VGT, HGT & XGT

Bi seminar 2004

Pavel S. Novichkov,1 Marina V. Omelchenko,2,3 Mikhail S. Gelfand,4,5 Andrei A. Mironov,1Yuri I. Wolf,3 and Eugene V. Koonin3*

"Genome-Wide Molecular Clock and Horizontal Gene Transfer in Bacterial Evolution"

JOURNAL OF BACTERIOLOGY, Oct. 2004, p. 6575–6585

Introduction

- Klassikaline molekulaarse kella konseptsioon on kehtiv, kui järjestus(ed) evolutsioneeru(vad)b ühtlase kirusega; nende funktsioon ei muutu.
- Molekulaarse kella konseptsiooni aluseks on neutraalne evolutsiooniteooria.
- Tegelikuses varieerub molekulaarne kell erinevat fuktsiooni kandvate järjestuste vahel ca 2 suurusjärku.
- Molekulaarse kella hüpotees on siiski kehtiv ortoloogide korral.

Introduction

- Peaaegu-neutraalne evolutsiooni teooria ennustab molekularse kella suuremat dispersiooni. Arvestab ka väikeste deletsioonide fikseerumise ja populatiooni suuruse mõjuga.
- Paljudel juhtudel ei jaotu evolutsiooni kiirused ortoloogide vahel H_0 kohaselt - Poisson jaotuse alusel.
- Põhjused võivad olla
  - Liini spetsiifiline evolutsiooni kiirenemine
  - Funktsiooni muutused
  - Valiku surve lõdvenemine
  - Mutatsiooniline positiivne surve

Introduction HGT

- Bakteritel on täheldatud küllaltki sagedast horisontaalset geenide ülekannet (HGT)
- HGT on eelistatud just “kaugete” liikide vahel.
- HGT sündmusi võiks klassifitseerida:
  - Fülogeneetilisele liinile uue geeni omandamine
  - Paraloogi omandamine
  - Xenoloogne geeni asendamine (XGT)

Introduction HGT & XGT

- HGT sündmusi on võimalik kindlaks teha phylogeneetiliste puude võrdlemisel.
- Võrreldakse organismi puud geeni puuga ja leitakse kõrvalekalded, kas harude asetuses või siis ebaloomulikud pikkused.

Analüüsitud materjal

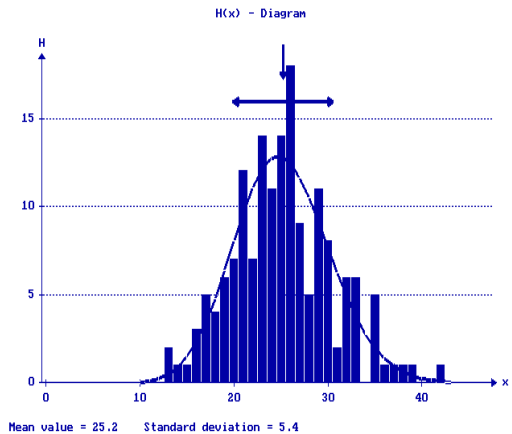
- μ-*Proteobacteria* *6 liiki*, *COG 563 valku*
  - *Escherichia coli* K-12, *Haemophilus influenzae*, *Pasteurella multocida*, *Salmonella enterica* serovar Typhimurium LT2, *Vibrio cholerae*, and *Yersinia pestis*.
- α-*Proteobacteria* *7 liiki*, *COG 274 valku*
  - *Agrobacterium tumefaciens* C58 Cereon, *Brucella melitensis*, *Caulobacter crescentus* CB15, *Mesorhizobium loti*, *Rickettsia conorii*, *Rickettsia prowazekii*, and *Sinorhizobium meliloti*.
- *Bacillus-Clostridium* *8 liiki*, *COG 234 valku*
  - *Bacillus halodurans*, *Bacillus subtilis*, *Clostridium acetobutylicum*, *Listeria innocua*, *Lactococcus lactis*, *Staphylococcus aureus* N315, *Streptococcus pneumoniae* TIGR4, and *Streptococcus pyogenes* M1 GAS.
- 21 liigi peale kokku 114 COG

Analüüsi vahendid

- Iga COG perekonna jaoks arvutati ML kaugused.
  - Programm PAML (JTT maatriks).
  - Asendusmudelit korrigeeriti vastavalt jälgitud aminohapete sagedusele ja Gamma-jaotuse parameetriga α =1,0.
- Fülogeneetiliste puude arvutamiseks kasutati PUZZLE paketti ja ML meetodid

Analüüsi vahendid

- Genoomide vaheliste kauguste arvutamiseks kasutati 114 COG'i perekonna kauguste jaotusest saadud mediaani väärtusi.
- Puude arvutamiseks kasutati PHYLIP paketi programme
  - Neghbor-joining NEIGHBOUR
  - Vähim-ruutude FITCH
- *Limited tree analysis* – meetod, mis jagab puu kahte, erinevat evolutsioonilist mudelit peegeldavasse ossa



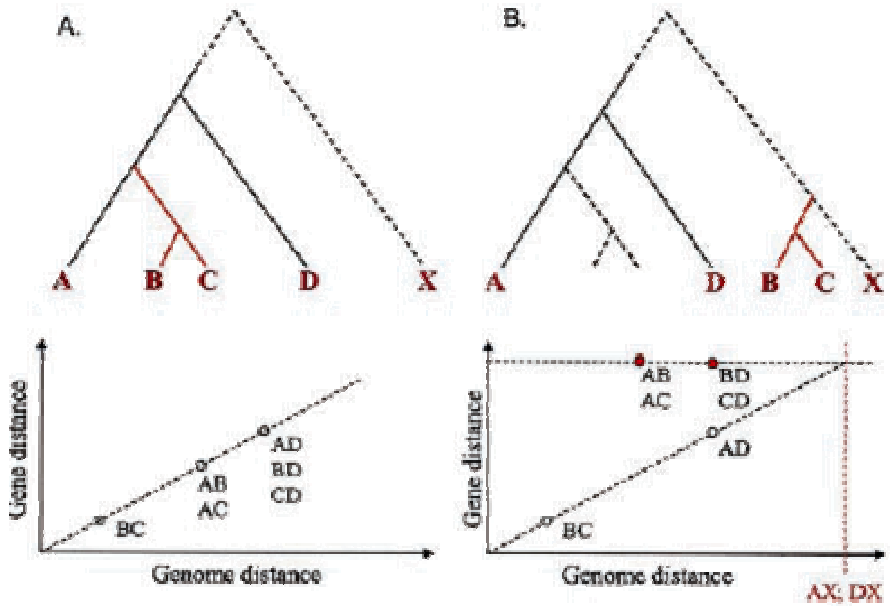
Teooria VGT ja/või HGT

A. *Clock-like* evolutsiooni mudel
B. Kui osad geenid on horisontaalselt üle tulnud.

Ideaalsel juhul

$d_{AB} = vD_{AB}$

v on vastava COG grupi erinev evolutsiooni kiirus võrreldes genoomide evolutsiooni kiirusega

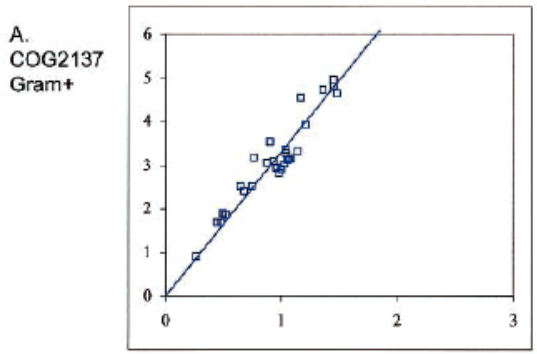


Vertikaal teljel on genoomide vaheline kaugus D_{AB} ja horisontaal teljel geenide vaheline kaugus d_{AB}

Seda ideaalpilti (A.) lõhuvad teatud evolutsioonilised sündmused. Diagonaalilt hälbivate punktide (valkude) kauguse mõõtmiseks kasutatakse distantsi D*

Vertikaalsele edasikandumise mudel

Olgu meil N genoomi {G}, millest igaüks on esindatud ühe COG'i liikmega. Siis nende paaride vahel distantsi mõõt COG(d_{ij}) ja kaugus genoomide vahel on D_{ij} üle kõikide paaride ([I,J]) genoomidest {G} ja paridest N'=N(N-1)/2. Minimiseerides standard viga (*square error*) üle kõikide paaaride,



saame optimaalse v väärtuse,

$$E^2 = \sum_{[I,J]} e_{IJ}^2 = \sum_{[I,J]} \left(\frac{d_{IJ} - vD_{IJ}}{\sqrt{vD_{IJ}}} \right)^2$$

$$v = \sqrt{\sum_{[I,J]} \frac{d_{IJ}^2}{D_{IJ}} / \sum_{[I,J]} D_{IJ}}$$

mudeli sobivus vea (s^2) ja eij varieeruvuse (u^2)

$$s^2 = \frac{1}{(N'-1)} \sum_{[I,J]} \left(\frac{d_{IJ}}{\sqrt{vD_{IJ}}} - \sqrt{vD_{IJ}} \right)^2 \qquad (4b)$$

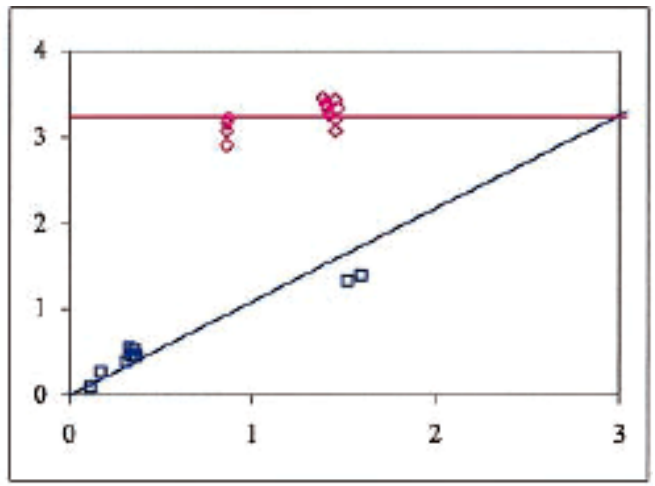
$$u^2 = \frac{1}{(N'-1)} \sum_{[I,J]} (d_{IJ} - vD_{IJ})^2$$

Statistilised mudelid

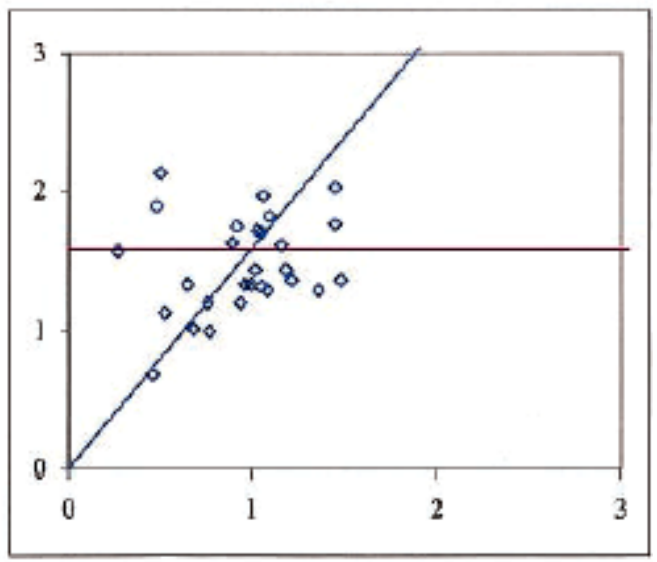
B.Kui COG rühmas on HGT, siis jagunevad distantsid kahte rühma
- Vertikaalset pärandumist peegaldavad [I,J]
- Horisontaalset ülekannet peegeldavad [K, L]

C. Juhusliku müra mudel
- Kui COG'i kuuluvate geenide kaugused ei koreleeru genoomide kaugustega ega grupeeru kumbagi mudelisse

B.
COG0276
α



C.
COG0801
Gram+



Andmete analüüs

- Kõik COG grupid analüüsiti kolme mudeli põhjal
  a) Noisy data, non clock-like evolution
  b) Lihtne molekulaarse kella mudel
  c) Üksikud kõrvalekalded molekulaarse kella mudelist
     - Kahe grupi leidmiseks tuli puu jagada kahte ossa ja leida mudelisse sobivus eraldi mõlemale osale
     - Jagati kõikvõimalike harude pealt, ja lõpuks jäeti alles grupid, mis vastasid kõige paremini mudelitele.
     - Väljaspoolse päritolu hindamiseks arvutati ka suhteline ülekande kaugus $D_T = (D*/\max(D_{KL})$
     - Kui $D_T > 1$ siis on tõenäoliselt tegu HTG'a

Statistilised testid

- Test 1
  - H0 the data do not follow either of the two clock-like models
  - H1 the data fit either the simple-clock or single-transfer model
- Test 2
  - H0 the data fit the simple-clock model
  - H1 (the data fit the single-transfer model

$$F_C = u^2{}_N / \min(u^2{}_C, u^2{}_T)$$

If the value of F_C exceeded the critical level at the 0.05 level of significance H0 was rejected. (1.94 to 2.64 sõltuvalt analüüsitavast grupist)

$$F_T = s_C^2 / s_T^2$$

If the value of *F*T exceeded the critical level at the 0.05 level of significance H0 was rejected (1.95 to 2.66 sõltuvalt analüüsitavast grupist)

Tulemused

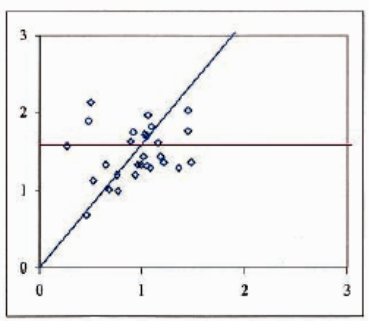
- Analüüsiti kolme suremat bakterite rühma
  μ-*Proteobacteria*
  α-*Proteobacteria*
  madala G+C sisaldusega Bacillus-Clostridium
- Igast rühmast võeti analüüsiks COG'd mida oli antud rühma liikides ainult üks koopia per genoom. Vastavalt siis:
  - μ- *6 liiki*, *COG 563*
  - α-*7 liiki*, *COG 274*
  - *Low G+C 8 liiki*, *COG 234*
  - 21 liigi peale kokku **COG 114**

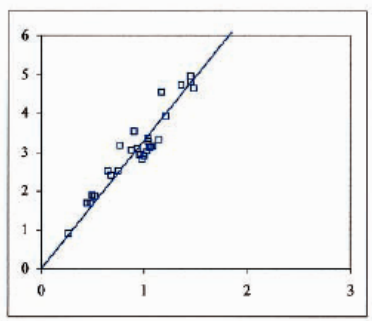
COG'e jaotus kolme mudeli vahel

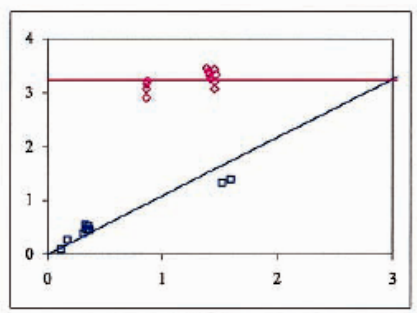
TABLE 1. Alternative evolutionary models for COGs

Lineage	No. (%) of COGs[a]				Total no.
	Fit to the models				
	Random scatter	Clock-like, no anomalies	Split into two subsets		
			$D_T \leq 1$	$D_T > 1$	
γ-*Proteobacteria*	8 (1.4)	387 (68.7)	42 (7.5)	126 (22.4)	563
α-*Proteobacteria*	4 (1.5)	193 (70.4)	37 (13.5)	40 (14.5)	274
Gram positive	21 (9.0)	173 (73.9)	9 (3.8)	31 (13.2)	234

[a] Number of COGs (percentage of the total number of COGs analyzed) for the given lineage.







Suhtelised evolutsiooni kiirused

TABLE 2. Relative evolution rates

Group	Relative rate			Ratio of fastest to slowest		
	Minimum	Median	Maximum	All	Conserved[a]	90%[b]
γ-*Proteobacteria*	0.10	1.40	9.32	97.6	35.9	12.0
α-*Proteobacteria*	0.21	1.08	4.17	19.4	18.4	5.6
Gram positive	0.19	0.97	4.00	13.0	9.4	6.9

[a] Only the proteins from the 108 conserved COGs that fit one of the two evolutionary models were included.

[b] Ratio after removal of 5% of the fastest-evolving proteins and 5% of the slowest-evolving proteins.

Evolutsioonikiiruste jaotus

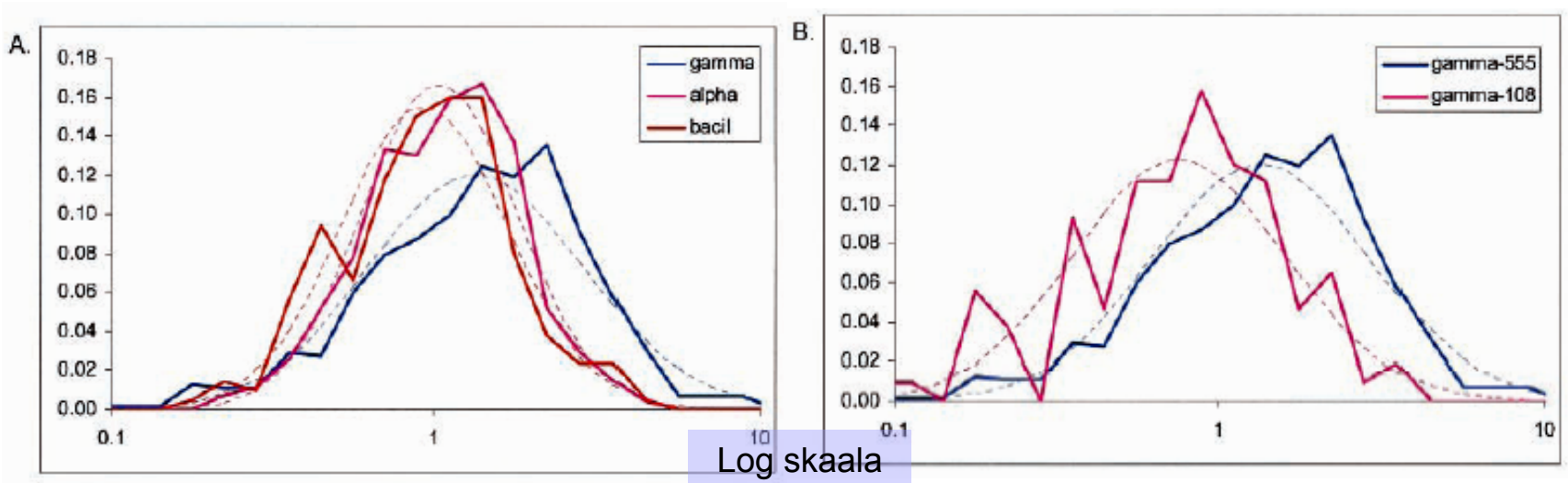


FIG. 3. Distributions of the relative evolutionary rates in the three bacterial lineages analyzed and in the conserved set of COGs represented in all lineages.
(A) Distributions of the relative evolutionary rates in the three lineages. gamma, -*Proteobacteria*; alpha, -*Proteobacteria*; bacil, *Bacillus-Clostridium* group.
(B) Distributions of therelative evolutionary rates in the full set of -*Proteobacteria* (**gamma- 555**) and in the subset of COGs represented in all lineages (**gamma- 108**). The distributions for the three bacterial lineages and the corresponding log-normal approximations (dashed lines) are color coded. The scale for the horizontal axis is logarithmic.

Molekulaarse kella anomaaliad

TABLE 4. COGs with most pronounced deviations from the clock-like model

COG	Function (gene)	Proposed split	F_C[a]	F_T[b]	D_T	v	Tree-Puzzle ELW[c]	Bootstrap support (%)[d]	Conclusions
COG2171	Tetrahydrodipicolinate *N*-succinyltransferase (*dapD*)	(*V. cholerae*), (γ-*Proteobacteria*)	1,050.2	263.1	29.72	0.287	1.0000	88	XGD in *Vibrio* and *Pseudomonas*; probable multiple HGT
COG0137	Argininosuccinate synthase (*argG*)	(*V. cholerae*), (γ-*Proteobacteria*)	3,835.1	578.8	11.81	0.505	1.0000	100	Uncertain; extremely long branch for a subset of γ-*Proteobacteria*; the remaining γ-*Proteobacteria* either cluster with α-*Proteobacteria* (*P. aeruginosa*) or are scattered around the tree base (*Xylella fastidiosa*, *V. cholerae*, *Buchnera* sp.); possible combination of acceleration with HGT
COG0221	Inorganic pyrophosphatase (*ppa*)	(*H. influenzae*), (γ-*Proteobacteria*)	317.8	96.3	10.00	0.480	1.0000	97	Uncertain; long branch of *H. influenzae* and *Neisseria* at the base of the *Proteobacteria* are separated from βγ-*Proteobacteria* and attached to the bacterial root; acceleration in one of these and HGT between them?
COG0207	Thymidylate synthase (*thyA*)	(*E. coli*, *S. enterica* serovar Typhimurium, *Y. pestis*), (γ-*Proteobacteria*)	466.1	125.4	7.75	0.602	0.9685	98	XGD in *H. influenzae*, *P. multocida*, and *V. cholerae* from gram-positive, bacteria

----------lower part of the table 4 is removed ----------

Anomaaliad

- COG'd mis andisd *longest split distance* (D_T) analüüsiti eraldi, eesmärgiga kindlaks teha anomaalia põhjus
  - Täielik fülogeneetiline analüüs
  - Statistilisd testid
- Puude analüüsiga püüti iga anomalset juhtu klassifitseerida ühte kolmest võimalusest
  - HGT mille aluseks on XGD
  - Evolutsiooni kiirenemine antud liinis
  - Ebaselege evolutsiooni stsenaariumiga (5)

HG,T mis annab geeni duplikatsiooni ei tule agvesse, sest nad valisid valgud mida on genoomis vaid 1 koopia.

Tulemused

- Enamus HGT juhte oli seotud metaboolsete ensüümide geenidega.
- 5 top 30’st juhust olid aa-tRNA süntetaasid, mida üldiselt usutakse olevat “kaitstud” selliste sündmuste eest.
  - COG0060 Isoleucyl-tRNA süntetaas IleS (*Rikketsia ja C.acetobutylicum*)
  - COG1217 TypA (*C.acetobutylicum, Bacillus-Clostridium)* grupp

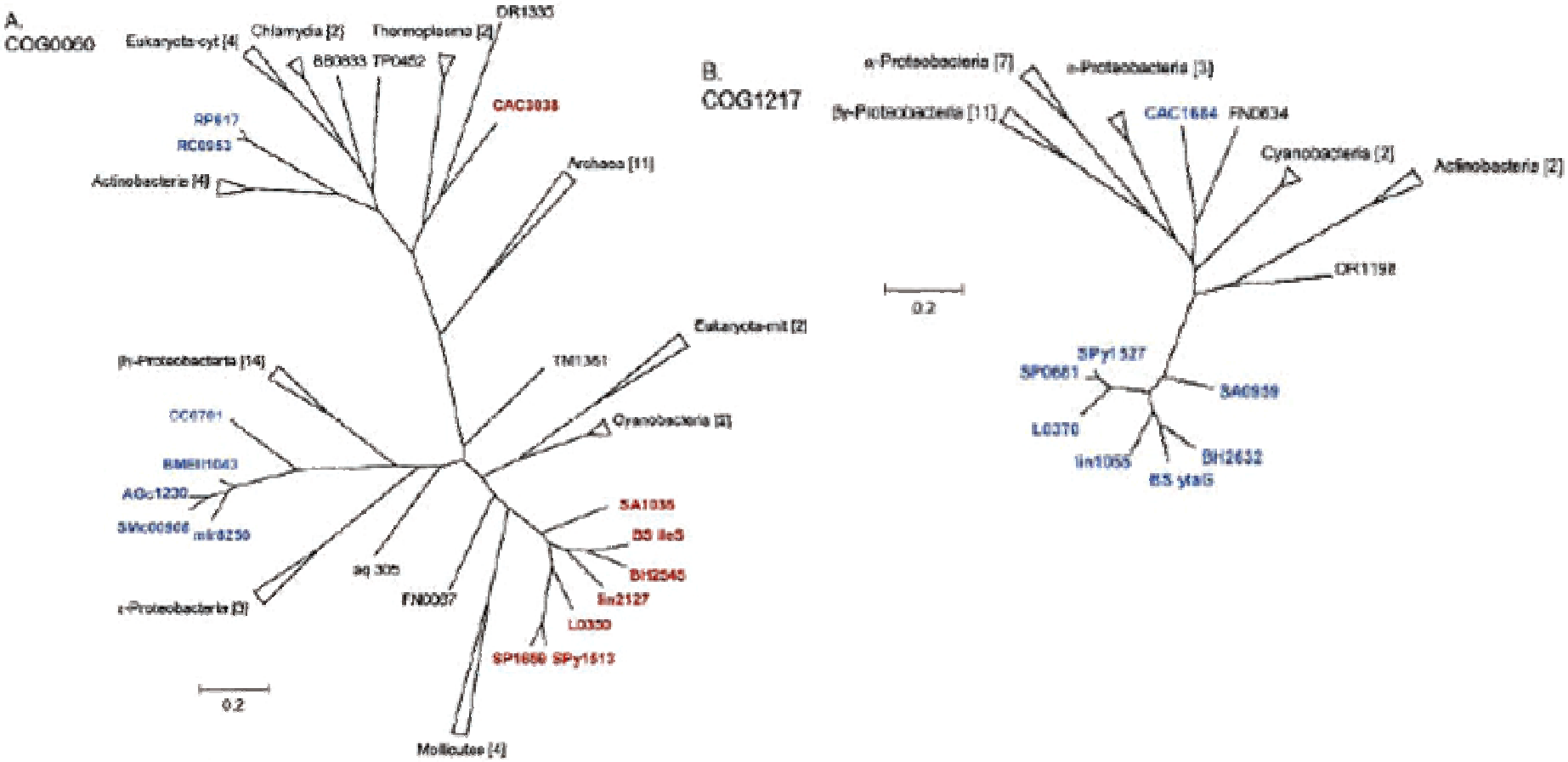


FIG. 5. Maximum-likelihood trees for detected cases of probable XGD. Colors indicate species belonging to the bacterial lineage(s) for which a significant deviation from the clock-like model was detected, namely, -*Proteobacteria* and gram-positive bacteria (A), gram-positive bacteria (B), -*Proteobacteria* (C), and -*Proteobacteria* (D). The triangles represent collapsed clades (the numbers of proteins are indicated in brackets). (A) COG0060 (isoleucyl-tRNA synthetase, IleS). The following proteins and species are included: RP617 of *R. prowazekii*, RC0953 of *R. conorii*,