Repetitive Regions in Proteins

---- Protein Repeats

Xue-Jia Hu

09/11/2014

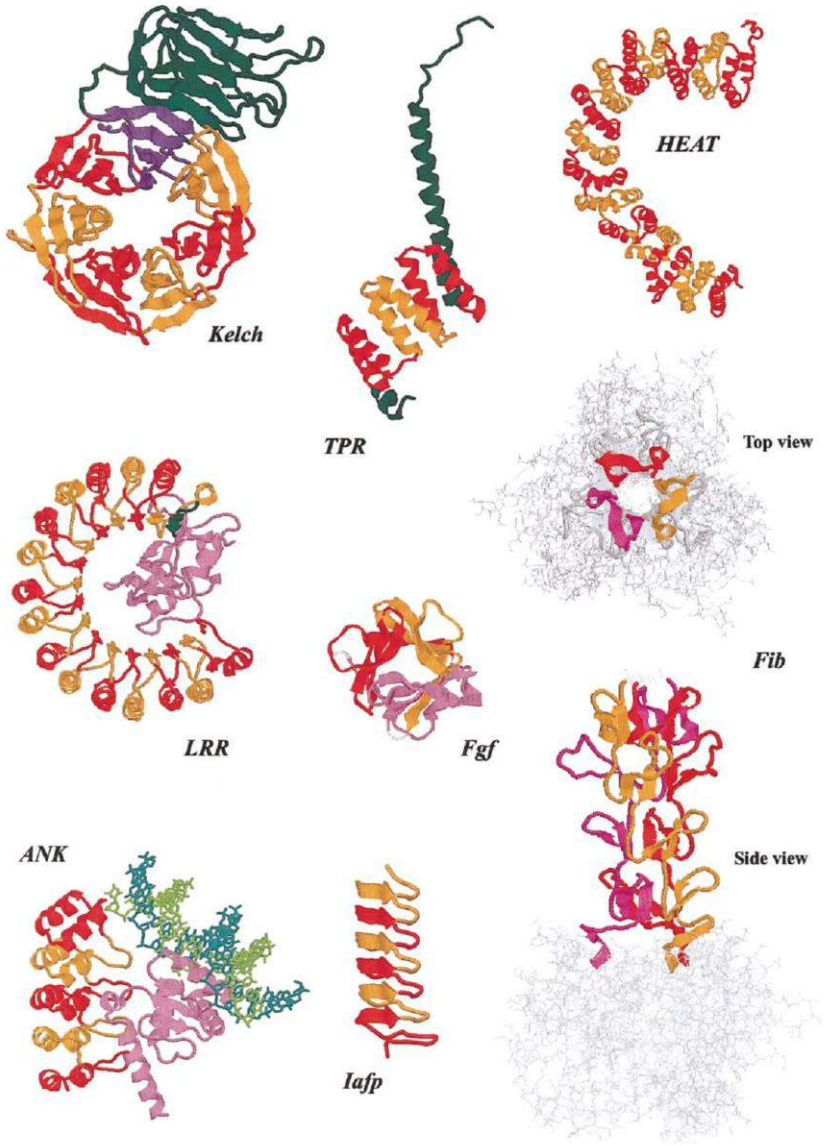
Content

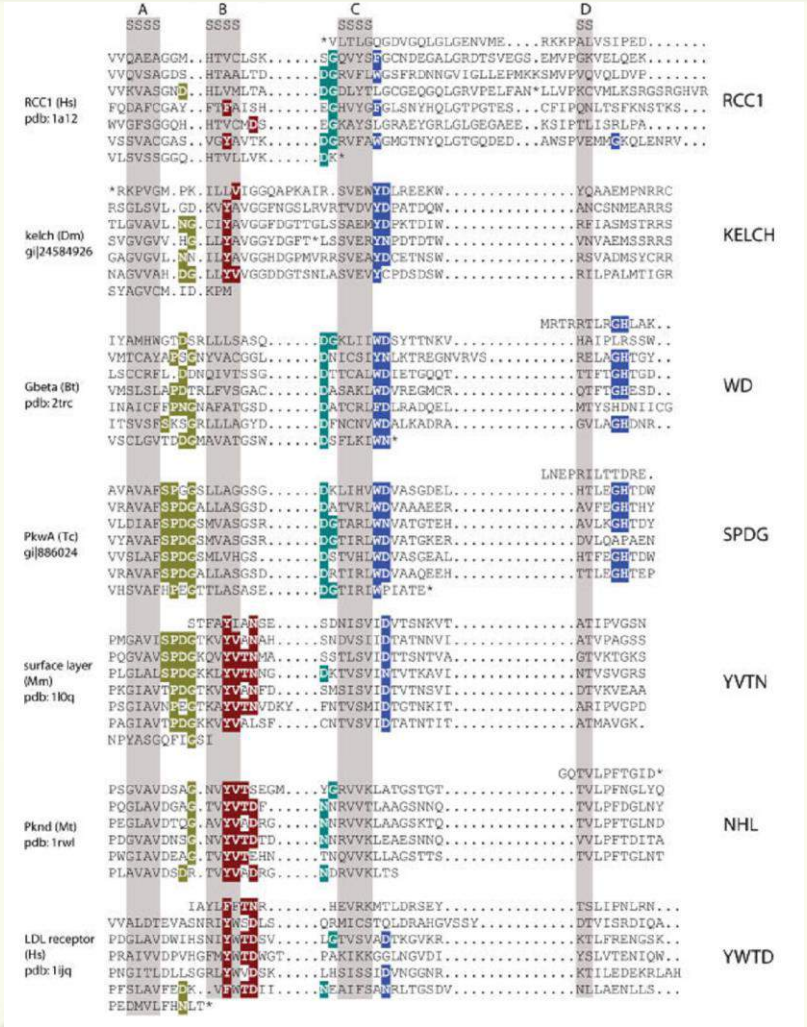
1. Background
2. The incidence of repetitive regions in proteins
3. Functions of repetitive regions
4. The formation of repetitive regions in proteins

Content

1. Background
2. The incidence of repetitive regions in proteins
3. Functions of repetitive regions
4. The formation of repetitive regions in proteins

Examples of repetitive regions in proteins





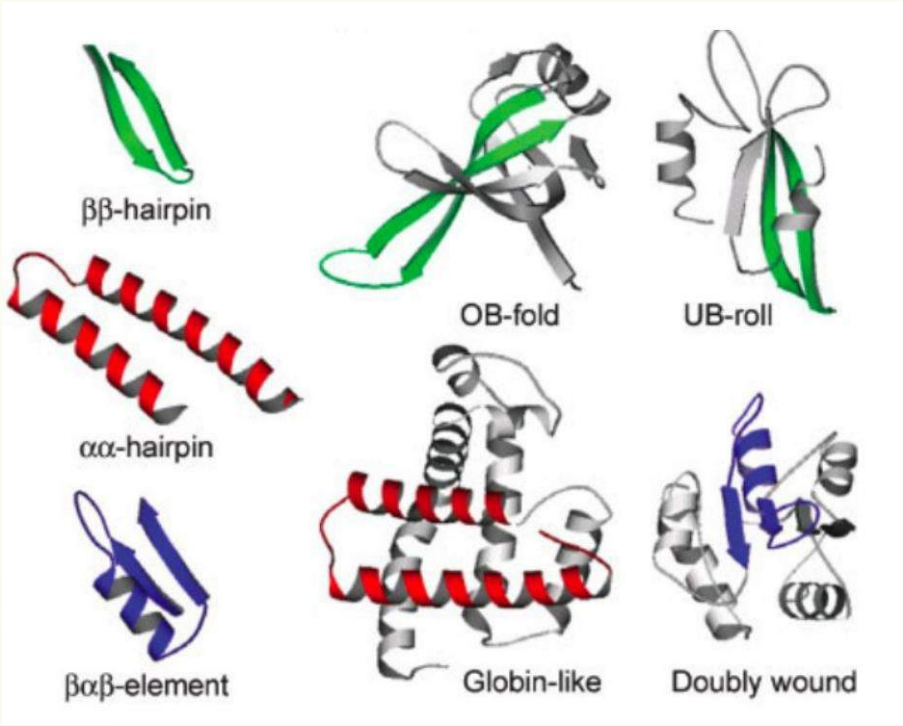
Andrade M. A.; et al. *JSB* **2001**, 117-131.
Andrei N. Lupas, et al. *Proteins*. **2008**, 795-803.

Why to study them?

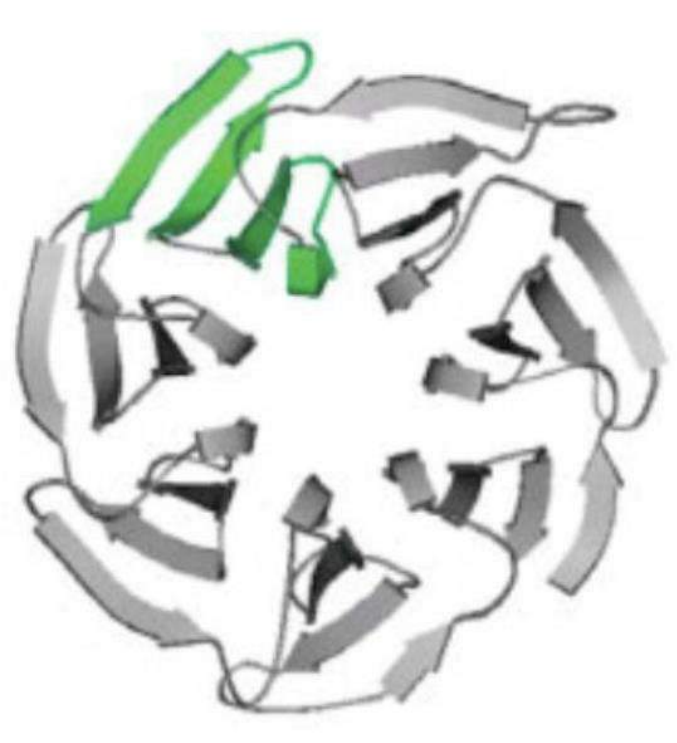
- At least 14% of all proteins contain repeats.
- About 25% of all eukaryotic proteins contain repeats.
- Internal repeats often correspond to structural or functional units in proteins
- Repetitive regions are found in many classes of proteins
  - Fibrous proteins
  - Solenoid proteins
  - Membrane proteins
  - Globular proteins
- There is a wide variety of functions of repeats, even within the same family

They can be decomposed into building blocks

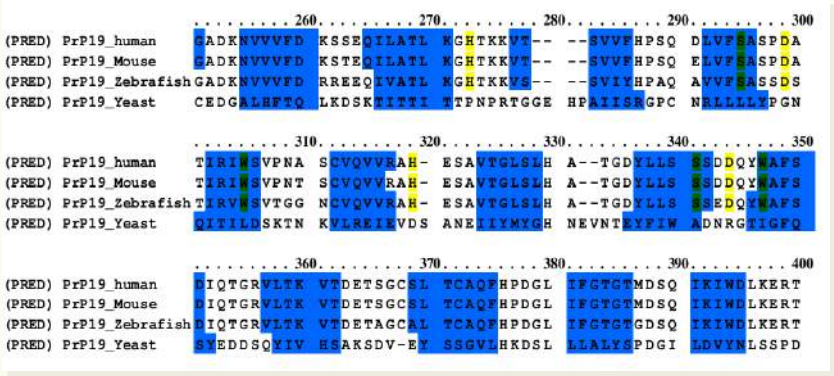
- Protein domains
  - Autonomously folding polypeptides
  - Modular reuse in different contexts
  - Classification schemes are based on domains not entire proteins
- Domains themselves are formed by a set of recurring super-secondary structure elements



Structural symmetry and sequence dissimilarity



Internal structural symmetry within domains



The identity of PrP19_human & PrP19_Yeast: 39%

- Even if the repeated domains have a well-defined and conserved structure, the sequence conservation is often low, with only a few conserved residues required for the correct fold.
- Their variable sequences and the variation in number of domains provide flexible binding to multiple binding partners.

Classification of repeats

Based on the length of the repeated unit:

- Short:
  - One amino acids (polyQ)
  - Three aa (Cold shock protein)
  - Seven aa (Coiled coils)
- Intermediate:
  - Super-secondary structure elements
  - (αα/ββ-hairpins, βαβ-elements, ca. 30-40 aa)
- Long:
  - Entire Domains (ca. 100-200 aa)

✓ Repeats number: variability

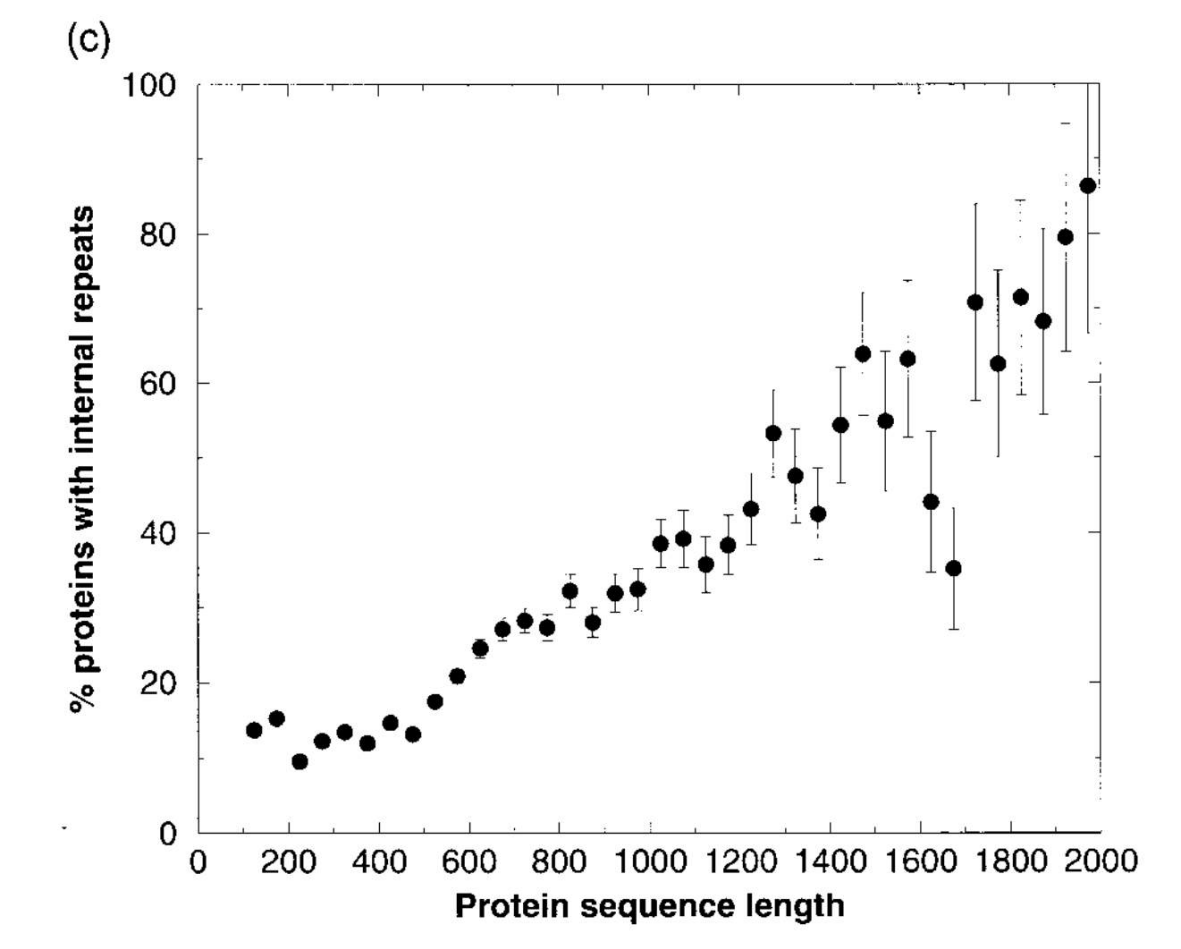
  ✓ The numbers of repeats can vary even between orthologues, indicating that rapid loss and/or gain of repeats occurs frequently in evolution.

Andrade M. A.; et al. *JSB* **2001**, 117-131.

Content

1. Background
2. The incidence of repetitive regions in proteins
3. Functions of repetitive regions
4. The formation of repetitive regions in proteins

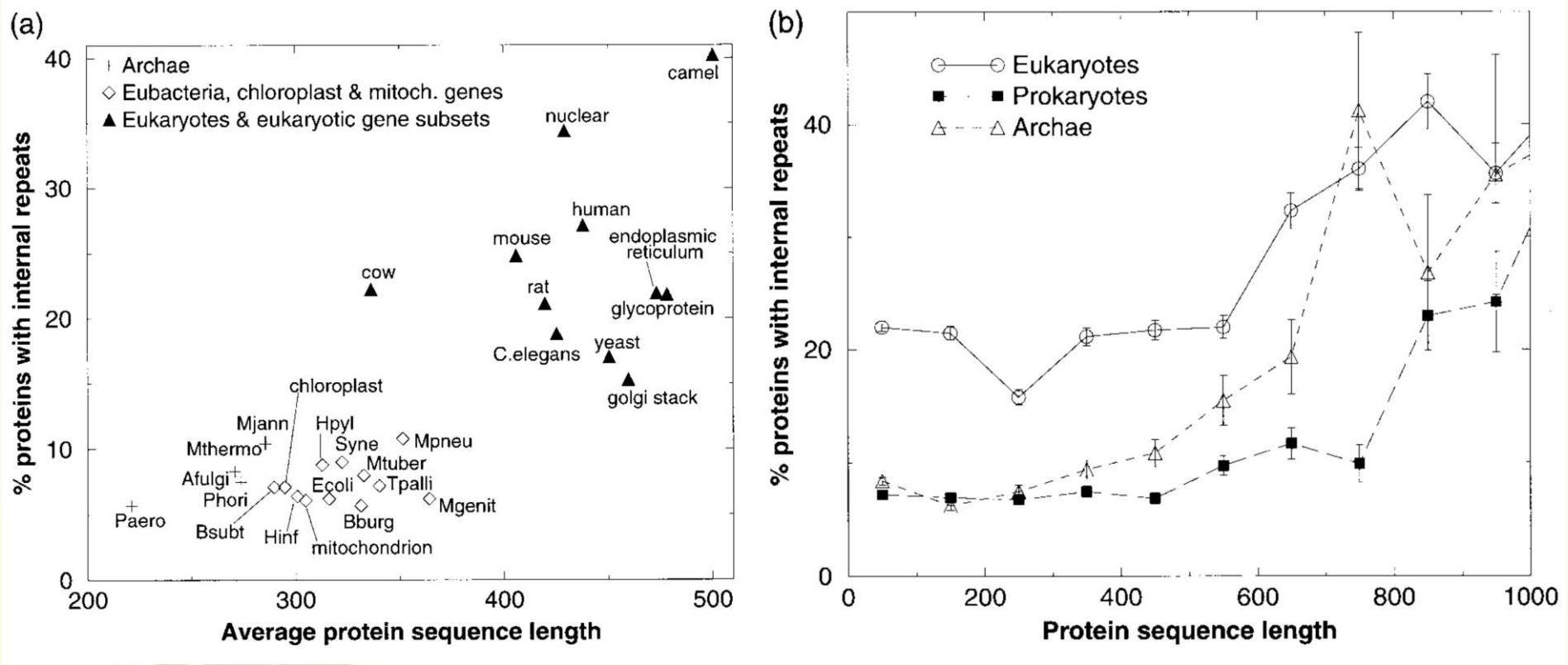
The incidence of internal repeats in proteins



Beyond 500 amino acid residues, a linear dependence upon sequence length is seen, suggesting that generation of internal repeats is an important mechanism for producing long proteins.

Marcotte E. M.; et al. *JMB* **1998**, 151-160.

The incidence of internal repeats in proteins

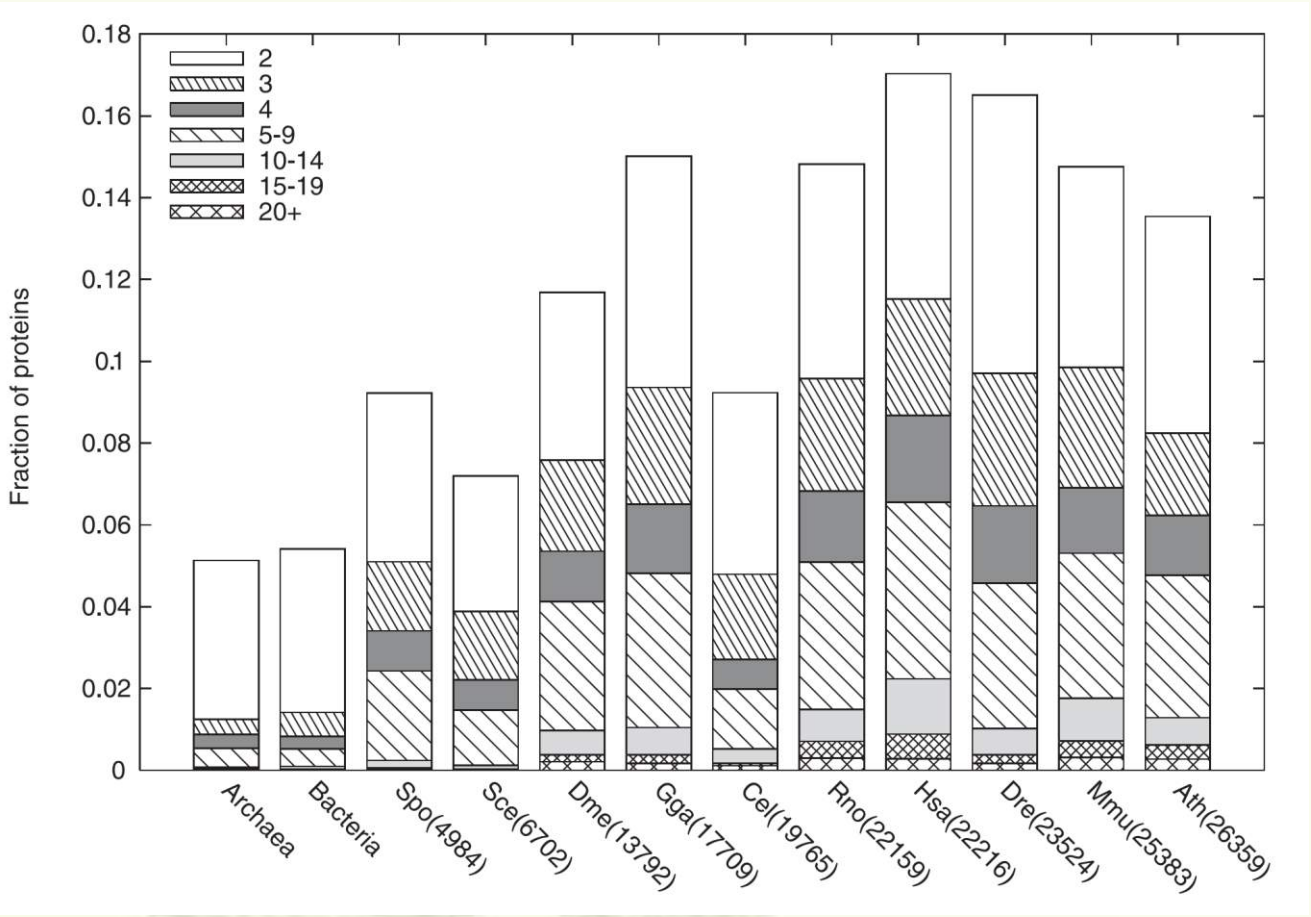


On average, eukaryotes have longer proteins than prokaryotes and archael organisms.

For proteins of the same length, eukaryotic proteins are approximately three times more likely to have internal duplications than prokaryotes, with archaea falling inbetween.

Marcotte E. M.; et al. *JMB* **1998**, 151-160.

Fraction of proteins



More complex organisms seem to require more domain repeats.

Consequently, the fraction of proteins with repeats is higher for species with large proteomes.

Fraction of proteins that contain a domain repeat in archaea, bacteria, yeast, and eight multicellular eukaryotes. The different patterns indicate the length of the repeat, whether it contains 2, 3, 4 domains, etc.

Bjorklund A. K.; et al. Plos Computational Biology. **2006**, 959-970.

Content

1. Background
2. The incidence of repetitive regions in proteins
3. Functions of repetitive regions
4. The formation of repetitive regions in proteins

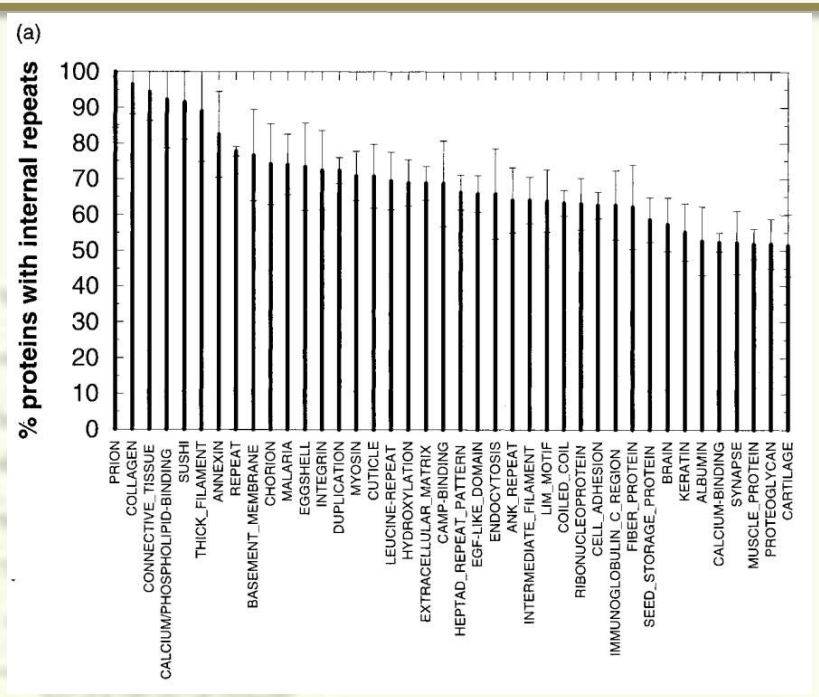
Functions of the repeats

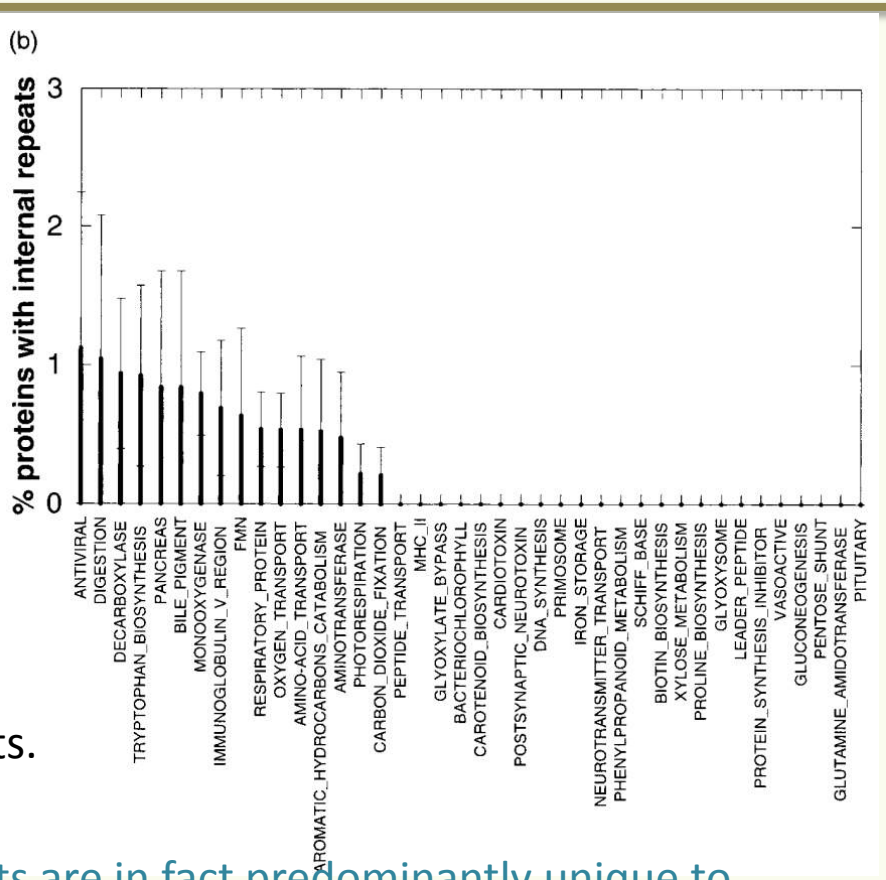
Repeat	Ref1	L	3D	PDB	Ref2	Distribution	Function	Pfam
Kelch	Neer *et al.* (1994)	40	β-Barrel	1gof	Ito *et al.* (1991)	Eukaryotic	Enzyme. Protein processing	PF01344
Fibroblast growth factor	Murzin *et al.* (1992)	40	β-Trefoil	2afg_A	Eriksson *et al.* (1993)	Eukaryotic–viral	Development	PF00167
Tetratrico-peptide repeats	Zhang *et al.* (1991)	34	α-α	1a17	Das *et al.* (1998)	Eukariotic–bacterial–archaeal	PPI	PF00515
Ankyrin	Lux *et al.* (1990)	33	α-α-β-Hairpin	1awc_B	Batchelor *et al.* (1998)	Eukaryotic–bacterial–viral	PPI	PF00023
HEAT	Andrade and Bork (1995)	47	α-α	1b3u_A	Groves *et al.* (1999)	Eukaryotic	PPI	None
Leucine-rich repeats	Kobe and Deisenhofer (1994)	20	α-β	1dfj_I	Kobe and Deisenhofer (1995)	Eukaryotic–bacterial	PPI	PF00560
β-Farnesyl transferase	Park *et al.* (1997)	42	α-Barrel	1ft2b	Park *et al.* (1997)	Eukaryotic	Enzyme. Protein processing	None
Adenovirus fiber protein	Green *et al.* (1983)	15	Triple β spiral	1qiu	van Raiij *et al.* (1999)	Viral	PPI. Binds to host receptor	None
Zein	Argos *et al.* (1982)	20	α-Helix (proposed)	Model	Matsushima *et al.* (1997)	Plants	Plant seed storage protein	PF01559
Bacterial glycosyl transferase	Wren (1991)	35	Unknown	None		Bacterial	Enzyme. Small molecules binding	None
Insect antifreeze protein	Graham *et al.* (1997)	12	β-sheet	1ezg_A	Liou *et al.* (2000)	Metazoa	Ice binding. Antifreeze	None
Ice nucleation protein	Gurian-Sherman and Lindow (1993)	16	Hairpin-loop	1ina	Tsuda *et al.* (1997)	Bacterial	Catalyst of ice formation	PF00818
Nebulin	Pfuhl *et al.* (1996)	35	α-Helix (proposed)	None		Metazoa	PPI. Binds to F-actin	PF00880
Notch/lin-12	Wharton *et al.* (1985)	31	Unknown	None		Metazoa	PPI. Lateral inhibition of development processes	PF00066
Plectin	Wiche *et al.* (1991)	38	Unknown	None		Metazoa	PPI. Cytoskeleton. Cell adhesion. Antigens	PF00681
Spectrin	Speicher and Marchesi (1984)	106	Three-helix bundle	1cun	Pascual *et al.* (1997)	Metazoa	PPI. Cell shape. Cytoskeleton	PF00435
Annexin	Barton *et al.* (1991)	60	Five-helix bundle	1ain	Weng *et al.* (1993)	Eukaryotic	Regulatory. Membrane fusion. Exocytosis	PF00191
Flocculin	Watari *et al.* (1994)	45	Unknown	None		*S. cerevisiae*	Regulatory of flocculation	PF00624
Major vault protein	Vasu *et al.* (1993)	52	Unknown	None		Eukaryotic	Multidrug resistance	PF01505

Repeat families commonly represent either enzymes or nonenzymes, but rarely both.

Andrade M. A.; et al. *JSB* **2001**, 117-131.

Functions of the repeats





(a)The 40 sets with the highest occurence of repeats.
(b) the 40 sets with the lowest occurence.

The classes of proteins most likely to contain repeats are in fact predominantly unique to eukaryotes. (connective tissue proteins, cytoskeletal proteins, ribonucleoproteins, muscle proteins, brain and synaptic proteins, and cell adhesion proteins…)

Ancient protein classes that are shared among eukaryotes and prokaryotes appear among the proteins least likely to have repeats. (proteins fromcentral metabolic pathways, proteins involved in sugar metabolism, DNA synthesis, transport, amino acid biosynthesis, and photosynthesis)

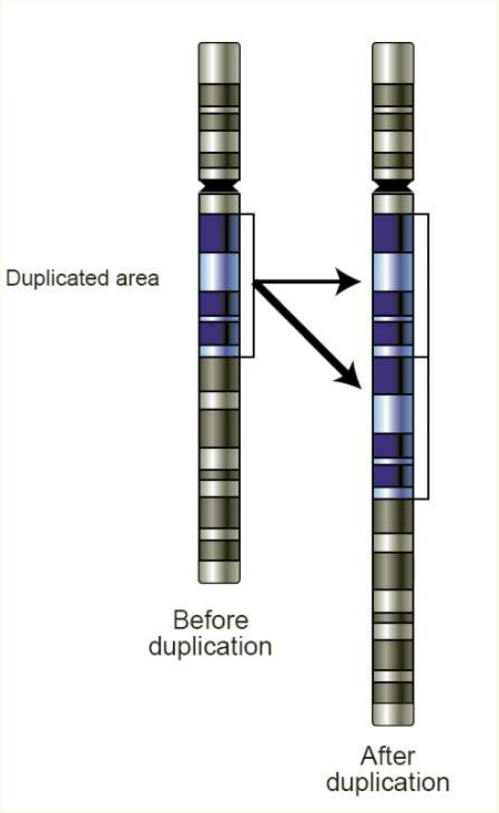
Most of repetitive regions formed after the prokaryotic-eukaryotic divergence.

Content

Protein diversification due to gene duplication

- The principle of modularity (duplication/amplification, recombination) operates at all levels of biological organization:
  - Full genomes
  - Entire Operons
  - Single Genes
  - (Domains and sub-domain sized fragments)

-> Effective path to increased complexity and a more adapted proteins, because the duplicated copy is free to evolve a novel function.



Duplication accompanied by fusion generates novel proteins

- Genetic Mechanisms:
  - Replication slippage
  - Illegitimate recombination
  - Crossover during sexual recombination (“exon shuffling”)
  - (Retro)-Transposition

-> These processes result in novel domain compositions, circularly permuted proteins (includes loss), or repetitive proteins

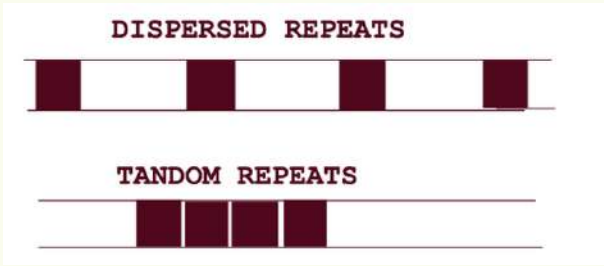
Repetitive nucleotide sequences

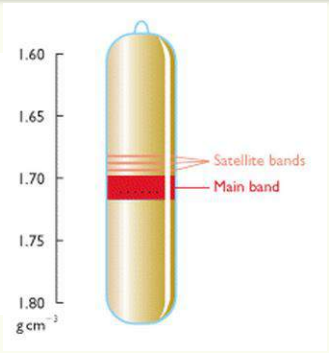
- Tandemly repeated DNA
  - Satellite sequences*

    (2-100bp)
  - Microsatellite sequences

    (arrays of 2-5bp nucleotide repeats)
  - Minisatellite sequences

    (tandem arrays of ~15bp repeats)
- Interspersed repeated DNA
- Transposable elements





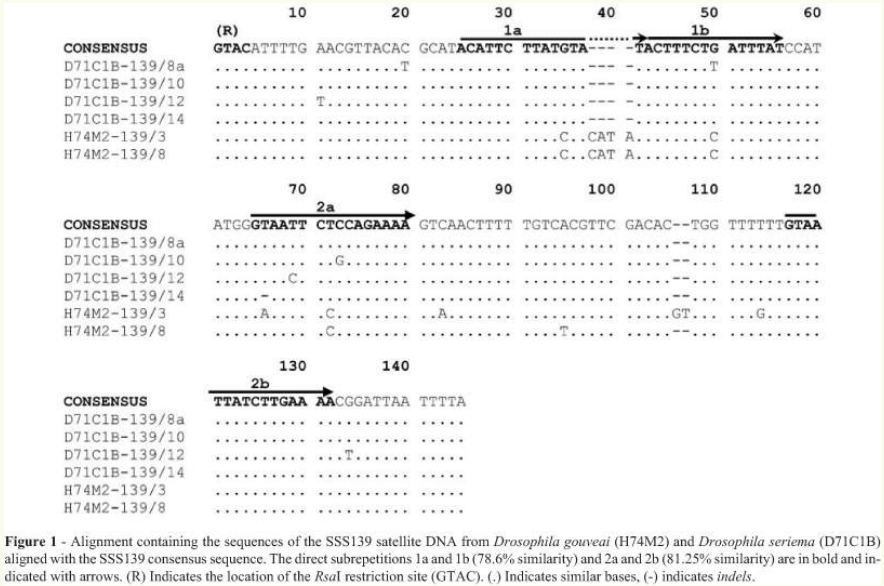
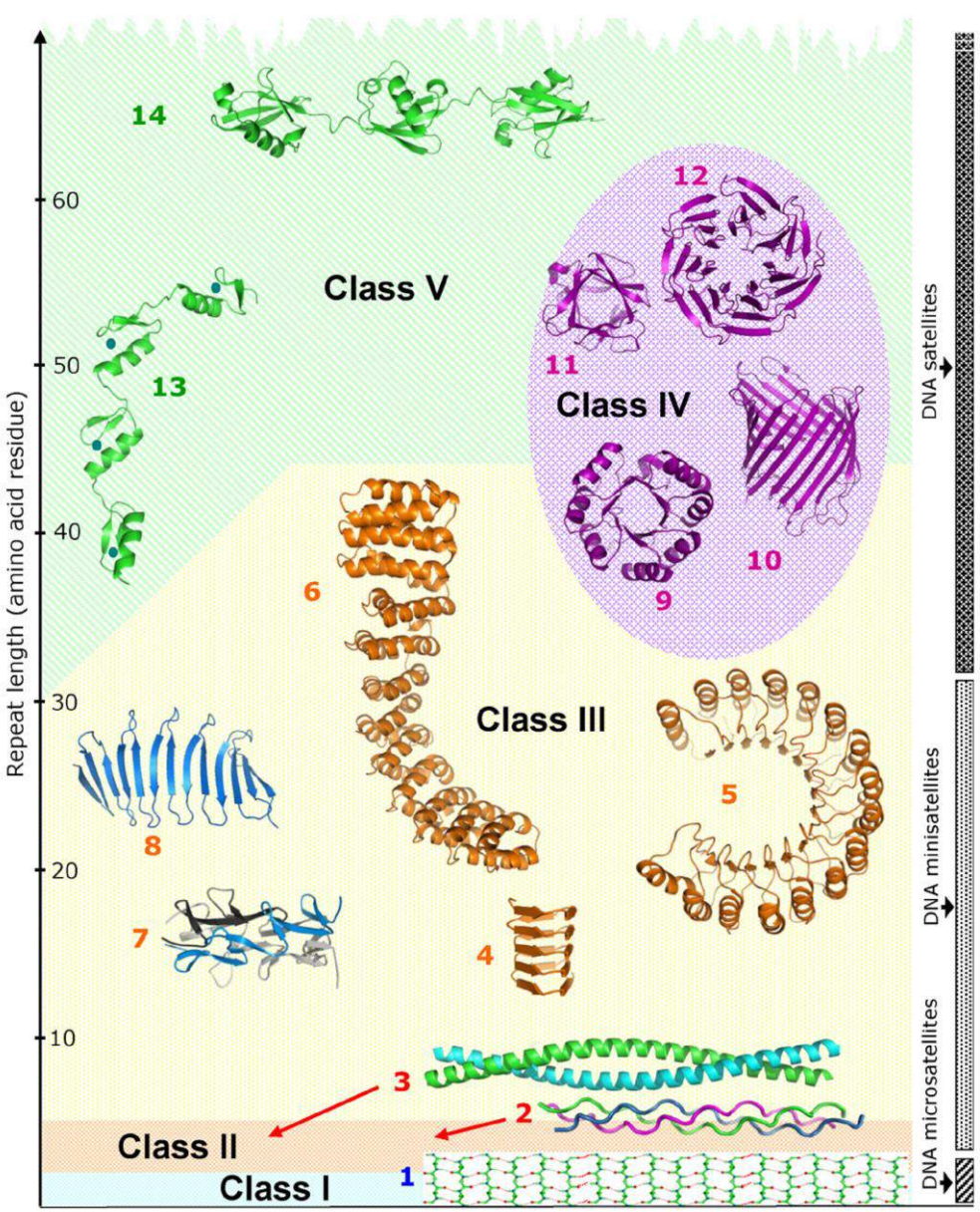


Figure 1 - Alignment containing the sequences of the SSS139 satellite DNA from *Drosophila gouveai* (H74M2) and *Drosophila seriema* (D71C1B) aligned with the SSS139 consensus sequence. The direct subrepetitions 1a and 1b (78.6% similarity) and 2a and 2b (81.25% similarity) are in bold and indicated with arrows. (R) Indicates the location of the *Rsa*I restriction site (GTAC). (.) Indicates similar bases, (-) indicates *indels*.

* The name "satellite DNA" refers to how repetitions of a short DNA sequence tend to produce a different frequency of the nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T), and thus have a different density from bulk DNA - such that they form a second or 'satellite' band when genomic DNA is separated on a densit gradient.

----Wikipedia

Franco F. F.; et al. *Genet. Mol. Biol.***2008**, 117-131.

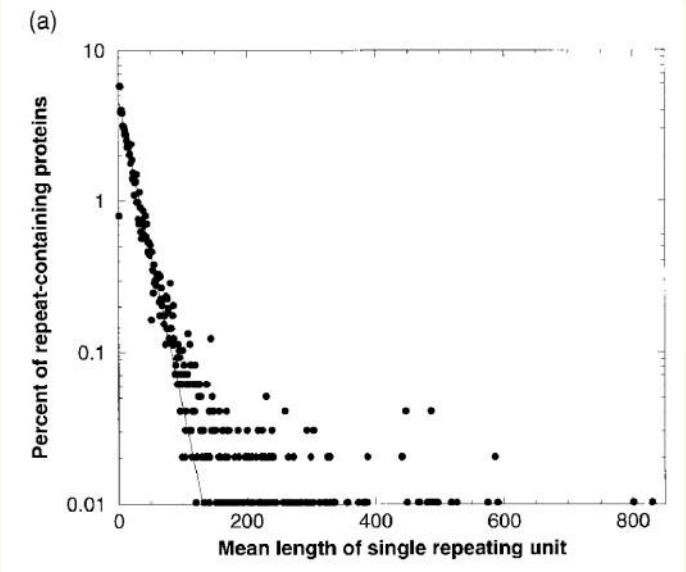


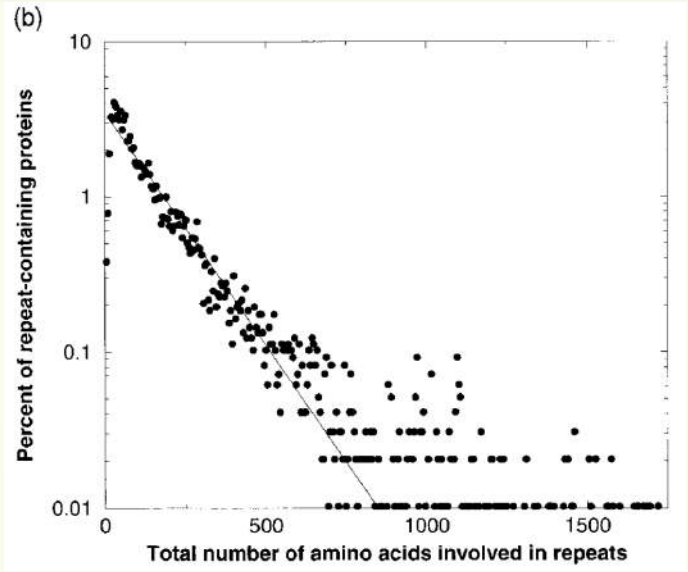
At the DNA level, the class I crystalline structures correspond to microsatellite repeats, class II and III structures are formed by minisatellites, and class IV and V repeats correspond to satellites.

The microsatellite, minisatellite and satellite loci may have different evolutionary mechanisms.

Kajave A. V. *J. str. Bio*. **2012**, 279-288.

Mechanism of repeat formation





$$P_{\text{recomb}} \propto e^{-\frac{n\langle E\rangle}{kT}}$$

E is the average energy of a nucleotide pair, and n is the fragment length.

(a) E = 0.016kT, on the order of 1/100 of the true melting energy per nucleotide pair

The mechanism producing repeats is far less sensitive to repeat length than would be expected if slippage and therefore duplex melting were the limiting factor. Instead, the result supports mechanisms such as recombination that show only weak length dependence.

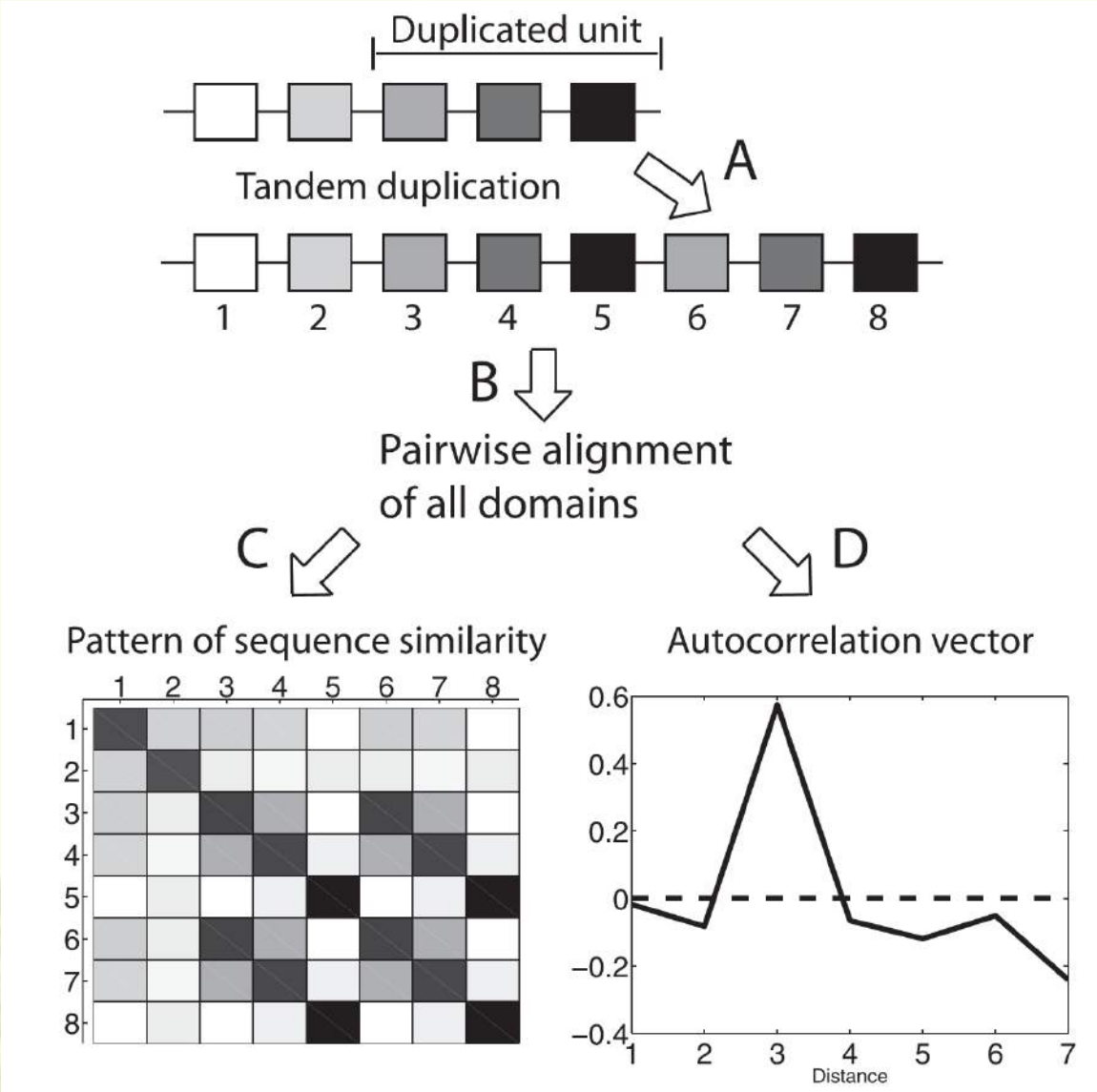
(b) E = 0.0023kT, approximately 1/10 of the energy required to form the single repeats

The expansion of repeats is much easier than the initial repeat formation.

Recombination rather than duplex melting or DNA hairpin formation may be the limiting mechanism underlying repeat formation.

Marcotte E. M.; et al. *JMB* **1998**, 151-160.

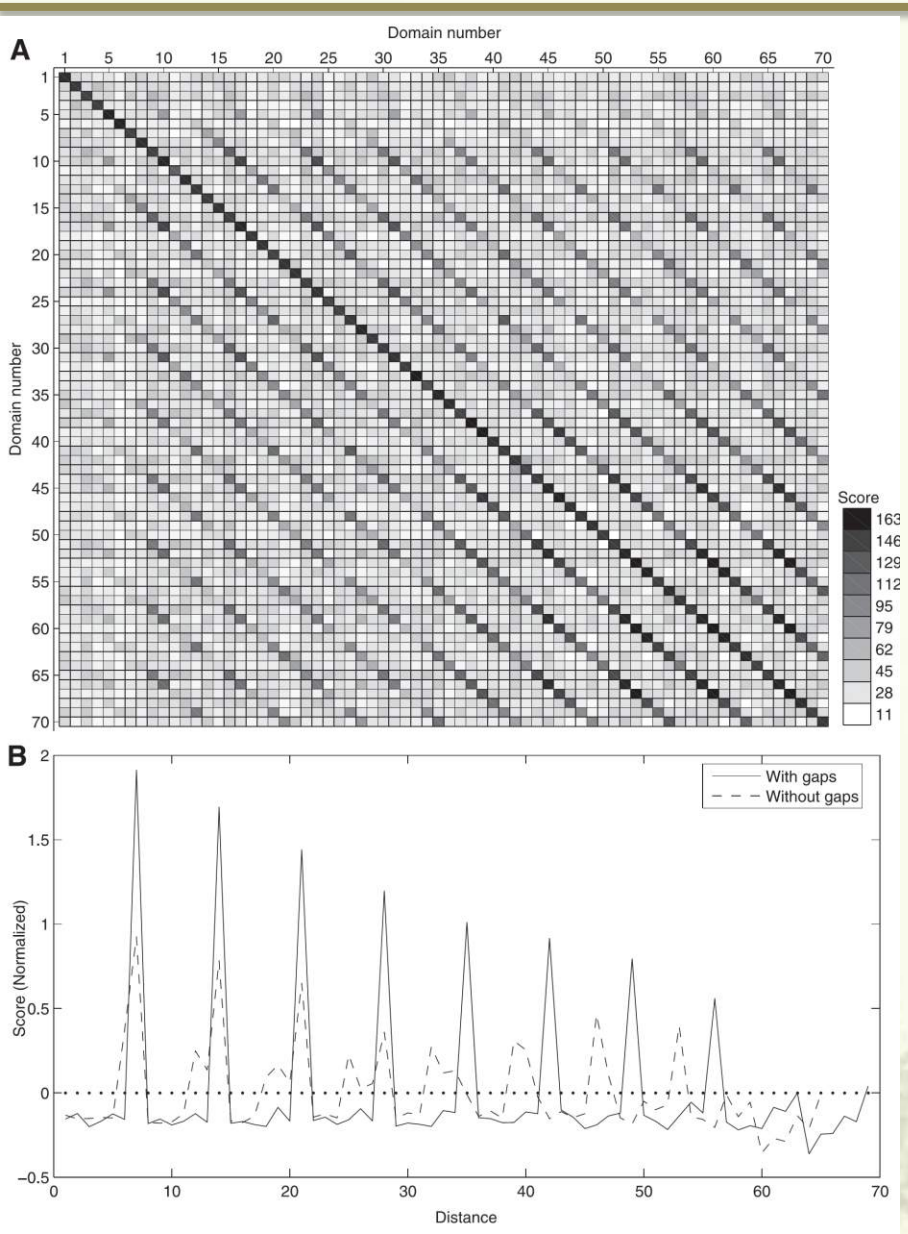
Expansion of protein domain repeats



The peaks in such a vector correspond to the most common sizes of duplication units in the evolution of the protein.

Bjorklund A. K.; et al. Plos Computational Biology. **2006**, 959-970.

Expansion of protein domain repeats

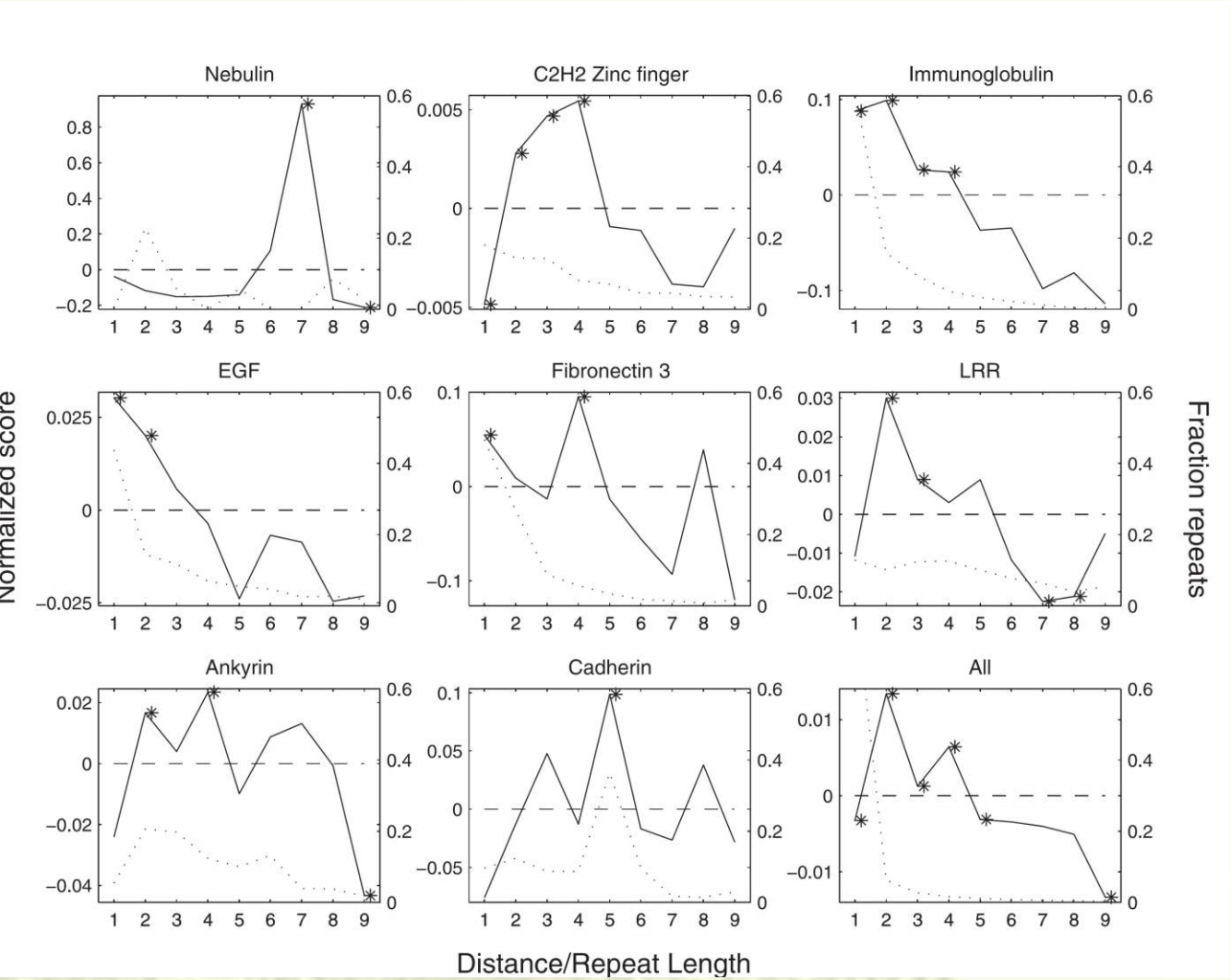


Pattern of internal domain duplications in the chicken protein ENSGALP00000020382, with 66 repeating Nebulin domains (Pfam)

The chicken nebulin protein has been duplicated with seven domains at a time, and similar patterns were seen in most nebulin proteins.

Bjorklund A. K.; et al. Plos Computational Biology. **2006**, 959-970.

Repeats duplication patterns



The domain repeats are most often created from the duplication of several domains at a time, while duplication of one domain appears to be less common.
The number of domains involved in each duplication event differs considerably within the domain families.

Bjorklund A. K.; et al. Plos Computational Biology. **2006**, 959-970.

Summary

- Repetitiveness of sequences and structures.
- About 25% of all eukaryotic proteins contain repeats.
- Internal repeats often correspond to structural or functional units in proteins
- Low sequence similarity, high structural symmetry, high function diversity.
- Repetitive regions are more common in eukaryotic proteins than in prokaryotic proteins. Most of them formed after the prokaryotic-eukaryotic divergence.
- Generation of internal repeats is an important mechanism for producing long proteins.
- Repetitve regions in proteins correspond to satellite repeated DNA.
- Recombination rather than duplex melting or DNA hairpin formation may be the limiting mechanism underlying repeat formation.
- Repeats are often expanded through duplications of several domains at a time, while the duplication of one domain is less common.

✓ The evolutionary mechanisms of repeats are not fully understood.


Thanks for attention!

Why is this stuff helpful for You?

About 25% of all eukaryotic proteins contain repeats

- Internal repeats often correspond to structural or functional units in proteins
- Therefore, methods capable of identifying diverged repeated segments

or domains at the sequence level can assist in:

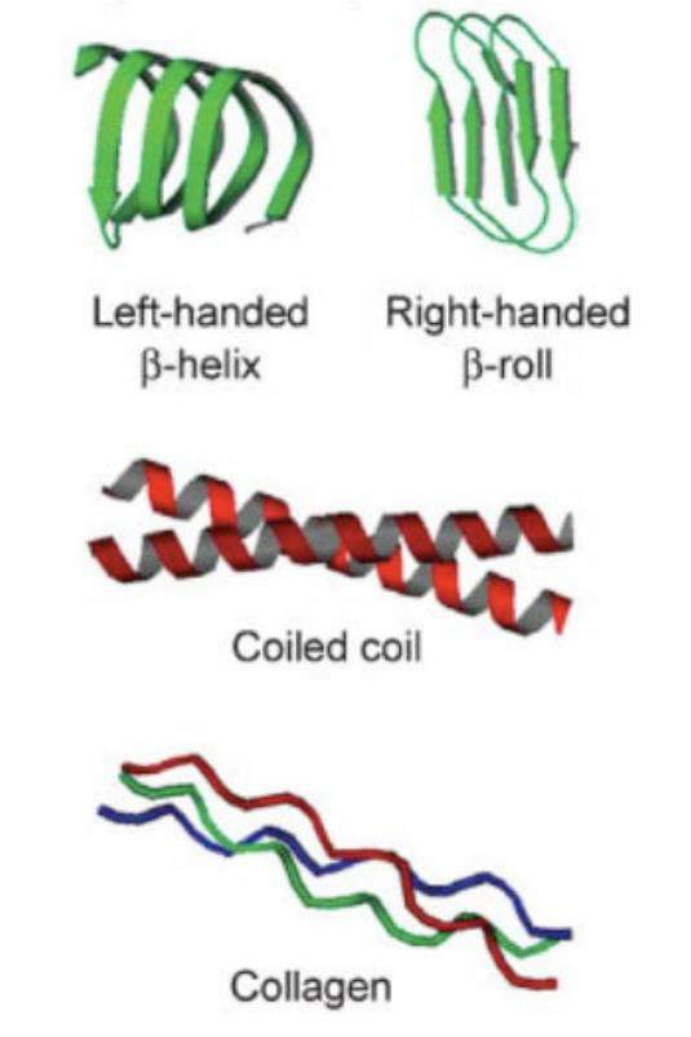
- predicting domain compositions
- predicting domain boundaries
- inferring hypotheses about function and mechanism
- investigating the evolution of the protein of interest

-> Design constructs in a smarter way and do more successful experiments

-> Get a deeper understanding of how the complexity of “modern”proteins evolved

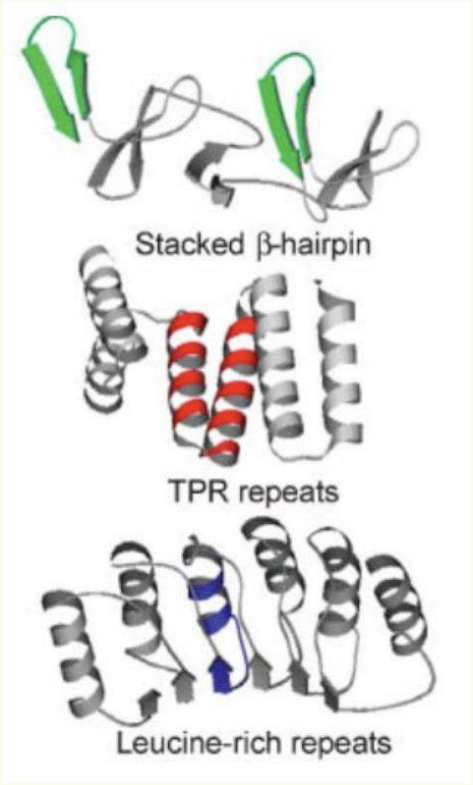
Fibrous proteins

abcdefg
LEEIVNQ
LNIYQSQ
VELIQQQ
MEAVRAT
ISELEIL
EKTLSDI
MESIKSQ
KNELEST
LQKMGEN
LRKITDI
MMKLSPQ
AEELLKK

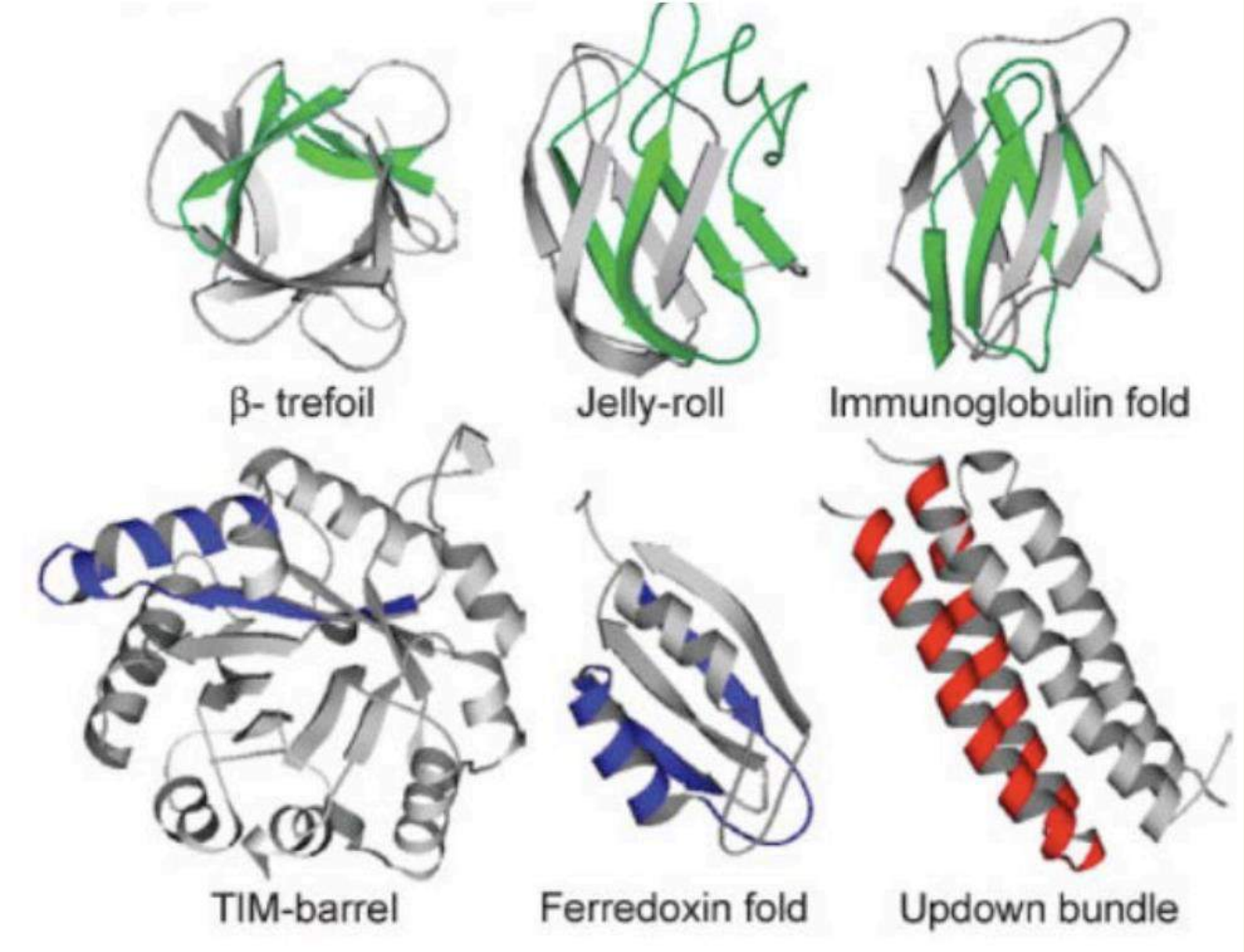


Monotonous repetition

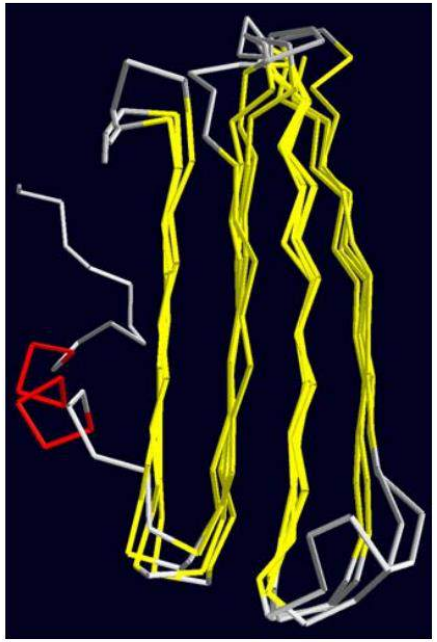
REYITSLDLSANELRDIDALSQKCCISVH
LEHLEKLELHQNALTSFPQQLCET
LKSLTHLDLHSNKFTSFPSYLLK
MSCIANLDVSRNDIGPSVVLDPTVK
CPTLKQFNLSYNQLSFVPENLTDV
VEKLEQLILEGNKISGICSPLR
LKELKILNLSKNHISSLSENFLEA
CPKVESFSARMNFLAAMPFL
PPSMTILKLSQNKFSCIPEAILN
LPHLRSLDMSSNDIQYLPGPAHWK
SLNLRELLFSHNQISILDLSEKAYL
WSRVEKLHLSHNKLKEIPPEIGC
LENLTSLDVSYNLELRSFPNEMGK
LSKIWDLPLDELHLNFDFK

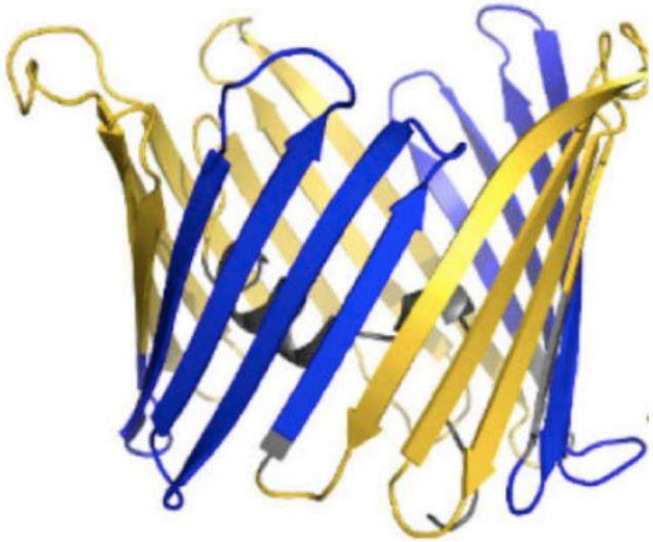


Some of the most frequently occurring protein folds have internal symmetry



Membrane proteins contain repetitive patterns





```
mavpptyadlgksardvftkgygf
---------------GLIKLDLKTKSE--NGLEFTSSGSANTETT---KVTGSLETKYRWTE
YGLTFTEKWNTD---NTLGTEITVEDQLARGLKLTFDSSFSPNTG---KKNAKIKTGYKR--
EHINLGCDMDFDIAGPSIRGALVLGY---EGWLAGYQMNFETAKS--RVTQSNFAVGYKT--
DEFQLHTNVNDG---TEFGGSIYQKVN--KKLETAVNLAWTAGNS---NTRFGIAAKYQID-
PDACFSAKVNNS---SLIGLGYTQTLK--PGIKLTLSALLDGKNVNAGGHKLGLGLEFQA
```

Repetitive patterns are found in different classes of proteins

- Fibrous proteins
- Solenoid proteins
- Membrane proteins
- Globular proteins

-> An evolutionary path from “simple”scaffold proteins to fully differentiated enzymes

-> Evolution of rather complex and well adapted molecules from smaller units

The secret of the evolutionary success of repetitive proteins

Problems: Only very few polypeptide sequences are capable of folding; protein folds are not very stable

- Repetition intrinsically promotes stability through the periodic recurrence of favorable interactions
- Modular reuse of already established components allows for a stepwise increase in complexity (emergence)

Table 1. Superfolds and the fraction of their residues contained in the supersecondary structure elements αα, ββ, βαβ[21]

Fold	Internal symmetry		Number of superfamilies (%)†	% Supersecondary structure content
	Sequence*	Structure		
β-trefoil	+	+	2 (0.1)	83
Jelly roll	–	+	17 (1.2)	47
Immunoglobin-like	–	+	55 (4.0)	67
TIM-barrel	+	+	28 (2.0)	82
Ferredoxin-like	+	+	65 (4.7)	38
Updown bundle	+	+	17 (1.2)	90
OB fold	–	–	16 (1.1)	77
UB-roll	–	–	16 (1.1)	55
Globin-like	–	–	4 (0.3)	88
Doubly wound	–	–	122 (8.8)	68
All superfolds			342 (24.7)	65
All folds			1386 (100)	62

ALKFTL AGHTKA VSSVKF SPNGE WLASSS ADK LIKIWG AYDG
KFEKTI SGHKLG ISDVAW SSDSN LLVSAS DDK TLKIWD VSSG
KCLKTL KGHSNY VFCCNF NPQSN LIVSGS FDE SVRIWD VKTG
KCLKTL PAHSDP VSAVHF NRDGS LIVSSS YDG LCRIWD TASG
QCLKTL IDDDNPP VSFVKF SPNGK YILAAT LDN TLKLWD YSKG
KCLKTY TGHKNEKY CIFANF SVTGGK WIVSGS EDN LVYIWN LQTK
EIVQKL QGHTDV VISTAC HPTEN IIASAA LENDK TIKLWK

VLLGRV PAHPDSR CWFLAW NPAGT LLASCG GDR RIRIWG TEGDSW
ICKSVL SEGHQRT VRKVAW SPCGN YLASAS FDA TTCIWK KNQDDF
ECVTTL EGHENE VKSVAW APSGN LLATCS RDK SVWVWE VDEED
EYECVS VLNSHTQD VKHVVW HPSQE LLASAS YDD TVKLYR EEEDDW
VCCATL EGHEST VWSLAF DPSGQ RLASCS DDR TVRIWR QYLPGNEQGVACSGSDPSW
KCICTL SGFHSRT IYDIAW CQLTG ALATAC GDD AIRVFQ EDPNSDPQQPTF
SLTAHL HQAHSQD VNCVAW NPKEPG LLASCS DDG EVAFWK