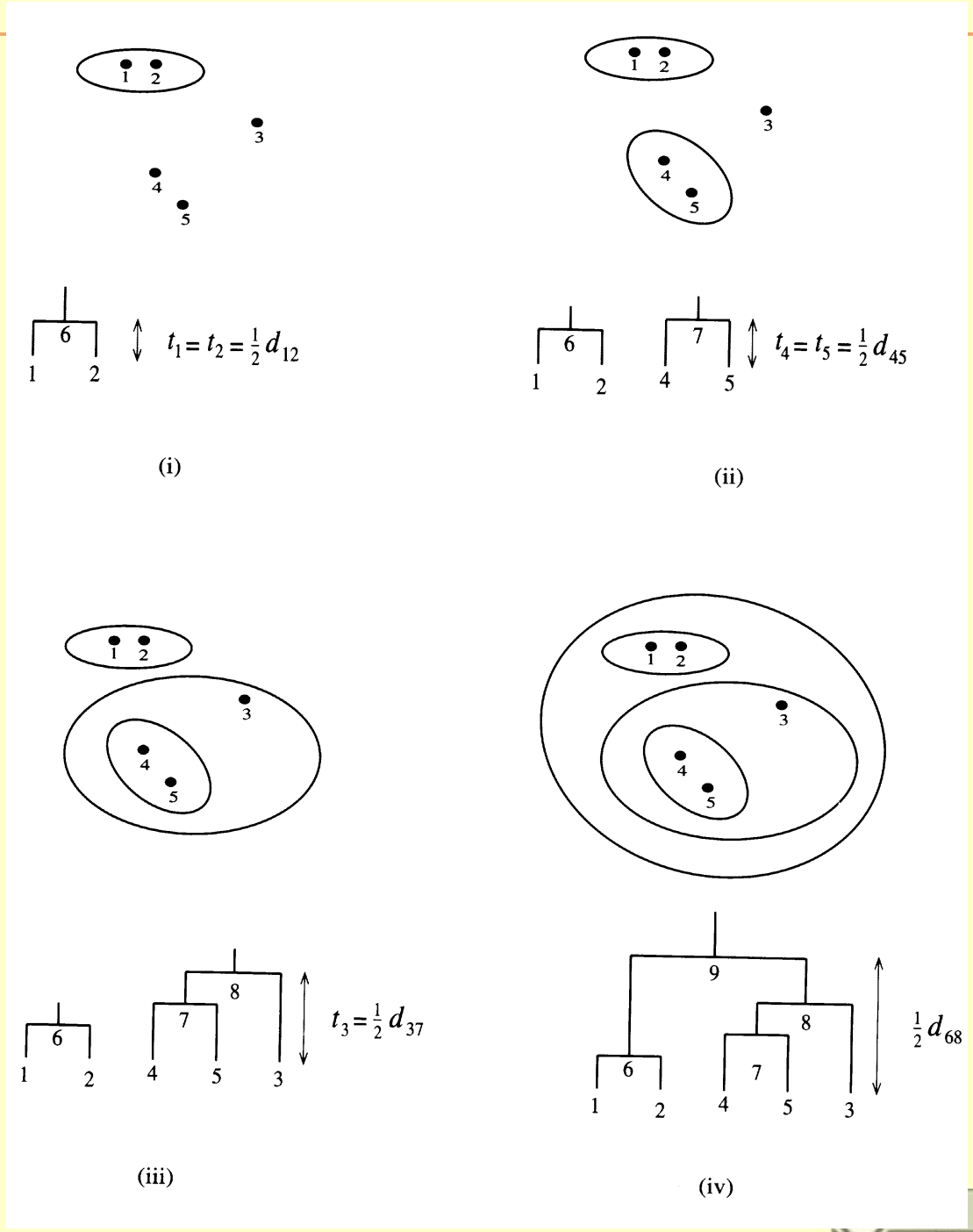
Where
D1,(34) = (D1,3+D1,4)/2
and
D2,(34) = (D2,3+D2,4)/2

UPGMA Dendrogram

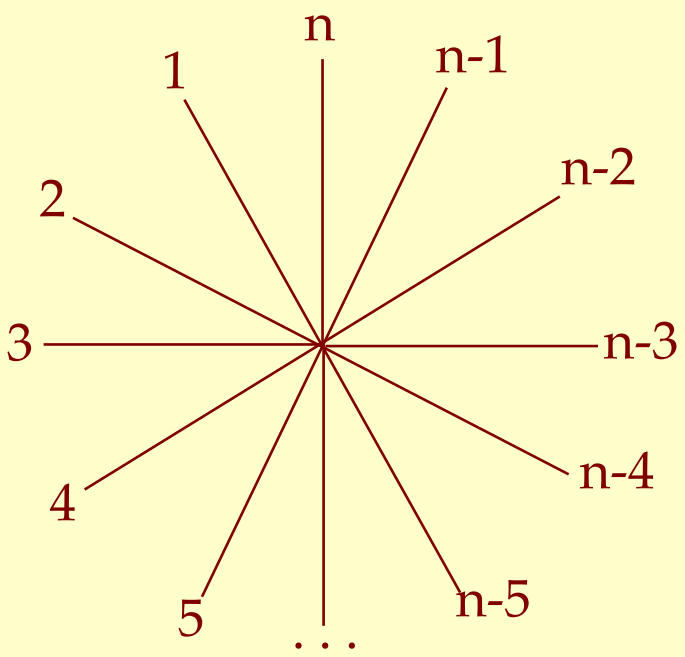


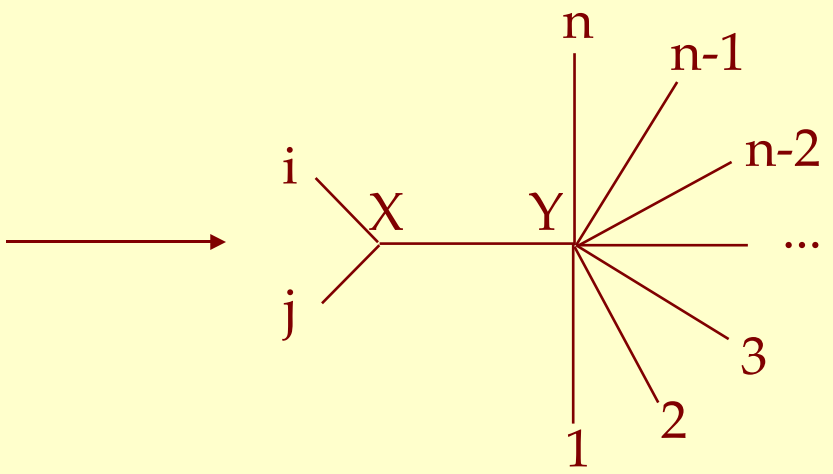


UPGMA Clustering



Neighbor Joining Method





For starlike tree $S_0 = Q/(n-1)$ where $Q = \sum_{i<j} D_{i,j}$

For nearest neighbor tree $S_{ij} = (B_{iX} + B_{jX}) + B_{XY} + \sum_{k \neq i,j} B_{kY}$

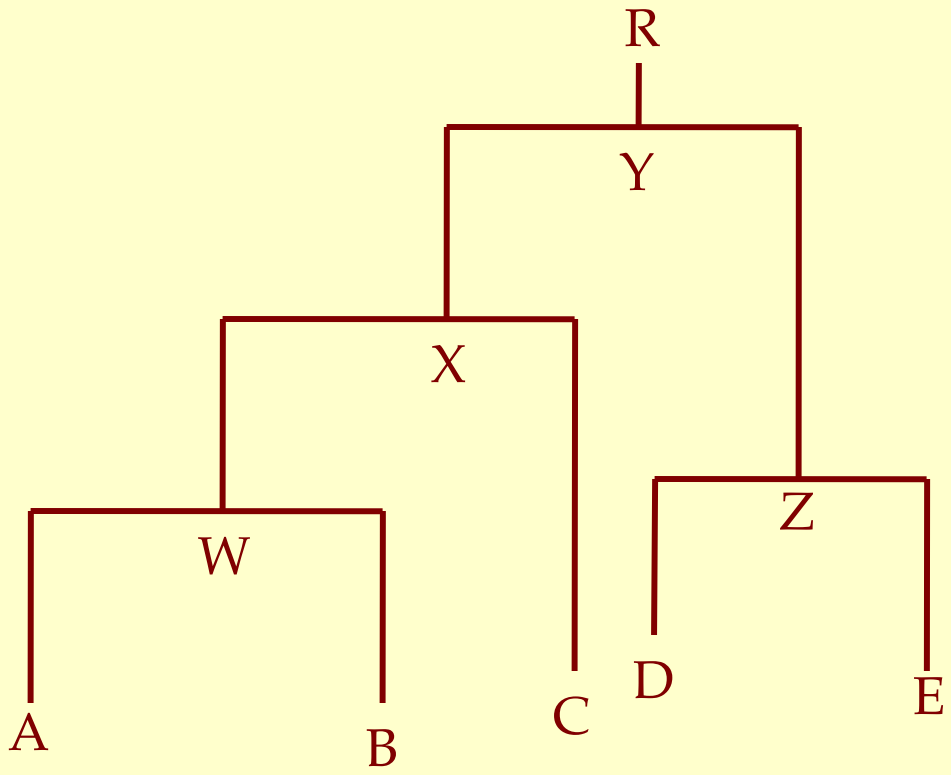
$D_{ij} = B_{iX} + B_{jX}$ $\quad D_{ik} = B_{iX} + B_{XY} + B_{kY}$ $\quad (k \neq i,j)$

$D_{kl} = B_{iY} + B_{jY}$ $\quad D_{jk} = B_{jX} + B_{XY} + B_{kY}$ $\quad (k,l \neq i,j)$

$$B_{XY} = \frac{Q - (n-1)D_{ij} - \frac{(n-1)}{(n-3)} \sum_{k,l \neq i,j} D_{kl}}{2(n-2)}$$

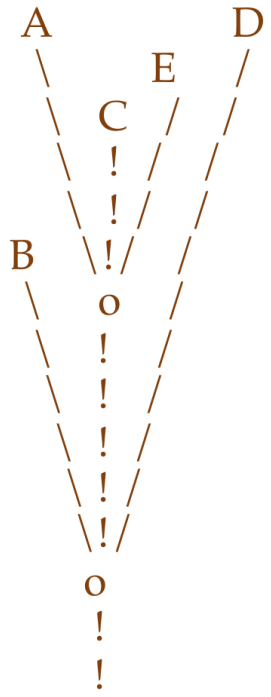
$$S_{ij} = \frac{D_{ij}}{2} + \frac{2\sum_{i<j} D_{ij} - \sum_j D_{ij} - \sum_i D_{ij}}{2(n-2)}$$

Nearest Neighbor Dendrogram



New Hampshire Standard Tree

If we have this rooted tree:



then the tree file is represented by the following sequence of printable characters, starting at the beginning of the file:

(B,(A,C,E),D);

(B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);

SeqWeb GrowTree Program

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot

SeqWeb v3.1

accelrys

Programs | Managers | Help Topics | Support

- Programs
- Comparison
- Database Searching
  - Similarity
  - Reference
- Evolution
- Mapping
- Pattern Recognition
- Primer Selection
- Protein Analysis
- Nucleic Acid Secondary Structure
- Translation
- Utilities

GrowTree

Construct Phylogenetic Trees from Peptide Sequences.

Input sequences:

Select From: Default | Project | Local File | Clipboard | Database

Sequence	Description	Type	Length	Range
myg_phyca	myg_phyca	P	153	1 .. 153
glb5_petma.pep	ID GLB5_PETMA STANDARD; PRT; 149 AA.	P	149	1 .. 149
hba_human	hba_human	P	141	1 .. 141
hba_horse.pep	ID HBA_HORSE STANDARD; PRT; 141 AA.	P	141	1 .. 141
hbb_horse.pep	ID HBB_HORSE STANDARD; PRT; 146 AA.	P	146	1 .. 146
lgb1_soybn.pep	- ID LGB1_SOYBN STANDARD; PRT; 143 AA.	P	143	1 .. 143
hbb_human	hbb_human	P	146	1 .. 146

Refresh | Clear

Input Parameters:

Distance Correction Method	uncorrected distance	○
	Jukes-Cantor distance	○
	Kimura distance	◉
Tree Construction Method	Neighbor joining	◉
	UPGMA	○

SeqWeb v 3.1

GrowTree Parameters

Input Parameters:

Distance Correction Method	uncorrected distance	○
	Jukes-Cantor distance	○
	Kimura distance	◉
Tree Construction Method	Neighbor joining	◉
	UPGMA	○

Select a sequence comparision matrix. This matrix determines how matches and mismatches are scored. The default penalites for gap creation and extension are given after each matrix name.

Scoring Matrix: blosum62

Set gap creation penalty: 8

Set gap extension penalty: 2

Limit the maximum input sequence range only when needed. Setting a higher limit allows you to align longer sequences while setting a lower limit allows you to add more and longer gaps to each sequence.

Maximum input sequence range: 5000 *(range 1 thru 700*

Limit the maximum number of gaps only when needed. Setting a higher limit allows you to add more and longer gap to each sequence while setting a lower limit allows you to align a greater number of sequences.

Maximum number of gap characters ('.' and '~') added to any sequence: 2000 *(range 0 thru 700*

Consider partial matches between degenerate symbols for uncorrected or Jukes-Cantor distance: ☐

Gap weight for uncorrected or Jukes-Cantor distance: 0.0 *(range 0.0 thru 2*

Report negative branch lengths as negative (instead of zero): ☐

Display Tree As:	phylogram (branch lengths proportional to distance)	◉
	cladogram (all branches the same length)	○

Run Reset

GrowTree Distances

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot

SeqWeb v3.1

Evolutionary Analysis Results

Genetic Distances

Calculated over: 9 to 157

Correction method: Kimura protein distance
Distances are: estimated number of substitutions per 100 amino acids
Symmatrix version 1
Number of matrices: 1

//
Matrix 1, dimension: 7

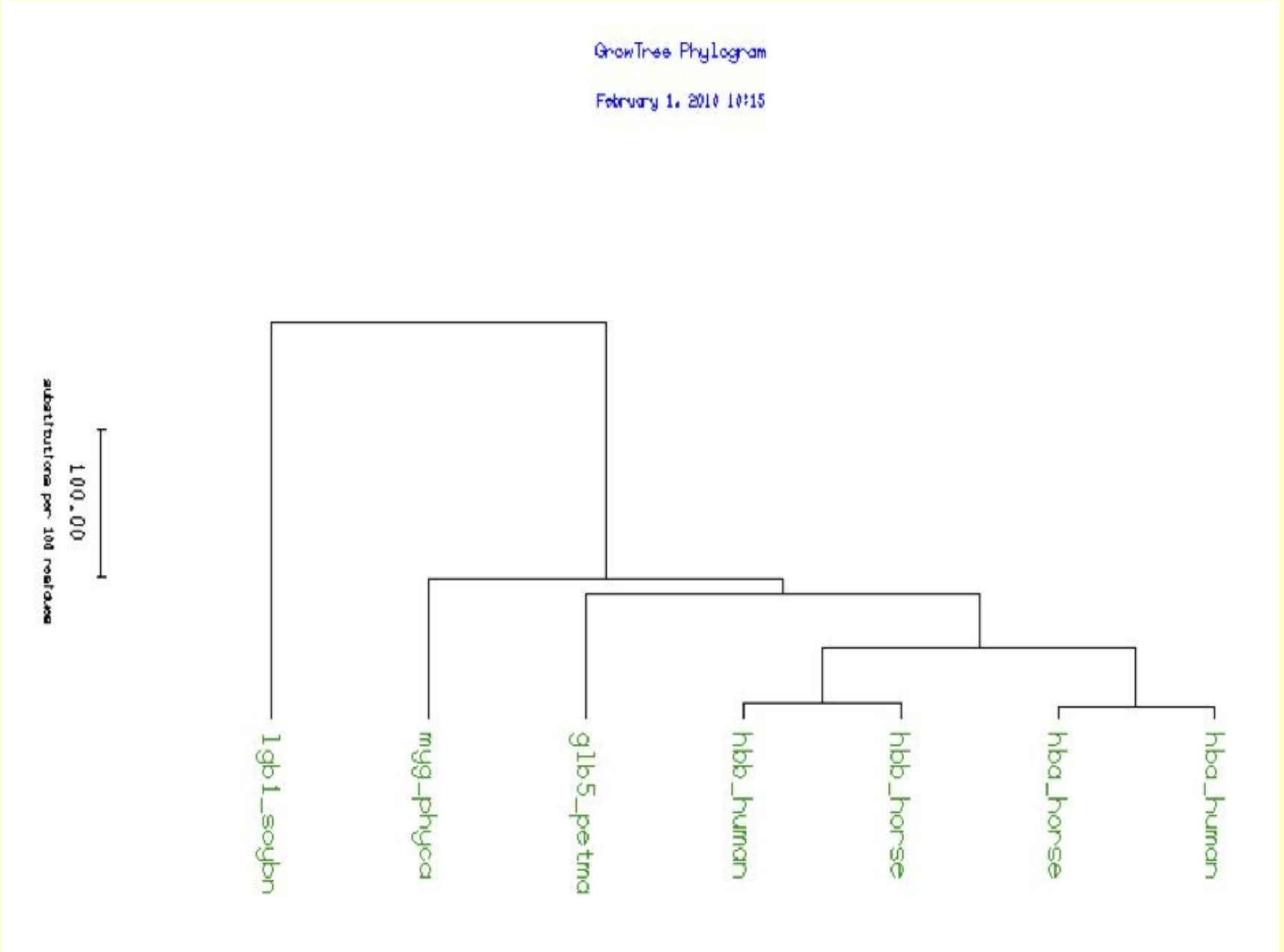
Key for column and row indices:

1 hba_human
2 hba_horse
3 hbb_horse
4 hbb_human
5 glb5_petma
6 myg_phyca
7 lgb1_soybn

Matrix 1: Part 1

	1	2	3	4	5	6	7
1	0.00	13.39	95.79	93.49	134.46	173.98	540.37
2		0.00	91.25	95.79	134.46	179.53	540.37
3			0.00	18.90	211.52	180.02	257.93
4				0.00	196.68	180.02	287.05
5					0.00	213.42	336.95
6						0.00	999.99
7							0.00

GrowTree Phylogram (UPGMA)



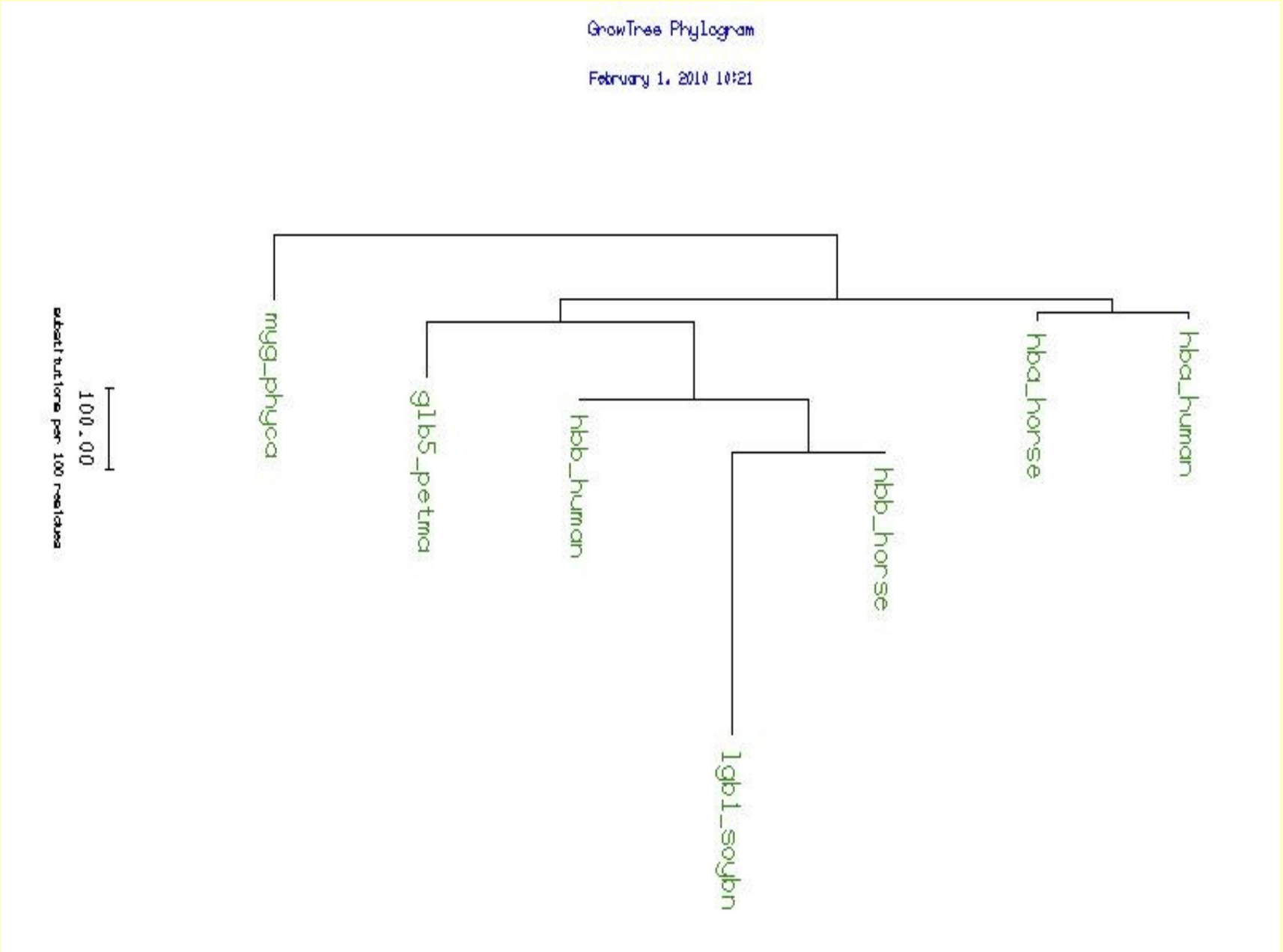
GrowTree Alignment

```
 Symbol comparison table: share_matrix:blosum62.cmp  CompCheck: 11

                    GapWeight: 8
                GapLengthWeight: 2

PileUp  MSF: 165  Type: P  February 1, 2010 10:15  Check: 6593 ..

 Name: hba_human          Len:   165  Check: 1231  Weight:  1.00
 Name: hba_horse          Len:   165  Check: 2167  Weight:  1.00
 Name: hbb_horse          Len:   165  Check: 9310  Weight:  1.00
 Name: hbb_human          Len:   165  Check:  208  Weight:  1.00
 Name: glb5_petma         Len:   165  Check: 2079  Weight:  1.00
 Name: myg_phyca          Len:   165  Check: 4320  Weight:  1.00
 Name: lgb1_soybn         Len:   165  Check: 7278  Weight:  1.00

//
                1                                                   5
   hba_human ~~~~~~~~~~~v lspadktnvk aawgkvgaha geygaealer mflsfpttk
   hba_horse ~~~~~~~~~~~V LSAADKTNVK AAWSKVGGHA GEYGAEALER MFLGFPTTK
   hbb_horse ~~~~~~~~~~VQ LSGEEKAAVL ALWDKV..NE EEVGGEALGR LLVVYPWTQ
   hbb_human ~~~~~~~~~~vh ltpeeksavt alwgkv..nv devggealgr llvvypwtq
  glb5_petma PIVDTGSVAP LSAAEKTKIR SAWAPVYSTY ETSGVDILVK FFTSTPAAQ
   myg_phyca ~~~~~~~~~~v lsegewqlvl hvwakveadv aghgqdilir lfkshpetl
  lgb1_soybn ~~~~~~~~~~ga ftekqealvs ssfeafkani pqysvvfyns ilekapaak

               51                                                  10
   hba_human yfphf.dlsh .....gsaqv kghgkkvada ltnavahvdd mpnalsals
   hba_horse YFPHF.DLSH .....GSAQV KAHGKKVGDA LTLAVGHLDD LPGALSNLS
   hbb_horse FFDSFGDLSN PGAVMGNPKV KAHGKKVLHS FGEGVHHLDN LKGTFAALS
   hbb_human ffesfgdlst pdavmgnpkv kahgkkvlga fsdglahldn lkgtfatls
  glb5_petma FFPKFKGLTT ADQLKKSADV RWHAERIINA VNDAVASMDD TEKMSMKLR
   myg_phyca kfdrfkhlkt eaemkasedl kkhgvtvlta lg...ailkk kghheaelk
  lgb1_soybn lfsflan... .gvdptnpkl tghaeklfal vrdsagql.k tngtvvada

              101                                                  15
   hba_human l...hahklr vdpvnfklls hcllvtlaah lpaeftpavh asldkflas
   hba_horse L...HAHKLR VDPVNFKLLS HCLLSTLAVH LPNDFTPAVH ASLDKFLSS
   hbb_horse L...HCDKLH VDPENFRLLG NVLVVVLARH FGKDFTPELQ ASYQKVVAG
   hbb_human l...hcdklh vdpenfrllg nvlvcvlahh fgkeftppvq aayqkvvag
  glb5_petma LSGKHAKSFQ VDPQYFKVLA AVIADTVA.. .......AGD AGFEKLMSM
   myg_phyca laqshatkhk ipikylefis eaiihvlhsr hpgdfgadaq gamnkalel
  lgb1_soybn lvsihaqkav tdpq.fvvvk eallktikea vggnwsdels sawevayde

              151             165
   hba_human stvltskyr~ ~~~~~~~
   hba_horse STVLTSKYR~ ~~~~~~~
   hbb_horse ANALAHKYH~ ~~~~~~~
   hbb_human analahkyh~ ~~~~~~~
  glb5_petma CILLRSAY~~ ~~~~~~~
   myg_phyca rkdiaakyke lgyqg
  lgb1_soybn aaaikka~~~~ ~~~~~~~
```

SeqWeb v 3.1

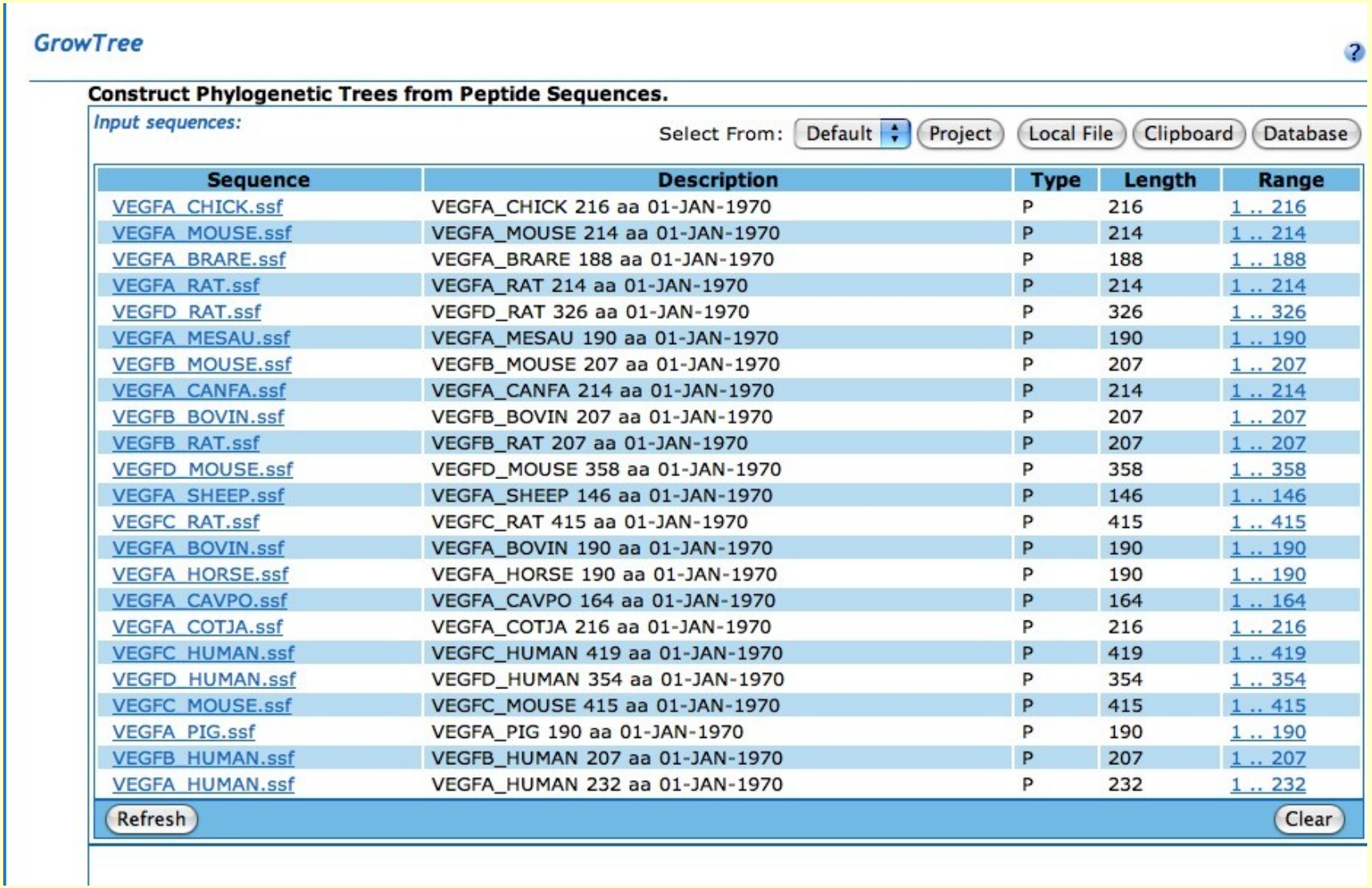
GrowTree Neighbor Joining Tree

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot



GrowTree VegF Input

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot



GrowTree

Construct Phylogenetic Trees from Peptide Sequences.

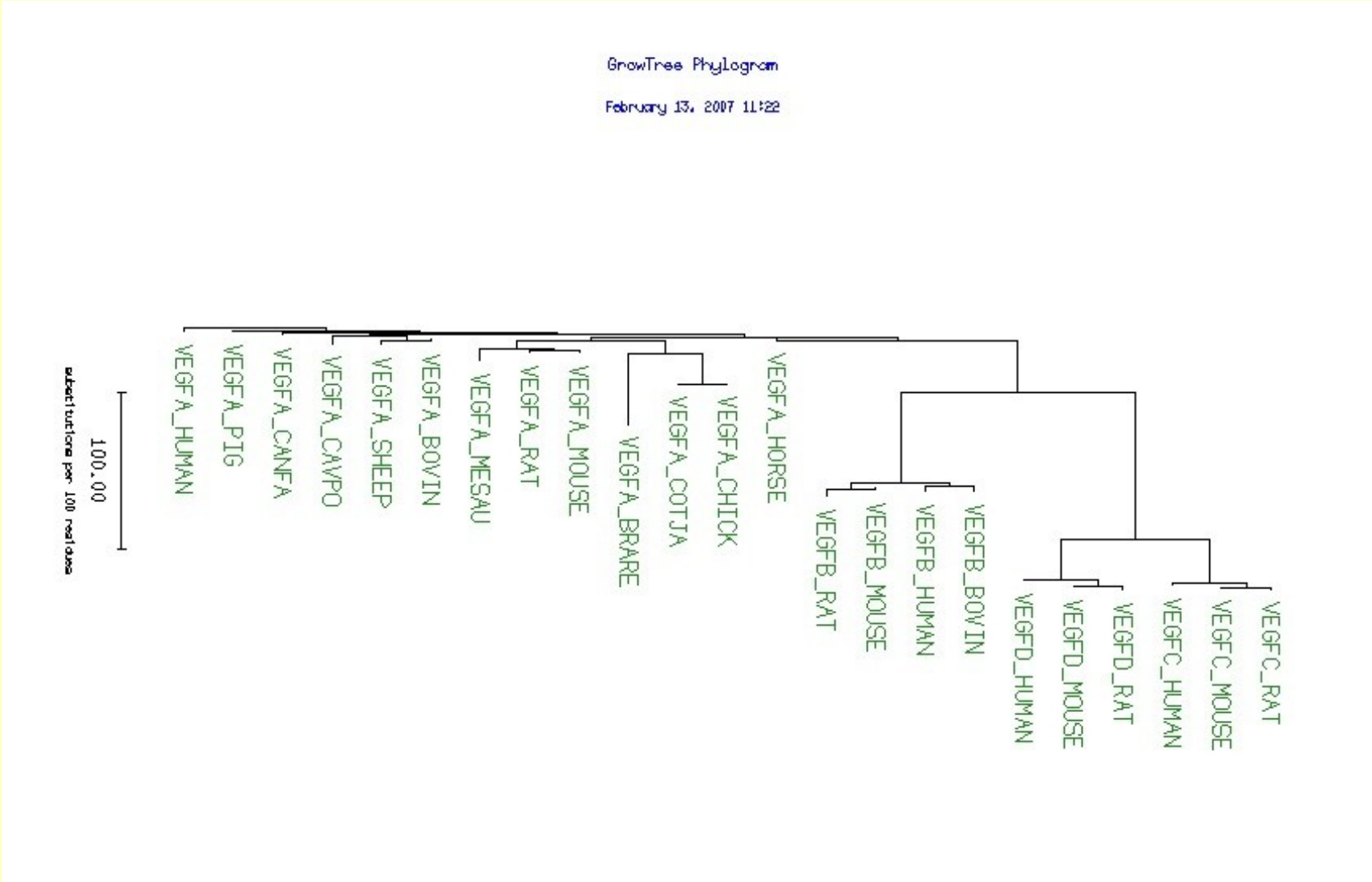
Input sequences:

Select From: Default | Project | Local File | Clipboard | Database

Sequence	Description	Type	Length	Range
VEGFA_CHICK.ssf	VEGFA_CHICK 216 aa 01-JAN-1970	P	216	1 .. 216
VEGFA_MOUSE.ssf	VEGFA_MOUSE 214 aa 01-JAN-1970	P	214	1 .. 214
VEGFA_BRARE.ssf	VEGFA_BRARE 188 aa 01-JAN-1970	P	188	1 .. 188
VEGFA_RAT.ssf	VEGFA_RAT 214 aa 01-JAN-1970	P	214	1 .. 214
VEGFD_RAT.ssf	VEGFD_RAT 326 aa 01-JAN-1970	P	326	1 .. 326
VEGFA_MESAU.ssf	VEGFA_MESAU 190 aa 01-JAN-1970	P	190	1 .. 190
VEGFB_MOUSE.ssf	VEGFB_MOUSE 207 aa 01-JAN-1970	P	207	1 .. 207
VEGFA_CANFA.ssf	VEGFA_CANFA 214 aa 01-JAN-1970	P	214	1 .. 214
VEGFB_BOVIN.ssf	VEGFB_BOVIN 207 aa 01-JAN-1970	P	207	1 .. 207
VEGFB_RAT.ssf	VEGFB_RAT 207 aa 01-JAN-1970	P	207	1 .. 207
VEGFD_MOUSE.ssf	VEGFD_MOUSE 358 aa 01-JAN-1970	P	358	1 .. 358
VEGFA_SHEEP.ssf	VEGFA_SHEEP 146 aa 01-JAN-1970	P	146	1 .. 146
VEGFC_RAT.ssf	VEGFC_RAT 415 aa 01-JAN-1970	P	415	1 .. 415
VEGFA_BOVIN.ssf	VEGFA_BOVIN 190 aa 01-JAN-1970	P	190	1 .. 190
VEGFA_HORSE.ssf	VEGFA_HORSE 190 aa 01-JAN-1970	P	190	1 .. 190
VEGFA_CAVPO.ssf	VEGFA_CAVPO 164 aa 01-JAN-1970	P	164	1 .. 164
VEGFA_COTJA.ssf	VEGFA_COTJA 216 aa 01-JAN-1970	P	216	1 .. 216
VEGFC_HUMAN.ssf	VEGFC_HUMAN 419 aa 01-JAN-1970	P	419	1 .. 419
VEGFD_HUMAN.ssf	VEGFD_HUMAN 354 aa 01-JAN-1970	P	354	1 .. 354
VEGFC_MOUSE.ssf	VEGFC_MOUSE 415 aa 01-JAN-1970	P	415	1 .. 415
VEGFA_PIG.ssf	VEGFA_PIG 190 aa 01-JAN-1970	P	190	1 .. 190
VEGFB_HUMAN.ssf	VEGFB_HUMAN 207 aa 01-JAN-1970	P	207	1 .. 207
VEGFA_HUMAN.ssf	VEGFA_HUMAN 232 aa 01-JAN-1970	P	232	1 .. 232

Refresh Clear

GrowTree VegF Neighbor Joining Tree



VegF Growth Factors

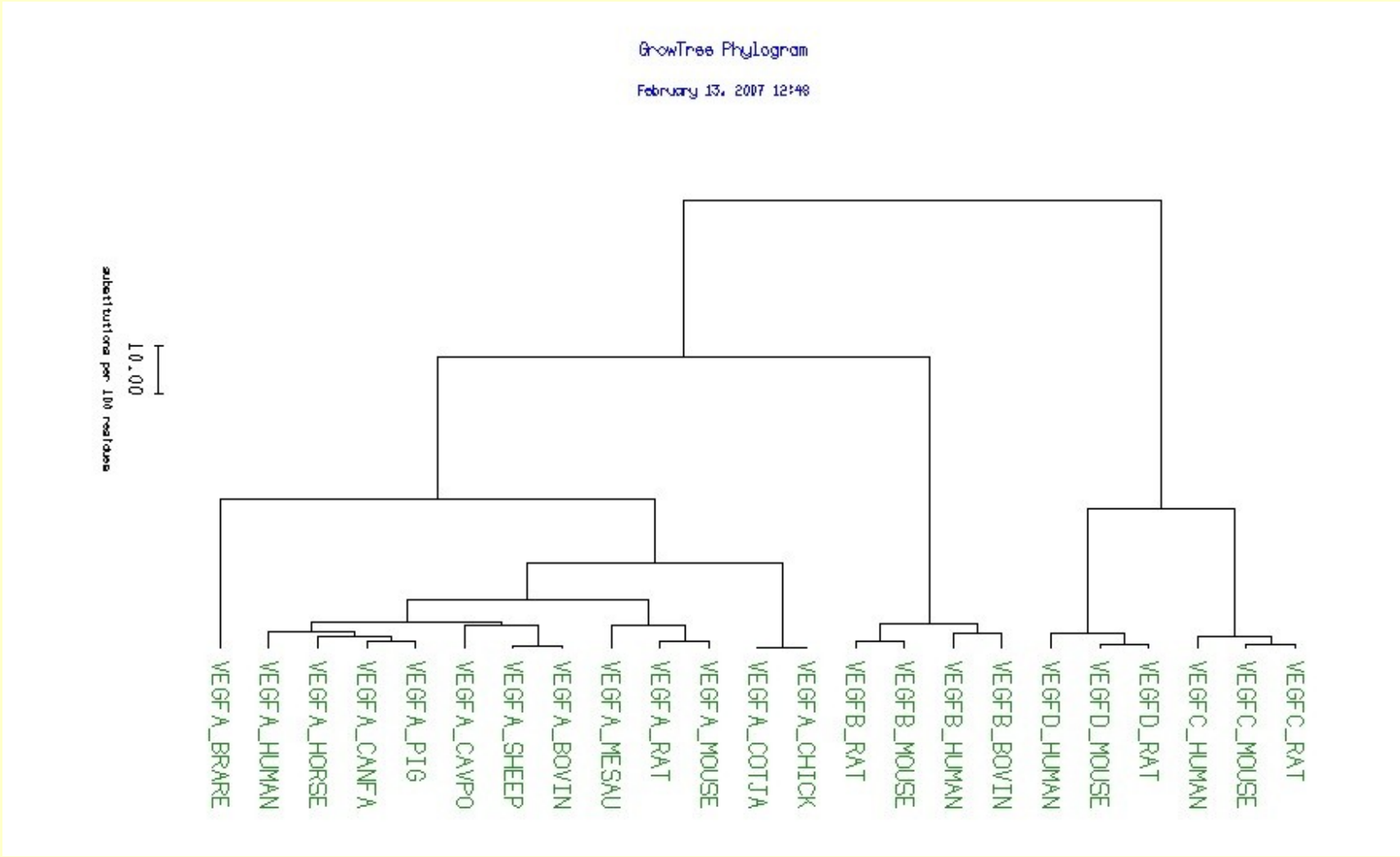
http://en.wikipedia.org/wiki/Vascular_endothelial_growth_factor

Comparison

Type	Function
VEGF-A	▪ Angiogenesis ▪ ↑ Migration of endothelial cells ▪ ↑ mitosis of endothelial cells ▪ ↑ Methane monooxygenase activity ▪ ↑ αvβ3 activity ▪ creation of blood vessel lumen ▪ creates fenestrations ▪ Chemotactic for macrophages and granulocytes ▪ Vasodilation (indirectly by NO release)
VEGF-B	Embryonic angiogenesis
VEGF-C	Lymphangiogenesis
VEGF-D	Needed for the development of lymphatic vasculature surrounding lung bronchioles
PIGF	Important for Vasculogenesis, Also needed for angiogenesis during ischemia, inflammation, wound healing, and cancer.

GrowTree VegF UPGMA Tree

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot



GrowTree VegF Alignment

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=evolution-prot