NON-GLOBULAR PROTEINS

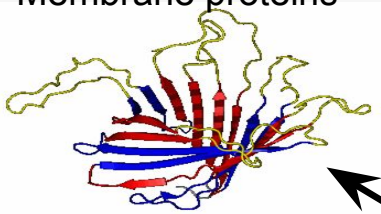
Master of Science in Data Science

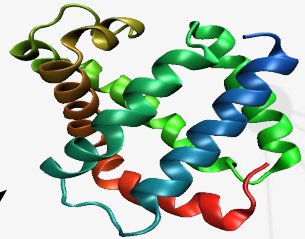
Damiano Piovesan

Non-globular proteins

DIPARTIMENTO MATEMATICA

Membrane proteins





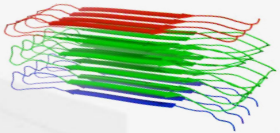
Globular proteins

Low complexity sequences

QQQQQQQQQQQQQ

Polypeptide chain

Intrinsically disordered proteins

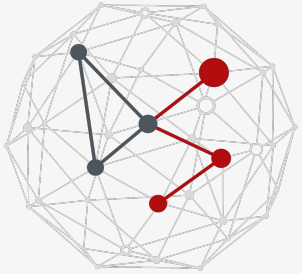


Aggregates, amyloids

Proteins with tandem repeats

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
MATEMATICA



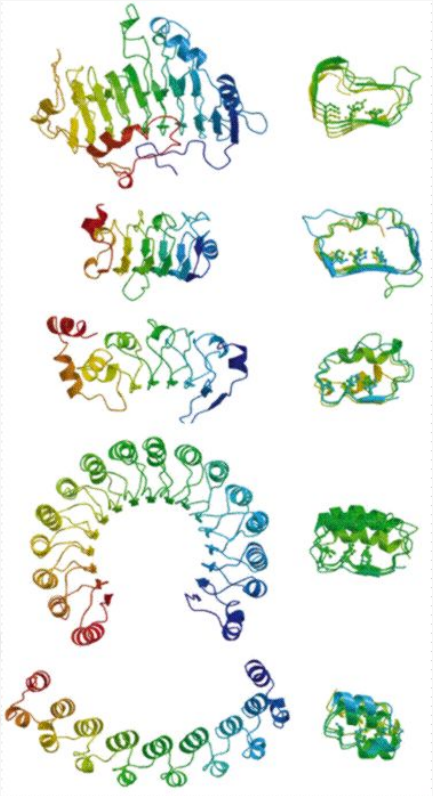
DATA SCIENCE
UNIVERSITY OF PADOVA

TANDEM REPEAT PROTEINS

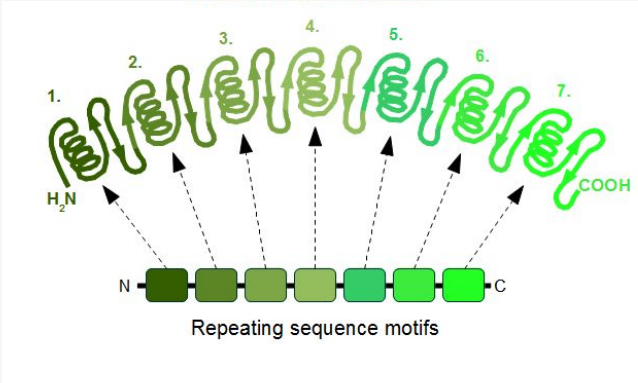
Master of Science in Data Science

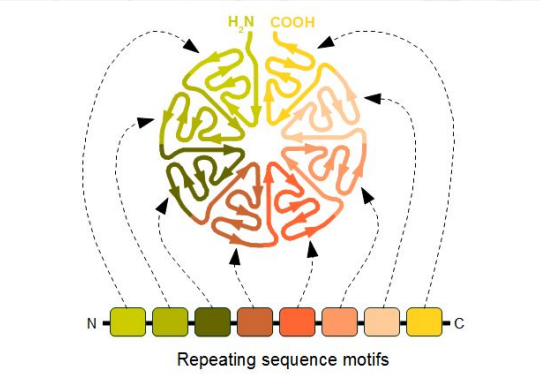
Damiano Piovesan

Tandem repeat proteins



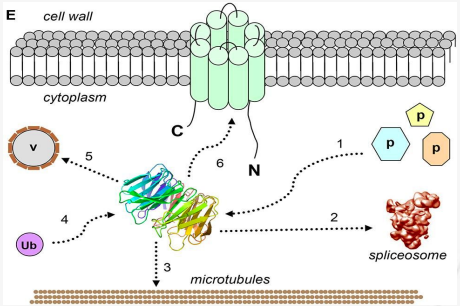
Tandem repeat proteins are the product of minimal folds from the repetition of simpler units



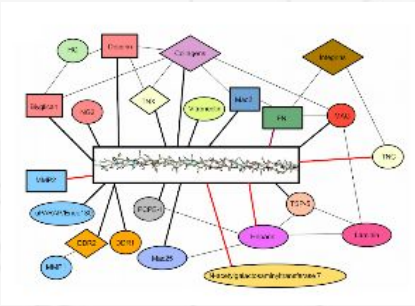


Function

- Extended structure, large surface
- Platform for protein-protein interactions

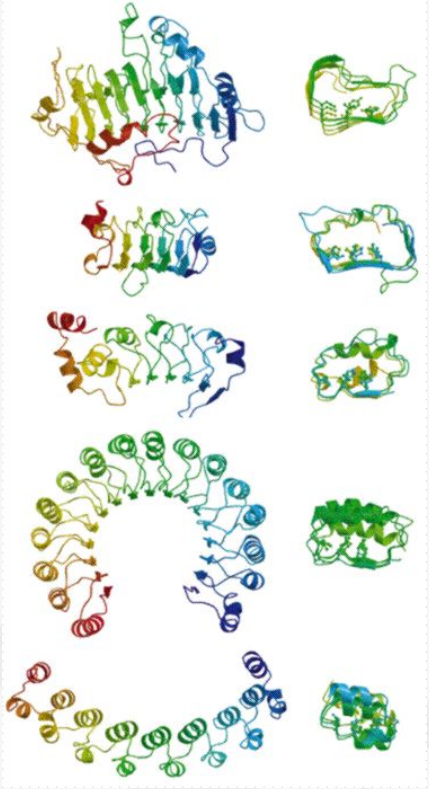


(G Guerreiro et al. Front.Plant.Sci. 2015)



(L Paladin et al, FEBS Lett. 2015)

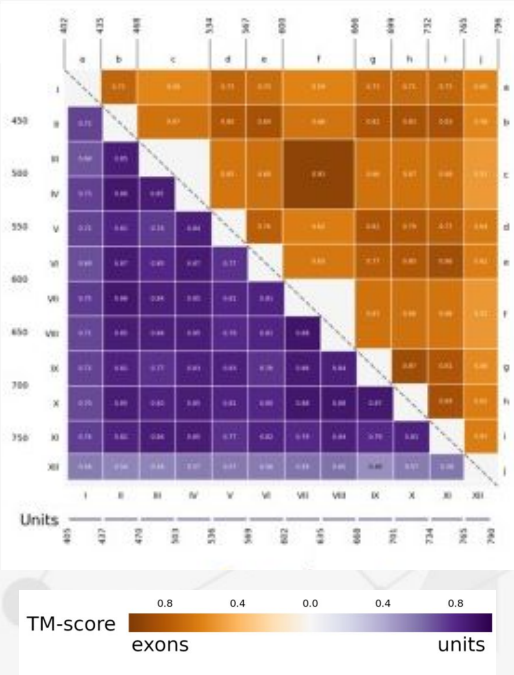
Evolution



- The majority of repeats are found in eukaryotes, probably due to *duplications*
- Many repeats are involved in diseases (e.g. *Huntington* or "*fragile X syndrome*"), especially when they give rise to fibril formation (e.g. *BSE*)
- Repeated sequences evolve more rapidly than the non-repeated ones

Repeat evolution

Exon boundaries of 3TSR chain E

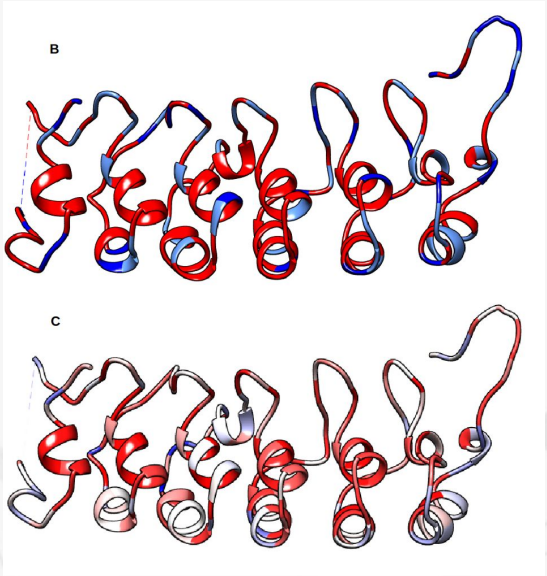


Paladin et al., JSB 2020

- Two definition of the units
  - Structural symmetry (RepeatsDB)
  - Exon boundaries
- Compare the two definitions to detect duplication events
  - Same boundaries
  - Exon duplication (2:1, ...)
  - Exon/RepeatsDB mismatch

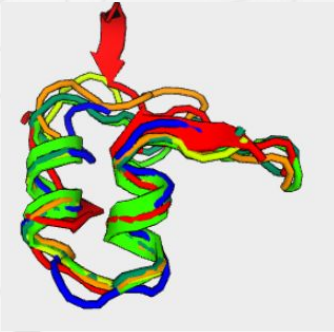
Properties

- High sequence variability
- Buried residues are more conserved



Hydrophobicity

- hydrophobic
- hydrophilic

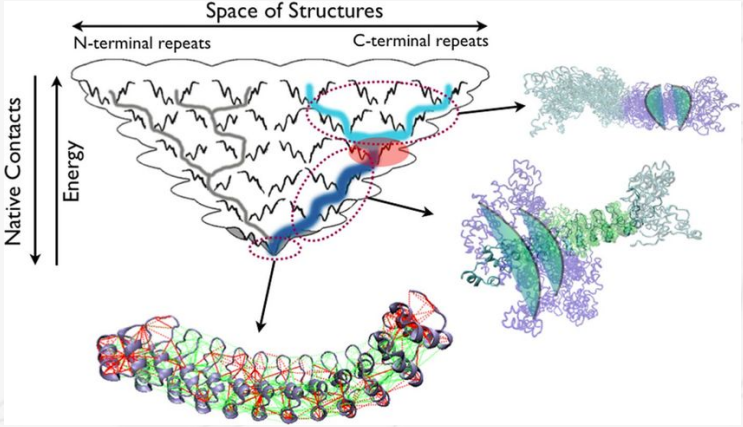


Conservation

- Conserved
- Variable

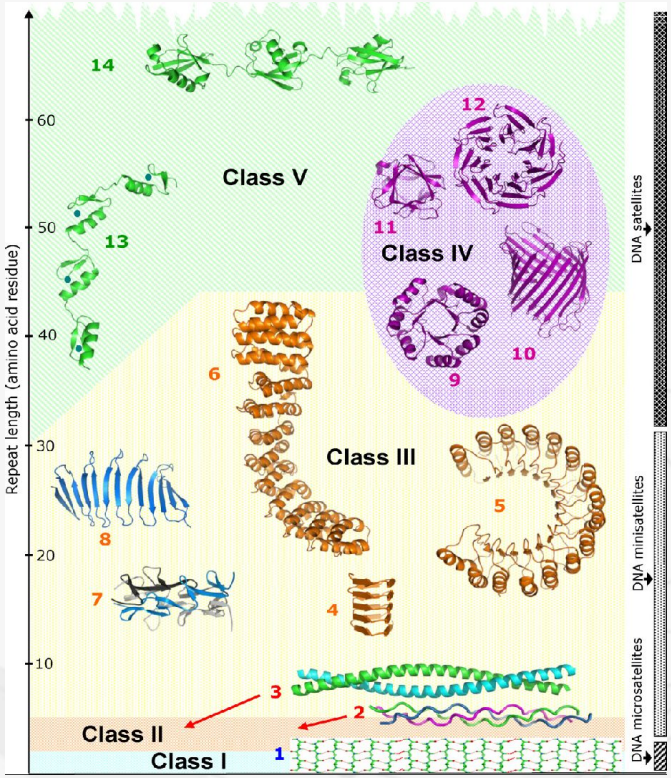
Folding

- Linear folding
- Local stabilizing interaction
- The first and last units are conserved and guarantee stability of the whole protein fold
- Central repeats can be deleted without leading to protein misfolding



(Ferreiro and Wolynes, PNAS, 2014)

Classification



(Kajava, J Struct Biol 2012)

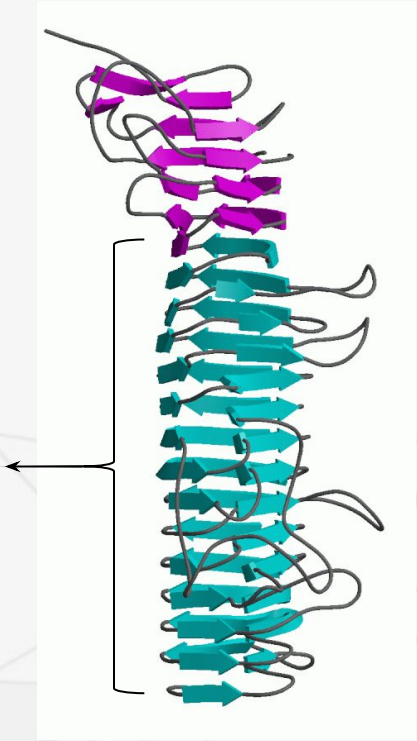
Broad classification based on periodicity

1. Aggregates
2. Coiled-coils and Collagen
3. Solenoids
4. Toroids
5. Beads on a string

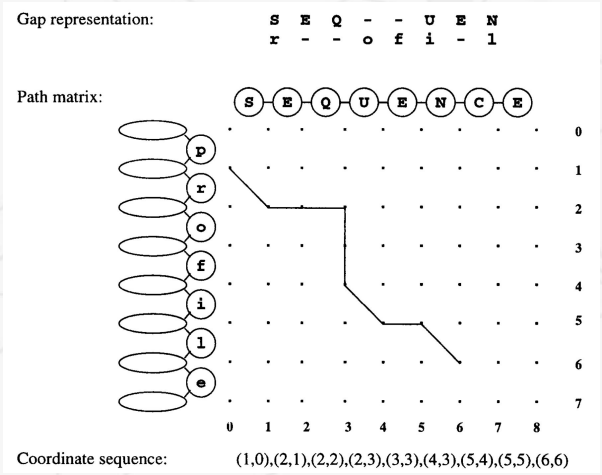
Pftools - Generalized sequence profiles

DIPARTIMENTO MATEMATICA

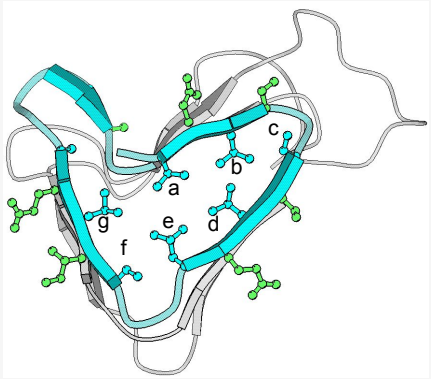
```
GILLENPAAELQFRNGSVTSSGQLSDDGIRRFLG
    TVTVKAGKLVADHATLANVGDTWDDDGI
      ALYVAGEQAQASIADSTLQGAG
       GVQIERGANVTVQRSAIVDG
GLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPA
     AVSVLGASELTLDGGHITGGRAA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWY
     GVDVSGSSVELAQSIVEAPELGA
     AIRVGRGARVTVSGGSLSAPHGN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

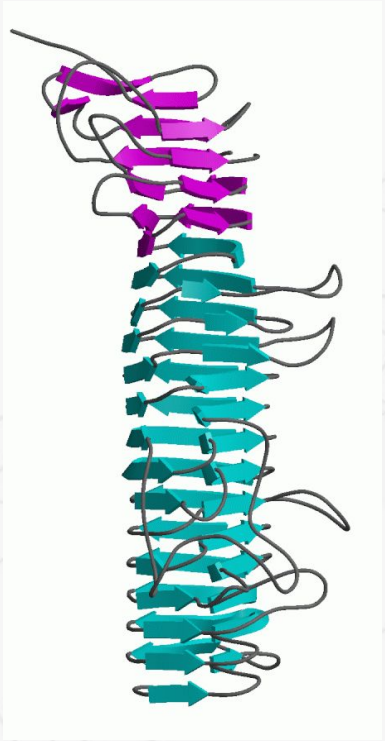


Pertactin from *Bordetella pertussis*



Bucher et al., 1996, Comput. Chem. 20, 3-23





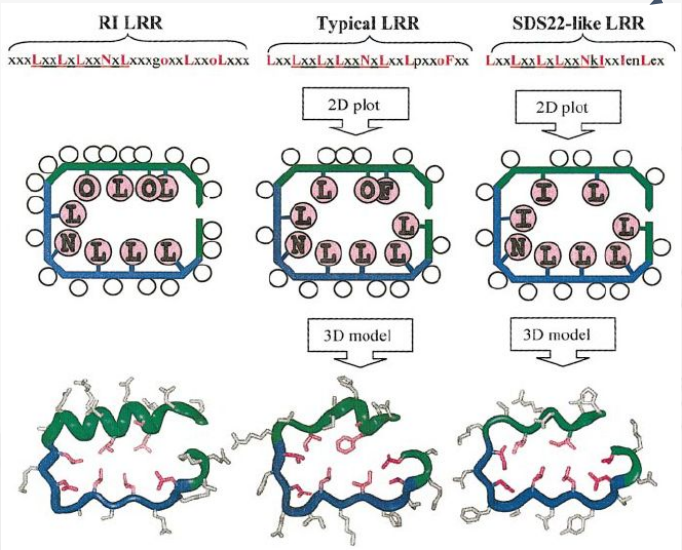
```
a b             c d e f g
↓ ↓             ↓ ↓ ↓ ↓ ↓
GILLENP----------AAELQFRN-GSVTS-SGQLSDDGIRRFLG
TVTVKA-----------GKLVADH-ATLAN-VGDTWDDDGI
ALYVAGEQ----------AQASIAD-STLQG-AG
GVQIERG----------ANVTVQR-SAIV-DG
GLHIGALQSLQPEDLPP-SRVVLRD-TNVTA-VPASGAPA
AVSVLGA-----------SELTLDG-GHITG-GRAA
GVAAMQG----------AVVHLQR-ATIR-RGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWY
GVDVSG-----------SSVELAQ-SIVEA-PELGA
AIRVGRG----------ARVTVSG-GSLSA-PHGN
VIETGGARRFAPQAAP--LSITLQAGAHA-QGKA
LLYRVLPEP---------VKLTLTGGADA-QG
DIVATELPSIPGTSIGP-LDVALASQARW-TG
●●x●xxx----------●x●x●xx-●x●-xx
```

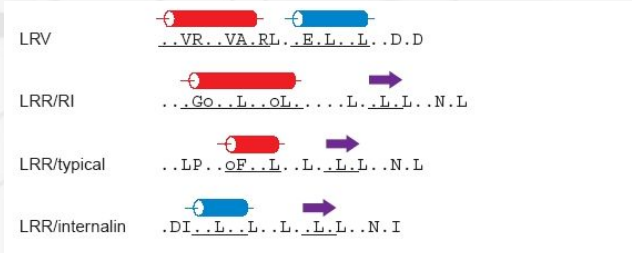
Bucher et al., 1996, Comput. Chem. 20, 3-23

Repeats prediction and modelling

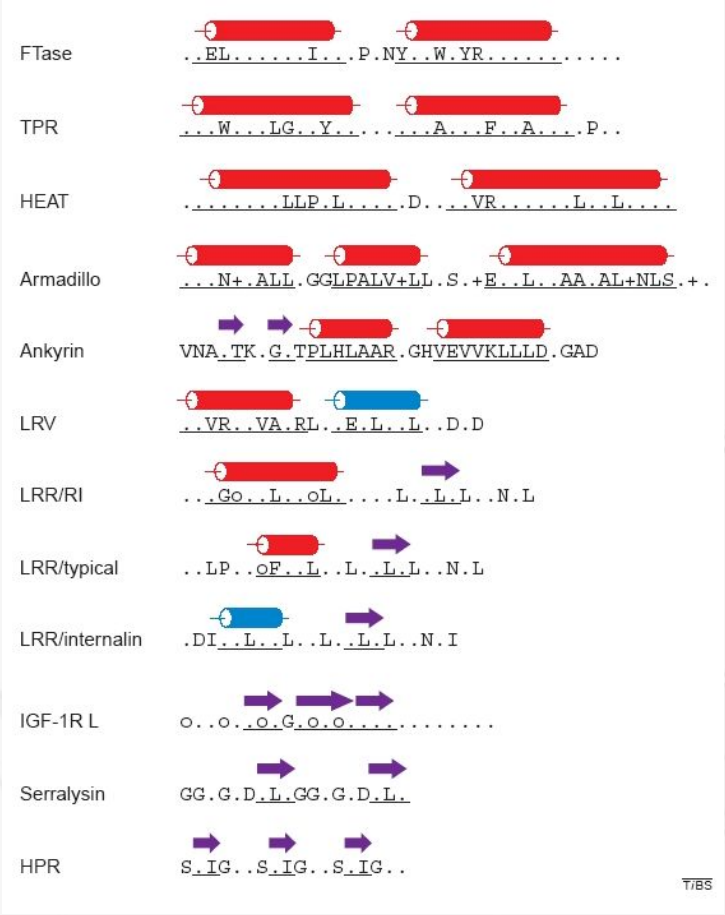
- Modelling is made easier by the repeats modularity, but it heavily depends on the alignments quality
- Standard methods do not exist (yet)

Leucine Rich Repeat





Consensus patterns



Consensus sequences in solenoid proteins

- red cylinders → α-helices
- blue cylinders → $_{3\text{-}10}$-helices
- magenta arrows → beta strands
- + → positively charged amino acid
- o → hydrophobic amino acid

Some statistics for LRR

DIPARTIMENTO MATEMATICA

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Section	Description	Number of hits
Pfam	Text fields for Pfam entries	577
Seq_info	Sequence description and species fields	1856
Pdb	HEADER and TITLE records from PDB entries	127
GO	Gene ontology IDs and terms	0
Interpro	InterPro entry abstracts	661

ID	Accession	Type	Number of sequences		Average length	Average %id	Average coverage	Has 3D	Change status	Description
			Seed	Full						
LRR19-TM	PF15176	Domain	9	67	103.90	48	43.54		New	Leucine-rich repeat family 19 TM domain
LRRC37AB_C	PF14914	Family	5	149	137.50	61	14.63		New	LRRC37A/B like protein 1 C-terminal domain
LRRCT	PF01463	Family	59	1112	25.80	34	6.10	✓	Changed	Leucine rich repeat C-terminal domain
LRRNT	PF01462	Family	26	4081	29.00	41	7.50	✓	Changed	Leucine rich repeat N-terminal domain
LRRNT_2	PF08263	Family	101	5573	41.00	31	5.53	✓	Changed	Leucine rich repeat N-terminal domain
LRR_1	PF00560	Repeat	2414	25597	23.30	33	5.68	✓	Changed	Leucine Rich Repeat
LRR_2	PF07723	Repeat	161	389	25.70	36	6.48		Changed	Leucine Rich Repeat
LRR_3	PF07725	Repeat	74	609	19.90	51	2.28		Changed	Leucine Rich Repeat
LRR_4	PF12799	Family	569	21961	44.60	29	11.83	✓	Changed	Leucine Rich repeats (2 copies)
LRR_5	PF13306	Family	205	12012	113.30	19	48.31	✓	Changed	Leucine rich repeats (6 copies)
LRR_6	PF13516	Repeat	889	30484	24.90	27	9.82	✓	Changed	Leucine Rich repeat
LRR_7	PF13504	Repeat	1228	2943	18.00	36	3.46	✓	Changed	Leucine rich repeat
LRR_8	PF13855	Repeat	93	70869	59.50	28	22.98	✓	Changed	Leucine rich repeat
LRR_9	PF14580	Repeat	6	1482	148.50	27	32.21	✓	New	Leucine-rich repeat
LRR_adjacent	PF08191	Family	29	1133	56.90	48	9.22	✓	Changed	LRR adjacent

BioComputing

Detection of tandem repeats from sequence

DIPARTIMENTO MATEMATICA

Type of method	*ab initio / a priori*	Properties of repeats	Rapidity
Fourier Transform analysis REPPER (Gruber et al., 2005)	*ab initio*	Long arrays without indels	+
Short string extension algorithms XSTREAM (Newman and Cooper, 2007) T-REKS (Jorda and Kajava, 2009)	*ab initio*	With indels and less than 15-20 residues	+
Sequence-sequence alignment RADAR (Heger and Holm, 2000) TRUST (Szklarczyk and Heringa, 2004)	*ab initio*	More than 15 residues. With indels	-
Hidden Markov Models (HHMs) or sequence profiles PFAM(Sonnhammer et al., 1998) SMART ((Schultz et al., 1998) BISMM library (Kajava and Steven, 2006)	requires *a priori* information	Long and strongly imperfect repeats	-
HMM-HMM or profile-profile comparisons HHREPID (Biegert and Soding, 2008)	*ab initio*	Long and strongly imperfect repeats	-
Sequence profile - Fourier and wavelet transforms REPETITA (Marsella et al., 2009), WAVELET (Vo et al., 2010)	requires *a priori* information	Long and strongly imperfect repeats	-

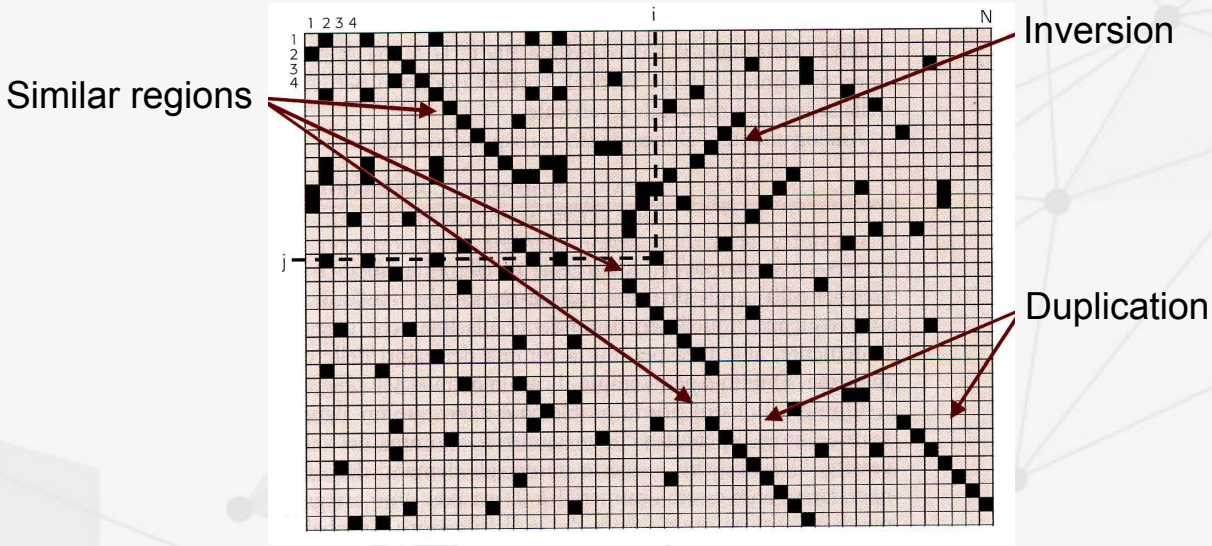
Kajava, (2012) *J. Struct. Biol.*

…from structure

Coordinate profile RAPHAEL (Walsh et al., 2012)	*ab initio*	Repeat period	+
Coordinates ReUPred / RepeatsDB-lite (Hirsh et al., 2016; Paladin et al. 2018)	*requires a priori* information	Repeated units and class	-

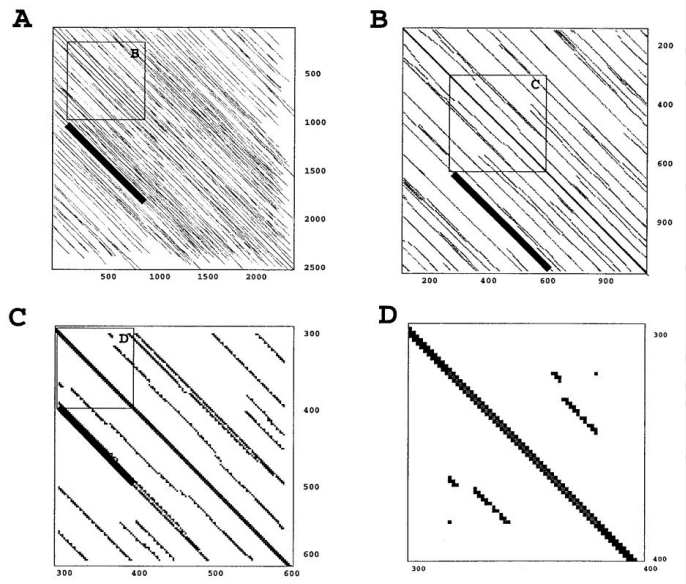
Repeats prediction

- **Dot plot** - Compare the sequence against itself, repeats can be identified as **duplications**
- Exact detection of repeats is hard if they are **degenerated**, units boundaries are ambiguous



RADAR - Rapid Automatic Detection and Alignment of Repeats

1. Compute **self-alignment** matrix
2. Choose **initial repeat** (highest local diagonal score, distance from main diagonal)
3. Iteratively identify **similar repeats** (unit Vs sequence)
4. Compute repeats unit **boundaries** (place units around a cylinder)
5. Remove not significative repeats (**Z-score**)
6. If **shorter repeats** can be determined, go to step 2
7. If the sequence contains **other repeats**, go to step 2

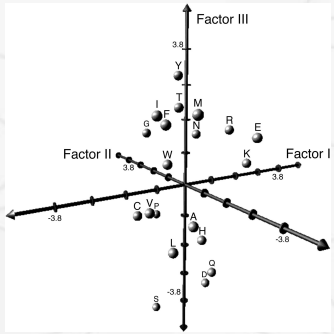


(Heger & Holm, 2000)

REPETITA

DIPARTIMENTO MATEMATICA

- Map sequence residues to the five features of the **Atchley's scale** *(a)*
  - Summarize variability among amino acids
  - Polarity, secondary structure, molecular volume, codon diversity, electrostatics charge
- Transform the sequence into **profiles**
  - **Probability *p*** of finding **amino acid *X*** in the Multiple Sequence Alignment (MSA) at **position *k***



Solving the protein sequence metric problem. Atchley et al., 2005, PNAS

$$f_a^k = \sum_X g_a(X) p_k(X) \qquad \text{with} \sum_X p_k(X) = 1 \qquad (k = 1, \ldots, N)$$

Atchley function — Sequence profile

REPETITA

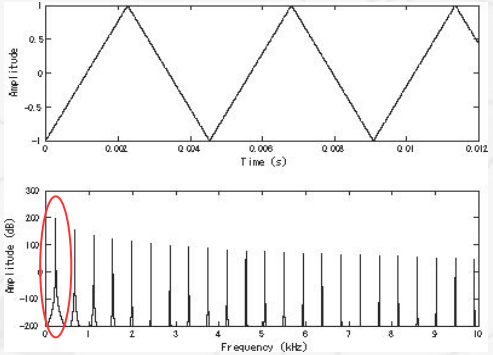
DIPARTIMENTO MATEMATICA

- Map to the frequency space using **discrete Fourier Transform (DFT)** to each of the five Atchley sequences, for each possible **fragment of size *n = 0, …, N - 1***, where N is the length of the sequence (F_a^n)
- Calculate the **angular frequency ω** and **period *T***

$$F_a = F\,[f_a]$$

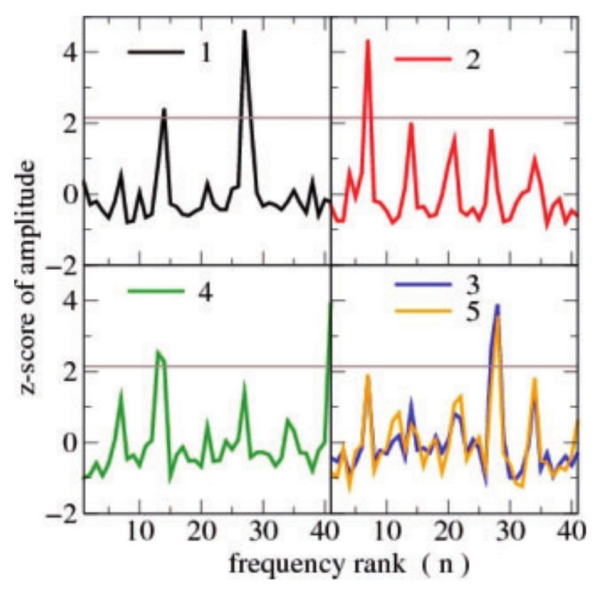
$$F_a^n = \frac{1}{\sqrt{2\pi N}} \sum_{k=0}^{N-1} f_a^k e^{-2\pi ik(n/N)}$$

$$\omega_n = \frac{2\pi n}{N} \qquad T_n = \frac{2\pi}{\omega_n} = \frac{N}{n}$$



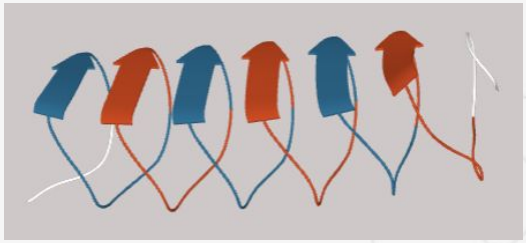
A periodic waveform (triangle wave) and its frequency spectrum, showing a "fundamental" frequency at 220 Hz followed by multiples (harmonics) of 220 Hz. (WikiPedia)

REPETITA



Atchley's scale profiles

PDB 1EZG



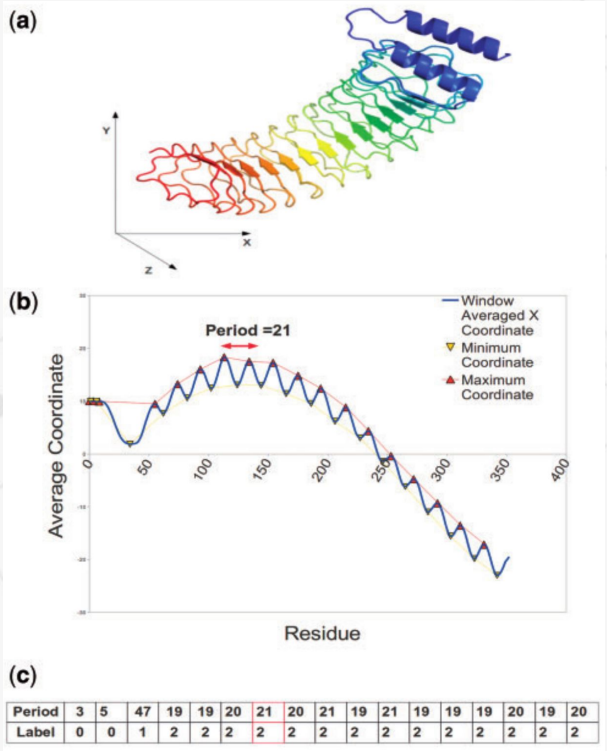
```
1   QCTGGADCTSCTGAC  TGCGNCPNAVTCTNS  QHCVKANTCTGSTDC
46  NTAQTCTNSKDCFEA  NTCTDSTNCYKATAC  TNSSGCPGH
```

- Peaks around the **fundamental frequency rank *n*=7** *(14, 21, 28, 35)*
- Corresponding to a **period *T* = *12***

RAPHAEL

DIPARTIMENTO MATEMATICA

- Coordinate profiles
  - Average C-alpha coordinates (windows 6 and 3)
- Labelling / period definition
  - Consecutive local maxima
- Scoring
  - Randomization (translation / rotation)
  - Frequent adjacent periods
  - Frequent periods separated by rarely occurring periods

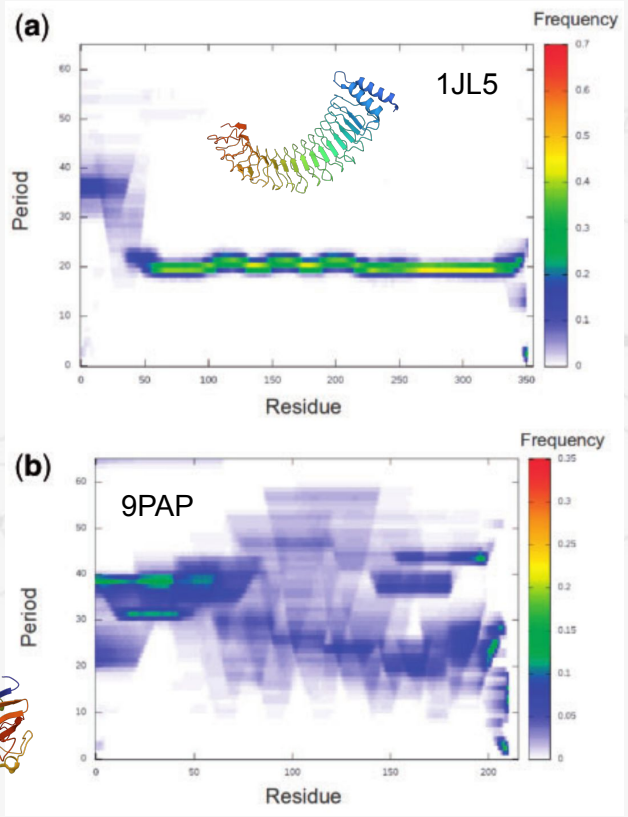


Period	3	5	47	19	19	20	21	20	21	19	21	19	19	19	20	19	20
Label	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2

RAPHAEL - Detect the period

DIPARTIMENTO MATEMATICA

- Period matrix
  - The frequency of period *k* of residue *j* given 200 random rotations and translations along each coordinate frame x, y and z
  - Y → Period length (max 60)
  - X → Period start index in the sequence (j)



RAPHAEL - Detect solenoids

DIPARTIMENTO MATEMATICA

- Solenoids, are usually elongated
  - Minimum distance between the first 40 residues and the last 40 residues
- Contacting residues in solenoids should have low sequence separation relative to globular proteins
  - The number of contacting residues at a sequence separation >55
- There should be regularity in sequence among the contacting residues (conversely, there should be large variance for non-solenoids)
  - Variance of the residue-wise contact order → how regular the sequence separation is for contacting residues

RepeatsDB

DIPARTIMENTO MATEMATICA

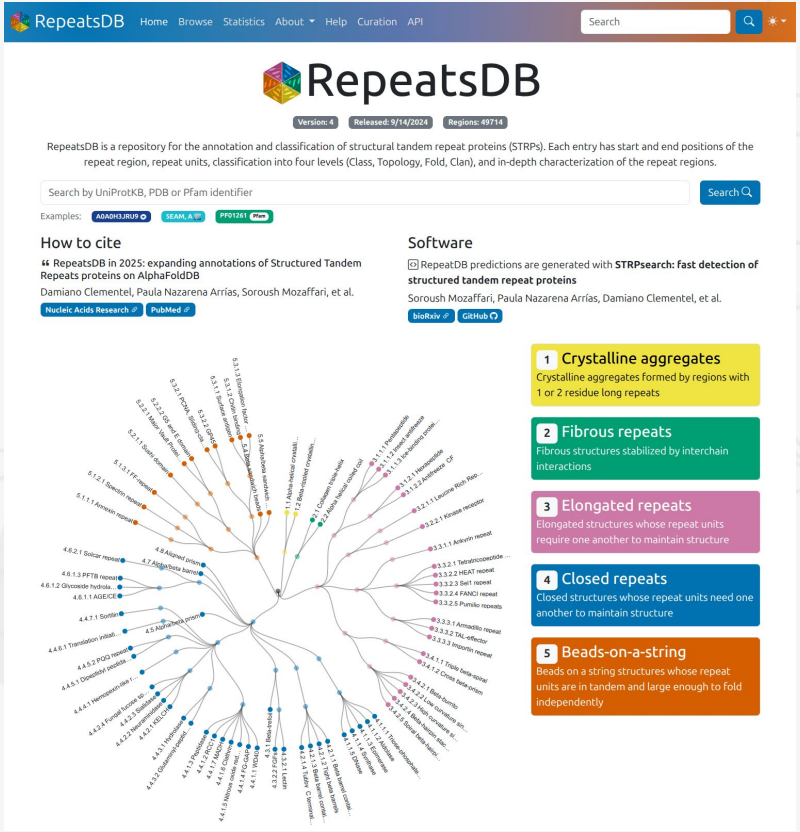
A database of repeat proteins

- Manual curation provide a high quality gold standard of repeat protein annotations
- Integration of RepeatsDB-lite predictions

In collaboration with

- Dr. Andrey V. Kajava (CNRS Montpellier)
- Prof. Miguel Andrade (University of Mainz)

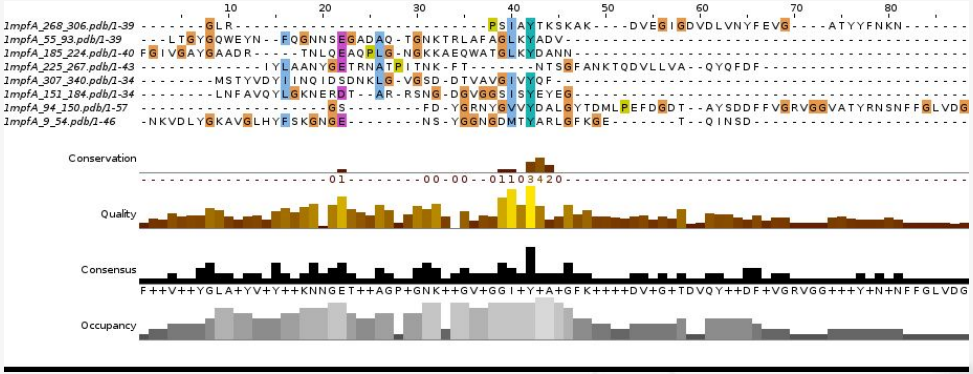
https://repeatsdb.org



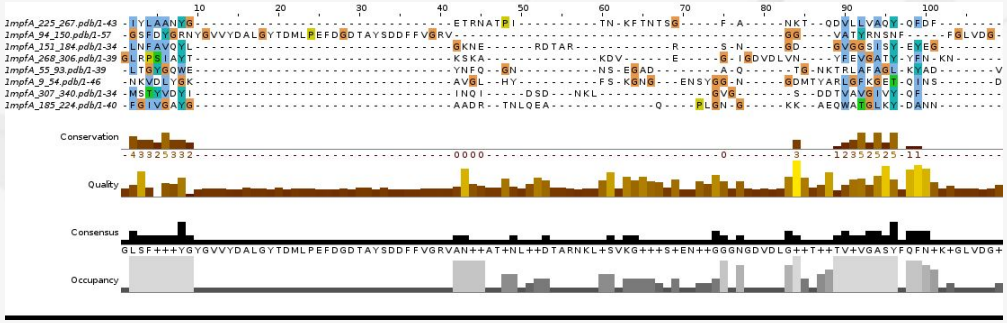
Structure Vs sequence alignments

DIPARTIMENTO MATEMATICA

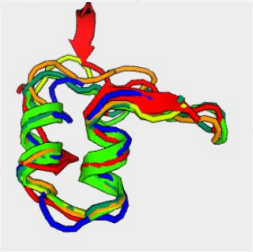
Sequence alignment



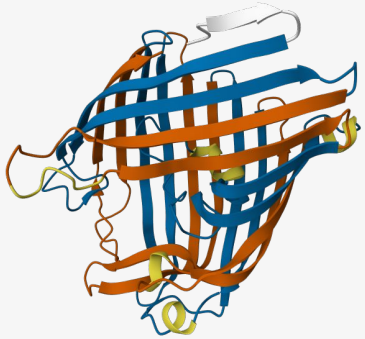
Structure alignment



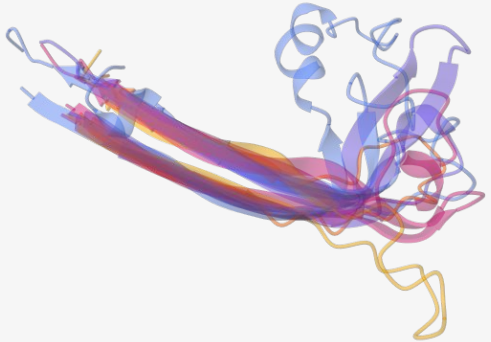
- Structure alignment
  - Higher conservation
  - Insertions
  - Biological significance



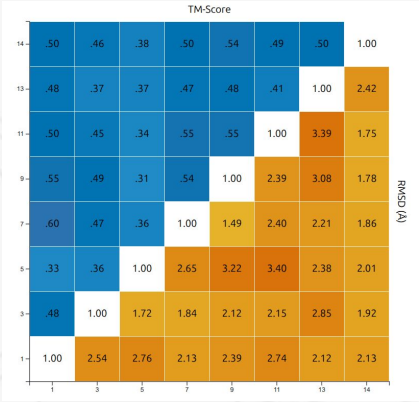
Intra-region unit similarity in RepeatsDB



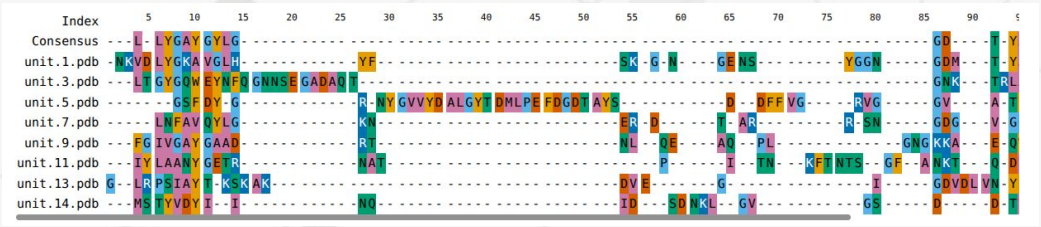
Matrix porin outer membrane protein f
Escherichia coli (strain K12)
PDB 1MPF, chain A



Multiple structural alignment
(Mustang)



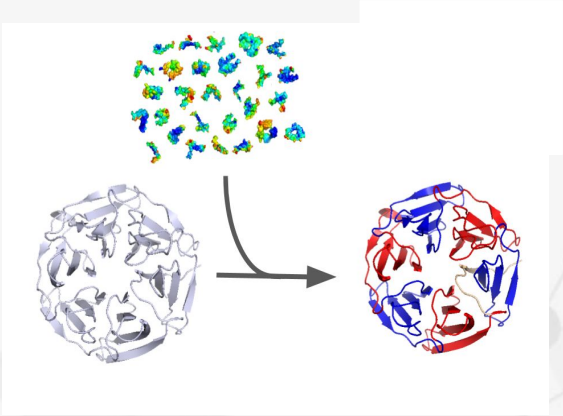
All-Vs-All pairwise structural alignment
(TM-align)



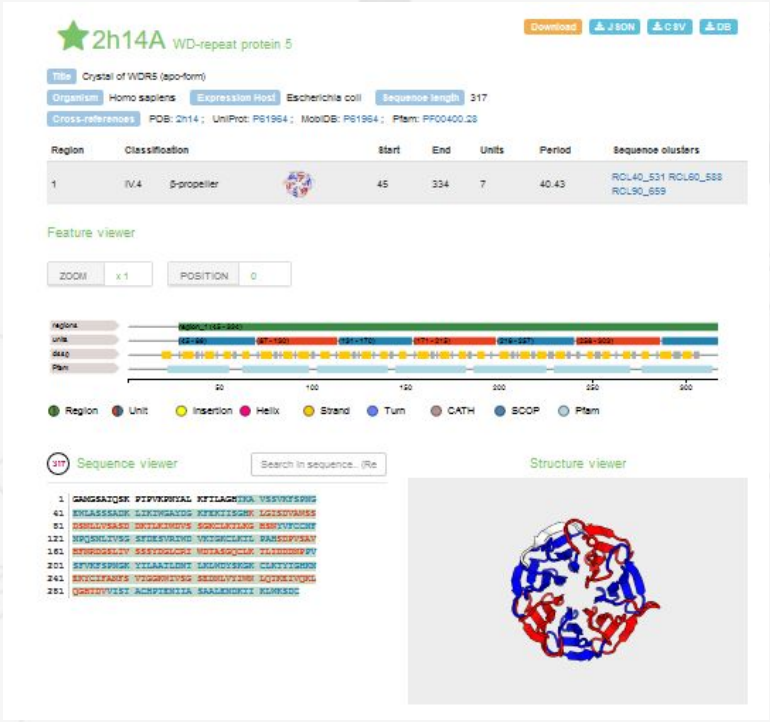
ReUPred / RepeatsDB-lite

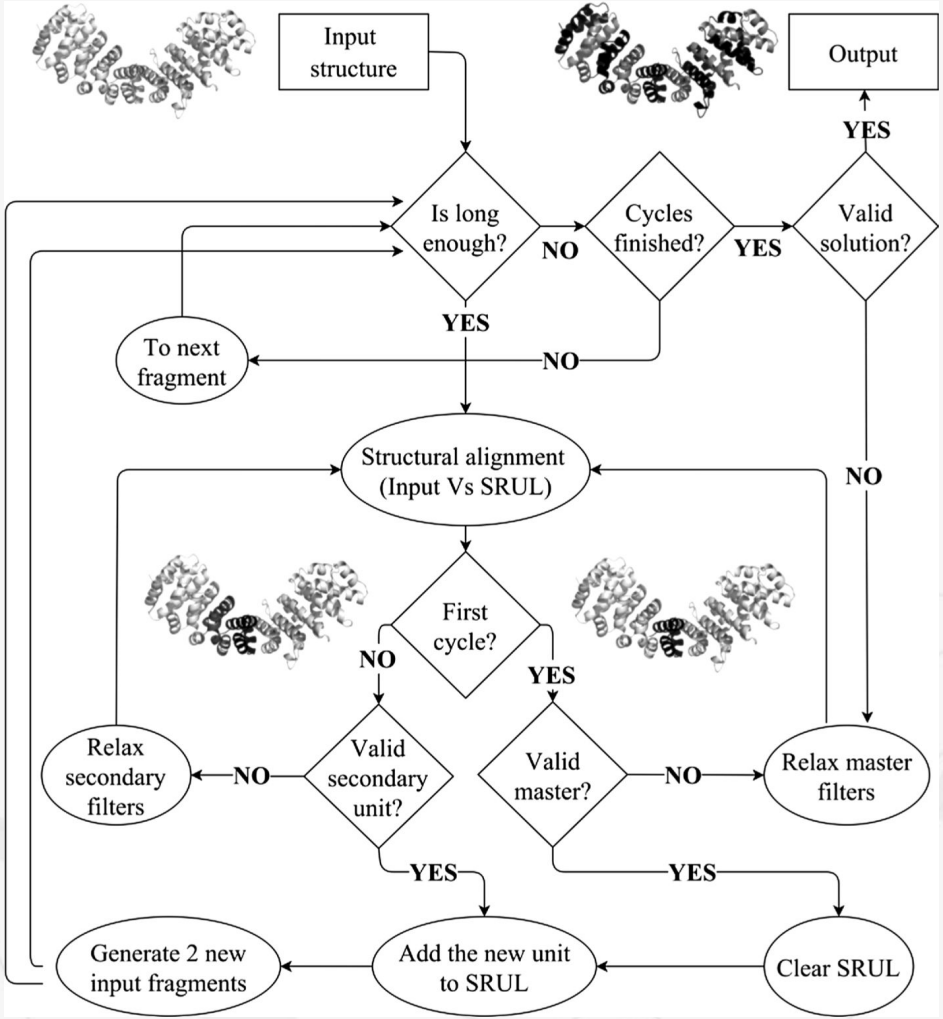
DIPARTIMENTO MATEMATICA

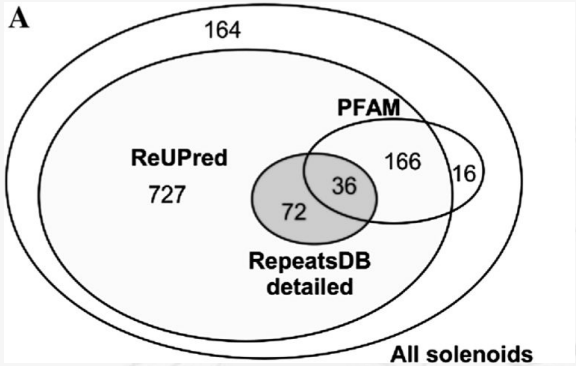
- Input - PDB structure
- Output - unit start/end, insertions, classification
- User feedback on data quality
- Results integrated into RepeatsDB

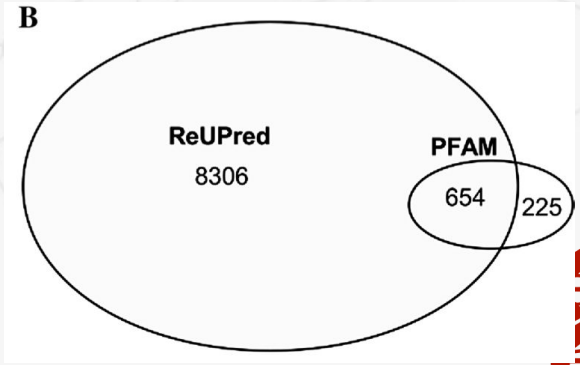


protein.bio.unipd.it/repeatsdb-lite

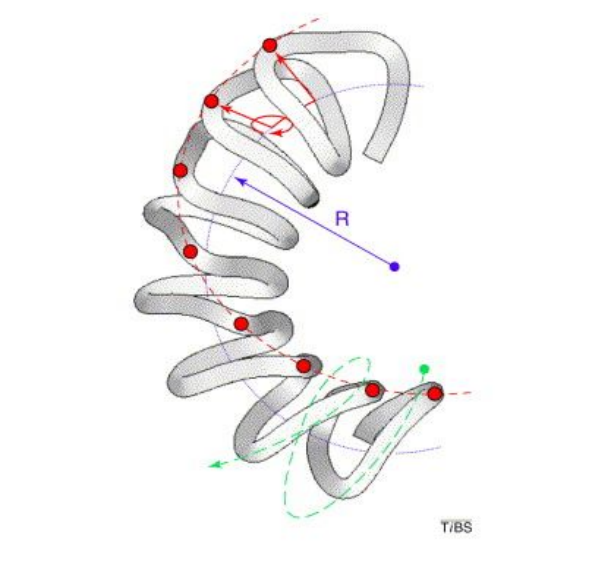


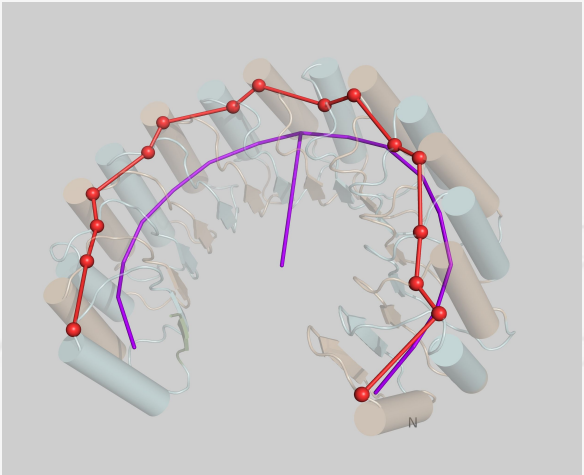

Input structure
Output
Is long enough?
Cycles finished?
Valid solution?
YES
NO
YES
YES
NO
To next fragment
NO
Structural alignment (Input Vs SRUL)
NO
First cycle?
NO
YES
Relax secondary filters
NO
Valid secondary unit?
Valid master?
NO
Relax master filters
YES
YES
Generate 2 new input fragments
Add the new unit to SRUL
Clear SRUL


A
164
PFAM
ReUPred
727
166
16
36
72
RepeatsDB detailed
All solenoids


B
ReUPred
8306
PFAM
654
225

Structural descriptors

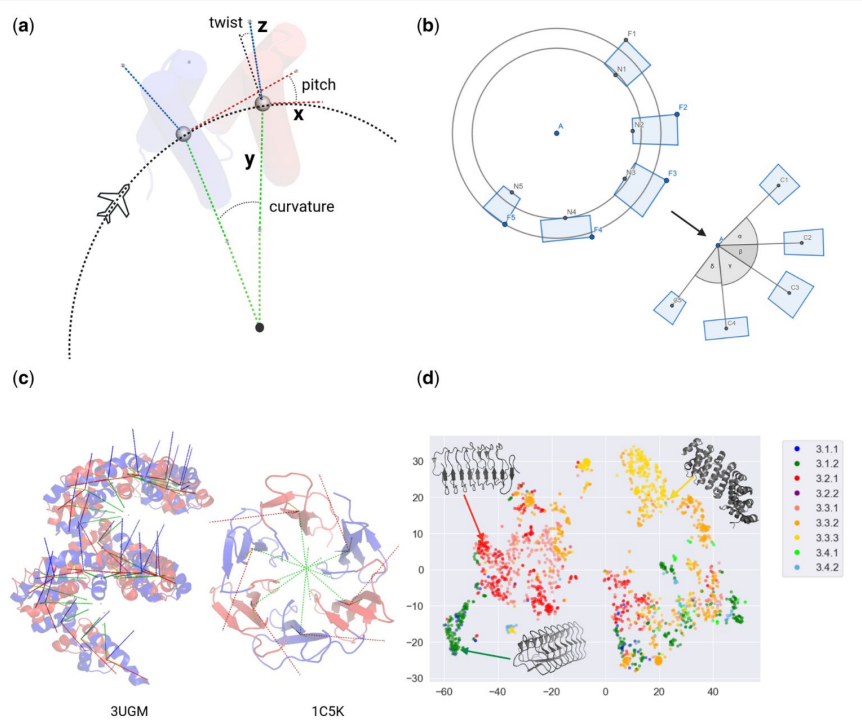




- Curvature
- Twist
- Handedness

Kobe B. and Kajava A. (2000)

Structural descriptors



Osmanli et al. **GeomeTRe: accurate calculation of geometrical descriptors of tandem repeat proteins** - Bioinformatics - 2025 - https://doi.org/10.1093/bioinformatics/btaf395