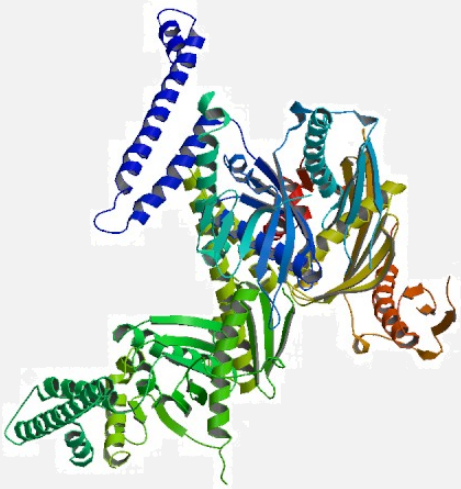
Analysis of translation termination sites in prokaryotes

Przemysław Biecek, Paweł Mackiewicz, Dorota Mackiewicz, Joanna Kiraga, Stanisław Cebrat

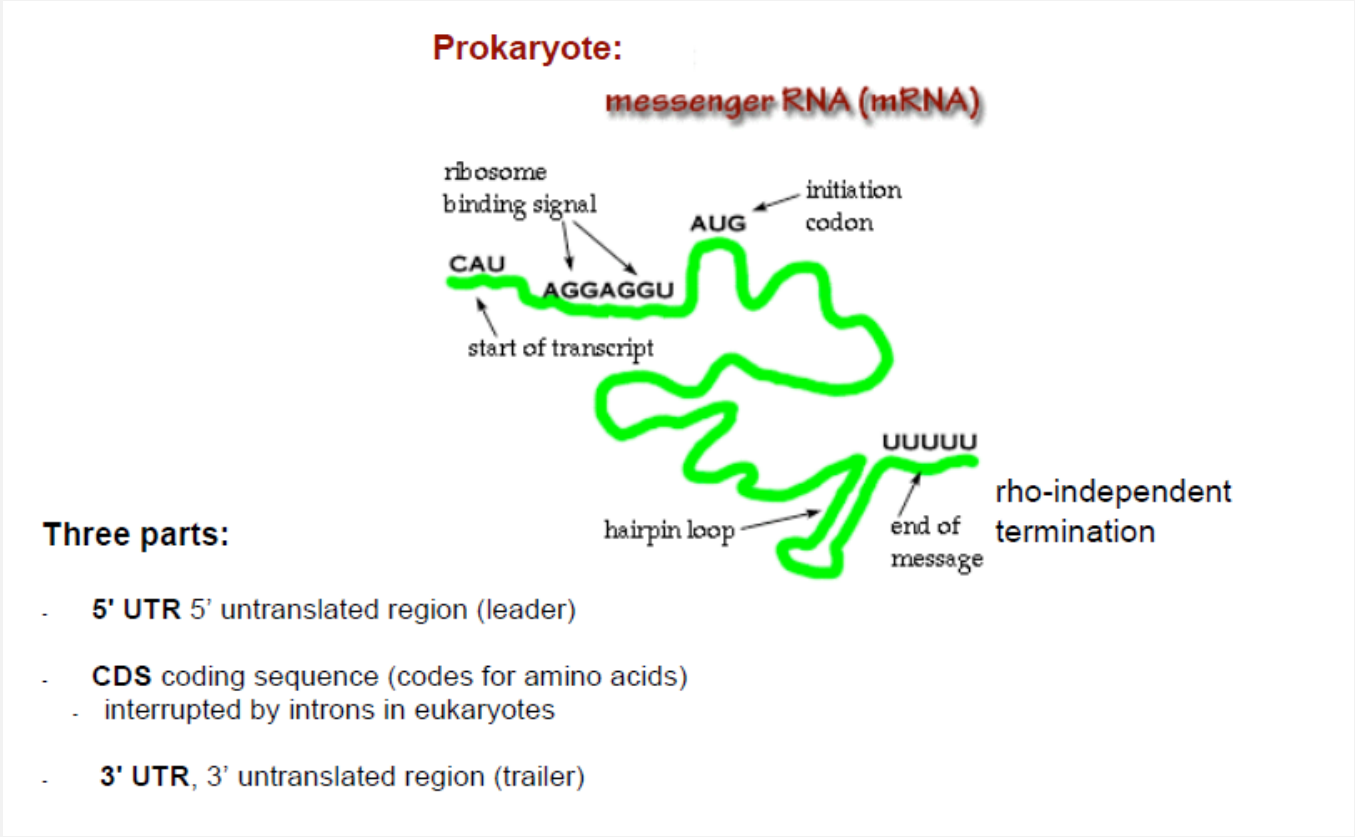
Wrocław/Warsaw University



Outline

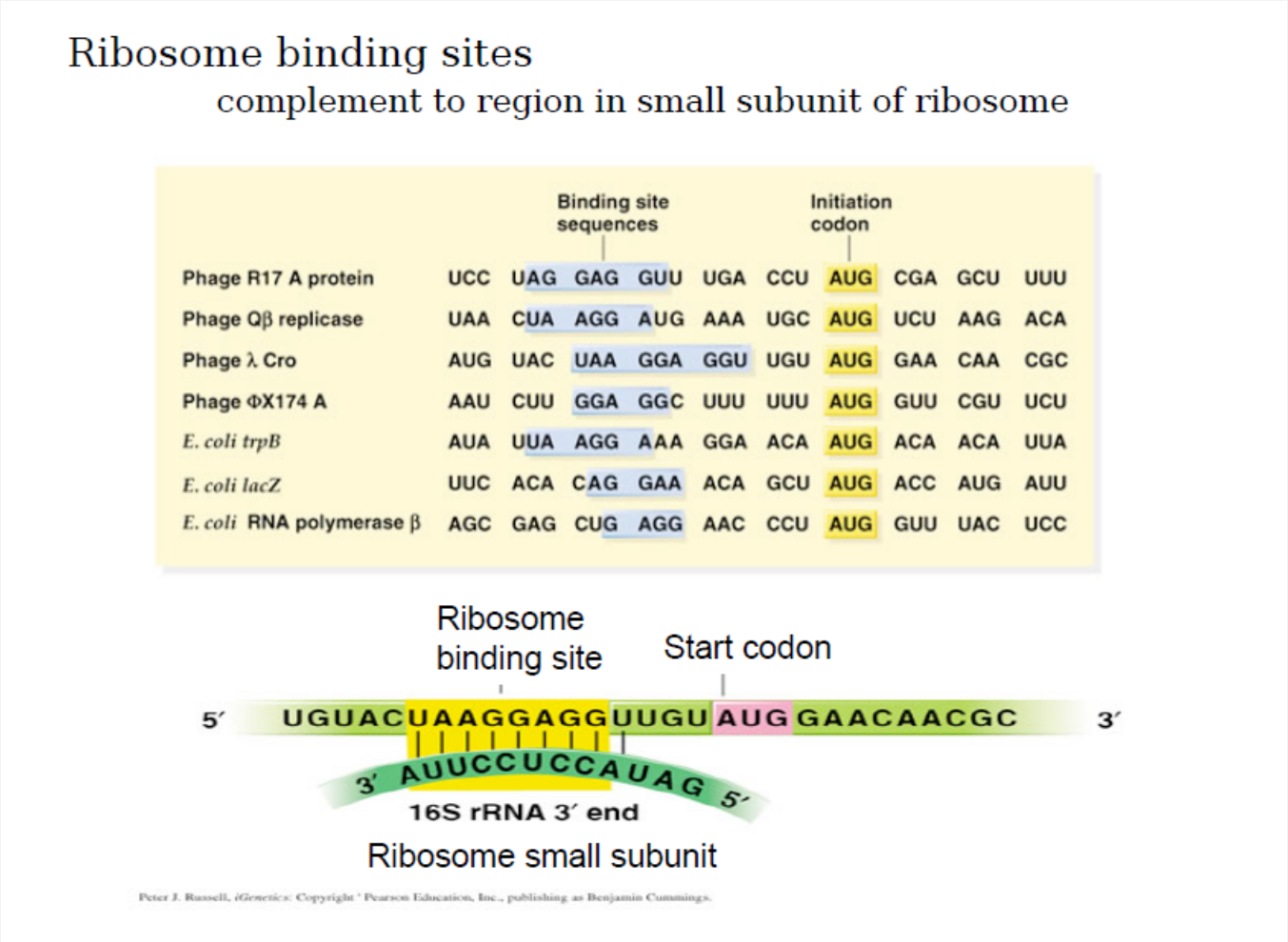
- Basic facts about translation
- Statistical background
- Our data set
- Results for
  - DNA composition, positions from -50 to -1,
  - codon composition, positions from -30 to -1,
  - codon composition before different stop codons,
  - codon composition vs. different GC content.

General dogma: DNA → mRNA → proteins.



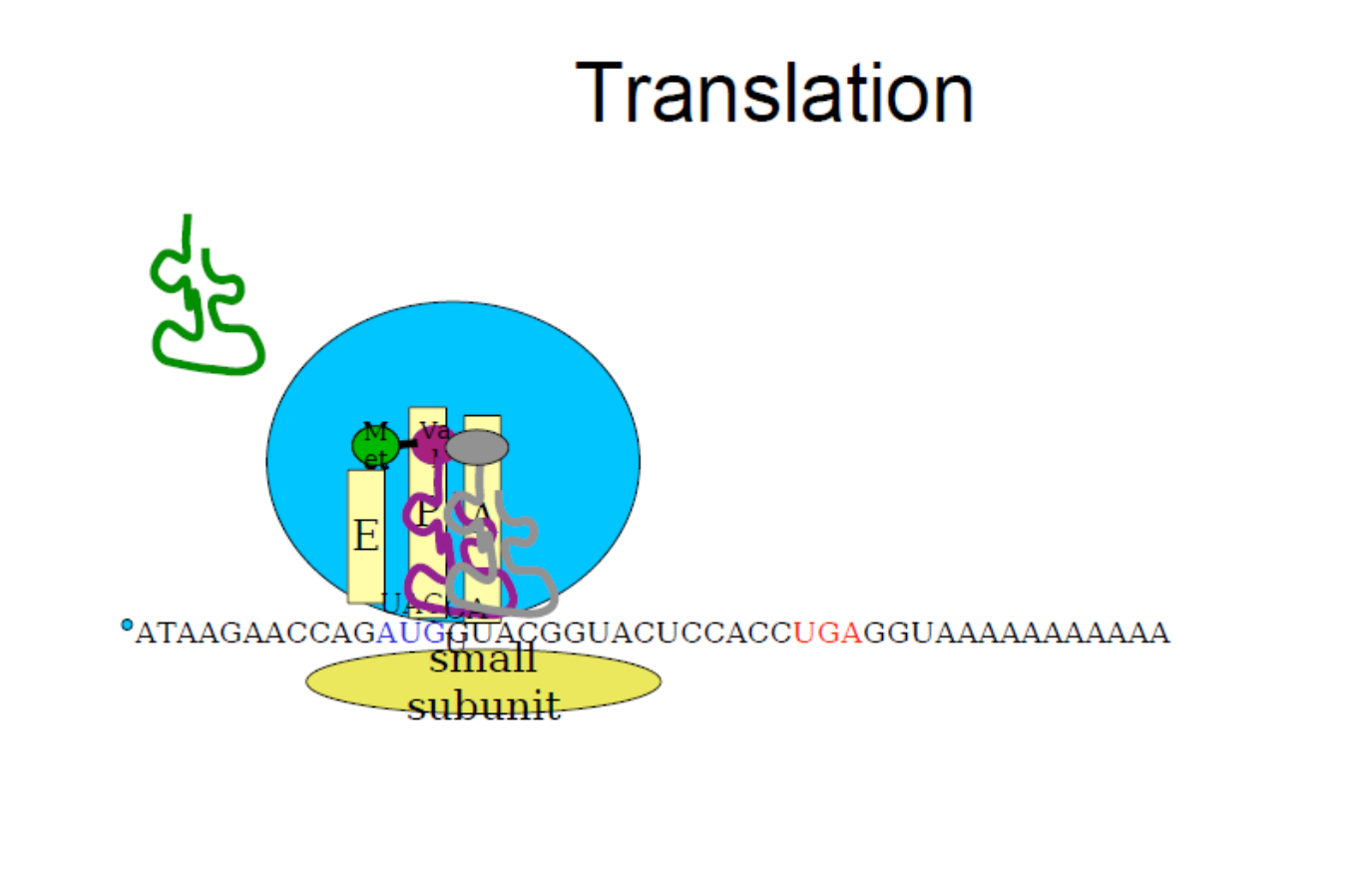
I. Initiation

Ribosom binds to Shine-Dalgarno sequence (AGGAGG).



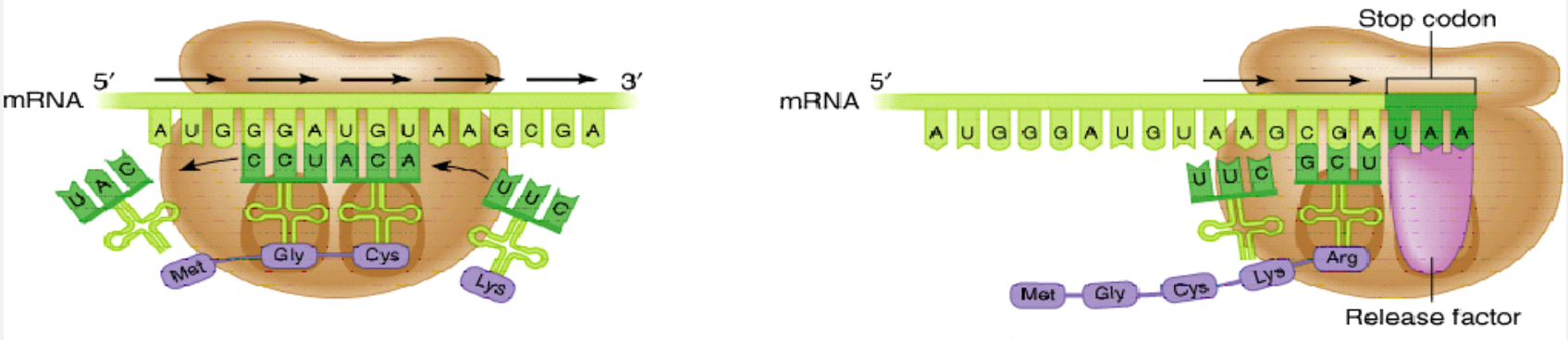
II. Elongation

tRNA with appropriate anticodon go to the ribosome site A.



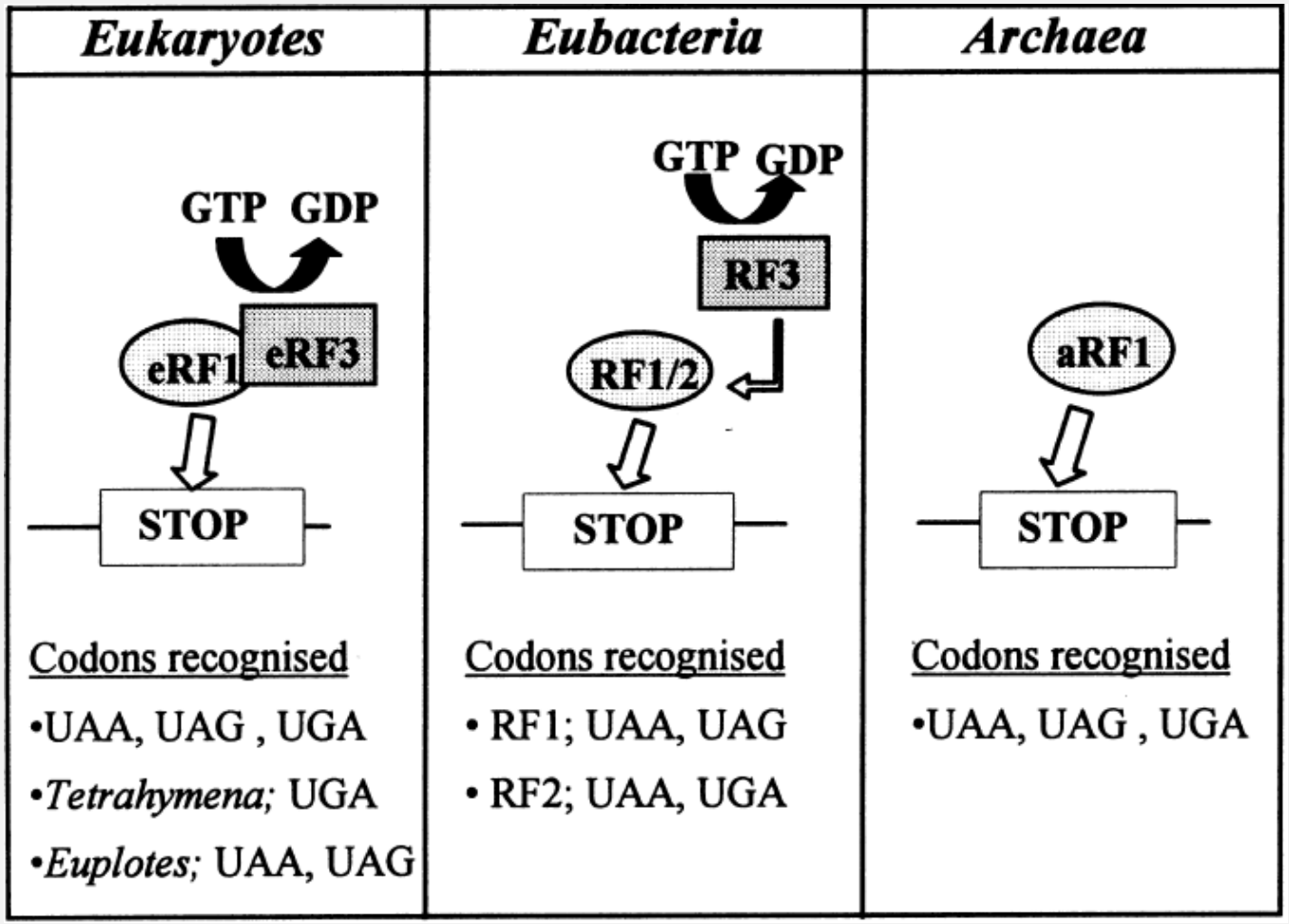
III. Termination

- Some release factor binds to stop codons.
- Peptide chain is released.
- In some cases ribosom start to translate next genes in other is dissociates.



Termination

- Stop codons may be different for different organisms (rare exceptions in small genomes).
- Release factors are different in different kingdoms.

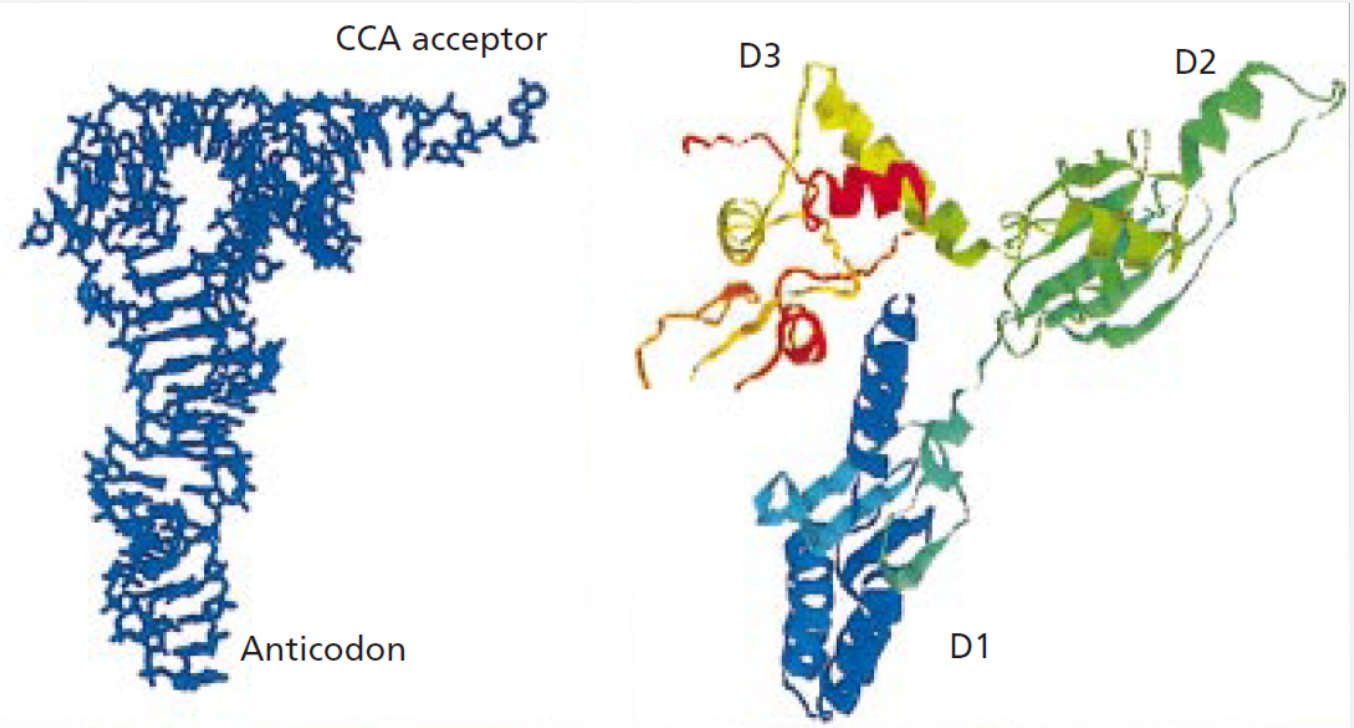


Genetic code

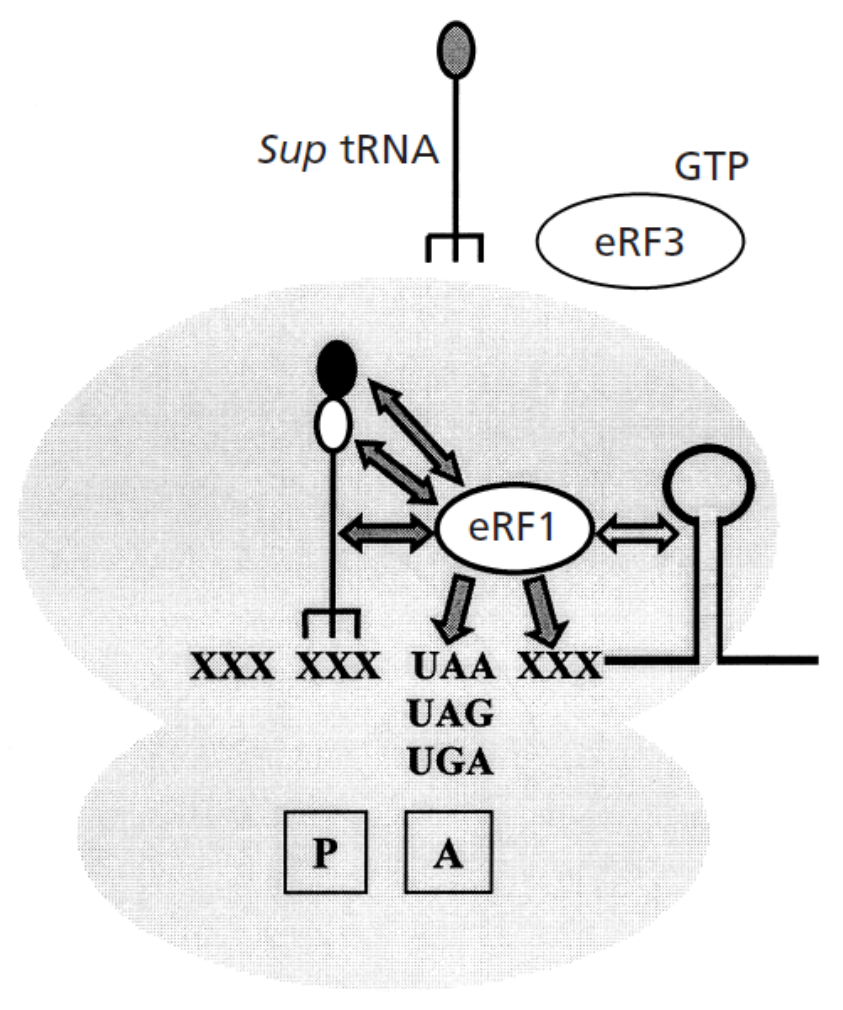
		2nd base			
		U	C	A	G
1st base	U	UUU (Phe/F)Phenylalanine	UCU (Ser/S)Serine	UAU (Tyr/Y)Tyrosine	UGU (Cys/C)Cysteine
		UUC (Phe/F)Phenylalanine	UCC (Ser/S)Serine	UAC (Tyr/Y)Tyrosine	UGC (Cys/C)Cysteine
		UUA (Leu/L)Leucine	UCA (Ser/S)Serine	UAA Ochre (*Stop*)	UGA Opal (*Stop*)
		UUG (Leu/L)Leucine	UCG (Ser/S)Serine	UAG Amber (*Stop*)	UGG (Trp/W)Tryptophan
	C	CUU (Leu/L)Leucine	CCU (Pro/P)Proline	CAU (His/H)Histidine	CGU (Arg/R)Arginine
		CUC (Leu/L)Leucine	CCC (Pro/P)Proline	CAC (His/H)Histidine	CGC (Arg/R)Arginine
		CUA (Leu/L)Leucine	CCA (Pro/P)Proline	CAA (Gln/Q)Glutamine	CGA (Arg/R)Arginine
		CUG (Leu/L)Leucine	CCG (Pro/P)Proline	CAG (Gln/Q)Glutamine	CGG (Arg/R)Arginine
	A	AUU (Ile/I)Isoleucine	ACU (Thr/T)Threonine	AAU (Asn/N)Asparagine	AGU (Ser/S)Serine
		AUC (Ile/I)Isoleucine	ACC (Thr/T)Threonine	AAC (Asn/N)Asparagine	AGC (Ser/S)Serine
		AUA (Ile/I)Isoleucine	ACA (Thr/T)Threonine	AAA (Lys/K)Lysine	AGA (Arg/R)Arginine
		AUG (Met/M)Methionine, *Start*[1]	ACG (Thr/T)Threonine	AAG (Lys/K)Lysine	AGG (Arg/R)Arginine
	G	GUU (Val/V)Valine	GCU (Ala/A)Alanine	GAU (Asp/D)Aspartic acid	GGU (Gly/G)Glycine
		GUC (Val/V)Valine	GCC (Ala/A)Alanine	GAC (Asp/D)Aspartic acid	GGC (Gly/G)Glycine
		GUA (Val/V)Valine	GCA (Ala/A)Alanine	GAA (Glu/E)Glutamic acid	GGA (Gly/G)Glycine
		GUG (Val/V)Valine	GCG (Ala/A)Alanine	GAG (Glu/E)Glutamic acid	GGG (Gly/G)Glycine

Release Factors

Release factors (1 and 2) are structurally similar to tRNA.



Context of stop codons



Context of the termination site

- last few amino acids,
- nucleotides before stop,
- nucleotides after stop,
- mRNA structure (hairpins, regions rich in nucleotide A).

Literature

- There is a lot of papers about stop translation signal. Usually only one organism is analyzed e.g. E. coli or other model organisms.
- RF speciation
  - RF1 recognise codons TAG and TAA,
  - RF2 recognise codons TGA and TAA.
- RF effectives
  - TAA (ochre) — the most frequent, strong signal, fails one per 1000 passes,
  - TGA (opal) — a weak signal of termination and fails one per 4 passes.
- Weak stops are natural source of alternative protein products!
- Tandem Stop Codons, higher than expected number of codon TAA in position +3 after stop.
  TSC are more frequent in highly expressed genes.

Our study

Comparative study for over 400 prokaryotic genomes (all prokaryotic genomes completed before VI 2007, ≈1 000 000 genes).

- Goal 1: To identify signals correlated with stop translation.
- Goal 2: To identify factors which may explain evolution of such signals.

Data set from NCBI

- genes from plastids and from chromosome are analyzed together,
- one candidate is chosen for a set of genomes with identical fourth-level taxonomic classification,
- .fna files with gene sequences and .ptt files with gene coordinates.

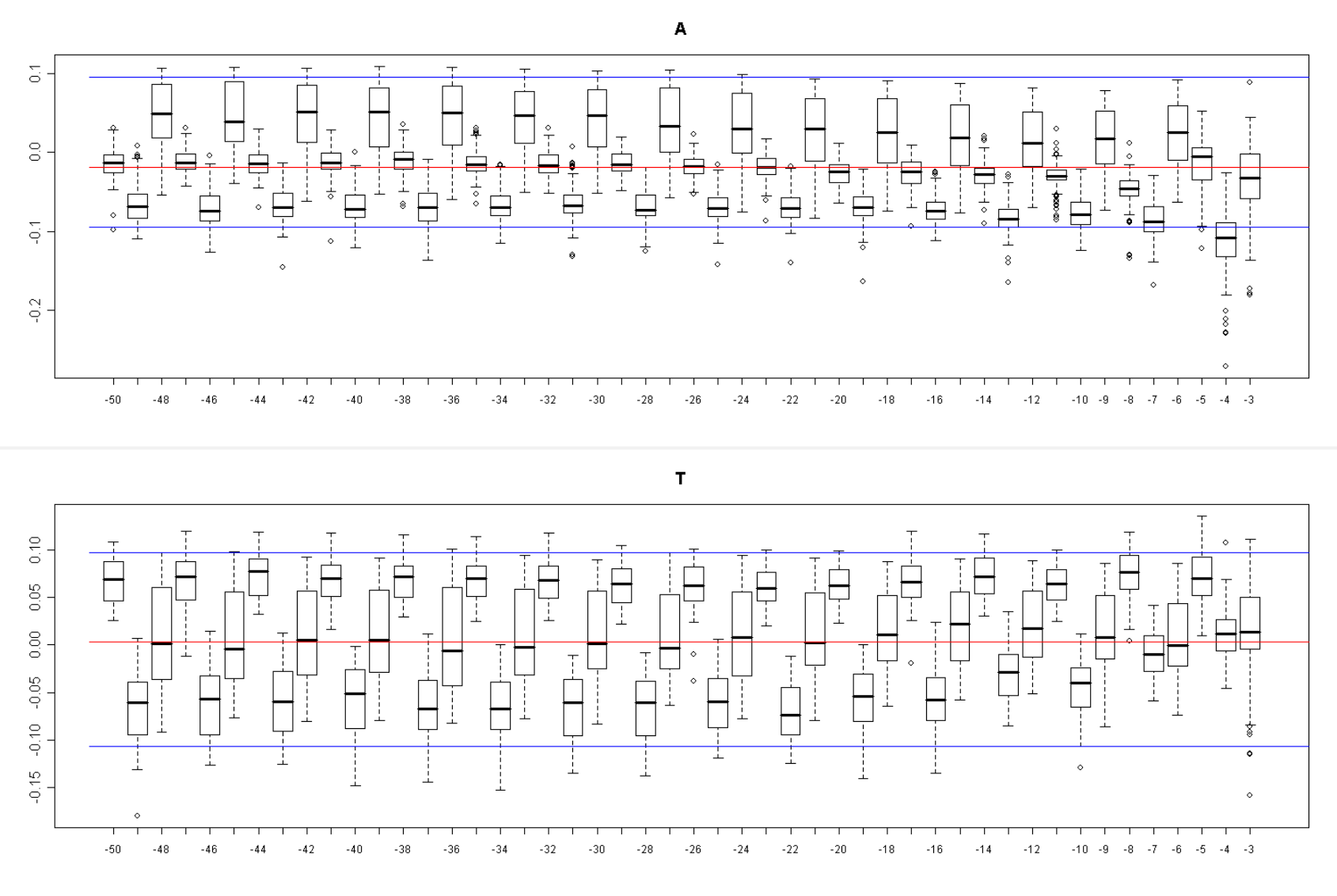
Composition skew

For every genome the composition skew is computed

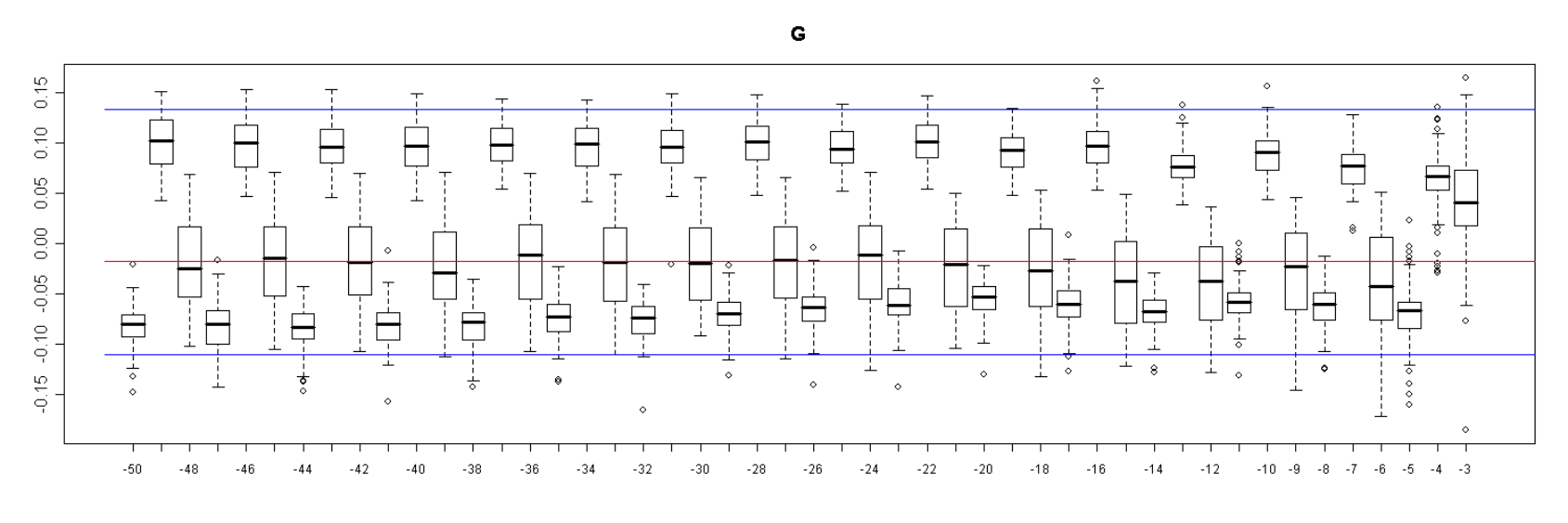
$$skew = \hat{\rho}_{-i,x} - \hat{\rho}_{intra,x},$$

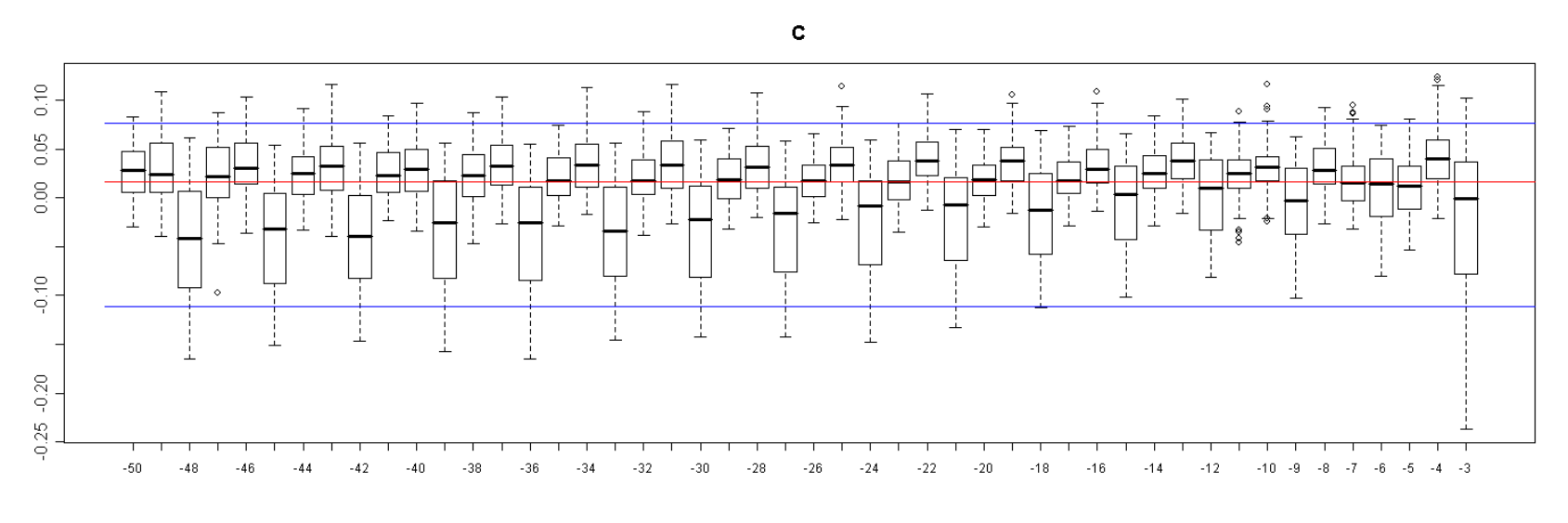
where $\hat{\rho}_{-i,x}$ is the codon/nucleotide x frequency at position $-i$ before stop, while $\hat{\rho}_{intra,x}$ is the intragene codon/nucleotide x frequency.

DNA rythm


A
0.1
0.0
-0.1
-0.2
-50 -48 -46 -44 -42 -40 -38 -36 -34 -32 -30 -28 -26 -24 -22 -20 -18 -16 -14 -12 -10 -9 -8 -7 -6 -5 -4 -3
T
0.10
0.05
0.00
-0.05
-0.10
-0.15
-50 -48 -46 -44 -42 -40 -38 -36 -34 -32 -30 -28 -26 -24 -22 -20 -18 -16 -14 -12 -10 -9 -8 -7 -6 -5 -4 -3

DNA rythm


G
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-50 -48 -46 -44 -42 -40 -38 -36 -34 -32 -30 -28 -26 -24 -22 -20 -18 -16 -14 -12 -10 -9 -8 -7 -6 -5 -4 -3


C
0.10
0.05
0.00
-0.10
-0.20
-0.25
-50 -48 -46 -44 -42 -40 -38 -36 -34 -32 -30 -28 -26 -24 -22 -20 -18 -16 -14 -12 -10 -9 -8 -7 -6 -5 -4 -3

DNA rythm

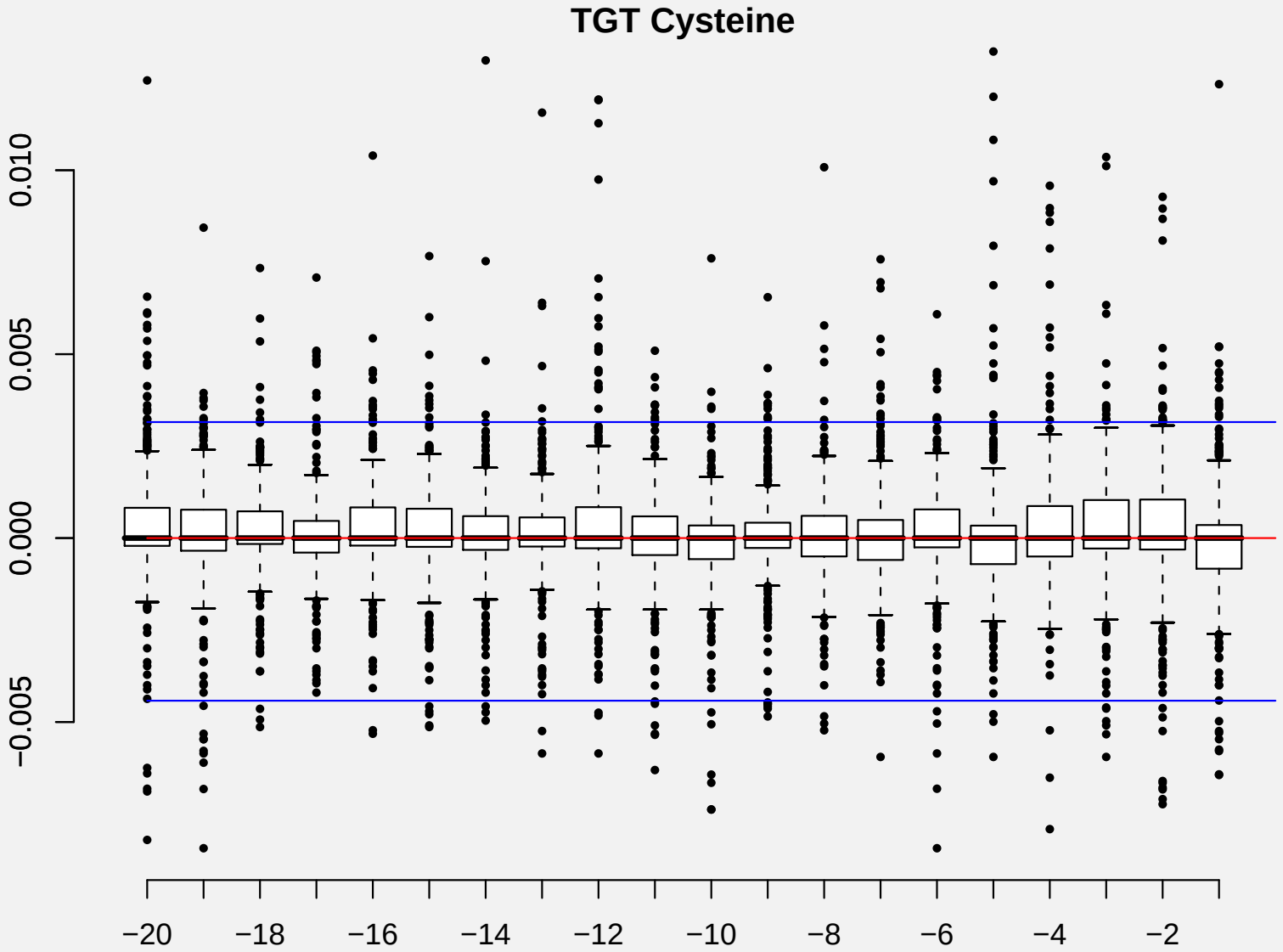
Changes in nucleotide composition

- loss of tri-nucleotide frequency pattern,
- nucleotide A is overrepresented,

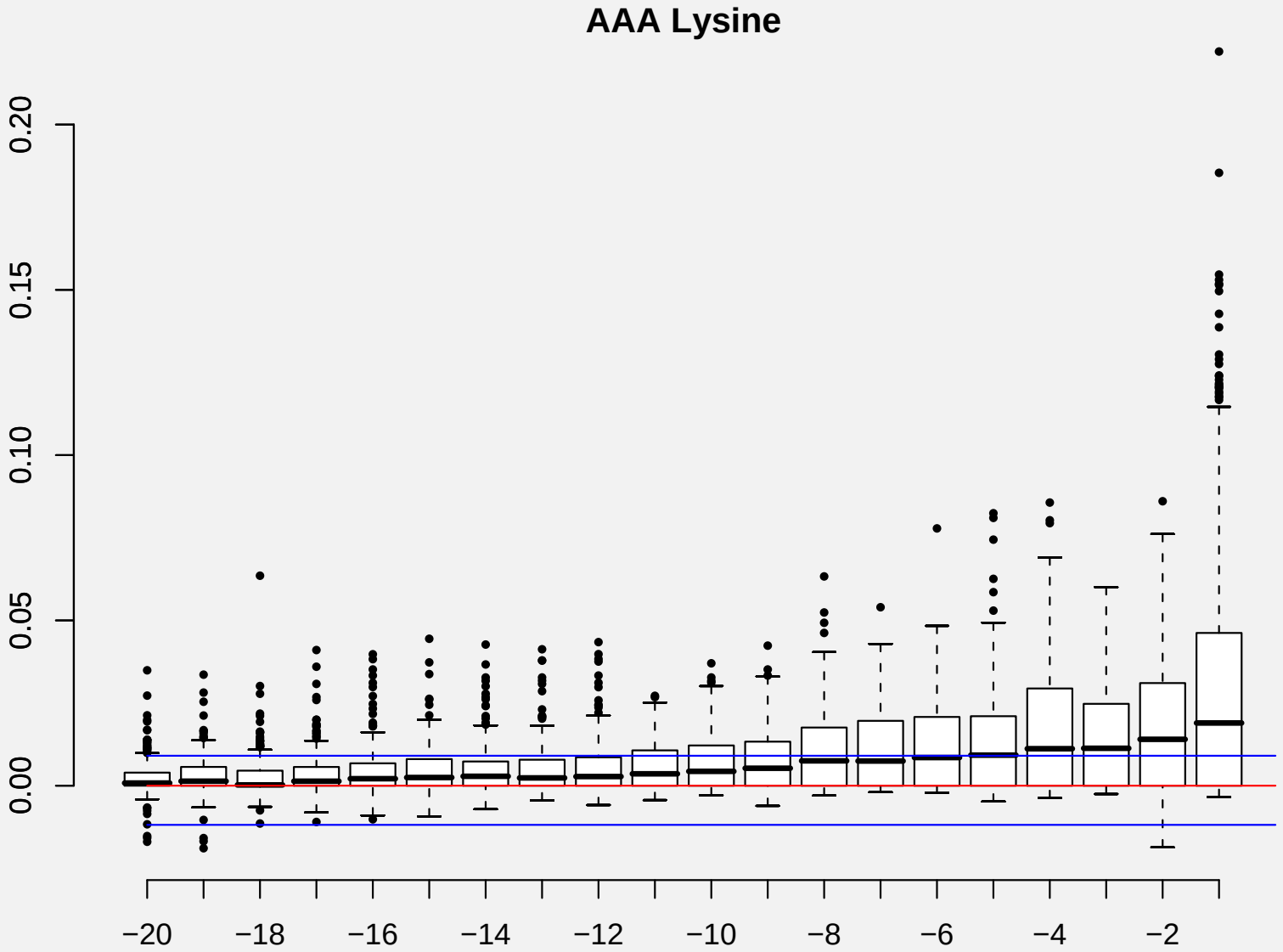
have impact on mRNA secondary structure and may prepare ribosome to termination.

We think that codons composition is more important than nucleotide composition.

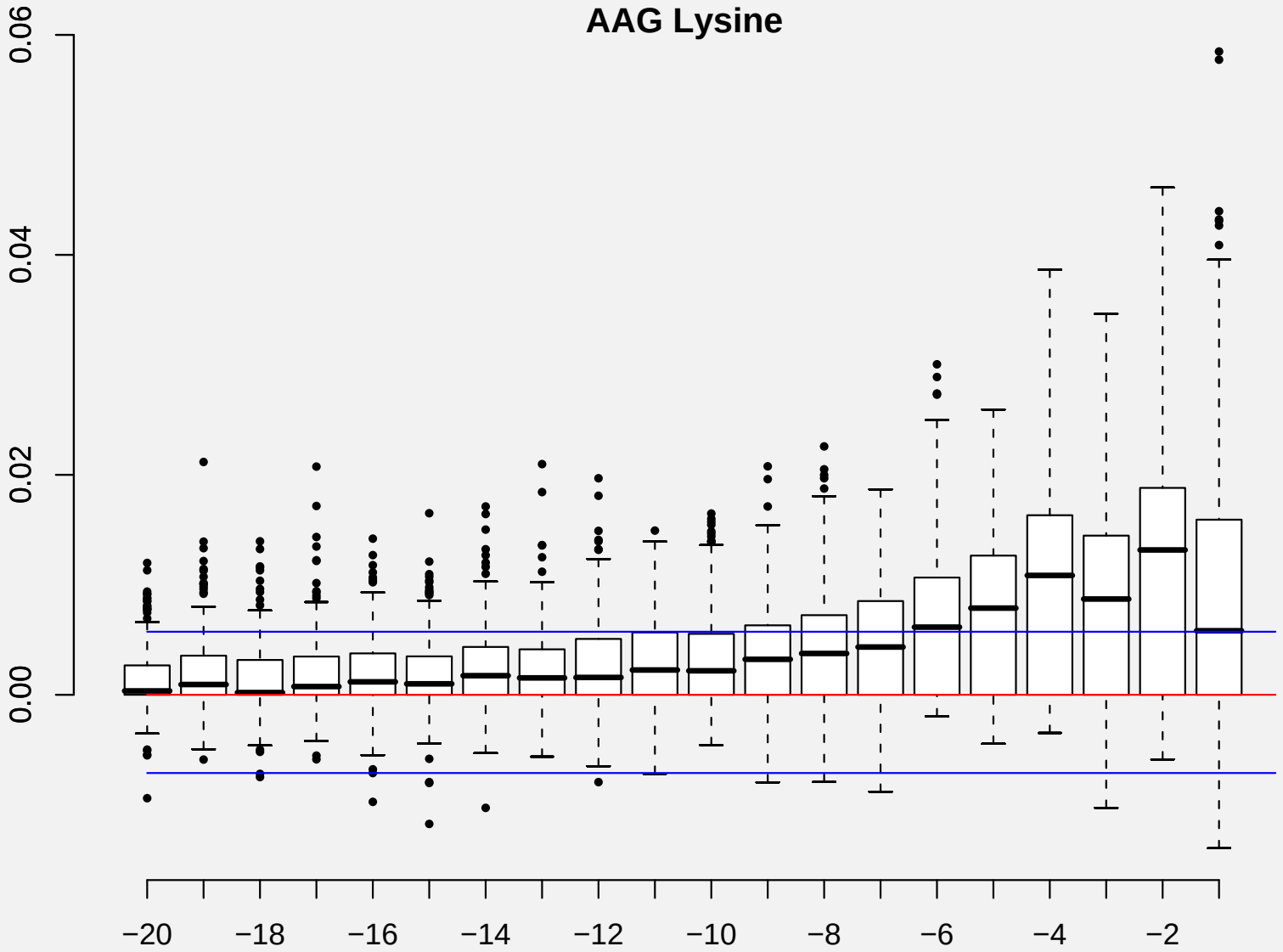
Codon composition



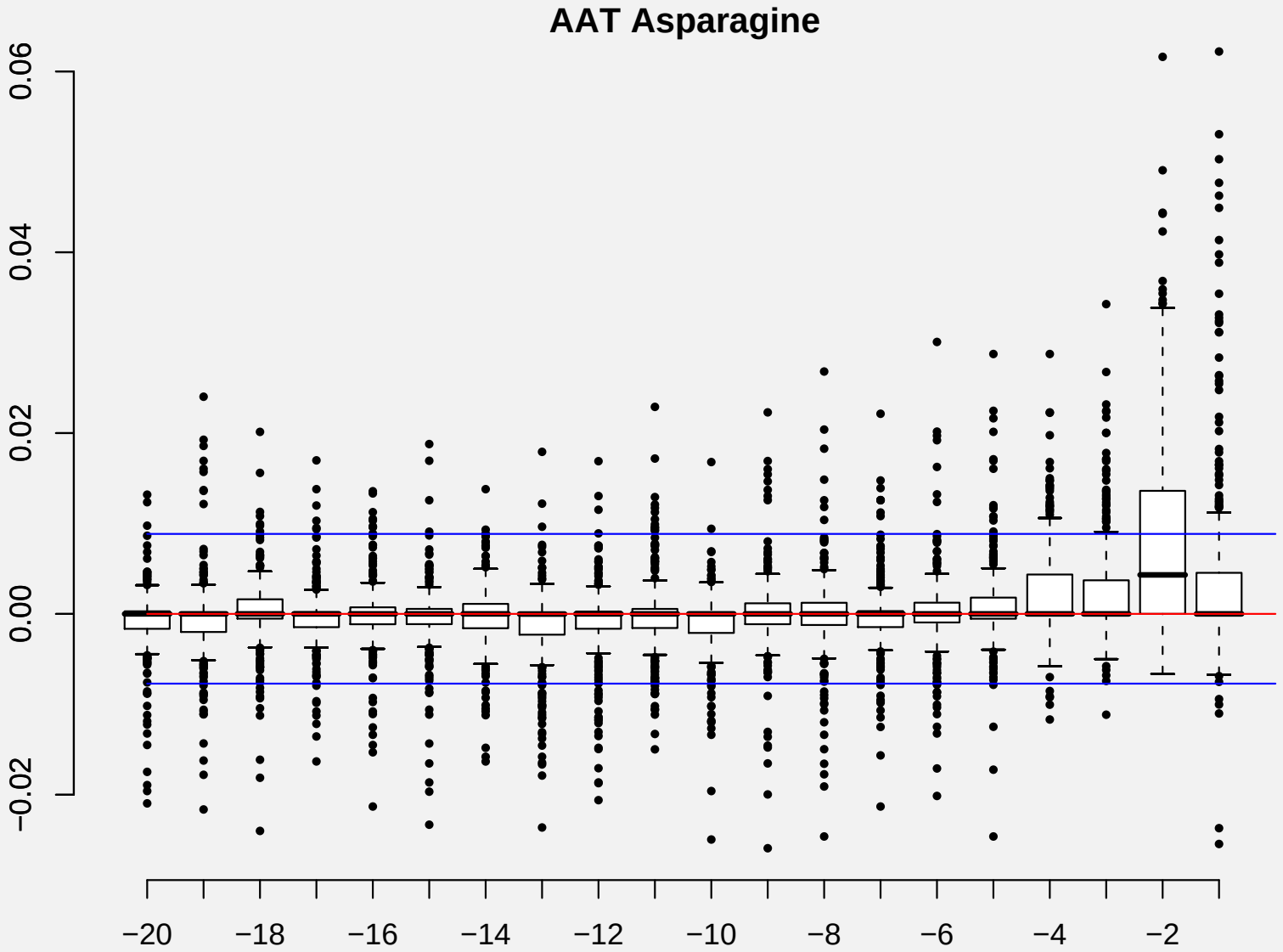
Codon composition



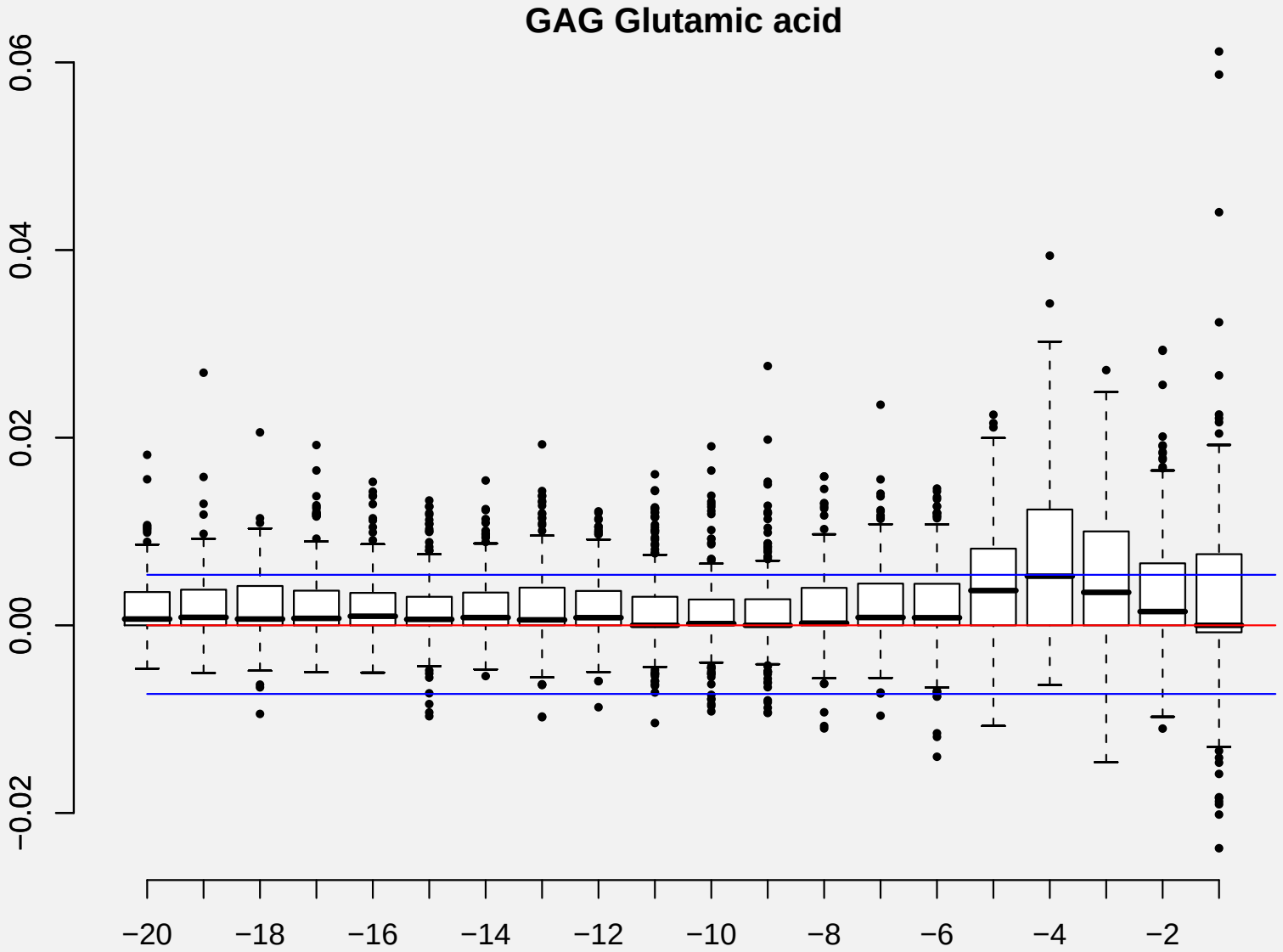
Codon composition



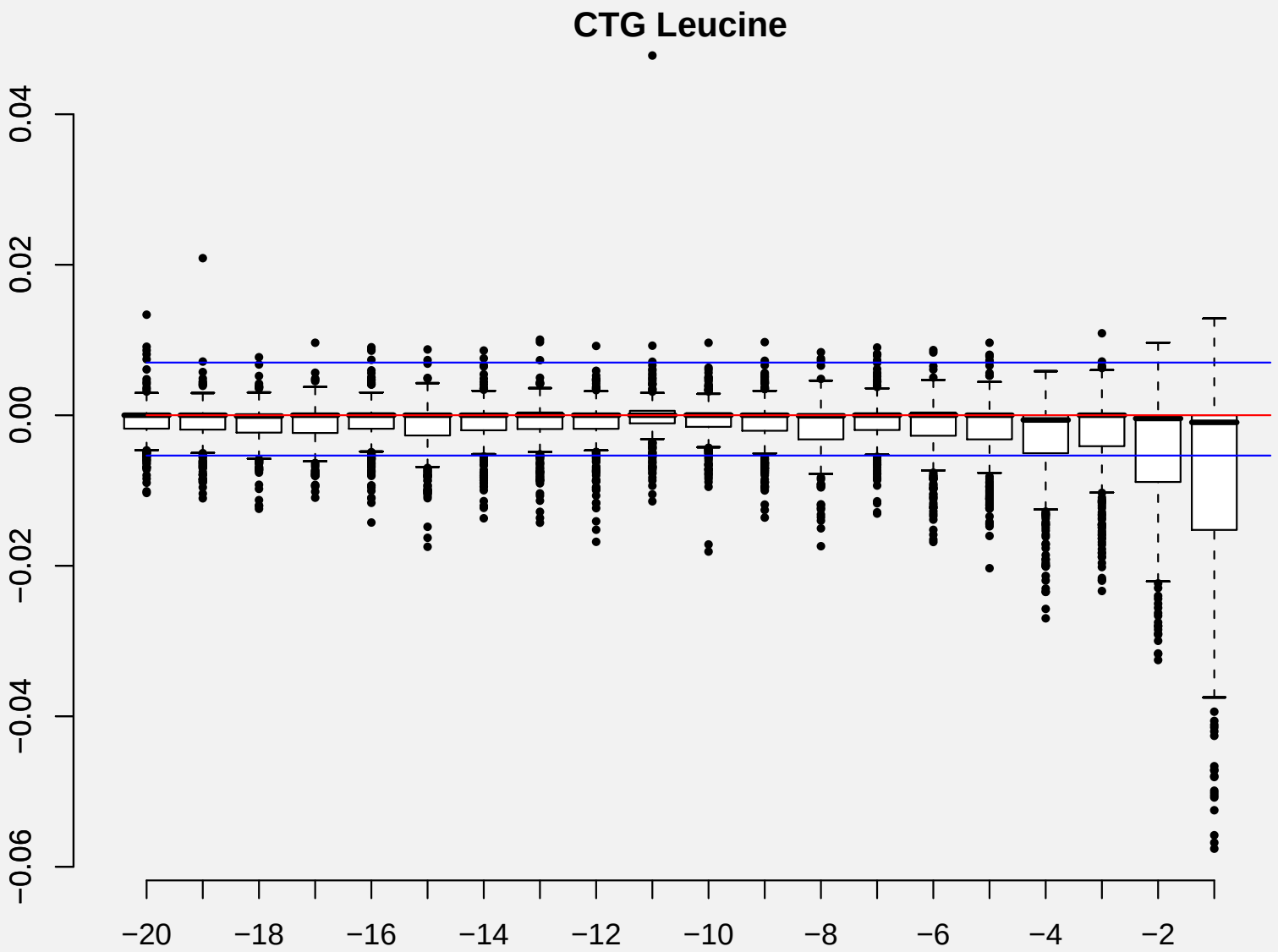
Codon composition



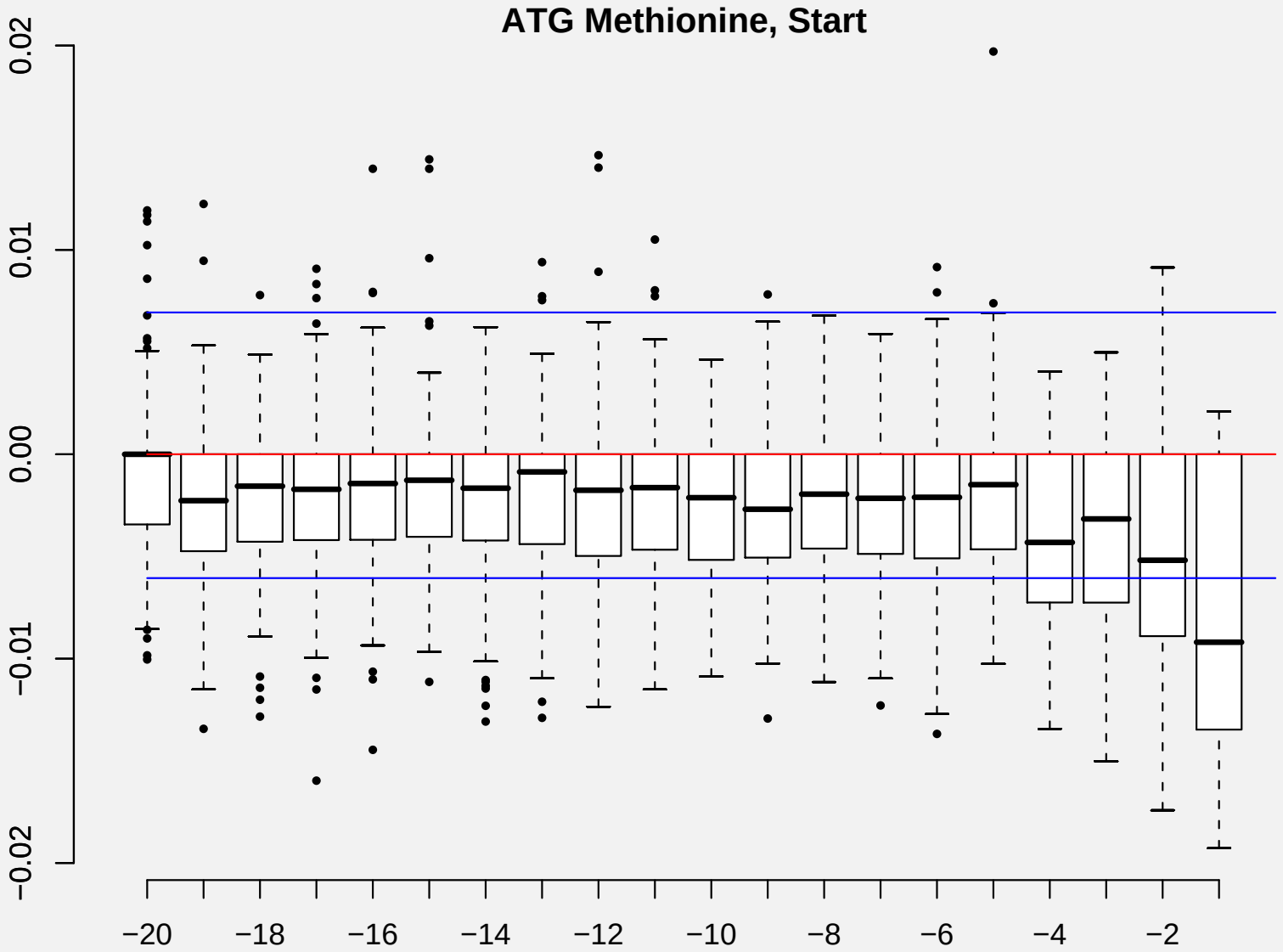
Codon composition



Codon composition



Codon composition



Codon composition

Changes in codons composition

- codon AAA is highly overrepresented,
- in some codons (e.g. AAG) strong signal is observed in more than one position,
- in most cases the strongest signal is in position -1,
- in some codons (e.g. AAT) in position -2 or (e.g. GAG) in position -4,
- some codons (e.g. ATG - start) are strongly underrepresented.

Genomes

Let's have a look on some genomes. We plot results for position -1. As before the codon skew is computed as:

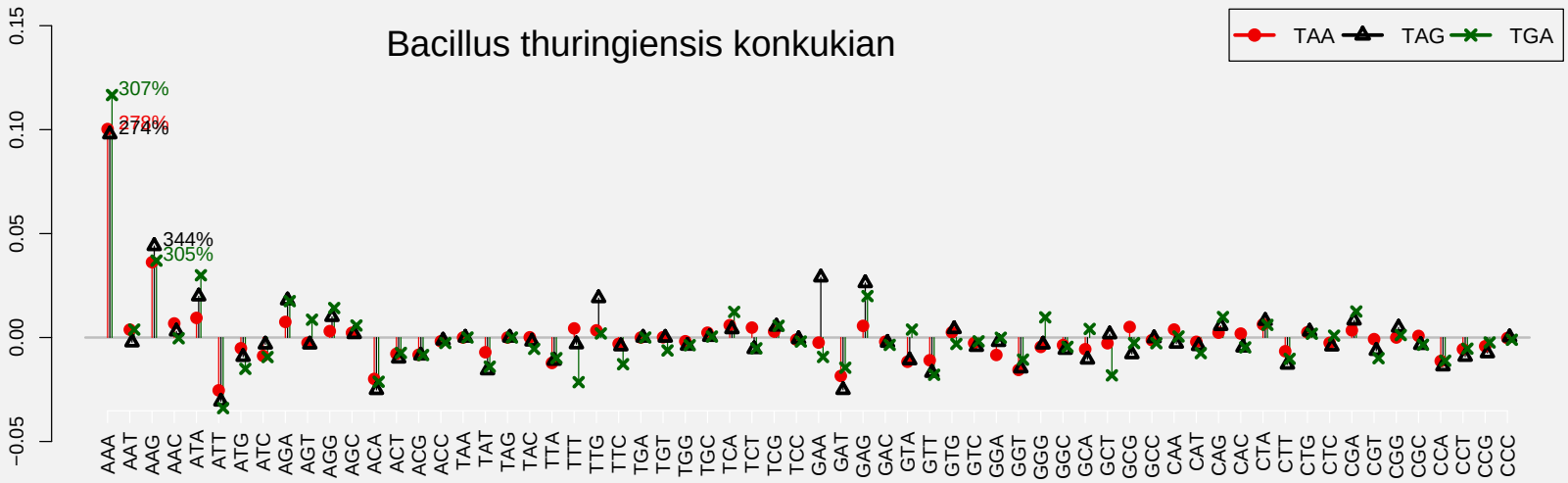
$$skew_x = \hat{\rho}_{-i,x} - \hat{\rho}_{intra,x},$$

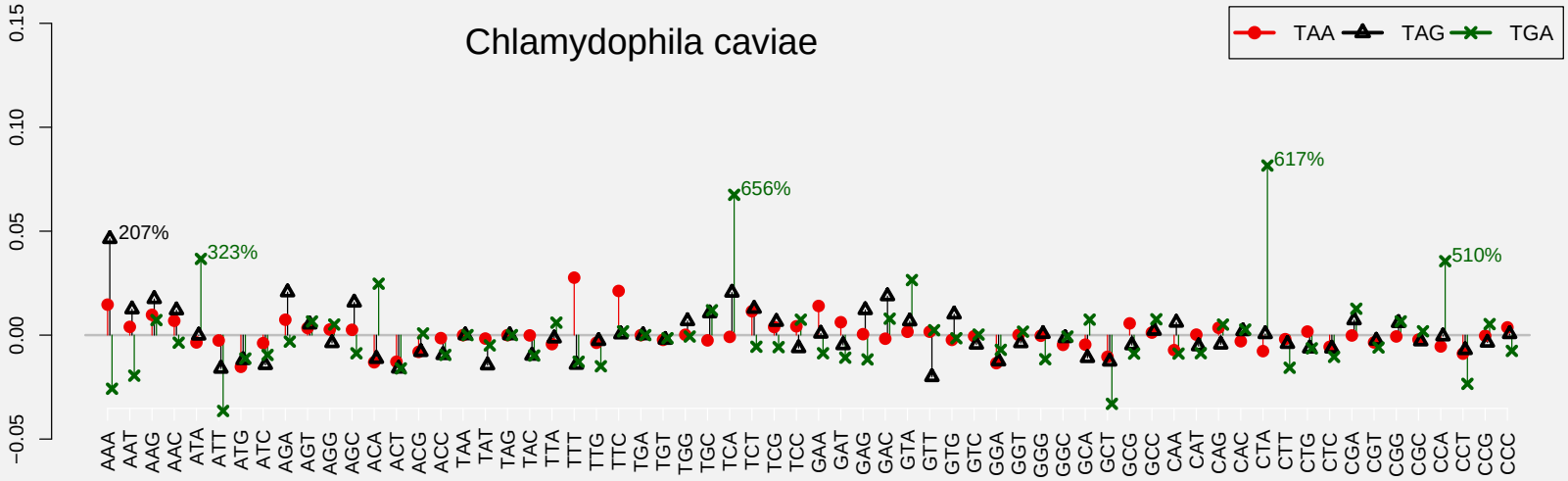
codon relative change is computed as

$$relative_x = \hat{\rho}_{-i,x} / \hat{\rho}_{intra,x},$$

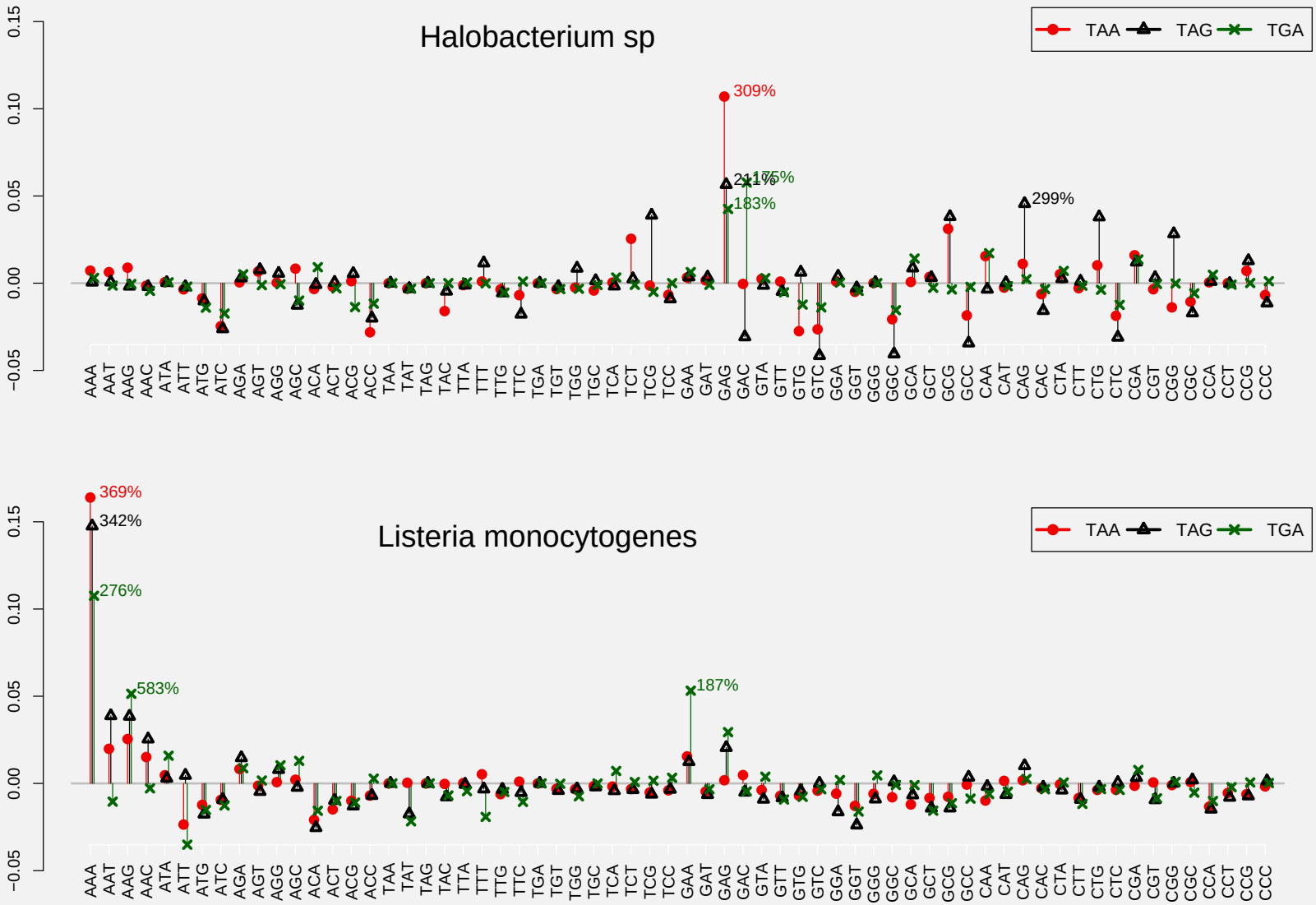
where $\hat{\rho}_{-i,x}$ is the codon x frequency at position $-i$ before stop, while $\hat{\rho}_{intra,x}$ is the codon x intragene frequency.

Genomes

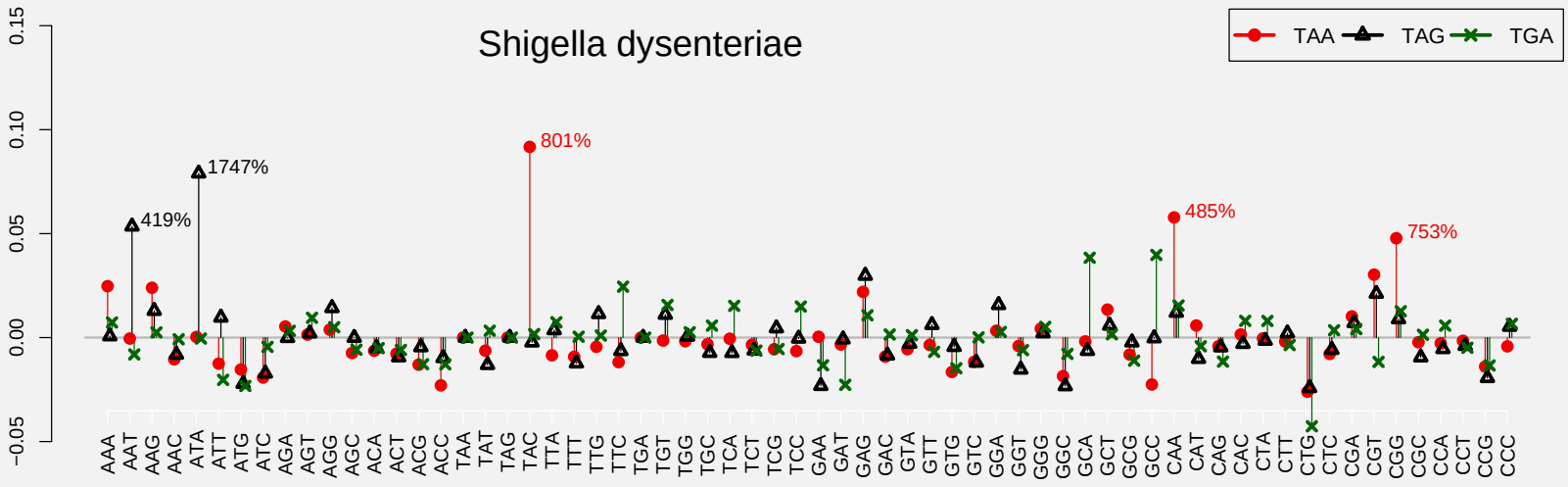

Bacillus thuringiensis konkukian
TAA
TAG
TGA
307%
278%
274%
344%
305%
0.15
0.10
0.05
0.00
−0.05
AAA AAT AAG AAC ATA ATT ATG ATC AGA AGT AGG AGC ACA ACT ACG ACC TAA TAT TAG TAC TTA TTT TTG TTC TGA TGT TGG TGC TCA TCT TCG TCC GAA GAT GAG GAC GTA GTT GTG GTC GGA GGT GGG GGC GCA GCT GCG GCC CAA CAT CAG CAC CTA CTT CTG CTC CGA CGT CGG CGC CCA CCT CCG CCC

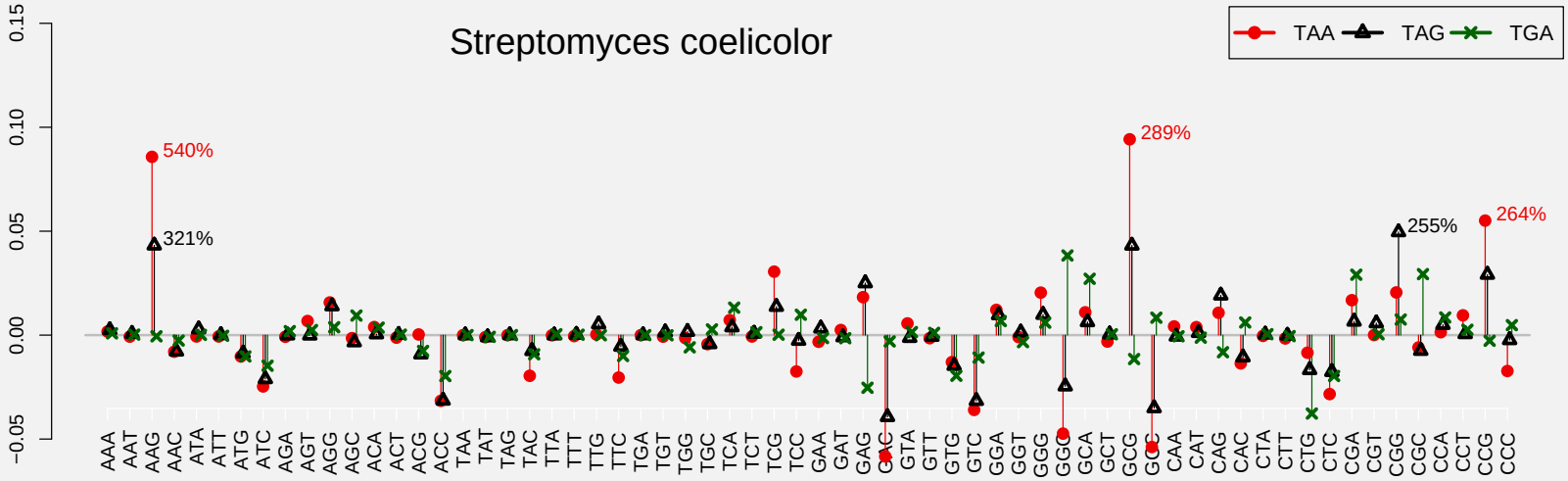

Chlamydophila caviae
TAA
TAG
TGA
207%
323%
656%
617%
510%
0.15
0.10
0.05
0.00
−0.05
AAA AAT AAG AAC ATA ATT ATG ATC AGA AGT AGG AGC ACA ACT ACG ACC TAA TAT TAG TAC TTA TTT TTG TTC TGA TGT TGG TGC TCA TCT TCG TCC GAA GAT GAG GAC GTA GTT GTG GTC GGA GGT GGG GGC GCA GCT GCG GCC CAA CAT CAG CAC CTA CTT CTG CTC CGA CGT CGG CGC CCA CCT CCG CCC

Genomes


Halobacterium sp
TAA
TAG
TGA
309%
211%
175%
183%
299%
Listeria monocytogenes
369%
342%
276%
583%
187%
0.15
0.10
0.05
0.00
−0.05
AAA AAT AAG AAC ATA ATT ATG ATC AGA AGT AGG AGC ACA ACT ACG ACC TAA TAT TAG TAC TTA TTT TTG TTC TGA TGT TGG TGC TCA TCT TCG TCC GAA GAT GAG GAC GTA GTT GTG GTC GGA GGT GGG GGC GCA GCT GCG GCC CAA CAT CAG CAC CTA CTT CTG CTC CGA CGT CGG CGC CCA CCT CCG CCC

Genomes


Shigella dysenteriae
TAA
TAG
TGA
0.15
0.10
0.05
0.00
−0.05
1747%
419%
801%
485%
753%
AAA AAT AAG AAC ATA ATT ATG ATC AGA AGT AGG AGC ACA ACT ACG ACC TAA TAT TAG TAC TTA TTT TTG TTC TGA TGT TGG TGC TCA TCT TCG TCC GAA GAT GAG GAC GTA GTT GTG GTC GGA GGT GGG GGC GCA GCT GCG GCC CAA CAT CAG CAC CTA CTT CTG CTC CGA CGT CGG CGC CCA CCT CCG CCC


Streptomyces coelicolor
TAA
TAG
TGA
0.15
0.10
0.05
0.00
−0.05
540%
321%
289%
255%
264%
AAA AAT AAG AAC ATA ATT ATG ATC AGA AGT AGG AGC ACA ACT ACG ACC TAA TAT TAG TAC TTA TTT TTG TTC TGA TGT TGG TGC TCA TCT TCG TCC GAA GAT GAG GAC GTA GTT GTG GTC GGA GGT GGG GGC GCA GCT GCG GCC CAA CAT CAG CAC CTA CTT CTG CTC CGA CGT CGG CGC CCA CCT CCG CCC

Genomes

In some genomes (e.g. Bacillus thuringiensis) there is no significant changes in codon overrepresentation before different stop codons.

In some genomes (e.g. Shigella dysenteriae) there is significant difference.

Why they behave differently?

Statistical background

The test of proportions was used to compare the frequency of given codon in position before a stop signal and in positions inside genes (intragenic). The frequency before stop signal was also compared with intergenic frequency, and in most cases results are similar.

The Bonferroni correction was applied to deal with the number of tests (for 400 genomes and 64 codons). For a particular test the very conservative significance level was used

$$\alpha = \frac{0.05}{400 * 64}.$$

While we test on such small significance level, all rejections are positive with high probability.

GC content

Genomes were divided into 4 groups according to their GC content.

GC content	**<37.5**	**37.5÷50**	**50÷62.5**	**>62.5**
AAA	90.57%	82.68%	80.61%	73.42%
AAC	23.58%	18.90%	9.18%	7.59%
AAG	57.54%	49.52%	42.85%	29.11%
AGA	26.42%	51.18%	66.33%	75.95%
AGG	27.36%	24.41%	32.65%	54.43%
TCA	5.66%	19.69%	65.31%	92.41%
GGA	0.94%	5.51%	44.90%	86.08%
GCA	0.94%	6.30%	56.12%	96.20%
CGA	14.15%	26.77%	90.82%	100.00%
CCA	0.94%	3.94%	47.96%	94.94%

GC content

- Some codons (e.g. GCA, CGA, TCA) are often overrepresented in position -1 in genomes with high GC content. Other codons (e.g. AAA, AAG) are often overrepresented in position -1 in genomes with low GC content! (no artefacts)
- The lower overrepresentation of AAA codon in genomes with high GC content may results from different mutational pressure.

Note that codon frequencies before stop were compared with intergenic or intragenic codon frequencies in the same genome.

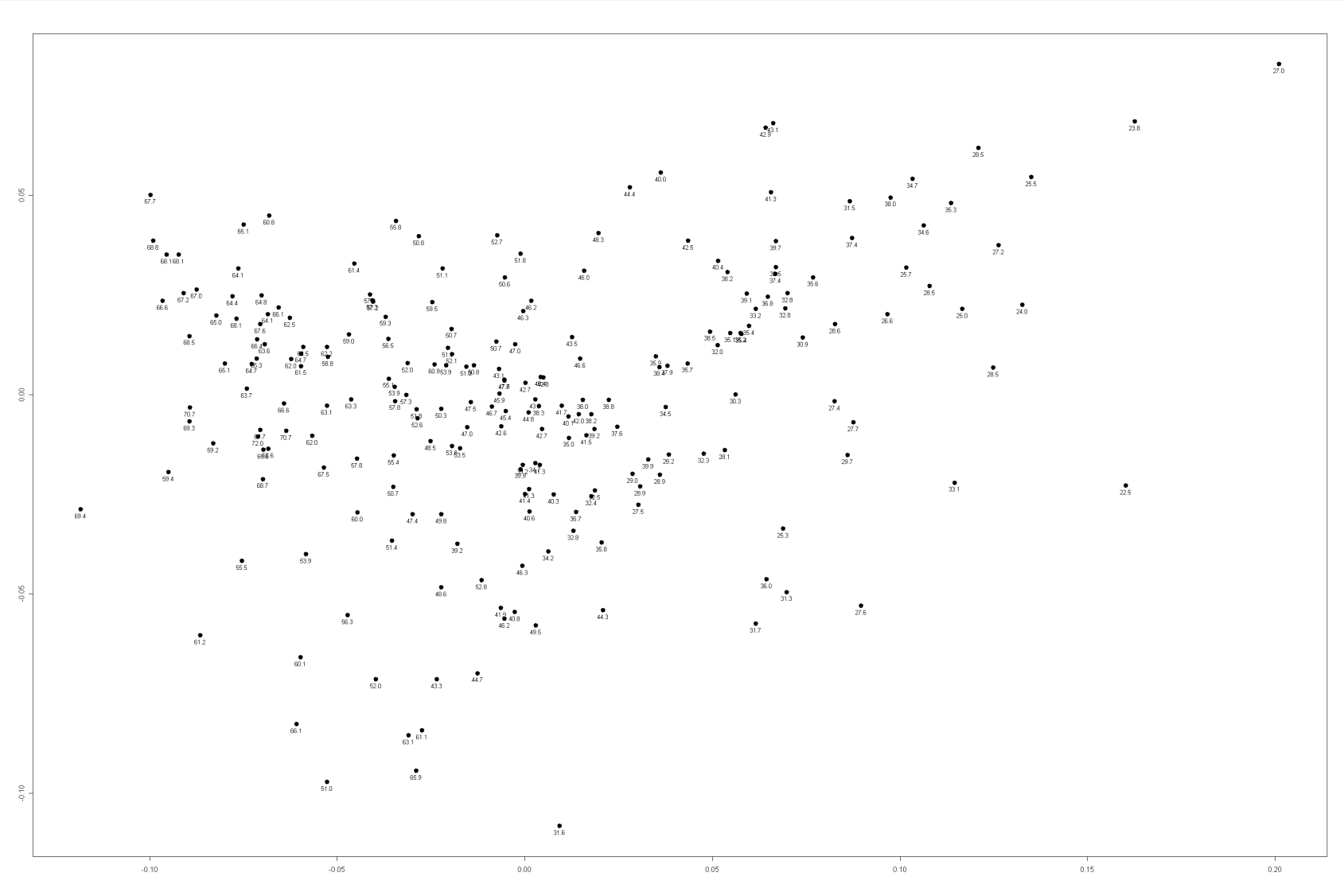
MDS

We use Multidimensional Scaling to visualize differences in codon composition for different genomes.

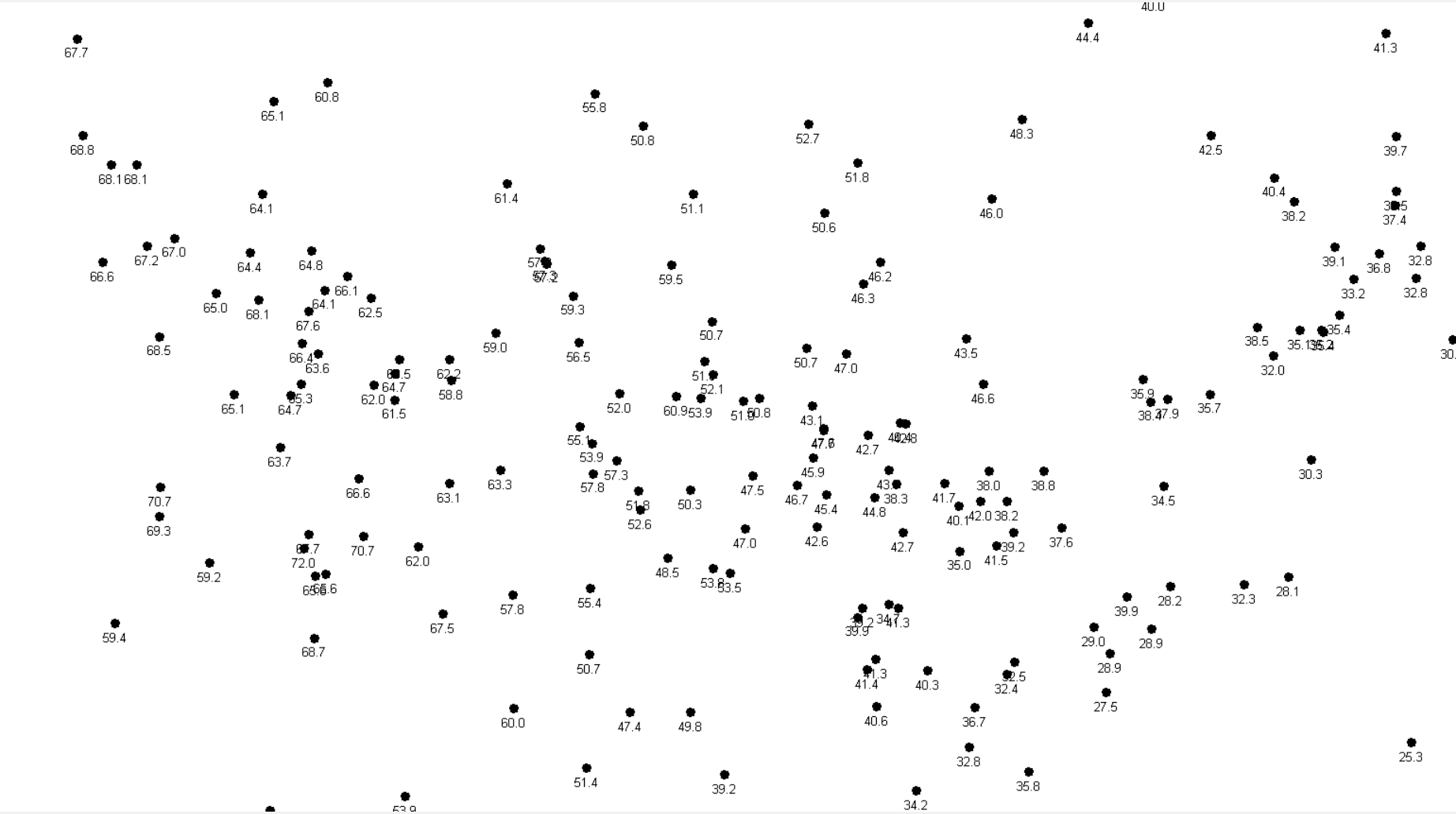
MDS is a set of related statistical techniques often used in data visualization or feature extraction problems.

Using MDS we find a new coordinates for genomes in 2D or 3D space. Coordinates are computed in a way to minimize differences between distances in the original parameter space (in our case it has 128 dimensions) and in the new parameter space.

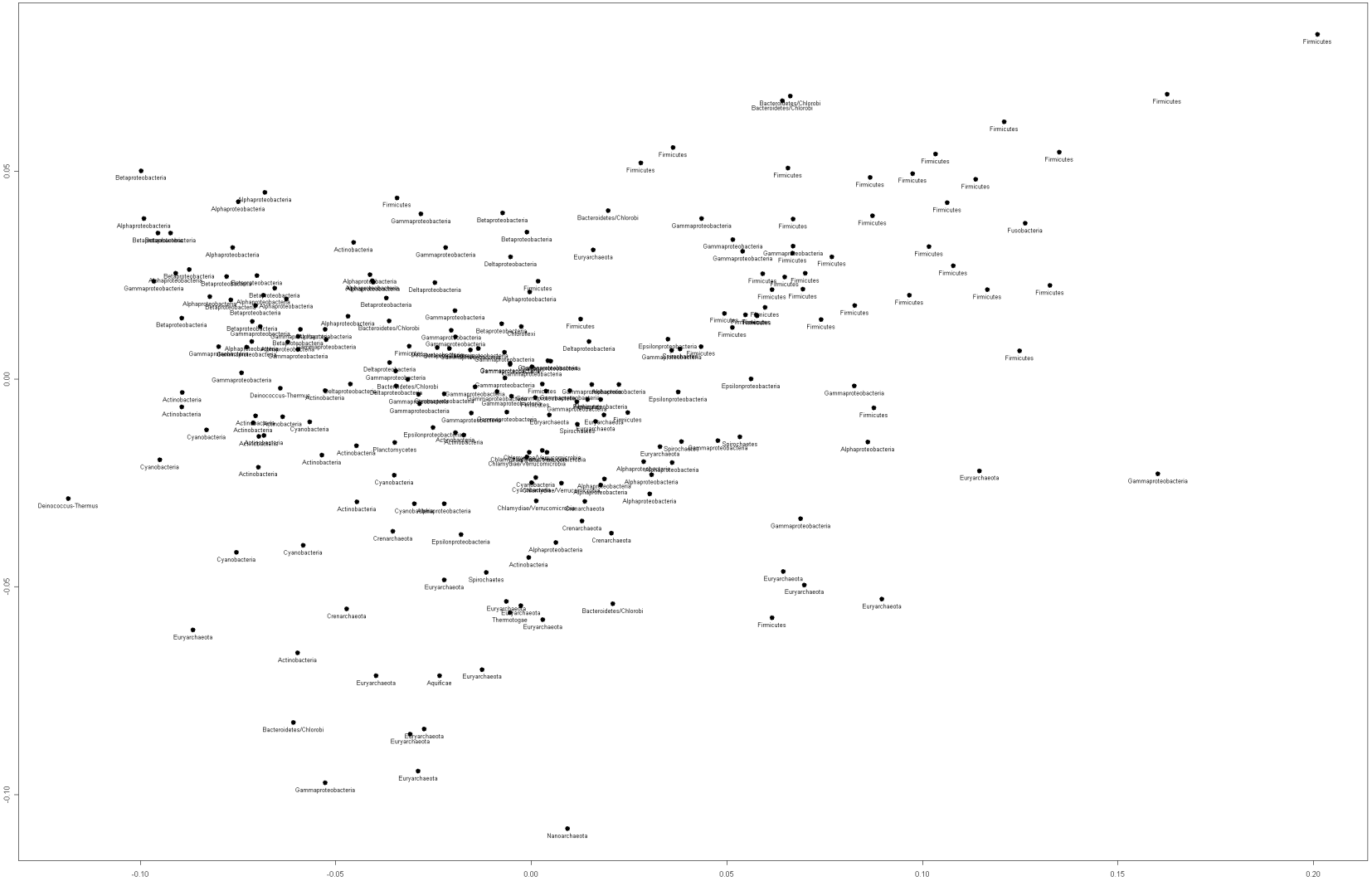
MDS for GC content



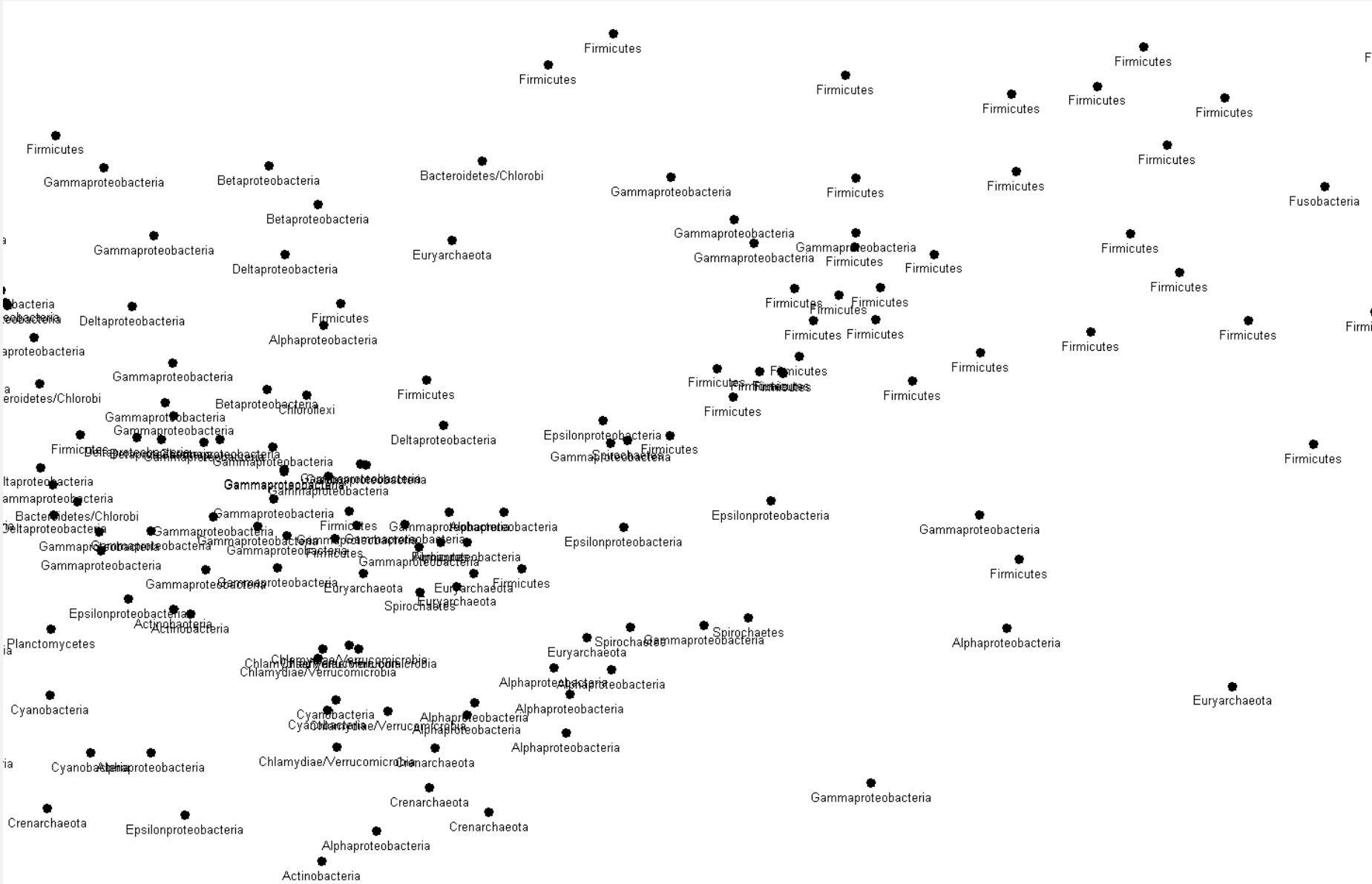
MDS for GC content



MDS for taxonomic groups



MDS for taxonomic groups



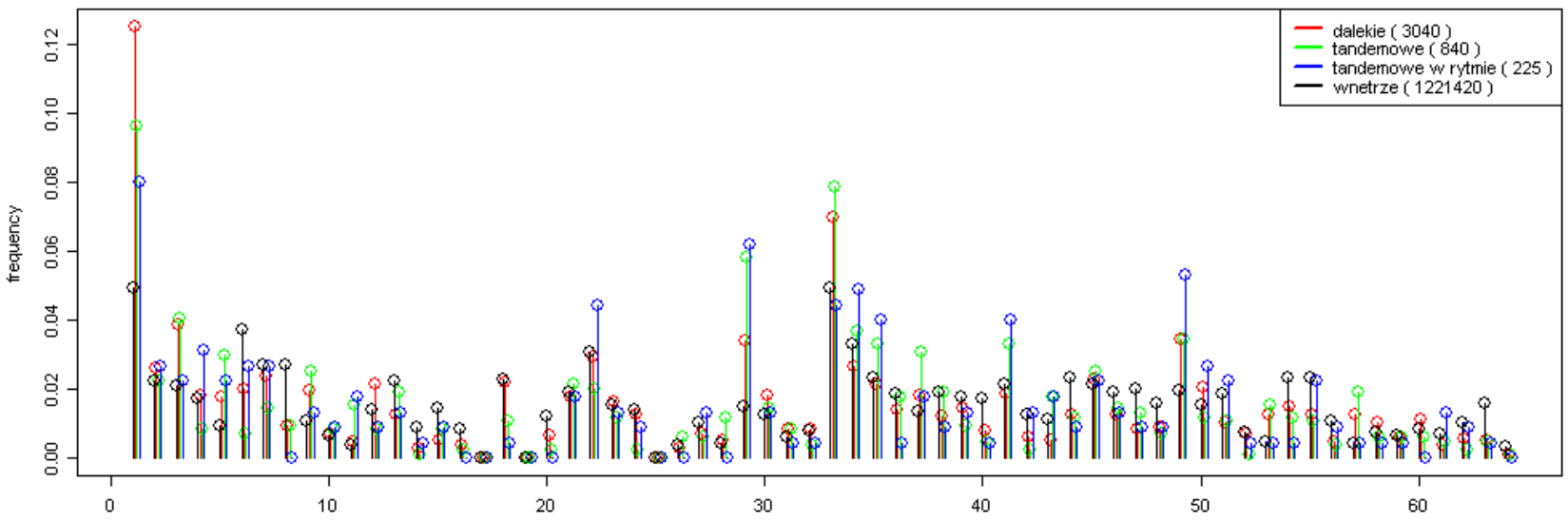
Other results

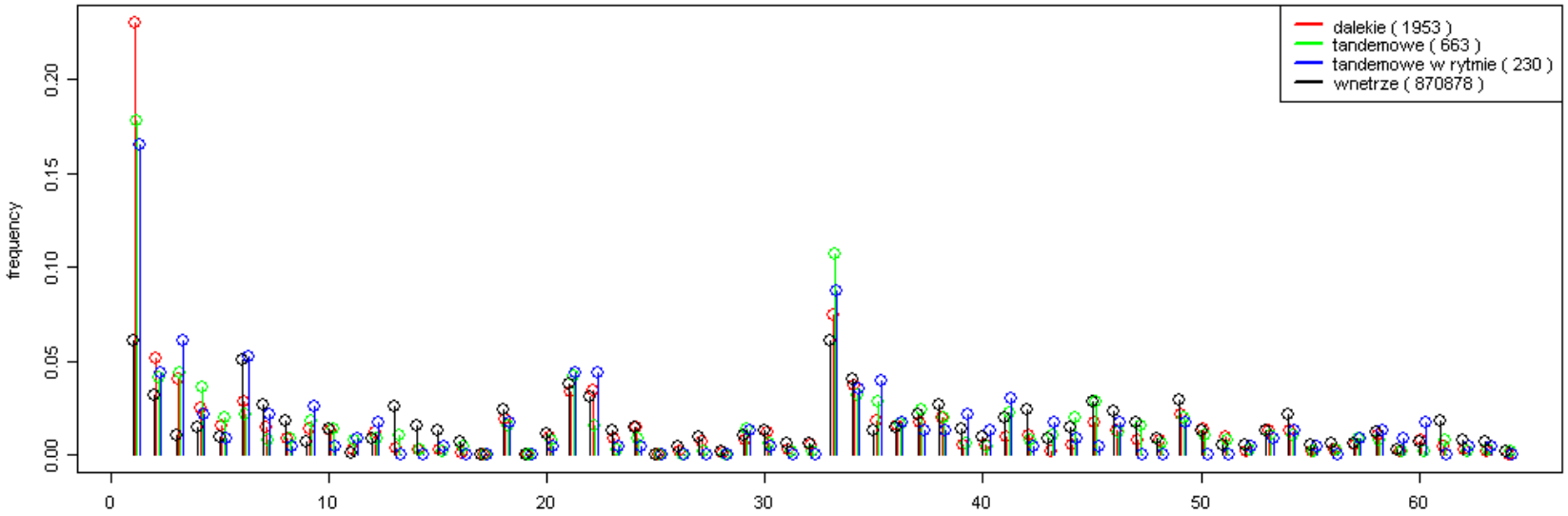
There is much more figures that may be presented

- results for Tandem genes,
- other genome properties (size, habitat, temp. range, etc.),
- results for amino acids composition,
- results for different isoelectric points (pI),
- ...

but they are not so interesting.

Other results





Summary

- Nucleotide and amino acid composition is changed before stop codon.
- Codons A**AA**, A**GA** and A**AG** are highly overrepresented.
- Codons G**AA**, G**GA** and G**AG** are overrepresented.
- Codon AUG is underrepresented.
- Differences in codon overrepresentation correlates with GC content and (possible) mutation pressure.