Inference and Analysis of the Relative Stability of Bacterial Chromosomes

Eduardo P. C. Rocha

Unité Génétique des Génomes Bactériens, Institut Pasteur, Paris, France and Atelier de BioInformatique, Université Pierre et Marie Curie (Paris VI), Paris, France

Mol. Biol. Evol. 23(3):513–522. 2006

April 10th 2006

Age

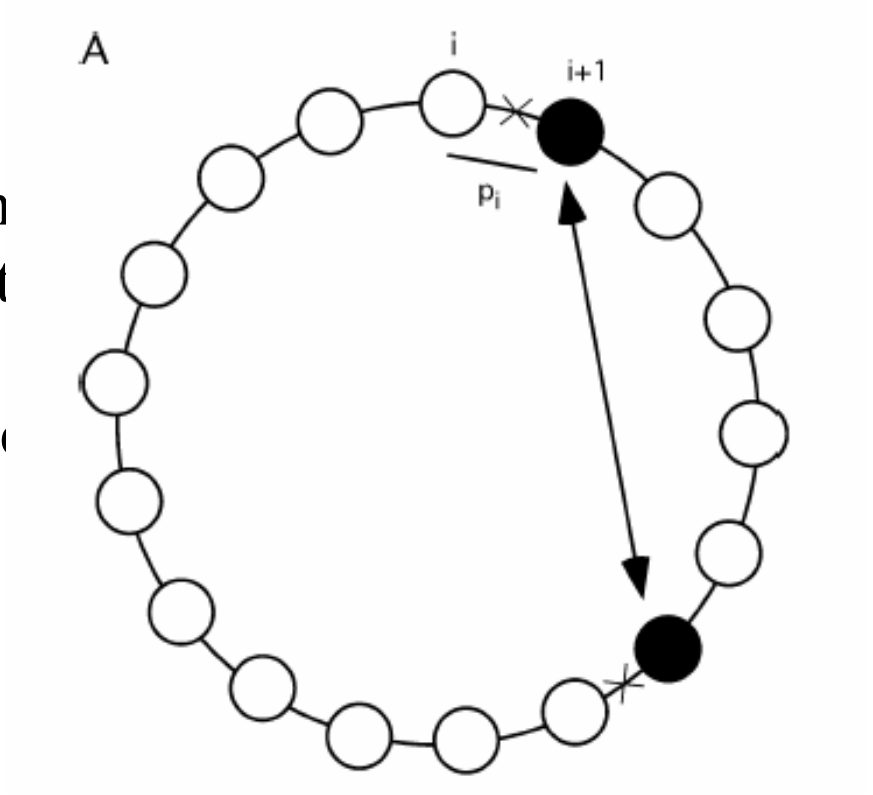
The stability of genomes results form a mutation-selection balance.

Gene expression shapes the chromosome organization by selecting the clustering of functionally related genes into operons and supraoperons. This allows the optimization of gene expression regulation.

-the preferential positioning of essential genes in the leading strand;

-highly expressed genes near the origin of replication in fast-growing bacteria

- Homologous recombination can target inverted repeated element (rDNA operons, insertion sequences (IS)) and lead to large inversions.
- IS and phages may additionally lead to chromosome rearrangements because of the activity of their recombinases.

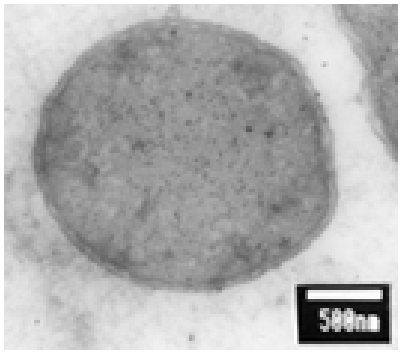


The frequency of rearrangement events and subsequent purging of deleterious ones by natural selection will determine the conservation of long term gene order in bacterial genomes.

Buchnera – the obligatory intracellular mutualist.

Among the most stable genomes, no large rearrangements apparent over 100 Myr (just some gene deletions by illegitimate recombination between small repeats).

Small genome (<0.7 Mb) with few repeated elements, no IS, few recombination genes.



Buchnera cell

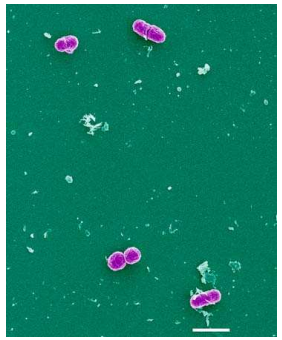
aphids

Chlamydia, Rickettsia – practically no large repeats

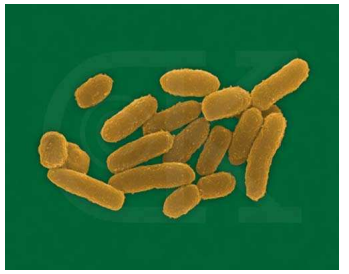
Among the least stable genomes are pathogens *Bordetella pertussis* and *Yersinia pestis*.

Very recently diverged from a more stable genome by losing some genes and by gaining a large number of IS.

A large number of rearrangements.



B.pertussis



Y.pestis

E.coli and *S.enterica* have intermediate genome stability.

Methodological approaches to the study of gene order:

1. Rearrangements distance as an estimation of the number of rearrangement events that took place since speciation
2. Empirical measures of distance based on pairwise comparisons of gene order

To calibrate bacterial genome stability by building an index of genome stability that takes into account the average loss of gene order through time.

To explore the potential association of the stability of genomes with their phylogenetic grouping, genome size and lifestyle.

126 genomes from six clades.

Orthologues were defined as unique reciprocal best hits with at least 40% similarity in amino acid sequence and less than 20% of difference in protein length.

The most distant comparisons in each clade still included a significant number of genes (>1000).

The evolutionary distances between bacteria were computed from the 16S rDNA subunit with Tree-Puzzle using (HKY)+Γ model.

The GOC is defined as the relative frequency with wich two contiguous genes in a genome have their respective orthologues also contiguous in the ordered list of orthologues of the other genome.

$$GOC = \frac{N_{orthologues,\ contiguous}}{N_{orthologues}}.$$

$0 \leq GOC \geq 1$

High values of GOC can be obtained if the genomes are very stable or if they have diverged very recently.

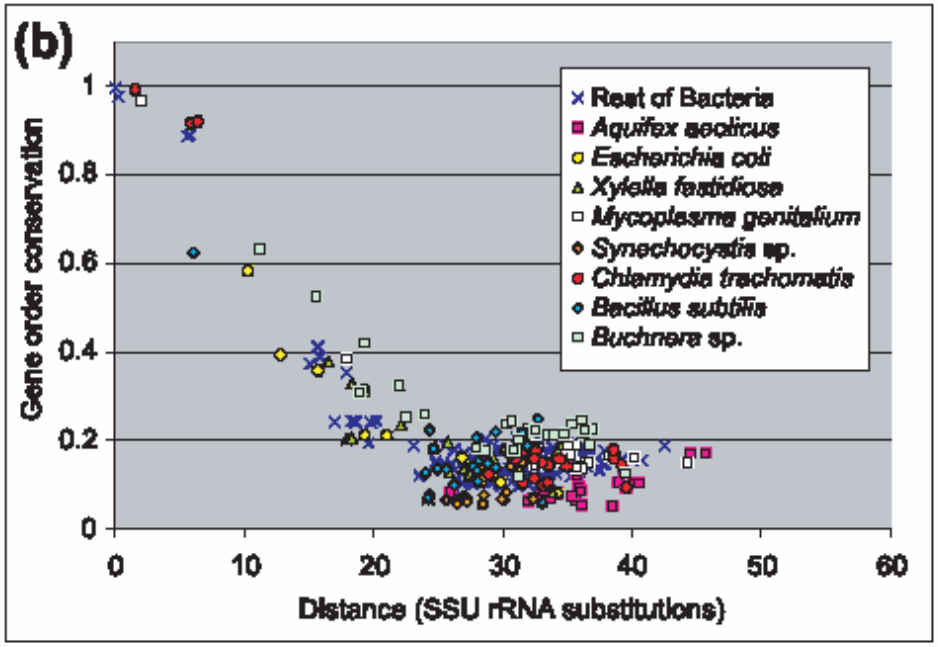
Different models to calculate the loss of GOC with time

$$GOC = \frac{2}{1 + e^{\alpha t}}. \qquad \text{(model 0)}$$

α - a parameter to be adjusted by regression

t - time

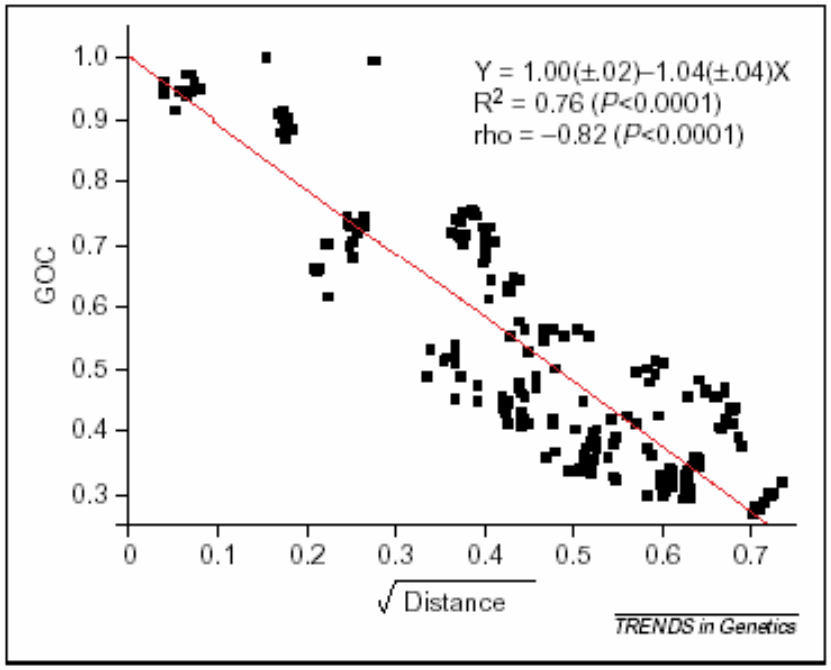
Sigmoid loss of GOC with time



Tamames 2001

Different models to calculate the loss of GOC with time

$$GOC = 1 - \sqrt{\alpha t}. \qquad \text{(model 1)}$$



Rocha 2003

For large enough t GOC will take negative values.

Different models to calculate the loss of GOC with time

$$\frac{1}{\mathrm{GOC}} = \alpha t + 1. \qquad \text{(model 2)}$$

Suyama and Bork 2001

arises from solving the following differential equation:

$$\frac{\mathrm{dGOC}}{\mathrm{d}t} = -\alpha \mathrm{GOC}^2.$$

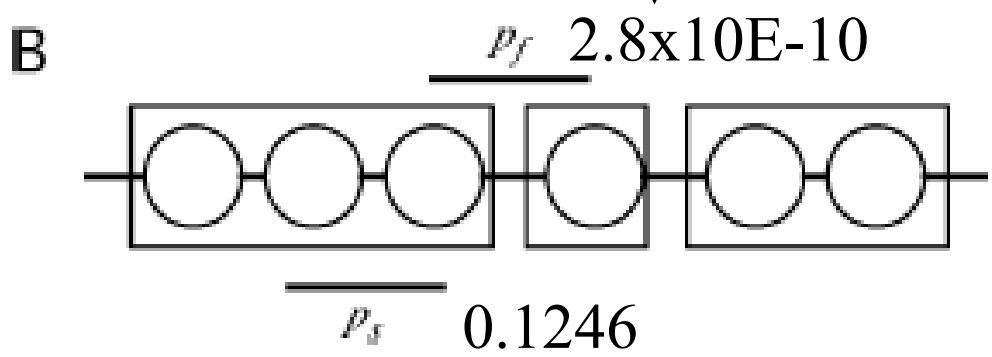
A decrease of GOC with time is negatively proportional to the square of the GOC at the given moment.

Different models to calculate the loss of GOC with time

Probabilistic model

$$GOC = p^t. \qquad \text{(model 3)}$$

probability of NOT being separated



$$GOC = P_f + P_s = \frac{(n_1 p_f^t + n_2 p_s^t)}{N}. \qquad \text{(model 4)}$$

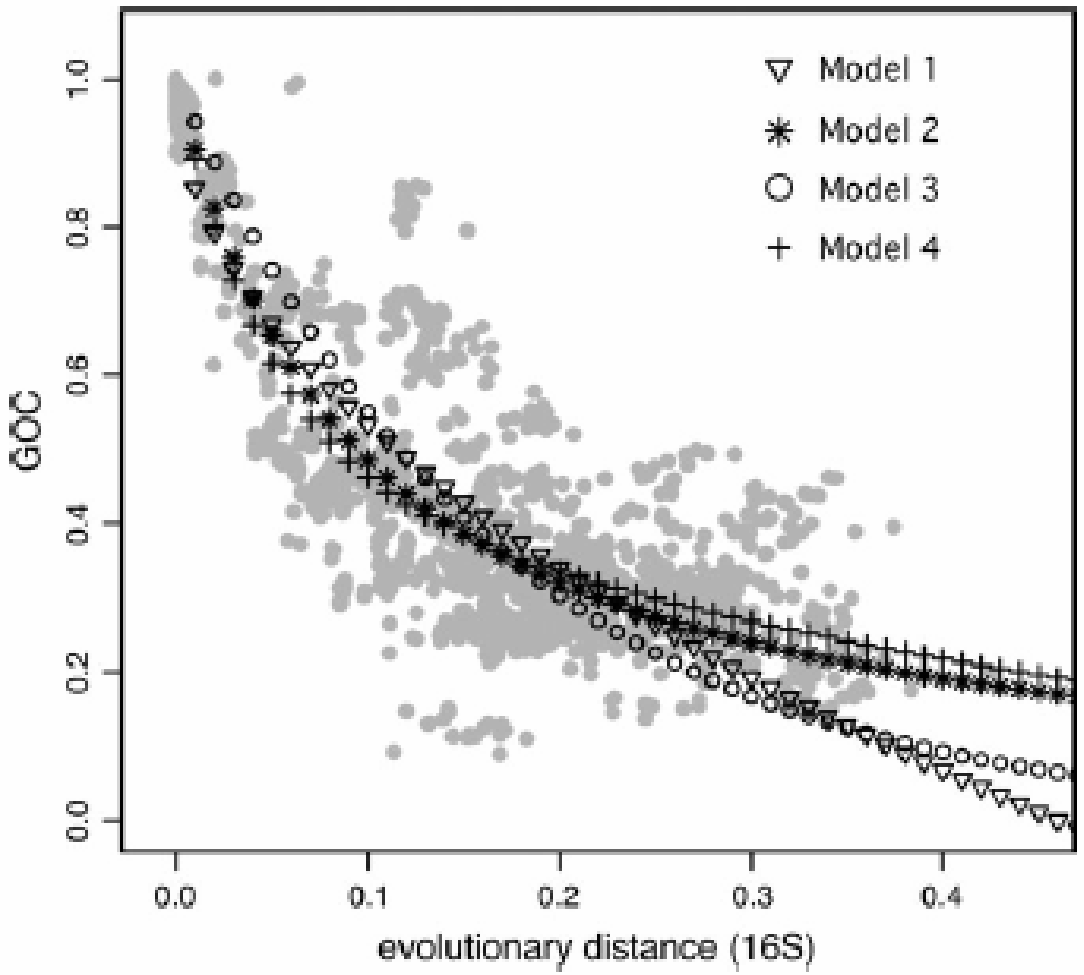
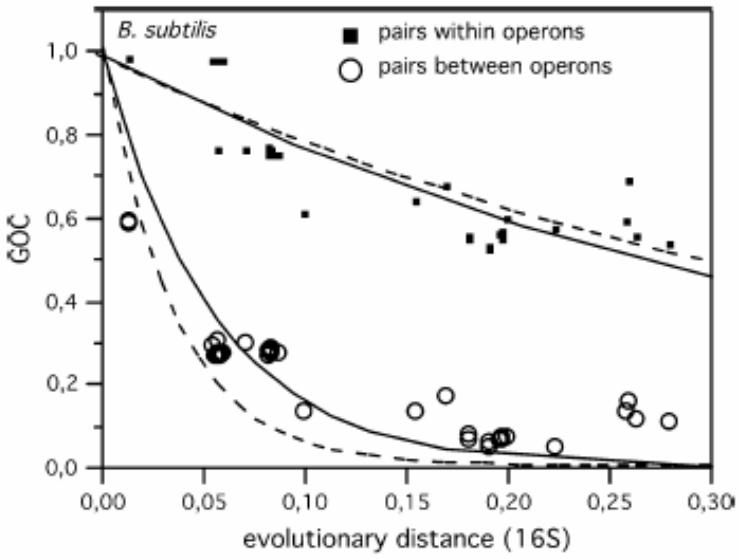
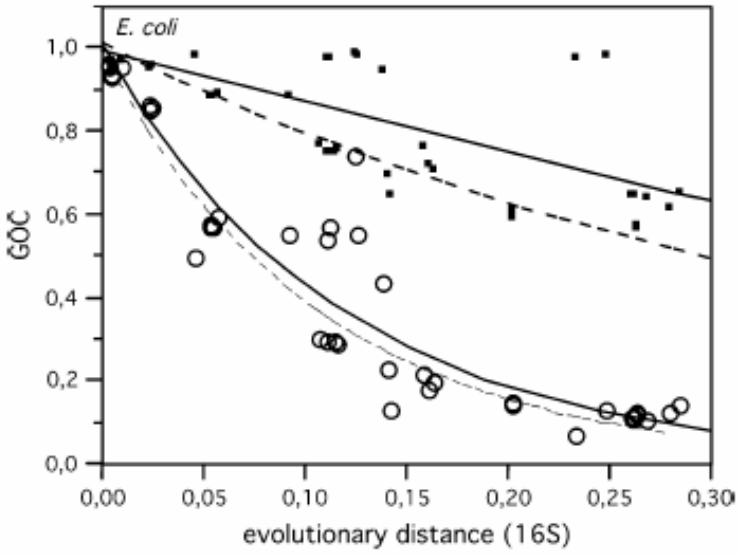


FIG. 2.—GOC between pairs of genomes in function of their evolutionary distance according to the 16S rRNA subunit. Four models are fitted by nonlinear regression. Model 1: $GOC = 1 - \alpha\sqrt{t}$, model 2: $1/GOC = \alpha t + 1$, model 3: $GOC = p'$, and model 4: $GOC = p'_t + p'_s$ where α's and p's are the estimated parameters and t is the evolutionary distance.





Operons were identified in *E.coli* and *B.subtilis* based on gene transcription sense and intergenic distance.

FIG. 3.—Loss of GOC when comparing *Bacillus subtilis* with other Firmicutes and *Escherichia coli* with other γ-proteobacteria. Two different analysis are indicated in each plot, one representing GOC for pairs of genes that are in the same operon in *B. subtilis* or *E. coli* (GOC_{op}, points) other for pairs of contiguous genes that are in contiguous operons (GOC_{betop}, circles). The regression lines were computed within each of the groups using as models $GOC=p'_{op}$ (points) and $GOC_{betop}=p'_{betop}$ (lines) and are represented as full lines. For comparison, are plotted the lines for the two groups when one regresses the bivariate model 4 on the GOC variable (dashed line) (as in fig. 3). The close fit between the two types of regressions suggests that genes separating fast (corresponding to the component p_f in model 4) are mostly associated with contiguous genes in contiguous operons, whereas genes separating slowly (corresponding to the component p_s in model 4) are mostly associated with contiguous genes within the same operon.

The selection acting upon operons is the major force acting against the disruption of gene order and is responsible for the particularly low level of rearrangements for pairs of orthologuos genes within an operon.

Bacteria separated by over 500 Myr still have conserved the contiguity of ~80% of the within-operon gene pairs.

The stability of each genome is the average deviation of the observed values of GOC in the pairwise comparisons from model 4 to the expected values given a certain phylogenetic distance.

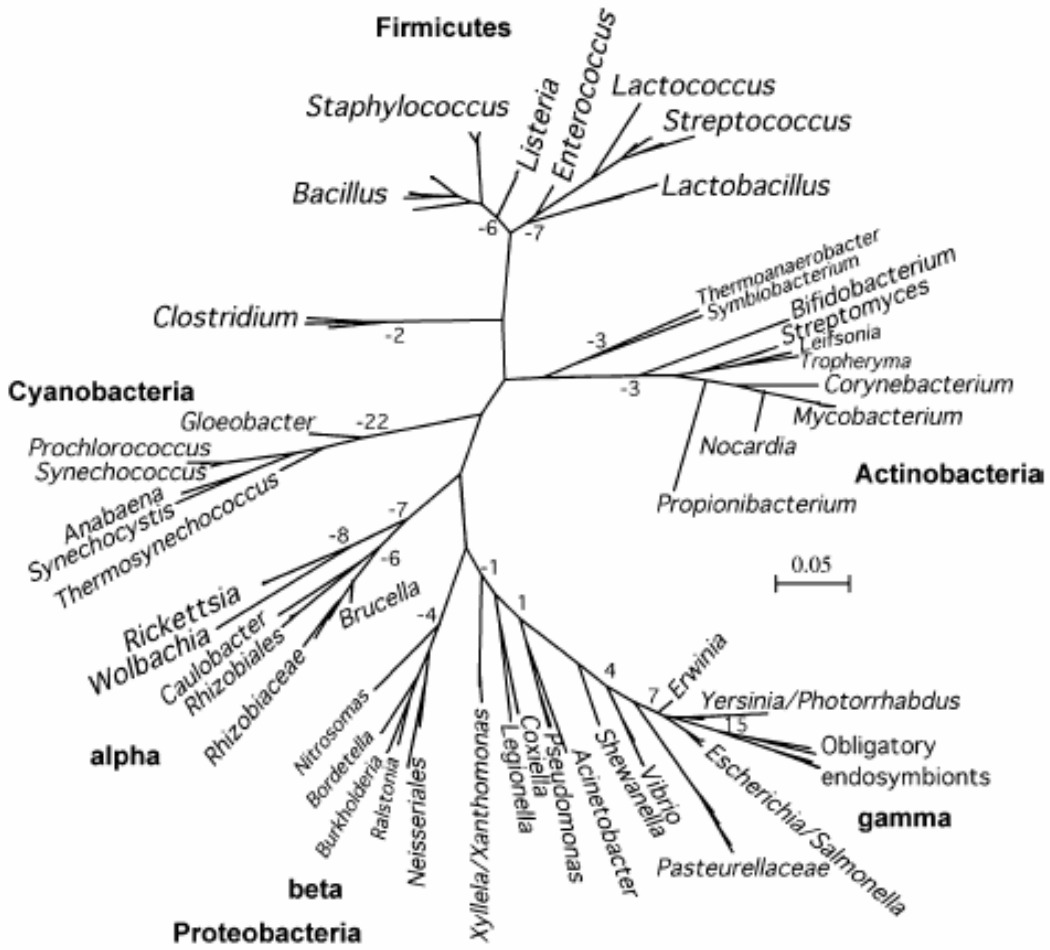


FIG. 4.—Unrooted phylogenetic tree of the species analyzed in this study. The numbers correspond to the relative stability of ancestral states (multiplied by 100), inferred with COMPARE (Martins and Hansen 1997), using the linear model (corresponding to Brownian evolution).

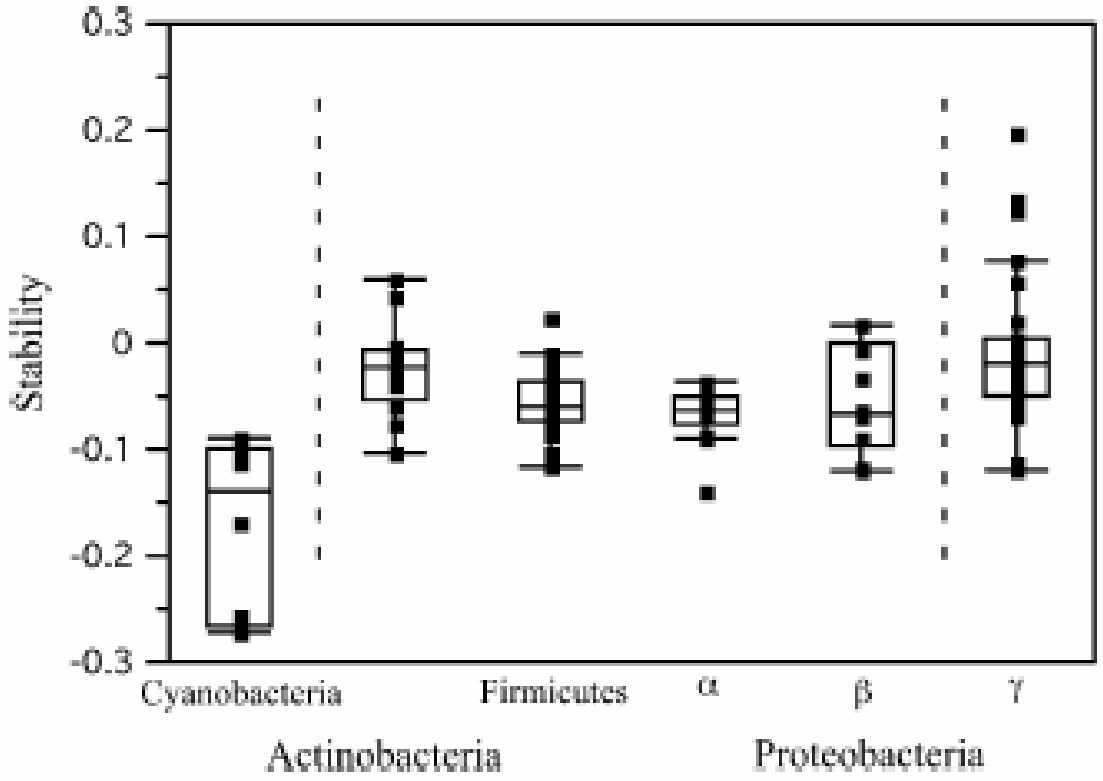


FIG. 5.—The variation of genome stability according to the phylogenetic classification of bacteria. The dashed lines separate the box plots of the groups with significantly different means ($P < 0.05$, Tukey-Kramer HSD test).

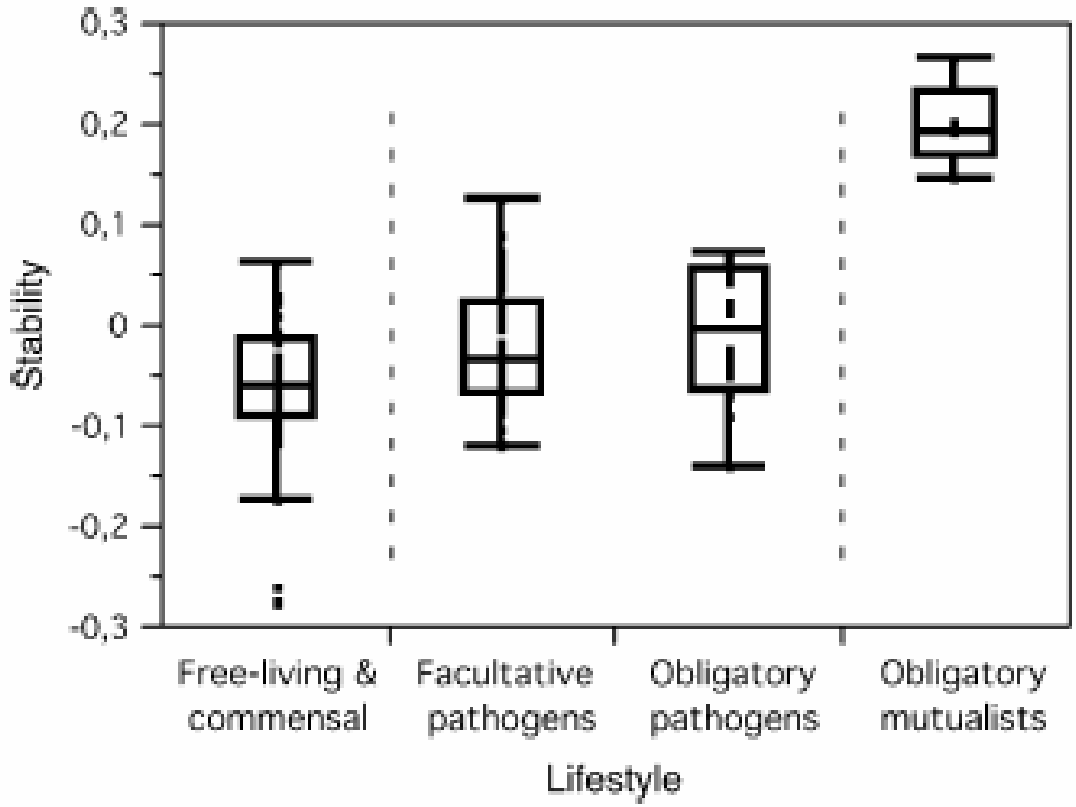


FIG. 6.—Genome stability classed in different groups according to lifestyle. The dashed lines separate the box plots of the groups with significantly different means ($P < 0.05$, Tukey-Kramer HSD test).

Kasutatud kirjandus

- Rocha (2006) Inference and stability of bacterial chromosomes. Mol.Biol.Evol. 23(3):513-522.
- Rocha (2003) DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet. 19:600-604.
- Tamames (2001) Evolution of gene order conservation in prokaryotes. Genome Biol. 2:0020.0021-0020.0011
- Suyama and Bork (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends genet. 17:10-13.