PROTEIN REPEATS :
FROM SEQUENCE TO STRUCTURE AND FUNCTION.

Dr Andrey Kajava

Group of Structural Bioinformatics and Molecular Modeling

Centre de Recherches de Biochimie Macromoléculaire, CNRS

Montpellier, FRANCE

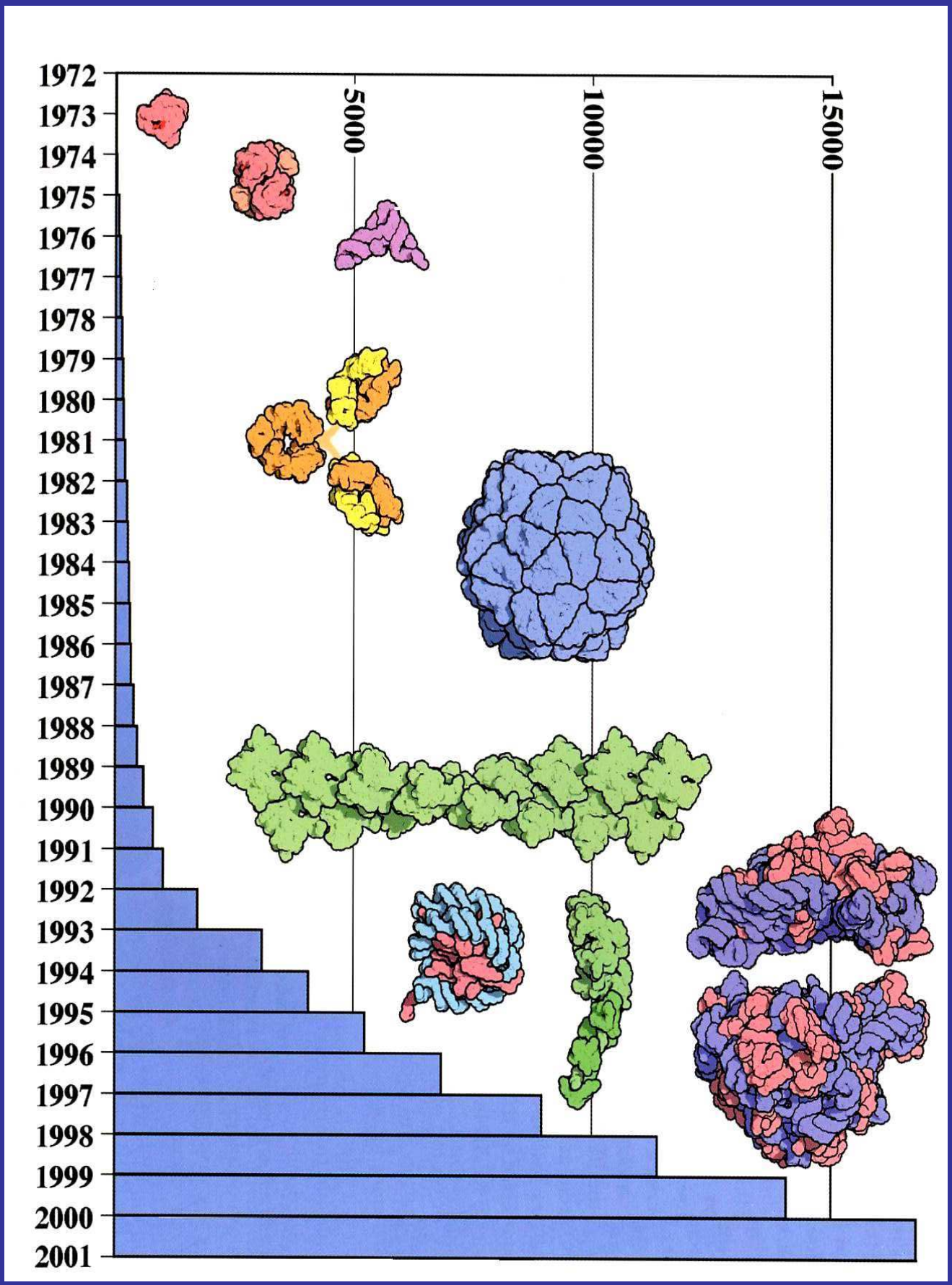
PROTEIN SEQUENCE – STRUCTURE - FUNCTION

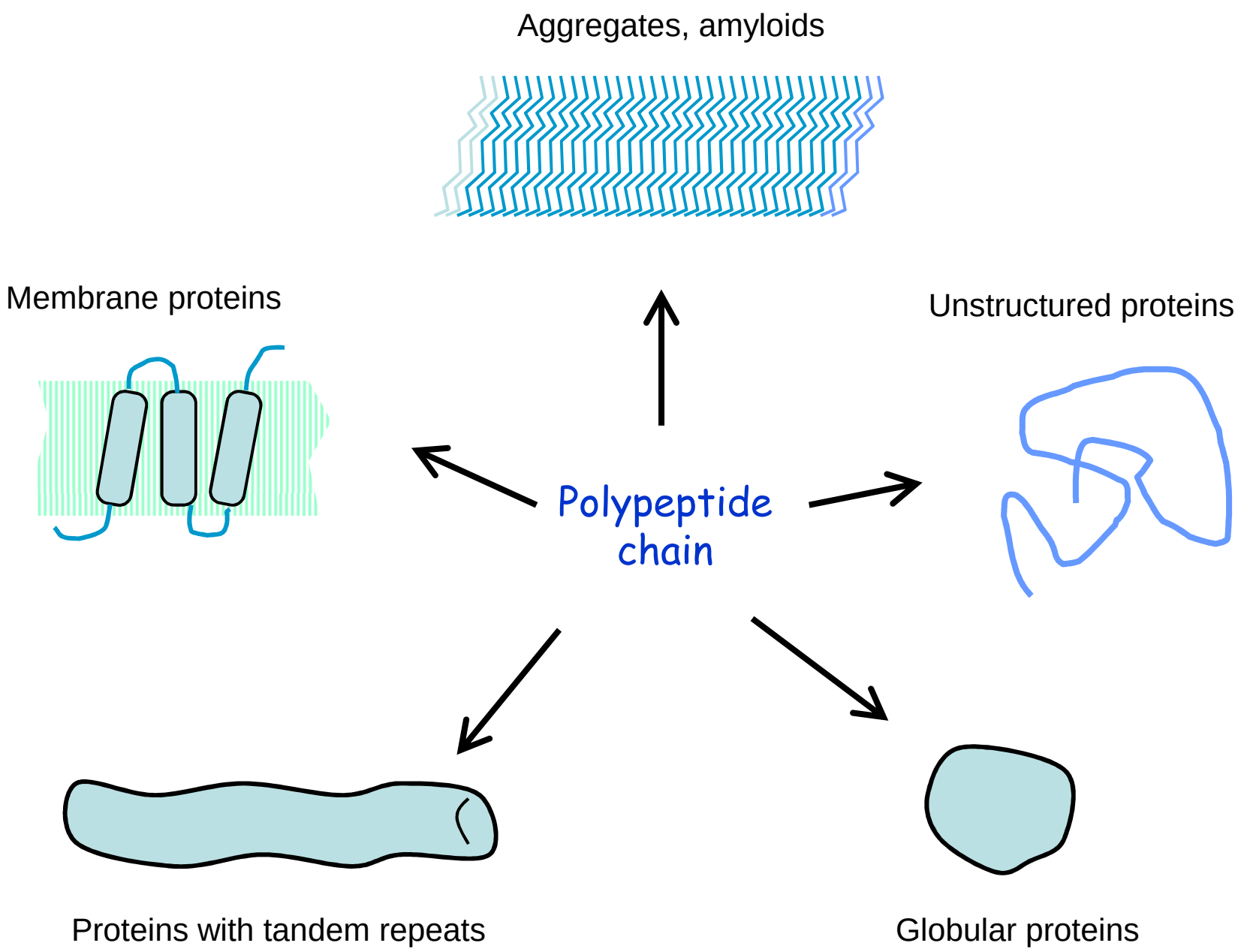
Principal activities:

Bioinformatics analysis of protein sequences and structures

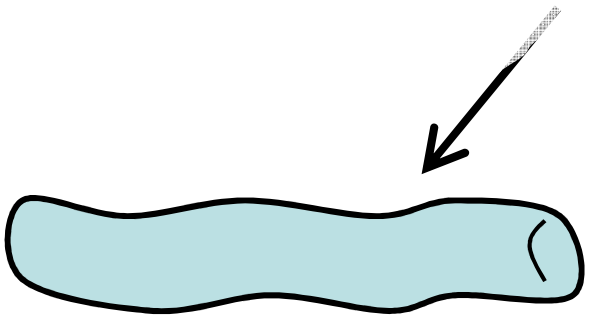
Structural prediction and molecular modelling

Molecular design

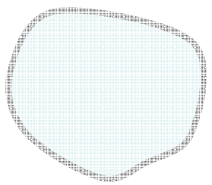

1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
5000
10000
15000


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

- ✓ Identification of protein repeats
- ✓ Analysis and Classification of the known 3D protein structures
- ✓ Structural prediction
- ✓ Experimental tests
- ✓ Evolution of proteins with repeats
- ✓ Applications in medicine, material science and nanotechnologies



Proteins with tandem repeats



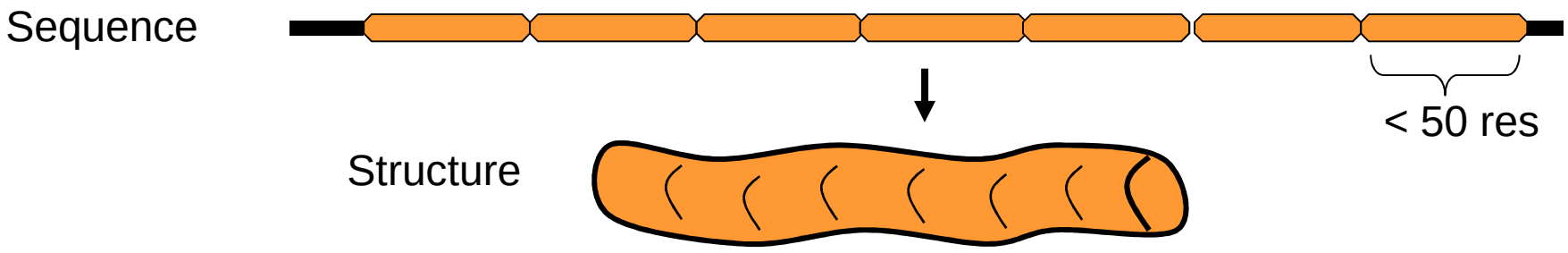
Globular proteins

Proteins with tandem repeats

Proteins with internal duplications represent a large portion of genomes

***E. coli* (7%), *S. cerevisiae* (17%),Human (27%)**
All SwissProt (14%)

Pellegrini et al. (1999) *Proteins* 35:440



only ~ 2% of known 3D structures

Difficulties of experimental (X-ray and NMR) determination of the 3D structure

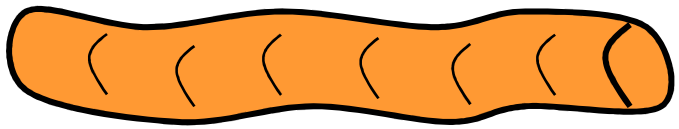
HYBRID APPROACHES TO OBTAIN 3D STRUCTURE

Bioinformatics analysis, structural prediction, molecular modeling

3D structure

It is possible to get a reliable 3D structural model based on sequence analysis!

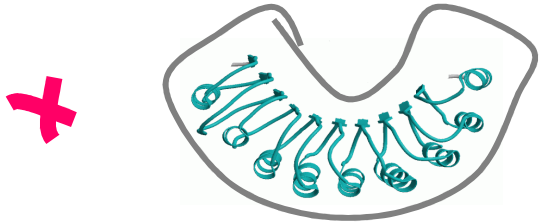
Incomplete experimental structural data (EM, CD, etc)



Prédiction et modélisation de protéines à séquences répétitives

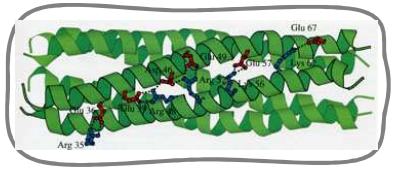
Leucine-rich repeat proteins

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



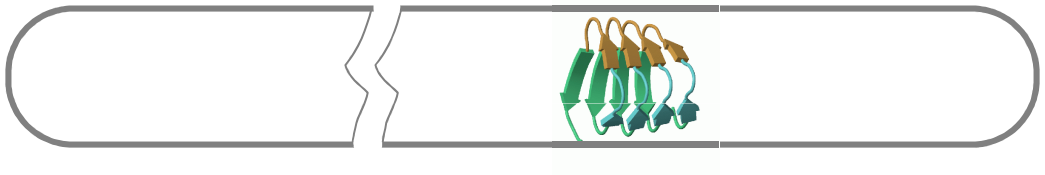
α-Helical Coiled coil pentamer of COMP

Kajava (1996) Proteins, .24, 218



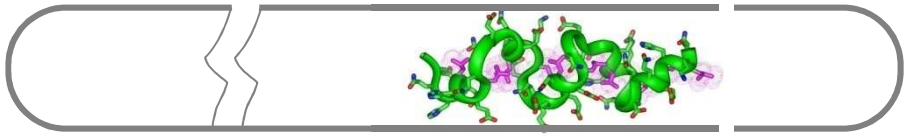
Filamentous Hemagglutinin Adhesin of Bordetella pertussis (56 nm long)

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

Structural-Functional Annotation of Genomes

Proteins with aperiodic sequences and globular domains

Proteins with tandem repeats

☺ Bioinformatcs tools ☹

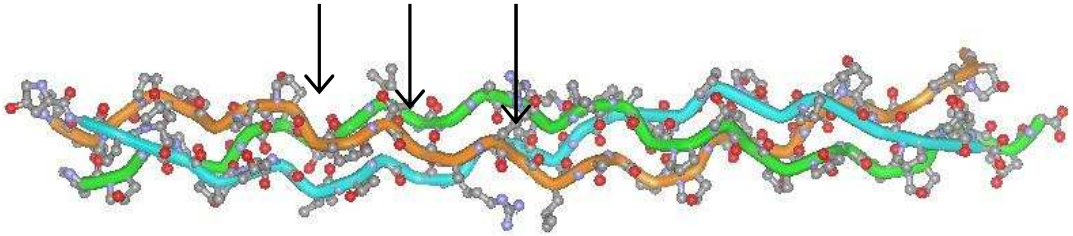
☹ *Ab initio* structural prediction ☺

IDENTIFICATION

OF PROTEIN REPEATS

PPGPPGPPGPPGPPGPPGPPGPPGPPGPPG

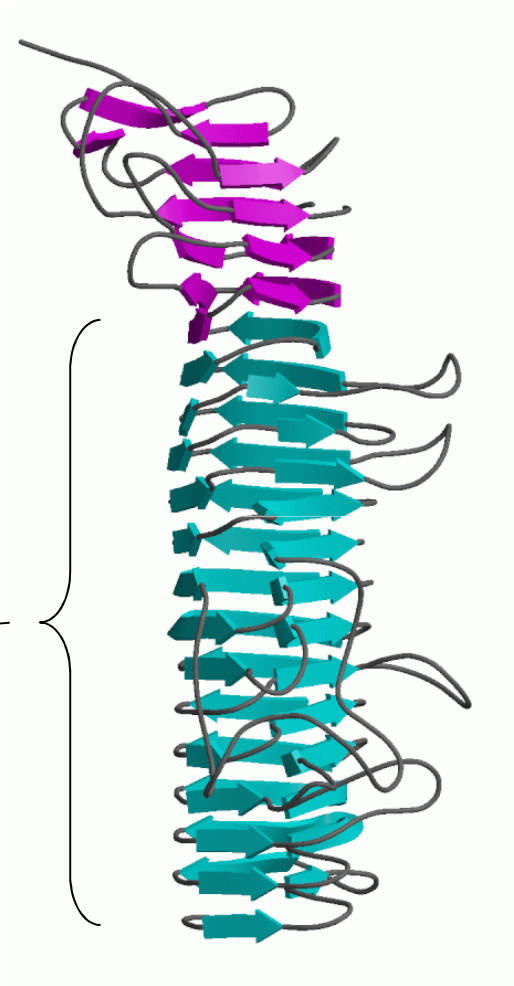
PPGPEGPPGITGARGLAGPPGPPGKPGPPG



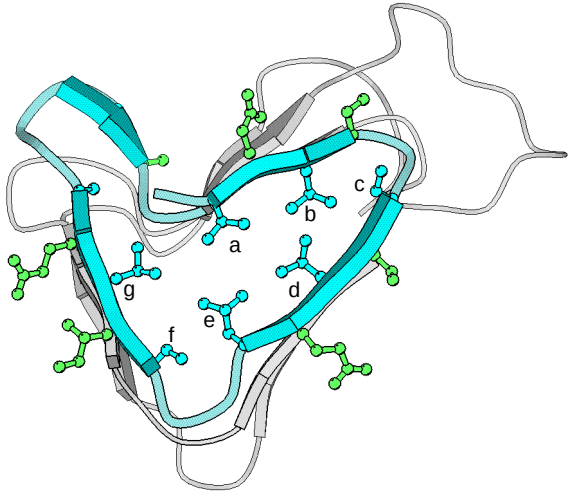
Collagen

Pertactin from *Bordetella pertussis*

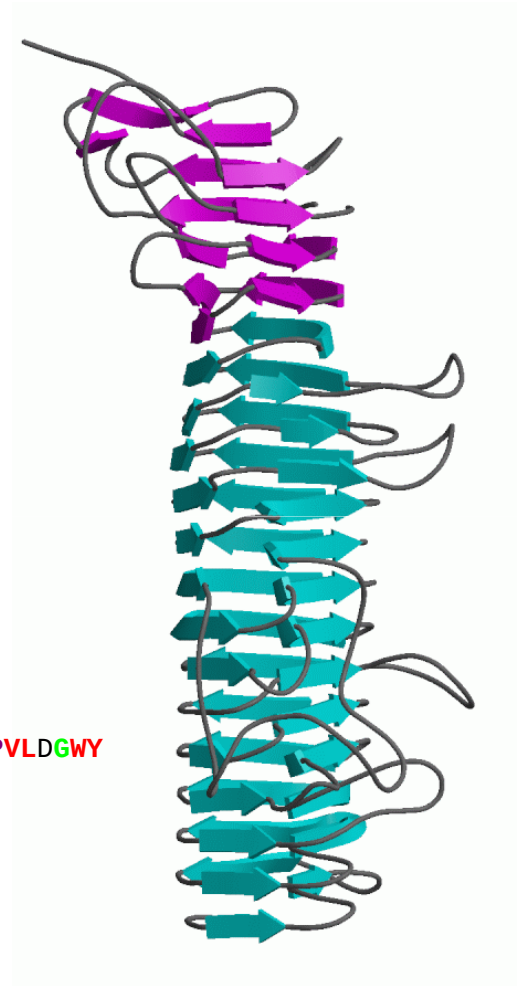
```
 GILLENPAAELQFRNGSVTSSGQLSDDGIRRFLG
    TVTVKAGKLVADHATLANVGDTWDDDGI
      ALYVAGEQAQASIADSTLQGAG
       GVQIERGANVTVQRSAIVDG
GLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPA
     AVSVLGASELTLDGGHITGGRAA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWY
    GVDVSGSSVELAQSIVEAPELGA
    AIRVGRGARVTVSGGSLSAPHGN
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 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	-

Kajava, (2011) *J. Struct. Biol.*

.:: BiSMM - Structural Bioinformatics and Molecular Modeling - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils ?

http://bioinfo.montp.cnrs.fr/?r=t-reks

Les plus visités À la une .:: BiSMM - Bioinforma... PubMed Home

.:: BiSMM - Structural Bioinform...

Structural Bioinformatics and Molecular Modeling
CRBM-CNRS Montpellier

[show login]

Home | Research | Publications | Tools | People

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.
```

Protein Repeat DataBase (PRDB)
http://bioinfo.montp.cnrs.fr /?r=repeatDB

http://bioinfo.montp.cnrs.fr/?r=repeatDB

Les plus visités À la une

.:: BiSMM - Structural Bi

release of 23 - Mar - 2009 containing 25914 sequences.

51685 repeats found in 33151 proteins of SwissProt release of 16 - Feb - 2009 containing 364403 sequences.

Help page

with tandem

Database	SwissProt database			
Kingdom	select the kingdom to sear			
Organism	select the organism to sea			
Number of repeats	from:	to:	repeats	
Repeat unit length	from:	to:	residues	
Tandem region length	from:	to:	residues	
Structure forming potential	from:	to:		
Level of perfection	from: 0.70	to: 1.00		
Consensus pattern				
Protein length	from:	to:	residues	
Keyword	All keywords			
Subcellular localization	select a general localizatio	select a specific localizatic		
Molecular function	select a general molecular	select a specific molecular		
Pfam domains	select a general domain	select a specific domain ki		
Gi-ref				
Redundancy filter	☐			
Output format	Html output			

QueryBox

There is no query saved.

View Repeats

Methods for detection of tandem repeats in proteins

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	-

Kajava, (2011) *J. Struct. Biol.*

Cargo recognition complex

a- Helical solenoid fold prediction for the N-terminal part of vps35 (orange in (d))

b- 2D class averages from negative stain electron microscopy

c- 2D projections of the full cargo recognition complex model (d) for comparison with the EM class averages in (b)

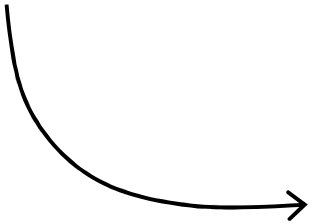
(Hierro et al., Nature, 2007)

Bar: 100Å

The α-solenoid fold extends the full length of Vps35 and Vps26 is bound at the opposite end from Vps29.

```
 ***   **        * *      ** * *   ***
GITLENPSS-------AAELQFRN-GSVTNSGQLSDGI
TITLKATSS--------AKLVADH-ASVANVGQTWDGI
```

```
 ***   **         * *      ** * *   ***
GITLENPSS-------AAELQFRN-GSVTNSGQLSDGI
TITLKATSS--------AKLVADH-ASVANVGQTWDGI
```



```
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY';
MA   /DISJOINT: DEFINITION=PROTECT; N1=1; N2=43;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE; R1=44.55; R2=-0.0035;
MA      R3=0.7386; R4=1.001; R5=0.208; TEXT='ZScore';
MA   /NORMALIZATION: MODE=2; FUNCTION=LINEAR; R1=0.0; R2=0.1;
MA      TEXT='OrigScore';
MA   /CUT_OFF: LEVEL=0; SCORE=90; N_SCORE=7.0; MODE=1;
MA   /DEFAULT: MI=-26; I=-3; IM=0; MD=-26; D=-3; DM=0;
MA   /M: SY='F';M=-2,-3,-3,-4,2,-3,-2,1,-2,0,-1,-2,-3,-3,-4,-2,-1,0,-5,2;
MA   /M: SY='I';M=-1,-5,-2,-3,-2,-3,0,1,1,-1,1,-1,-2,-1,1,-1,0,1,-4,-4;
MA   /M: SY='A';M=2,-3,1,0,-5,2,-2,-1,-1,-3,-2,1,1,0,-2,2,2,0,-8,-5;
MA   /M: SY='L';M=-3,-8,-5,-4,2,-6,-2,2,-4,6,4,-3,-3,-2,-3,-3,-2,1,-3,0;
MA   /M: SY='Y';M=-4,-2,-6,-6,9,-7,0,-1,-5,-1,-3,-3,-6,-5,-6,-4,-4,-4,-1,11;
MA   /M: SY='D';M=1,-6,3,3,-7,0,0,-2,-1,-4,-3,2,0,1,-2,0,0,-2,-9,-6;
MA   /M: SY='Y';M=-5,-3,-6,-6,10,-7,-1,-1,-2,-1,-2,-3,-6,-5,-5,-4,-4,-4,-1,11;
MA   /M: SY='K';M=-1,-6,1,1,-4,-2,0,-2,2,-3,-1,1,-1,1,1,0,0,-3,-7,-6;
MA   /M: SY='A';M=1,-4,1,0,-5,1,-1,-1,0,-3,-1,1,0,0,0,1,1,-1,-7,-6;
MA   /M: SY='R';M=0,-5,0,0,-5,-1,0,-1,1,-3,-1,1,0,1,1,0,0,-2,-5,-5;
MA      /I: MI=0; I=-2; MD=0; /M: SY='X'; M=0; D=-2;
MA   /M: SY='R';M=0,-5,1,1,-6,0,1,-2,1,-4,-2,1,0,1,2,1,0,-2,-5,-5;
MA   /M: SY='F';M=-3,-7,-6,-6,6,-5,-3,3,-2,5,3,-4,-5,-4,-5,-4,-3,1,-3,3;
MA   /M: SY='Q';M=-1,-6,0,0,-3,-2,1,-1,1,-2,0,0,-1,1,1,-1,0,-1,-6,-4;
MA   /M: SY='K';M=-1,-8,0,1,-3,-2,0,-2,3,-3,0,1,0,2,2,0,0,-3,-6,-6;
MA   /M: SY='G';M=2,-5,1,0,-7,7,-3,-4,-2,-6,-4,1,-1,-2,-4,2,0,-2,-10,-8;
MA   /M: SY='D';M=1,-7,5,4,-8,1,1,-3,0,-5,-3,2,-1,2,-2,0,0,-4,-10,-6;
MA   /M: SY='I';M=0,-5,-1,-2,-2,-2,-1,2,0,0,1,-1,-2,0,0,-1,0,1,-6,-5;
MA   /M: SY='L';M=-2,-6,-5,-5,3,-5,-3,4,-3,6,4,-4,-4,-3,-4,-3,-2,3,-5,0;
MA   /M: SY='Q';M=-1,-5,-1,-1,-3,-2,0,0,0,-2,-1,0,-1,0,0,-1,0,-1,-6,-3;
MA   /M: SY='V';M=0,-4,-3,-4,-1,-3,-3,5,-3,3,3,-2,-2,-2,-3,-2,0,5,-8,-4;
MA   /M: SY='L';M=-1,-6,-3,-3,-1,-3,-2,2,-3,3,2,-2,-2,-2,-3,-2,-1,2,-5,-3;
MA   /M: SY='D';M=0,-6,3,3,-6,0,1,-3,2,-5,-2,2,-1,2,1,0,0,-4,-7,-5;
MA   /M: SY='K';M=-1,-6,0,0,-2,-1,0,-3,3,-4,-1,1,-1,0,1,0,0,-3,-6,-4;
MA   /M: SY='N';M=1,-4,1,1,-5,0,0,-2,0,-3,-2,1,1,0,-1,1,1,-1,-7,-5;
MA      /I: MI=0; I=-1; MD=0; /M: SY='X'; M=0; D=-1;
MA   /M: SY='G';M=1,-5,0,0,-5,1,-2,-1,-2,-3,-2,0,0,-1,-2,0,0,-1,-8,-6;
MA   /M: SY='G';M=1,-6,3,3,-7,3,0,-4,-1,-5,-4,2,-1,1,-2,1,0,-3,-10,-6;
MA   /M: SY='W';M=-9,-12,-9,-11,1,-11,-4,-8,-5,-3,-6,-6,-8,-7,3,-4,-8,-9,26,0;
MA   /M: SY='W';M=-7,-9,-9,-9,0,-9,-4,-5,-5,-1,-4,-6,-7,-6,2,-3,-6,-6,18,-1;
MA   /M: SY='K';M=-1,-7,0,0,-3,-2,0,-2,2,-3,-1,1,-1,1,2,0,-1,-3,-5,-5;
MA   /M: SY='G';M=2,-3,0,-1,-6,3,-3,-2,-3,-4,-3,0,0,-2,-3,1,0,0,-10,-6;
MA   /M: SY='Q';M=-2,-6,0,0,-3,-3,1,-2,0,-2,-1,0,-2,1,1,-1,-1,-3,-5,-3;
MA   /M: SY='T';M=0,-4,-1,-1,-4,0,-2,0,-1,-2,0,0,-1,-1,-1,0,1,0,-7,-5;
MA   /M: SY='T';M=0,-5,0,0,-3,-1,-1,-1,1,-3,-1,1,-1,0,0,1,1,-1,-6,-4;
MA   /M: SY='G';M=0,-5,0,-1,-5,3,-2,-3,-1,-5,-3,0,-1,-1,-1,1,0,-2,-7,-6;
```

```
 ***   **        * *      ** * *   ***
GITLENPSS-------AAELQFRN-GSVTNSGQLSDGI
TITLKATSS--------AKLVADH-ASVANVGQTWDGI
ALYVAGEQ---------AQASIAD-STLQGAG
GVQIERG----------ANVTVQR-SAIVDG
GLHIGALQSLQPEDLPPSRVVLRD-TNVTAVPASGAPA
AVSVLGA----------SELTLDG-GHITGGRAA
GVAAMQG----------AVVHLQR-ATIRRGDAPAGG
GVDVSG-----------SSVELAQ-SIVEAPELGA
AIRVGRG----------ARVTVSG-GSLSAPHGN
VIETGGARRFAPQAAP-LSITLQAGAHAQGKA
LLYRVLPEP--------VKLTLTGGADAQG
DIVATELPSIPGTSIGPLDVALASQARWTG
```

```
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY';
MA   /DISJOINT: DEFINITION=PROTECT; N1=1; N2=43;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE; R1=44.55; R2=-0.0035;
MA      R3=0.7386; R4=1.001; R5=0.208; TEXT='ZScore';
MA   /NORMALIZATION: MODE=2; FUNCTION=LINEAR; R1=0.0; R2=0.1;
MA      TEXT='OrigScore';
MA   /CUT_OFF: LEVEL=0; SCORE=90; N_SCORE=7.0; MODE=1;
MA   /DEFAULT: MI=-26; I=-3; IM=0; MD=-26; D=-3; DM=0;
MA   /M: SY='F';M=-2,-3,-3,-4,2,-3,-2,1,-2,0,-1,-2,-3,-3,-4,-2,-1,0,-5,2;
MA   /M: SY='I';M=-1,-5,-2,-3,-2,-3,0,1,1,-1,1,-1,-2,-1,1,-1,0,1,-4,-4;
MA   /M: SY='A';M=2,-3,1,0,-5,2,-2,-1,-1,-3,-2,1,1,0,-2,2,2,0,-8,-5;
MA   /M: SY='L';M=-3,-8,-5,-4,2,-6,-2,2,-4,6,4,-3,-3,-2,-3,-3,-2,1,-3,0;
MA   /M: SY='Y';M=-4,-2,-6,-6,9,-7,0,-1,-5,-1,-3,-3,-6,-5,-6,-4,-4,-4,-1,11;
MA   /M: SY='D';M=1,-6,3,3,-7,0,0,-2,-1,-4,-3,2,0,1,-2,0,0,-2,-9,-6;
MA   /M: SY='Y';M=-5,-3,-6,-6,10,-7,-1,-1,-2,-1,-2,-3,-6,-5,-5,-4,-4,-4,-1,11;
MA   /M: SY='K';M=-1,-6,1,1,-4,-2,0,-2,2,-3,-1,1,-1,1,1,0,0,-3,-7,-6;
MA   /M: SY='A';M=1,-4,1,0,-5,1,-1,-1,0,-3,-1,1,0,0,0,1,1,-1,-7,-6;
MA   /M: SY='R';M=0,-5,0,0,-5,-1,0,-1,1,-3,-1,1,0,1,1,0,0,-2,-5,-5;
MA      /I: MI=0; I=-2; MD=0; /M: SY='X'; M=0; D=-2;
MA   /M: SY='R';M=0,-5,1,1,-6,0,1,-2,1,-4,-2,1,0,1,2,1,0,-2,-5,-5;
MA   /M: SY='F';M=-3,-7,-6,-6,6,-5,-3,3,-2,5,3,-4,-5,-4,-5,-4,-3,1,-3,3;
MA   /M: SY='Q';M=-1,-6,0,0,-3,-2,1,-1,1,-2,0,0,-1,1,1,-1,0,-1,-6,-4;
MA   /M: SY='K';M=-1,-8,0,1,-3,-2,0,-2,3,-3,0,1,0,2,2,0,0,-3,-6,-6;
MA   /M: SY='G';M=2,-5,1,0,-7,7,-3,-4,-2,-6,-4,1,-1,-2,-4,2,0,-2,-10,-8;
MA   /M: SY='D';M=1,-7,5,4,-8,1,1,-3,0,-5,-3,2,-1,2,-2,0,0,-4,-10,-6;
MA   /M: SY='I';M=0,-5,-1,-2,-2,-2,-1,2,0,0,1,-1,-2,0,0,-1,0,1,-6,-5;
MA   /M: SY='L';M=-2,-6,-5,-5,3,-5,-3,4,-3,6,4,-4,-4,-3,-4,-3,-2,3,-5,0;
MA   /M: SY='Q';M=-1,-5,-1,-1,-3,-2,0,0,0,-2,-1,0,-1,0,0,-1,0,-1,-6,-3;
MA   /M: SY='V';M=0,-4,-3,-4,-1,-3,-3,5,-3,3,3,-2,-2,-2,-3,-2,0,5,-8,-4;
MA   /M: SY='L';M=-1,-6,-3,-3,-1,-3,-2,2,-3,3,2,-2,-2,-2,-3,-2,-1,2,-5,-3;
MA   /M: SY='D';M=0,-6,3,3,-6,0,1,-3,2,-5,-2,2,-1,2,1,0,0,-4,-7,-5;
MA   /M: SY='K';M=-1,-6,0,0,-2,-1,0,-3,3,-4,-1,1,-1,0,1,0,0,-3,-6,-4;
MA   /M: SY='N';M=1,-4,1,1,-5,0,0,-2,0,-3,-2,1,1,0,-1,1,1,-1,-7,-5;
MA      /I: MI=0; I=-1; MD=0; /M: SY='X'; M=0; D=-1;
MA   /M: SY='G';M=1,-5,0,0,-5,1,-2,-1,-2,-3,-2,0,0,-1,-2,0,0,-1,-8,-6;
MA   /M: SY='G';M=1,-6,3,3,-7,3,0,-4,-1,-5,-4,2,-1,1,-2,1,0,-3,-10,-6;
MA   /M: SY='W';M=-9,-12,-9,-11,1,-11,-4,-8,-5,-3,-6,-6,-8,-7,3,-4,-8,-9,26,0;
MA   /M: SY='W';M=-7,-9,-9,-9,0,-9,-4,-5,-5,-1,-4,-6,-7,-6,2,-3,-6,-6,18,-1;
MA   /M: SY='K';M=-1,-7,0,0,-3,-2,0,-2,2,-3,-1,1,-1,1,2,0,-1,-3,-5,-5;
MA   /M: SY='G';M=2,-3,0,-1,-6,3,-3,-2,-3,-4,-3,0,0,-2,-3,1,0,0,-10,-6;
MA   /M: SY='Q';M=-2,-6,0,0,-3,-3,1,-2,0,-2,-1,0,-2,1,1,-1,-1,-3,-5,-3;
MA   /M: SY='T';M=0,-4,-1,-1,-4,0,-2,0,-1,-2,0,0,-1,-1,-1,0,1,0,-7,-5;
MA   /M: SY='T';M=0,-5,0,0,-3,-1,-1,-1,1,-3,-1,1,-1,0,0,1,1,-1,-6,-4;
MA   /M: SY='G';M=0,-5,0,-1,-5,3,-2,-3,-1,-5,-3,0,-1,-1,-1,1,0,-2,-7,-6;
```

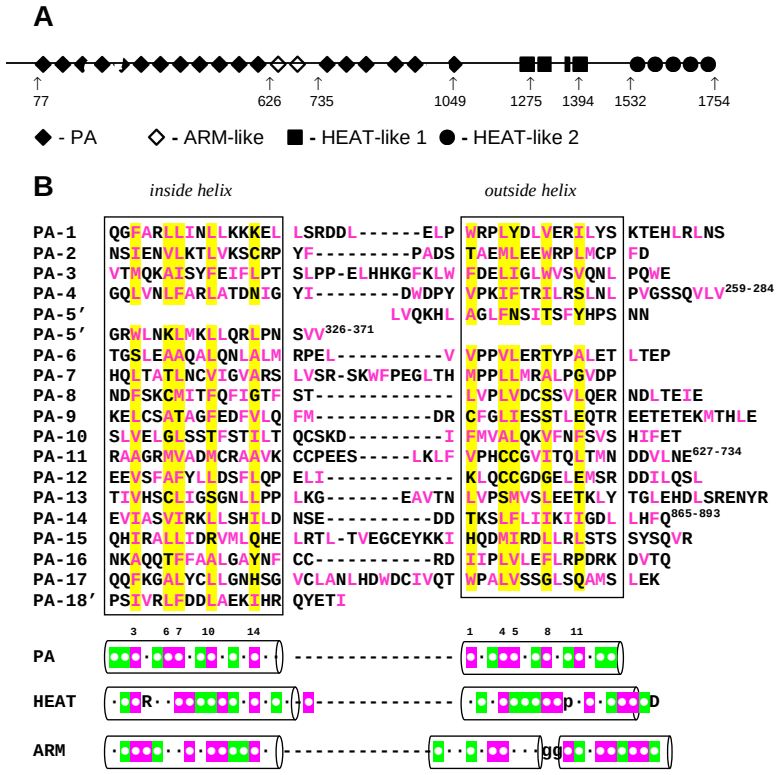
Sequence profile search

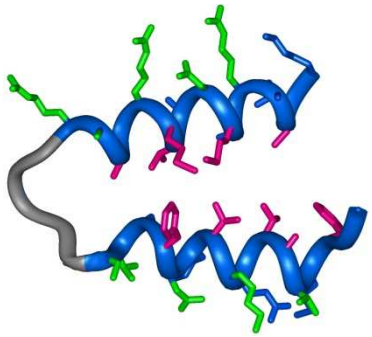
Prosite and Pfam collections of motifs http://hits.isb-sib.ch/cgi-bin/PFSCAN;

CRBM collection of protein repeats: http://bioinfo.montp.cnrs.fr

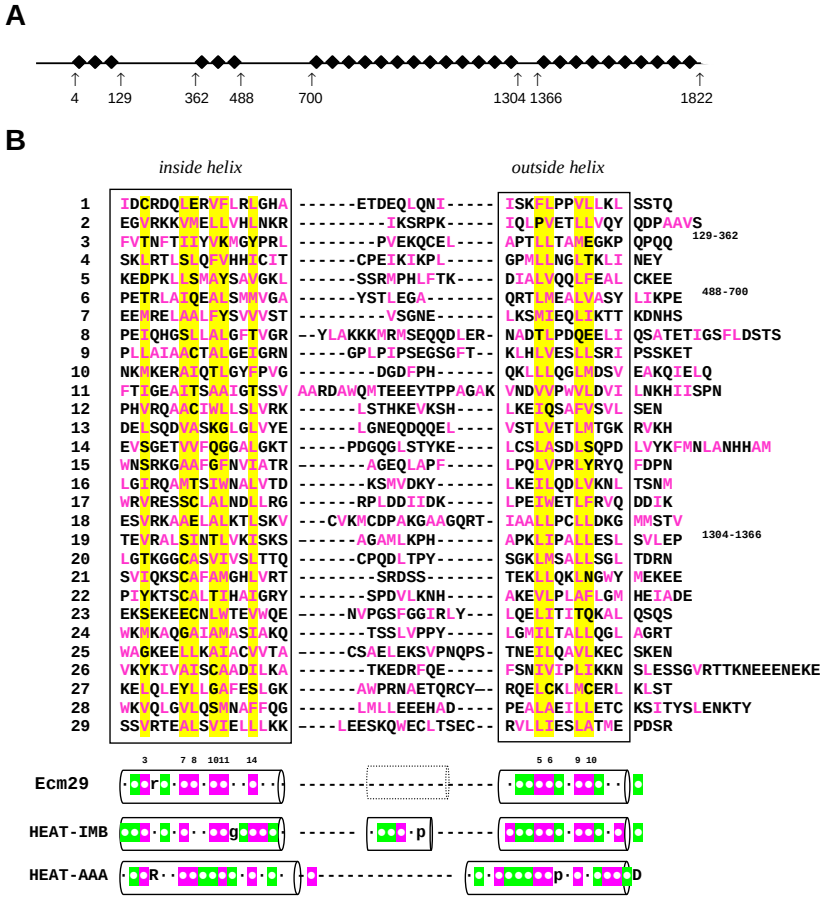
New HEAT-like repeat motifs in proteins regulating proteasome structure and function

Nuclear proteasome activator Blm10 (PA200)





Ecm29



Kajava, A.V., Gorbea, C., Ortega, J., Rechsteiner M. and A. C. Steven (2004) *J. Struct. Biol.* 146,425

: BiSMM - Bioinformatique Structurale et Modélisation Moléculaire - Mozilla Firefox

Edit View Go Bookmarks Tools Help

http://bioinfo.montp.cnrs.fr/index.php?p=pfscan.php&r=profiles Go myhits protein

RBM FRE2593 PubMed bibliovie BiBliovie - l'IST pour le... .:: BiSMM - Bioinformat...

Bioinformatique Structurale et Modélisation Moléculaire
CRBM-CNRS Montpellier

Accueil | Recherche | Publications | Outils | Equipe

Scan protein sequence against CRBM libraries of sequence profiles

Profiles are constructed by using program package *pftools* (Bucher et al. (1996) Comput. Chem. 20, 3-23.) This page uses Pfscan module of *pftools* that allows to find known protein motifs in a given sequence.

Library "**REPEATS**" contains profiles of the known tandem repeats. Each profile spans several (normally 3) repeats and this affords a more sensitive probe of noisy data than a single repeat (e.g. Kajava (1998) J.Mol.Biol. 277, 519-527). Detail information about profiles can be found in their **annotations**.
Library "**PROTEIN DOMAINS**" contains the other sequence motifs.

Paste your protein sequence in FASTA format:

```
>Protein A
KKKKGMNVCSRTYIPPPPPWWAFDKKKKGMNVCSRTYIPPPPPWWAFD
MNVCSRTYIPPPPPWWAFD
```

Select profile categories to search in :
- [x] Repeats
- [x] Protein domains

Or upload a FASTA sequence file : Browse... Submit

admin You're logged as kajava Log out

Pfscan [Help]

Query results :

Matching profile : LRR_SDS22 - SDS22+-like LRR profile.

?	**E-value = 6.19E-01 ; N_score = 7.535 ; Position : 116 - 198**
Profile	NNNRIKKIENLEA---------LPNLEHLBLNNNRIKKIENLEA-----LPNLEHLBLNNNRIKKIENLEA---LPNLEHLBL-
Query	TLKQLSLSENLLHgnipqelglLNRLVYLDLGSNRLNGSIPVQLfcngsSSSLQYIDLSNNSLTGEIPLNYhchLKELRFLLL-
?	**E-value = 8.50E-01 ; N_score = 7.397 ; Position : 305 - 376**
Profile	NNNRIKKIENLEA--LPNLEHLBLNNNRIKKIENLEA--LPNLEHLBLNNNRIKKIENLEA--LPNLEHLBL-
Query	DQNRIHGSIPPEIsnLLNLTLLNLSSNLLSGPIPRELckLSKLERVYLSNNHLTGEIPMELgdIPRLGLLDV-
?	**E-value = 4.32E-01 ; N_score = 7.691 ; Position : 378 - 446**
Profile	NNNRIKKIENLEA-LPNLEHLBLNNNRIKKIENLEA--LPNLEHLBLNNNRIKKIENLEALPNLEHLBL-
Query	RNNLSGSIPDSFGnLSQLRRLLLYGNHLSGTVPQSLgkCINLEILDLSHNNLTGTIPVEVVSNLRNLKL-

Matching profile : LRR_TYPICAL - Typical LRR profile.

?	**E-value = 1.63E-02 ; N_score = 9.114 ; Position : 175 - 247**
Profile	NHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLF-RGLPNLEHLYL-
Query	NNSLTGEIPLNYHCHLKELRFLLLWSNKLTGTVPSSLSNSTNLKWMDLESNMLSGELPSQViSKMPQLQFLYL-
?	**E-value = 3.59E-05 ; N_score = 11.771 ; Position : 353 - 424**
Profile	NHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYL-
Query	SNNHLTGEIPMELGDIPRLGLLDVSRNNLSGSIPDSFGNLSQLRRLLLYGNHLSGTVPQSLGKCINLEILDL-
?	**E-value = 5.16E-02 ; N_score = 8.614 ; Position : 281 - 352**
Profile	NHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYL-
Query	GNSLGGEITSSVRHLSVNLVQIHLDQNRIHGSIPPEISNLLNLTLLNLSSNLLSGPIPRELCKLSKLERVYL-
?	**E-value = 4.41E-04 ; N_score = 10.682 ; Position : 475 - 546**
Profile	NHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYLNHNQLTRVPDHLFRGLPNLEHLYL-
Query	SSNELSGKIPPQLGSCIALEHLNLSRNGFSSTLPSSLGQLPYLKELDVSFNRLTGAIPPSFQQSSTLKHLNF-

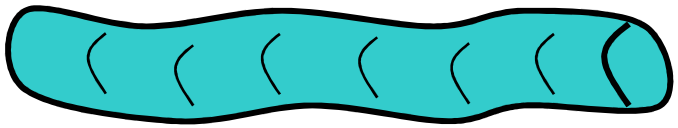
Matching profile : LRR_PS - Plant specific LRR profile.

!	**E-value = 1.53E-11 ; N_score = 18.143 ; Position : 98 - 173**
Profile	SNNRLSGEIPESLGSLKN-LQRLDLSNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKN---LQRLDL-
Query	SRNFFVGKIPPEIGSLHEtLKQLSLSENLLHGNIPQELGLLNRLVYLDLGSNRLNGSIPVQLFCNGSsssLQYIDL-
!	**E-value = 5.24E-10 ; N_score = 16.607 ; Position : 174 - 247**
Profile	SNNRLSGEIPESLGSLKN-LQRLDLSNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKN-LQRLDL-
Query	SNNSLTGEIPLNYHCHLKeLRFLLLWSNKLTGTVPSSLSNSTNLKWMDLESNMLSGELPSQVISKMPqLQFLYL-
!	**E-value = 1.02E-13 ; N_score = 20.317 ; Position : 353 - 424**
Profile	SNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKNLQRLDL-
Query	SNNHLTGEIPMELGDIPRLGLLDVSRNNLSGSIPDSFGNLSQLRRLLLYGNHLSGTVPQSLGKCINLEILDL-
!	**E-value = 2.60E-10 ; N_score = 16.911 ; Position : 280 - 352**
Profile	SNNRLSGEIPESLGSLKN-LQRLDLSNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKNLQRLDL-
Query	AGNSLGGEITSSVRHLSVnLVQIHLDQNRIHGSIPPEISNLLNLTLLNLSSNLLSGPIPRELCKLSKLERVYL-
!	**E-value = 3.27E-12 ; N_score = 18.812 ; Position : 475 - 546**
Profile	SNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKNLQRLDLSNNRLSGEIPESLGSLKNLQRLDL-
Query	SSNELSGKIPPQLGSCIALEHLNLSRNGFSSTLPSSLGQLPYLKELDVSFNRLTGAIPPSFQQSSTLKHLNF-

Visualisation des résultats du module pfscan

From sequence to 3D structure

IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?



Structured

Unstructured

Protein tandem repeats:

the more perfect the less structured

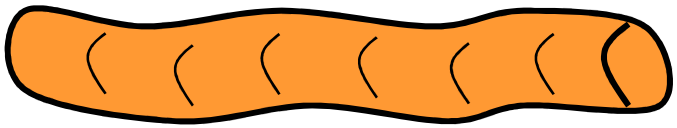
	P_{sim} =0.7-0.8	P_{sim} =0.8-0.9	P_{sim} =0.9-1
ID Ratio – VSL2	80.4%	88.6%	88.9%
ID Ratio – IUpred	56.0%	62.7%	67.2%
ID Ratio – FoldIndex	62.4%	68.6%	70.3%
ID Ratio – TopIDP	85.6%	88.8%	91.1%

Jorda et al. FEBS Journal 2010.

IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?

Polypeptide
with tandem repeats

3D structure ?

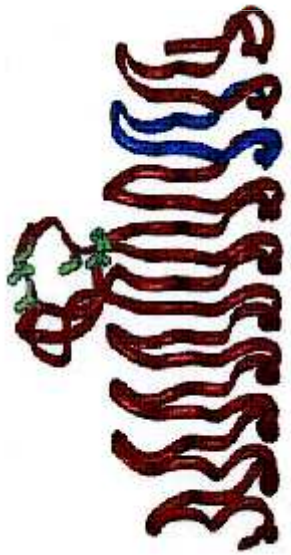


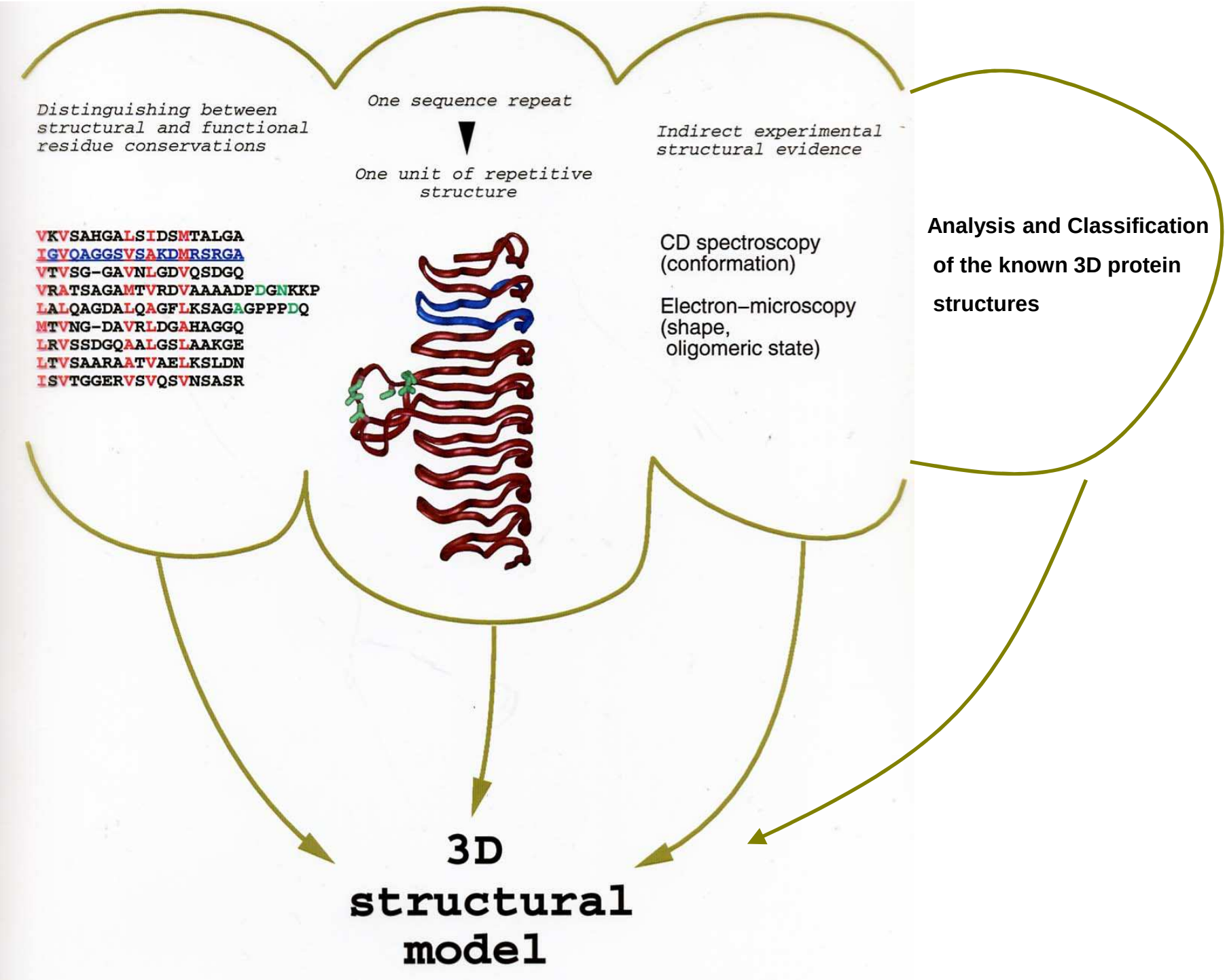
Unstructured

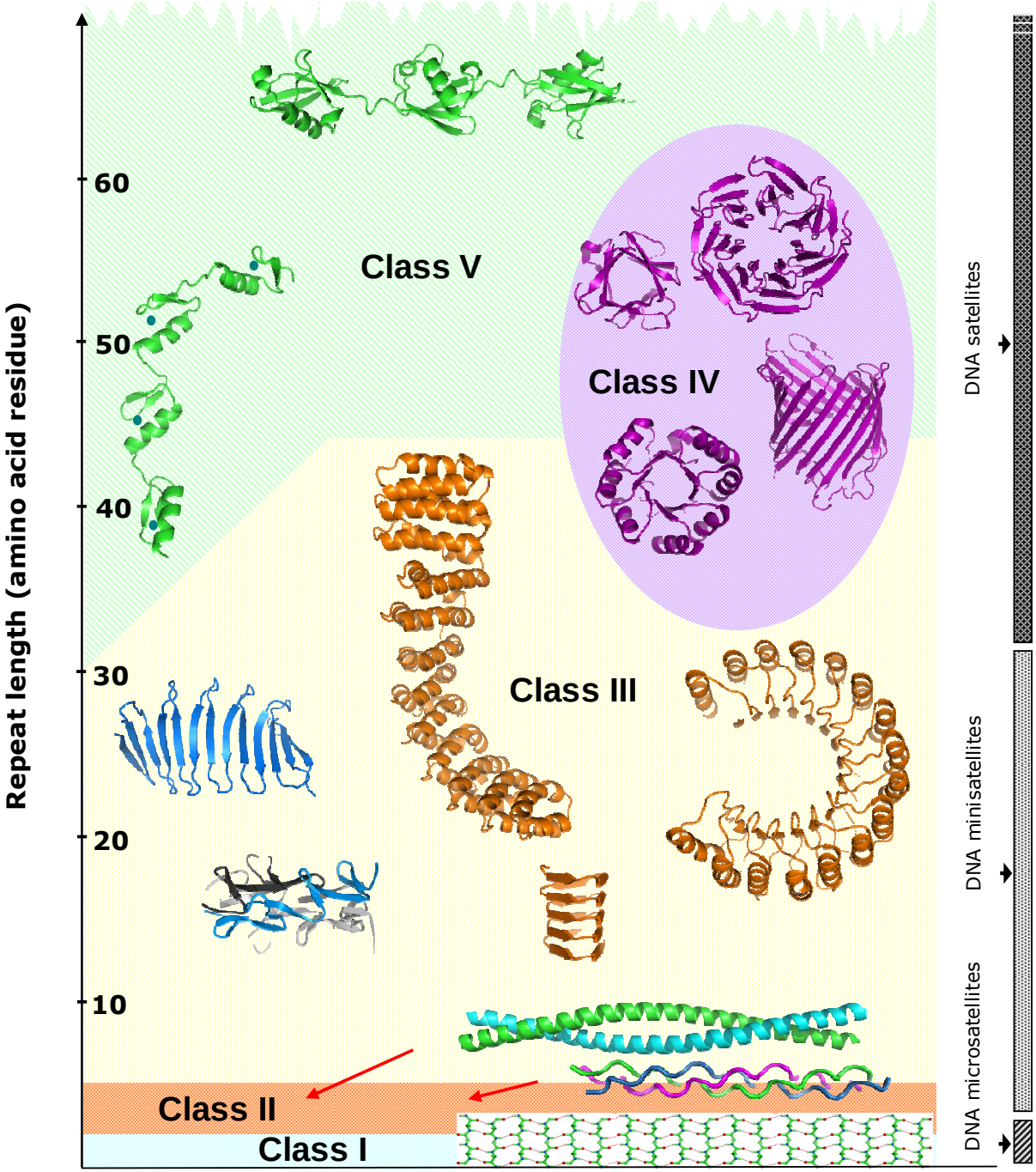
Distinguishing between structural and functional residue conservations

```
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



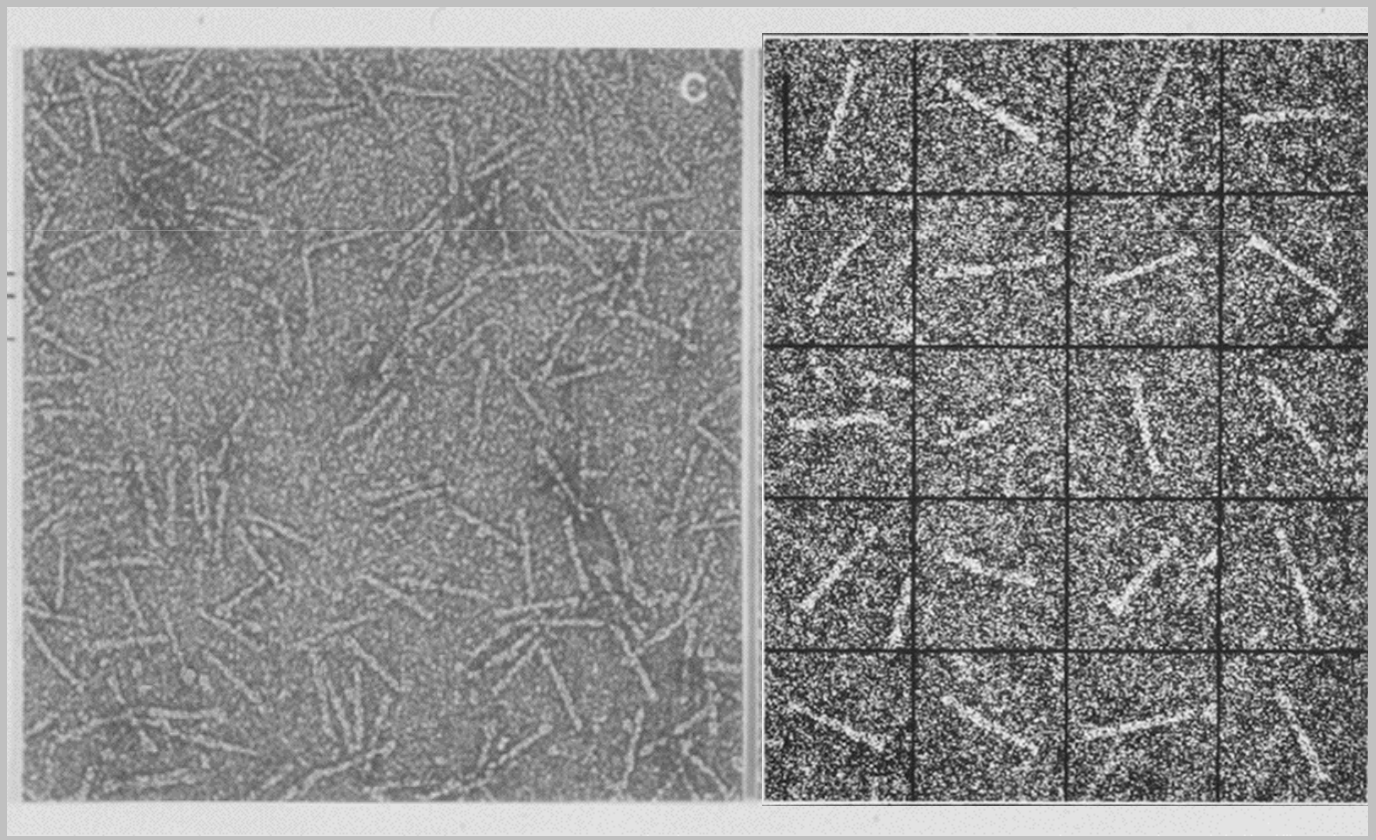
Analysis and Classification of the known 3D protein structures

Kajava, (2011) *J. Struct. Biol.* (in press)

Filamentous Haemagglutinin adhesin

major virulance factor of *Bordatella pertussis*,

etiological agent of whooping cough



EM
negatively stained

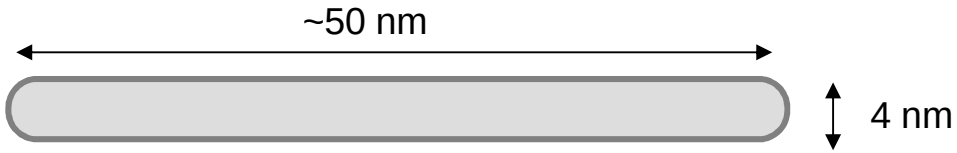
↓

Rod-like shape
50 x 4 nm

Makhov, Hannah, Brennan, Trus, Kocsis, Conway, Wingfield, Simon, Steven *J.Mol.Biol.* (1994) 241, 110

Filamentous Haemagglutinin adhesin (FHA) of *Bordetella pertussis*

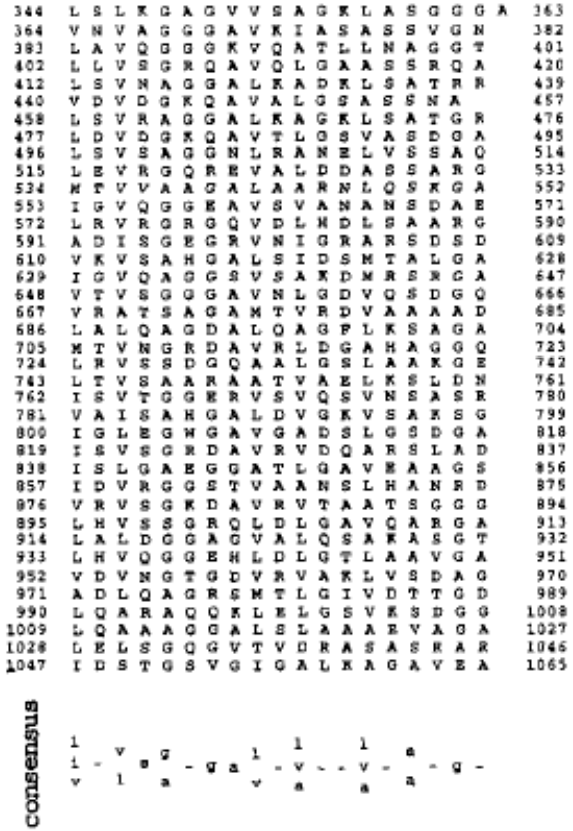
2165 residues sequence



- Rod-like shape according to EM

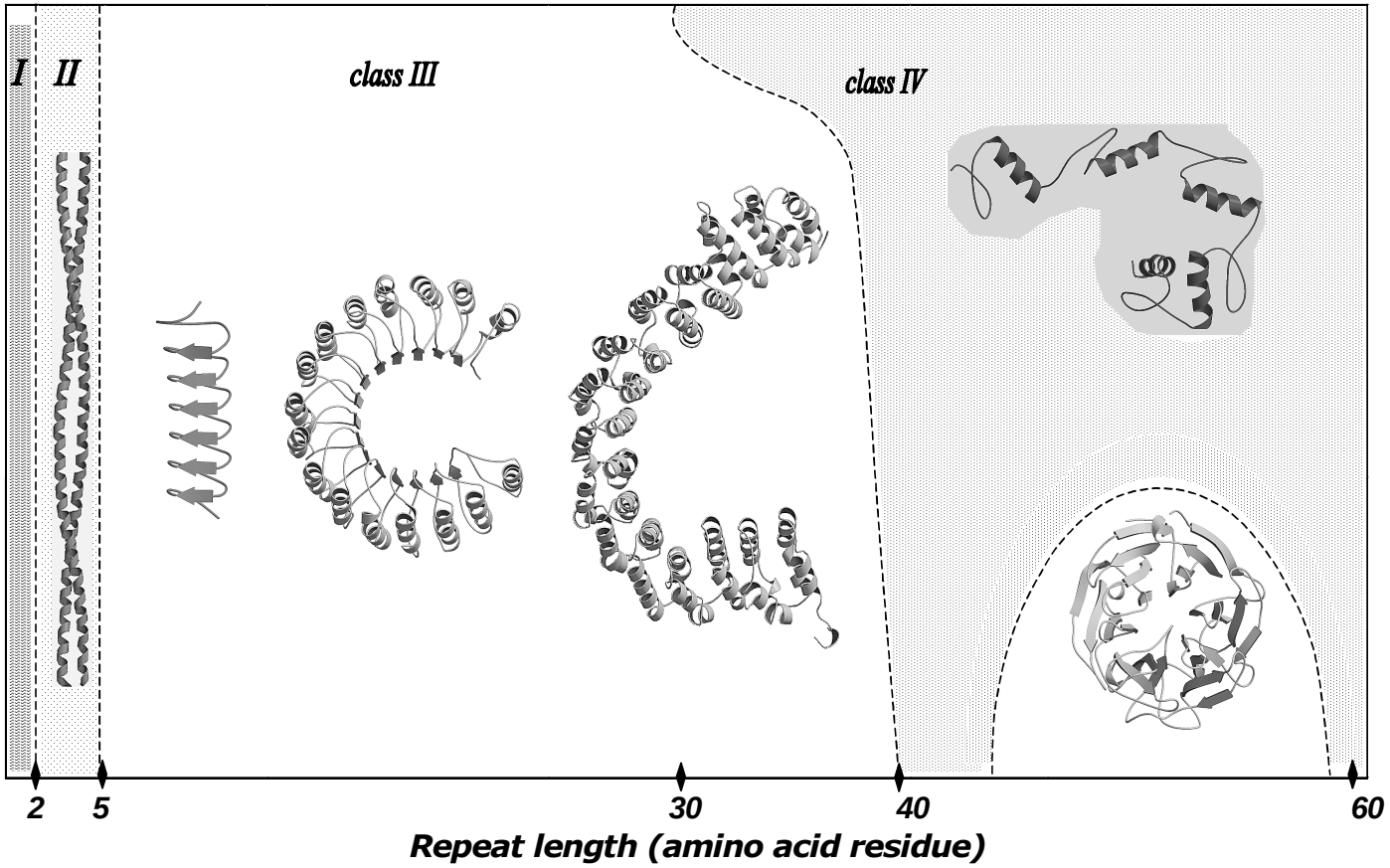
- β-structural protein
  according to circular dichroism spectroscopy mesurements

19-residue repeats



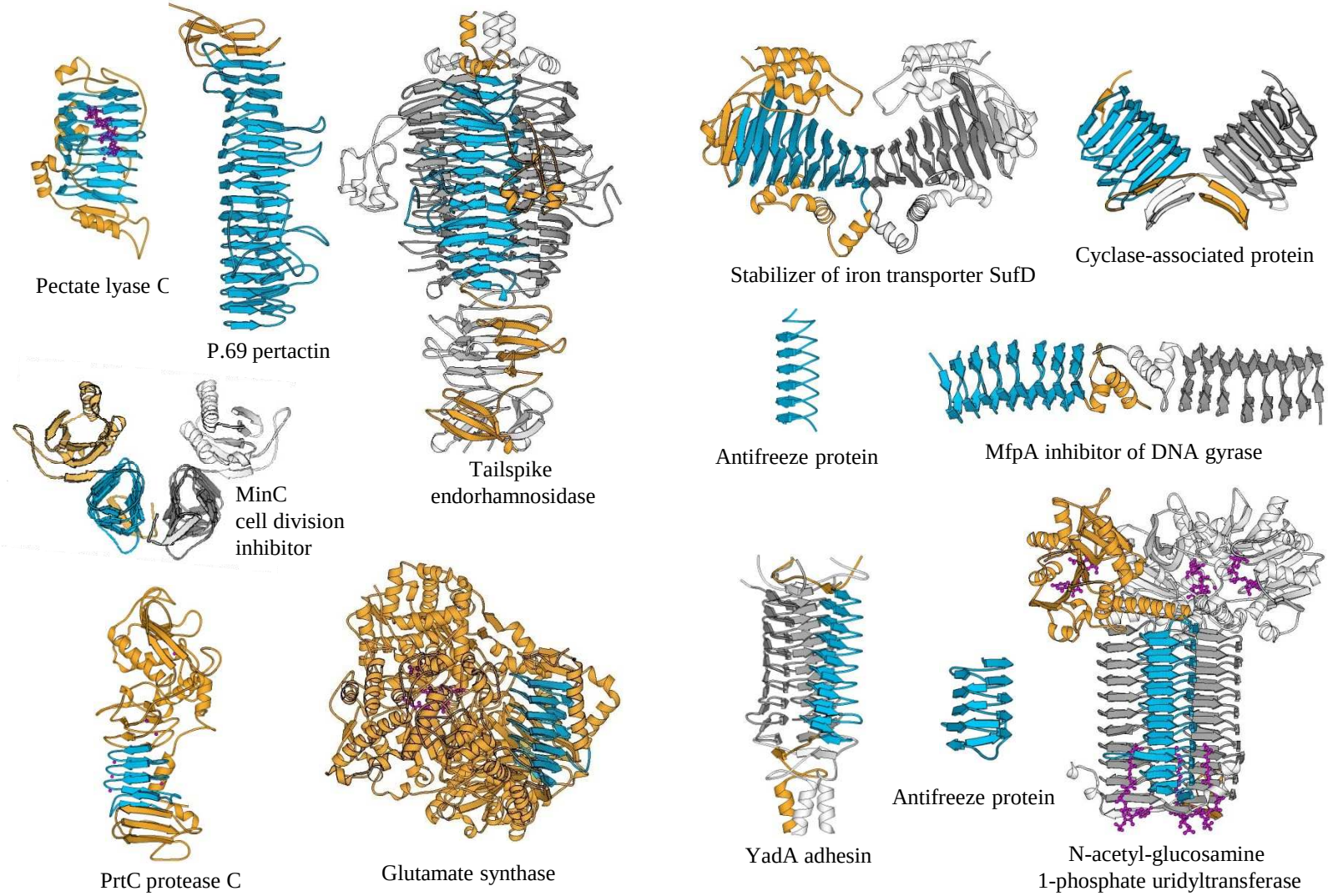
Makhov, Hannah, Brennan, Trus, Kocsis, Conway, Wingfield, Simon, Steven *J.Mol.Biol.* (1994) 241, 110

WHAT CAN REPEAT LENGTH TELL US ABOUT ITS STRUCTURE?



A.V. Kajava (2001) *J. Struct. Biol.* **134**:132

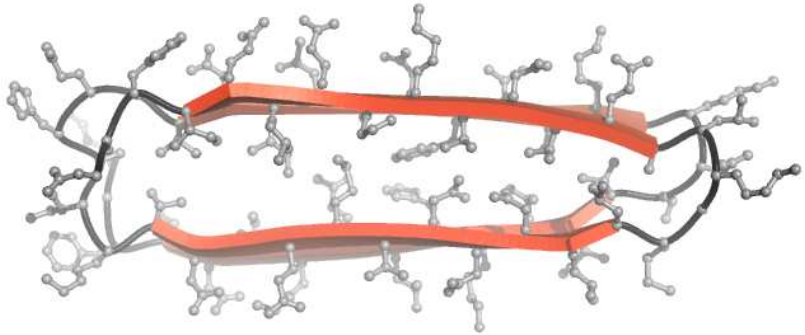
The known structures of β-solenoid proteins



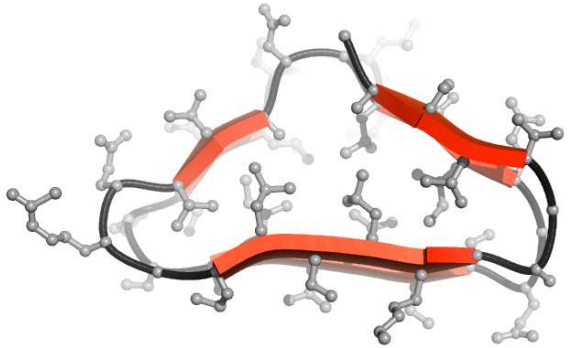
Kajava and Steven (2006) *Advances in Protein Chemistry*, 73, 55-96

Classification of beta-solenoids

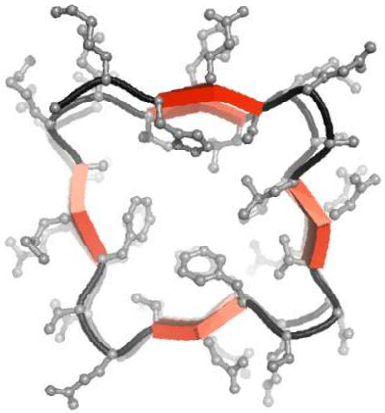
Cross-sectional shapes



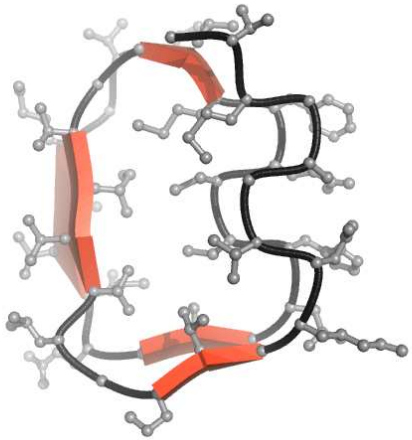
O-type



T-type



R-type



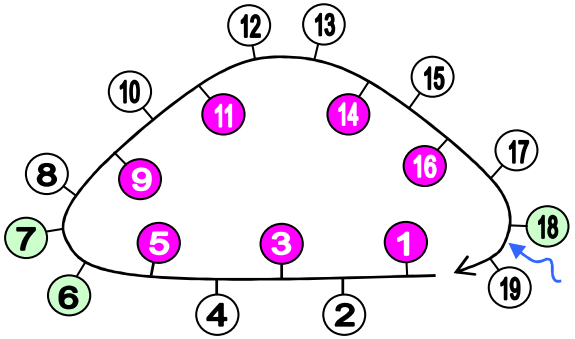
B-type

L-type

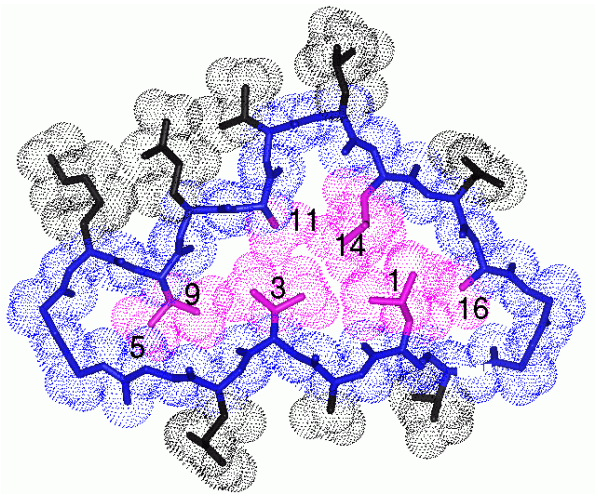
Kajava and Steven (2006) *Advances in Protein Chemistry*, 73, 55-96

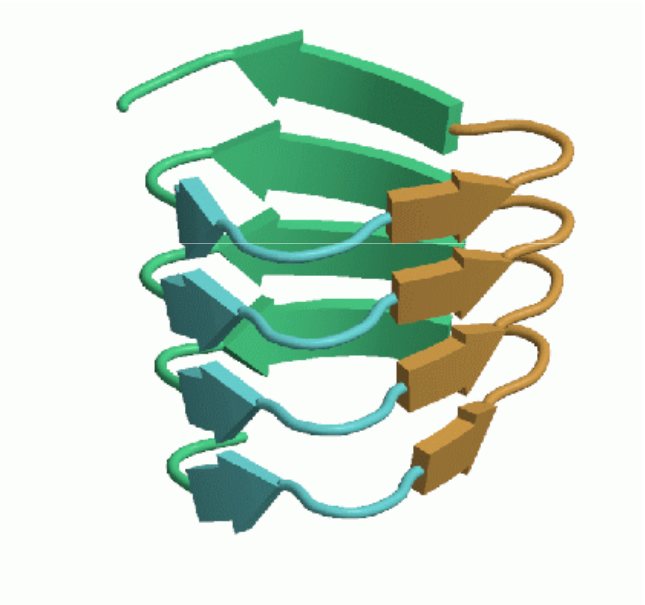
```
Repeat 1     V N V A G G G A V K I A S A S S V G - N
Repeat 2     L A V Q A G G K V Q A T L L N A G G - T
Repeat 3     L L V S A R Q S V Q L G A L S A R Q - A
Repeat 4     L S V N A G G A L K A D K L S A T G S R

consensus    L x V x A G G x V x L x x L x A x G - x

positions    1   3   5   7   9  11  13  15  17    19
```

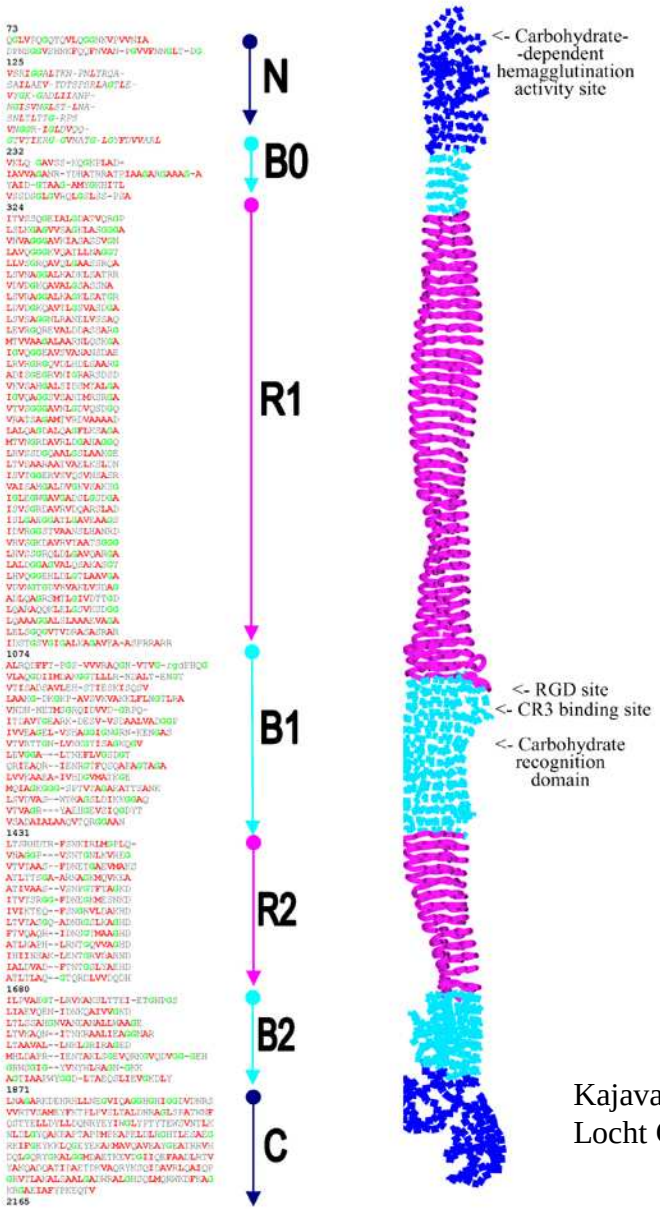
2D plot



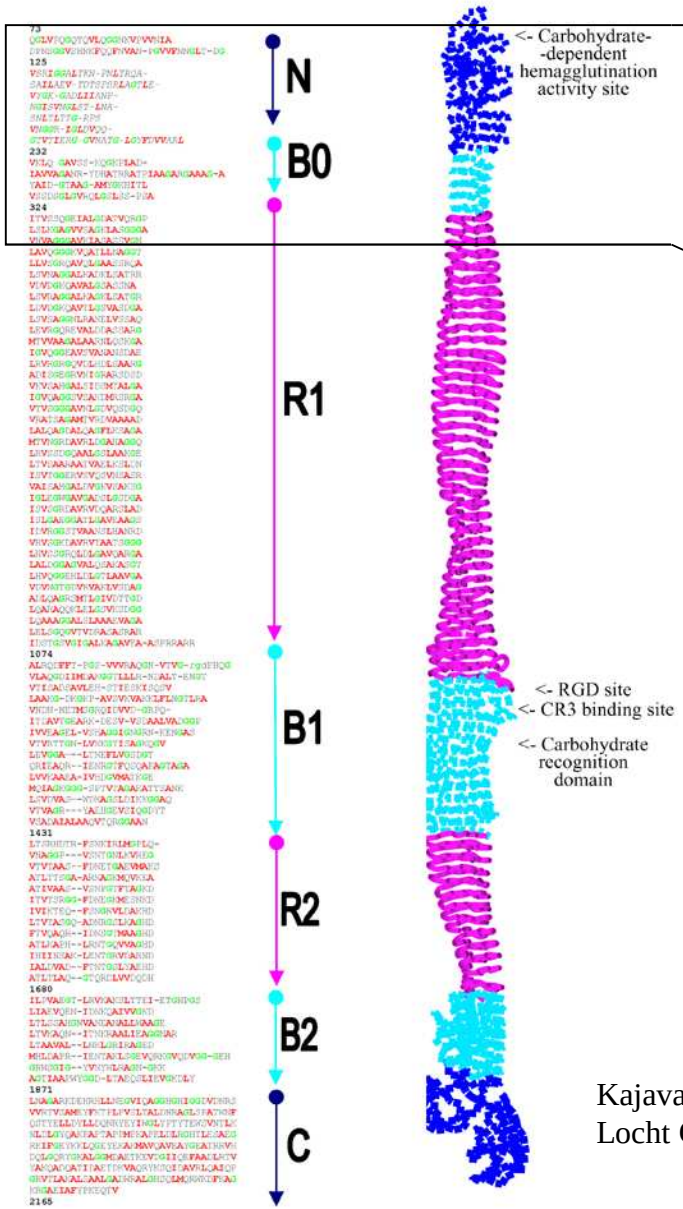
3D structure

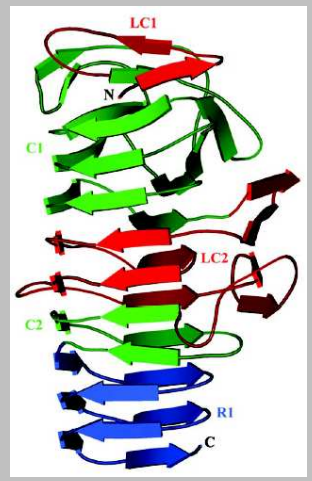






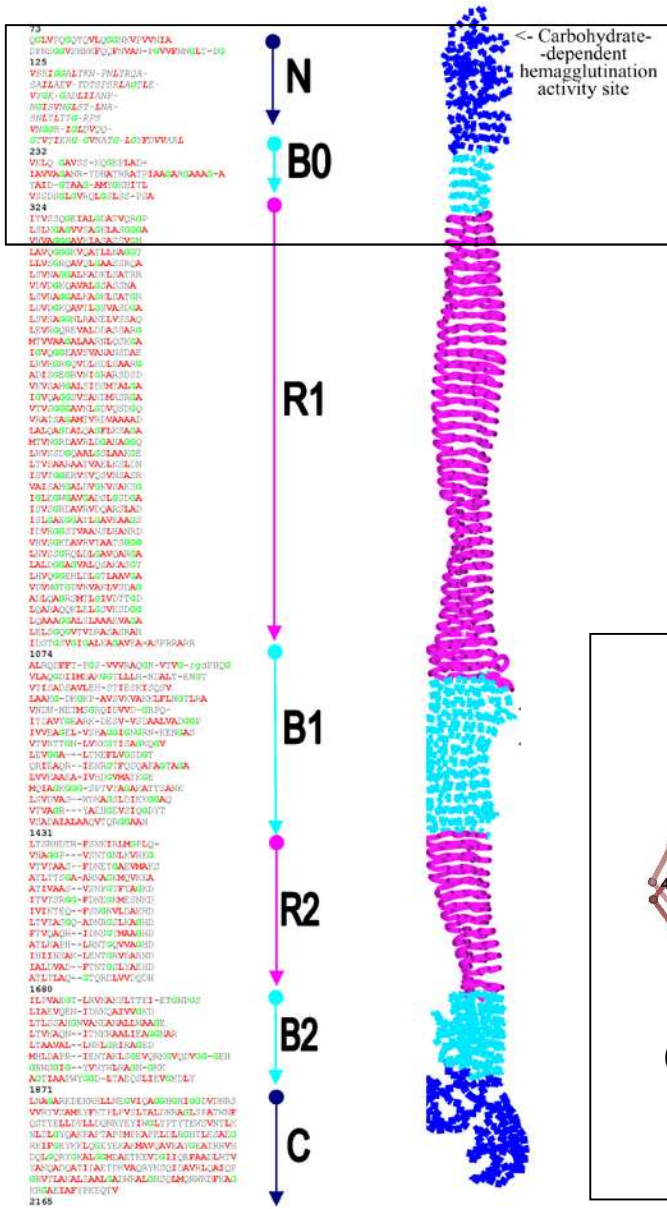
Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Locht C, Steven AC. *Mol Microbiol*. 2001; 42(2):279

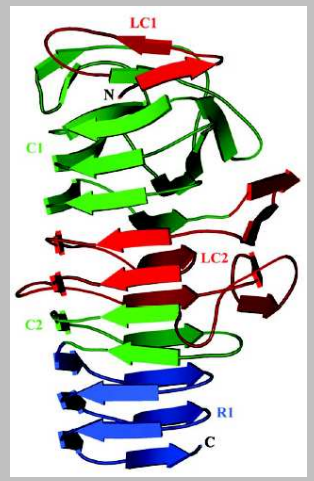


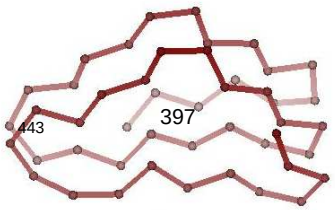


Clantin, Hodak, Willery, Locht, Jocob-Dubuisson and Villeret *PNAS* 2004; 101: 6194

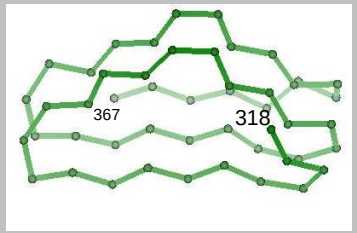
Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Locht C, Steven AC. *Mol Microbiol*. 2001; 42(2):279







Model
(Kajava et al., 2001)

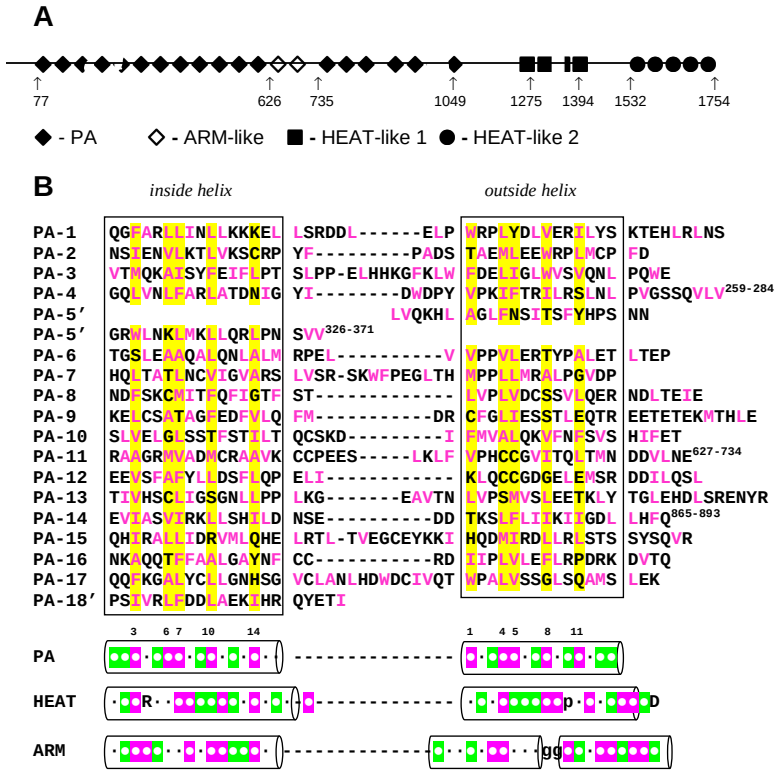


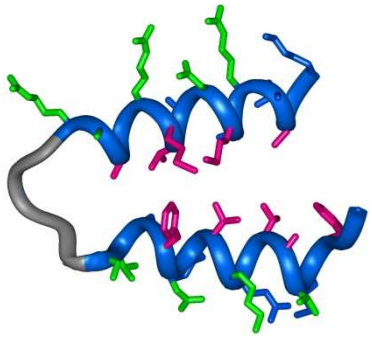
Crystal structure
(Clintin et al., 2004

RMS deviation of C_α atoms is 1.1 Å

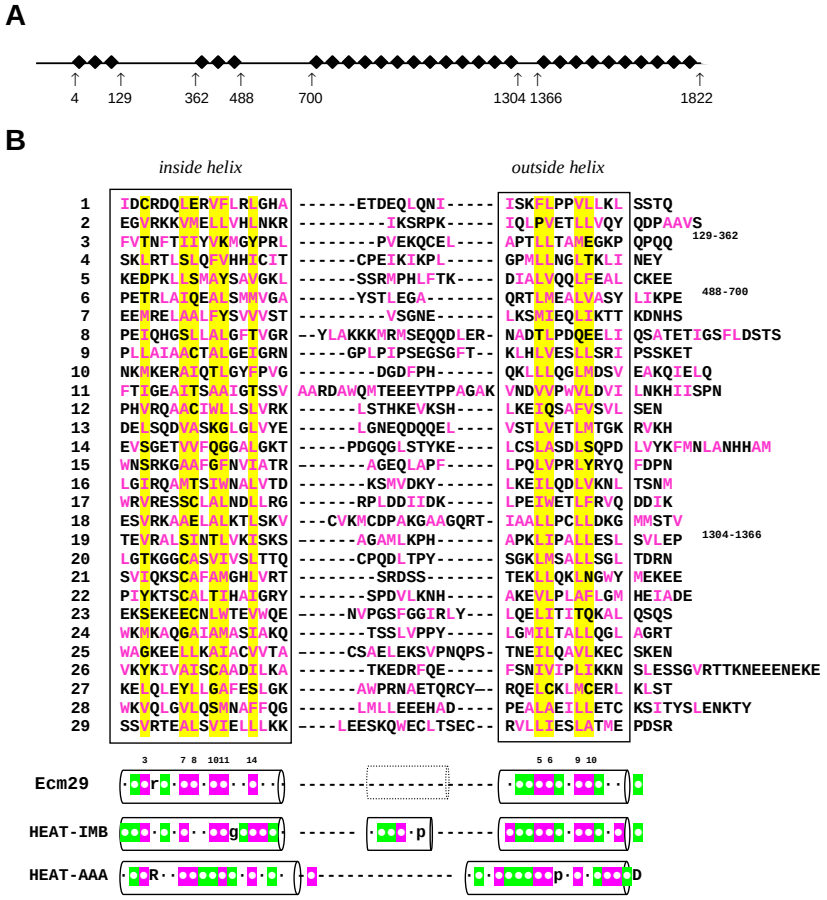
New HEAT-like repeat motifs in proteins regulating proteasome structure and function

Nuclear proteasome activator Blm10 (PA200)

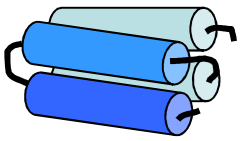




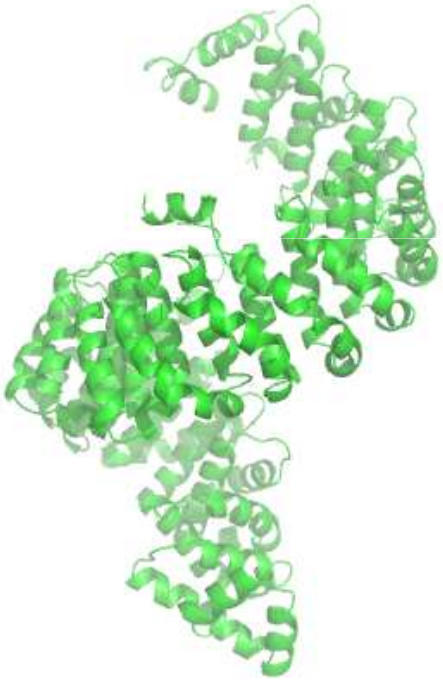
Ecm29



Kajava, A.V., Gorbea, C., Ortega, J., Rechsteiner M. and A. C. Steven (2004) *J. Struct. Biol.* 146,425

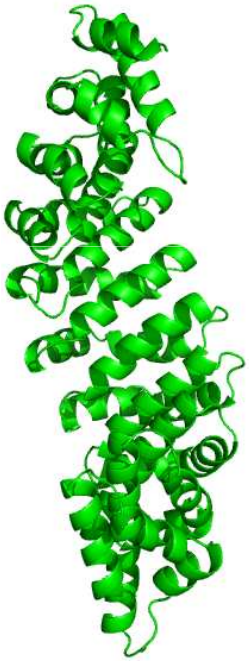


?

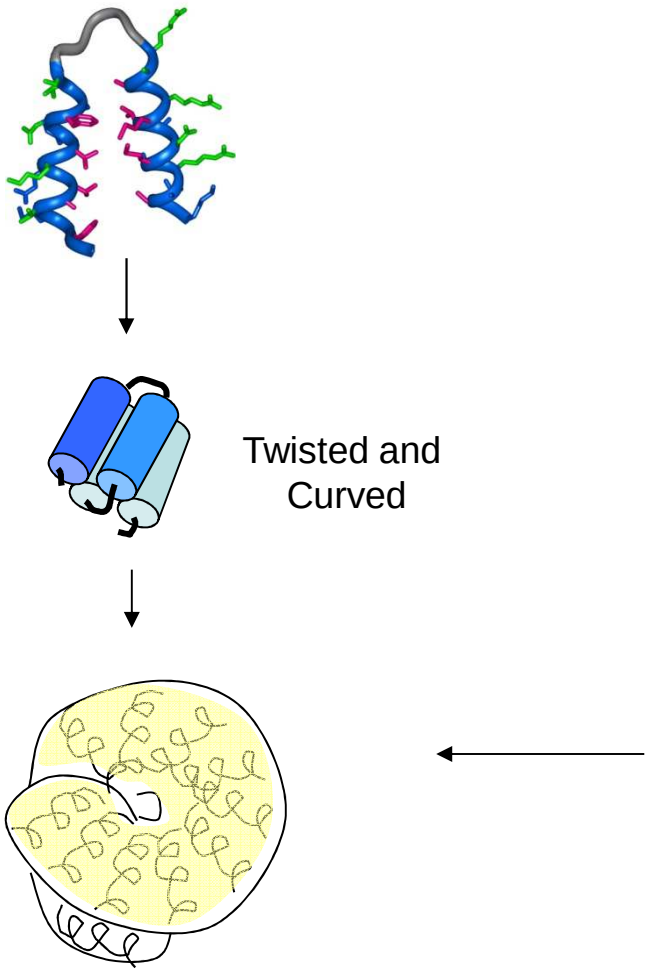


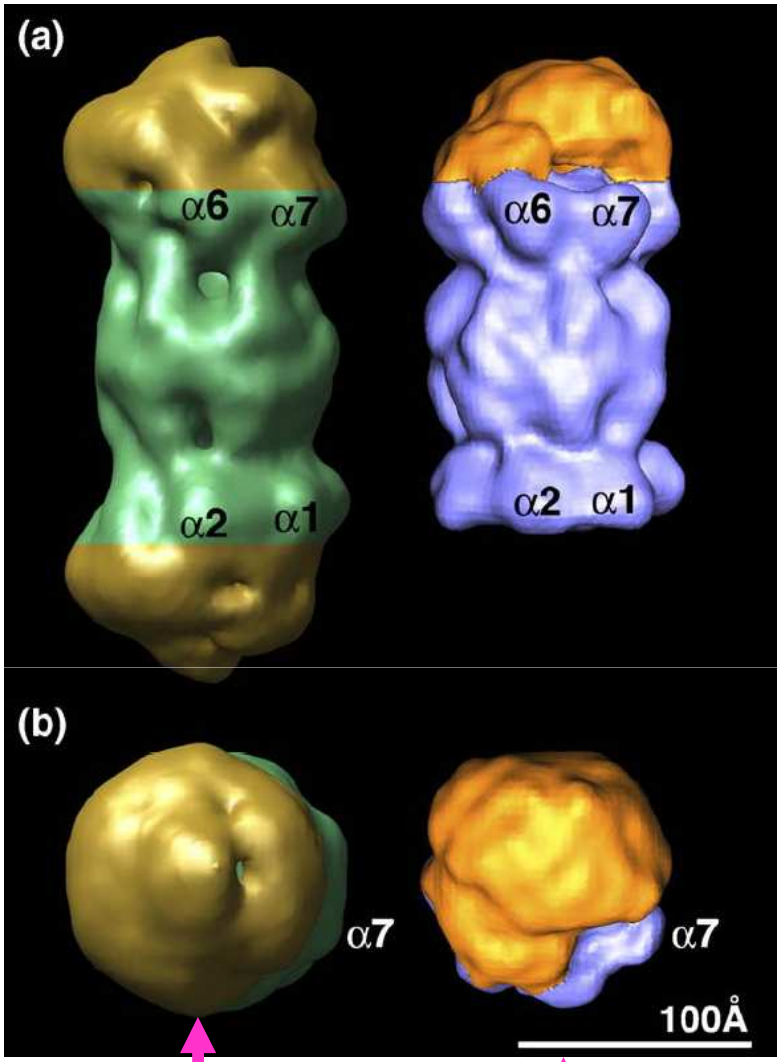
Twisted and curved

Curved



Twisted





Blm10–20 S Proteasome
Iwanczyk, Sadre-Bazzaz, Ferrell, Kondrashkina, Formosa, Hill and Ortega
J. Mol. Biol. (2006) 363, 648–659

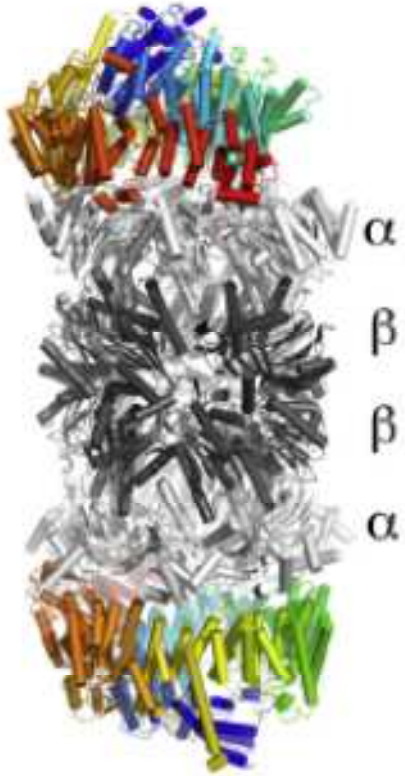
PA200 activator- 20S proteasome
Ortega, Heymann, Kajava, Ustrell, Rechsteiner & Steven (2005). J. Mol. Biol. 346, 1221–1227.

Structure of a Blm10 Complex Reveals Common Mechanisms for Proteasome Binding and Gate Opening

Kianoush Sadre-Bazzaz,[1] Frank G. Whitby,[1] Howard Robinson,[2] Tim Formosa,[1] and Christopher P. Hill[1,*]
[1]Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, UT 84112-5650, USA
[2]Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000, USA
*Correspondence: chris@biochem.utah.edu
DOI 10.1016/j.molcel.2010.02.002

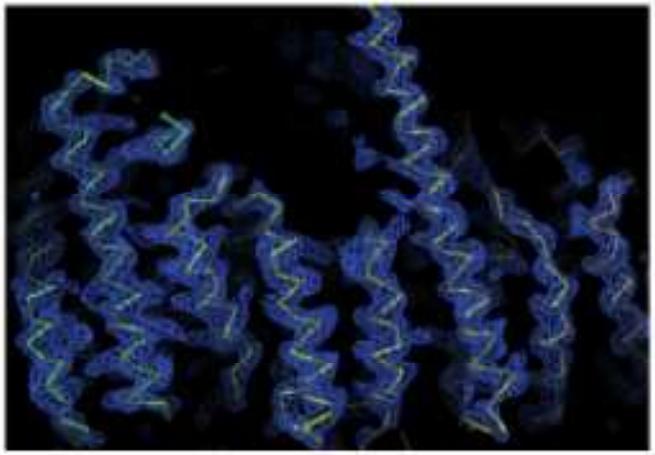
March, 2010

A



B

C



Structure of the Blm10:Proteasome Complex

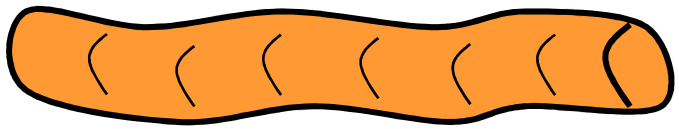
HYBRID APPROACHES TO OBTAIN 3D STRUCTURE

Bioinformatics analysis, structural prediction, molecular modeling

3D structure

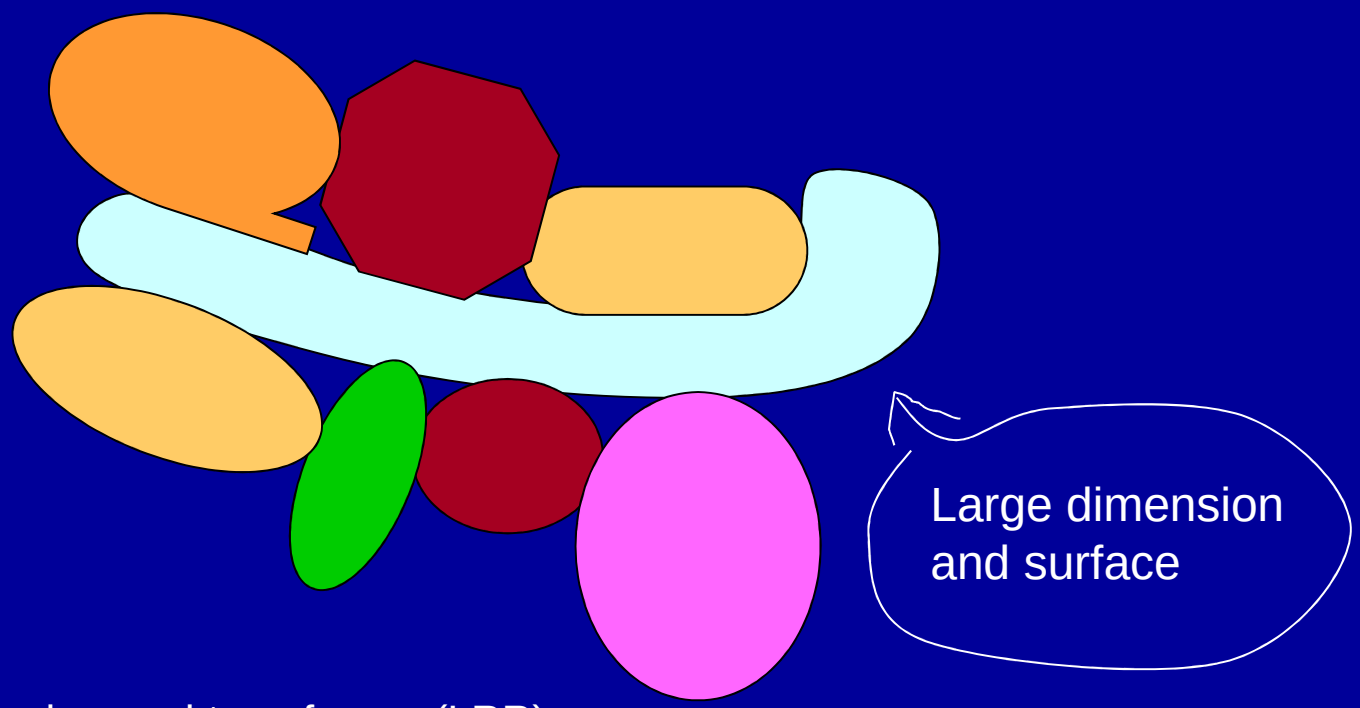
It is possible to get a reliable 3D structural model based on sequence analysis!

Incomplete experimental structural data (EM, CD, etc)



Functions
of proteins with repeats

Protein-protein interactions



Rab geranylgeranyl transferase (LRR)

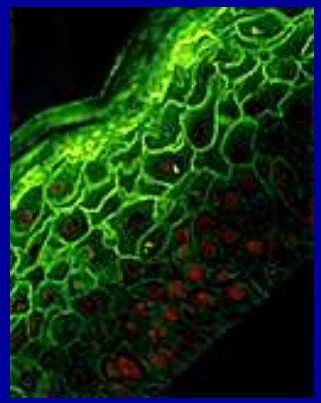
Anaphase-promoting complex (TPR)

Cleavage Stimulatory Factor (HAT repeats)

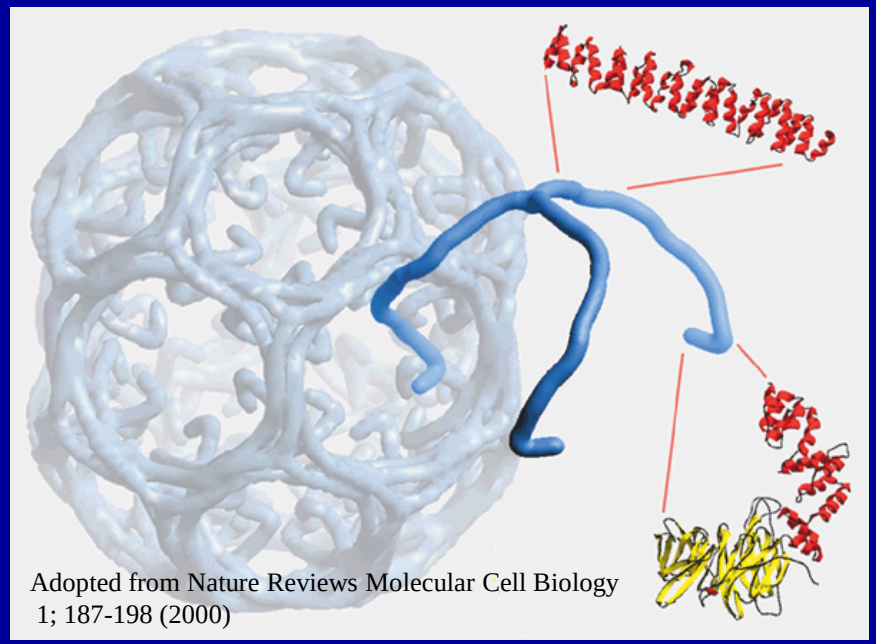
Structural proteins (silk, collagen, etc)

Building blocks for open-work lattices

Involucrin
(cell envelope)

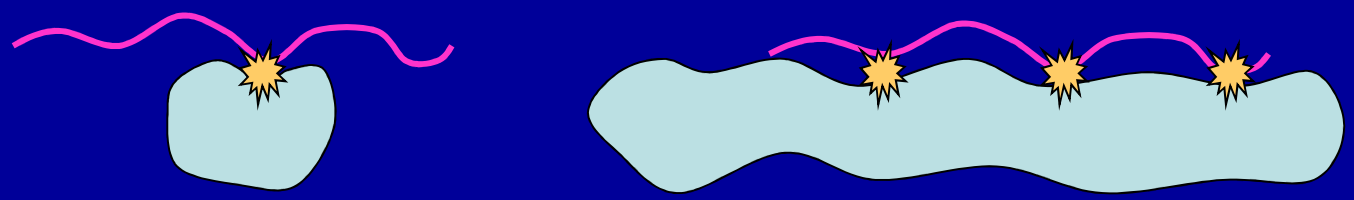


Clathrin

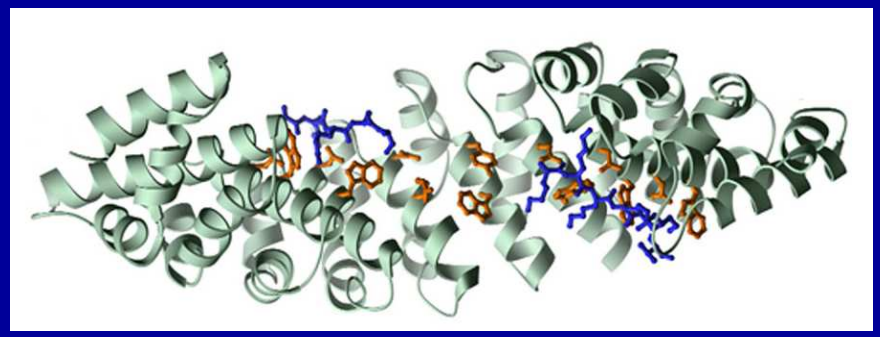


Adopted from Nature Reviews Molecular Cell Biology 1; 187-198 (2000)

Multivalent binding

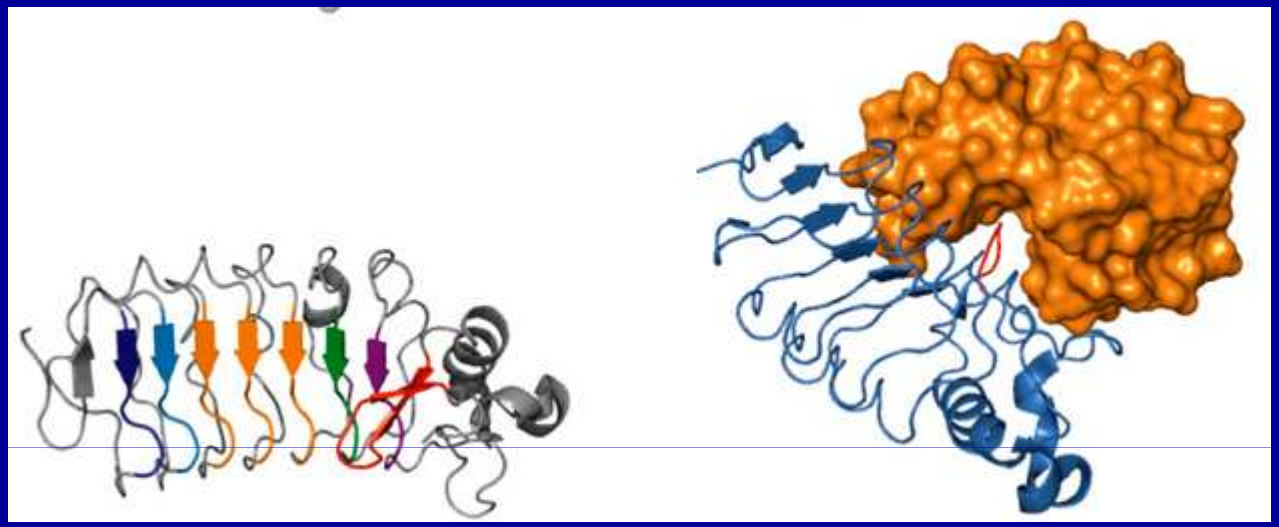


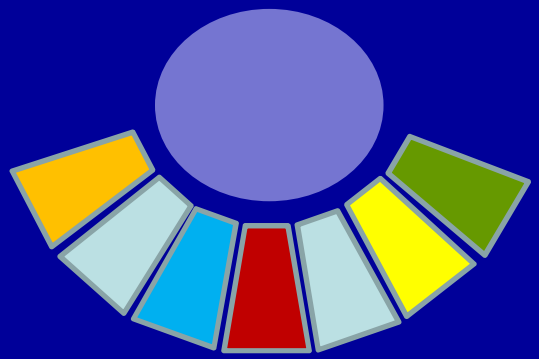
Karyopherin alpha
and nuclear localization signal
(Conti et al., Cell, Vol 94, 193-204)

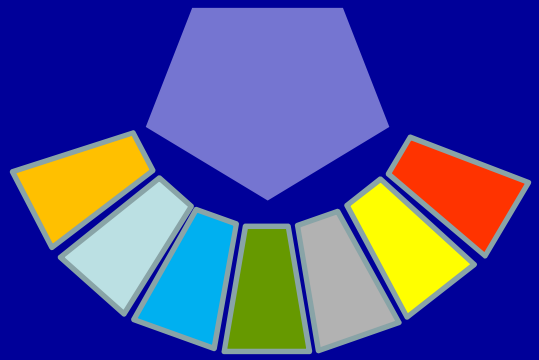


Bacterial outer surface
proteins (pertactin, FHA)

Adaptive immune system: the variable lymphocyte receptors of jawless vertebrates.







COLLABORATORS

Julien JORDA, CRBM, Montpellier
Jerome Hennetin, CRBM, Montpellier
Berangere Jullian CRBM, Montpellier

Bostjan Kobe, University of Queensland Brisbane, Australia

John M. Squire, Imperial College London, UK

David Parry, Massey University, New Zealand

G. Corradin University of Lausanne, Switzerland

M. Anisimova, ETH, Zurich, Switzerland

Alasdair Steven NIAMS, NIH, USA

James Hurley, NIDDK, NIH, USA

Aitor Hierro CIC bioGUNE, Bilbao, Spain