Lognormal Pattern

of Exon size distributions

in Eukaryotic genomes

Yaroslav Ryabov

Outline

- Global and local approaches in analysis of genetic information
- Vocabulary of contemporary genetics:
  Exons and Introns
- What we can learn from exon size distributions?
  Lognormal pattern and two clases of exons
- How we can model exon size distributions of real genomes?
- What could be the biological reason for observed pattern of exons size distributions?
- Conclusions

Analyzing Genetic Information

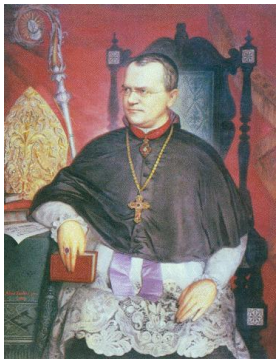
Global approach

analyzing properties of entire genome

Recombination of inherited properties

Frequency of mutations

Size of genome *etc*.



Georg Johann Mendel
1866

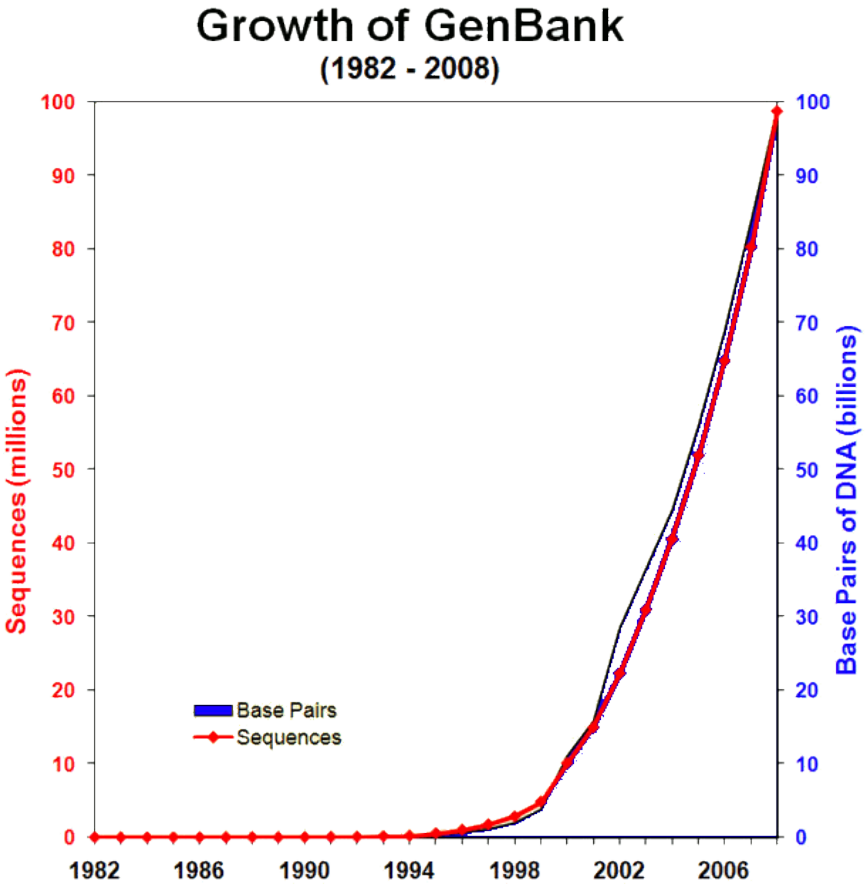
Local approach

analyzing details of nucleotide sequence

Watson and Crick
1953
Marshall Nirenberg
1968

Basic **L**ocal
Alignment **S**earch **T**ool
1990

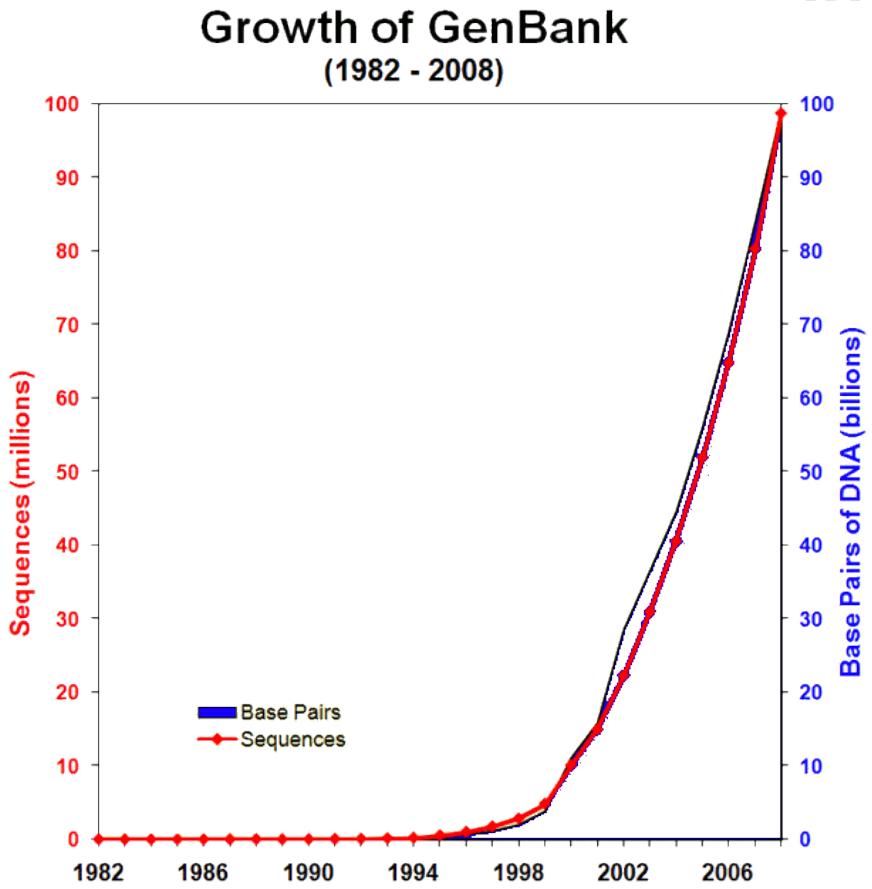
Analyzing Genetic Information in Post-Genomic Era



e!Ensembl

More than 60 animal genomes with complete annotations

Analyzing Genetic Information in Post-Genomic Era



e!Ensembl

More than 60 animal genomes with complete annotations

Back to Global Approach ?

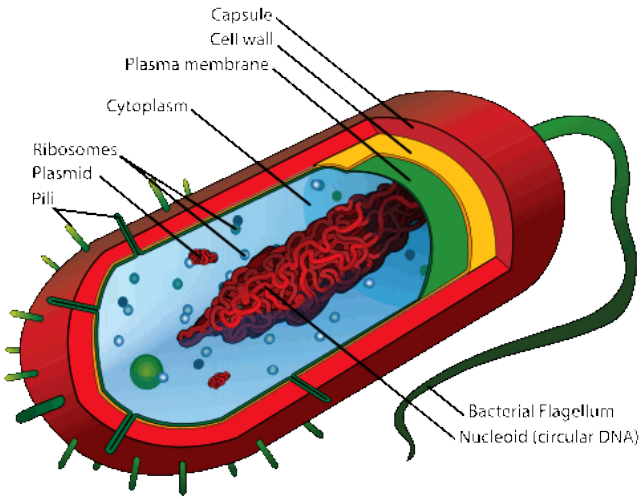
Prokaryote and *Eukaryote*

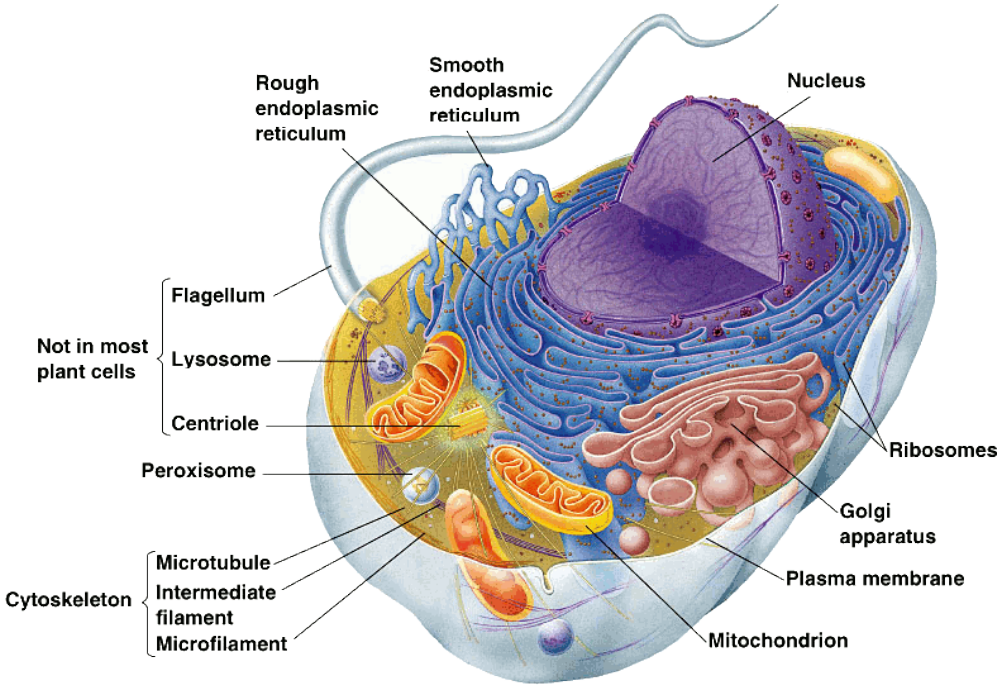
pro + karyon

before + nucleos

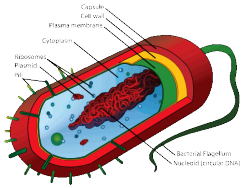
eu + karyon

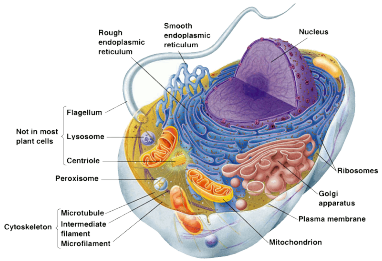
good + nucleos





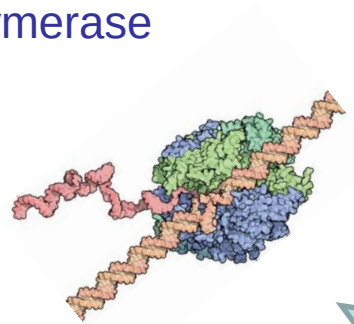
Gene expression

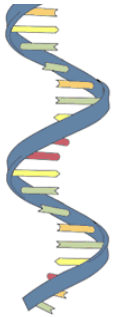




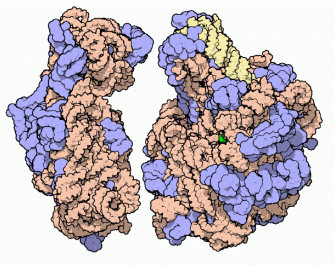
mRNA

Transcription with RNA polymerase



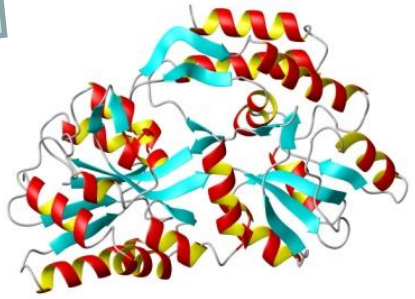


Translation with ribosome

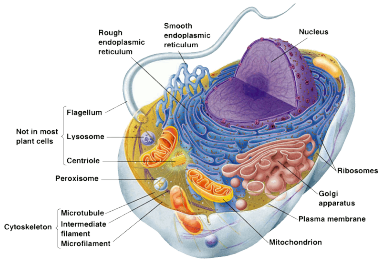


DNA

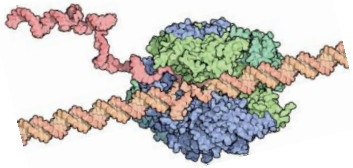
Protein



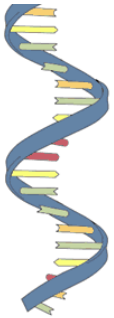
Splicing in Eukaryotes



Transcription with RNA polymerase

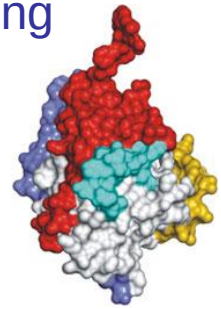


mRNA



DNA

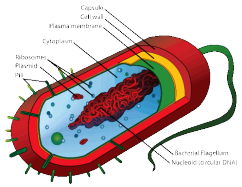
Splicing

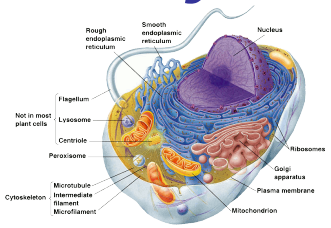


Exons

Introns

Prokaryote *vs* *Eukaryote*





Most of DNA code is used to produce some cellular product

Substantial fraction "silent" DNA regions

Exons **Introns**

Short genomes: ~ 1 000 exons

Long exons: ~ 1 000 base pairs

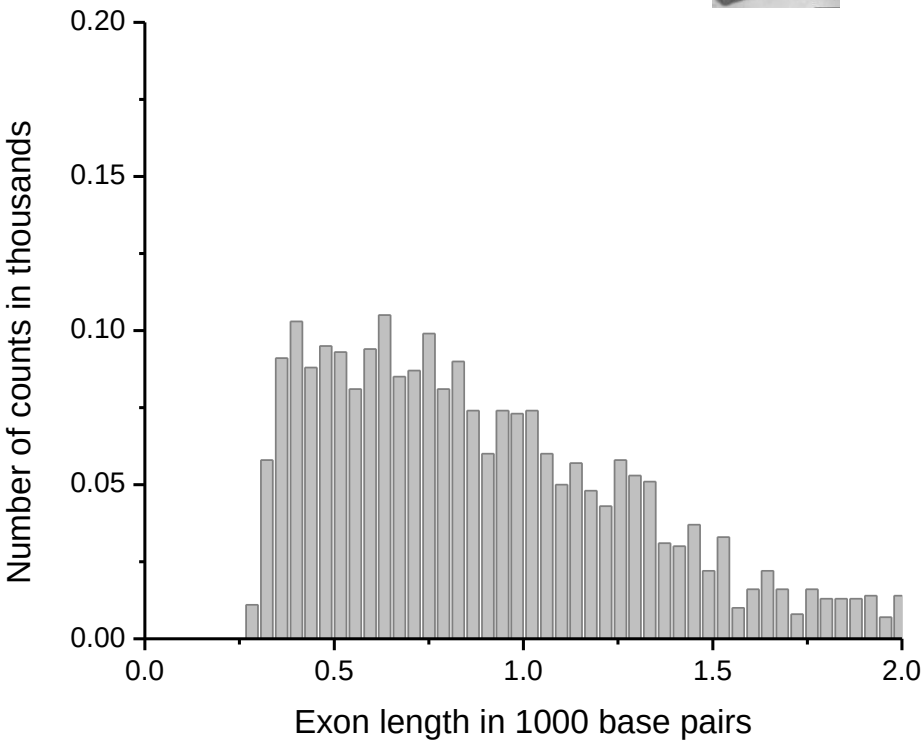
Long genomes: ~ 100 000 exons

Short exons: ~ 100 base pairs

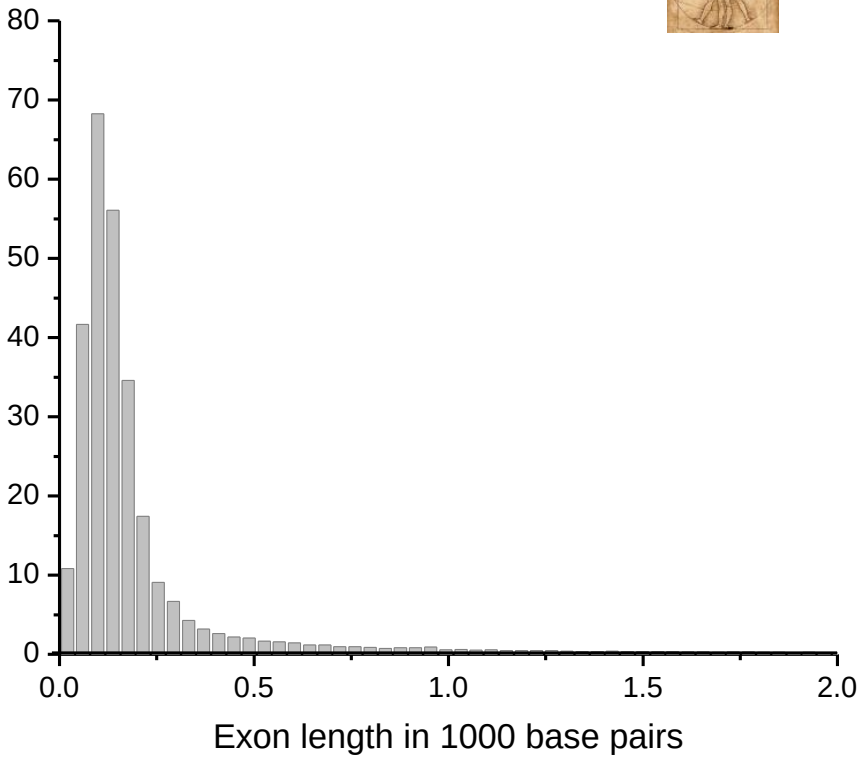
Prokaryote *vs* ***Eukaryote***

Exon size distributions

Flavobacteria
Bacterium



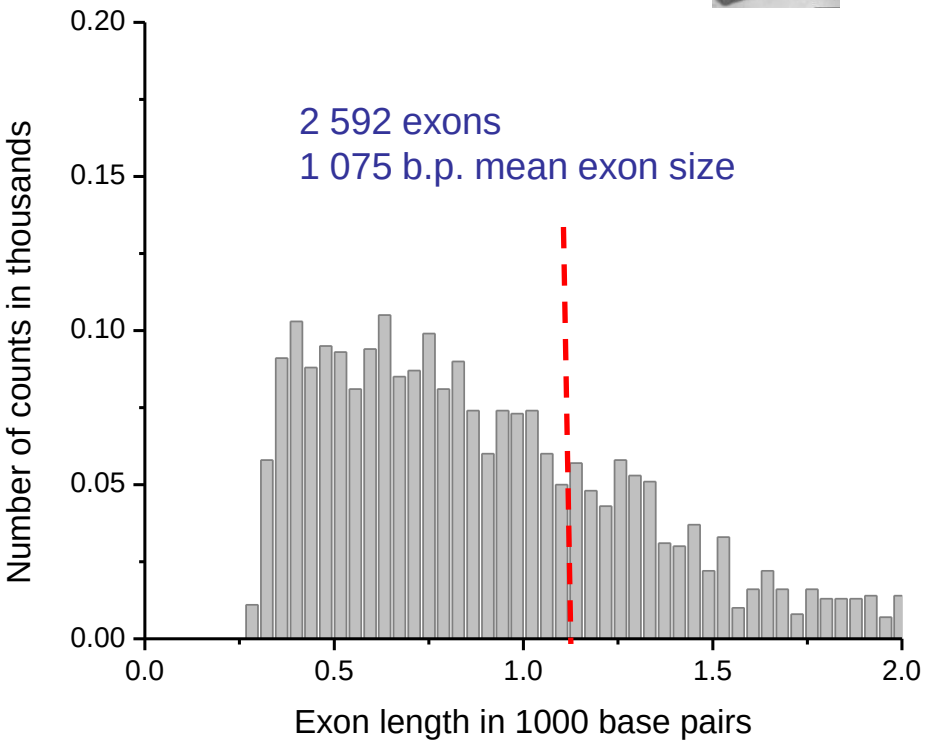
Homo Sapiens



What we can learn from exon size distributions ?

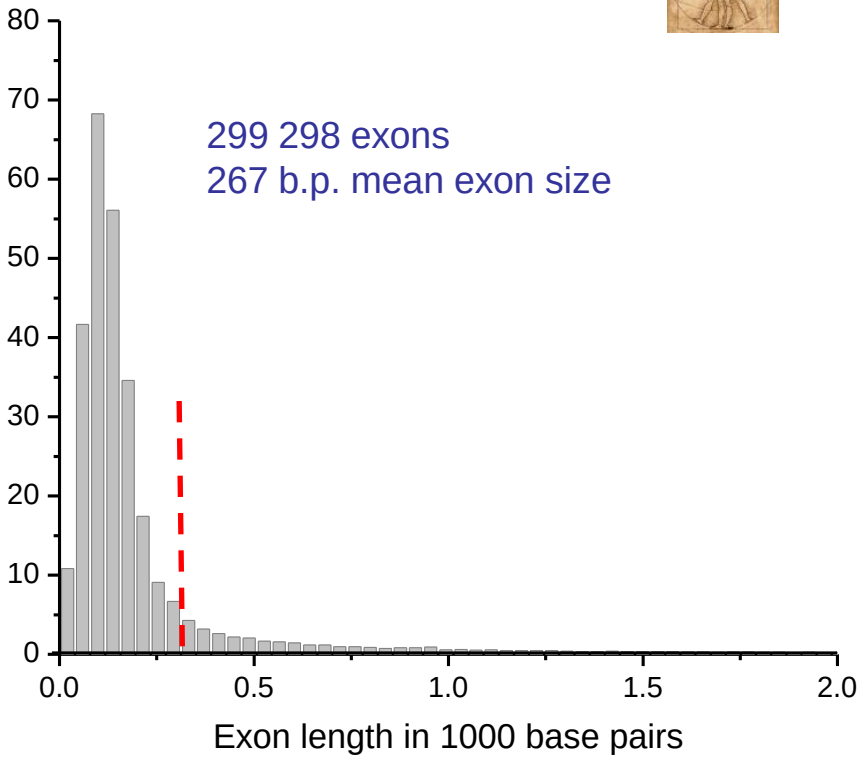
Flavobacteria
Bacterium

2 592 exons
1 075 b.p. mean exon size



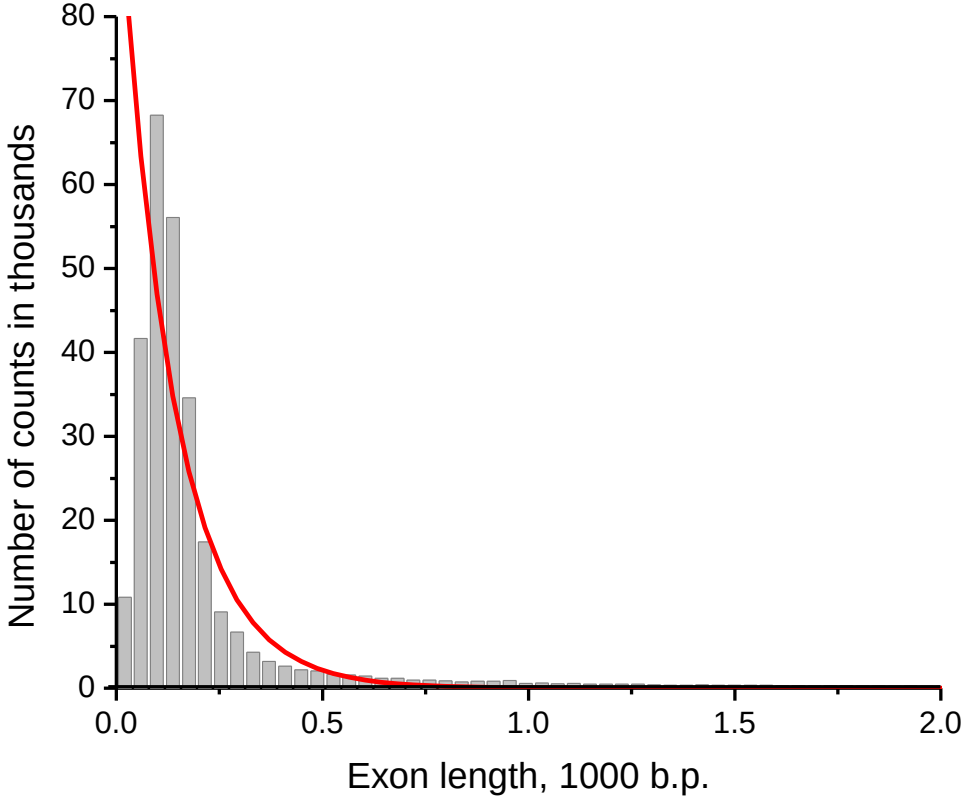
Homo Sapiens

299 298 exons
267 b.p. mean exon size



Poisson Process

λ = *Const* frequency of splitting



Assumption:

Probability to split DNA at any location and any time is *Constant*

Consequence:

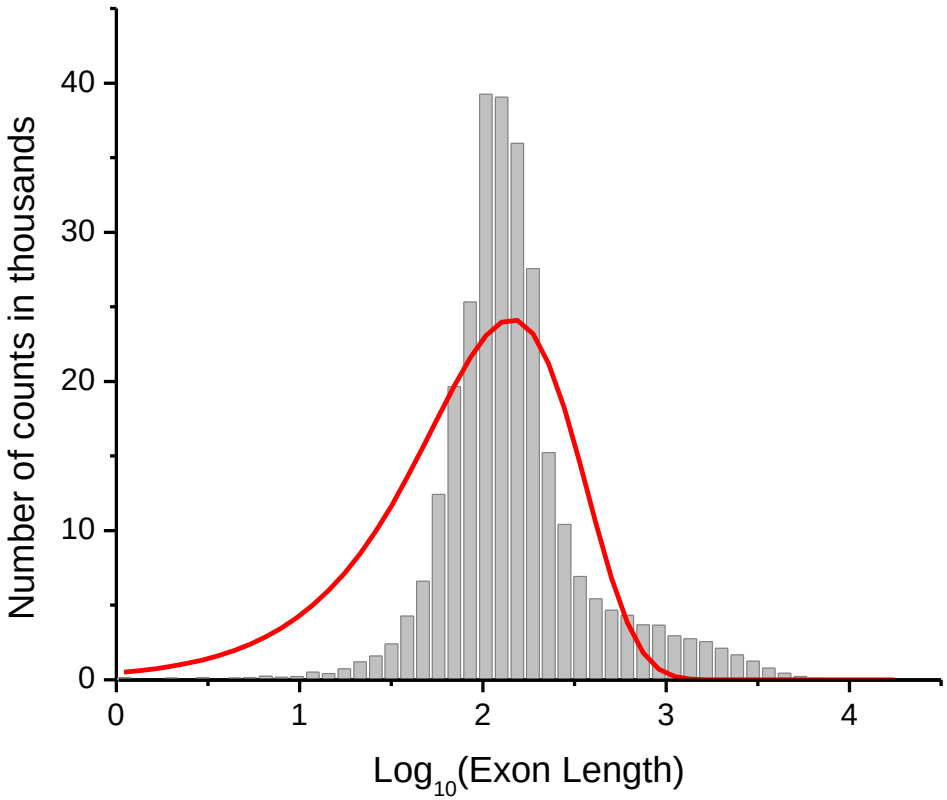
The lengths of intervals between splitting points obey

Exponential distribution

$$\frac{dN}{dE} = n_0 \lambda e^{-\lambda E}$$

Poisson Process

λ = *Const* frequency of splitting



Assumption:

Probability to split DNA at any location and any time is *Constant*

Consequence:

The lengths of intervals between splitting points obey

Exponential distribution

$$\frac{dN}{dE} = n_0 \lambda e^{-\lambda E}$$

or in logarithm scale

$$\frac{dN}{d\log E} = n_0 \lambda e^{\log E - \lambda e^{\log E}}$$

Poisson Process

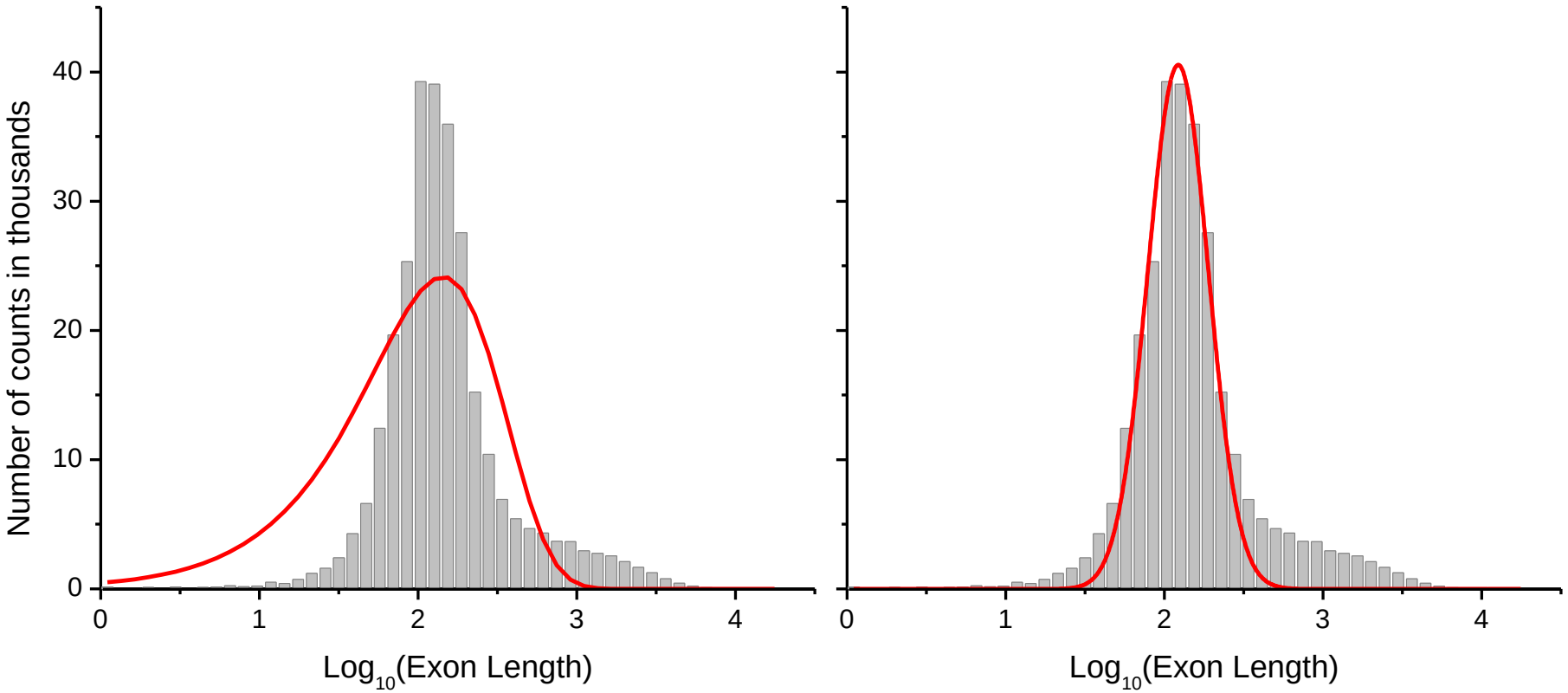
Exponential distribution

$$\frac{dN}{d\log E} = n_0 \lambda e^{\log E - \lambda e^{\log E}}$$

Kolmogoroff Process

Lognormal distribution

$$\frac{dN}{d\log E} = \frac{n_0}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log E - M)^2}{2\sigma^2}}$$



Lognormal distribution

is a consequence of **Central Limit Theorem**

Sum
of random variables

$$\Sigma = \xi_1 + \xi_2 + \xi_3 + \xi_4 + \cdots$$

Normal distribution

$$\frac{dP}{d\Sigma} = \frac{1}{\sqrt{2\pi\sigma_\Sigma^2}} e^{-\frac{(\Sigma - \langle\Sigma\rangle)^2}{2\sigma_\Sigma^2}}$$

Product
of random variables

$$\Pi = \eta_1 \cdot \eta_2 \cdot \eta_3 \cdot \eta_4 \cdot \ldots$$

Lognormal distribution

$$\frac{dP}{d\log\Pi} = \frac{1}{\sqrt{2\pi\sigma_\Pi^2}} e^{-\frac{(\log\Pi - \langle\log\Pi\rangle)^2}{2\sigma_\Pi^2}}$$

$$\log\Pi = \log\eta_1 + \log\eta_2 + \log\eta_3 + \log\eta_4 + \ldots$$

Kolmogoroff process (1941)

Kolmogoroff process (1941)

$i = 0$

$i = 1$

E_0

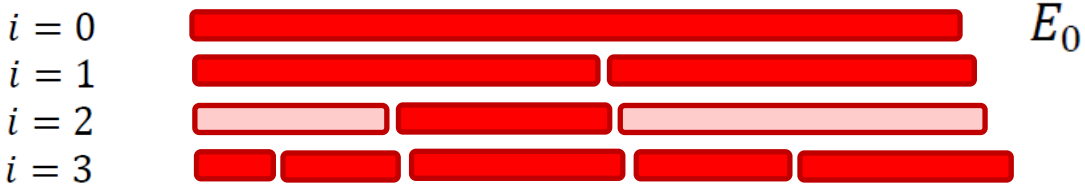
Kolmogoroff process (1941)

E_0

Assumption:

Probability to split any exon is *Independent of exon size*

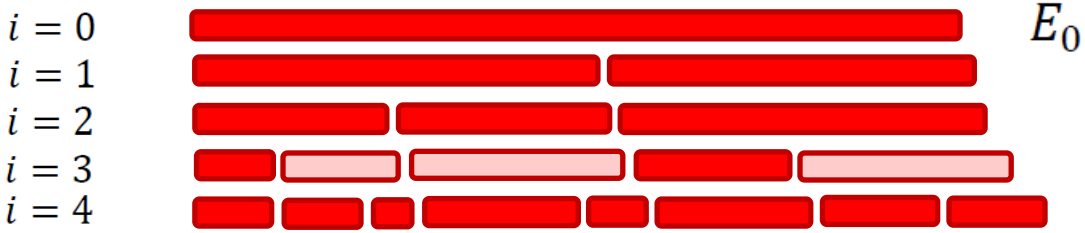
Kolmogoroff process (1941)



Assumption:

Probability to split any exon is *Independent of exon size*

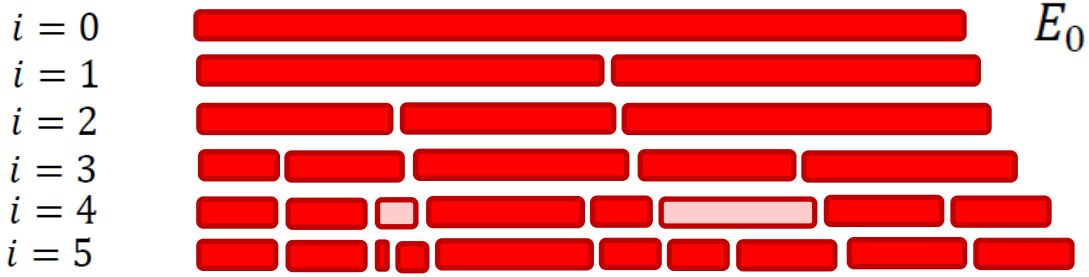
Kolmogoroff process (1941)



Assumption:

Probability to split any exon is *Independent of exon size*

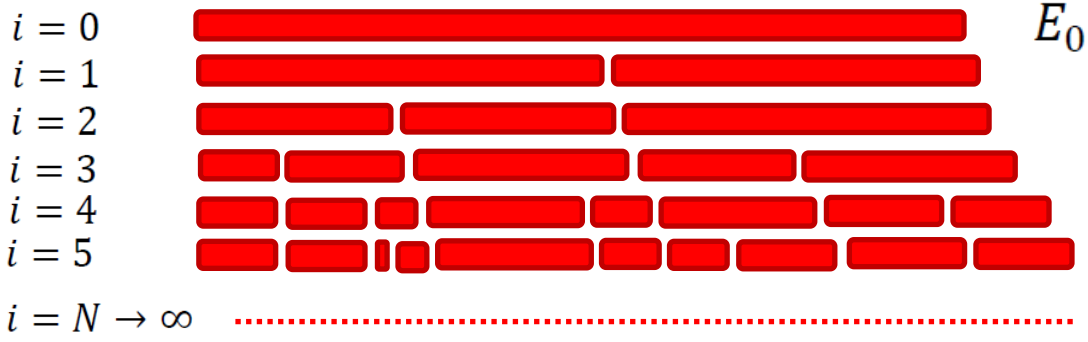
Kolmogoroff process (1941)

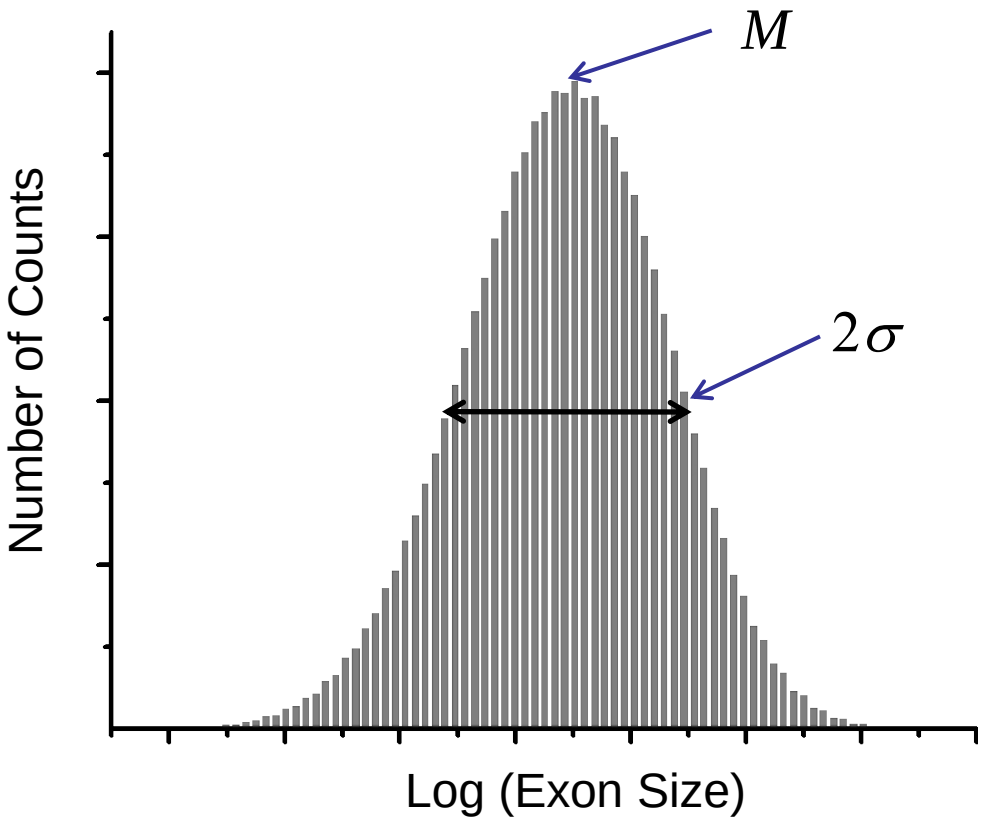


Assumption:

Probability to split any exon is *Independent of exon size*

Kolmogoroff process (1941)





Assumption:

Probability to split any exon is *Independent of exon size*

Consequence:

The lengths of exons obey

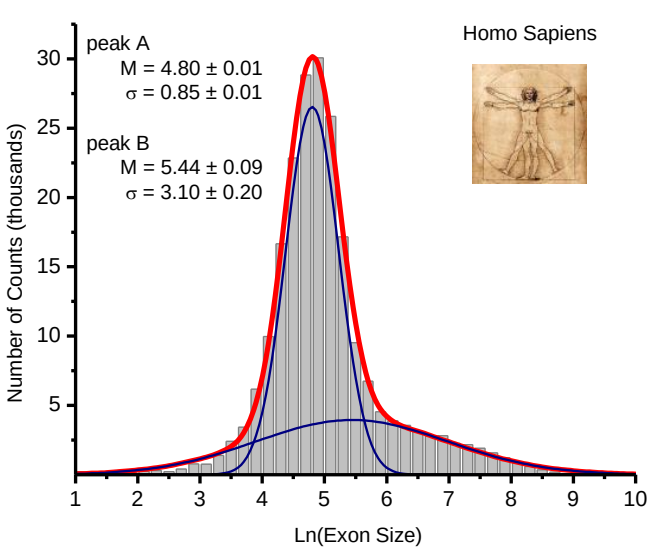
Lognormal distribution

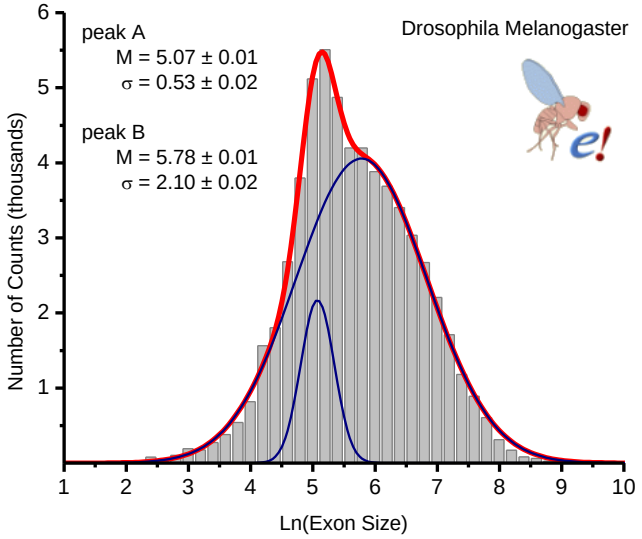
$$\frac{dP}{d\log E} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log E - M)^2}{2\sigma^2}}$$

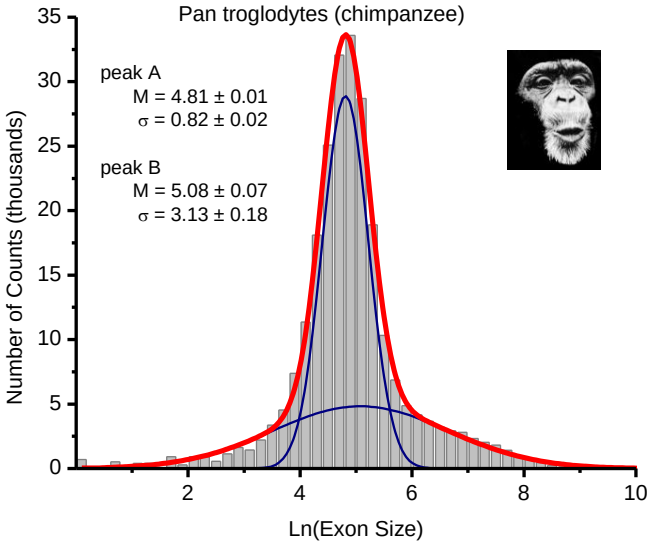
$$\log E_0 - M \sim N$$

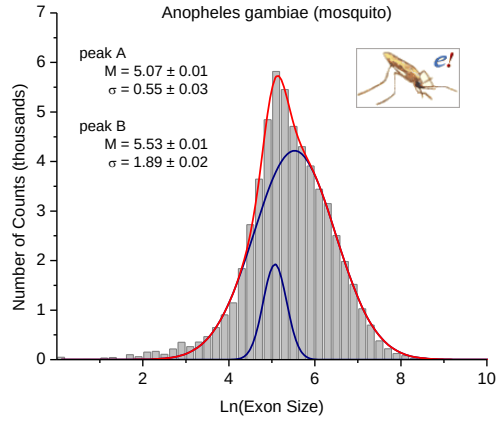
$$\sigma^2 \sim N$$

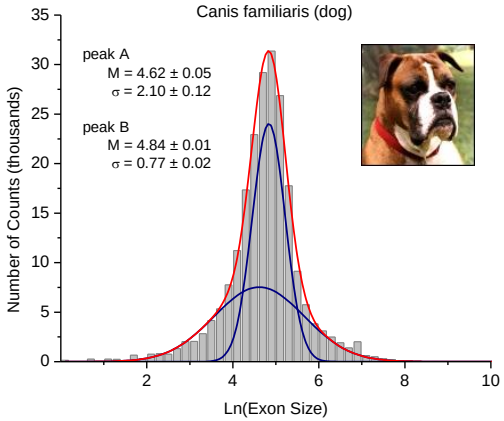
Real Genomes: Two Lognormal Peaks

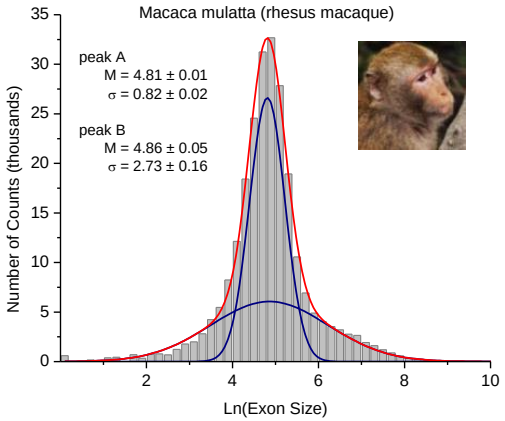


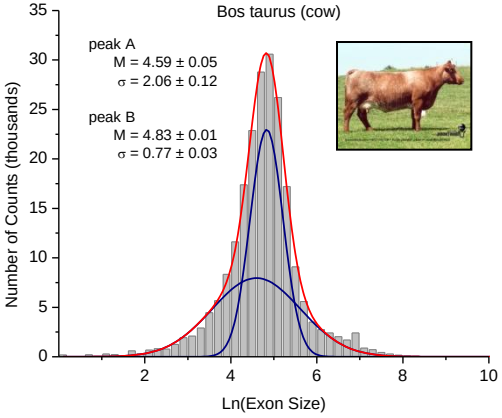


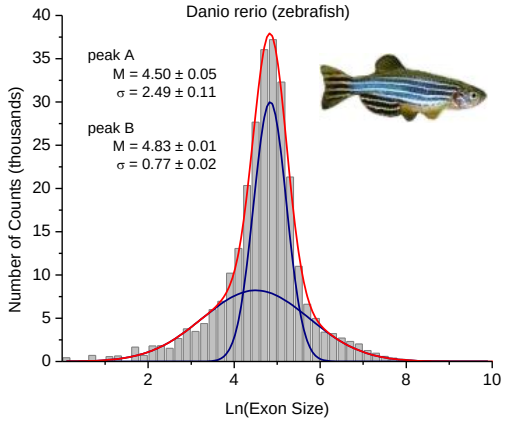


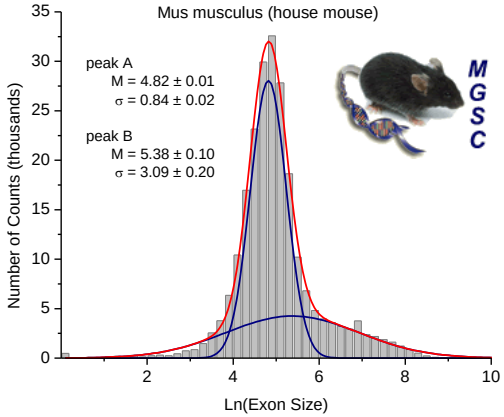

Anopheles gambiae (mosquito)
peak A
M = 5.07 ± 0.01
σ = 0.55 ± 0.03
peak B
M = 5.53 ± 0.01
σ = 1.89 ± 0.02
Number of Counts (thousands)
Ln(Exon Size)

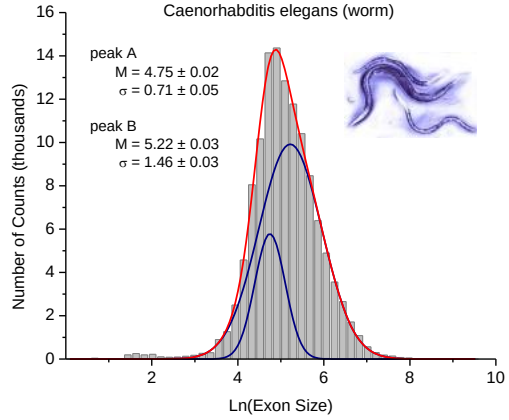

Canis familiaris (dog)
peak A
M = 4.62 ± 0.05
σ = 2.10 ± 0.12
peak B
M = 4.84 ± 0.01
σ = 0.77 ± 0.02
Number of Counts (thousands)
Ln(Exon Size)

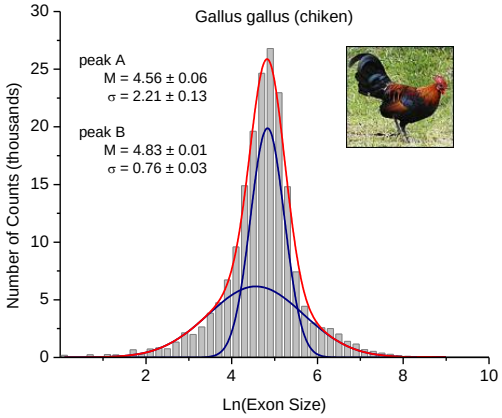

Macaca mulatta (rhesus macaque)
peak A
M = 4.81 ± 0.01
σ = 0.82 ± 0.02
peak B
M = 4.86 ± 0.05
σ = 2.73 ± 0.16
Number of Counts (thousands)
Ln(Exon Size)

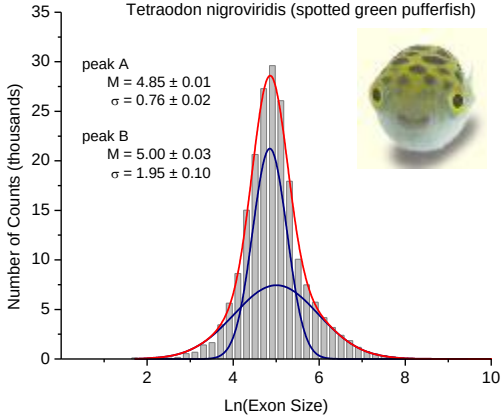

Bos taurus (cow)
peak A
M = 4.59 ± 0.05
σ = 2.06 ± 0.12
peak B
M = 4.83 ± 0.01
σ = 0.77 ± 0.03
Number of Counts (thousands)
Ln(Exon Size)


Danio rerio (zebrafish)
peak A
M = 4.50 ± 0.05
σ = 2.49 ± 0.11
peak B
M = 4.83 ± 0.01
σ = 0.77 ± 0.02
Number of Counts (thousands)
Ln(Exon Size)

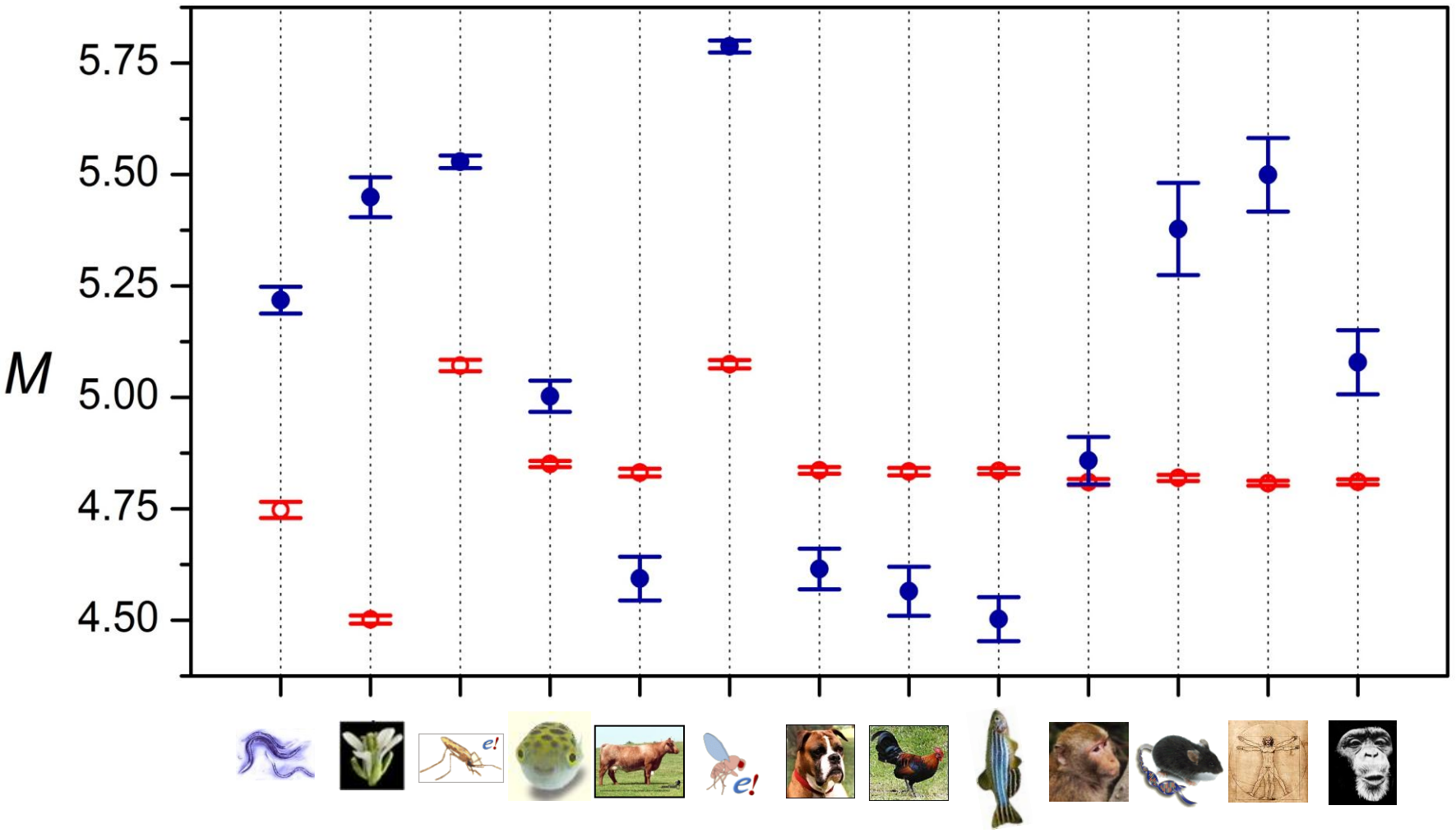

Mus musculus (house mouse)
peak A
M = 4.82 ± 0.01
σ = 0.84 ± 0.02
peak B
M = 5.38 ± 0.10
σ = 3.09 ± 0.20
MGSC
Number of Counts (thousands)
Ln(Exon Size)


Caenorhabditis elegans (worm)
peak A
M = 4.75 ± 0.02
σ = 0.71 ± 0.05
peak B
M = 5.22 ± 0.03
σ = 1.46 ± 0.03
Number of Counts (thousands)
Ln(Exon Size)


Gallus gallus (chiken)
peak A
M = 4.56 ± 0.06
σ = 2.21 ± 0.13
peak B
M = 4.83 ± 0.01
σ = 0.76 ± 0.03
Number of Counts (thousands)
Ln(Exon Size)

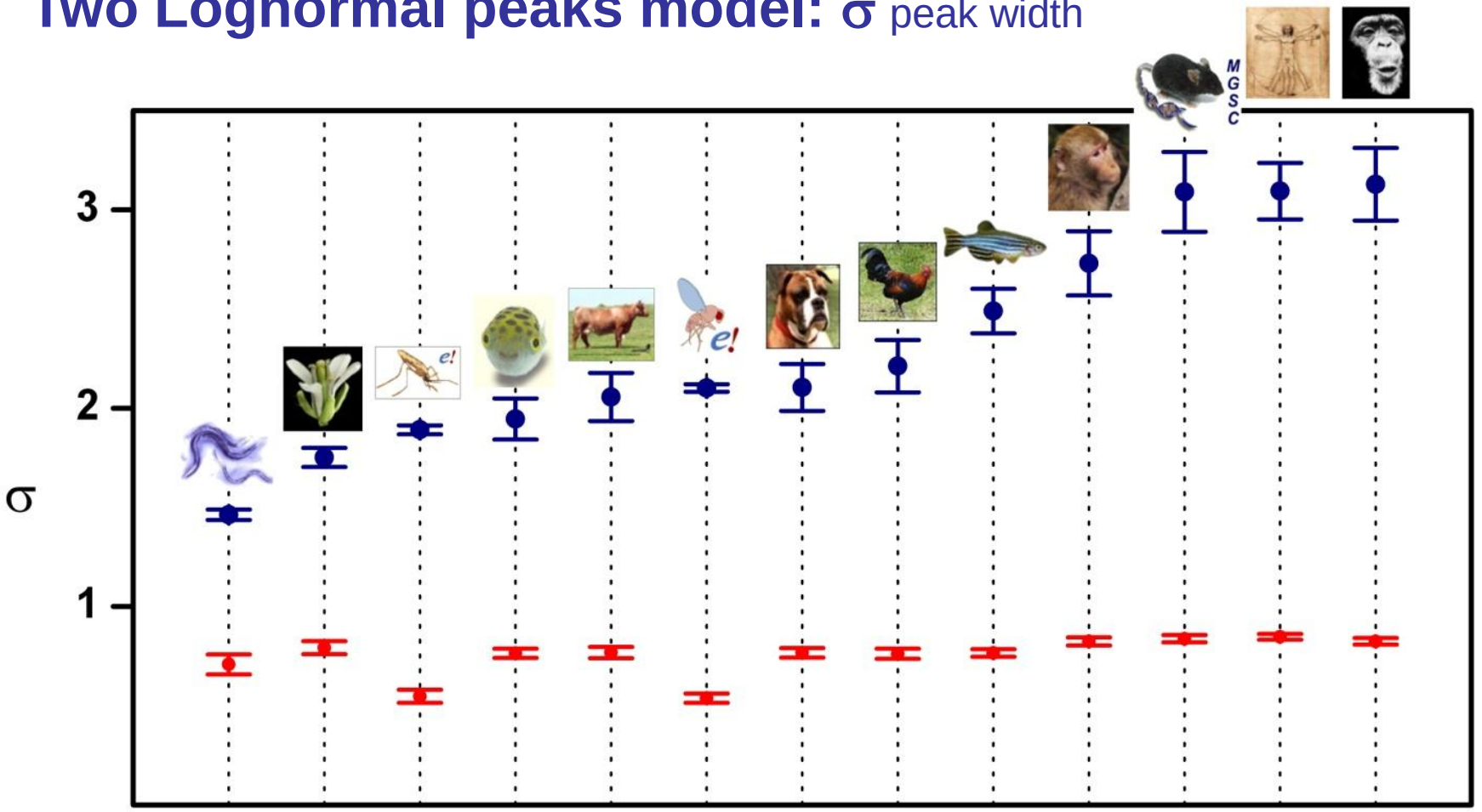

Tetraodon nigroviridis (spotted green pufferfish)
peak A
M = 4.85 ± 0.01
σ = 0.76 ± 0.02
peak B
M = 5.00 ± 0.03
σ = 1.95 ± 0.10
Number of Counts (thousands)
Ln(Exon Size)

Parameters of
Two Lognormal peaks model: *M* location of peak maximum



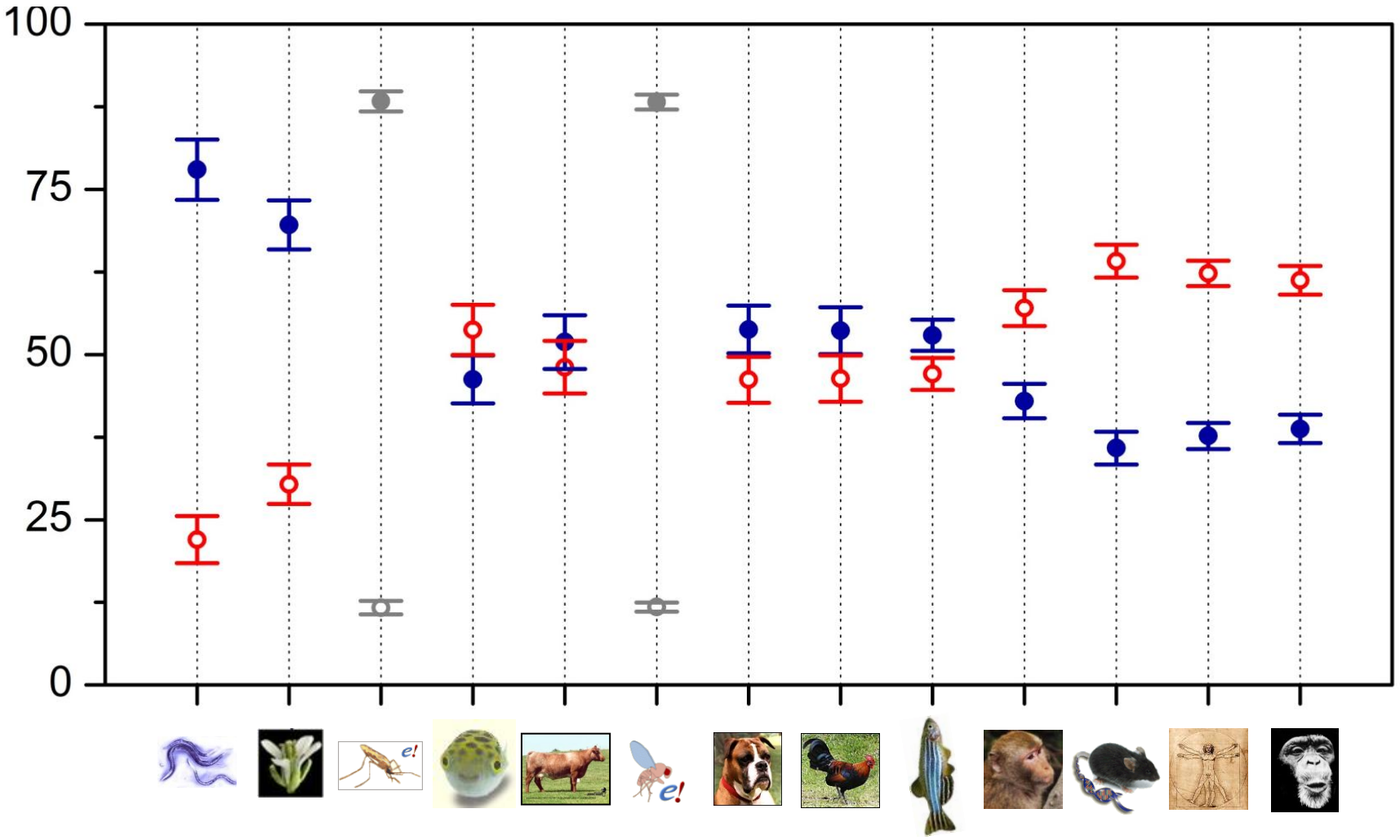
Ryabov & Gribskov, *Nucleic Acids Research*, **36**, 2756-2763 (2008)

Parameters of
Two Lognormal peaks model: σ peak width



Ryabov & Gribskov, *Nucleic Acids Research*, **36**, 2756-2763 (2008)

Parameters of
Two Lognormal peaks model: peak area Narrow and Wide, %



Ryabov & Gribskov, *Nucleic Acids Research*, **36**, 2756-2763 (2008)

Summary of Observed Facts

- Exon lengths distributions of studied eukaryotic genomes can be fitted by Two Lognormal Distributions
- Parameters of those two peaks follow two distinctive patterns: changes of peak width and relative peak area correlate with complexity of species.
- This may indicate presence of two different classes of exons with different evolutionary histories

How we can model exon size distributions of real genomes ?

Growth of total genome length

Duplications in genome code

Merging exons together (intron loss)

Exon loss

Parameters of Kolmogoroff Process

as model parameters for elementary “exon modification” event

If any exon of length l_i has probability p to be modified during time interval Δt

and

$Q(k)$ average number of exons with sizes $l_{i+1} \leq k\, l_i$

$$\int_0^\infty |\ln(k)|^3 dQ(k) < \infty$$

with

$$A = \frac{1}{Q(\infty)} \int_0^\infty \ln(k)\, dQ(k)$$

$$B^2 = \frac{1}{Q(\infty)} \int_0^\infty (\ln(k) - A)^2\, dQ(k)$$

Then the process converges to a lognormal distribution with

$$M(t) = M_0 + A \frac{p\, t}{\Delta t}$$

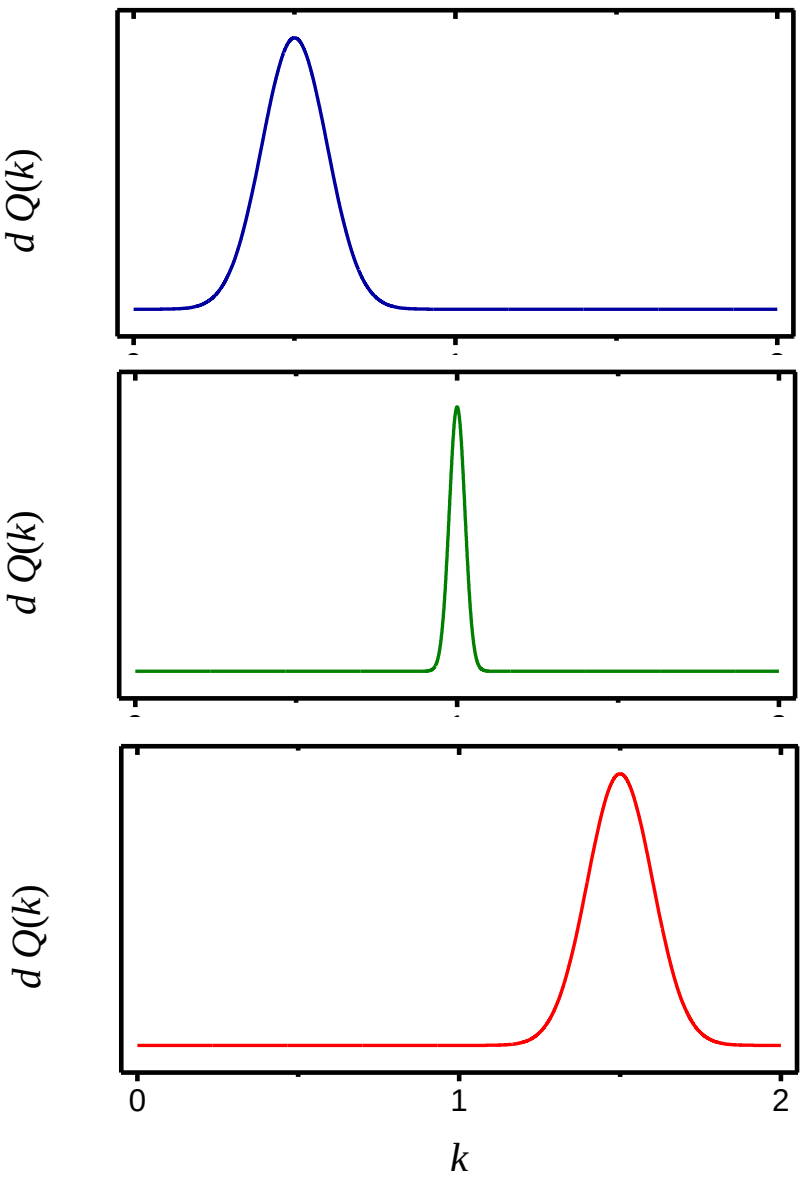
peak position

$$\sigma^2(t) = \sigma_0^2 + B^2 \frac{p\, t}{\Delta t}$$

peak width

Parameters of Q(k) function

$dQ(k)$ distribution



$A < 0$

$$A = \frac{1}{Q(\infty)} \int_0^{\infty} \ln(k)\, dQ(k)$$

Exons splitting

(decreasing Exon length)

$A \approx 0$

Exons duplicating

$A > 0$

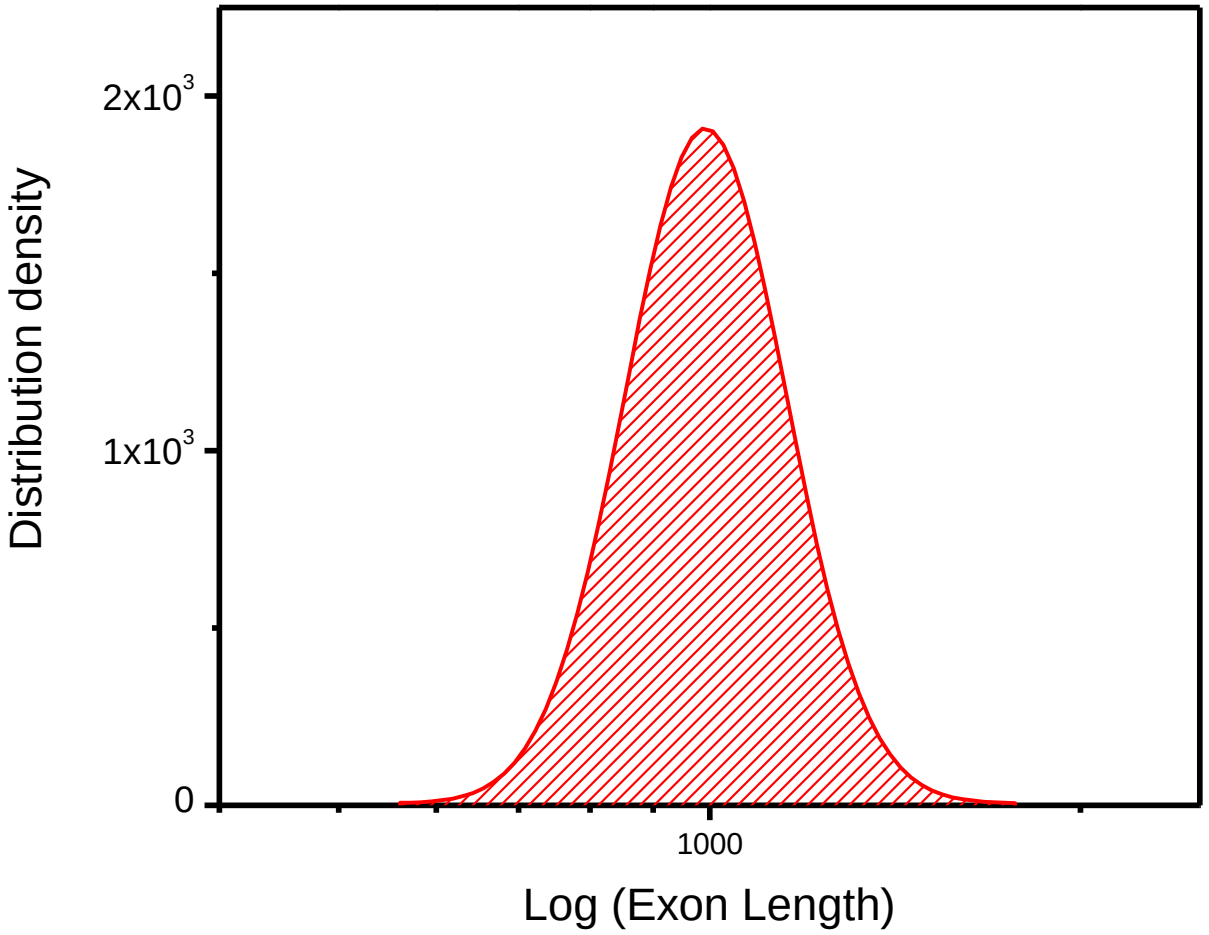
Increasing Exon length

Modeling exon size distributions in real genomes

Initial exon size distribution

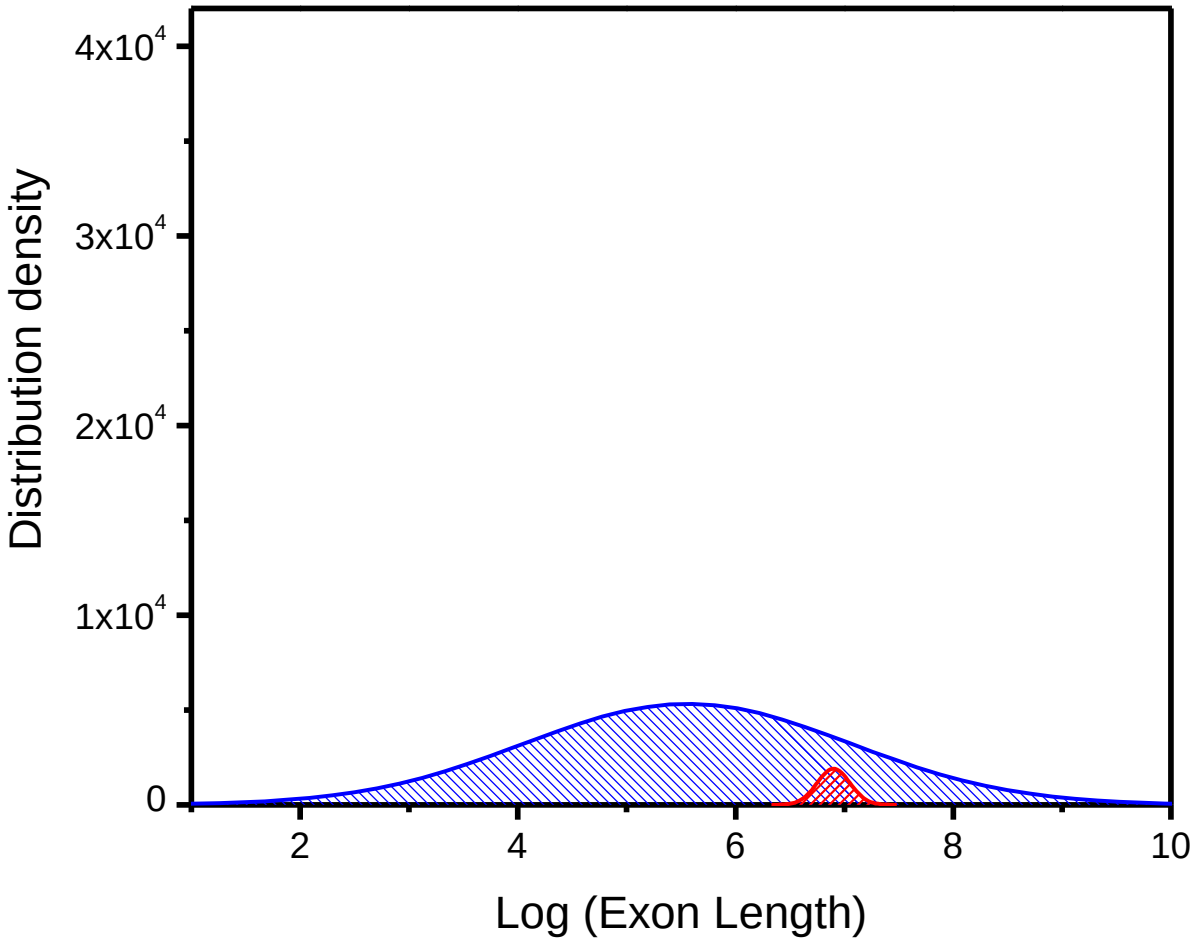
A mockup of a bacterial genome with 500 exons

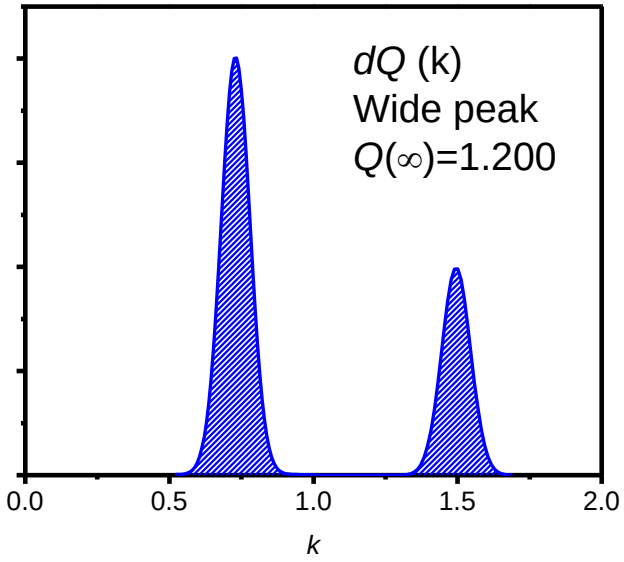
Having 1000 bp mean exons length



Modeling exon size distributions in real genomes

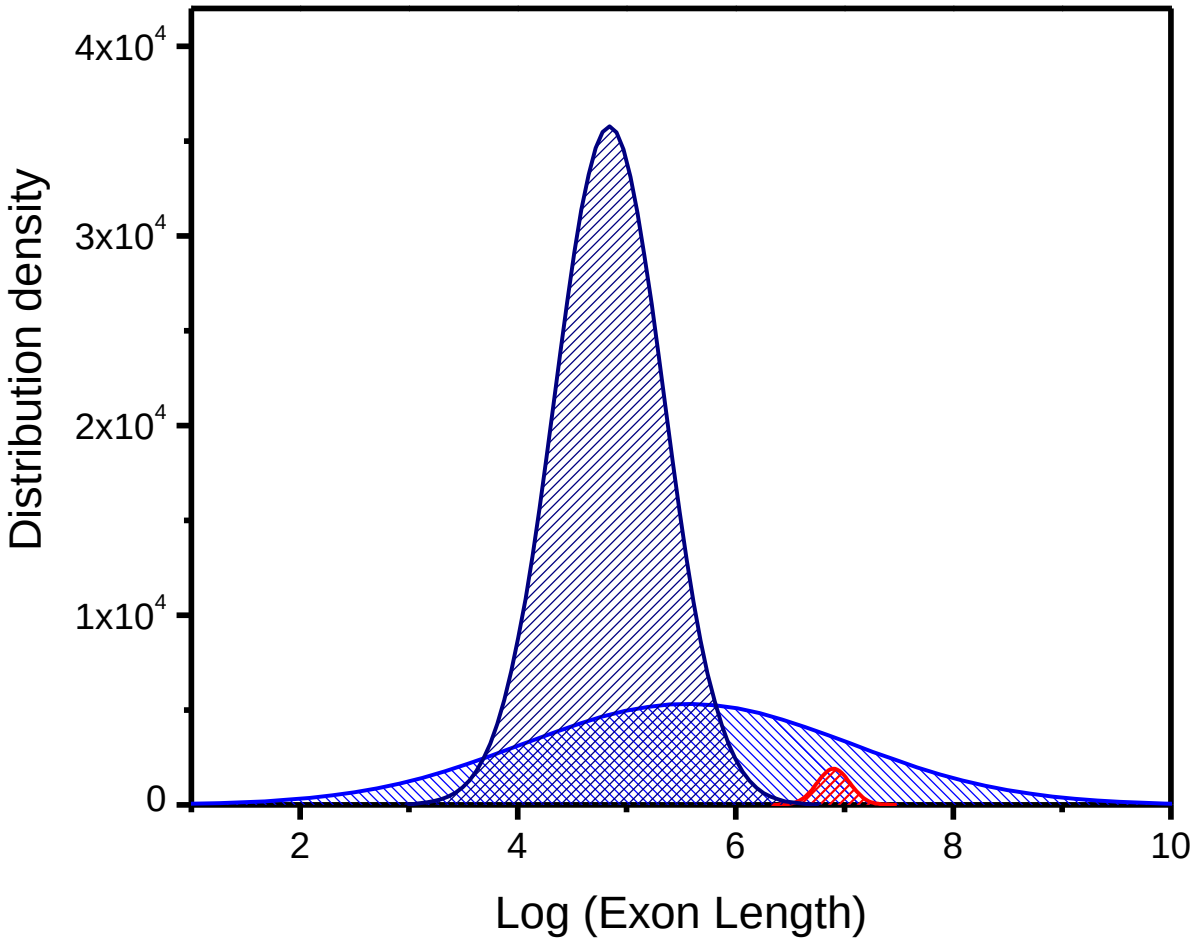
15 000 time steps for p=0.001

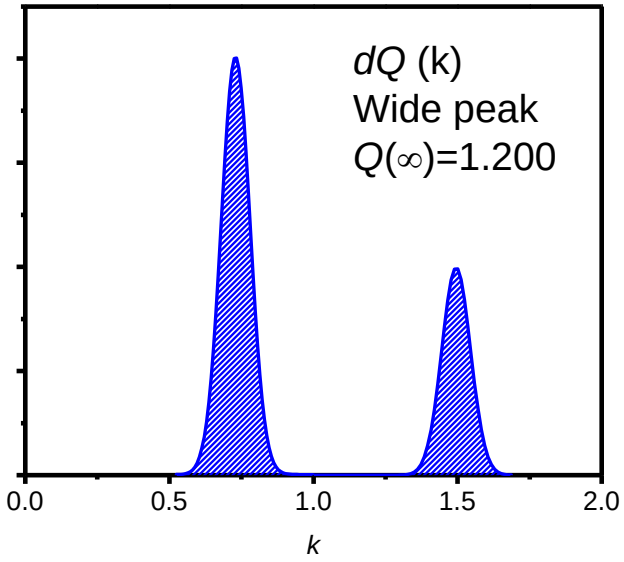


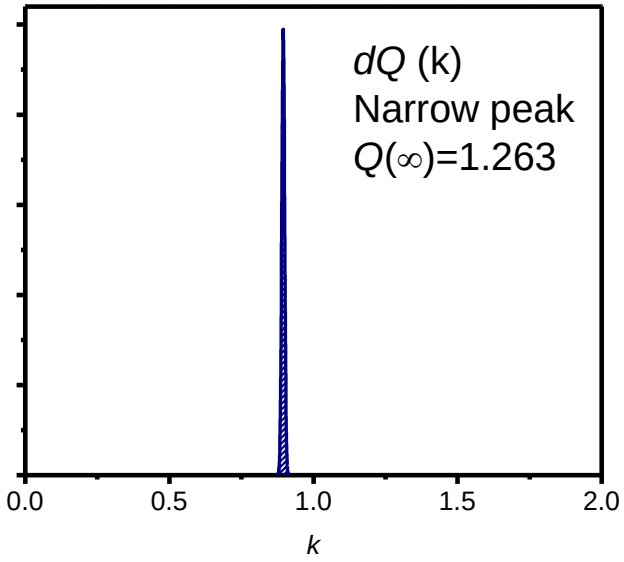


Modeling exon size distributions in real genomes

15 000 time steps for p=0.001

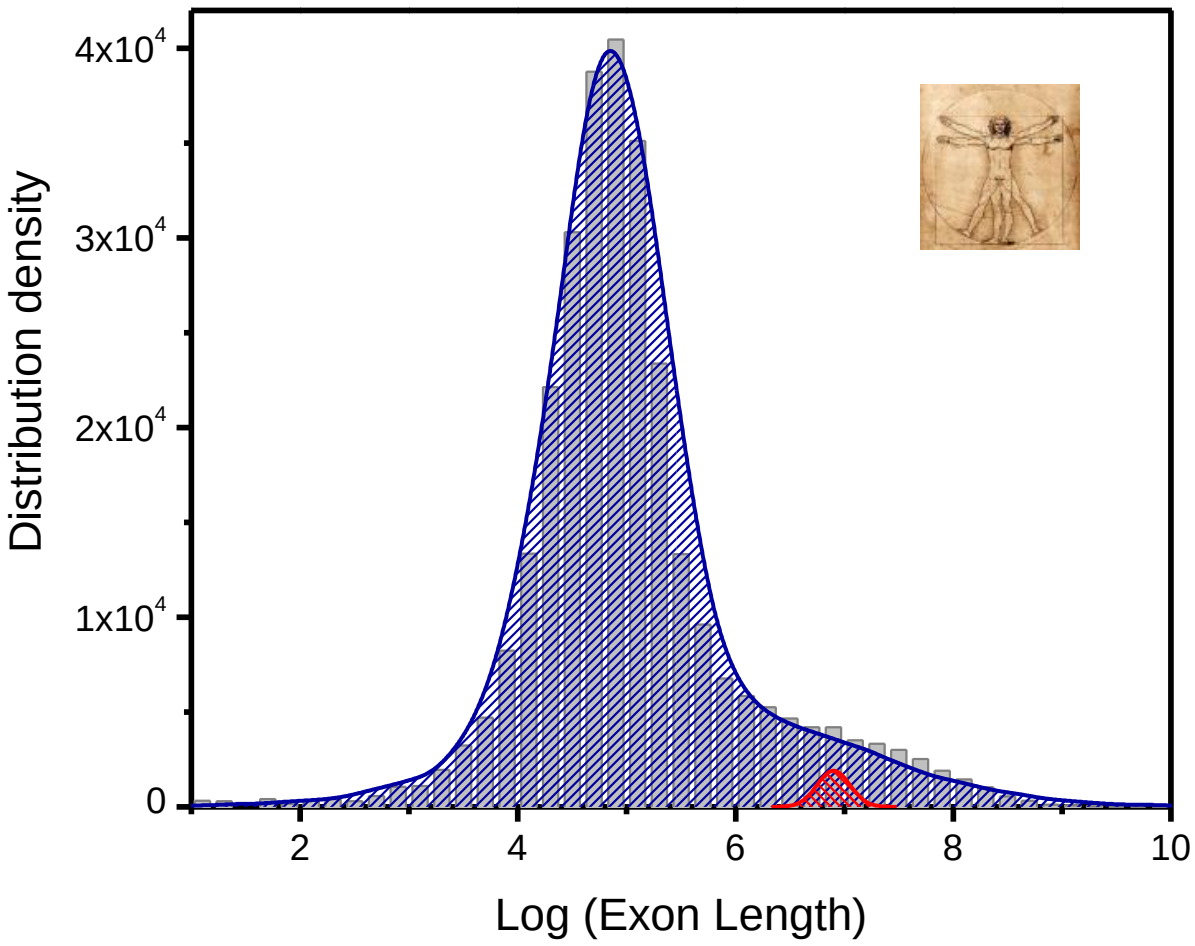


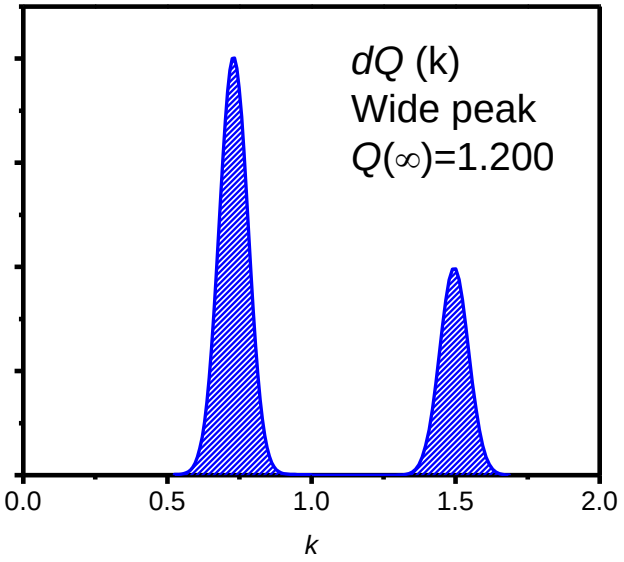


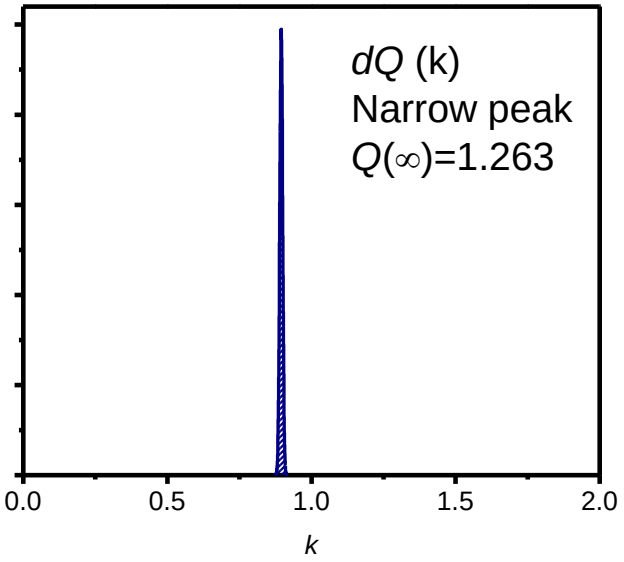


Modeling exon size distributions in real genomes

15 000 time steps for p=0.001

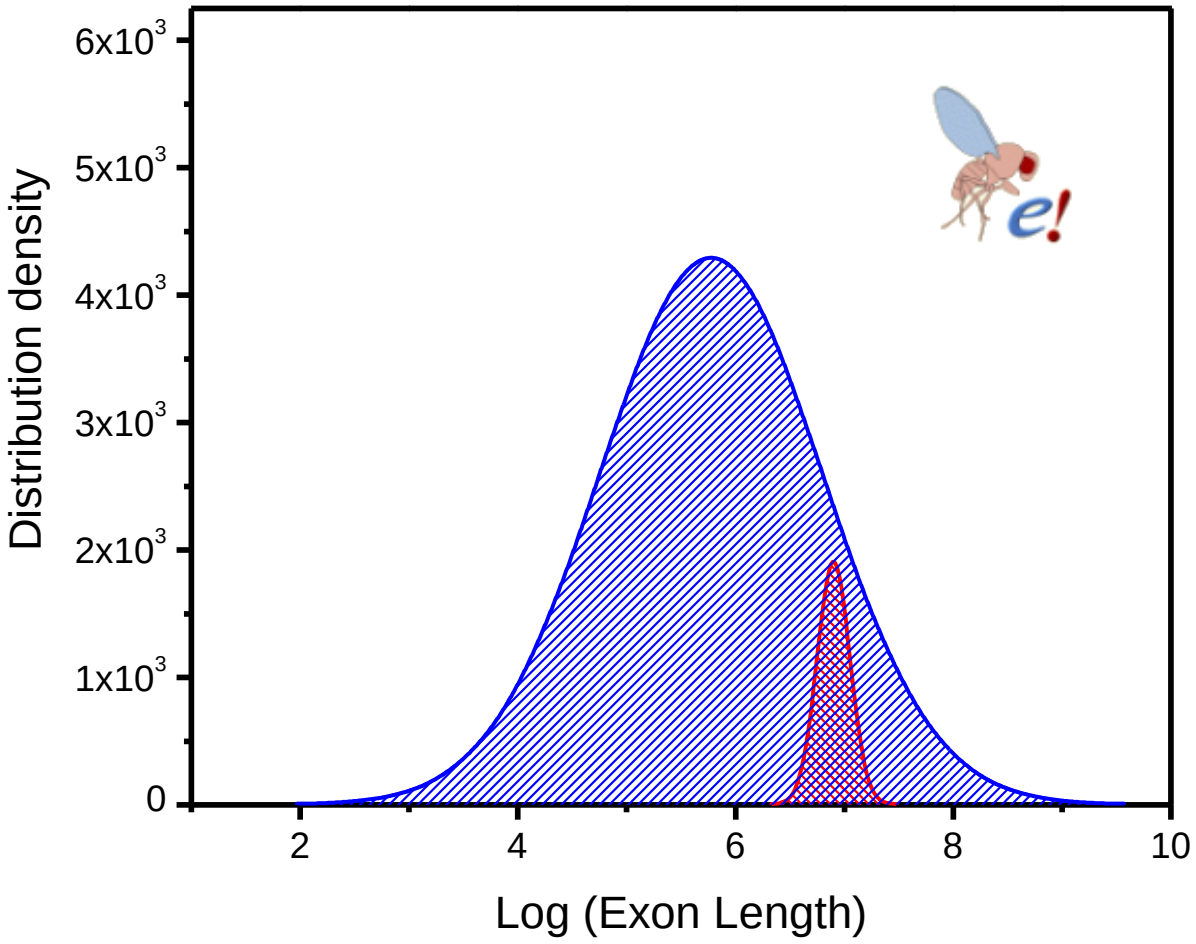


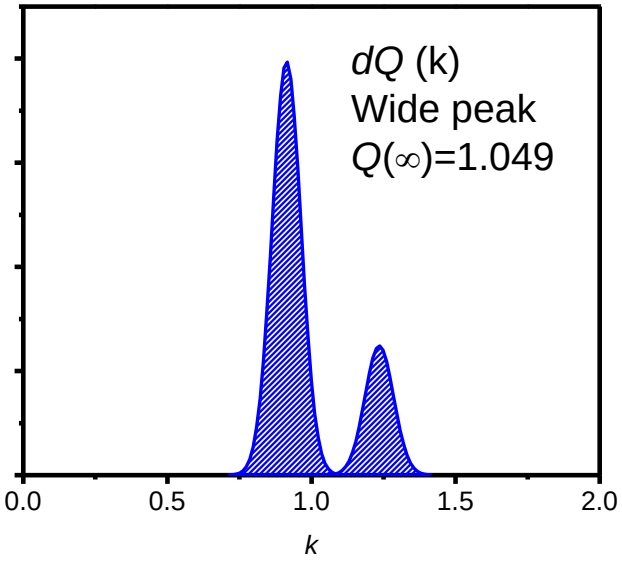




Modeling exon size distributions in real genomes

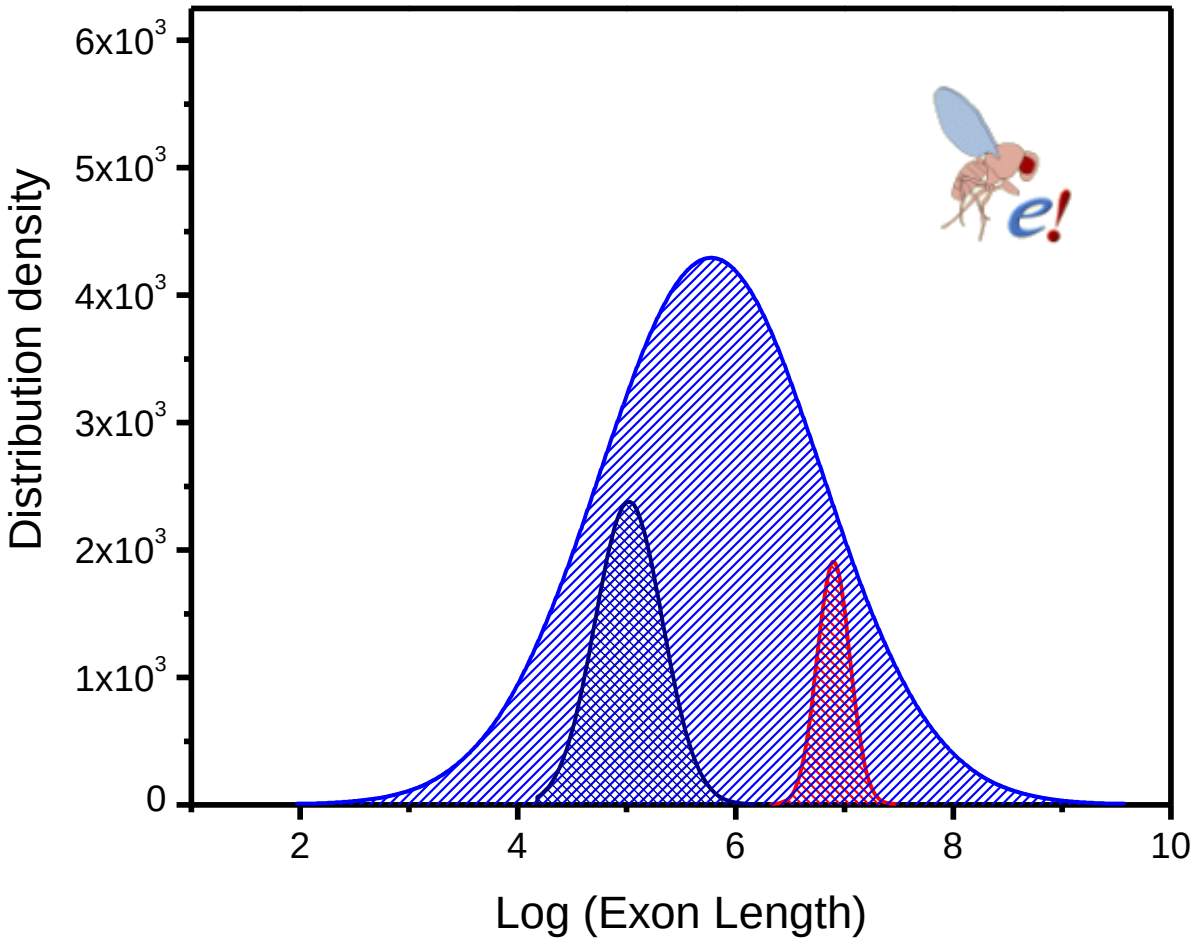
50 000 time steps for p=0.001

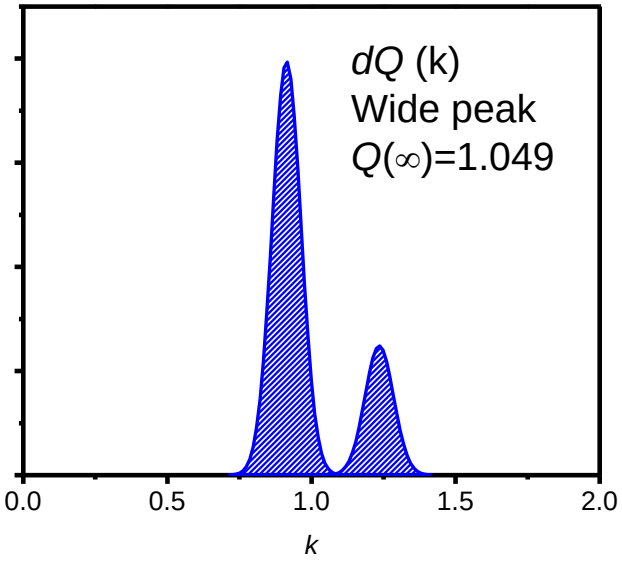


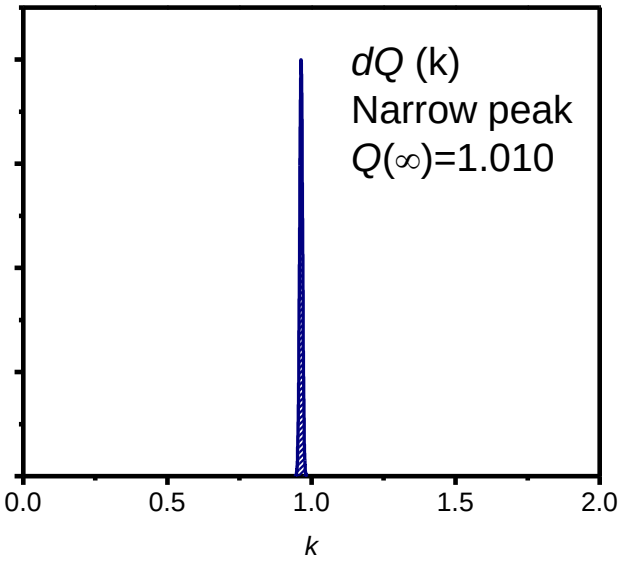


Modeling exon size distributions in real genomes

50 000 time steps for p=0.001

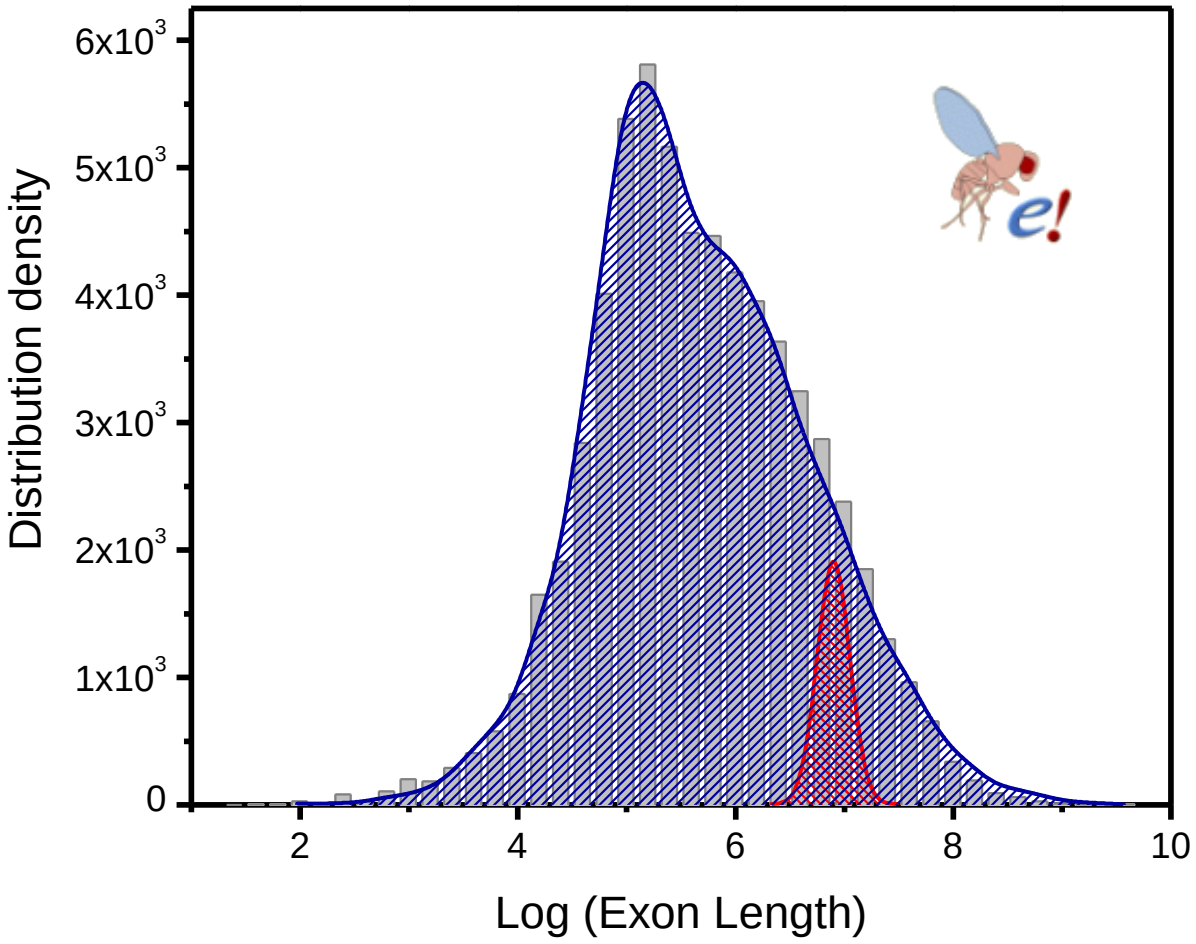


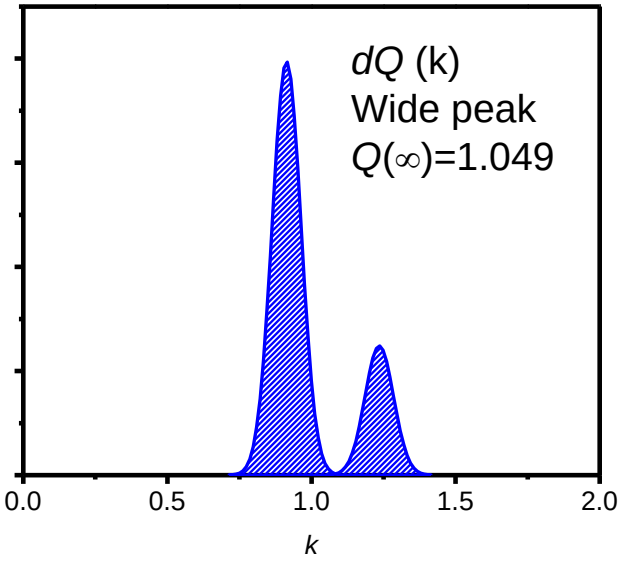


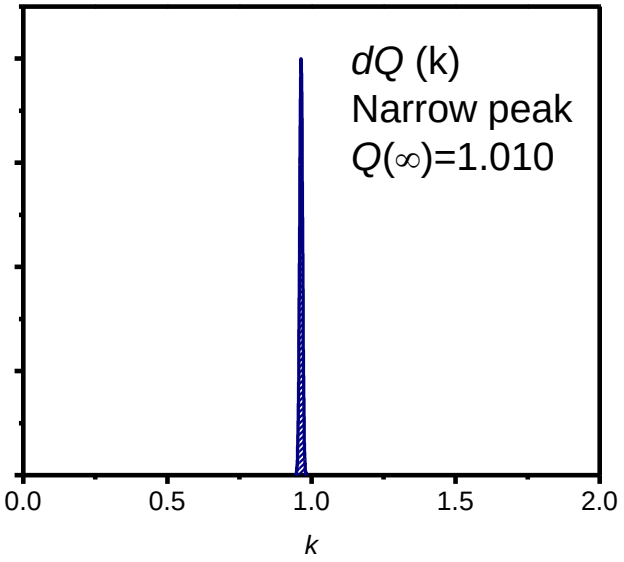


Modeling exon size distributions in real genomes

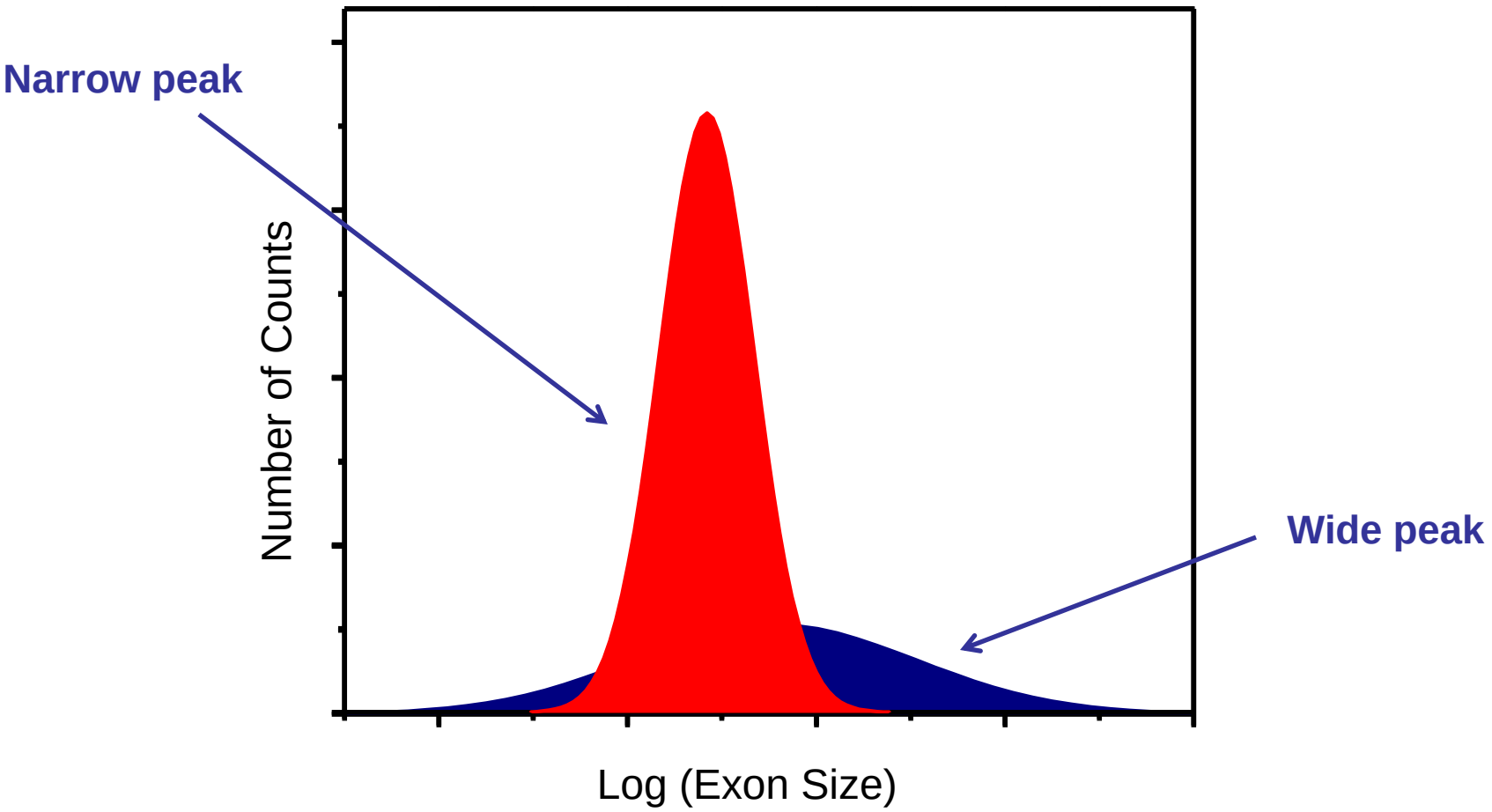
50 000 time steps for p=0.001





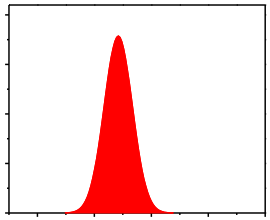


What cloud be the biological reason for two exons peaks?



What could be the biological reason for two exons peaks?

Narrow peak

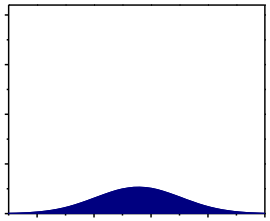


Holds approximately
Constant position

Has approximately
Constant width

Has greater relative
occupation for
Complex organisms

Wide peak

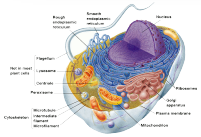


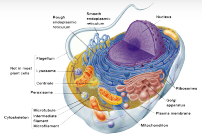
Changes position

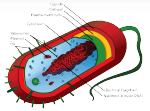
Width increases with
increasing complexity
of and organism

Has greater relative
occupation for
Simple organisms

Two ways of Gene Expression Regulation





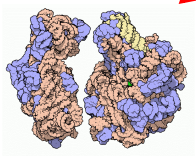


Alterative Splicing

Introns

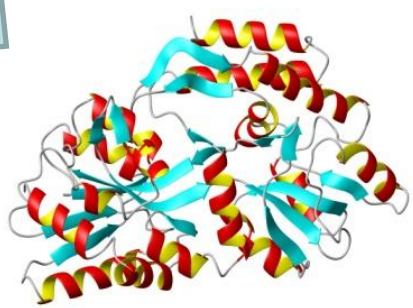
mRNA

Untranslated regions of mRNA

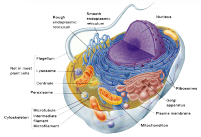


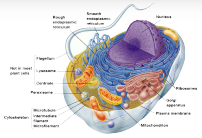
DNA

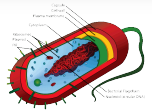
Protein



Two ways of Gene Expression Regulation



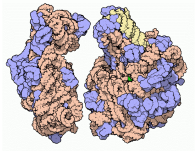




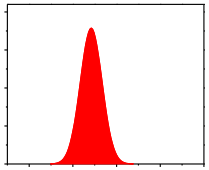
mRNA

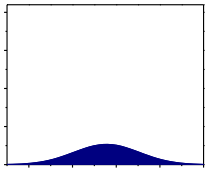
Untranslated regions of mRNA

Introns

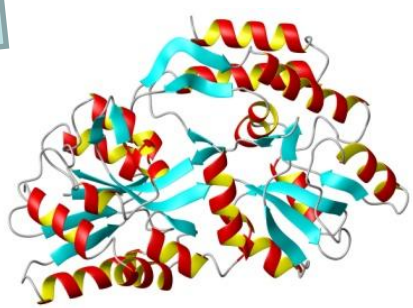


DNA





Protein



Conclusions

- Analysis of global properties of eukaryotic genomes reveals two distinct peaks in statistical distribution of exon sizes

- The observed peak could be fitted by a sum of two lognormal distributions which may imply that they originated by two different exon splitting pathways described in the general frameworks of Kolmogoroff splitting process

- Two observed peaks of exons could be correlated with the phenomenon of alternative splicing and with exons contributing into untranslated regions of mRNA. This suggests that the observed separation of exons in two different classes may be originated from two different ways of protein expression regulation.

Acknowledgments

Michael Gribskov

Alexander Berezhkovskii