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

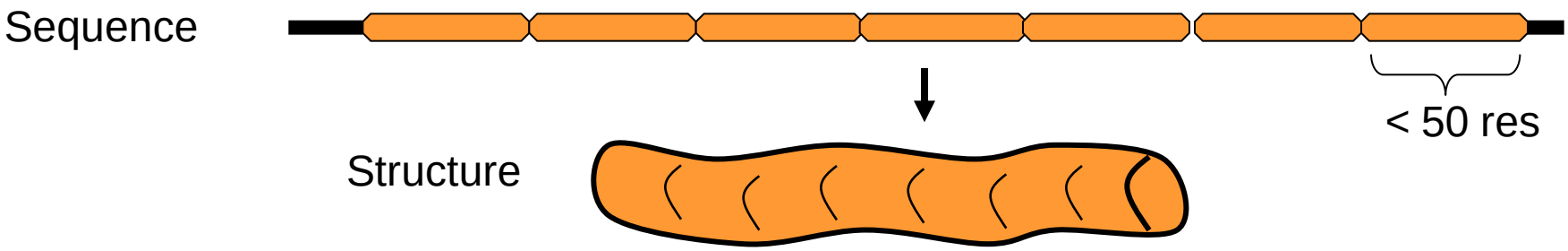
- ✓ **Structural prediction**
- ✓ **Analysis and Classification of the known 3D protein structures**
- ✓ **Identification of protein repeats**
- ✓ Experimental tests
- ✓ Evolution of proteins with repeats
- ✓ Applications in medicine, material science and nanotechnologies

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



only ~ 2% of known 3D structures

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

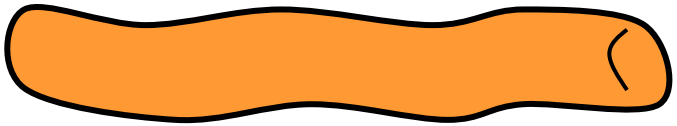
HYBRID APPROACHES TO OBTAIN 3D STRUCTURE

Bioinformatics analysis, structural prediction, molecular modeling

3D structure

It is possible to get a reliable 3D structural model without X-ray, NMR data and the known structures of homologous proteins!

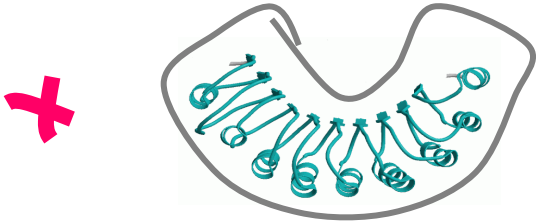
Incomplete experimental structural data (EM, CD, etc)



Prédiction et modélisation de protéines à séquences répétitives

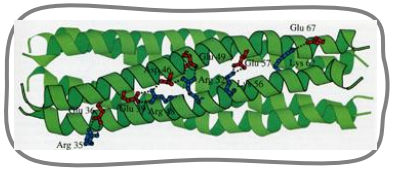
Leucine-rich repeat proteins

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

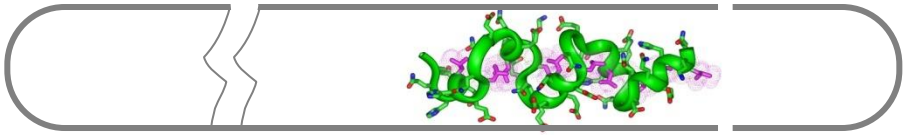


Filamentous Hemagglutinin Adhesin of Bordetella pertussis (56 nm long)

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

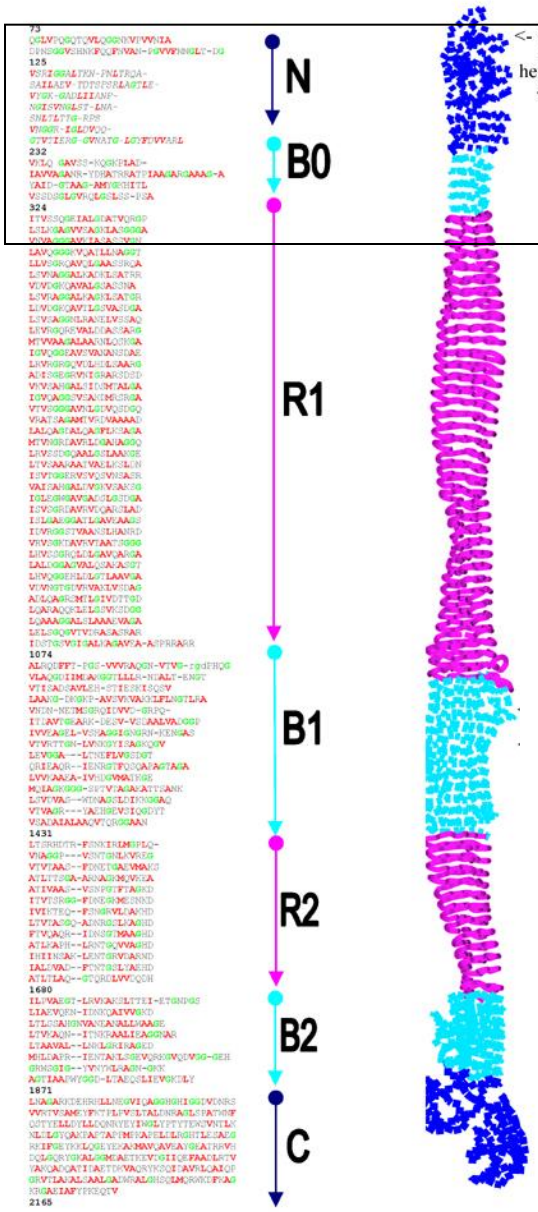
Human involucrin (46 nm long)

Kajava (2000) FEBS Lett. 473, 127

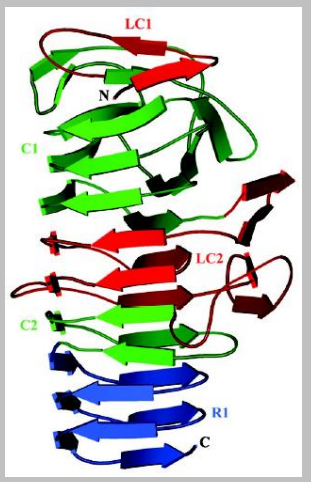


Rpn1 and Rpn2 subunits of eukaryotic proteasome

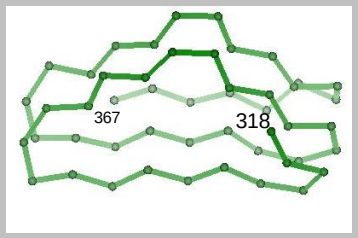
Kajava (2002) J.Biol.Chem. 277, 49791



<- Carbohydrate-
-dependent
hemagglutination
activity site



Model
(Kajava et al., 2001)

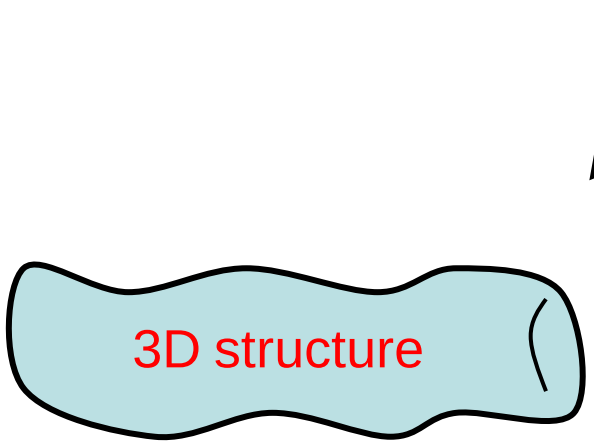


Crystal structure
(Clintin et al., 2004

RMS deviation of C_α atoms is 1.1 Å

IS A PROTEIN WITH REPEATS

STRUCTURED OR UNSTRUCTURED?



Structured

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

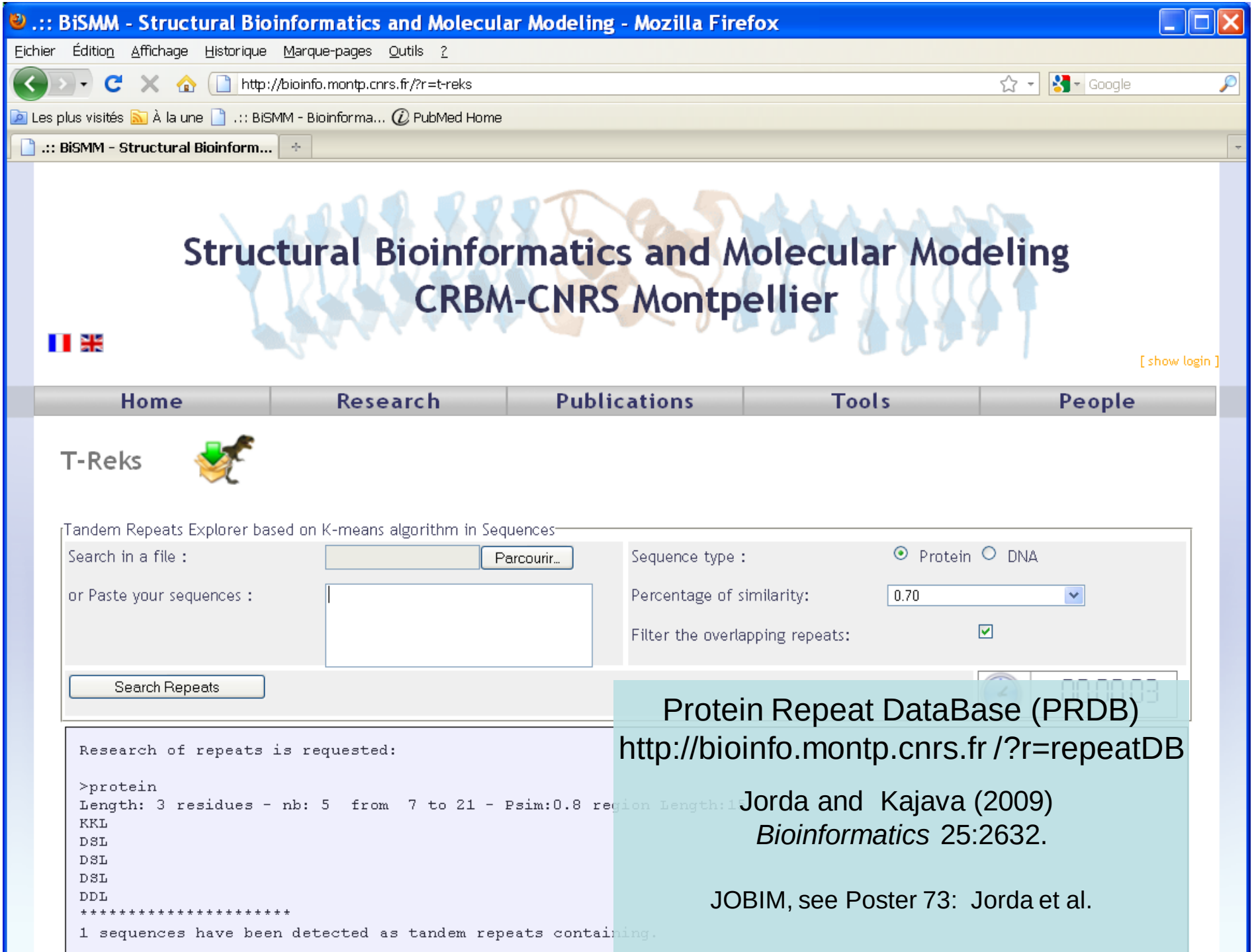
XSTREAM

Newman and Cooper, 2007

Estimation of edit distance between strings

TRED

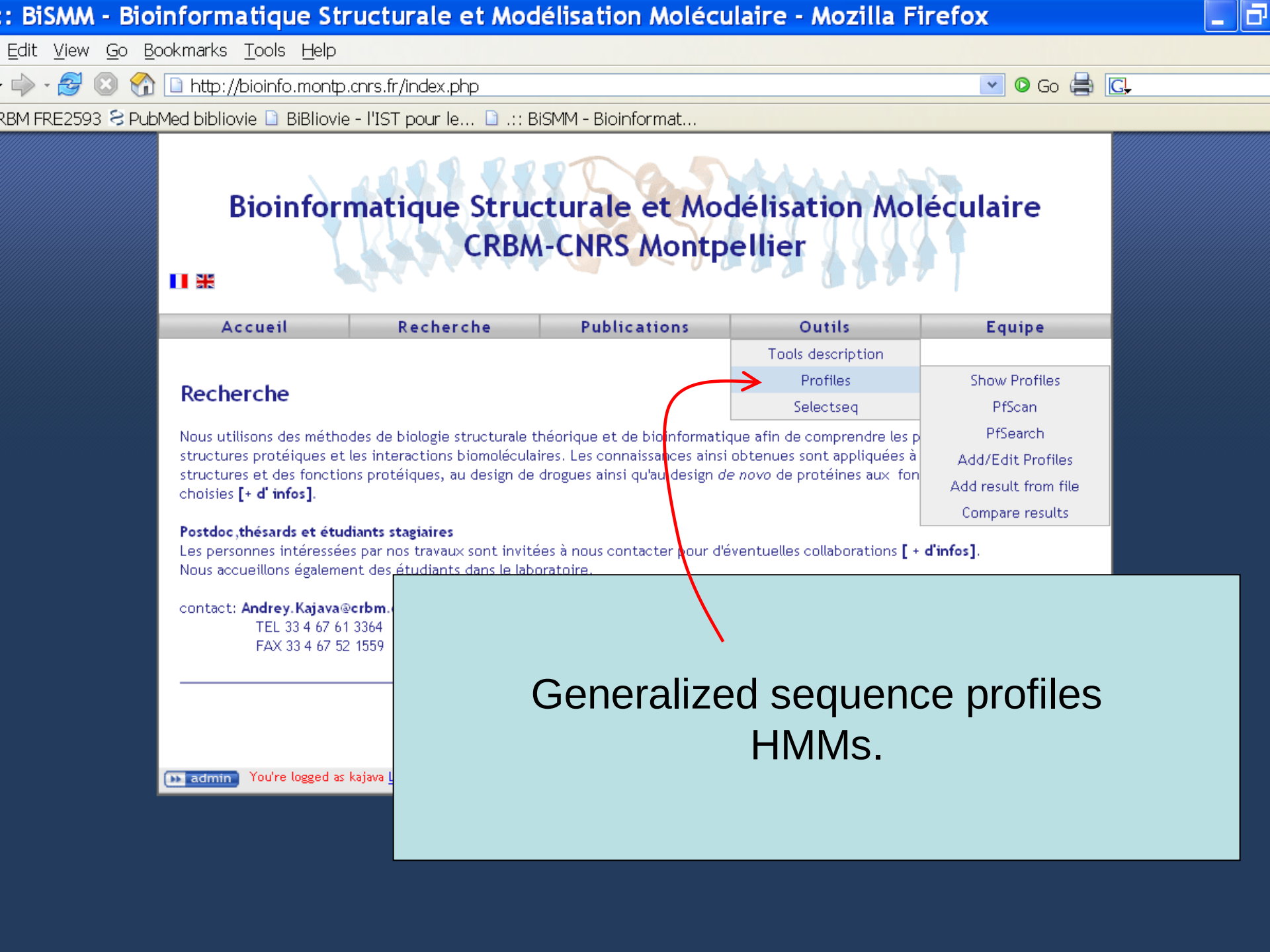
Sokol et al. 2007



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.


:: BiSMM - Bioinformatique Structurale et Modélisation Moléculaire - Mozilla Firefox
Edit View Go Bookmarks Tools Help
http://bioinfo.montp.cnrs.fr/index.php
Go
RBM FRE2593 PubMed bibliovie BiBliovie - l'IST pour le... .:: BiSMM - Bioinformat...
Bioinformatique Structurale et Modélisation Moléculaire
CRBM-CNRS Montpellier
Accueil
Recherche
Publications
Outils
Equipe
Tools description
Profiles
Selectseq
Show Profiles
PfScan
PfSearch
Add/Edit Profiles
Add result from file
Compare results
Recherche
Nous utilisons des méthodes de biologie structurale théorique et de bioinformatique afin de comprendre les p
structures protéiques et les interactions biomoléculaires. Les connaissances ainsi obtenues sont appliquées à
structures et des fonctions protéiques, au design de drogues ainsi qu'au design de novo de protéines aux fon
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.
TEL 33 4 67 61 3364
FAX 33 4 67 52 1559
admin
You're logged as kajava
Generalized sequence profiles
HMMs.

Structural-Functional Annotation of Genomes

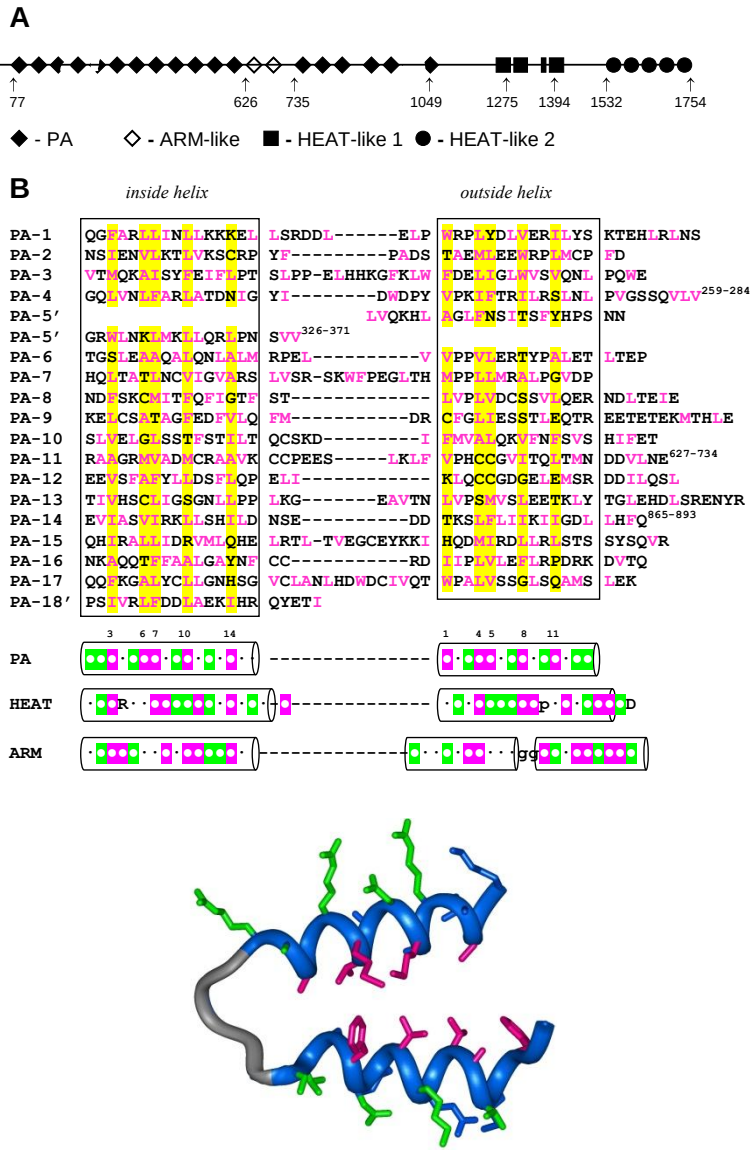
Proteins with aperiodic sequences and globular domains

Proteins with tandem repeats

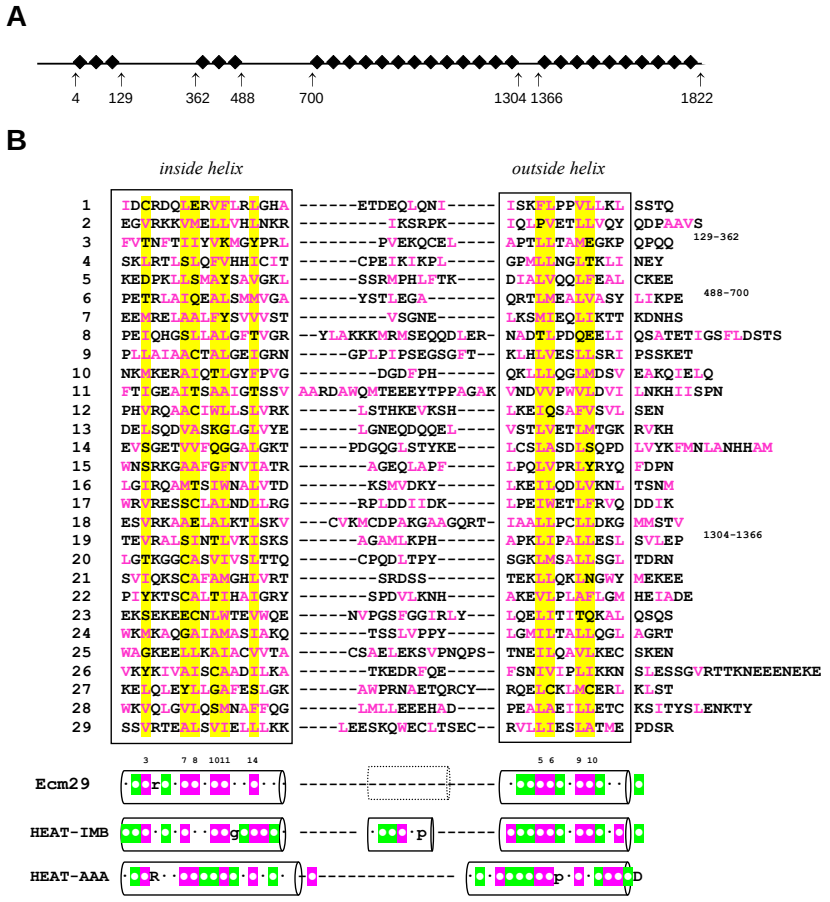
☺	Bioinformatcs tools	☹
☹	*Ab initio* structural prediction	☺

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

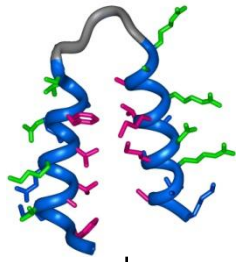
Nuclear proteasome activator Blm10 (PA200)

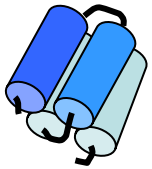


Ecm29

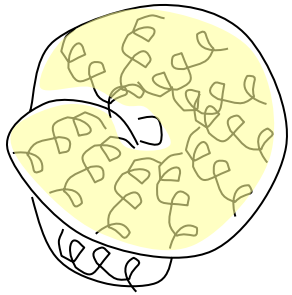


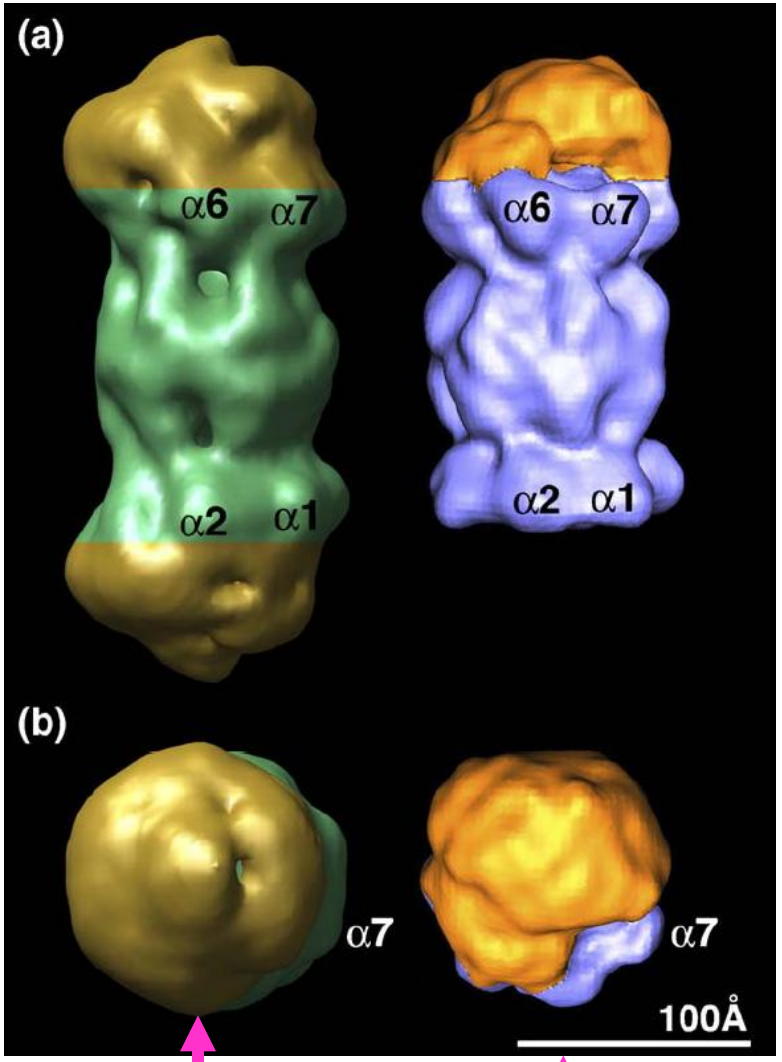
Kajava, A.V., Gorbea, C., Ortega, J., Rechsteiner M. and A. C. Steven (2004) *J. Struct. Biol.* 146,425





Twisted and
Curved





Blm10–20 S Proteasome

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

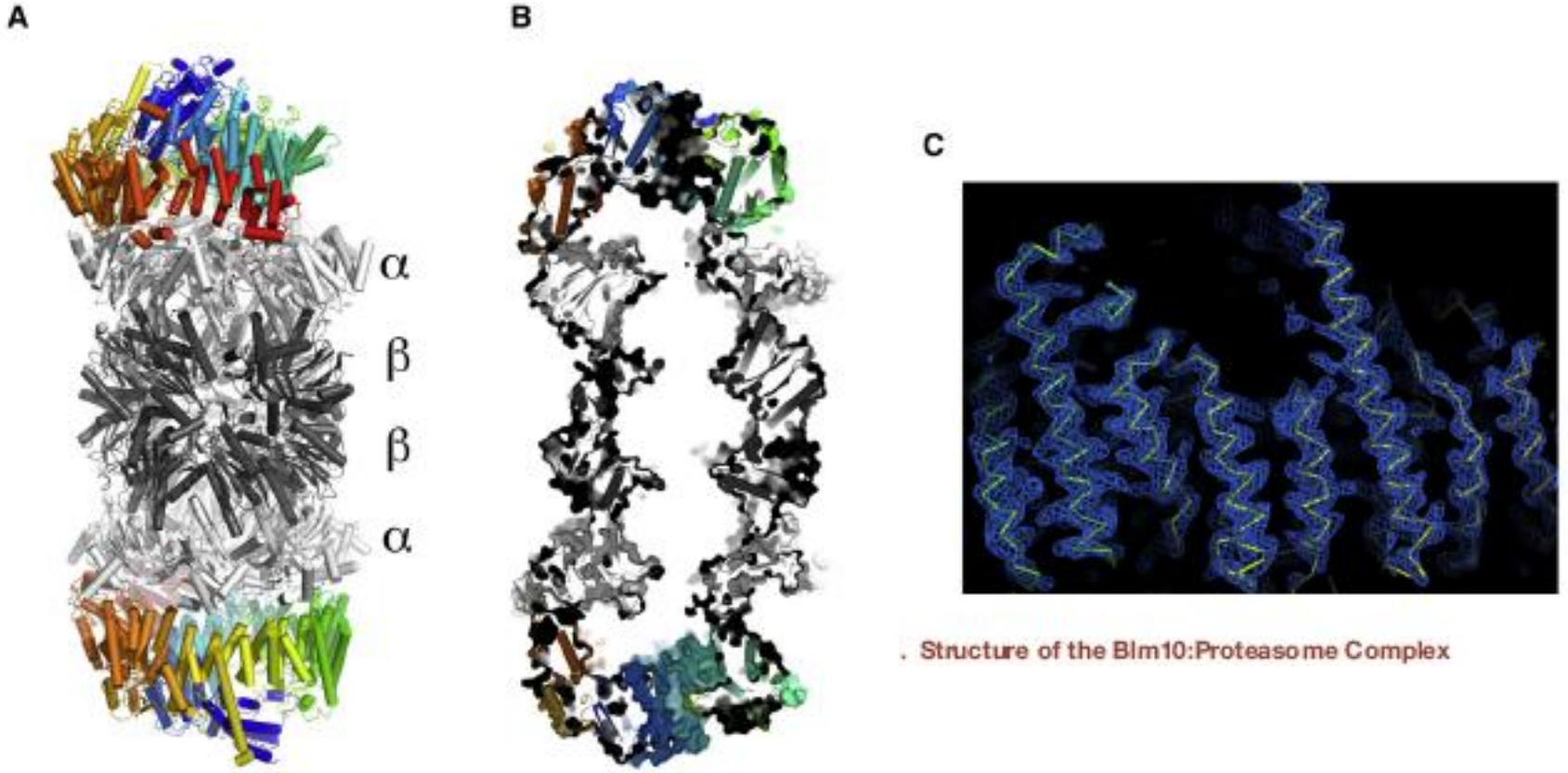
PA200 activator- 20S proteasome

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



. 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

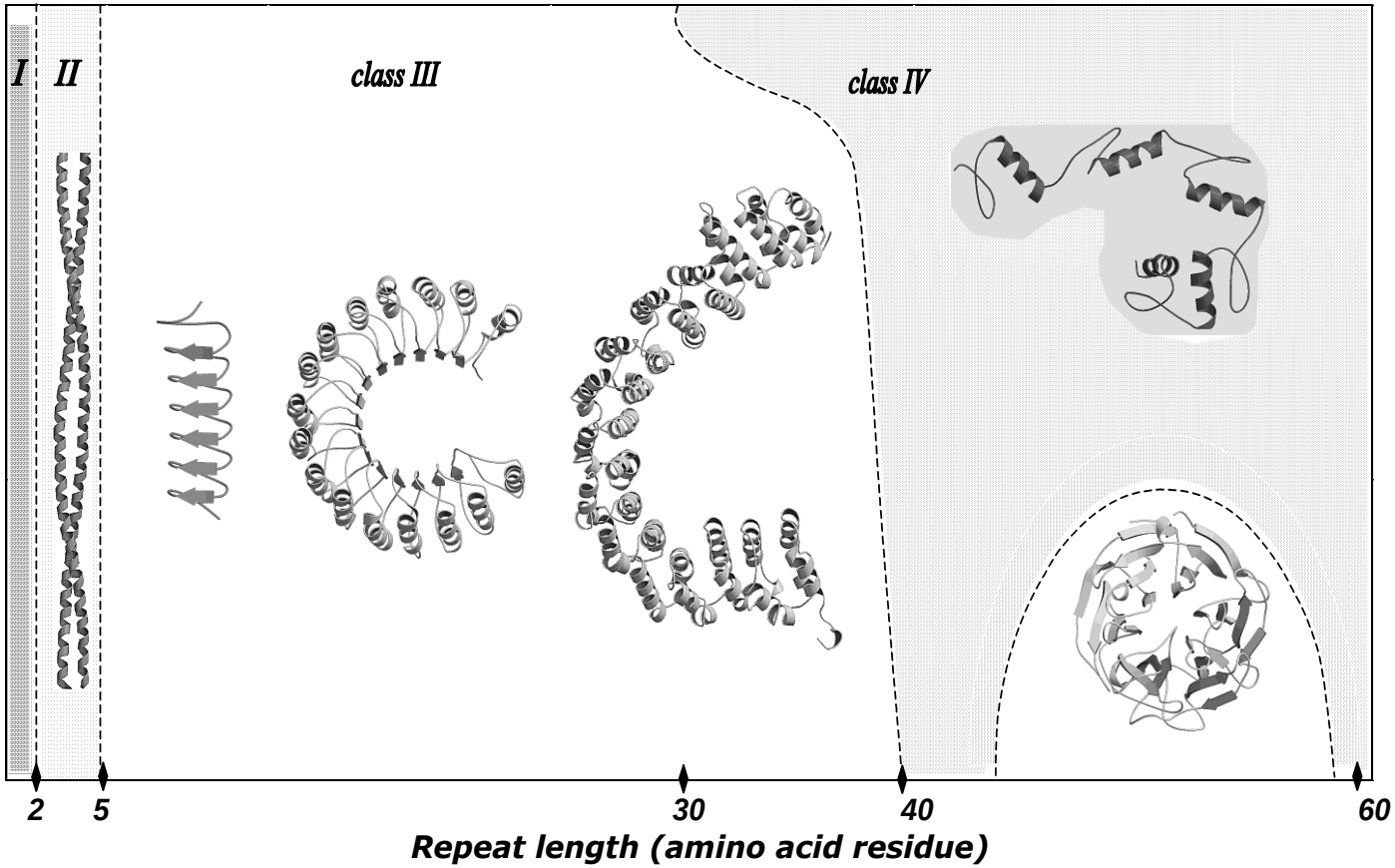
James Hurley, NIDDK, NIH, USA

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

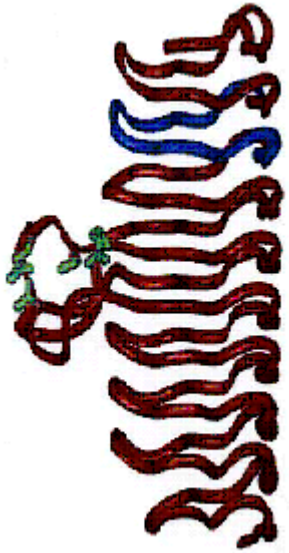
James Hurley, NIDDK, NIH, USA

Aitor Hierro CIC bioGUNE, Bilbao, Spain

Distinguishing between structural and functional residue conservations

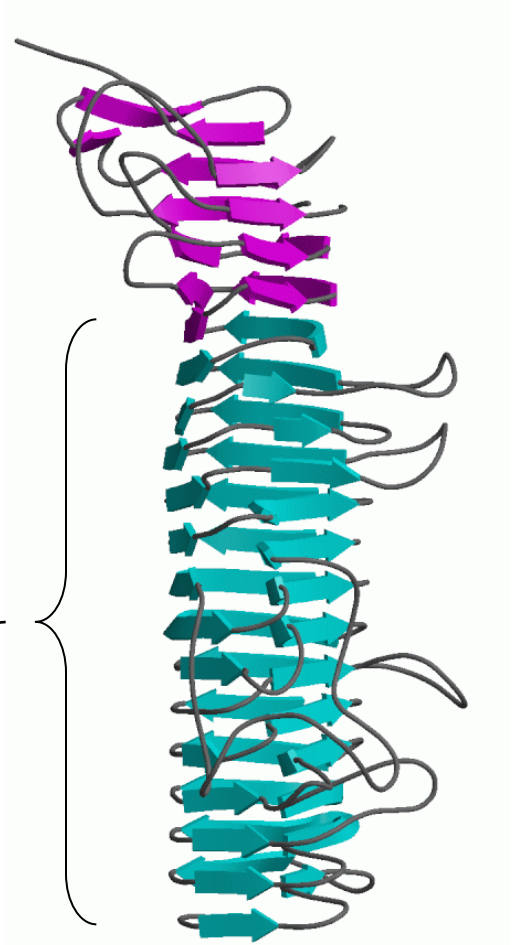
```
protein_human   VKVSAHGALSIDSMTALGAIGVQAGGSVSAKDMRSRGAVTVSG.GAVN
protein_rat     VHLNAHGALTIKTMYSGNHISVQAGSHVSAREMHQSAFVTVHCAGSVN
protein_yeast   VKVSFQSSLSIDSMTALGAIGVVSSGSVDAKDMRSRGAVWVSG.GAVK

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

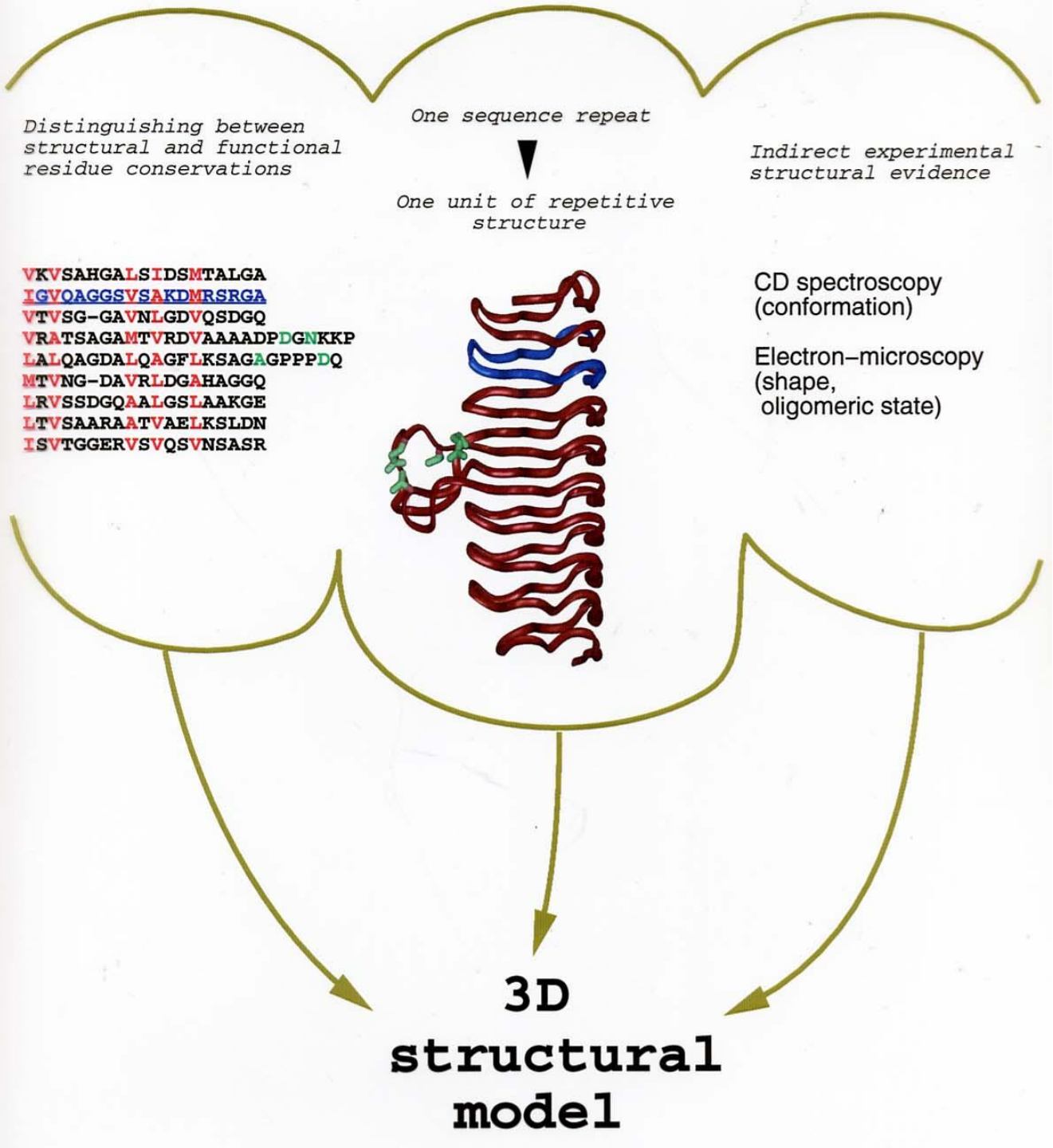
protein_human



Pertactin from *Bordetella pertussis*

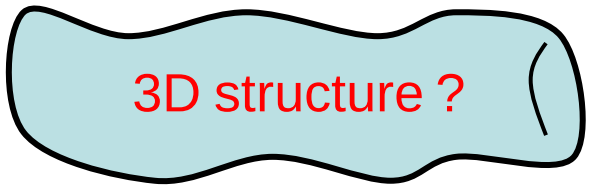
GILLENPAAELQFRNGSVTSSGQLSDDGIRRFLG
TVTVKAGKLVADHATLANVGDTWDDDGI
ALYVAGEQAQASIADSTLQGAG
GVQIERGANVTVQRSAIVDG
GLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPA
AVSVLGASELTLDGGHITGGRAA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWY
GVDVSGSSVELAQSIVEAPELGA
AIRVGRGARVTVSGGSLSAPHGN
VIETGGARRFAPQAAPLSITLQAGAHAQGKA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




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)
3D structural model

IS A PROTEIN WITH REPEATS

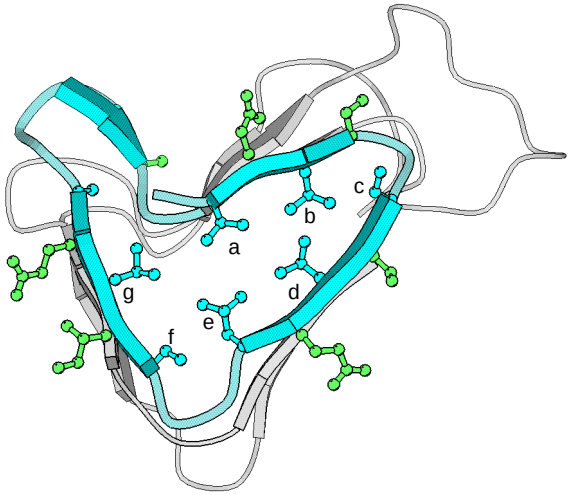
STRUCTURED OR UNSTRUCTURED?



Structured

Unstructured

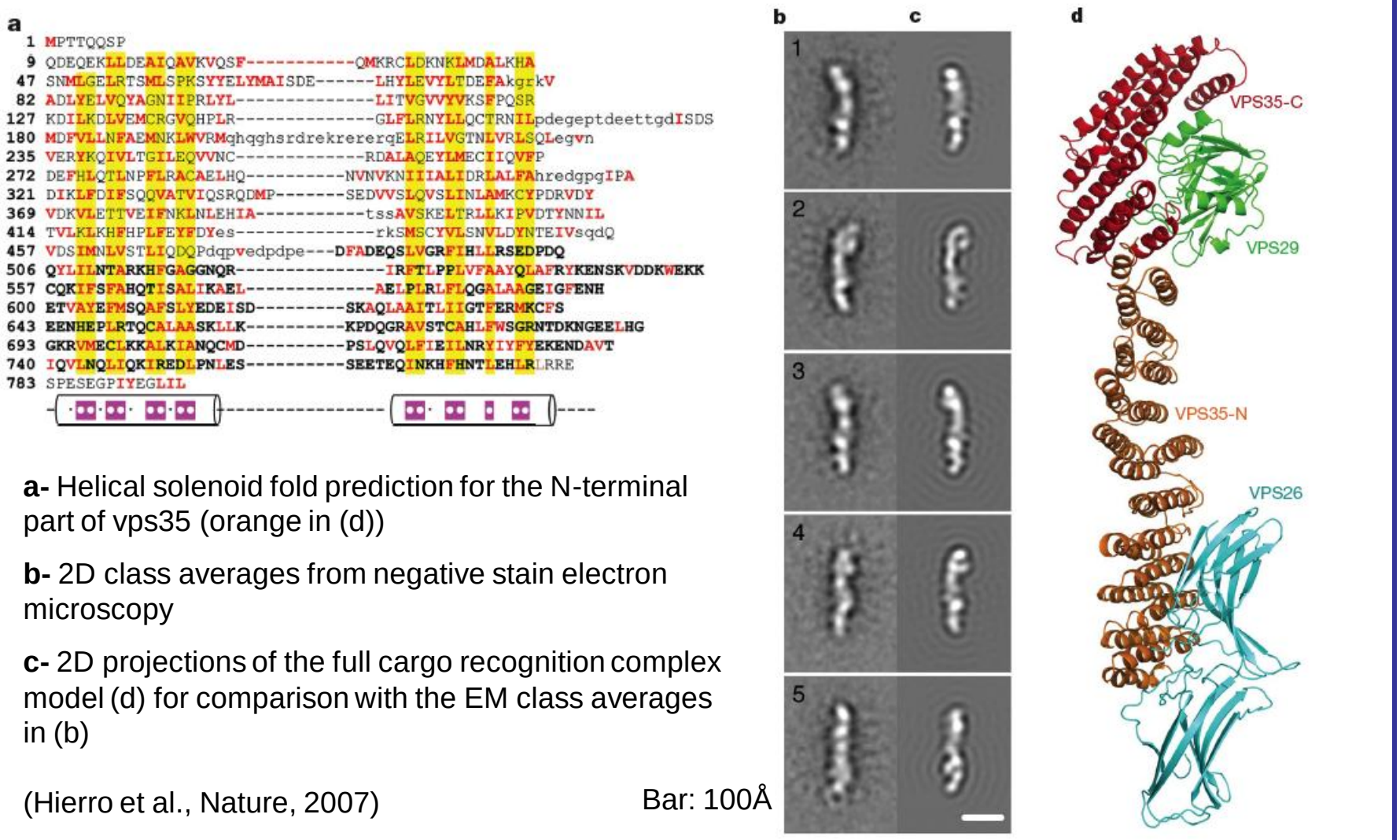
Pertactin from *Bordetella pertussis*




```
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)

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)

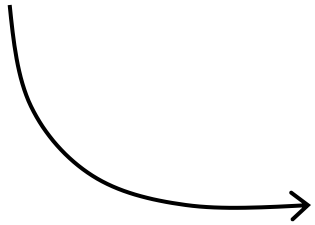
(Hierro et al., Nature, 2007)

Bar: 100Å

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

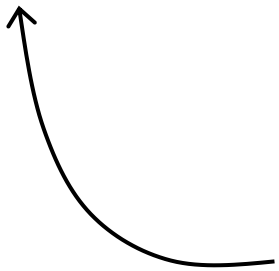
```
 ***   **         * *      ** * *   ***
GITLENPSS-------AAELQFRN-GSVTNSGQLSDGI
TITLKATSS--------AKLVADH-ASVANVGQTWDGI
```

```
 ***    **          * *        ** * *    ***
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
```

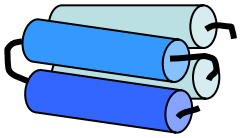


```
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY';
MA   /DISJOINT: DEFINITION=PROTECT; N1=1; N2=43;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE; R1=44.55; R2=-0.0035;
MA      R3=0.7386; R4=1.001; R5=0.208; TEXT='ZScore';
MA   /NORMALIZATION: MODE=2; FUNCTION=LINEAR; R1=0.0; R2=0.1;
MA    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

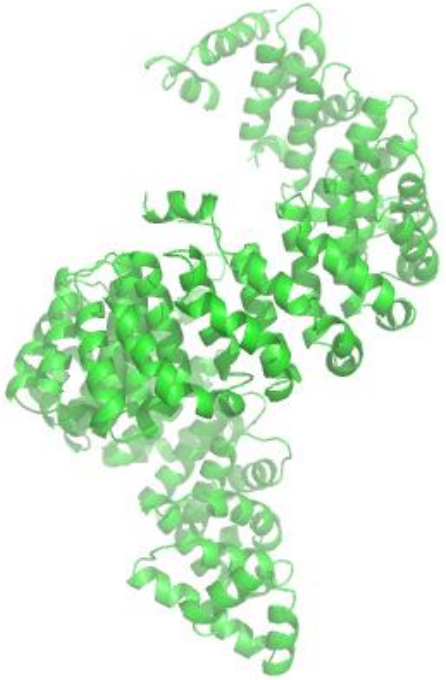
Sequence profile search

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

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

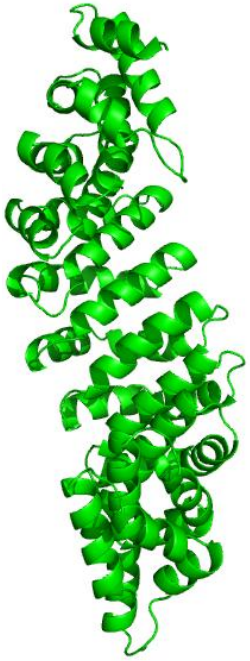


?



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)

ANDREI LUPAS AND WOLFGANG BAUMEISTER

Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany.
Email: lupas@vms.biochem.mpg.de

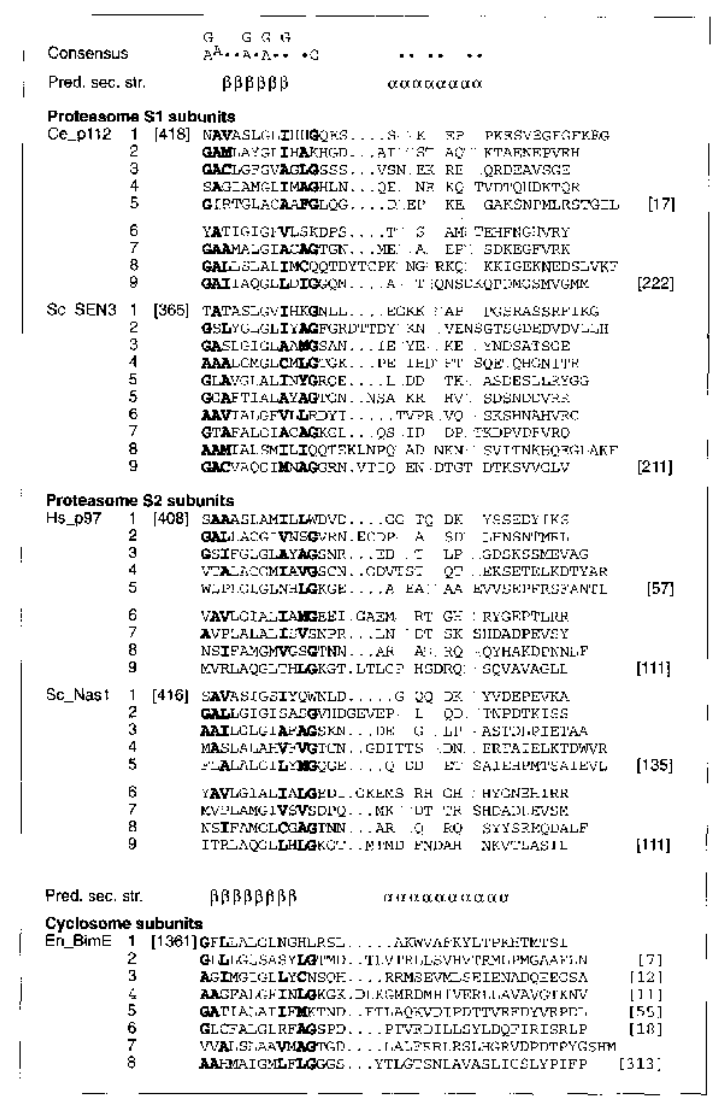
KAY HOFMANN

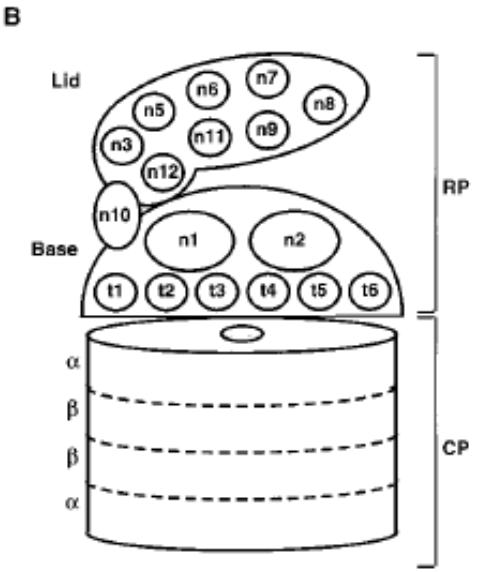
Swiss Institute for Experimental Cancer Research, Chemin des Boveresses 155, CH-1066 Epalinges s/Lausanne, Switzerland.

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 peptides[2].

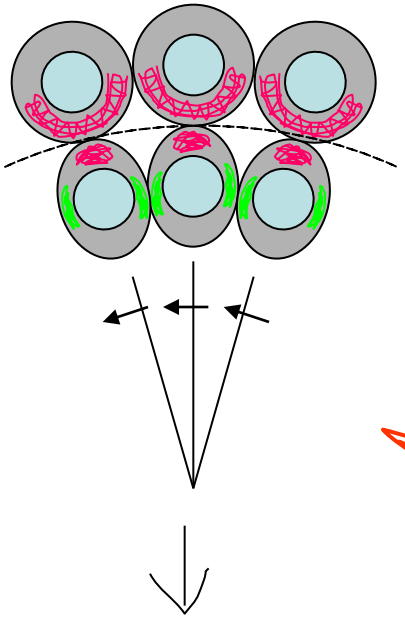
The caps consist of approximately 20 different subunits between 25 and 110 kDa[3,4]. The two largest subunits are S1 (also called p112 in mammals and Sen3 in yeast) and S2 (p97 in mammals, Nas1 in yeast). Their functions are unclear, as is the case for most of the subunits of the 19S complex. Sen3, the S1 subunit of yeast, was identified as a factor required for the degradation of Sen1, a protein involved in tRNA splicing[5]. TRAP-2 (55.1), 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 proteins[8]. Although S1 and S2 sequences have diverged significantly (30–45% sequence identity), human p97 can suppress the phenotype of Nas1 mutants[8], 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 subunits[8] and an alignment using MACAW[9] shows them to be related in a region of about 400



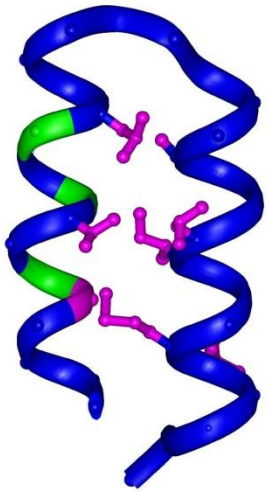


Molecular modelling of Rpn1 and Rpn2 subunits of eukaryotic proteasome

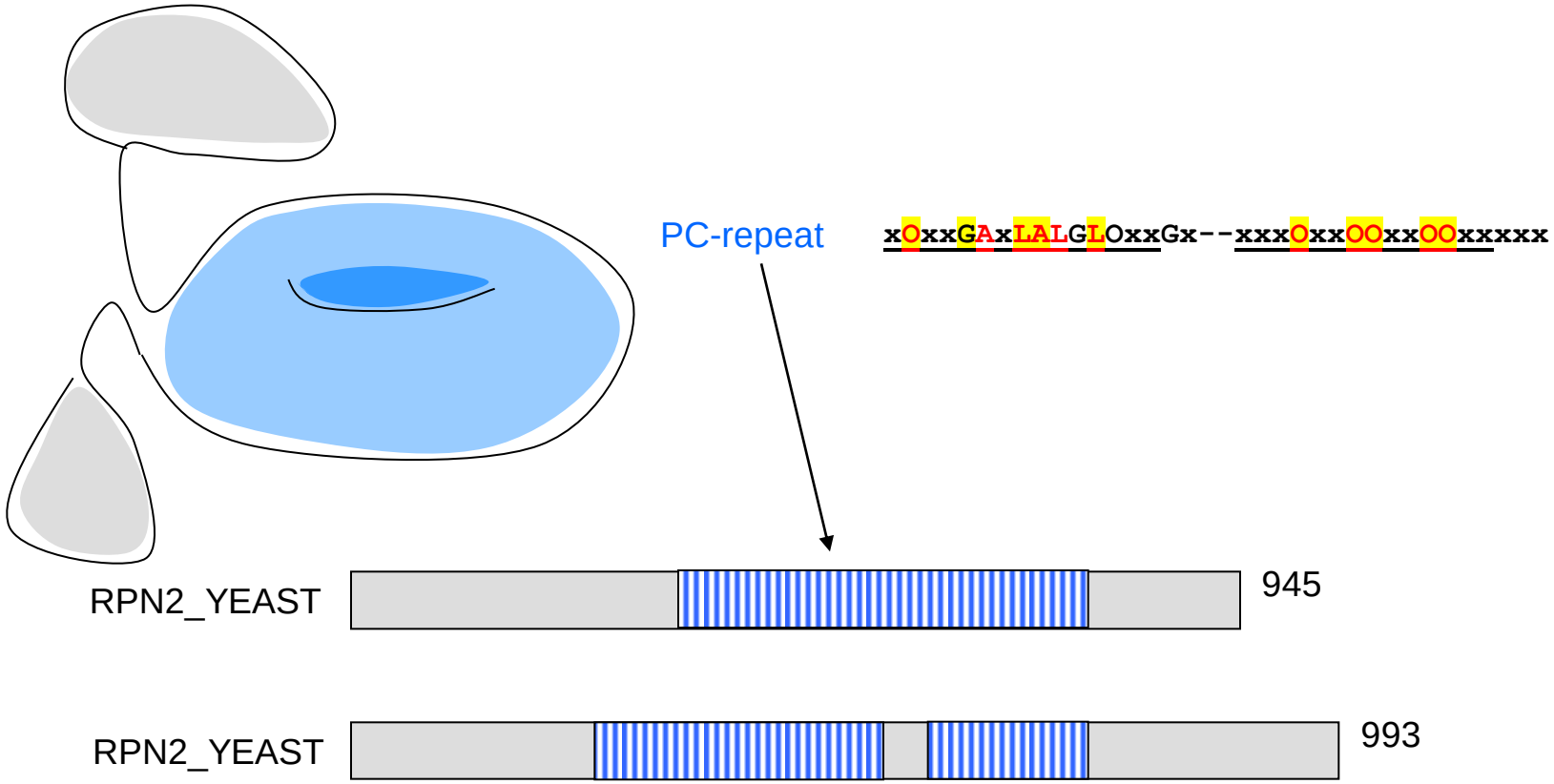


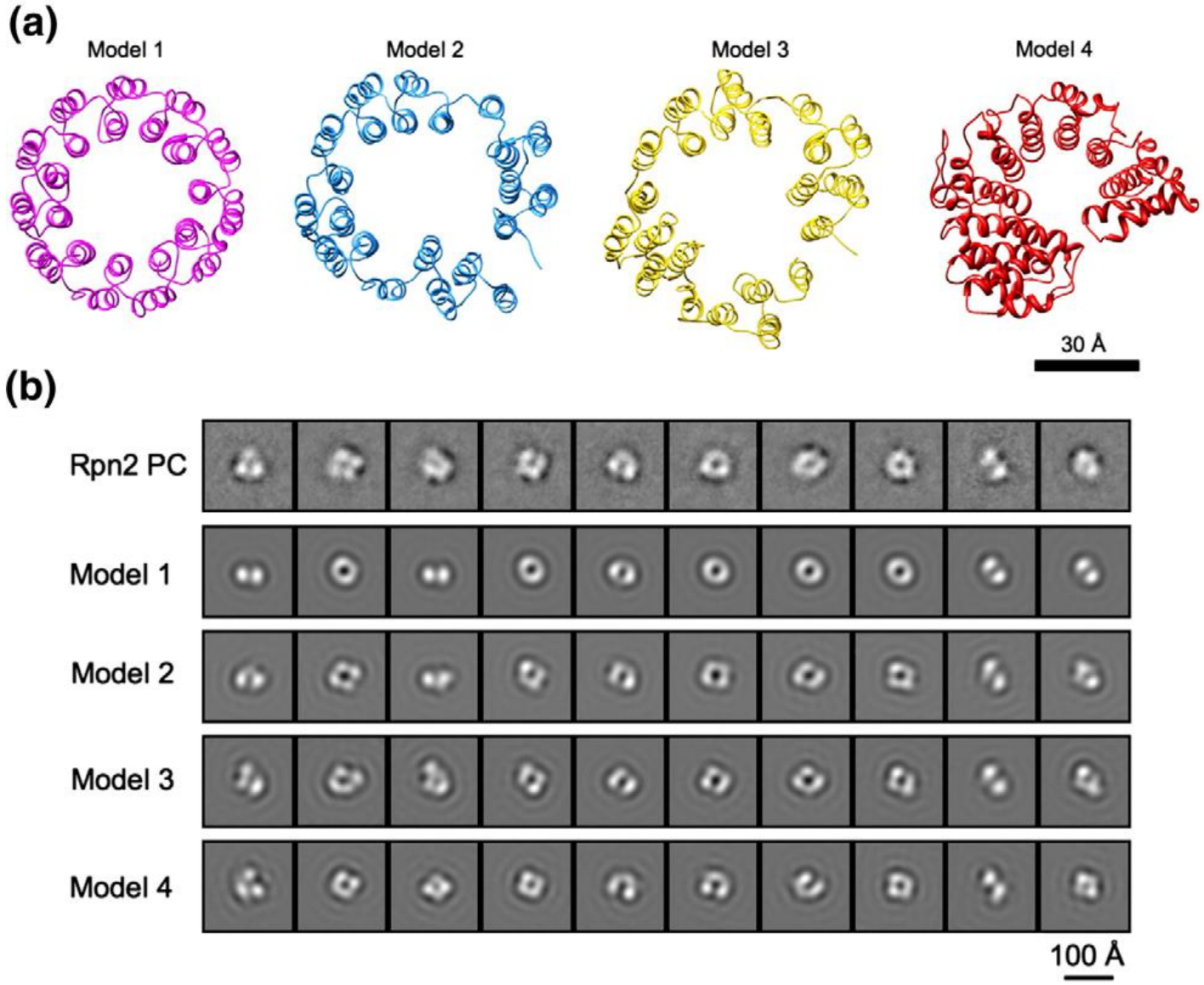
RPN2_YEAST (residues 309-712)

```
                                                IDIGLLNKSKSSLDGK
1*  FSLFHTAVSVANGFMHAGTTDNSFIKANLPWLGKA
2*  QNWAKFTATASLGVIHKGN--
                                          LLEGKKVMAPYLPGSRAS
1   SRFIKGGSLYGLGLIYAGF--GRDTTDYLKNIIVENSGTSGDED
2   VDVLLHGASLGIGLAAMGSA-NIEVYEALKEVLYND
3   SATSGEAAALGMGLCMLGTGKPEAIHDMFTYSQETQ
4   HGNITRGLAVGLALINYGR--QELADDLITKMLASD
5   ESLLRYGGAFTIALAYAGTGNNSAVKRLLHVAVSDS
6   NDDVRRAAVIALGFVLLRD--YTTVPRIVQLLSKSH
7   NAHVRCGTAFALGIACAGKG-LQSAIDVLDPLTKDP
8   VDFVRQAAMIALSMILIQQ--TEKLNPQVADINKNFLSVITNKHQ
9   EGLAKFGACVAQGIMNAGGRN
```



Identification of new Rpn1 and Rpn2 repeats

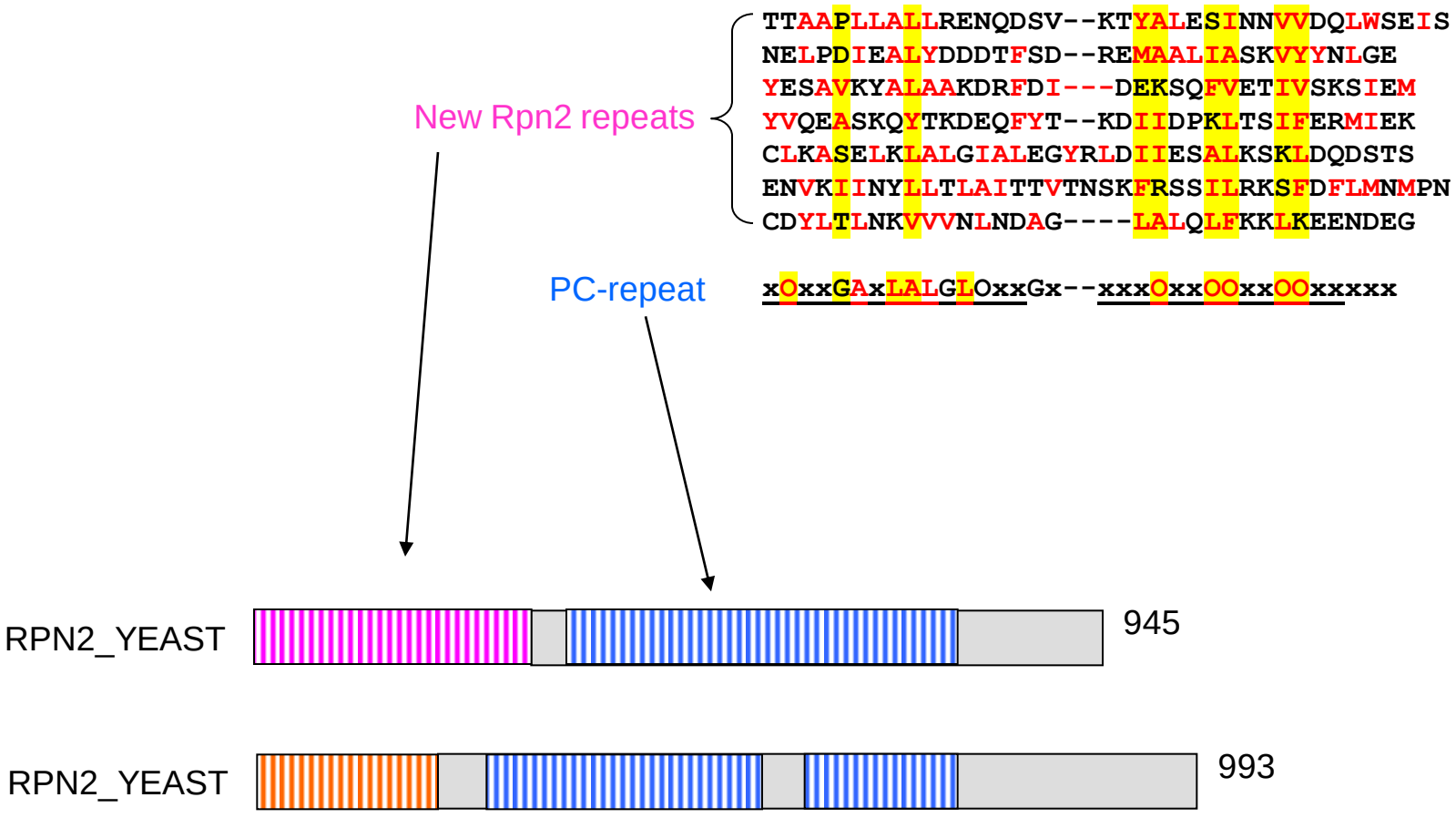


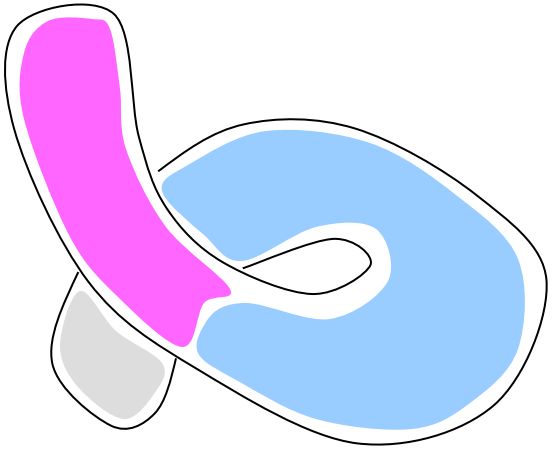


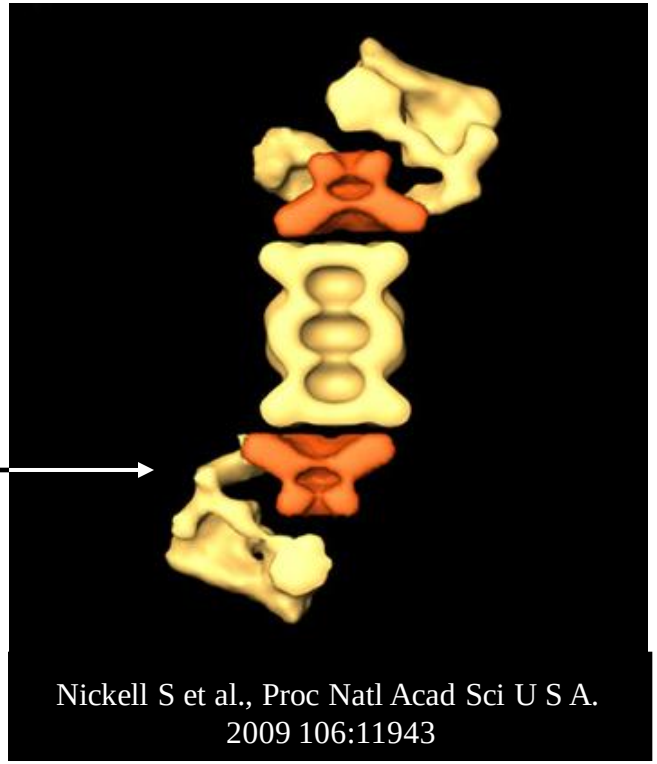
Alpha-helix from CD spectroscopy

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

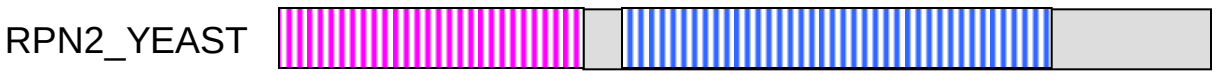
Identification of new Rpn1 and Rpn2 repeats



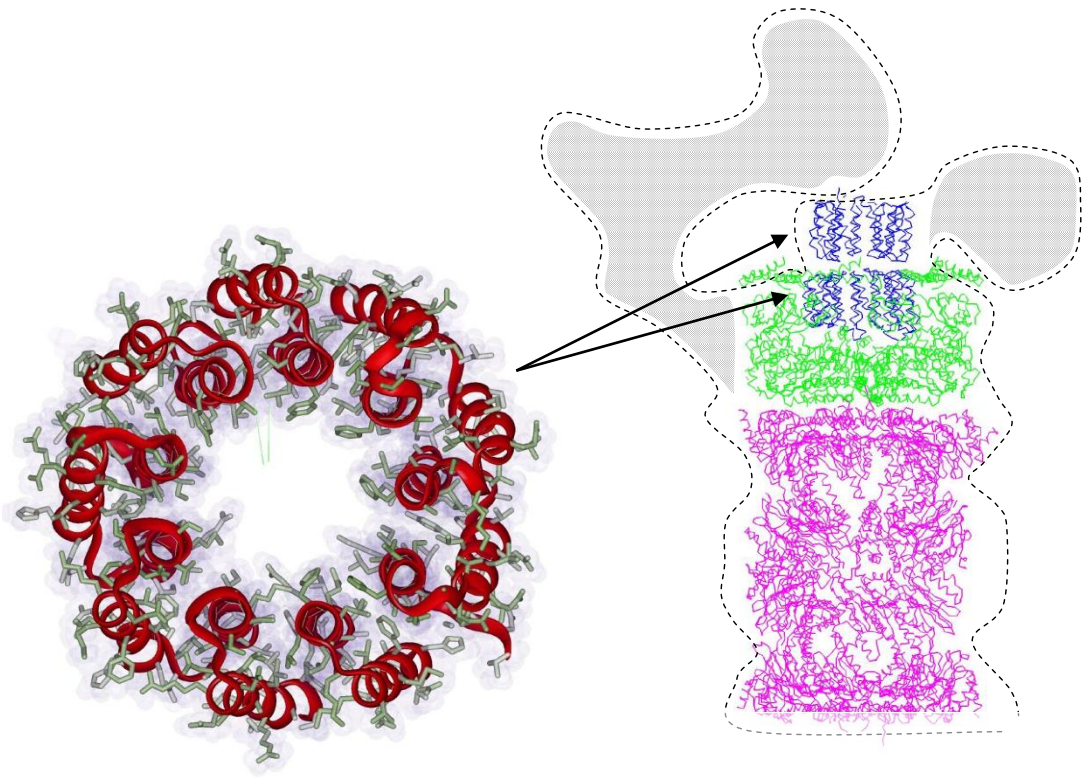




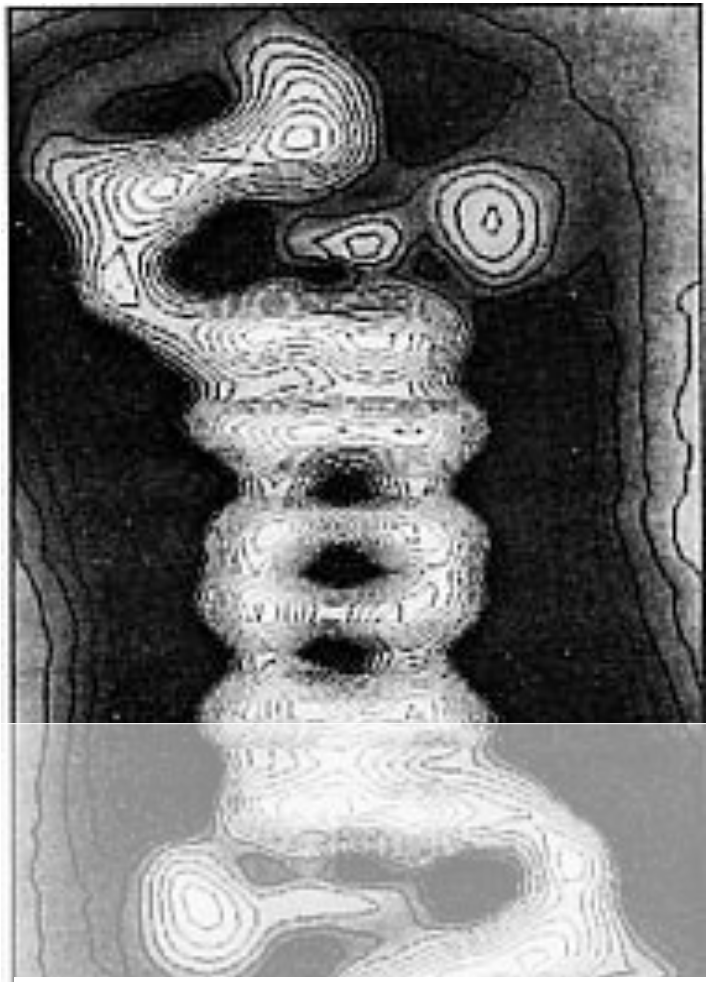
Nickell S et al., Proc Natl Acad Sci U S A. 2009 106:11943



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)

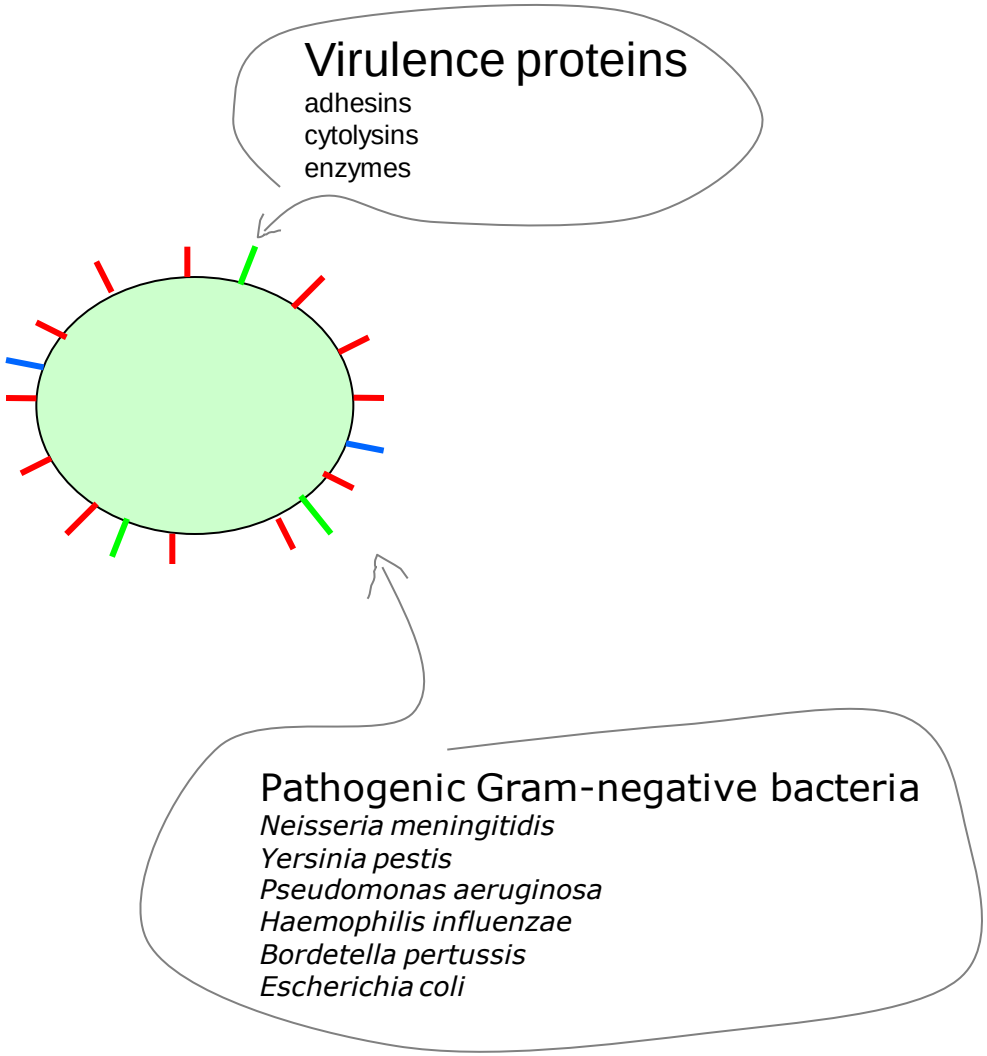


Table 1

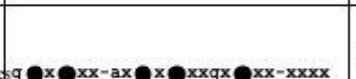
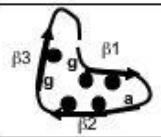
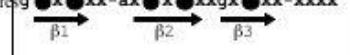
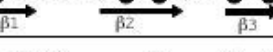
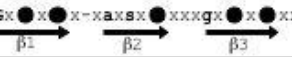
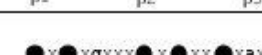
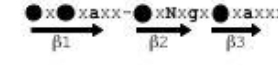
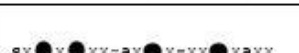
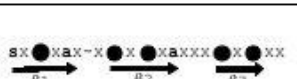
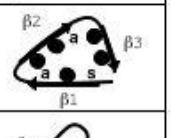
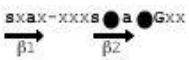
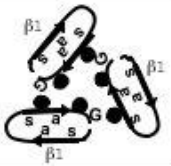
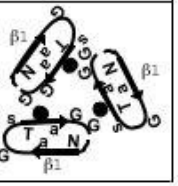
№	Representative protein	Repeat length	Consensus sequence of repeat	Coil of β-solenoid
L1	Serum resistance protein brkA (*B. pertussis*)	22-26 res		
L2	Slr1753 protein (*Synechocystis sp*)	23-28 res		
L3	AGRL_3085 protein (*Agrobacterium tumefaciens*)	25-27 res		
T1	FHA protein *B. pertussis*	18-19 res		
T2	FHA protein *B. pertussis*	19-20 res		
T3	HBP protein *E.coli*	18-20 res		
T4	TibA protein *E.coli*	18-19 res		
T5	YapA protein *Y.pestis*	18-19 res		
T6	Hap protein *Haemophilus influenzae*	19-20 res		
T7	Hemagg.-hemolysin related protein *E.coli*	20-22 res		

Table 1 (continuation)

№	Representative protein	Repeat length	Consensus sequence of repeat	Coil of β-solenoid
T8	LSPA1 *Haemophilus* ducreyi	20-21 res		
O1	XadA protein *Xylella fastidiosa*	13-14 res		
O2	UspA2H protein *Moraxella catarrahalis*	15 res		

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

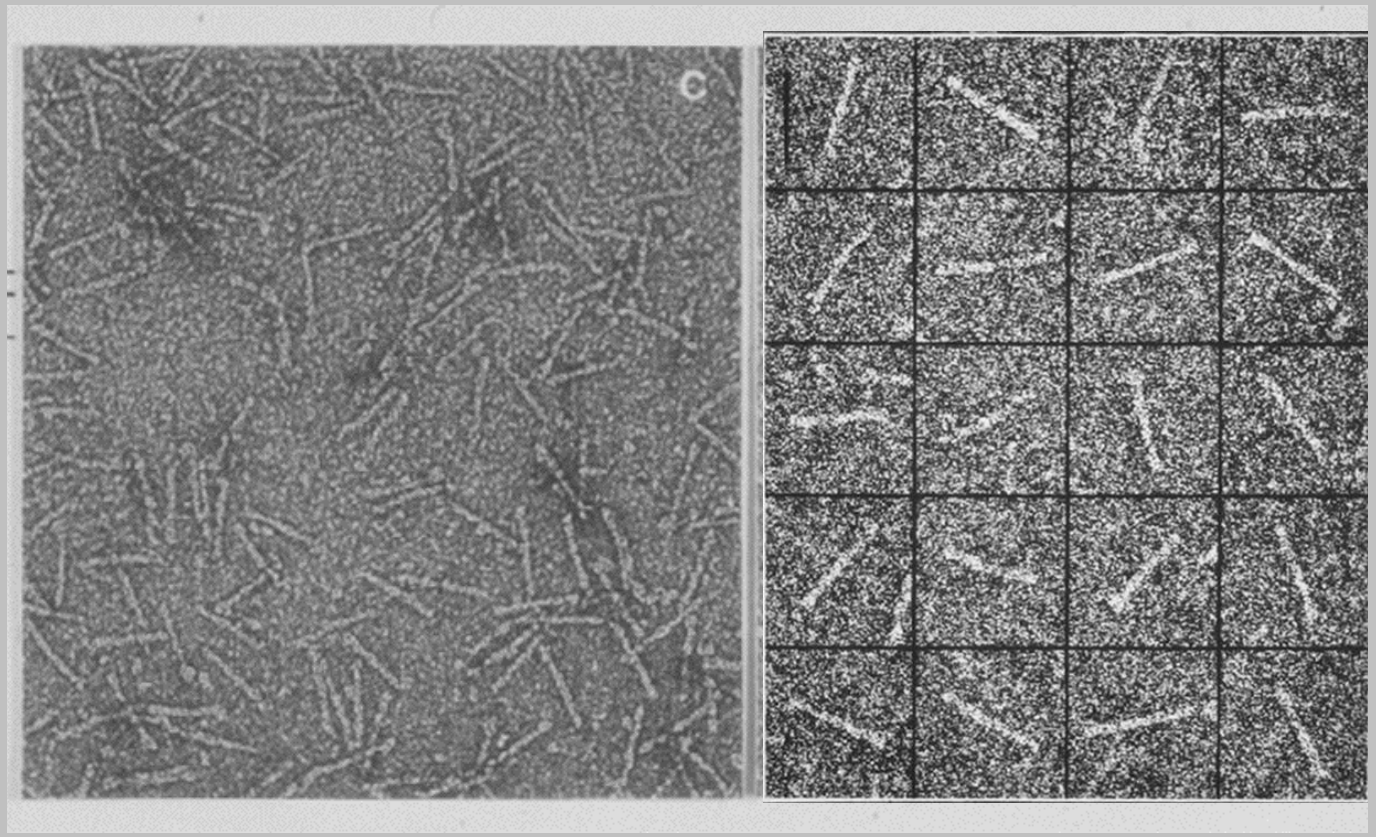
	P_{sim} =0.7-0.8	P_{sim} =0.8-0.9	P_{sim} =0.9-1
ID Ratio – VSL2	80.4%	88.6%	88.9%
ID Ratio – IUpred	56.0%	62.7%	67.2%
ID Ratio – FoldIndex	62.4%	68.6%	70.3%
ID Ratio – TopIDP	85.6%	88.8%	91.1%

Jorda et al. FEBS Journal 2010 (in press).

Filamentous Haemagglutinin adhesin

major virulance factor of *Bordatella pertussis*,

etiological agent of whooping cough



EM
negatively stained

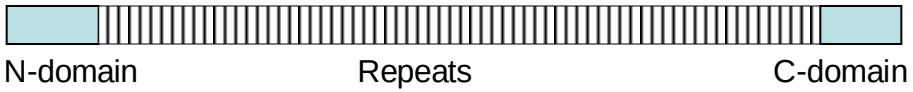
↓

Rod-like shape
50 x 4 nm

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

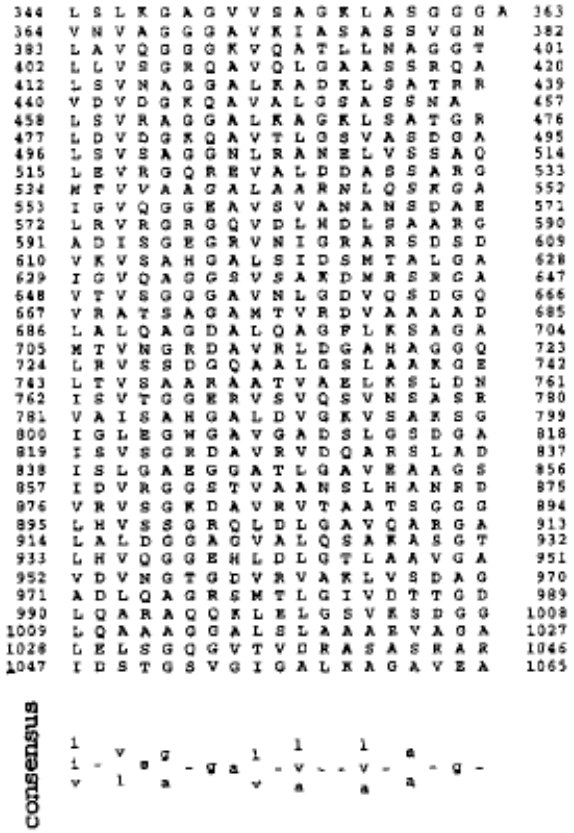
Filamentous Haemagglutinin adhesin (FHA) of *Bordetella pertussis*

2165 residues sequence



~50 nm

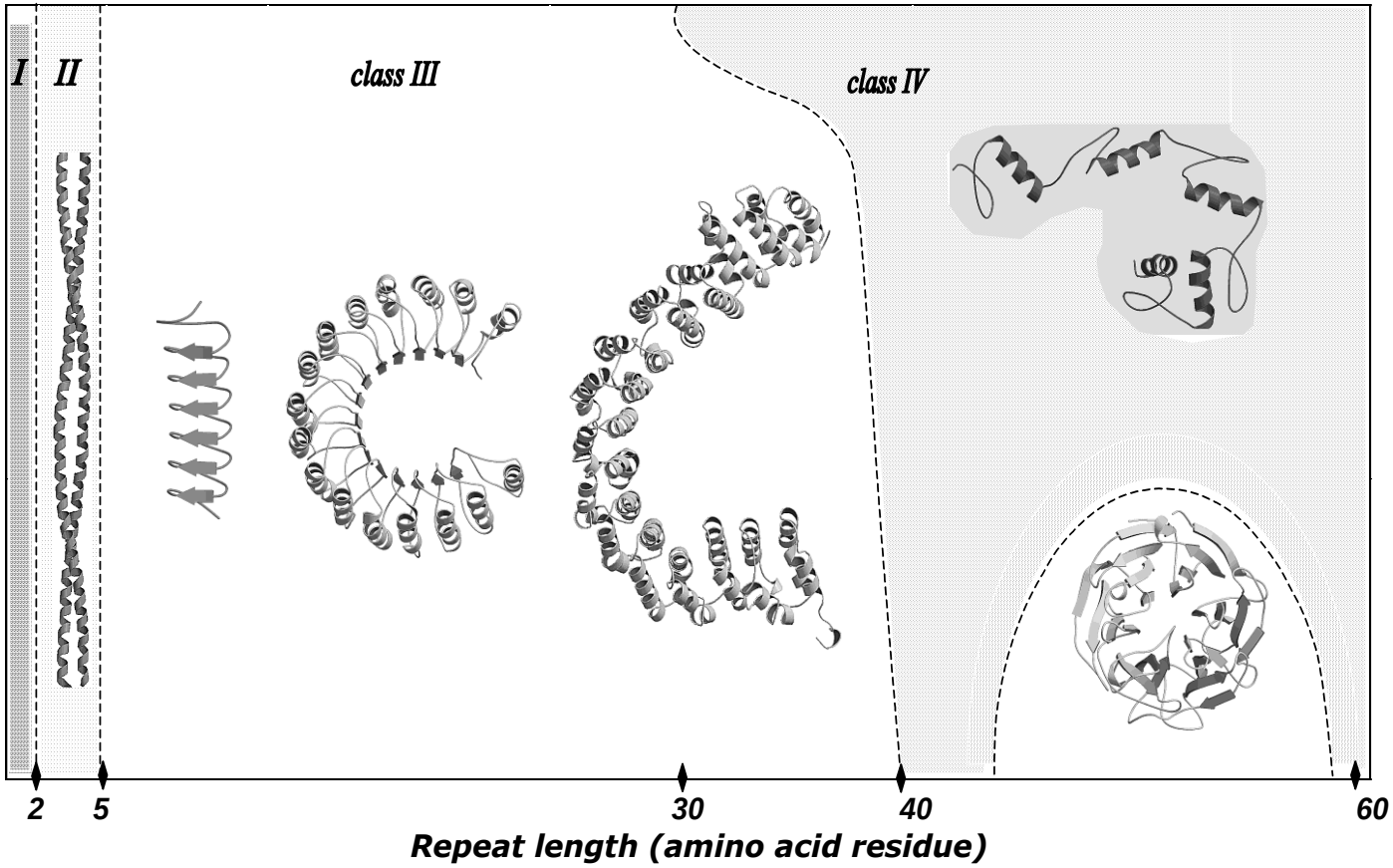
19-residue repeats



- Rod-like shape according to EM
- β-structural protein
  according to circular dichroism spectroscopy mesurements

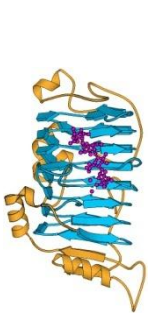
Makhov, Hannah, Brennan, Trus, Kocsis, Conway, Wingfield, Simon, Steven *J.Mol.Biol.* (1994) 241, 110

WHAT CAN REPEAT LENGTH TELL US ABOUT ITS STRUCTURE?

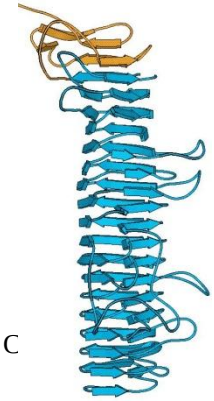


A.V. Kajava (2001) *J. Struct. Biol*. **134**:132

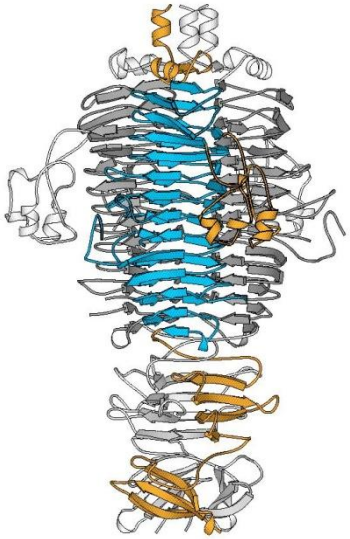
The known structures of β-solenoid proteins



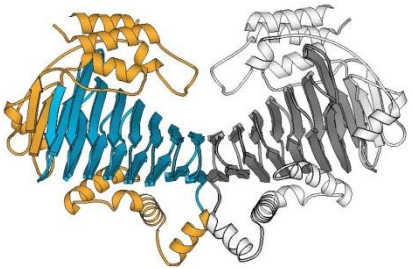
Pectate lyase C



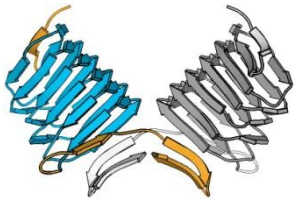
P.69 pertactin



Tailspike endorhamnosidase

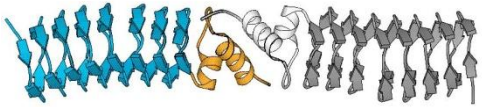


Stabilizer of iron transporter SufD

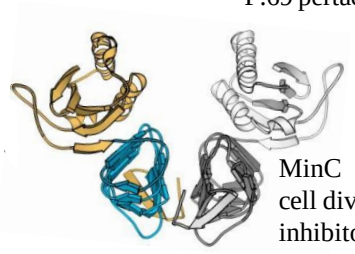


Cyclase-associated protein

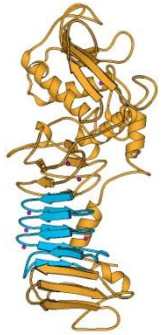
Antifreeze protein



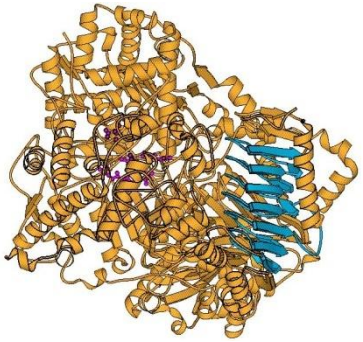
MfpA inhibitor of DNA gyrase



MinC cell division inhibitor

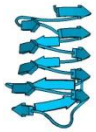


PrtC protease C

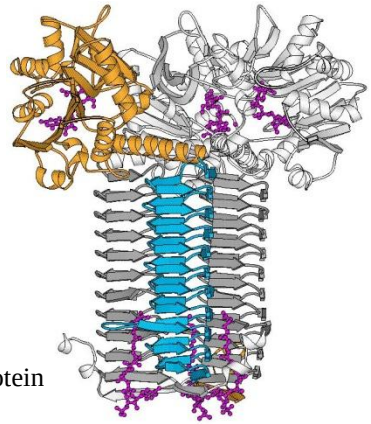


Glutamate synthase

YadA adhesin



Antifreeze protein

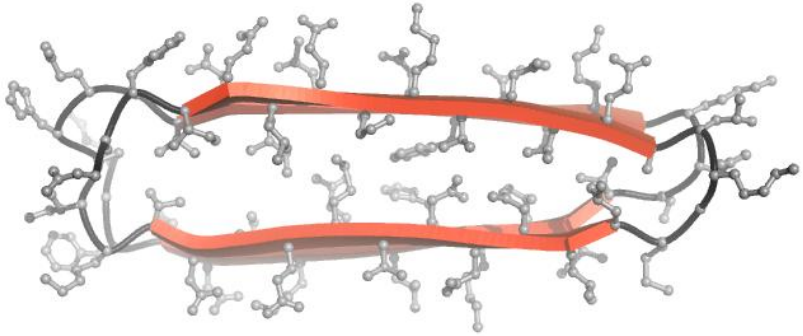


N-acetyl-glucosamine 1-phosphate uridyltransferase

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

Classification of beta-solenoids

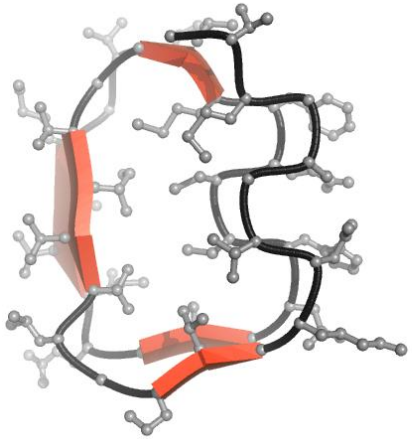
Cross-sectional shapes



O-type

T-type

R-type



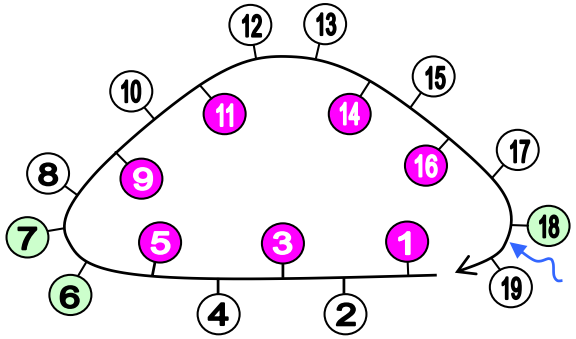
B-type

L-type

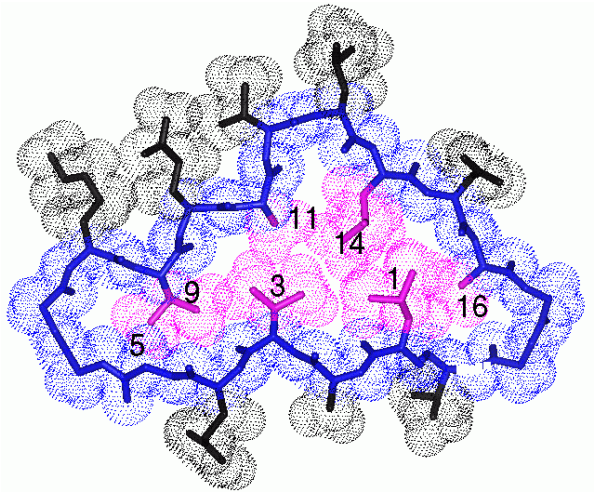
Kajava and Steven (2006) *Advances in Protein Chemistry*, 73, 55-96

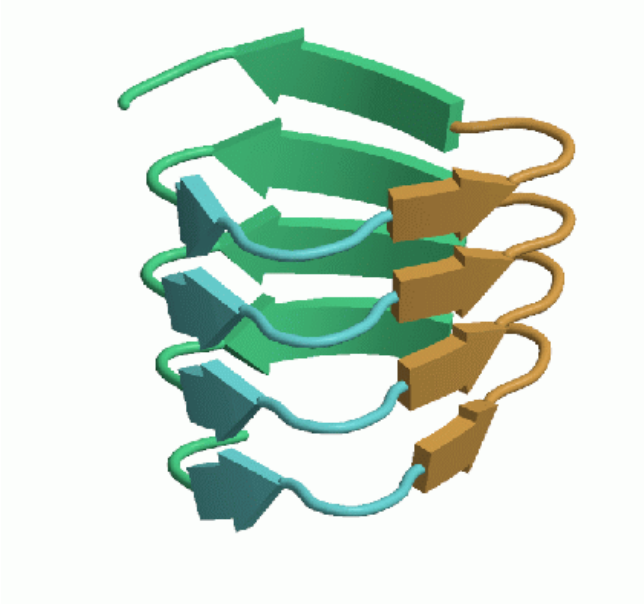
```
Repeat 1    V N V A G G G A V K I A S A S S V G - N
Repeat 2    L A V Q A G G K V Q A T L L N A G G - T
Repeat 3    L L V S A R Q S V Q L G A L S A R Q - A
Repeat 4    L S V N A G G A L K A D K L S A T G S R

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
```

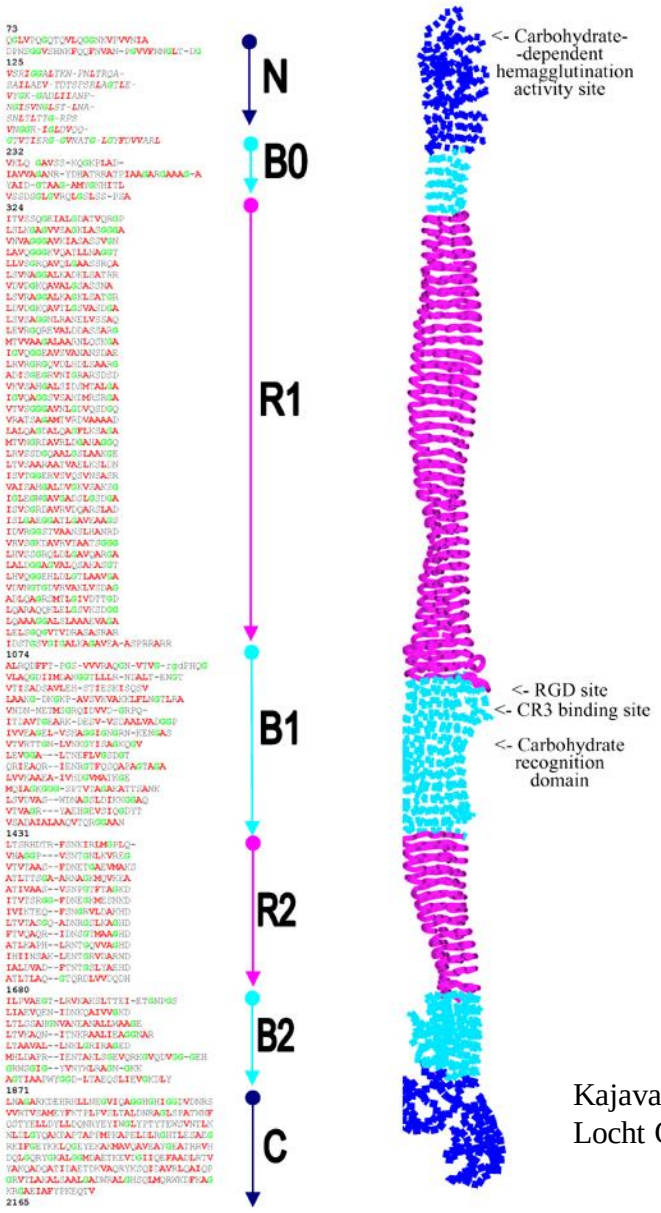
2D plot



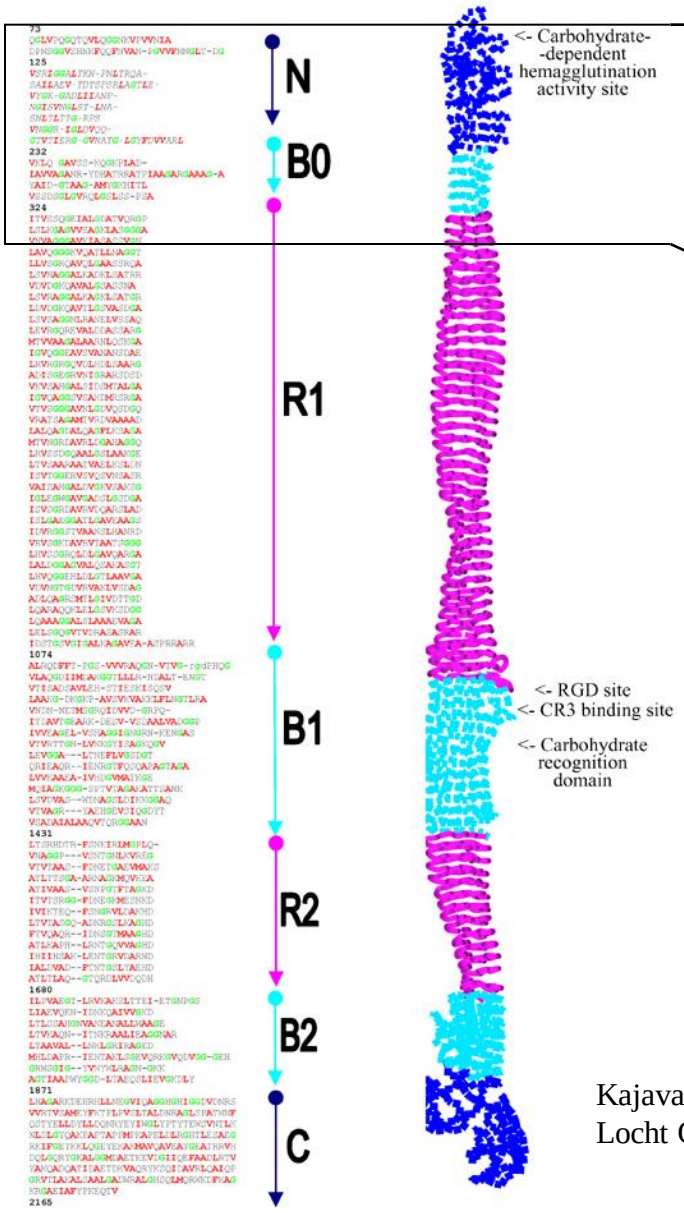
3D structure

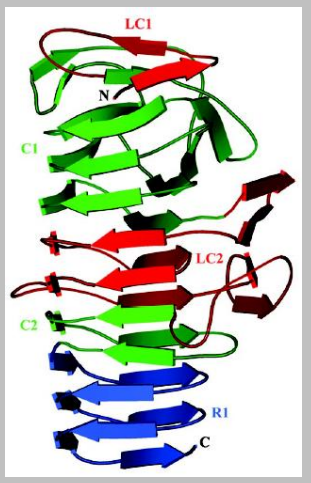






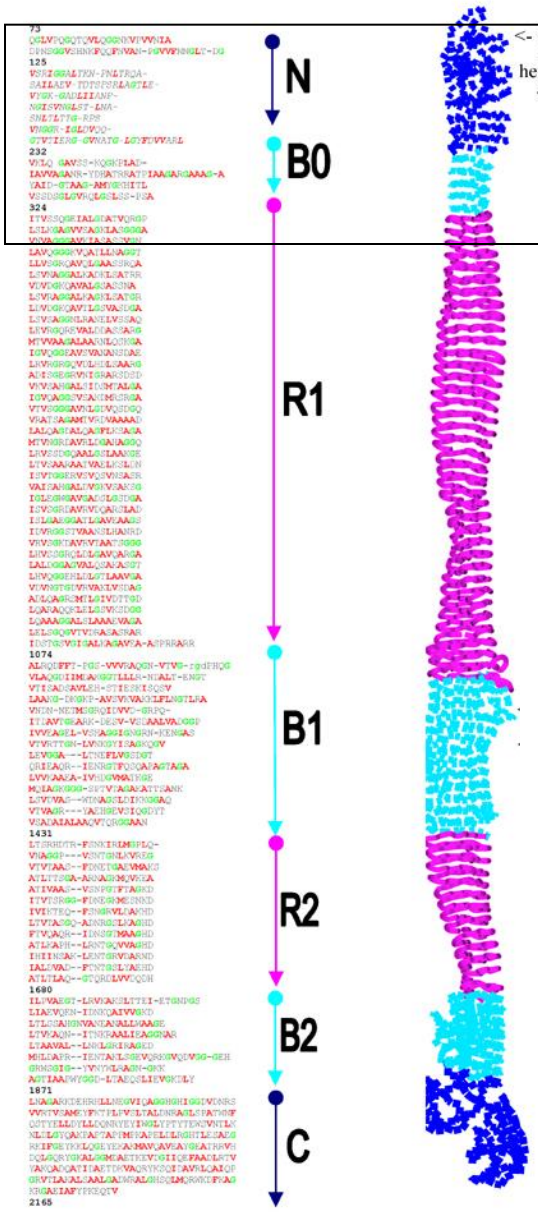
Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Locht C, Steven AC. *Mol Microbiol*. 2001; 42(2):279



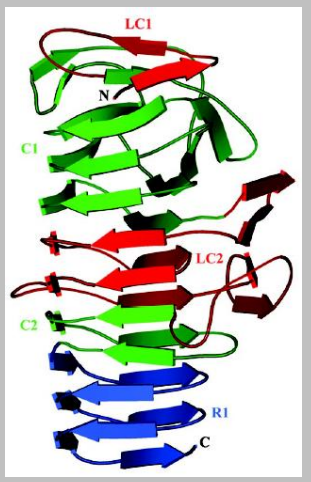


Clantin, Hodak, Willery, Locht, Jocob-Dubuisson and Villeret *PNAS* 2004; 101: 6194

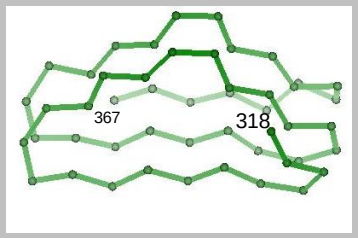
Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Locht C, Steven AC. *Mol Microbiol*. 2001; 42(2):279



<- 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.OO 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

Jorda J and Kajava A.V. (2009) *Bioinformatics*. 25(20):2632-8.