Estimation of evolutionary distance based on K-string composition

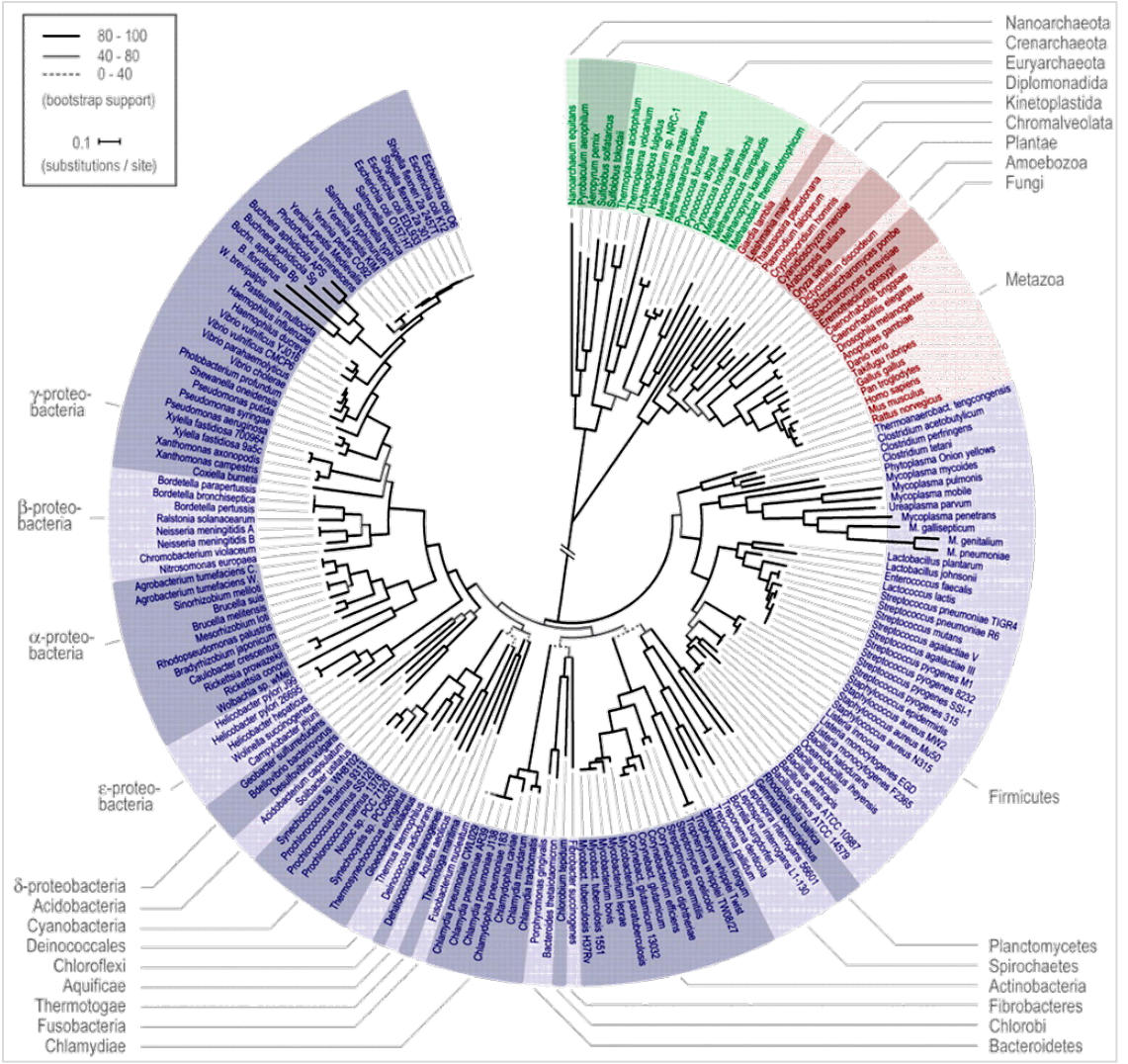
Qiang Li

Department of Combinatorics and Geometry

CAS-MPG Partner Institute for Computational Biology

Difficulties in phylogenomics

- Species trees versus gene trees
- Too few common genes
  - "The tree of one percent"
- Model selection
  - A science, or an art?
- Computational complexity
- …



A tree based on 31 genes (Ciccarelli '06)

Phylogenetics above sequence-level

- Rare genomic change
- Gene content & gene order
- *K*-string composition

"In summary, the methods for inferring trees based on whole-genome features are at an early stage of their development, which might be comparable to that of sequence-based methods in the early 1970s. In particular, they generally lack a global probabilistic modeling."

(Philippe 2005)

CVTree: Compositional vector

- "Random background": (K – 2)-order Markov chain

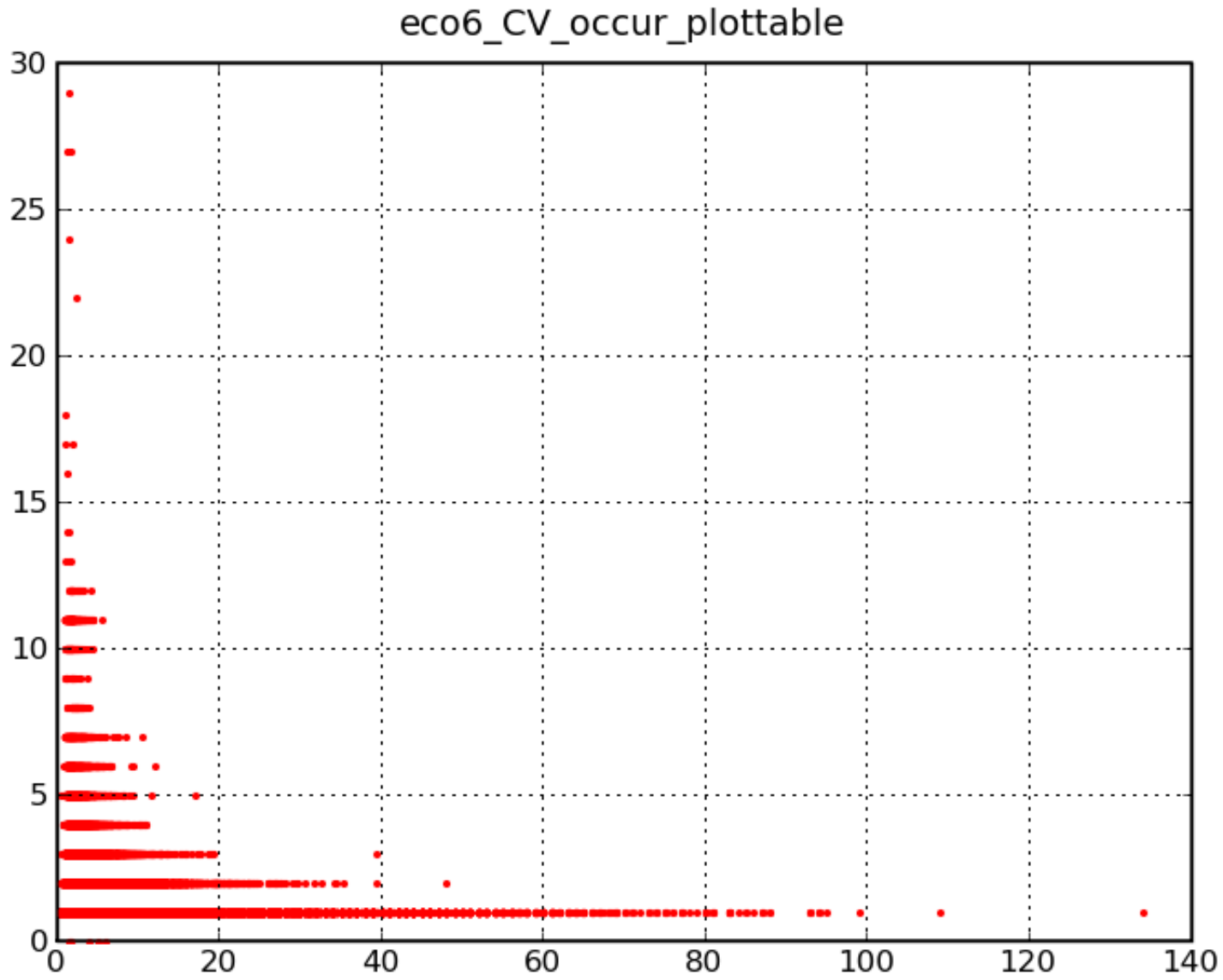
$$f^0(\alpha_1 \cdots \alpha_K) = \frac{f(\alpha_1 \cdots \alpha_{K-1}) f(\alpha_2 \cdots \alpha_K)}{f(\alpha_2 \cdots \alpha_{K-1})}$$

- "Subtraction"

$$a_i = \begin{cases} \dfrac{f(i) - f^0(i)}{f^0(i)}, & f^0(i) \neq 0 \\ 0, & f^0(i) = 0 \end{cases} \qquad i \in \Sigma^K$$

- Distance (dissimilarity)

$$d(\mathbf{a}, \mathbf{b}) = \frac{1 - \cos(\mathbf{a}, \mathbf{b})}{2} \in [0,1]$$



f versus f^0

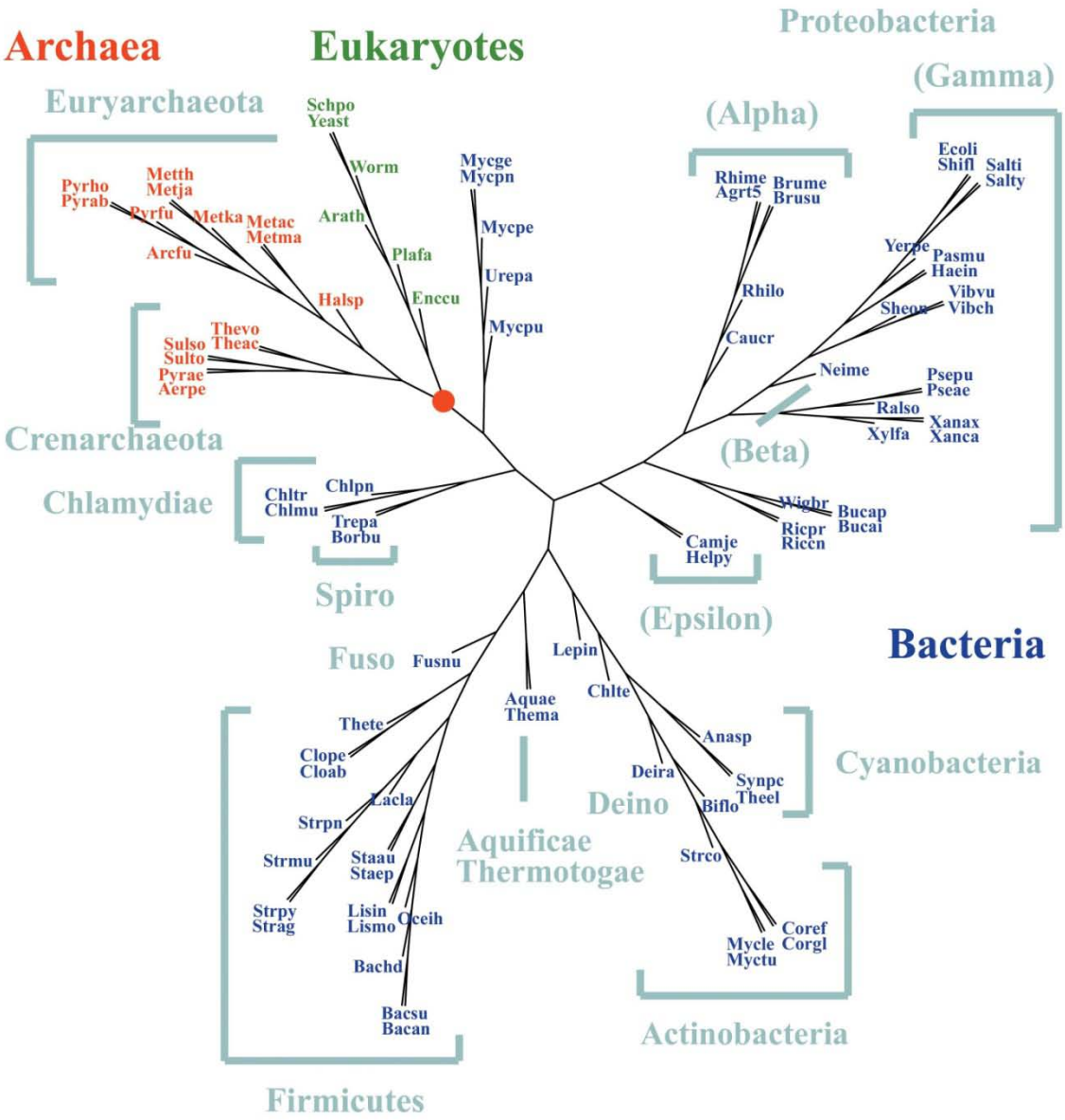
Courtesy of Alexander Grossmann

CVTree

- simple
- efficient
- parameter free
- truly whole genomic

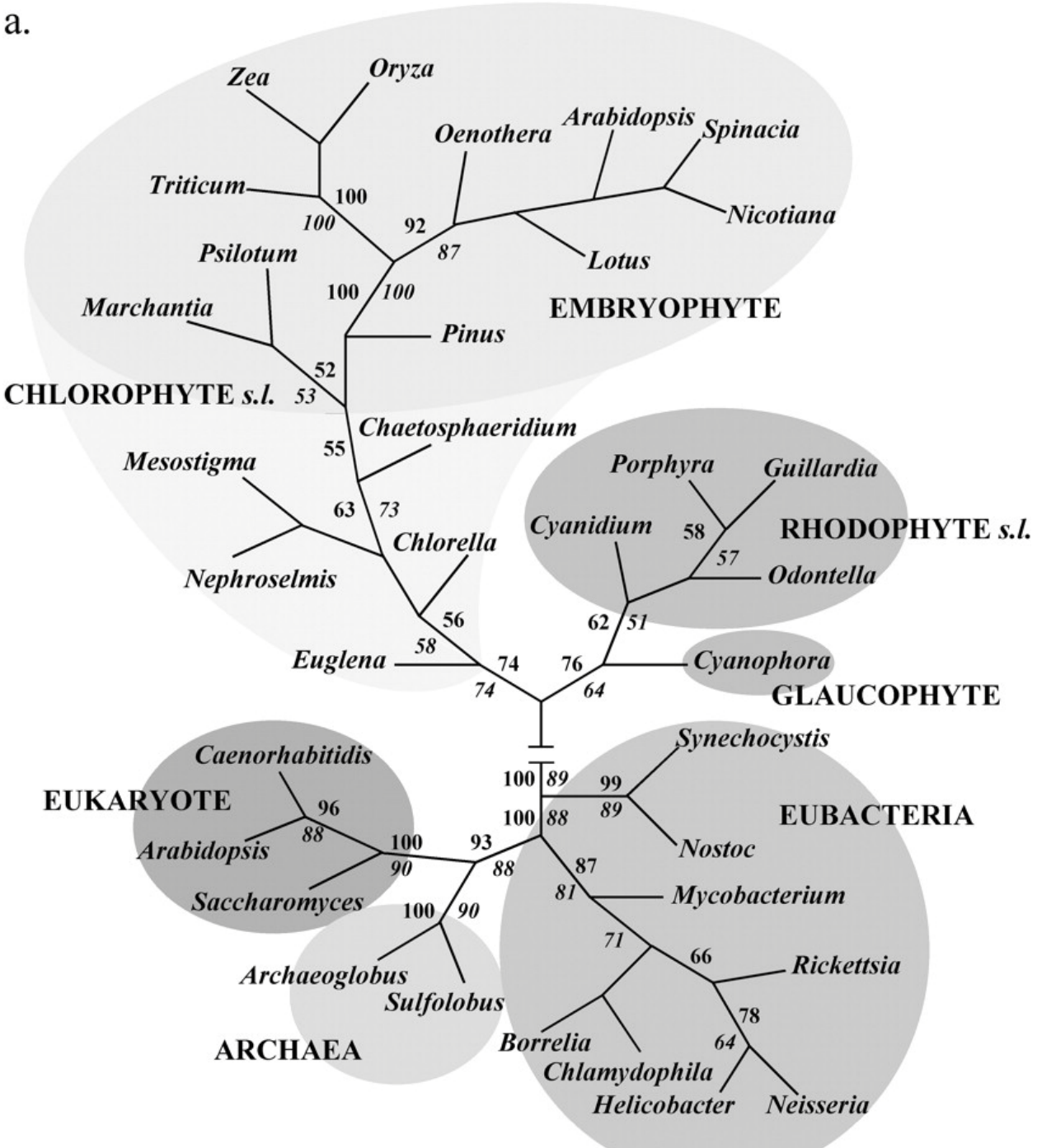
…

- mysterious

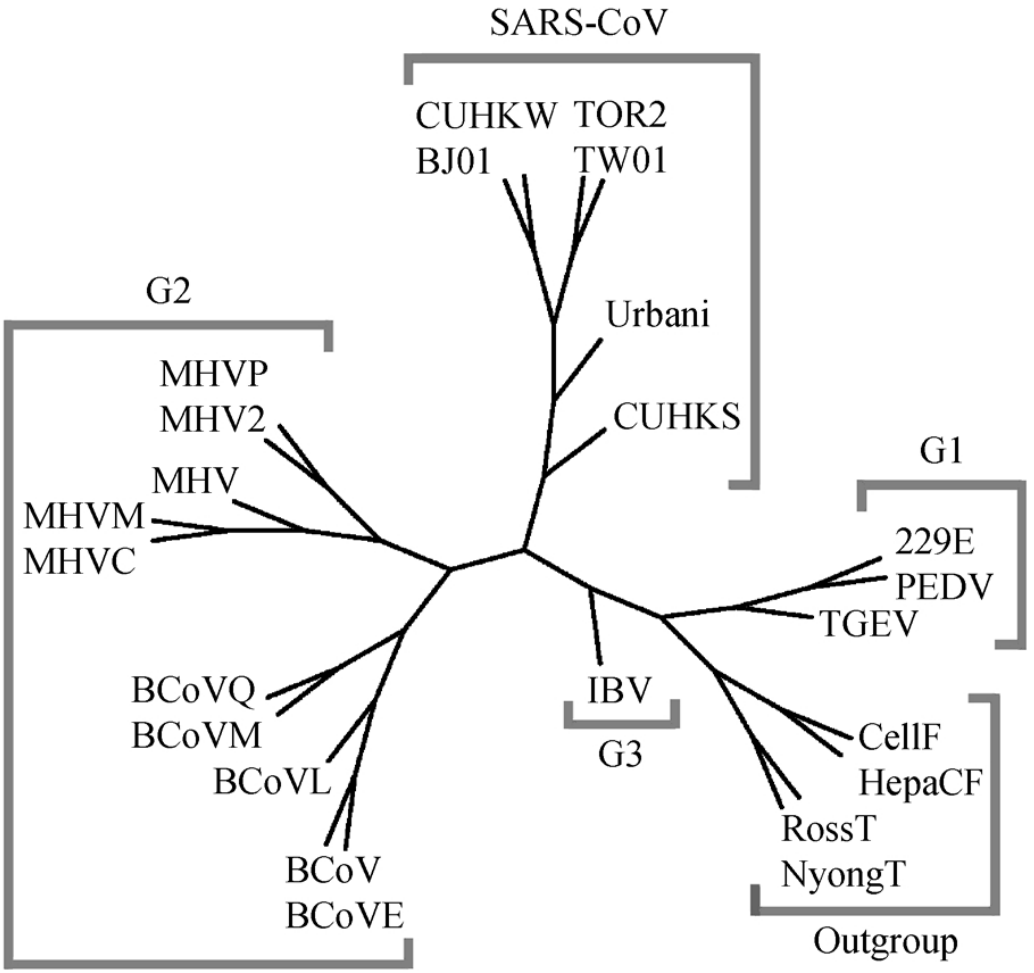


(Qi *et al* ’04)

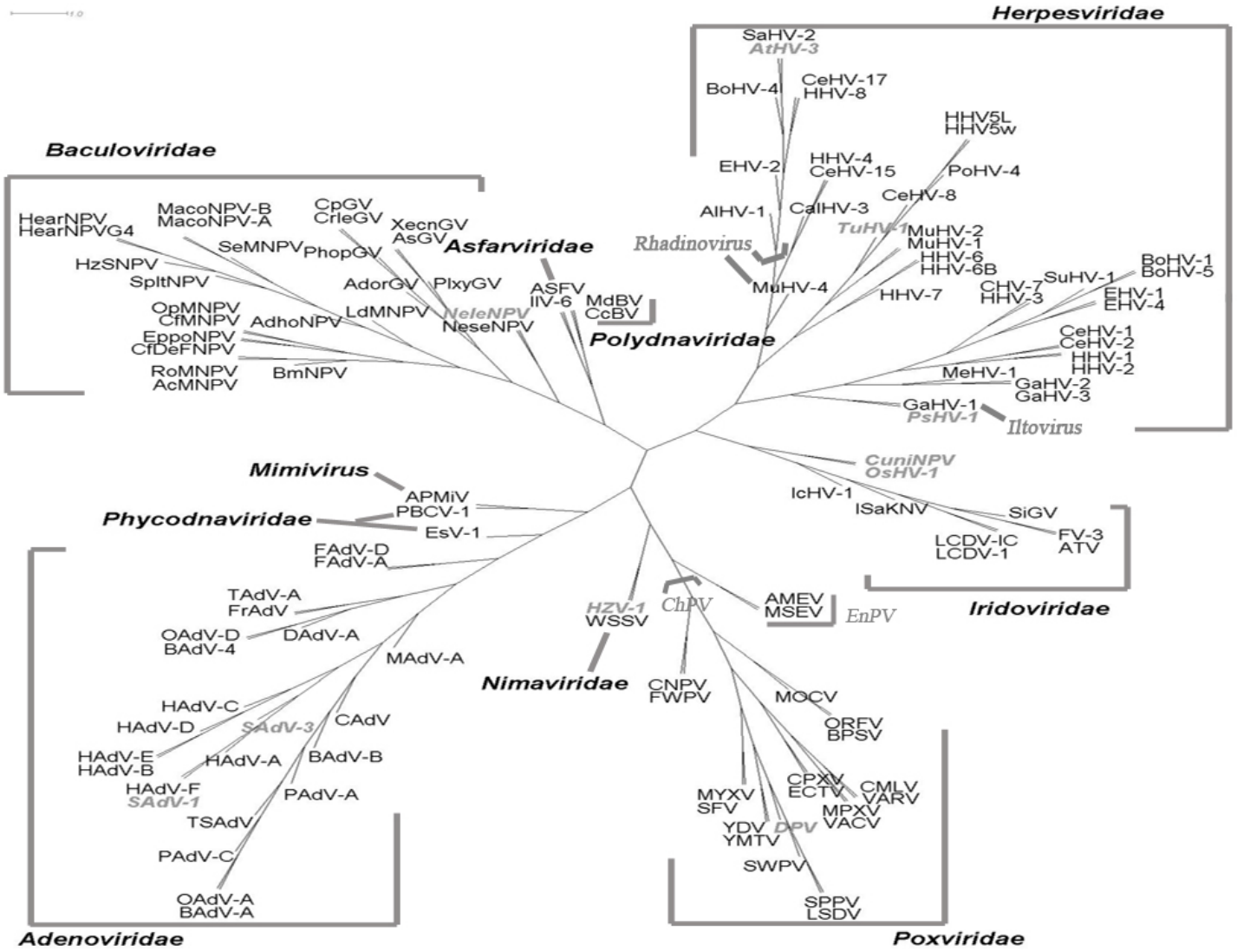
CVTree of chloroplasts



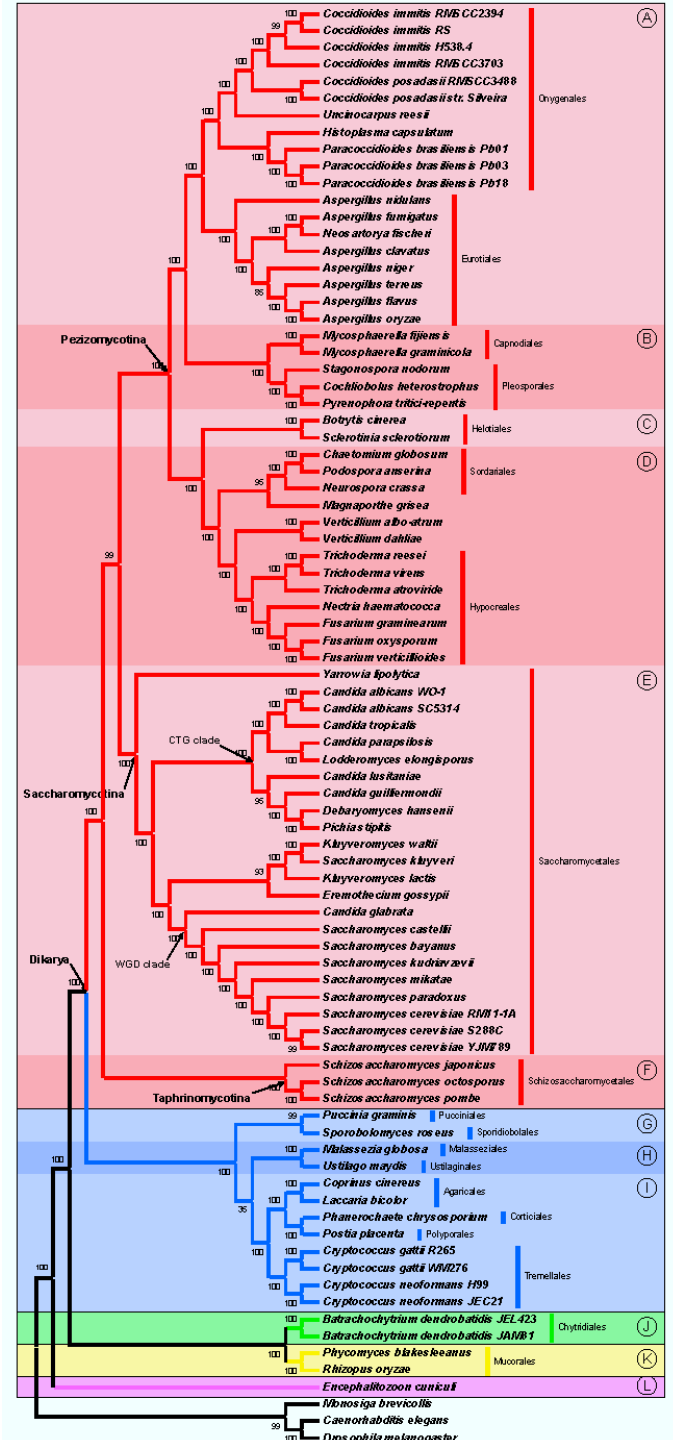
CVTree of coronavirus

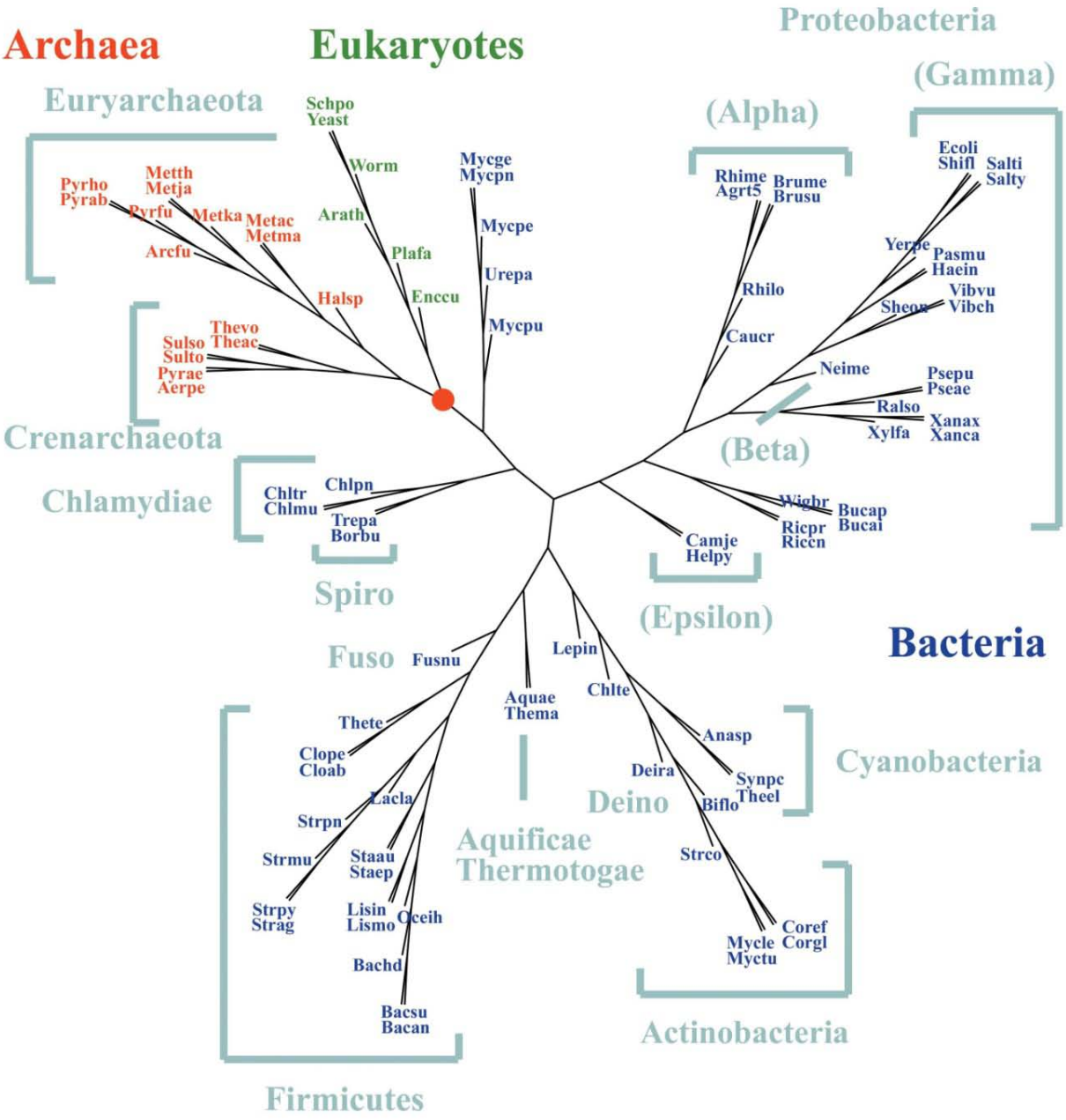


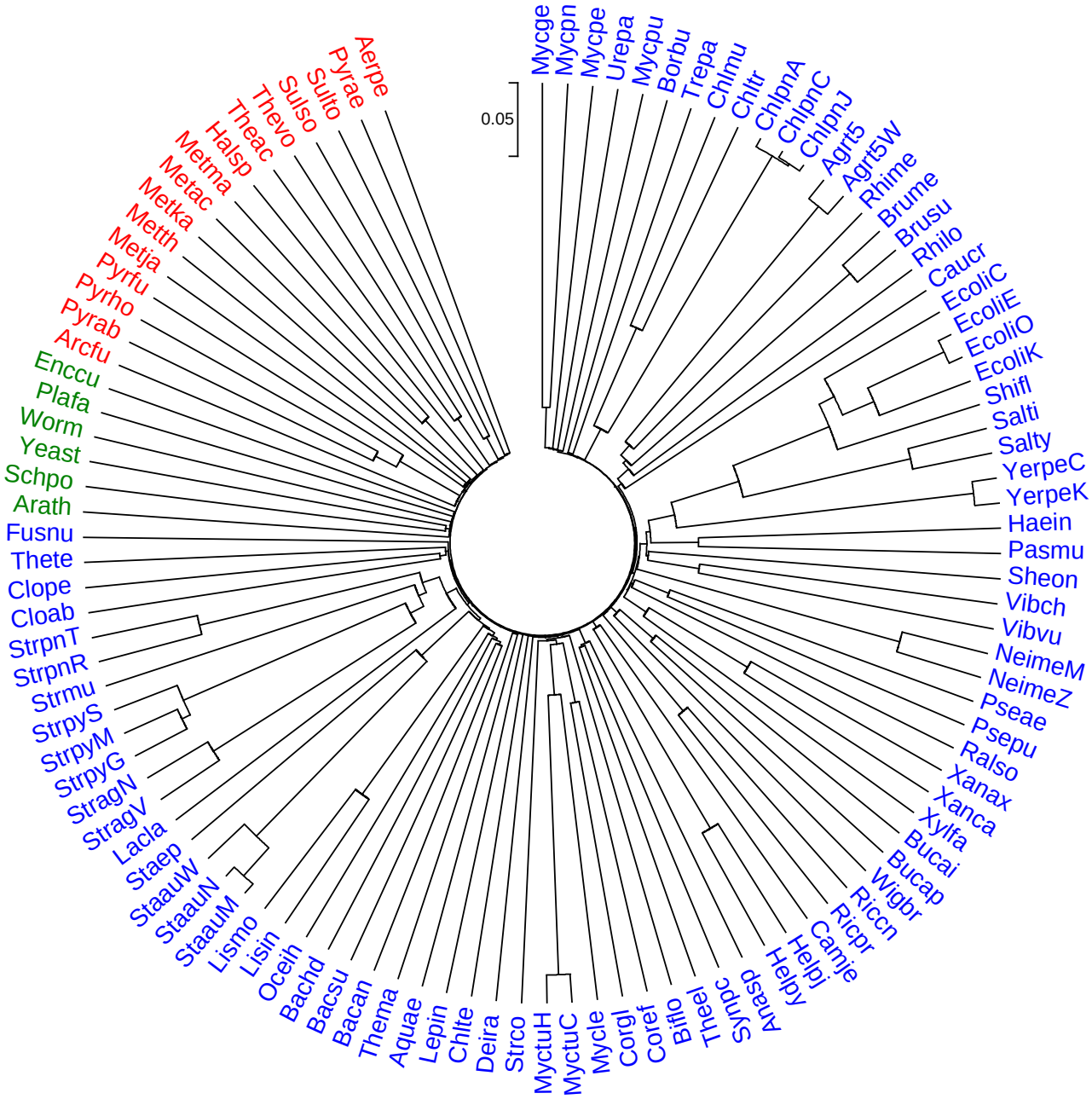
CVTree of dsDNA viruses



CVTree of fungi

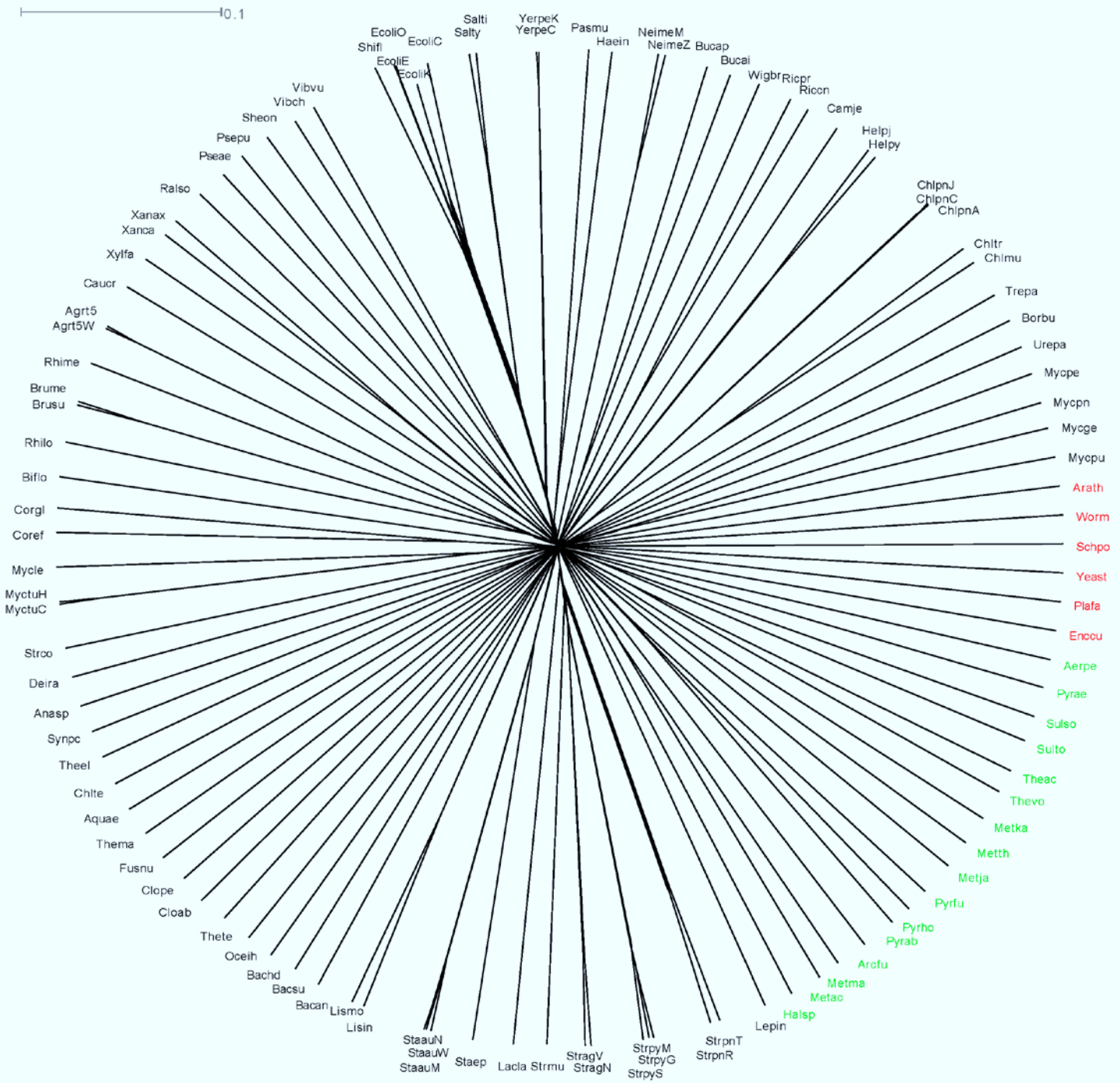


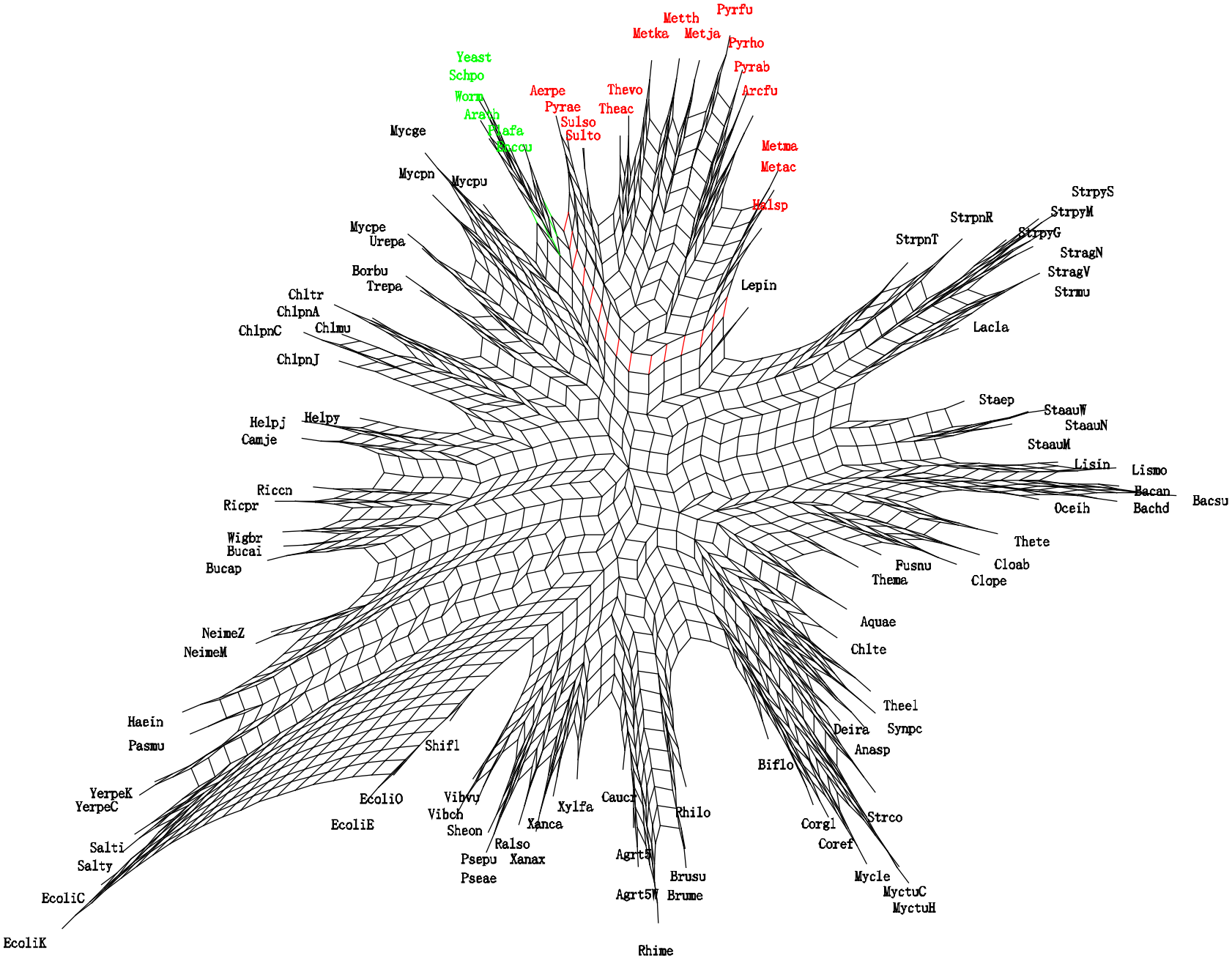

Archaea
Eukaryotes
Proteobacteria
(Gamma)
(Alpha)
Euryarchaeota
Schpo
Yeast
Worm
Arath
Plafa
Enccu
Mycge
Mycpn
Mycpe
Urepa
Mycpu
Metth
Metja
Pyrho
Pyrab
Pyrfu
Metka
Metac
Metma
Arcfu
Halsp
Thevo
Sulso
Theac
Sulto
Pyrae
Aerpe
Crenarchaeota
Rhime
Agrt5
Brume
Brusu
Rhilo
Caucr
Ecoli
Shifl
Salti
Salty
Yerpe
Pasmu
Haein
Vibvu
Sheon
Vibch
Neime
Psepu
Pseae
Ralso
Xylfa
Xanax
Xanca
(Beta)
Chlamydiae
Chltr
Chlmu
Chlpn
Trepa
Borbu
Spiro
Wigbr
Ricpr
Riccn
Bucap
Bucai
Camje
Helpy
(Epsilon)
Bacteria
Fuso
Fusnu
Lepin
Aquae
Thema
Chlte
Thete
Clope
Cloab
Lacla
Strpn
Strmu
Staau
Staep
Strpy
Strag
Lisin
Lismo
Oceih
Bachd
Bacsu
Bacan
Firmicutes
Aquificae
Thermotogae
Deira
Deino
Anasp
Synpc
Theel
Biflo
Cyanobacteria
Strco
Coref
Corgl
Mycle
Myctu
Actinobacteria



The tree plotted to scale

"Concentrating on the topology of the trees in the first place, we did not scale the branch lengths on the tree. However, these lengths should reflect evolution rates in terms of *K*-string composition changes. The calibration of branch lengths is further complicated by the overlapping nature of the *K*-strings when $K \geq 2$."


0.1
EcoliO
Shifl
EcoliC
EcoliE
EcoliK
Salti
Salty
YerpeK
YerpeC
Pasmu
Haein
NeimeM
NeimeZ
Bucap
Bucai
Wigbr
Ricpr
Ricon
Camje
Helpj
Helpy
ChlpnJ
ChlpnC
ChlpnA
Chltr
Chlmu
Trepa
Borbu
Urepa
Mycpe
Mycpn
Mycge
Mycpu
Arath
Worm
Schpo
Yeast
Plafa
Enccu
Aerpe
Pyrae
Sulso
Sulto
Theac
Thevo
Metka
Metth
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Metma
Metac
Halsp
Lepin
StrpnT
StrpnR
StrpyM
StrpyG
StrpyS
StragV
StragN
Lacla
Strmu
Staep
StaauN
StaauW
StaauM
Bacan
Lismo
Lisin
Bachd
Bacsu
Oceih
Thete
Cloab
Clope
Fusnu
Thema
Aquae
Chlte
Theel
Synpc
Anasp
Deira
Strco
MyctuH
MyctuC
Mycle
Coref
Corgl
Biflo
Rhilo
Brume
Brusu
Rhime
Agrt5
Agrt5W
Caucr
Xylfa
Xanax
Xanca
Ralso
Psepu
Pseae
Sheon
Vibch
Vibvu


Metth
Pyrfu
Metka
Metja
Pyrho
Yeast
Schpo
Pyrab
Worm
Arcfu
Aerpe
Thevo
Arath
Pyrae
Theac
Plafa
Sulso
Mycge
Sulto
Saccu
Metma
Mycpn
Mycpu
Metac
Halsp
StrpyS
StrpnR
StrpyM
Mycpe
StrpnT
StrpyG
Urepa
StragN
Borbu
StragV
Lepin
Trepa
Strmu
Chltr
ChlpnA
Lacla
ChlpnC
Chlmu
ChlpnJ
Staep
Helpj
Helpy
StaauW
Camje
StaauN
StaauM
Lisin
Lismo
Riccn
Bacan
Ricpr
Bacsu
Oceih
Bachd
Wigbr
Thete
Bucai
Pusnu
Cloab
Bucap
Thema
Clope
Aquae
NeimeZ
Chlte
NeimeM
Theel
Haein
Deira
Sympc
Pasmu
Anasp
Shifl
Biflo
YerpeK
EcoliO
Vibvu
Caucr
Xylfa
Strco
YerpeC
EcoliE
Vibch
Rhilo
Sheon
Xanca
Corgl
Salti
Ralso
Coref
Psepu
Xanax
Agrt5W
Salty
Brusu
Mycle
Pseae
MyctuC
EcoliC
Agrt5W
Brume
MyctuH
EcoliK
Rhime

- “However, the methods proposed so far are extremely crude. Oligonucleotide or oligopeptide frequencies are transformed into distances without any underlying model of evolution. It is nevertheless remarkable that something considered as a bias in standard sequence-based methods (Lockhart *et al*. 1994) contains a phylogenetic signal, but it is not yet clear whether accurate methods can be developed to extract it.”

  (Philippe *et al*. 2005)

- “Why, for example, does the *K*-string complement of a proteome yield a tree that is similar to sequence-based trees?” (Snel *et al*. 2005)

A heuristic probabilistic model

- Transition of frequency distribution:

  Assume that the evolution process is stationary, and the average mutation rate per site is 1 – α. Let

  $L = |\mathrm{S}| - K + 1$, then the conditional distribution

$$f\left(\mathrm{S}^{(t+1)}(\mathbf{i}) \middle| \mathrm{S}^{(t)}(\mathbf{i}) = n\right) = \mathrm{binomial}(n, \alpha^K)$$

$$* \mathrm{binomial}\left(L, (1-\alpha^K)\frac{\mathrm{E}\,\mathrm{S}(\mathbf{i})}{L}\right)$$

- Time evolution of expectation

$$\mathrm{E}\left(\mathrm{S}^{(t')}(\mathbf{i})\middle|\mathrm{S}^{(t)}(\mathbf{i})=n\right)=\mathrm{E}\,\mathrm{S}(\mathbf{i})+\alpha^{K(t'-t)}(n-\mathrm{E}\,\mathrm{S}(\mathbf{i}))$$

Calibration of CVTree

- Generalized CV

$$w(\mathrm{S}, \mathbf{i}) = (\mathrm{S}(\mathbf{i}) - \mathrm{S}^0(\mathbf{i}))c(\mathbf{i})$$

where $\mathrm{S}^0(\mathbf{i})$ is an unbiased estimator of $\mathrm{S}(\mathbf{i})$, and $c(\mathbf{i})$ is arbitrary "normalization" factor.

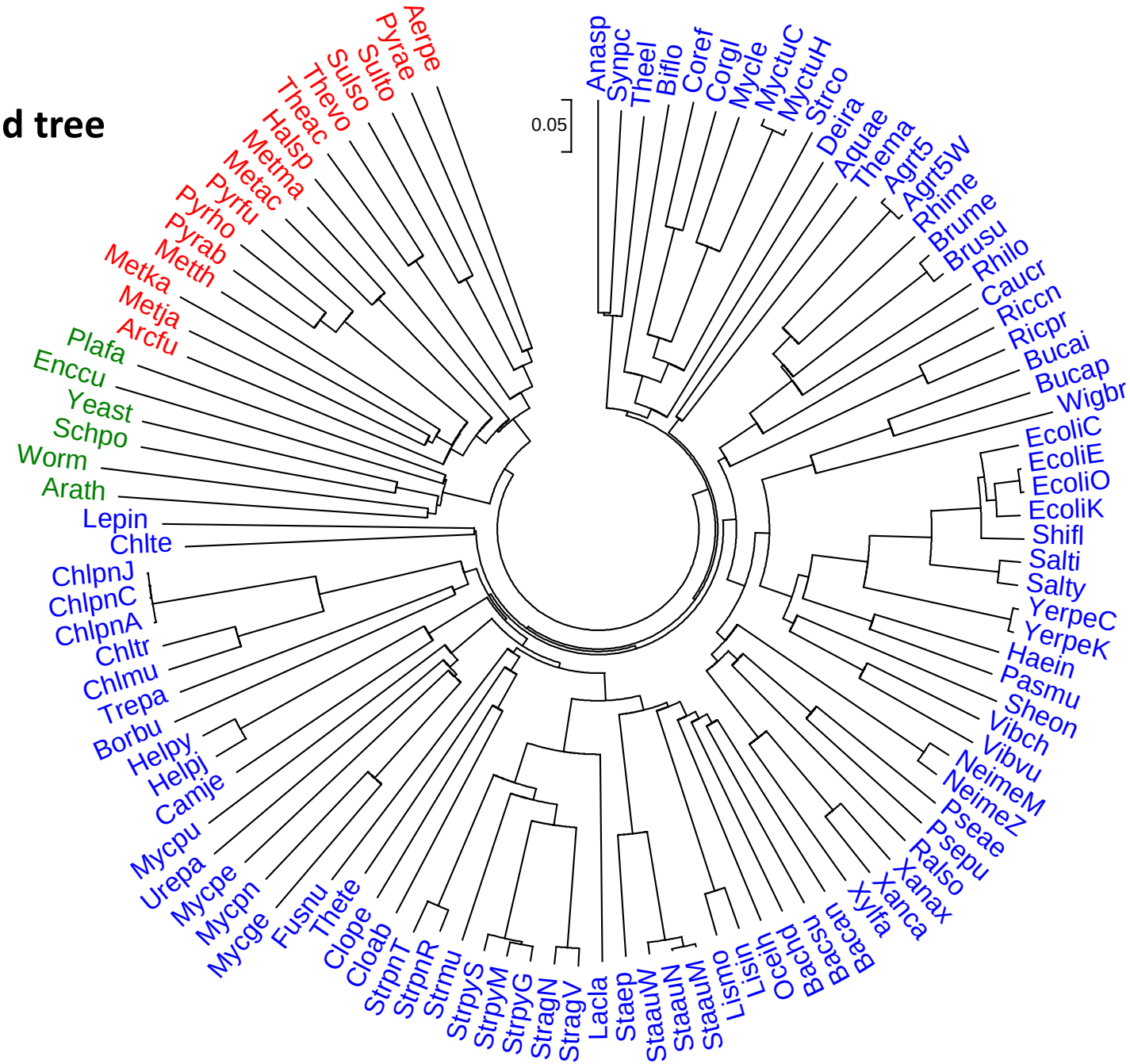
- Angle between vectors vs. sequence identity

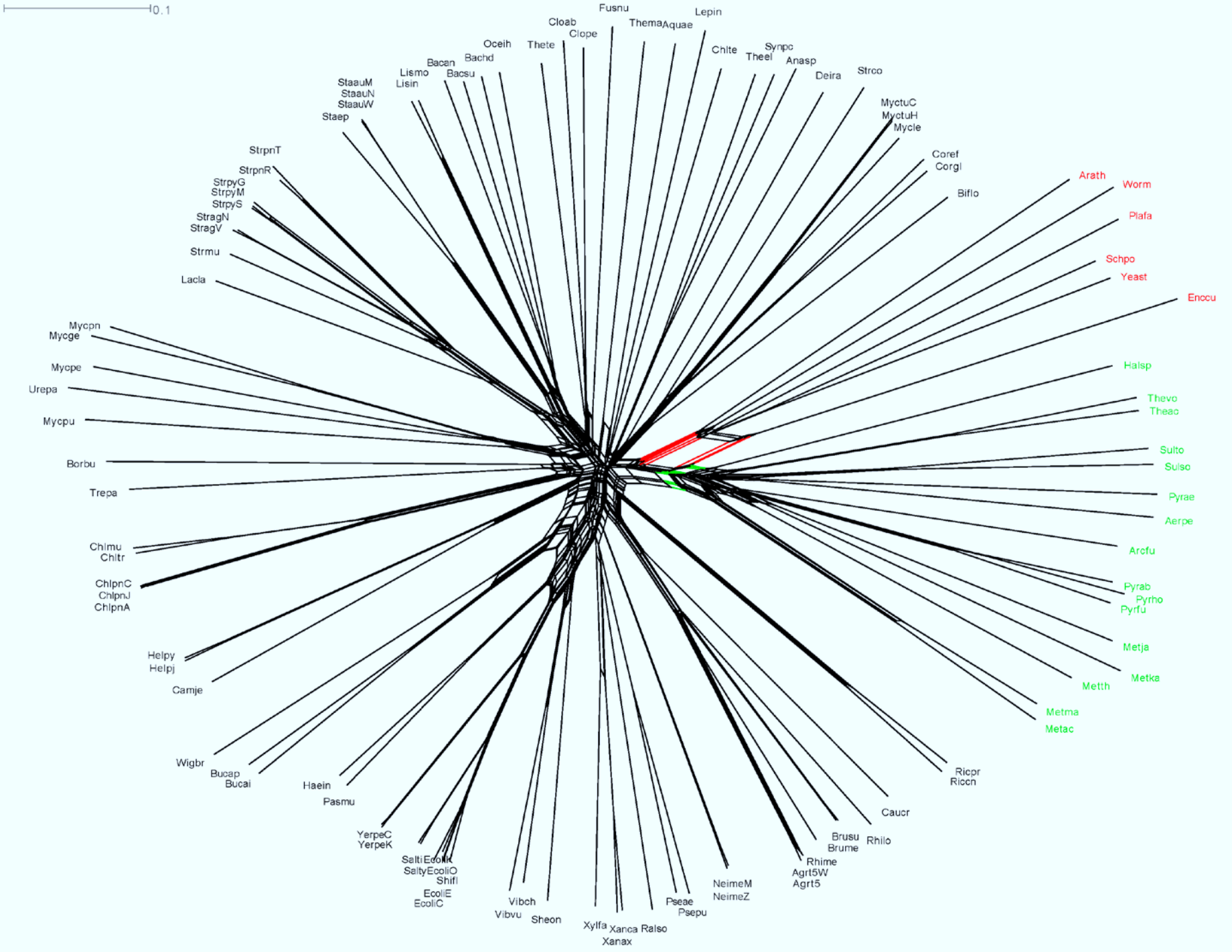
$$\mathrm{E}\cos(w(\mathrm{S}), w(\mathrm{S}')) = q^K$$

- Calibration

$$\hat{p}(\mathrm{S},\mathrm{S}') = 1 - \cos^{1/K}\left(w(\mathrm{S}), w(\mathrm{S}')\right)$$

The calibrated tree




0.1
Fusnu
Lepin
Cloab
Thema
Aquae
Clope
Oceih
Thete
Chlte
Synpc
Theel
Anasp
Deira
Strco
Bacan
Bachd
Lismo
Bacsu
Lisin
StaauM
StaauN
StaauW
Staep
MyctuC
MyctuH
Mycle
StrpnT
StrpnR
Coref
Corgl
StrpyG
StrpyM
StrpyS
Biflo
StragN
StragV
Strmu
Lacla
Arath
Worm
Plafa
Schpo
Yeast
Enccu
Mycpn
Mycge
Halsp
Mycpe
Urepa
Thevo
Theac
Mycpu
Sulto
Sulso
Borbu
Trepa
Pyrae
Aerpe
Chlmu
Chltr
Arcfu
ChlpnC
ChlpnJ
ChlpnA
Pyrab
Pyrho
Pyrfu
Helpy
Helpj
Metja
Camje
Metka
Metth
Metma
Metac
Wigbr
Bucap
Bucai
Ricpr
Riccn
Haein
Pasmu
Caucr
YerpeC
YerpeK
Brusu
Brume
Rhilo
SaltiEcoliK
SaltyEcoliO
ShifI
Rhime
Agrt5W
Agrt5
EcoliE
EcoliC
NeimeM
NeimeZ
Vibch
Vibvu
Pseae
Psepu
Sheon
Xylfa
Xanca
Ralso
Xanax

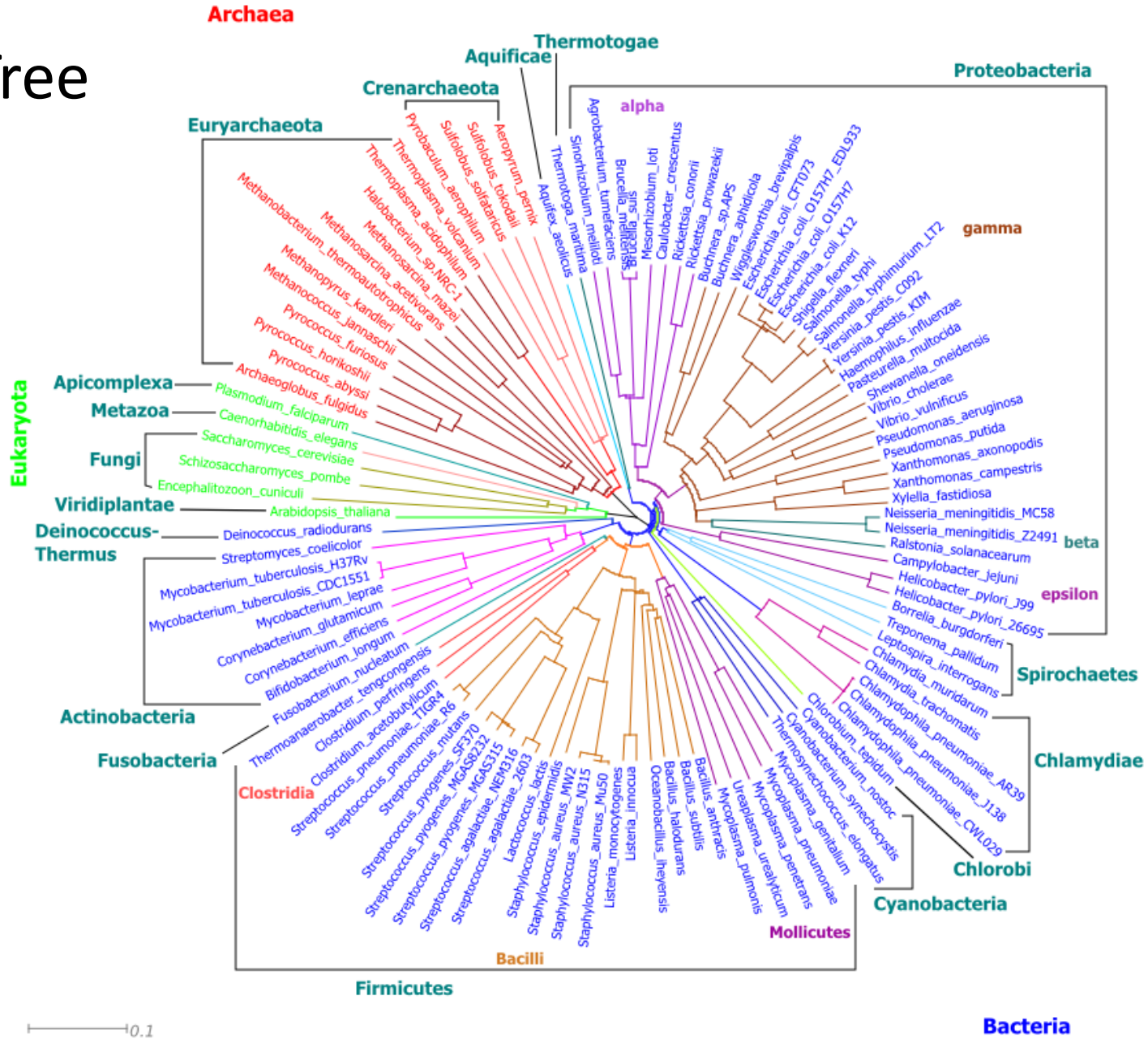
Using pre-/absence of words

- Any sequence S as above can be represented also by just the set $X(\mathrm{S})$ of its K-strings. In a "gedanken alignment", ignoring repeats and collisions, we may put

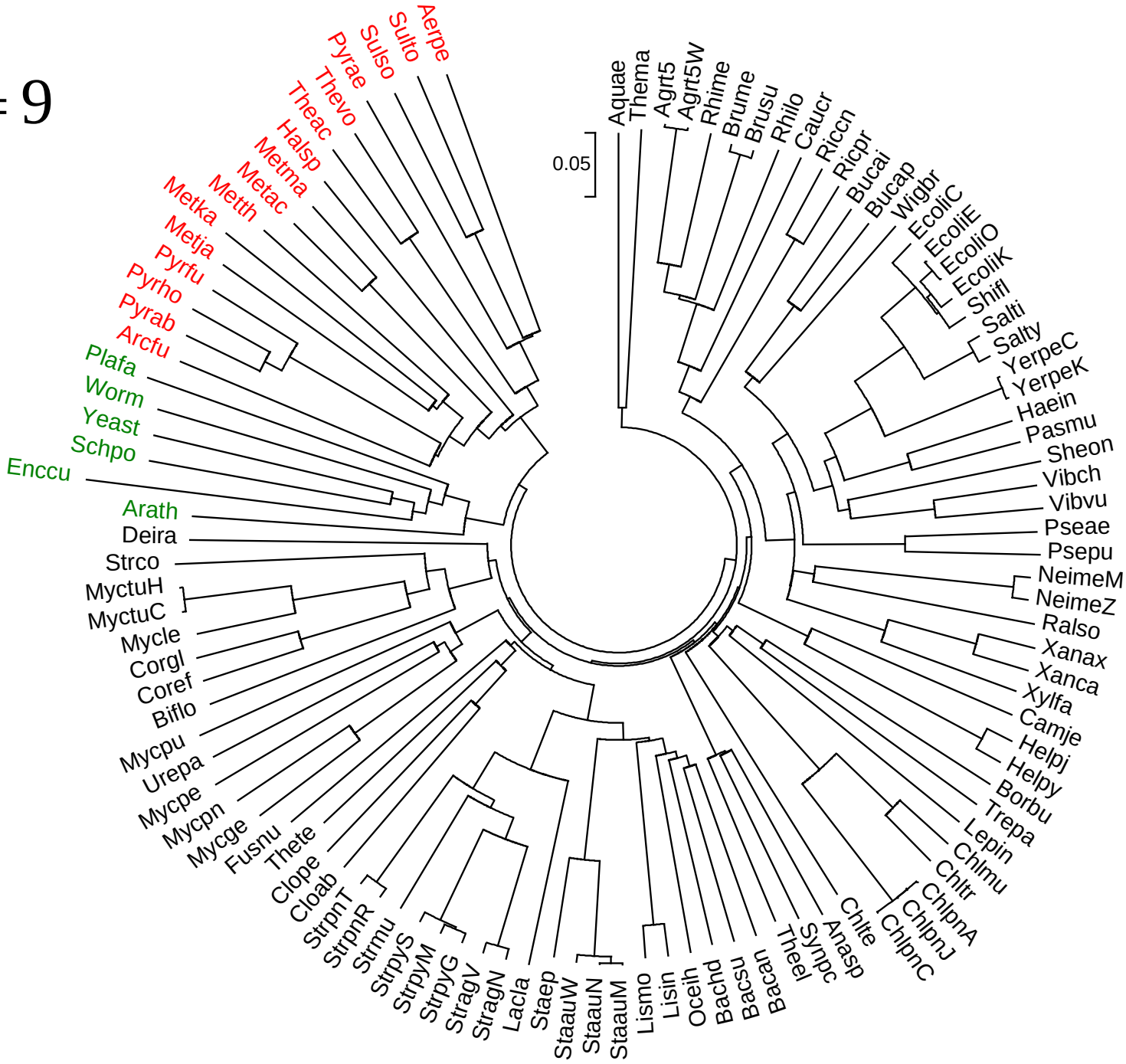
$$Q^{(K)}(\mathrm{S},\mathrm{S}') := \frac{1}{2}\left(\frac{\left|X(\mathrm{S}) \bigcap X(\mathrm{S}')\right|}{\left|X(\mathrm{S})\right|} + \frac{\left|X(\mathrm{S}) \bigcap X(\mathrm{S}')\right|}{\left|X(\mathrm{S}')\right|} \right)$$

and $$q^{(K)}(\mathrm{S},\mathrm{S}') := Q^{(K)}(\mathrm{S},\mathrm{S}')^{1/K}$$

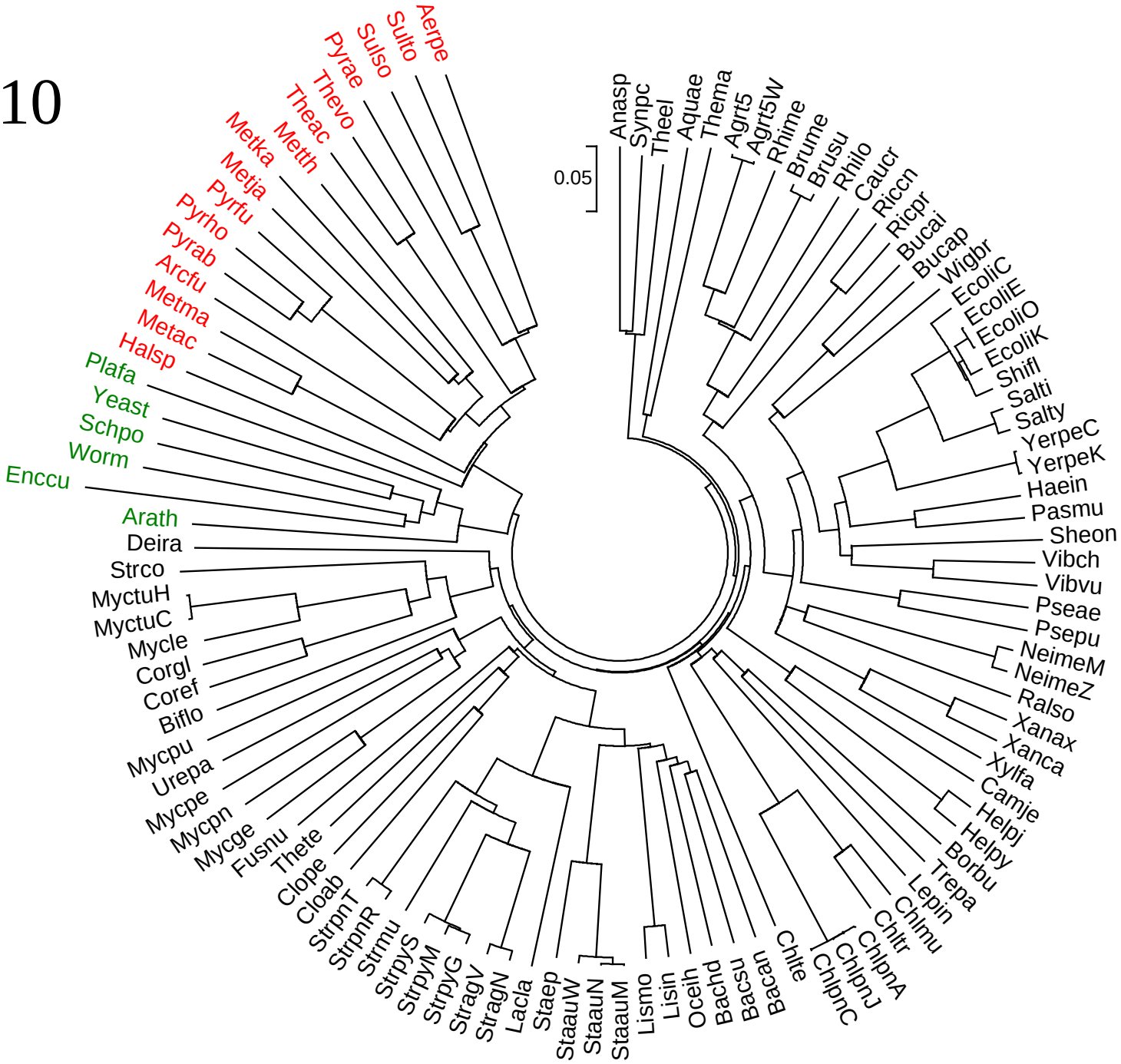
$K = 9$ Tree


Archaea
Thermotogae
Aquificae
Crenarchaeota
Euryarchaeota
Proteobacteria
alpha
gamma
beta
epsilon
Eukaryota
Apicomplexa
Metazoa
Fungi
Viridiplantae
Deinococcus-Thermus
Actinobacteria
Fusobacteria
Clostridia
Bacilli
Firmicutes
Mollicutes
Spirochaetes
Chlamydiae
Chlorobi
Cyanobacteria
Bacteria
Aeropyrum_pernix
Sulfolobus_solfataricus
Sulfolobus_tokodaii
Pyrobaculum_aerophilum
Thermoplasma_volcanium
Thermoplasma_acidophilum
Halobacterium_sp.NRC-1
Methanosarcina_acetivorans
Methanosarcina_mazei
Methanobacterium_thermoautotrophicus
Methanopyrus_kandleri
Methanococcus_jannaschii
Pyrococcus_furiosus
Pyrococcus_horikoshii
Pyrococcus_abyssi
Archaeoglobus_fulgidus
Aquifex_aeolicus
Thermotoga_maritima
Sinorhizobium_meliloti
Agrobacterium_tumefaciens
Brucella_melitensis
Brucella_suis
Mesorhizobium_loti
Caulobacter_crescentus
Rickettsia_conorii
Rickettsia_prowazekii
Buchnera_sp.APS
Buchnera_aphidicola
Wigglesworthia_brevipalpis
Escherichia_coli_CFT073
Escherichia_coli_O157H7_EDL933
Escherichia_coli_O157H7
Escherichia_coli_K12
Shigella_flexneri
Salmonella_typhi
Salmonella_typhimurium_LT2
Yersinia_pestis_C092
Yersinia_pestis_KIM
Haemophilus_influenzae
Haemophilus_multocida
Pasteurella_multocida
Shewanella_oneidensis
Vibrio_cholerae
Vibrio_vulnificus
Pseudomonas_aeruginosa
Pseudomonas_putida
Xanthomonas_axonopodis
Xanthomonas_campestris
Xylella_fastidiosa
Neisseria_meningitidis_MC58
Neisseria_meningitidis_Z2491
Ralstonia_solanacearum
Campylobacter_jejuni
Helicobacter_pylori_J99
Helicobacter_pylori_26695
Borrelia_burgdorferi
Treponema_pallidum
Leptospira_interrogans
Chlamydia_muridarum
Chlamydia_trachomatis
Chlamydophila_pneumoniae_AR39
Chlamydophila_pneumoniae_J138
Chlamydophila_pneumoniae_CWL029
Chlorobium_tepidum
Cyanobacterium_nostoc
Thermosynechococcus_elongatus
Mycoplasma_genitalium
Mycoplasma_pneumoniae
Mycoplasma_penetrans
Ureaplasma_urealyticum
Mycoplasma_pulmonis
Bacillus_anthracis
Bacillus_subtilis
Bacillus_halodurans
Oceanobacillus_iheyensis
Listeria_innocua
Listeria_monocytogenes
Staphylococcus_aureus_Mu50
Staphylococcus_aureus_N315
Staphylococcus_aureus_MW2
Staphylococcus_epidermidis
Lactococcus_lactis
Streptococcus_agalactiae_2603
Streptococcus_agalactiae_NEM316
Streptococcus_agalactiae_MGAS8232
Streptococcus_pyogenes_MGAS315
Streptococcus_pyogenes_SF370
Streptococcus_pneumoniae_TIGR4
Streptococcus_pneumoniae_R6
Streptococcus_mutans
Clostridium_perfringens
Clostridium_acetobutylicum
Thermoanaerobacter_tengcongensis
Fusobacterium_nucleatum
Bifidobacterium_longum
Corynebacterium_efficiens
Corynebacterium_glutamicum
Mycobacterium_leprae
Mycobacterium_tuberculosis_CDC1551
Mycobacterium_tuberculosis_H37Rv
Streptomyces_coelicolor
Deinococcus_radiodurans
Arabidopsis_thaliana
Encephalitozoon_cuniculi
Schizosaccharomyces_pombe
Saccharomyces_cerevisiae
Caenorhabitidis_elegans
Plasmodium_falciparum
0.1

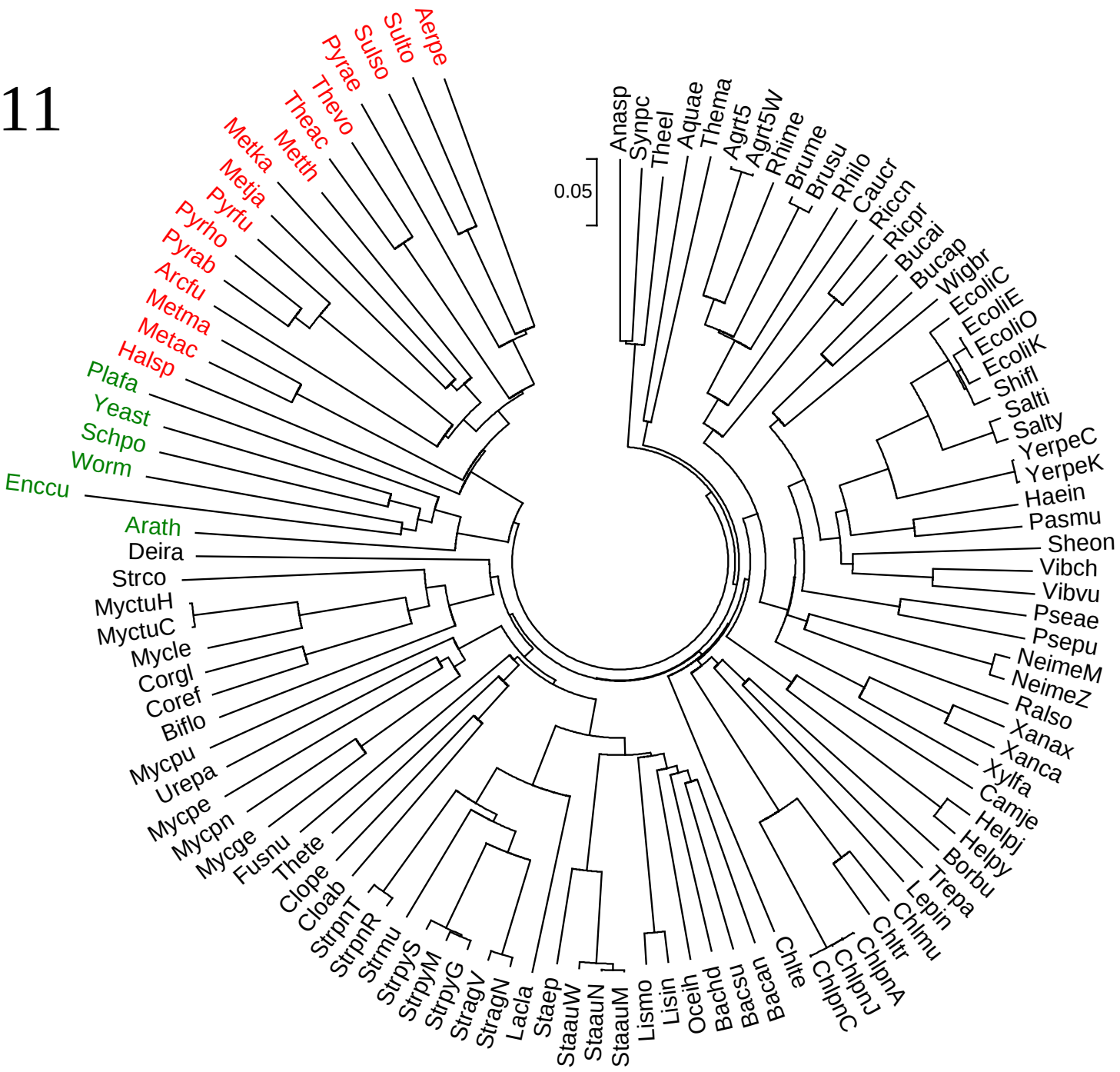
$K = 9$


0.05
Aerpe
Sulto
Sulso
Pyrae
Thevo
Theac
Halsp
Metma
Metac
Metth
Metka
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Plafa
Worm
Yeast
Schpo
Enccu
Arath
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo
Mycpu
Urepa
Mycpe
Mycpn
Mycge
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragV
StragN
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Theel
Synpc
Anasp
Chlte
ChlpnA
ChlpnJ
ChlpnC
Chltr
Chlmu
Lepin
Trepa
Borbu
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
NeimeZ
NeimeM
Psepu
Pseae
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Thema
Aquae

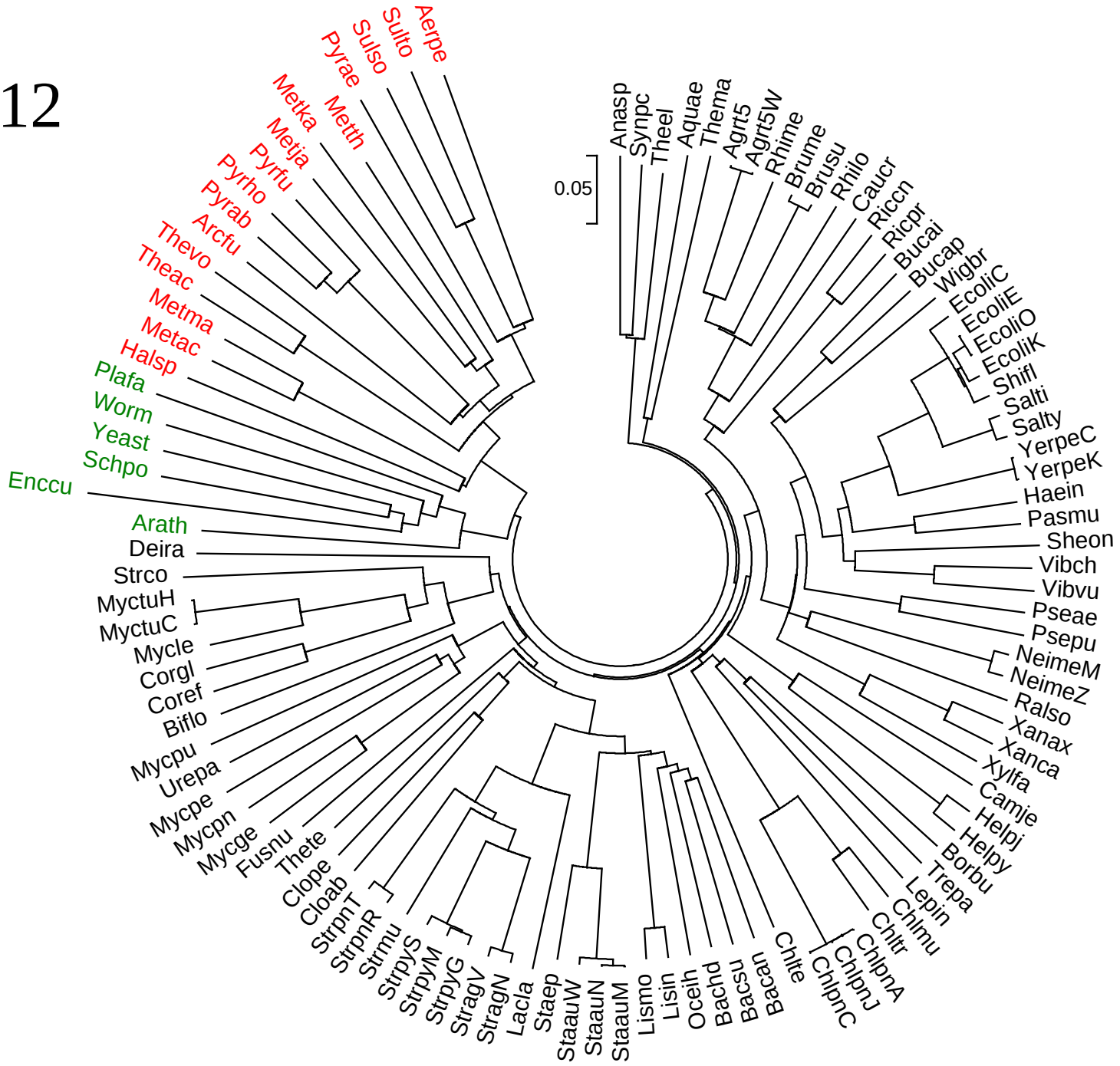
$K = 10$


0.05
Aerpe
Sulto
Sulso
Pyrae
Thevo
Theac
Metth
Metka
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Metma
Metac
Halsp
Plafa
Yeast
Schpo
Worm
Enccu
Arath
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo
Mycpu
Urepa
Mycpe
Mycpn
Mycge
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragV
StragN
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Chlte
ChlpnC
ChlpnJ
ChlpnA
Chltr
Chlmu
Lepin
Trepa
Borbu
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
NeimeZ
NeimeM
Psepu
Pseae
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Thema
Aquae
Theel
Synpc
Anasp

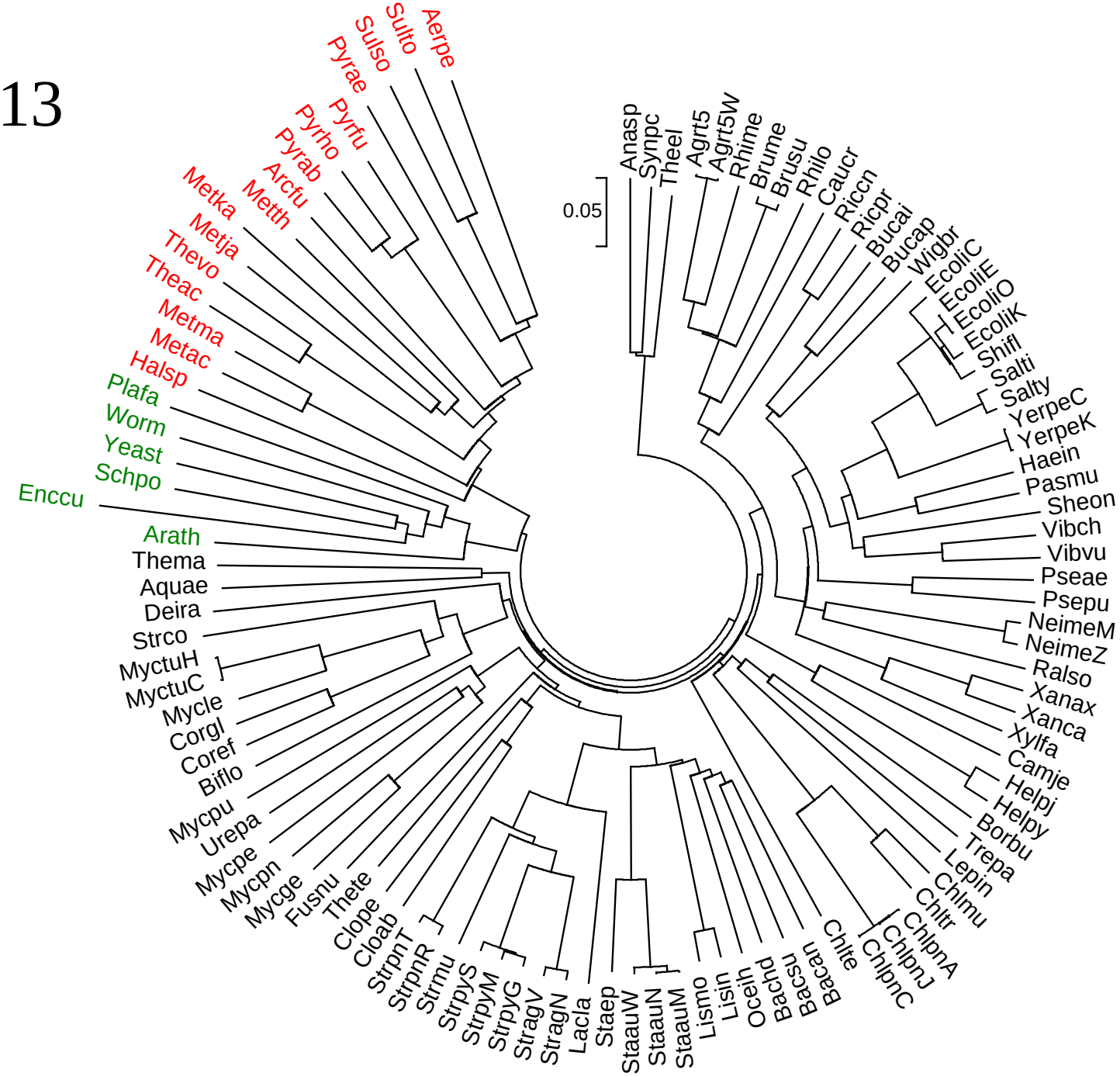
$K = 11$


0.05
Aerpe
Sulto
Sulso
Pyrae
Thevo
Theac
Metth
Metka
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Metma
Metac
Halsp
Plafa
Yeast
Schpo
Worm
Enccu
Arath
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo
Mycpu
Urepa
Mycpe
Mycpn
Mycge
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragV
StragN
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Chlte
ChlpnC
ChlpnJ
ChlpnA
Chltr
Chlmu
Lepin
Trepa
Borbu
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
NeimeZ
NeimeM
Psepu
Pseae
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Thema
Aquae
Theel
Synpc
Anasp

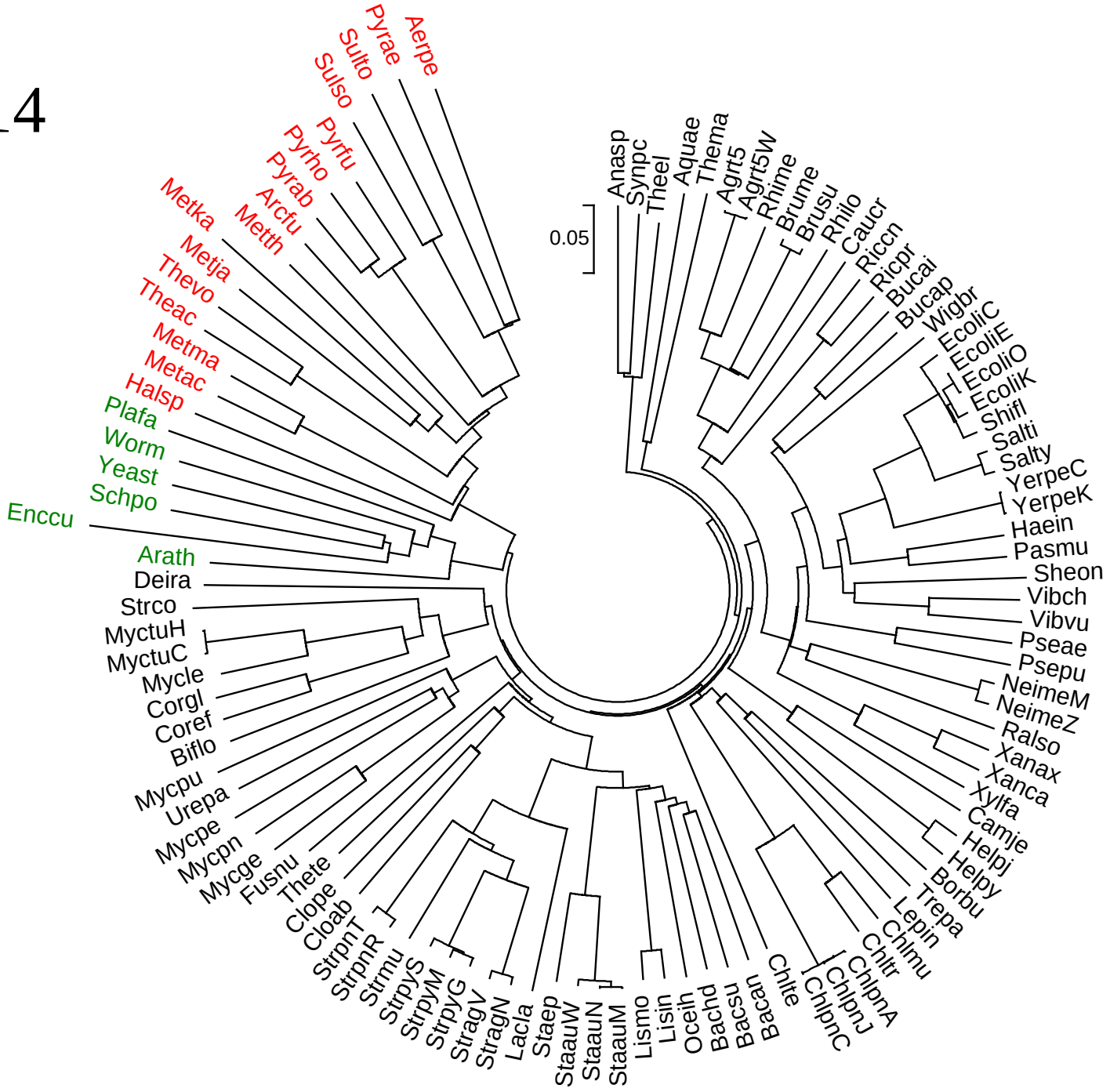
$K = 12$

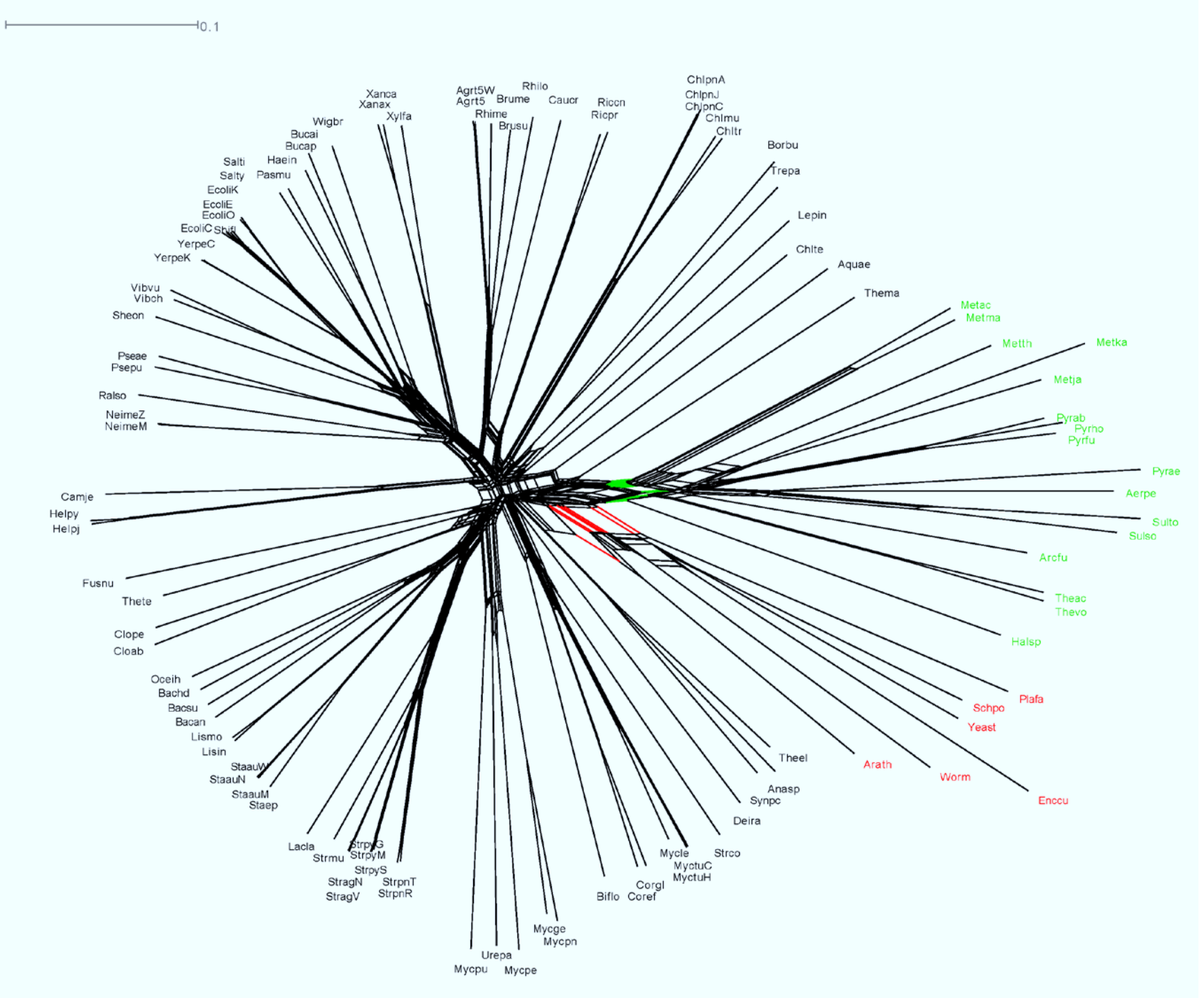

0.05
Aerpe
Sulto
Sulso
Pyrae
Metth
Metka
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Thevo
Theac
Metma
Metac
Halsp
Plafa
Worm
Yeast
Schpo
Enccu
Arath
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo
Mycpu
Urepa
Mycpe
Mycpn
Mycge
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragV
StragN
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Chlte
ChlpnC
ChlpnJ
ChlpnA
Chltr
Chlmu
Lepin
Trepa
Borbu
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
NeimeZ
NeimeM
Psepu
Pseae
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Thema
Aquae
Theel
Synpc
Anasp

$K = 13$


0.05
Aerpe
Sulto
Sulso
Pyrae
Pyrfu
Pyrho
Pyrab
Arcfu
Metth
Metka
Metja
Thevo
Theac
Metma
Metac
Halsp
Plafa
Worm
Yeast
Schpo
Enccu
Arath
Thema
Aquae
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo
Mycpu
Urepa
Mycpe
Mycpn
Mycge
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragV
StragN
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Chlte
ChlpnA
ChlpnJ
ChlpnC
Chltr
Chlmu
Lepin
Trepa
Borbu
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
NeimeZ
NeimeM
Psepu
Pseae
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Theel
Synpc
Anasp

$K = 14$


0.05
Anasp
Synpc
Theel
Aquae
Thema
Agrt5
Agrt5W
Rhime
Brume
Brusu
Rhilo
Caucr
Riccn
Ricpr
Bucai
Bucap
Wigbr
EcoliC
EcoliE
EcoliO
EcoliK
Shifl
Salti
Salty
YerpeC
YerpeK
Haein
Pasmu
Sheon
Vibch
Vibvu
Pseae
Psepu
NeimeM
NeimeZ
Ralso
Xanax
Xanca
Xylfa
Camje
Helpj
Helpy
Borbu
Trepa
Lepin
Chlmu
Chltr
ChlpnA
ChlpnJ
ChlpnC
Chlte
Bacan
Bacsu
Bachd
Oceih
Lisin
Lismo
StaauM
StaauN
StaauW
Staep
Lacla
StragN
StragV
StrpyG
StrpyM
StrpyS
Strmu
StrpnR
StrpnT
Cloab
Clope
Thete
Fusnu
Mycge
Mycpn
Mycpe
Urepa
Mycpu
Biflo
Coref
Corgl
Mycle
MyctuC
MyctuH
Strco
Deira
Arath
Enccu
Schpo
Yeast
Worm
Plafa
Halsp
Metac
Metma
Theac
Thevo
Metja
Metka
Metth
Arcfu
Pyrab
Pyrho
Pyrfu
Sulso
Sulto
Pyrae
Aerpe


0.1
ChlpnA
ChlpnJ
ChlpnC
Chlmu
Chltr
Agrt5W
Agrt5
Brume
Rhilo
Caucr
Rhime
Brusu
Ricen
Ricpr
Xanca
Xanax
Xylfa
Wigbr
Bucai
Bucap
Haein
Pasmu
Salti
Salty
EcoliK
EcoliE
EcoliO
EcoliC
Shfl
YerpeC
YerpeK
Vibvu
Vibch
Sheon
Pseae
Psepu
Ralso
NeimeZ
NeimeM
Camje
Helpy
Helpj
Fusnu
Thete
Clope
Cloab
Oceih
Bachd
Bacsu
Bacan
Lismo
Lisin
StaauW
StaauN
StaauM
Staep
Lacla
Strmu
StrpyG
StrpyM
StrpyS
StragN
StragV
StrpnT
StrpnR
Mycpu
Urepa
Mycpe
Mycge
Mycpn
Biflo
Corgl
Coref
Mycle
MyctuC
MyctuH
Strco
Deira
Synpc
Anasp
Theel
Borbu
Trepa
Lepin
Chlte
Aquae
Thema
Metac
Metma
Metth
Metka
Metja
Pyrab
Pyrho
Pyrfu
Pyrae
Aerpe
Sulso
Sulto
Arcfu
Theac
Thevo
Halsp
Plafa
Schpo
Yeast
Arath
Worm
Enccu

Using pre-/absence of words

- Any sequence S as above can be represented also by just the set $X(\mathrm{S})$ of its K-strings. In a "gedanken alignment", ignoring repeats and collisions, we may put

$$Q^{(K)}(\mathrm{S},\mathrm{S}') := \frac{1}{2}\left(\frac{\left|X(\mathrm{S}) \bigcap X(\mathrm{S}')\right|}{\left|X(\mathrm{S})\right|} + \frac{\left|X(\mathrm{S}) \bigcap X(\mathrm{S}')\right|}{\left|X(\mathrm{S}')\right|} \right)$$

and $$q^{(K)}(\mathrm{S},\mathrm{S}') := Q^{(K)}(\mathrm{S},\mathrm{S}')^{1/K}$$

or $$:= Q^{(K+1)}(\mathrm{S},\mathrm{S}') / Q^{(K)}(\mathrm{S},\mathrm{S}')$$

Making up alignments

- Pattern:

  1: key, must match

  0: space, to watch

- "Align" pairwise:

  For keys occur in both sequences, concatenate letters in spaces.

- Sequence-based distance

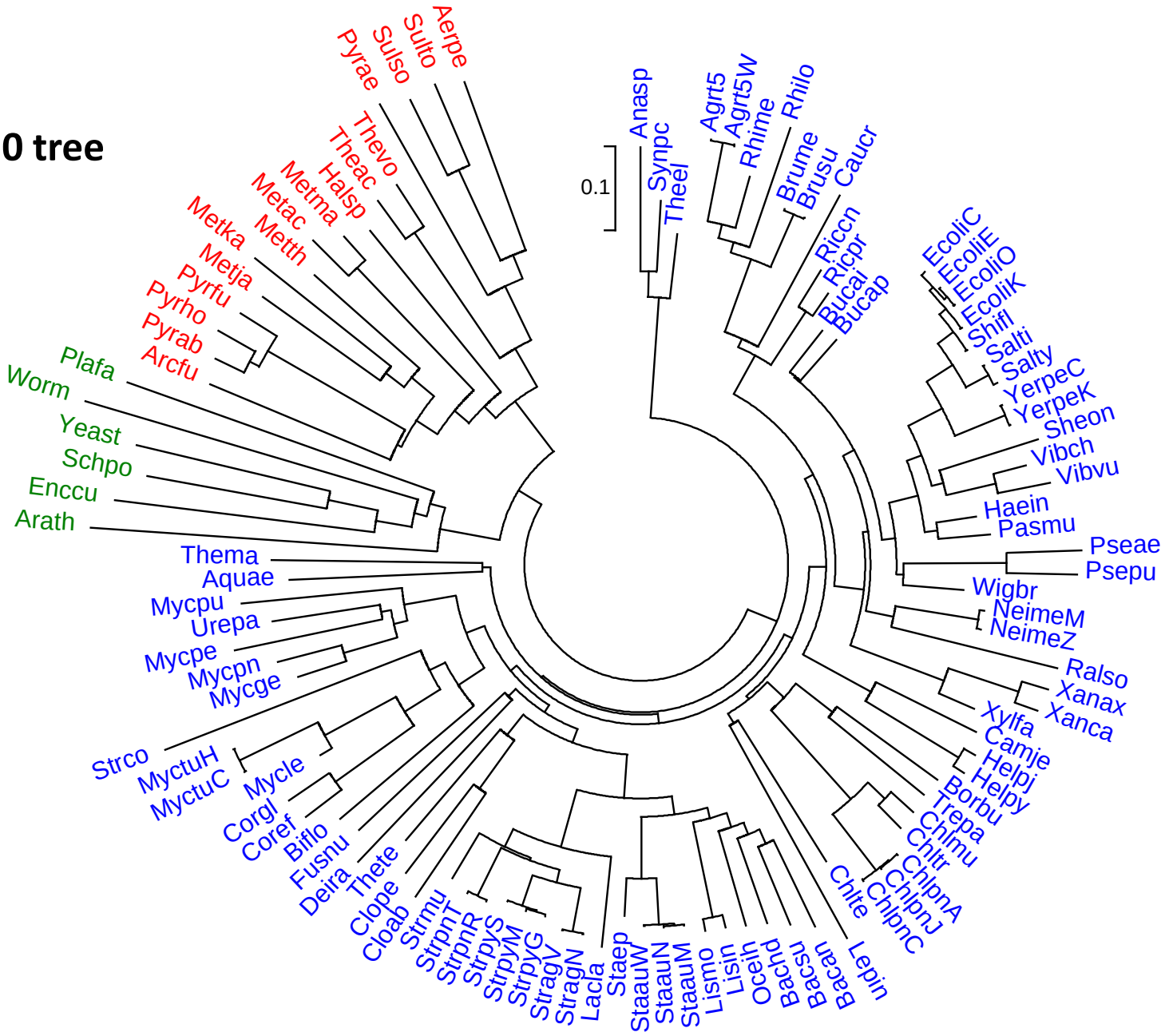
111110

X: VKAAWG,HAGEYG…
Y: VKAAWS,HAGEYG…

GG…
SG…

$$\hat{p} = n_d / n$$

011111110 tree



The missing words approach

- Probabilities of double- vs. single-missing

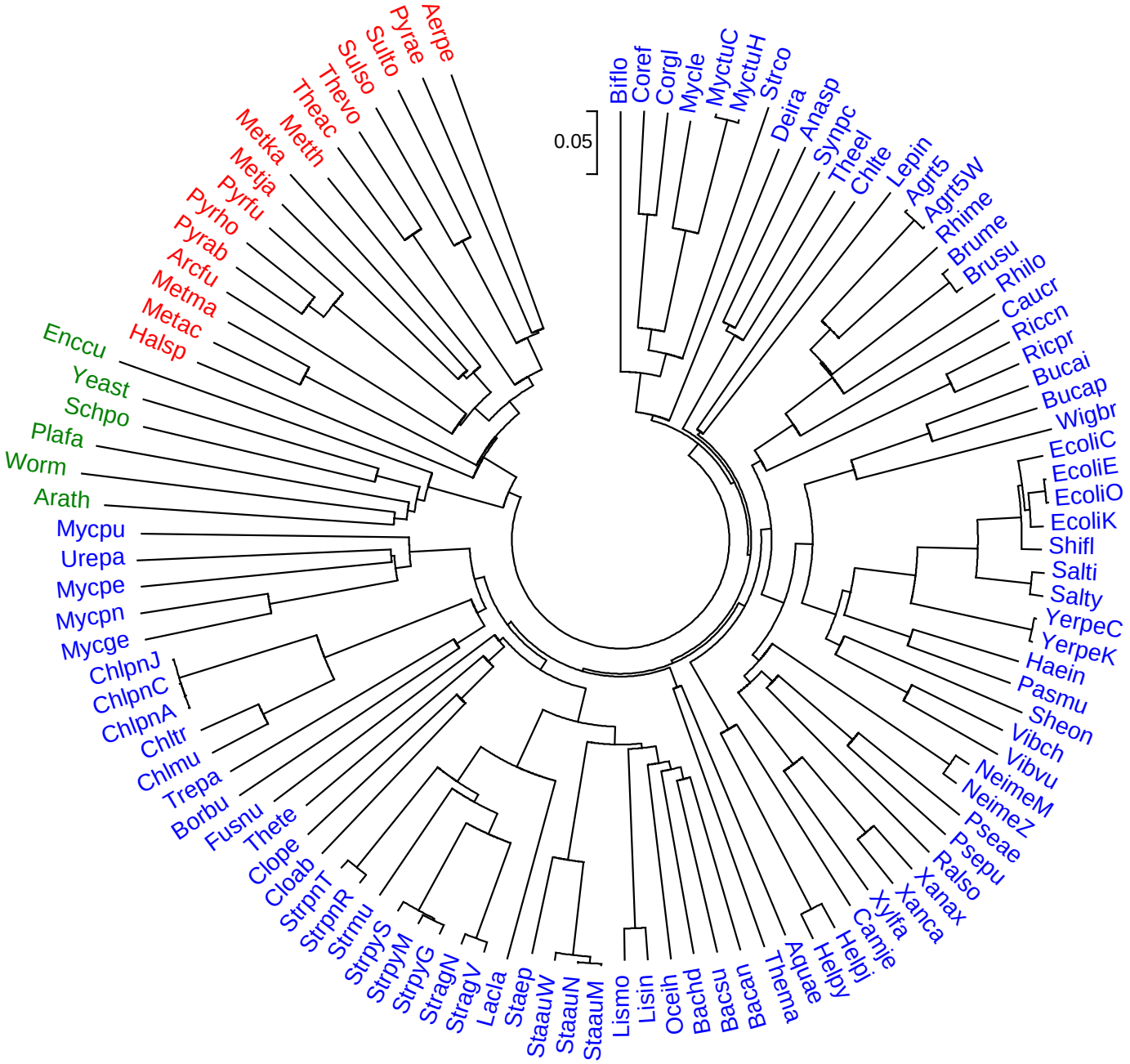
$$p_0 = u^{2-q^K} \Rightarrow \hat{q} = \left(2 - \frac{\ln p_0}{\ln u}\right)^{1/K}$$

- Estimation of probabilities

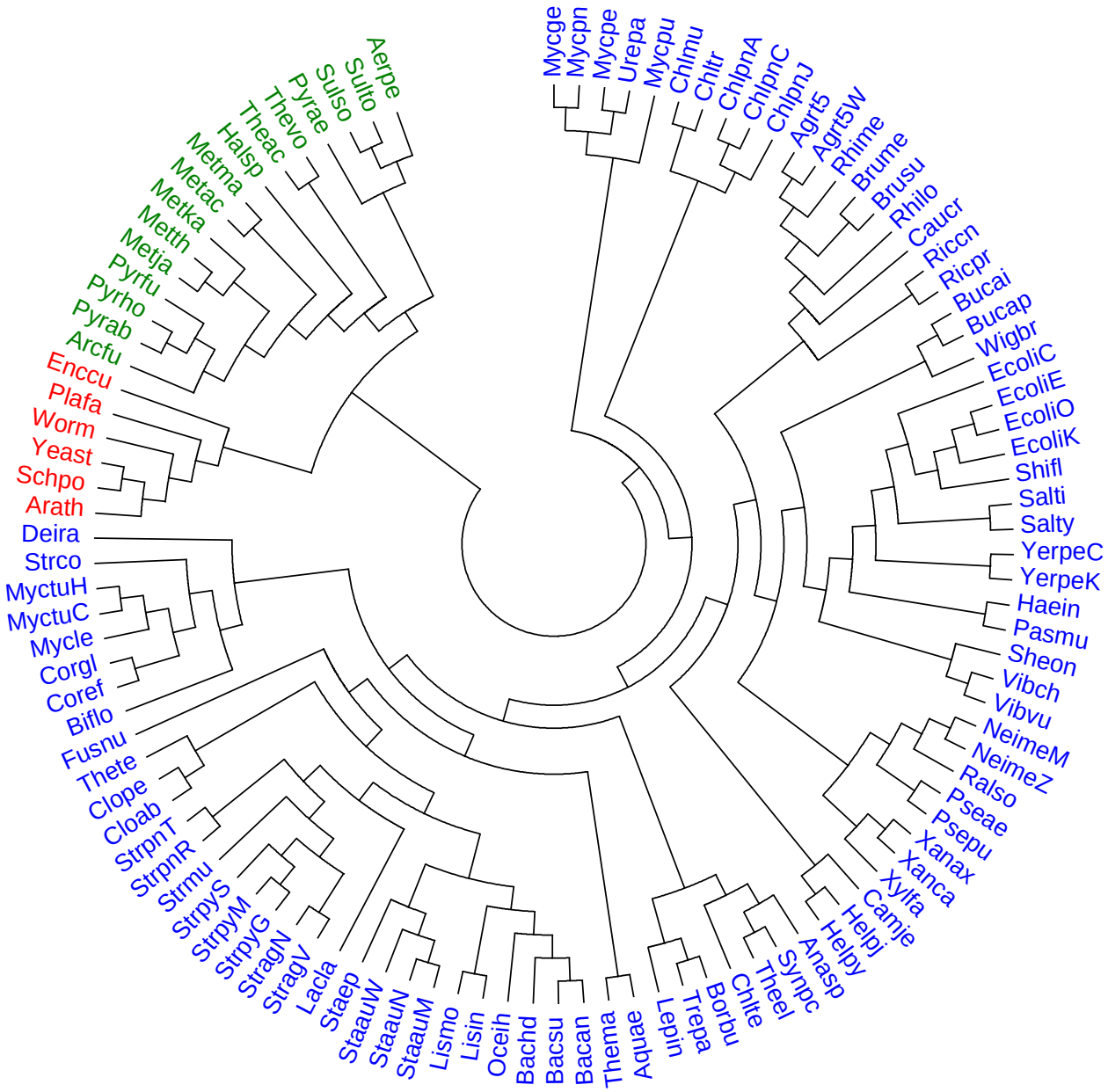
$$\hat{p}_0 = \frac{n_0}{|\Sigma|^K}$$

$$\hat{u} = \sqrt{\frac{n_0 + n_{\bar{B}}}{|\Sigma|^K} \cdot \frac{n_0 + n_{\bar{A}}}{|\Sigma|^K}}$$

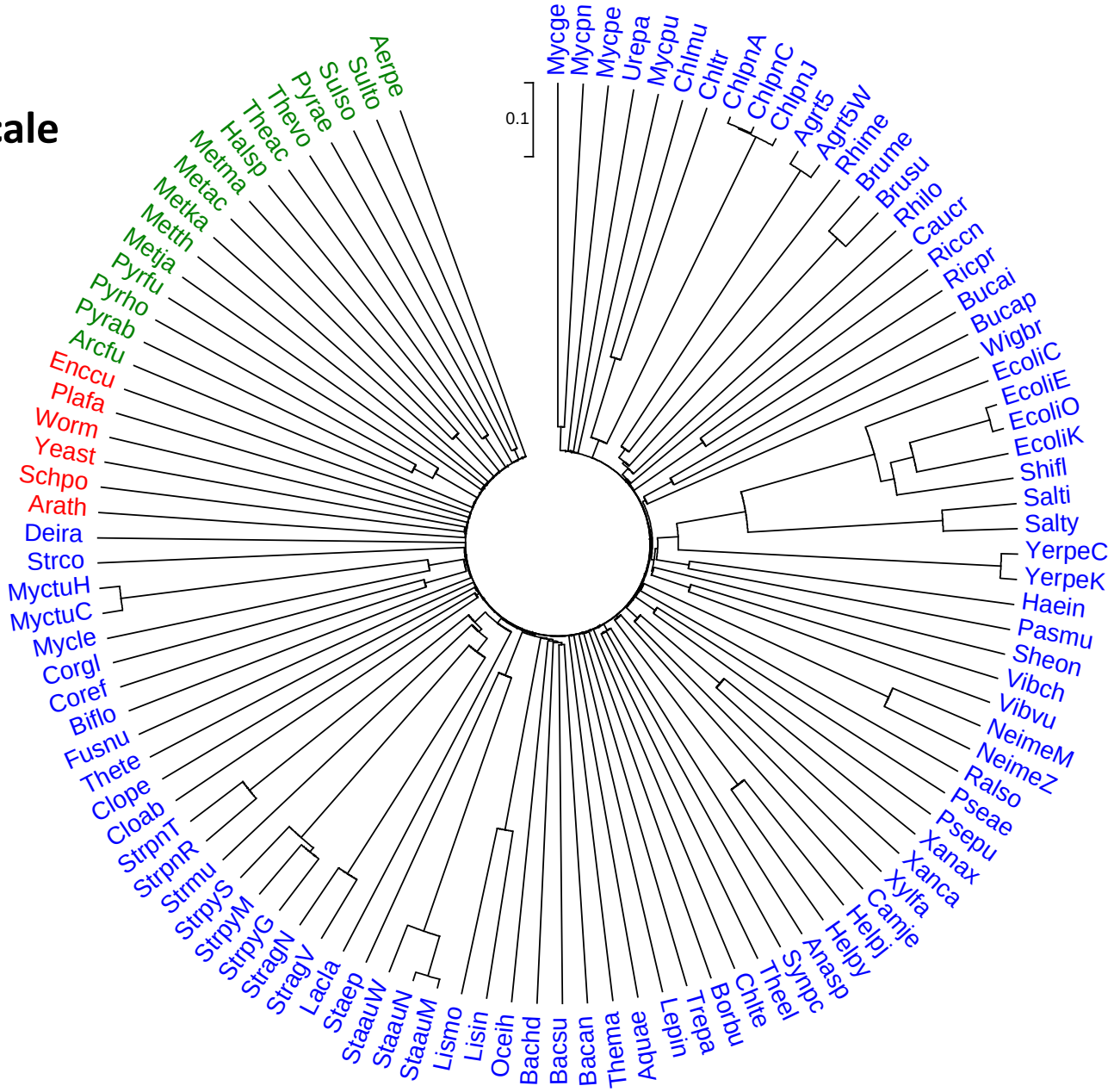
MW9 Tree


0.05
Aerpe
Pyrae
Sulto
Sulso
Thevo
Theac
Metth
Metka
Metja
Pyrfu
Pyrho
Pyrab
Arcfu
Metma
Metac
Halsp
Enccu
Yeast
Schpo
Plafa
Worm
Arath
Mycpu
Urepa
Mycpe
Mycpn
Mycge
ChlpnJ
ChlpnC
ChlpnA
Chltr
Chlmu
Trepa
Borbu
Fusnu
Thete
Clope
Cloab
StrpnT
StrpnR
Strmu
StrpyS
StrpyM
StrpyG
StragN
StragV
Lacla
Staep
StaauW
StaauN
StaauM
Lismo
Lisin
Oceih
Bachd
Bacsu
Bacan
Thema
Aquae
Helpy
Helpj
Camje
Xylfa
Xanca
Xanax
Ralso
Psepu
Pseae
NeimeZ
NeimeM
Vibvu
Vibch
Sheon
Pasmu
Haein
YerpeK
YerpeC
Salty
Salti
Shifl
EcoliK
EcoliO
EcoliE
EcoliC
Wigbr
Bucap
Bucai
Ricpr
Riccn
Caucr
Rhilo
Brusu
Brume
Rhime
Agrt5W
Agrt5
Lepin
Chlte
Theel
Synpc
Anasp
Deira
Strco
MyctuH
MyctuC
Mycle
Corgl
Coref
Biflo

A tree based on "Jensen-Shannon" distance $JSh(S, S')$, the "symmetrized" Kullback-Leibler distance between the "induced K-string probabilities"

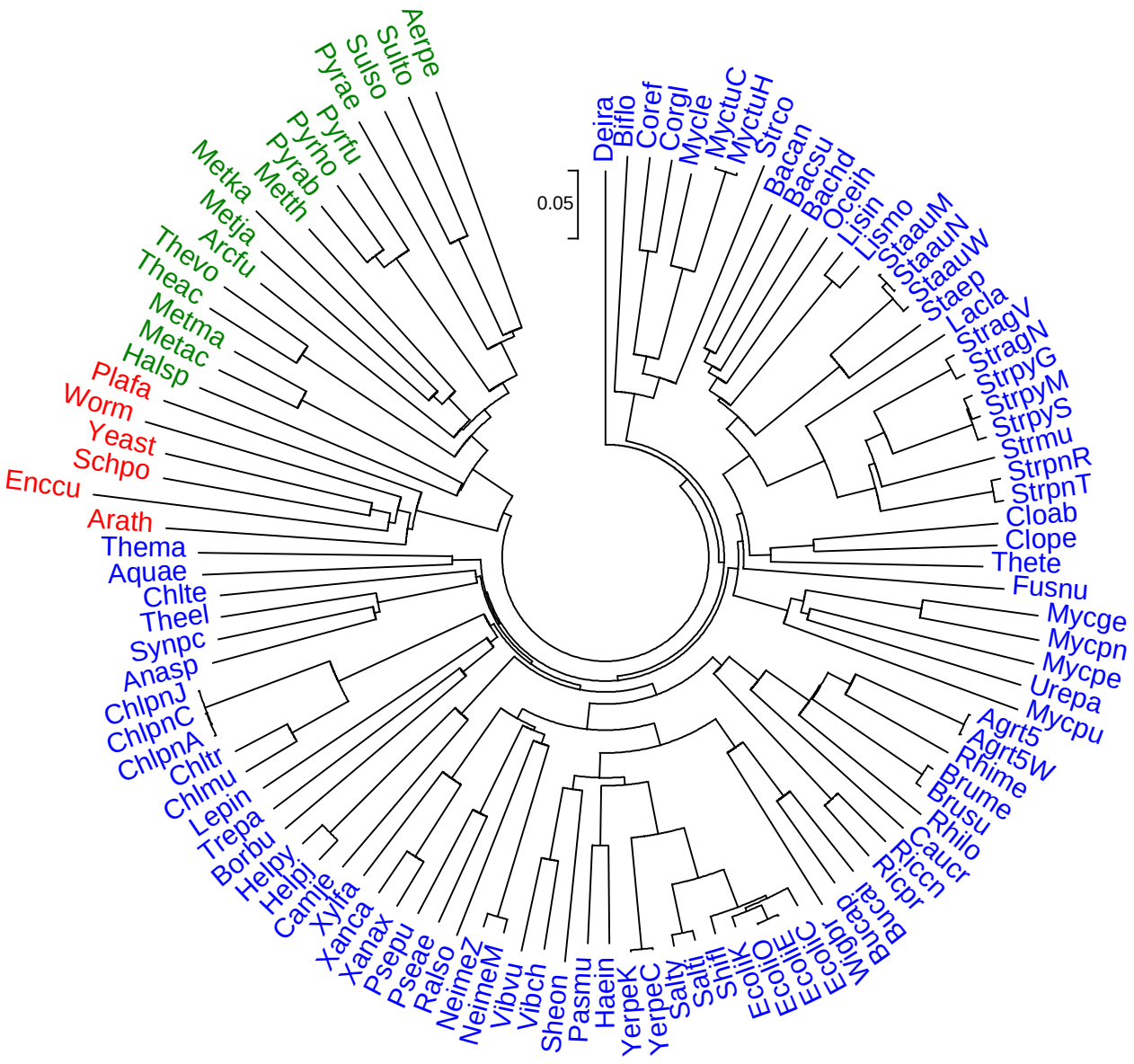


The tree drawn to scale

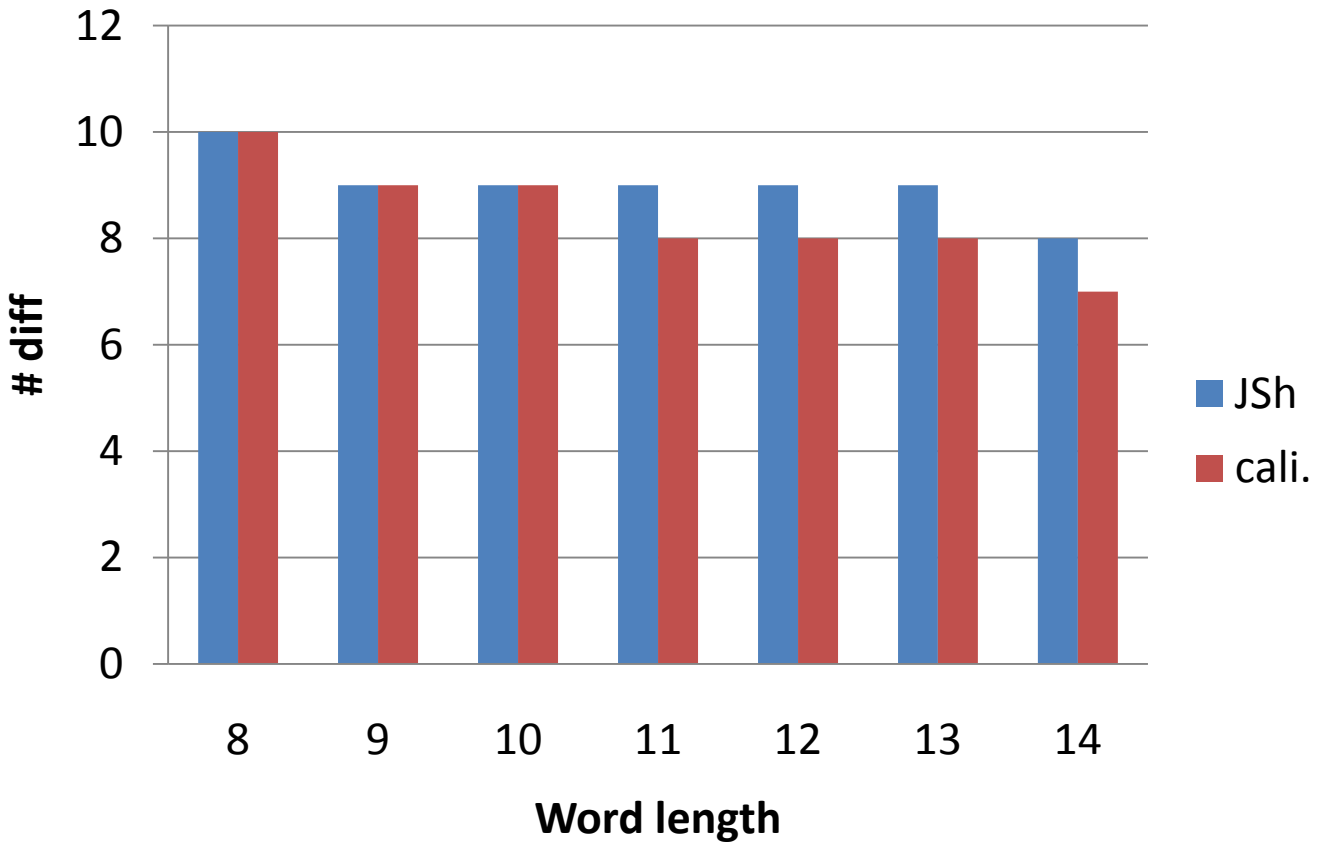


The tree using the calibrated distance

$$1-\left(1-JSh(\mathrm{S},\mathrm{S}')\right)^{1/K}$$



Comparison with “standard” taxonomy



Outlook

- More realistic evolutionary dynamics
- Taking account of the variance of the frequency distribution
  - Better construction of composition vector
  - Optimal value of K

Acknowledgement

- Bailin Hao
- Andreas Dress
- Zhao Xu

Thank You!