

PROTEINS WITH TANDEM REPEATS

Dr Andrey Kajava

Group of Structural Bioinformatics and Molecular Modeling

Centre de Recherches de Biochimie Macromoléculaire, CNRS

Montpellier, FRANCE

PROTEIN SEQUENCE – STRUCTURE - FUNCTION

Proteins with tandem repeats

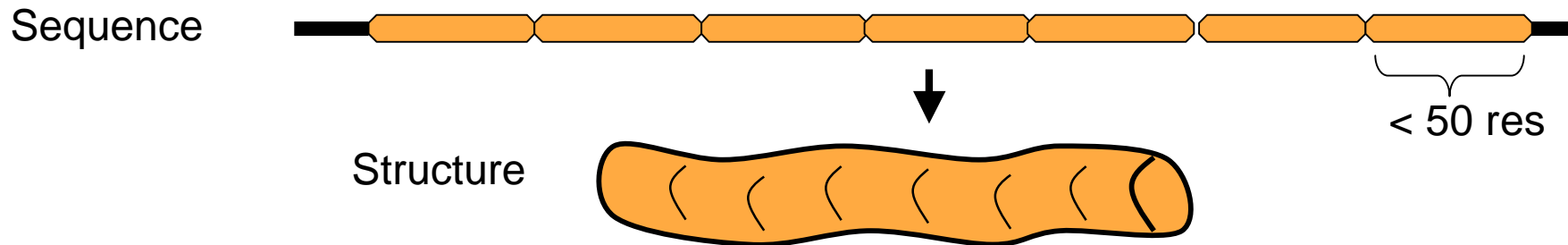
- ✓ Identification of protein repeats
- ✓ Analysis and Classification of the known 3D protein structures
- ✓ Structural prediction
- ✓ Experimental tests
- ✓ Evolution of proteins with repeats
- ✓ Applications in medicine, material science and nanotechnologies

Proteins with tandem repeats

Proteins with internal duplications represent a large portion of genomes

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

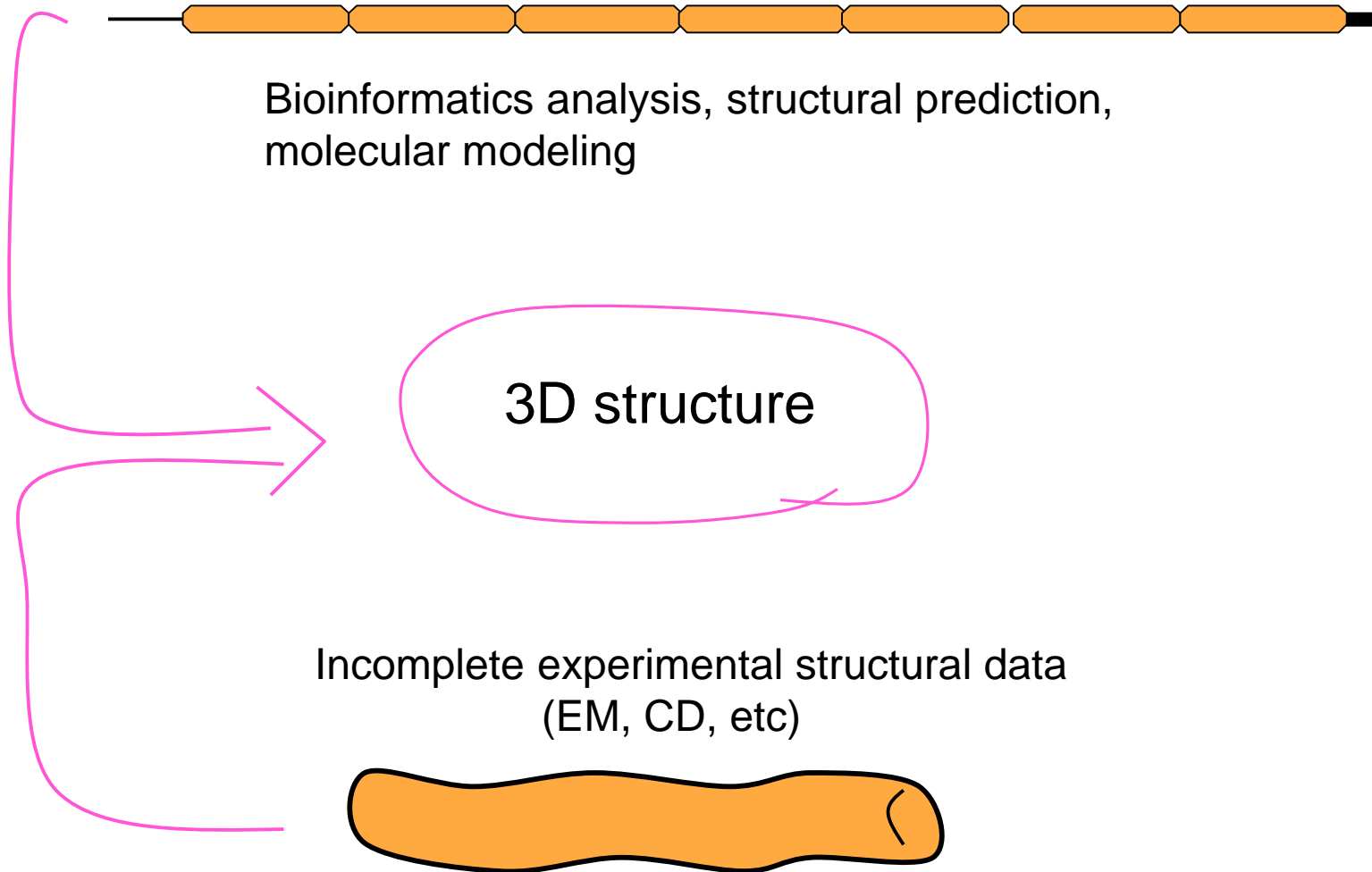
Pellegrini et al. (1999) *Proteins* 35:440



only ~ 2% of known 3D structures

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

HYBRID APPROACHES TO OBTAIN 3D STRUCTURE



Proteins with tandem repeats

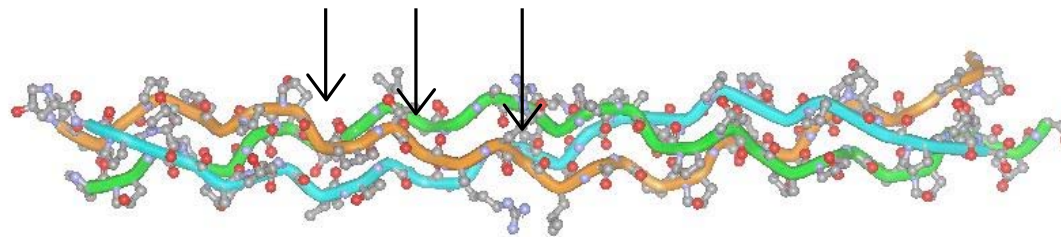
PROTEIN SEQUENCE – STRUCTURE - FUNCTION



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

**IDENTIFICATION
OF PROTEIN REPEATS**

PPGPPGPPGPPGPPGPPGPPGPPGPPGPPG
PPGPEGGPPGITGARGLAGPPGPPGKPPGPPG



Collagen

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

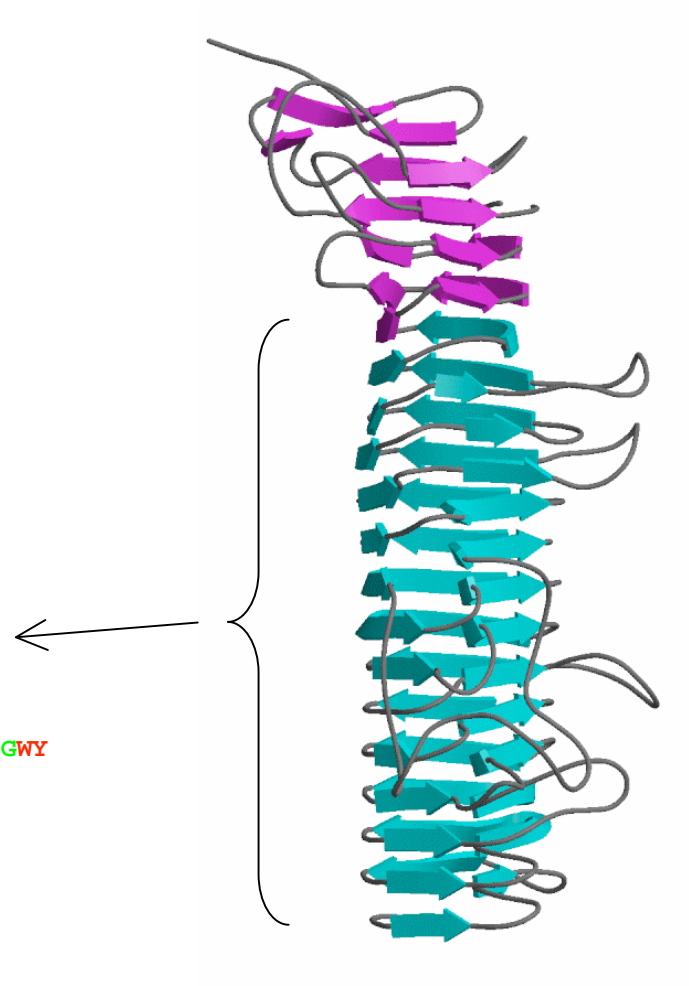
Estimation of edit distance between strings

TRED

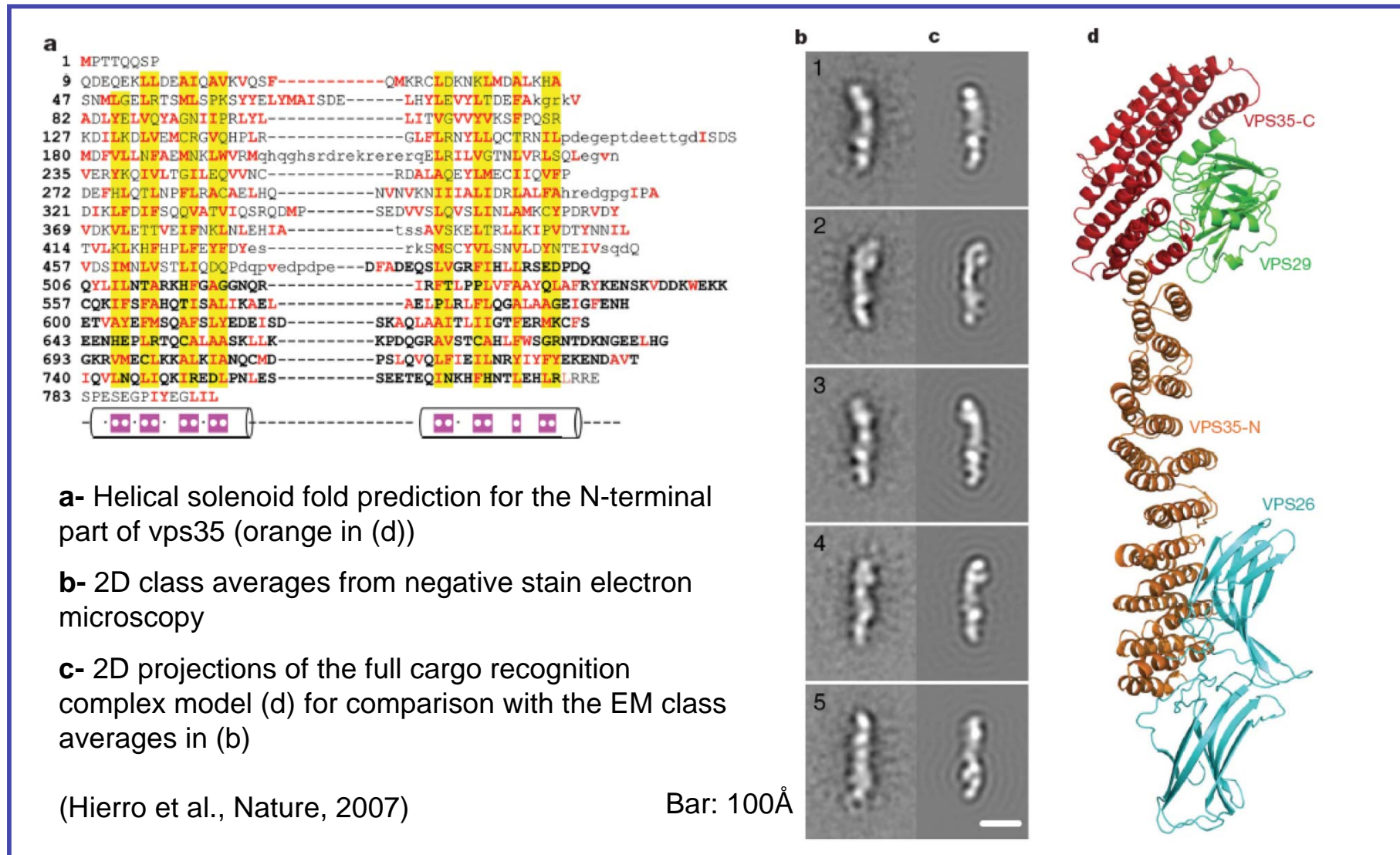
Sokol et al. 2007

Pertactin from *Bordetella pertussis*

GILLENPAAELQFRNGSVTSSGQLSDDGIIRRFLLG
TVTVKAGKLVADHATLANVGDWDDDGI
ALYVAGEQAQASIA DSTLQGAG
GVQIERGANVTVQRS AIVDG
GLHIGALQSLQPEDLPPSRVVL RDTNVTAVPASGAPA
AVSVLGA SELTLDGGHITGGR AA
GVAAMQGAVVHLQRATIRRGDAPAGGAVPGGAVPGGAVPGGFGPGGFGPVL DGWY
GVDVSGSSVELAQS IVEAPELGA
AIRVGRGARVTVSGGSLSAPHGN
VIETGGARRFAPQAAPLSITLQAGAH AQGKA
LLYRVLPEPVKLTLTGGADAQG
DIVATELPSIPGTSIGPLDVALASQARWTG



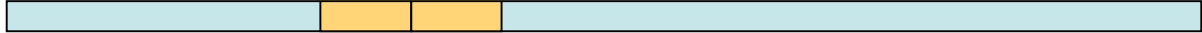
Cargo recognition complex



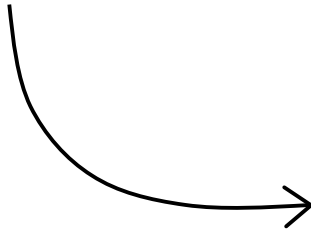
The α -solenoid fold extends the full length of Vps35 and Vps26 is bound at the opposite end from Vps29.



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



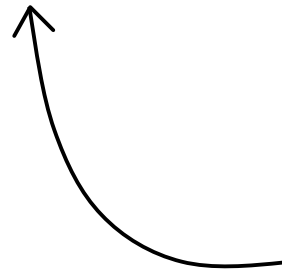
*** ** * * * * *
GITLENPSS-----AAELQFRN-GSVTNSGQLSDGI
TITLKATSS-----AKLVADH-ASVANVGQTDWGI



```
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,-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,-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,-1,1,0,0,1,1,-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-STLQAGAG
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,-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,-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,-1,1,0,0,1,1,-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;
```

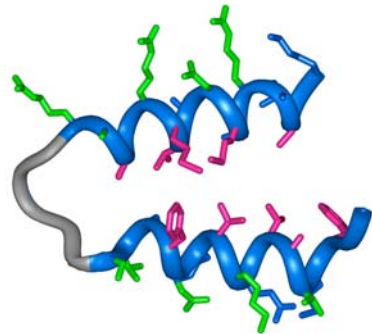
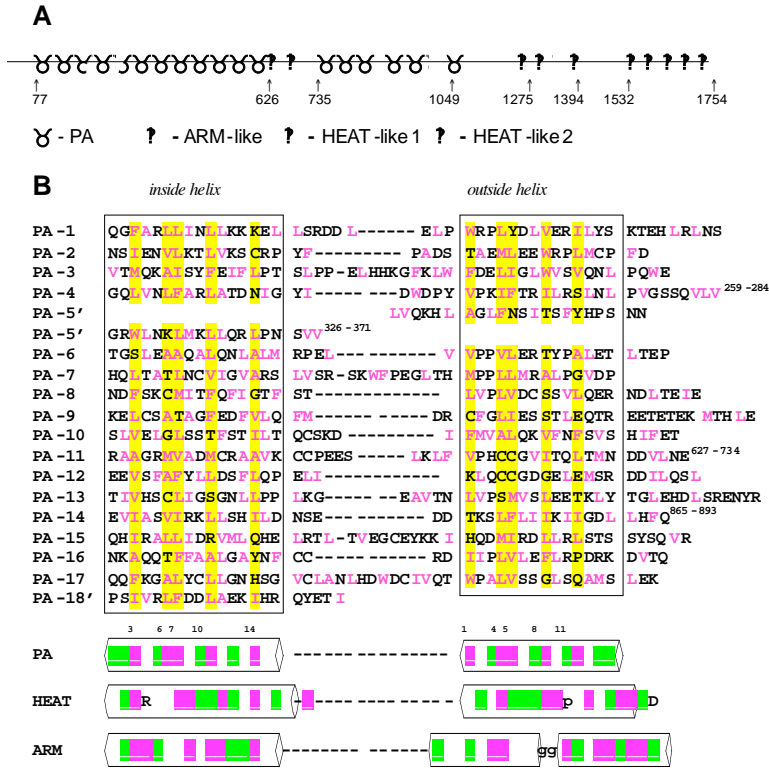
Sequence profile search

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

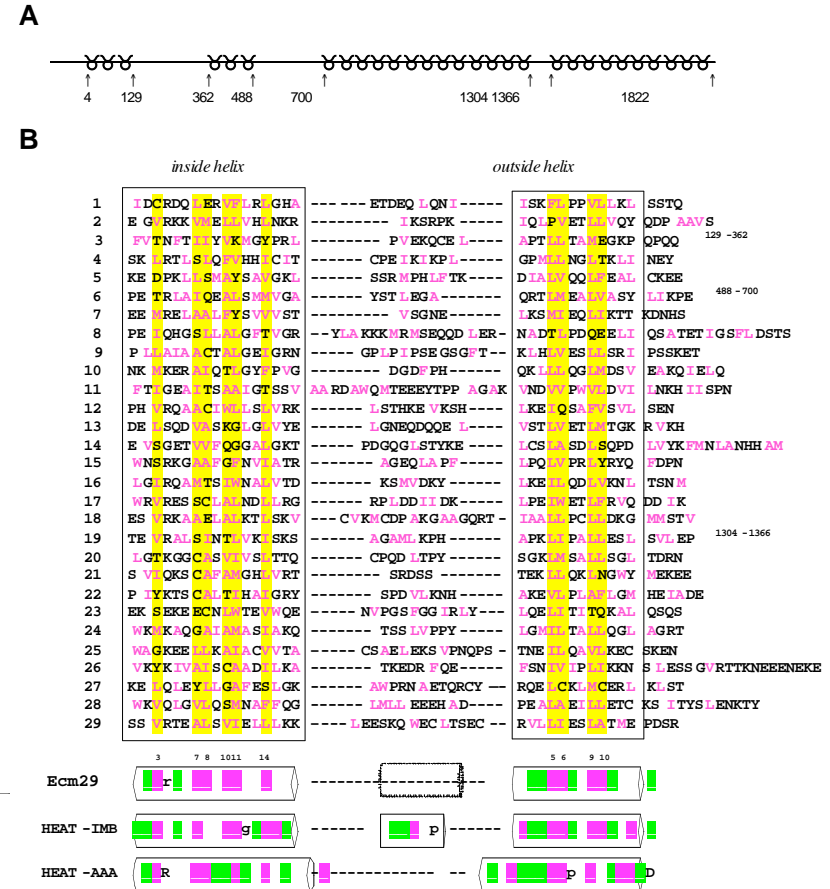
CRBM collection of protein repeats: <http://bioinfo.montp.cnrs.fr>

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

Nuclear proteasome activator PA200



Ecm29



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

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.cnrs.fr

















TEL 33 4 67 61 3364

FAX 33 4 67 52 1559

[Retour] [Accueil du CRBM]

La bibliothèque de profils

Profiles in database [\[Add a profile\]](#)

Action	Name	Category	Theme	Description	Results	Author	Last modification	Validated
	1VH4_v5	Repeats	beta solenoid	beta solenoid profile based on SufD protein from E.Coli (3 repeats)		Laurent Guth	Tue 06 Jun 2006	Yes
	LRR_BACT_20	Repeats	Leucine Rich Repeat	Bacterial 20-residue LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_CC	Repeats	Leucine Rich Repeat	Cysteine-containing LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_PS	Repeats	Leucine Rich Repeat	Plant specific LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_RI	Repeats	Leucine Rich Repeat	Ribonuclease inhibitor-like LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_SDS22	Repeats	Leucine Rich Repeat	SDS22+-like LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_TP	Repeats	Leucine Rich Repeat	Treponema pallidum (Tp) LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes
	LRR_TYPICAL	Repeats	Leucine Rich Repeat	Typical LRR profile.		Andrey Kajava	Wed 26 Apr 2006	Yes

login :

password :



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
KKKKGMMNVCSTRYIPPPPPWAFDKKKKGMMNVCSTRYIPPPPPWAFD
MNVCSR...YIPPPPPWAFD
```

Select profile categories to search in :

- Repeats
- Protein domains

Or upload a FASTA sequence file :

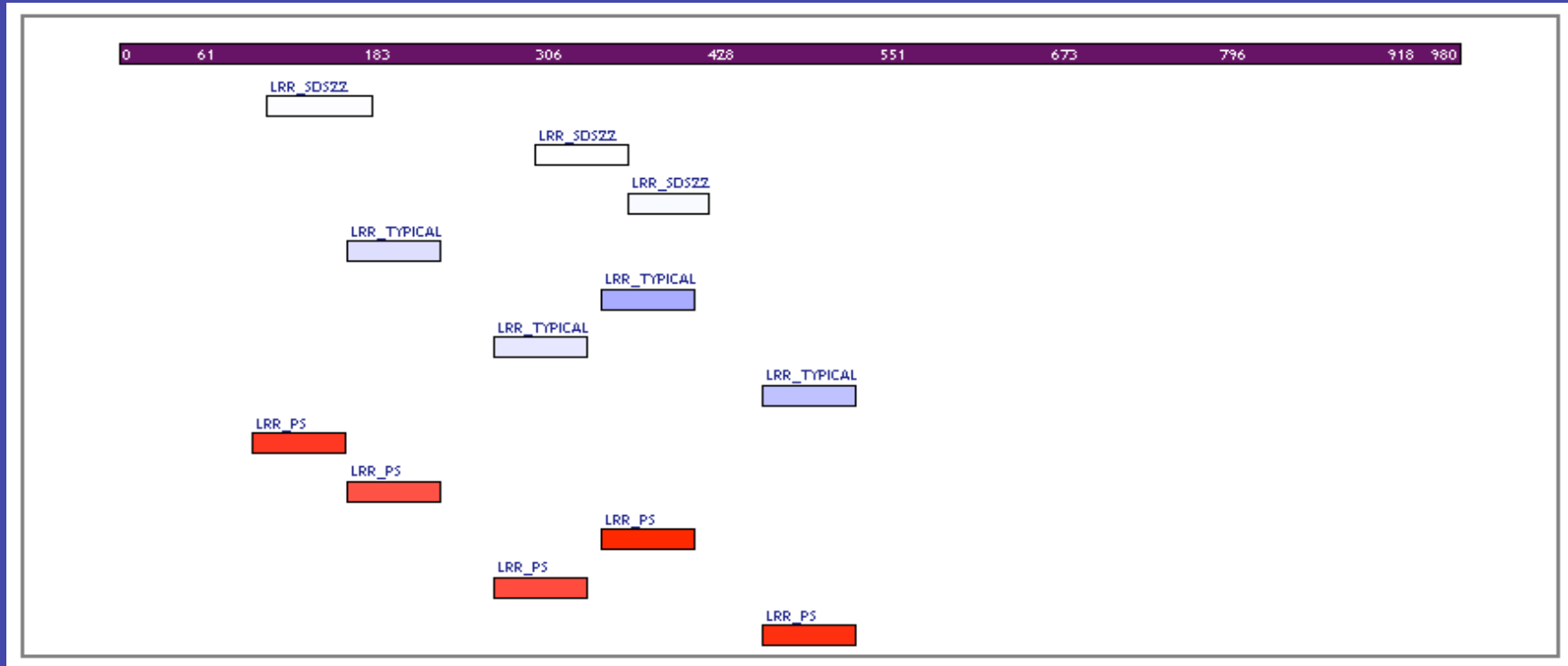
Pfscan [Help]

Query results :

Matching profile : LRR_SDS22 - SDS22+-like LRR profile.	
?	E-value = 6.19E-01 ; N_score = 7.535 ; Position : 116 - 198
Profile	NNNR I KK I ENLEA-----LPN L EH L BLNNNR I KK I ENLEA-----LPN L EH L BLNNNR I KK I ENLEA---LPN L EH L BL-
Query	TLKQ L SL S EN L LHGn i pq e l g LN R L V Y L DL G SN R L G S I P V Q L f c ng s SS S L Q Y I DL S NN S L T GE I PL N Y h ch L KE L RF L L L -
?	E-value = 8.50E-01 ; N_score = 7.397 ; Position : 305 - 376
Profile	NNNR I KK I ENLEA--LPN L EH L BLNNNR I KK I ENLEA--LPN L EH L BLNNNR I KK I ENLEA--LPN L EH L BL-
Query	D Q N R I H G S I P P E I s n L L N L T L L NS N L S G I P P R E L c k L S K L E R V Y L SN H L T GE I P M E L g d I P R L G L L D V-
?	E-value = 4.32E-01 ; N_score = 7.691 ; Position : 378 - 446
Profile	NNNR I KK I ENLEA--LPN L EH L BLNNNR I KK I ENLEA--LPN L EH L BLNNNR I KK I ENLEA L PN L EH L BL-
Query	R N N L SG S I P DS F G n L S Q L R R L L L Y G N H L S G T V P Q S L g k C I N L E I I D L S H N L T GT I P V E V W S N L R N L K L-
Matching profile : LRR_TYPICAL - Typical LRR profile.	
?	E-value = 1.63E-02 ; N_score = 9.114 ; Position : 175 - 247
Profile	N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L -R G L P N L EH L Y L -
Query	N S L T GE I PL N Y h ch L KE L RF L L L W S N K L T GT V P S L S N S T N L K W D L S N M L S GE L P S Q V i S K M P Q L Q F L Y L -
?	E-value = 3.59E-05 ; N_score = 11.771 ; Position : 353 - 424
Profile	N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L -
Query	S N N H L T GE I P M E L GD I P R L G L L D V S R N N L S G S I P DS F G N L S Q L R R L L L Y G N H L S G T V P Q S L G K C I N L E I I D L-
?	E-value = 5.16E-02 ; N_score = 8.614 ; Position : 281 - 352
Profile	N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L -
Query	G N S L GG E I T S S V R H L S V N L V Q I H L D Q N R I H G S I P P E I S N L L N L T L L N L S N L S G I P R E L C K L S K L E R V Y L -
?	E-value = 4.41E-04 ; N_score = 10.682 ; Position : 475 - 546
Profile	N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L N H N Q L T R V PD H L F R G LP N L E H L Y L -
Query	S S N E L S G K I P P Q L G S C I A L E H L N L S R N G F S S T L P S S L Q L P Y L K E L D V S F N R L T G A I P P S F Q Q S S T L K H L N F-
Matching profile : LRR_PS - Plant specific LRR profile.	
!	E-value = 1.53E-11 ; N_score = 18.143 ; Position : 98 - 173
Profile	S N N R L S GE I P E S L G S L K N--L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N ---L Q R L D L -
Query	S R N F F V G K I P P E I G S L H e t L K Q L S L S EN L L H G N I P Q E L G L N R L V Y DL L G S N R L G S I P V Q L F C N G S s s L Q Y I D L-
!	E-value = 5.24E-10 ; N_score = 16.607 ; Position : 174 - 247
Profile	S N N R L S GE I P E S L G S L K N--L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N --L Q R L D L -
Query	S N N S L T GE I PL N Y h ch L ke L RF L L L W S N K L T GT V P S L S N S T N L K W D L S N M L S GE L P S Q V i S K M P q L Q F L Y L -
!	E-value = 1.02E-13 ; N_score = 20.317 ; Position : 353 - 424
Profile	S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L-
Query	S N N H L T GE I P M E L GD I P R L G L L D V S R N N L S G S I P DS F G N L S Q L R R L L L Y G N H L S G T V P Q S L G K C I N L E I I D L-
!	E-value = 2.60E-10 ; N_score = 16.911 ; Position : 280 - 352
Profile	S N N R L S GE I P E S L G S L K N--L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L -
Query	A G N S L G GE I T S S V R H L S V n L V Q I H L D Q N R I H G S I P P E I S N L L N L T L L N L S N L S G I P R E L C K L S K L E R V Y L-
!	E-value = 3.27E-12 ; N_score = 18.812 ; Position : 475 - 546
Profile	S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L S N N R L S GE I P E S L G S L K N L Q R L D L-
Query	S S N E L S G K I P P Q L G S C I A L E H L N L S R N G F S S T L P S S L Q L P Y L K E L D V S F N R L T G A I P P S F Q Q S S T L K H L N F-

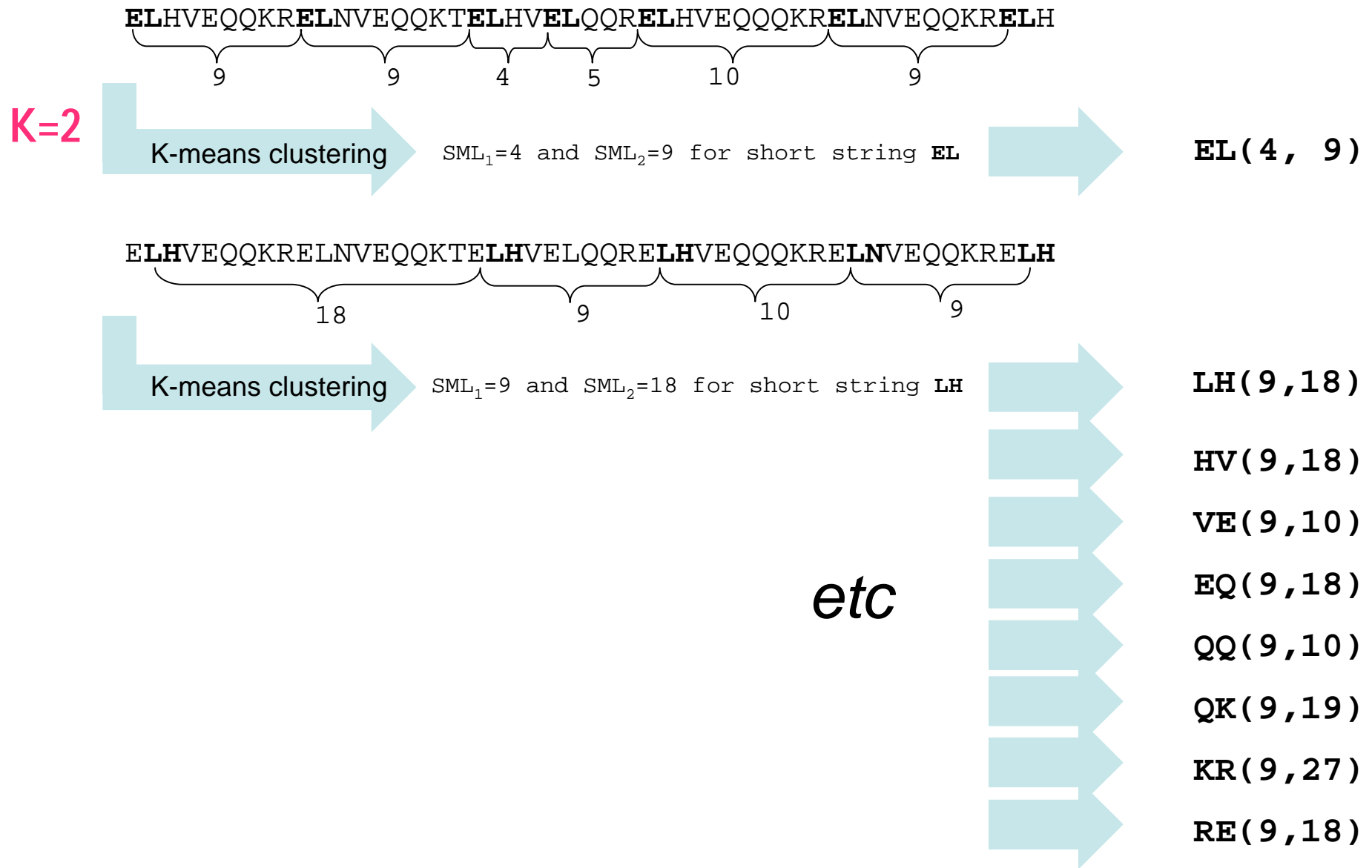
Visualisation des résultats du module pfscan

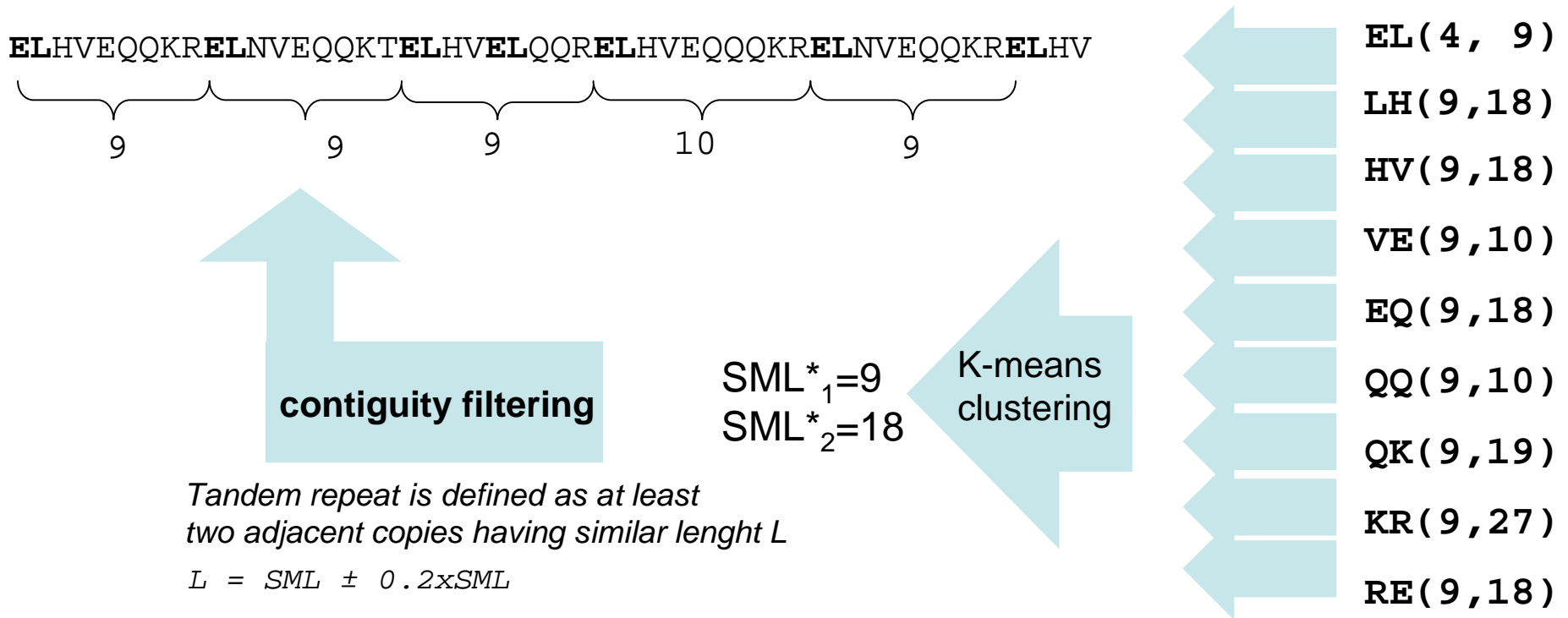
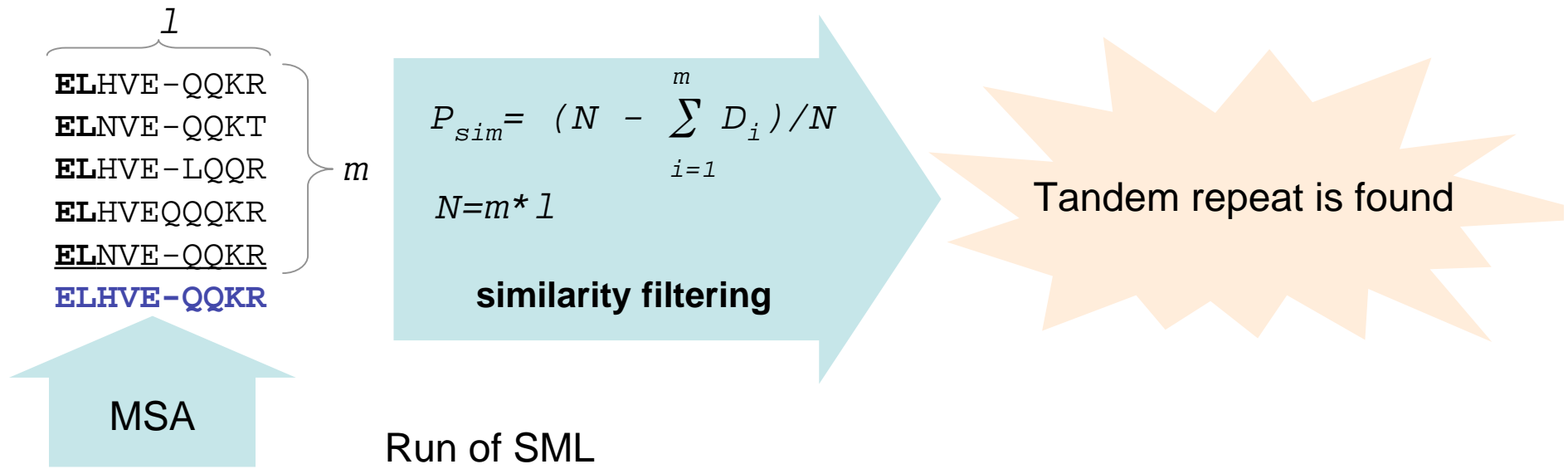
Module pfscan (suite)



Positions relatives des motifs sur la séquence d'une protéine soumise.

T-REKS: identification of tandem repeats based on clustering of lengths between identical short strings by using a K-means algorithm





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	4m	21324	11h50
INTREP ²	863	25m	19405**	22h20
TRED ³	866	4m	14499	16h10
XSTREAM ⁴	418	40s	2040	8m

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

T-REKS can be applied to the nucleotide sequences

Comparison of repeats found by our program and Tandem Repeats Finder in the Human Frataxin gene intron 1*.

T-REKS ¹ / TRF ²			
Start	End	Copy Length	Copy Number
827 / 822	854	14 / 14	2 / 2.4
1199 / -	1212 / -	2 / -	7 / -
1229 / -	1255 / -	2 / -	13 / -
1760 / 1787	1847 / 1874	44 / 44	2 / 2
2167 / -	2184 / -	1 / -	18 / -
2185 / 2183	2207 / 2211	3 / 3	8 / 9.7
2387 / -	2406 / -	7 / -	3 / -

¹ T-REKS parameters: **SS length = 4res; K=20**

² Benson, 1999.

* additional repeats identified by T-REKS are in bold.

Bioinformatique Structurale et Modélisation Moléculaire CRBM-CNRS Montpellier



[\[show login \]](#)

[Accueil](#)

[Recherche](#)

[Publications](#)

[Outils](#)

[Equipe](#)

T-Reks

Tandem Repeats Explorer based on K-means algorithm in Sequences

[\[Download T-Reks\]](#)

Search in a database:

clusters for K-means:

OR

Percentage of similarity:

Search in a file :

Filter the overlapping repeats:

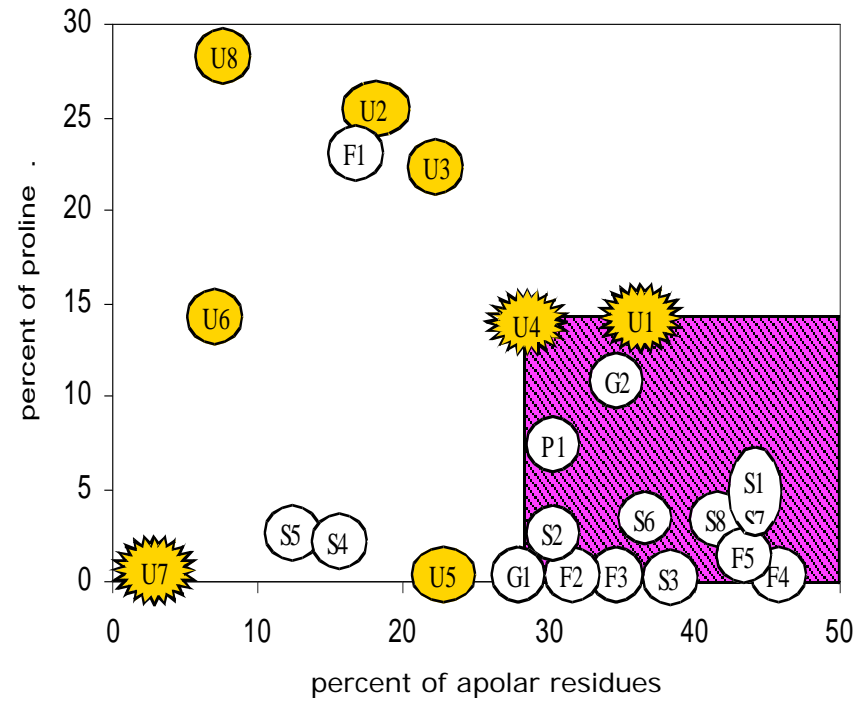


Database of protein repeats

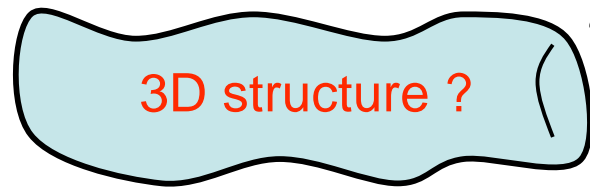
Large scale, systematic analysis of genomes

From sequence to 3D structure

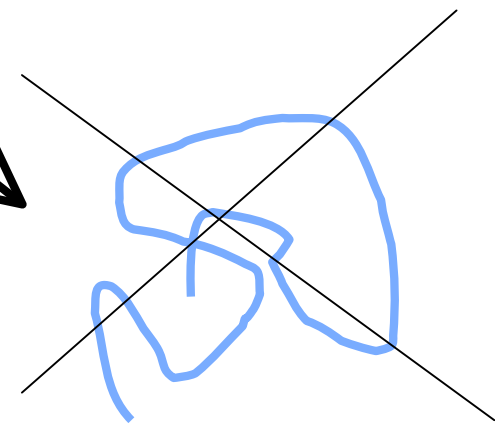
IS A PROTEIN WITH REPEATS STRUCTURED OR UNSTRUCTURED?



Polypeptide
with tandem repeats



Structured

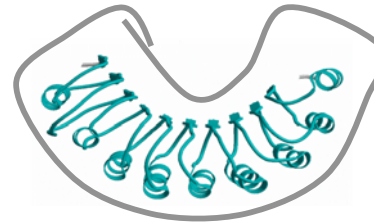


Nonstructured

Prédiction et modélisation de protéines à séquences répétitives

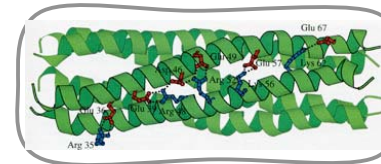
Leucine-rich repeat proteins

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



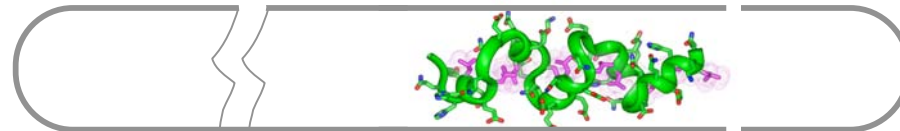
Filamentous Hemagglutinin Adhesin of Bordetella pertussis (56 nm long)

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



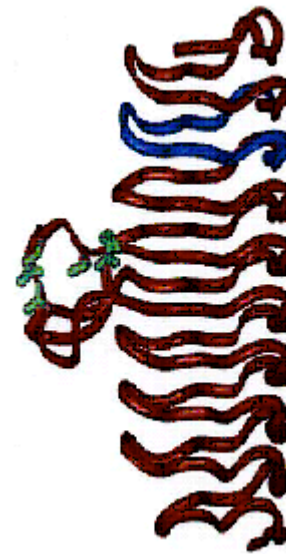
Distinguishing between structural and functional residue conservations

protein_human VKVSAHGALSIDSMTALGAIQVQAGGSVSAKDMRSRGAVTVSG.GAVN
 protein_rat VHLNAHGALTIKTMYSGNHISVQAGSHVSAREMHQSAFVTVHCAGSVN
 protein_yeast VKVSFQSSLIDSMTALGAIQVSSGSVDAKDMRSRGAVWVSG.GAVK

LGDVQSDGQ.VRATSAGAMTVRDVAAADPDGNKKPLALQAGDALQAGFLKSAGAGPPPDQM...
 LGDVQSWGQFVHASDGFMTVRDVSYRDGDPNRYTLGLQAGHALQAYYLRSSSA..NDQM...
 LAAVNNDGQ.VRATSAGAMCVWDVAAQDPDGNKKPLALSSGDGLKAGFLKSAGAGPPPDLM...

protein_human

VKVSAHGALSIDSMTALGA
 IQVQAGGSVSAKDMRSRGA
 VTVSG-GAVNLGDVQSDGQ
 VRATSAGAMTVRDVAAAADPDGNKKP
 LALQAGDALQAGFLKSAGAGPPPDQ
 MTVNG-DAVRLDGAHAGGQ
 LRVSSDGQAALGSLAAKGE
 LTVSAARAATVAELKSLDN
 ISVTGGERVSVQSVNSASR



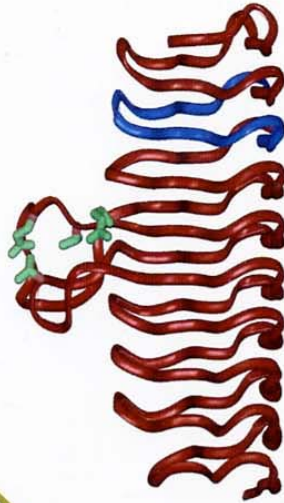
Distinguishing between structural and functional residue conservations

VK**V**SAHGAL**S**IDSMTALGA
IG**V**OAGGS**V**SA**K**DMRSRGA
VT**V**SG-GAV**N**LGD**V**QSDGQ
VR**A**TSAGAM**T**VR**D**VAAA**A**DE**P**DGN**K**KP
L**A**LQAGD**A**LQ**A**G**F**L**K**SAG**A**G**P**PP**D**Q
M**T**V**N**G-DA**V**RLD**G**A**H**AG**G**Q
L**R**VSSD**G**Q**A**AL**G**S**L**A**A**K**G**E
L**T**V**S**A**A**R**A**A**T**V**A**E**L**K**S**L**D**N
I**S**V**T**G**G**E**R**V**S**V**Q**S**V**N**S**A**S**R

One sequence repeat



One unit of repetitive structure



Indirect experimental structural evidence

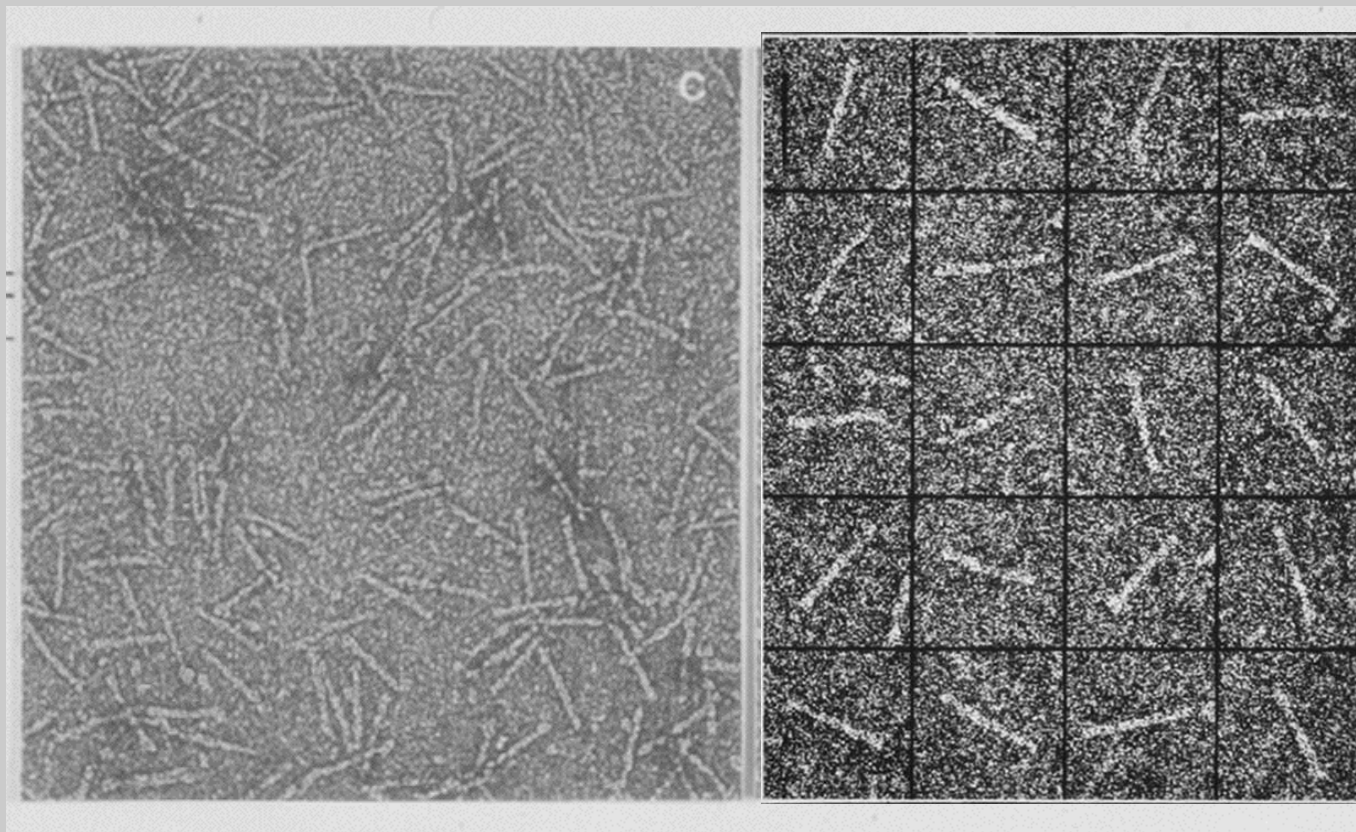
CD spectroscopy
(conformation)

Electron-microscopy
(shape,
oligomeric state)

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

**3D
structural
model**

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

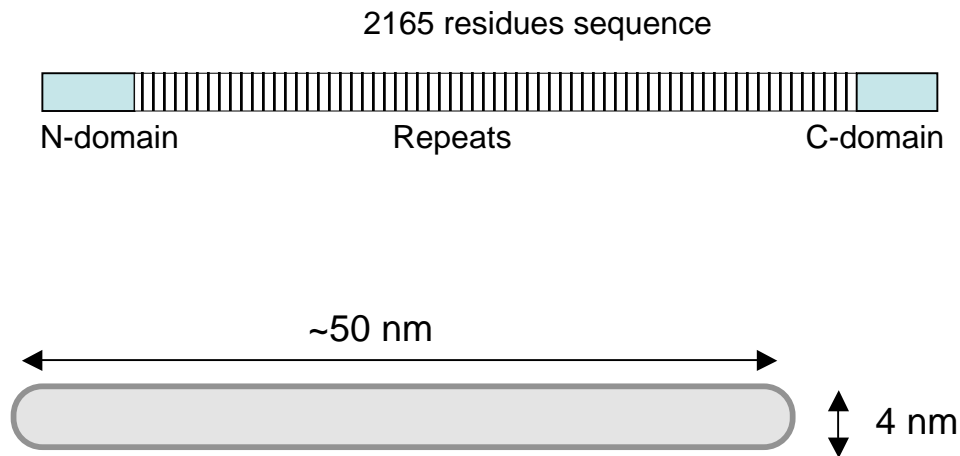


EM
negatively stained



Rod-like shape
50 x 4 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

```

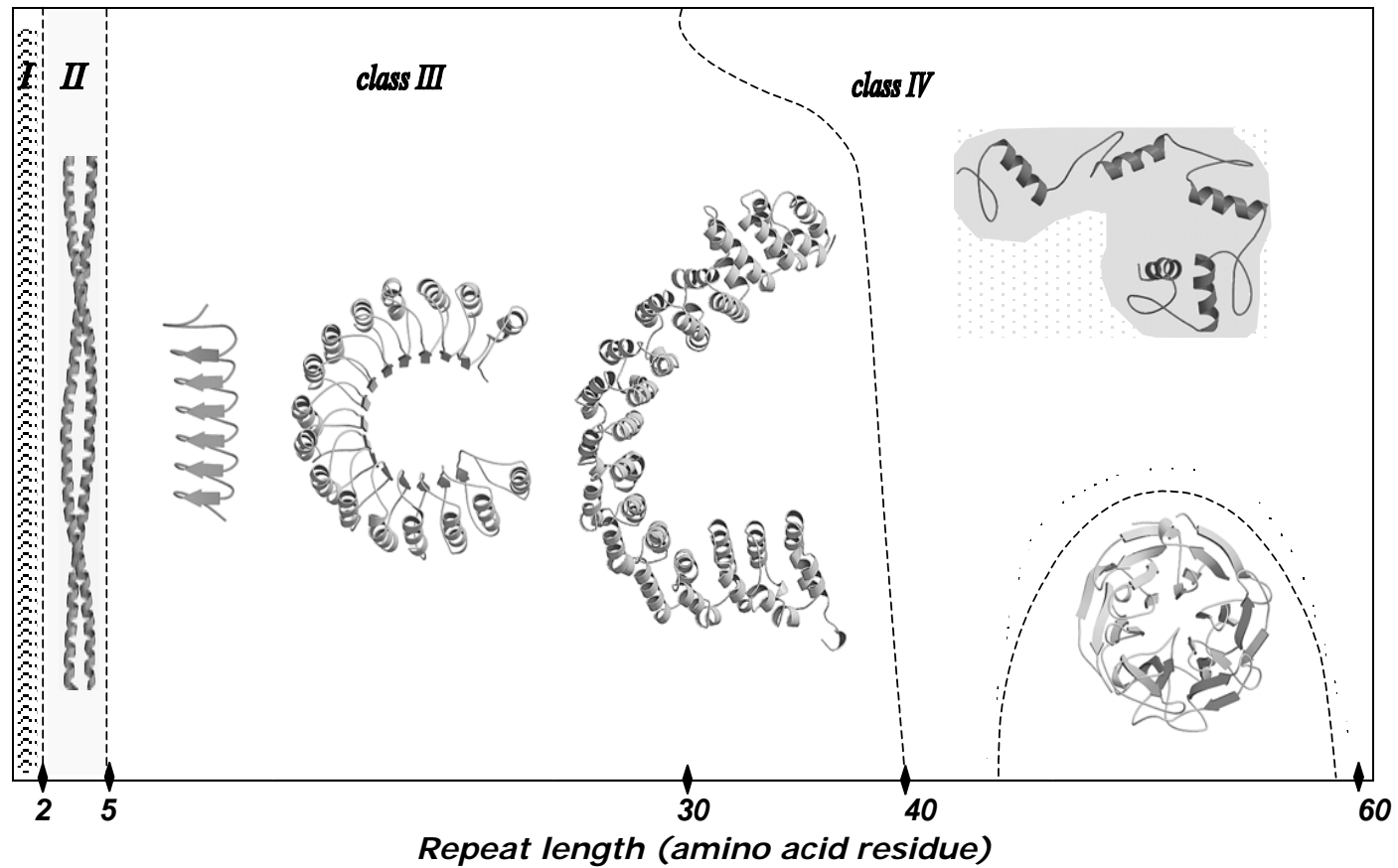
344 L S L K G A G V V S A G K L A S S V G G A 363
364 V N V A G G G A V K I A S S V G G N 382
383 L A V Q G G G K V Q L T L L N A G G T 401
402 L L V S G R Q A V Q L G A S S R Q A 420
412 L S V N A G G A L K A D K L S A T R R 439
440 V D V D G K Q A V A L G S A S S N A 457
458 L S V R A G G A L K A G K L S A T G R 476
477 L D V D G K Q A V T L O S S V A S D G A 495
496 L S V S A G G N L R A N E L V S S A Q 514
515 L E V R G Q R R V A L D D A S S A R G 533
534 M T V V A A C 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 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 K D M R S R G A 647
648 V T V S G G A V N 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 P L K S A G A 704
705 M T V N G R D A V R L D G A H A G G Q 723
724 L H 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 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 H G A V G A D S L G S D G A 818
819 I S V 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 Q A V E 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 G K D A V R V T A A T S C G G 894
895 L H V S S G R Q L D L G A V Q A R G A 913
914 L A L D G G A G V A L Q S A K A S G T 932
933 L H V Q G G E H L D L Q T L A A V G A 951
952 V D V N G T G D V R V A K L V S D A G 970
971 A D L Q A G R S M T L G I V D T T G D 989
990 L Q A R A Q Q K L E L G S V V K S D G G 1008
1009 L Q A A A Q G A L S L A A E V A G A 1027
1028 L E L S 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 G A L K A G A V E A 1065
    
```

consensus

```

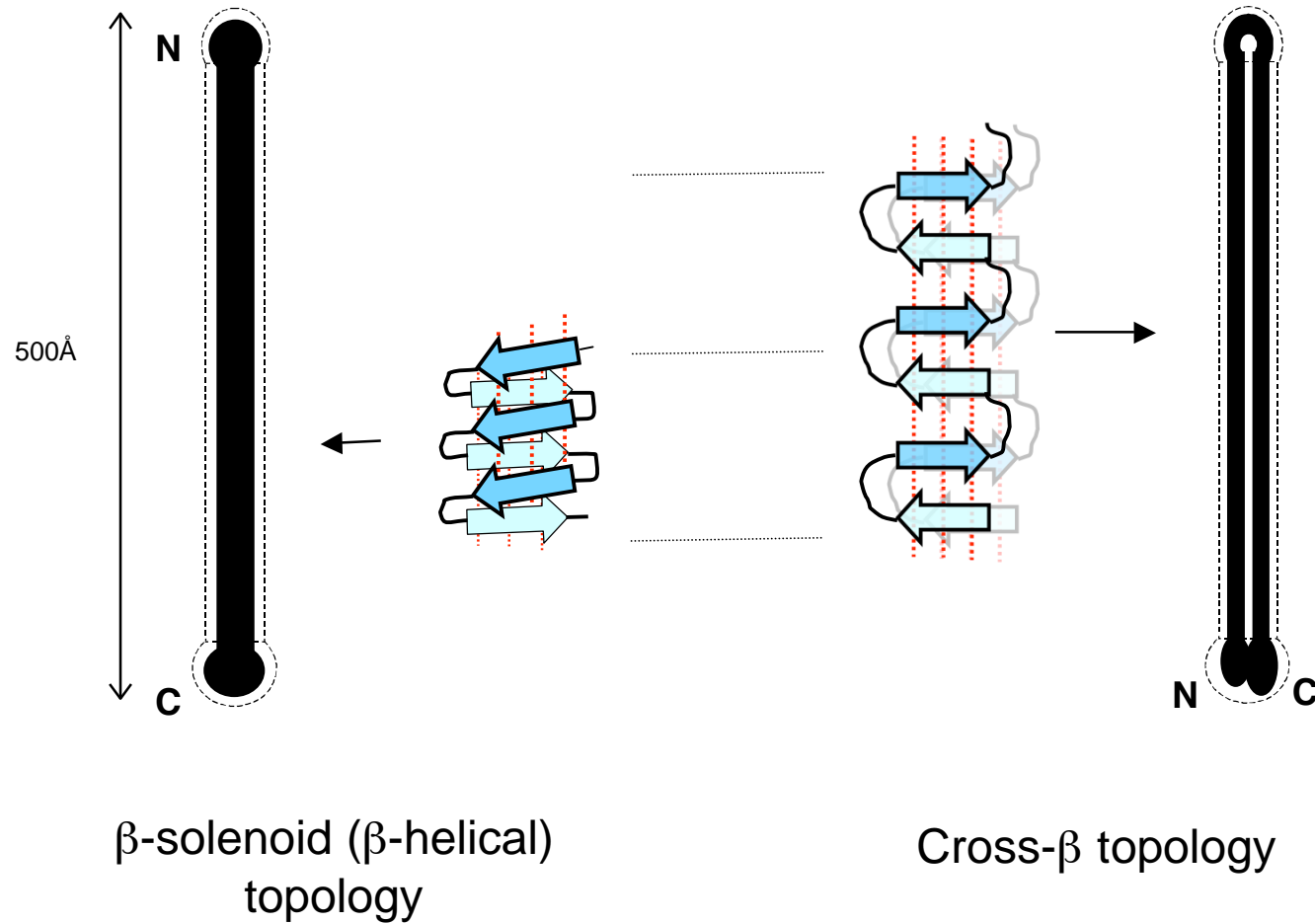
1 - v - g - g a l - l - l - a - g -
v - l - a - g a v - a - - v - a - g -
    
```

WHAT CAN REPEAT LENGTH TELL US ABOUT ITS STRUCTURE?

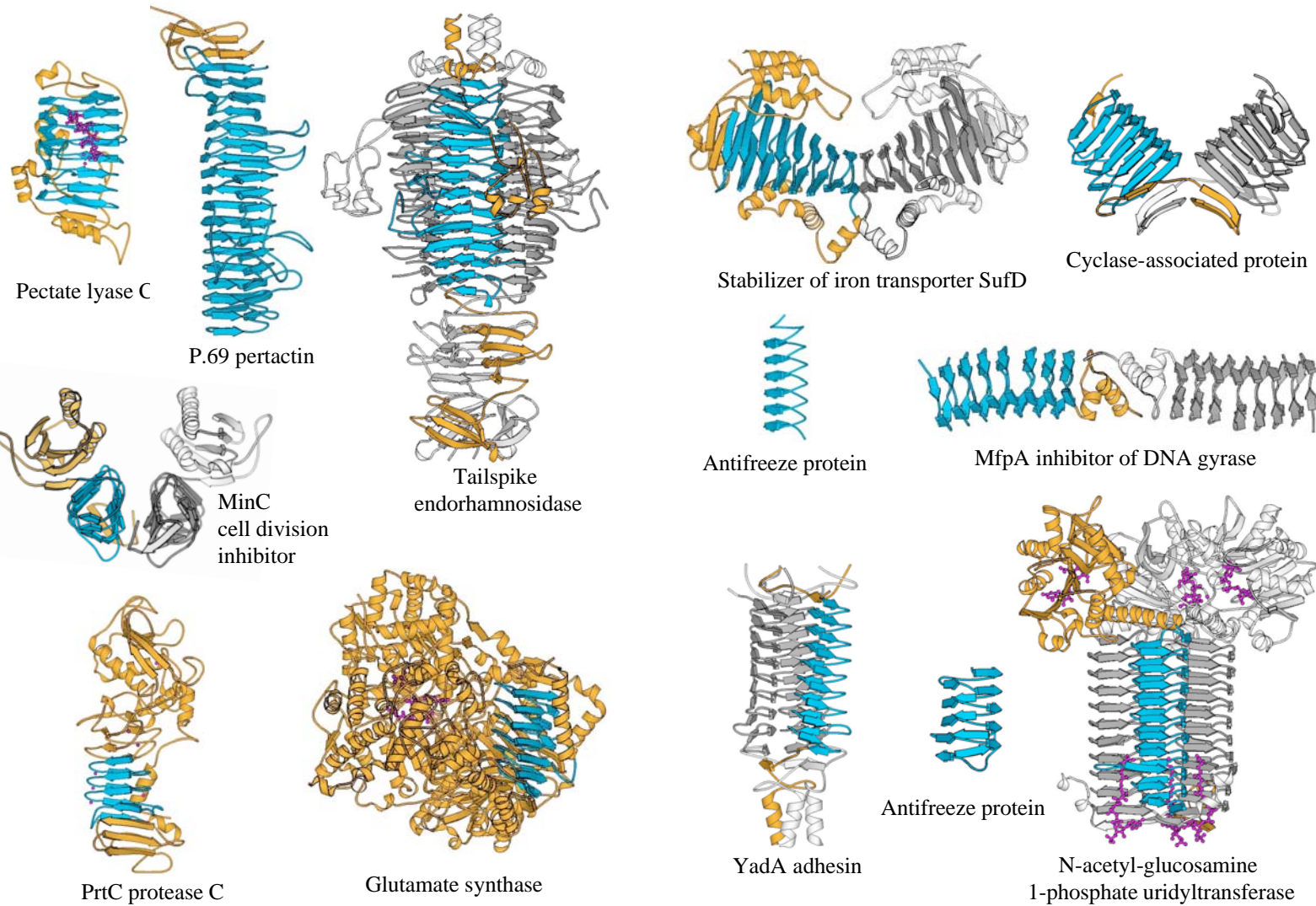


A.V. Kajava (2001) *J. Struct. Biol.* **134**:132

Topology of 3D structure of FHA from *Bordatella pertussis*

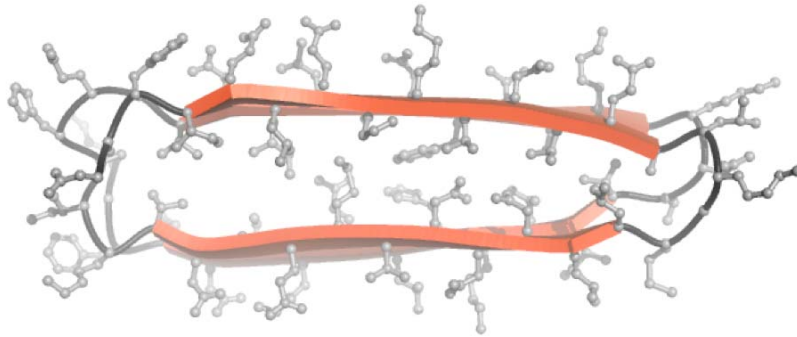


The known structures of β -solenoid proteins

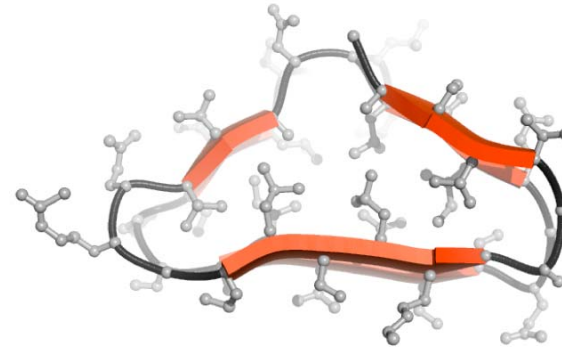


Classification of beta-solenoids

Cross-sectional shapes



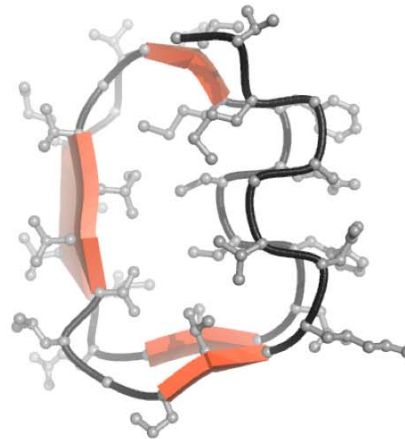
O-type



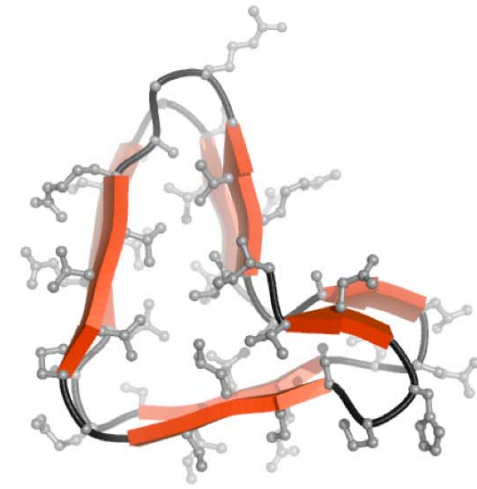
T-type



R-type



B-type



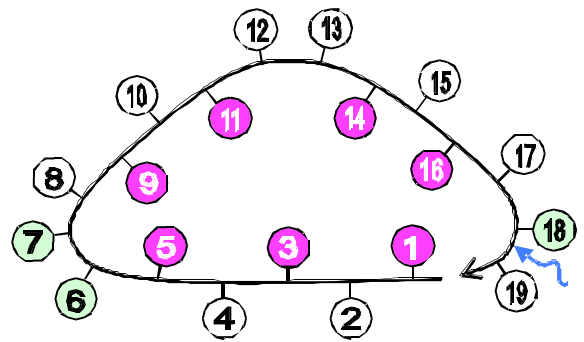
L-type

Repeat 1 V N V A G G G A V K I A S A S S 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

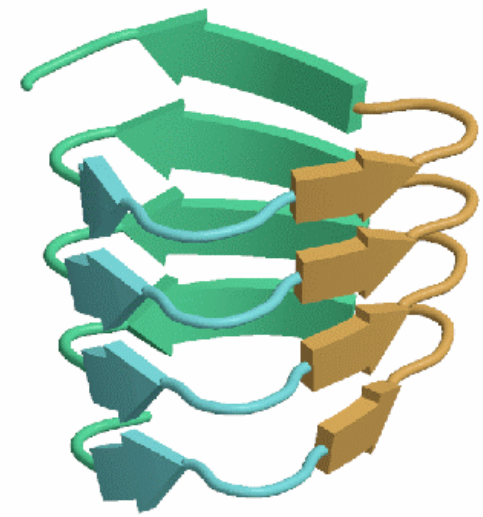
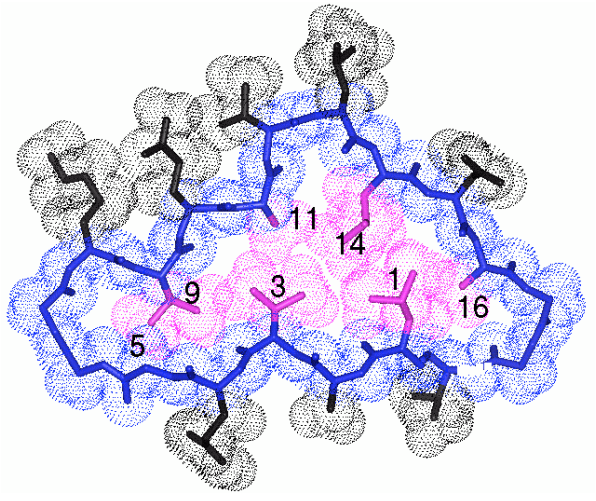
consensus L x V x A G G x V x L x x L x A x G - x

position s 1 3 5 7 9 11 13 15 17 19

2D plot



3D structure



FHA is a member of a large family of autotransporter proteins

(Over 1000 proteins)

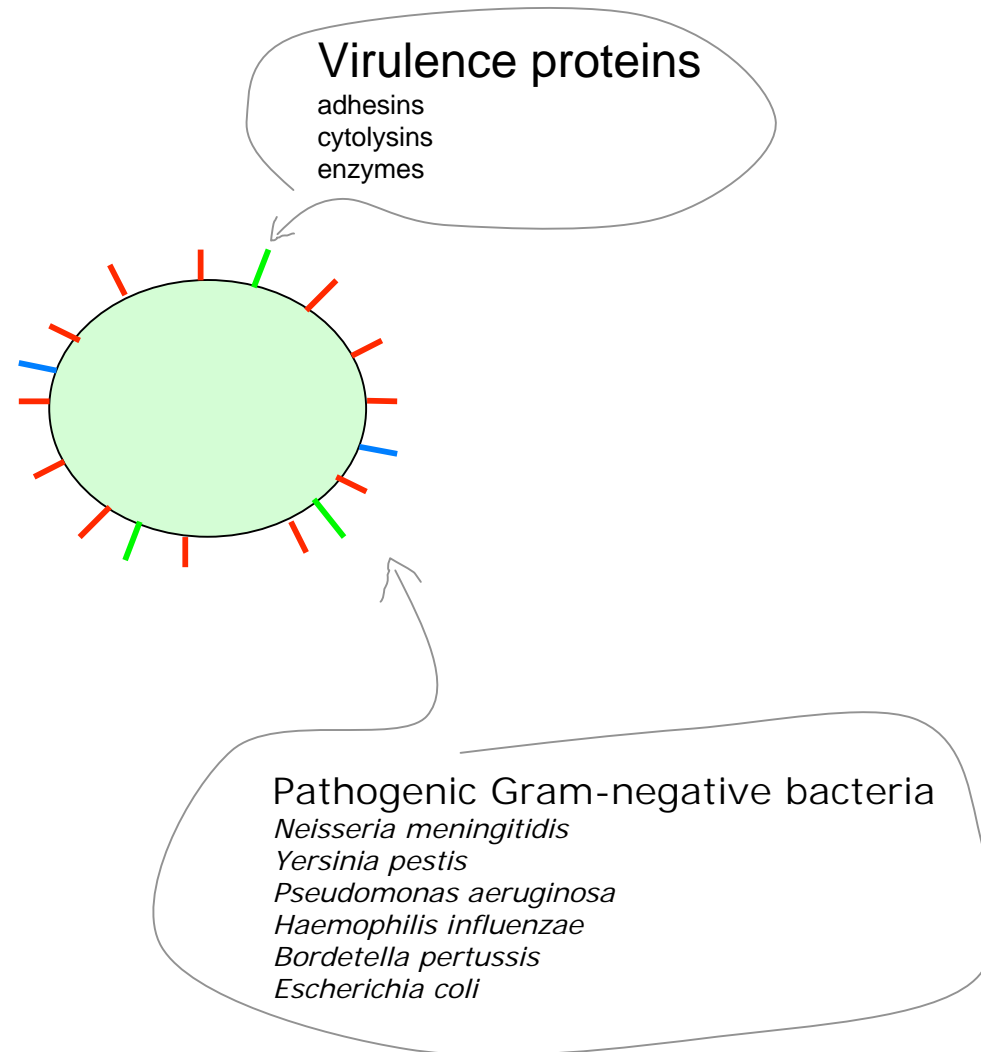


Table 1

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 $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
L2	Slr1753 protein (<i>Synechocystis sp.</i>)	23-28 res	x ● xtxxxxx - Gx ● x ● xaxx - ● x ● xx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
L3	AGRL_3085 protein (<i>Agrobacterium tumefaciens</i>)	25-27 res	xxGx ● x ● x - xaxsx ● xxxgx ● x ● xxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T1	FHA protein <i>B. pertussis</i>	18-19 res	● x ● xgx ● x ● xx ● axx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T2	FHA protein <i>B. pertussis</i>	19-20 res	● x ● xaxx - ● xNngx ● xaxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T3	HBP protein <i>E. coli</i>	18-20 res	s ● x ● x ● xx - ax ● x - xx ● axx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T4	TibA protein <i>E. coli</i>	18-19 res	g ● xQx ● x - xxgxaxxTx ● xxg $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T5	YapA protein <i>Y. pestis</i>	18-19 res	g ● xFx ● x - xxaxaxxtx ● xxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T6	Hap protein <i>Haemophilus influenzae</i>	19-20 res	s ● x ● x ● xxax ● xgx ● x ● xxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
T7	Hemagg.-hemolysin related protein <i>E. coli</i>	20-22 res	g ● x ● xsx - x ● x ● xsx - gx ● xNxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	

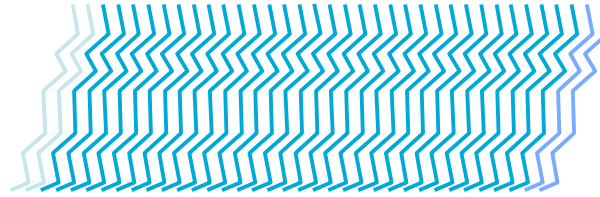
Table 1 (continuation)

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 $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$ $\xrightarrow{\beta_3}$	
O1	XadA protein <i>Xylella fastidiosa</i>	13-14 res	s ● xax - xxxs ● a ● Gxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$	
O2	UspA2H protein <i>Moraxella catarrhalis</i>	15 res	N ● xax - GxxsT ● aGGxx $\xrightarrow{\beta_1}$ $\xrightarrow{\beta_2}$	

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

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

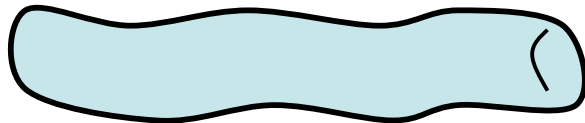
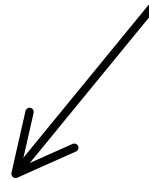
Aggregates, amyloids



Polypeptide
with tandem repeats



Nonstructured



Structured

Amyloid and prion fibrils

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

β -amyloid (Alzheimer's disease)

HET-s prion

Poly(Q) tracts (Huntingtin disease)

α -synuclein (Parkinson's disease)

Tau protein (Alzheimer's disease)

Prion domains of yeast proteins Sup35



β -solenoid models

Superpleated β -structural model

β -amyloid
Petkova et al. 2002

HET-s prion
Ritter et al. 2005

Ure2p prion, amylin
Kajava et al. 2004, 2005

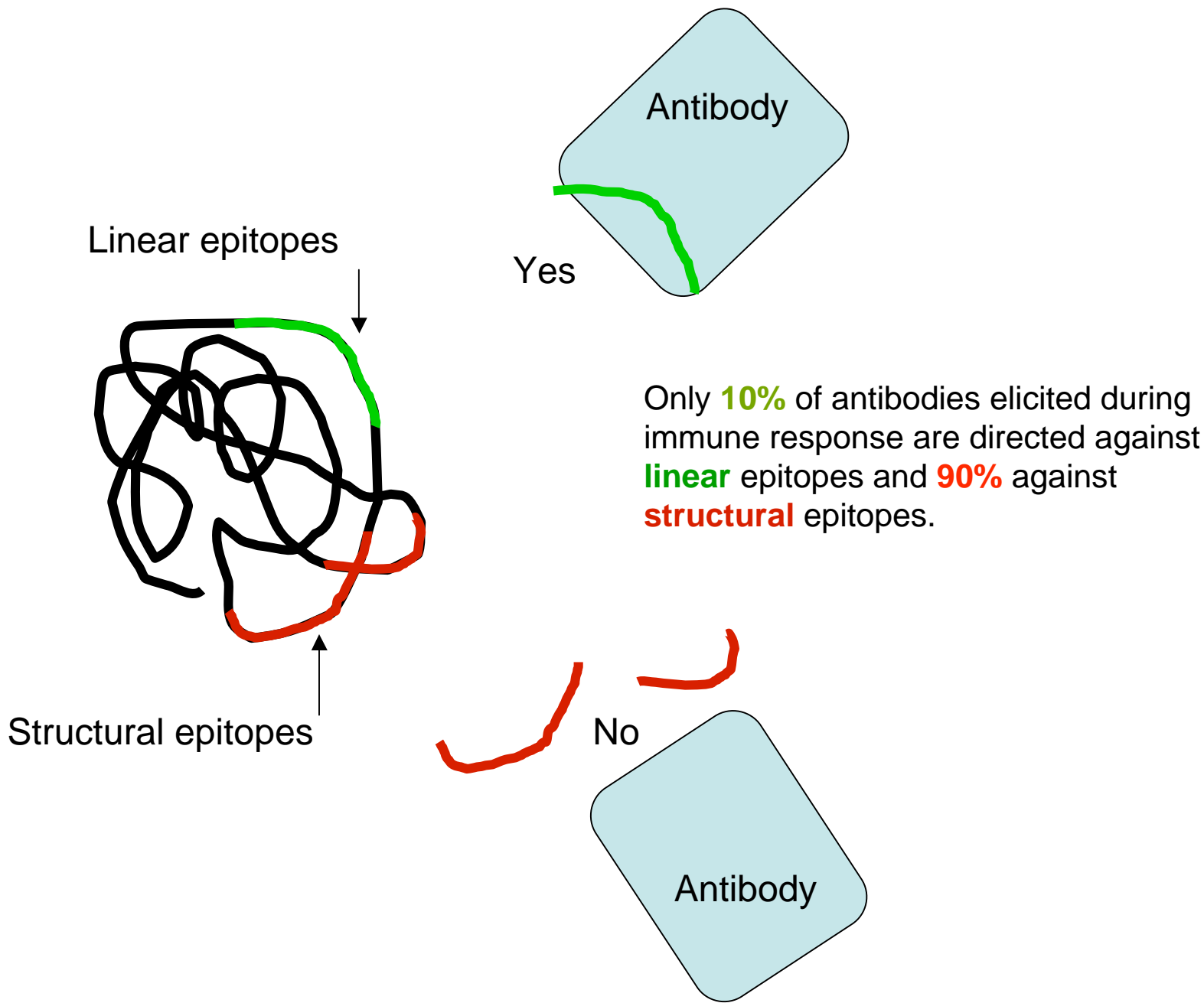
The 3D structure of amyloid fibrils and beta-solenoids

Rational design of inhibitors
of fibrillogenesis

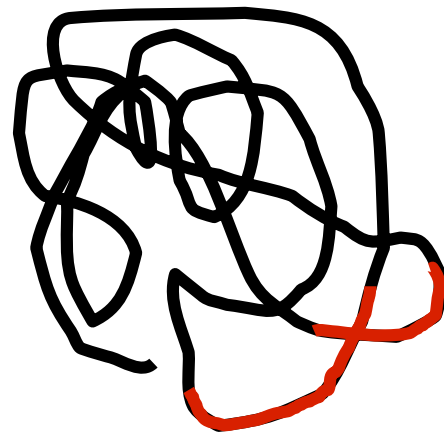
Prediction of amyloidogenicity
of proteins

Applications in medicine

**Protein structure based strategies
for antigen discovery and vaccine development**



Structure-based strategies for development of vaccine



Structural epitopes

The whole protein as a vaccine?



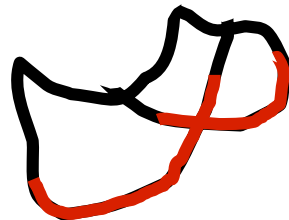
Expression of recombinant proteins



☹ Misfolding

☹ Costly and poorly adapted to high-throughput screening

Mimicry of “native” structural epitopes by designing mini-proteins



Peptide synthesis

- ☺ Fast and cost effective
- ☹ Limited size of peptides (> 50 res).



A challenge is to find >50 residue fragments that being taken separately from the protein will fold into the “native” structure.



Elicit Ab reactive with “native” structural epitopes

Strategies for development of vaccine

Conventional

Killed or attenuated in vitro-grown pathogens
(☹️ not all pathogens are grown in vitro)

or purified components of pathogens



Antigen selection



Clone genes



Tests and Vaccine development

☹️ 5-15 years

Bioinformatics + Peptide synthesis

(Corradin, Villard and Kajava (2007) *Endocrine, Metabolic & Immune Disorders - Drug Targets*, 7, 259)



Sequence motif of widespread, small and stable protein domain located on the pathogen surface



Bioinformatics analysis of genome



Vaccine candidates



Peptide synthesis



😊 2-3 years

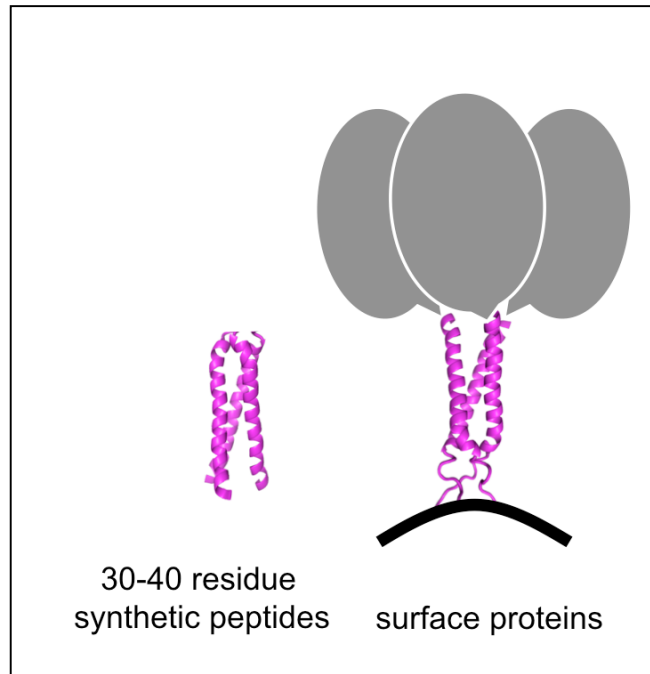
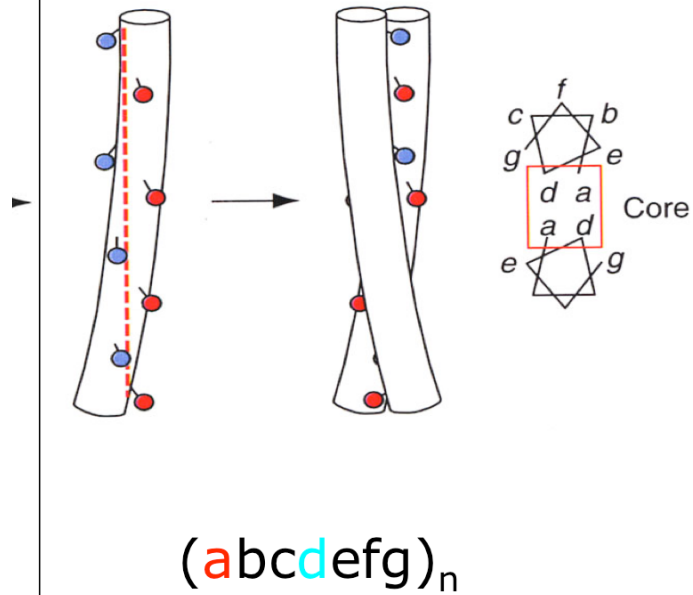
Strategies for development of vaccine

Malaria



Genome

Alpha-helical coiled coil domains, well-defined motif, widespread small and stable



Structural Bioinformatics

Peptide synthesis

Tests and Vaccine development

Strategies for development of vaccine

Malaria



Genome

**95 peptides (30-40 residues)
were synthesized**

All peptides are recognized (ELISA)

All 18 peptide specific antibodies are positive in IFA

12 out of 18 peptide specific antibodies are inhibitory in ADCI

Some peptides are immunogenic in CBF1 mice and Ab + in IFA

About 10 peptides suitable for preclinical development



A powerful approach to select new antigen

Villard et al. PLoS ONE. 2007 Jul 25;2(7):e645.u



→ **Structural
Bioinformatics**



Peptide synthesis



**Tests and Vaccine
development**

CRBM, CNRS, France

Jerome Hannetin

Berangere Jullian

Maria Kondratova

Arunachalam Jothi

Julien JORDA

Structural classification of proteins with repeats

Bostjan Kobe, University of Queensland Brisbane, Australia

John M. Squire, Imperial College London, UK

David Parry, Massey University, New Zealand

Vaccines

G. Corradin University of Lausanne, Switzerland

Structural studies of proteins with repeats and amyloids

Alasdair Steven,

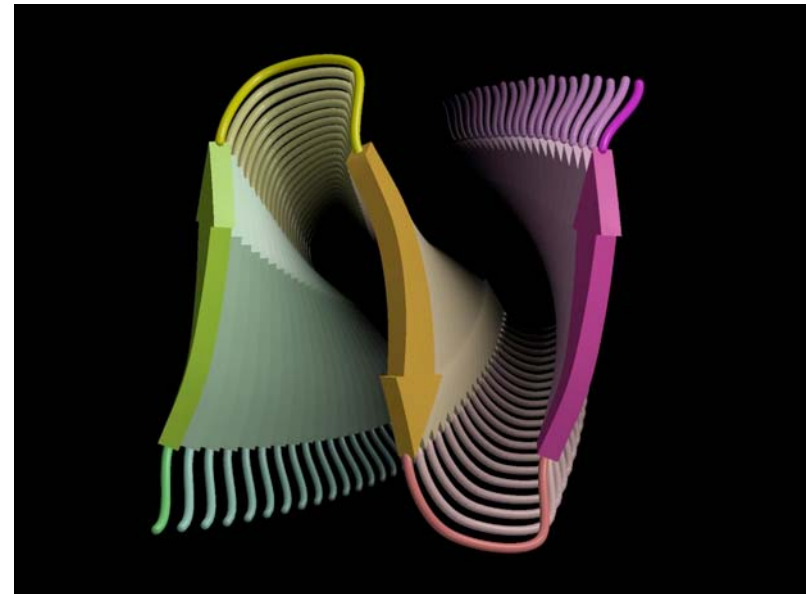
Laboratory of Structural Biology, NIAMS, NIH, USA

Amyloid fibrils of human amylin

Human amylin is the major component of pancreatic amyloid deposits found in ~ 90% of persons with non-insulin-dependent (type 2) diabetes mellitus.

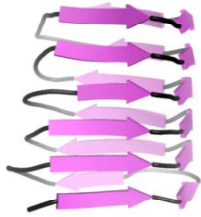
Experimental data on structural arrangement of hA fibrils.

1. By EM, diameter of a protofilament is 4.5 - 5.5 nm (Goldsbury et al. 1999; Makin and Serpell, 2004)
2. By X-ray, fibrils have a cross-beta structure with reflections 0.47 and 0.95 nm (Makin and Serpell, 2004)
3. STEM mass-per-unit-length data are consistent with one molecule per one beta-layer (Goldsbury et al., 1997)
4. By EPR-spectroscopy, parallel β -strands within fibril (Jayasinghe and Langen 2004)



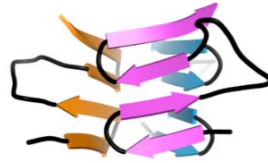
Pathogenic folds

Beta-structural fibrous proteins



β -solenoids

(in virulence factors of gram- bacteria)

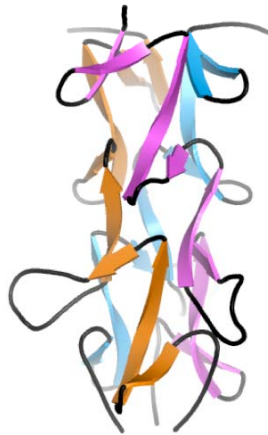


Cross- β -prism



Triple-stranded β -solenoid

(in bacteriophage tail proteins)



Triple β -spiral

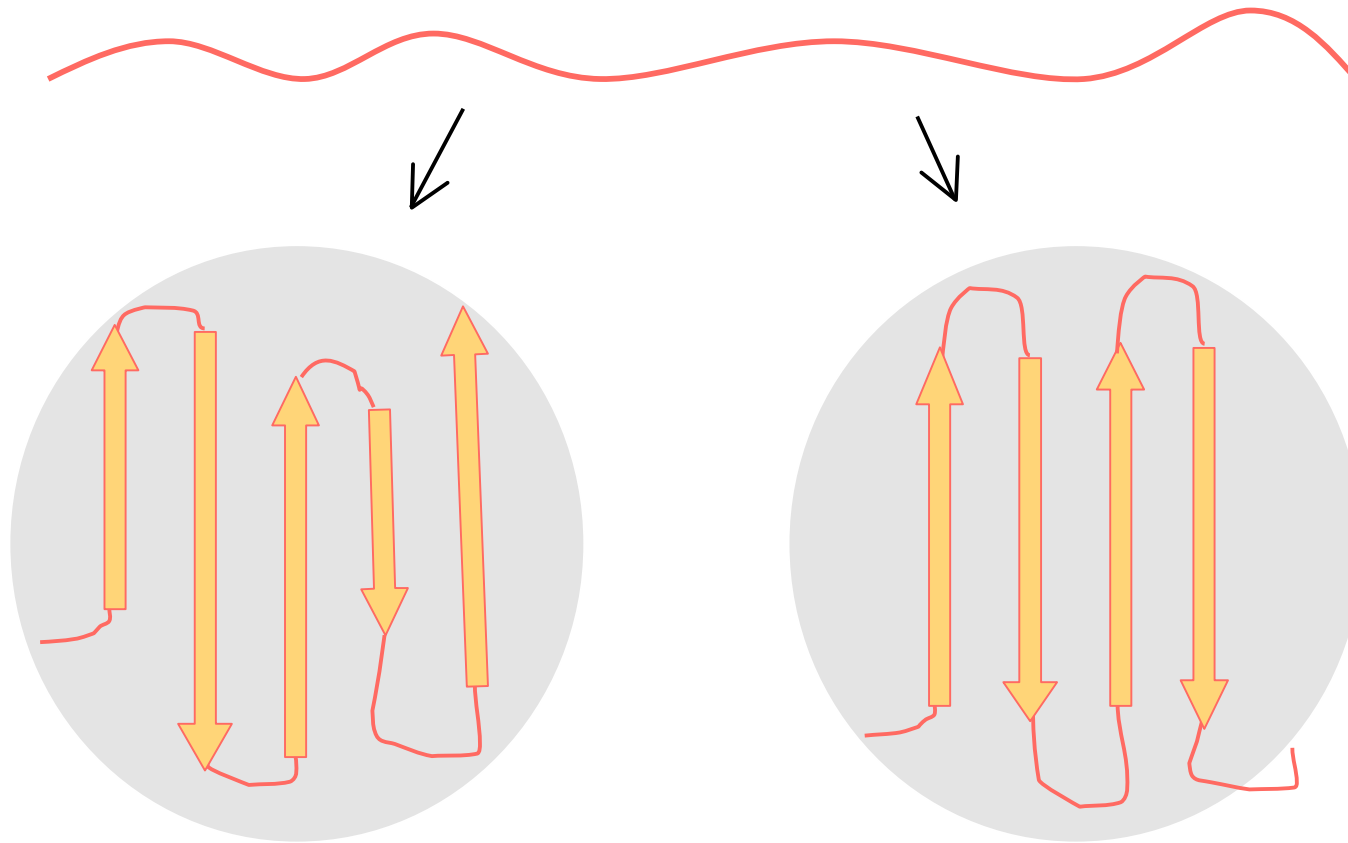
(in adenovirus fibers)



Spiral β -hairpin staircase

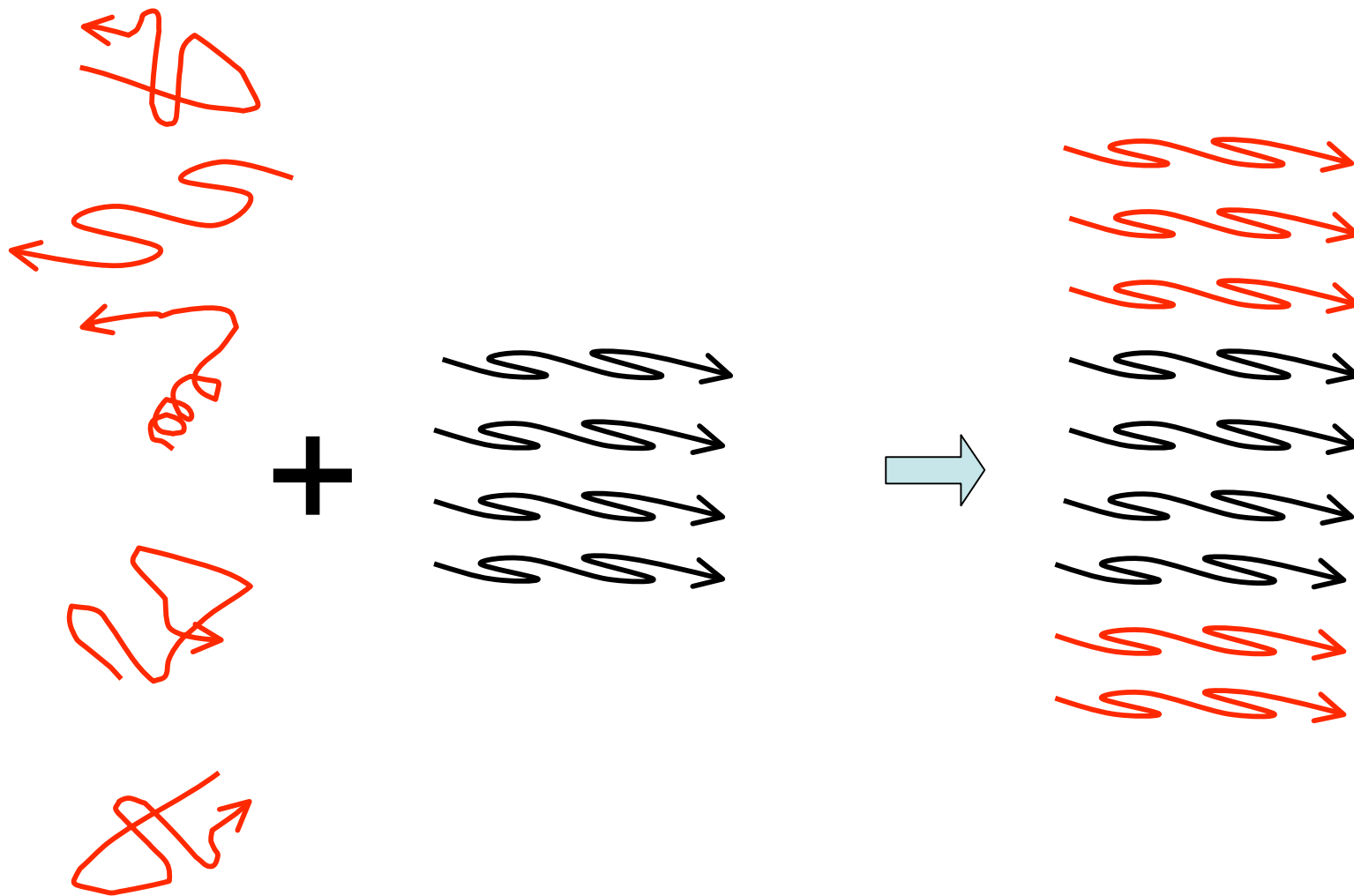
(in surface proteins of gram+ bacteria
and their bacteriophages)

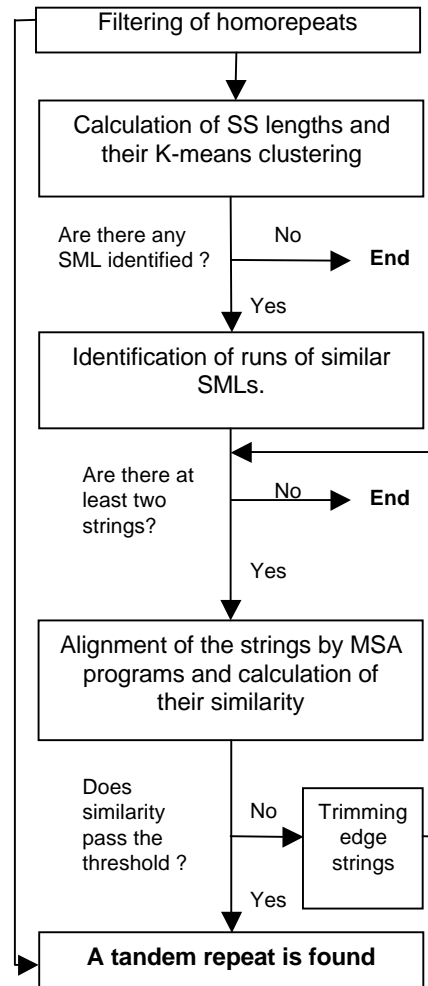
“Plasticity” of parallel superpleated beta-structures



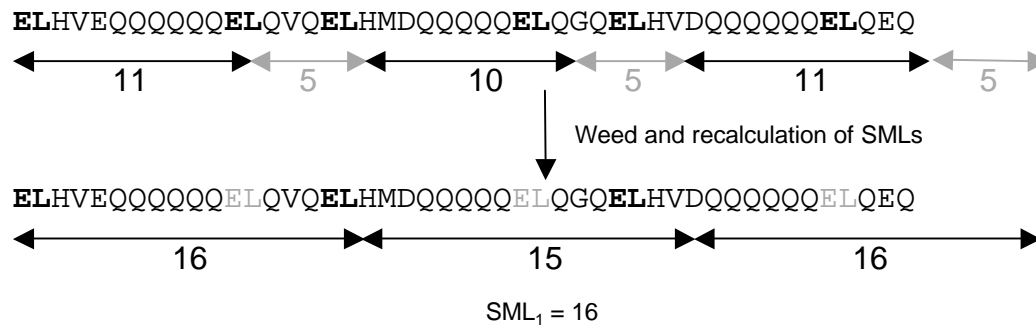
Shift of b-strands
Size of b-strands and loops
Number of b-strands

Parallel superpleated beta-structures as “infectious agent”.
Template-assisted fibril grow.

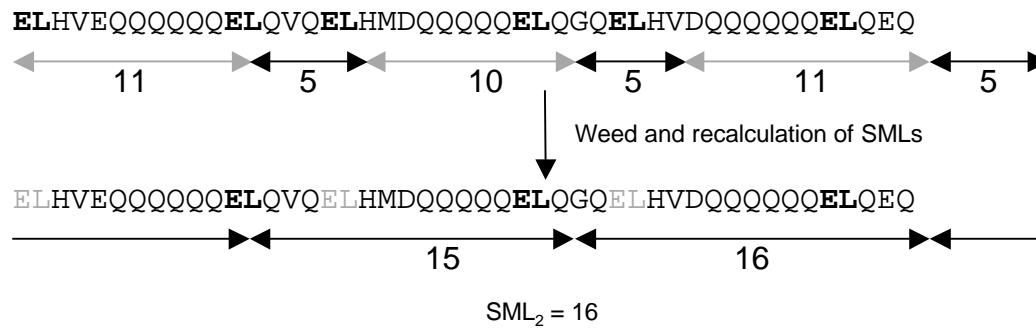


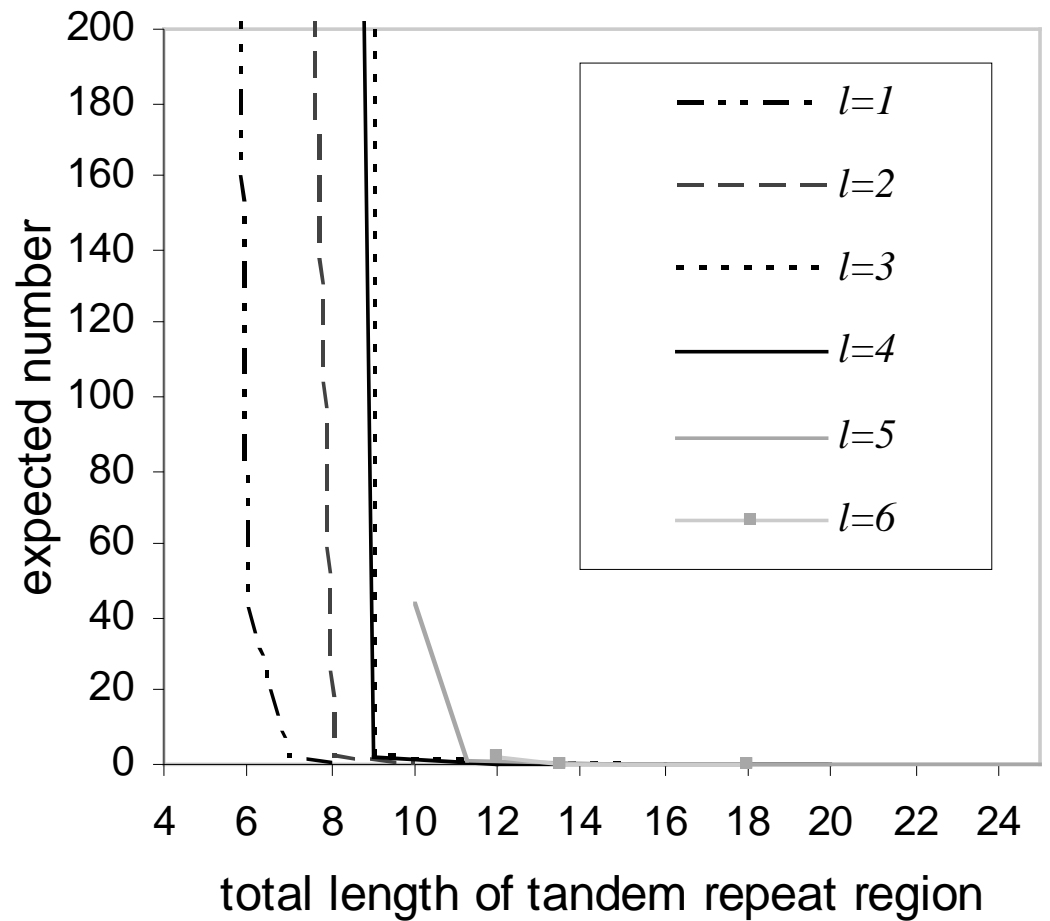


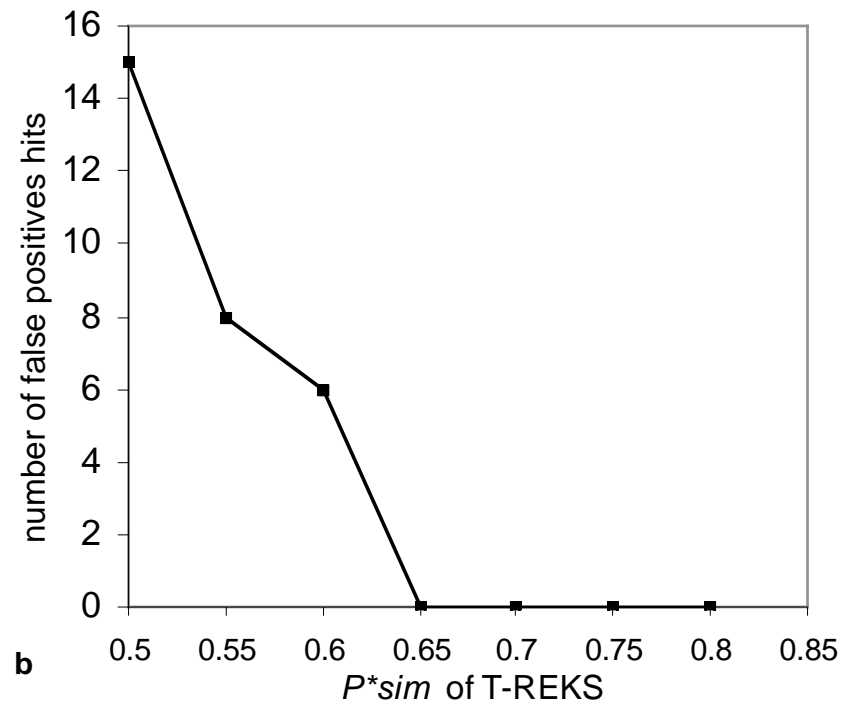
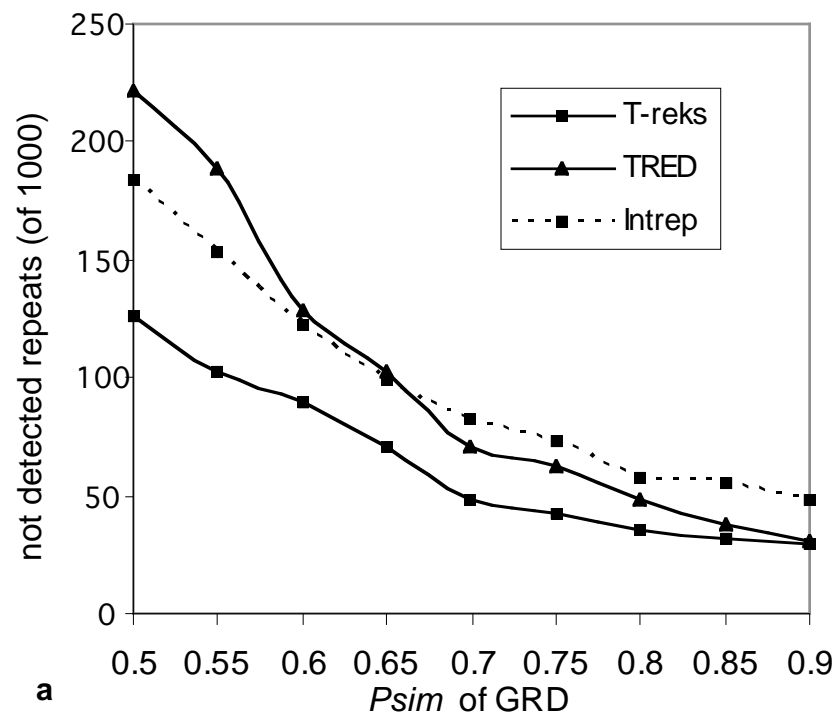
•Based on $SML_1 = 11$:

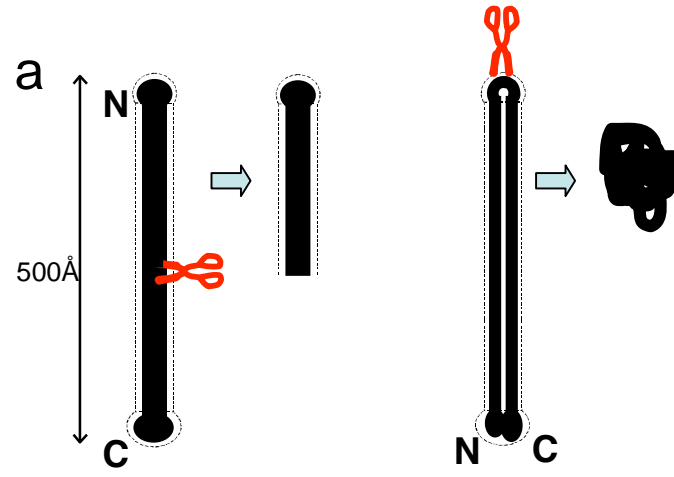


•Based on $SML_2 = 5$:

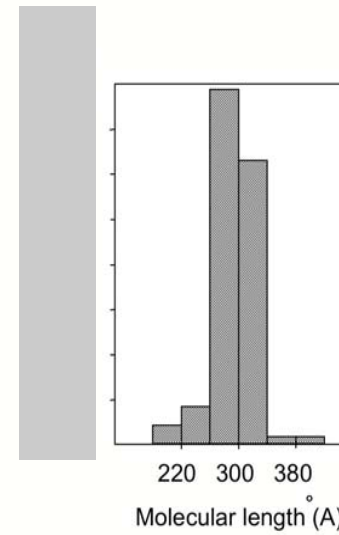
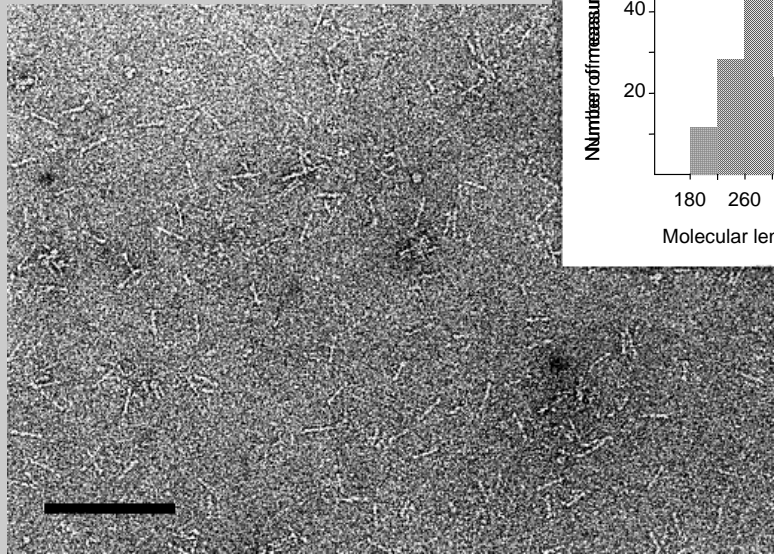




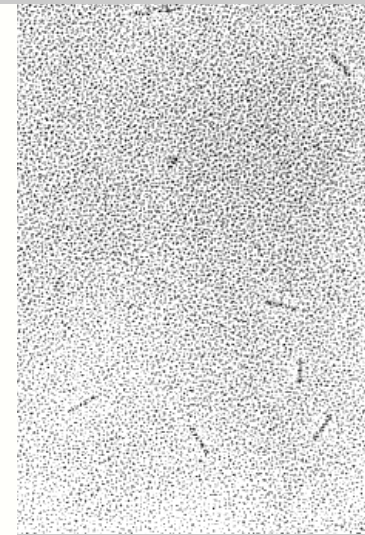




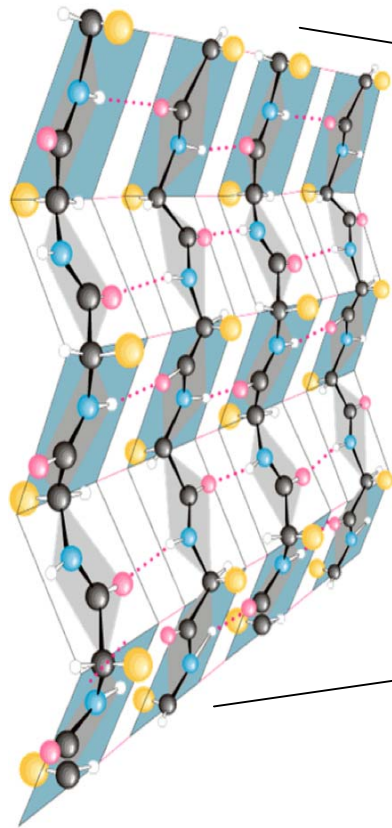
Negative staining of FHA44



Rotary Pt shadowing

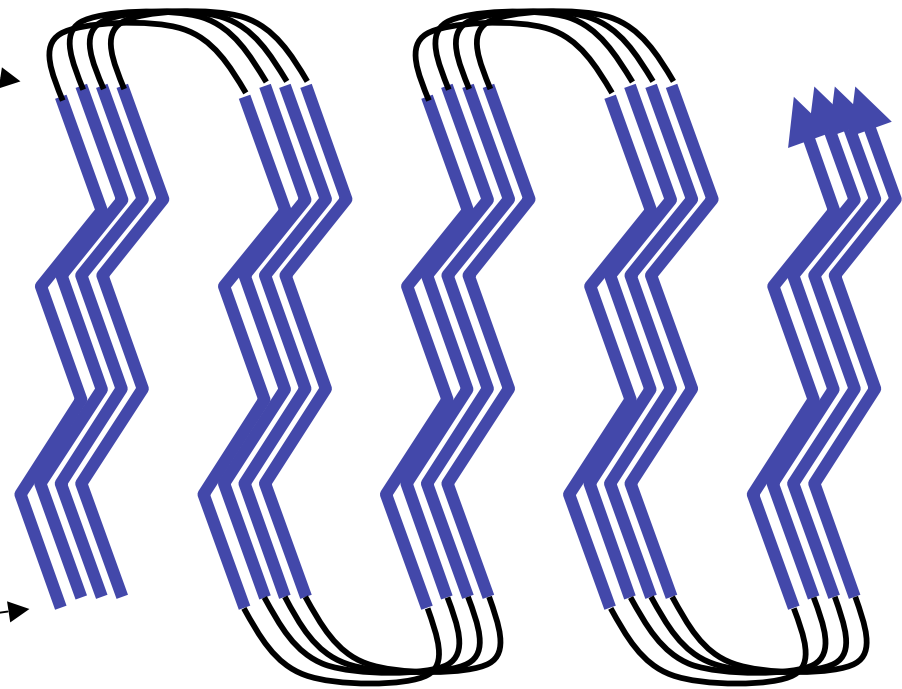


Kajava A, Cheng N, Kessel M, Simon M, Willery E, Jacob-Dubuisson, F, Loch C, Steven AC.
Mol Microbiol. 2001; 42(2):279



Canonical pleated β -structure

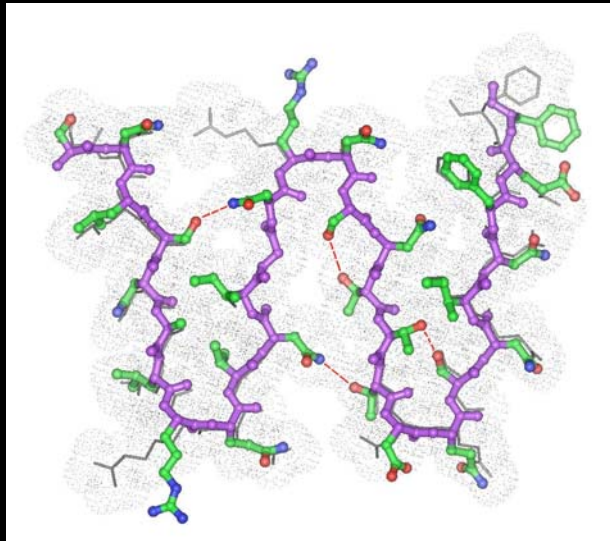
Adapted from Biochemistry 2nd Ed. by Garrett and Grisham



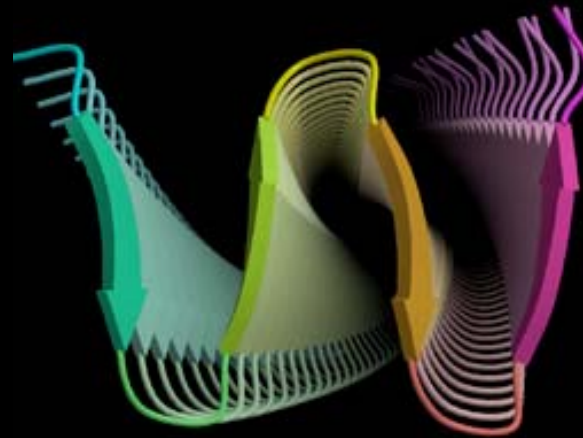
Superpleated β -structure

Structural fold for Ure2p prion domain

STEM + EM + X-ray fiber diffr + ssNMR

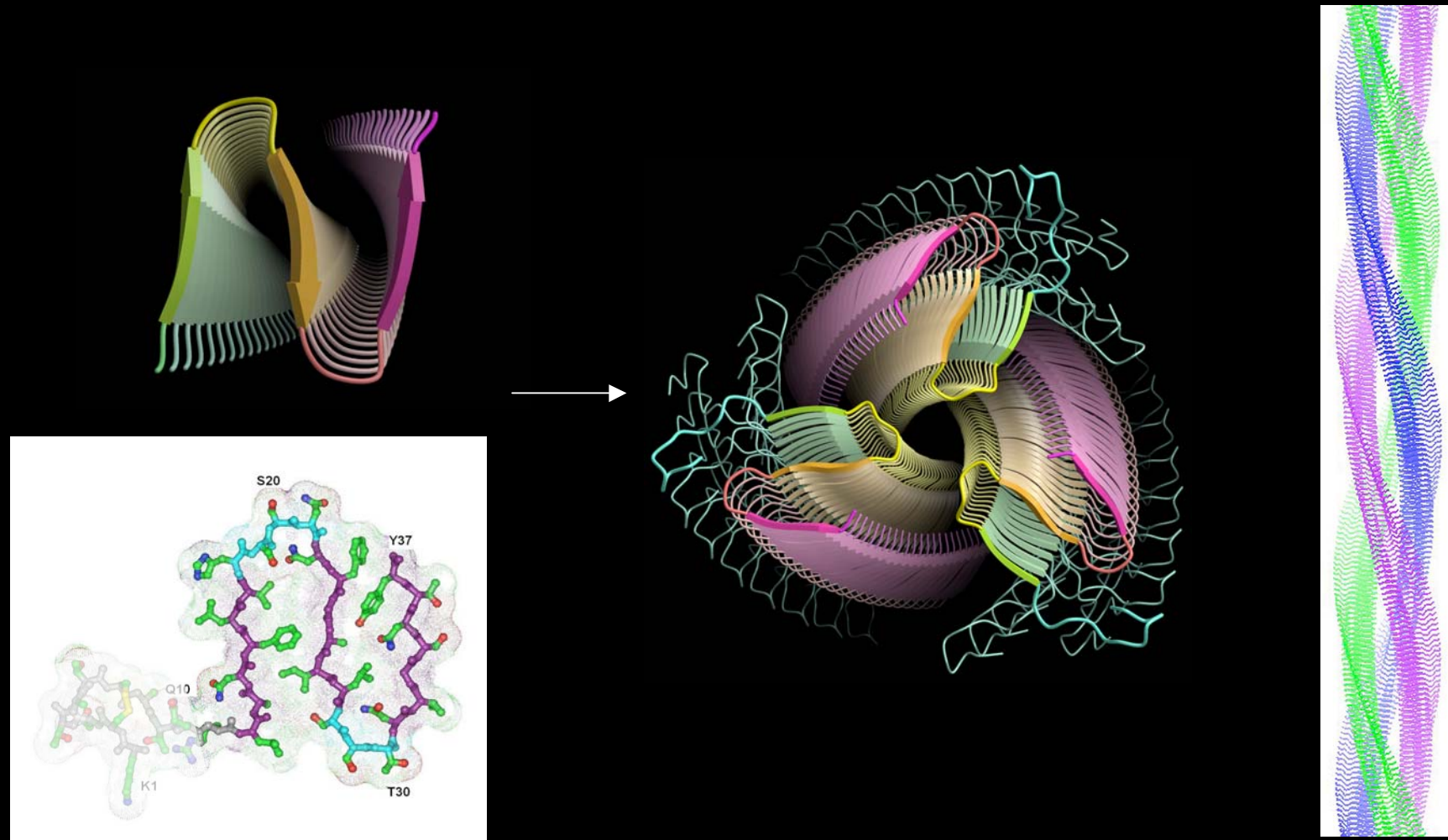


Ure2p(10-39)

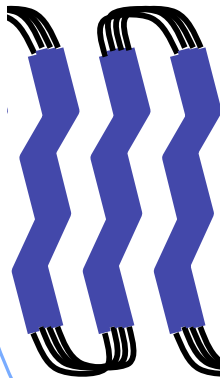


Amyloid Fibrils of Human Amylin

Human amylin is the major component of pancreatic amyloid deposits found in ~ 90% of persons with non-insulin-dependent (type 2) diabetes mellitus.



Applicability of the superpleated β -structure to other amyloids



Poly(Q) tracts (**Huntingtin disease**)

α -synuclein (**Parkinson's disease**)

(Der-Sarkissian et al., 2003, JBC, 278, 37530)

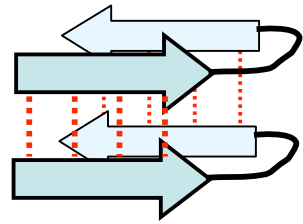
Tau protein (**Alzheimer's disease**)

(Margittai and Langen, 2004, PNAS, 101, 10278)

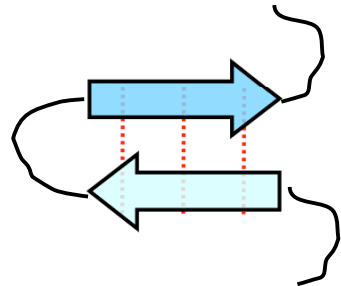
Prion domains of yeast proteins Sup35

(Shewmaker et al., PNAS. 2006103(52):19754)

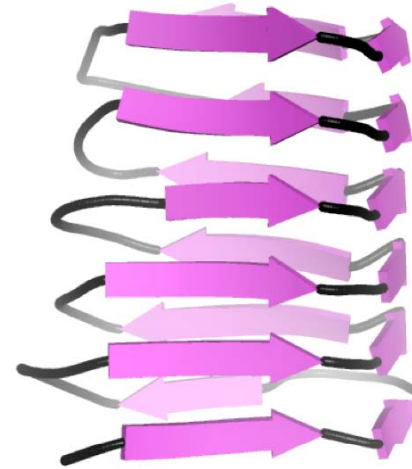
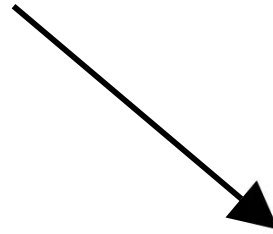
Prediction of amyloidogenicity of proteins



β -arches

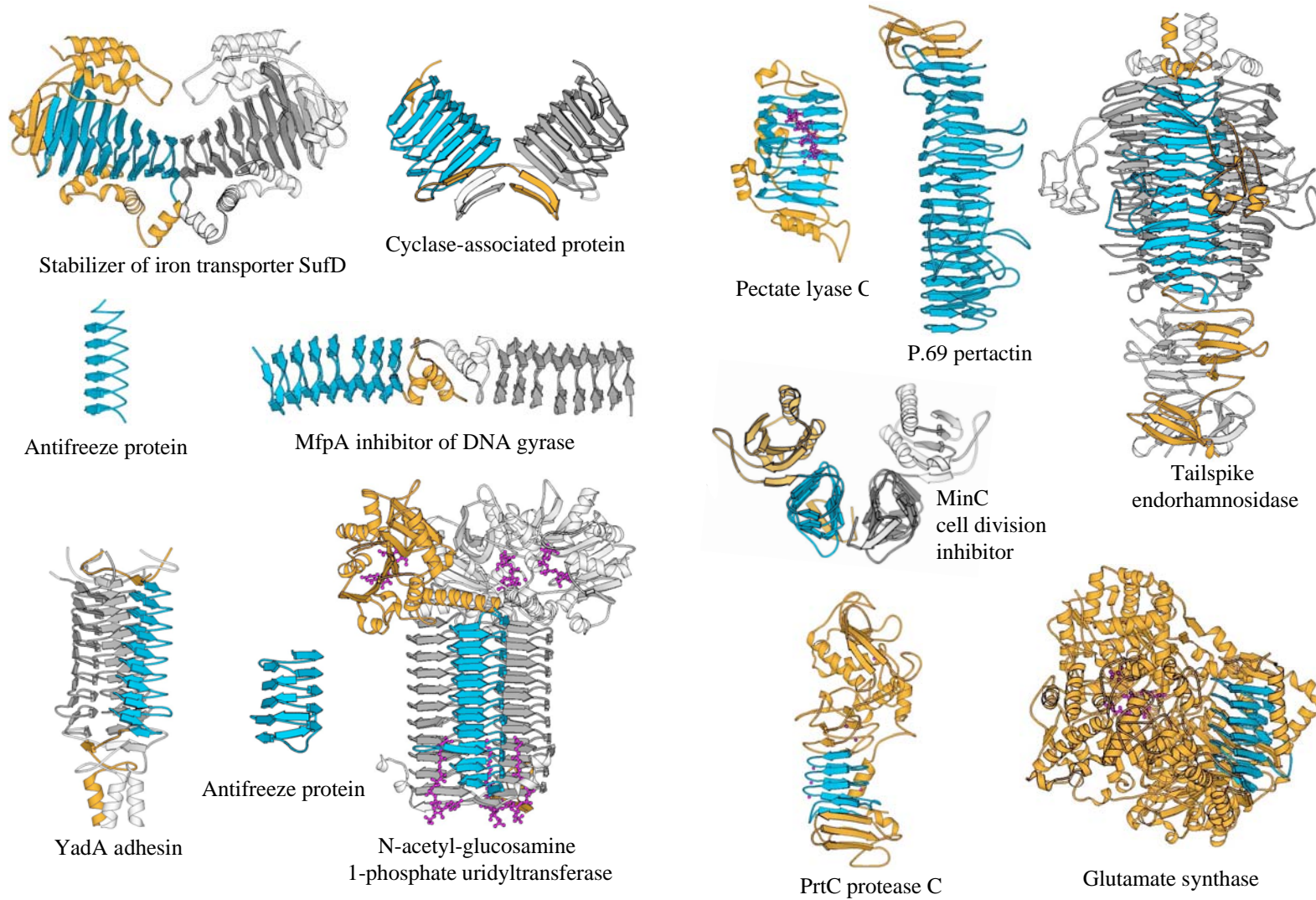


β -hairpin



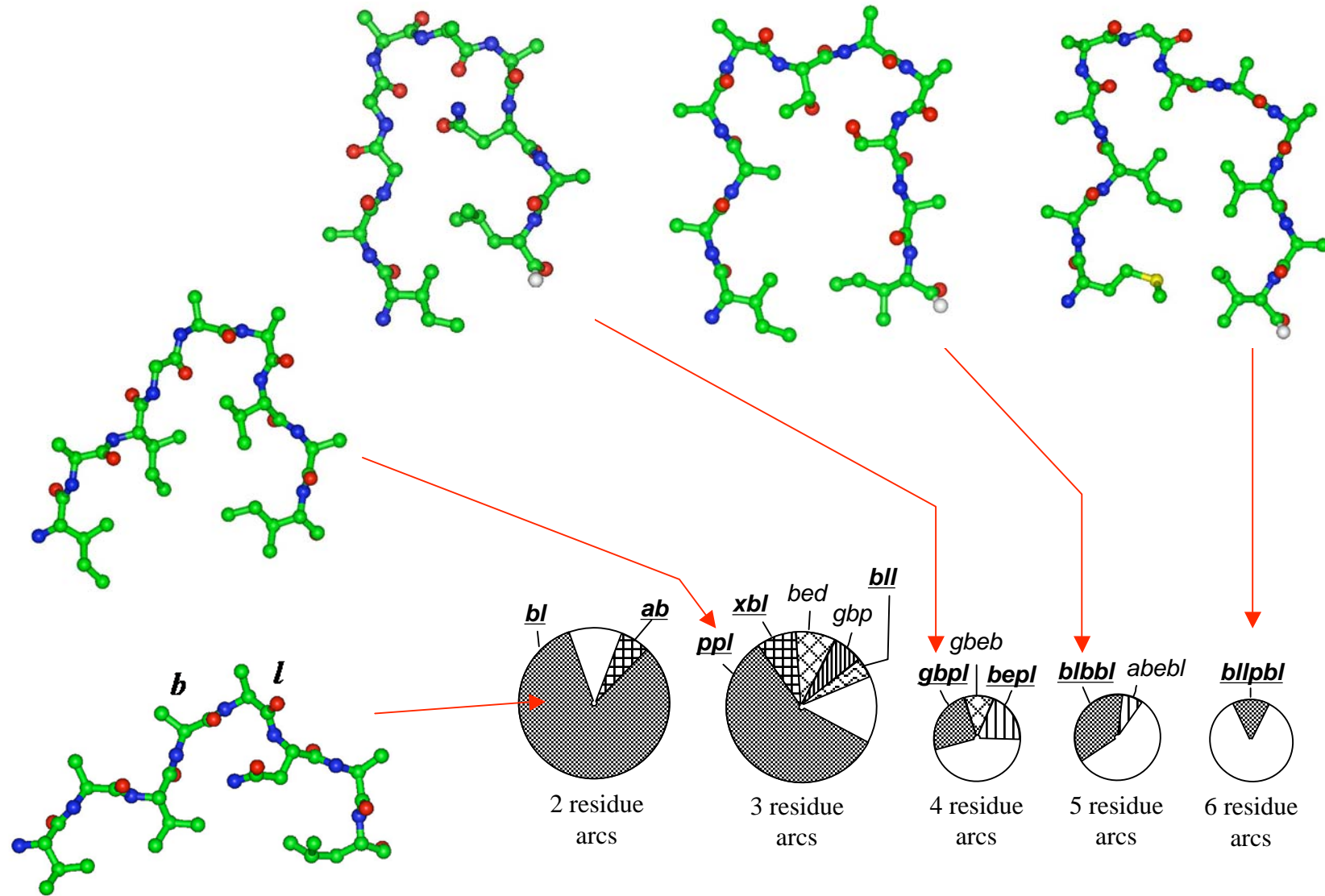
β -solenoid

Known β -solenoids



A.V. Kajava and A.C. Steven –Beta-helices, beta-rolls and the other beta-solenoid proteins (2006) *Advances in Protein Chemistry* 73:55-96.

Standard conformations of β -arches



Standard conformations of beta-arches in beta-solenoid proteins
 Hennetin, Julien, Stevenne and Kajava (2006) J.Mol.Biol., 358, 1094