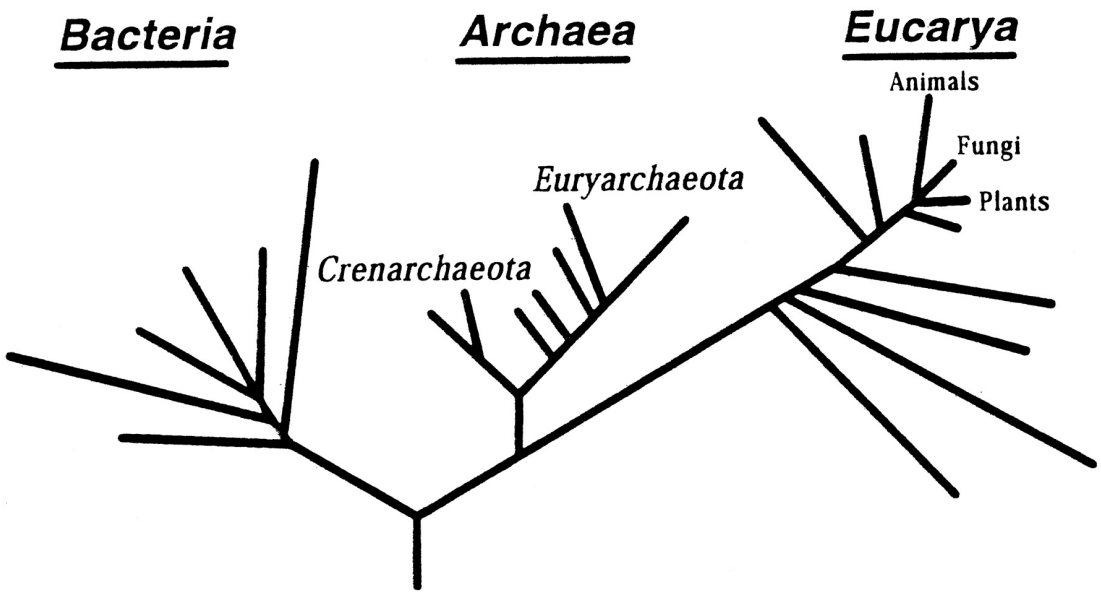
The Tree of Life as we know it, revised

Peter G. Foster

March 20, 2009

The Tree of Life, as we know it



The 'three-domains tree'

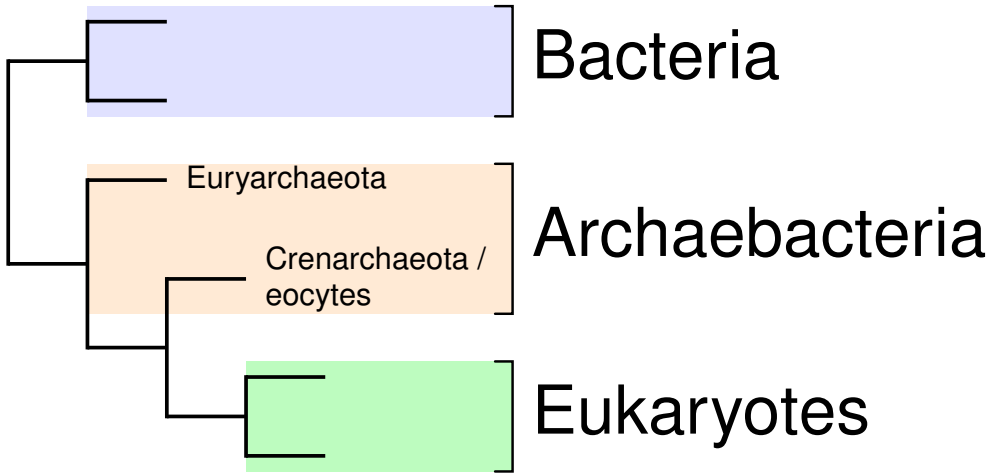
Woese Tree of Life

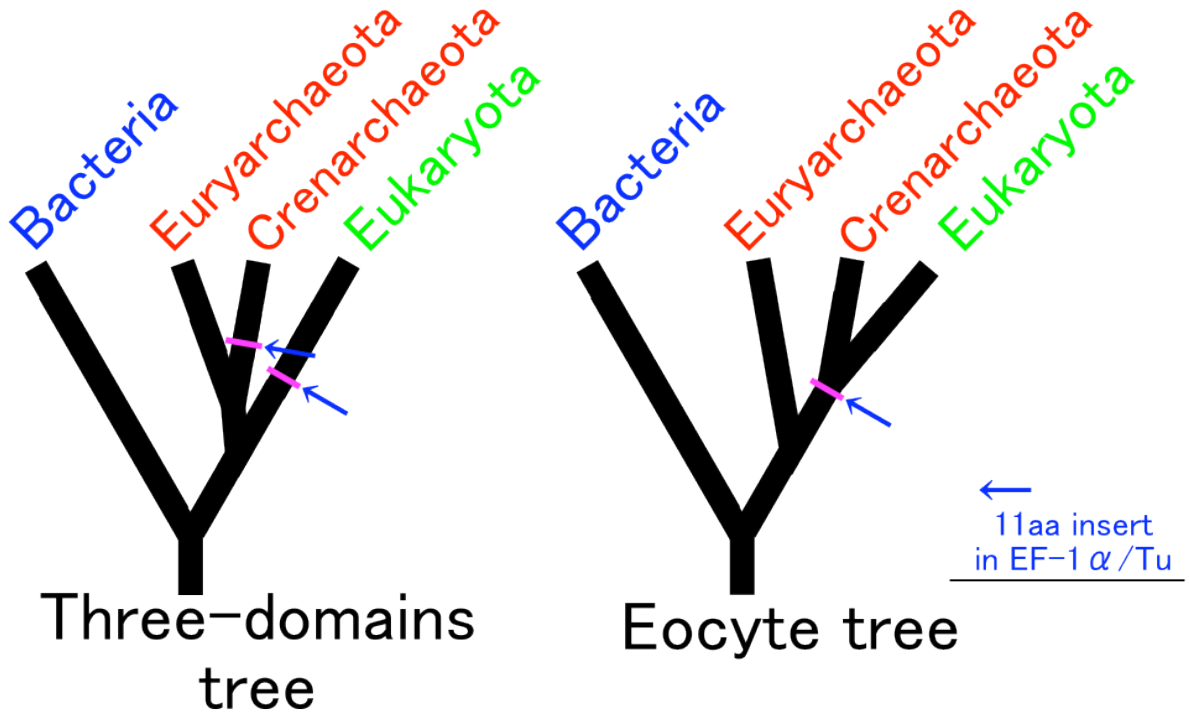
- Based on SSU rRNA genes
- Unrooted tree 1987
- Root inferred by gene duplications
- 1990 picture
- It is now dogma

The eocyte hypothesis

- Proposal that eocytes and eukaryotes are close relatives
- Proposed by Jim Lake and colleagues 1984, based on ribosome structure.
- Further evidence 1988 and 1992
- Generally ignored, and overshadowed by the Woese tree

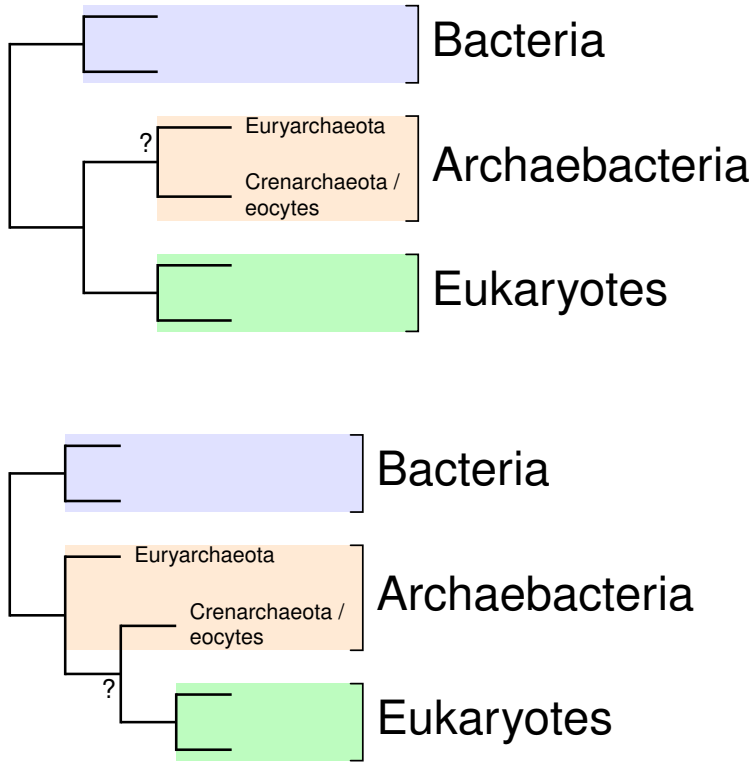
The eocyte hypothesis tree



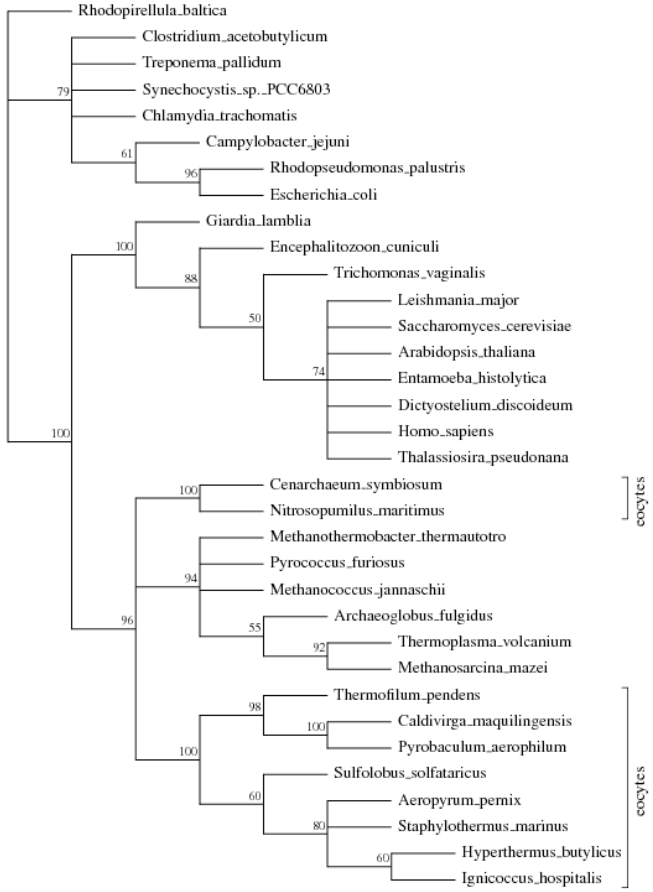


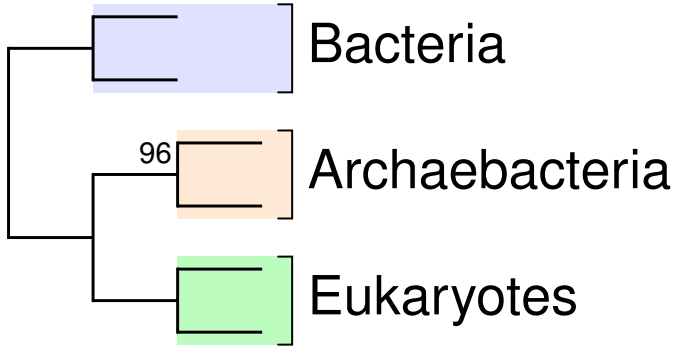
(from Wikipedia)

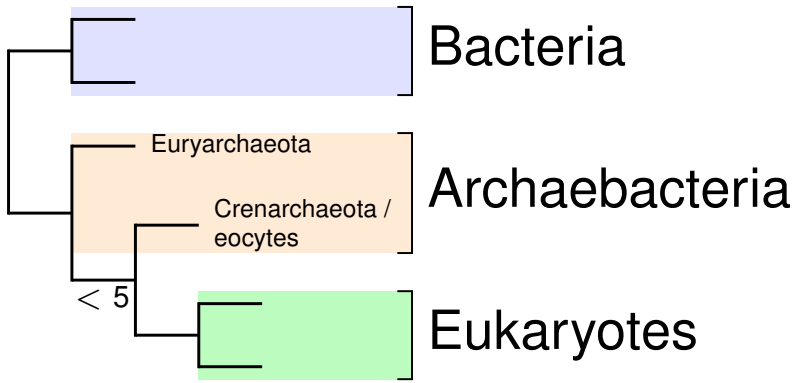
- Re-examine support for these trees of life
- Use better methods
- Use similar data: SSU and LSU rRNA
- Mask it conservatively
- Remove constant sites and singletons
  - They do not contribute to topological resolution
  - Constant sites have a different composition
  - They are difficult to model
- 34 taxa, 1045 chars
- Unrooted



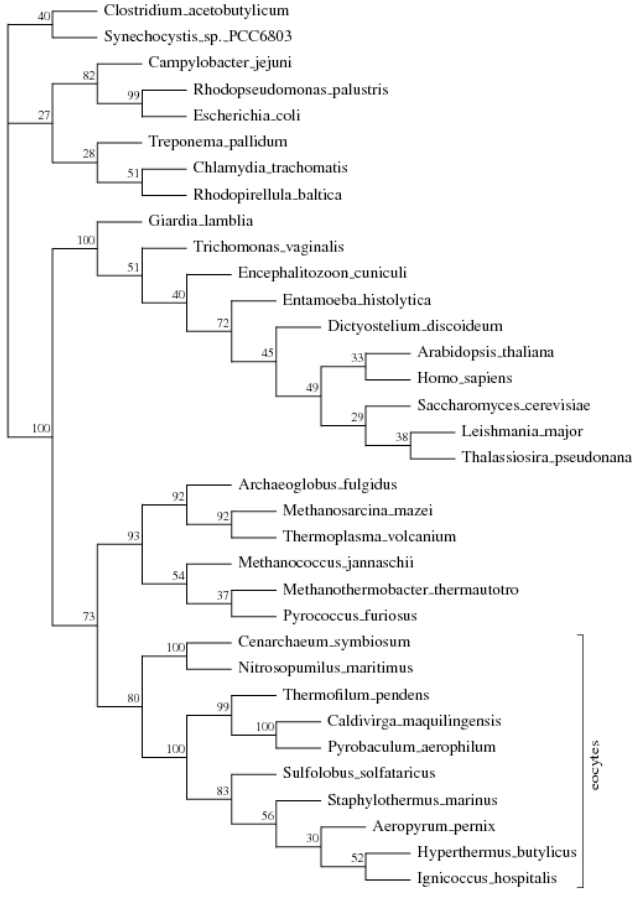
The parsimony tree

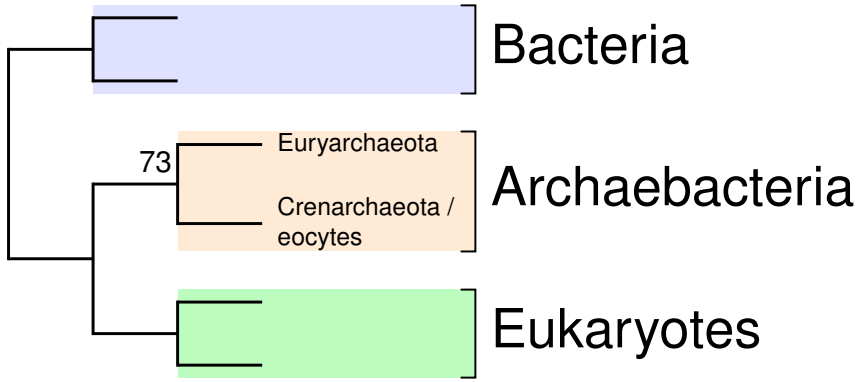


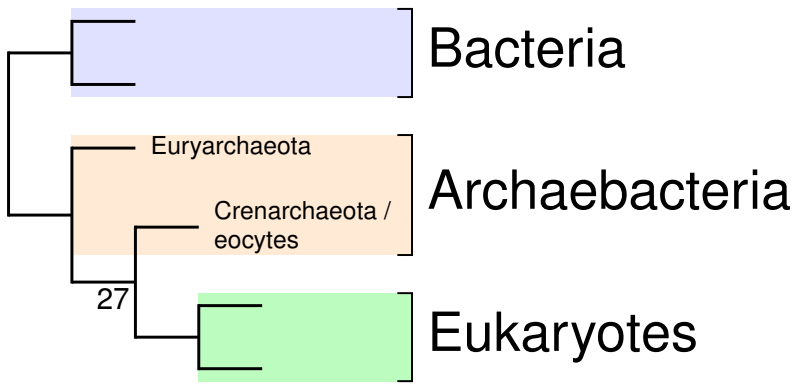




The ML tree from RAxML





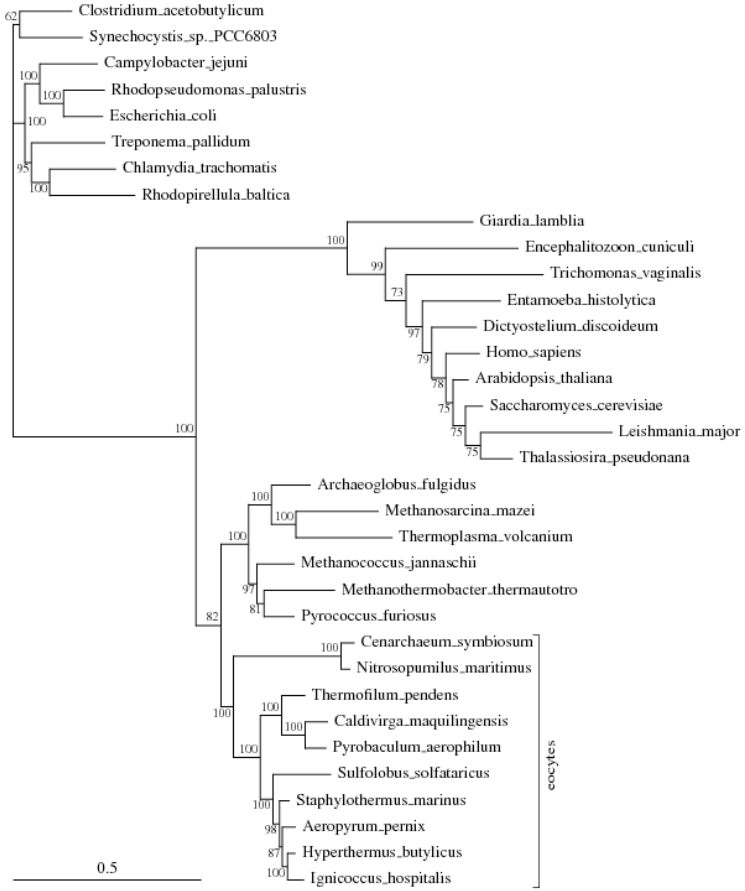


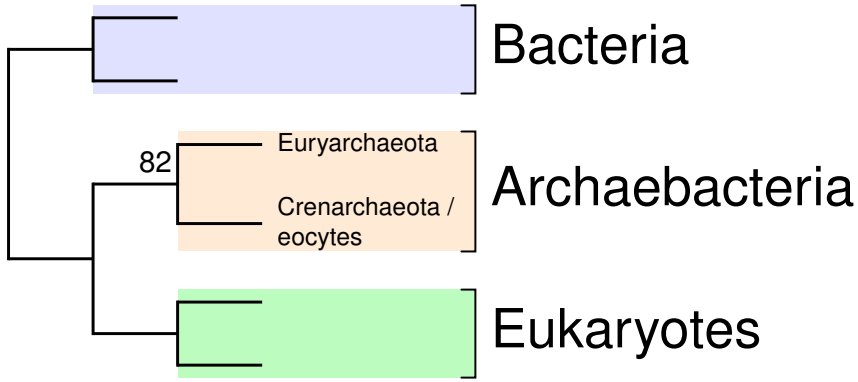
Model using partitioned data

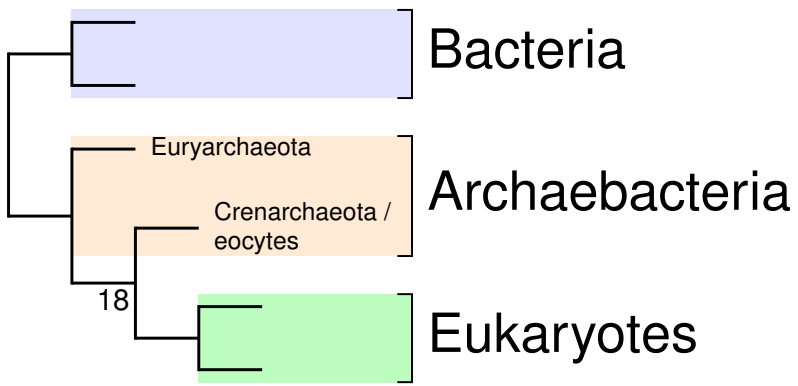
- We are using 2 genes, SSU & LSU rRNA
- If the genes differ in evolutionary dynamics, we can give them separate models

	log Likelihood
one partition	-21758.4
two partitions	-21748.3

The Bayesian tree from MrBayes







Accommodate covarion evolution

- Fitch and Markowitz's "concomitantly variable codons" or "covarion" hypothesis 1970.
- This proposes that at any given time, some sites are invariable due to functional or structural constraints, but that as mutations are fixed elsewhere in the sequence these constraints may change, so that sites that were previously invariable may become variable and vice versa.
  - Biologically realistic
- A simple form of 'heterotachy'
- Modelled with a 2-state on-off switch

Covarion

	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
covarion	-21493	99	1

- So there is lots of evidence for covarion evolution

So what is the problem?

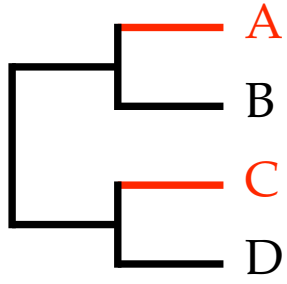
- It appears to be a robust conclusion
- However, the data are compositionally heterogeneous
  - across the tree
  - across the data

The data are heterogeneous over the tree

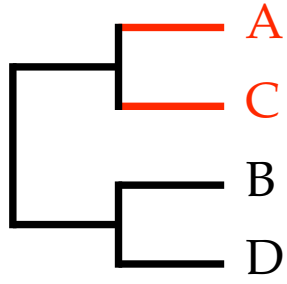
- Test for homogeneity of composition with a χ^2 test
- Both genes fail completely ($P = 0.0$)
- That test suffers from a high probability of Type II error, so if it fails that test then it must be *really* bad

Compositional attraction

- If unrelated taxa have the same compositional bias, they tend to attract on a tree
- For example, if we have 2 T-rich sequences, many T's will be apposed to each other
  - Easily mistaken for recent common ancestry



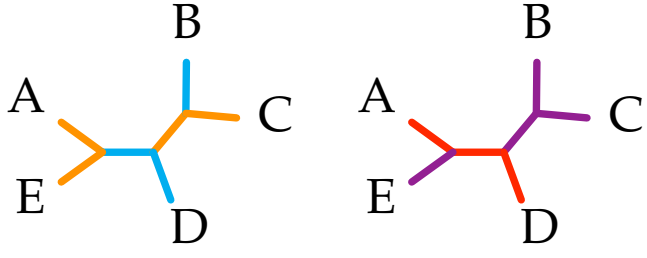
True tree



Recovered tree

Accommodate across-tree composition heterogeneity with the NDCH model

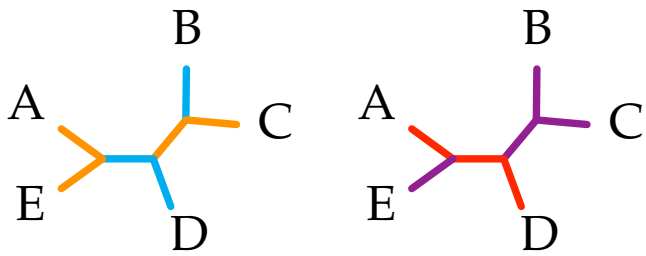
- The Node-discrete composition heterogeneity model accommodates composition heterogeneity over the tree
- You choose a small number of composition vectors, and each branch has one of them at a time



partition 1 partition 2

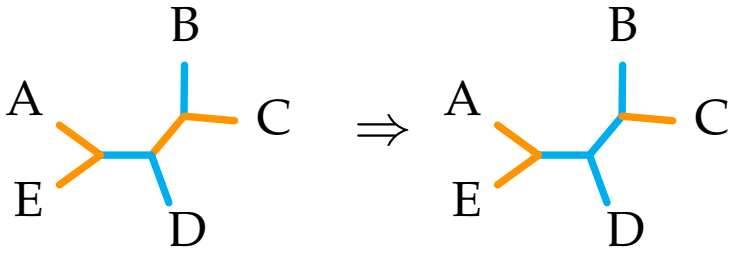
NDCH model

		A	C	G	T
partition 1	orange	0.2	0.3	0.4	0.1
	blue	0.1	0.2	0.3	0.4
partition 2	red	0.3	0.4	0.1	0.2
	violet	0.4	0.1	0.2	0.3

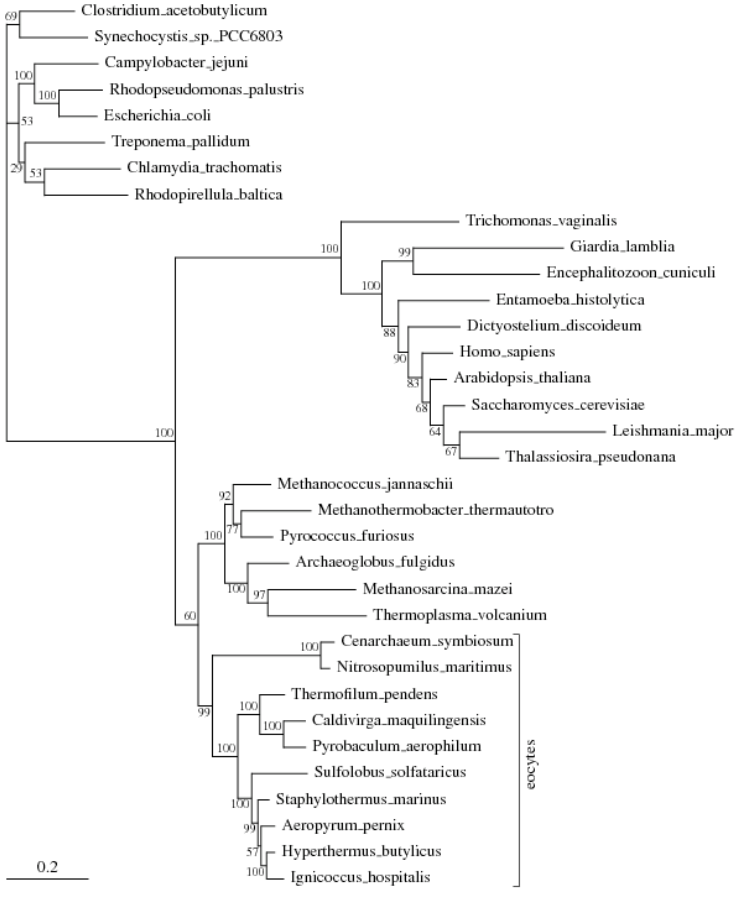


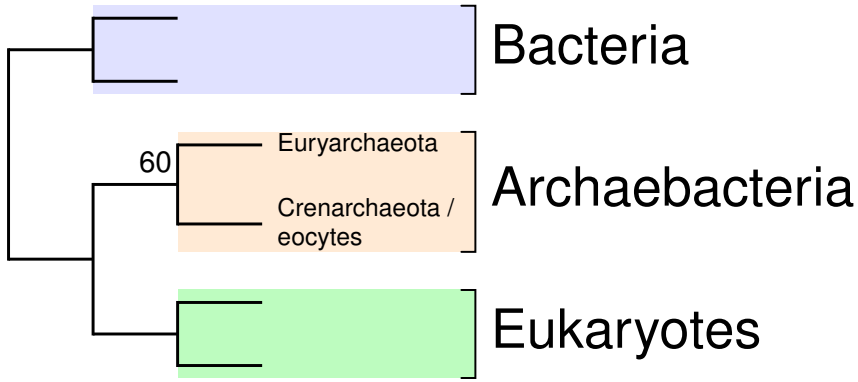
partition 1 partition 2

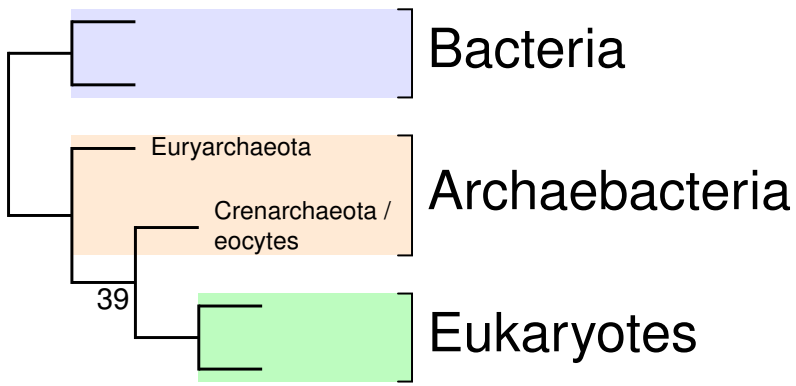
Bayesian NDCH model in an MCMC



NDCH(2,2)

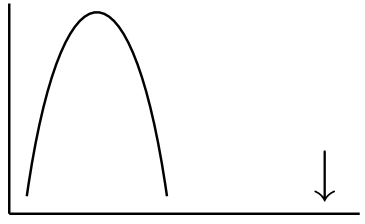

Clostridium_acetobutylicum
Synechocystis_sp._PCC6803
Campylobacter_jejuni
Rhodopseudomonas_palustris
Escherichia_coli
Treponema_pallidum
Chlamydia_trachomatis
Rhodopirellula_baltica
Trichomonas_vaginalis
Giardia_lamblia
Encephalitozoon_cuniculi
Entamoeba_histolytica
Dictyostelium_discoideum
Homo_sapiens
Arabidopsis_thaliana
Saccharomyces_cerevisiae
Leishmania_major
Thalassiosira_pseudonana
Methanococcus_jannaschii
Methanothermobacter_thermautotro
Pyrococcus_furiosus
Archaeoglobus_fulgidus
Methanosarcina_mazei
Thermoplasma_volcanium
Cenarchaeum_symbiosum
Nitrosopumilus_maritimus
Thermofilum_pendens
Caldivirga_maquilingensis
Pyrobaculum_aerophilum
Sulfolobus_solfataricus
Staphylothermus_marinus
Aeropyrum_pernix
Hyperthermus_butylicus
Ignicoccus_hospitalis
eocytes
0.2


Bacteria
60
Euryarchaeota
Crenarchaeota / eocytes
Archaebacteria
Eukaryotes

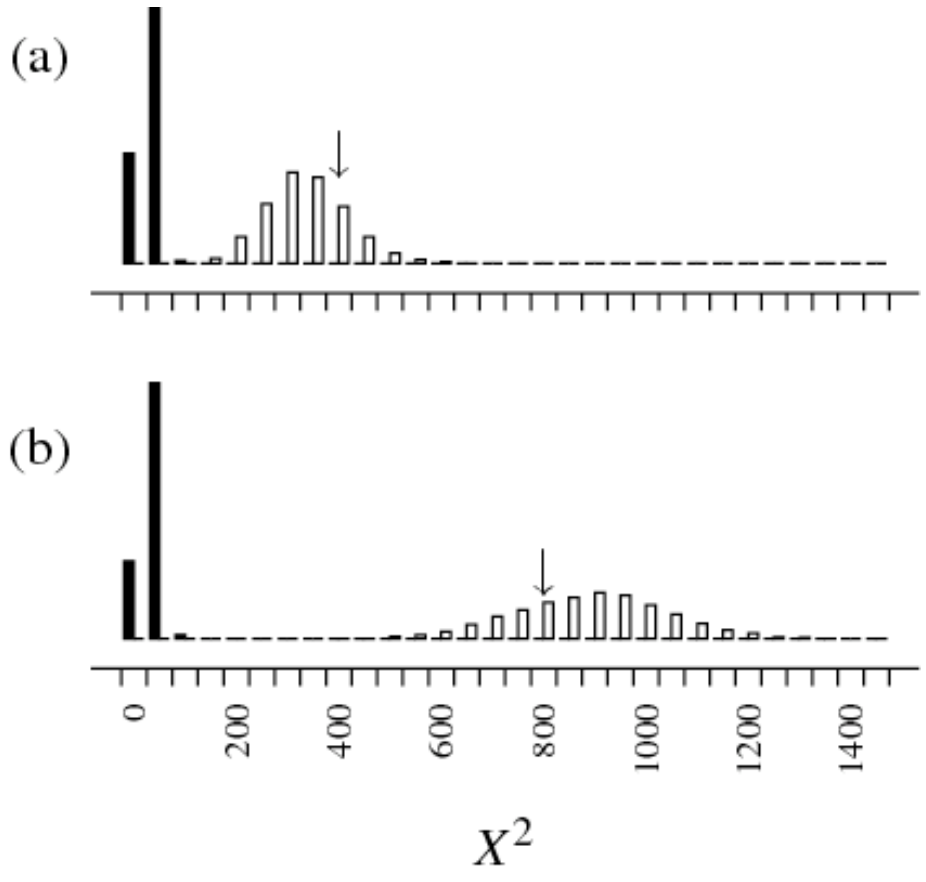

Bacteria
Euryarchaeota
Crenarchaeota / eocytes
Archaebacteria
39
Eukaryotes

NDCH(2,2) has a better fit

- Use posterior predictive simulation
- Use X^2 as the test quantity, to show compositional heterogeneity
- The original data has a X^2 value
- Simulate data based on the posterior samples
- If the X^2 from the original data falls within the simulation distribution, then the model fits (by this test)



NDCH(2,2) has a better fit



NDCH(2,2) has a better fit

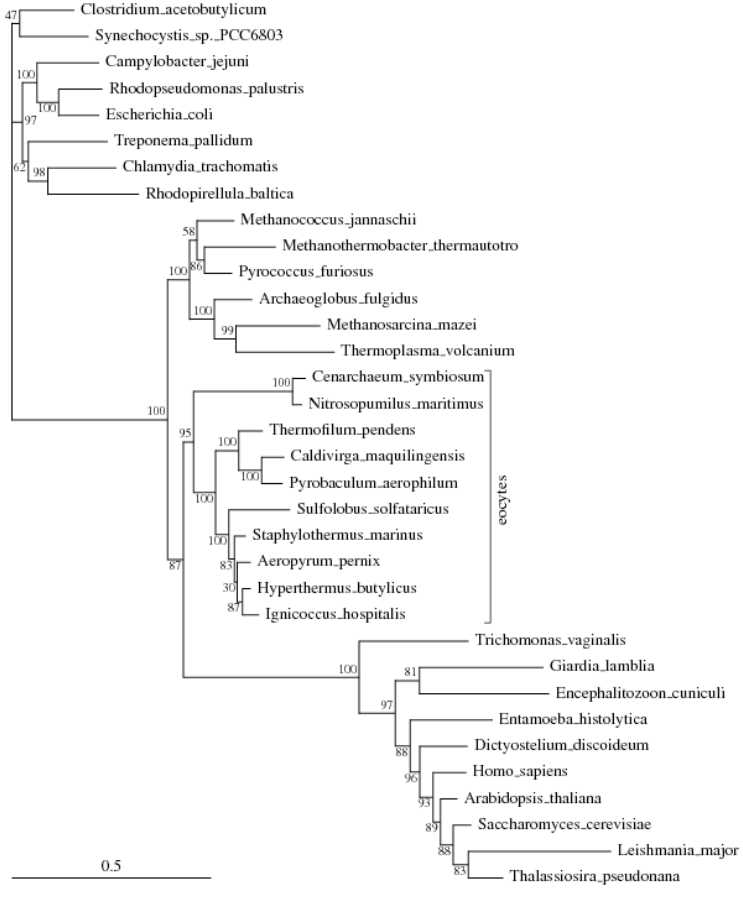
	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
NDCH(2,2)	-21373	60	39

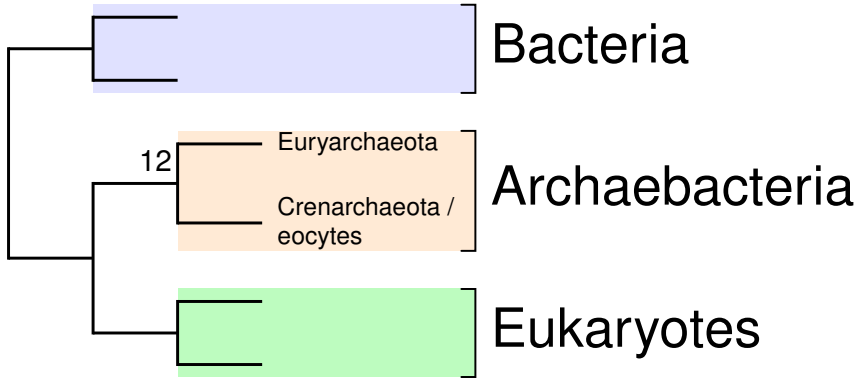
NDCH(2,2) has a better fit

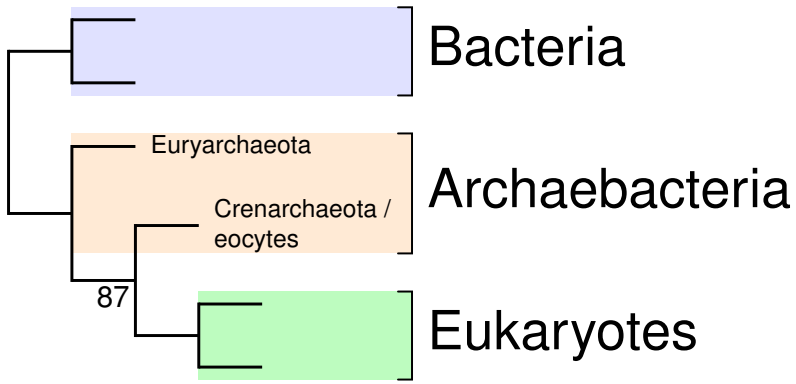
	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
covarion	-21493	99	1
NDCH(2,2)	-21373	60	39

The improvement here is bigger than the improvement that was given by the covarion model

NDCH(4,4)


Clostridium_acetobutylicum
Synechocystis_sp._PCC6803
Campylobacter_jejuni
Rhodopseudomonas_palustris
Escherichia_coli
Treponema_pallidum
Chlamydia_trachomatis
Rhodopirellula_baltica
Methanococcus_jannaschii
Methanothermobacter_thermautotro
Pyrococcus_furiosus
Archaeoglobus_fulgidus
Methanosarcina_mazei
Thermoplasma_volcanium
Cenarchaeum_symbiosum
Nitrosopumilus_maritimus
Thermofilum_pendens
Caldivirga_maquilingensis
Pyrobaculum_aerophilum
Sulfolobus_solfataricus
Staphylothermus_marinus
Aeropyrum_pernix
Hyperthermus_butylicus
Ignicoccus_hospitalis
eocytes
Trichomonas_vaginalis
Giardia_lamblia
Encephalitozoon_cuniculi
Entamoeba_histolytica
Dictyostelium_discoideum
Homo_sapiens
Arabidopsis_thaliana
Saccharomyces_cerevisiae
Leishmania_major
Thalassiosira_pseudonana
0.5


Bacteria
12
Euryarchaeota
Crenarchaeota /
eocytes
Archaebacteria
Eukaryotes


Bacteria
Euryarchaeota
Crenarchaeota /
eocytes
Archaebacteria
87
Eukaryotes

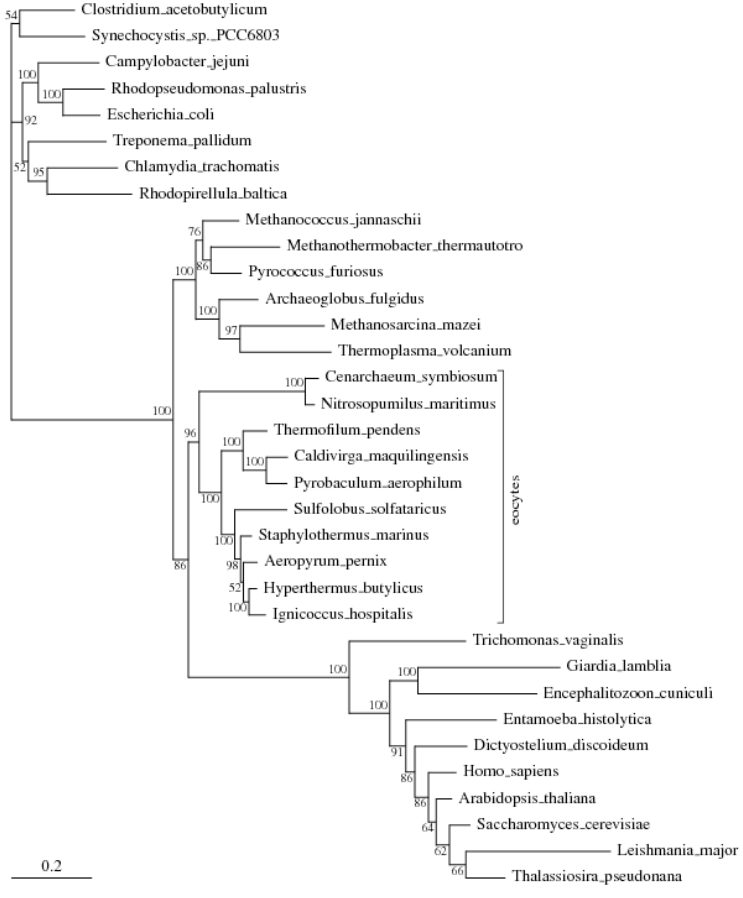
NDCH(4,4) has a better fit

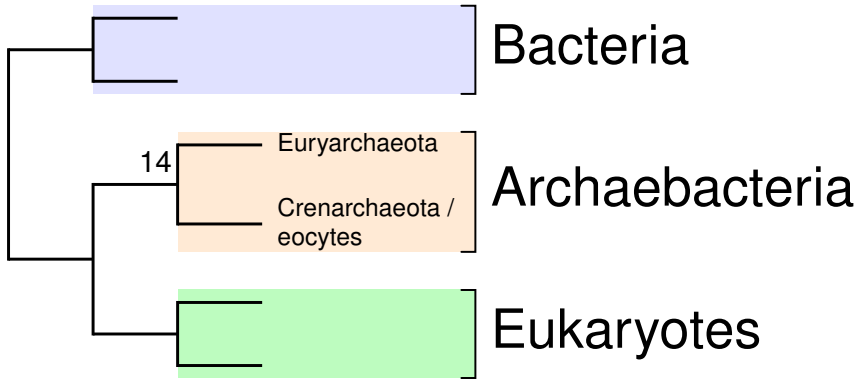
	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
NDCH(2,2)	-21373	60	39
NDCH(4,4)	-21291	12	87

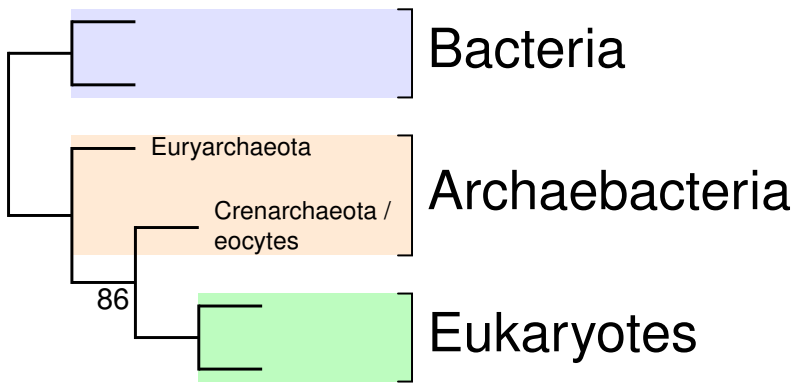
Accommodate across-tree rate matrix heterogeneity with the NDRH model

- Node discrete rate heterogeneity model.
- Like the NDCH model, but allows 2 or more independent GTR rate matrices over the tree
- Also separate in each data partition

NDCH(2,2), NDRH(2,2)



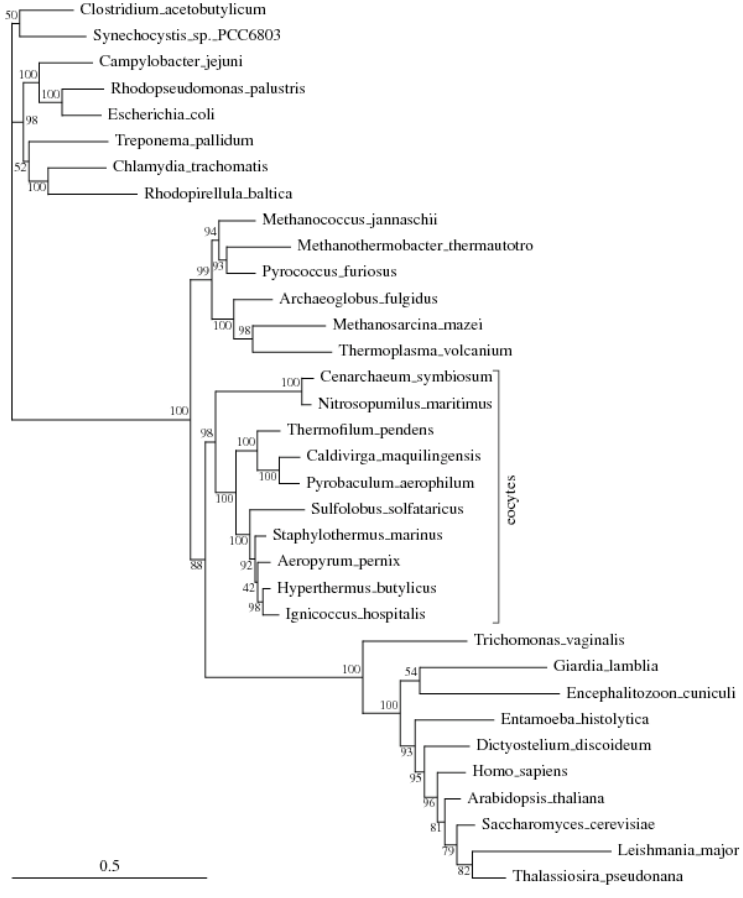


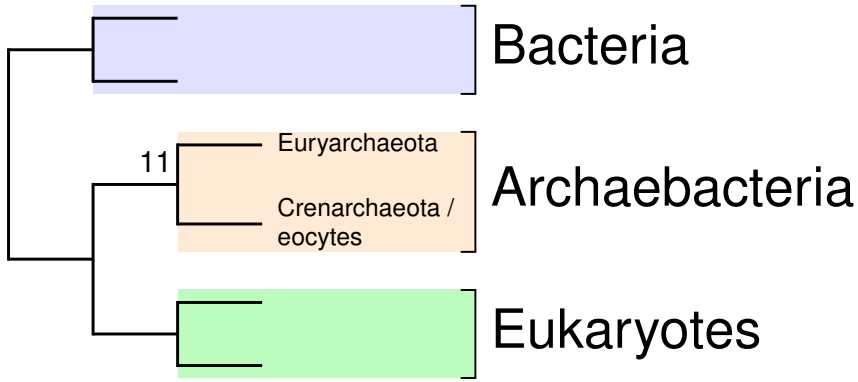


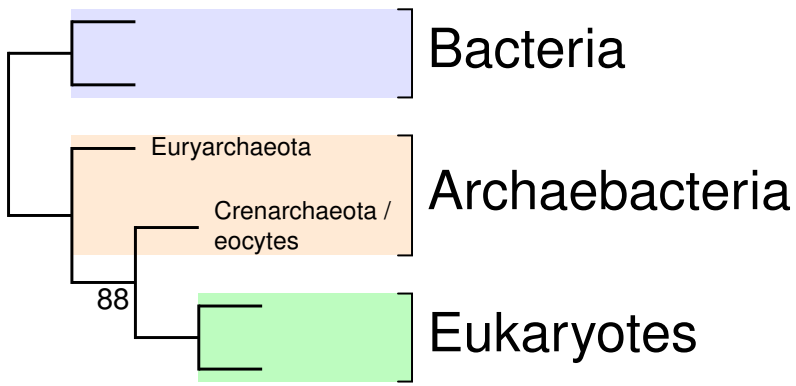
NDCH(2,2), NDRH(2,2)

	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
NDCH(2,2)	-21373	60	39
NDCH(4,4)	-21291	12	87
NDCH(2,2), NDRH(2,2)	-21288	14	86

NDCH(4,4), NDRH(2,2)







NDCH(4,4), NDRH(2,2)

	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
NDCH(2,2)	-21373	60	39
NDCH(4,4)	-21291	12	87
NDCH(2,2), NDRH(2,2)	-21288	14	86
NDCH(4,4), NDRH(2,2)	-21221	11	88

Accommodate across-data composition heterogeneity with the CAT model

- The CAT model was originally made for protein
- It was noticed that sites in real data had composition profiles that did not reflect what you would expect based on a rate matrix such as the JTT, WAG, ...
- Sites often had simpler AA comps, as if their composition was restricted.
- Model this with a mixture model of simple poisson processes — CAT
- Lartillot and Philippe 2004

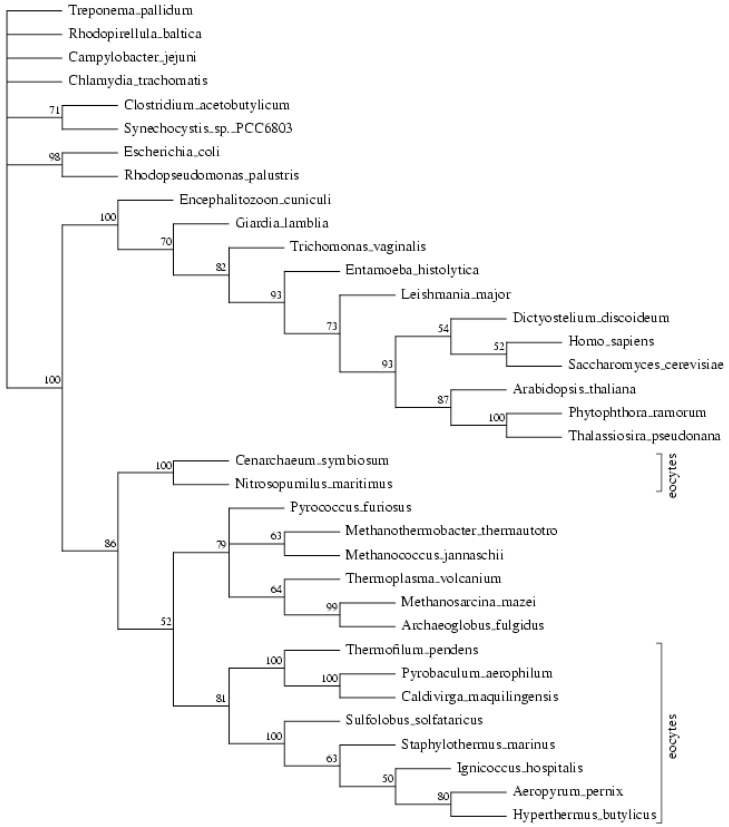
CAT

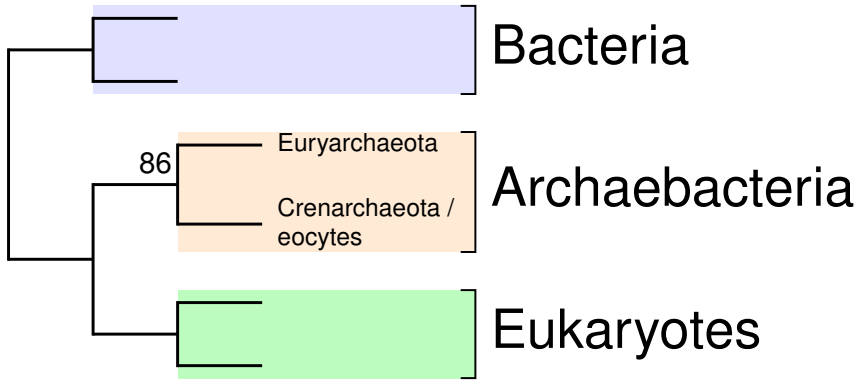
	log marginal likelihood	3-domains	eocyte
homog	-21717	82	18
NDCH(2,2)	-21373	60	39
NDCH(4,4)	-21291	12	87
NDCH(2,2), NDRH(2,2)	-21288	14	86
NDCH(4,4), NDRH(2,2)	-21221	11	88
CAT	-19948	0	100

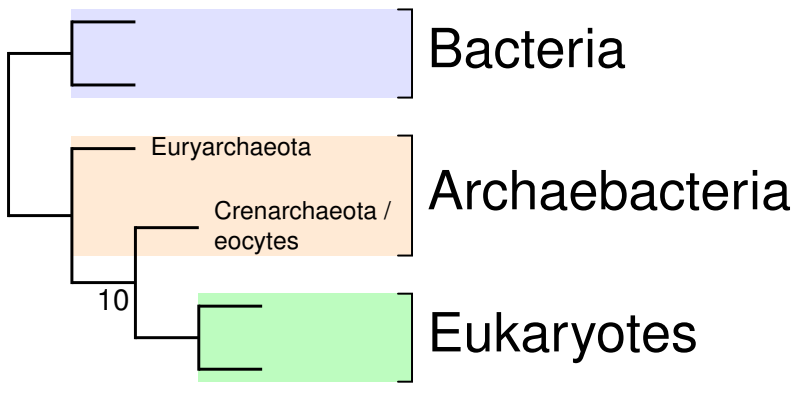
Using protein data

- 35 taxa, 41 proteins
- conservatively masked
- Constant sites and singletons removed
- Some analyses used Dayhoff recoded data
- 5222 sites
- 4008 sites when Dayhoff-recoded

Protein MP



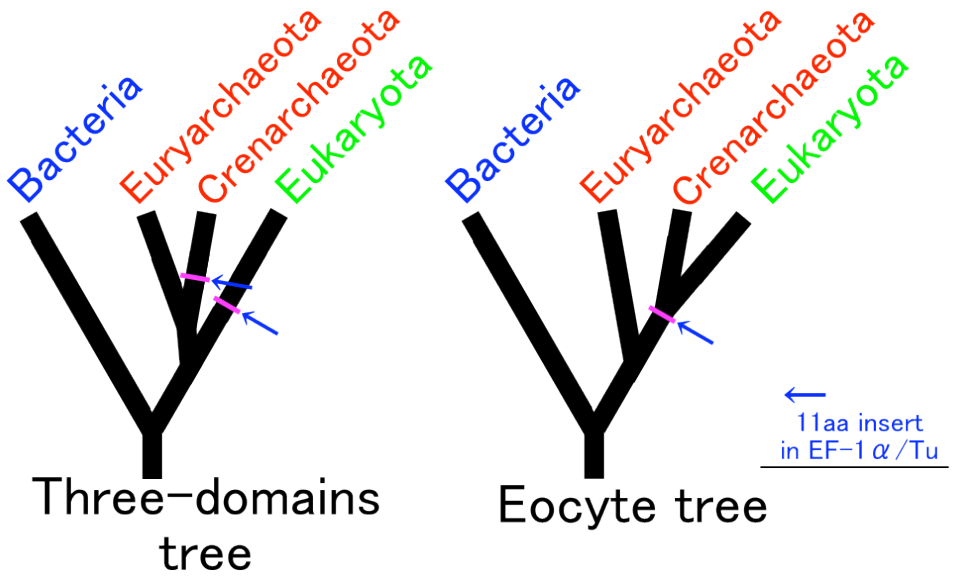




(88% when using Dayhoff-recoded data)

Other protein analyses

	log marginal likelihood	eocyte
MP		10
WAG+G	-246692	99
WAG+G+Covarion	-246690	100
CAT+G	-220644	100
MP		6
GTR+G	-106068	98
NDCH(14)	-105488	99
CAT+G	-98756	100



- rRNA genes do not support the 3-domains tree
  - The eocyte tree is better supported
- Multi-gene protein analysis supports the eocyte tree
- Eukaryotes originated within the archaebacteria
  - are not a “primordial” lineage

Collaborators

- Cymon Cox
- Martin Embley
- Robert Hirt
- Simon Harris