



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Repeats and composition bias

Miguel Andrade

Faculty of Biology,
Johannes Gutenberg University
Mainz, Germany
andrade@uni-mainz.de

Repeats

Frequency

14% proteins contains repeats (Marcotte et al, 1999)

1: Single amino acid repeats.

E.g. QQQQQQQQQQQQQQ

2: Longer imperfect tandem repeats.
Assemble in structure.

Definition repeats

Sequence, length, imperfect, tandem

MRAVVKSPIMCHEKSPSVCSPLNMTSSVCSPAGINSVSSTTASF
GSFPVHSPITQGTPLTCSPNVENRGSRSHSPAHASNVGSPLSSP
LSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINN

Definition repeats

Sequence, length, imperfect, tandem

MRAVVK**SP**IMCHEKSPSVC**SP**LNMTSSVC**SP**AGINSVSSTTASF
GSFPVH**SP**ITQGTPLTC**SP**NVENRGSRSH**SP**AHASNVGSPLS**SP**
LSSMKSSIS**SP**PSHCSVKSPVS**SP**NNVTLRSSVS**SP**ANINN

Definition repeats

Sequence, length, imperfect, tandem

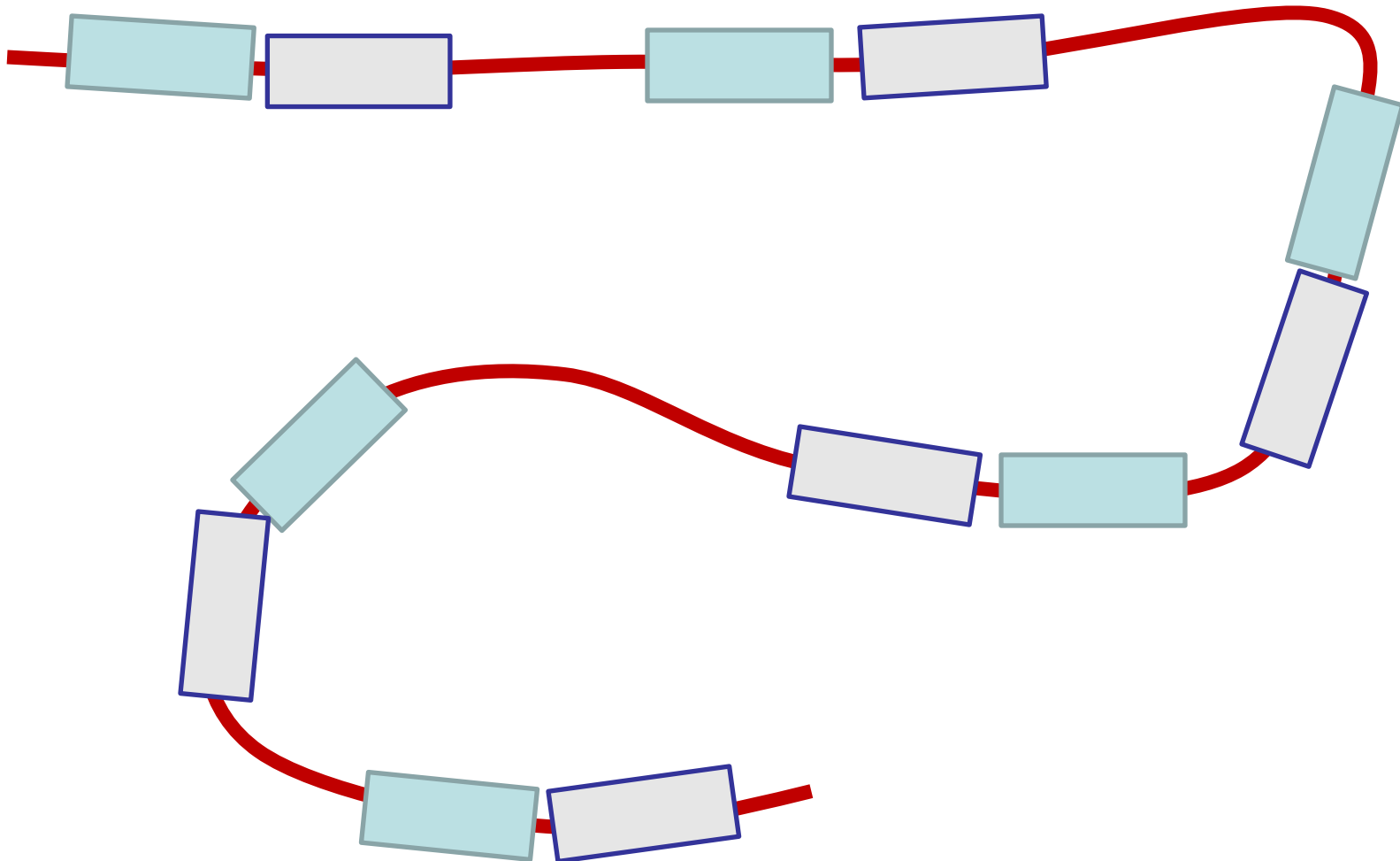
MRAVVK	SP	IM	CHE
KSPSVC	SP	LN	
MTSSVC	SP	AG	INSVSSTTASF
GSFPVH	SP	IT	Q
GTPLTC	SP	NV	EN
RGSRSH	SP	AH	ASN
VGSPLS	SP	LS	S
MKSSIS	SP	PS	HCS
VKSPVS	SP	NN	VT
LRSSVS	SP	AN	INN

Definition repeats

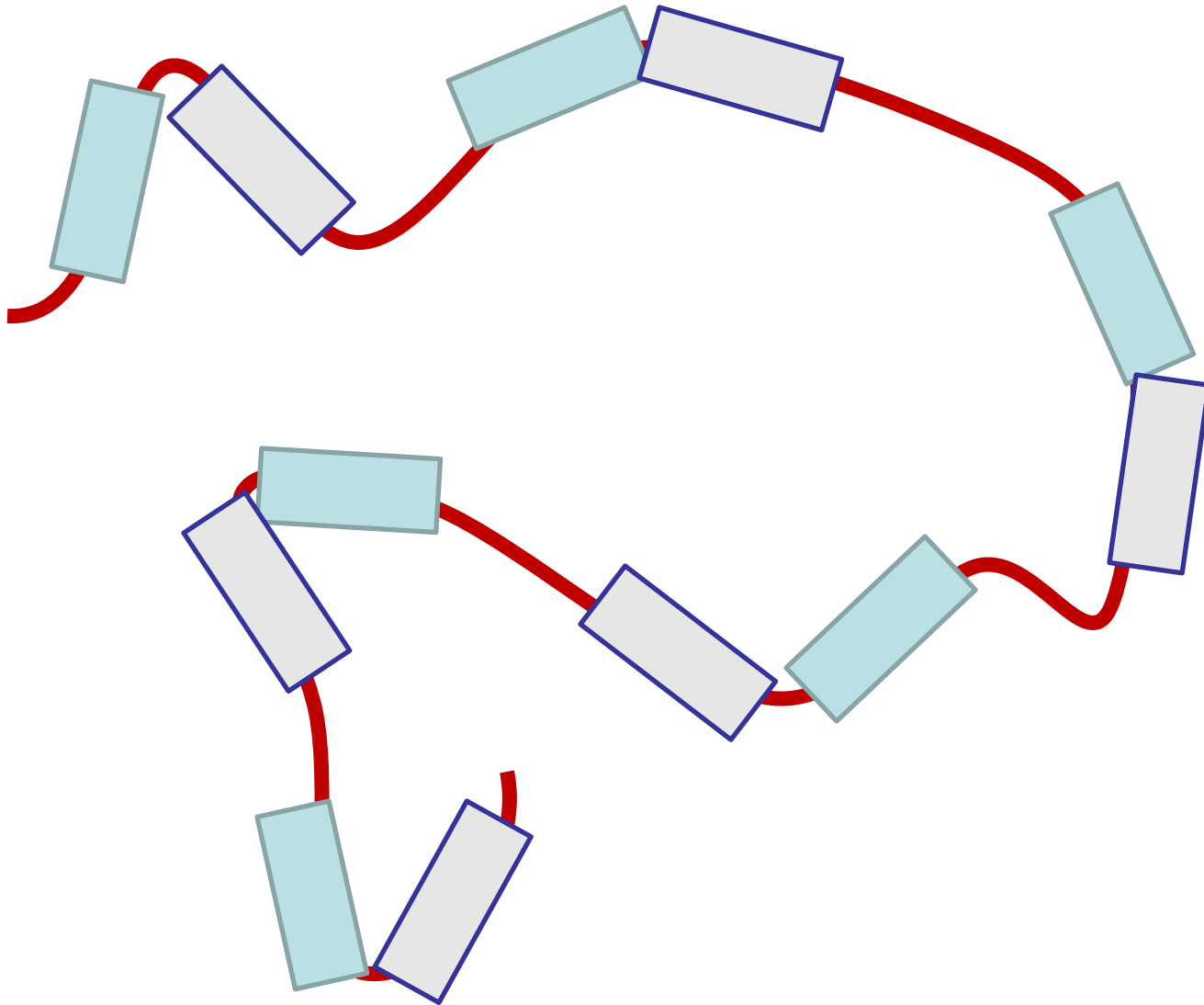
Sequence, length, imperfect, tandem

MRAV	V	K	SP	IM	CHE
KSPSV	C	SP	LN		
MT	S	V	C	SP	AG
INS	V	S	S	T	T
AS	F				
GS	F	V	H	SP	IT
Q					
GT	P	L	T	C	SP
EN					
RG	S	R	S	H	SP
ASN					
VG	S	P	L	S	SP
LS					
MK	S	S	I	S	SP
PS					
HCS					
VK	S	P	V	S	SP
NN					
VT					
LR	S	S	V	S	SP
AN					
INN					

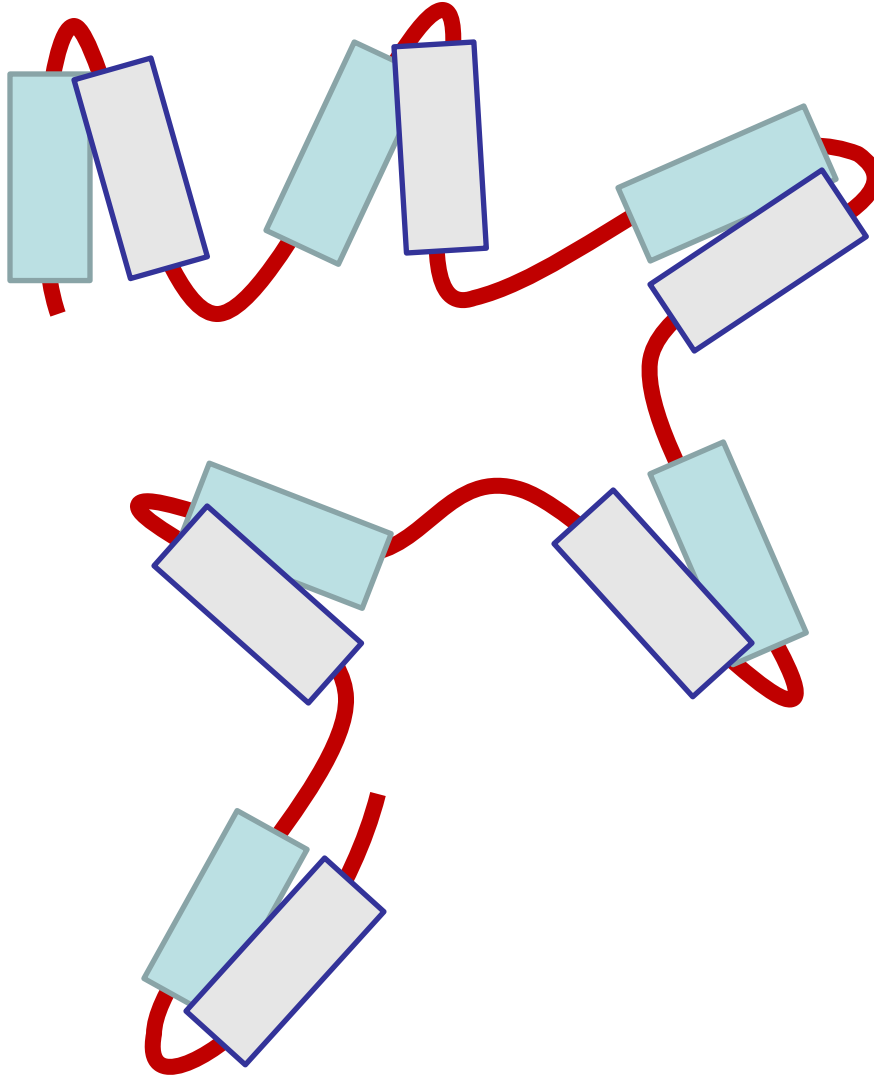
Tandem repeats fold together



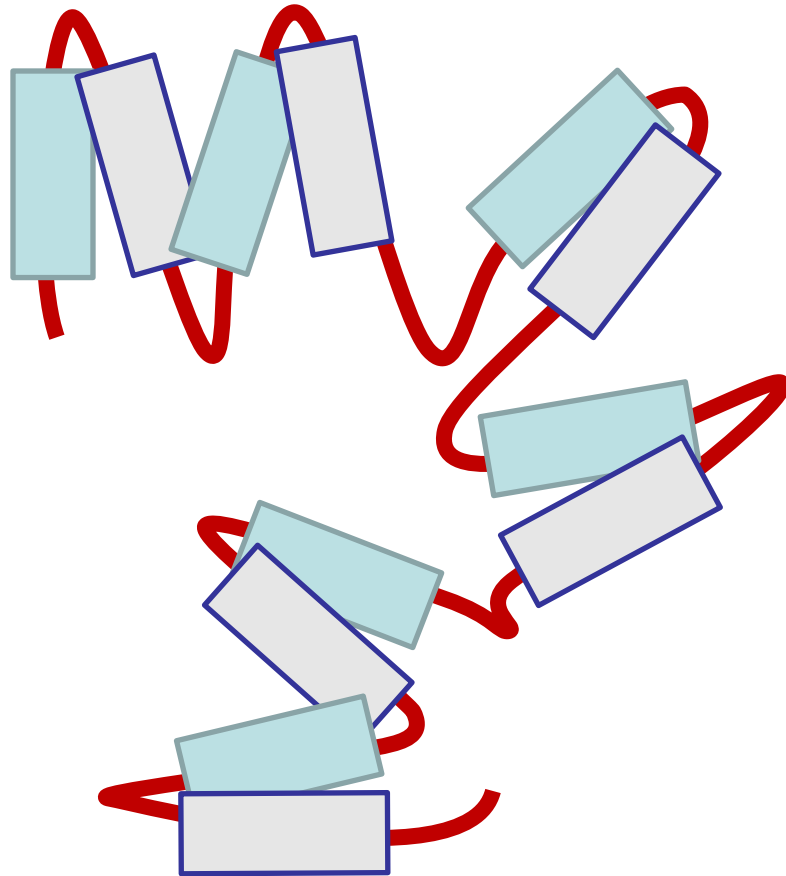
Tandem repeats fold together



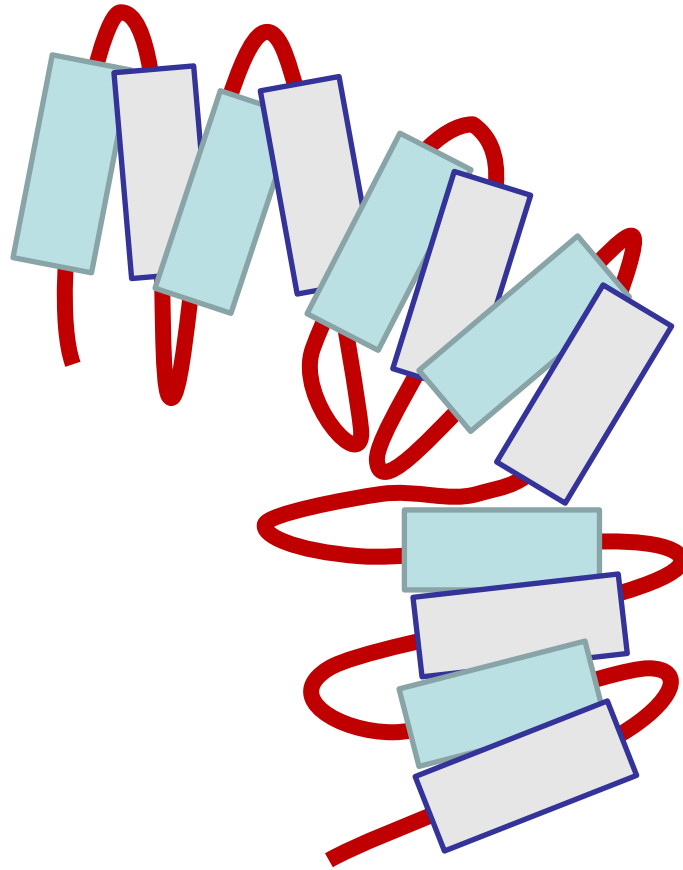
Tandem repeats fold together



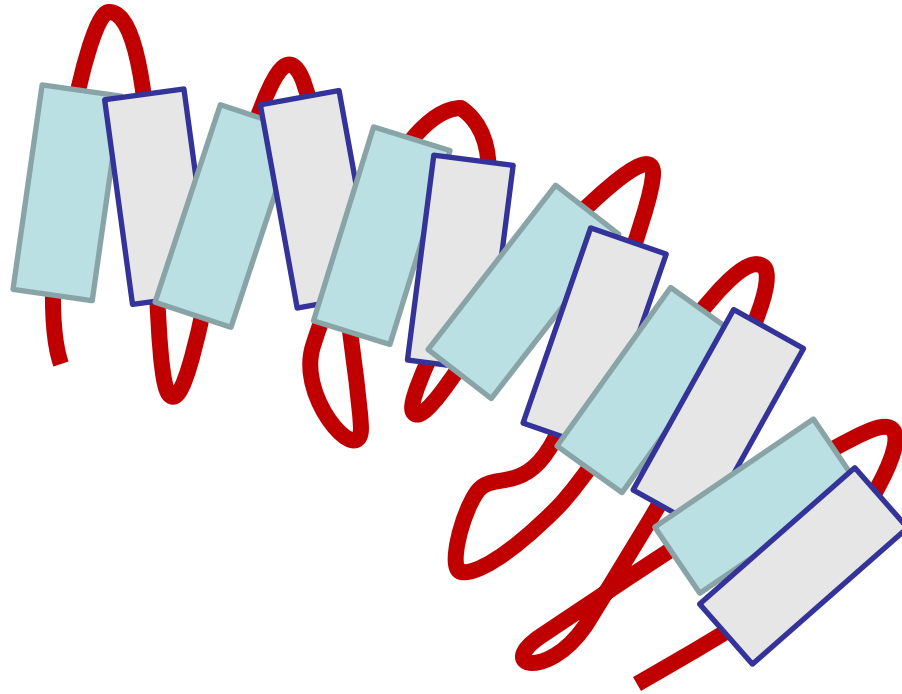
Tandem repeats fold together



Tandem repeats fold together



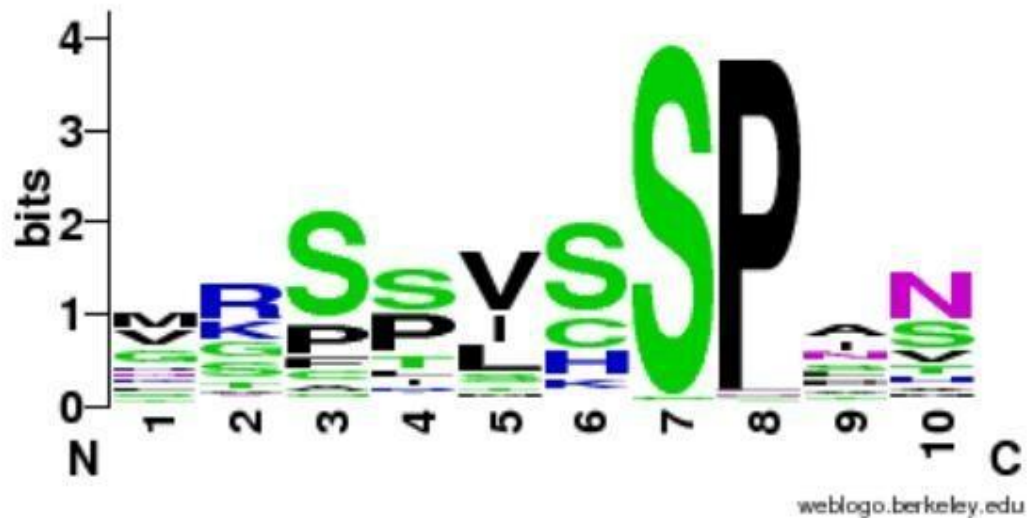
Tandem repeats fold together



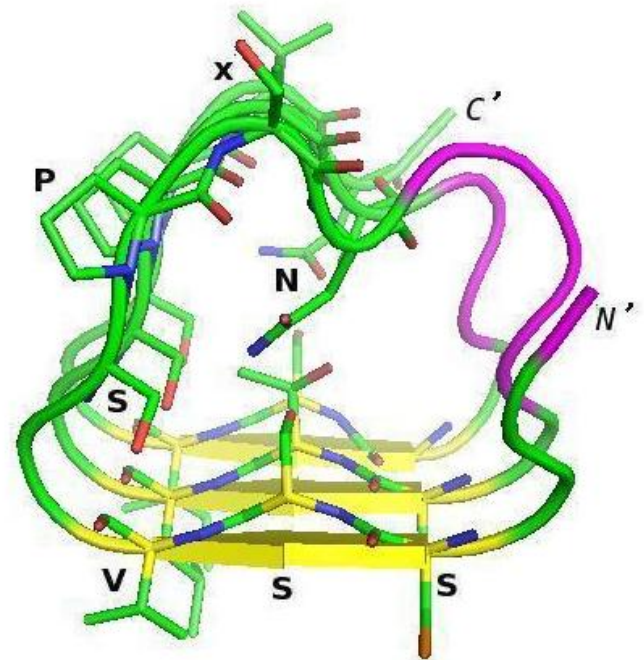
Definition repeats

Sequence, length, imperfect, tandem

MRAV	VK	SP	IM	CHE
KSPSV	C	SP	LN	
MT	SS	VC	SP	AG
INS	V	S	S	T
T	A	S	F	
GSFP	V	H	SP	IT
Q				
GTPL	T	C	SP	NV
EN				
RG	S	R	SH	SP
AH				ASN
VG	S	P	L	SS
LS				S
MK	S	S	I	SS
PS				HCS
VK	S	P	V	SS
NN				VT
LR	S	V	S	SP
AN				INN



<http://weblogo.berkeley.edu>



(Vlassi et al, 2013)

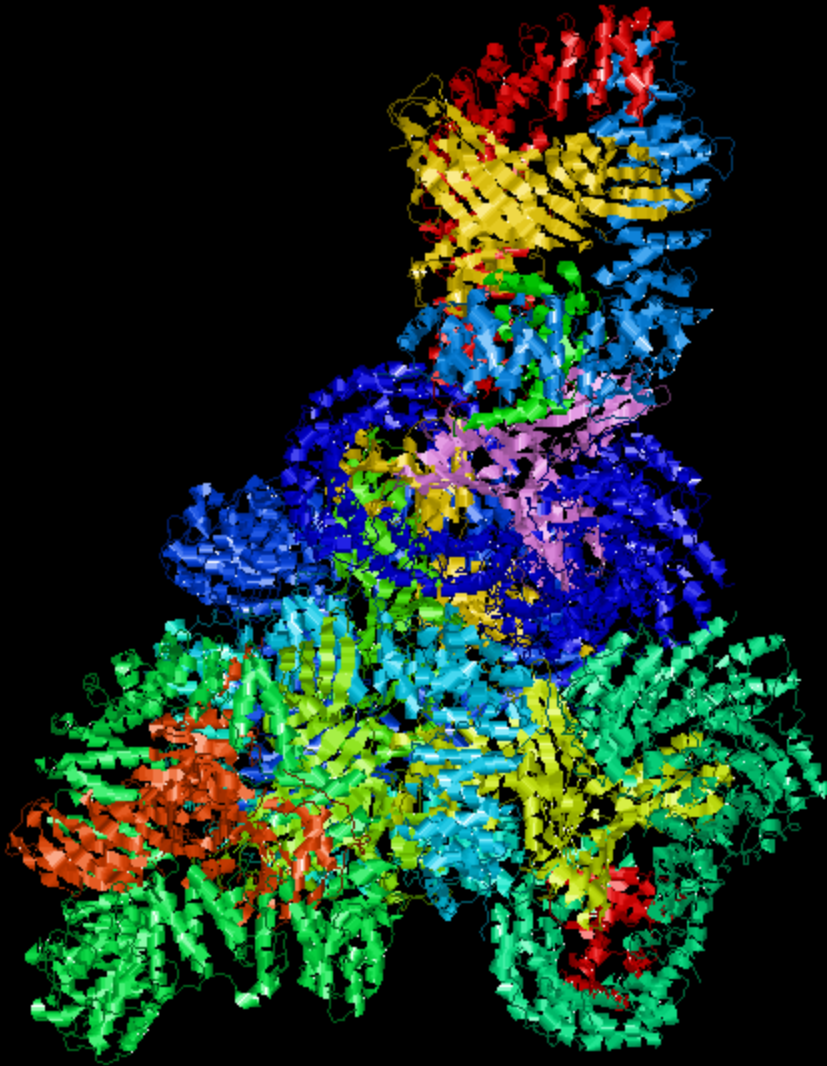
A subunit PP2A structure



PDB:1b3u

Groves et al. (1999) *Cell*

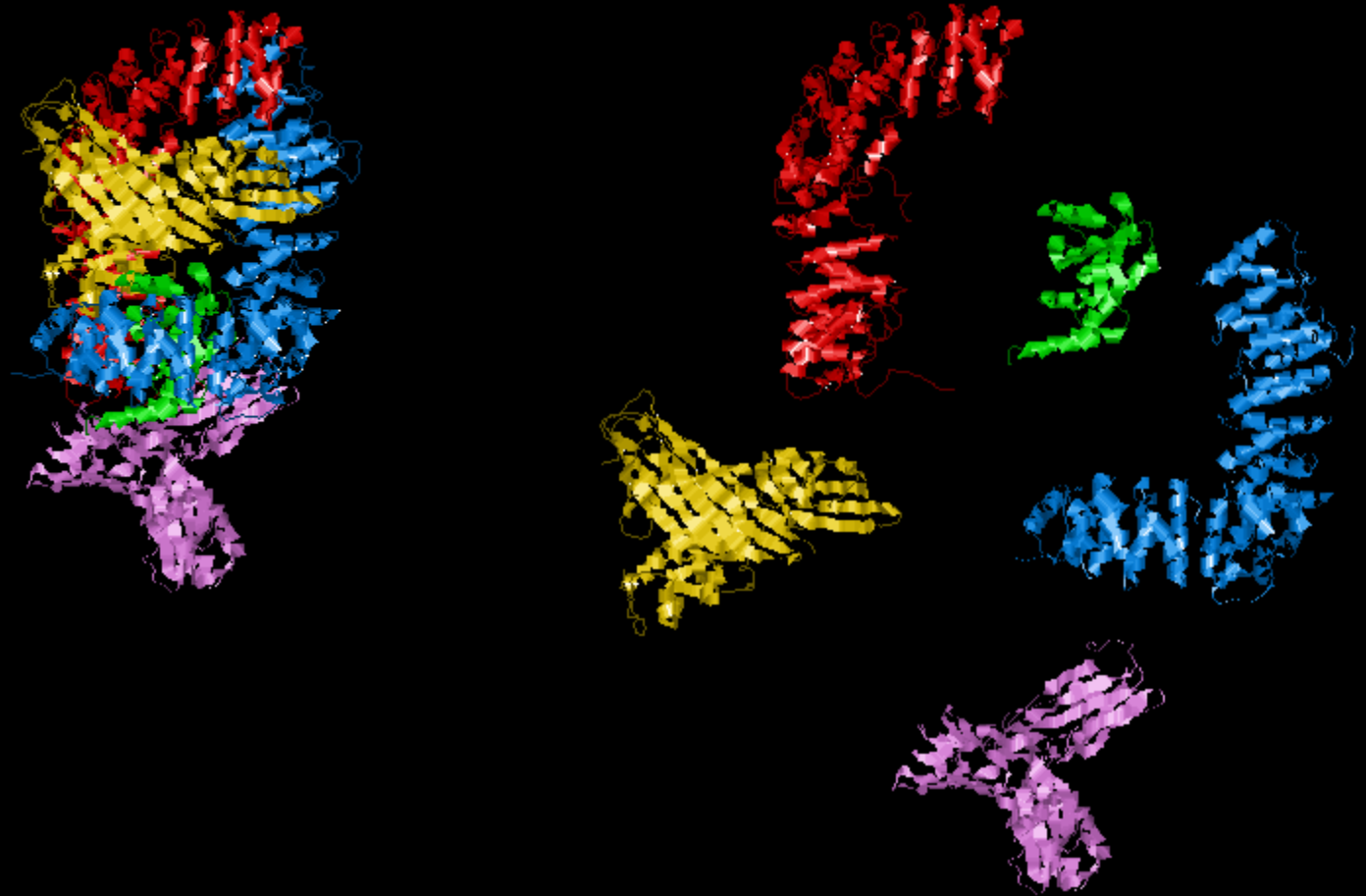
Ap1 Clathrin Adaptor Core



PDB:1w63

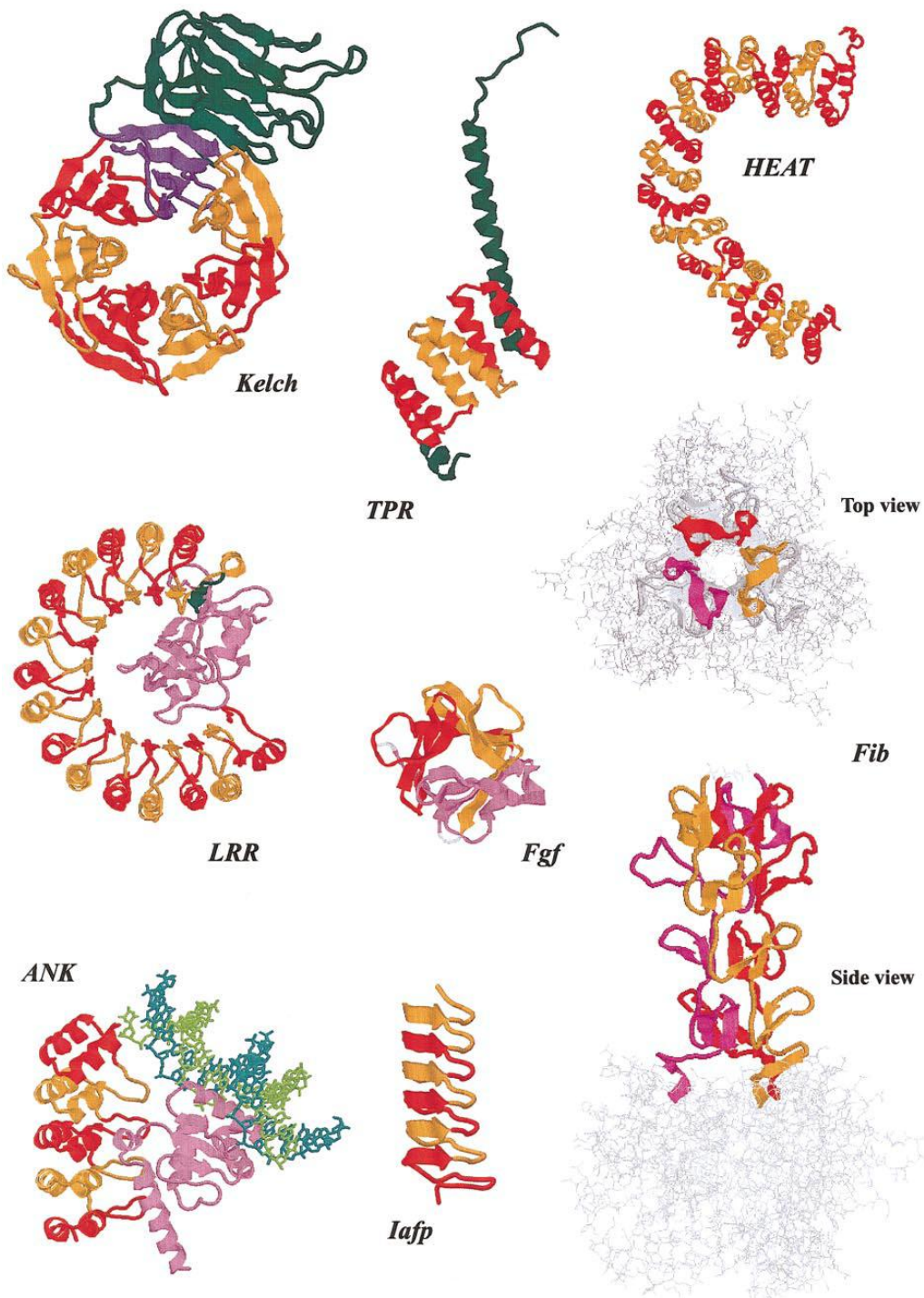
Heldwein et al. (2004) *PNAS*

Ap1 Clathrin Adaptor Core



PDB:1w63

Heldwein et al. (2004) *PNAS*



Andrade et al. (2001)
J Struct Biol

Detecting repeats in sequence

Often not straightforward.

N-terminal human Huntingtin.

Can you find **repeats**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
```

```
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP  
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQH SRRTQYFYSWLLNVLLGLLV  
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS PSAEQLVQVYEL  
TLHHTQH QDHN VVTGALELLQQLFRTPPPELLQT LTAVGGIGQLTAAKEESGGRSRSGSI  
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSES RSDVSSSALTASVKDEISGELAA  
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV  
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE  
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG  
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL
```

Detecting repeats in sequence

Often not straightforward.
N-terminal human Huntingtin.
Can you find **repeats**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSVDVSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLDNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detecting repeats in sequence

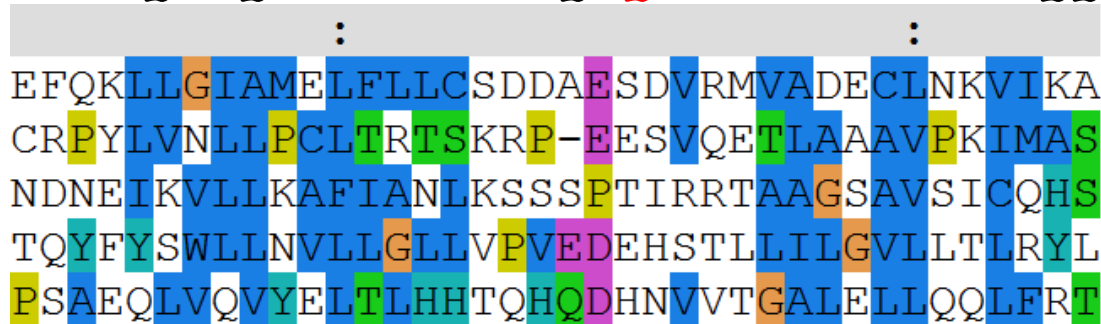
Often not straightforward.
N-terminal human Huntingtin.
Can you find **repeats**?

EFQKLLGIAMELFLLCSD**DA**ESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKR**P**-EESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKS**SS**PTIRRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVP**VE**DEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQ**HQ**DHNVVTGALELLQQLFRT

Detecting repeats in sequence

Often not straightforward.
N-terminal human Huntingtin.
Can you find **repeats**?

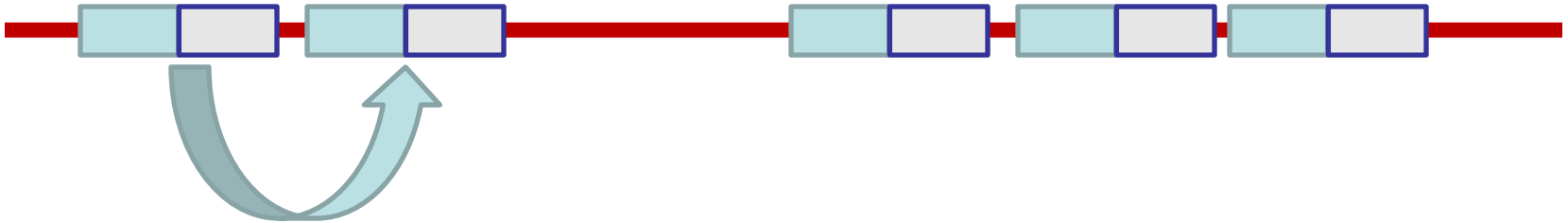
EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKSSSPTIRRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQHGDHNVVTGALELLQQLFRT



Detecting repeats in sequence

Dotplots

Comparing a sequence against itself



Detecting repeats in sequence

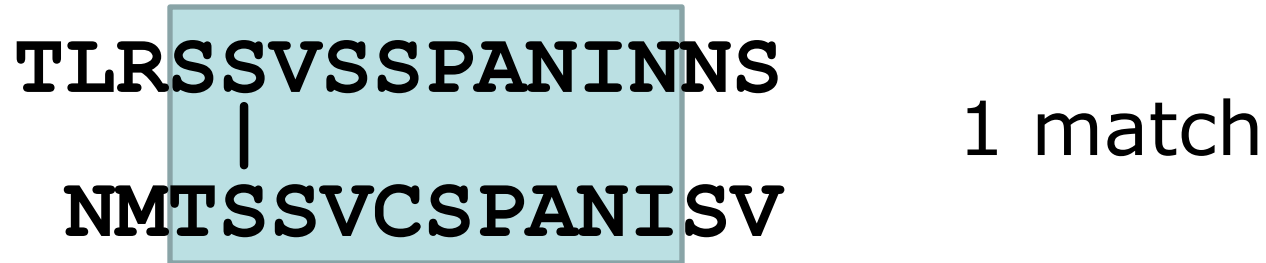
Dotplots

TLRSSVSSPANINNS

NMTSSVCSPANISV

Detecting repeats in sequence

Dotplots



Detection of repeats

Dotplots



TLRSSVSSPANINNS
| | | | | | | |
NMTSSVCSPANISV

8 matches

Detection of repeats

Dotplots



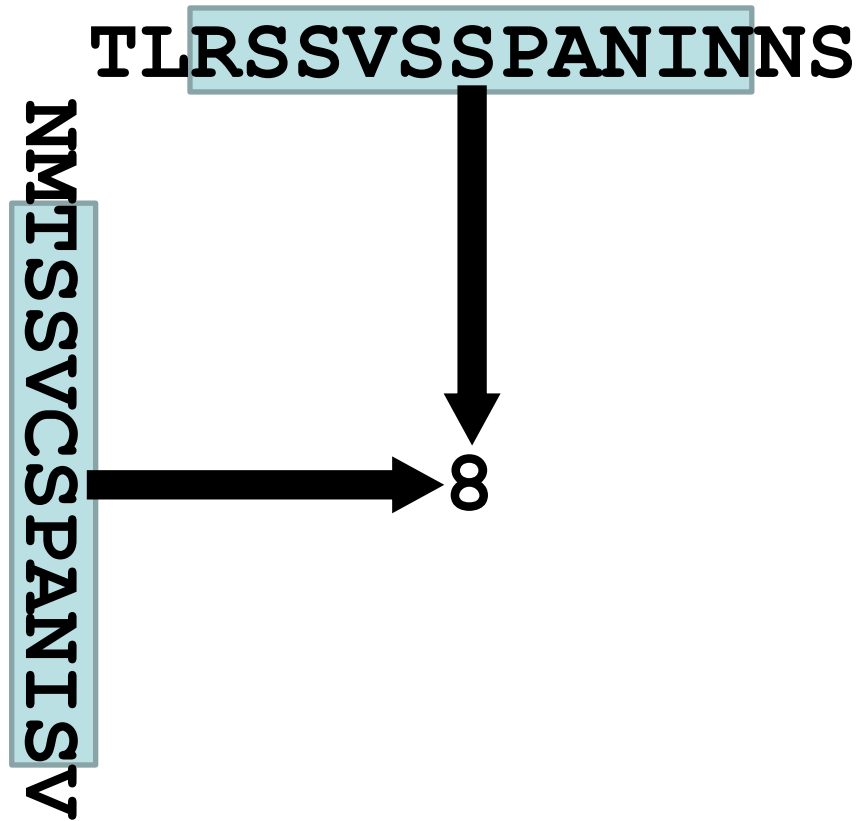
Detecting repeats in sequence

Dotplots



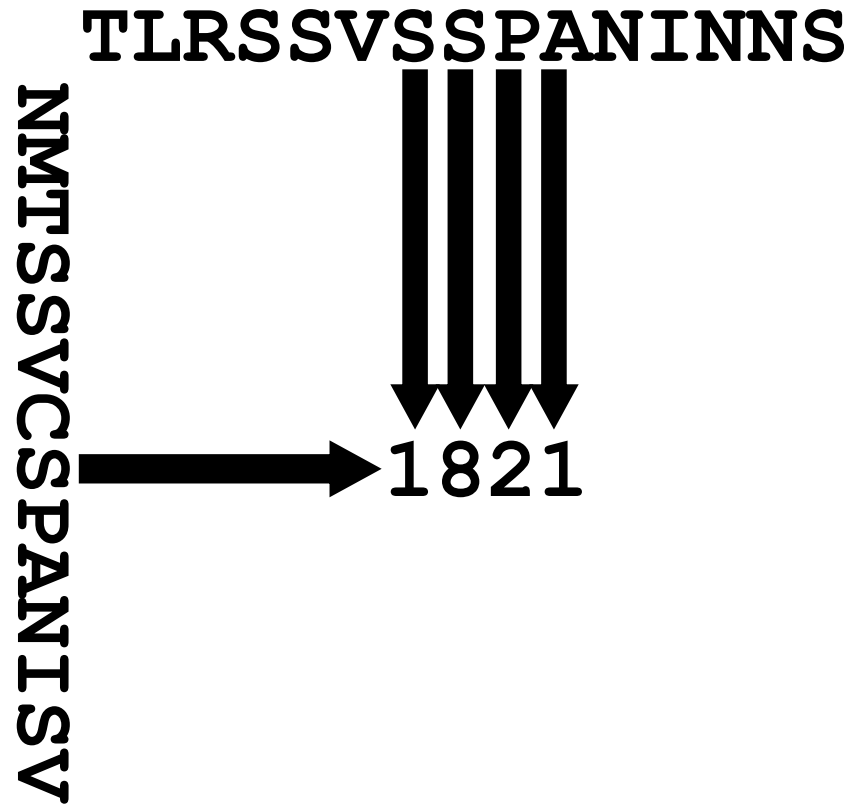
Detecting repeats in sequence

Dotplots



Detecting repeats in sequence

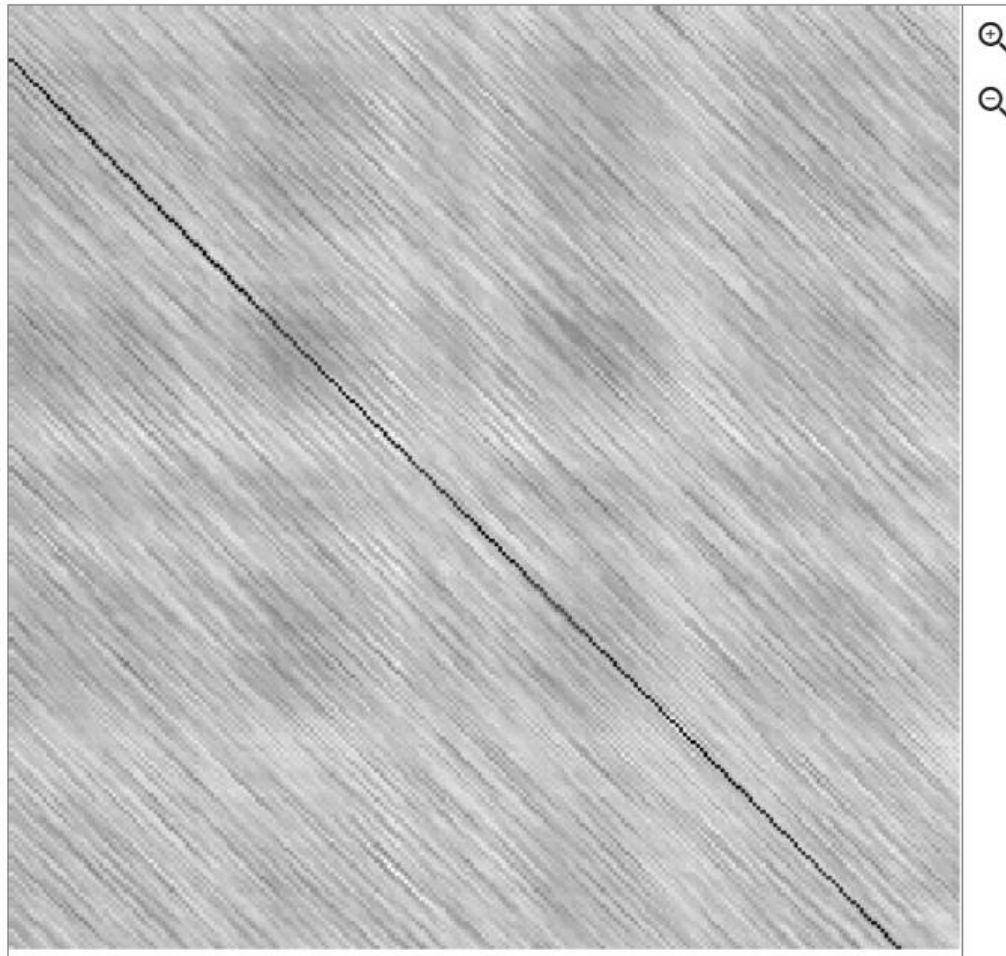
Dotplots



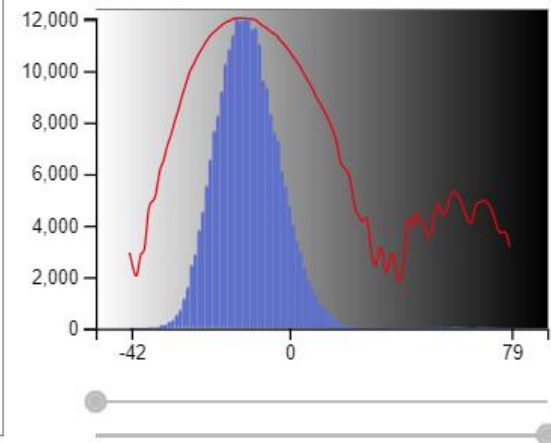
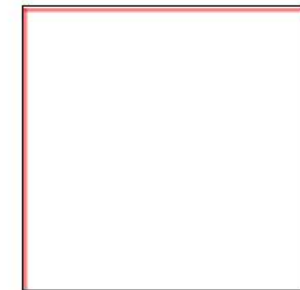


Sequence 1

Sequence 2



[246 x 244] # Score at (1:M, 1:R) : -8



Seq1:1

[

MTMDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHE
RKPLQTPTPIRRLWTMDTSELVQKAKLAEQAERYDDM

Exercise 1. Using Dotlet with the human mineralocorticoid receptor (MR)

- Go to the Dotlet web page:

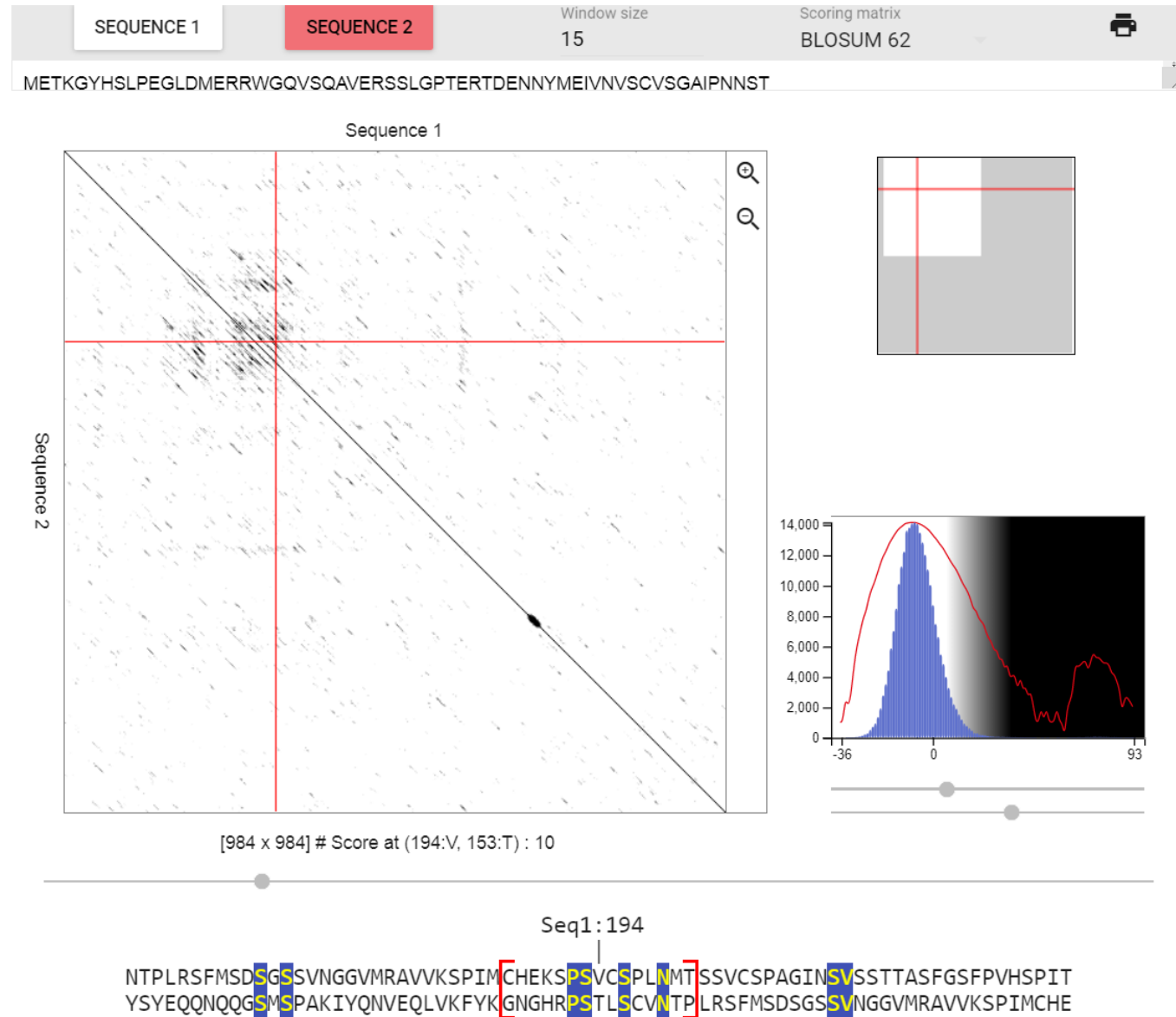
<http://dotlet.vital-it.ch>

- Click on the input button and paste the sequence of the human mineralocorticoid receptor (UniProt id P08235)

- Try to find combinations of parameters that show patterns in the dot plot (Hint: You can adjust this finely using the arrows)

- Find repetitions clicking in the diagonal patterns

Exercise 1. Using Dotlet with the human mineralocorticoid receptor (MR)



Detecting repeats in sequence

Using a multiple sequence alignment helps.
Conserved repeated patterns

```

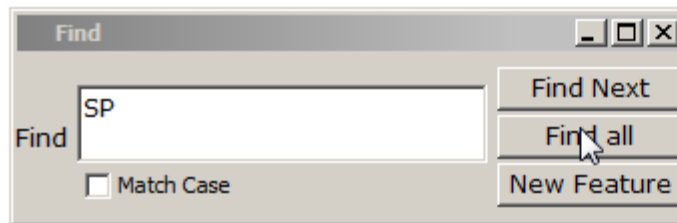
      240      250      260      270      280      290      300      310      320
mr_danio/1-970  - - - TYF - - DSDCP - TLDSATSSLTHCQHTSPNICSPVKSSIVGSPPLPSPLSVMKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTNG
mr_rutilus/1-973 - - - TYF - - DSDCP - SLASASTNLTGQHTSPNTCSPVKSSMVGSPPLASPLSVMKSPVSSPRSIGSVRSPLSC - - NTNMRSSVSSPTTNG
mr_cyprinus/1-971 - - - TFF - - DSDCP - SLASTHTNLIQGQHTSPNTCSPVKSSVVGSPPLASPLSVIKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTYG
mr_oryzias/1-994  TCFGPPQCSAVSSPVSQTS CAATLANIKRRNSVTCSPVESCTVGSPPLTSPNLNIMRSPMSSPHSMSSVRSPPCSTTCNIRSSVSSPT - - -
mr_takifugu/1-991 MCFGPMCSSVSSPVSQTS CAATLANIKRRNSATCSPVESSTVGSPPLTSPNLNIMRSPISSPQSMSSVRSPPCSTTCNIRSSVSSPT - - -
mr_oreochromis/1-994 TCFAPLCSSVSSPVSQTS CAATLANIKRRNSVTCSPVESSTVGSPPLTSPNLNIMRSPMSSPQSMSSVRSPPCSTTCNIRSSVSSPT - - -
mr_xenopus/1-979  - - FGNF - - TVHSPVNQVTPKSCSPHTDNRC SIAHSP - - AGTVES - PLSSPVSSMRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI -
mr_anolis/1-990   - - FGNF - - TVSSPVNQGTPLSCSPNIENRGSMLHSPPHASNMGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
mr_alligator/1-985 - - FGNF - - VVNSPINQGTPLSCSPNIENRGSMLHSPAHASNVGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
mr_teniopygia/1-981 - - FGNF - - SMHSPMGQGTPLSRSPNVENRGSMLHSPAHISNVGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANL -
mr_gallus/1-986   - - FGNF - - AMHSPIGQGTPLSRSPNVESRGSMLHSPAHVSNVGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
mr_monodelphis/1-993 - - FGSF - - PVHSPITQGTPLPCSPNVENRSSVSHSPAHASNVGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNVTMRSSVSSPANIN -
mr_mus/1-980     - - FGSF - - PVHSPITQGTSLTCSPSVENRGSRSHPVHASNVGS - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN -
mr_rattus/1-981  - - FGSF - - PVHSPITQGTSLTCSPSVENRGSRSHPHASNVS - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN -
mr_homo/1-984    - - FGSF - - PVHSPITQGTPLTCSRNAENRGSRSHPAHASNVGS - PLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN -
mr_equus/1-984   - - FGNF - - TVHSPITQGTPLTCSPNVENRGSRSHPAHASNVGS - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN -
```

JalView with Regular Expression searches

Detecting repeats in sequence

Using a multiple sequence alignment helps
Conserved repeated patterns

```
      240      250      260      270      280      290      300      310      320
nr_danio/1-970  - - - TYF - - DSDCP - TLDSATSSLTHCQHTSPNICSPVKSSIVGSPPLPSPLSVMKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTNG
nr_rutilus/1-973  - - - TYF - - DSDCP - SLASASTNLTQGHHTSPNTCSPVKSSMVGSPPLASPLSVMKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTNG
nr_cyprinus/1-971  - - - TFF - - DSDCP - SLASTHTNL IQGQHTSPNTCSPVKSSVVGSPPLASPLSVIKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTYG
nr_oryzias/1-994  TCFGPPQCSAVSSPVVSQTSCAATLANIKRRNSVTCSPVESCTVGSPPLTSPLNIMRSPMSSPHSMSSVRSPSCSTTCNIRSSVSSPT - - -
nr_takifugu/1-991  MCFGPMCSSVSSPVVSQTSCASTLPNIKRRNSATCSPVESSTVGSPPLTSPLNIMRSPISSPQSMSSVRSPSCSTTTCNIRSSVSSPT - - -
nr_oreochromis/1-994 TCFAPLCSSVSSPVVSQTSCAATLANIKRRNSVTCSPVESSTVGSPPLTSPLNVMRSPMSSPQSMSSVRSPSCSTTCNIRSSVSSPT - - -
nr_xenopus/1-979  - - FGNF - - TVHSPVNQVTPKSCSPHTDNRCIAHSP - - AGTVES - PLSSPVSSMRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI -
nr_anolis/1-990   - - FGNF - - TVSSPVNQGTPLSCSPNIENRGSMMLHSPPHASNMGSP - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
nr_alligator/1-985 - - FGNF - - VVNSPINQGTPLSCSPNIENRGSMMLHSPAHASNVGSP - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
nr_teniopygia/1-981 - - FGNF - - SMHSPMGQGTPLSRSPNVENRGSMMLHSPAHISNVGSP - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANL -
nr_gallus/1-986   - - FGNF - - AMHSPIGQGTPLSRSPNVESRGSMLHSPAHVSNVGSP - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
nr_monodelphis/1-993 - - FGSF - - PVHSPITQGTPLPCSPNVENRSSVSHSPAHASNVGSP - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNVTMRSSVSSPANIN -
nr_mus/1-980     - - FGSF - - PVHSPITQGTSLTCSPSVENRGRSRHSPVHASNVGSP - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN -
nr_rattus/1-981   - - FGSF - - PVHSPITQGTSLTCSPSVENRGRSRHSPTHASNVGSP - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN -
nr_homo/1-984    - - FGSF - - PVHSPITQGTPLTCSNAENRGRSRHSPAHASNVGSP - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN -
nr_equus/1-984   - - FGNF - - TVHSPITQGTPLTCSNVENRGRSRHSPAHASNVGSP - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
```



JalView with Regular Expression searches

Detecting repeats in sequence

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

Detecting repeats in sequence

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

- Regular Expressions:

[LS]P.A

matches L or S, followed by P, followed by
anything, followed by A

Detecting repeats in sequence

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

- Regular Expressions:

[LS]P.A

matches L or S, followed by P, followed by
anything, followed by A

Which one is not matched?

- LPTA, SPAA, LPPA, LPAP, SPLA**

Detecting repeats in sequence

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

- Regular Expressions:

[LS]P.A

matches L or S, followed by P, followed by
anything, followed by A

Which one is not matched?

- **LPTA**, **SPAA**, **LPPA**, **LPAP**, **SPLA**

Exercise 2. Using JalView with a MSA of the MR with orthologs

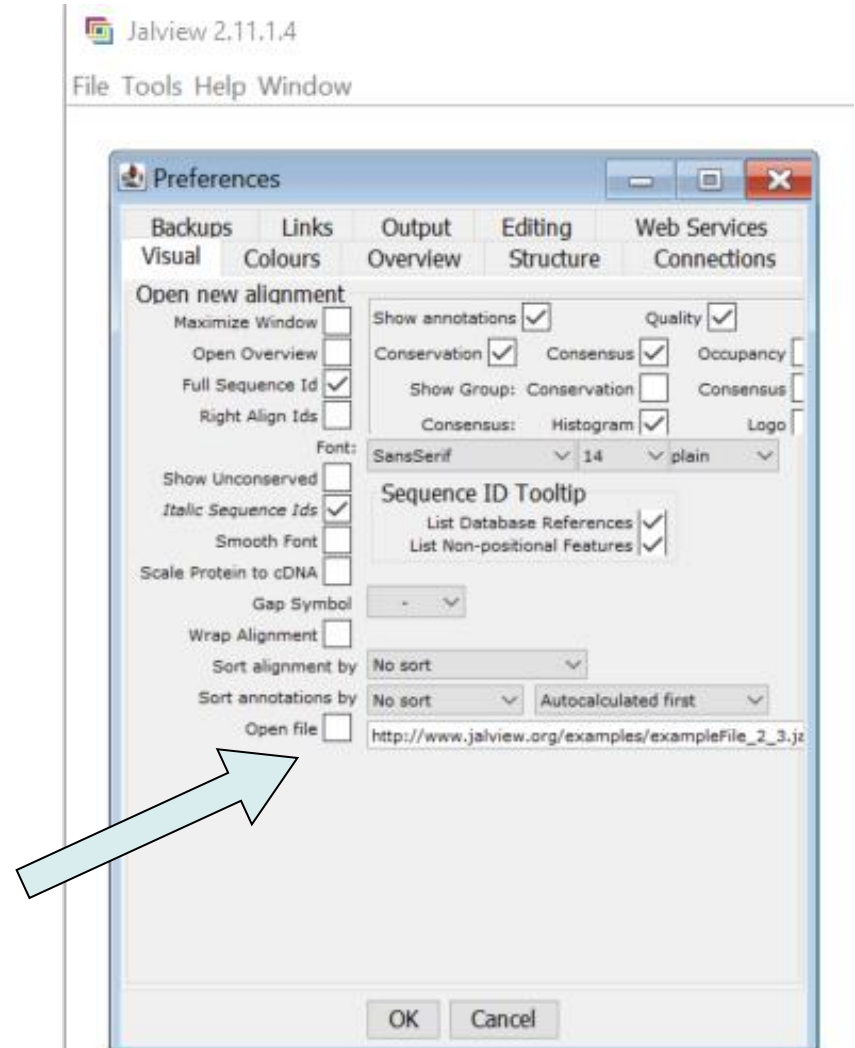
- Load the multiple sequence alignment of the MR in JalView: MR1_fasta.txt (from URL: https://cbdm.uni-mainz.de/files/2015/02/MR1_fasta.txt)
- Use the "Select > find" (or Ctrl+F) option with a regular expression and mark all matches (**click the "Find all" option!**)
- Search the motif "SP". Look for regions with repeats. Would you correct the alignment based on the hits? Experiment with other motifs.

First time running JalView?

Remove annoying start:

Go to
Tools > Preferences > Visual

Un-tick option "Open file"

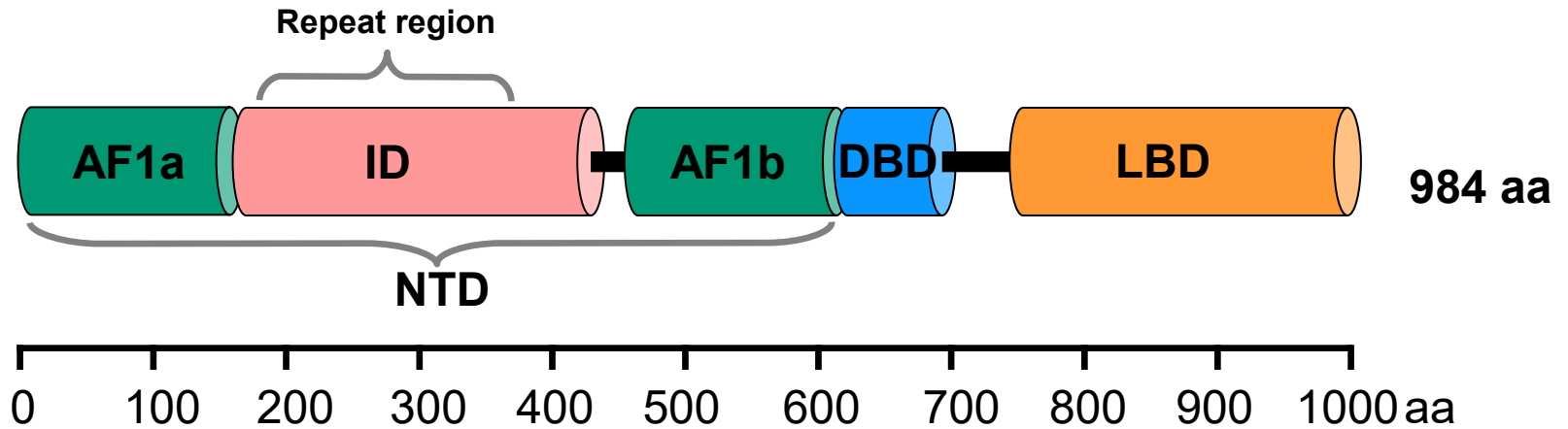


	#T1	#T2	#T3	#T4	#T5	#T6	#T7
..170.. ...180.. ...190.. ...200.. ...210.. ...220.. ...230.. ...240.. ...250.. ...260.. ..		*		*	*		*
158508572 Hsapiens	SGSSVNGGVMRAIVVKSPIMCHEKSPSVCSPLN	MTSSVCSPAGINSVSSTTA-SF	GSFPVHSPITQGTPLTCSNVENRGSRSHPAHA-SNVGSPLSPLSS				
31324675 Cjacchus	SGSSVNGGVMRAIVVKSPIMCHEKSPSVCSPLN	MTSSVCSPAGINSVSSTTA-SF	GSFPVHSPITQGTPLTCSNVENRGSRSHPAHA-SNVGSPLSPLSS				
126331313 Mdomestica	SGNSVNGSIMRSIVKSPIMCHEKSPSVCSPLN	MNSSVCSPAGINSVSSTTA-NF	GSFPVHSPITQGTPLPCSPNVENRSSVSHPAHA-SNVGSPLSPLSS				
73978292 Clupus	SGSSVNGGVMRAIVVKSPIMCHEKSPSVCSPLN	MTSSVCSPAGISSVSSTSA-SF	GSFTVHSPITQGTPLTCSNVENRGSRSHPAH---VGSPLSPLSS				
301763180 Amelanoleuca	SGSSVNGGVMRAIVVKSPIMCHEKSPSVCSPLN	MTSSVCSPAGISSVSSTSA-SF	GSFTVHSPITQGTPLTCSNVENRGSRSHPAHA-SNVGSPLSPLSS				
6981208 Rnorvegicus	SAASMNGGALRAIVKSPIIICHEKSSSVSSPLN	MASSVCSPVGINSMSSSTT-SF	GSFPVHSPITQGTSLTCSNVENRGSRSHPHTHA-SNVGSPLSPLSS				
144227212 Mmusculus	SGTSMNGGALRAIVKSPIIICHEKSPSVCSPLN	MPSSVCSPAGINSMSSTT-SF	GSFPVHSPITQGTSLTCSNVENRGSRSHPVHA-SNVGSPLSPLSS				
148224443 Xlaevis	SDKSMNGKIKSNTVKSPLSYSSEKNLVSGSPSV	MALPVCSPGTGISSTSCST--TF	GNFTVHSPVNCVTPKSCSPHTDNRCIAHSPAGT---VESPLSSPVSS				
327274009 Acarolinensis	CANSMNGSIMPIMKNPRITQERSPDPQSQS	MTSSVCSPGINSVTSTPTPTNF	GNFTVSSPVNCGTPLSCSPNIENRGSMLHSPPHA-SNMGSPLSPLSS				
115529242 Tguttata	SGSTVNGGAMHTIVKSPIMCQKSPSGCSPQN	MASSVCSPAGVNSMSSTTA-SF	GNFMSHSPMCGTPLSRSPNVENRGSMLHSPAH-SNVGSPLSPLSS				
225936142 Ggallus	SGSAMNGGAMHAVVKSPIVCQKSPSGCSPQN	MASSVCSPAGVNSVSSTTA-NF	GNFAMHSPICGTPLSRSPNVESRGSMLHSPAHV-SNVGSPLSPLSS				
239923135 Rrutilus	GSHQPSGGPQECAVVSAVPSGMAVLSGSSDGS	GPGPMSSPTGHNMVSTTSPTYFDSDCPSLASASTNLTQGH--TSPNTCSPVKSSMVGSPPLASPLSV					
154240734 Drerio	GAQLPNGGPQECVVSDSVPSALVTALSSSTD-	GSCPMSSPTGHNMVSTTSPTYFSDCPTLDSATSSLTHCQH--TSPNICSPVKSSIVGSPPLPLSV					
		#F1	#F2			#F3	#F4

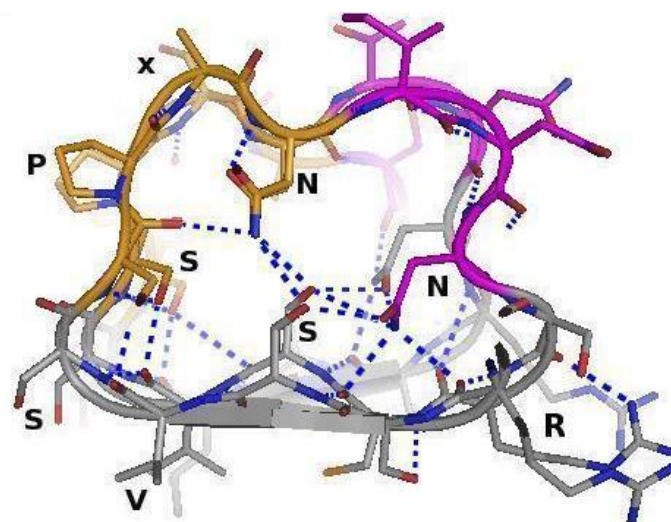
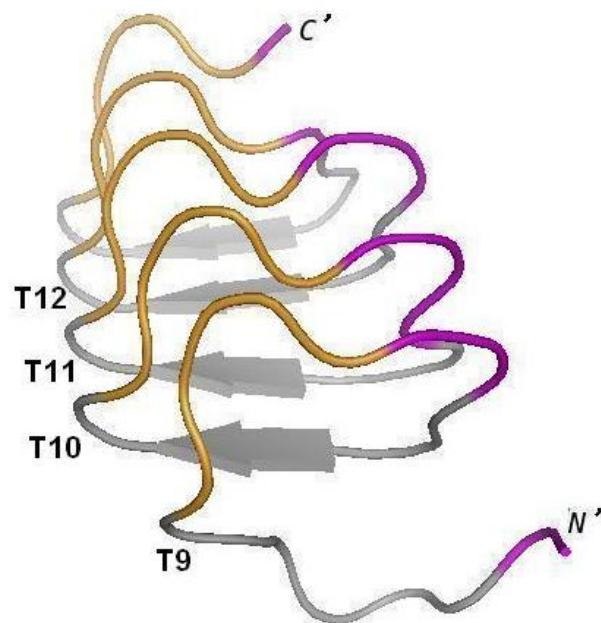
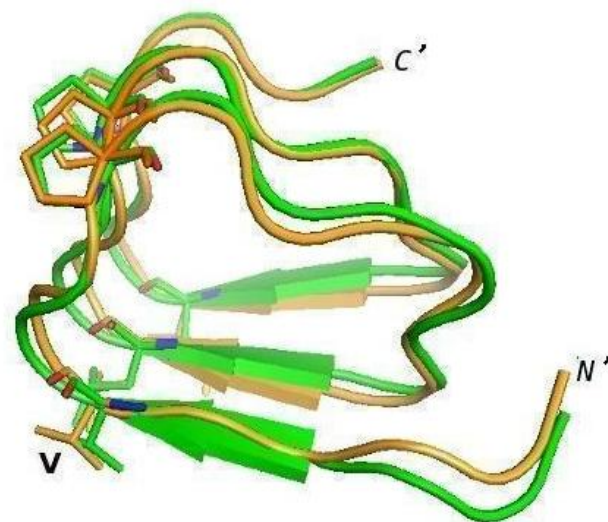
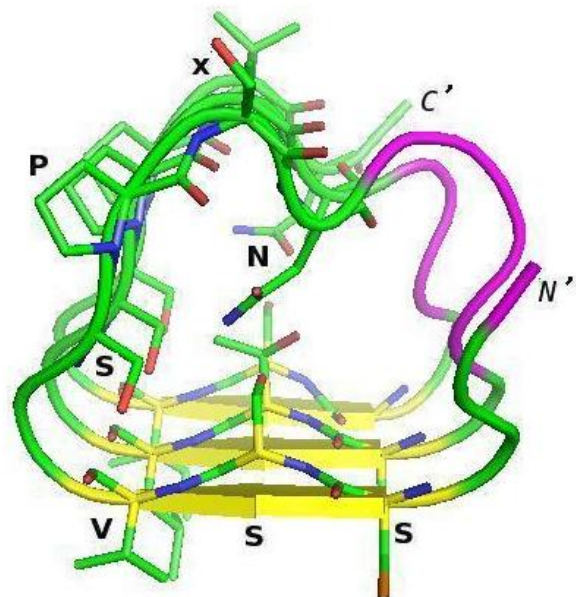
	#T8	#T9	#T10	#T11	#T12	#T13	#T14	#T15
..270.. ...280.. ...290.. ...300.. ...310.. ...320.. ...330.. ...340.. ...350.. ...360.. ..		*	*					*
158508572 Hsapiens	MKSSISSPPSHCSVKSPVSPNNVTLRSSVSSPANIN	NSRCSVSSPSNTN	NRSTLSSPAAS	TVGSICSPVNAF	FSYTASGTSAGSSTLRDVVPSPDTQE--			
31324675 Cjacchus	MKSSISSPPSHCSVKSPVSPNNVTLRSSVSSPANIN	NSRCSVSSPSNTN	NRSTLSSPAAS	TVGSICSPVNAF	FSYTASGTSAGSSTLRDVVPSPDTQE--			
126331313 Mdomestica	MKSPISSPPSHCSVKSPVSPNNVTLRSSVSSPANIN	NSRCSISSPSNTN	NRSTLSSPAAS	TVGSICSPVNAF	FSFASATSVGSSSTQDIVSPETNE--			
73978292 Clupus	MKSPISSPPSHCSVKSPVSPNNVTLRSSVSSPANIN	NSRCSVSSPSNTN	NRSTLSSPTAS	TVGSICSPVNAF	FSYTASGTPAGSSVARDVVPSPDTQE--			
301763180 Amelanoleuca	MKSPISSPPSHCSVKSPVSPNNVTLRSSVSSPANIN	NSRCSVSSPSNTN	NRSTLSSPTAS	TVGSICSPVNAF	FSYTASGTPAGSSAARDVVPSPDTQE--			
6981208 Rnorvegicus	MKSPISSPPSHCSVKSPVSPNNVTLRSSVSSPANLN	NSRCSVSSPSNTN	NRSTLSSPTAS	TVGSIGSPISNAF	SYATSGASAGAGAIQDVVPSPDTHE--			
144227212 Mmusculus	MKSPISSPPSHCSVKSPVSPNNVTLRSSVSSPANLN	NSRCSVSSPSNTN	NRSTLSSPTAS	TVGSIGSPISNAF	SYTTSAGASAGAGAIQDMVPSPDTHE--			
148224443 Xlaevis	MRSPISSPPSHASLKSPVSPNNITVRPSVSPGMI	NTRSSLSSPSNAN	NRSTISSPAAS	SMGSSICSPASSTLGLFLPGVMPTDG--GTASDISAETKD-				
327274009 Acarolinensis	MKSPISSPPSHCSVKSPVSPNNITMRSSVSSPANM	NSRCSIASPSNAN	NRSTLSSPAVS	TVGSSMCSPVNGALSFPPSSPTVPGPSGTQDIVSPDTKD-				
115529242 Tguttata	MKSPISSPPSHCSVKSPVSPNNITMRSSVSSPANL	NSKSSIASPSNAN	NRSTRSSPAVS	TVGSSICSPVNSIGFLPSGTPVGPSRSQDTPVPSPETKD-				
225936142 Ggallus	MKSPISSPPSHCSVKSPVSPNNITMRSSVSSPANM	NSRSSIASPSNAN	NRSTRSSPAVSTV	GGSSICSPVNSLGLFPASSTPAGPSRSQDTPVPSPETKD-				
239923135 Rrutilus	MKSPVSSPRSIGVRSPLSCNTN	MRSSVSSPTTNGGNTCN	IRPSISSPPTV		GSMAMSSPRNRSSRGFVSSPPSGLGLVQNDVNSPESRE--			
154240734 Drerio	MKSPVSSPRSIGVRSPLSCNTN	MRSSVSSPTTNGGNTCN	IKPSISSPPTA		GSMMSSPRNRSSRGFVSSPPSGLGLVQNDVNSPESRE--			
	#F5	#F6	#F7	#F8		#F9	#F10	#F11

Vlassi et al. (2013) BMC Struct. Biol.

Mineralocorticoid receptor



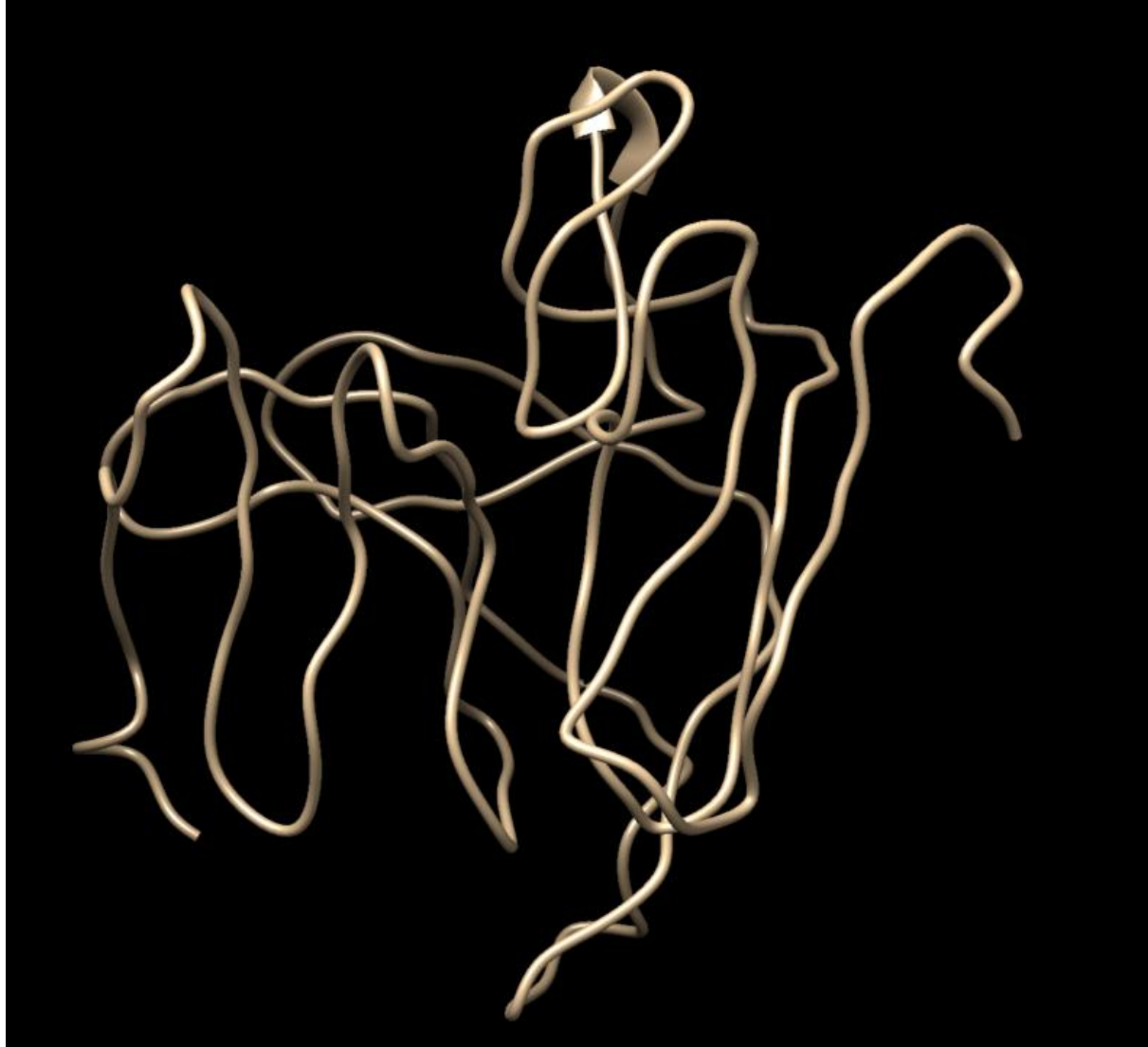
Vlassi *et al.* (2013) *BMC Struct. Biol.*



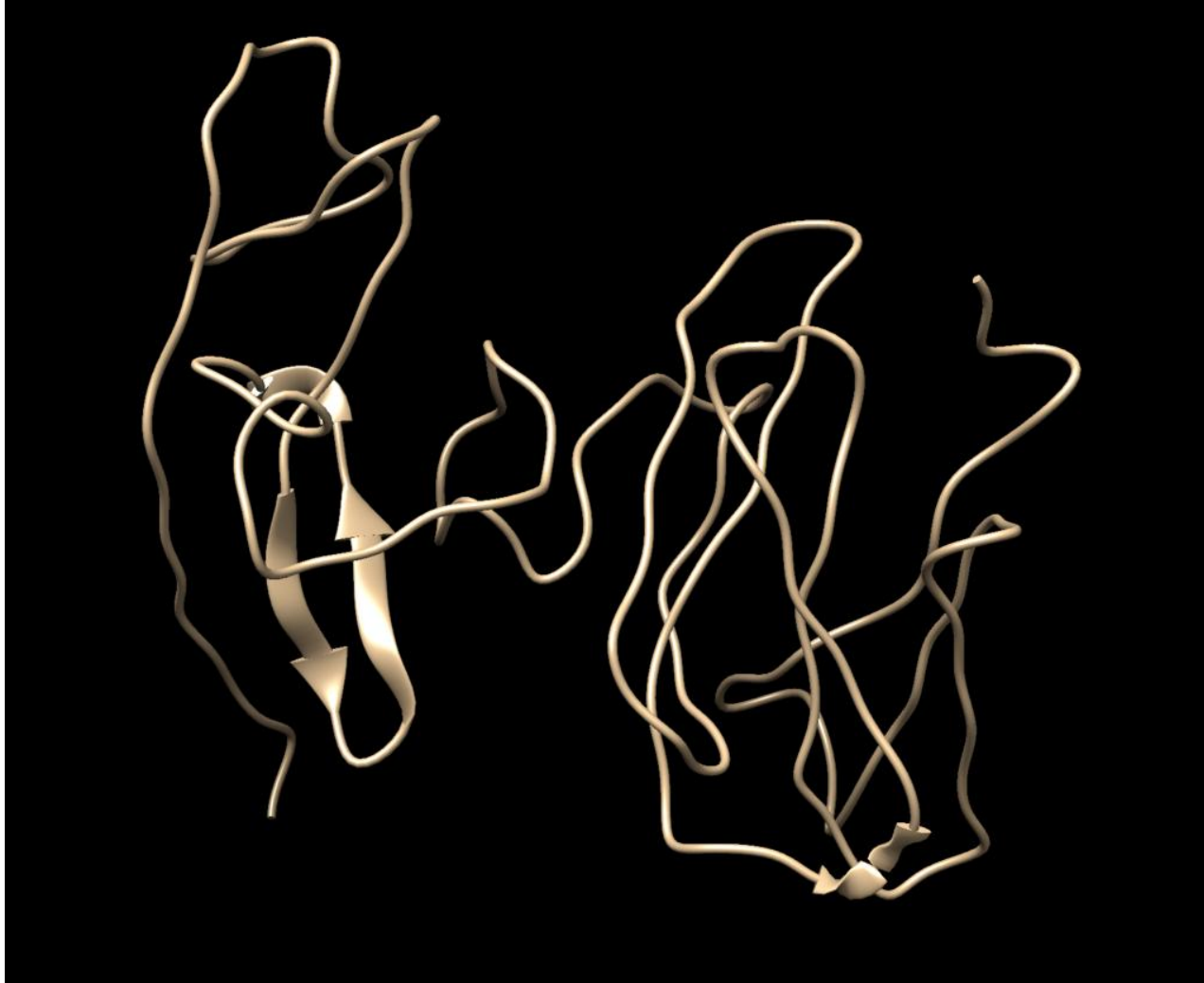
Modelling the repeat region with AlphaFold...



Modelling with CI-Tasser...

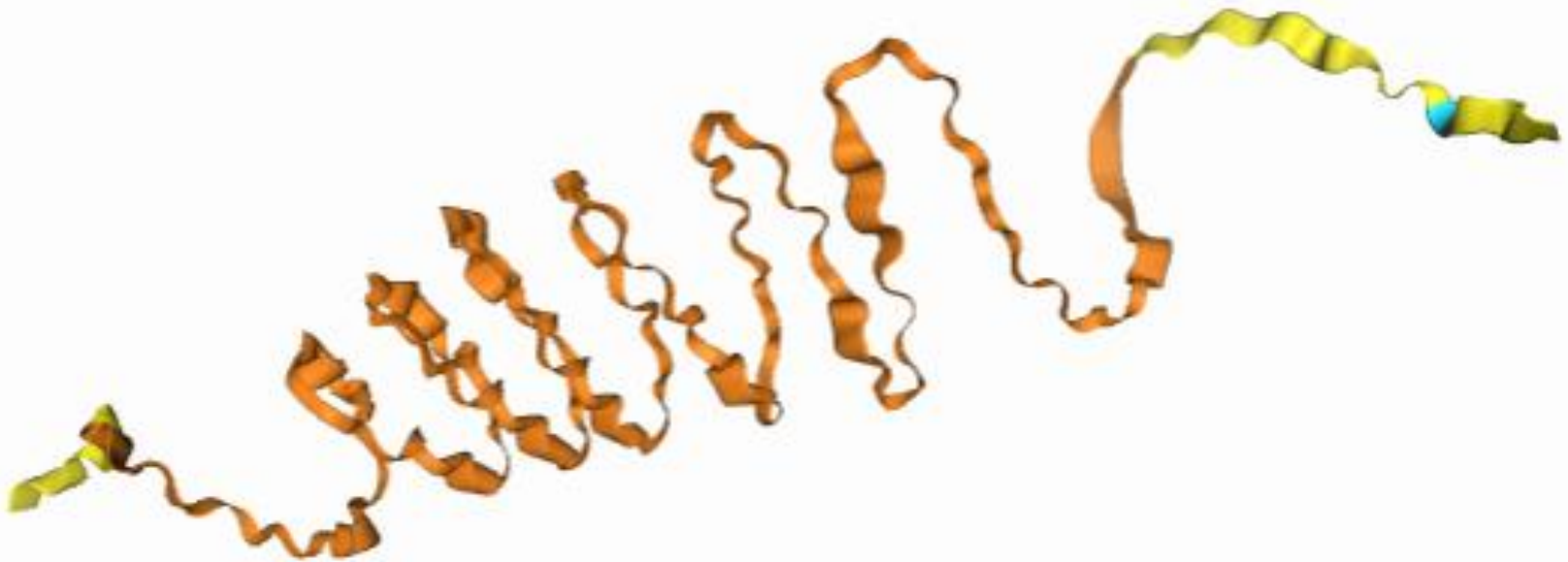


Modelling with trRosetta...



Modelling with AlphaFold3...

27 Nov 2024



Composition bias

Detection of CBRs

Compositionally biased regions (CBRs)

High frequency of one or two amino acids in a region.

E.g. **QLPQPPPAQPLLPPQ**

Also:

Single amino acid repeats (homorepeats)

E.g. **QHQHQHQHQHQHQHQ** (polyQ)

Detection of CBRs

Often straightforward.

N-terminal human Huntingtin.

Can you find **CBRs**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP
LLPQPQPPPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQH SRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS PSAEQLVQVYEL
TLHHTQH QDHN VVTGALELLQQLFRTPPPELLQT LTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection of CBRs

Often straightforward.

N-terminal human Huntingtin.

Can you find **CBRs**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
```

```
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP  
LLPQPQPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  
RSLRAALWRFAELAHVLRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV  
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPPSAEQLVQVYEL  
TLHHTQHGDHNVVTGALELLQQQLFRTPPPELLQTTLTAVGGIGQLTAAKEESGGRSRSGSI  
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSVDVSSSALTASVKDEISGELAA  
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV  
PSDPAMDLDNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE  
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG  
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL
```

Detection of CBRs

Often straightforward.

N-terminal human Huntingtin.

Can you find **CBRs**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
```

```
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP  
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQH SRRTQYFYSWLLNVLLGLLV  
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS PSAEQLVQVYEL  
TLHHTQH QDHN VVTGALELLQQLFRTPPPELLQTTLTAVGGIGQLTAAKEESGGRSRSGSI  
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA  
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV  
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE  
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG  
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection of CBRs

Often straightforward.

N-terminal human Huntingtin.

Can you find **CBRs**?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
```

```
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPQAQP  
LLPQPQPPPPPPPPGPAAVEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQH SRRTQYFYSWLLNVLLGLLV  
PVEDEHSTLLILGVLLTRYLVLLQQQVKDTSLKGSFGVTRKEMEVS PSAEQLVQVYEL  
TLHHTQH QDHN VVTGALELLQQLFRTPPPELLQTTLTAVGGIGQLTAAKEESGGRSRSGSI  
VELIAGGGSSCSPVLSRKQKGKVLLGEEEAEDDESRSDVSSSALTASVKDEISGELAA  
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV  
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE  
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG  
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Function of CBRs

Conservation => Function

Length, amino acid type not necessarily conserved

Frequency: 1 in 3 proteins contains a compositionally biased region (Wootton, 1994), ~11% conserved (Sim and Creamer, 2004)

Function of CBRs

Conservation => Function

Length, amino acid type not necessarily conserved

Functions:

Passive: linkers

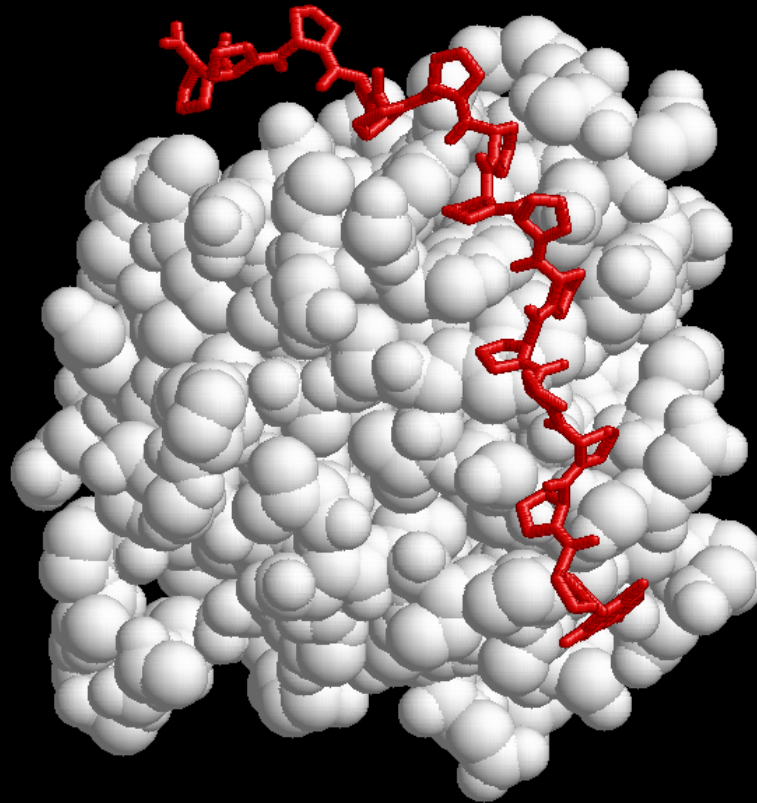
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

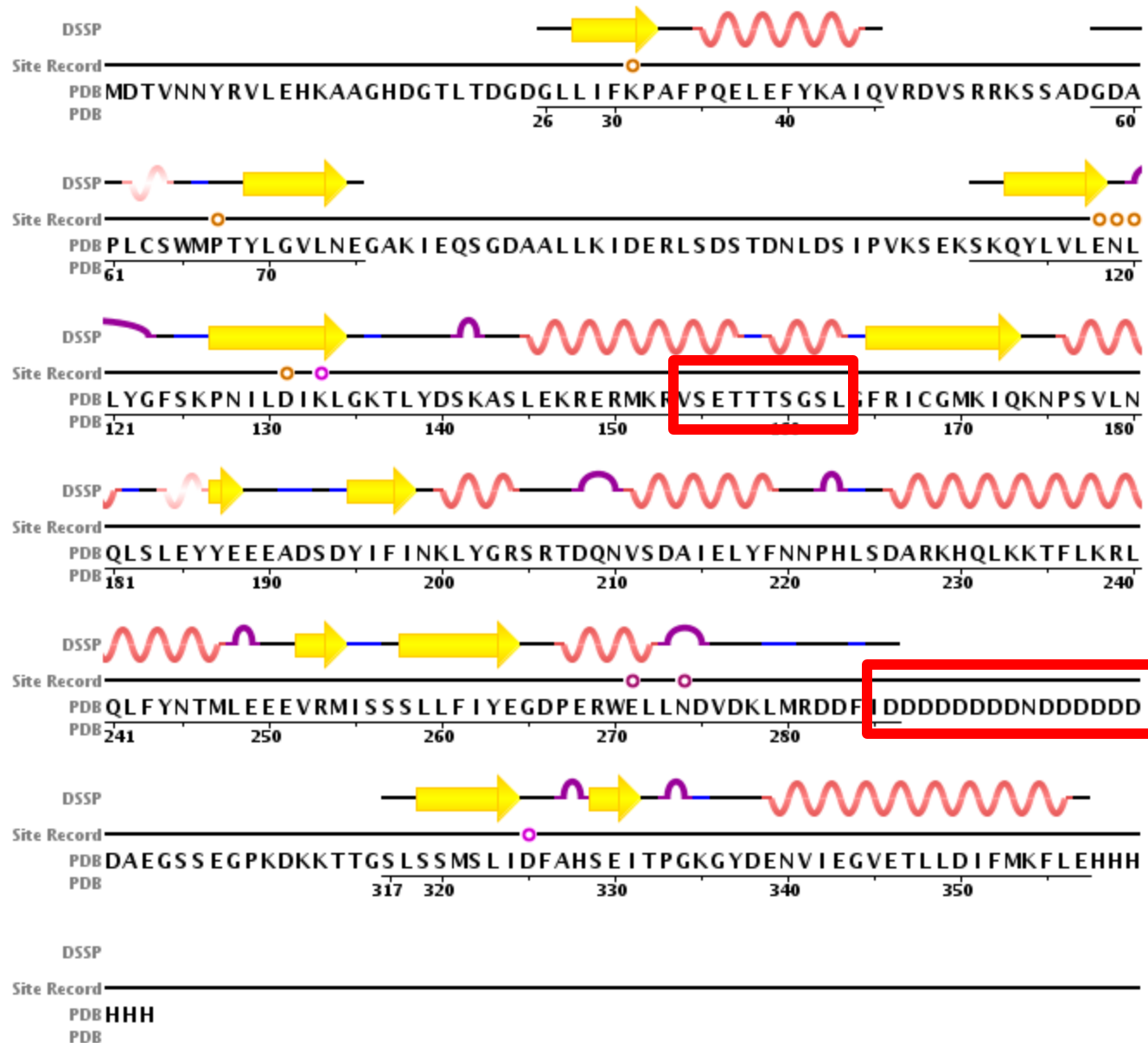
Structure of CBRs

Often variable or flexible: do not easily crystallize

1CJF: profilin bound to polyP



2IF8: Inositol Phosphate Multikinase Ipk2

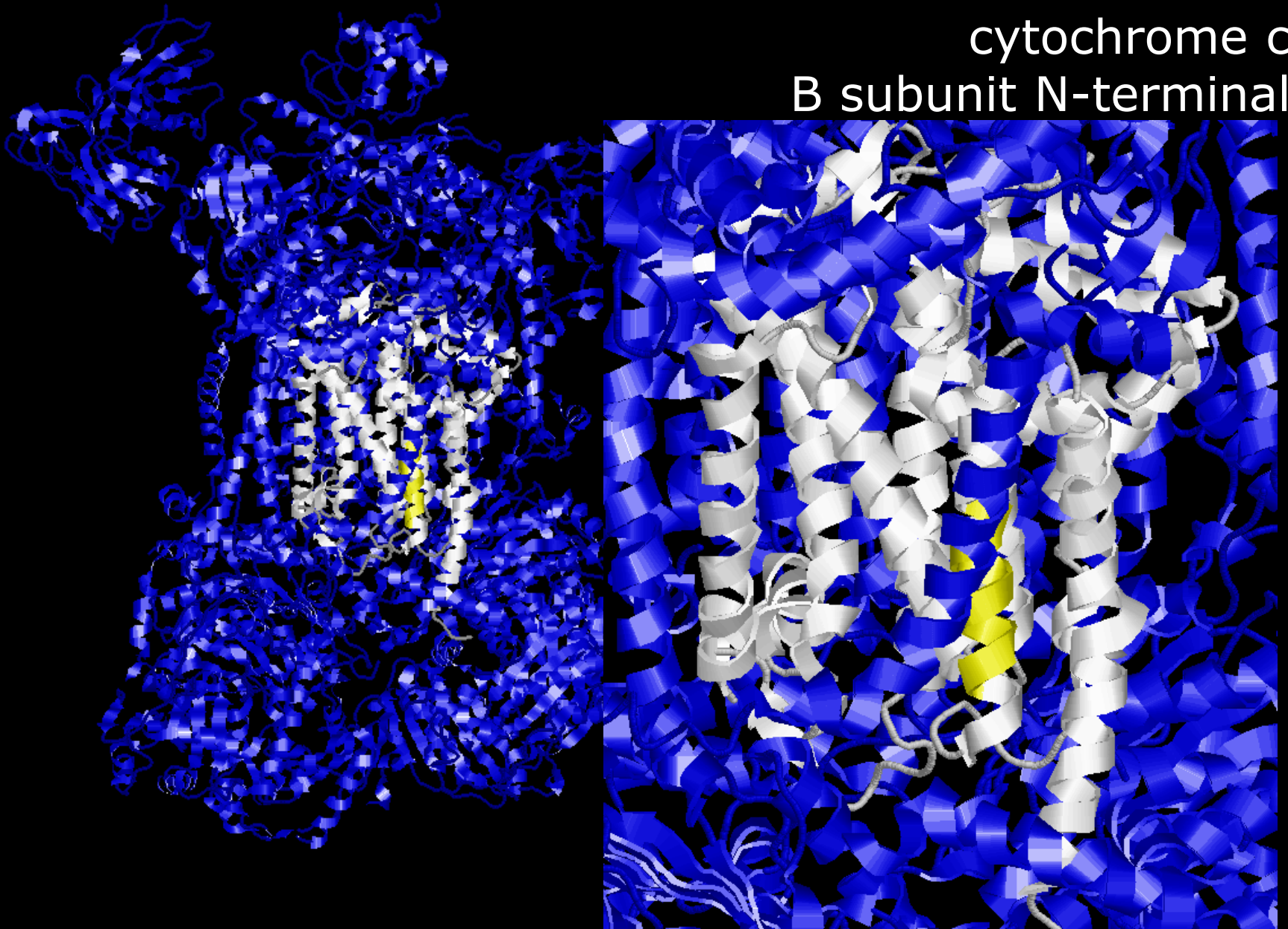


2IF8: Inositol Phosphate Multikinase Ipk2

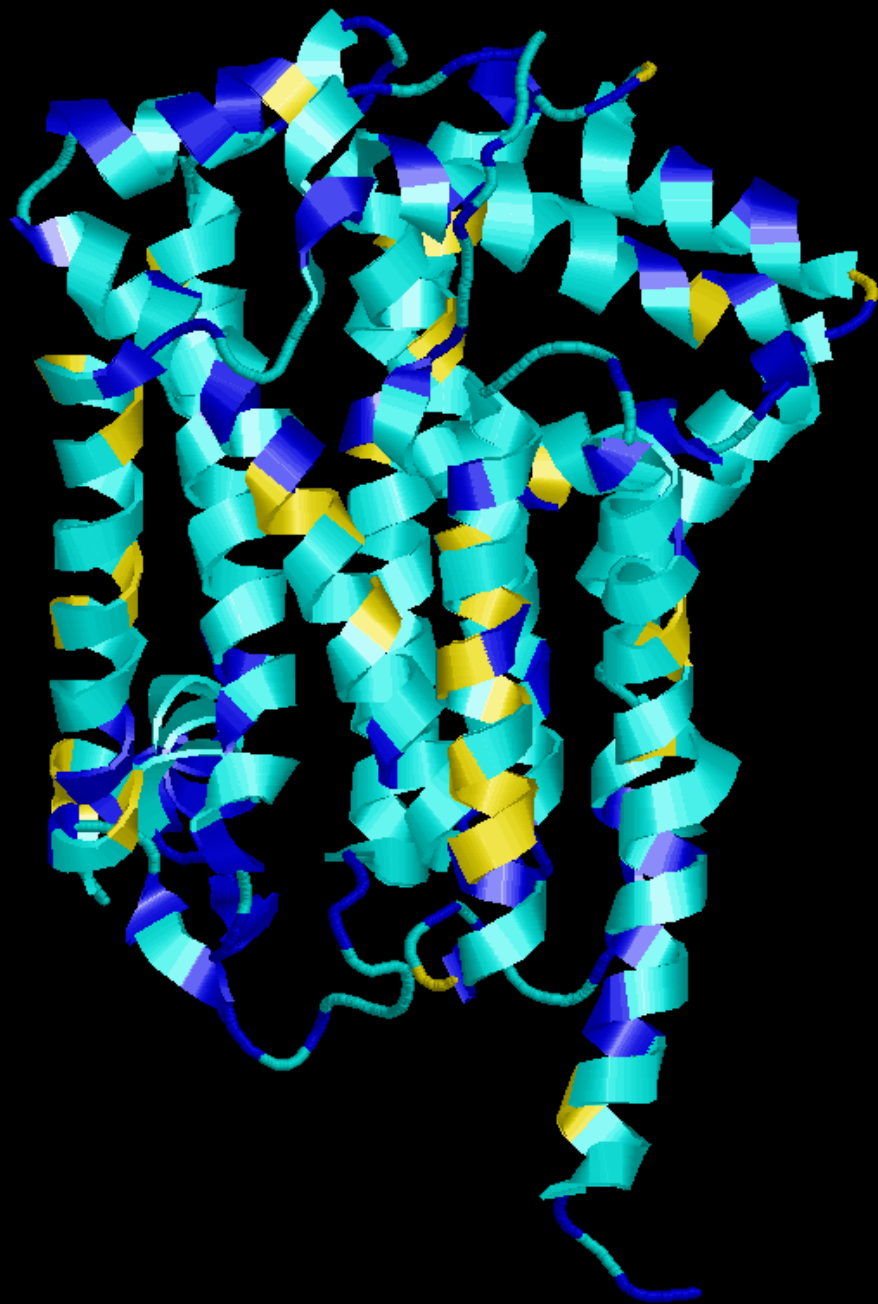


RVSETTSGSL

2CX5: mitochondrial
cytochrome c
B subunit N-terminal



2CX5: mitochondrial
cytochrome c
B subunit N-terminal



EEFFIIEVNE



Amino acid homorepeats (polyX)

Distribution is not random:

Eukaryota:

Most common: poly-Q, poly-N, poly-A, poly-S, poly-G

Prokaryota:

Most common: poly-S, poly-G, poly-A, poly-P

Relatively rare: poly-Q, poly-N

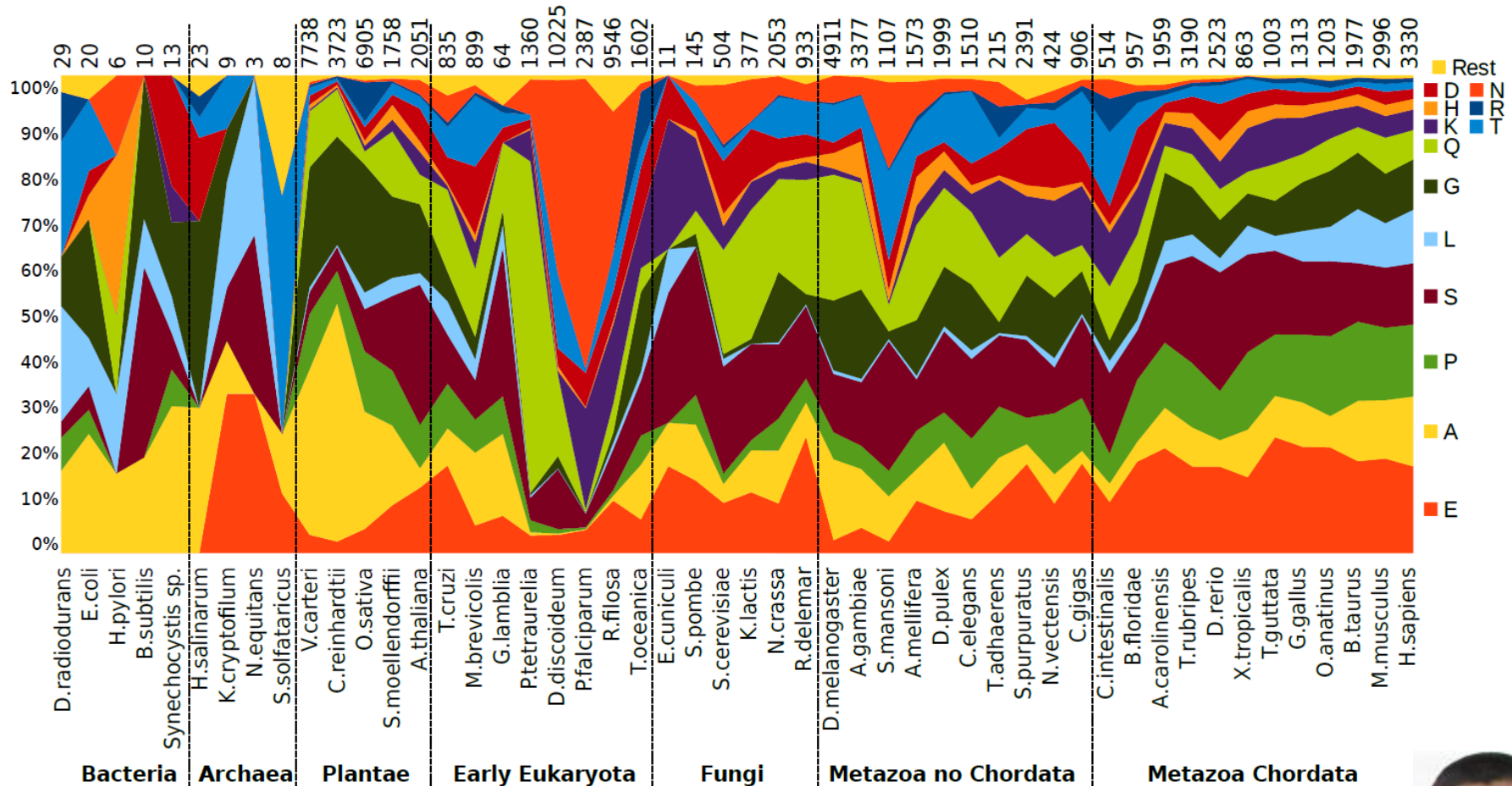
Very rare or absent in both eukaryota and prokaryota:

Poly-I, Poly-M, Poly-W, Poly-C, Poly-Y

Toxicity of long stretches of hydrophobic residues.

(Faux et al 2005)

Amino acid homorepeats (polyX)



Pablo
Mier

Homorepeats are frequent but difficult to characterize

Pablo
Mier



e.g. polyQ:

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQP

- 10% of human proteins have homorepeats
- low sequence conservation
- not possible to predict function by homology

Homorepeats need to be studied in context

Exercise 3. Search for a polyQ insertion in the MR family

- Open in Jalview the alignment of the mineralocorticoid receptor: MR1_fasta.txt
- Find a polyQ insertion.

Do you see any other biased region nearby?