Diagnosing phylogenetic conflict between genes and proteins: Evidence for the origin of plastids

Blaise Li Peter G. Foster T. Martin Embley Cymon J. Cox

SMBE Meeting - Dublin 2012 - 24/06/2012

The questions

- Archaeplastida (plants) = Rhodophyta (red algae) + Glaucophyta + Viridiplantae (green plants)

The questions

- Archaeplastida (plants) = Rhodophyta (red algae) + Glaucophyta + Viridiplantae (green plants)
  → Characterised by a plastid, with endosymbiotic origin within cyanobacteria

[1] Notation adapted from Criscuolo and Gribaldo (2011)

The questions

- Archaeplastida (plants) = Rhodophyta (red algae) + Glaucophyta + Viridiplantae (green plants)
  → Characterised by a plastid, with endosymbiotic origin within cyanobacteria
- Groups of cyanobacteria already identified: NOST-1, OSC-2, SPM-3, SO-6, GBACT, UNIT+[1]

[1] Notation adapted from Criscuolo and Gribaldo (2011)

The questions

- Archaeplastida (plants) = Rhodophyta (red algae) + Glaucophyta + Viridiplantae (green plants)
  $\rightarrow$ Characterised by a plastid, with endosymbiotic origin within cyanobacteria
- Groups of cyanobacteria already identified: NOST-1, OSC-2, SPM-3, SO-6, GBACT, UNIT+[1]
- What are the relationships between plastids and cyanobacteria?

[1] Notation adapted from Criscuolo and Gribaldo (2011)

The questions

- Archaeplastida (plants) = Rhodophyta (red algae) + Glaucophyta + Viridiplantae (green plants)
  → Characterised by a plastid, with endosymbiotic origin within cyanobacteria
- Groups of cyanobacteria already identified: NOST-1, OSC-2, SPM-3, SO-6, GBACT, UNIT+[1]
- What are the relationships between plastids and cyanobacteria?
- Why do the gene tree and the protein tree give a different answer?

[1] Notation adapted from Criscuolo and Gribaldo (2011)

Dataset

- 42 taxa, including 8 outgoup (non-cyano)bacteria, 16 cyanobacteria, and plastids from 1 Glaucophyta, 4 Rhodophyta and 13 Viridiplantae

Dataset

- 42 taxa, including 8 outgoup (non-cyano)bacteria, 16 cyanobacteria, and plastids from 1 Glaucophyta, 4 Rhodophyta and 13 Viridiplantae
- 75 protein-coding genes, but 452 missing sequences (i.e. 14% overall, and up to 38 genes missing for one of the outgroup taxa)

Dataset

- 42 taxa, including 8 outgoup (non-cyano)bacteria, 16 cyanobacteria, and plastids from 1 Glaucophyta, 4 Rhodophyta and 13 Viridiplantae
- 75 protein-coding genes, but 452 missing sequences (i.e. 14% overall, and up to 38 genes missing for one of the outgroup taxa)
- Concatenated dataset (cg75)

Analysis method

- Nucleotide dataset (cg75)

Analysis method

- Nucleotide dataset (cg75) and amino-acid translation (cp75)

Analysis method

- Nucleotide dataset (cg75) and amino-acid translation (cp75)
- Various recodings of cg75 using IUPAC ambiguity codes to remove signal associated with some synonymous substitutions.

Analysis method

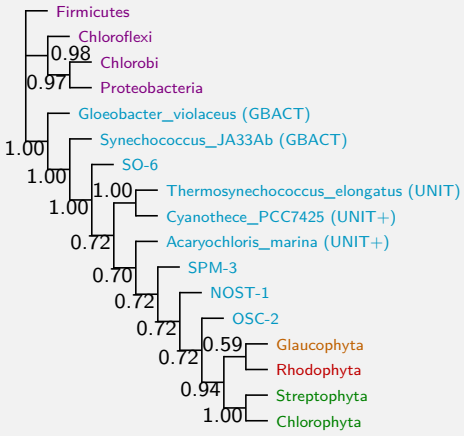
- Nucleotide dataset (cg75) and amino-acid translation (cp75)
- Various recodings of cg75 using IUPAC ambiguity codes to remove signal associated with some synonymous substitutions. e.g for His: CAC $\leftrightarrow$ CAT (synonymous codons) $\rightarrow$ CAY (degenerate codon)

Analysis method

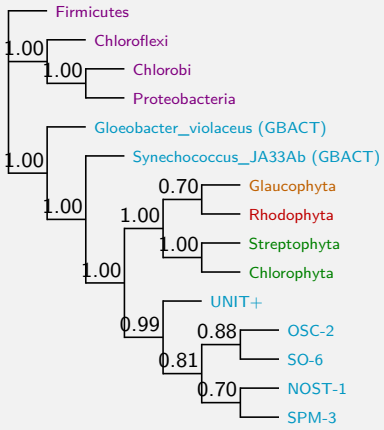
- Nucleotide dataset (cg75) and amino-acid translation (cp75)
- Various recodings of cg75 using IUPAC ambiguity codes to remove signal associated with some synonymous substitutions. e.g for His: CAC ↔ CAT (synonymous codons) → CAY (degenerate codon)
- Maximum likelihood bootstrap analyses (200 resamplings) using RAxML, with GTR+I+Γ (or LG+I+Γ for cp75, chosen using ProtTest)

Unrecoded ML bootstrap analyses

translation



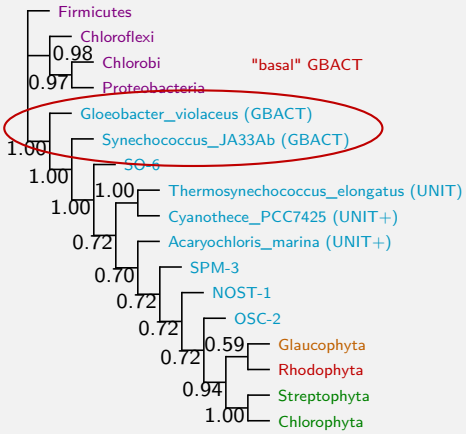
cg75



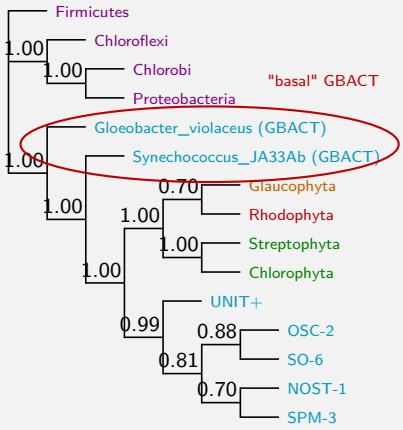
cp75

Unrecoded ML bootstrap analyses

translation

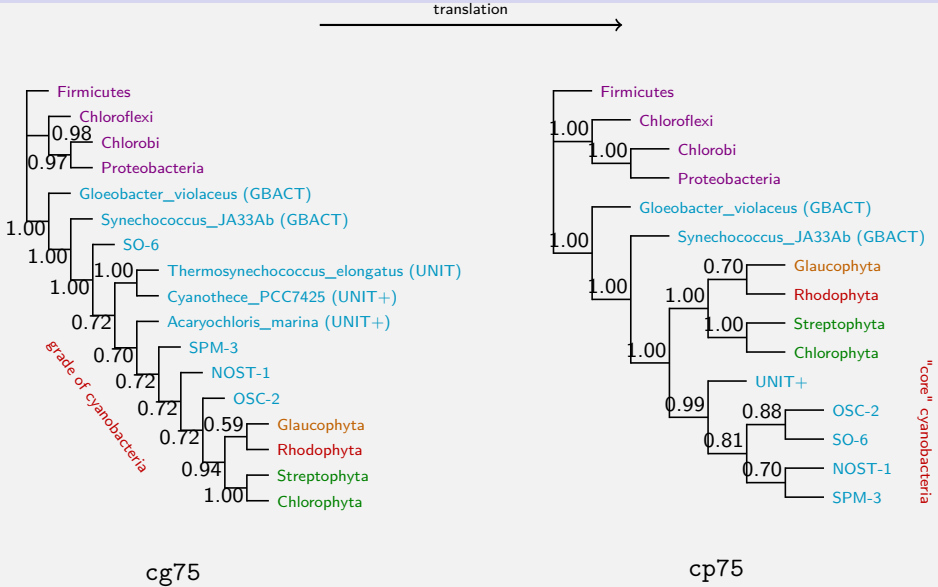


cg75



cp75

Unrecoded ML bootstrap analyses



Unrecoded ML bootstrap analyses

- cp75 is a direct translation of cg75

Unrecoded ML bootstrap analyses

- cp75 is a direct translation of cg75
  $\rightarrow$ The trees should be the same.

Unrecoded ML bootstrap analyses

- cp75 is a direct translation of cg75
  $\rightarrow$ The trees should be the same.
- But the analyses conflict in the identification of the plastid sister-group.

Unrecoded ML bootstrap analyses

- cp75 is a direct translation of cg75
  $\rightarrow$ The trees should be the same.
- But the analyses conflict in the identification of the plastid sister-group.
  $\rightarrow$ Something is not well modelled.

Unrecoded ML bootstrap analyses

- cp75 is a direct translation of cg75
  $\rightarrow$ The trees should be the same.
- But the analyses conflict in the identification of the plastid sister-group.
  $\rightarrow$ Something is not well modelled.
- Data recoding experiments to identify some causes of model mis-specification.

Degenerating synonymous 3rd codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT

Degenerating synonymous 3rd codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

```
ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT
```

Degenerating synonymous 3rd codon positions

	T		C		A		G	
T	TTY TTY	Phe	TCN TCN	Ser	TAY TAY	Tyr	TGY TGY	Cys
	TTN TTN	Leu	TCN TCN		TAR TAR	Ter	TGR	Ter
							TGG	Trp
C	CTN CTN CTN CTN	Leu	CCN CCN CCN CCN	Pro	CAY CAY	His	CGN CGN	Arg
					CAR CAR	Gln	CGN CGN	
A	ATH ATH ATH	Ile	ACN ACN ACN ACN	Thr	AAY AAY	Asn	AGN AGN	Ser
	ATG	Met			AAR AAR	Lys	AGN AGN	Arg
G	GTN GTN GTN GTN	Val	GCN GCN GCN GCN	Ala	GAY GAY	Asp	GGN GGN	Gly
					GAR GAR	Glu	GGN GGN	

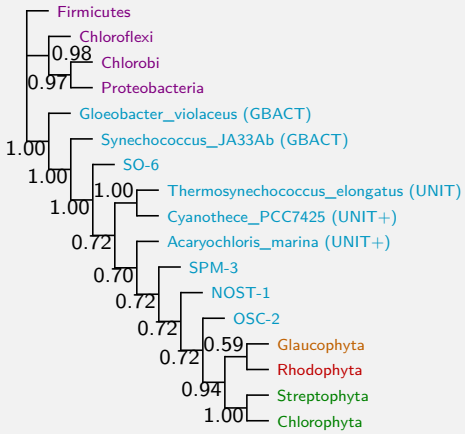
Example:

```
ATGAARCARTTNCTNGARGCNGGNGTNCAYTTY
TTNTCNGARCTNTTNGAYGCNAGNGCNCAYATH
TTNGARGAYATGATHCARAGNGGNATGCAYTTY
CTNGARCARATGCTNGAYGTNGGNGTNCAYTTY
TTNCARTCNATGCTNGARGCNGGNGTNCAYTTY
TTNGARGCNCTNTTNGARGCNGGNGTNCAYTTY
TTNGARGARATGATGGARGCNGGNATHCAYTTY
TTNGARCARATGATGGARGCNGGNGTNCAYTTY
ATHCARCARTTNTTNGARGCNGGNGTNCAYCTN
ATHGARGARTTNTTNGARGCNGGNGTNCAYTTY
TTNGARCARATGTTNGAYGCNGGNGTNCAYTTY
TTNCARAARATGATHGARGCNGGNGTNCAYTTY
CTNTCNGARATGATGGARGCNGGNGCNCAYTTY
CTNCCNCARATGCTNGARGCNGGNGTNCAYTTY
TTNGARGARATGATGGARGCNGGNGTNCAYTTY
TTNGCNGARTTNCTNGARGCNGGNGTNCAYTTY
```

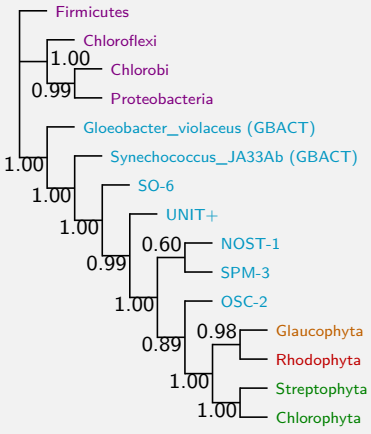
Degenerating synonymous 3rd codon positions

degenerate at 3rd pos.

(27.35% recoded)

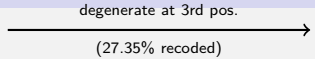


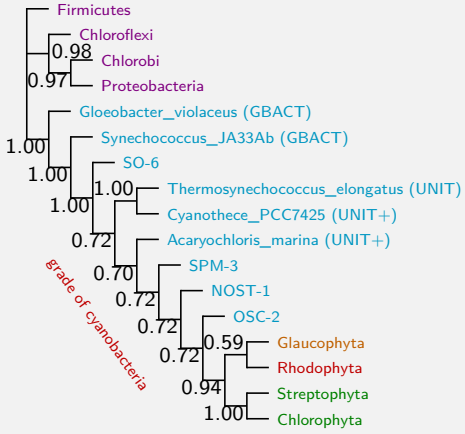
cg75



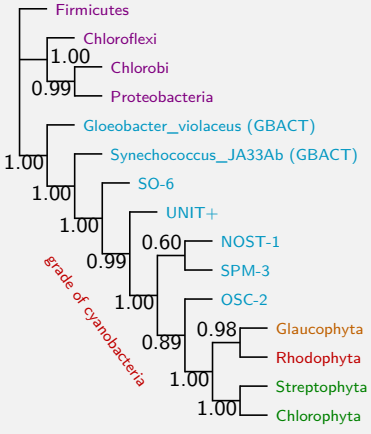
cg75_degenerate3

Degenerating synonymous 3rd codon positions





cg75



cg75_degenerate3

Degenerating synonymous 3rd codon positions

- No "core" cyanobacteria monophyly.

Degenerating synonymous 3rd codon positions

- No "core" cyanobacteria monophyly.
  $\rightarrow$ Synonymous substitutions at 3rd codon position are not the only cause of discrepancy between cp75 and cg75.

Degenerating synonymous 3rd codon positions

- No "core" cyanobacteria monophyly.
  $\rightarrow$ Synonymous substitutions at 3rd codon position are not the only cause of discrepancy between cp75 and cg75.
- But there are synonymous substitutions also at 1st and 2nd position:

Degenerating synonymous 3rd codon positions

- No "core" cyanobacteria monophyly.
  $\rightarrow$ Synonymous substitutions at 3rd codon position are not the only cause of discrepancy between cp75 and cg75.
- But there are synonymous substitutions also at 1st and 2nd position:
  e.g for Leu: (TTA $\leftrightarrow$ CTA) $\rightarrow$ YTA
  e.g for Arg: (CGA $\leftrightarrow$ AGA) $\rightarrow$ MGA
  e.g for Ser: (TCT $\leftrightarrow$ AGA) $\rightarrow$ WST

Degenerating synonymous 1st and 2nd codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT

Degenerating synonymous 1st and 2nd codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT

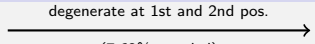
Degenerating synonymous 1st and 2nd codon positions

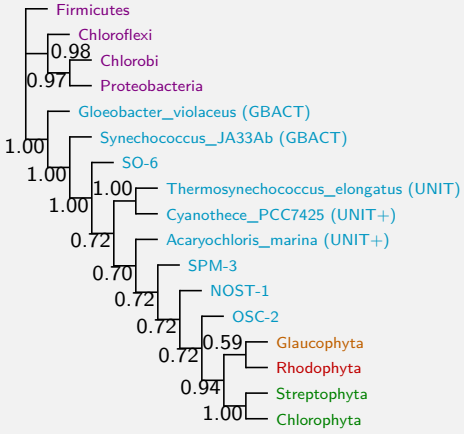
	T		C		A		G	
T	TTT	Phe	WST	Ser	TAT	Tyr	TGT	Cys
	TTC		WSC		TAC		TGC	
	YTA	Leu	WSA		TAA	Ter	TGA	Ter
	YTG		WSG		TAG		TGG	Trp
C	YTT	Leu	CCT	Pro	CAT	His	MGT	Arg
	YTC		CCC		CAC		MGC	
	YTA		CCA		CAA	Gln	MGA	
	YTG		CCG		CAG		MGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	WST	Ser
	ATC		ACC		AAC		WSC	
	ATA		ACA		AAA	Lys	MGA	Arg
	ATG	Met	ACG		AAG		MGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

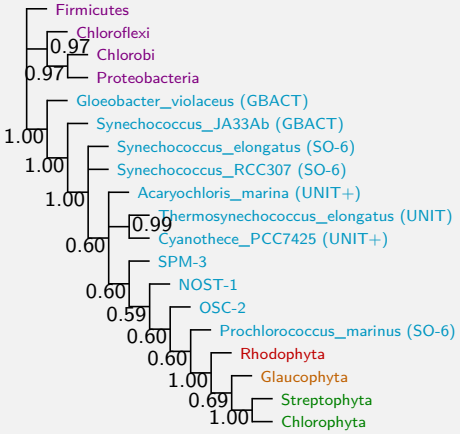
```
ATGAAACAGYTGYTGGAAGCCGGTGTTCACTTC
YTAWSAGAAYTTYTAGATGCCWSCGCTCACATA
YTAGAAGATATGATTCAAWSTGGAATGCATTTT
YTCGAACAAATGYTAGATGTAGGTGTTCATTTT
YTACAGWSAATGYTTGAAGCTGGTGTTCACTTT
YTAGAAGCAYTTYTAGAGGCTGGTGTTCATTTT
YTGGAAGAAATGATGGAAGCGGGTATTCATTTC
YTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAAYTAYTGGAAGCAGGAGTCCATYTG
ATTGAAGAAYTGYTAGAAGCTGGCGTGCATTTC
YTAGAACAAATGYTAGATGCAGGTGTACATTTT
YTACAAAAAATGATTGAAGCTGGTGTTCATTTT
YTAWSAGAAATGATGGAAGCTGGTGCTCATTTC
YTGCCGCAAATGYTGGAAGCCGGTGTCCATTTT
YTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
YTAGCAGAAYTAYTAGAAGCGGGCGTTCATTTT
```

Degenerating synonymous 1st and 2nd codon positions



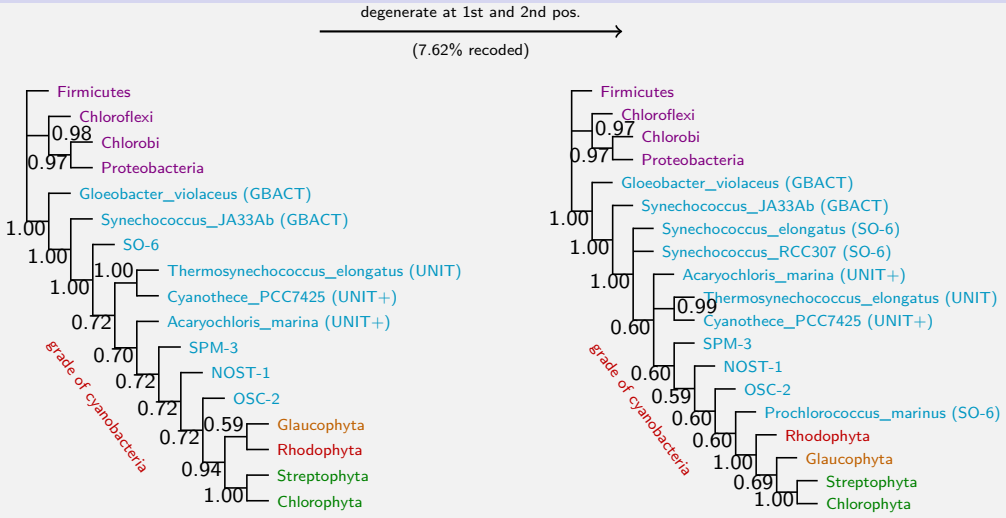


cg75



cg75_degenerate12

Degenerating synonymous 1st and 2nd codon positions



Degenerating synonymous 1st and 2nd codon positions

- Again no "core" cyanobacteria monophyly.

Degenerating synonymous 1st and 2nd codon positions

- Again no "core" cyanobacteria monophyly.
  $\rightarrow$ Synonymous substitutions at 1st and 2nd codon position are not the only cause of discrepancy between cp75 and cg75.

Degenerating synonymous 1st and 2nd codon positions

- Again no "core" cyanobacteria monophyly.
  $\rightarrow$ Synonymous substitutions at 1st and 2nd codon position are not the only cause of discrepancy between cp75 and cg75.
- Let's try to neutralize all synonymous substitutions...

Degenerating all synonymous codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT

Degenerating all synonymous codon positions

	T		C		A		G	
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	Ter	TGA	Ter
	TTG		TCG		TAG		TGG	Trp
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

Example:

```
ATGAAACAGTTGCTGGAAGCCGGTGTTCACTTC
TTATCAGAACTTTTAGATGCCAGCGCTCACATA
TTAGAAGATATGATTCAAAGTGGAATGCATTTT
CTCGAACAAATGCTAGATGTAGGTGTTCATTTT
TTACAGTCAATGCTTGAAGCTGGTGTTCACTTT
TTAGAAGCACTTTTAGAGGCTGGTGTTCATTTT
TTGGAAGAAATGATGGAAGCGGGTATTCATTTC
TTGGAACAAATGATGGAAGCAGGAGTCCATTTT
ATTCAGCAATTATTGGAAGCAGGAGTCCATCTG
ATTGAAGAATTGTTAGAAGCTGGCGTGCATTTC
TTAGAACAAATGTTAGATGCAGGTGTACATTTT
TTACAAAAAATGATTGAAGCTGGTGTTCATTTT
CTATCAGAAATGATGGAAGCTGGTGCTCATTTC
CTGCCGCAAATGCTGGAAGCCGGTGTCCATTTT
TTAGAAGAAATGATGGAAGCAGGGGTCCATTTT
TTAGCAGAATTACTAGAAGCGGGCGTTCATTTT
```

Degenerating all synonymous codon positions

	T		C		A		G	
T	TTY	Phe	WSN	Ser	TAY	Tyr	TGY	Cys
	TTY		WSN		TAY		TGY	
	YTN	Leu	WSN		TAR	Ter	TGR	Ter
	YTN		WSN		TAR		TGG	Trp
C	YTN	Leu	CCN	Pro	CAY	His	MGN	Arg
	YTN		CCN		CAY		MGN	
	YTN		CCN		CAR	Gln	MGN	
	YTN		CCN		CAR		MGN	
A	ATH	Ile	ACN	Thr	AAY	Asn	WSN	Ser
	ATH		ACN		AAY		WSN	
	ATH		ACN		AAR	Lys	MGN	Arg
	ATG	Met	ACN		AAR		MGN	
G	GTN	Val	GCN	Ala	GAY	Asp	GGN	Gly
	GTN		GCN		GAY		GGN	
	GTN		GCN		GAR	Glu	GGN	
	GTN		GCN		GAR		GGN	

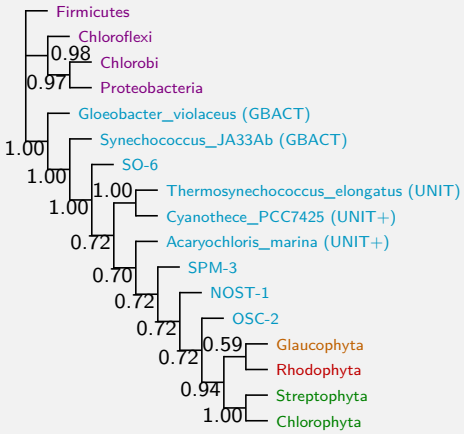
Example:

```
ATGAARCARYTNYTNGARGCNGGNGTNCAYTTY
YTNWSNGARYTNYTNGAYGCNWSNGCNCAYATH
YTNGARGAYATGATHCARWSNGGNATGCAYTTY
YTNGARCARATGYTNGAYGTNGGNGTNCAYTTY
YTNCARWSNATGYTNGARGCNGGNGTNCAYTTY
YTNGARGCNYTNYTNGARGCNGGNGTNCAYTTY
YTNGARGARATGATGGARGCNGGNATHCAYTTY
YTNGARCARATGATGGARGCNGGNGTNCAYTTY
ATHCARCARYTNYTNGARGCNGGNGTNCAYYTN
ATHGARGARYTNYTNGARGCNGGNGTNCAYTTY
YTNGARCARATGYTNGAYGCNGGNGTNCAYTTY
YTNCARAARATGATHGARGCNGGNGTNCAYTTY
YTNWSNGARATGATGGARGCNGGNGCNCAYTTY
YTNCCNCARATGYTNGARGCNGGNGTNCAYTTY
YTNGARGARATGATGGARGCNGGNGTNCAYTTY
YTNGCNGARYTNYTNGARGCNGGNGTNCAYTTY
```

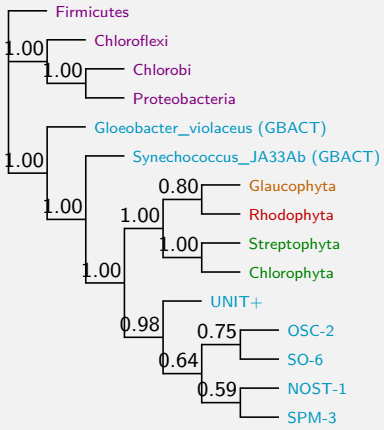
Degenerating all synonymous codon positions

degenerate at 1st, 2nd and 3rd pos.

(34.97% recoded)



cg75

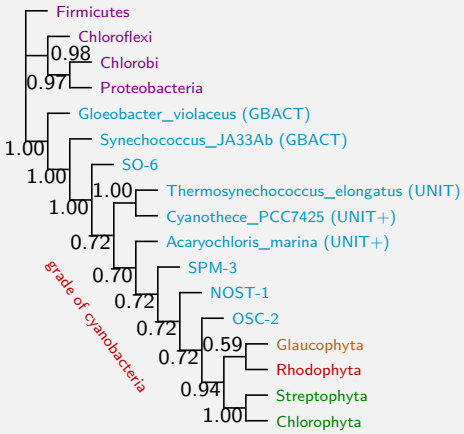


cg75_degenerate123

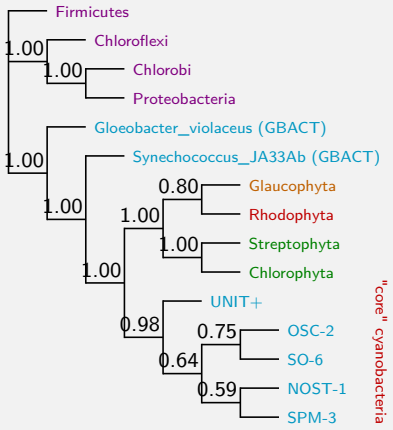
Degenerating all synonymous codon positions

degenerate at 1st, 2nd and 3rd pos.

(34.97% recoded)



cg75



cg75_degenerate123

Degenerating all synonymous codon positions

- Same topology as with the protein data

Degenerating all synonymous codon positions

- Same topology as with the protein data
- Synonymous substitutions are responsible for the incongruence between nucleotide and amino-acid data.

Degenerating all synonymous codon positions

- Same topology as with the protein data
- Synonymous substitutions are responsible for the incongruence between nucleotide and amino-acid data.
- This could be explained by effects of biases in codon usage and DNA composition:

Degenerating all synonymous codon positions

- Same topology as with the protein data
- Synonymous substitutions are responsible for the incongruence between nucleotide and amino-acid data.
- This could be explained by effects of biases in codon usage and DNA composition:
  - Codon preference / composition biases: "allowed" within a synonymy class (neutral at the protein level and above)

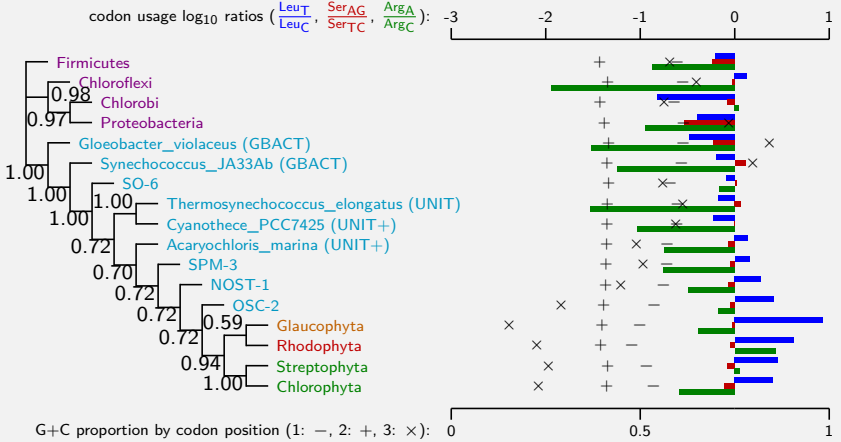
Degenerating all synonymous codon positions

- Same topology as with the protein data
- Synonymous substitutions are responsible for the incongruence between nucleotide and amino-acid data.
- This could be explained by effects of biases in codon usage and DNA composition:
  - Codon preference / composition biases: "allowed" within a synonymy class (neutral at the protein level and above)
    $\rightarrow$ Convergence between taxa sharing the same biases

Degenerating all synonymous codon positions

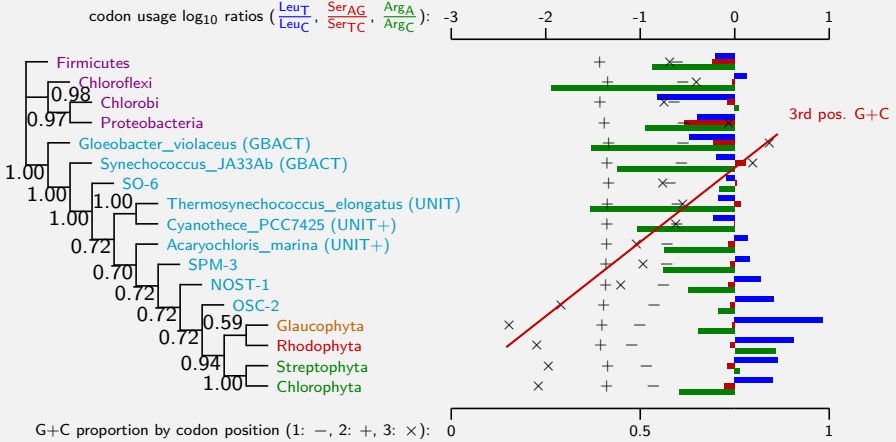
- Same topology as with the protein data
- Synonymous substitutions are responsible for the incongruence between nucleotide and amino-acid data.
- This could be explained by effects of biases in codon usage and DNA composition:
  - Codon preference / composition biases: "allowed" within a synonymy class (neutral at the protein level and above)
    $\rightarrow$ Convergence between taxa sharing the same biases
  - Let's visualize this...

Effects of composition and codon usage



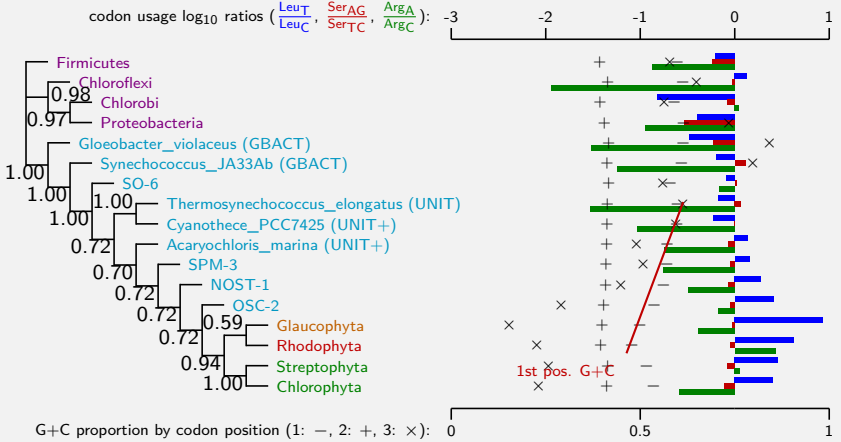
cg75

Effects of composition and codon usage



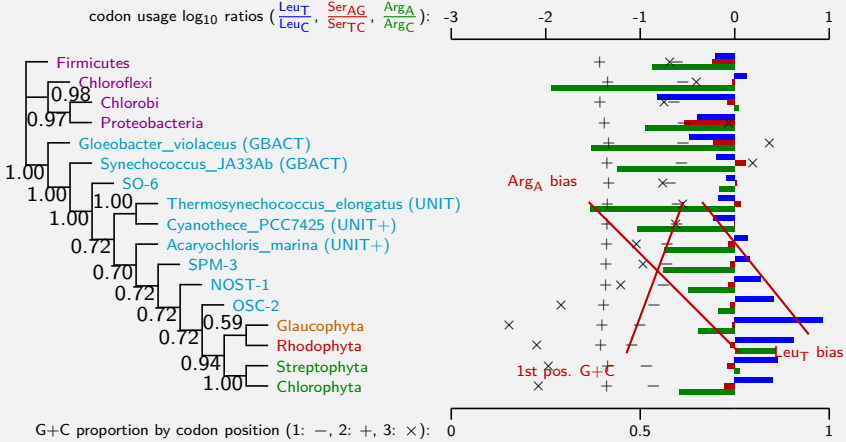
cg75

Effects of composition and codon usage



cg75

Effects of composition and codon usage



cg75

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.
  $\rightarrow$ Nucleotide data are prone to reconstruction errors for large scale relationships (more time for convergence).

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.
  $\rightarrow$ Nucleotide data are prone to reconstruction errors for large scale relationships (more time for convergence).
  $\rightarrow$ We favour the hypothesis of a monophyletic "core" cyanobacteria clade sister to plastids (protein topology).

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.
  $\rightarrow$ Nucleotide data are prone to reconstruction errors for large scale relationships (more time for convergence).
  $\rightarrow$ We favour the hypothesis of a monophyletic "core" cyanobacteria clade sister to plastids (protein topology).
- But information potentially useful for small scale resolution is lost in translated / recoded data

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.
  $\rightarrow$ Nucleotide data are prone to reconstruction errors for large scale relationships (more time for convergence).
  $\rightarrow$ We favour the hypothesis of a monophyletic "core" cyanobacteria clade sister to plastids (protein topology).
- But information potentially useful for small scale resolution is lost in translated / recoded data
- The use of more advanced models could turn misleading signal into useful signal.

The conclusions

- Synonymous substitutions enable convergence in codon usage and composition biases.
  $\rightarrow$ Nucleotide data are prone to reconstruction errors for large scale relationships (more time for convergence).
  $\rightarrow$ We favour the hypothesis of a monophyletic "core" cyanobacteria clade sister to plastids (protein topology).
- But information potentially useful for small scale resolution is lost in translated / recoded data
- The use of more advanced models could turn misleading signal into useful signal.
  $\rightarrow$ Could we then get both large and small scale resolution using un-recoded nucleotide data?

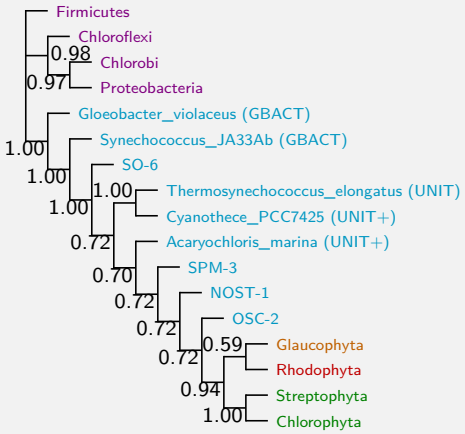
Thanks for your attention

- This work was supported by a Fundação para a Ciência e a Tecnologia (FCT, Portugal) grant to Cymon J. Cox, Centro de Ciencias do Mar (CCMAR) - CIMAR-Lab. Assoc., (PTDC/BIA-BCM/099565/2008).

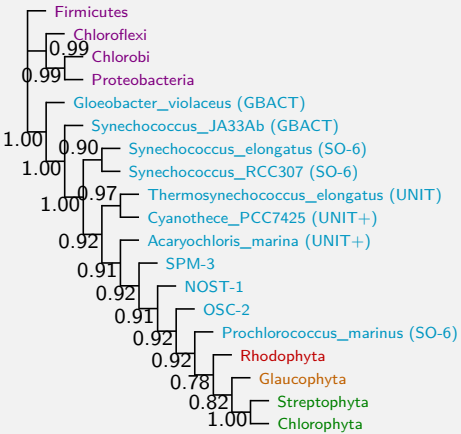
- Contact: `blaise.li@normalesup.org`

Effects of serine synonymous substitutions

degenerate serines (at pos. 1 and 2) →

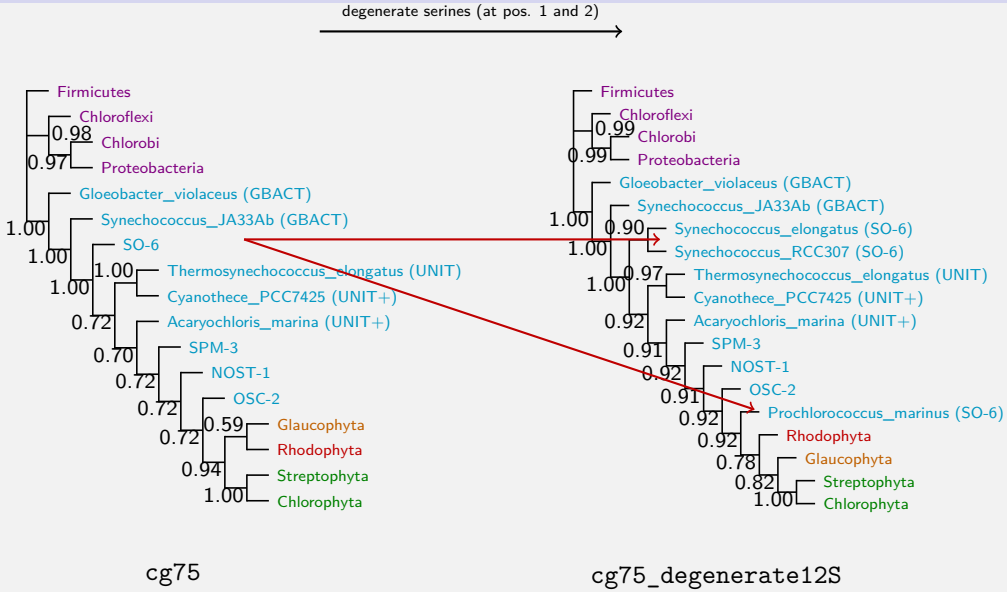


cg75

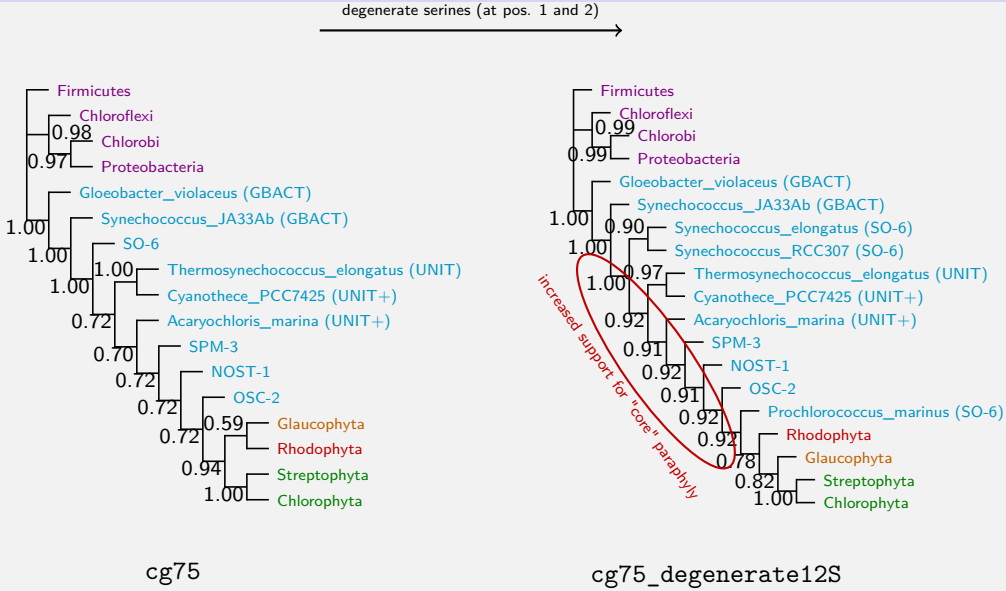


cg75_degenerate12S

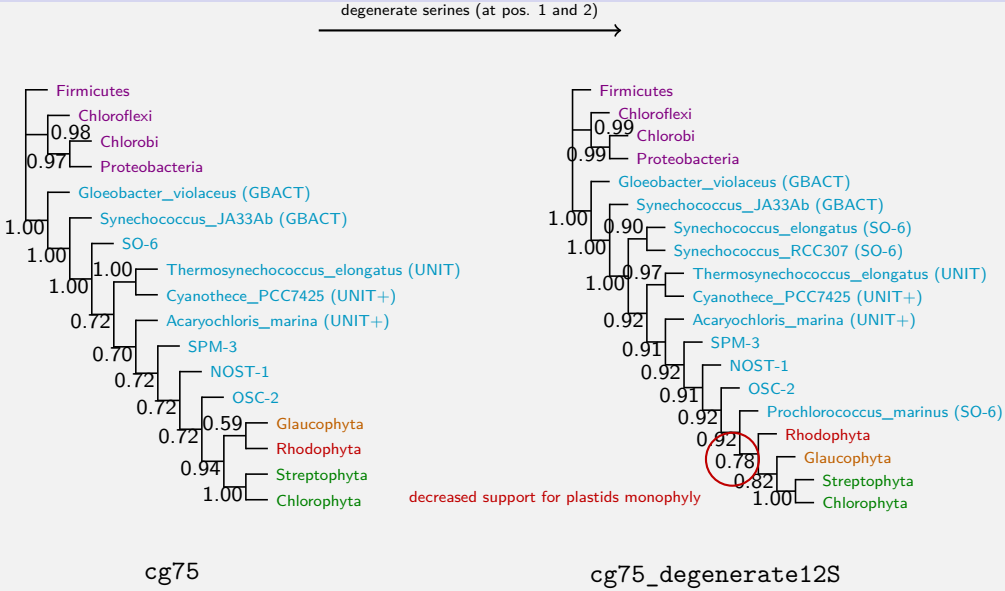
Effects of serine synonymous substitutions



Effects of serine synonymous substitutions



Effects of serine synonymous substitutions



Effects of serine synonymous substitutions

- Serine signal:

Effects of serine synonymous substitutions

- Serine signal:
  - contributes to SO-6 and plastid monophyly

Effects of serine synonymous substitutions

- Serine signal:
  - contributes to SO-6 and plastid monophyly
  - decreases the support of "core" cyanobacteria paraphyly

Effects of serine synonymous substitutions

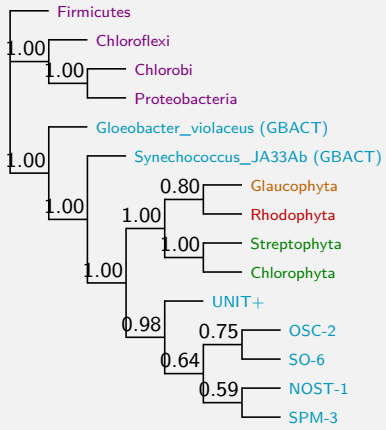
- Serine signal:
  - contributes to SO-6 and plastid monophyly
  - decreases the support of "core" cyanobacteria paraphyly
- Serine signal is not present in translated datasets

Effects of serine synonymous substitutions

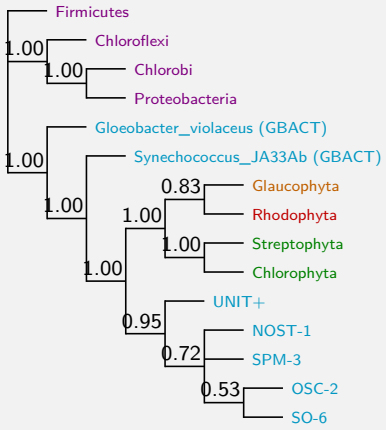
- Serine signal:
  - contributes to SO-6 and plastid monophyly
  - decreases the support of "core" cyanobacteria paraphyly
- Serine signal is not present in translated datasets
- Maybe we should keep this signal in the nucleotide dataset?

Effects of serine synonymous substitutions

restore serine codons



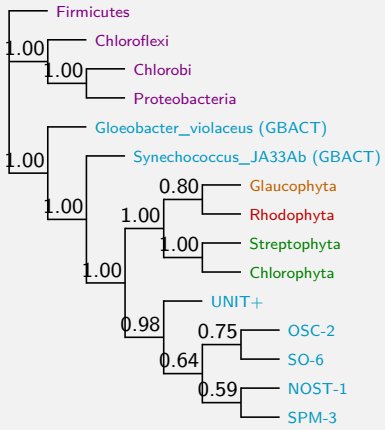
cg75_degenerateLRS3



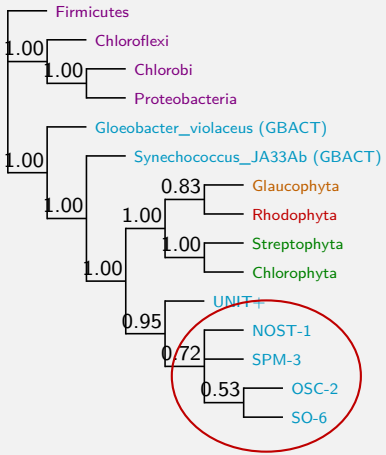
cg75_degenerateLR3

Effects of serine synonymous substitutions

restore serine codons

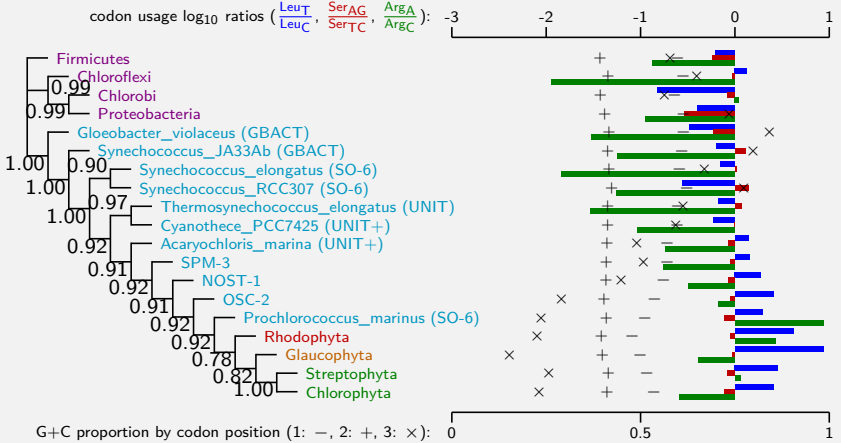


cg75_degenerateLRS3



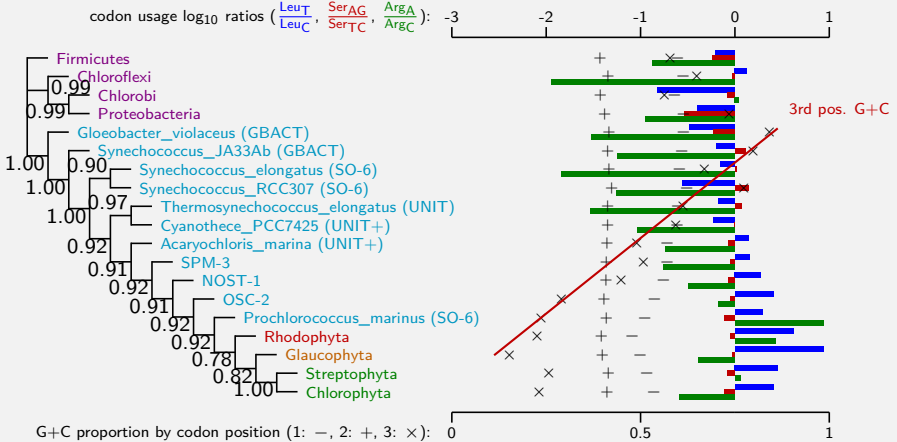
cg75_degenerateLR3

Effects of composition and codon usage



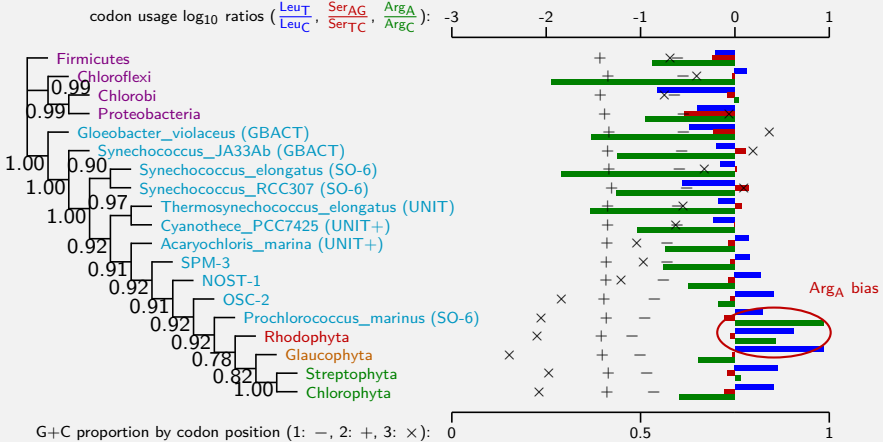
cg75_degenerate12S

Effects of composition and codon usage



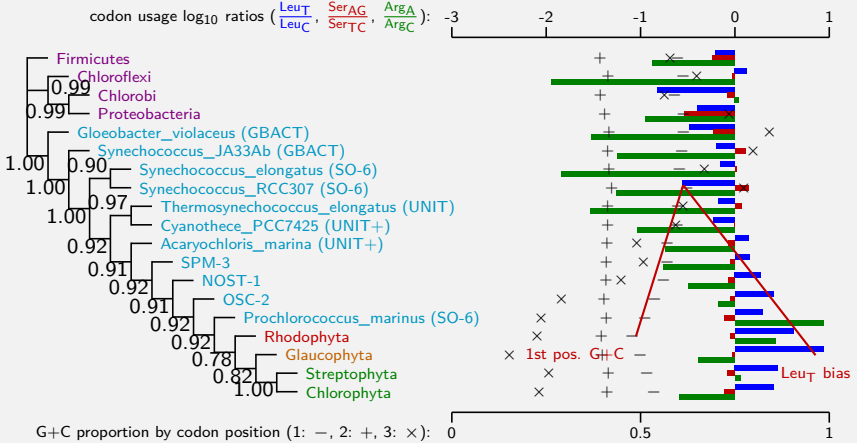
cg75_degenerate12S

Effects of composition and codon usage



cg75_degenerate12S

Effects of composition and codon usage



cg75_degenerate12S