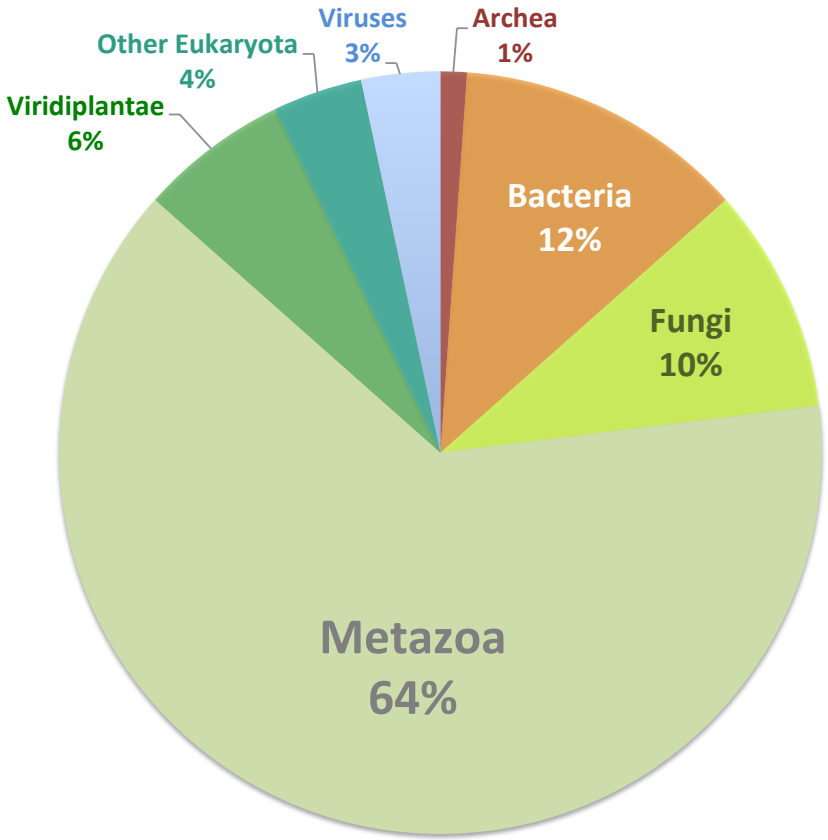
Evolution of proteins with repeats

Maria Anisimova

Computational Biochemistry Research Group

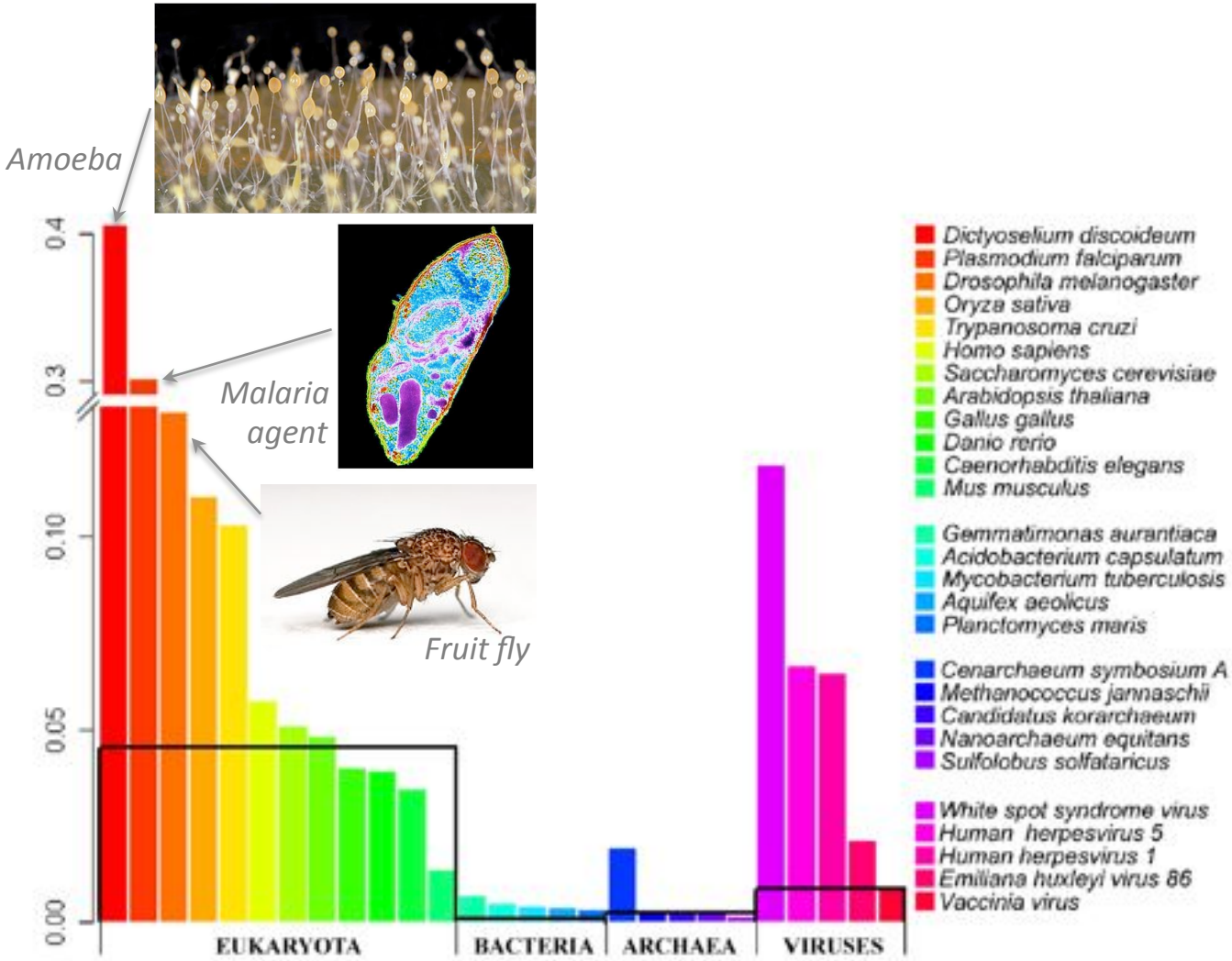
ETH Zurich

Proteins with amino acid motifs arranged in tandem are found in all three kingdoms of life



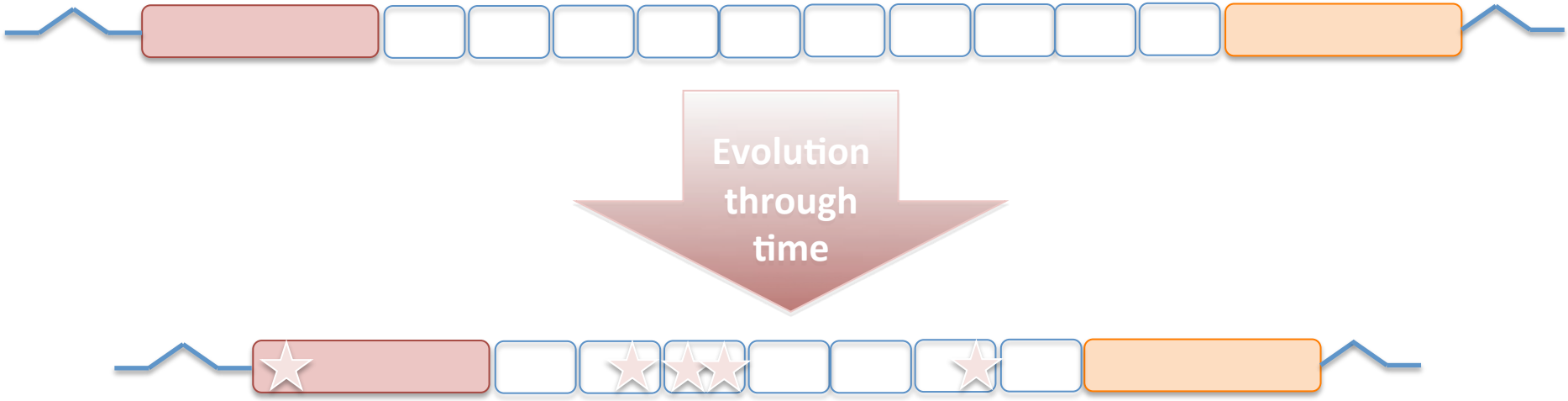
Distribution in SwissProt (based on Andrade et al. 2001)

Frequency of proteins with homorepeats: NCBI RefSeq



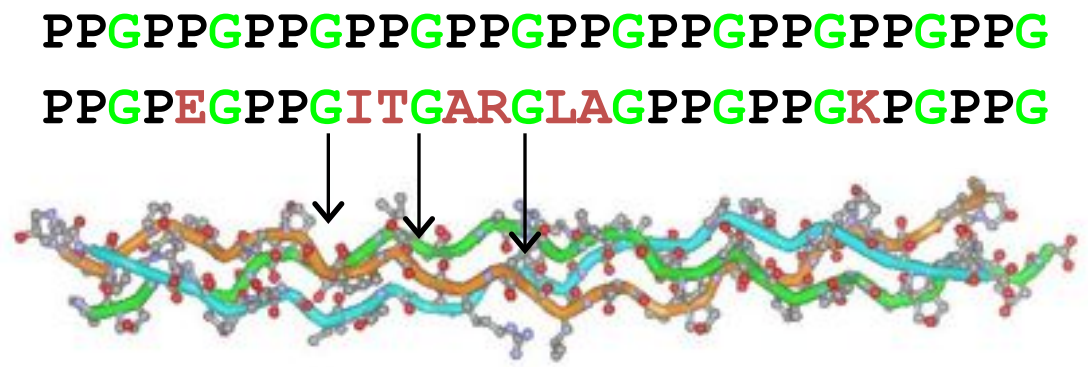
from Jorda and Kajava 2010

> 14% of all proteins contain tandem repeats (TR)
Marcotte et al. 1999



Perfect (more recent) repeats are easier to detect

Example 1: Collagen



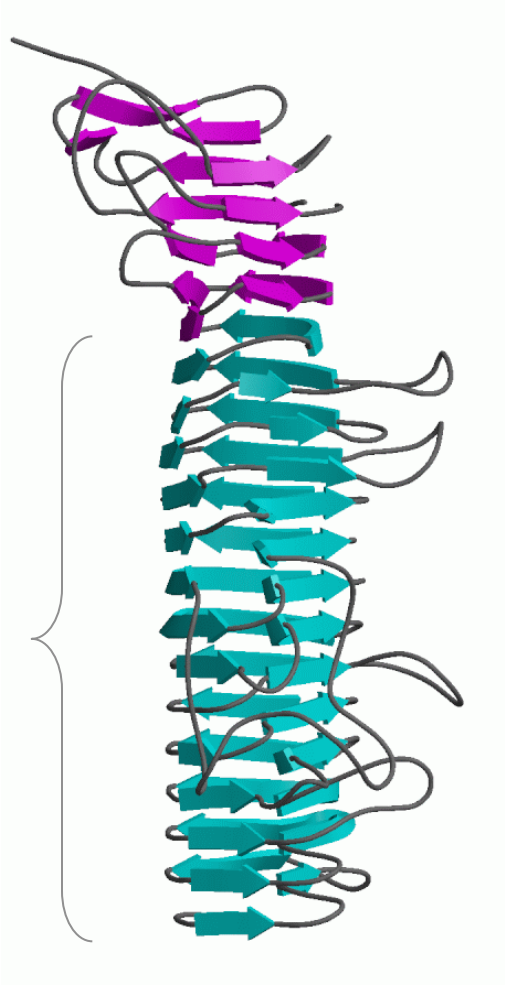
Main component of connective tissue
Most abundant protein in mammals
Only found in animals
~ 30% of body proteins

Example 2: Pertactin from *Bordetella pertussis*

Protective immunity to Bordetella infections

Includes to two TR regions

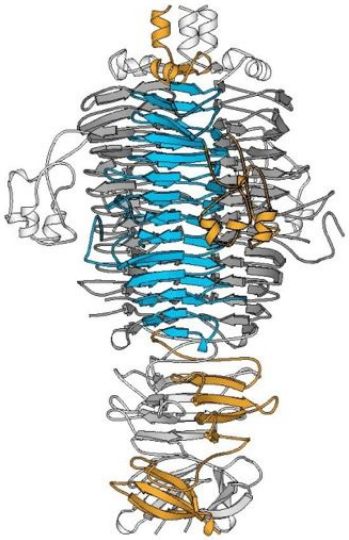
```
  GILLENPAAELQFRNGSVTSSGQLSDDGIRRFLG
     TVTVKAGKLVADHATLANVGDTWDDDGI
       ALYVAGEQAQASIADSTLQGAG
        GVQIERGANVTVQRSAIVDG
GLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPA
      AVSVLGASELTLDGGHITGGRAA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWY
     GVDVSGSSVELAQSIVEAPELGA
     AIRVGRGARVTVSGGSLSAPHGN
VIETGGARRFAPQAAPLSITLQAGAHAQGKA
       LLYRVLPEPVKLTLTGGADAQG
    DIVATELPSIPGTSIGPLDVALASQARWTG
```



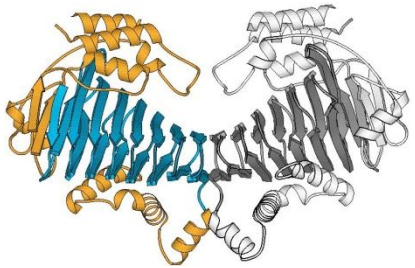
Example 3: β-solenoid proteins

Pectate lyase C

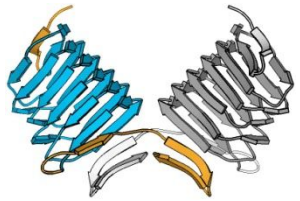
P.69 pertactin



Tailspike endorhamnosidase

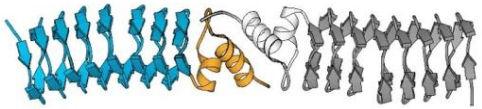


Stabilizer of iron transporter SufD

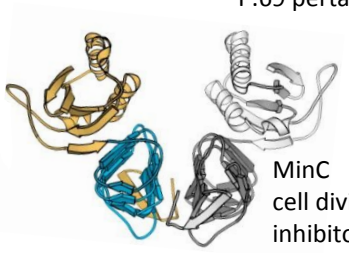


Cyclase-associated protein

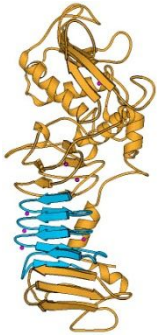
Antifreeze protein



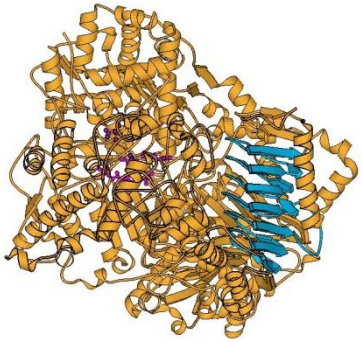
MfpA inhibitor of DNA gyrase



MinC cell division inhibitor

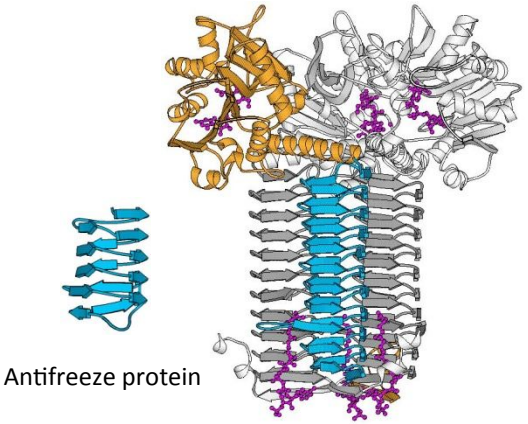


PrtC protease C



Glutamate synthase

YadA adhesin



Antifreeze protein

N-acetyl-glucosamine 1-phosphate uridyltransferase

Kajava and Steven 2006

Example 4: Homorepeats

Single-amino acid repeats

Local concentration of physico-chemical property

Strong potential for forming molecular interactions

(polyA and polyG important for extracellular structural properties; polyH, polyK, polyG proteins used in biotechnology)

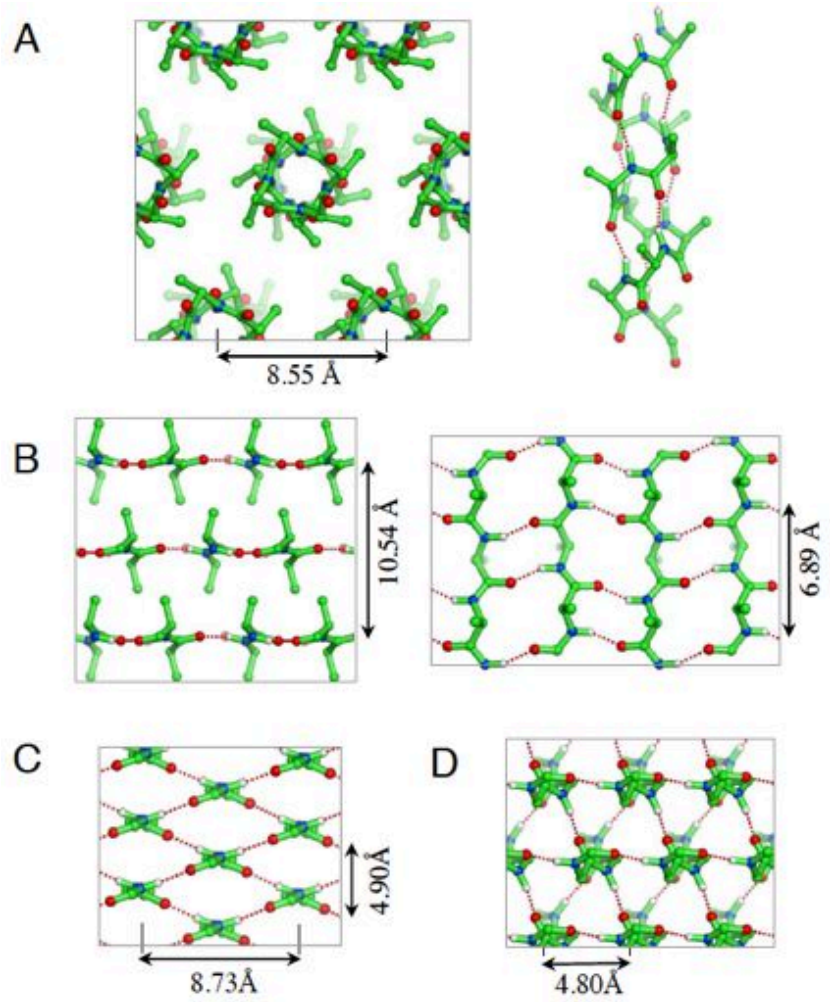
High length polymorphism

> 40% associated with disease

polyQ → neurodegenerative disorders (Huntington's disease)

polyA → epilepsy, mental retardation, etc.

Homorepeats: Types of crystal structures



Axial projections:

(A) PolyA α-helixes

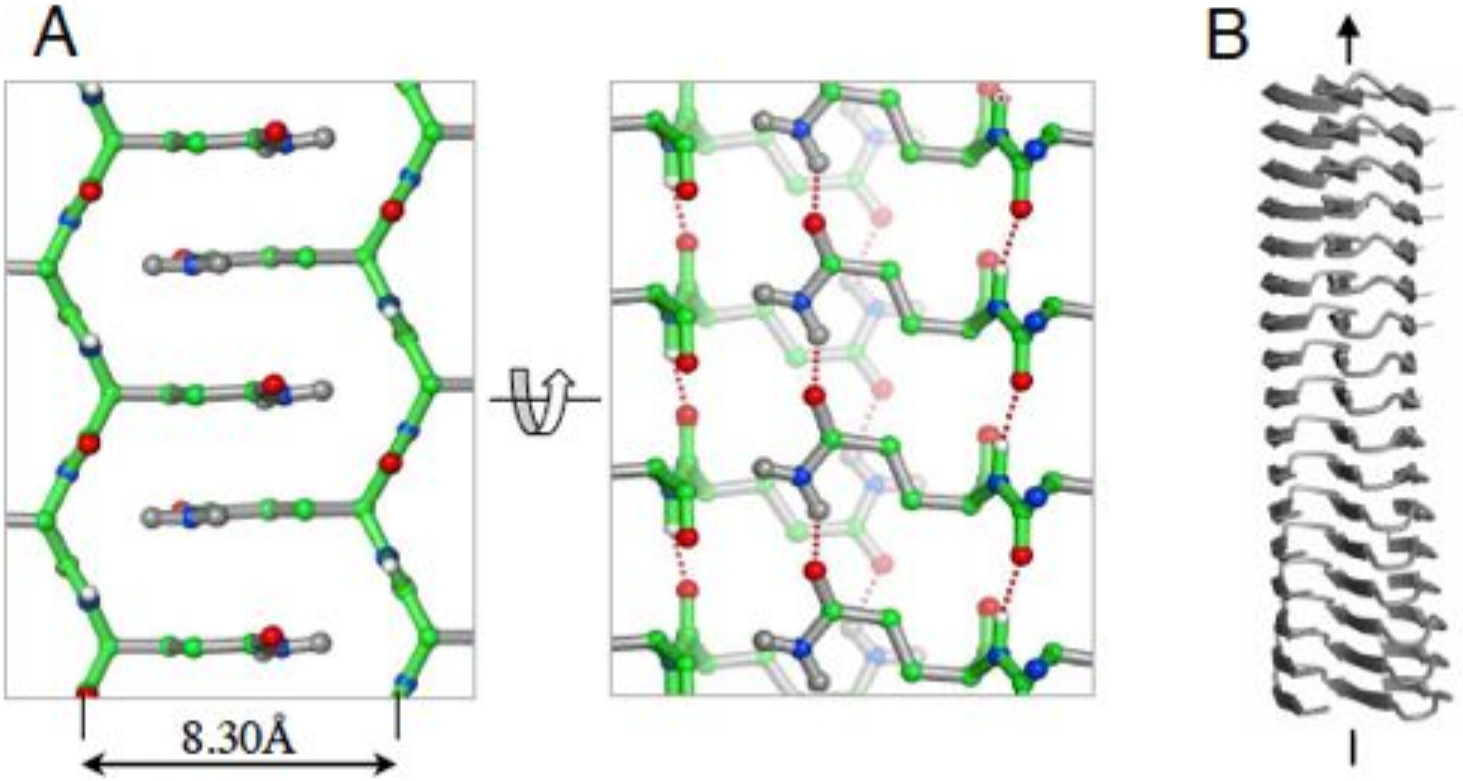
(B) PolyA: Antiparallel β-structure

(C) PolyG form I

(D) PolyG form II

from Jorda and Kajava 2010

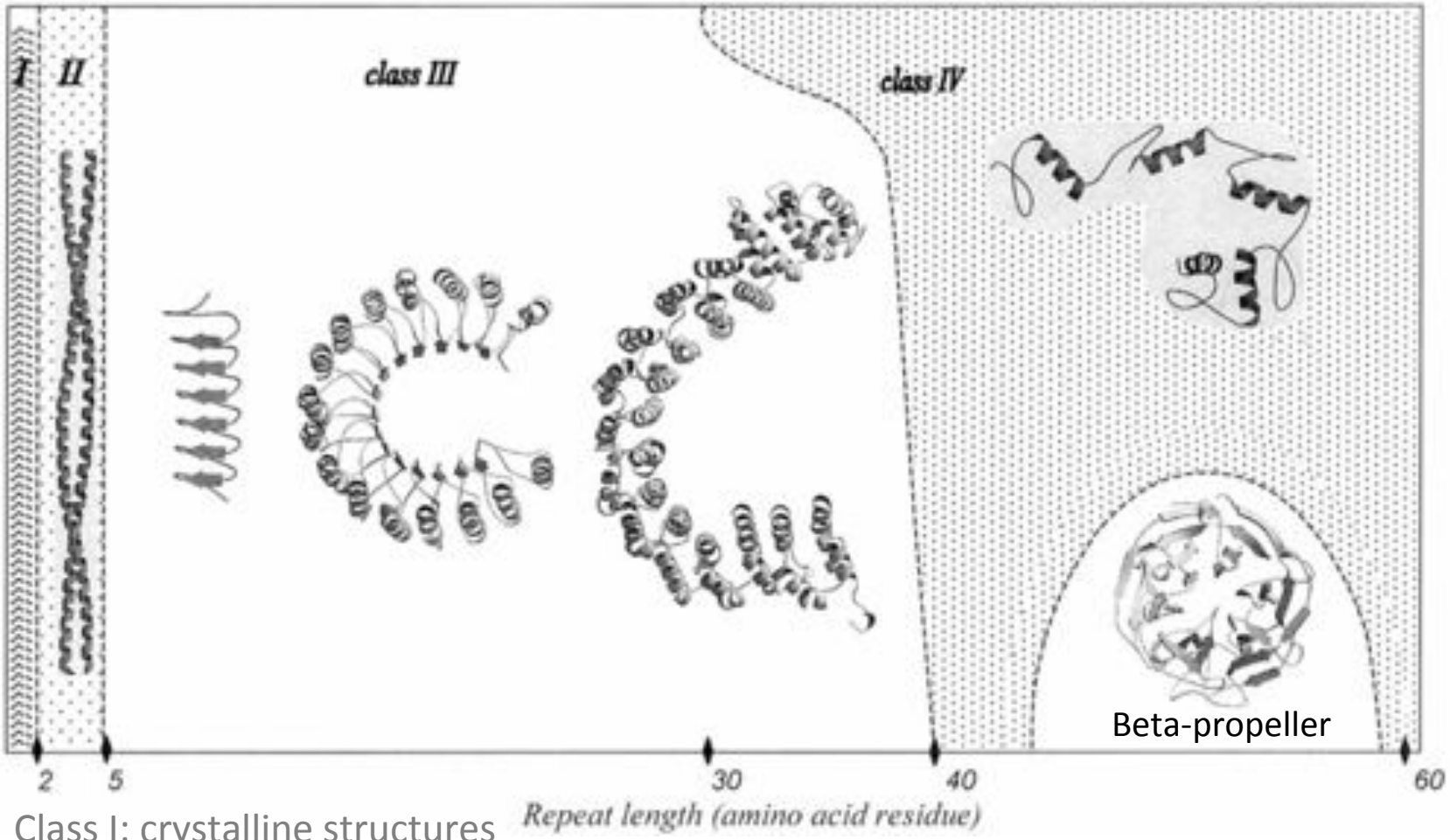
Homorepeats: Types of crystal structures



PolyQ: crystallite (A) and fibril-like assemblies (B)

from Jorda and Kajava 2010

TRs: Structure vs unit length



Kajava 2001

Class I: crystalline structures
Class II: fibrous proteins
Class III: solenoid proteins
Class IV: domain-forming repeats

TRs are often disordered

Proteins with TRs are more likely to be disordered (Dunker et al. 2002)

Proteins with intrinsic disorder are more likely to contain repeats (Szalkowski & Anisimova, submitted)

TRs are often disordered

the FEBS Journal

Protein tandem repeats – the more perfect, the less structured

Julien Jorda[1], Bin Xue[2,3], Vladimir N. Uversky[2,3,4,5] and Andrey V. Kajava[1]

Table 2. Analysis of intrinsic disorder distribution in TRs and TR-containing proteins.

	P_{sim} = 0.7–0.8	P_{sim} = 0.8–0.9	P_{sim} = 0.9–1	Homorepeats
TRs				
Total no.	34 286	5519	1382	5259
Average length	25.5	41.0	59.1	13.8
Intrinsic disorder ratio (%): VSL2	80.4	88.6	88.9	98.4
Intrinsic disorder ratio (%): IUPRED	56.0	62.7	67.2	86.5
Intrinsic disorder ratio (%): FOLDINDEX	62.4	68.6	70.3	79.9
Intrinsic disorder ratio (%): TOPIDP	85.6	88.8	91.1	74.4

Why study repeat proteins?

A large fraction of all proteins

Enhanced binding properties

Mediate protein-protein interactions

Play role in evolution of disordered regions

Frequently associated with disease

Prokaryotes: in virulence-related genes

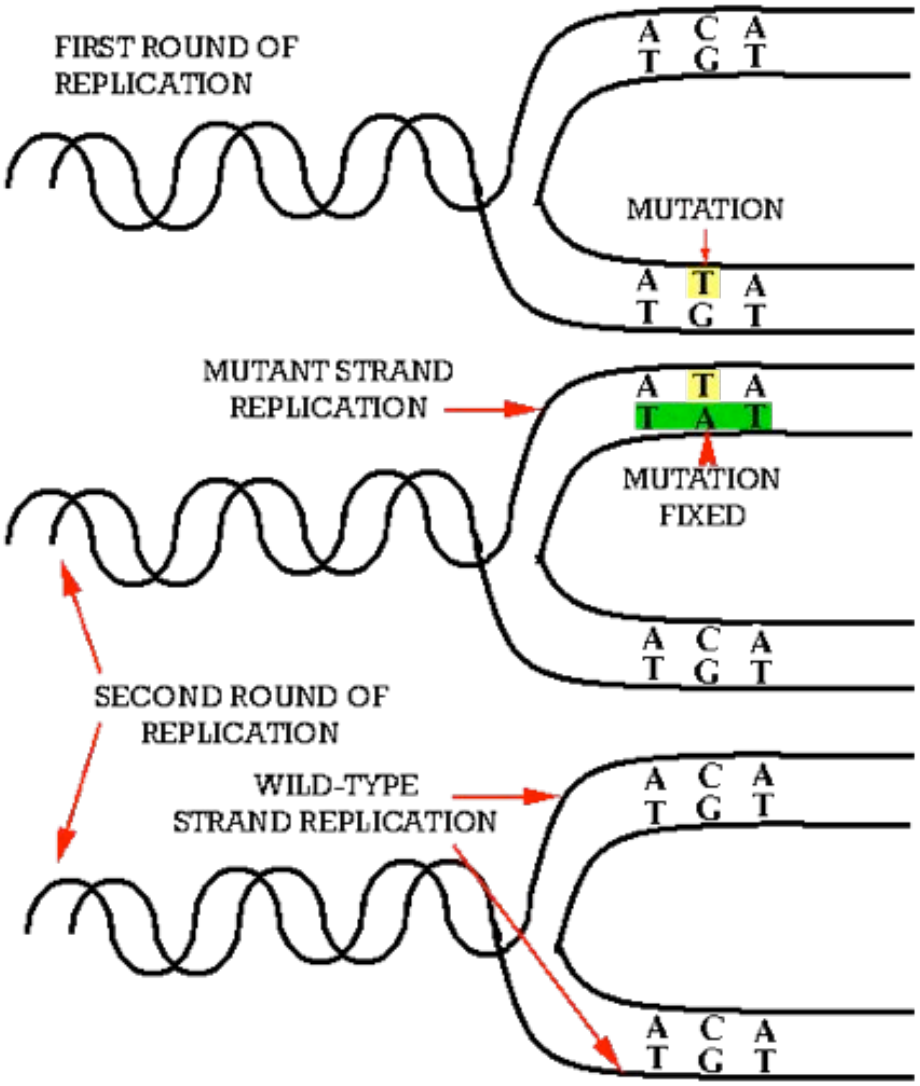
How do proteins with TRs evolve?

- Point mutation (incl. indels)
- Replication slippage
- Intragenic duplications
- Intragenic recombination

(e.g., Marcotte et al 1999)

- *de novo* creation

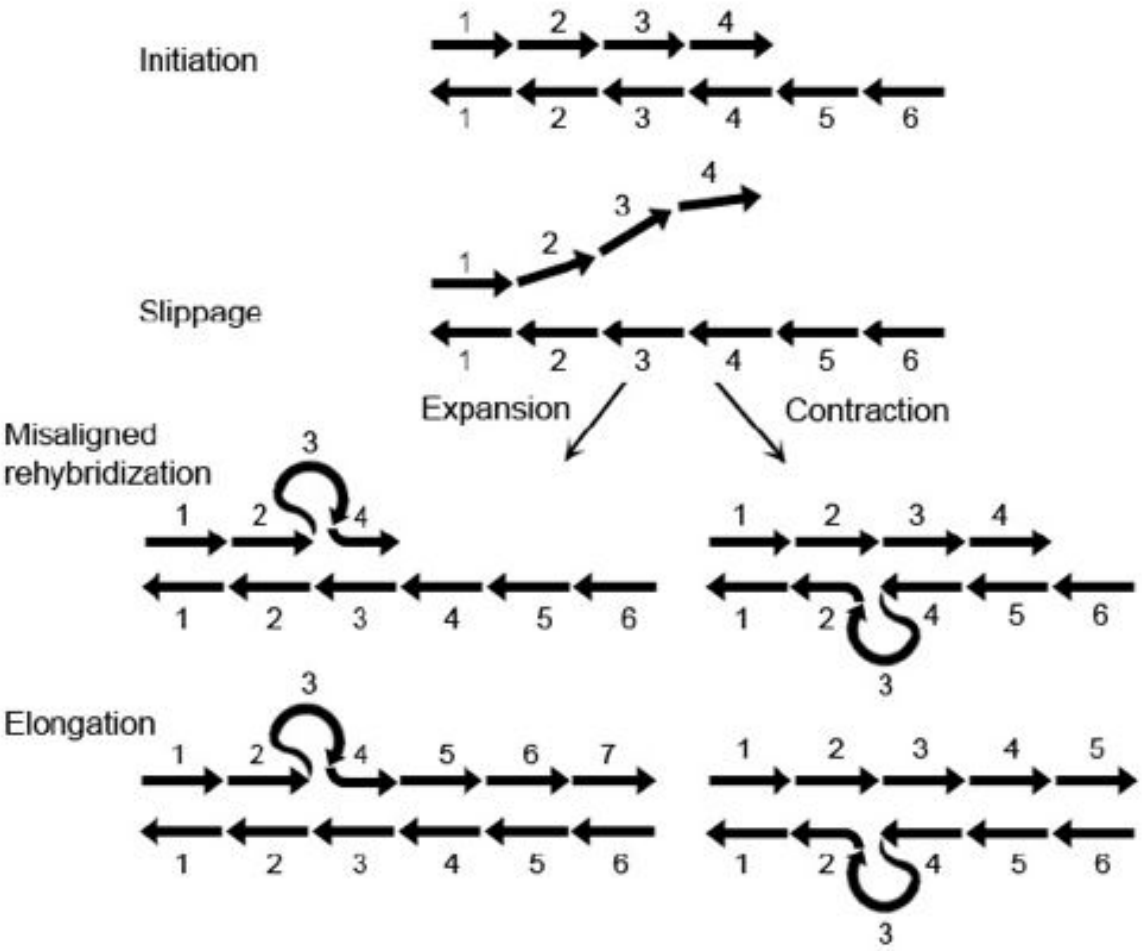
Point mutation

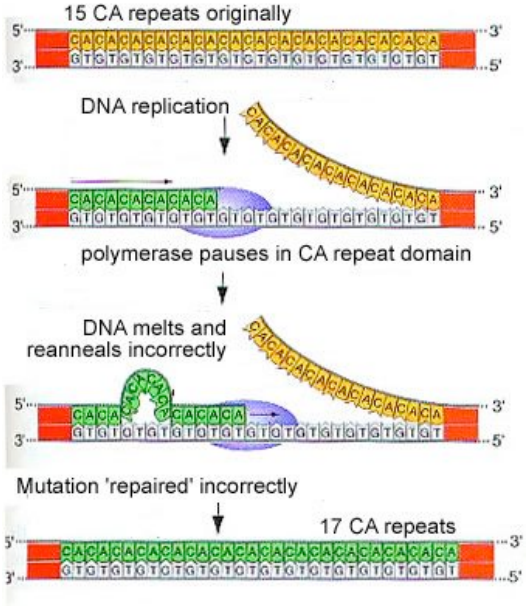


Markov process of character substitutions

No realistic models for indels

Replication slippage



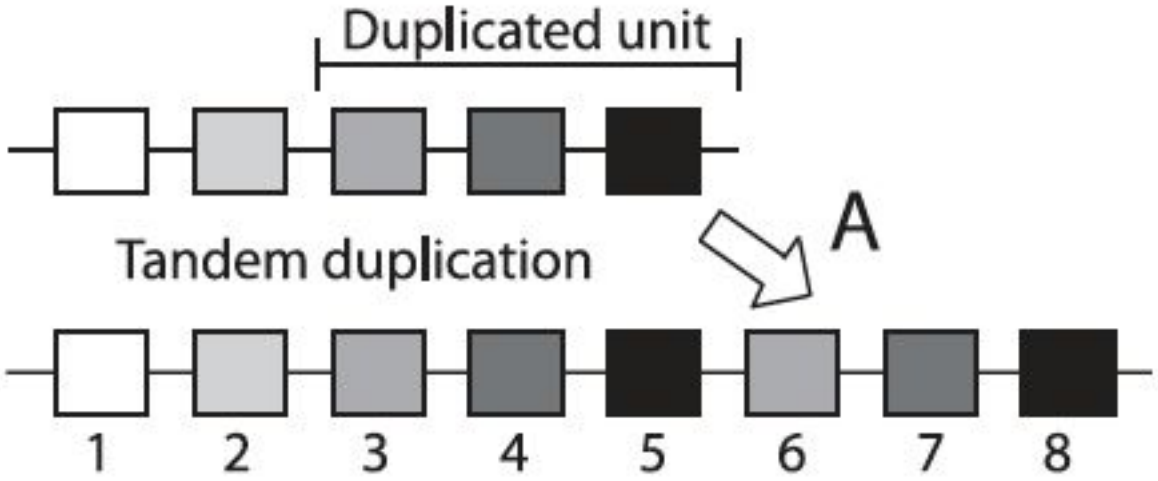


From Jorda & Kajava 2010

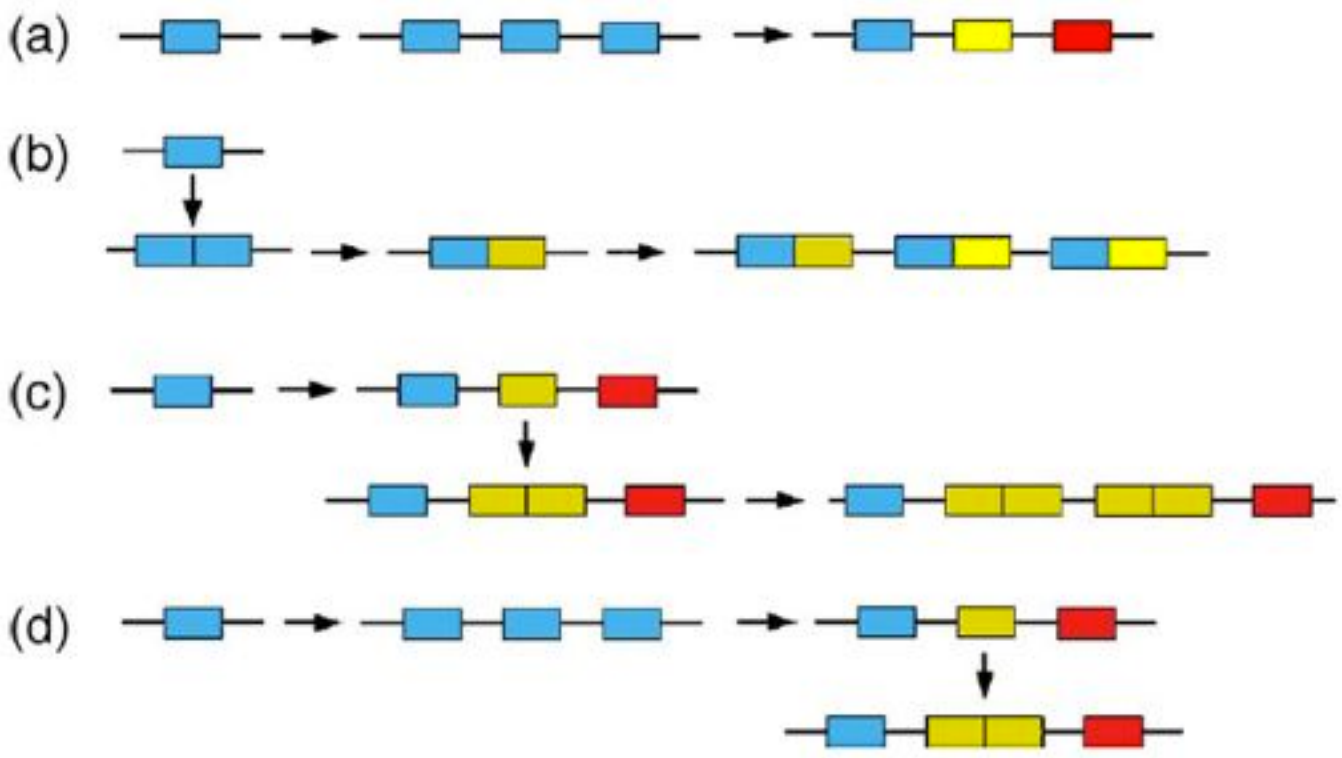
http://virtuallaboratory.colorado.edu/Biofundamentals/lectureNotes/Topic4-3_Mutation%20and%20Duplication.htm

Intragenic duplications

A process similar to gene duplication/loss



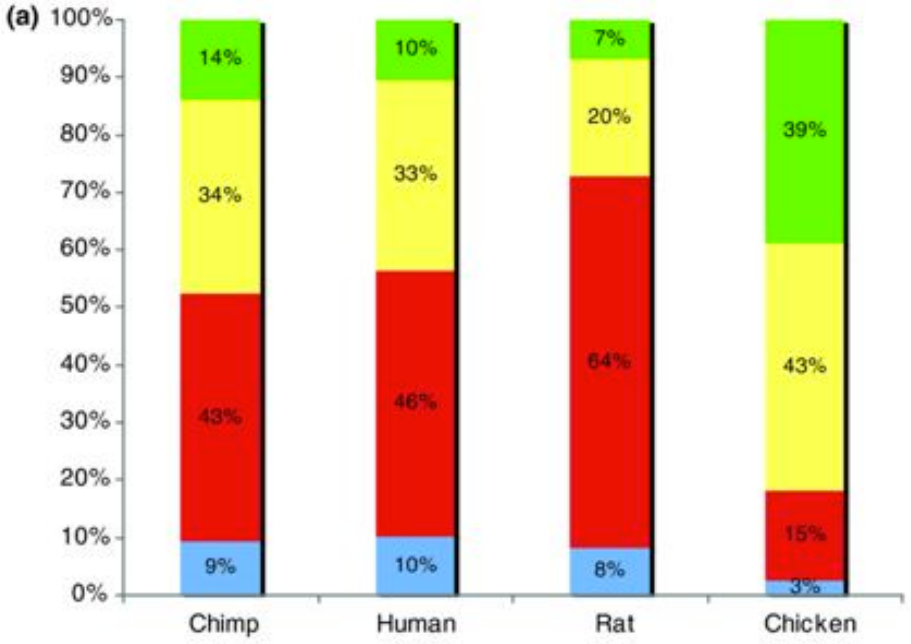
Intron facilitated repeat duplications



Introns have contributed to the greater abundance of repeat protein genes in eukaryotic vs prokaryotic organisms

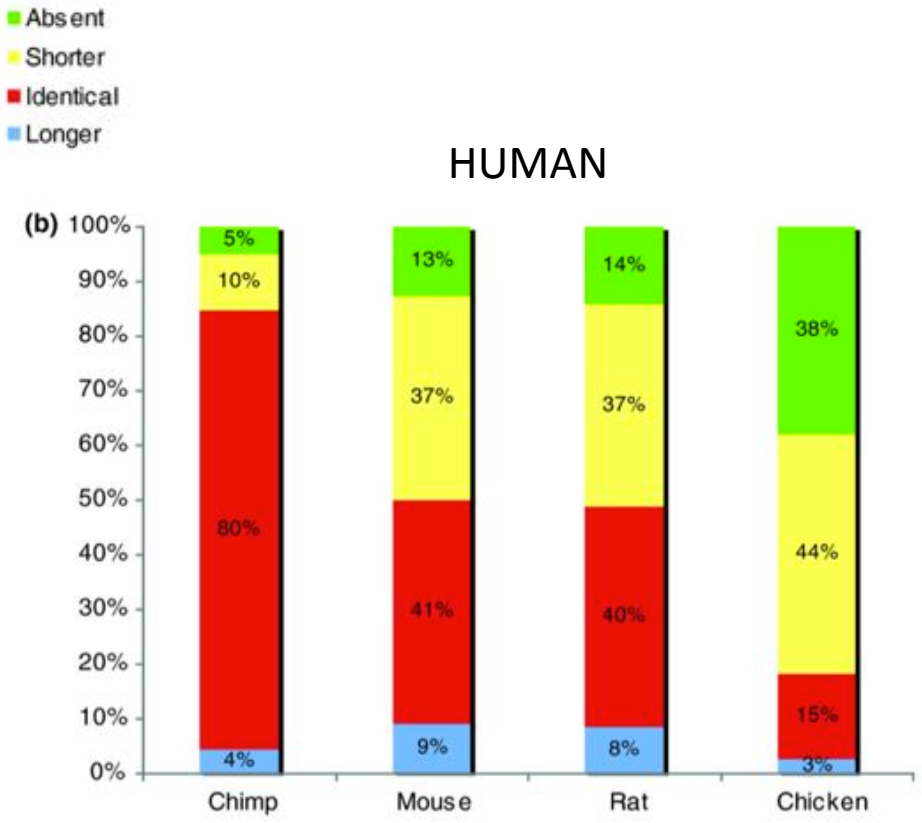
Street, Rose and Barrik 2006

Conservation of TRs



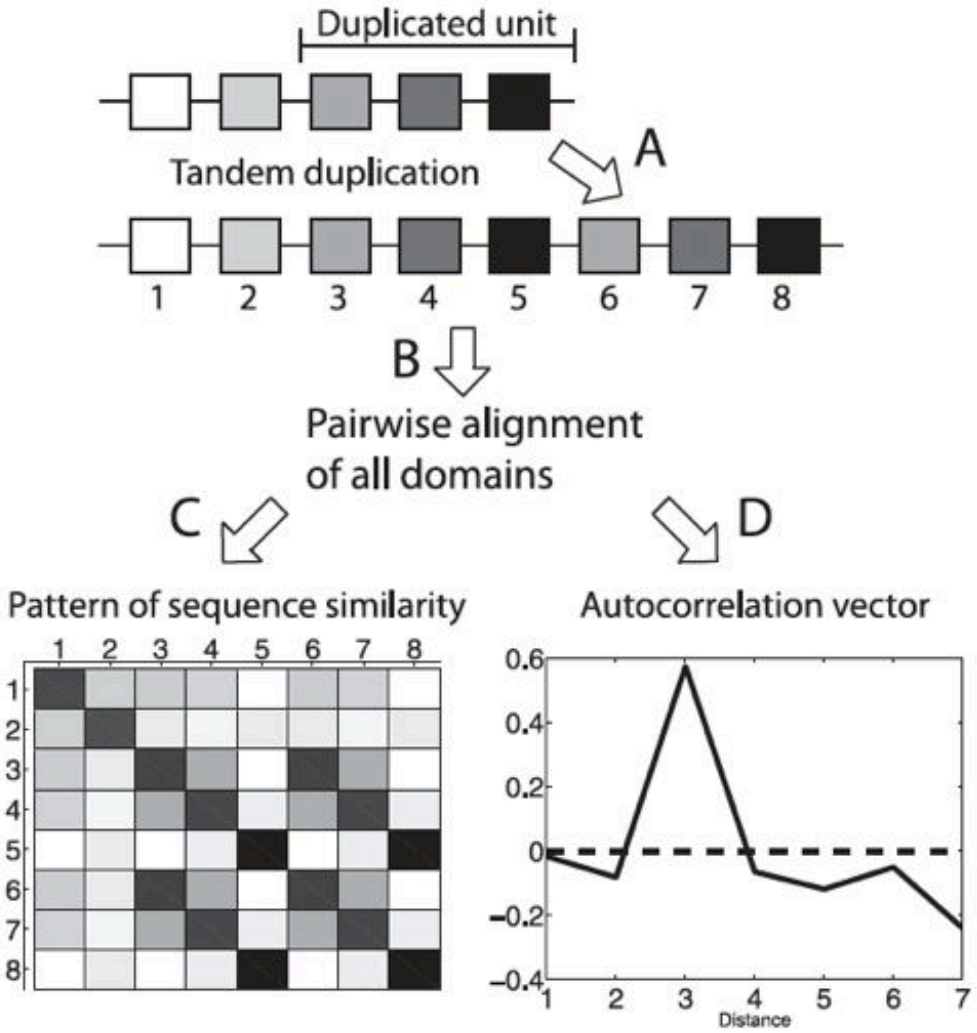
MOUSE

Species were tested for presence of human or mouse TRs



From Simon and Hancock 2009

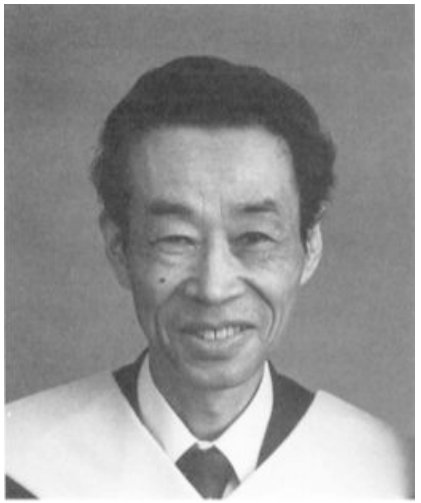
Conservation patterns of TRs



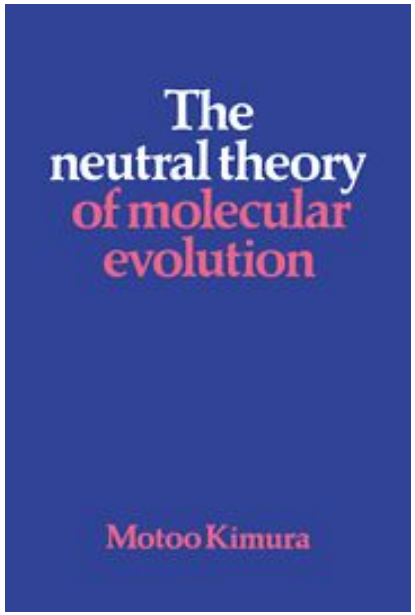
Bjorklund, Ekman, Elofsson 2006

Neutrality theory

Kimura 1968, King & Jukes 1969



Majority of molecular changes in evolution are due *to random fixation of neutral or nearly neutral mutations*, and not caused by positive selection of advantageous alleles nor by balancing selection



Provides a falsifiable null hypothesis when testing for selection

Does selection shape TRs?

Selection may act on:
Point mutations in TRs?
Selection on TR number?
Selection on TR composition?

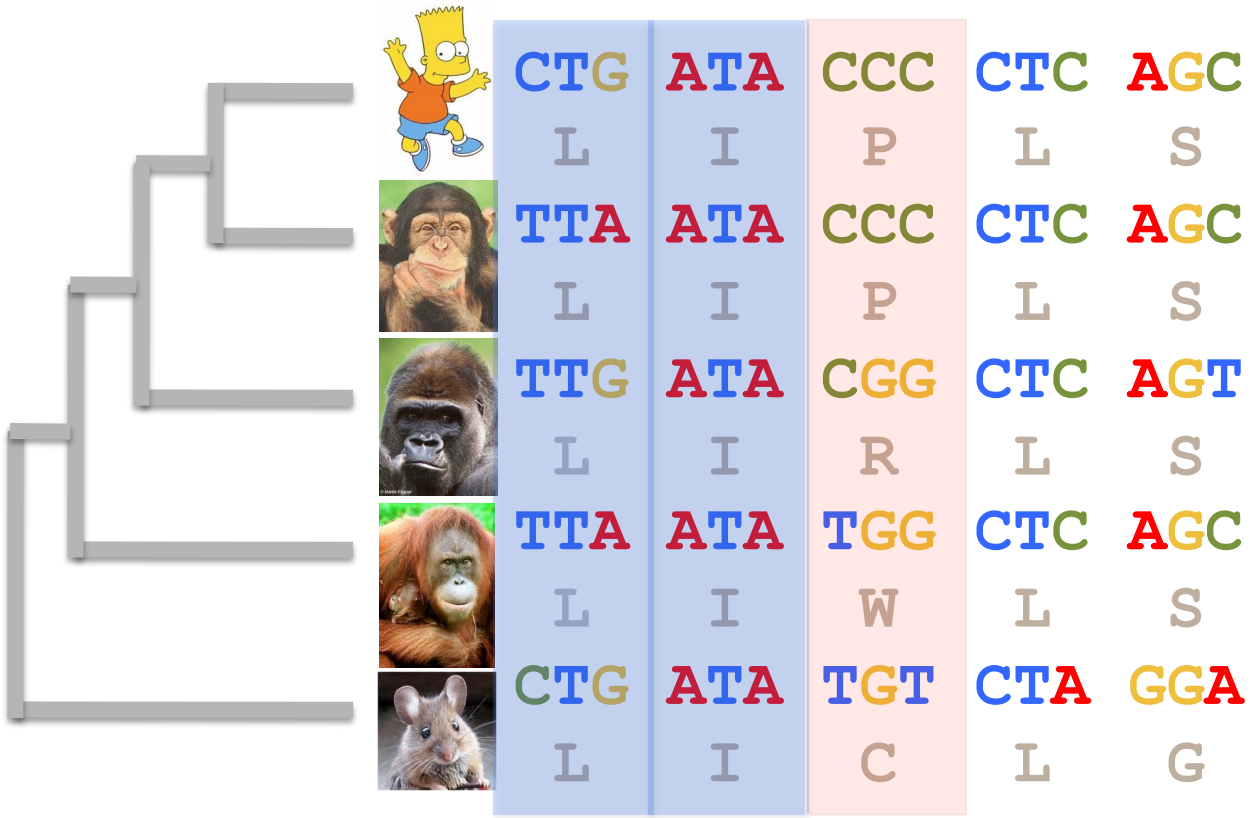
Null hypothesis: neutrality

Selection on point mutations

Neutrality tests (mutational spectrum)

Codon-model-based tests of selection

Measuring selection on protein



synonymous rate: d_S nonsynonymous rate: d_N

$\omega = d_N/d_S > 1$ positive selection

$\omega < 1$ negative selection

Markov model of codon substitution

Instantaneous substitution matrix $Q = \{q_{ij}\}$:

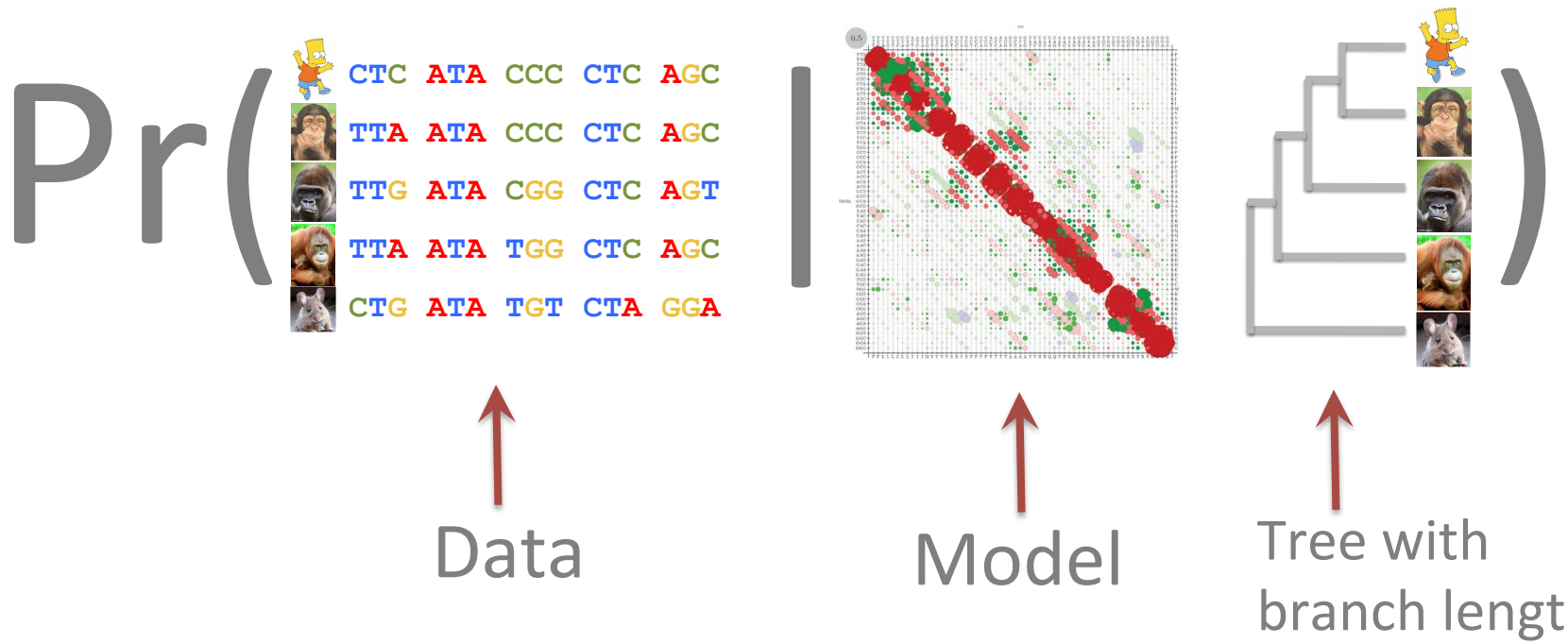
Type of change	GY-type model
2 or 3 nt changes	0
Synonymous transition	π_j
Synonymous transversion	$\kappa\pi_j$
Nonsynonymous transition	$\omega\pi_j$
Nonsynonymous transversion	$\omega\kappa\pi_j$

$\omega_s = d_N/d_S$ (selection on protein)
κ = transition/transversion ratio
π_j = frequency of codon j

Likelihood calculation on a phylogeny

Transition probability matrix over time t: $P(t) = e^{Qt}$

Using P(t) a likelihood L(Data) can be constructed:



Parameters optimized by maximum likelihood

Likelihood ratio test of selection

Model 0 no positive selection

(H0: ω is always ≤1)

Model 1 allows positive selection

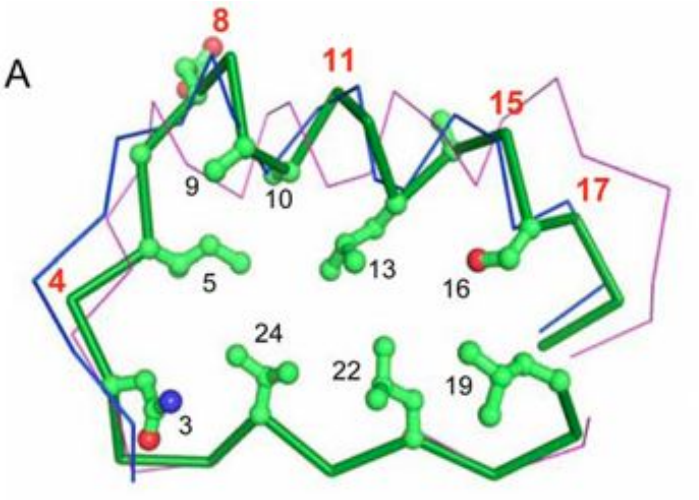
(H1: ω >1 for some sites or lineages)

M0 and M1 are nested!

LRT statistic: $2\Delta\ell = 2(\ell_1 - \ell_0) \sim \chi^2_{d.f.}$

d.f. = difference in numbers of parameters

Case study 1: LRR proteins as receptors



Bacterial GALA (type III effectors) acquired from host plants by LGT: residues under positive selection are found on the convex side of horse-shoe & involved in binding

From Kajava, Anisimova, Peeters (2008)

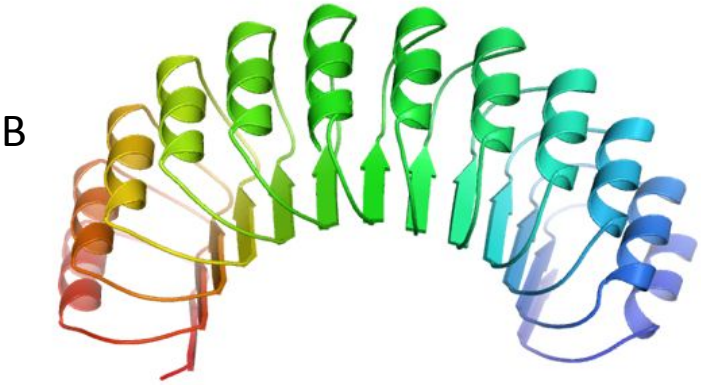


Figure 2. Structural model of GALA-LRR. (A) Cα-trace superposition of a modeled GALA-LRR and the known CC-LRR from human Skp2 protein [10] and RI-LRR from porcine ribonuclease inhibitor [46]. GALA-LRR model is shown in a ball-and-stick representation, CC-LRR is shown by a blue trace and RI-LRR by a magenta trace. Numbering of the conserved GALA-LRR residues is taken from Figure 1. Numbers in red point to positions inferred to be under positive selection. The carbon atoms are in green, oxygen in red, nitrogen in blue. (B) A ribbon diagram of a structural model of the C-terminal LRR domain of GALA4 type III effector protein from *R. solanacearum* (strain MolK2, region 170 to 460, accession code ZP_00946474). The figure was generated with Pymol [47]. The atomic coordinates of the model are available on request.

Case study 2: Homorepeats (Faux et al 2007)

Data: 13 species: 3 mammals, 2 fish, 1 bird, 3 insects, worm, yeast, malaria, 1 plant

(1) Codon bias, but no no GC or AT bias relative to the organism's transcriptome

(2) single-nt transversions from the homocodon are unusually common:

mechanism of reducing the rate of slippage?

Case study 2: Homorepeats (Faux et al 2007)

Alignment = Repeat + repeat Flanks + Shadow

Applied codon models to test:

$$H0: \omega_S = \omega_F = \omega_R \text{ vs } H1: \omega_S = \omega_F \neq \omega_R$$

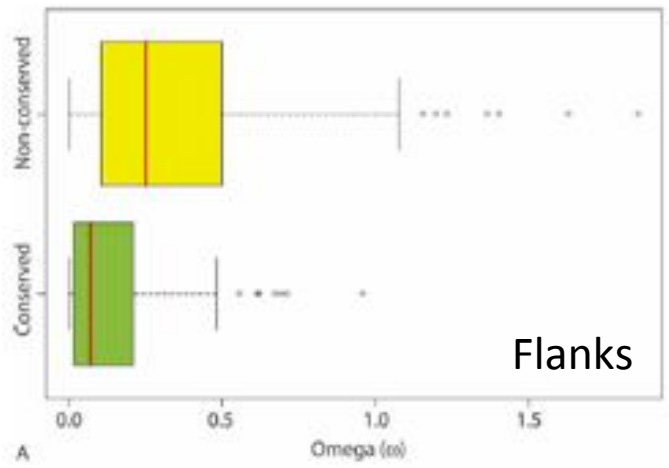
$$H0: \omega_S = \omega_F \neq \omega_R \text{ vs } H1: \omega_S \neq \omega_F \neq \omega_R$$

$$H0: \omega_S \neq \omega_F \neq \omega_R \text{ vs } H1: \omega_S \neq \omega_{F\_NC} \neq \omega_{F\_C} \neq \omega_R$$

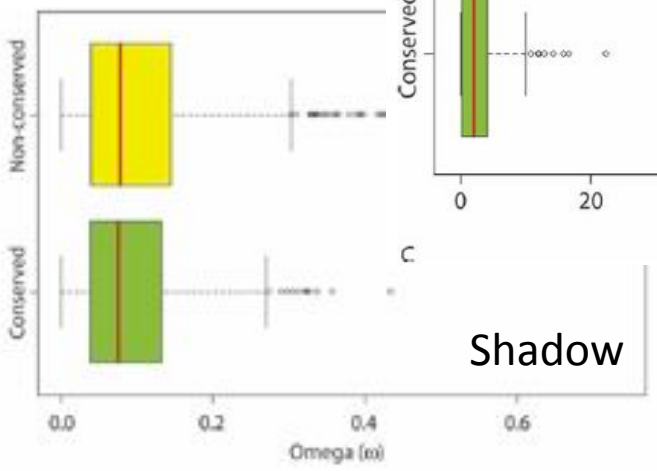
Case study 2: Homorepeats (Faux et al 2007)

$\omega_S > \omega_F$: conserved across species homopeptides lie within regions that are under stronger purifying selection in contrast to nonconserved homopeptides

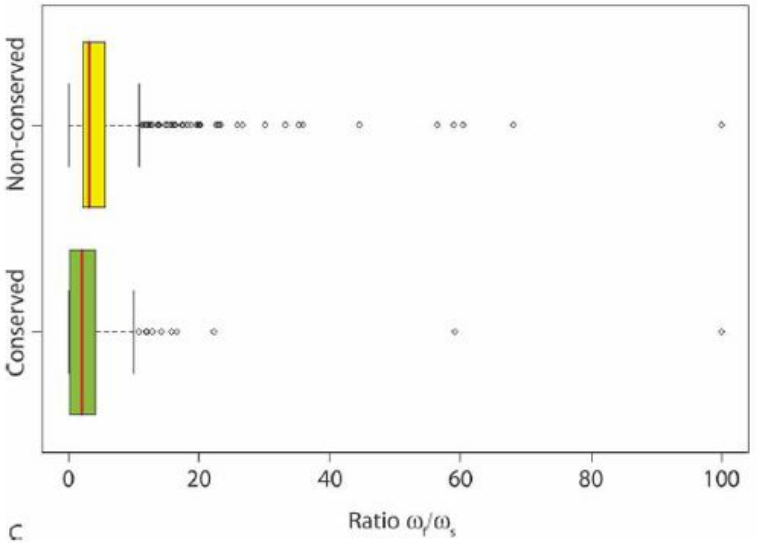
ω_R? - unreliable



Flanks



Shadow



Case study 3: selection on repeat number

28.3% repeats conserved in vertebrates

Mularoni et al 2010 detected repeats using "in-house Python program"

Repeat conservation: same amino acid motif and overlapping repeat regions

What is the null expectation?

Mularoni et al 2010 used non-coding sequences with matching DNA composition (consecutive repetitions of 4 or more trinucleotides)

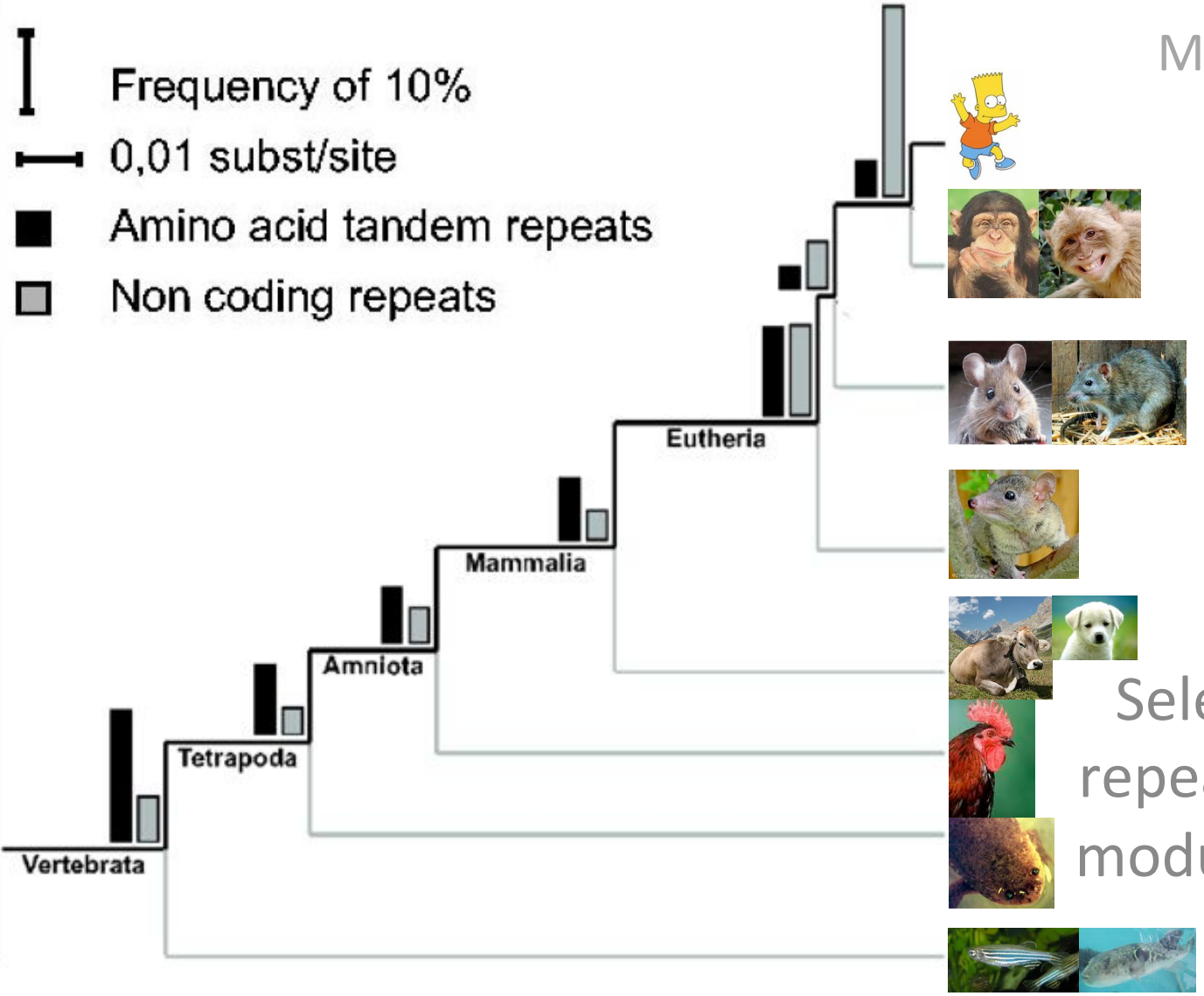
Case study 3: selection on repeat number

From Mularoni et al 2010

AA I triplet	Amino acid tandem repeat size				
	N	Max	Mean	Median	Standard Deviation
A I GCN	361	15	5.04	4	2.02
E I GAR	545	15	4.95	4	1.72
G I GGN	211	16	5.03	4	1.12
K I AAR	224	9	4.5	4	0.93
L I TTR CTN	402	11	4.34	4	0.75
P I CCN	446	21	4.88	4	1.83
Q I CAR	158	40	6.15	4	5
S I AGY TCN	160	42	4.67	4	2.26
other	495	14	4.45	4	1.25
total	3417	42	4.8	4	2.02

Table 1. Characteristics of human amino acid tandem repeats. Repeats of size 4 or longer present in human proteins from a 6,477 vertebrate protein orthologous dataset.

Case study 3: selection on repeat number

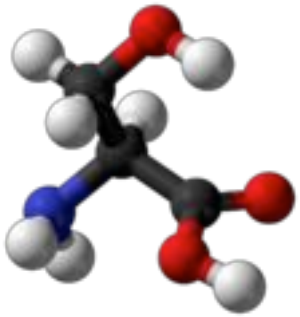


Mularoni et al 2010

Selection increases repeat retention rate, modulates repeat size

Case study 4:
Selection or slippage? (Huntley & Golding 2006)

Data: orthologs to proteins with serine homopolymers in human (including Serine rich regions)



Ser (S)

{ AG**T**, AG**C**, TC**A**, TC**C**, TC**T**, TC**G**}

Model states: {AGY, TCN, Nonserine (*), -}

	AGY	TCN	*	-
AGY	$1-\alpha-\beta-\gamma$	α	γ	β
TCN	α	$1-\alpha-\beta-\gamma$	γ	β
*	γ	γ	$1-2\gamma-\delta$	δ
-	β	β	δ	$1-2\beta-\delta$

High slippage -> high β; high selection on protein -> high α

Case study 4:
Selection or slippage?

Likelihood ratio tests (testing if $\alpha=0$ and if $\beta=0$) suggest that not all selection is due to slippage

In some proteins: no slippage but selection (Huntley and Golding 2006)

Other studies

Codon models (PAML) were used by Huntley & Clark 2007:

Presence of TRs is associated with the increase of the evolutionary rate in the embedding sequence,

and elevation of positive selection

(Data: 12 Drosophila)

Haerty and Golding 2010:

Difference in exposure to selection between alternatively and constitutively spliced exons leads to a significant difference in size and codon composition in such exons:

Slippage is in balance with point mutations due to selection

Other studies

Cruz, Roux & Robinson-Rechavi 2009

No correlation of number of TRs with selective pressure (estimated from codon models)

In mammals weak correlations are found but can be explained by CG composition

Potential problems

Total data set

Inaccurate repeat identification

Representation of null hypothesis

How are homorepeats treated?

(often include low complexity regions,
vary in min. length threshold)

What type of selection is tested for?

Questions for review

What are the common trends of the evolution of TR proteins?

Is there evidence of selection operating on:
point mutation in TRs
repeat number/length?

Do homorepeats exhibit special trends
(different from other TRs)?

Is current knowledge about TRs dominated by homorepeats?
If so – is this justified?

In what organisms TRs are “understudied”?