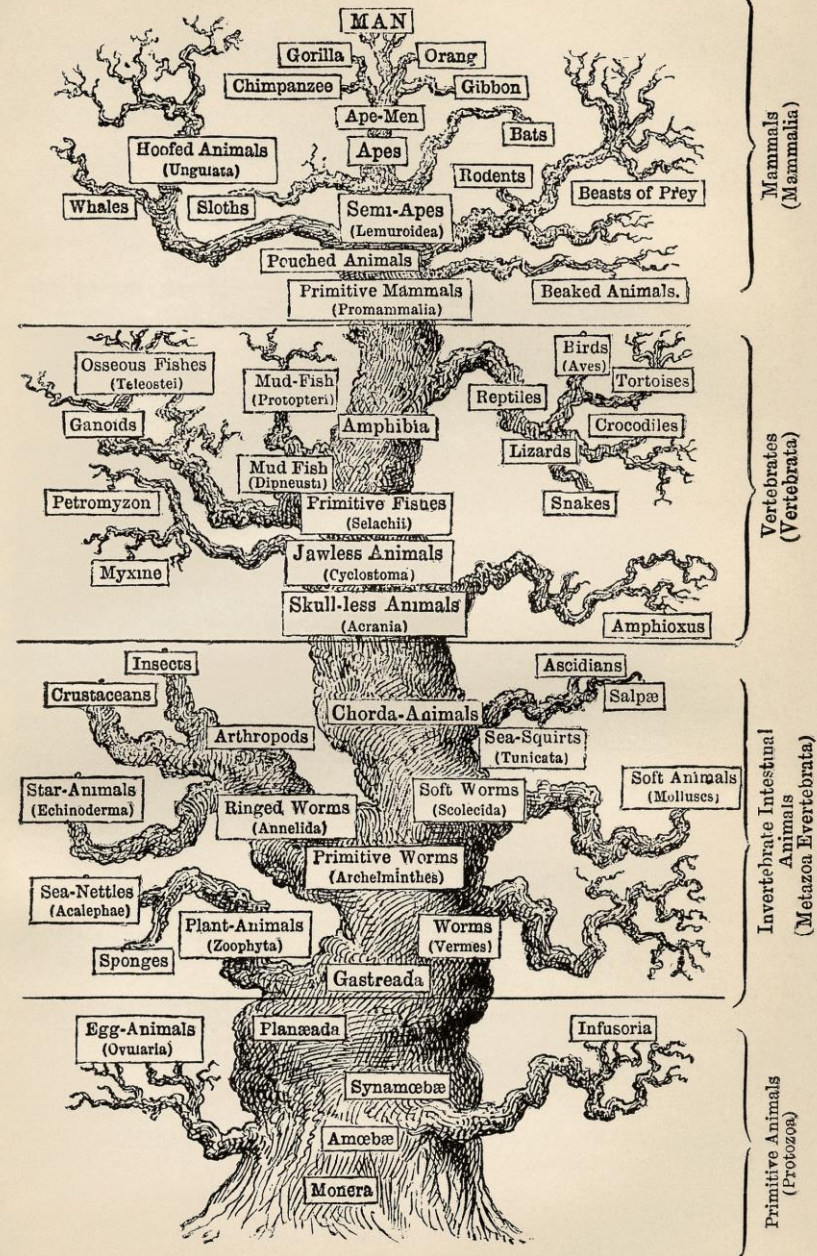


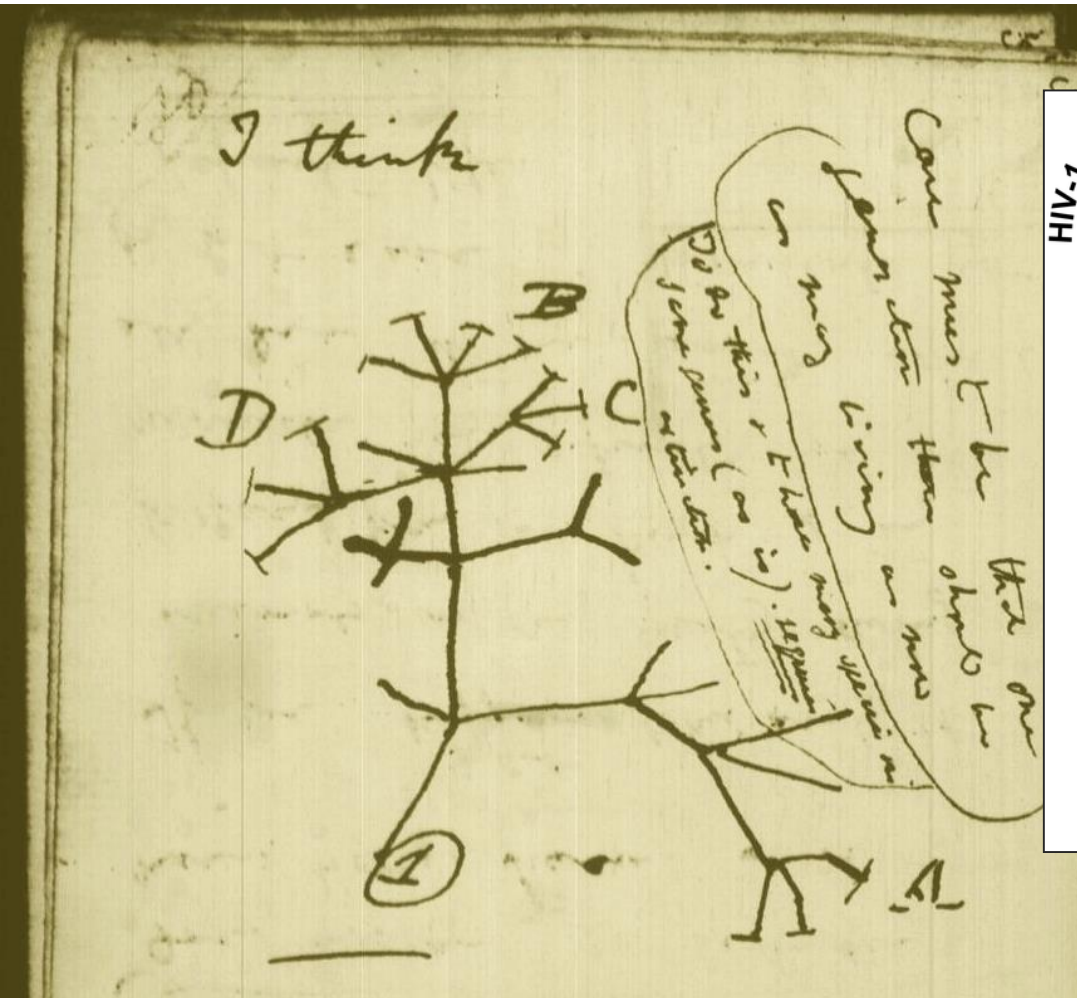
Phylogeny Reconstruction

- a quick overview -

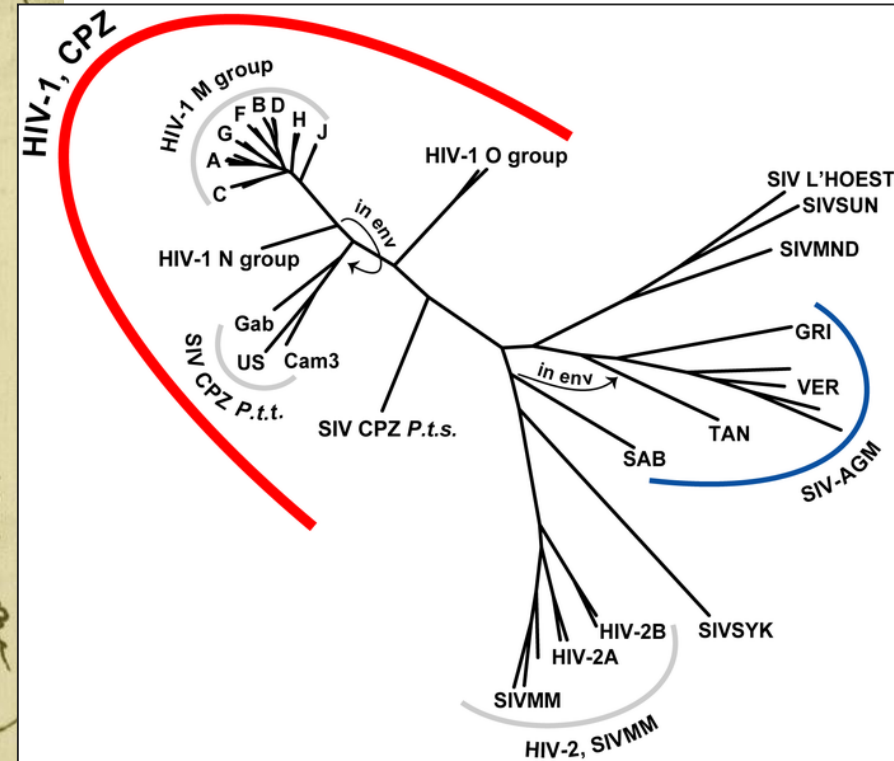
PEDIGREE OF MAN.



Some iconic phylogenetic trees

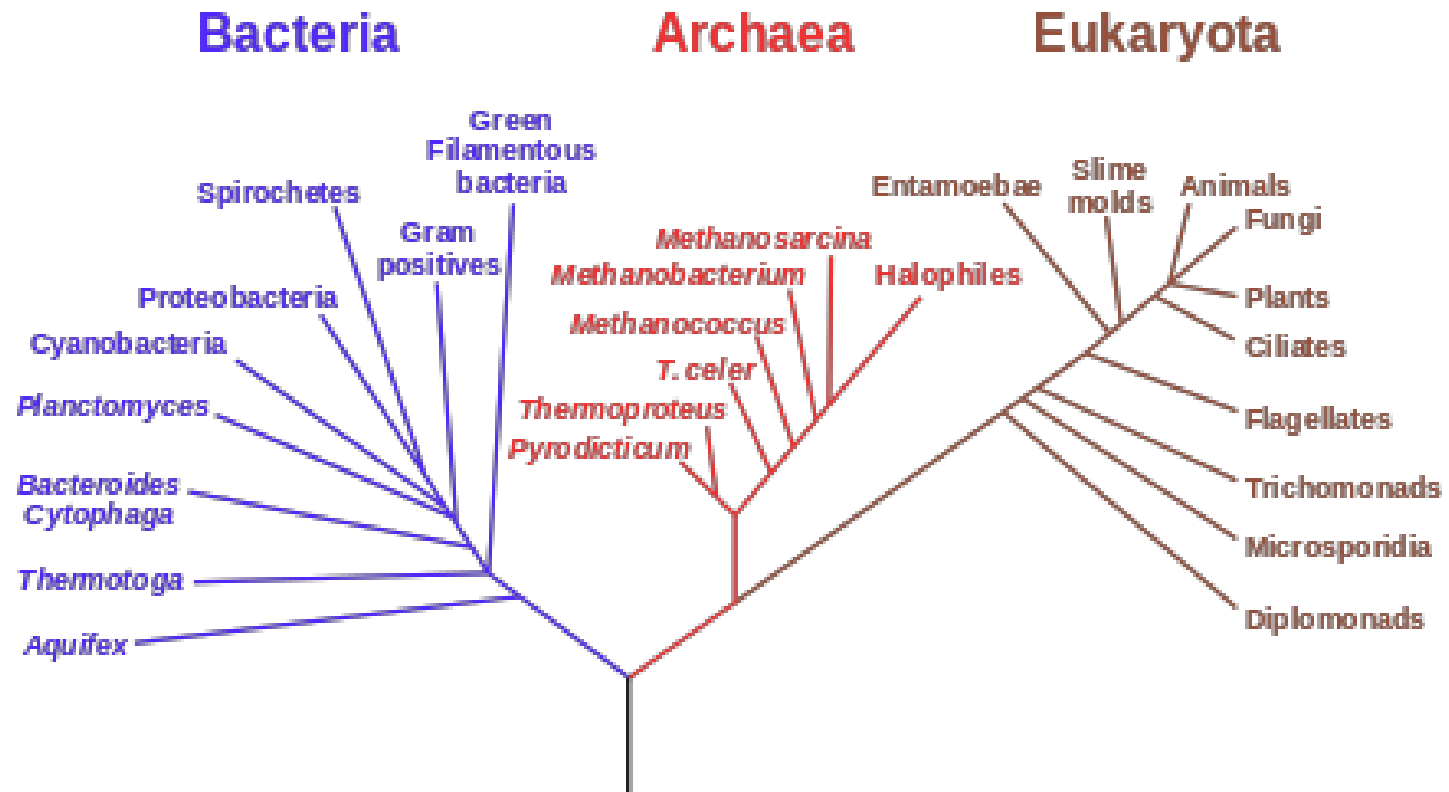


Charles Darwin, *personal notebook*



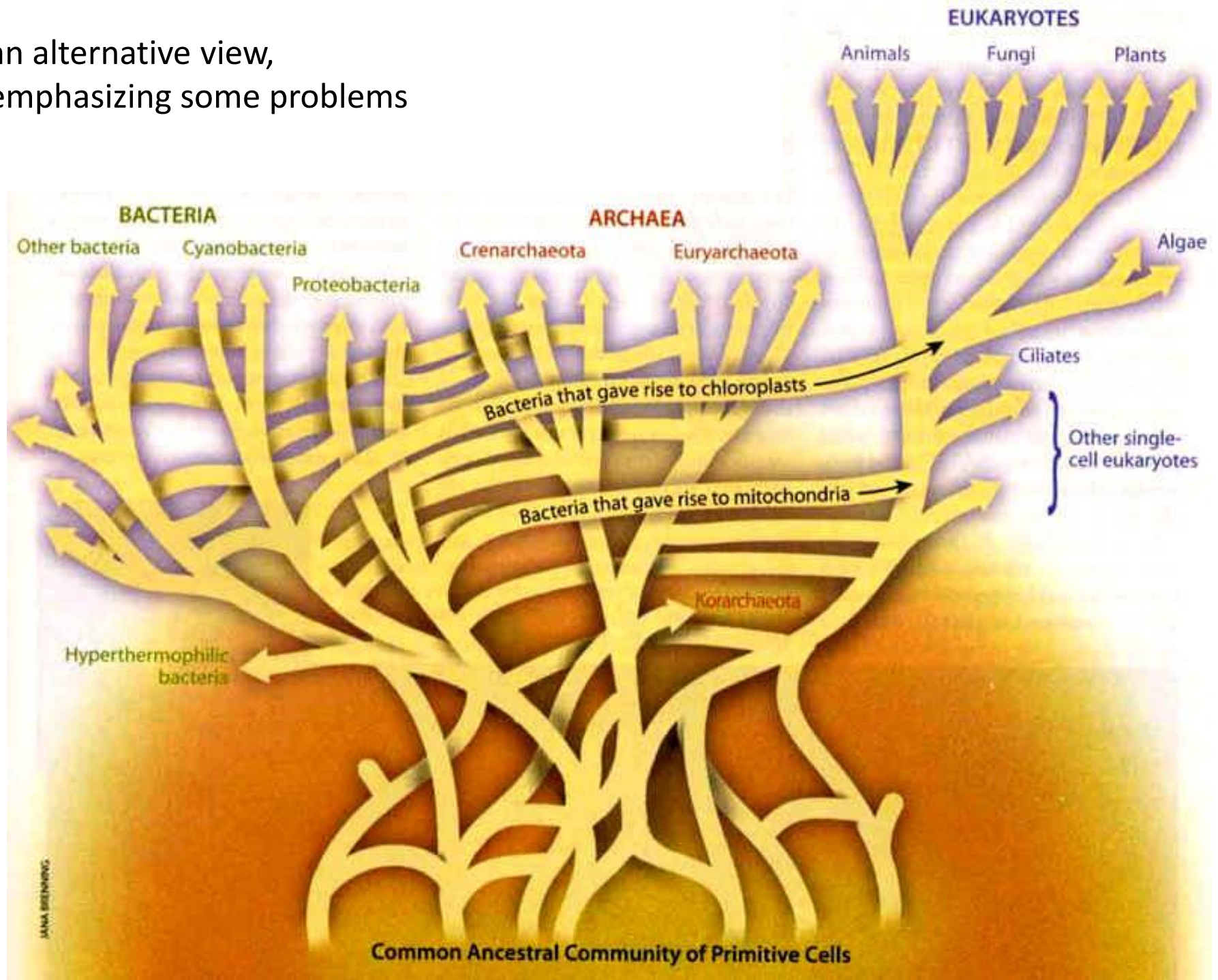
HIV (AIDS),
and closely related
animal viruses.

Some iconic phylogenetic trees





current, conventional version of the 'tree of life'

an alternative view,
emphasizing some problems



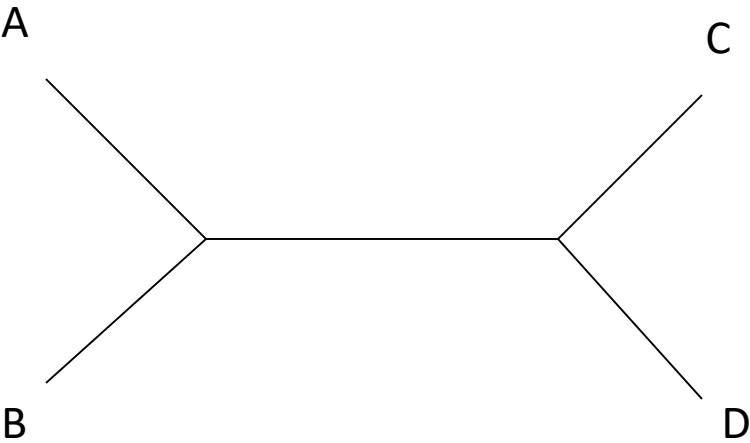
Generating phylogenetic trees

- from gene/protein sequences -

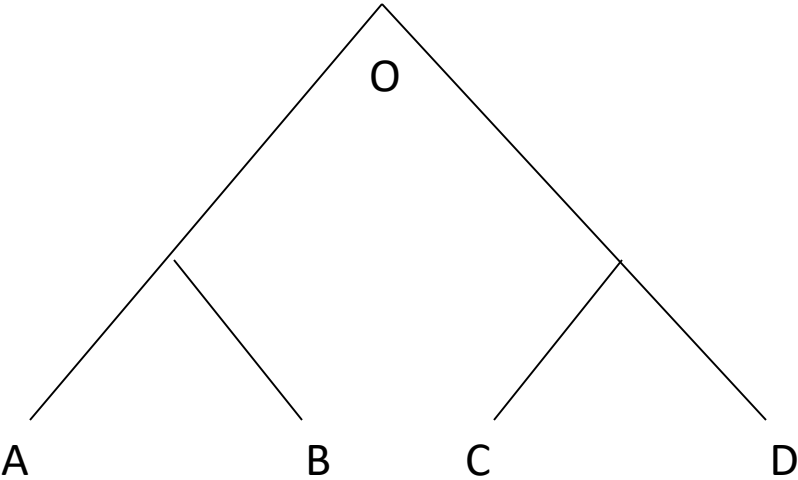
- Phenetic: trees are constructed based on observed characteristics directly, not on evolutionary history  Distance methods
- Cladistic: trees are constructed based on fitting observed characteristics to some model of evolutionary history  Parsimony and Maximum Likelihood methods

Numer of topologies for m taxa

Unrooted tree



Rooted Tree



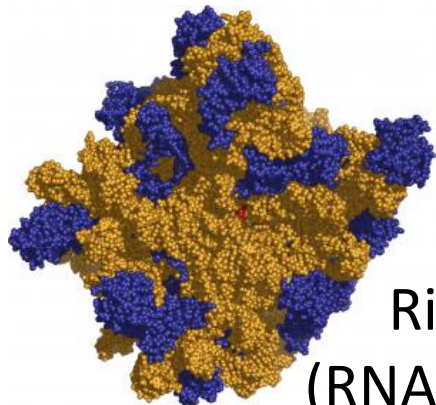
M	Rooted tree	UnRooted Tree
	$(2m-3)! / 2^{m-2}(m-2)!$	$(2m-5)! / 2^{m-3}(m-3)!$

2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025

Which genes to use ?

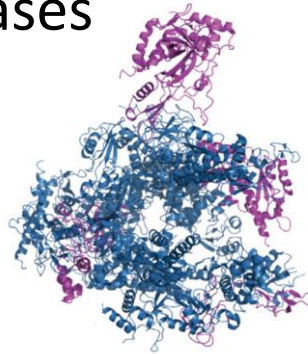
suitable marker genes ...

- ... should occur in every organism**
- ... should rarely undergo horizontal transfer**
- ... should be evolving 'slowly'**
- ... should only occur in one copy per genome**
- ... should function in a process that sees no change**



Ribosomes
(RNA or proteins)

Polymerases



But, for recent events:
fast-evolving genes

Example for a phenetic technique: UPGMA

(Unweighted Pair Group Method with Arithmetic Mean)

1) Alignment

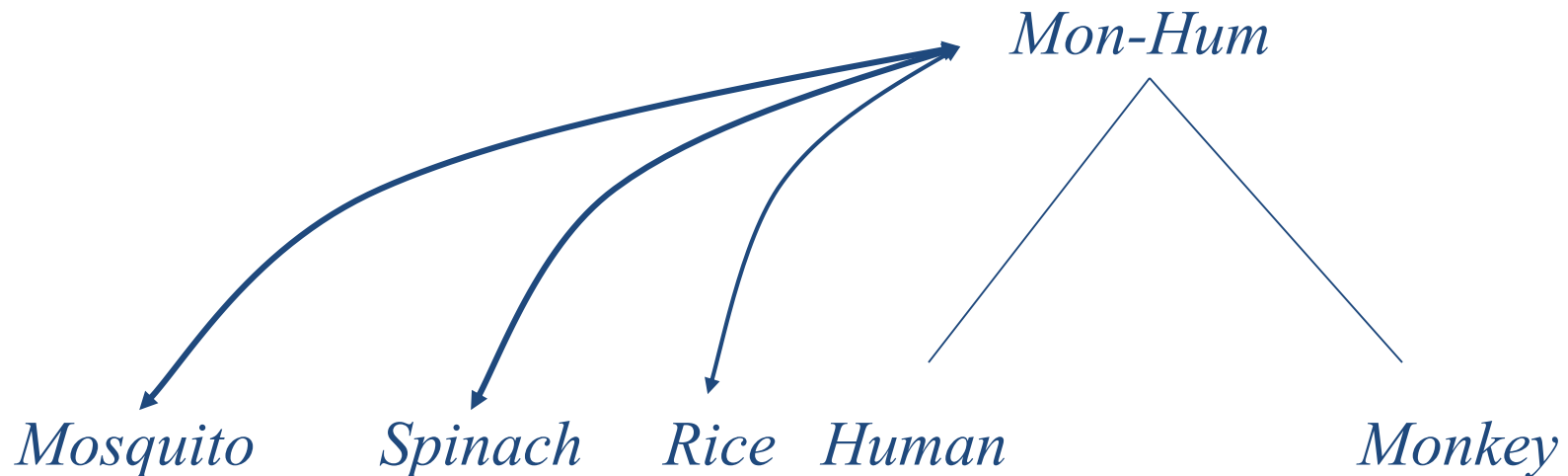
```
FMSLDEVIIVNSGDLILEAFVCMKDNIGGGPVVEGPNKKLVGVSIRDO--IRFLLRPDLF-SNFRQITVMEFMKTIIGS--(15)--GSPDASLGSVIDSASRIITHRIYVVDGQGVVTLRDVISDFI
MIRPSRLVKVRHDEPALKAFRLMRKRGGVGPVVDHAG-KPTGSIIMIKD--VKHLLASDAN-RDYRTLTAEFIANARQ--(10)--CKKEESIKEIFKLDAEKRIYVVDGQGLITLTDIAKLV
LMKCKLVKVNEDOPVLKAFRLMRKRGGVGPVMDTSGTKAIGNISIRDO--VQYLLTAPNIY-KDYRTITAKDFLTAVRQ--(17)--CRRDDEVKDIILKLDSEKTHRIYVIDKGVITLTDIISKLV
KASNRQLRTRSRSTPLNSCLDILLEDVSSIPIVDVNG-ALLDVVSLSD-----IMALGKN-DVYTRIIELEQVIVNEHAL--(14)--CLSTSTFLVLEQLSAPGVRRVVVIEPRGIIISLRDAFTLI
KASNRQLRTRSRSTPLNSCLDILLEDVSSIPIVDVNG-ALLDVVSLSD-----IMALGKN-DVYTRIIELEQVIVNEHAL--(14)--CLSTSTFLVLEQLSAPGVRRVVVIEPRGIIISLRDAFTLI
KASNRQLRTRSRSTPLNSCLDILLEDVSSIPIVDVNG-ALLDVVSLSD-----IMALGKN-DVYTRIIELEQVIVNEHAL--(14)--CLSTSTFLVLEQLSAPGVRRVVVIEPRGIIISLRDAFTLI
KASNRQLRTRSRSTPLNSCLDILLEDVSSIPIVDVNG-ALLDVVSLSD-----IMALGKN-DVYTRIIELEQVIVNEHAL--(14)--CLSTSTFLVLEQLSAPGVRRVVVIEPRGIIISLRDAFTLI
VYVSSKIAVLDAFLPVKQAFIMHDEGLSLVPLWDDQQQTVTGMLTASDFVILRLKLRNIRTLGHEELEMHSVSAWKEA--(20)--VKDSNLRDVALAIRNEISSVPIFKRSGLATLPGIVKFC
TVGKPEVVELHHTDLDAAARAJAASPEGAVPVWPPSARFLGMSALD--IATFVAASGVGDRAMAAVGGEVQPNPGL--(03)--VDPGTRLIDALDLMKQG-VIRFLVRKNGAWRGISKRFVSVLV
IMSKDHIKIKIYEDERVLOAFRLMRKRIGGIPVIERNSEKPVGNISLRD--VQFLLTAPNIY-HDYRSITTKNFLVSVRE--(18)--CTKNHTLKEILMLDAEKIIRIYVVDGFLITLTDIARLV
FMSNEVIEIESEELILEAFVRRMRDNNIGGLPVVEGLNKKIVGNIMDO--IRYLLLOPEMF-SNFRQITVKSFAFKIAT--(10)--CRPDLTGSGVINSLASRSVHRVYVAAAGGVITLTDVISDFV
ESSSKPLATLRHAGLSALALLVQAVSSIPVVDND-SLIDIVRSR--ITA--LAKD--KAYAQIHLDDMVHQAL--(20)--CLRSDSLVKVMERLANPGVRLVIVEAGGIIISLSDVFFLL
GAVNDSVIAITERITVSNAINVMKCALLNAVPIVIAQEDHLOLVNRRHRKVIGTFSATDL--KGCRLPELQTLPLAL--(20)--CGVEITMEAEIKVVTRGVHRVWMDQQGVVSLTDIIRSLR
RLRLSKALTIPOHTIYVEACRMAARVDAVLLTDSNA-LLCGILDKD--ITTRVIAREL--KLEETPVSKVMTRNPLF--(00)--VLSDTLAVEALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLTKAVTIPEGITVAEACRMAARVDAVLLTDANG-LLSGIIVDKD--IAKRVIAGEG--RVEQITISKIMTRTPVF--(00)--VMSDTLAI-EALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLAKALTPEATSVSEACRMAALKRVDAALLTDSNG-MLSGIILAEQ--ISGRVIAEGL--RPDETINAKAMTRNPVF--(00)--VMSNSPAI-EALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLSKALTIPEGITVSEACRMAARVDAVLLTDAQG-LLSGIIVDKD--VATRVVAGEG--RVEQITISKIMTRNPTI--(00)--AMSDTLAI-EALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLSKALTVPATTIYEAACRMAASRVDAALLTDSNE-MLCGIILDKD--IATRVIAQEL--NVEETPVSKVMTRKNPMF--(00)--VLSDTLAVEALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLSKALTINGITVFDACRMAARVDAVLLTDSNA-LLSGIIVDKD--IATRVIAEGL--RPETLVSKVMTRNPIF--(00)--VTSDSLAI-EALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLSKALTIPEGITVFDACRMAARVDAVLLTDSNA-LLSGIIVDKD--IATRVIAEGL--RPDETLVSKVMTRNPIF--(00)--VTSDSLAI-EALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLSKALTVPOSTILFEACRMAARVDAVLLTDSNA-LLCGIILDRD--IATKVIKQGL--NLEETPVSKVMTRKNPVF--(00)--VLSDTLAVEALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
IKRLCKALTVPDSTILFEACRMAARVDAVLLTDSNA-LLCGIILDRD--IATKVIKQGL--NLEETPVSKVMTRKNPVF--(00)--VLSDTLAVEALCKMVGGKFRHLPVVENGEVIALLDIAKOLY
```

2) Distance matrix

PAM	Spinach	Rice	Mosquito	Monkey	Human
Spinach	0.0	84.9	105.6	90.8	86.3
Rice	84.9	0.0	117.8	122.4	122.6
Mosquito	105.6	117.8	0.0	84.7	80.8
Monkey	90.8	122.4	84.7	0.0	3.3
Human	86.3	122.6	80.8	3.3	0.0

First Step

PAM distance 3.3 (Human - Monkey) is the minimum. So we'll join Human and Monkey to MonHum and we'll calculate the new distances.



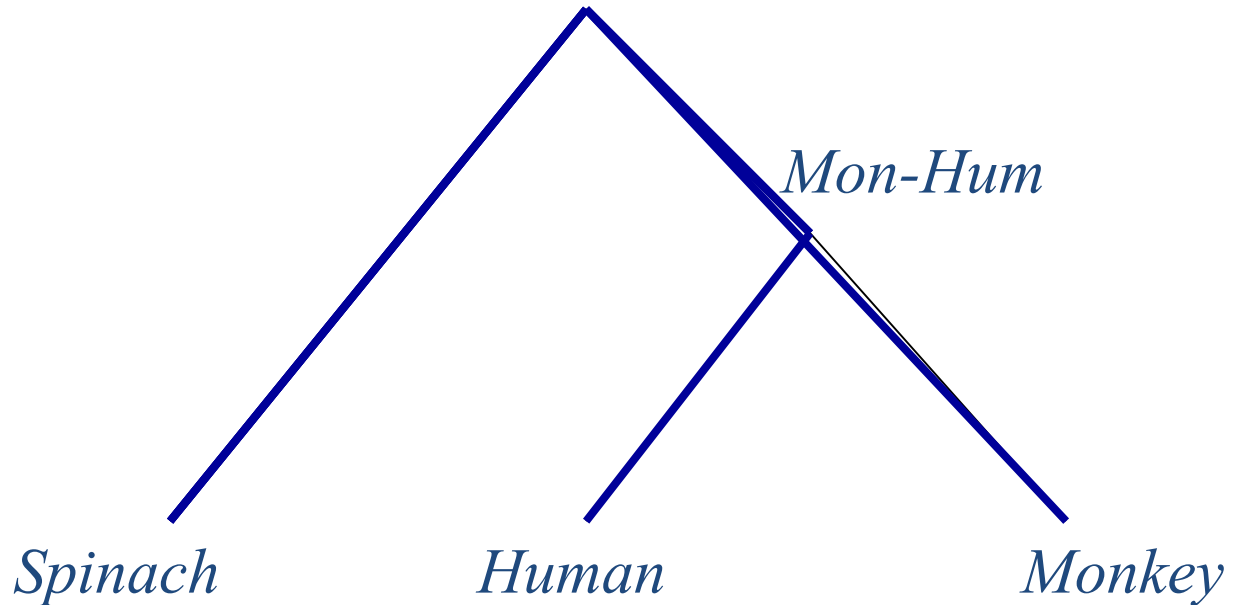
Calculation of new distances

After we have joined two species in a subtree we have to compute the distances from every other node to the new subtree. We do this with a simple average of distances:

$Dist[Spinach, MonHum]$

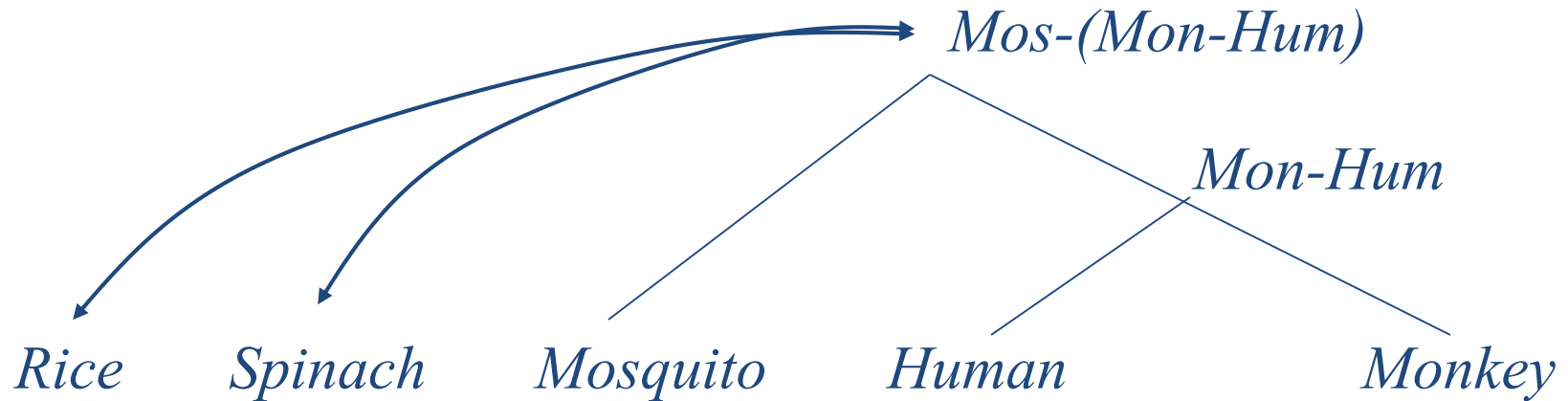
$$= (Dist[Spinach, Monkey] + Dist[Spinach, Human])/2$$

$$= (90.8 + 86.3)/2 = 88.55$$



Next cycle

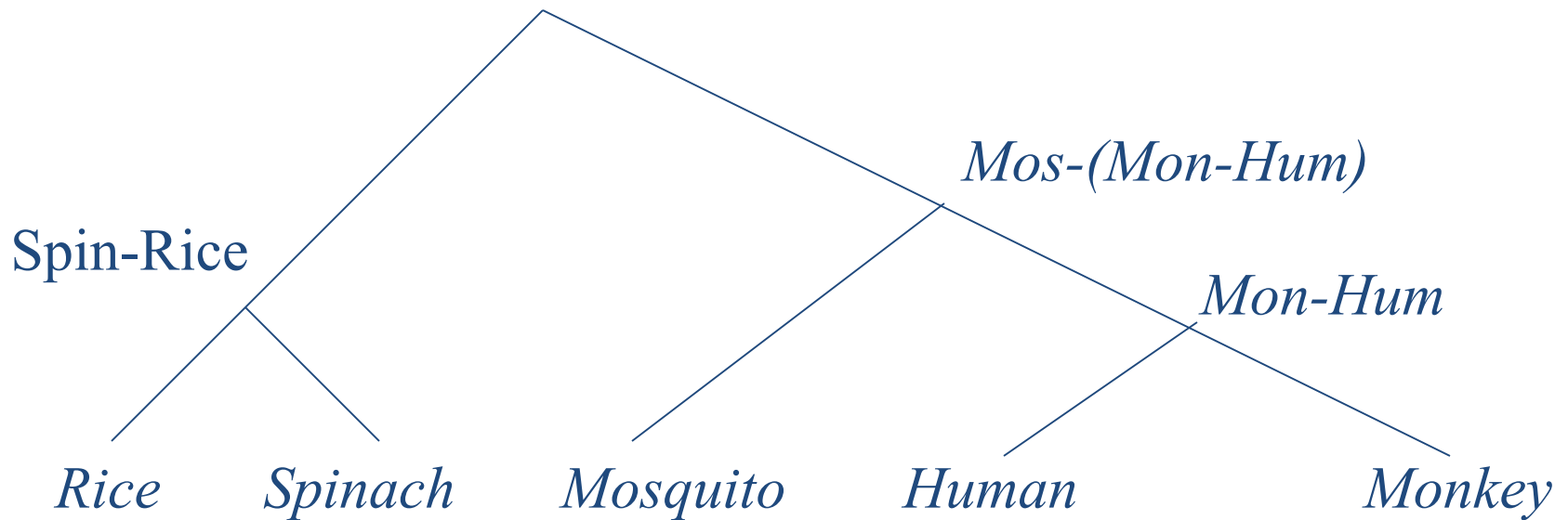
PAM	Spinach	Rice	Mosquito	MonHum
Spinach	0.0	84.9	105.6	88.6
Rice	84.9	0.0	117.8	122.5
Mosquito	105.6	117.8	0.0	82.8
MonHum	88.6	122.5	82.8	0.0



Last joining

PAM	SpinRice	MosMonHum
Spinach	0.0	108.7
MosMonHum	108.7	0.0

(Spin-Rice)-(Mos-(Mon-Hum))



Example for a cladistic technique: Maximum Likelihood

- The likelihood is the probability of the data given the model
- The probability of observing the data under the assumed model will change depending on the parameter values of the model.
- The aim of maximum likelihood is to choose the value of the parameter that maximizes the probability of finding the data.

What is an evolutionary model in this context ?

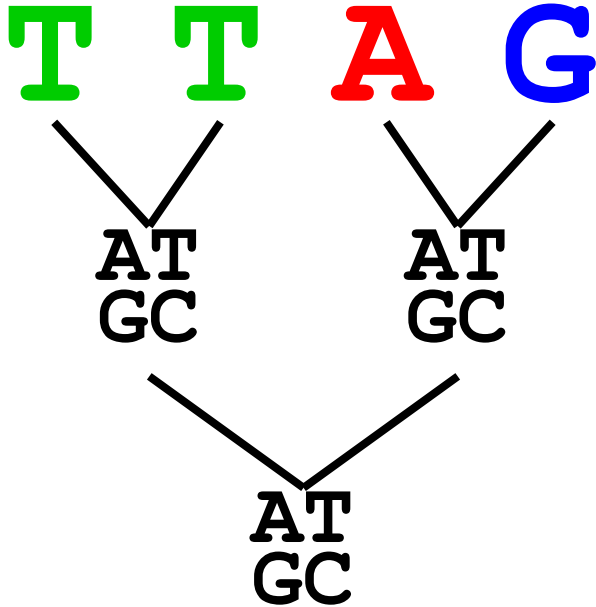
“an empirical matrix describing the relative rates of amino acid replacements”

Dayhoff matrix (Dayhoff et al., 1978)
JTT matrix (Jones et al., 1992)
mtREV matrix (Adachi and Hasegawa, 1996)
WAG matrix (Whelan and Goldman, 2001).

Typically, the model has additional free ‘parameters’:

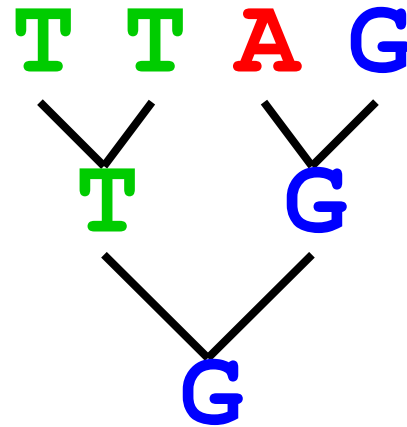
- The rate of evolution can vary across parts of the tree
- The rate of evolution can vary from site to site in the protein

How is maximum likelihood computed ?



1) Image all ancestral possibilities and evolutionary paths.

2) Compute the likelihood of each path



$$L(\text{path}) = L(\text{root}) \times \prod L(\text{branches})$$

$$= P(G \rightarrow T) P(G \rightarrow G) P(G \rightarrow A) P(G \rightarrow G) [\dots]$$

3) multiply all likelihoods over all possible paths

4) throughout, do not forget to optimize all free parameters

5) Repeat for each tree topology, identify the one with best Likelihood

How do we verify a tree?

Difficult ! Very few trees are actually known with certainty

a) Simulation

b) Bootstrapping

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
A T A G C C A T A G C A A C C T
A T A C C C A T G A C A A C G A
A T A C C C A T A G C A A C C A
A T A G C C A T A G C A A C G A
A T C C C C A T A G C A A C C T

The real multiple alignment

2 7 4 9 11 4 16 5
T A G A C G T C
T A C G C C A C
T A C A C C A C
T A G A C G A C
T A C A C C T C

New alignment

