Novosibirsk, 2016

# DARK PROTEOME


Dr Andrey Kajava

Group of Structural Bioinformatics and Molecular Modeling

Centre de Recherches de Biochimie Macromoléculaire, CNRS

Montpellier, FRANCE

**PRINCIPAL INTEREST**

**PROTEIN SEQUENCE -> 3D STRUCTURE -> FUNCTION**

**CHALLENGES AND MOTIVATIONS**

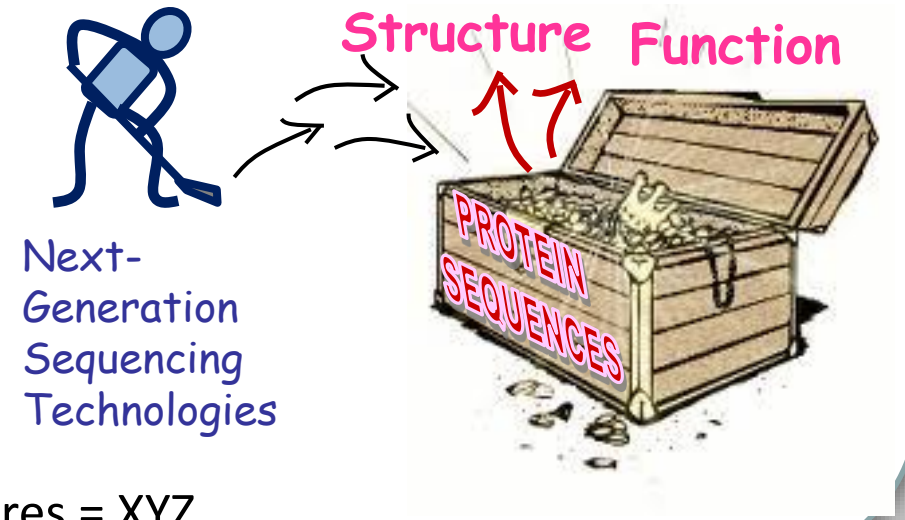
✓ **Unresolved fundumental problems**: protein folding, protein interactions

---

✓ **Dramatic growth of data**
(genomes, proteomes)
that need to be understood

✓ **Unprecedentedly strong need in bioinformatics**
for multi-proteome analysis

✓ **Well suited for computational methods:** Sequences=text, Structures = XYZ



# PRINCIPAL ACTIVITIES

## FORMATICS
## MOLECULAR MODELLING

Development of bioinformatics tools

Bioinformatics analysis of
protein sequences and structures

Prediction and modeling
of protein structures

## COLLABORATIONS WITH EXPERIMENTALISTS

Structure-based
design of experiments
and interpretation of results

Homology modelling

Molecular design

Institut de biologie computationnelle

# AXE3: Structural and functional annotation of proteomes

LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique)
Laurent Bréhélin

CRBM (Centre de Recherche de Biochimie Macromoléculaire)
Andrey Kajava

LBCM (Laboratoire de Biologie Cellulaire et Moléculaire)
Emmanuel Cornillot

CBS (Centre de Biochimie Structurale)
Jerome Gracy

CPBS (Centre d'étude d'agents pathogènes et biotech pour la santé)
Laurent Chaloin

I3M (Montpellier Institute of Mathematics and Modeling)
Pierre Pudio

CIRAD UMR AGAP
Jean-François Dufayard

Institut de biologie computationnelle

# AXE3: Structural and functional annotation of proteomes

## Raison d'être

Making sense of millions of **protein sequences** requires information about their **3D structure** as well as about their **functional relationships**.

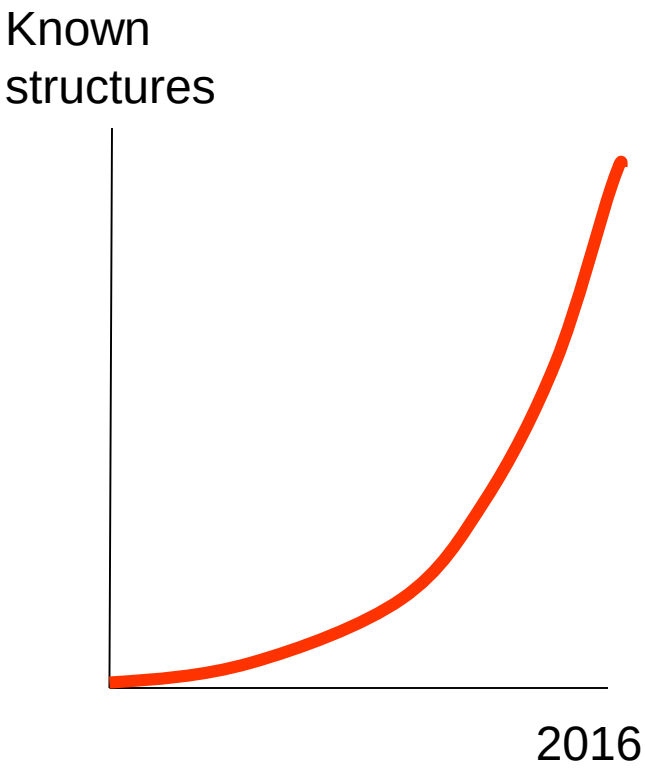
Today, the growth of the sequencing data significantly exceeds the growth of capacities to analyze these data.

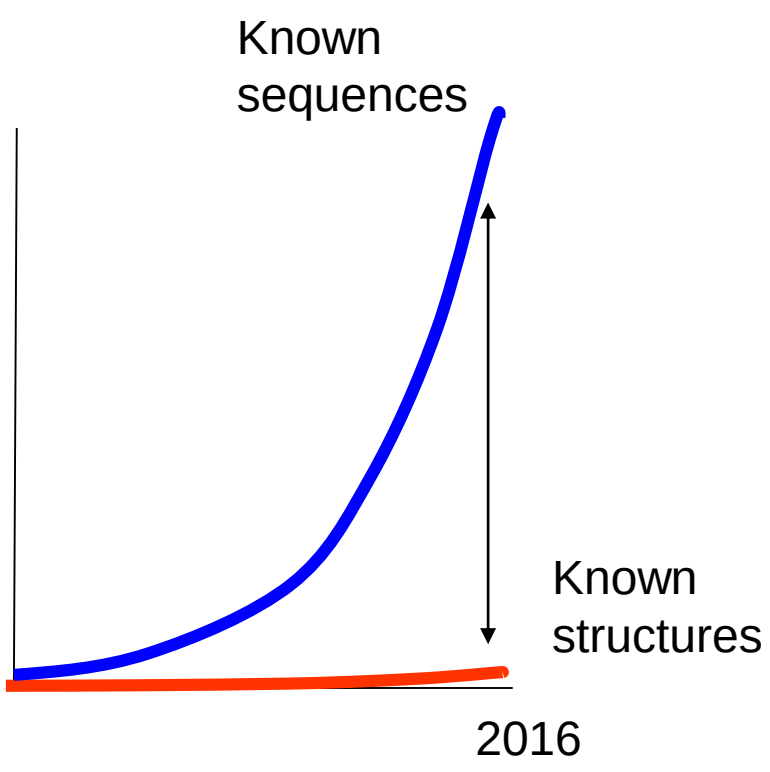
In line with the needs, our AXE3 deal with the development of new algorithms, software, data integration tools to implement the processing chains to analyze proteome data.

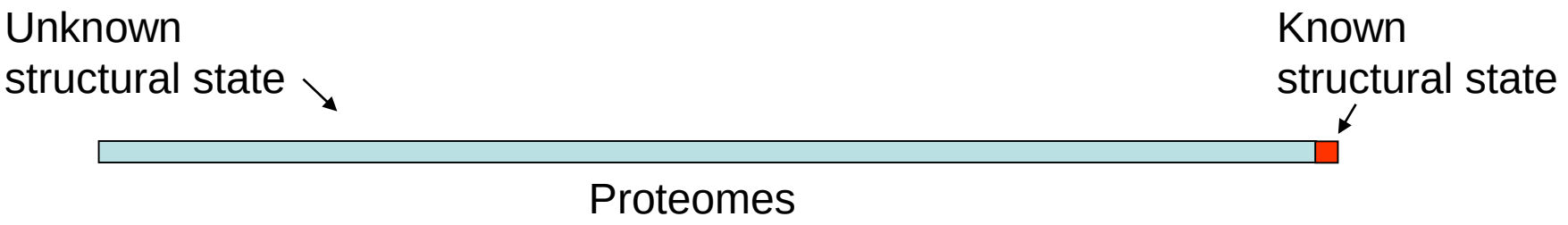

Non-globular proteins - from sequence to structure, function and application in molecular physiopathology.

27 EU countries

# PROTEIN SEQUENCE -> 3D STRUCTURE -> FUNCTION



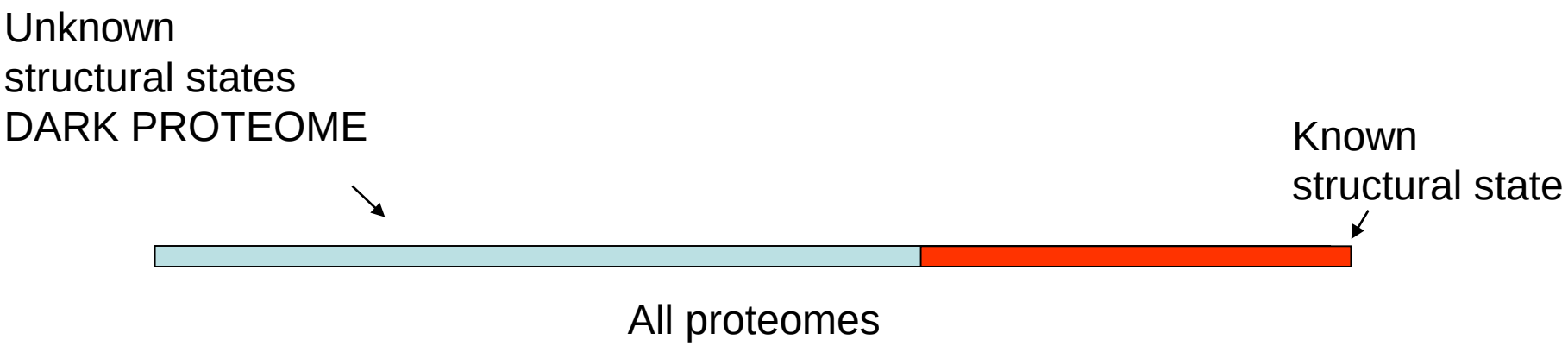



# PROTEIN SEQUENCE -> 3D STRUCTURE -> FUNCTION

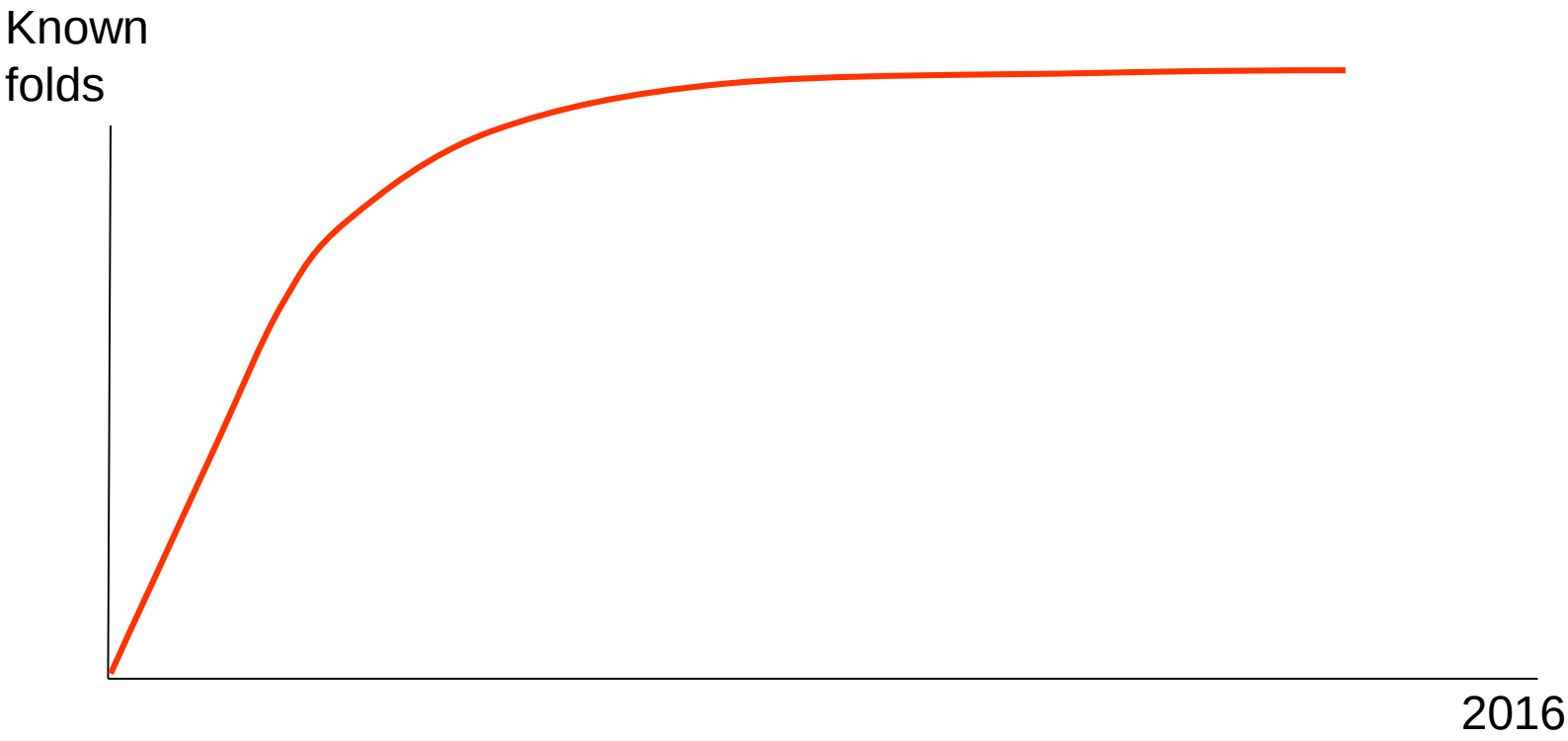
Homology-based prediction of 3D structures

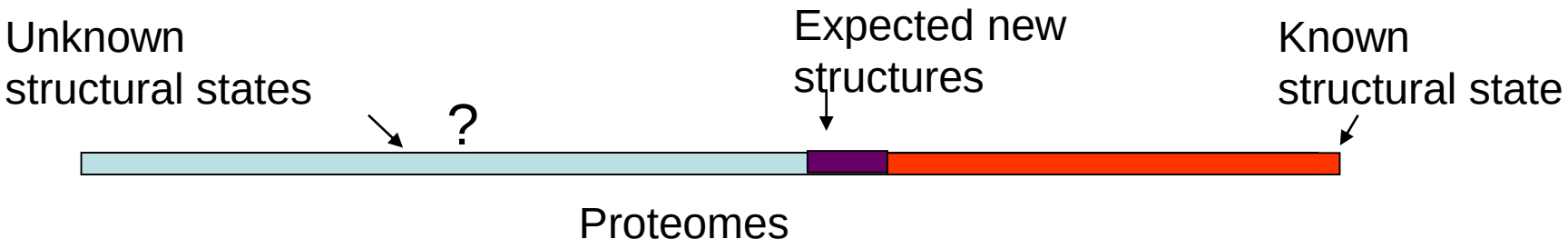
Proteins with similar sequences have similar 3D structure.°

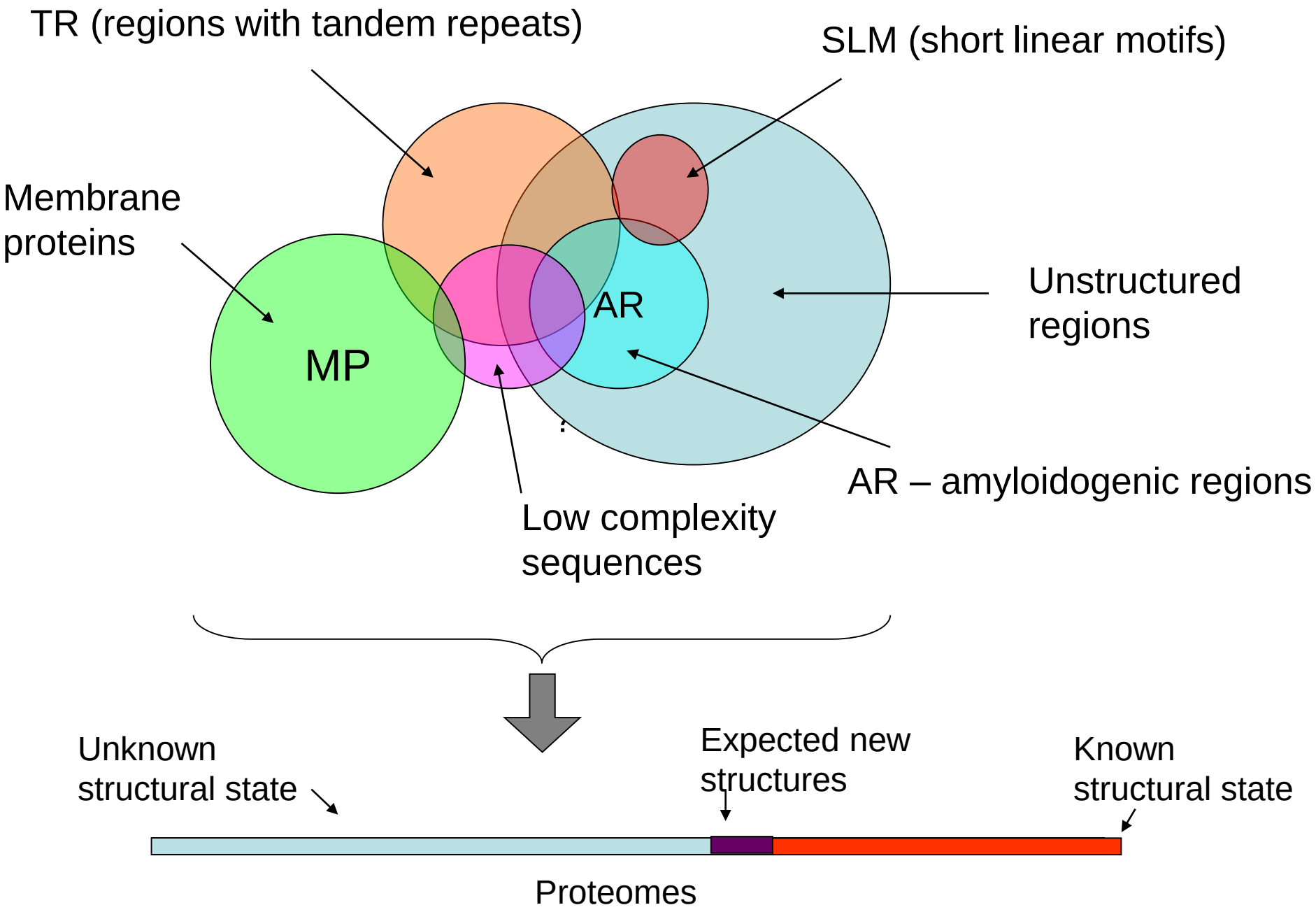
(BLAST HMMs HMM vs HMM)

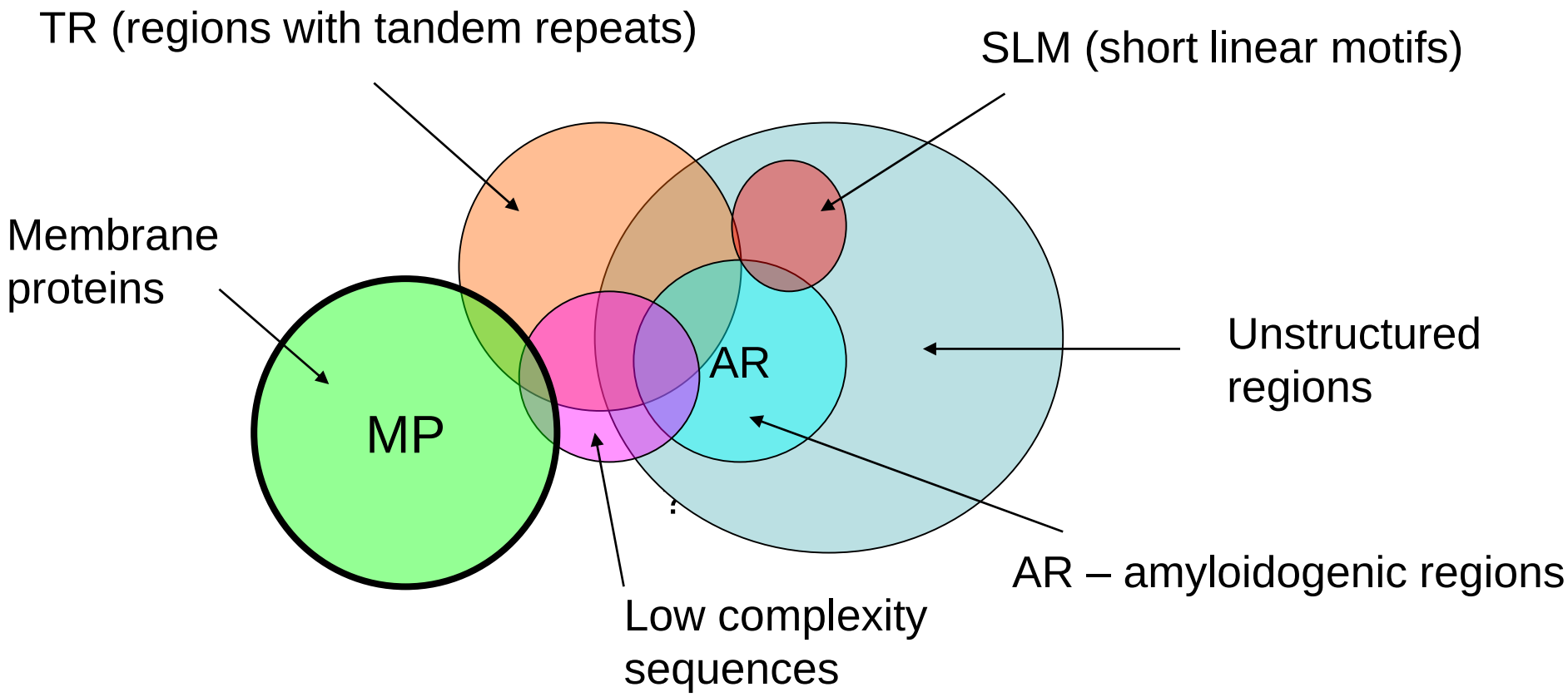


# PROTEIN SEQUENCE -> 3D STRUCTURE -> FUNCTION

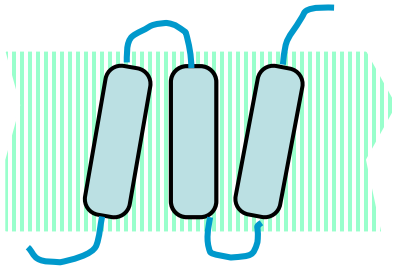


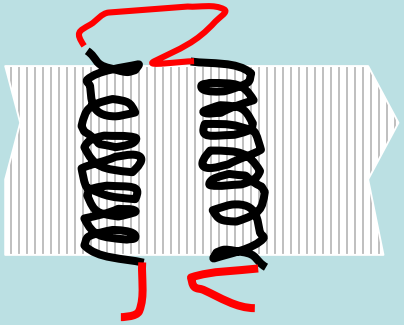



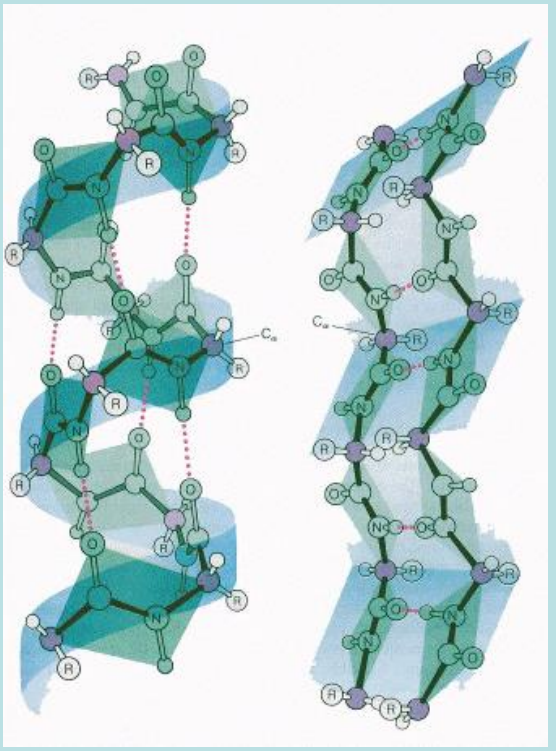

TR (regions with tandem repeats)
SLM (short linear motifs)
Membrane proteins
MP
AR
Unstructured regions
AR – amyloidogenic regions
Low complexity sequences
Unknown structural state
Expected new structures
Known structural state
Proteomes

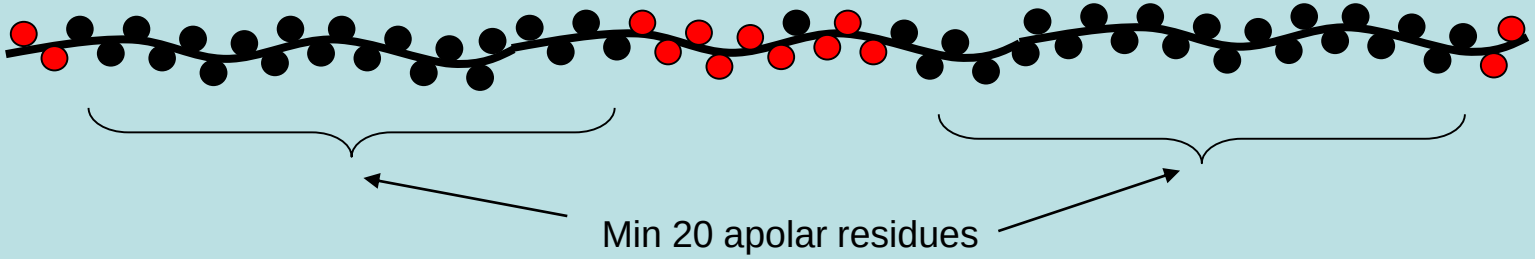

TR (regions with tandem repeats)
SLM (short linear motifs)
Membrane proteins
MP
AR
Unstructured regions
AR – amyloidogenic regions
Low complexity sequences

Protéines membranaires



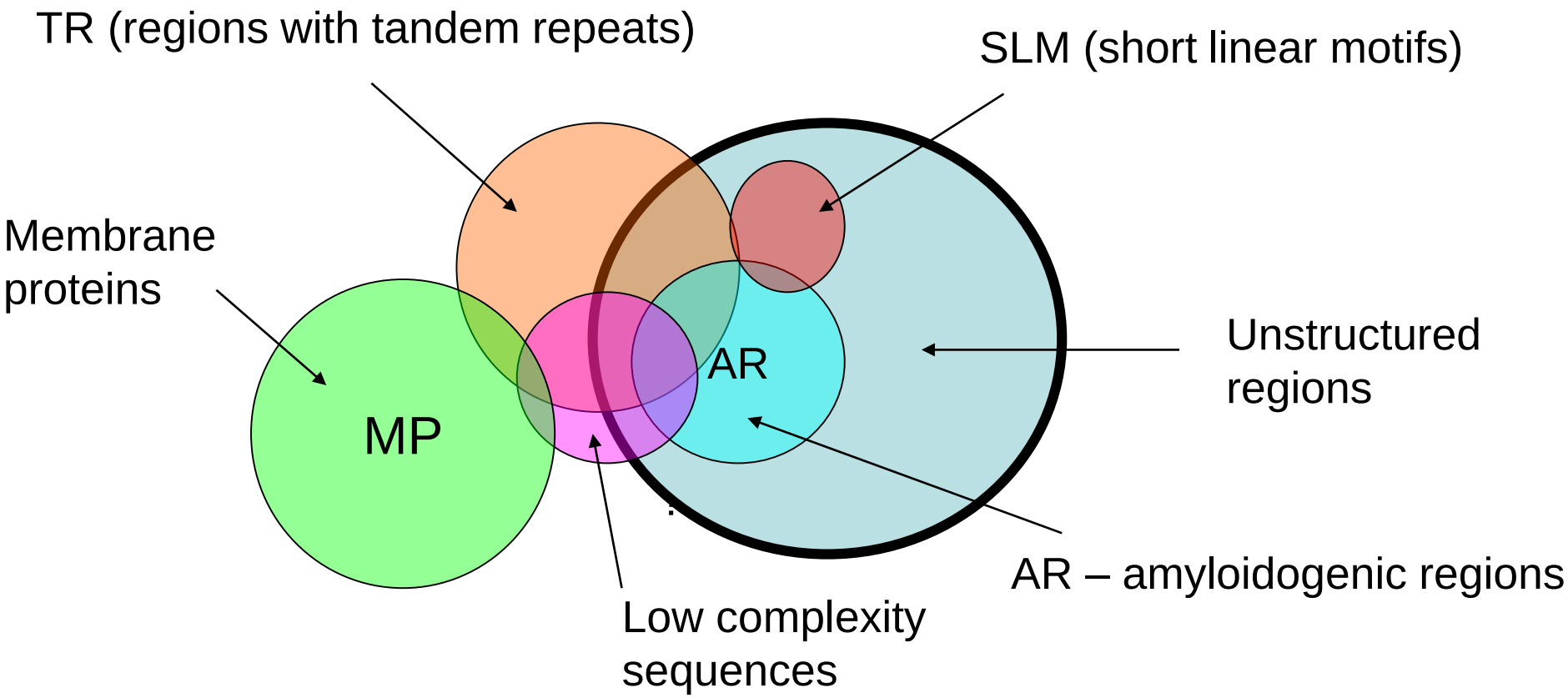




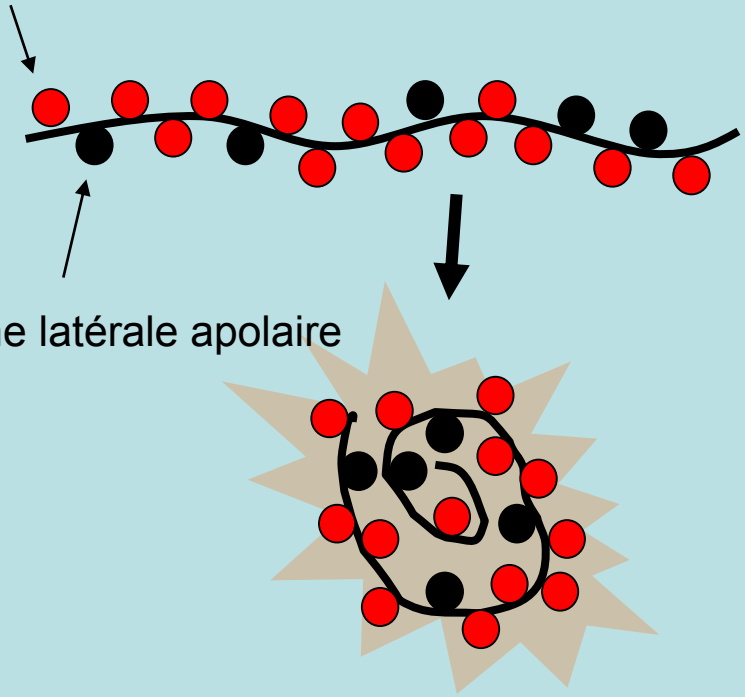



Logiciels: TMHMM, TMpred

Problèmes: compactage latéral et orientation
protéines à feuillets bêta


TR (regions with tandem repeats)
SLM (short linear motifs)
Membrane proteins
MP
AR
Unstructured regions
AR – amyloidogenic regions
Low complexity sequences

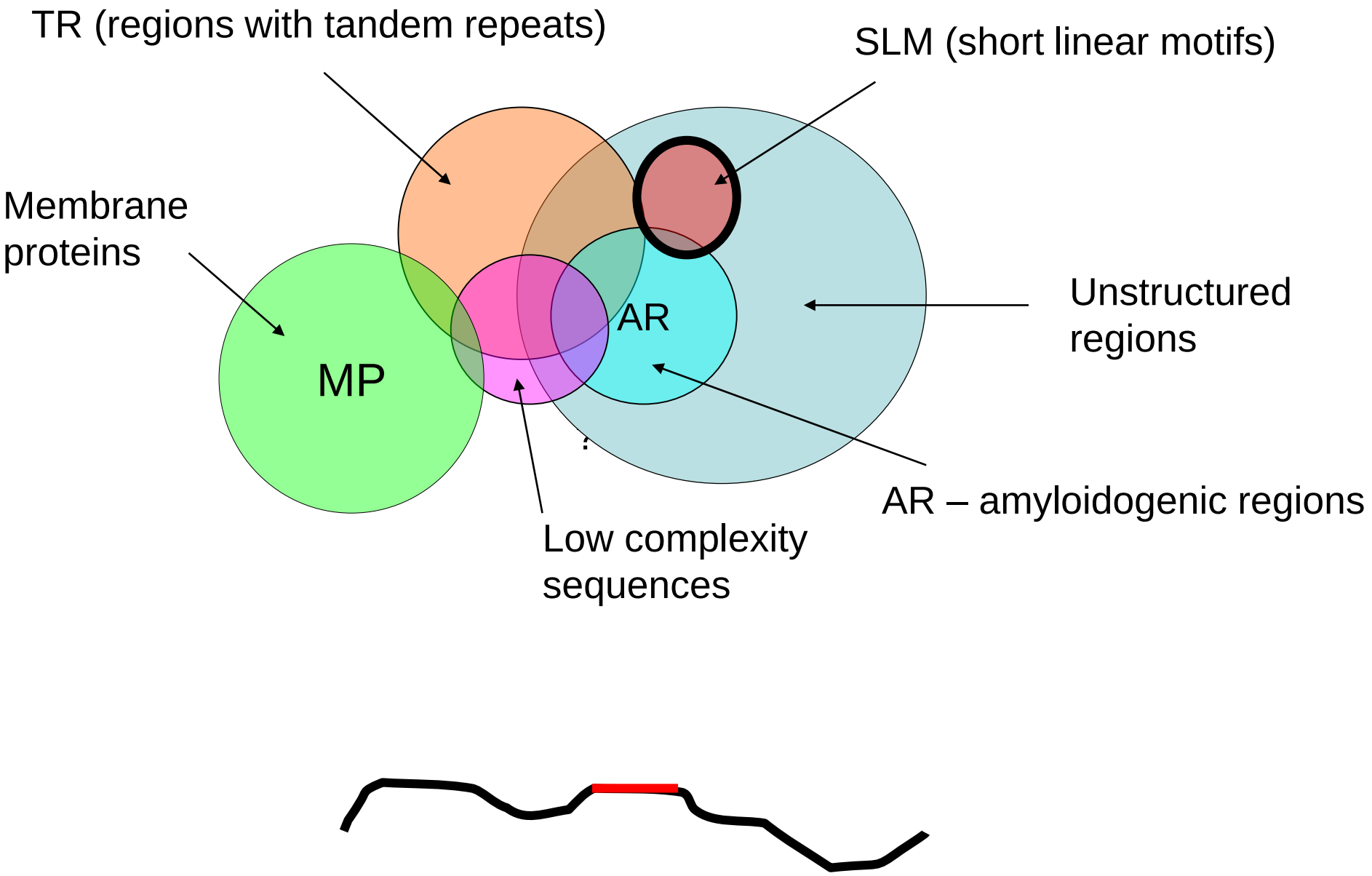
Chaîne latérale polaire

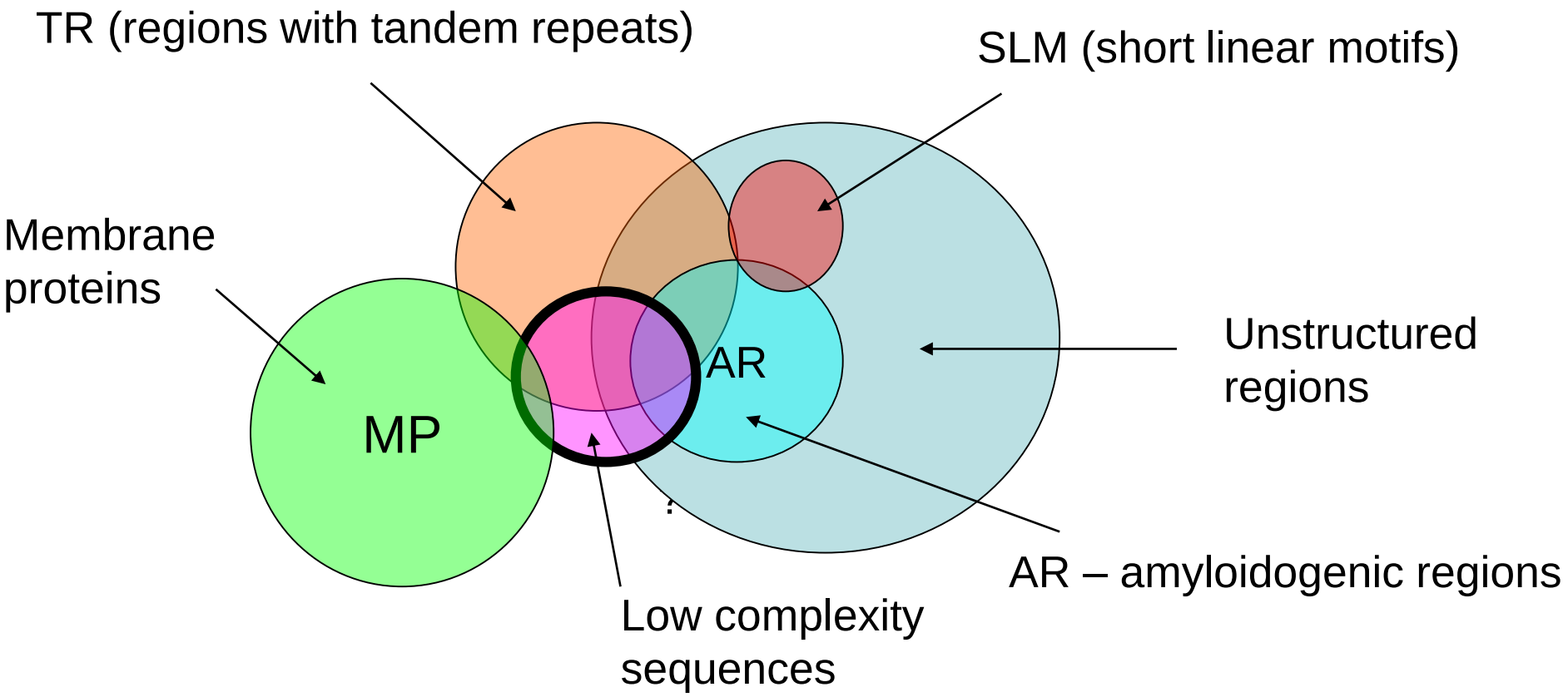


Chaîne latérale apolaire

Protéines non structurées

Software and database on web-page:
http://www.disprot.org/


TR (regions with tandem repeats)
SLM (short linear motifs)
Membrane proteins
MP
AR
Unstructured regions
AR – amyloidogenic regions
Low complexity sequences




**AAA**K**AAAAAA**S**AAAAAAA**M**AAAAAA**

SDFKHGRTEACWLMKHNVFYTIPLM

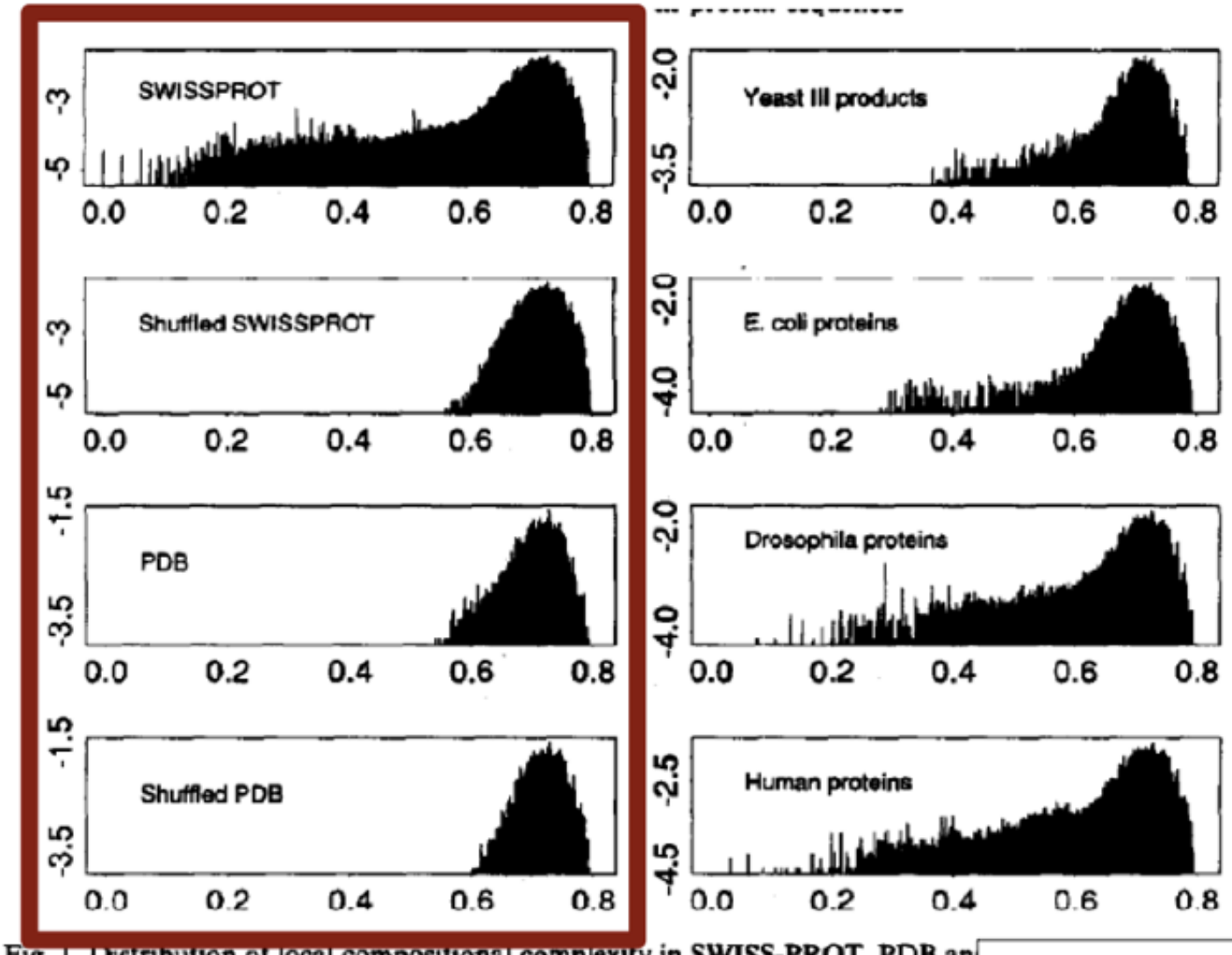


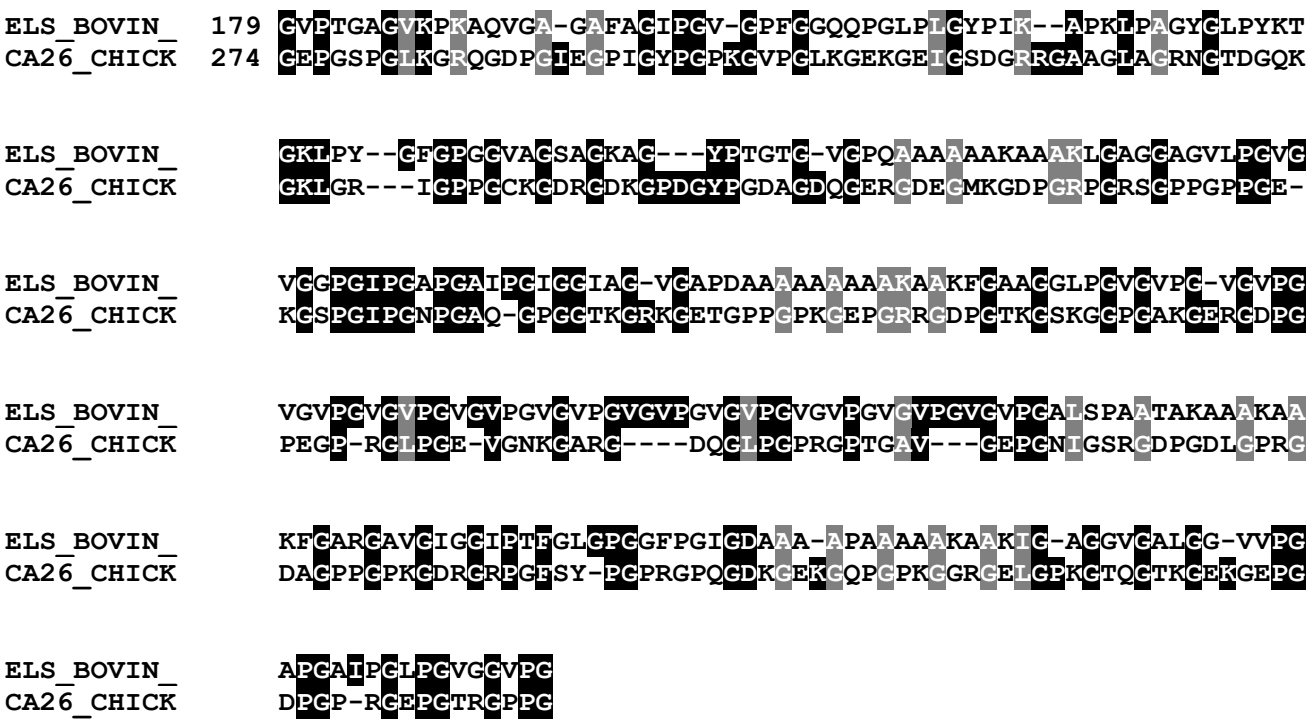

Fig. 1. Distribution of local compositional complexity in SWISS-PROT, PDB an
of SWISS-PROT. The horizontal axis of each plot is Complexity $K_1$ [equatio
windows of length $L = 40$ moved in steps of one residue along all the sequences i
The vertical axis is $-\log_{20}(\textit{frequency})$ where *frequency* is the fraction of the tot
value of $K_1$. For $L = 40$, there are 35,251 possible values o

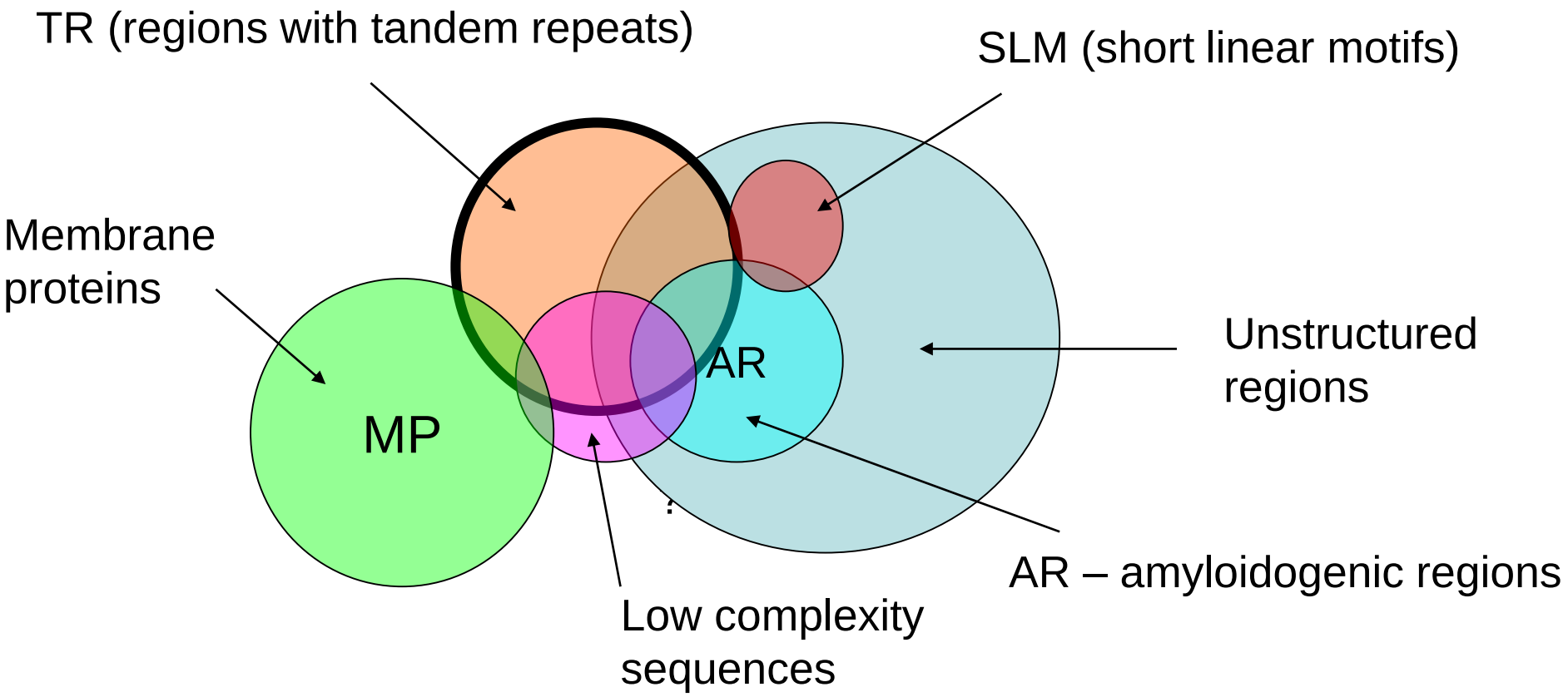
JOHN C. WOOTTO
*Computers Chem.* Vol. 18, No. 3, pp. 269–285, 1
Elsevier Science Ltd. Printed in Great Brit

# "Faux Amis" of homology modeling

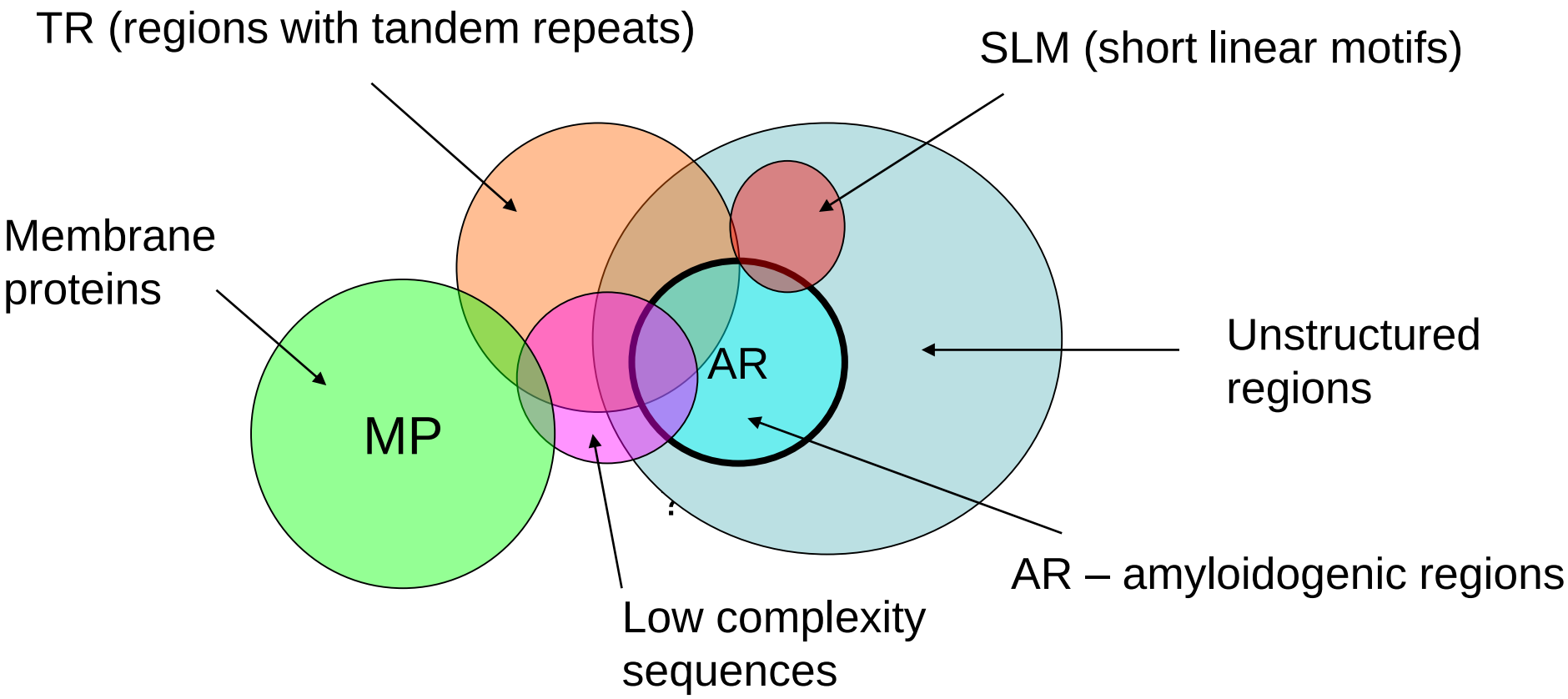
## 1. Low-complexity sequences

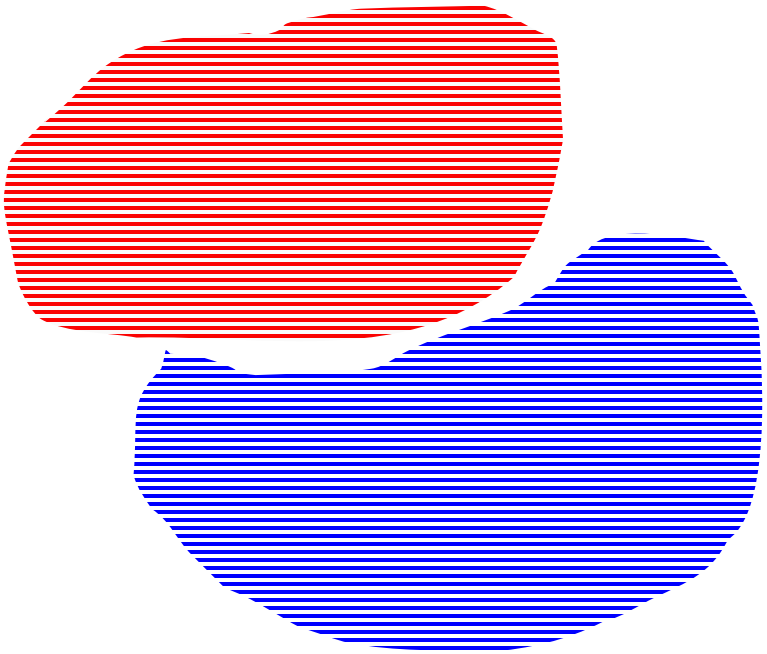
Alignment of 301-residue fragments of elastin and type VI collagen with 30% of residue identity and E-value $10^{-4}$.



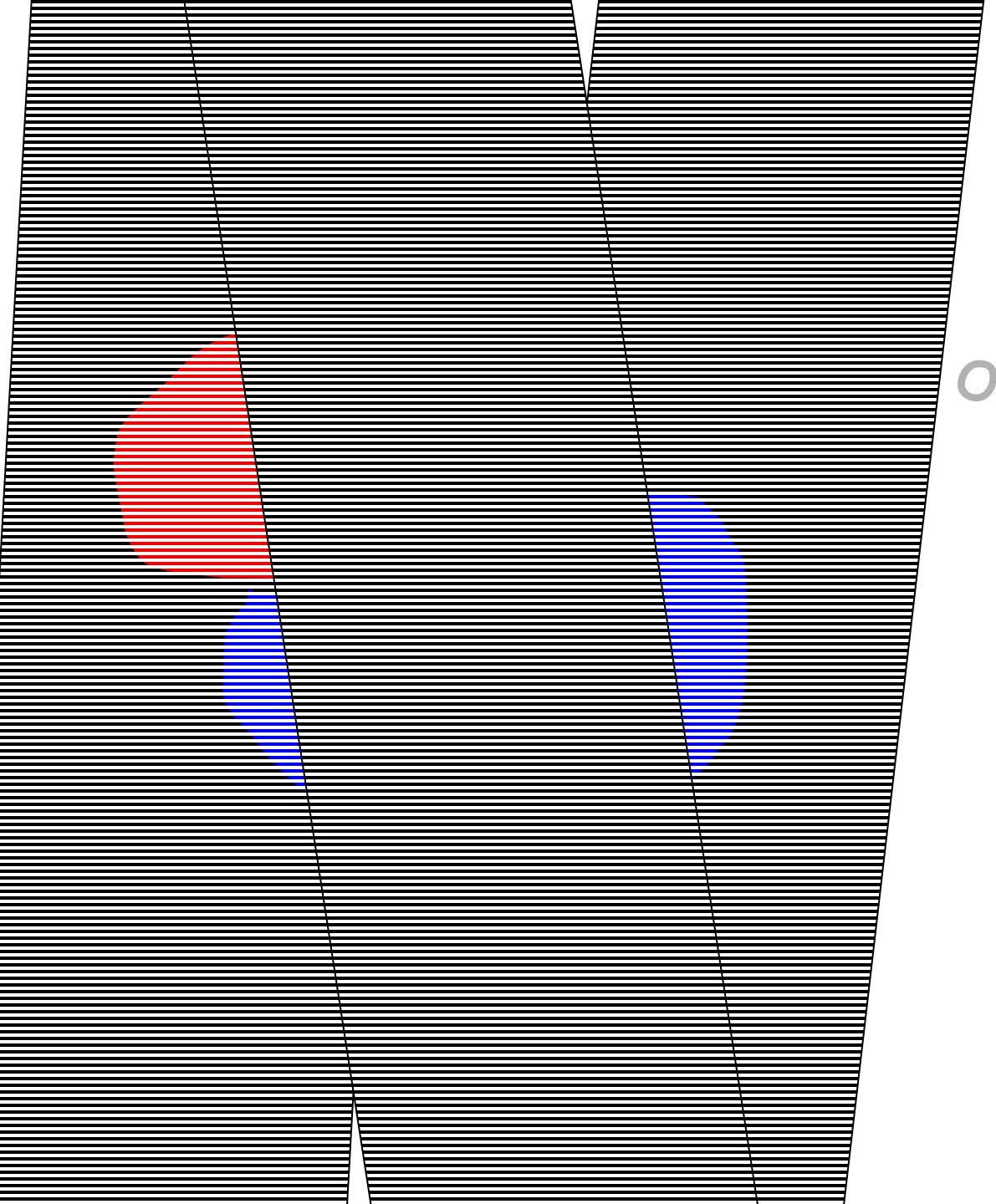



**PPGPPGPPGPPGPPGPPGPPGPPGPPGPPG**


TR (regions with tandem repeats)
SLM (short linear motifs)
Membrane proteins
MP
AR
Unstructured regions
AR – amyloidogenic regions
Low complexity sequences



Limited size
and
Optimal stability
of
Protein
Structures



Limited size
and
Optimal stability
of
Proteins
Structures

**Stable structures**
**of**
**Unlimited size**

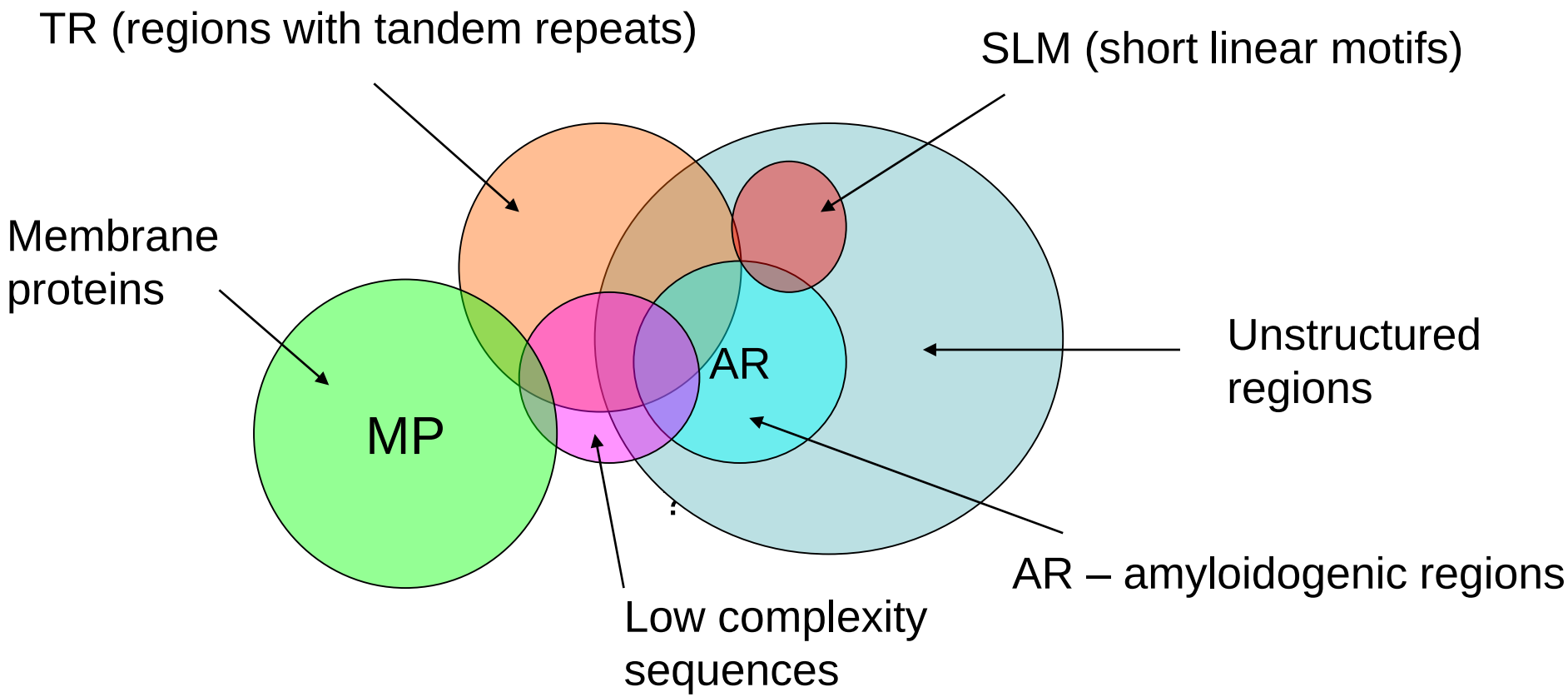
# AMYLOID FIBRILS

- ✓ Presence of amyloid fibrils is connected with serious neurodegenerative diseases, including Alzheimer's disease, Parkinson's desease, Huntington's disease, and also the transmissible prion diseases.

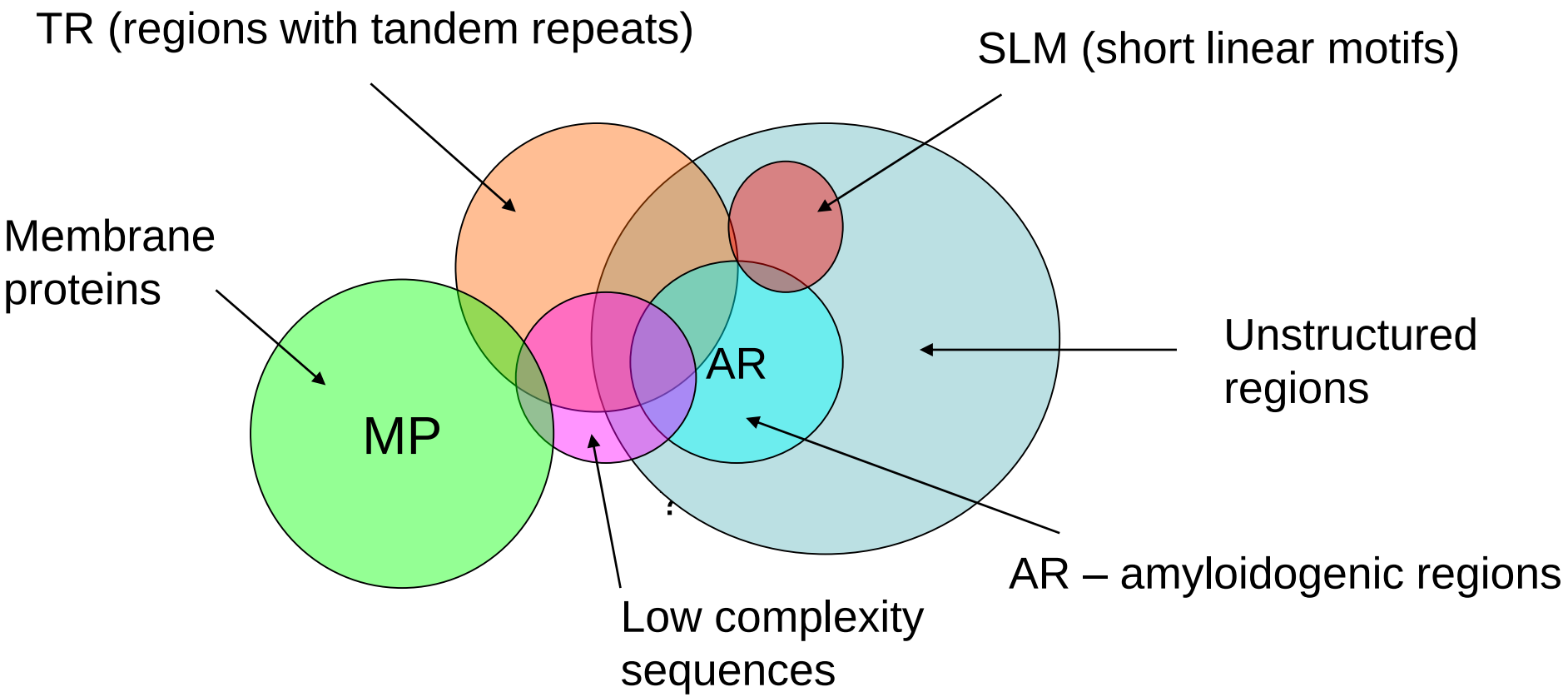
- ✓ Other amyloidosis (type 2 diabetes, cancer – p53, etc)

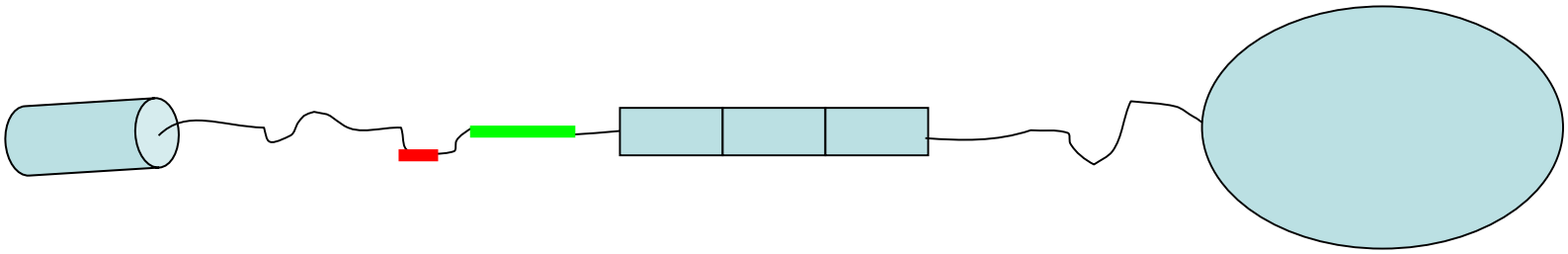
- ✓ **Functional amyloids:**
  Bacterial biofilms
  Bacterial surfaces
  Necrosis

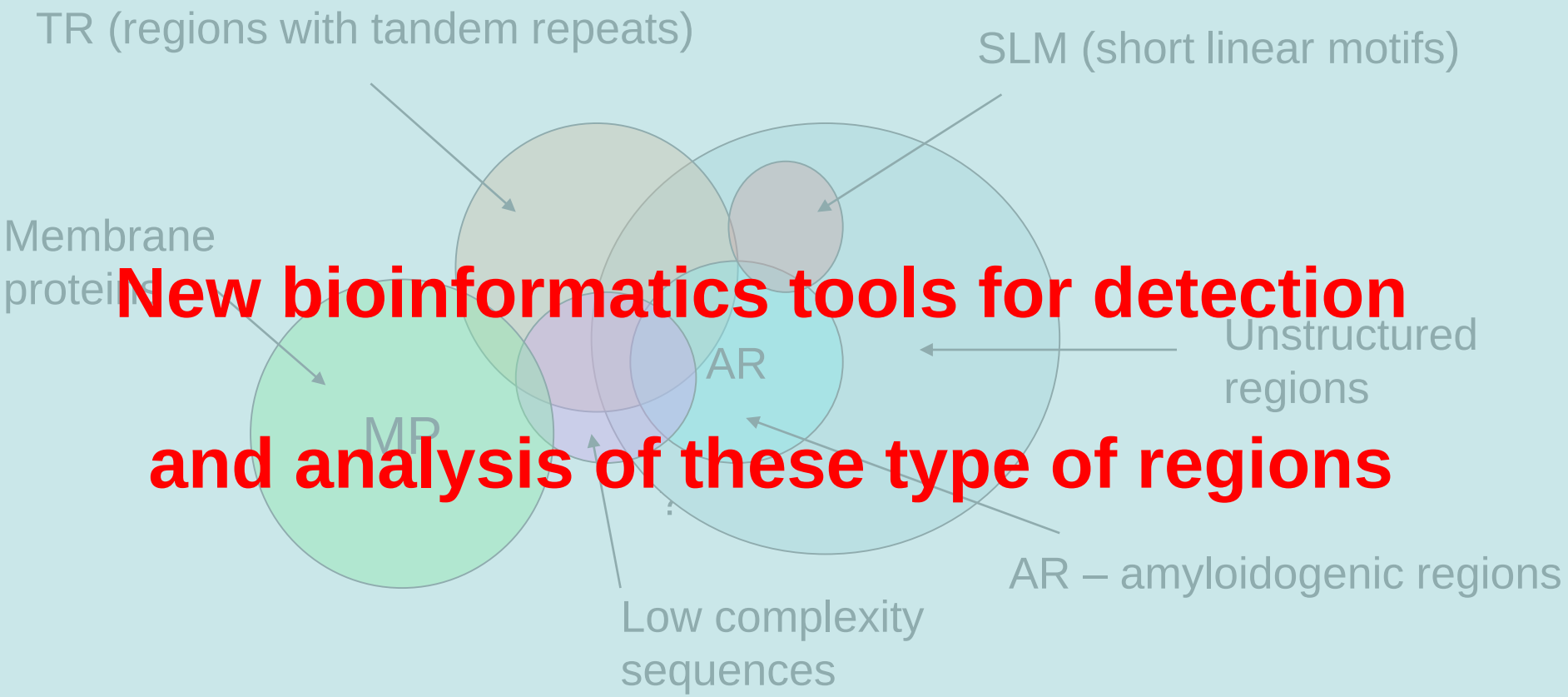
# Relationships between these states



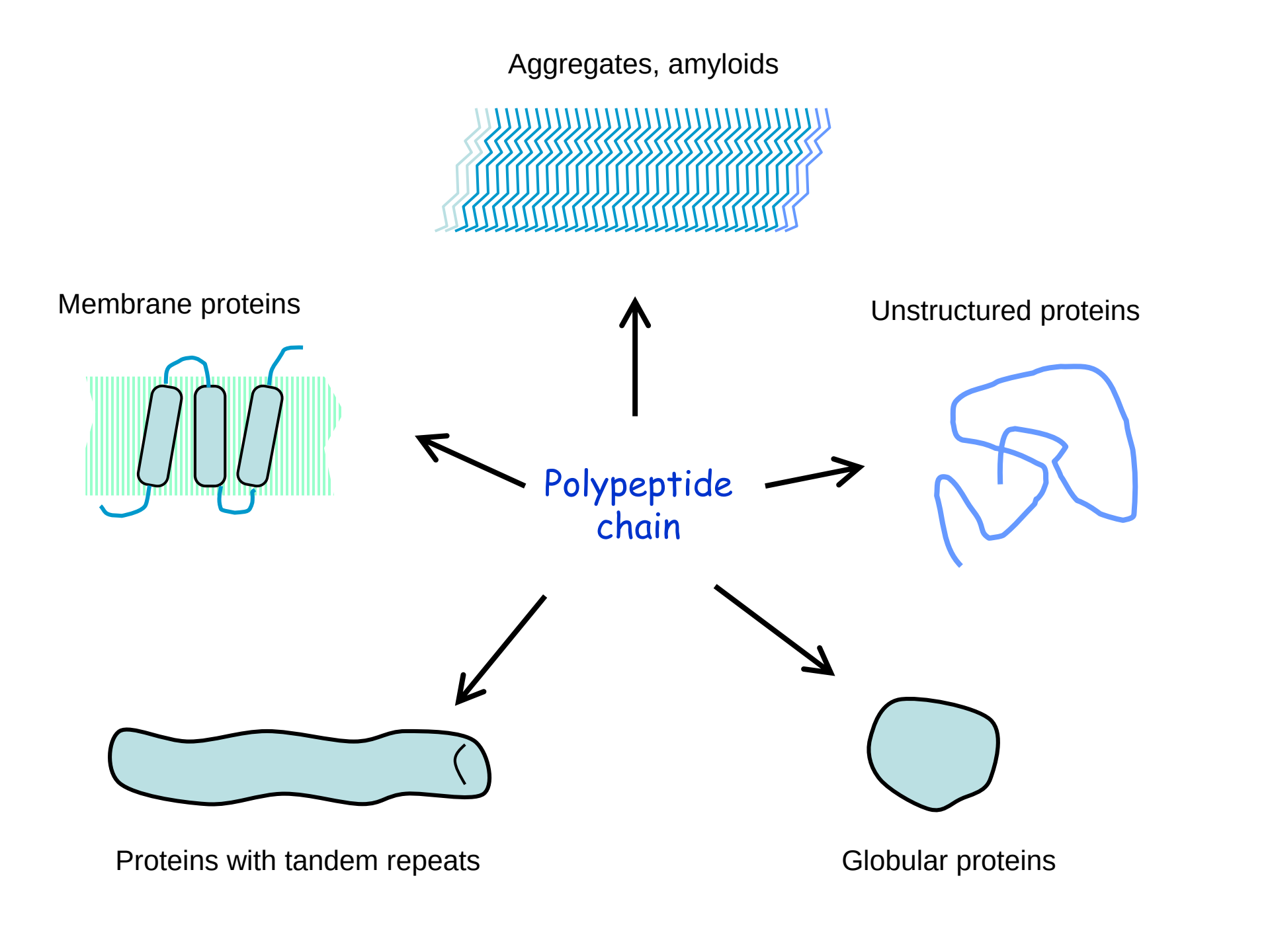
# Relationships between these states

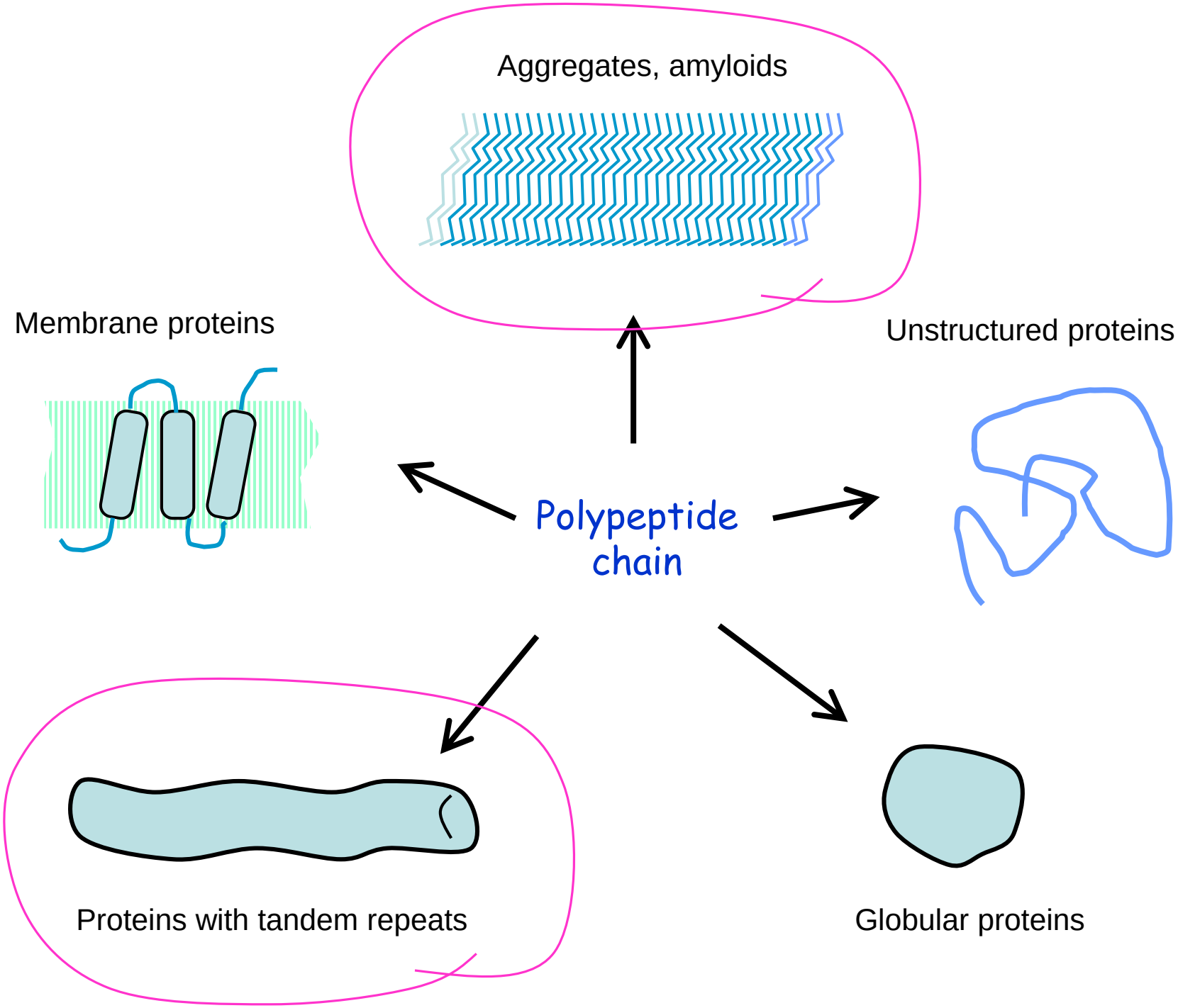





**New bioinformatics tools for detection and analysis of these type of regions**

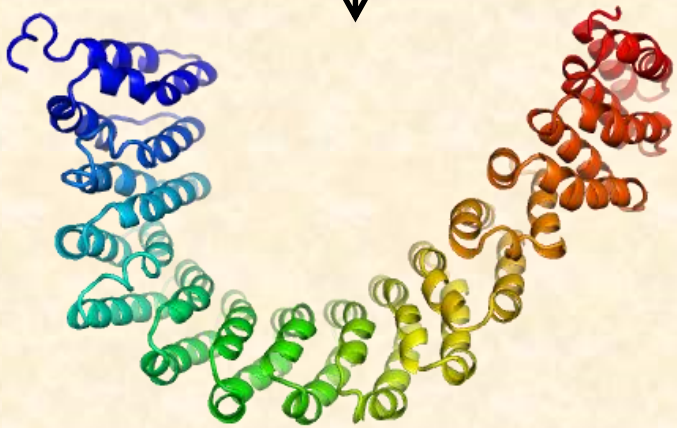

Aggregates, amyloids
Membrane proteins
Unstructured proteins
Polypeptide chain
Proteins with tandem repeats
Globular proteins


Aggregates, amyloids
Membrane proteins
Unstructured proteins
Polypeptide chain
Proteins with tandem repeats
Globular proteins

# PROTEINS WITH TANDEM REPEATS

< 50 res

Structure



Proteins with internal duplications represent a large portion of proteomes

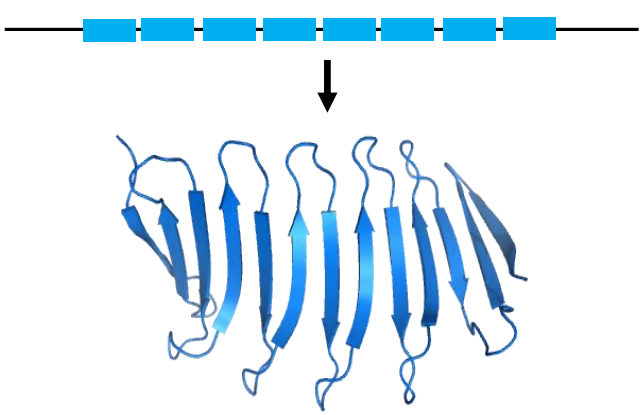
*Homo sapiens (30%)*

*Plasmodium falciparum (60%)*

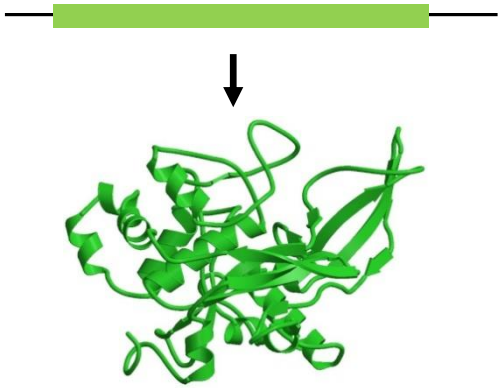
# PROTEINS WITH TANDEM REPEATS

Bioinformatics tools for analysis of
relationship between sequence – structure – function

Tandem repeats

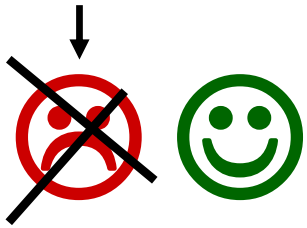


Globular domains (aperiodic sequences)

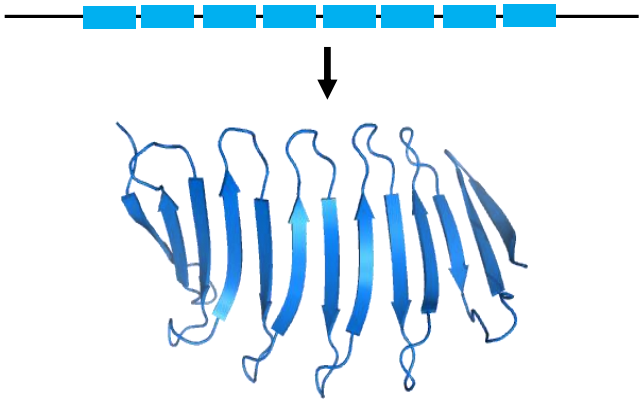


# PROTEINS WITH TANDEM REPEATS

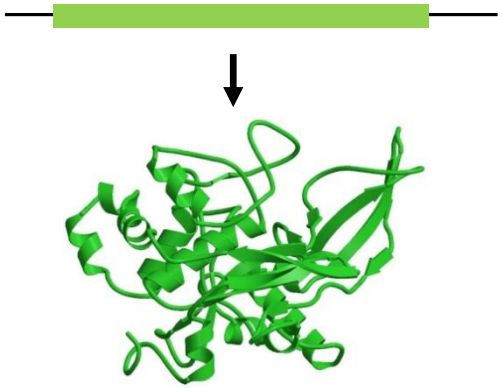
Bioinformatics tools for analysis of
relationship between sequence – structure – function



Tandem repeats



Globular domains (aperiodic sequences)



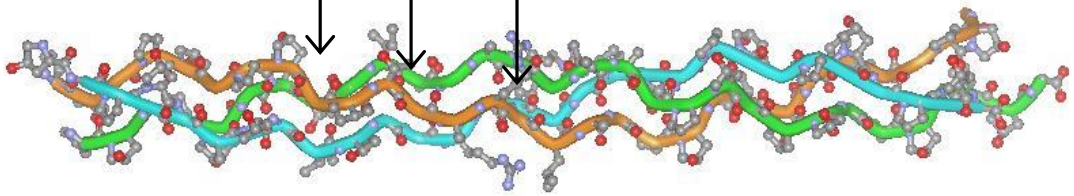
Our objective is to fill this gap

# IDENTIFICATION

# OF PROTEIN REPEATS

PPGPPGPPGPPGPPGPPGPPGPPGPPGPPGPPG

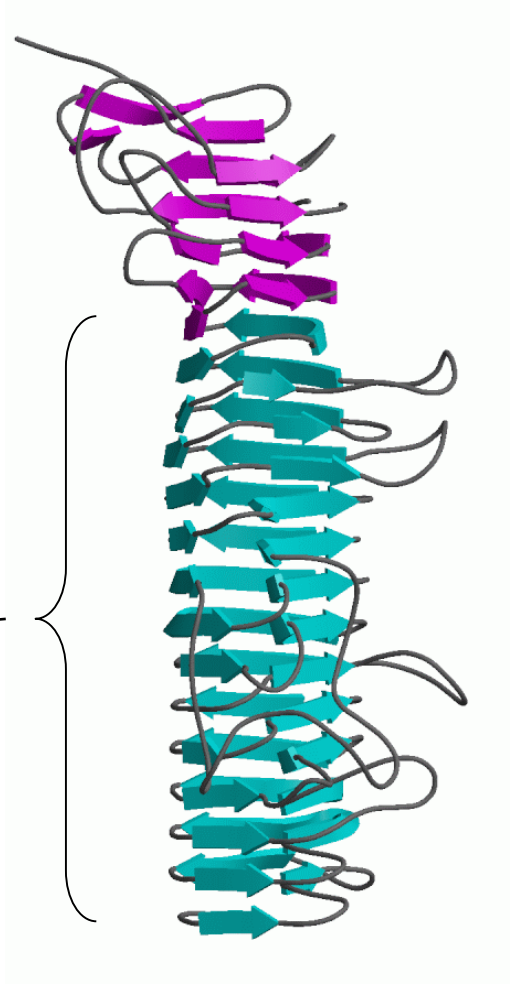
PPGPEGPPGITGARGLAGPPGPPGKPGPPG



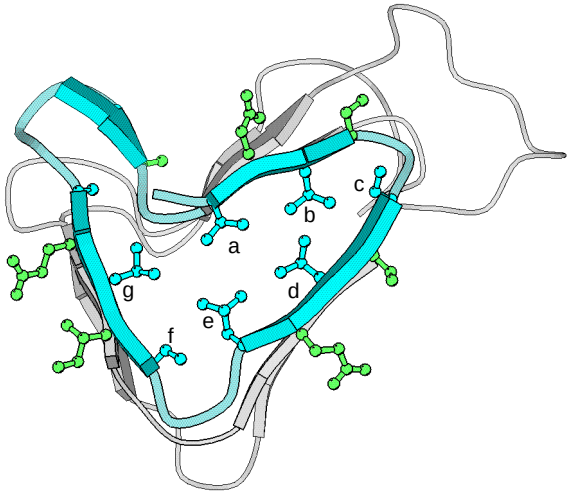
*Collagen*

# Pertactin from *Bordetella pertussis*

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



# Pertactin from *Bordetella pertussis*




```
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
```

Generalized sequence profiles implemented in *pftools* (Bucher et al., 1996, Comput. Chem. 20, 3-23)

# Methods for detection of tandem repeats in proteins

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

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

# Structural Bioinformatics and Molecular Modeling
## CRBM-CNRS Montpellier



## T-Reks

Tandem Repeats Explorer based on K-means algorithm in Sequences

Search in a file : Parcourir...

or Paste your sequences :

Sequence type : Protein DNA

Percentage of similarity: 0.70

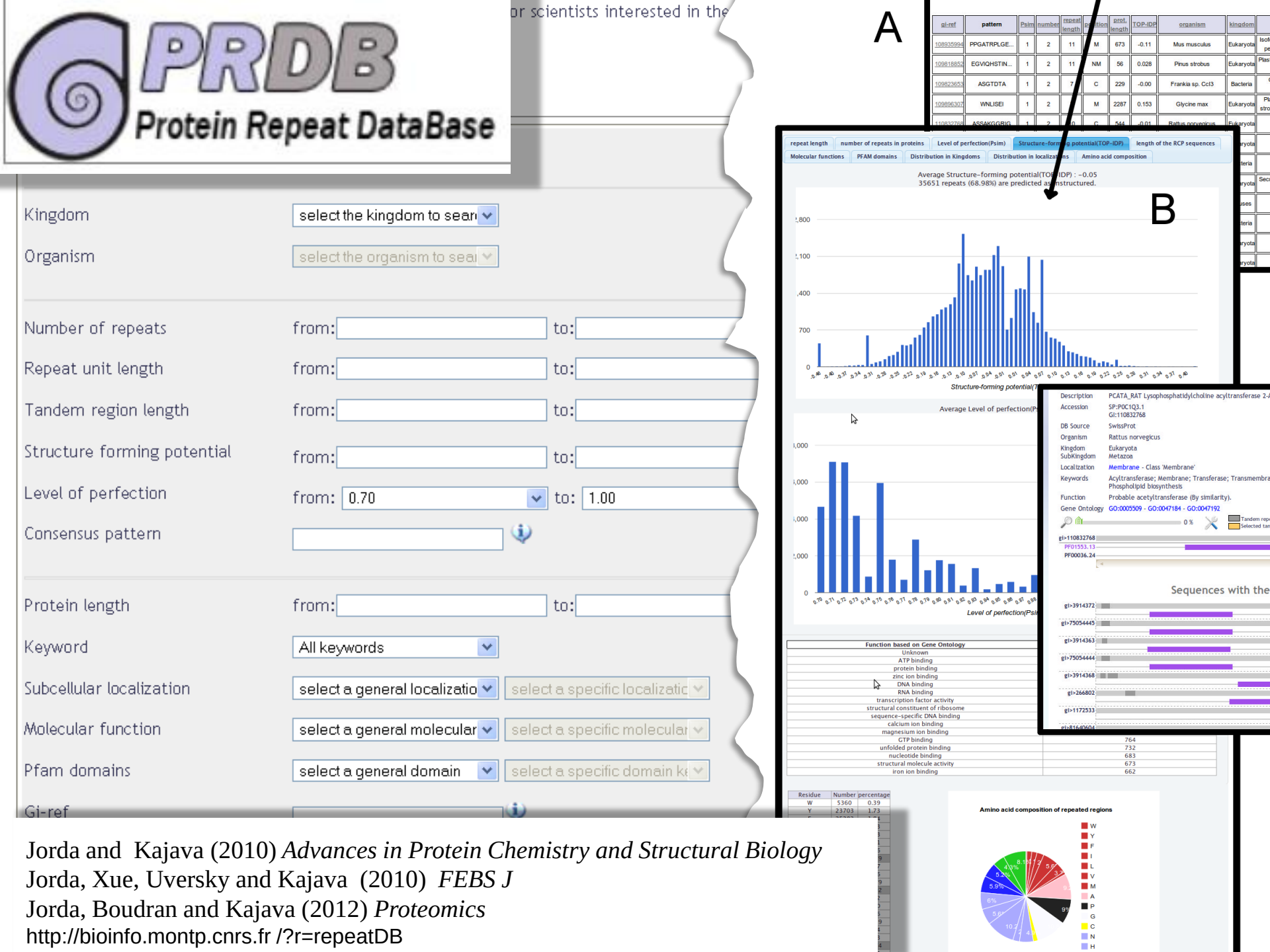
Filter the overlapping repeats:

Search Repeats

00:00:03

```
Research of repeats is requested:

>protein
Length: 3 residues - nb: 5  from  7 to 21 - Psim:0.8 region Length:15
KKL
DSL
DSL
DSL
DDL
**********
1 sequences have been detected as tandem repeats containing.
```




Jorda and Kajava (2010) *Advances in Protein Chemistry and Structural Biology*
Jorda, Xue, Uversky and Kajava (2010) *FEBS J*
Jorda, Boudran and Kajava (2012) *Proteomics*
http://bioinfo.montp.cnrs.fr /?r=repeatDB

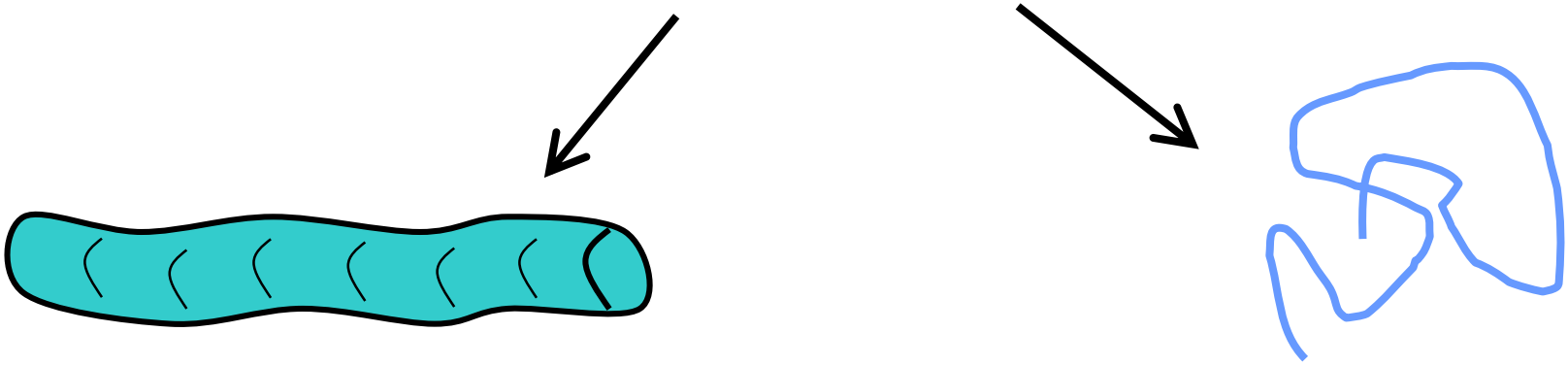
# Methods for detection of tandem repeats in proteins

| Type of method | *ab initio / a priori* | Properties of repeats | Rapidity |
|---|---|---|---|
| **Fourier Transform analysis**<br><br>REPPER (Gruber et al., 2005) | *ab initio* | Long arrays without indels | + |
| **Short string extension algorithms**<br><br>XSTREAM (Newman and Cooper, 2007)<br>T-REKS (Jorda and Kajava, 2009) | *ab initio* | With indels and less than 15-20 residues | + |
| **Sequence-sequence alignment**<br><br>RADAR ((Heger and Holm, 2000)<br>TRUST (Szklarczyk and Heringa, 2004) | *ab initio* | More than 15 residues. With indels. | - |
| **Hidden Markov Models (HHMs) or sequence profiles**<br><br>PFAM(Sonnhammer et al., 1998)<br>SMART ((Schultz et al., 1998)<br>BISMM library (Kajava and Steven, 2006) | requires *a priori* information | Long and strongly imperfect repeats | - |
| **HMM-HMM or profile-profile comparisons**<br><br>HHREPID (Biegert and Soding, 2008) | *ab initio* | Long and strongly imperfect repeats | - |
| **Sequence profile - Fourier and wavelet transforms**<br><br>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 sequence to 3D structure

## IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?



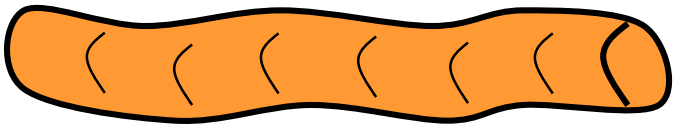
Structured

Unstructured

# IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?

Polypeptide
with tandem repeats

3D structure ?



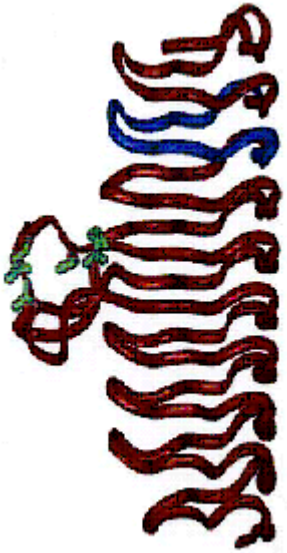
Unstructured

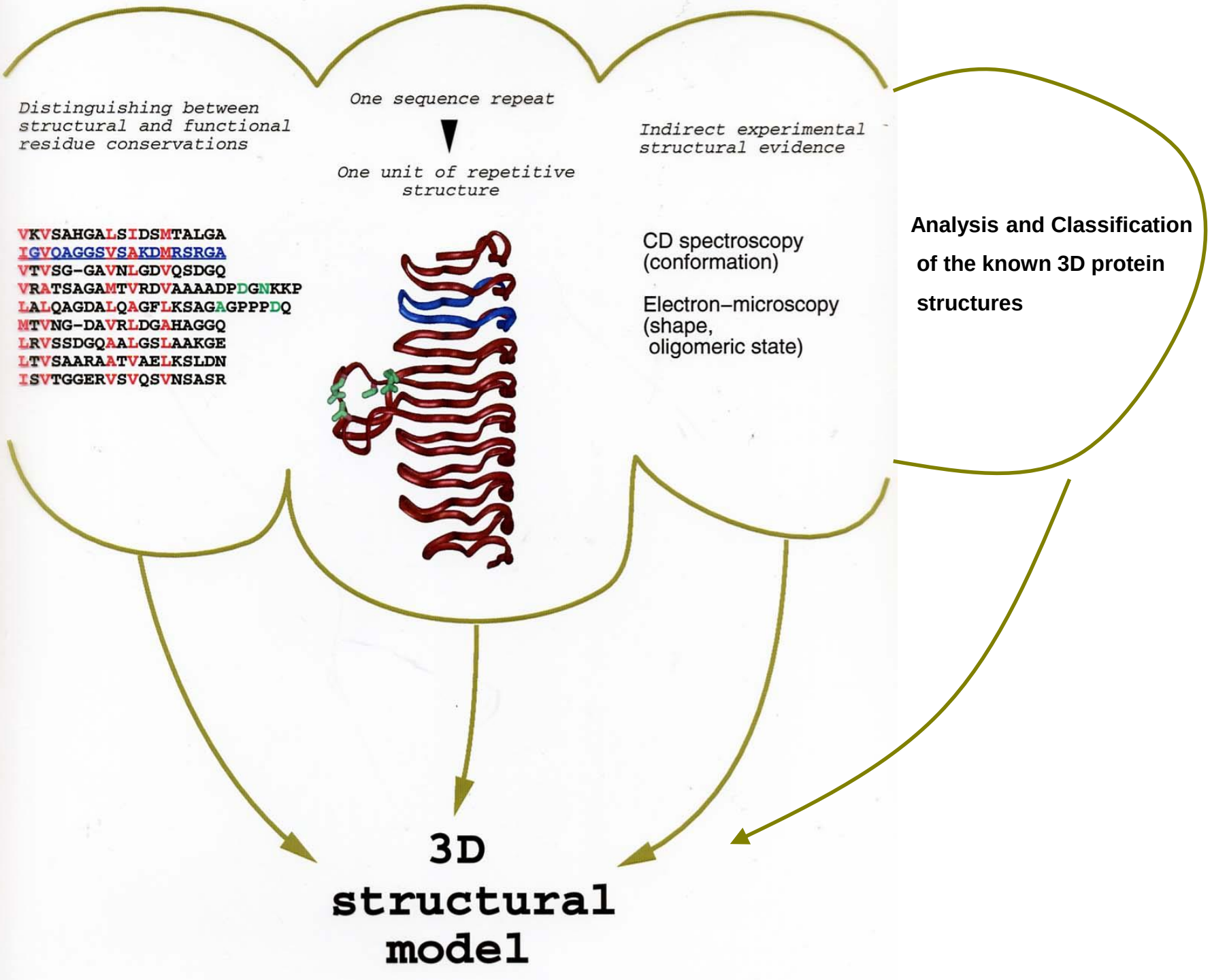
# Distinguishing between structural and functional residue conservations

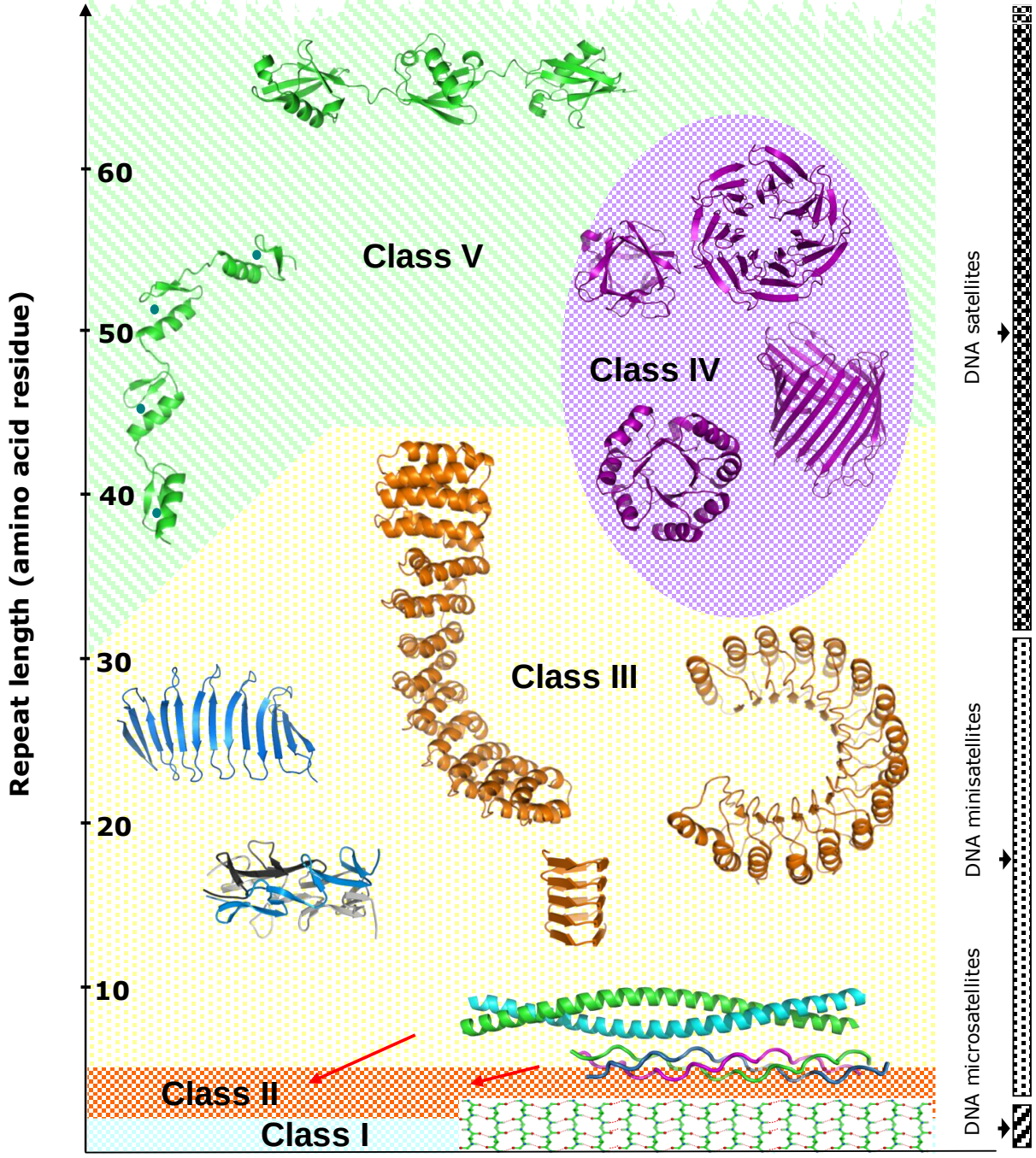
```
protein_human   VKVSAHGALSIDSMTALGAIGVQAGGSVSAKDMRSRGAVTVSG.GAVN
protein_rat     VHLNAHGALTIKTMYSGNHISVQAGSHVSAREMHQSAFVTVHCAGSVN
protein_yeast   VKVSFQSSLSIDSMTALGAIGVVSSGSVDAKDMRSRGAVWVSG.GAVK

LGDVQSDGQ.VRATSAGAMTVRDVAAADPDGNKKPLALQAGDALQAGFLKSAGAGPPPDQM...
LGDVQSWGQFVHASDGFCMTVRDVSYRDGDPNRYTLGLQAGHALQAYYLRSSSAA..NDQM...
LAAVNNDGQ.VRATSAGAMCVWDVAAQDPDGNKKPLALSSGDGLKAGFLKSAGAGPPPDLM...
```

protein_human


Distinguishing between structural and functional residue conservations
One sequence repeat
One unit of repetitive structure
Indirect experimental structural evidence
VKVSAHGALSIDSMTALGA
IGVQAGGSVSAKDMRSRGA
VTVSG-GAVNLGDVQSDGQ
VRATSAGAMTVRDVAAAADPDGNKKP
LALQAGDALQAGFLKSAGAGPPPDQ
MTVNG-DAVRLDGAHAGGQ
LRVSSDGQAALGSLAAKGE
LTVSAARAATVAELKSLDN
ISVTGGERVSVQSVNSASR
CD spectroscopy (conformation)
Electron–microscopy (shape, oligomeric state)
Analysis and Classification of the known 3D protein structures
3D structural model



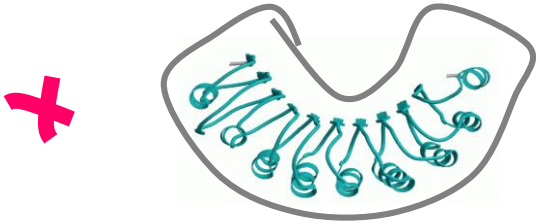

**Analysis and Classification of the known 3D protein structures**

Kajava (2012) *J. Struct. Biol.* 179:279–288.

# Prédiction et modélisation de protéines à séquences répétitives

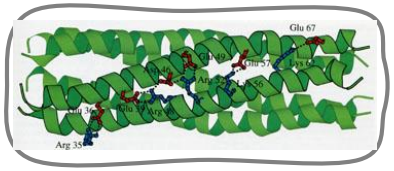
Leucine-rich repeat proteins

Kajava et al., (1995) Structure, 3, 863
Kajava (1998) J.Mol.Biol. 277,519



α-Helical Coiled coil pentamer of COMP

Kajava (1996) Proteins, .24, 218

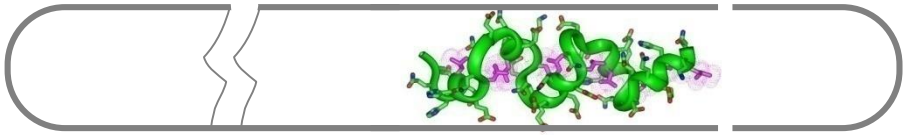



Filamentous Hemagglutinin Adhesin of Bordetella pertussis (56 nm long)

Kajava et al. (2001) Mol. Microbiology, 42, 279

Human involucrin (46 nm long)

Kajava (2000) FEBS Lett. 473, 127

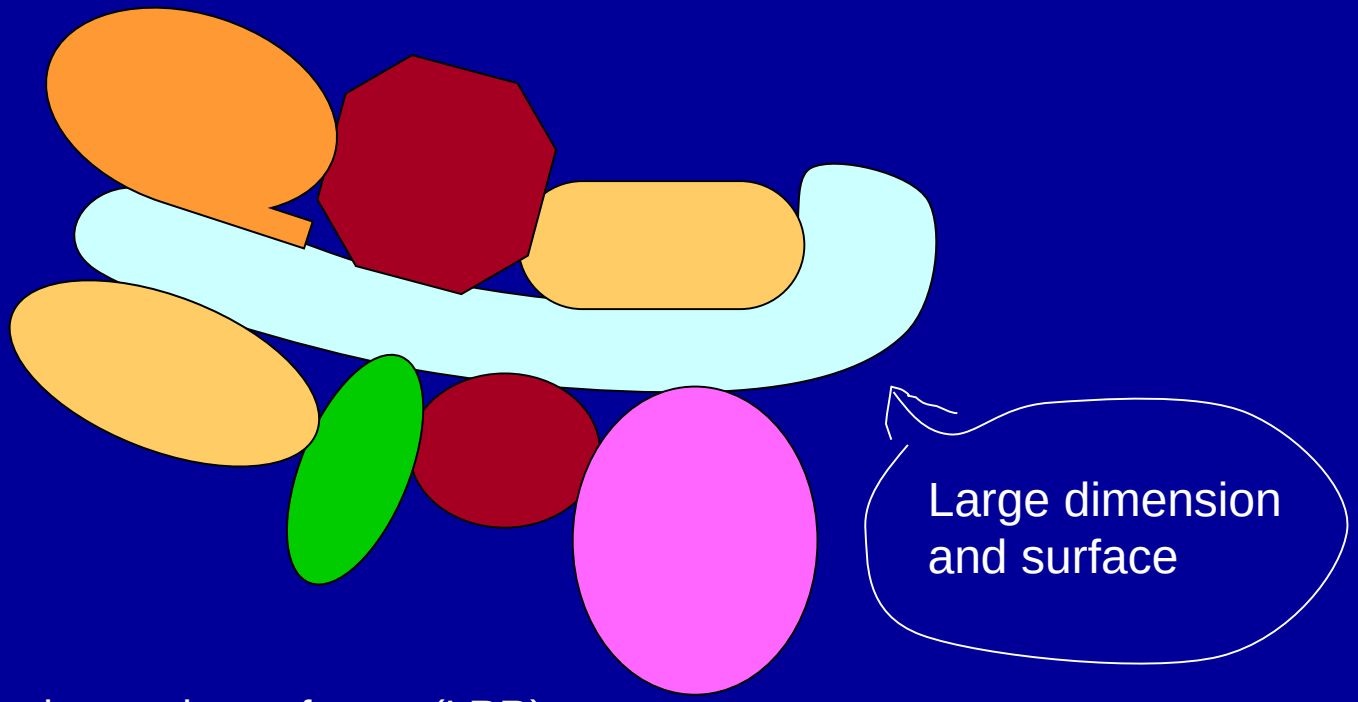


Rpn1 and Rpn2 subunits of eukaryotic proteasome

Kajava (2002) J.Biol.Chem. 277, 49791

# Functions of proteins with tandem repeats

# Protein-protein interactions




Rab geranylgeranyl transferase (LRR)

Anaphase-promoting complex (TPR)

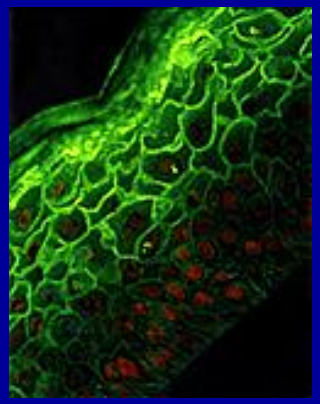
Cleavage Stimulatory Factor  (HAT repeats)

# Structural proteins

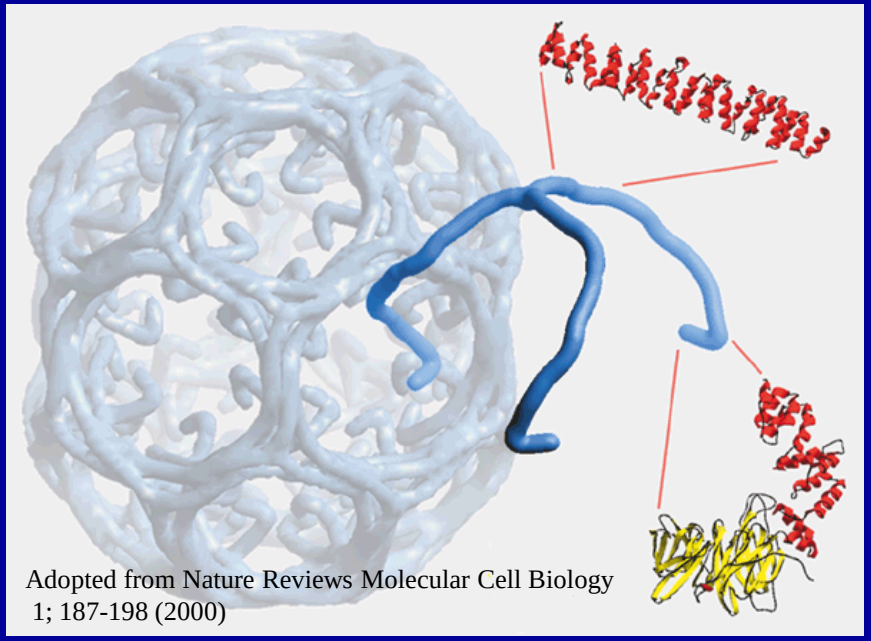
Fibrous proteins: silk, collagen, etc

Building blocks for open-work lattices

Involucrin
(cell envelope)

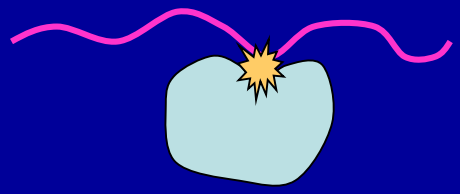


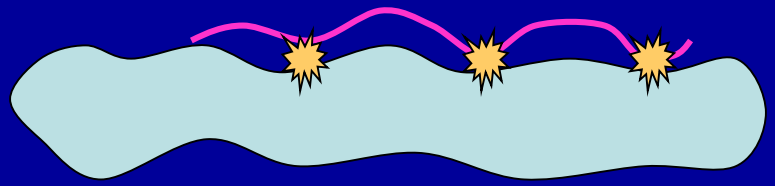
Clathrin



Adopted from Nature Reviews Molecular Cell Biology 1; 187-198 (2000)

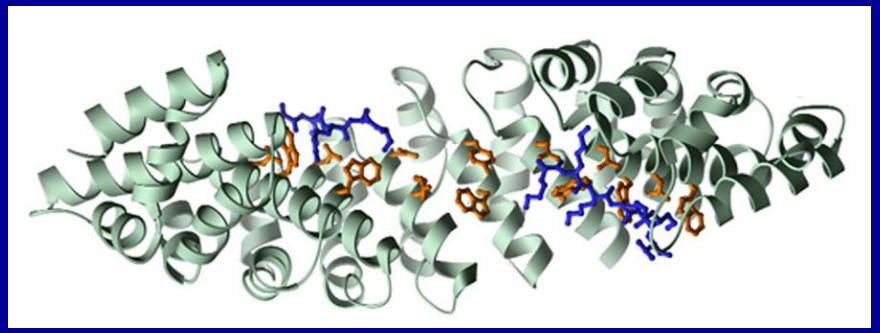
# Multivalent binding





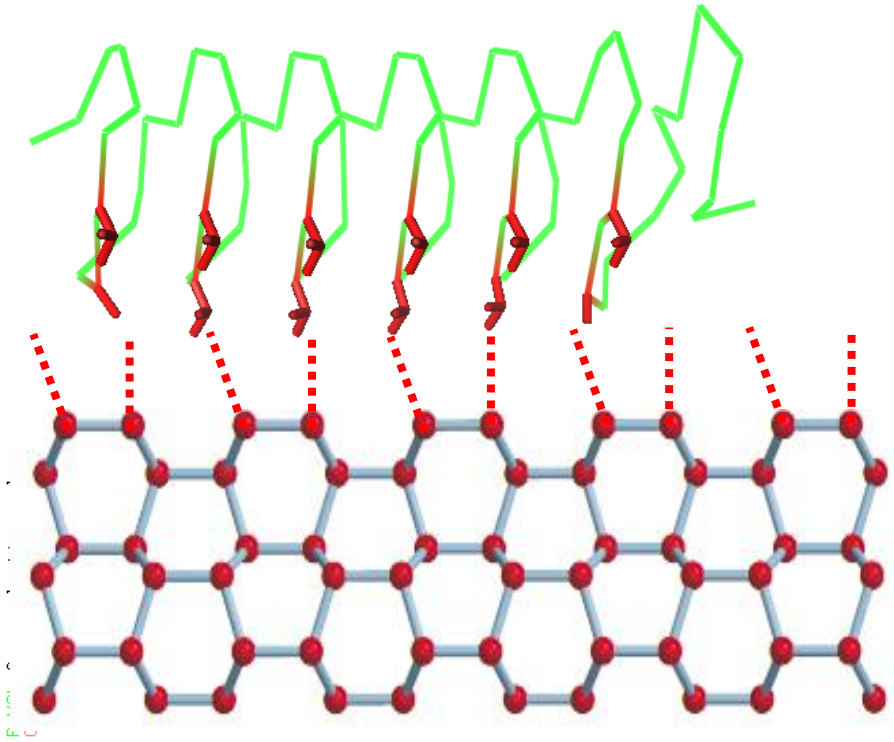
Karyopherin alpha
and peptide with nuclear
localization signal

(Conti et al., Cell, Vol 94, 193-204)



Bacterial outer surface
proteins (pertactin, FHA)

# Antifreeze proteins

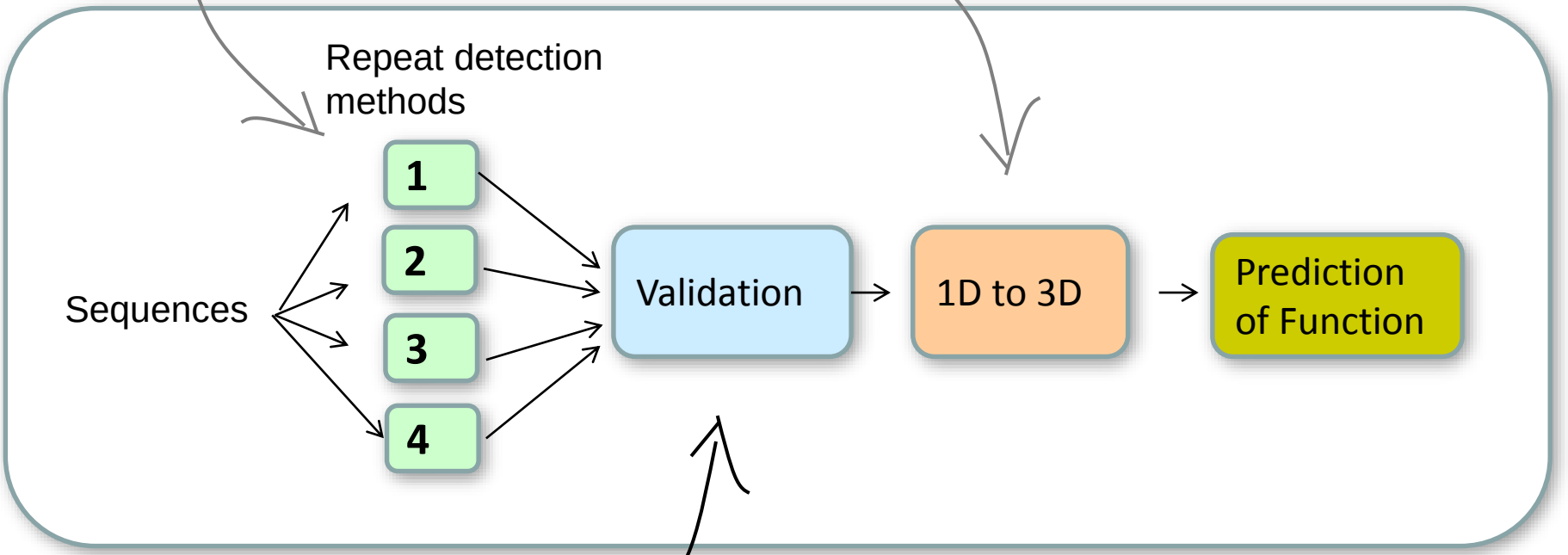


**Antifreeze protein**
**from *Tenebrio molitor***

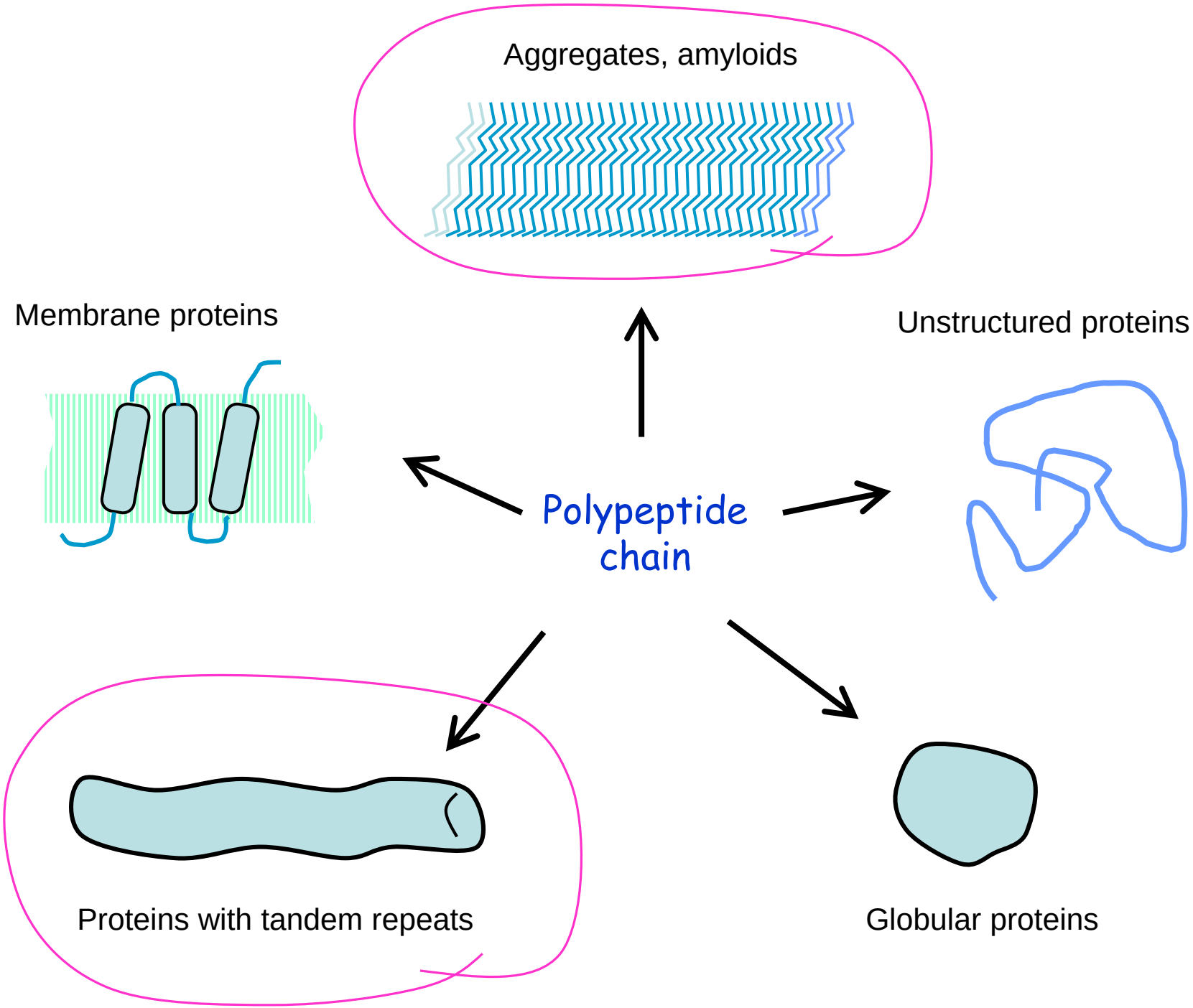
# Pipeline for large-scale identification of protein tandem repeats and prediction of their structure and function

**T-REKS**
Jorda and Kajava. (2009) *Bioinformatics*

Protein Tandem Repeats: **from Sequence to Structure**
Kajava, (2012) *J. Struct. Biol.*




**Repeat or not repeat?**—Statistical validation of tandem repeat prediction in genomic sequences.
Schaper, Kajava, Hauser and Anisimova, (2013) *Nucl. Acids Res*


Aggregates, amyloids
Membrane proteins
Unstructured proteins
Polypeptide chain
Proteins with tandem repeats
Globular proteins

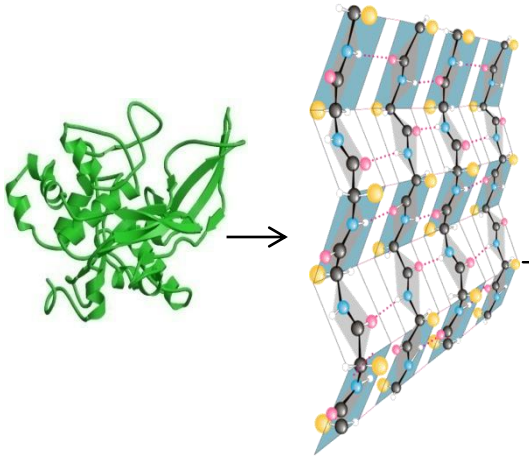
# AMYLOID FIBRILS

Presence of amyloid fibrils is connected with serious **neurodegenerative diseases**, including Alzheimer's, Parkinson's, Huntington's diseases, and also the transmissible prion diseases.

**Also more and more functional amyloids**

# Existing methods to predict amyloids from amino acid sequences

AA composition

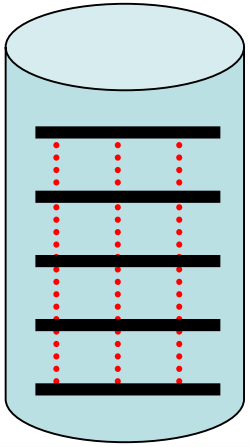



| Name | Basic approach | Server/Website |
|---|---|---|
| AGGRESCAN | Composition of amino acids | http://bioinf.uab.es/aggrescan/ |
| FoldAmyloid | Composition of amino acids | http://bioinfo.protres.ru/fold-amyloid/oga.cgi |
| Zyggregator | Amino acid aggregation propensities and properties of β-structural conformation | http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php |
| TANGO | Properties of β-structural conformation | http://tango.crg.es/ |
| PASTA | Pairwise interactions within the β-sheets | http://protein.bio.unipd.it/pasta/ |
| BetaScan | Pairwise interactions within the β-sheets | http://groups.csail.mit.edu/cb/betascan/betascan.html |
| 3D Profile method (ZipperDB) | Amyloid-like structures of short peptides | http://services.mbi.ucla.edu/zipperdb/submit |
| Waltz | Amyloid-like structures of short peptides | http://waltz.switchlab.org/ |

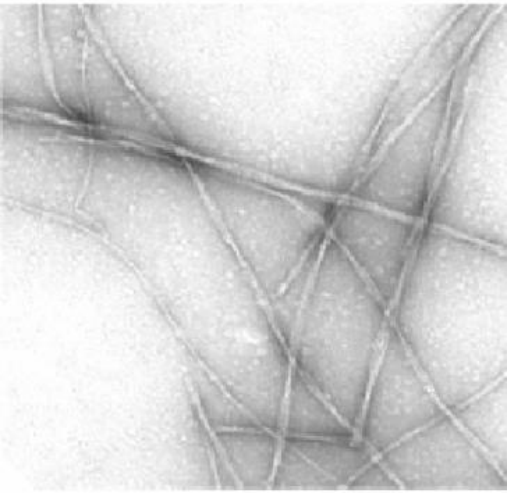
**Reviewed in Ahmed and Kajava (2013) FEBS Lett**

## BEFORE 2000

From X-ray diffraction
« cross-beta »structures



fiber axis
and
direction
of H-bonds



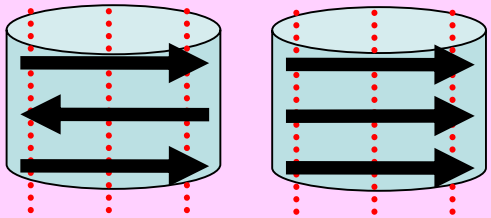
rom EM
straight, unbranched fibrils
4 to 15 nm in diameter

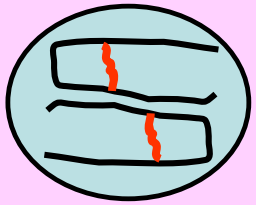
## NEW METHODS AFTER 2000

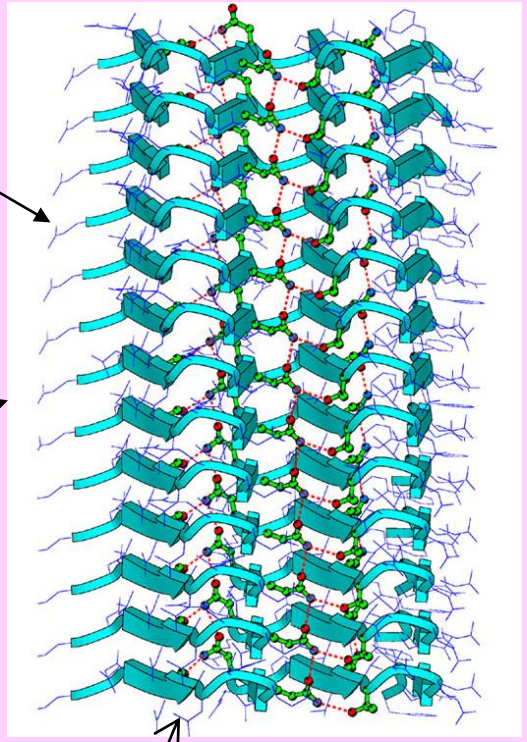
✓ Cryo-EM

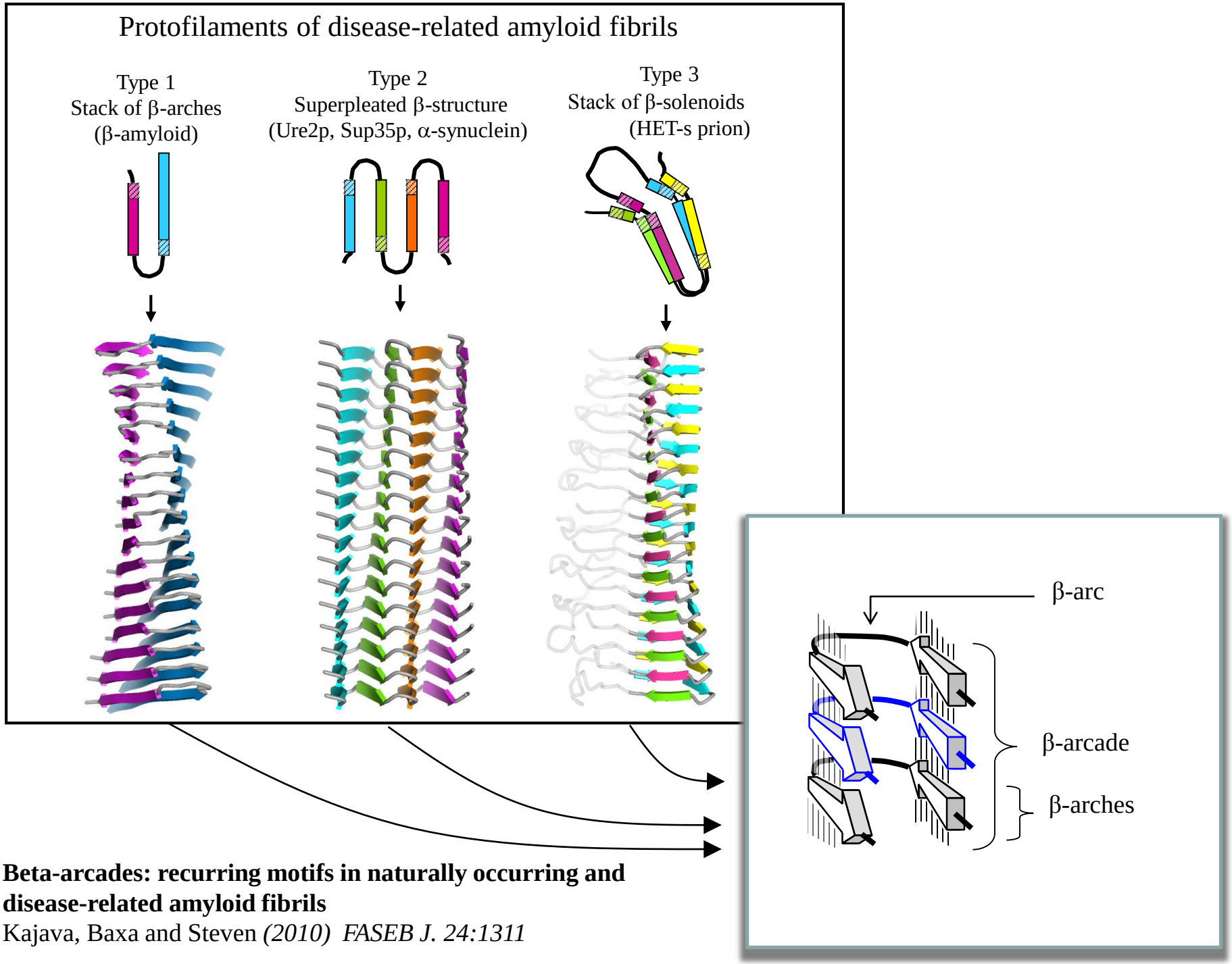
✓ STEM ( number of peptides
in fibril cross-section)

✓ ssNMR,
EPR spectroscopy



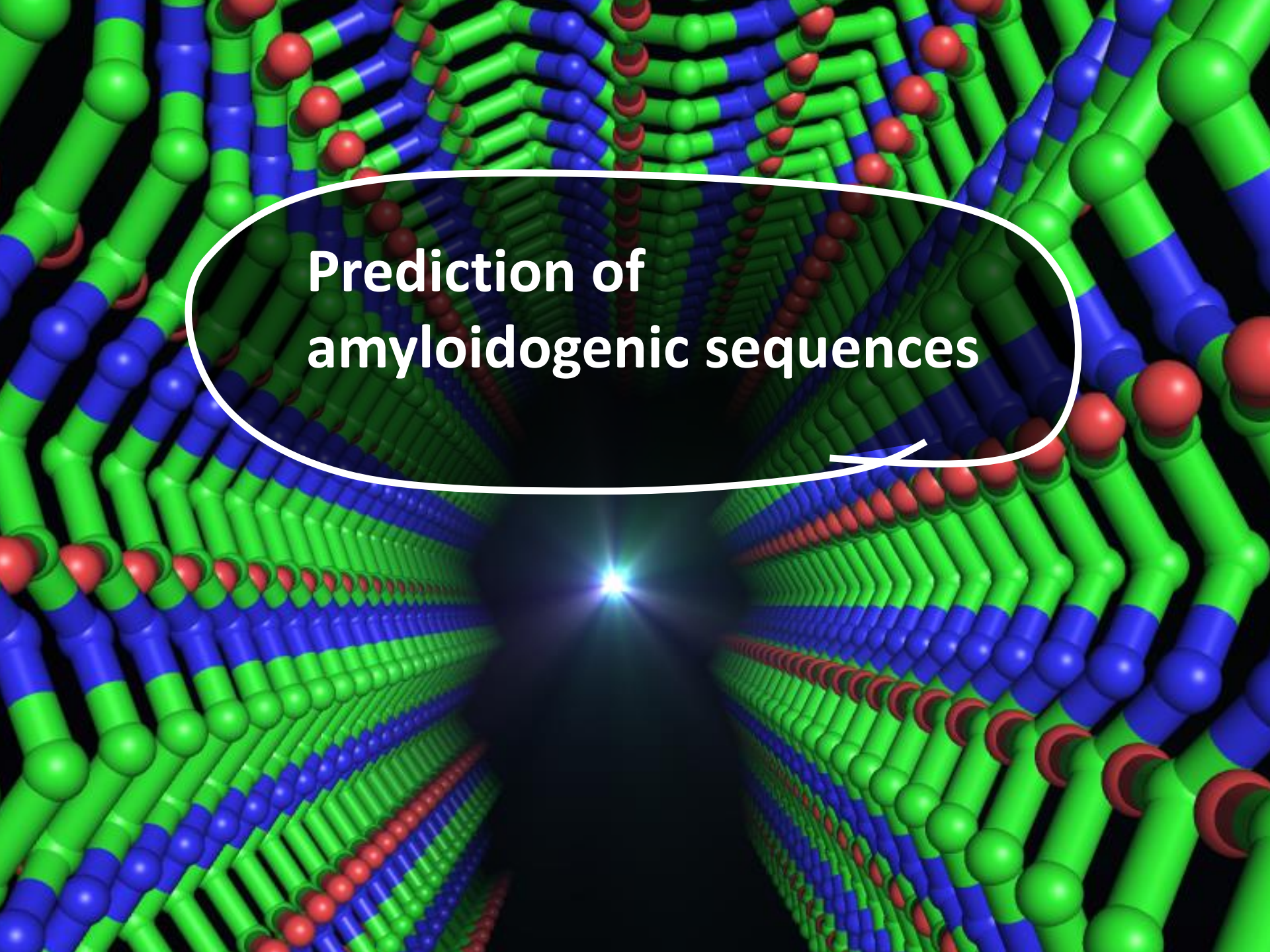






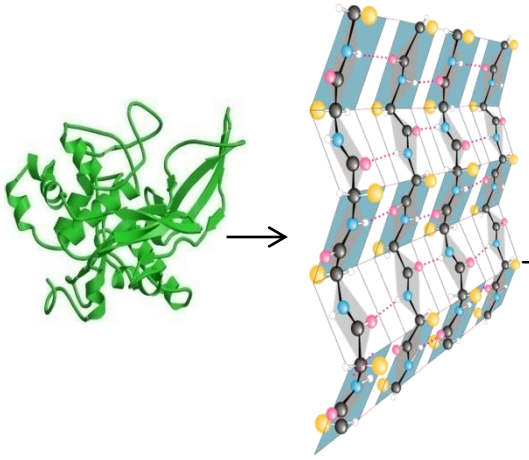

**Beta-arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils**

Kajava, Baxa and Steven *(2010) FASEB J. 24:1311*


Prediction of
amyloidogenic sequences

# Existing methods to predict amyloids from amino acid sequences

AA composition




| Name | Basic approach | Server/Website |
|---|---|---|
| AGGRESCAN | Composition of amino acids | http://bioinf.uab.es/aggrescan/ |
| FoldAmyloid | Composition of amino acids | http://bioinfo.protres.ru/fold-amyloid/oga.cgi |
| Zyggregator | Amino acid aggregation propensities and properties of β-structural conformation | http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php |
| TANGO | Properties of β-structural conformation | http://tango.crg.es/ |
| PASTA | Pairwise interactions within the β-sheets | http://protein.bio.unipd.it/pasta/ |
| BetaScan | Pairwise interactions within the β-sheets | http://groups.csail.mit.edu/cb/betascan/betascan.html |
| 3D Profile method (ZipperDB) | Amyloid-like structures of short peptides | http://services.mbi.ucla.edu/zipperdb/submit |
| Waltz | Amyloid-like structures of short peptides | http://waltz.switchlab.org/ |

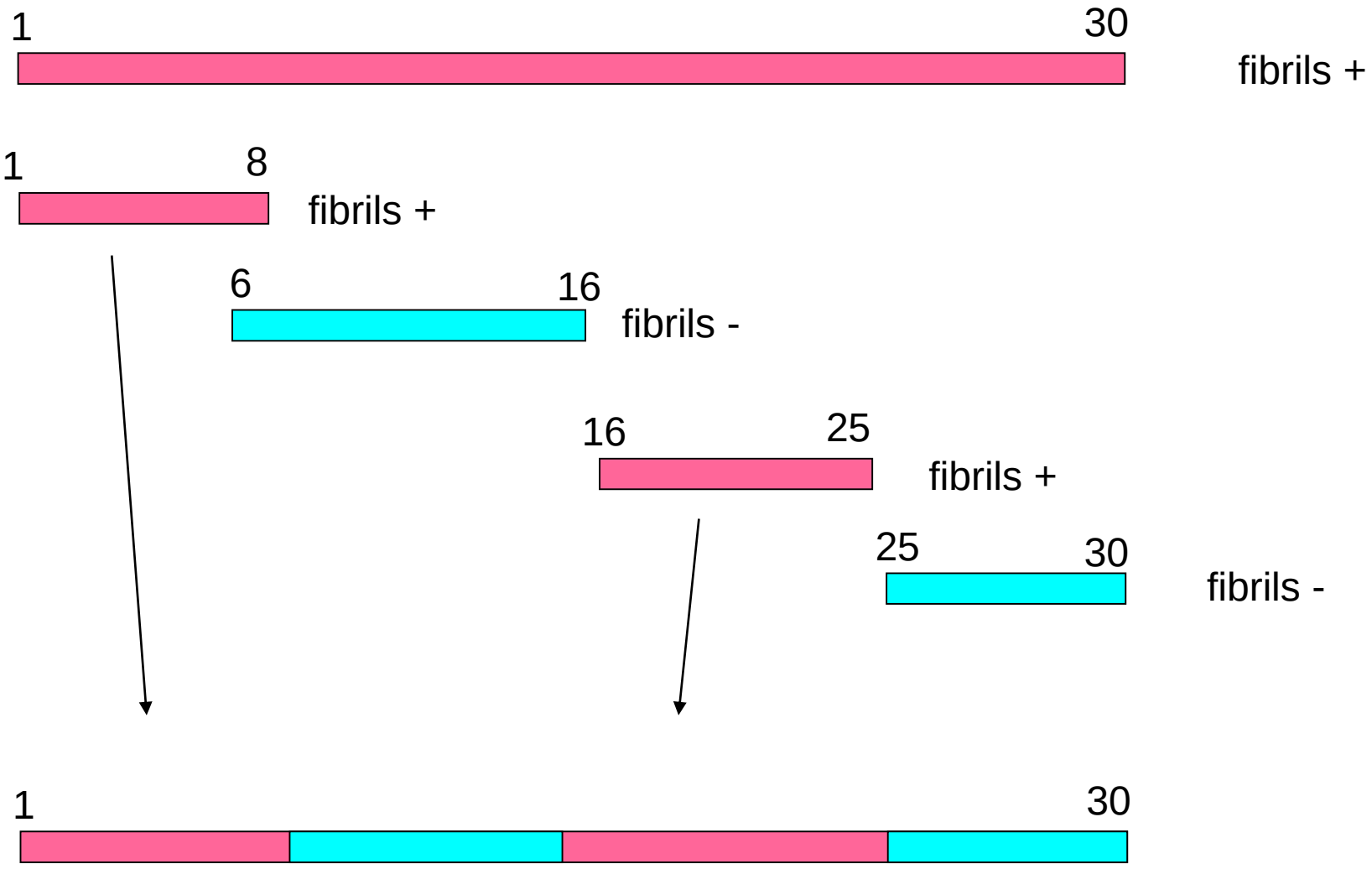
**Reviewed in Ahmed and Kajava (2013) FEBS Lett**

These algorithms assume that a

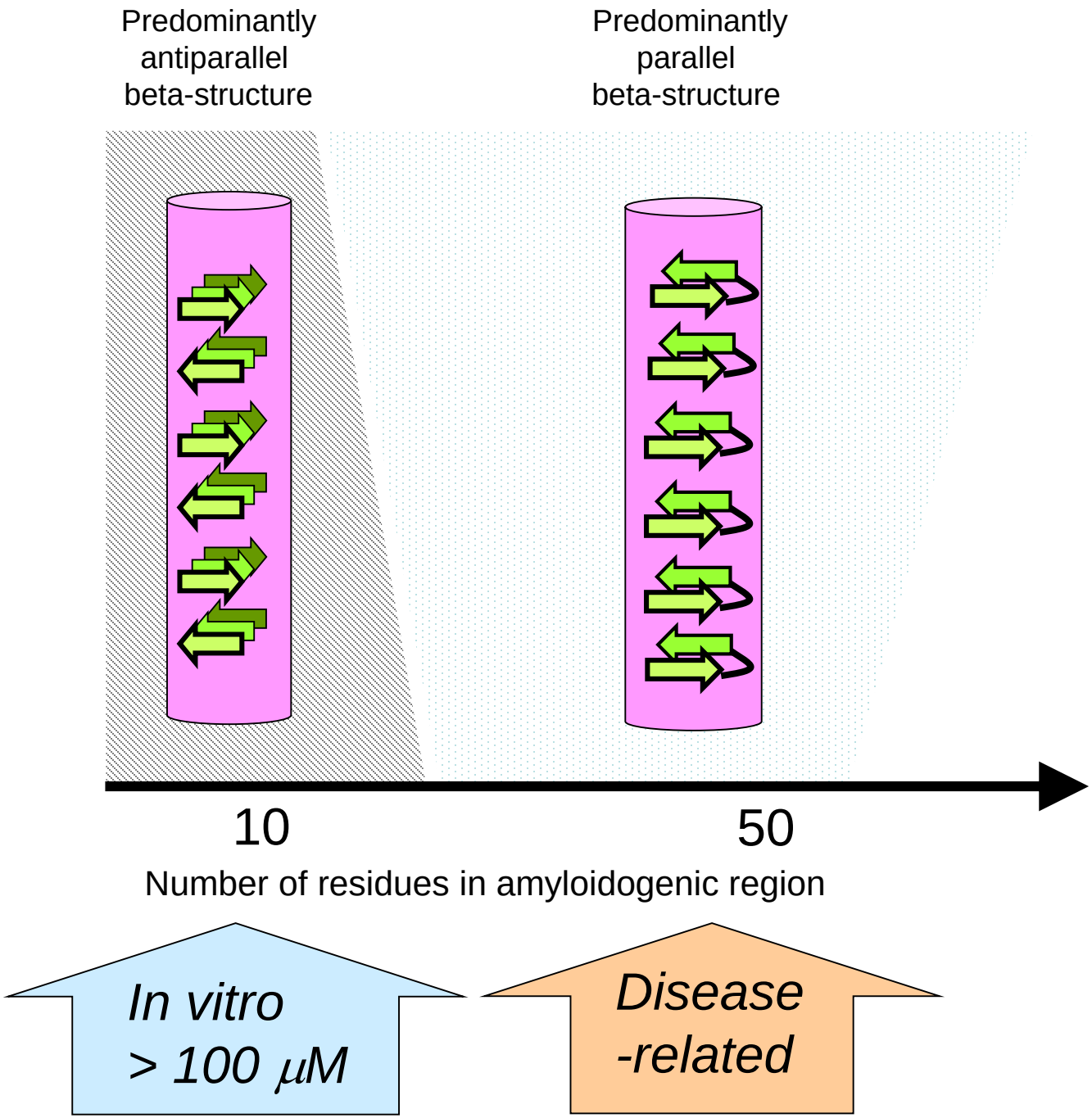
**short sequence (about 6 residues) is sufficient**

to trigger the amyloid formation

# Short vs long peptides



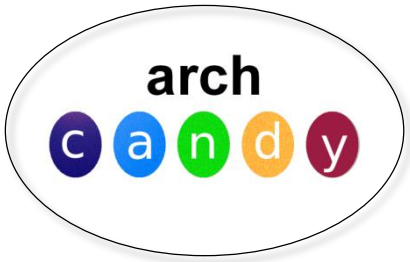

Assumption: short regions of long peptides determine their amyloidogenicity

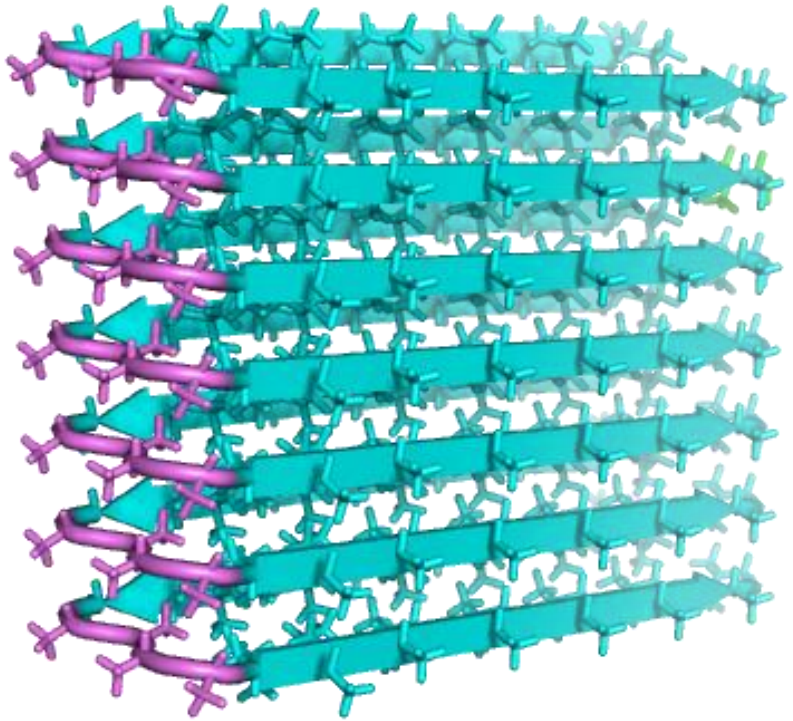

Predominantly antiparallel beta-structure
Predominantly parallel beta-structure
10
50
Number of residues in amyloidogenic region
In vitro
> 100 μM
Disease
-related

# Structure-based approach to predict amyloidogenicity

An approach based on the assumption that sequences that are able to form β-arcades are amyloidogenic.

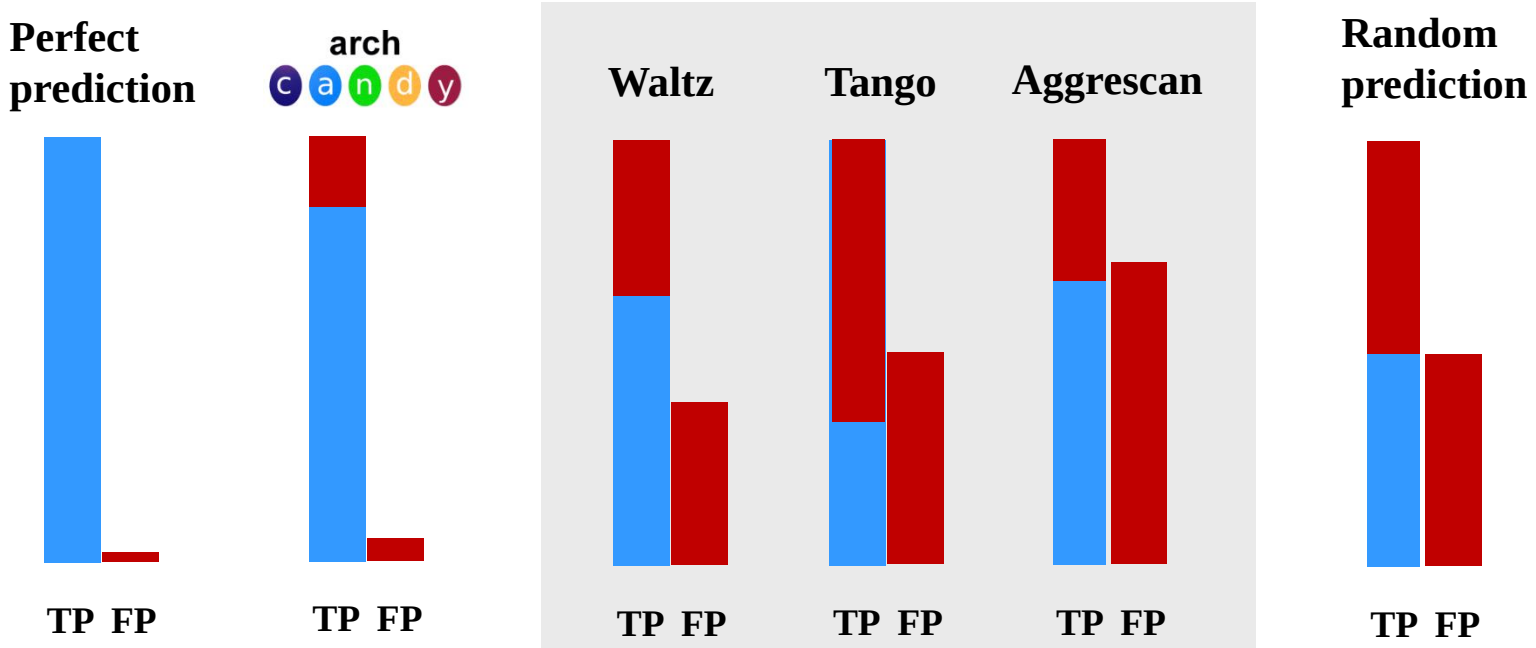
↓



Ahmed, Znassi, Château and Kajava (2015)
*Alzheimers & Dementia*

# Performance of existing methods on datasets of proteins




wrong prediction

correct prediction

**TP** – True Positive, **FP** – False Positive

# ArchCandy APPLICATIONS

**Personal medicine:** patient-specific assessment of risk to develop age related diseases

Comparative analysis of proteomes, search for new prions and amyloidogenic proteins

**Biotechnology:** recommendations to make proteins soluble, new materials