

Structural analysis of proteins with tandem repeats by hybrid approaches

Dr Andrey Kajava

Group of Structural Bioinformatics and Molecular Modeling

Centre de Recherches de Biochimie Macromoléculaire, CNRS

Montpellier, FRANCE

Proteins with tandem repeats

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

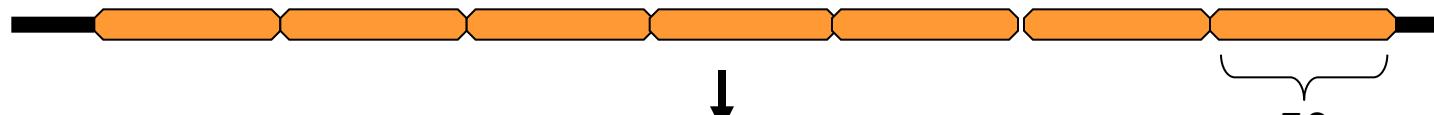
Proteins with tandem repeats

Proteins with internal duplications represent a large portion of genomes

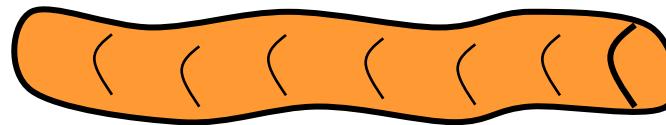
E. coli (7%), *S. cerevisiae* (17%), Human (27%)
All SwissProt (14%)

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

Sequence



Structure

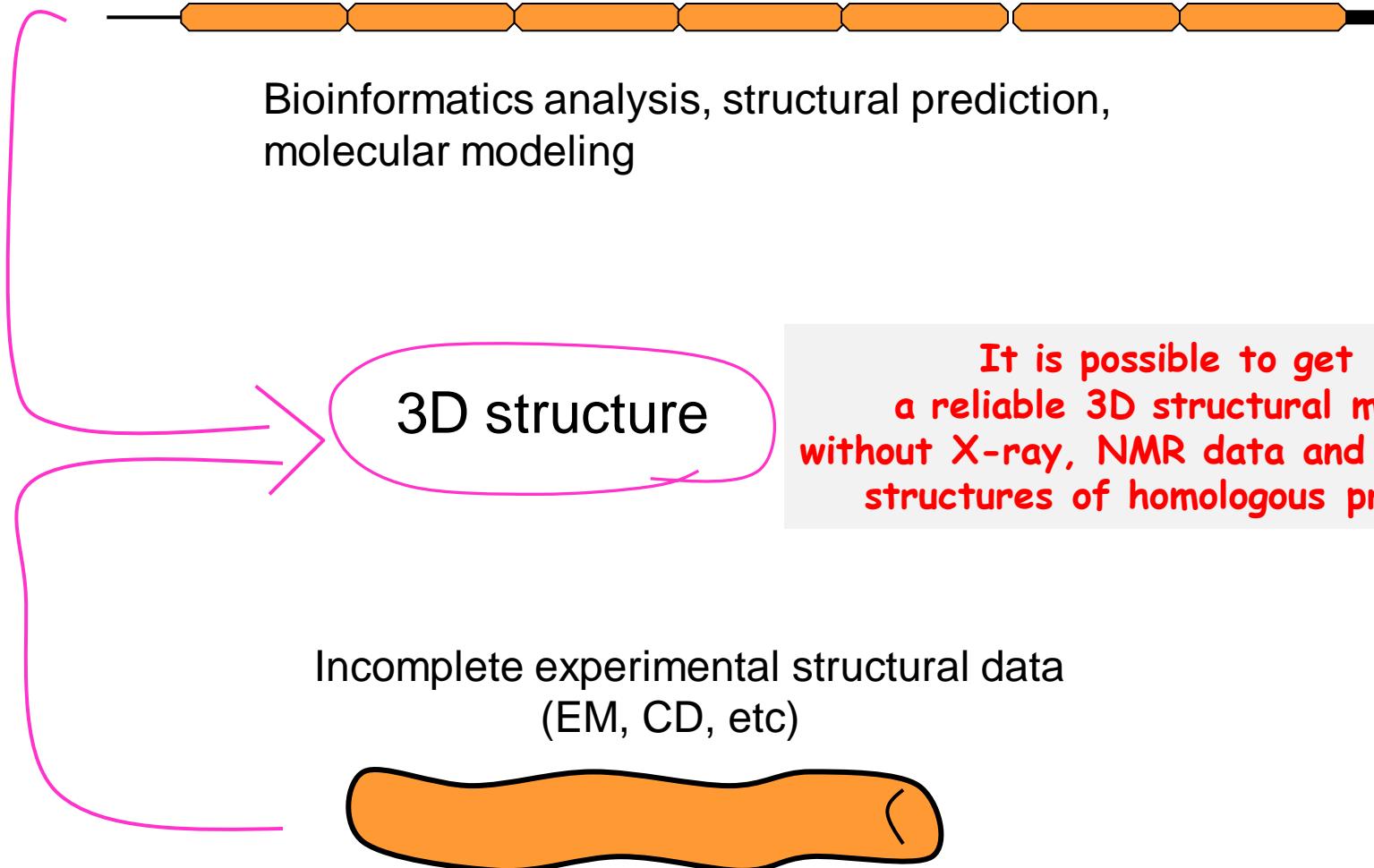


< 50 res

only ~ 2% of known 3D structures

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

HYBRID APPROACHES TO OBTAIN 3D STRUCTURE

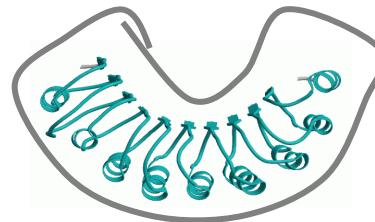


Prédiction et modélisation de protéines à séquences répétitives

Leucine-rich repeat proteins

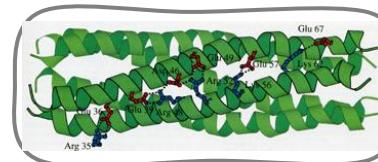
Kajava et al., (1995) Structure, 3, 863

Kajava (1998) J.Mol.Biol. 277, 519



α -Helical Coiled coil pentamer of COMP

Kajava (1996) Proteins, .24, 218



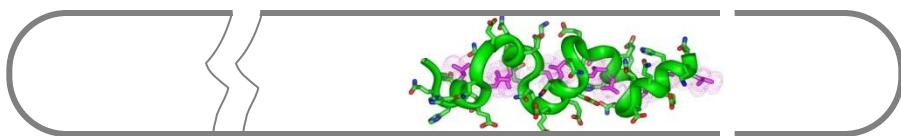
Filamentous Hemagglutinin Adhesin
of Bordetella pertussis (56 nm long)

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

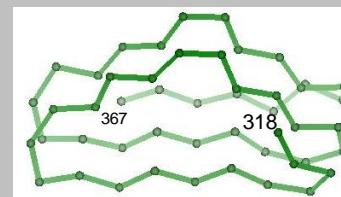




Carbohydrate-dependent hemagglutination activity site



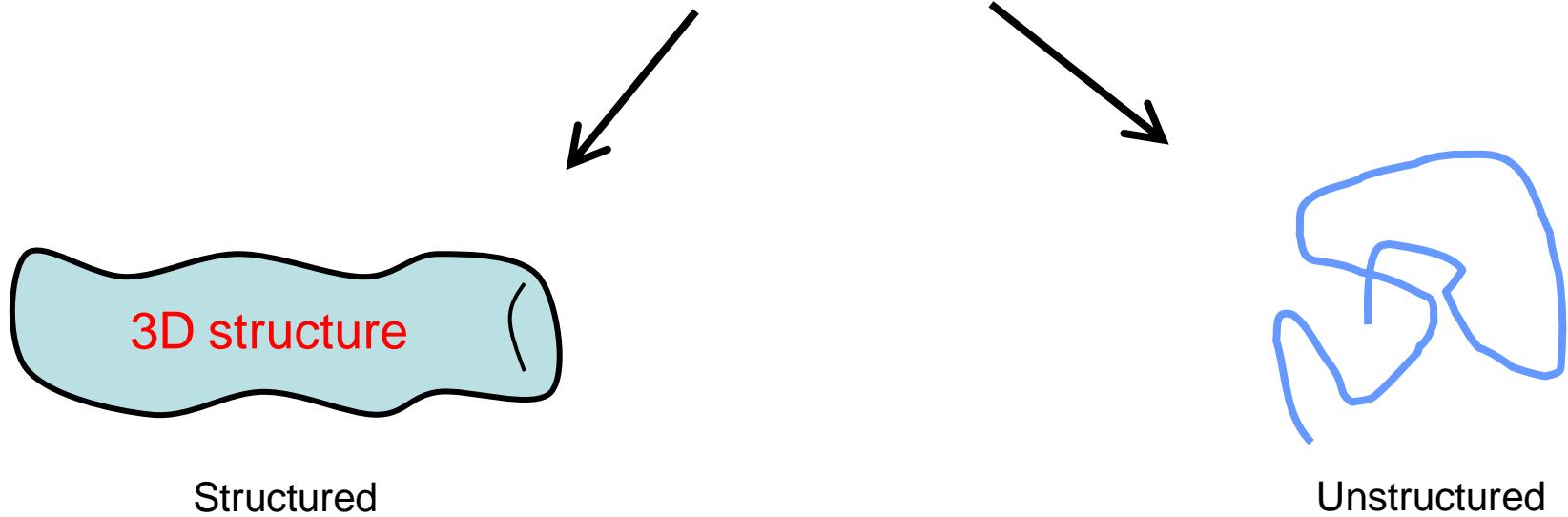
Model (Kajava et al., 2001)



Crystal structure (Clintin et al., 2004)

RMS deviation of C_α atoms is 1.1 Å

IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?



Tandem repeat regions in proteins: the more perfect the less structured

Jorda Xue, Uversky and Kajava (2010) *FEBS J* 277:2673–2682

JOBIM, see Poster 73: Jorda et al.

Proteins with tandem repeats

- ✓ Structural prediction
- ✓ Analysis and Classification of the known 3D protein structures
- ✓ **Identification of protein repeats**

Experimental tests

- ✓ Evolution of proteins with repeats
- ✓ Applications in medicine, material science and nanotechnologies
- ✓

Repeat detection in protein sequences

Self-alignment algorithms

REPRO

George RA. and Heringa J. (2000) *Trends Biochem. Sci.* **25**, 515
<http://mathbio.nimr.mrc.ac.uk/~rgeorge/repro/>

RADAR

Heger A, Holm L. (2000) *Proteins* 2000 Nov 1;41(2):224-237
<http://www.ebi.ac.uk/Radar/>

Internal Repeat Finder

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. (1999) *J Mol Biol* 293, 151
<http://www.doe-mbi.ucla.edu/Services/Repeats/>

Short string extension algorithm

XSTREAM

Newman and Cooper, 2007

Estimation of edit distance between strings

TRED

Sokol et al. 2007

Eichier Édition Affichage Historique Marque-pages Outils ?



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



Google



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...

Sequence type :

 Protein DNA

or Paste your sequences :

Percentage of similarity:

0.70

Filter the overlapping repeats:



Search Repeats

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>

Jorda and Kajava (2009)
Bioinformatics 25:2632.

JOBIM, see Poster 73: Jorda et al.

Bioinformatique Structurale et Modélisation Moléculaire

CRBM-CNRS Montpellier

[Accueil](#)[Recherche](#)[Publications](#)[Outils](#)[Equipe](#)

Recherche

Nous utilisons des méthodes de biologie structurale théorique et de bioinformatique afin de comprendre les structures protéiques et les interactions biomoléculaires. Les connaissances ainsi obtenues sont appliquées à la conception de drogues ainsi qu'au design de *novo* de protéines aux fonctions choisies [[+ d'infos](#)].

Postdoc, thésards et étudiants stagiaires

Les personnes intéressées par nos travaux sont invitées à nous contacter pour d'éventuelles collaborations [[+ d'infos](#)]. Nous accueillons également des étudiants dans le laboratoire.

contact: Andrey.Kajava@crbm.fr

TEL 33 4 67 61 3364

FAX 33 4 67 52 1559

Tools description

Profiles

Selectseq

Show Profiles

PfScan

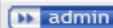
PfSearch

Add/Edit Profiles

Add result from file

Compare results

Generalized sequence profiles
HMMs.



You're logged as kajava

Structural-Functional Annotation of Genomes

Proteins with aperiodic sequences and globular domains

Proteins with tandem repeats



Bioinformatics tools

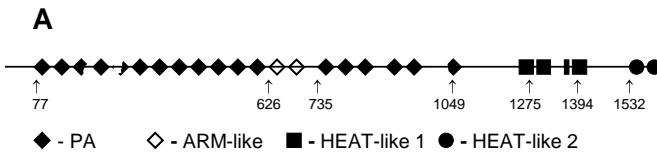


Ab initio structural prediction



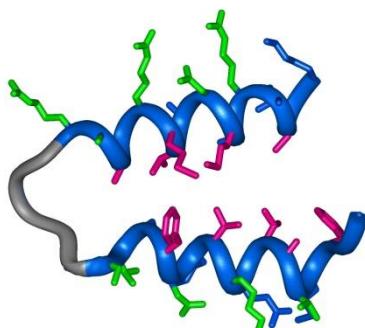
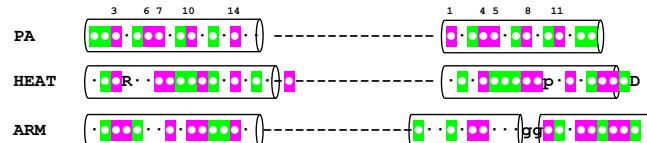
New HEAT-like repeat motifs in proteins regulating proteasome structure and function

Nuclear proteasome activator Blm10 (PA200)



B

	inside helix	outside helix
PA-1	QG ⁷⁷ FARLLINELKKKEL	LSRDDL-----ELP
PA-2	NSIENVLTIVKSCRP	YF-----PADS
PA-3	VTMQRKVISYEIFLPT	SLPP-ELHHKGFKLW
PA-4	GQLVNLFARQATDNING	YI-----DWDPY
PA-5'	GRWLNLKLMKILQRPN	SVV ³²⁶⁻³⁷¹
PA-6	TGSLERAAQAIQNLAIM	RPEL-----V
PA-7	HQLTATLNCIGVRS	LVSR-SKWFPEGLTH
PA-8	NDFSKCMITHPIQFIGTF	ST-----LVP
PA-9	KELCSATAGEDFWLQ	FM-----DR
PA-10	SLVELGLSSSTFSTILT	QCSKD-----I
PA-11	RAAGRMVADCRAAVK	CCPEES-----LKLP
PA-12	EEVSFAFYLIIDSFPQ	ELI-----LQ
PA-13	TIVHSCLIGSGNLLPP	IKG-----EAVTN
PA-14	EVIASVIRKLSHLD	NSE-----DD
PA-15	QHERRAIDIDRVMLQHE	IRTL-TVEGCEYKKK
PA-16	NKAQQTFFAFLGAYNF	CC-----RD
PA-17	QQPKGRLYCHLGNSG	VCLANLHDWDICIVQT
PA-18'	PSIVRLEFDDELAEKHR	QYETI

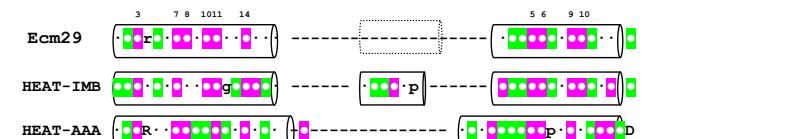


Ecm29

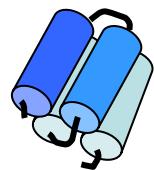
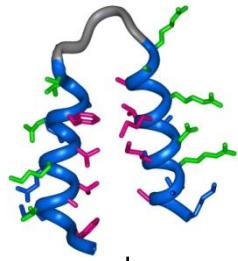


B

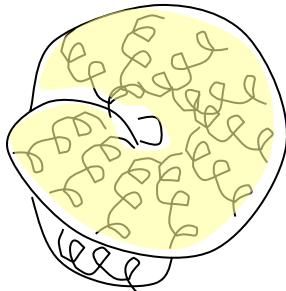
	inside helix	outside helix
1	IDCRDQLERVFIRIGHA	-----ETDEQLQNT-----
2	EGVRKKVNEELVHINKR	-----IKSRPK-----
3	FVTNITIYVVKGYPRL	-----PVEKQCEL-----
4	SKERPLLSQPHHICIT	-----APTLLTAEKGKP-----
5	KEDPKLILSMAYSAYGKL	-----GPMILNGTKL-----
6	PETRLLAIQEAISMMVGA	-----DIALVQQLPEAL-----
7	EEWRELAALFYSVVST	-----QRTLMALVASY-----
8	PEIQHQHSSLAGFTVGR	-----LIKSMTEOIXKT-----
9	PILATAACTAEGIAGR	-----KDNHS-----
10	NKMKERAOTGVYPVG	-----NADTLPDQEELI-----
11	FTIGEAITSAAGTSSV	-----QSATETIGSFLDSTS-----
12	PHVROAACIWLISLVRK	-----PSKSKET-----
13	DELSODVNSKGIGLYVE	-----OKLLIQQGNDSV-----
14	EVSGETVVFQGGALGKT	-----EAKQIELQ-----
15	WNSRKGAAFGIVIAITR	-----VNDVPNVLDVI-----
16	LGIRQAMTISINNAVTD	-----LNKHIIISP-----
17	WRVRESSCLANNDLRG	-----LKEQSABFVSV-----
18	ESVPRKAELAKTLSKV	-----SEN-----
19	TEVRAISINTVZKISKV	-----VSTLGETVNTGK-----
20	IGTKGGCIVIVSLTTQ	-----RVK-----
21	SIVIKSCACAGHVVRT	-----LGE-----
22	PVYKTSCAITIHAGRY	-----LQET-----
23	ERSEKEECNLATEVWQE	-----TSS-----
24	WKQKAQGAIAMPASIAQ	-----LVP-----
25	WAGKEELVKAICAVVTA	-----LPP-----
26	WYKIKIVAVSCAADILKA	-----LQ-----
27	KEIQLEYLIGAFESIGK	-----QK-----
28	WKVQLGVQSNWAFQG	-----LQ-----
29	SSYRTEALSYELLKK	-----Q-----



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

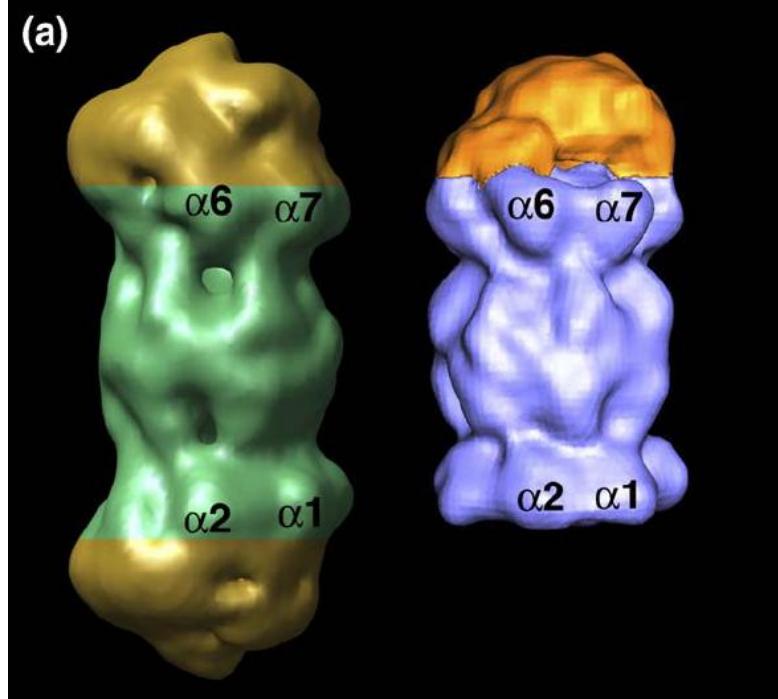


Twisted and
Curved

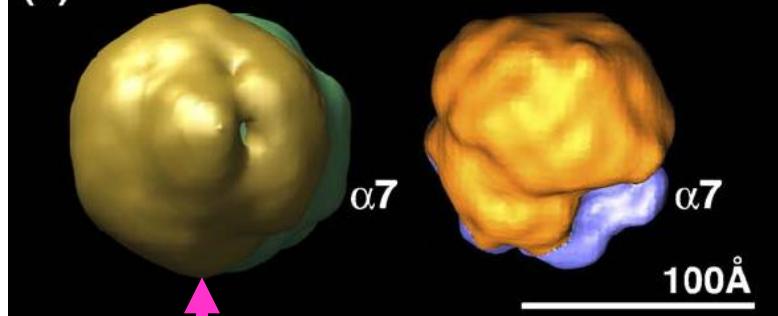


BIm10-20 S Proteasome

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



(b)



PA200 activator- 20S proteasome

Ortega, Heymann, Kajava, Ustell, 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,*}

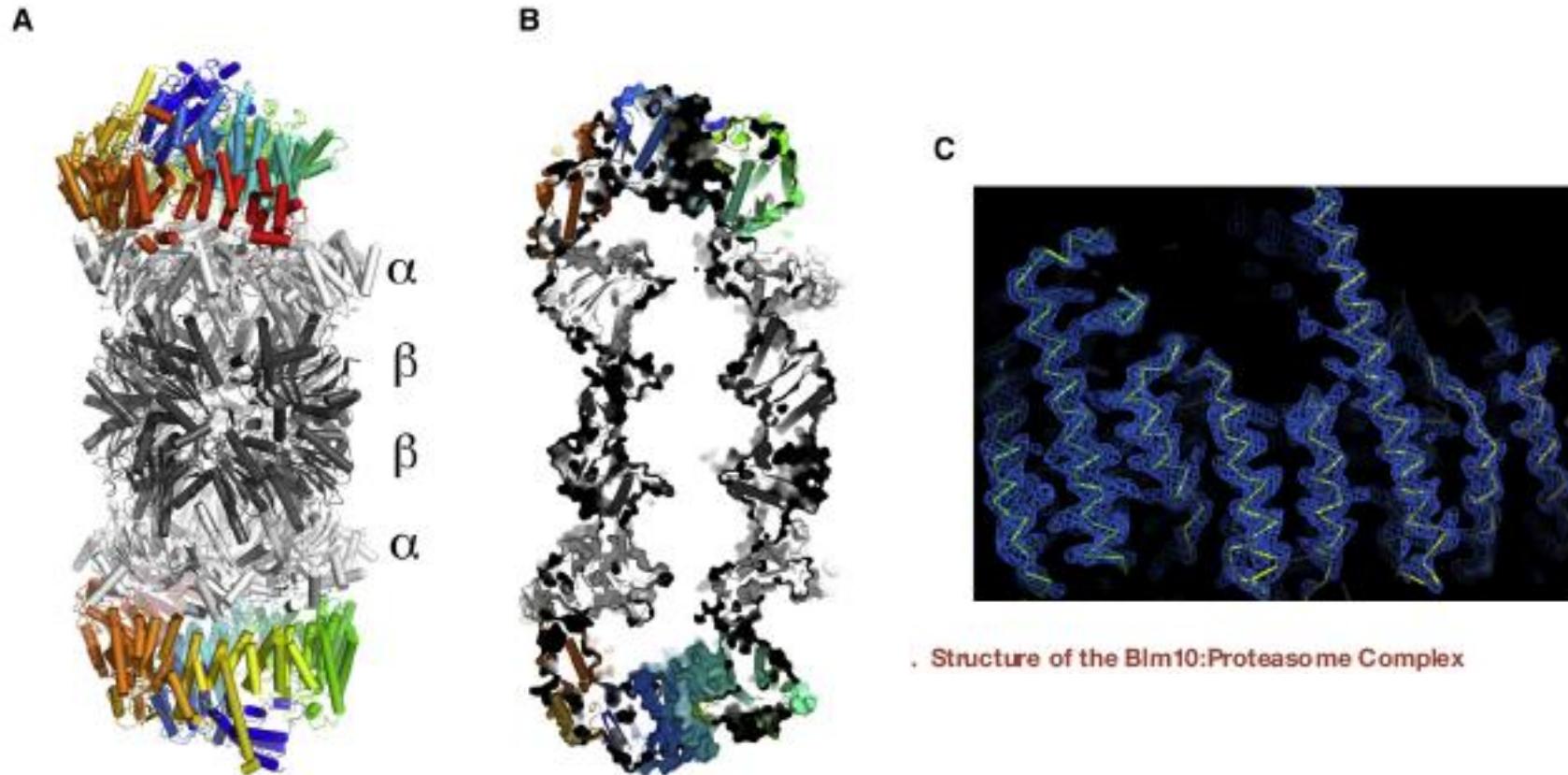
¹Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, UT 84112-5650, USA

²Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000, USA

*Correspondence: chris@biochem.utah.edu

DOI 10.1016/j.molcel.2010.02.002

March, 2010



. Structure of the Blm10:Proteasome Complex



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

S. Potekhin, Institute of Protein Research, Russia

Alasdair Steven NIAMS, NIH, USA

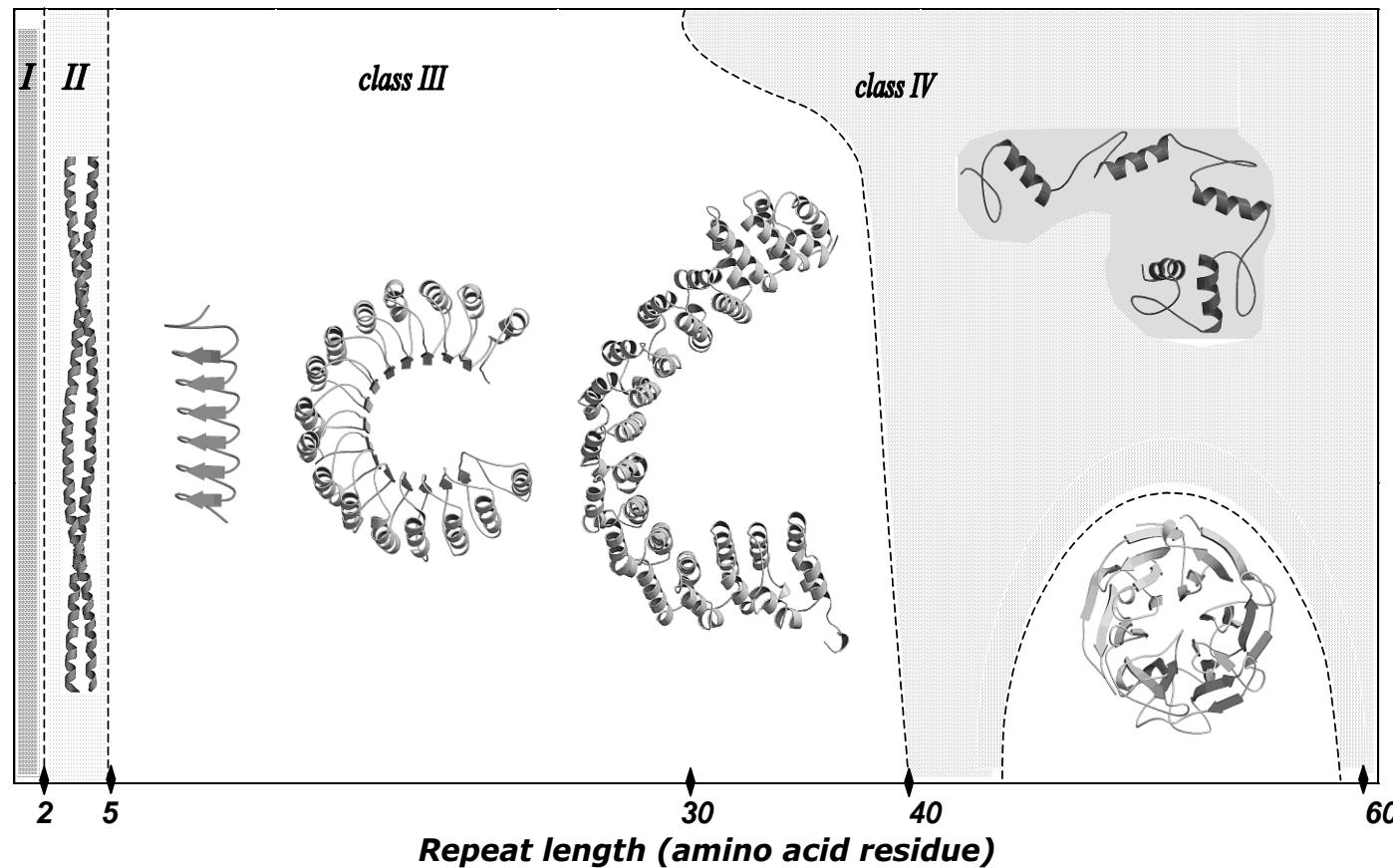
James Hurley, NIDDK, NIH, USA

Aitor Hierro CIC bioGUNE, Bilbao, Spain

Proteins with tandem repeats

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

WHAT CAN REPEAT LENGTH TELL US ABOUT ITS STRUCTURE?



A.V. Kajava (2001) *J. Struct. Biology*, 134:132-144

Solenoid proteins

Kobe and Kajava (2000) *Trends Biochem. Sci.* 25:509.

LRR proteins

Kobe and Kajava (2001) *Current Opinion in Structural Biology* 11:725.

Beta-solenoid proteins

Kajava and Steven (2006) *Advances in Protein Chemistry* 73:55.

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

S. Potekhin, Institute of Protein Research, Russia

Alasdair Steven NIAMS, NIH, USA

James Hurley, NIDDK, NIH, USA

Aitor Hierro CIC bioGUNE, Bilbao, Spain

Distinguishing between structural and functional residue conservations

protein_human VKVSAHGALSIDSMTALGAIGVQAGGSVSAKDMRSRGA~~VTVSG~~.GAVN
protein_rat VHLNAHGALT~~I~~KTMYSGNHISVQAGSHVSAREMHQSAFVT~~V~~H~~CAGSVN~~
protein_yeast VKVSFQSSLIDSMTALGAIGVVSSGSVD~~A~~KDMRSRGA~~VWVSG~~.GAVK

LGDVQSDGQ . VRATSAGAMTVRDVAAADPDGNKKPLALQAGDALQAGFLKSAGAG~~PPP~~DQM...
LGDVQSWGQFVH~~A~~SDGFCMTVRDVSYRDGDPNRYTLGLQAGHALQAYYLRSSSAA ..NDQM...
LAAVNNDGQ . VRATSAGAMCVWDVAAQDPDGNKKPLALSSGDGLKAGFLKSAGAG~~PPP~~DLM...

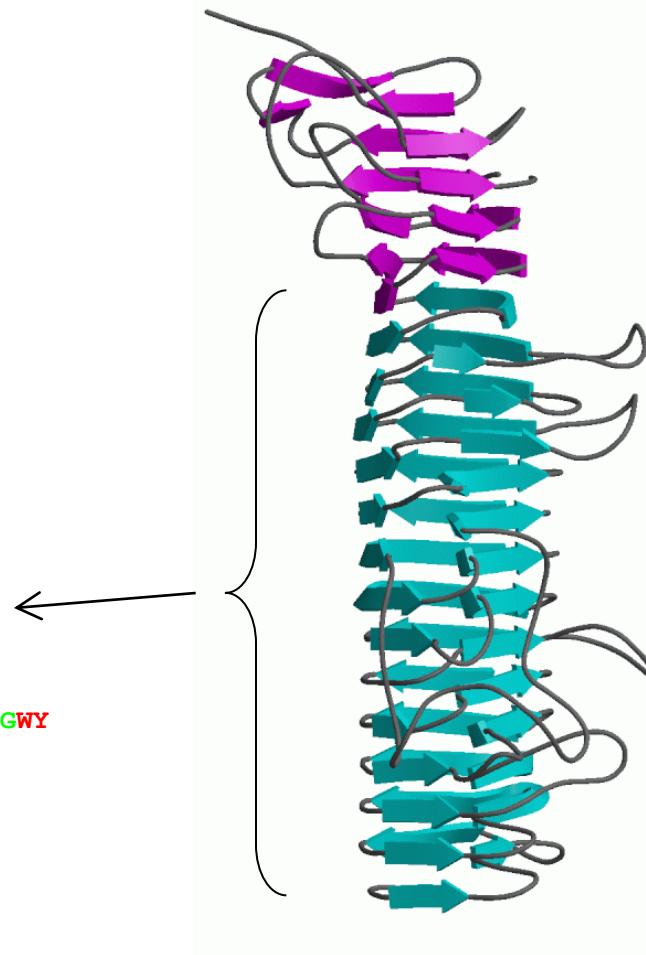
protein_human

VKVS~~A~~HGAL~~S~~IDSMTALGA
IGVQAGGSVS~~A~~KDMRSRGA
VTVSG-GA~~VNL~~G~~D~~VQSDGQ
VRATSAGAMTVRDVAAAADP~~D~~GNKKP
LALQAGDALQAGFLKSAGAG~~PPP~~DQ
MTVNG-DA~~VRL~~DGA~~H~~AGGQ
LRVSSDGQ~~A~~ALGSLAAKGE
LTVSAARAATV~~A~~ELKSLDN
ISVTGGERVS~~V~~QSVNSASR



Pertactin from *Bordetella pertussis*

GILLENPAAELQFRNGSVTSSGQLSDDGIRRFIG
TVTVKAGKLVADHATLANVGDTWDDDG
ALYVAGEQAQASIAADSTLQGAG
GVQIERGANVTVQRSAIVDG
GLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPA
AVSVLGAASELTLDGGHITGGRRAA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGF
GPGGFGPVLGDGWY
GVDVSGSSVELAQSIVEAPELGA
AIRVGRGARVTVSGGSLSAPHGN
VIETGGARRFAPQAAPLSITLQAGAHAQGKA
LLYRVLPEPVKLTGTGGADAQG
DIVATELPSIPGTSIGPLDVALASQARWTG



Distinguishing between structural and functional residue conservations

One sequence repeat

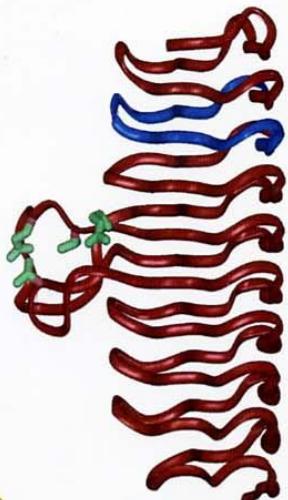
One unit of repetitive structure

VKVSAGALSIDSMTALGA
IGVOAGGSVSAKDMRSRGA
VTVSG-GAVNLGKVQSDGQ
VRATSAGAMTVRDVAAAADPDGNKKP
LALQAGDALQAGFLKSAGAGPPPDQ
MTVNG-DAVRLDGAHAGGQ
LRVSSDGQAALGSLAAKGE
LTVSAARAATVVAELKSLDN
ISVTGGERVSQSVNSASR

Indirect experimental structural evidence

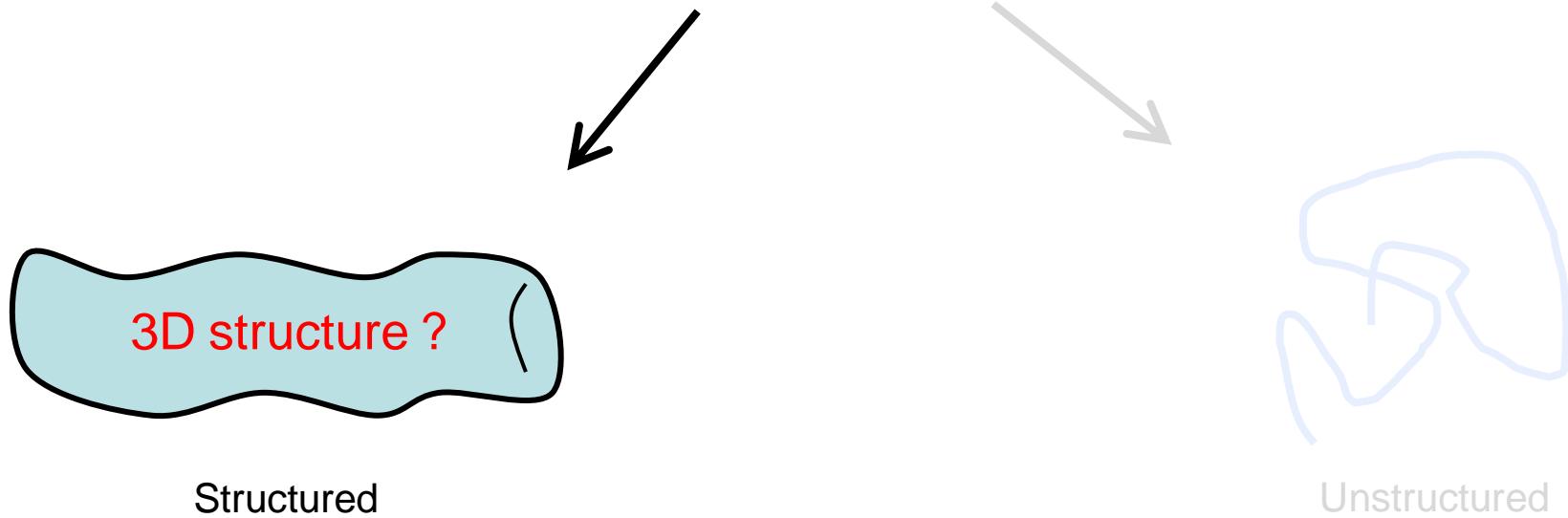
CD spectroscopy
(conformation)

Electron-microscopy
(shape,
oligomeric state)

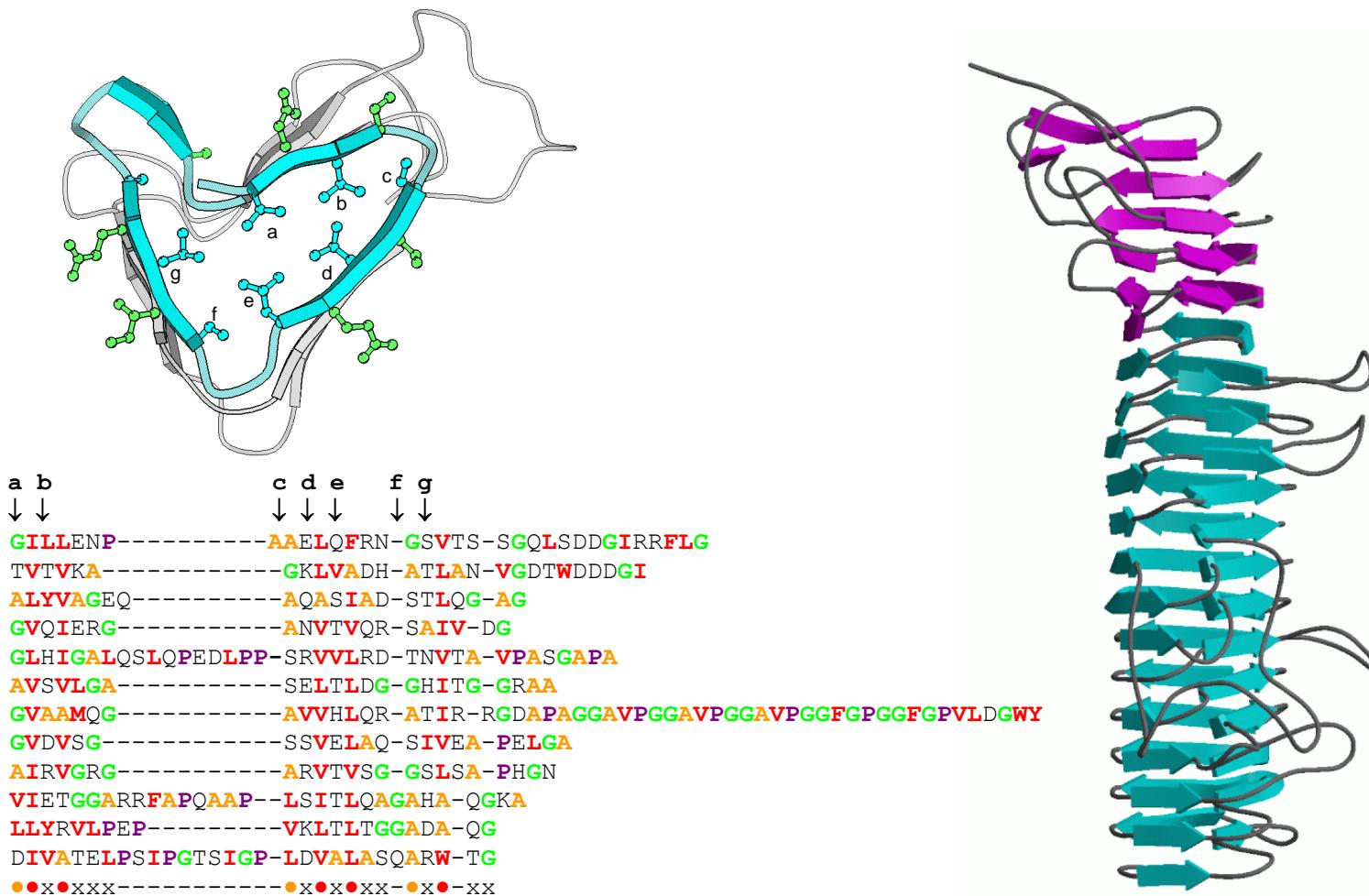


**3D
structural
model**

IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?



Pertactin from *Bordetella pertussis*



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

Cargo recognition complex



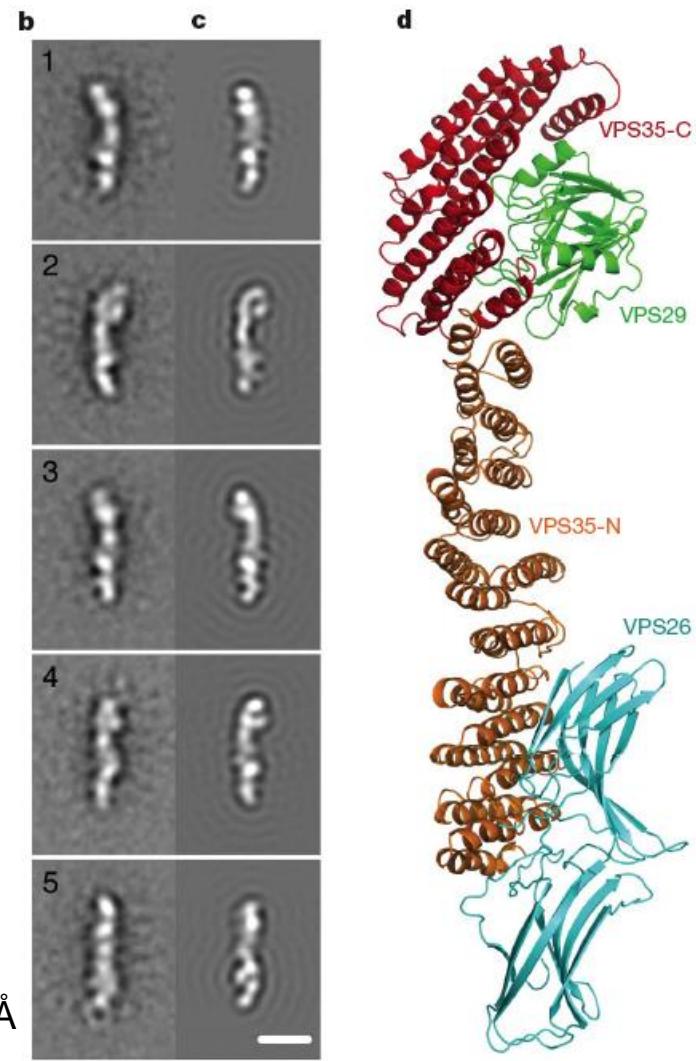
a- Helical solenoid fold prediction for the N-terminal part of vps35 (orange in (d))

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



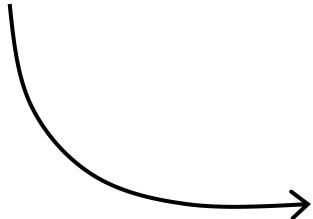
The α -solenoid fold extends the full length of Vps35 and Vps26 is bound at the opposite end from Vps29.



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



*** * * * * ***
GITLENPPS-----AAELQFRN-GSVTNSNGQLSDGI
TITLKATSS-----AKLVADH-ASVANVGQTWDGI

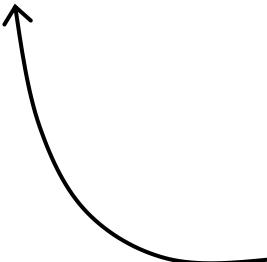


```

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,-3,1,-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,-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,-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,-1,0,-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,-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,-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,-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,-2,-3,-1,-3,0,-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,-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,-1,1,0,-2,-7,-6;
```



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

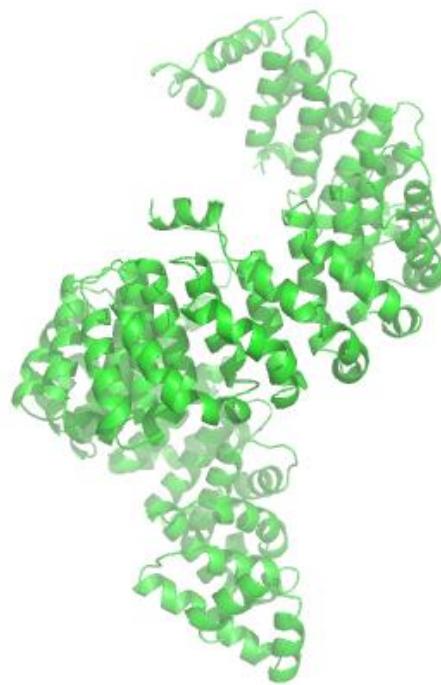
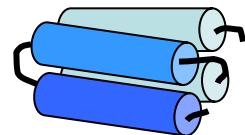


```
MA /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY';
MA /DISJOINT: DEFINITION=PROTECT; N1=1; N2=43;
MA /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE; R1=44.55; R2=-0.0035;
MA     R3=0.7386; R4=1.001; R5=0.208; TEXT='ZScore';
MA /NORMALIZATION: MODE=2; FUNCTION=LINEAR; R1=0.0; R2=0.1;
MA     TEXT='OrigScore';
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,-4,2,-3,-2,1,-2,0,-1,-2,-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,-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,-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,-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,-1,2,0,0,-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,-3,3,3,-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,-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,-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='W'; 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,-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,-2,-3,-1,-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,-1,1,0,-2,-7,-6;
```

Sequence profile search

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

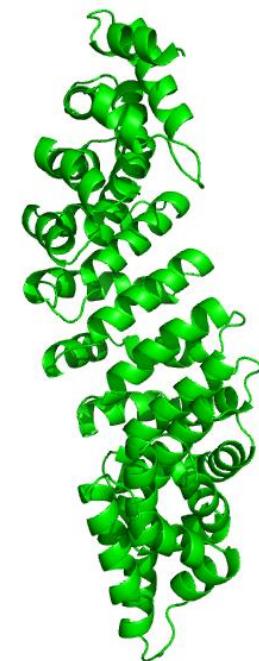
CRBM collection of protein repeats: <http://bioinfo.montp.cnrs.fr>



Twisted and curved



Curved



Twisted

Large subunits of
19S activator of proteasome

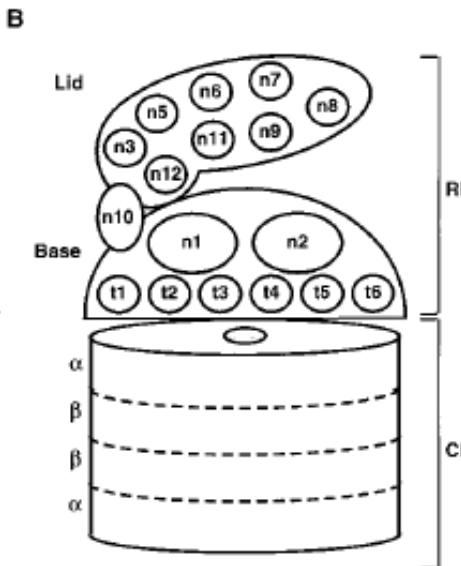
A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex)

The 26S proteasome is a large ATP-dependent proteolytic complex of eukaryotes, which represents the central protease of the ubiquitin-dependent pathway of protein degradation^{1,2}. It consists of two asymmetric 19S caps flanking a barrel-shaped 20S core. The caps are thought to recognize proteins targeted for degradation, unfold them and thread them into the 20S core, where they are degraded to peptides³.

The caps consist of approximately 20 different subunits between 25 and 110 kDa^{3,4}. The two largest subunits are S₁ (also called p112 in mammals and Sen3 in yeast) and S₂ (p97 in mammals, Nas1 in yeast). Their functions are unclear, as is the case for most of the subunits of the 19S complex. Sen3, the S₁ subunit of yeast, was identified as a factor required for the degradation of Sen1, a protein involved in tRNA splicing⁵. TRAP-2 (55.11), the S₂ subunit of humans, was identified by two-hybrid screening as a factor interacting with the intracellular domain of the tumor necrosis factor receptor^{6,7}. Recently, disruption of Nas1, the S₂ subunit of yeast, was found to result in the accumulation of polyubiquitinated protein(s)⁸. Although S₁ and S₂ sequences have diverged significantly (30–45% sequence identity), human p97 can suppress the phenotype of Nas1 mutants⁹, indicating that the human S₂ subunit can be incorporated into the yeast 26S complex.

BLAST searches reveal a low level of similarity between S1 and S2 subunits⁶ and an alignment using MACAW⁹ shows them to be related in a region of about 400

		G G G	
Consensus		A A A A A . . . C	* * * *
Pred. sec. str.		ββββββ	αααααααα
Proteasome S1 subunits			
Ce_p112	1	[418] NAVASLGQHIIIGRS . . . S . K . RP . PKRSVGFEGPKKG 2 GAMAYGIIHAKHGD . . . AT . ST . AQ . KTAEEFREVHR 3 GAGCIGGVGAVCGSS . . . VSN . EK . RE . QRDEAVSGE 4 SAGCAGLMLIAGHLN . . . QE . NF . KC . TVTDQIDKTKR 5 GIFTGLACAGLQG . . . D . E . P . KE . GAKSNPMLRSTGEL [17] 6 YATIGGIVLSPKDP . . . T . S . AM . PEHNGIVRY 7 GAMALGIGACAGTCN . . . ME . A . EP . SDKEGFVRK 8 GALESLALIMCQDITCPK NG . RKC . KKIGERNEEDSLEVK 9 GATAQQLLGLIQQ . . . A . T . NSQSAQGMSGMVGMK [222]	
Sc SEN3	1	[365] TATASLGVIHKGNL . . . ECKK . PAP . INGRASSRPIKG 2 GLSLVG . GLTIAGPGRDUTY . KN . VENSHTGQGMDVEDVLLH 3 GASLGICLAMGSAN . . . IE . YE . . . PNEASATCOP 4 AAAACGKGLCMG . GK . . . PE . IED . PT . SQE . QHGNITHE 5 GLAVGIALVINGRC . . . L . DD . TK . ASDESLLKGG 6 GCAFTIALAYAGUON . . . NSA . KR . HV . SDSINULVRA 7 AAATAGCFVLLPDY . . . T . VPR . VO . SRSNHNAHVR 8 GTFAFLCAGCAGKL . . . Q . ID . DP . EKOPDFVFR 9 AAMALSMILQQTFSKLNPQ . AD . NKN . . . SVTTHKQHGIARE [211]	
Proteasome S2 subunits			
Hs_p97	1	[408] SAMASLAMILWDVC . . . CG . TC . DK . VSSEDIYIKG 2 GALI . AGC . VNSGQRN . ECP . A . SD . DFNNSNPMRL 3 GSIFUGLGLAYAGGNR . . . ED . . . LP . . . GDCKSSMEVAG 4 VPAACOMMAGVSCN . . . GDVTST . QC . EKESTERLDTYAR 5 WLPICLGLNHLRKE . . . A . EAA . AA . EKVVEPRPSTANIL [57] 6 VAVLCLALIAJGEEI . GAEM . RT . GH . RNEGETPLRU 7 AVPLALALIIVVNPR . . . LN . DT . SK . SHADAPPEVSY 8 NSIFPAMGMVGSTHN . . . AR . AR . RQ . QYHAKDPPNLF 9 KVRLAQGCLTHLGKT . LTLCP . HSDHQ . SCQAVACLL [111]	
Sc_Nast	1	[416] SAVASIGSIYQWNLD . . . G . QQ . DK . VYDEPEVKA 2 GALLGIGIGAEGVWIDGEVPP . L . QD . TKEPDITKISS 3 AALLGCLIAPIAGSKN . . . DW . G . LP . ASTON . EIEAA 4 MASLALAHFV . VGTIN . . . GDITTS . DN . ERDAIELKTDWUR 5 FLALRLCLILYMGCGE . . . Q . DD . EP . SAIEHPPMTSAEVL [135] 6 VAVLCLALIAJGEEI . . . CKES . RH . CH . HYCNELIRK 7 KVPLANGIIVSVSDPQ . . . MK . DT . CR . SHGADLEVR 8 NSIFPAMCLOCAGTNN . . . AR . Q . EO . SYNSRQDQALEF 9 ITPLAQGCLHLHGKCI . . . NMID . FNDAH . NKVTLASIL [111]	
Pred. sec. str.		βββββββ	αααααααααααα
Cyclosome subunits			
En_BimE	1	[1361] GFLALALGCLNLHRLSL . . . AKWVAFKYLTPRKTETSI 2 G . L . L . A . S . A . Y . L . G . M . D . . . T . V . I . V . H . L . R . H . V . T . M . P . M . G . A . P . N [7] 3 AGMIGLGLLYCNSQE . . . RRMSEVPLCIFIENADCEGSA [121] 4 AACPNAGLKHNLQKX . DLKEMGRDMH . IVEHLA . AVA . VGH . KNV [111] 5 GATIALATIERTND . . . FTIAQKQVTPDITPTTVYVYRPU [55] 6 GLCFCALCIREASPD . . . PTIVDILLYLQLDQFIRISRLP [18] 7 VVALGAAVMAFGTG . . . ALPFRRLSLHGRVDDPMPYGHSM [18] 8 AARMAGMLFLGGGS . . . YTGLTSNLAVASLCSLYPIPP [313]	



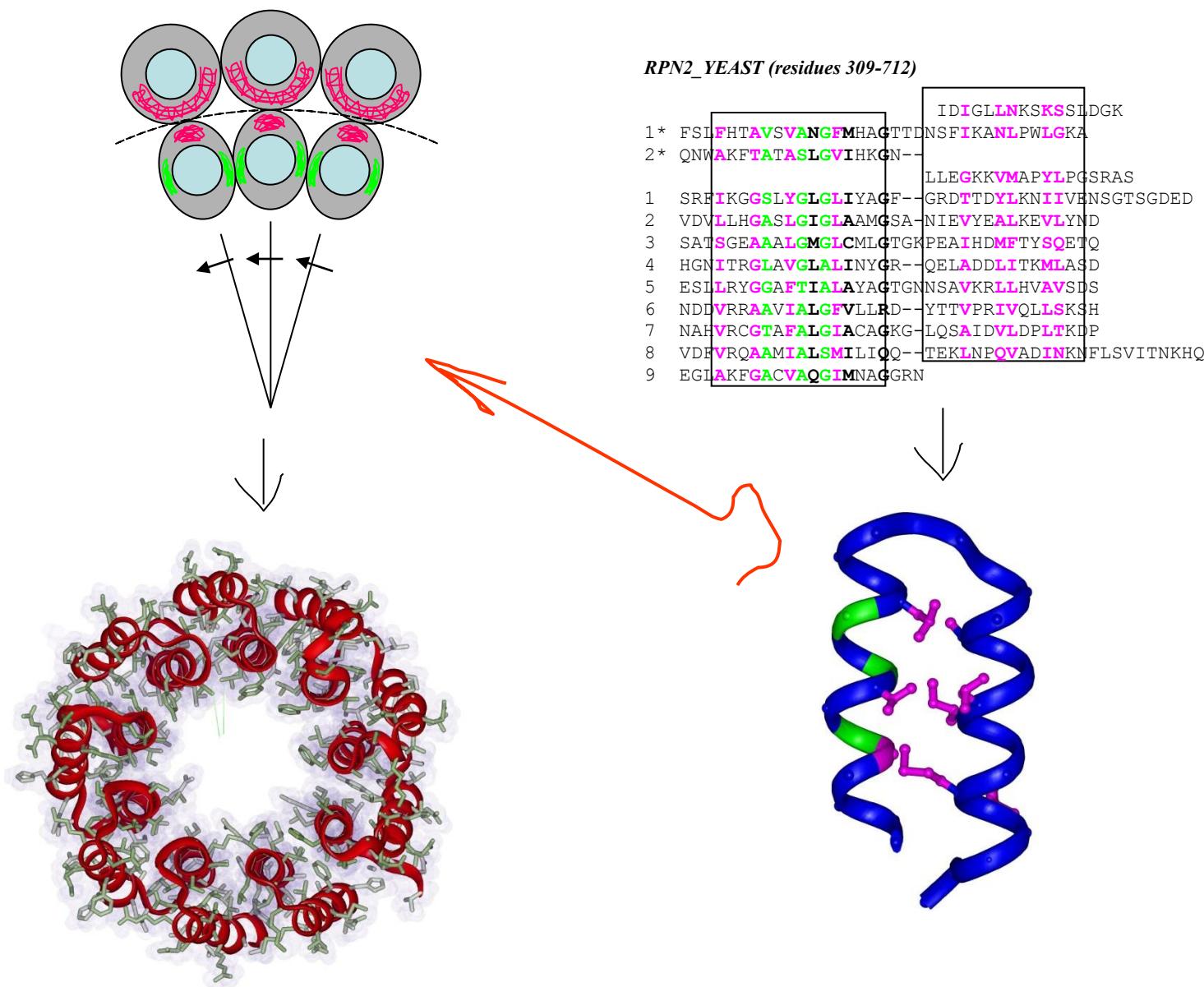
ANDREI LUPAS AND
WOLFGANG BAUMEISTER

Max-Planck-Institut für Biochemie, D-8215
Martinsried, Germany.
Email: lupas@vms.biochem.mpg.de

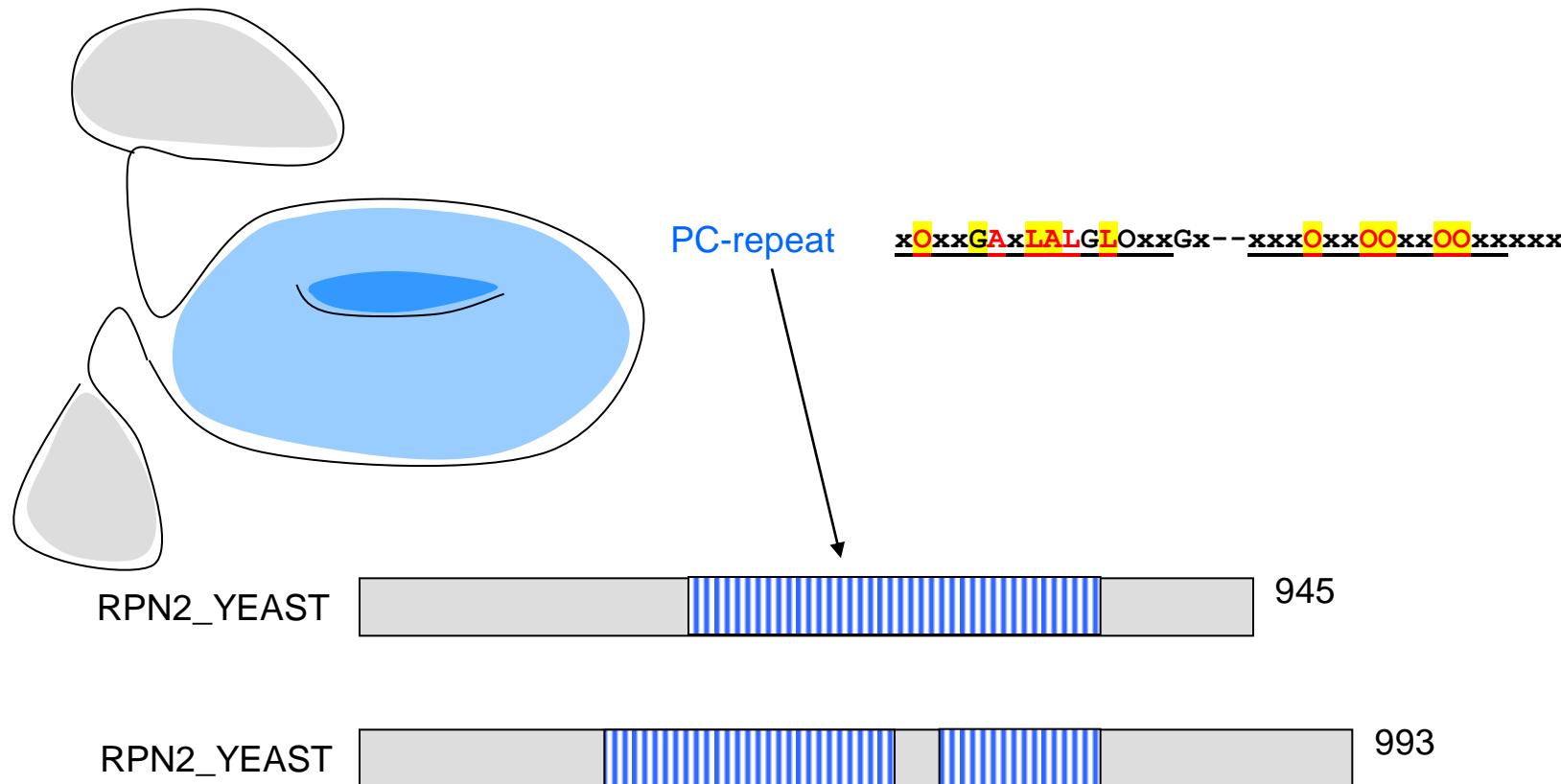
KAY HOFMANN

Swiss Institute for Experimental Cancer
Research, Chemin des Boveresses 155,
CH-1066 Epalinges s/Lausanne, Switzerland

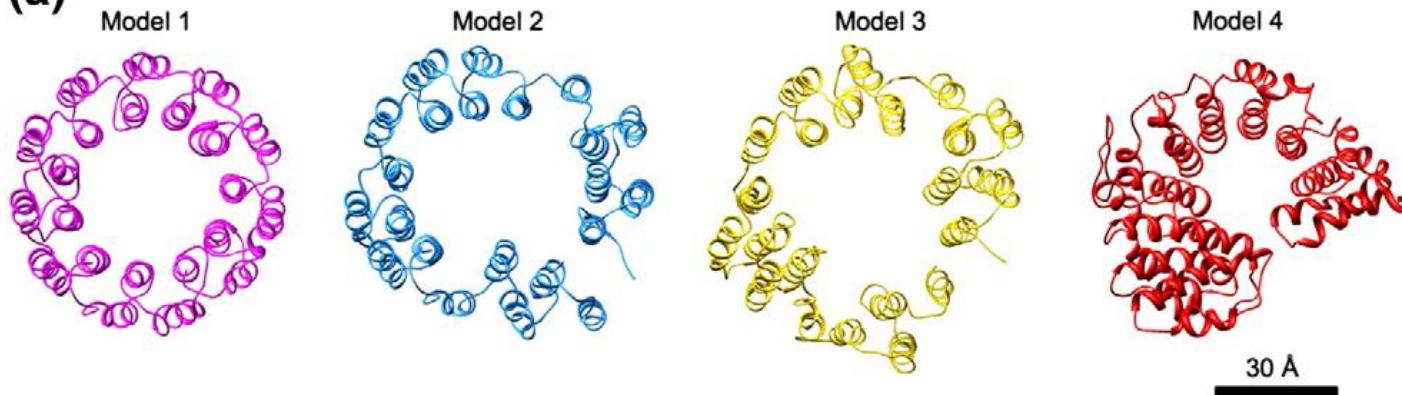
Molecular modelling of Rpn1 and Rpn2 subunits of eukaryotic proteasome



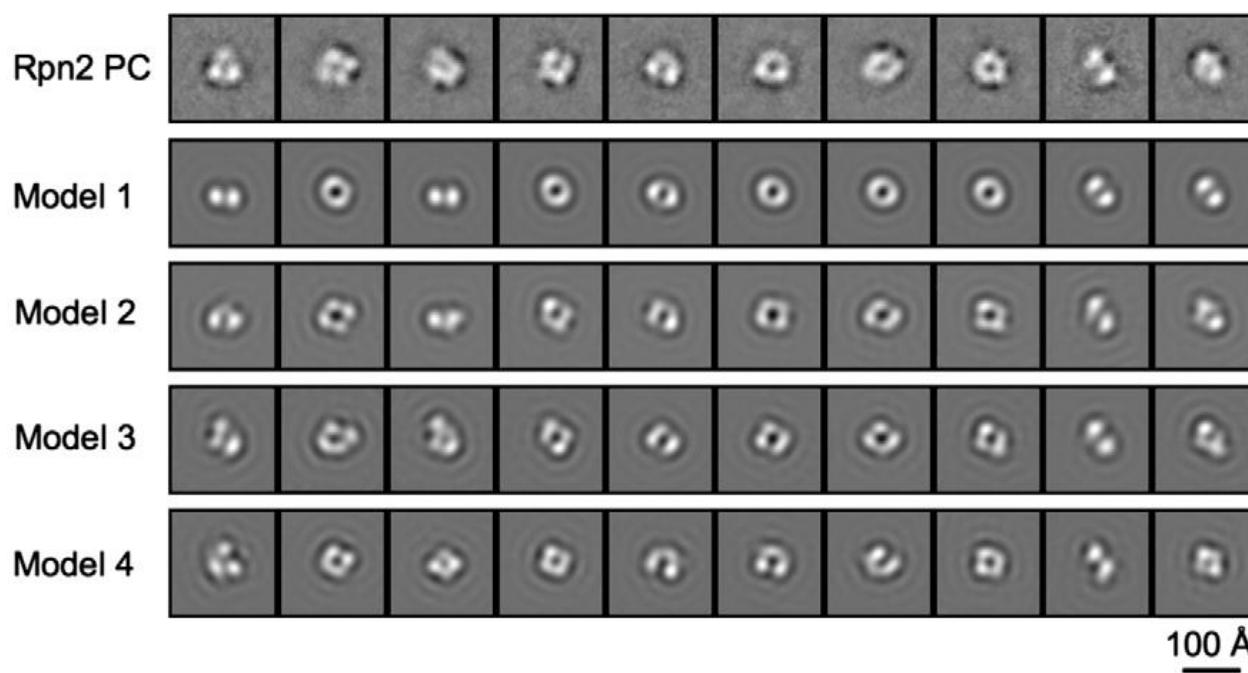
Identification of new Rpn1 and Rpn2 repeats



(a)



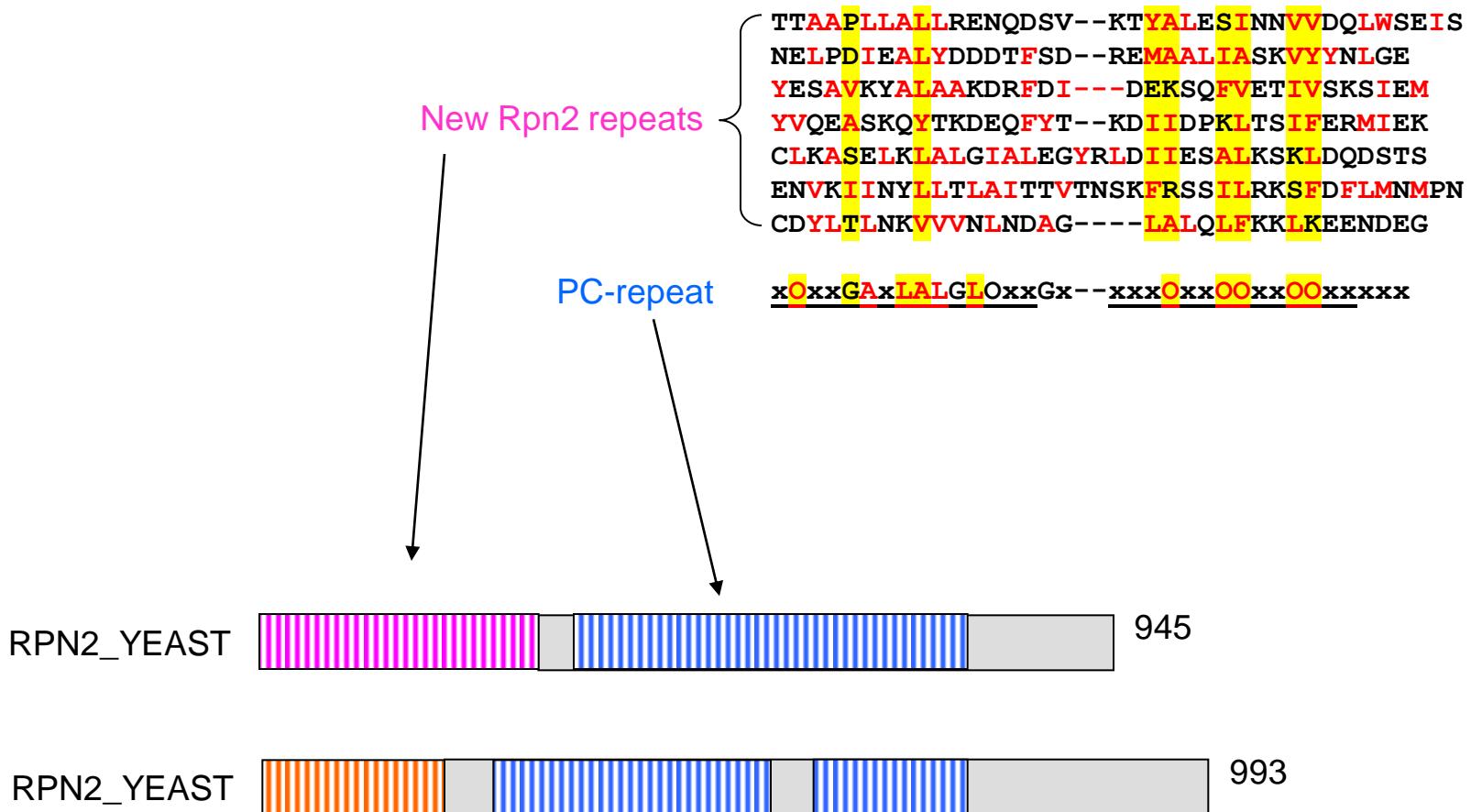
(b)

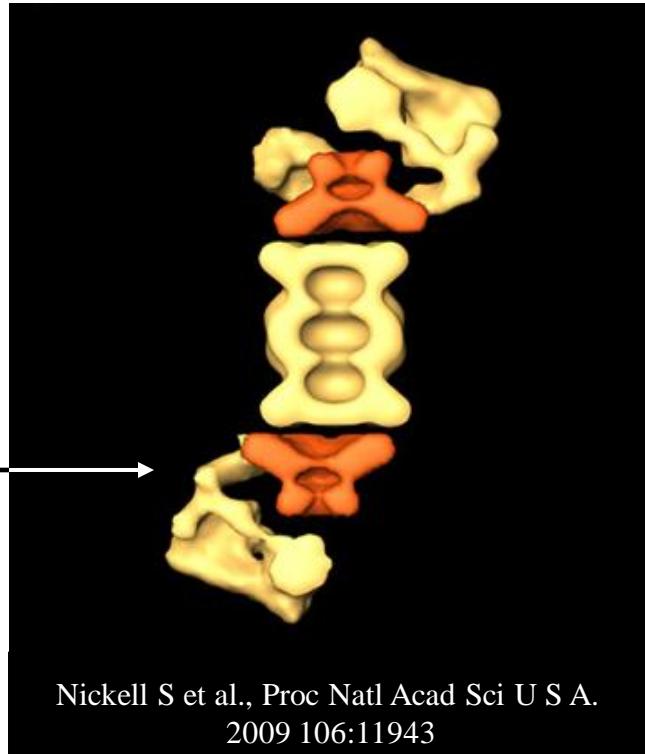
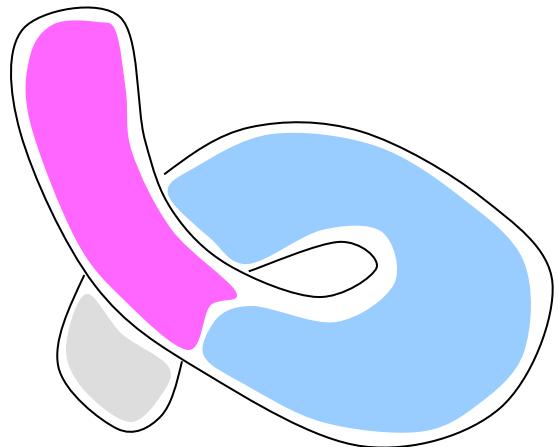


Alpha-helix from CD spectroscopy

Effantin G, Rosenzweig R, Glickman MH, Steven AC.
J Mol Biol. 2009 Mar 13;386(5):1204.

Identification of new Rpn1 and Rpn2 repeats



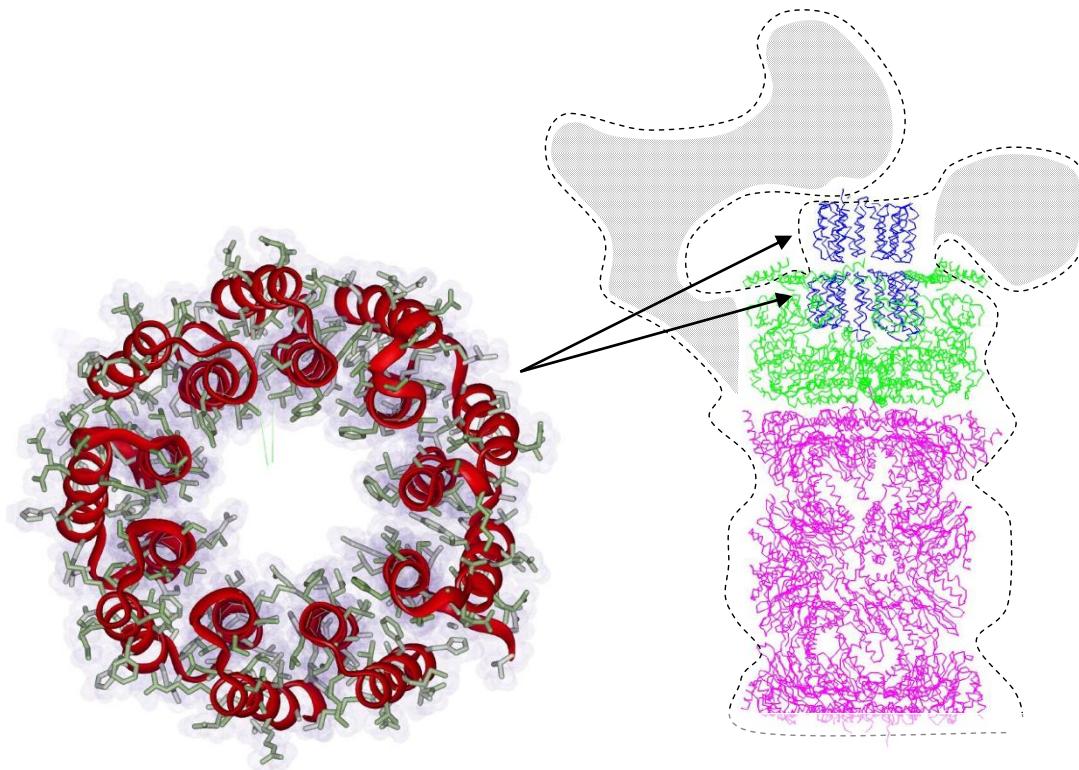


Nickell S et al., Proc Natl Acad Sci U S A.
2009 106:11943

RPN2_YEAST 945

RPN2_YEAST 993

Structural models of Rpn1 and Rpn2 subunits of eukaryotic proteasome



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



Glickman et al., (1998) Cell, 94, 615

FHA is a member of a large family of autotransporter proteins

(Over 1000 proteins)

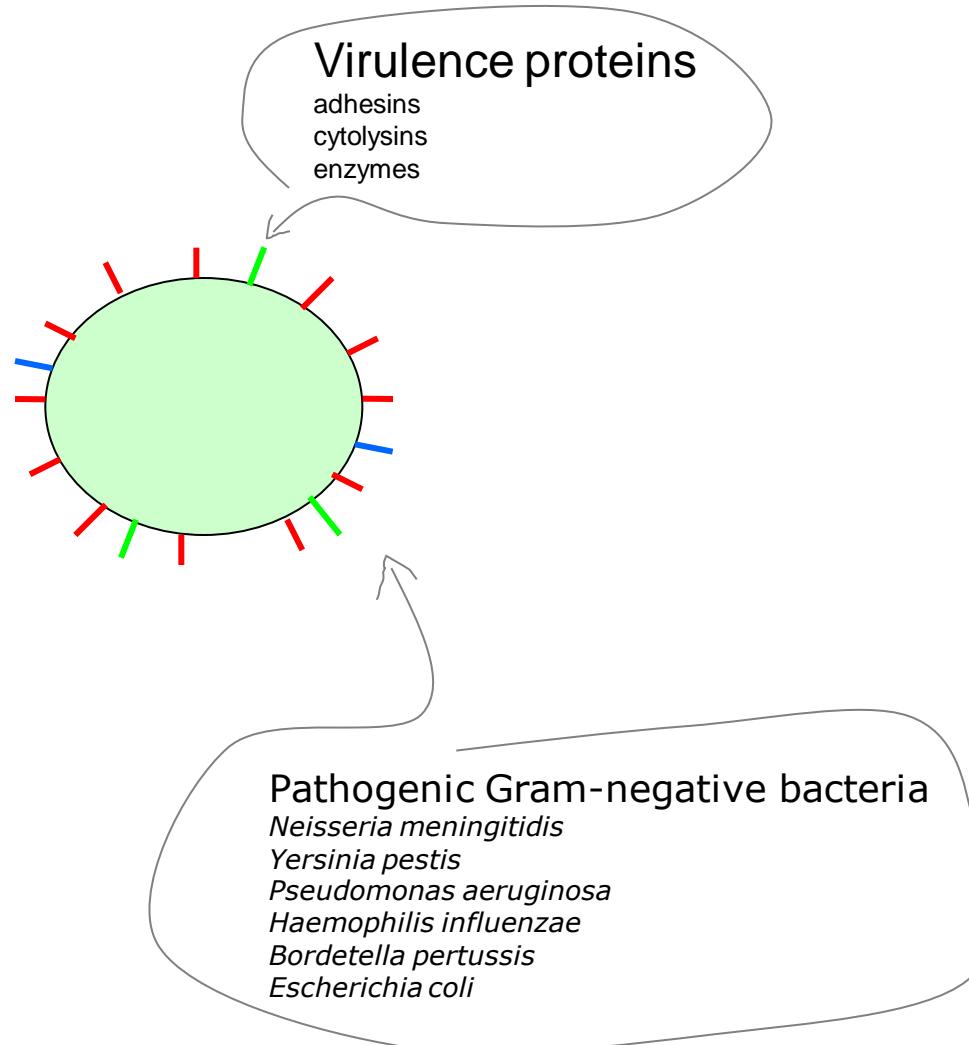


Table 1

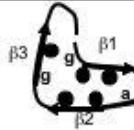
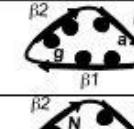
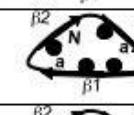
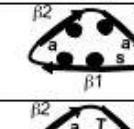
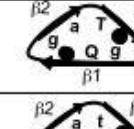
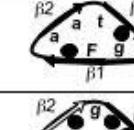
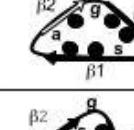
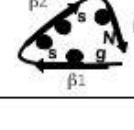
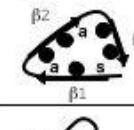
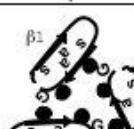
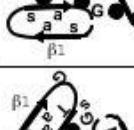
Nº	Representative protein	Repeat length	Consensus sequence of repeat	Coil of β -solenoid
L1	Serum resistance protein brkA (<i>B. pertussis</i>)	22-26 res	g●x●xx-ax●x●xxgx●xx-xxxx β1 → β2 → β3 →	
L2	Slr1753 protein (<i>Synechocystis sp.</i>)	23-28 res	x●xtxxxxx-Gx●x●xaxx●x●xx β1 → β2 → β3 →	
L3	AGRL_3085 protein (<i>Agrobacterium tumefaciens</i>)	25-27 res	xxGx●x●x-xaxsx●xxxgx●x●xxx β1 → β2 → β3 →	
T1	FHA protein <i>B. pertussis</i>	18-19 res	●x●xgxxxx●x●xx●xaxx β1 → β2 → β3 →	
T2	FHA protein <i>B. pertussis</i>	19-20 res	●x●xaxx-●xNxgx●xaxxx β1 → β2 → β3 →	
T3	HBP protein <i>E. coli</i>	18-20 res	s●x●x●xx-ax●x●xx●xaxx β1 → β2 → β3 →	
T4	TibA protein <i>E. coli</i>	18-19 res	gxQx●x●xxgxaxxxTx●xxg β1 → β2 → β3 →	
T5	YapA protein <i>Y. pestis</i>	18-19 res	gxFx●x●xxaxaxxxtx●xxx β1 → β2 → β3 →	
T6	Hap protein <i>Haemophilus influenzae</i>	19-20 res	s●x●x●xxax●xgx●x●xxx β1 → β2 → β3 →	
T7	Hemagg.-hemolysin related protein <i>E. coli</i>	20-22 res	gx●xsx-x●x●xsx-gx●xNxx β1 → β2 → β3 →	

Table 1 (continuation)

Nº	Representative protein	Repeat length	Consensus sequence of repeat	Coil of β -solenoid
T8	LSPA1 <i>Haemophilus ducreyi</i>	20-21 res	s●x●xax-x●x●xaxxx●x●xx β1 → β2 → β3 →	
O1	XadA protein <i>Xylella fastidiosa</i>	13-14 res	s●x●xax-xxx●a●Gxx β1 → β2 →	
O2	UspA2H protein <i>Moraxella catarrhalis</i>	15 res	Nxax-GxxST●aGGxx β1 → β2 →	

Beta-solenoids are found in about 500 of 1000 AT and TPS proteins

Kajava and Steven (2006) J.Struct.Biol. 155,306.

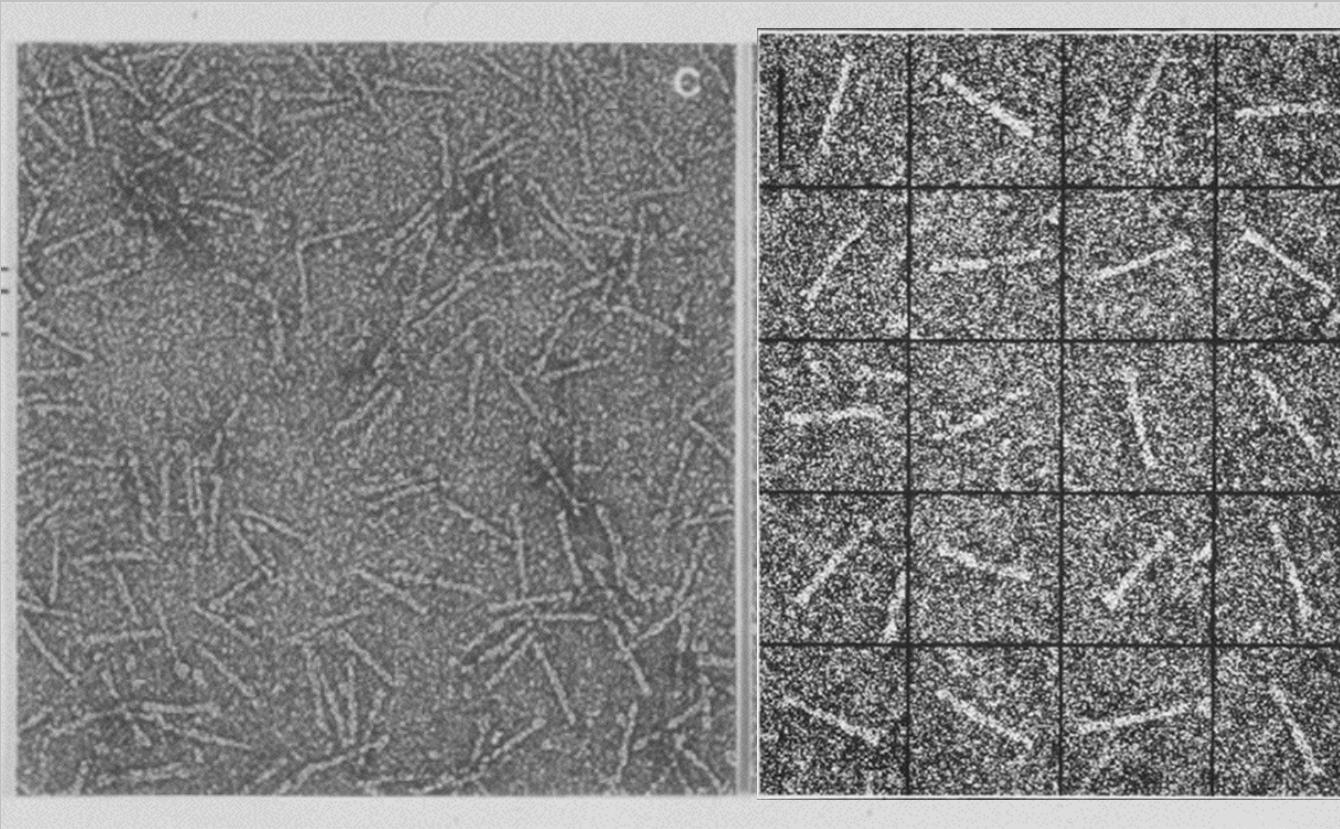
Tandem repeat regions in proteins: the more perfect the less structured

	$P_{sim} = 0.7-0.8$	$P_{sim} = 0.8-0.9$	$P_{sim} = 0.9-1$
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 (in press).

Filamentous Haemagglutinin adhesin

major virulence factor of *Bordatella pertussis*,
etiological agent of whooping cough

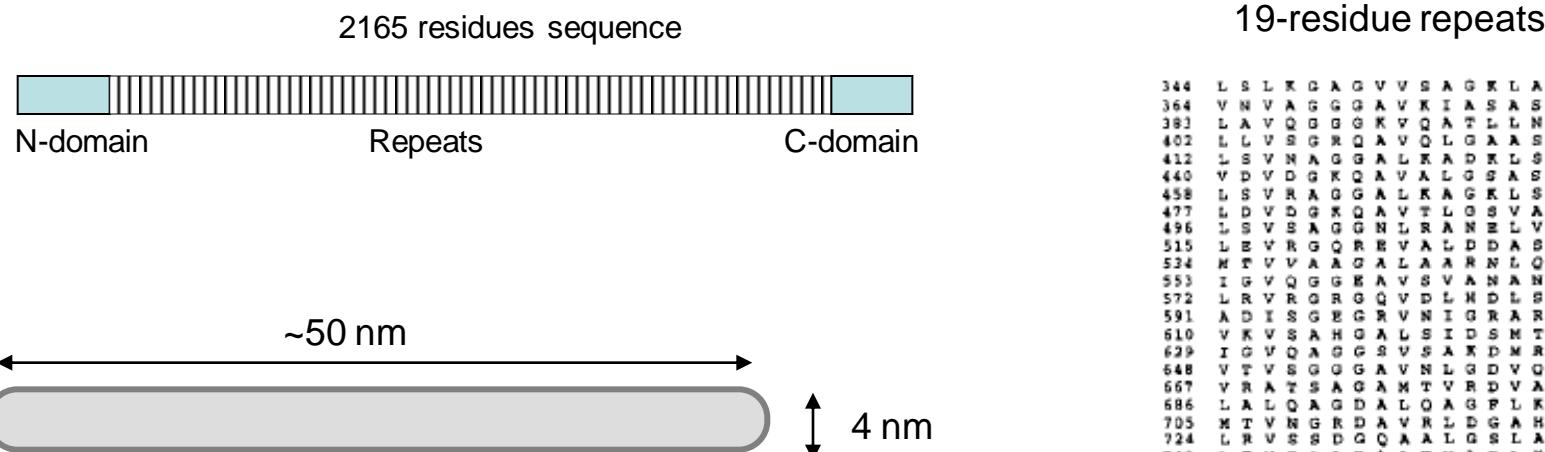


EM
negatively stained



Rod-like shape
 $50 \times 4 \text{ nm}$

Filamentous Haemagglutinin adhesin (FHA) of *Bordetella pertussis*



Rod-like shape according to EM



β -structural protein
according to circular dichroism spectroscopy measurements

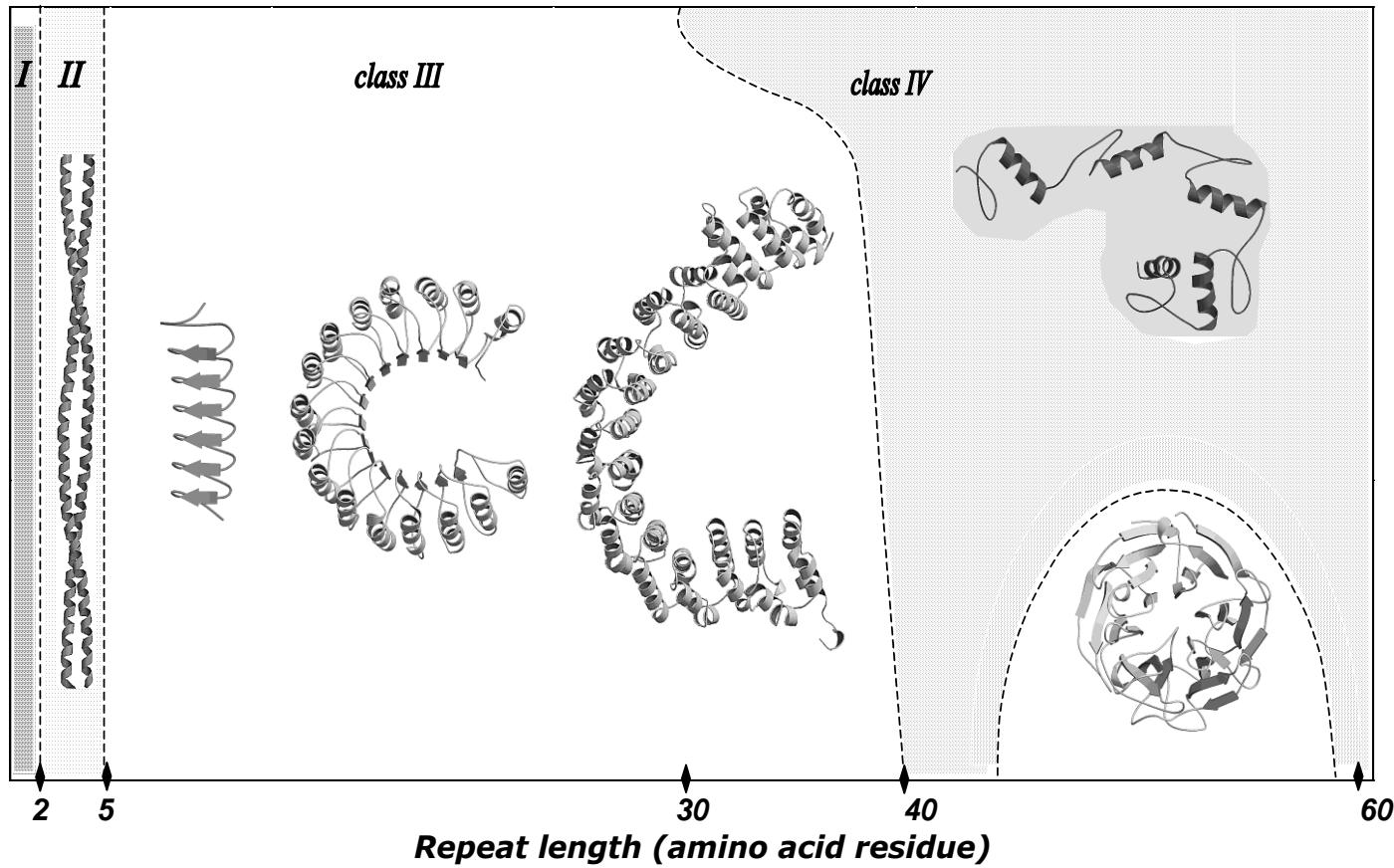
19-residue repeats

344	L S L K G A G V V S A G K L A S G G G A	363
364	V N V A G G G G A V K I A S A S S V G N	382
383	L A V Q G G G K V Q A T L L N A G G T	401
402	L D V S G G Q A V Q O L G A A S S R Q A	420
412	L S V N A G G A L L R A D P L S A T R R	439
440	V D V D G K Q A V A L L G S A S S N A	457
458	L S V R A G G A L L K A G K L S A T G R	476
477	L D V D G K Q A V T L G S V A S D G A	495
496	L S V S A G G G N L R A M E L V S S A Q	514
515	L E V V R G Q R E V A L D D A S S A R G	533
534	H F V V A A G A L A A R N L Q S R G A	552
553	I G V Q G G E A V S V A N A N S D A E	571
572	L R V R G R G Q V D L H D L S A A R G	590
591	A D I S G E G G R V N I G R A R S D S D	609
610	V K V S A H G A L S I D S M T A L G A	628
629	I G V Q A G G S V S A X D M R S R G A	647
648	V T V S G G G A V N V L G D V Q S D G Q	666
667	V R A T S A G A M T V R D V A A A A D	685
686	L A L Q A G D A L Q A G F L K S A G A	704
705	M T V W G R D A V R L D G A H A G G Q	723
724	L R V S S D Q Q A A L G S L A A K G E	742
743	L T V S A A R A A T V A E L K S L D N	761
762	I S V T G G E R V V S V Q S V N S A S R	780
781	V A I S A H G A L D V G K V S A K S G	799
800	I G L E G W G A V G A D S L G S D G A	818
819	I S V S S G R D A V R V D Q A R S L A D	837
838	I S L G A E G G A T L G A V R A A G S	856
857	I D V R G G S T V A A N S L H A N R D	875
876	V R V S S G K D A V R V T A A T S G G G	894
895	L H V S S S G R Q L D L G A V Q A R G A	913
914	L A L D G Q A G Q V A L Q S A K A S G T	932
933	L H V Q G G E H L D D L G T L A A V G A	951
952	V D V N G F G D V R V A K L V S D A G	970
971	A D L Q A G R E M T L G I V D T T G D	989
990	L Q A R A A Q Q K L E L G S V R S D G G	1008
1009	L Q A A A A Q Q G A L S L A A R V Q A	1027
1028	D E L S G G Q G V T V D R A S A S R A R	1046
1047	I D S T G S V G I Q A L K A G A V E A	1065

consensus

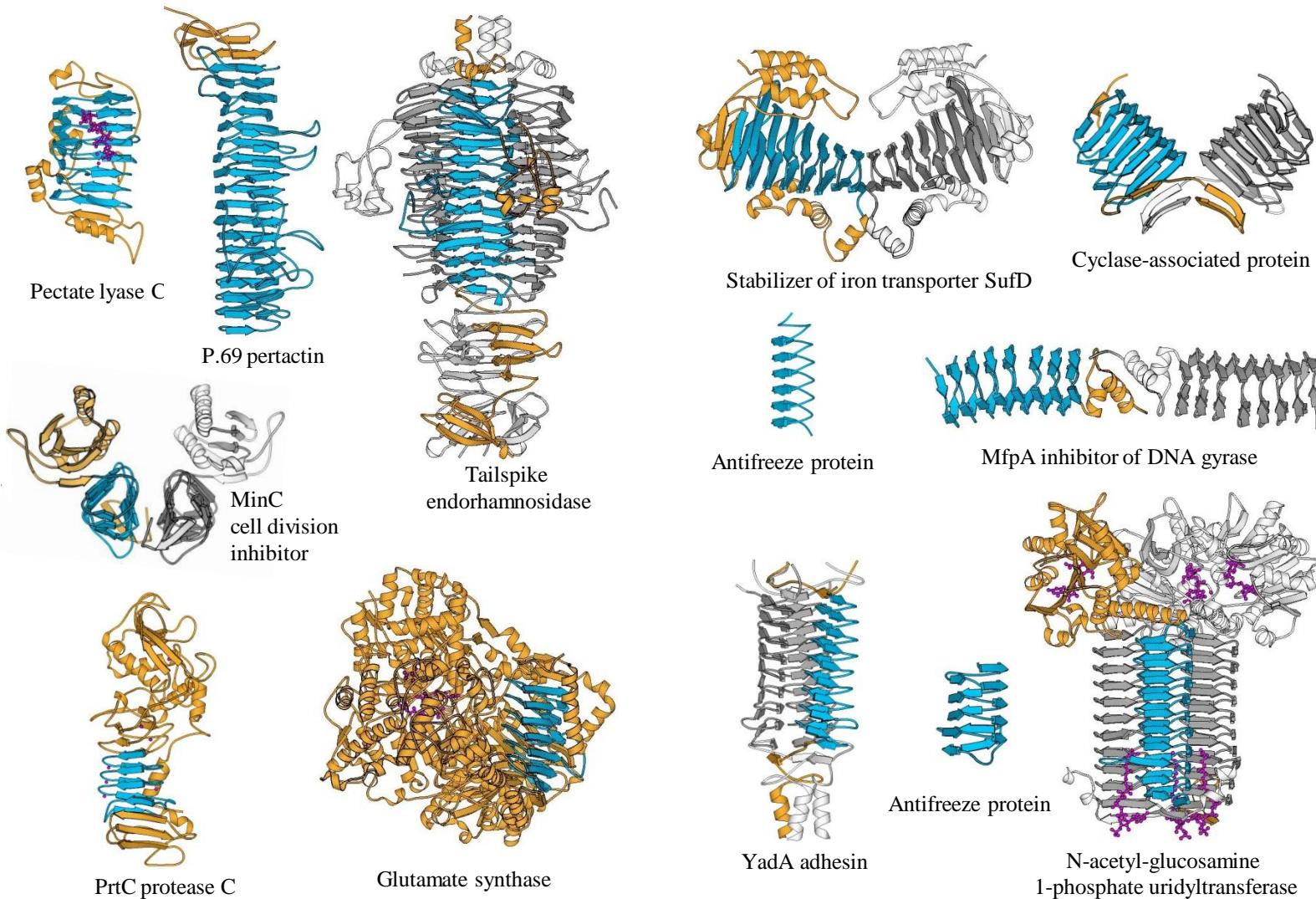
1	v	s	g	h	l	1	1	1
i	=	-	-	-	1	-	a	-
v	l	a	v	a	v	-	a	g

WHAT CAN REPEAT LENGTH TELL US ABOUT ITS STRUCTURE?



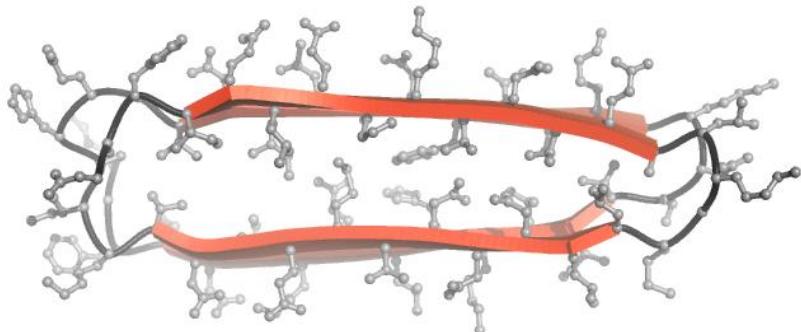
A.V. Kajava (2001) *J. Struct. Biol.* **134**:132

The known structures of β -solenoid proteins



Classification of beta-solenoids

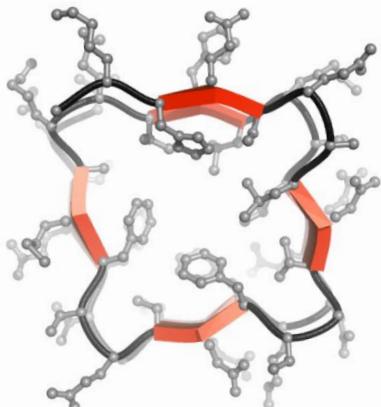
Cross-sectional shapes



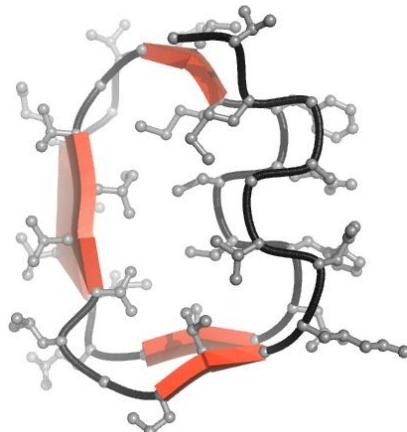
O-type



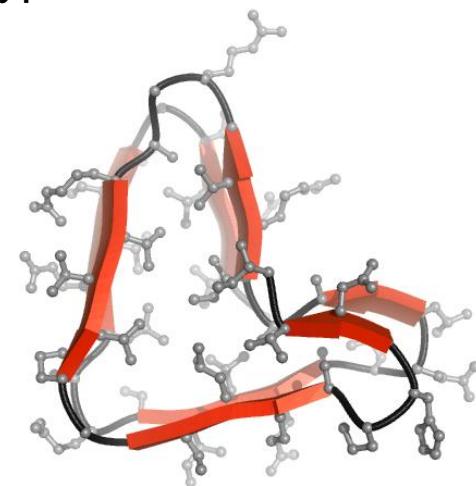
T-type



R-type



B-type

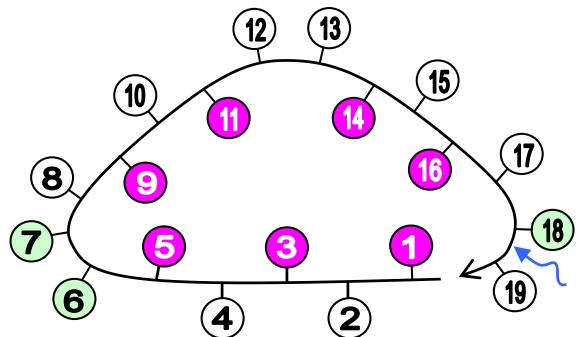


L-type

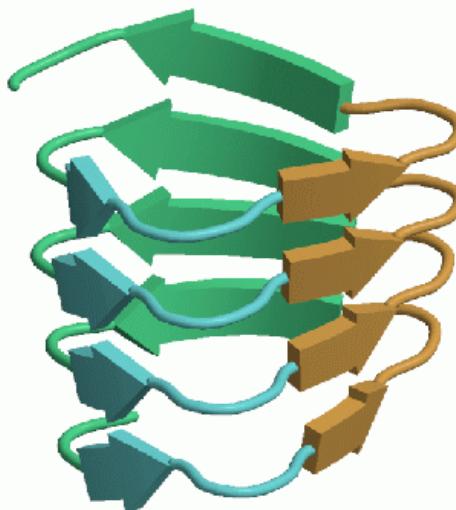
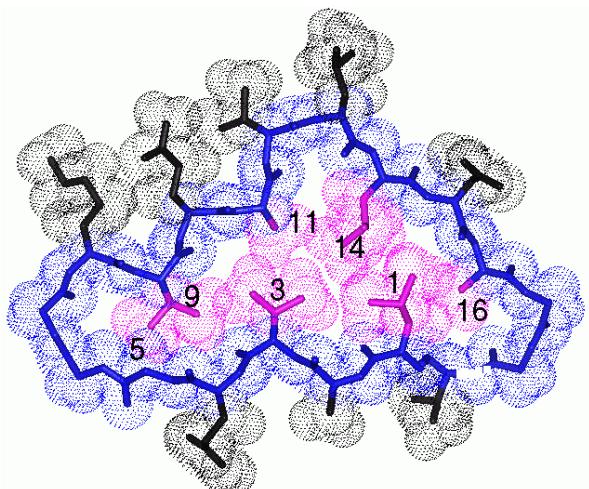
Repeat 1 V N V A G G G A V K I A S A S S S V G - N
 Repeat 2 L A V Q A G G K V Q A T L L N A G G - T
 Repeat 3 L L V S A R Q S V Q L G A L S A R Q - A
 Repeat 4 L S V N A G G A L K A D K L S A T G S R

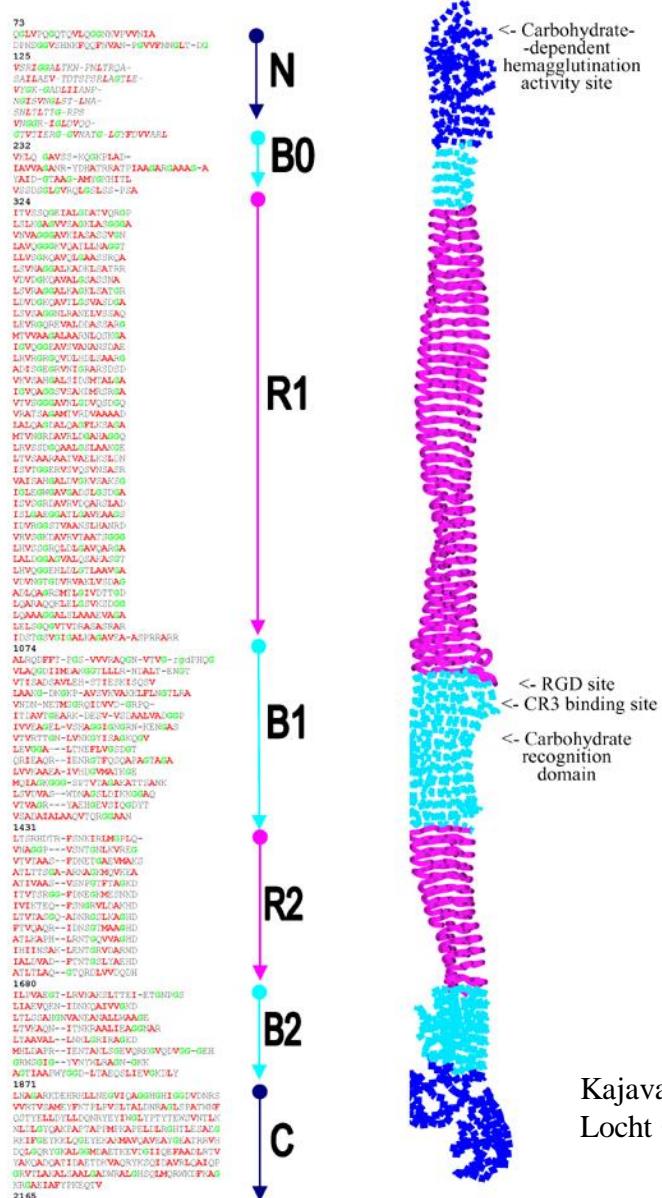
consensus
 positions 1 3 5 7 9 11 13 15 17 19

2D plot

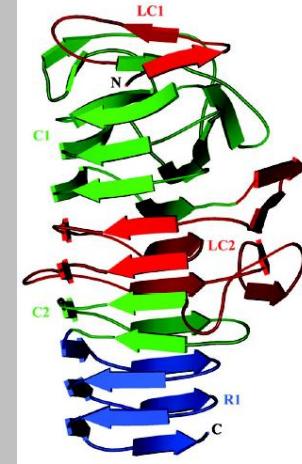
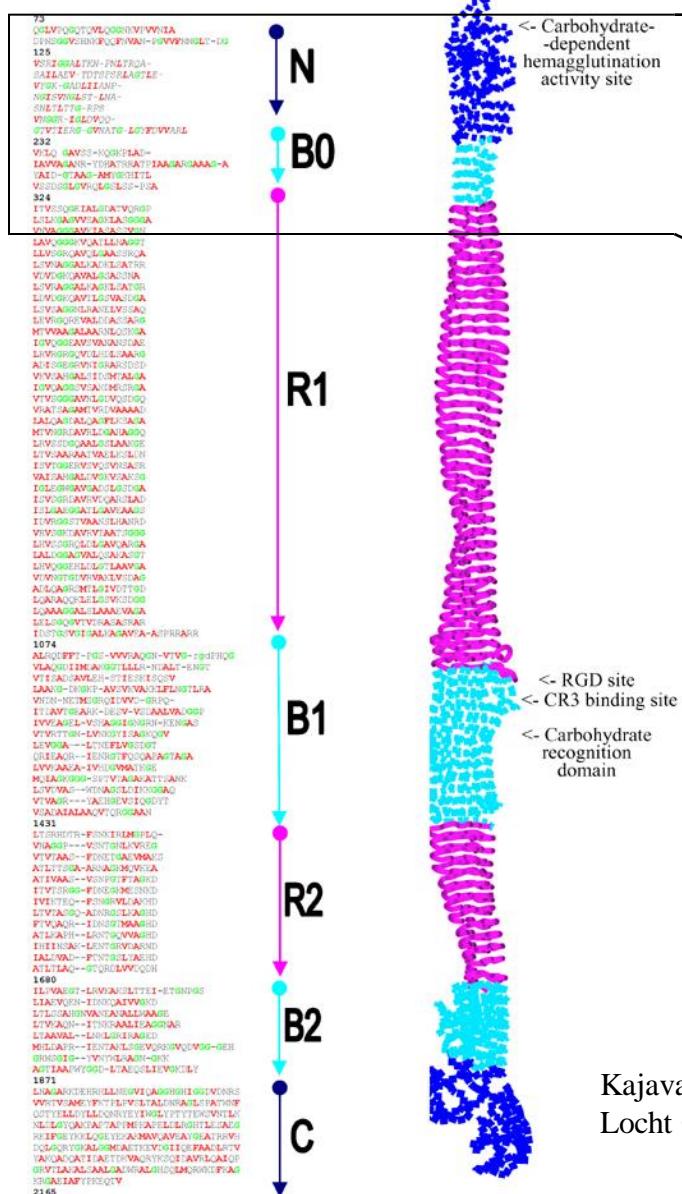


3D structure





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

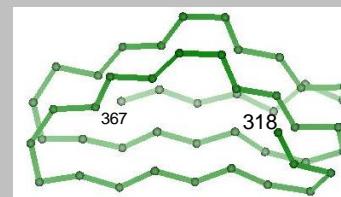
Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Locht C, Steven AC. *Mol Microbiol*. 2001; 42(2):279



• <- Carbohydrate-dependent hemagglutination activity site



Model
(Kajava et al., 2001)



Crystal structure (Clintin et al., 2004)

RMS deviation of C_α atoms is 1.1 Å

Benchmark of T-REKS, INTREP, TRED and XSTREAM programs executed on two databanks of protein sequences

TRIPS (893 sequences with tandem repeats)		SWISSPROT (342391 sequences)		
	Sequences identified*	Execution time	Sequences identified*	Execution time
T-REKS ¹	889	2m	21324	5h50
INTREP ²	863	25m	19405**	22h20
TRED ³	866	4m	14499	16h10
XSTREAM ⁴	818	40s	19040	10m

Benchmark has been performed with a Personal Computer Pentium 4 3.00 GHz and 2Gb of RAM.

**Sometimes, the number of identified tandem repeats exceeds the number of sequences due to ability of programs to find several tandem repeats in the same sequence.*

*** INTREP results include both tandem and interspersed repeats.*

¹ T-REKS parameters K=10; $P^*_{sim}=0.65$

² Marcotte et al., 1999; ³ Sokol and Benson, 2007; ⁴ Newman and Cooper, 2007