



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.

2: Longer imperfect tandem repeats.
Assemble in structure.

Definition repeats

Sequence, long, imperfect, tandem

MRAVVKSPIMCHEKSPSVCSPLNMTSSVCS PAGINSVSSTTASF
GSFPVHSPITQGTPLTCSPNVENRGSRS HSPA HASNVGSPLSSP
LSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINN

Definition repeats

Sequence, long, imperfect, tandem

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

Definition repeats

Sequence, long, imperfect, tandem

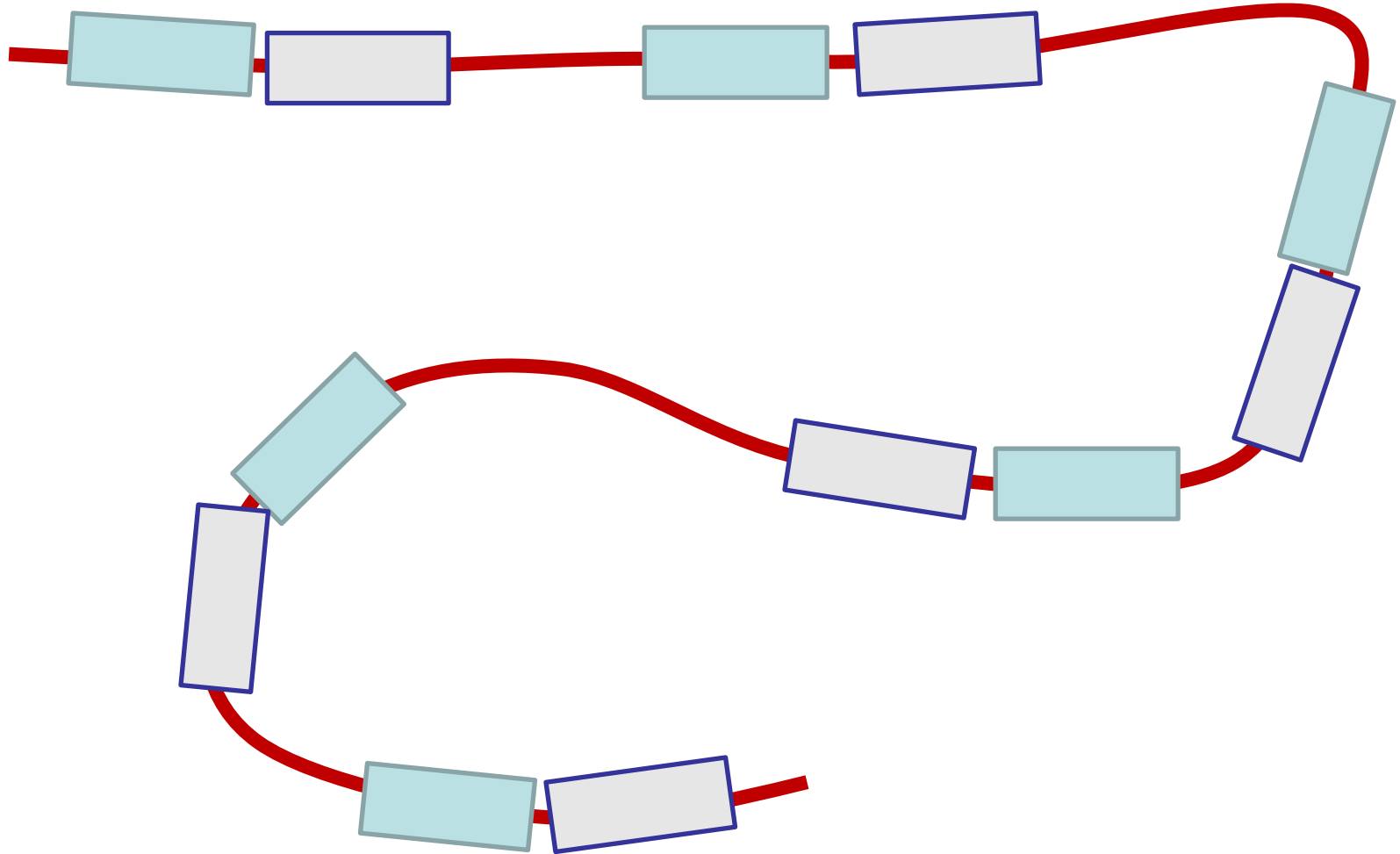
MRAVVK**SP**IM CHE
KSPSVC**SP**LN
MTSSVC**SP**AG INSVSSTTASF
GSFPVH**SP**IT Q
GTPLTC**SP**NV EN
RGSRSH**SP**AH ASN
VGSPLS**SP**LS S
MKSSIS**SP**PS HCS
VKSPVS**SP**NN VT
LRSSVS**SP**AN INN

Definition repeats

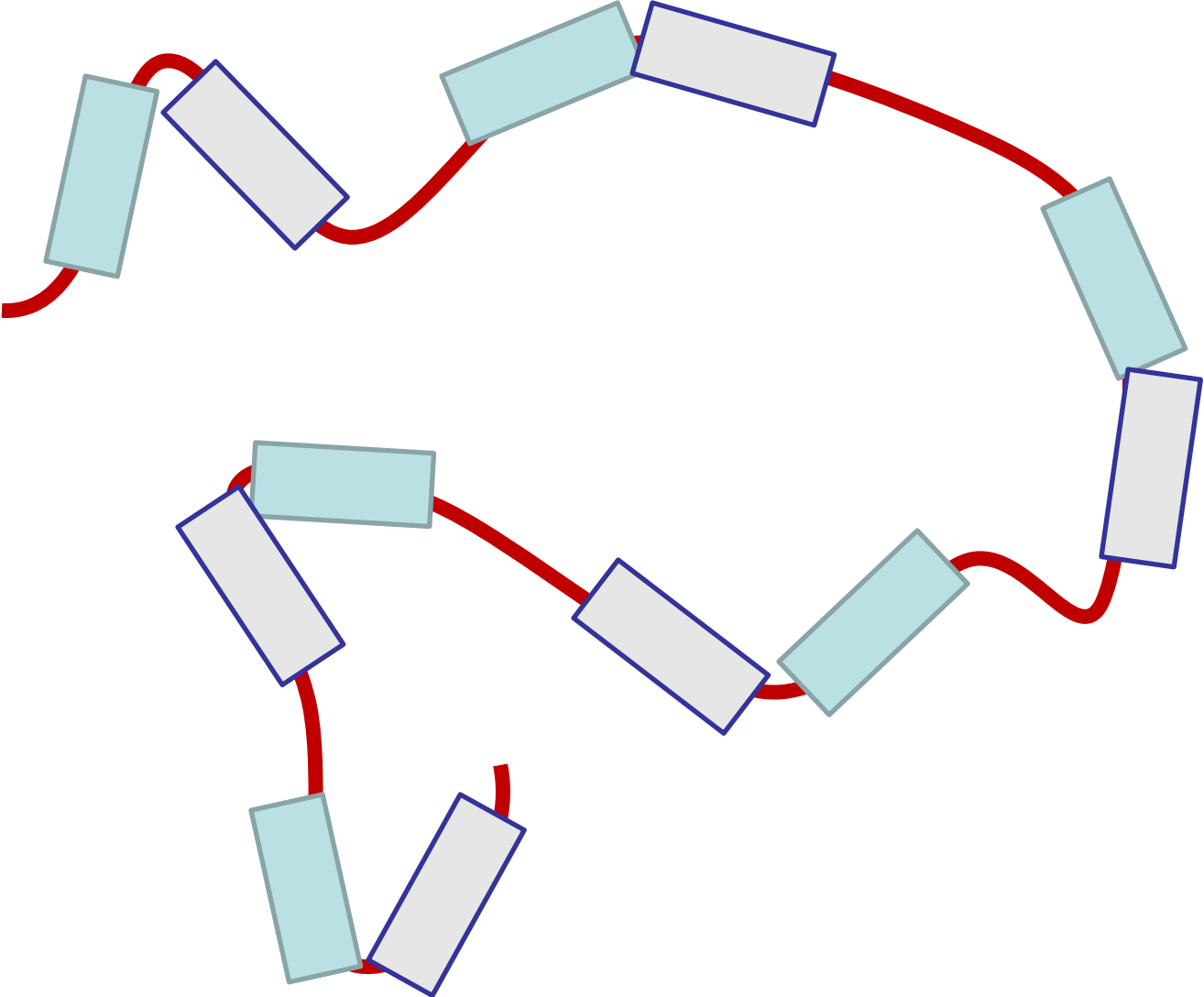
Sequence, long, imperfect, tandem

MRAV**V**K**SP**IM CHE
KSPSVC**SP**LN
MT**S****V**C**SP**AG INSVSSTTASF
GSFP**V**H**SP**IT Q
GTPLTC**SP**NV EN
RG**S**RS**H****SP**AH ASN
VG**S**PL**S****SP**LS S
MK**S**SI**S****SP**PS HCS
VK**S**PV**S****SP**NN VT
LR**S**SV**S****SP**AN INN

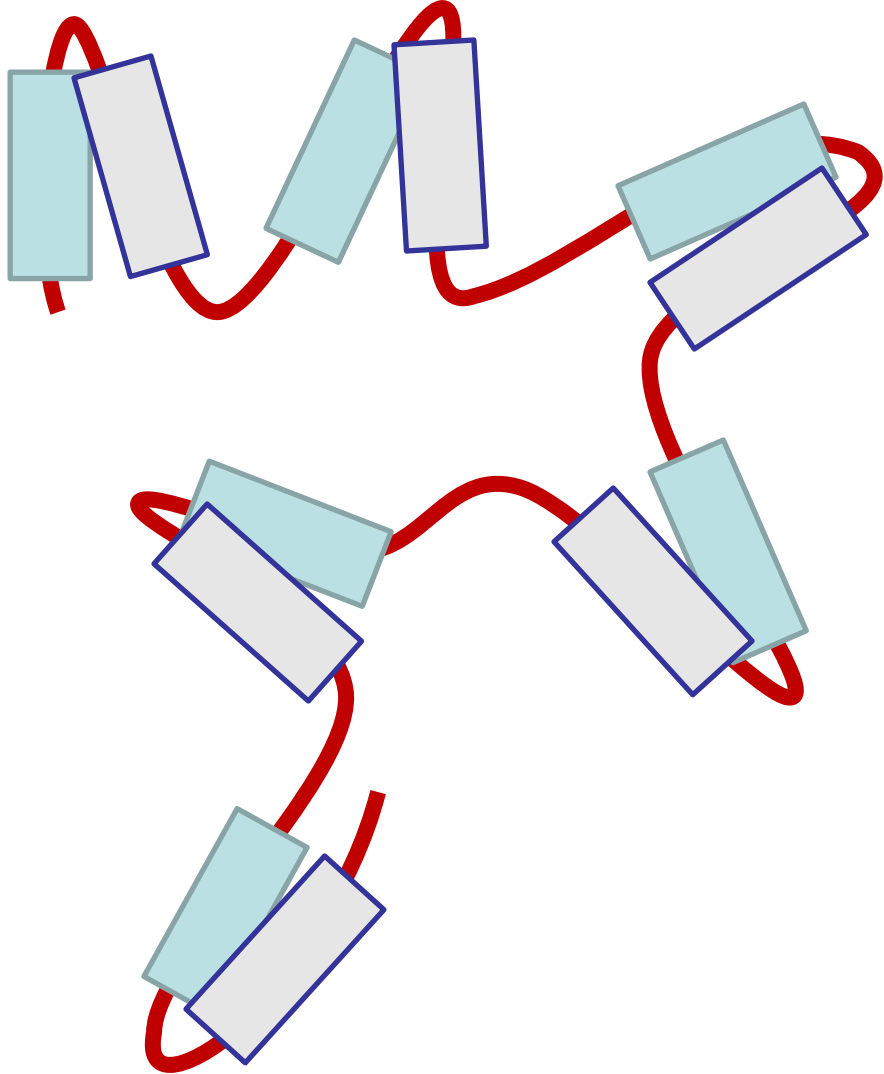
Tandem repeats fold together



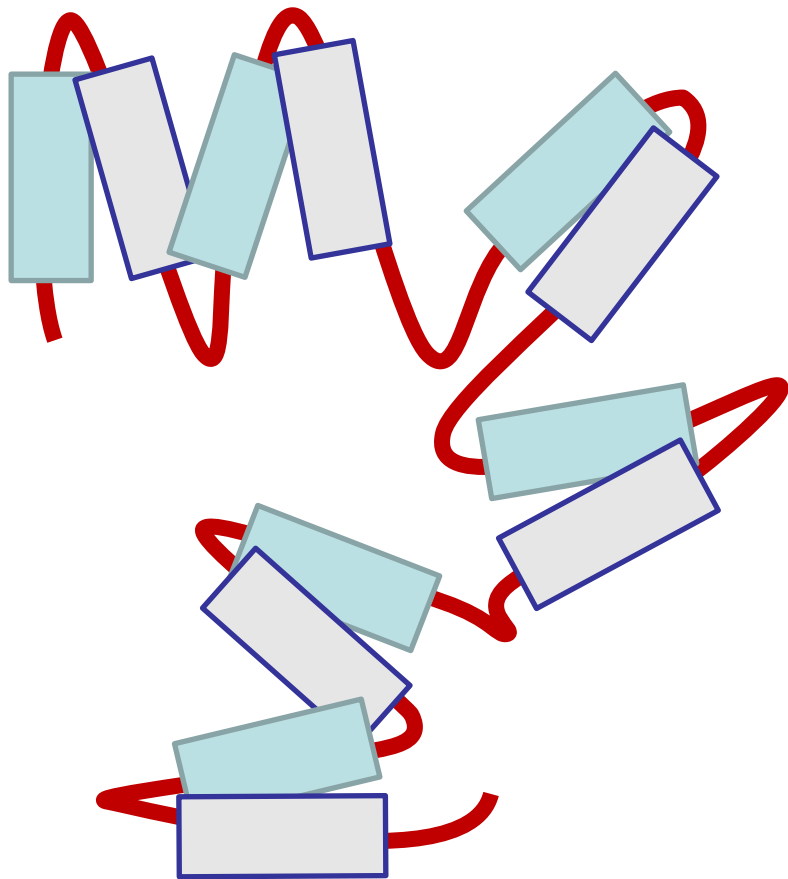
Tandem repeats fold together



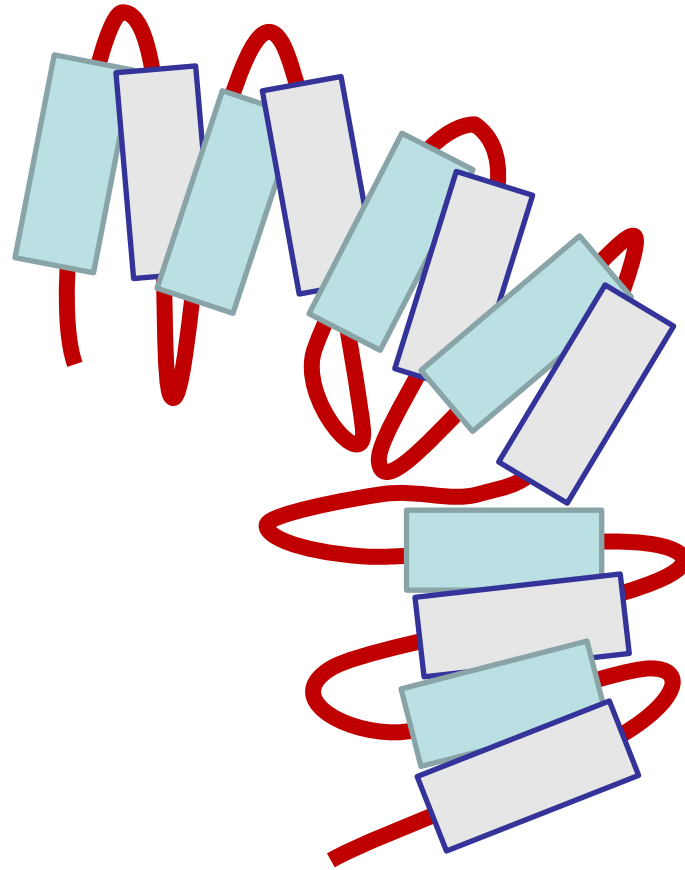
Tandem repeats fold together



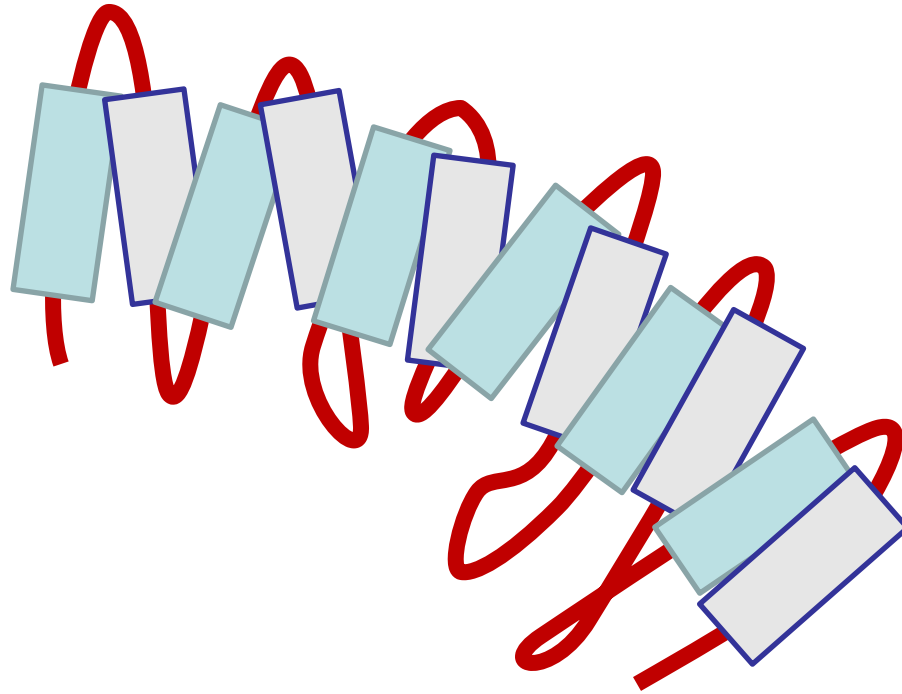
Tandem repeats fold together



Tandem repeats fold together



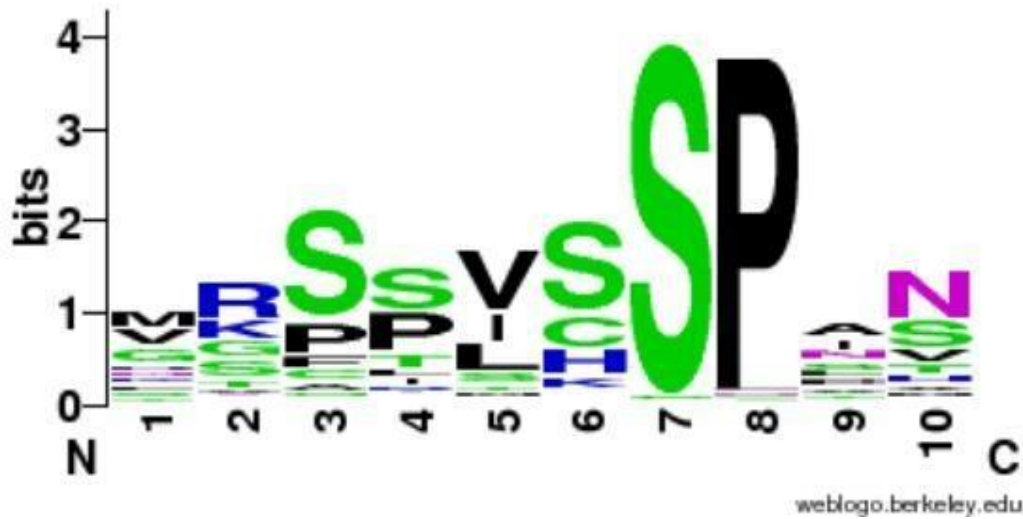
Tandem repeats fold together



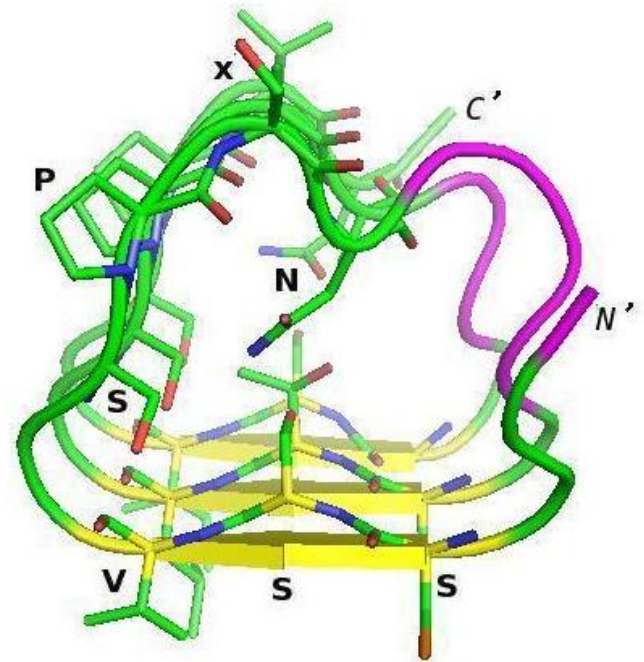
Definition repeats

Sequence, long, imperfect, tandem

MRAV**V**K**SP**IM CHE
KSPSVC**SP**LN
MT**S****V**C**SP**AG INSVSSTTASF
GSFP**V**H**SP**IT Q
GTPLTC**SP**NV EN
RG**S**RS**H****SP**AH ASN
VG**S**PL**S****SP**LS S
MK**S**SI**S****SP**PS HCS
VK**S**P**V****S****SP**NN VT
LR**S****V****S****SP**AN INN



<http://weblogo.berkeley.edu>



(Vlassi et al, 2013)

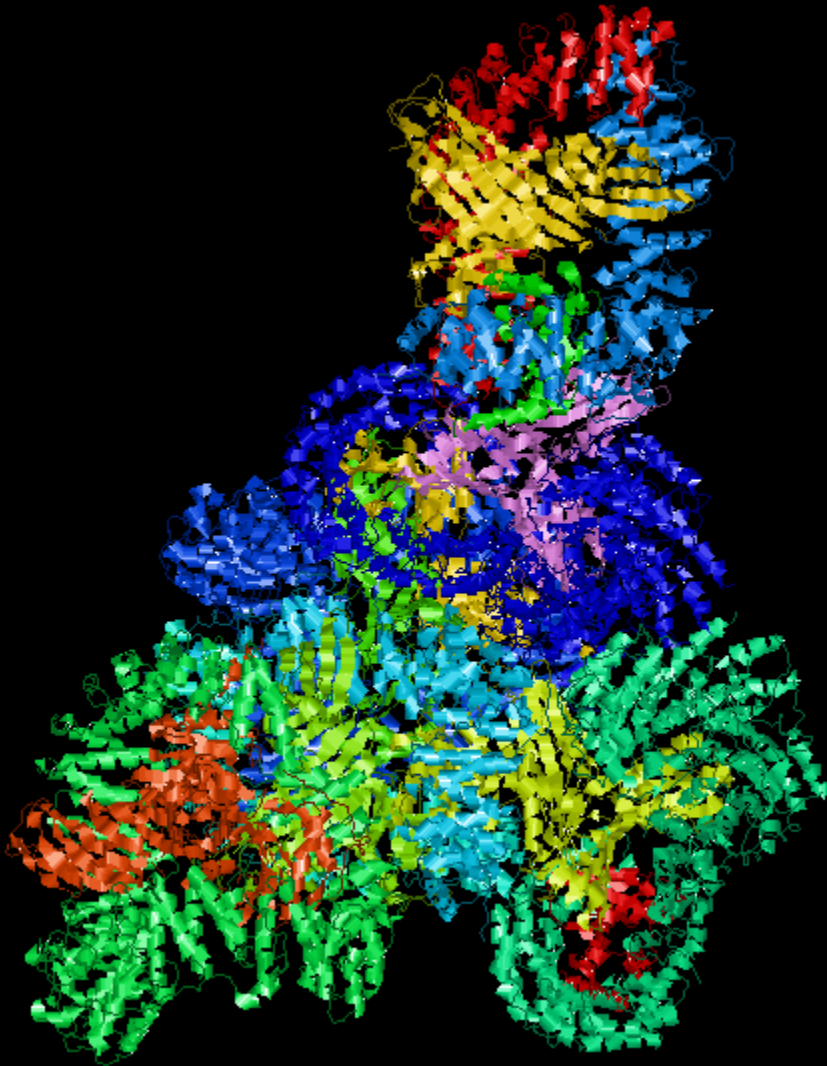
A subunit PP2A structure



PDB:1b3u

Groves et al. (1999) *Cell*

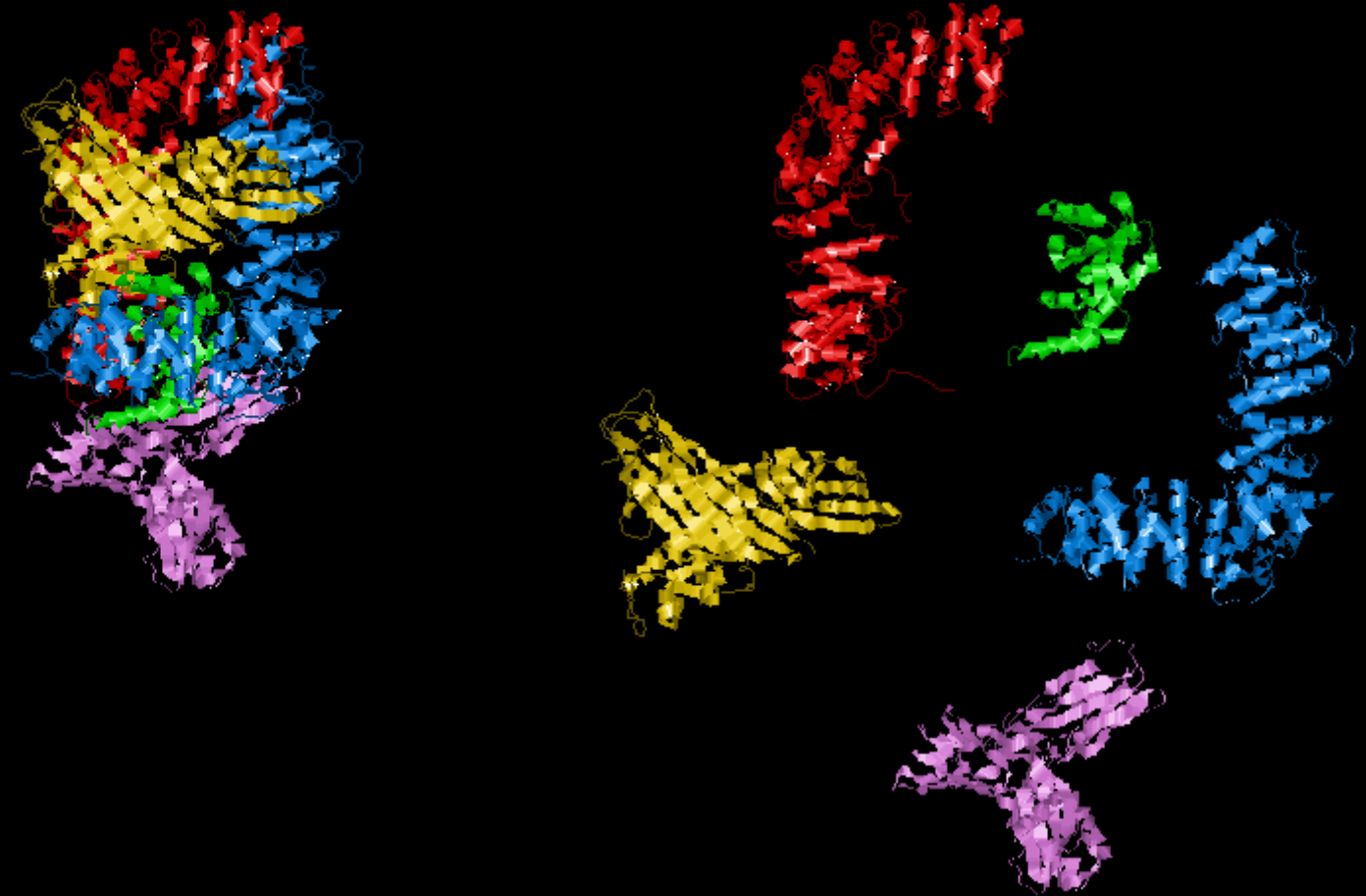
Ap1 Clathrin Adaptor Core



PDB:1w63

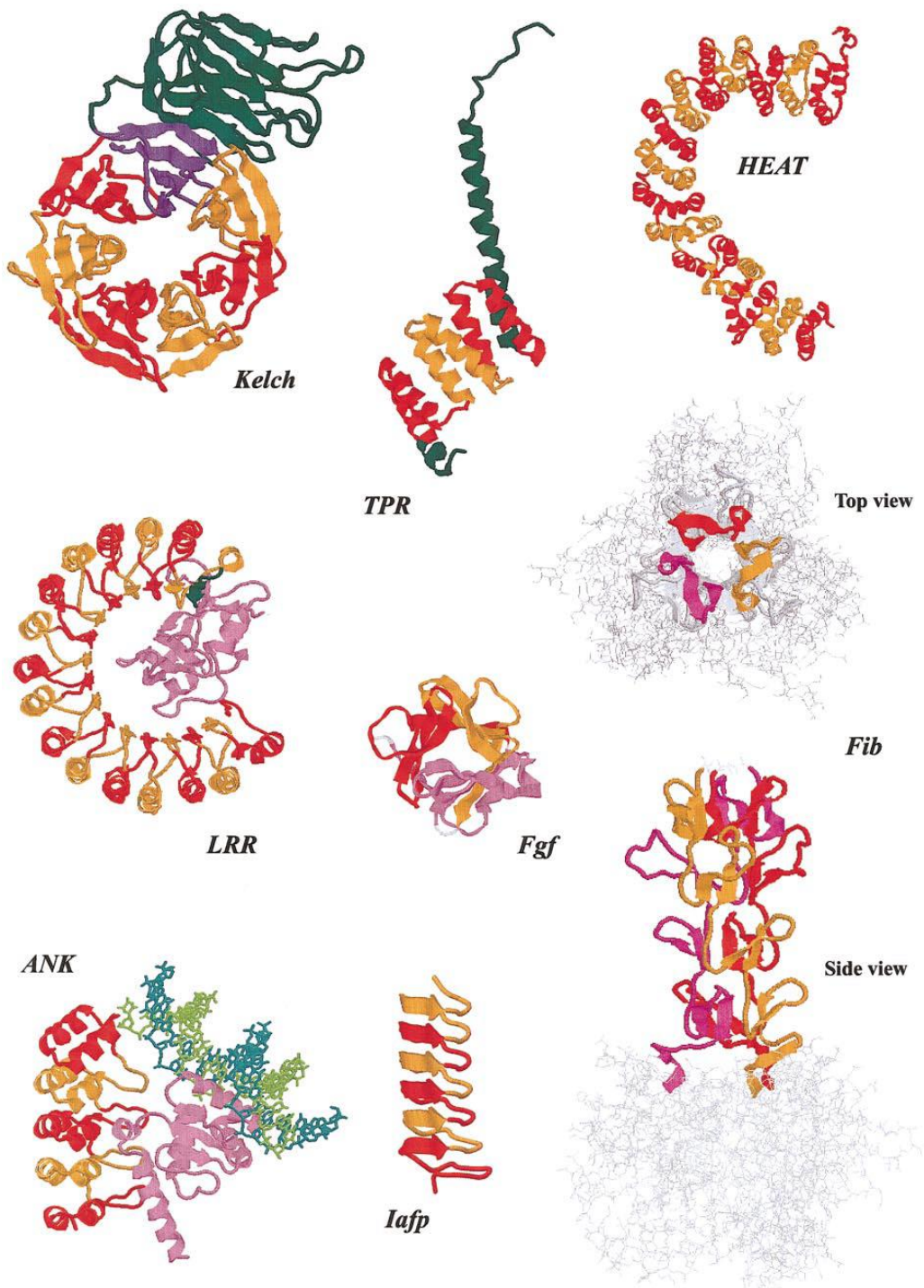
Heldwein et al. (2004) *PNAS*

Ap1 Clathrin Adaptor Core



PDB:1w63

Heldwein et al. (2004) *PNAS*



Andrade et al. (2001)
J Struct Biol

Definition CBRs

Perfect repeat: QQQQQQQQQQQQ

Imperfect: QQQQPQQQQQQ

Amino acid type: DDDDEEEDEDEED

Compositionally biased regions (CBRs)

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

Detection CBRs

Sometimes straightforward.
N-terminal human Huntingtin.
How many **CBRs** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHRRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPPSAEQLVQVYEL
TLHHTQHGDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASC DLTSSATDGDEEDILSHSSSQVSAV
PSDPAMD LNDGTQASSPISDSSQTTEGPD SAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection CBRs

Sometimes straightforward.
N-terminal human Huntingtin.
How many **CBRs** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSF QQQQQQQQQQQQQQQQQQQQQQ PPPPPPPPPPQLPQPPPQAQP
LLPQPQ PPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHRRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHGDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLDNDGTQASSPISDSSQTTEGPD SAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICCSIL
```

Detection CBRs

Sometimes straightforward.
N-terminal human Huntingtin.
How many **CBRs** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPPSAEQLVQVYEL
TLHHTQHGDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLDNDGTQASSPISDSSQTTEGPD SAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHL LKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection CBRs

Sometimes straightforward.
N-terminal human Huntingtin.
How many **CBRs** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHRRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASC DLTSSATDGDEEDILSHSSSQVSAV
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHL LKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```


Detection repeats

Sometimes straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLDNDGTQASSPISDSSQTTEGPD SAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

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

```
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPPQAQP  
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV  
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL  
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI  
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA  
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASC DLTSSATDGDEEDILSHSSSQVSAV  
PSDPAMD LNDGTQASSPISDSSQT TTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  
EDEEATGILPDEASEAFRNSSMALQQAHL LKNMSHCRQPSDSSVDKFVLRDEATEPGDQE  
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLPDRDVRVSVKALALSCVG  
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLIC SIL
```

Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

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

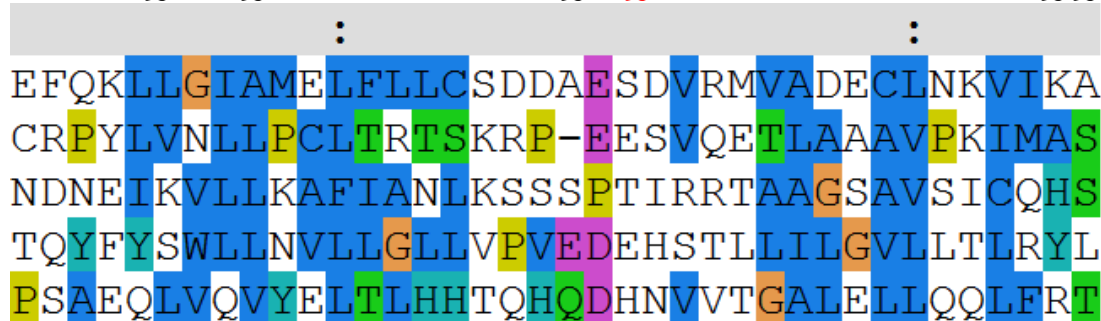
Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

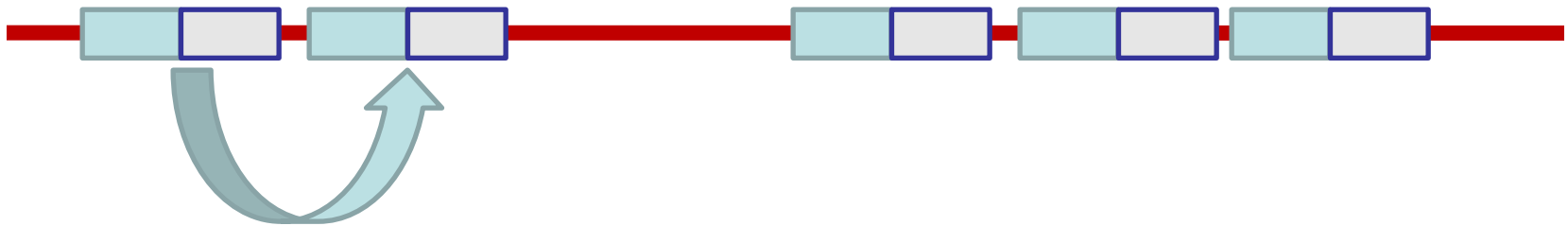
EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKSSPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQHGDHNVVTGALELLQQLFRT



Detection of repeats

Dotplots

Comparing a sequence against itself



Detection of repeats

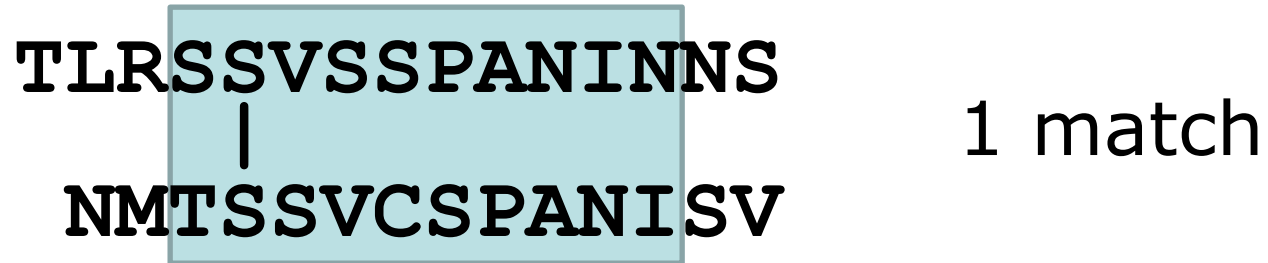
Dotplots

TLRSSVSSPANINNS

NMTSSVCSPANISV

Detection of repeats

Dotplots



Detection of repeats

Dotplots



TLRSSVSSPANINNS
| | | | | | | |
NMTSSVCSPANISV

8 matches

Detection of repeats

Dotplots



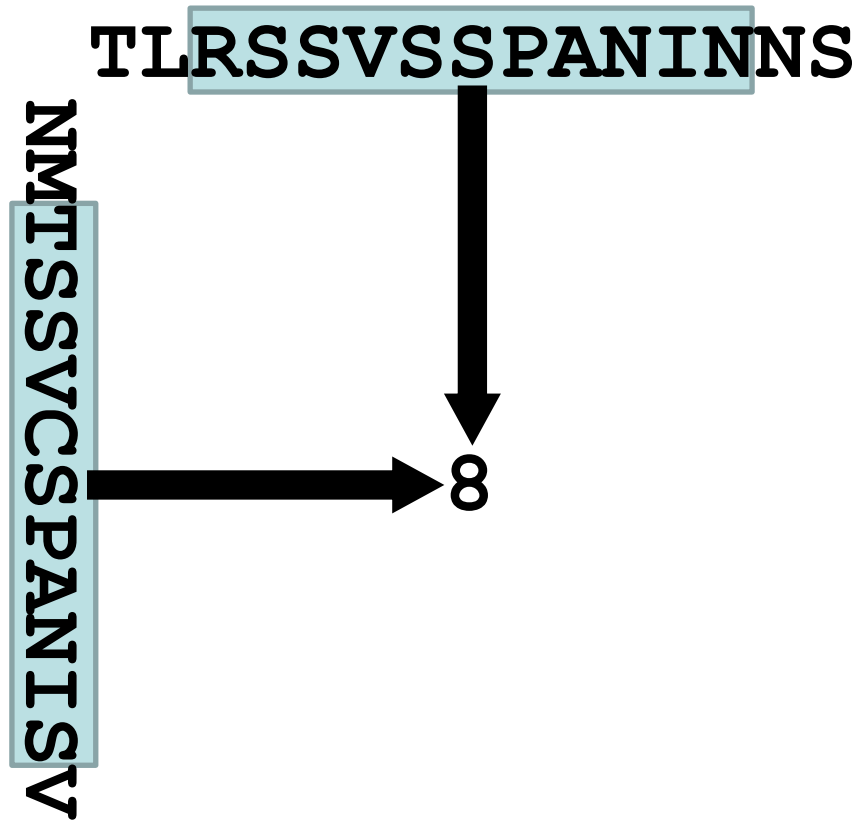
Detection of repeats

Dotplots



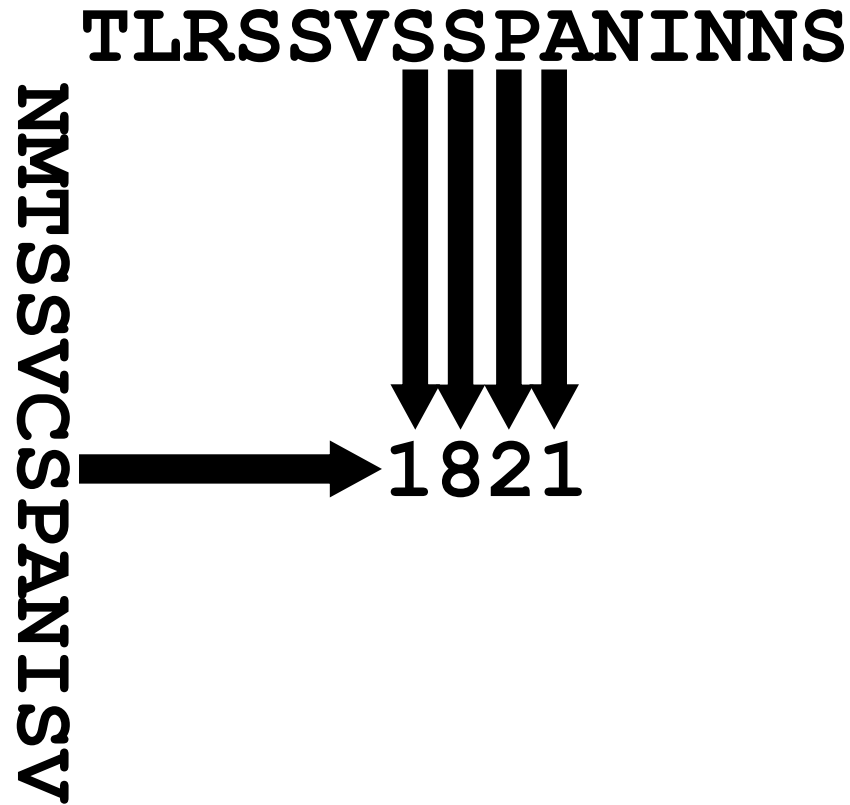
Detection of repeats

Dotplots



Detection of repeats

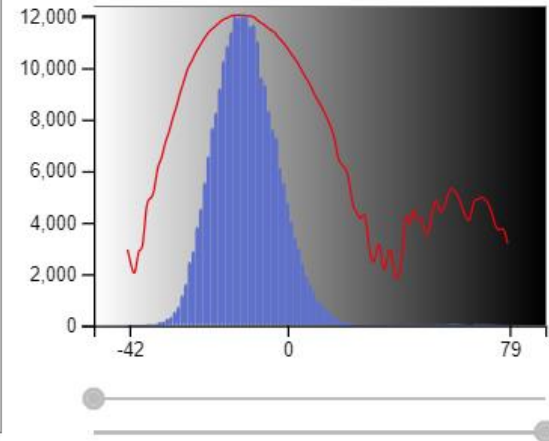
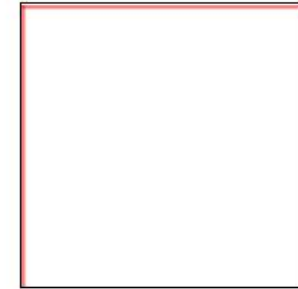
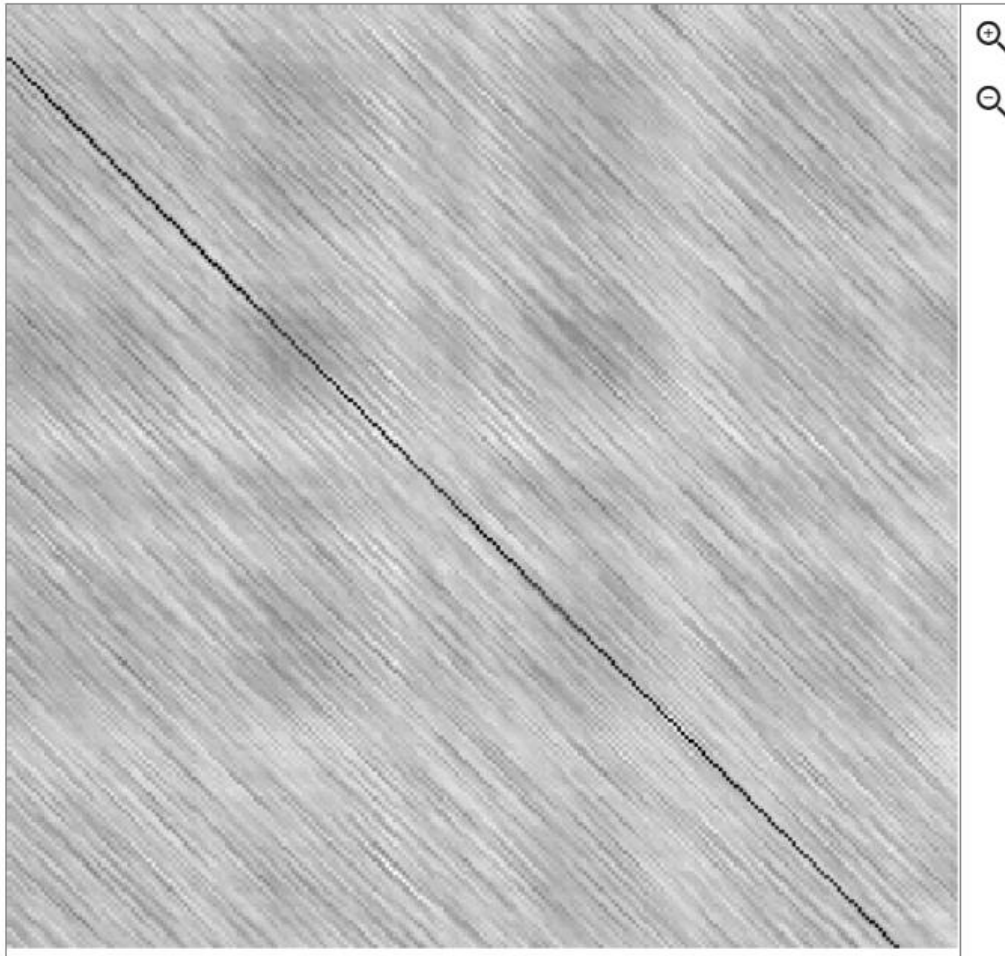
Dotplots





Sequence 1

Sequence 2



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

Seq1:1



MTMDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHE
RKPLQTPTPIRRLWTMDTSELVQAKLAEQAERYDDM



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)

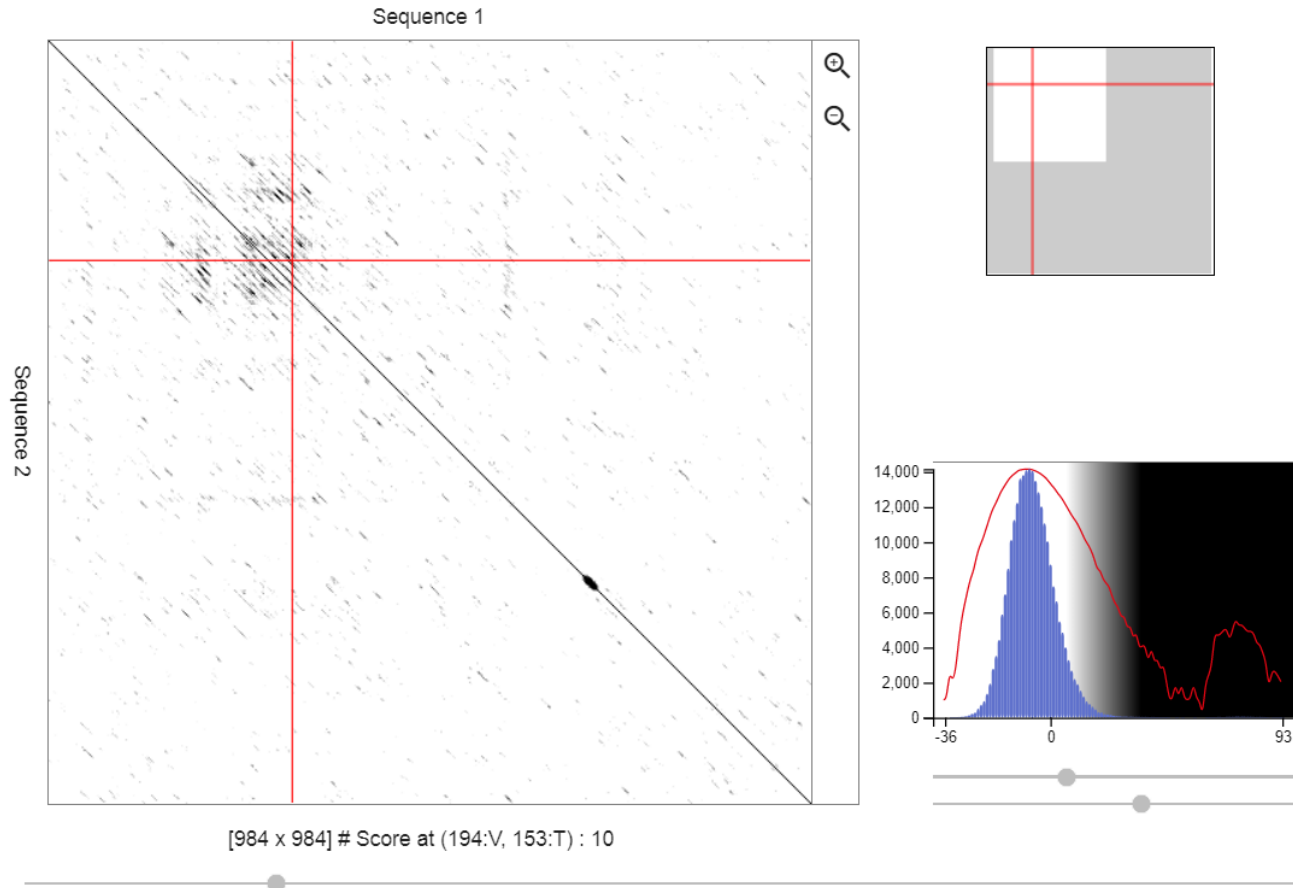
- Try to find combinations of parameters that show patterns in the dot plot (Hint: You can adjust this finely using the arrows)

- Find repetitions clicking in the diagonal patterns

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

SEQUENCE 1 **SEQUENCE 2** Window size 15 Scoring matrix BLOSUM 62 

METKGYHSLPEGLDMERRWGQVSQAVERSLGPRTERTDENNYMEIVNVSCVSGAIPNNST



Seq1:194

NTPLRSFMSD**SGS**SVNGGVMRAVVKSPIM**CHEKS****PSVCS**PLNMTSSVCSPAGIN**SV**SSTTASFGSFPVHSPIT
YSYEQQNQ**QGS**SPAKIYQNVEQLVKFYK**GN**GHR**PS**TLSCV**NTP**LRSFMSD**SGS**SVNGGVMRAVVKSPIM**CHE**

Detection of repeats

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 - SLASASTNL TQGHHTSPNTCSPVKSSMVGSPPLASPLSVMKSPVSSPRSIGSVRSPLSC - - NTNMRSSVSSPTTNG
mr_cyprinus/1-971  - - - TFF - - DSDCP - SLASTHTNL IQGQHTSPNTCSPVKSSVVGSPPLASPLSVIKSPVSSPHSIGSVRSPLSC - - NTNMRSSVSSPTTYG
mr_oryzias/1-994  TCFGPPQCSAVSSPVSQTS CAATLANIKRRNSVTCS PVESCTV GSPPLTSPLNIMRSPMSSPHSMSSVRSPPSCSTTCNIRSSVSSPT - - -
mr_takifugu/1-991  MCFGPMCSSVSSPVSQTS CASTLPNIKRRNSATCS PVESSTV GSPPLTSPLNIMRSPISSPQSMSSVRSPPSCSTTTSNIRSSVSSPT - - -
mr_oreochromis/1-994  TCFAPLCSSVSSPVSQTS CAATLANIKRRNSVTCS PVESSTV GSPPLTSPLNVMRSPMSSPQSMSSVRSPPSCSTTCNIRSSVSSPT - - -
mr_xenopus/1-979  - - FGNF - - TVHSPVNQVTPKSCSPHTDNRC SIAHSP - - AGTVES - PLSSPVSSMRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI -
mr_anolis/1-990   - - FGNF - - TVSSPVNQGTPLSCSPNIENRGSMLHSPPHASNMGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
mr_alligator/1-985  - - FGNF - - VVNSPINQGTPLSCSPNIENRGSMLHSPAHASNVGS - PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM -
mr_taeeniopygia/1-981  - - FGNF - - SMHSPMGGTPLSRSPNVENRGSMLHSPAHISNVGS - 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 - - PVHSPITQGTSLTCSPSVENRGSRSHPHSPHASNVGS - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN -
mr_homo/1-984    - - FGSF - - PVHSPITQGTPLTCS PNAENRGSRSHPAHASNVGS - PLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN -
mr_equus/1-984   - - FGNF - - TVHSPITQGTPLTCS PNVENRGSRSHPAHASNVGS - PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
```

JalView with Regular Expression searches

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns

```
      240      250      260      270      280      290      300      310      320
nr_danio/1-970  - - - TYF - - DSDCP - TLDSATSSLTHCQHT SP NIC SP VKSSIVG SP PLP SP LSVMK SP VSS SP HSIGSVR SP LSC - - NTNMRSSVS SP TTNG
nr_rutilus/1-973  - - - TYF - - DSDCP - SLASASTNLTQGHHT SP NTCS SP VKSSMVG SP PLAS SP LSVMK SP VSS SP RSIGSVR SP LSC - - NTNMRSSVS SP TTNG
nr_cyprinus/1-971  - - - TFF - - DSDCP - SLASTHTNL IQGQHT SP NTCS SP VKSSVVG SP PLAS SP LSVIK SP VSS SP HSIGSVR SP LSC - - NTNMRSSVS SP TTYG
nr_oryzias/1-994  TCFGPPQCSAVS SP VSQTSCAATLANIKRRNSVTC SP VESCTVG SP PLTS SP LNIMR SP MSS SP HSMSSVR SP PSCSTTCNIRSSVS SP T - - -
nr_takifugu/1-991  MCFGPMCSVSS SP VSQTSCASTLPNIKRRNSATC SP VESSTVG SP PLTS SP LNIMR SP ISS PQSMSSVR SP PSCSTTSNIRSSVS SP T - - -
nr_oreochromis/1-994 TCFAPLCSSVSS SP VSQTSCAATLANIKRRNSVTC SP VESSTVG SP PLTS SP LNVMR SP MSS SP PQSMSSVR SP PSCSTTCNIRSSVS SP T - - -
nr_xenopus/1-979  - - - FGNF - - TVHSPVNQVTPKSC SP HTDNRC SIAHS P - - AGTVES - PLSS PVSSMR SP ISS PPSHASLK SP VSS SP PNNITVRPSVSS SP GNI -
nr_anolis/1-990  - - - FGNF - - TVSSPVNQGTPLSC SP NIENRGSMLH SP PHASNMGS - PLSS PISSMK SP ISS PPSHCSVK SP VSS SP PNNITMRSSVS SP PANM -
nr_alligator/1-985  - - - FGNF - - VVNSP INQGTPLSC SP NIENRGSMLH SP PAHASNVGS - PLSS PISSMK SP ISS PPSHCSVK SP VSS SP PNNITMRSSVS SP PANM -
nr_taeniopygia/1-981  - - - FGNF - - SMHSP MGQGTPLSR SP NVENRGSMLH SP PAHISNVGS - PLSS PISSMK SP ISS PPSHCSVK SP VSS SP PNNITMRSSVS SP PANL -
nr_gallus/1-986  - - - FGNF - - AMHSP IGQGTPLSR SP NVESRGSMLH SP PAHVSNVGS - PLSS PISSMK SP ISS PPSHCSVK SP VSS SP PNNITMRSSVS SP PANM -
nr_monodelphis/1-993  - - - FGSF - - PVHSP ITQGTPLPC SP NVENRSSVSH SP PAHASNVGS - PLSS PISSMK SP ISS PPSHCSVK SP VSS SP PNNVTMRSSVS SP PANIN -
nr_mus/1-980  - - - FGSF - - PVHSP ITQGTSLTC SP SVENRGRSH SP VHASNVGS - PLSS PLSSMK SP ISS PPSHCSVK SP VSS SP PNNVPLRSSVS SP PANLN -
nr_rattus/1-981  - - - FGSF - - PVHSP ITQGTSLTC SP SVENRGRSH SP THASNVGS - PLSS PLSSMK SP ISS PPSHCSVK SP VSS SP PNNVPLRSSVS SP PANLN -
nr_homo/1-984  - - - FGSF - - PVHSP ITQGTPLTC SP NAENRGRSH SP PAHASNVGS - PLSS PLSSMK SSISS PPSHCSVK SP VSS SP PNNVTLRSSVS SP PANIN -
nr_equus/1-984  - - - FGNF - - TVHSP ITQGTPLTC SP NVENRGRSH SP PAHASNVGS - PLSS PLSSMK SP ISS PPSHCSVK SP VSS SP PNNVTLRSSVS SP PANIN
```



JalView with Regular Expression searches

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns

JaView with Regular Expression searches

- Regular Expressions:

[LS]P.A

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

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns

JaView 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**

Detection of repeats

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

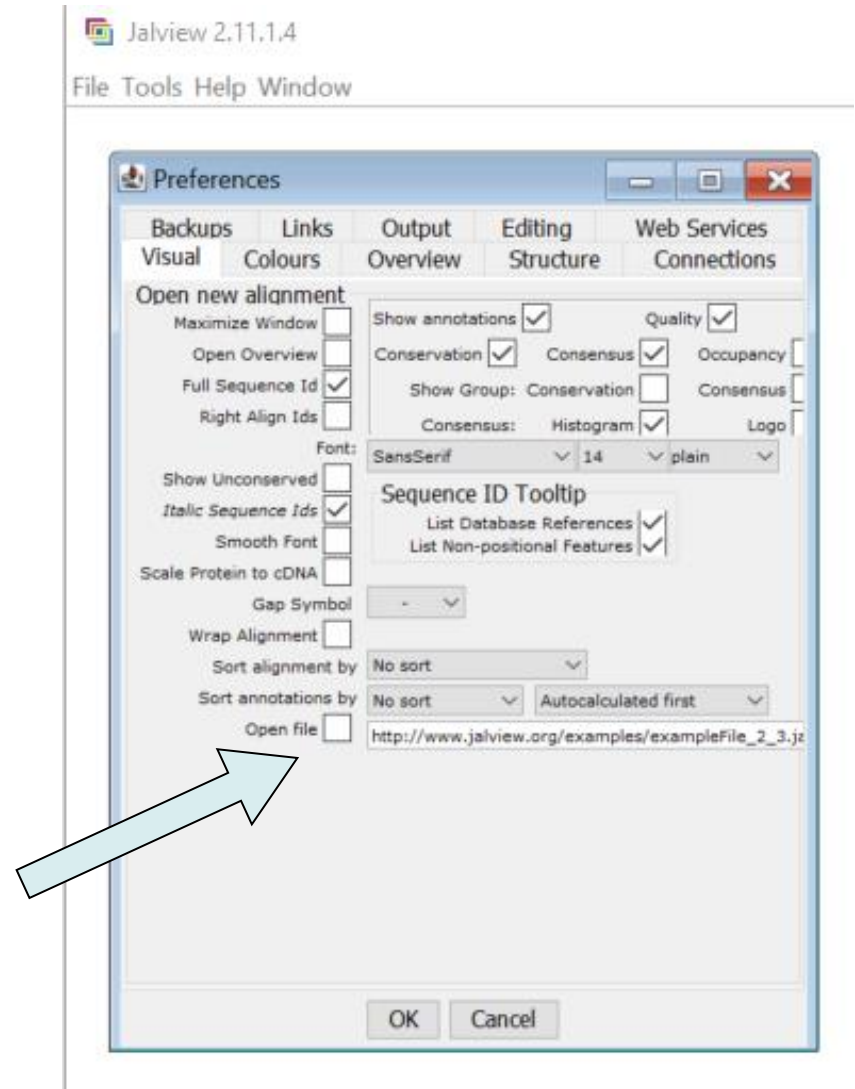
- Load the multiple sequence alignment of the MR in JalView: MR1_fasta.txt (from URL: https://cbdm.uni-mainz.de/files/2015/02/MR1_fasta.txt)
- Use the "Select > find" (of Ctrl+F) option with a regular expression and mark all matches (**click the "Find all" option!**)
- Try to find the expression that matches more repeats. How many repeats do you see? How long are they? Would you correct the alignment based on these findings?

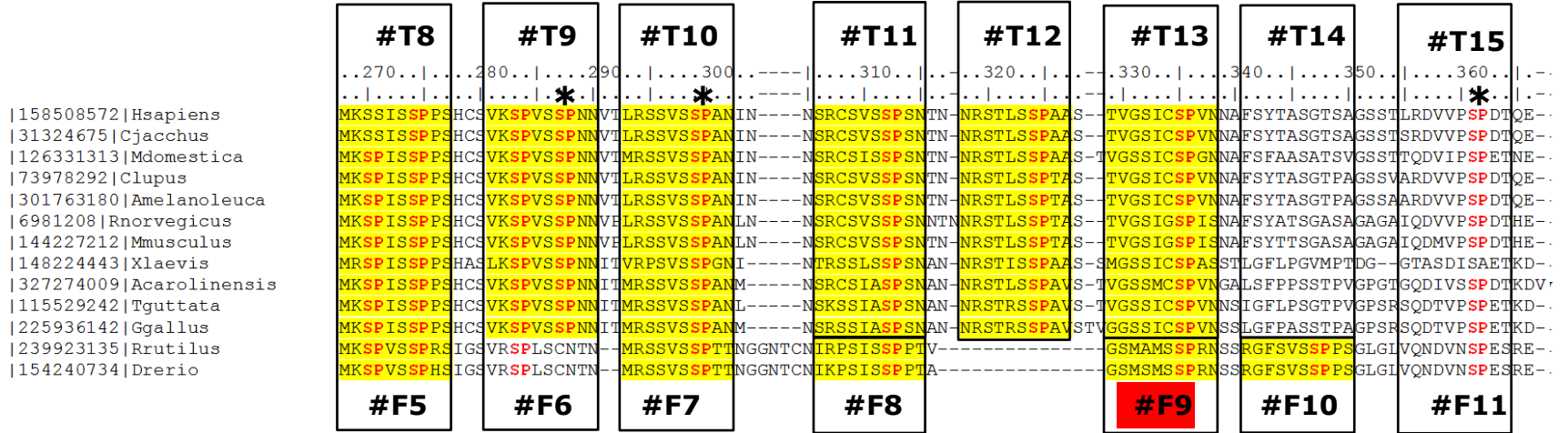
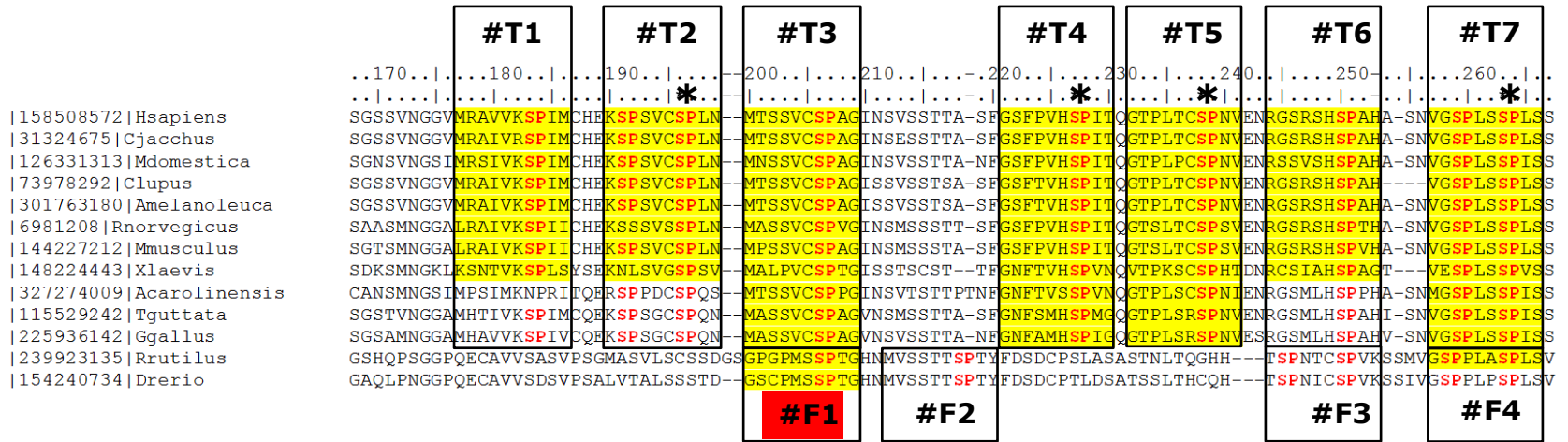
First time running JalView?

Remove annoying start:

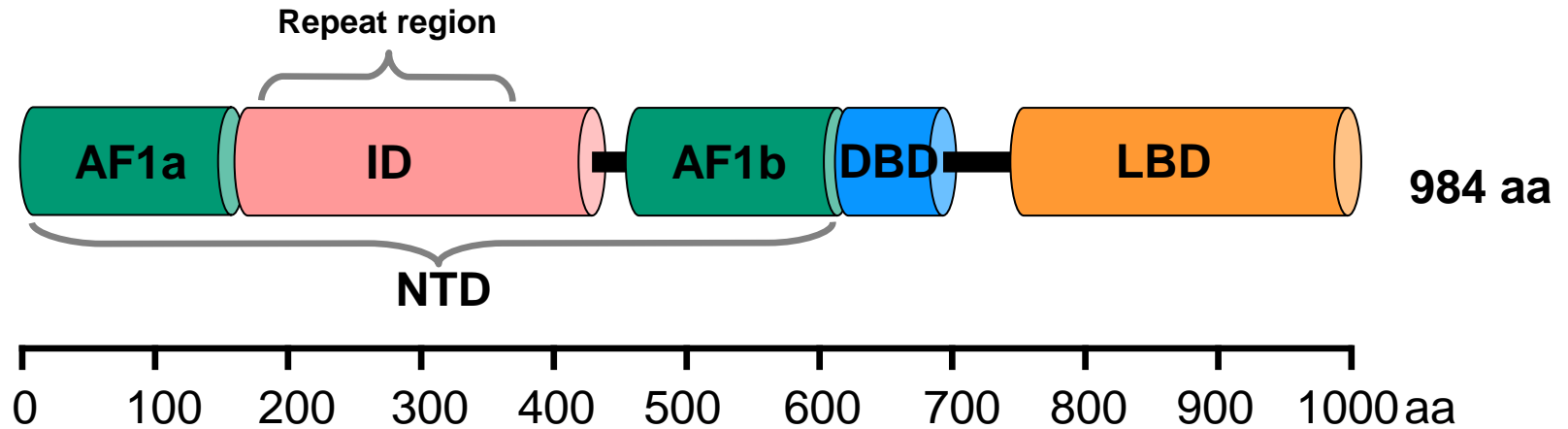
Go to
Tools > Preferences > Visual

Un-tick option "Open file"

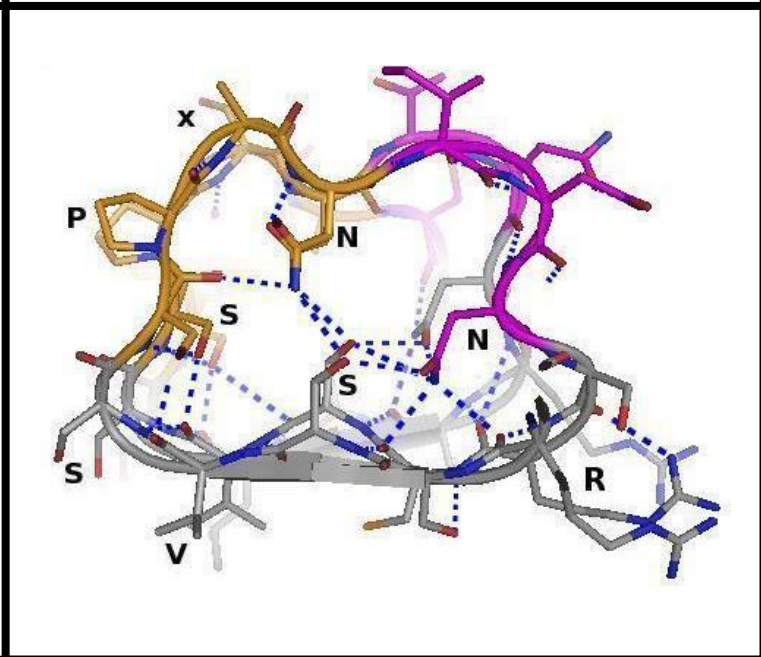
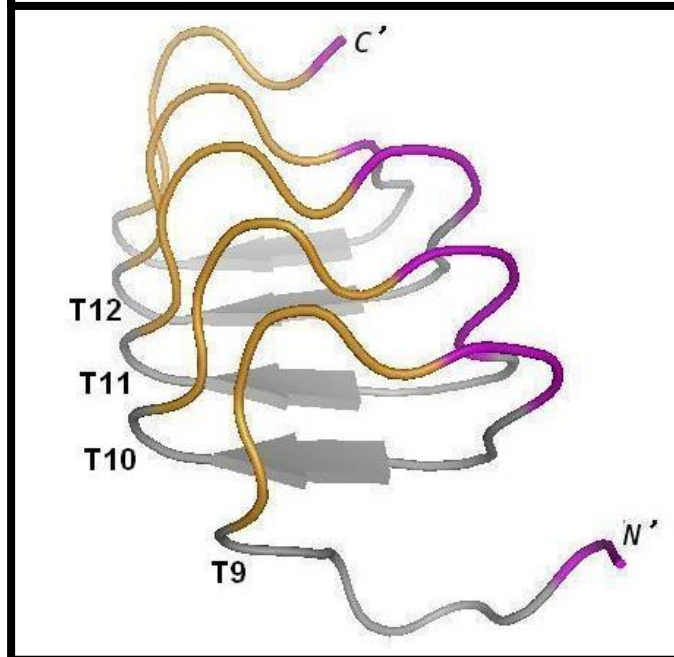
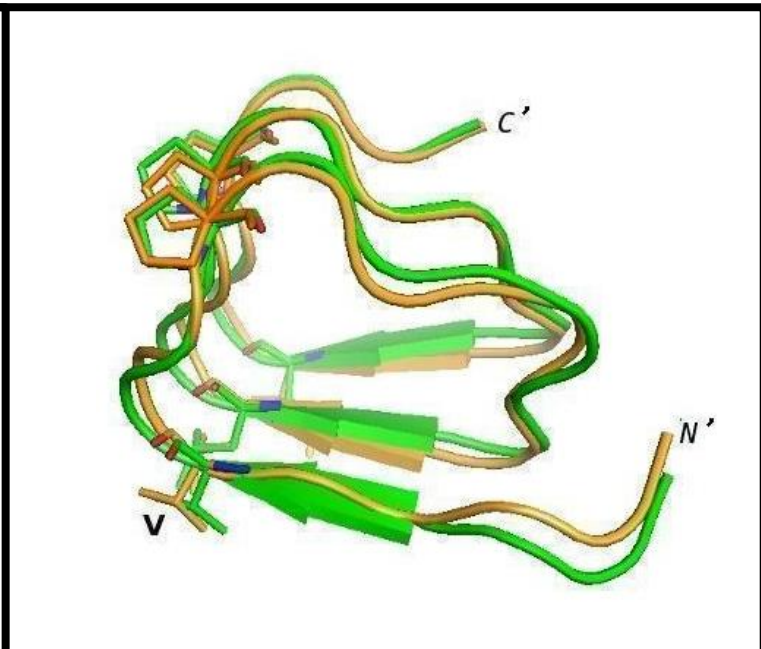
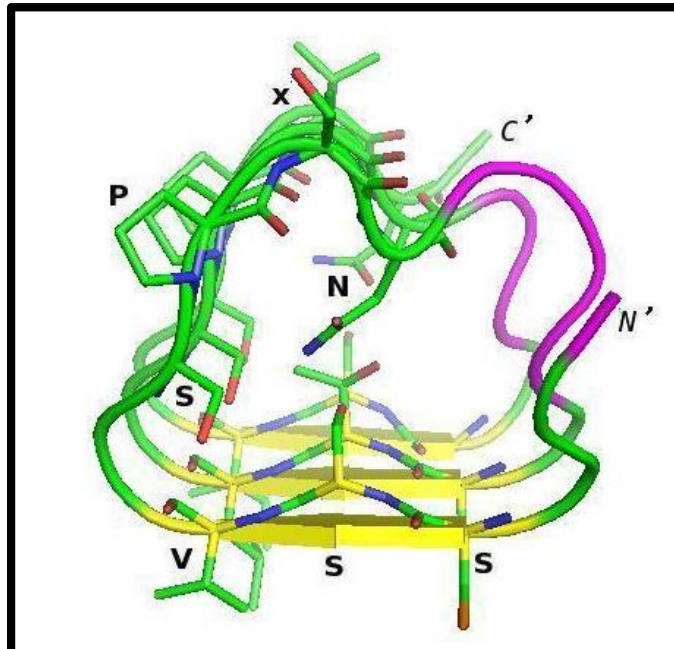




Mineralocorticoid receptor



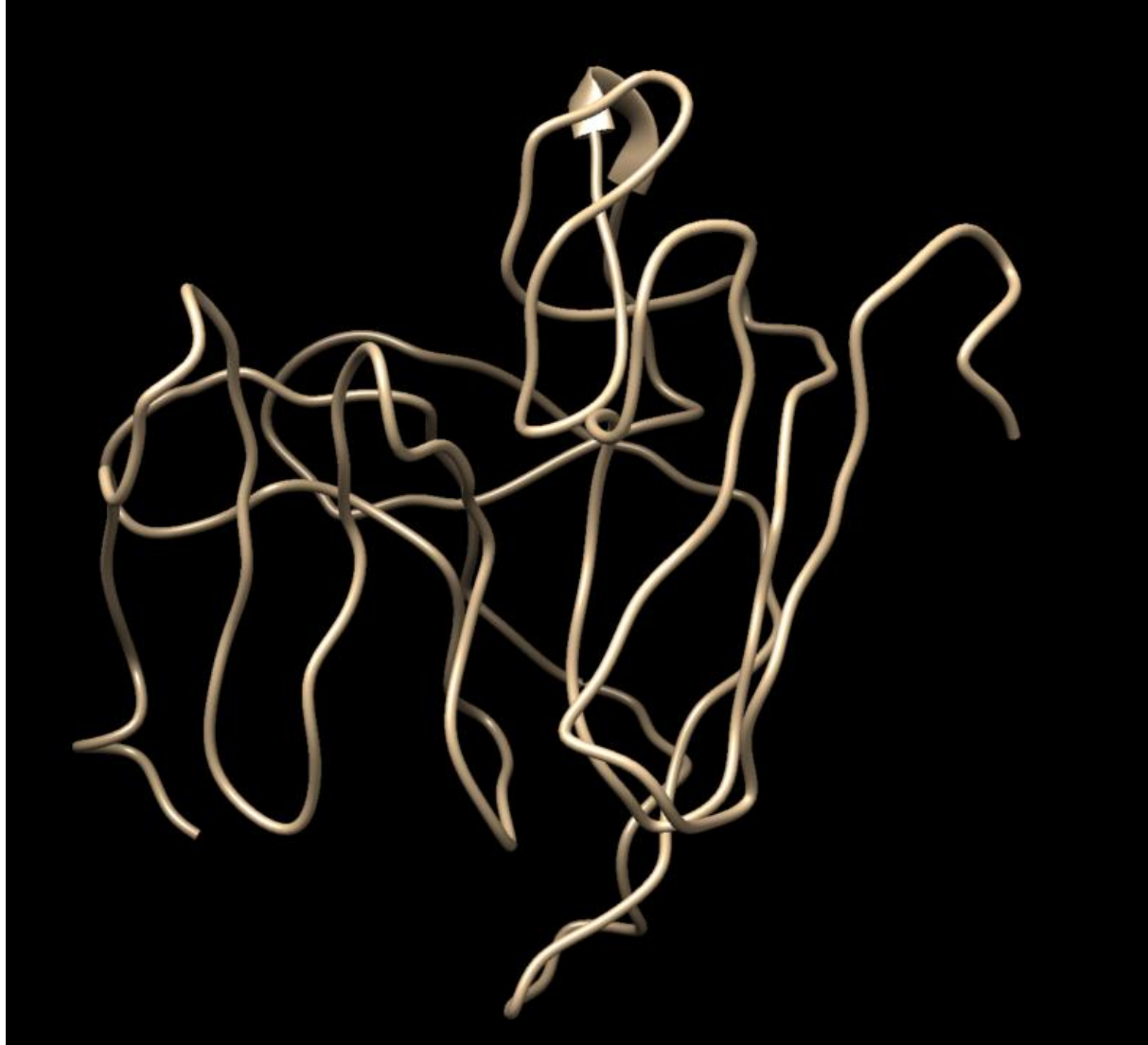
Vlassi *et al.* (2013) *BMC Struct. Biol.*



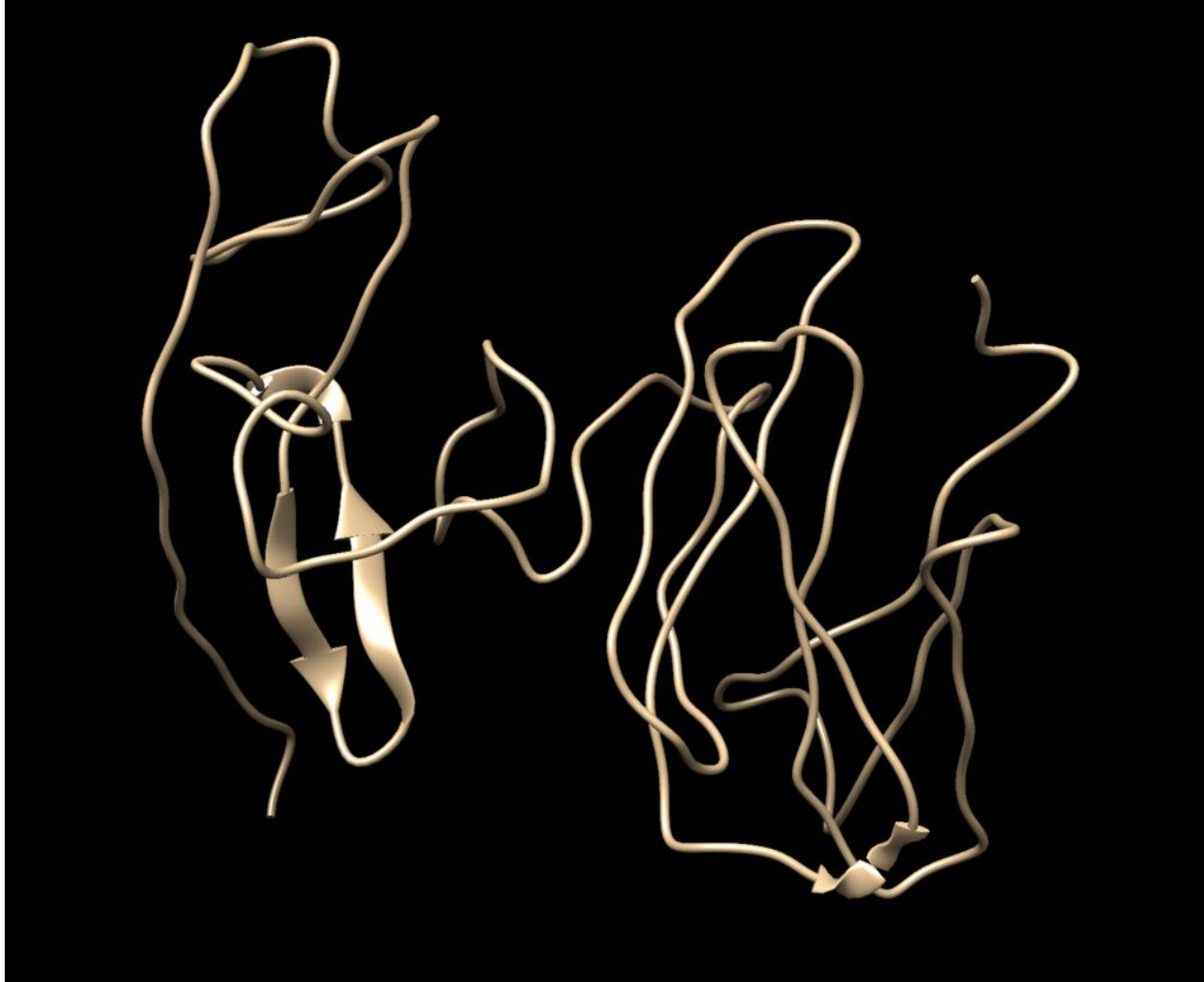
Modelling with AlphaFold...



Modelling with CI-Tasser...



Modelling with trRosetta...



Composition bias

Definition

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

1: Single amino acid repeats.

2: Longer imperfect tandem repeats.
Assemble in structure.

Definition CBRs

Perfect repeat: QQQQQQQQQQQQ

Imperfect: QQQQPQQQQQQ

Amino acid type: DDDDEEEDEDEED

Compositionally biased regions (CBRs)

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

Particular case of low complexity region

Function 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 CBRs

Conservation => Function

Length, amino acid type not necessarily conserved

Functions:

Passive: linkers

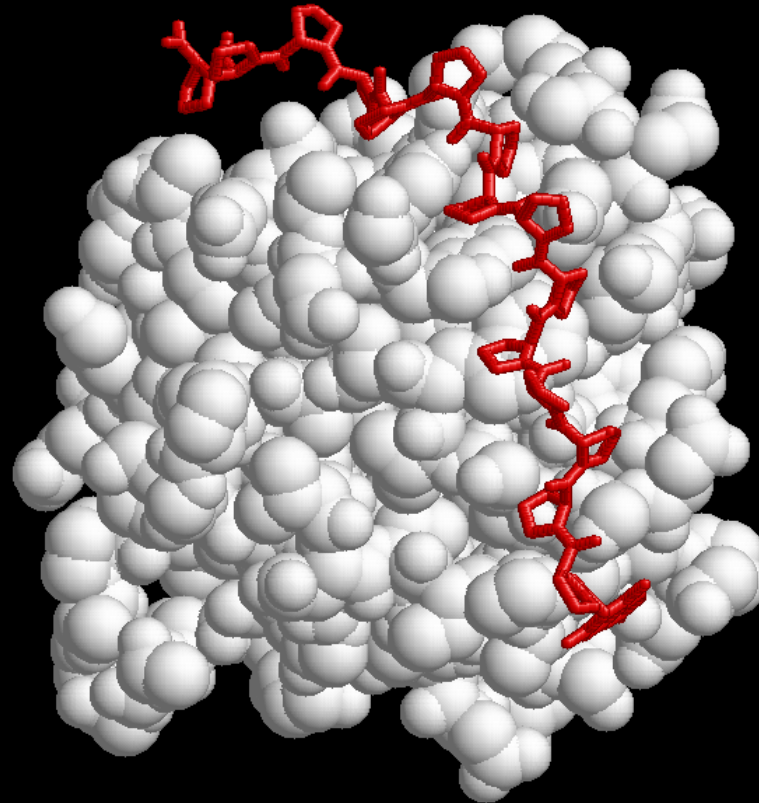
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

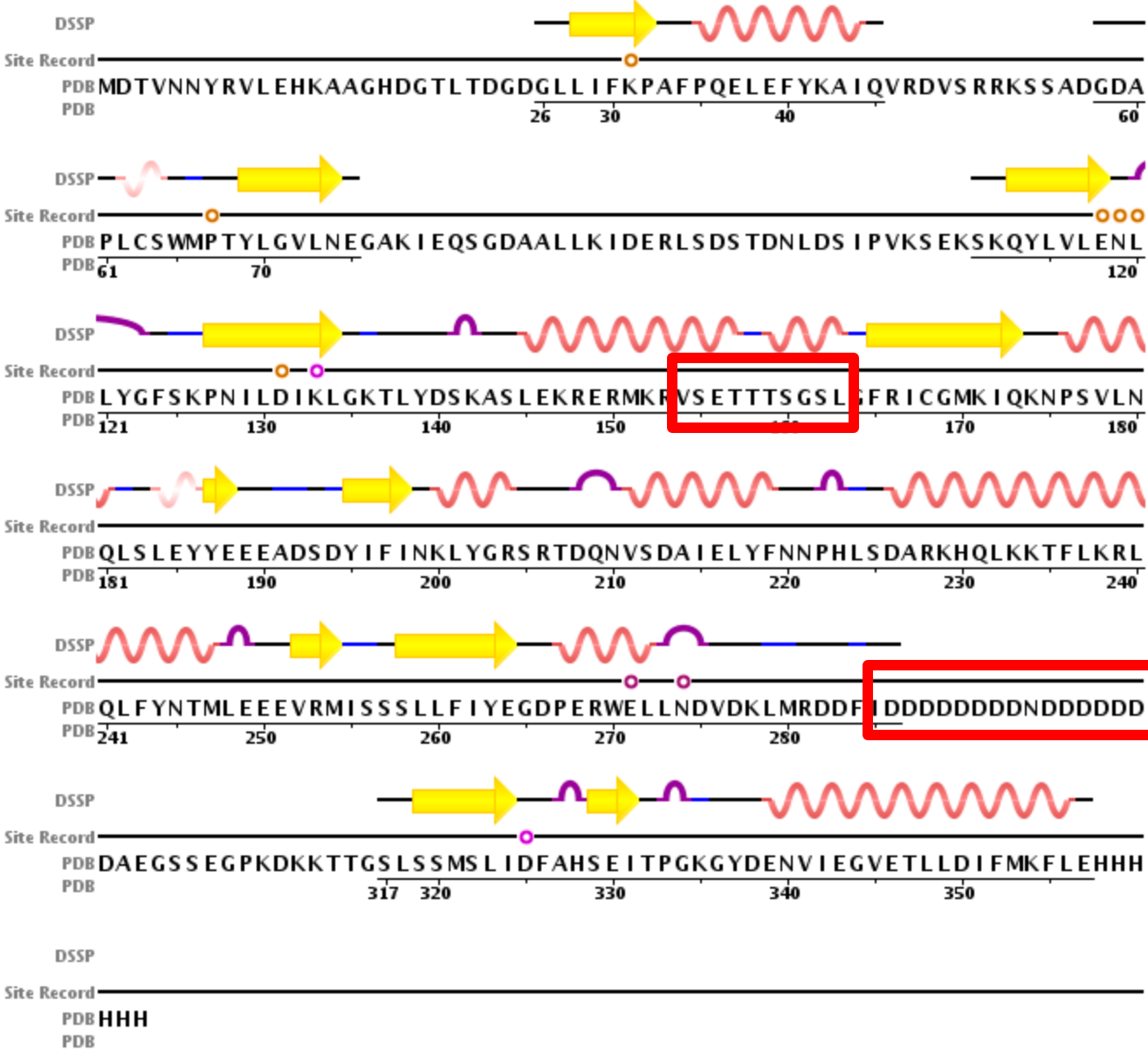
Structure of CBRs

Often variable or flexible: do not easily crystallize

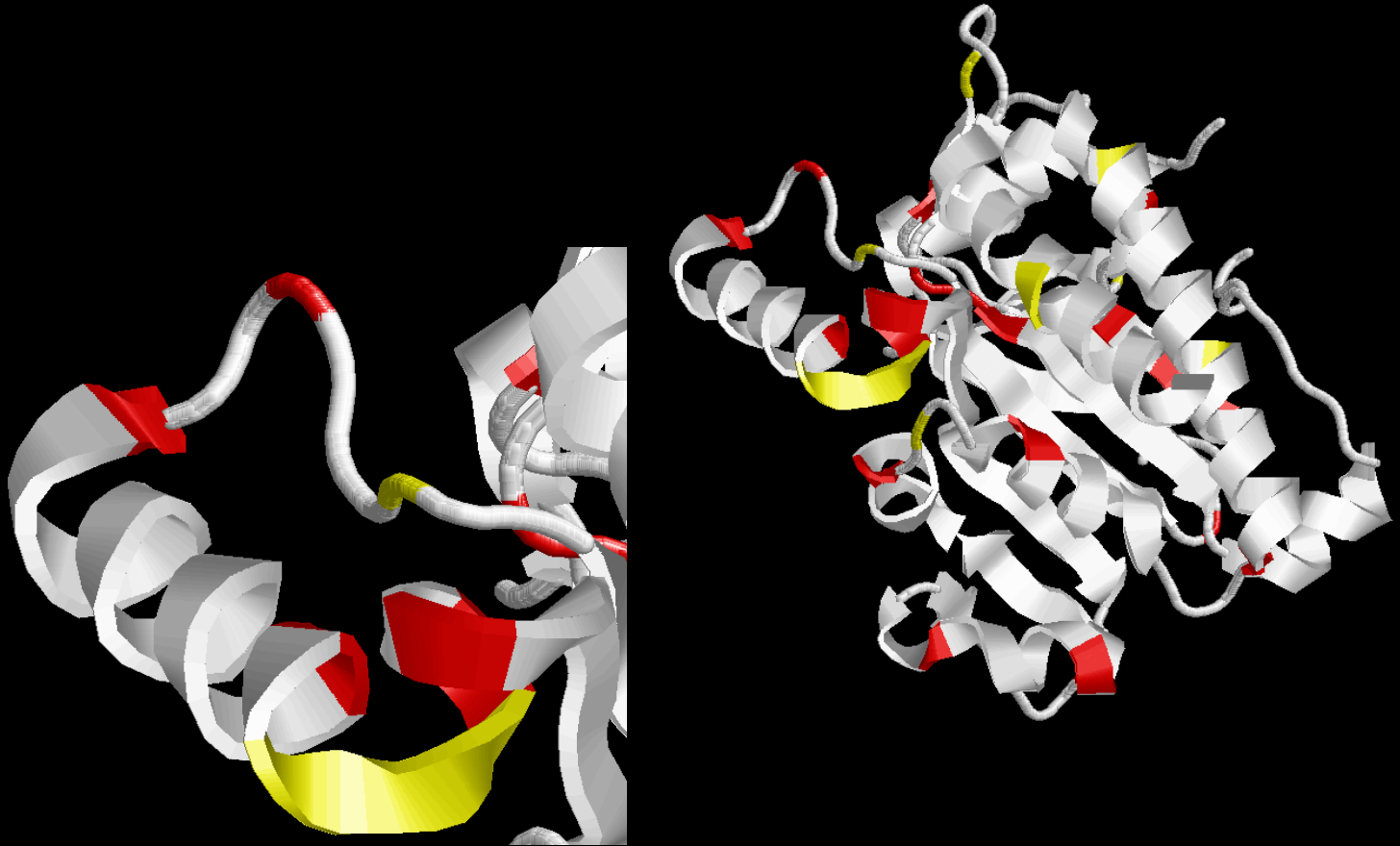
1CJF: profilin bound to polyP



2IF8: Inositol Phosphate Multikinase Ipk2

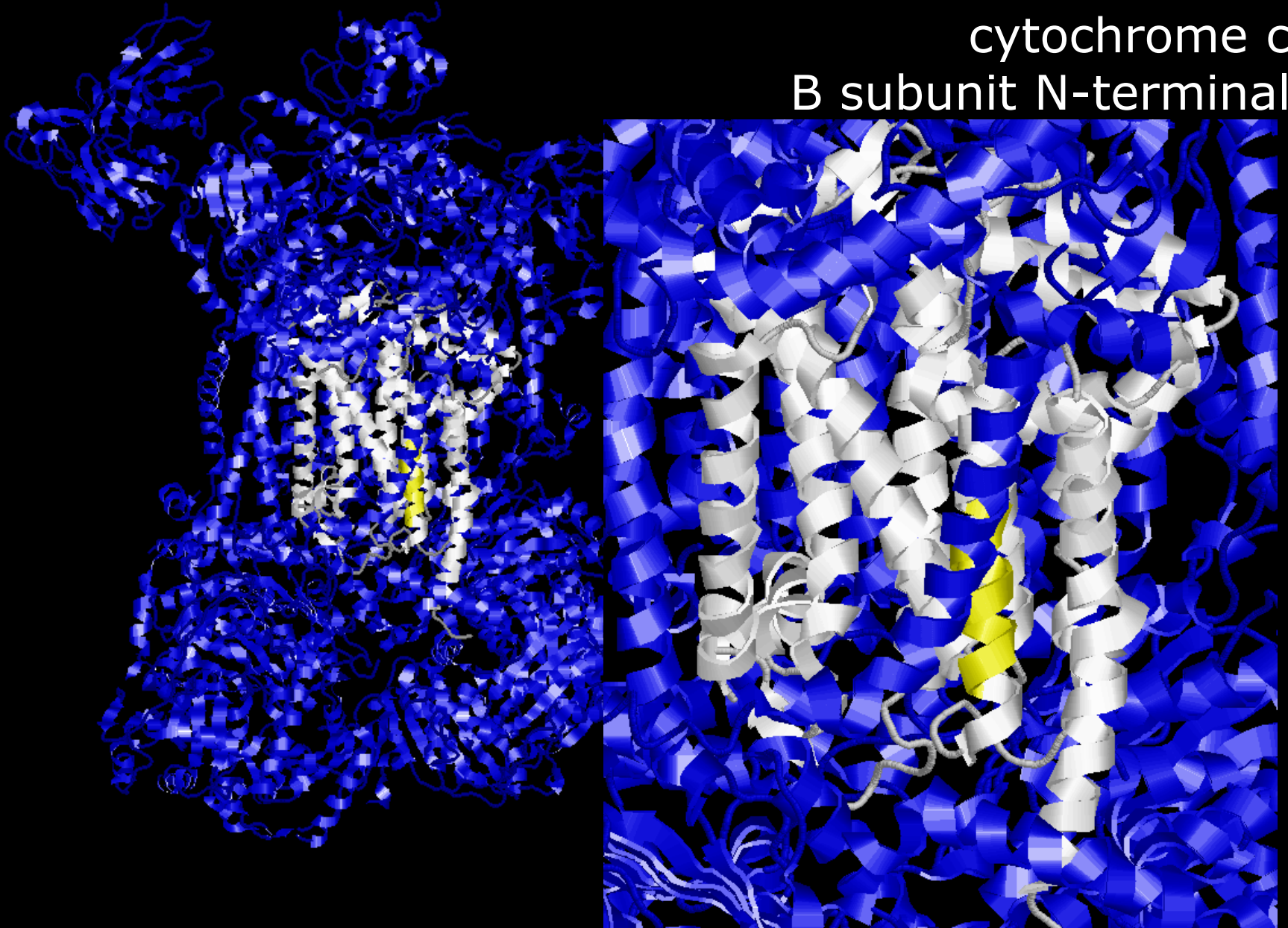


2IF8: Inositol Phosphate Multikinase Ipk2

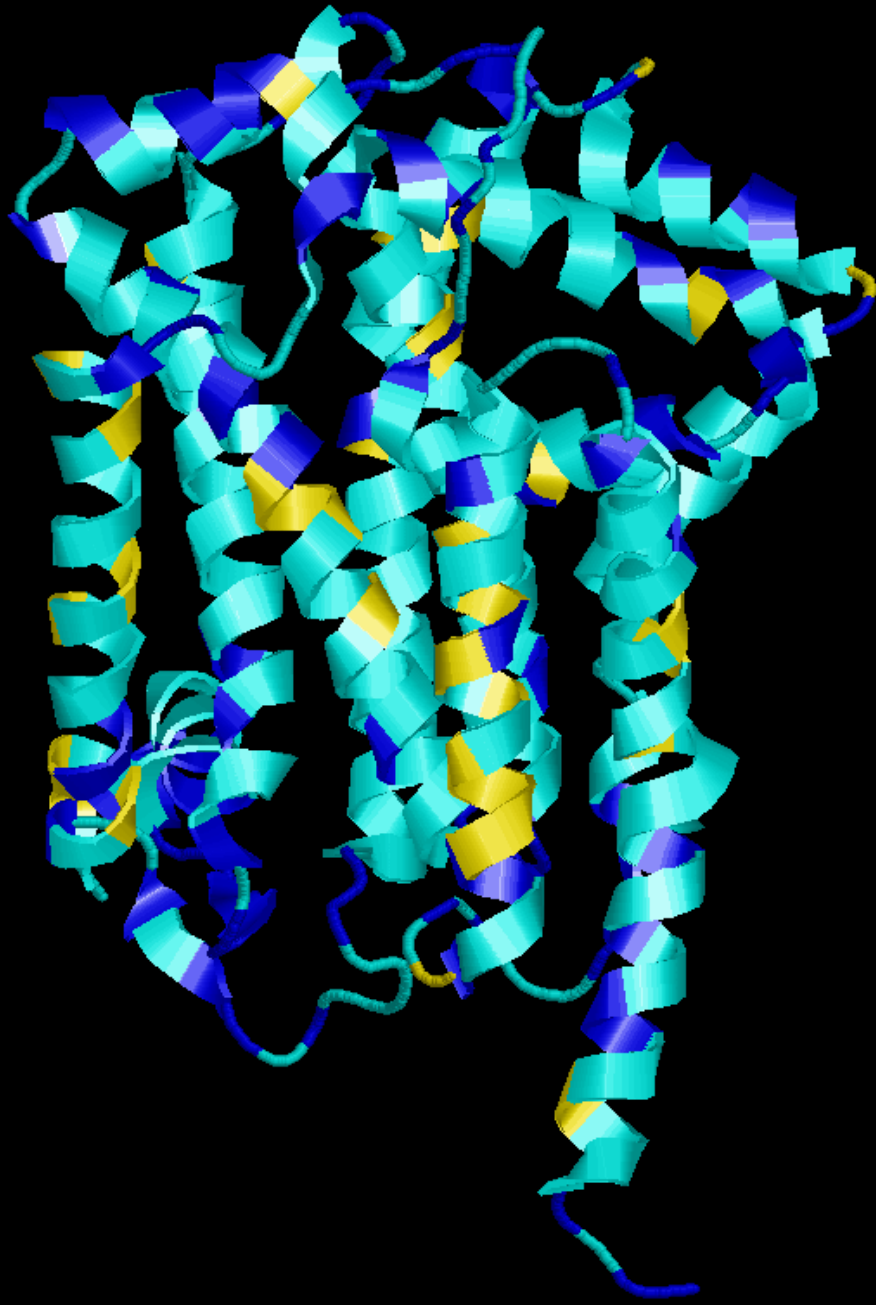


RVSE**TTT**SGSL

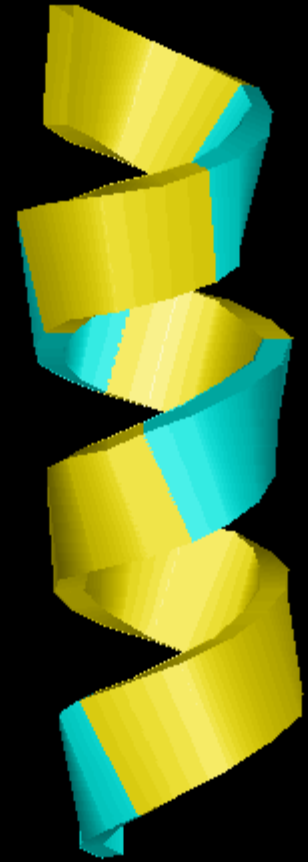
2CX5: mitochondrial
cytochrome c
B subunit N-terminal



2CX5: mitochondrial
cytochrome c
B subunit N-terminal



**FEFFIEVNE
FEFFIEVNE**



Amino acid repeats

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

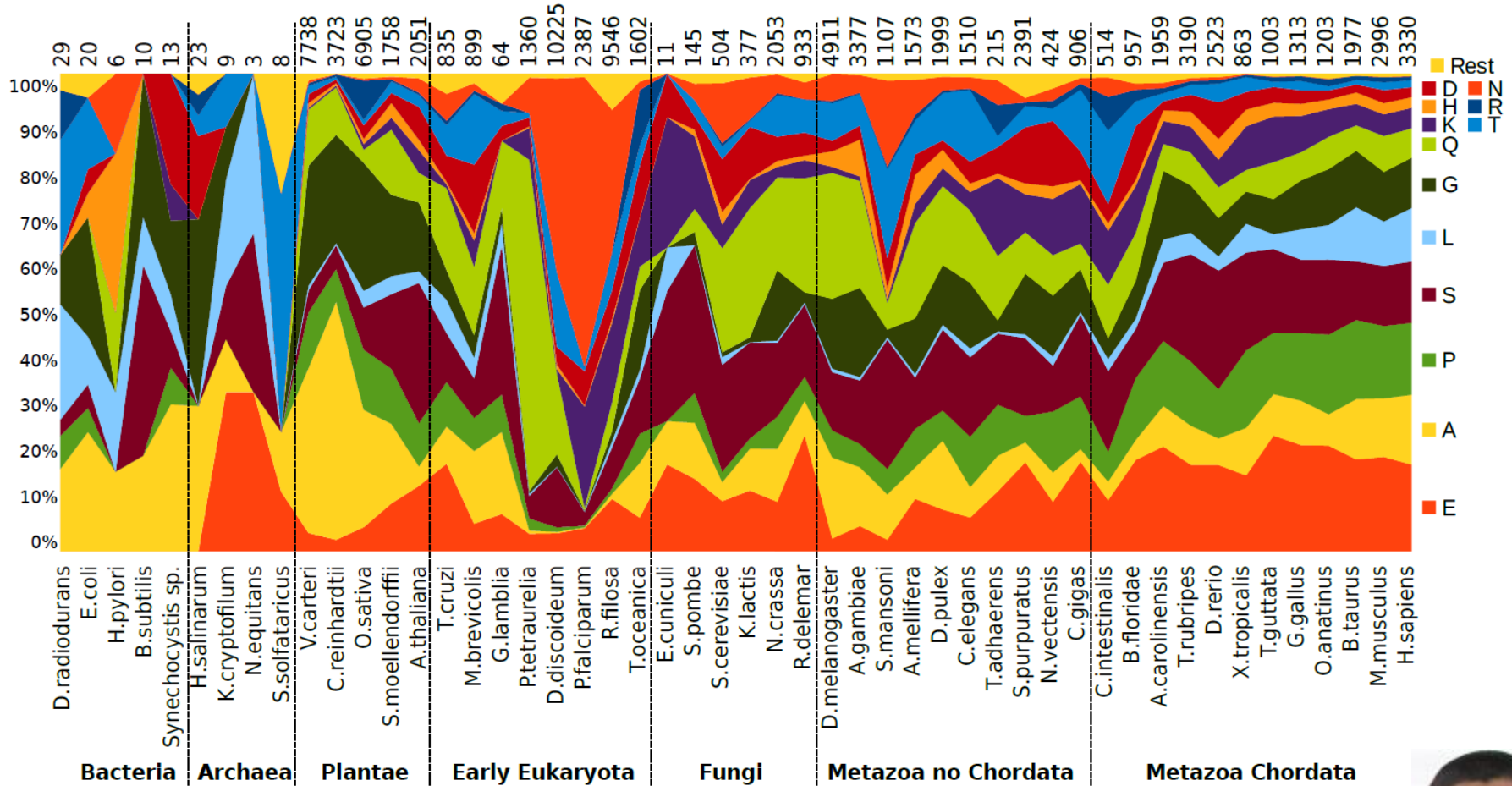
Very rare or absent in both eukaryota and prokaryota:

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

Toxicity of long stretches of hydrophobic residues.

(Faux et al 2005)

Amino acid repeats



Pablo Mier

Homorepeats are frequent but difficult to characterize



Pablo
Mier

e.g. polyQ:

MATLEKLMKAFESLKSF **QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ** **PPPPPPPPPP**QLPQP

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

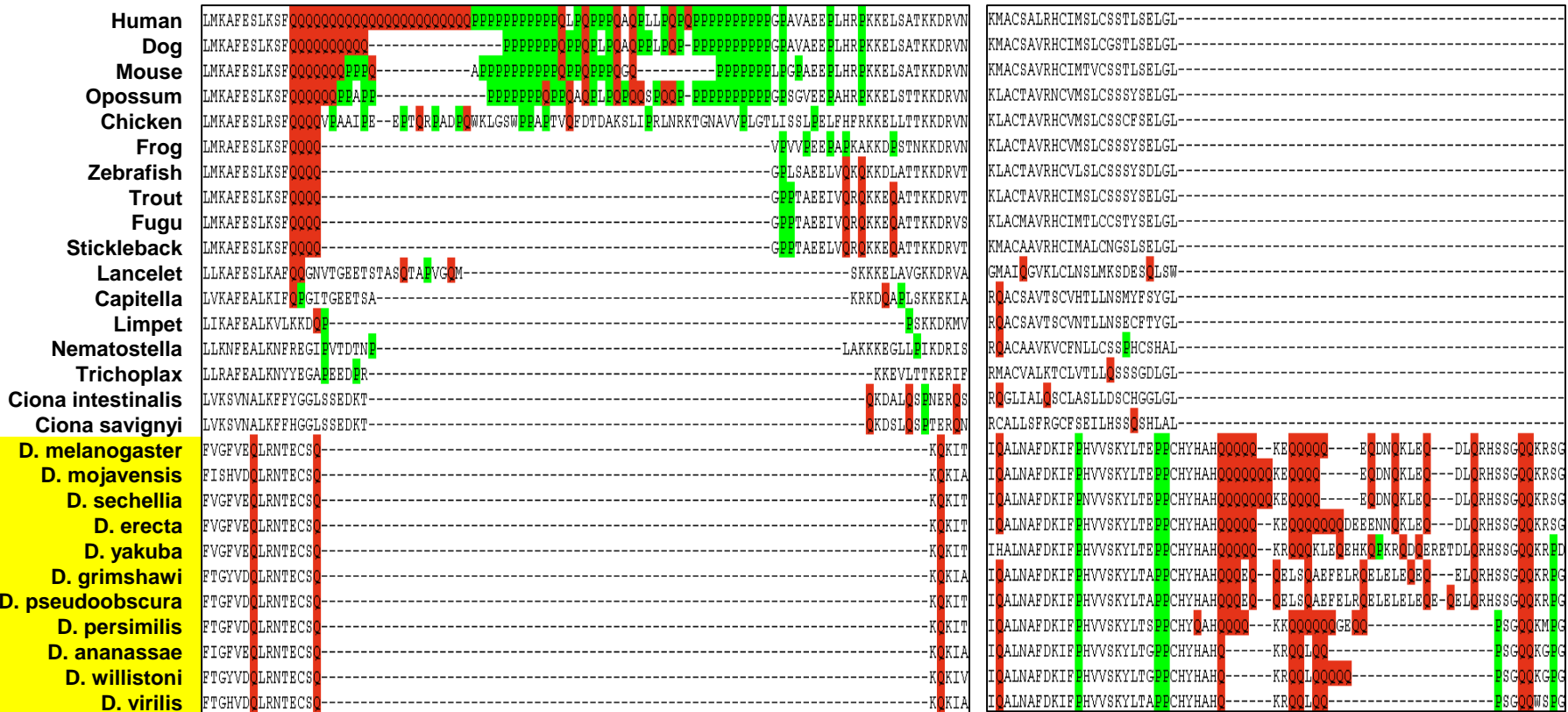
Homorepeats need to be studied in context

Function of polyQ

Martin Schaefer



polyQ in Huntingtin



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?