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

MRAVVKSPIMCHEKSPSVCSPLNMTSSVCSPAGINSVSSTTASF
GSFPVHSPITQGTPLTCSPNVENRGSRSHSPAHASNVGSPLSSP
LSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINN

Definition repeats

Sequence, long, 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, long, imperfect, tandem

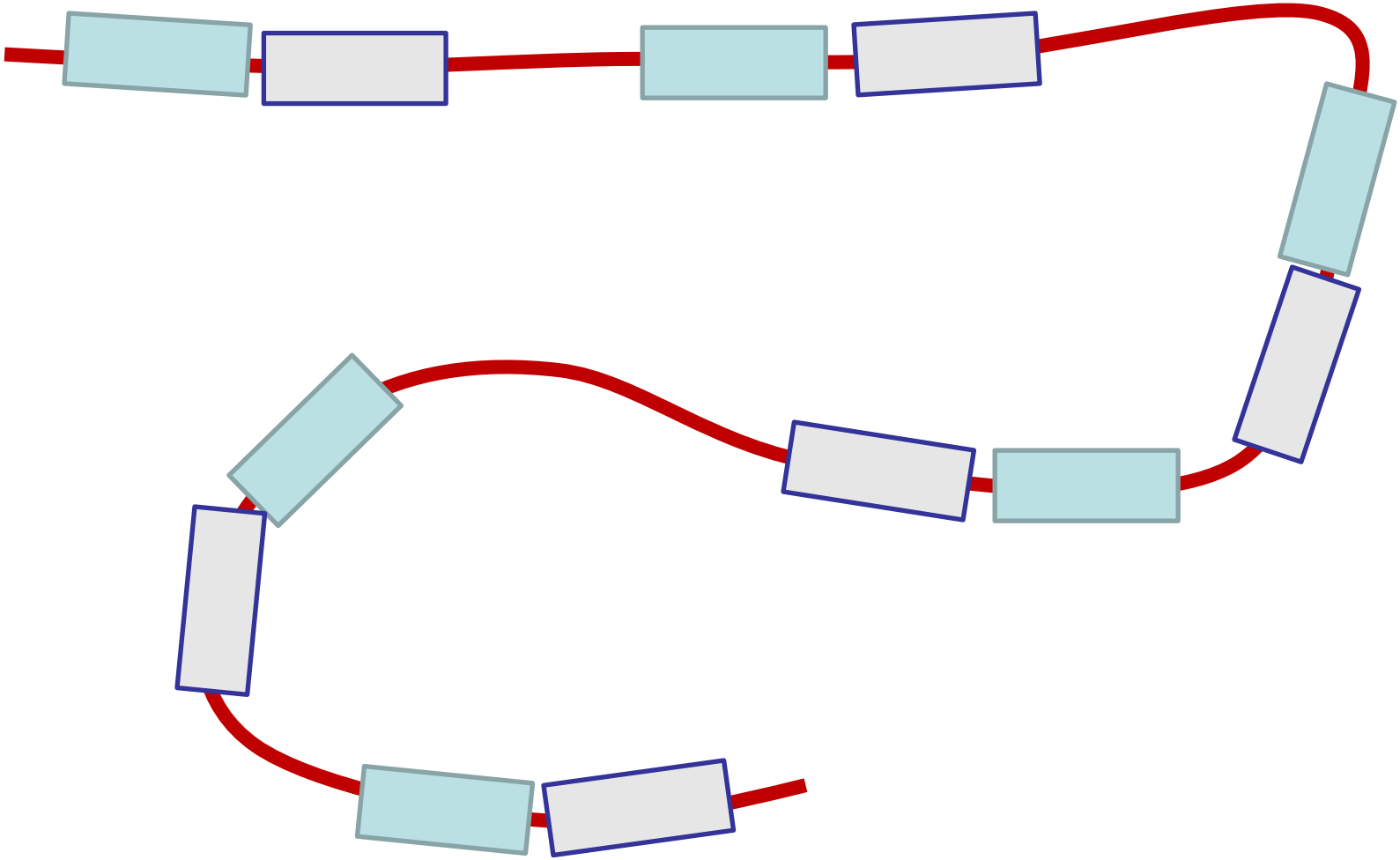
MRAVVK**SP**IM CHE
KSPSVC**SP**LN
MTSSVC**SP**AG INSVSSTTASF
GSFPVH**SP**IT Q
GTPLTC**SP**NV EN
RGSRSH**SP**AH ASN
VGSPLS**SP**LS S
MKSSIS**SP**PS HCS
VKSPVS**SP**NN VT
LRSSVS**SP**AN INN

Definition repeats

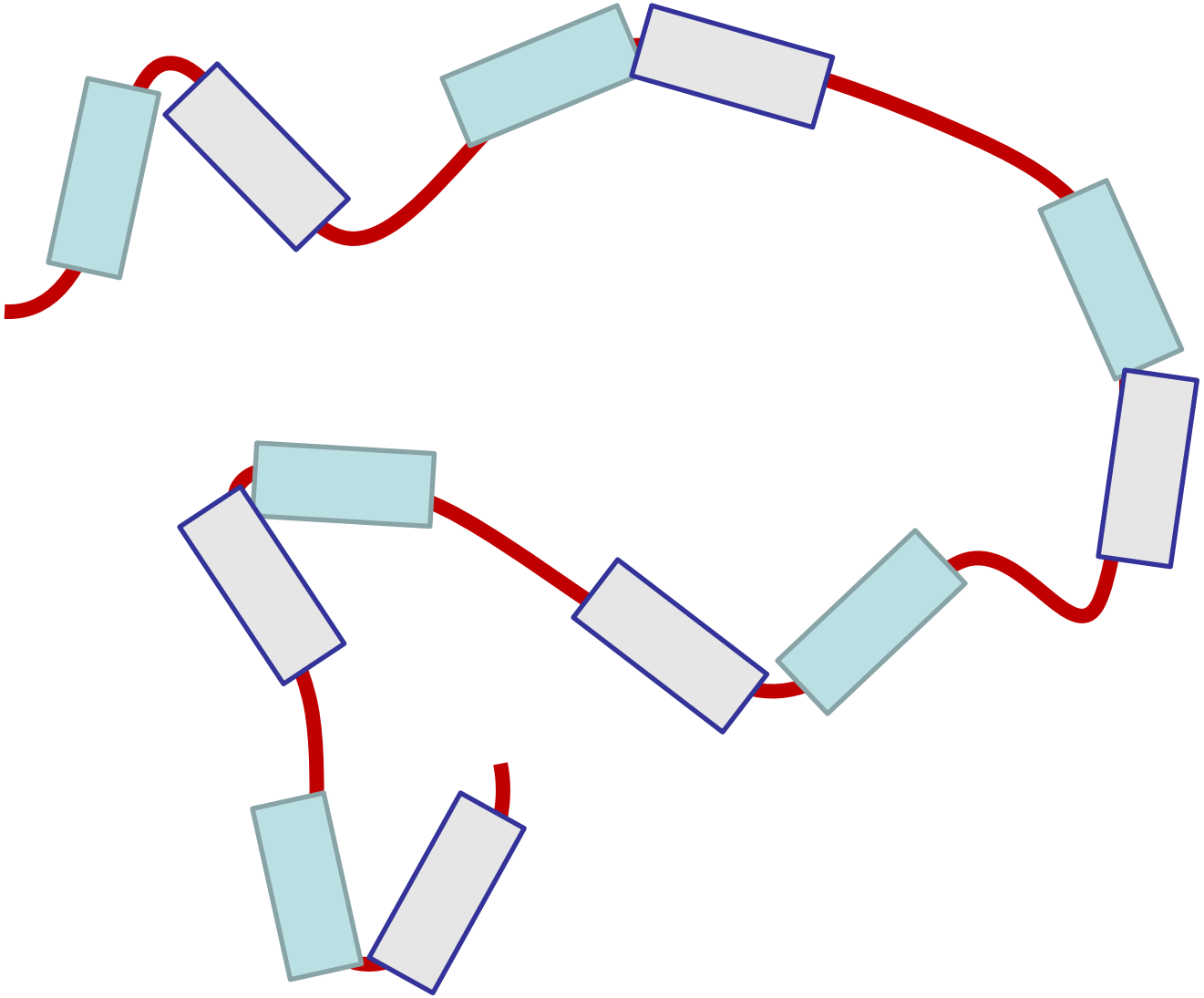
Sequence, long, imperfect, tandem

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

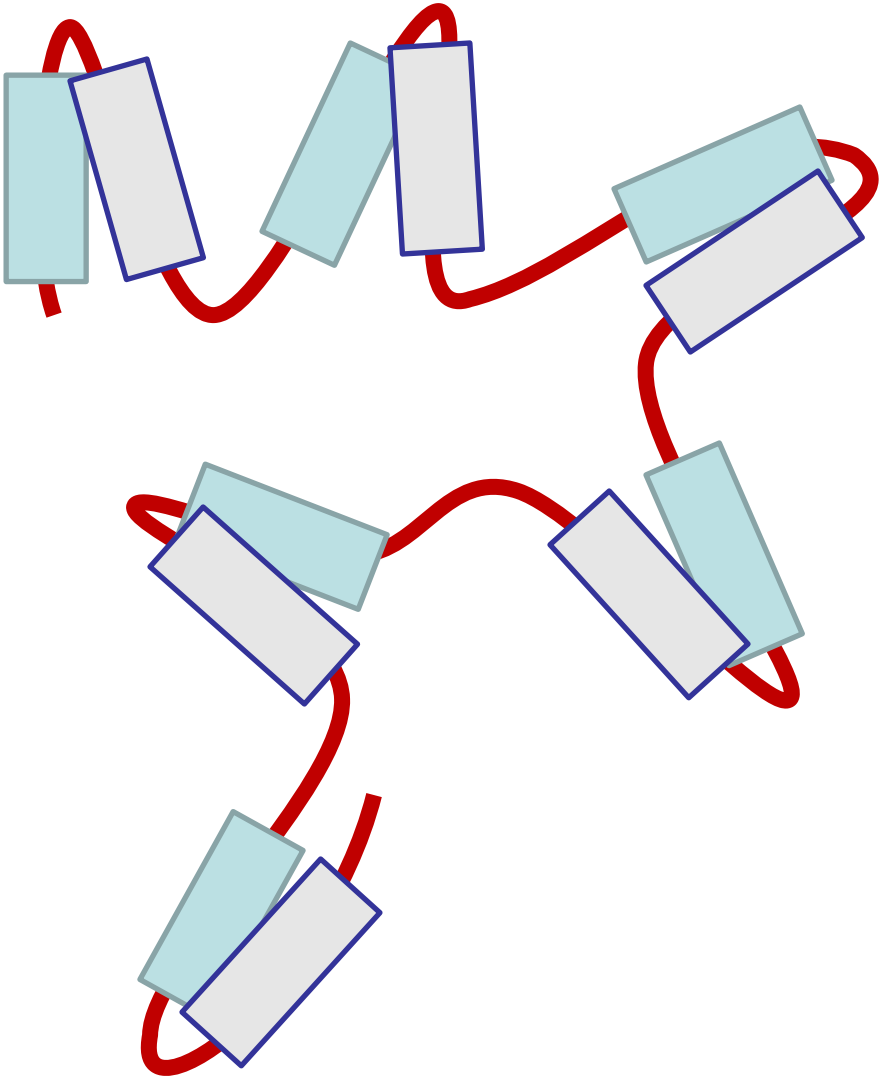
Tandem repeats fold together



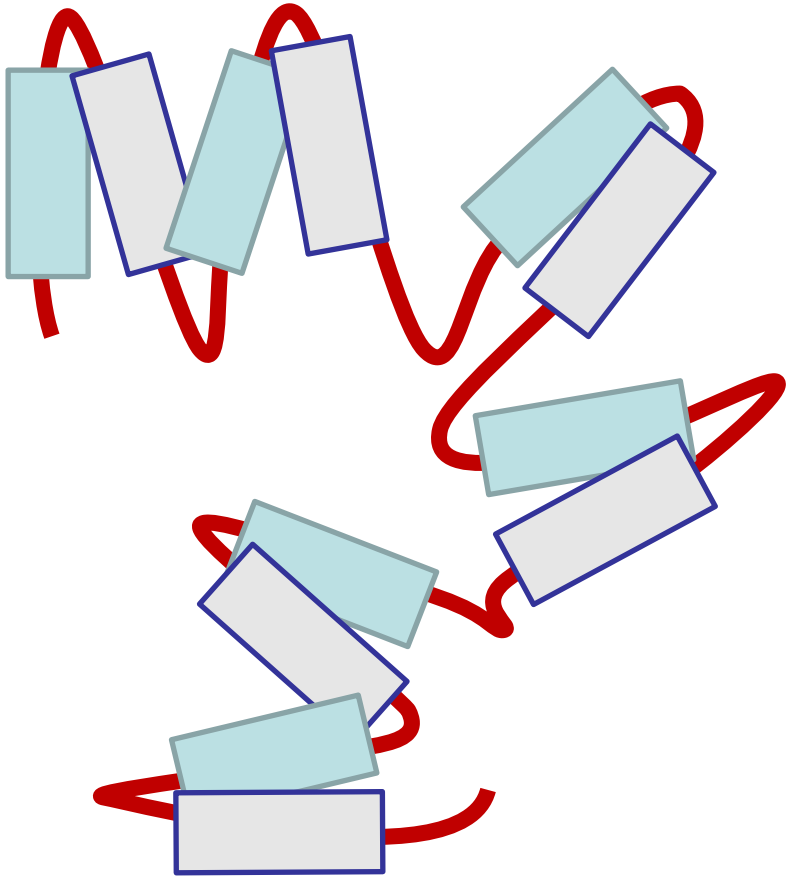
Tandem repeats fold together



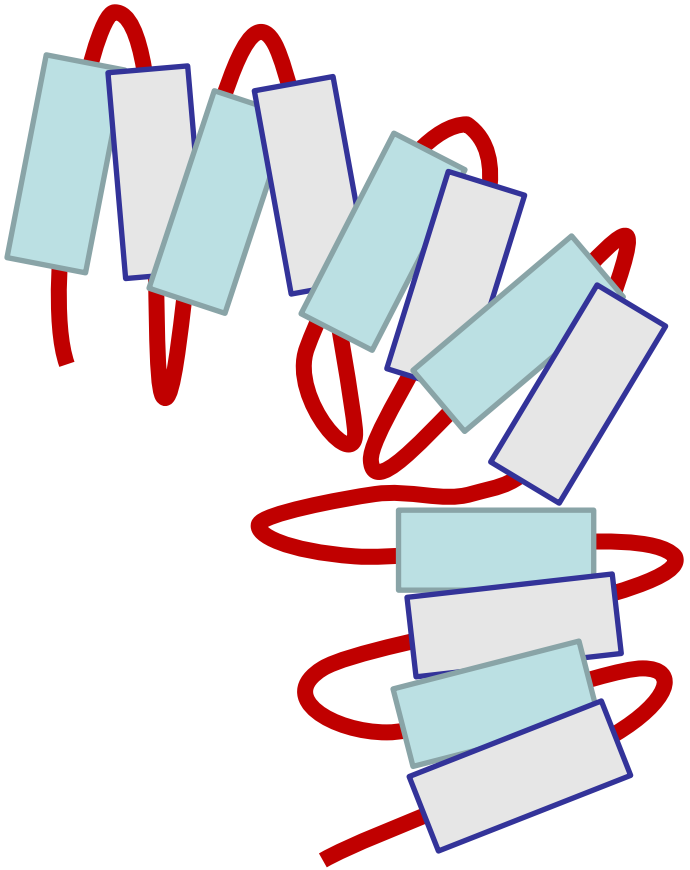
Tandem repeats fold together



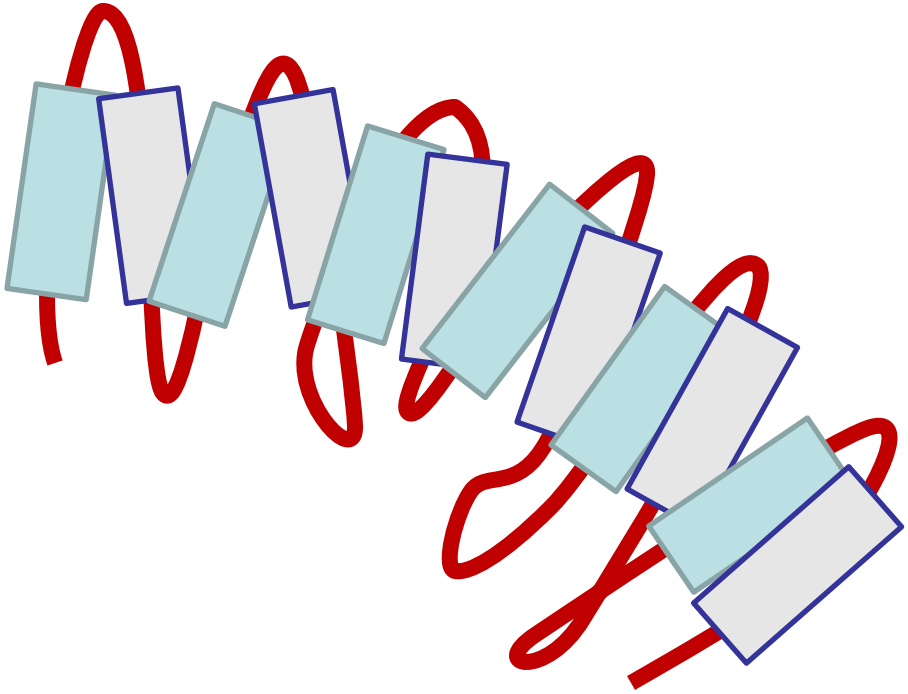
Tandem repeats fold together



Tandem repeats fold together



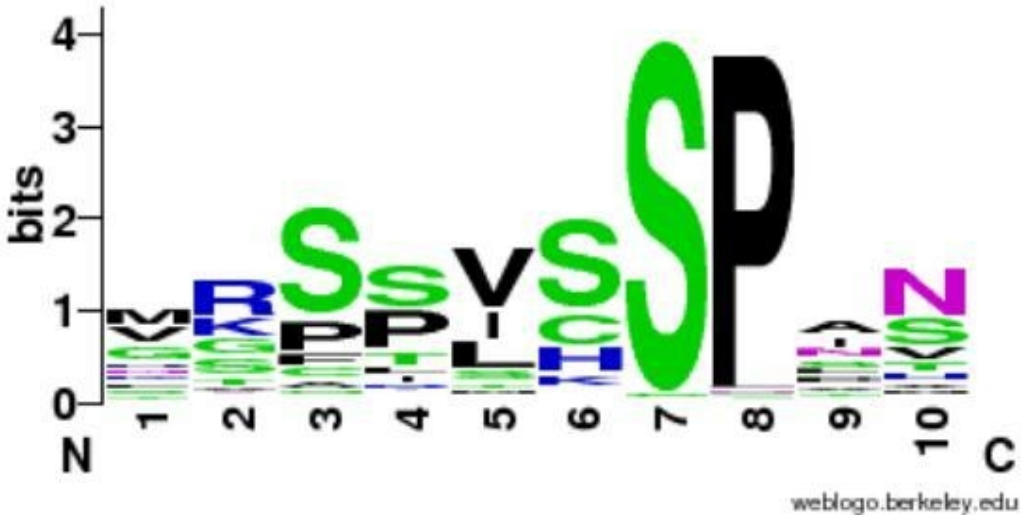
Tandem repeats fold together



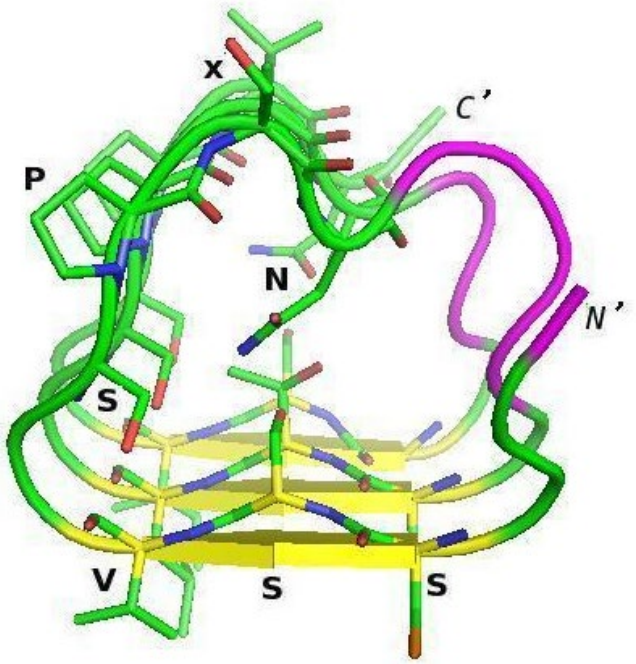
Definition repeats

Sequence, long, imperfect, tandem

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



http://weblogo.berkeley.edu

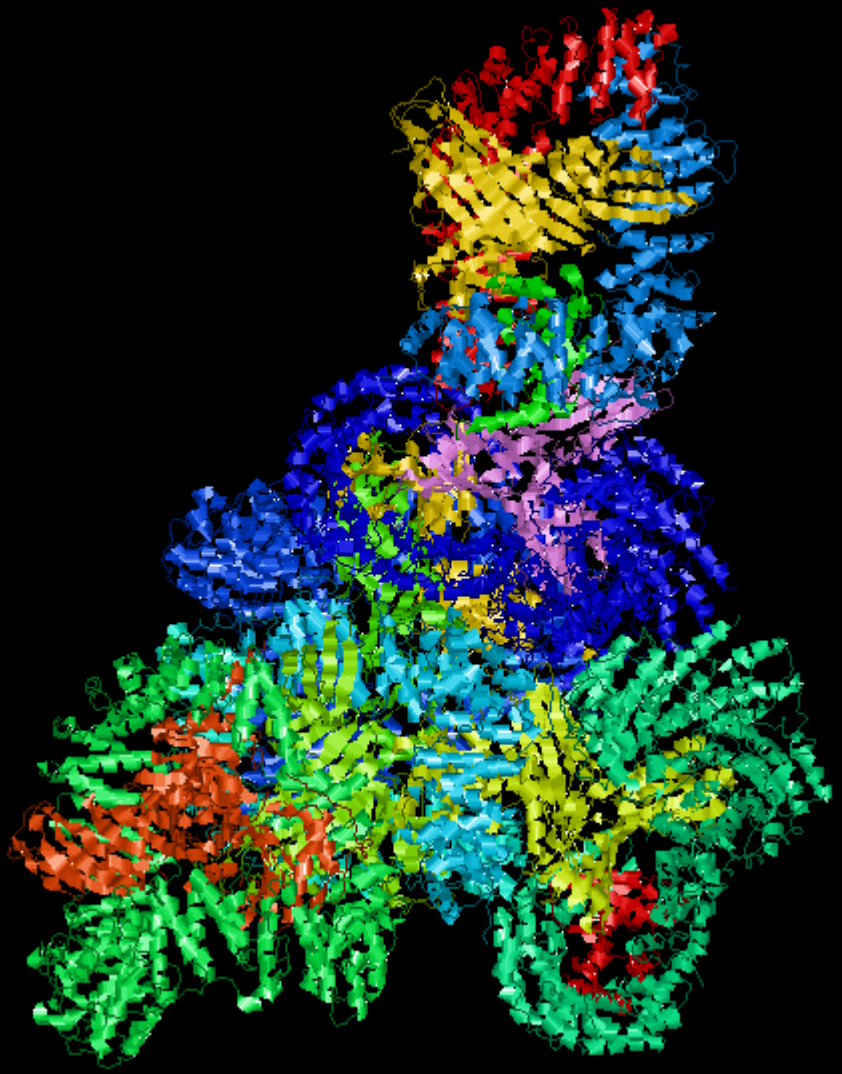


(Vlassi et al, 2013)

A subunit PP2A structure

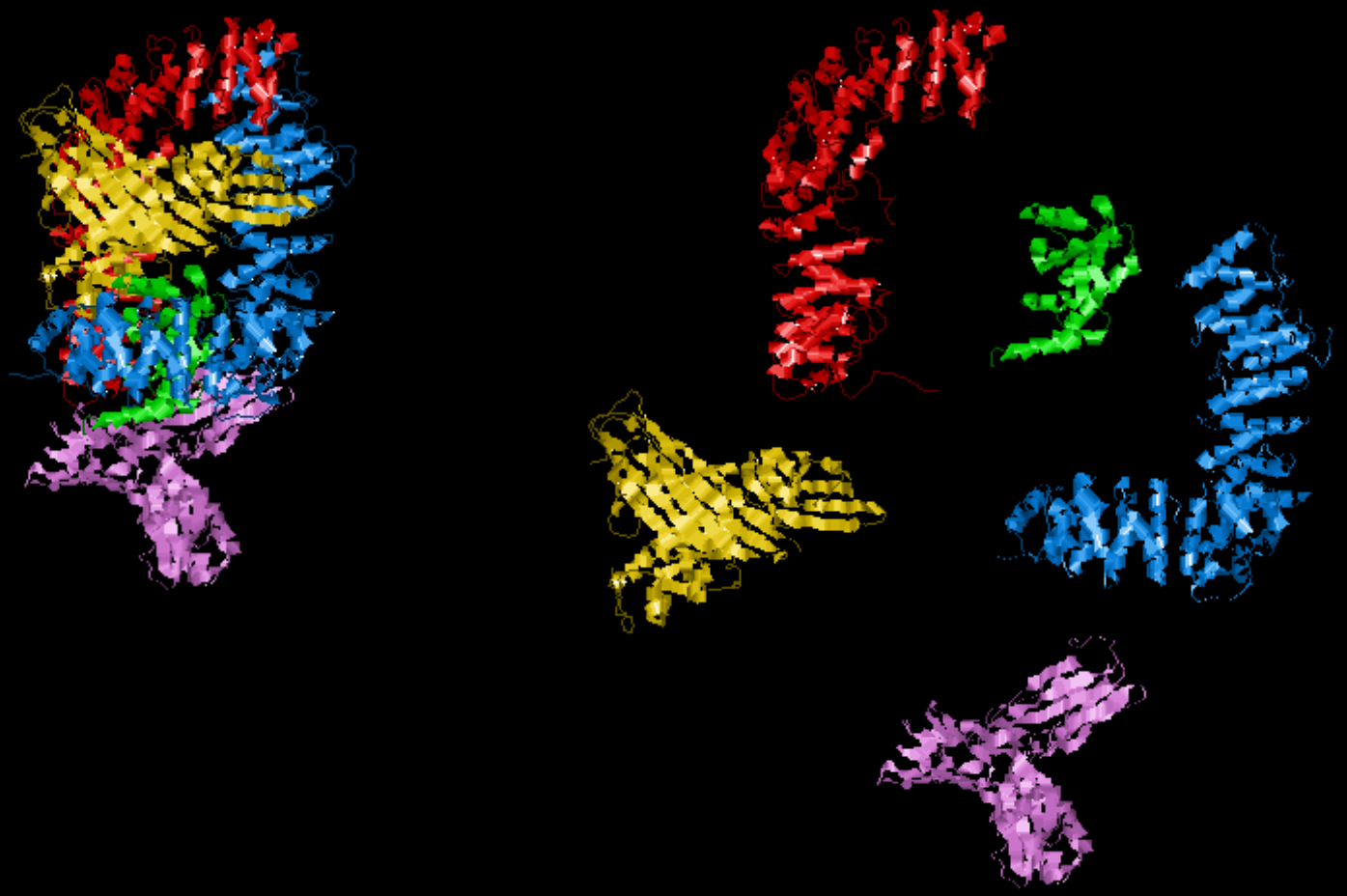
PDB:1b3u
Groves et al. (1999) *Cell*

Ap1 Clathrin Adaptor Core



PDB:1w63
Heldwein et al. (2004) *PNAS*

Ap1 Clathrin Adaptor Core

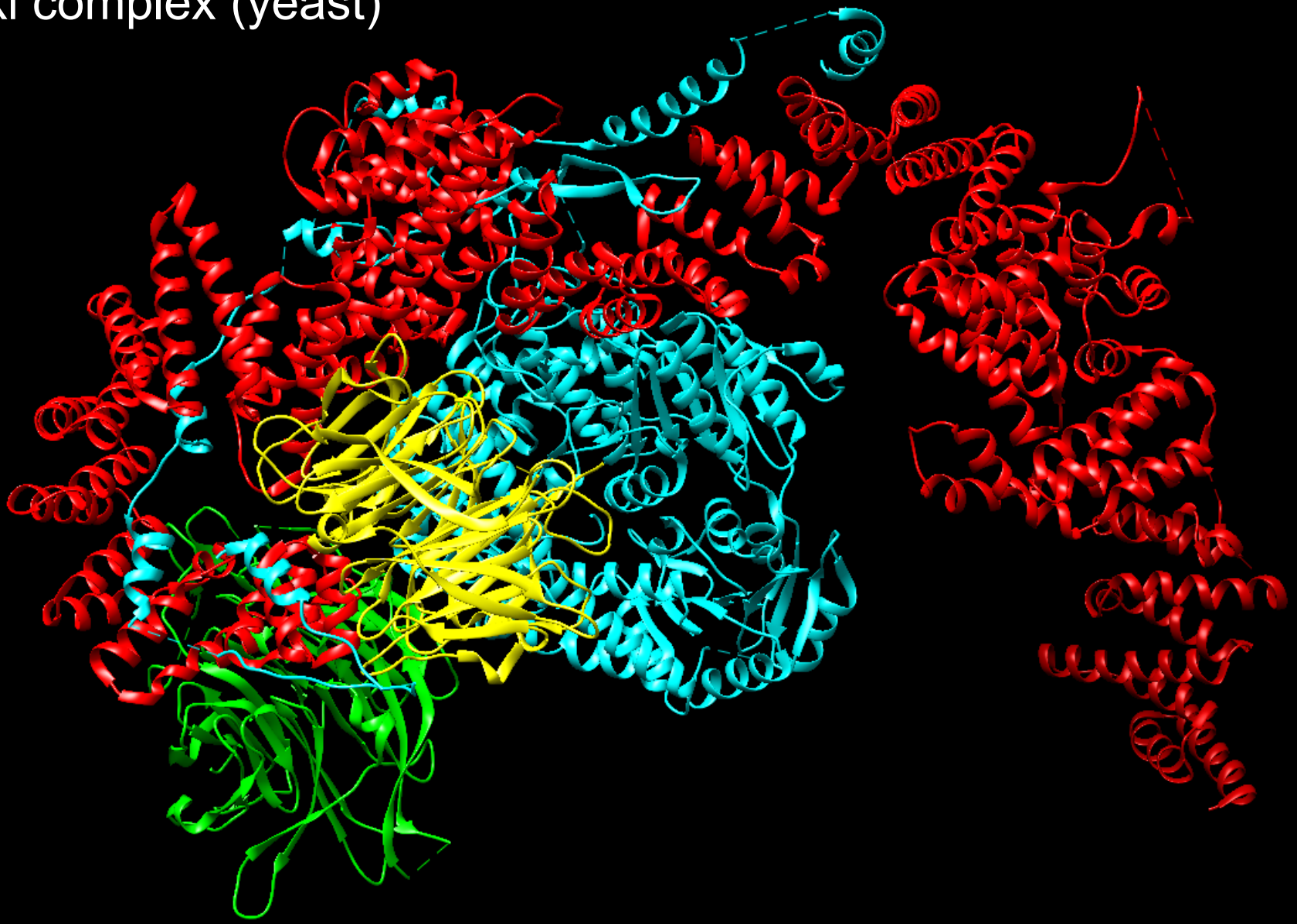


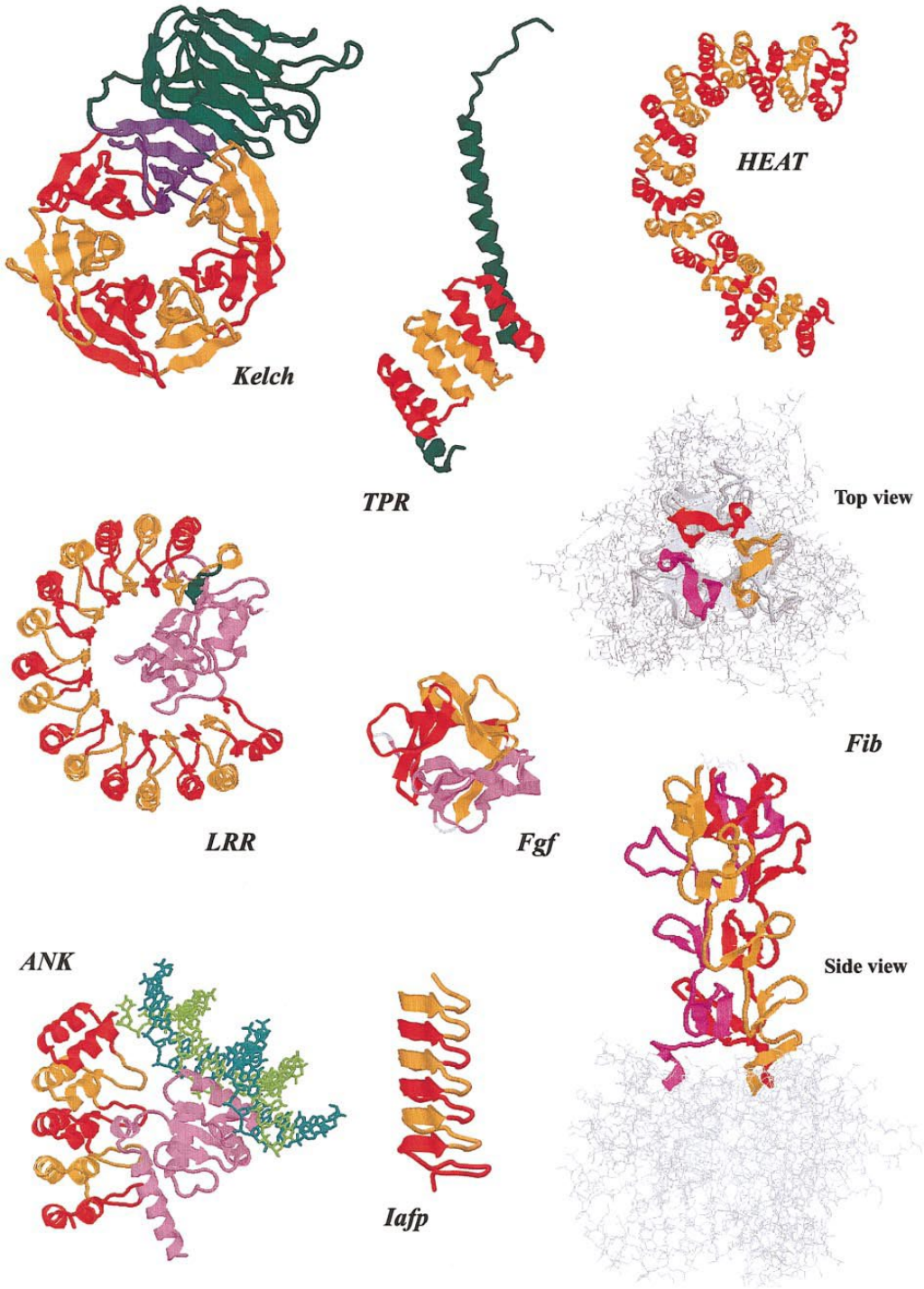
PDB:1w63
Heldwein et al. (2004) *PNAS*

i-TASSER model of *D. melanogaster* thr protein

Based on PDB 4BUJ chain B

PDB 4BUJ
Ski complex (yeast)





Andrade et al. (2001)
J Struct Biol

Definition CBRs

Perfect repeat: QQQQQQQQQQQ
Imperfect: QQQQPQQQQQQ
Amino acid type: DDDDDEEEDEDEED

Compositionally biased regions (CBRs)

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

Particular case of low complexity region

Detection CBRs

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

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL
```

Detection CBRs

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

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSF**QQQQQQQQQQQQQQQQQQQQ**PPPPPPPPPPPQLPQPPPQAQP
LLPQPQ**PPPPPPPPPP**GPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

Detection CBRs

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

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSF**QQQQQQQQQQQQQQQQQQQQ**PPPPPPPPPPQLPQPPPQAQP
LLPQPQ**PPPPPPPPPP**GPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHST**LLILGVLL**T**L**R**YLV**P**LL**QQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLG**EEE**AL**EDD**S**E**SRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQ**D**
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

Detection CBRs

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

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSF**QQQQQQQQQQQQQQQQQQQQQ**PPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

Detection repeats

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

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL
```

Detection repeats

Often NOT straightforward.
N-terminal human Huntingtin.
How many **repeats** can you find?

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSP**E**
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQK**CRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMAS**FG
NFA**NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS**RR**TQYFYSWLLNVLLGLLV**
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS**PSAEQLVQVYEL**
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

Detection repeats

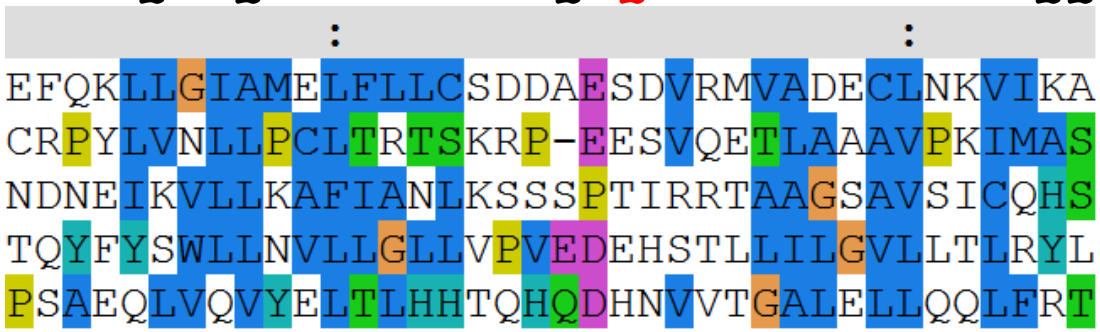
Often NOT straightforward.
N-terminal human Huntingtin.
How many **repeats** can you find?

```
EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKRP-EESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRT
```

Detection repeats

Often NOT straightforward.
N-terminal human Huntingtin.
How many **repeats** can you find?

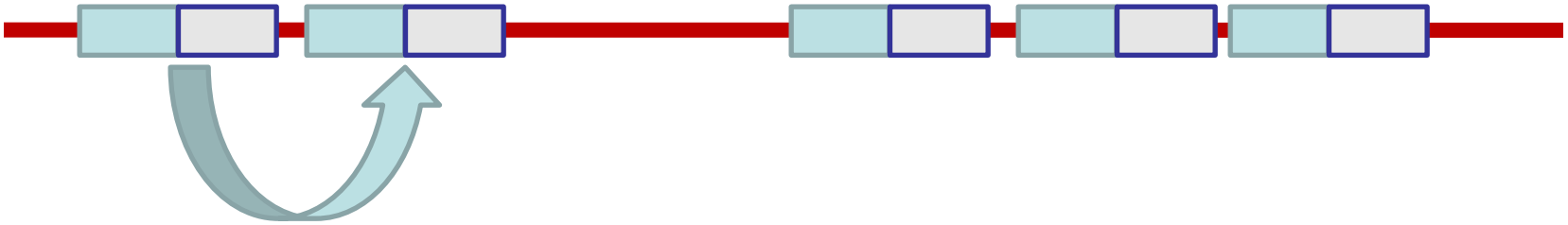
EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKRP-EEESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRT



Detection of repeats

Dotplots

Comparing a sequence against itself



Detection of repeats

Dotplots

Detection of repeats

Dotplots

1 match

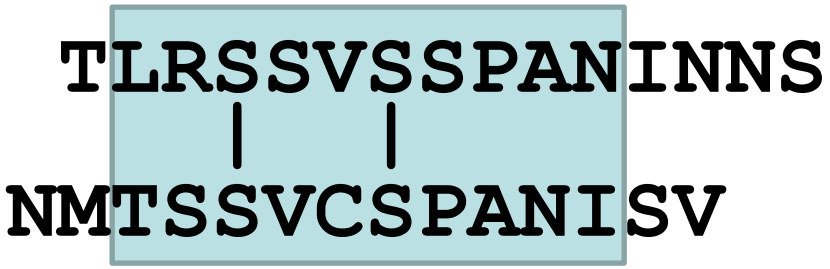
Detection of repeats

Dotplots

8 matches

Detection of repeats

Dotplots



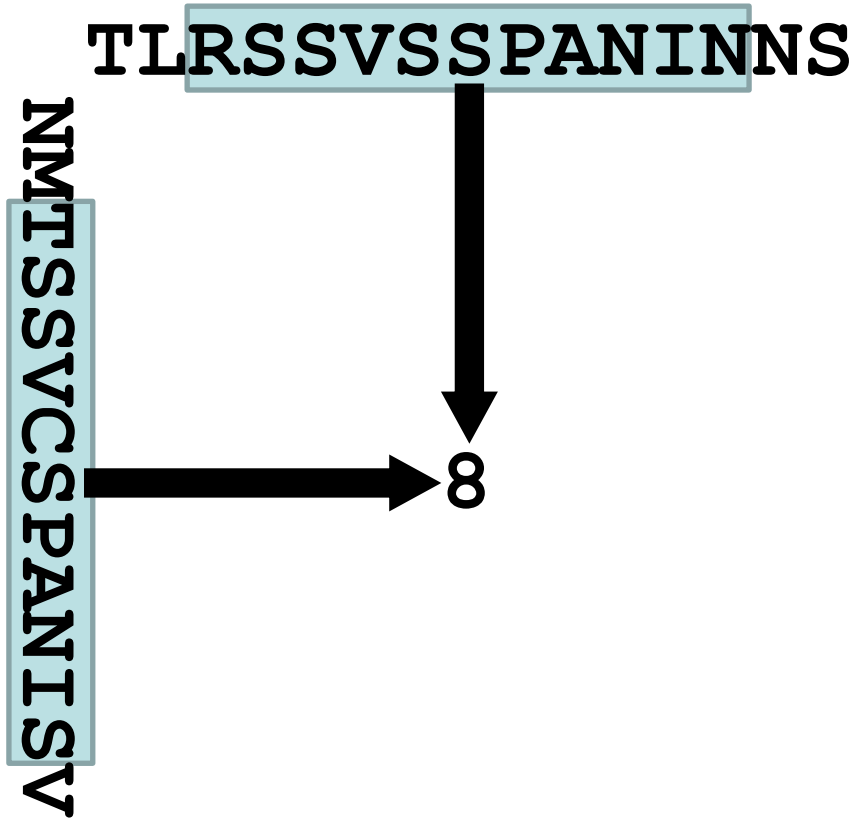
2 matches

Detection of repeats

Dotplots

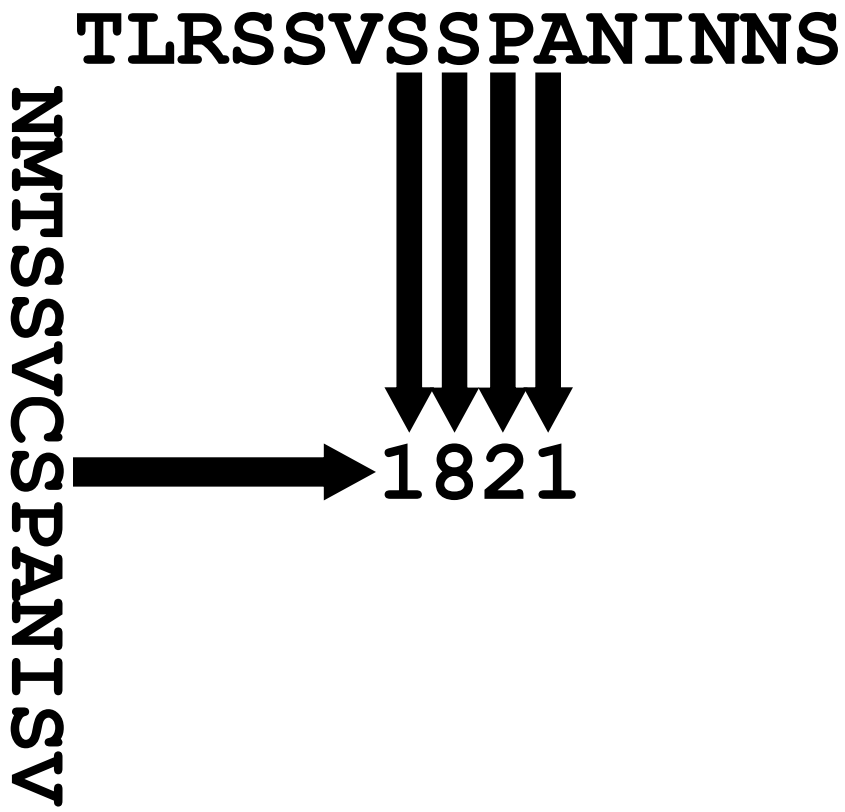
Detection of repeats

Dotplots

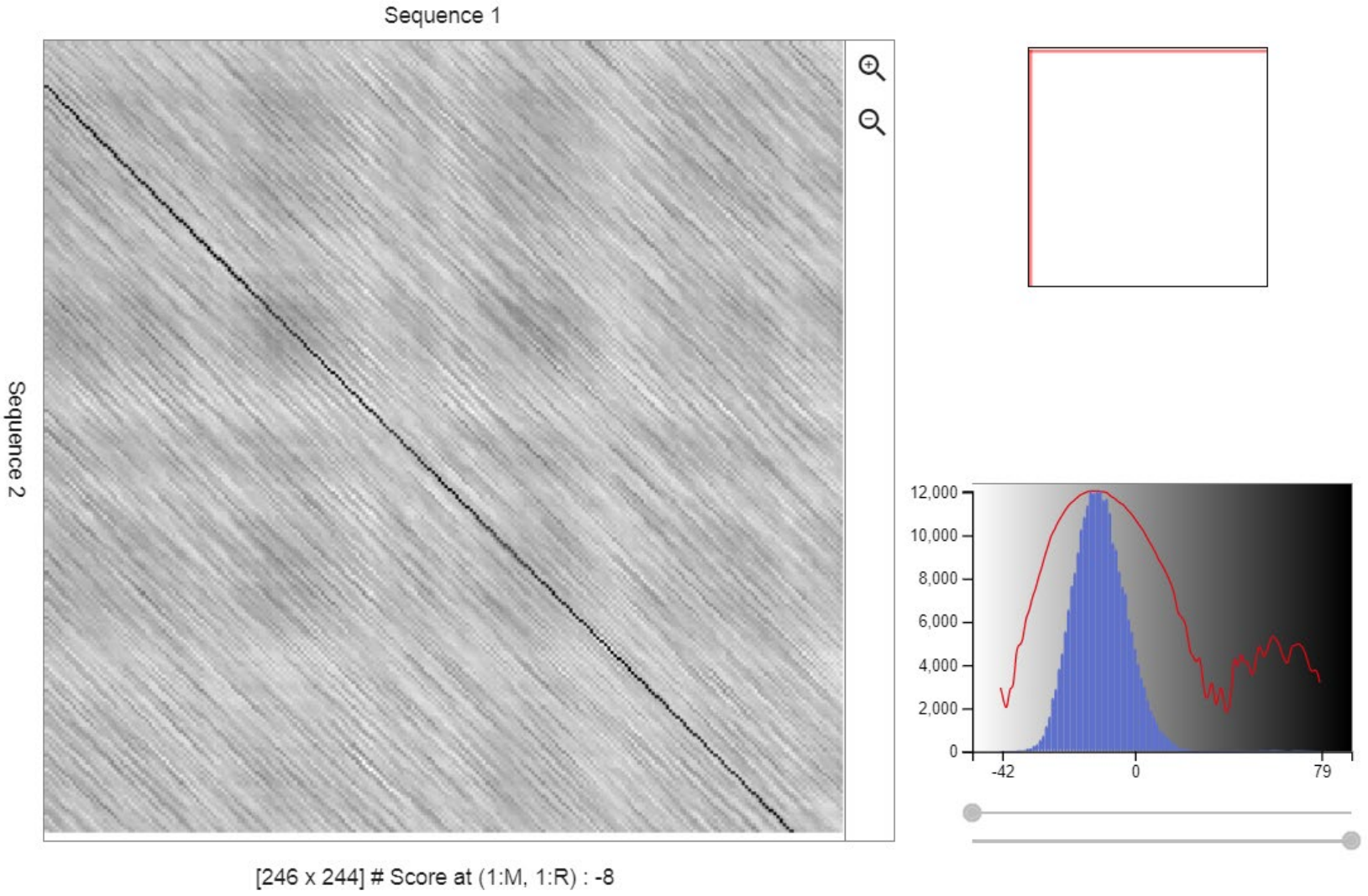


Detection of repeats

Dotplots



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



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

•Go to the Dotlet web page:
http://dotlet.vital-it.ch

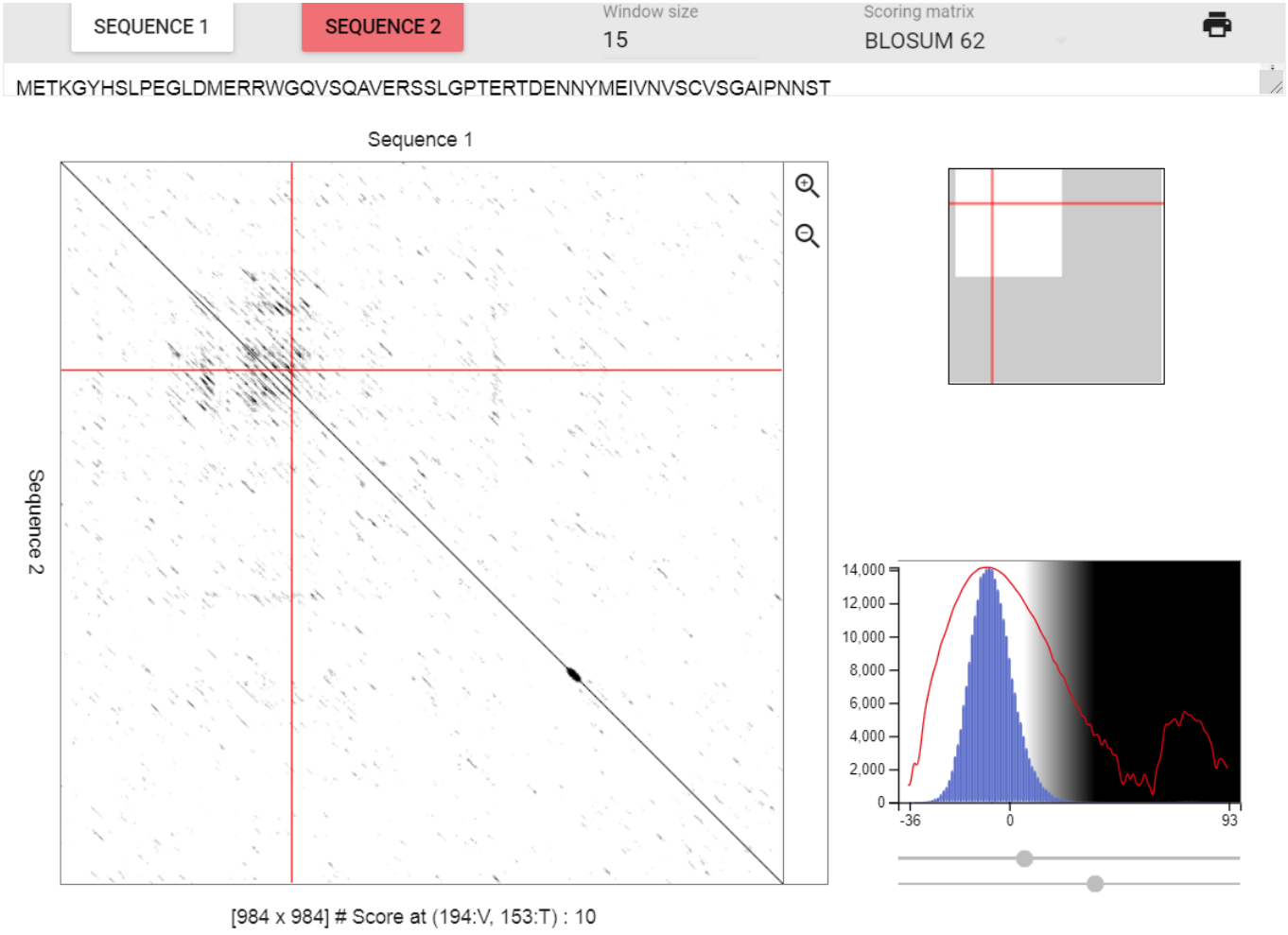
•Click on the input button and paste the sequence of the human mineralocorticoid receptor (UniProt id P08235)

•Click on the "compute" button

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

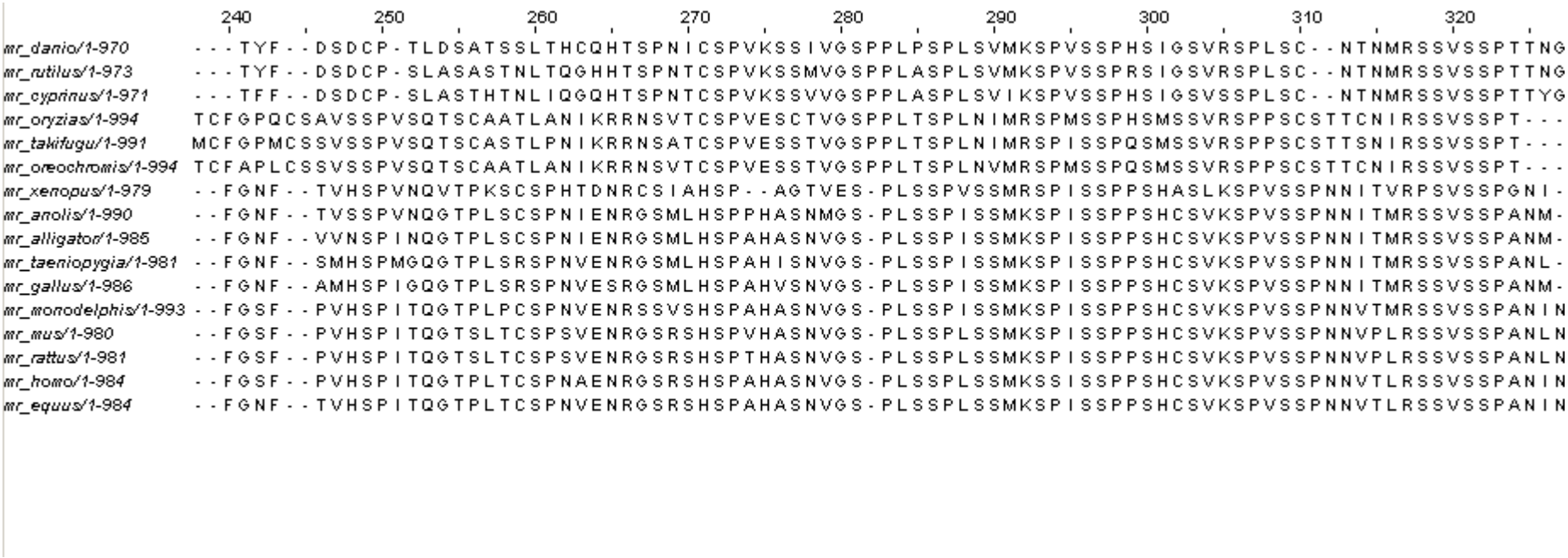
•Find repetitions clicking in the diagonal patterns

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



Detection of repeats

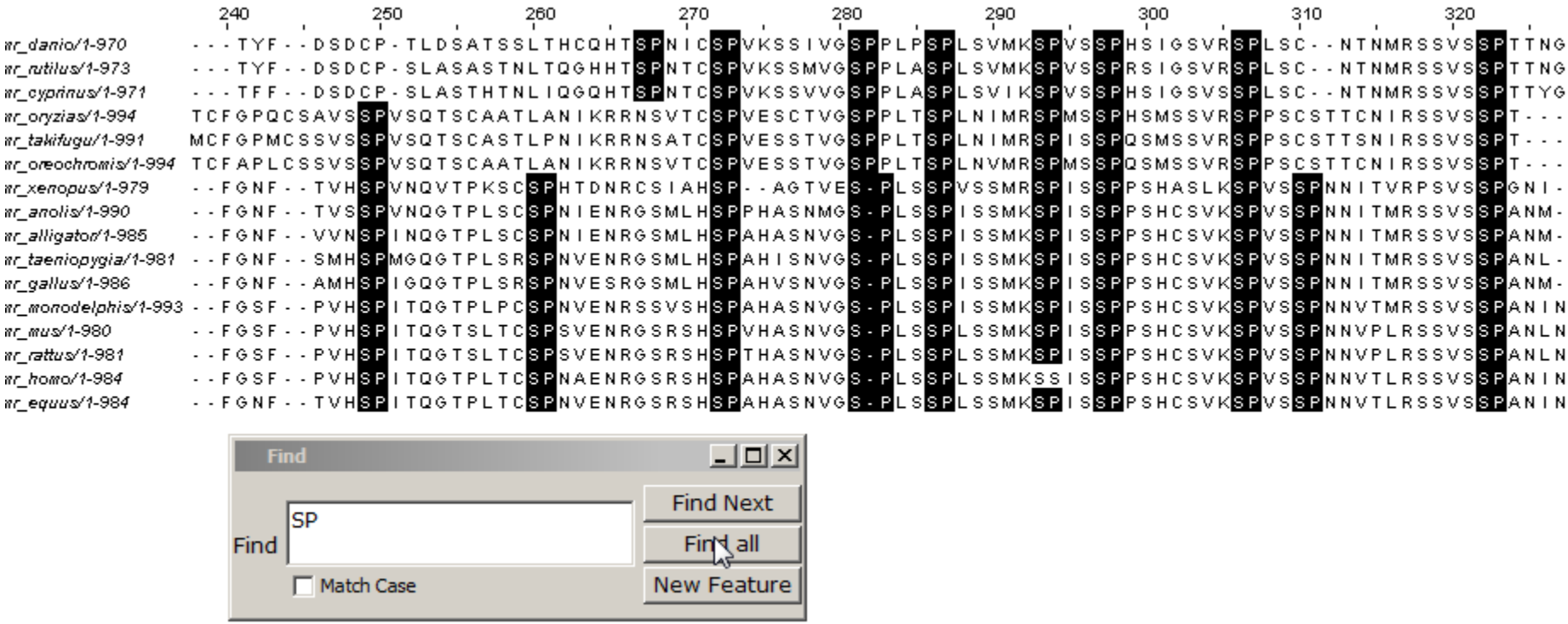
Using a multiple sequence alignment helps.
Conserved repeated patterns



JalView with Regular Expression searches

Detection of repeats

Using a multiple sequence alignment helps Conserved repeated patterns



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

JalView 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

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`**

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!**)

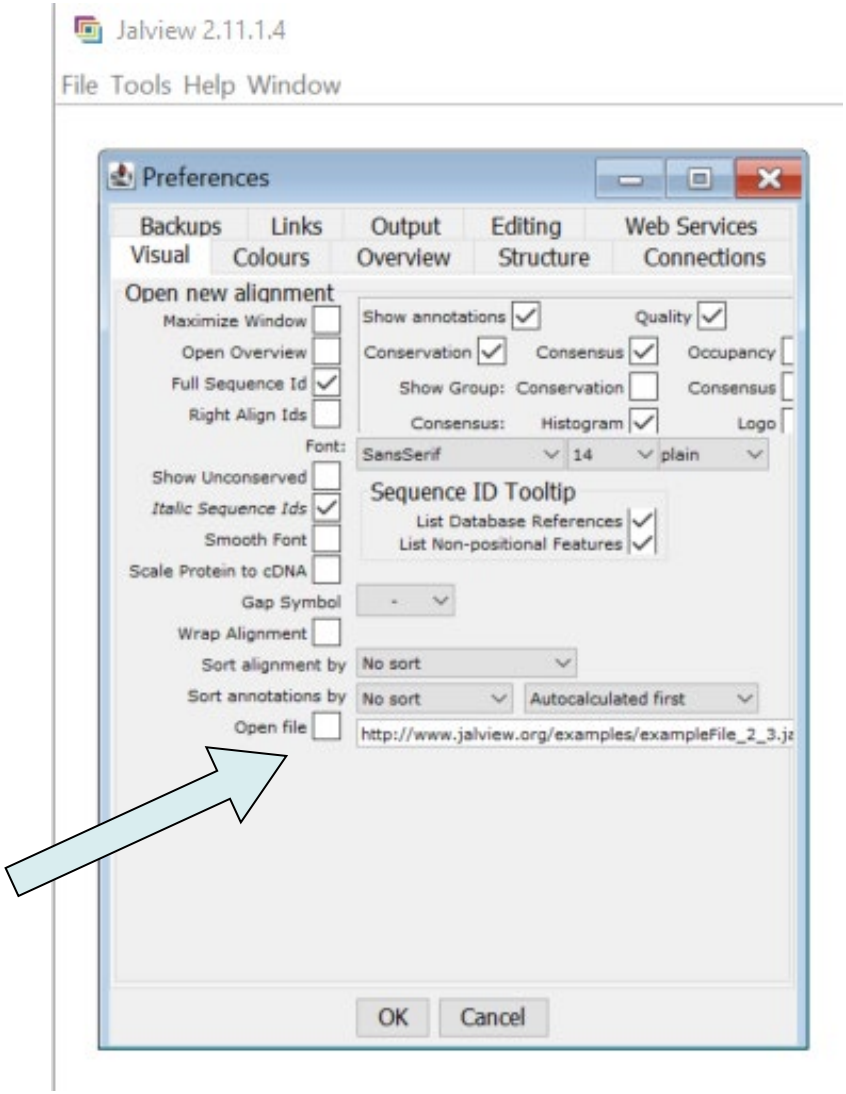
- Try to find the expression that matches more repeats. How many repeats do you see? How long are they? Would you correct the alignment based on these findings?

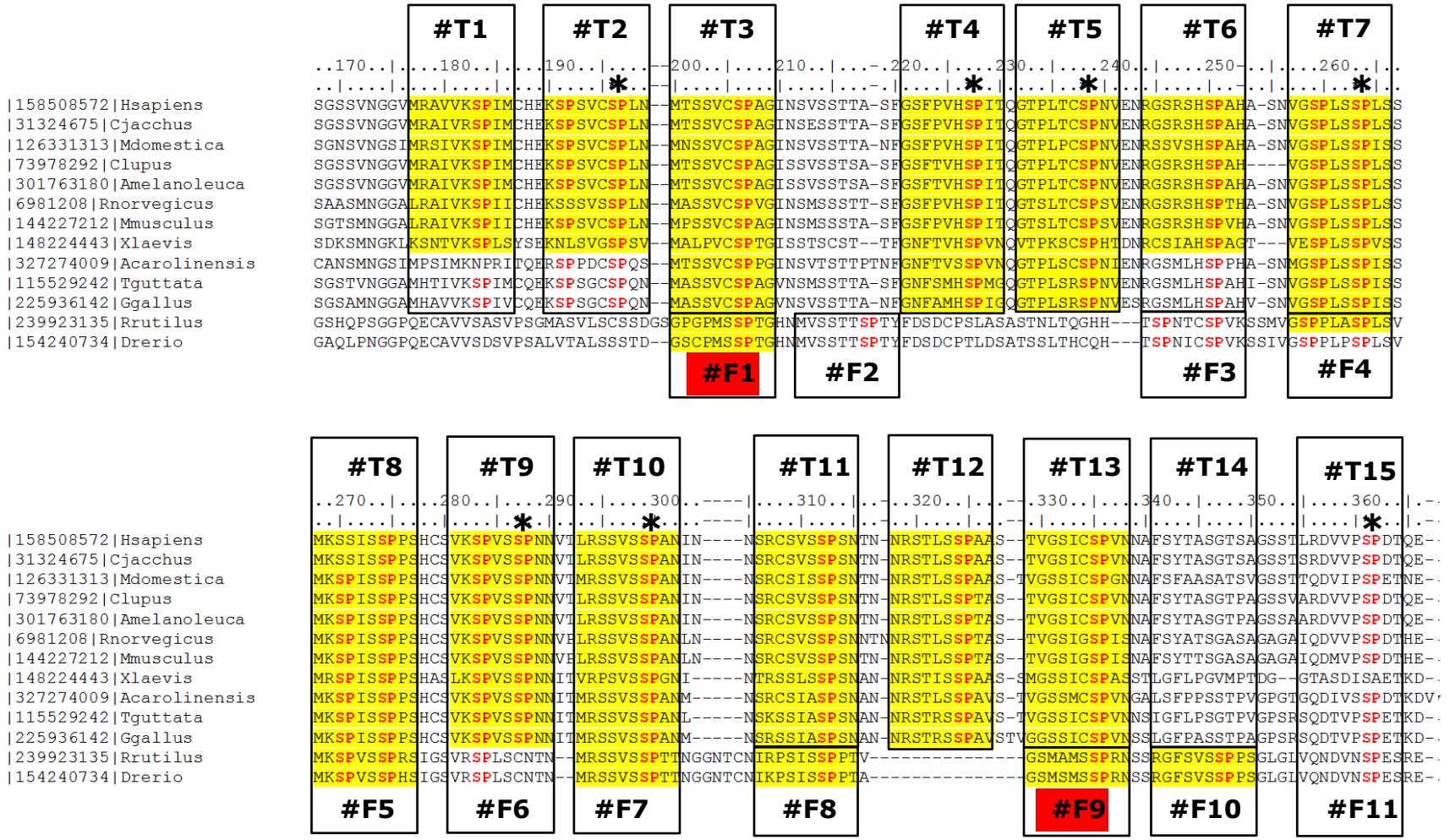
First time running JalView?

Remove annoying start:

Go to
Tools > Preferences > Visual

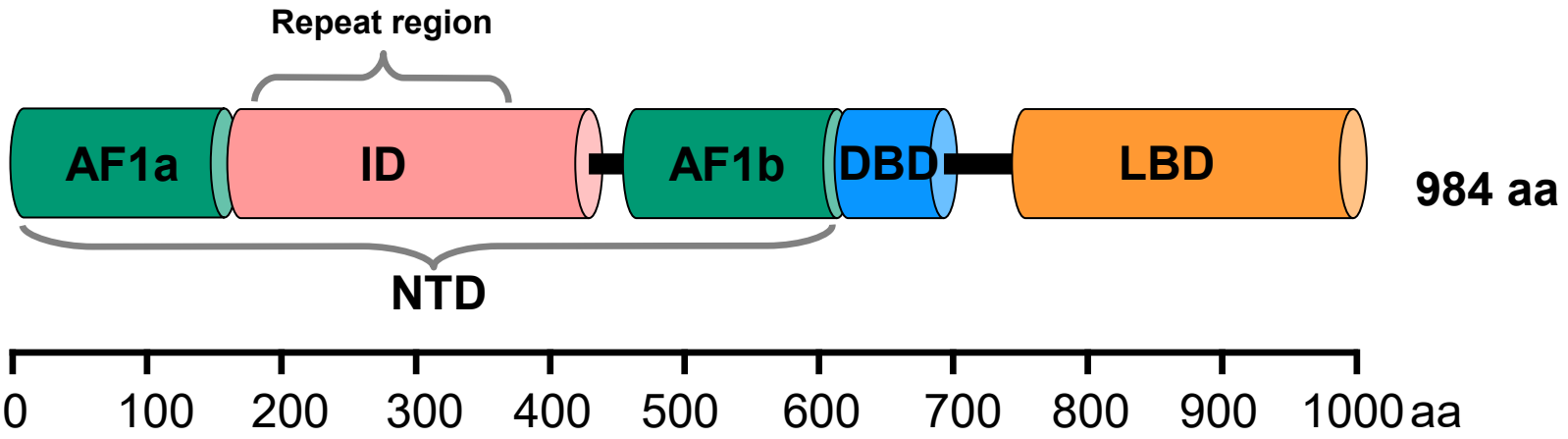
Un-tick option “Open file”



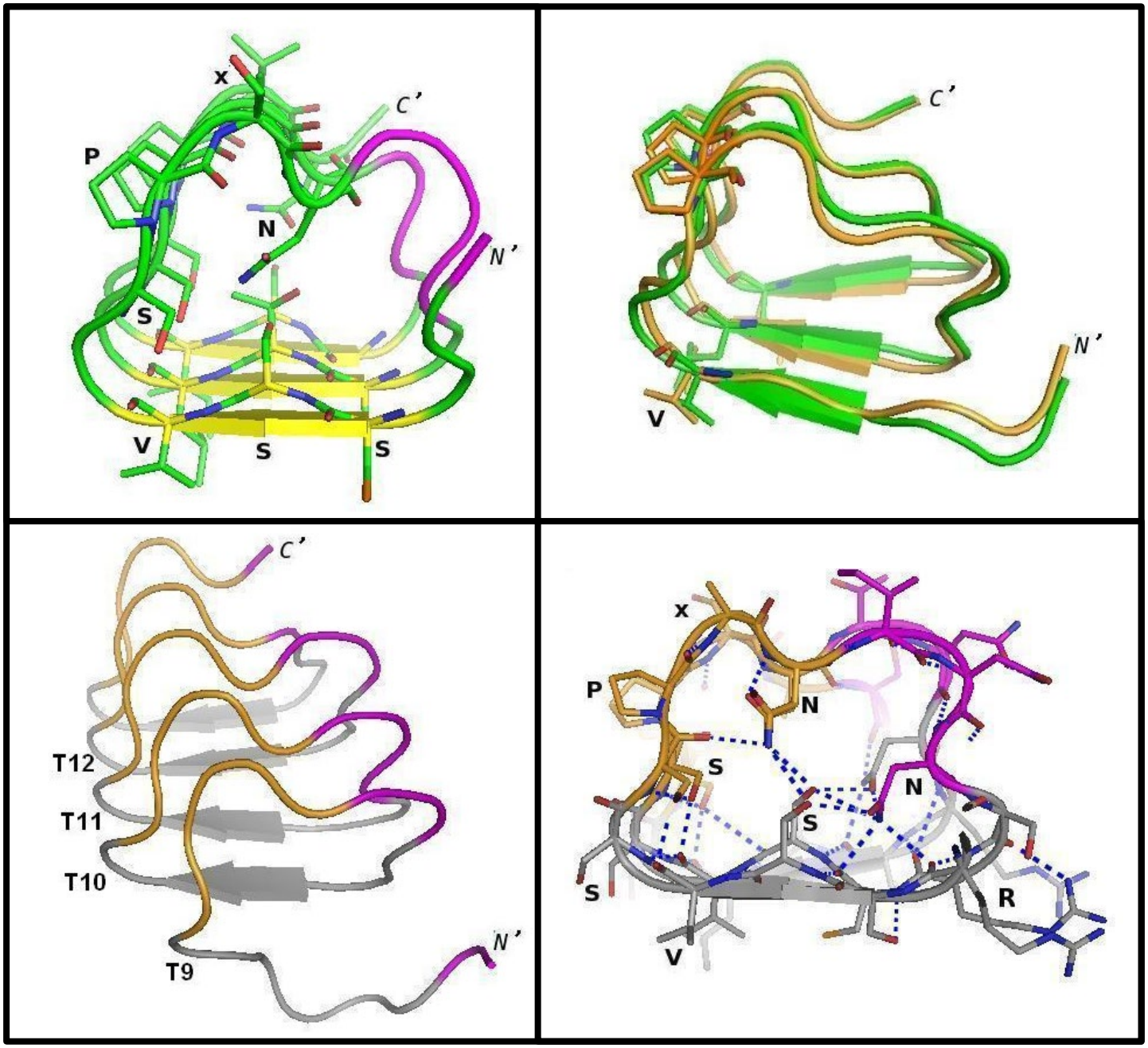


Vlassi *et al.* (2013) *BMC Struct. Biol.*

Mineralocorticoid receptor



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


x
C'
P
N
N'
S
V
S
S
C'
N'
V
C'
T12
T11
T10
T9
N'
x
P
N
S
N
S
S
R
V

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: QQQQQQQQQQQ
Imperfect: QQQQPQQQQQQ
Amino acid type: DDDDDEEEDEDEED

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

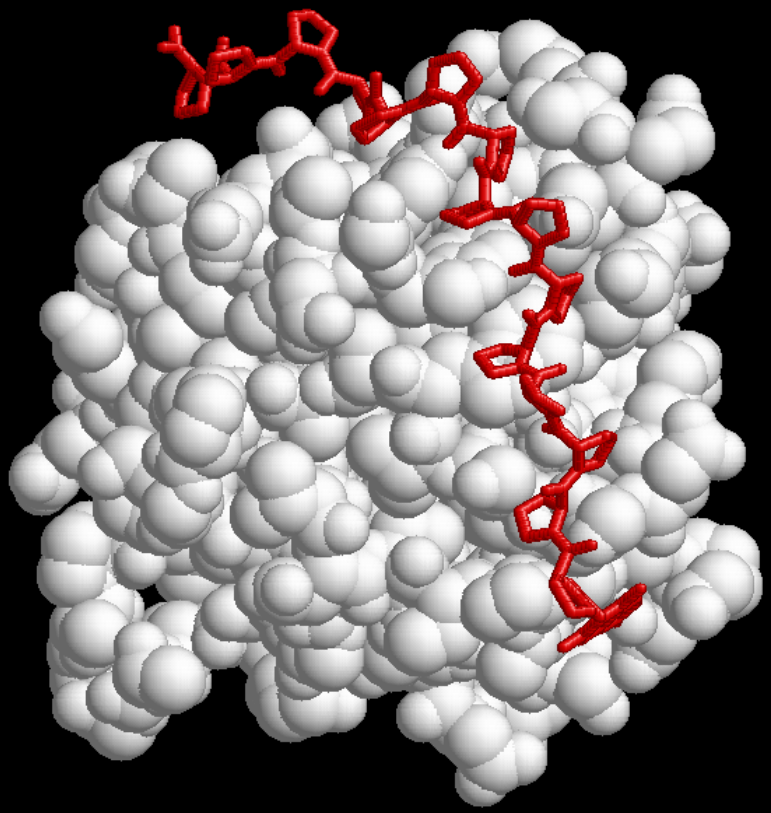
Functions:
Passive: linkers
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

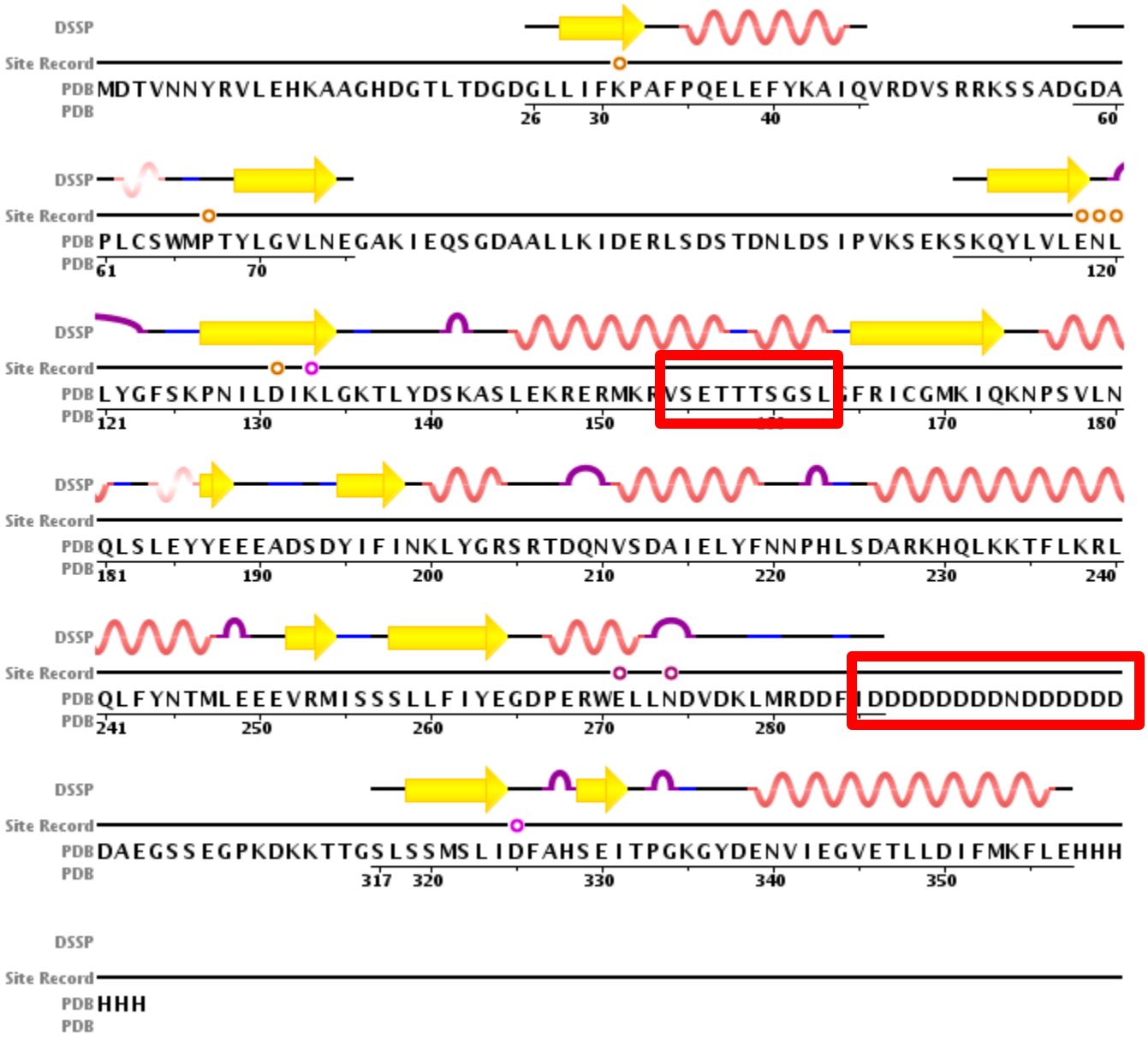
Structure of CBRs

Often variable or flexible: do not easily crystalize

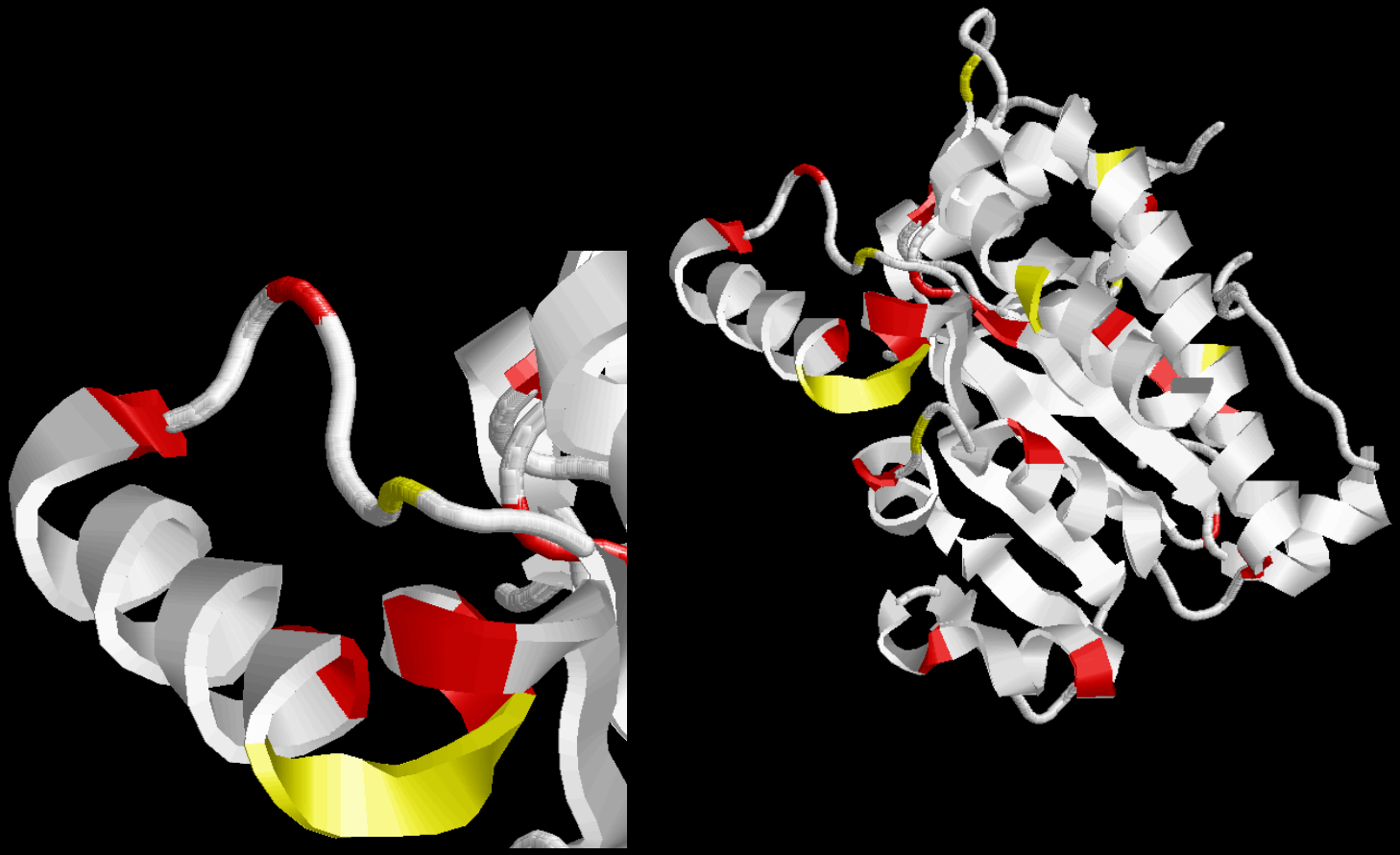
1CJF: profilin bound to polyP



2IF8: Inositol Phosphate Multikinase Ipk2

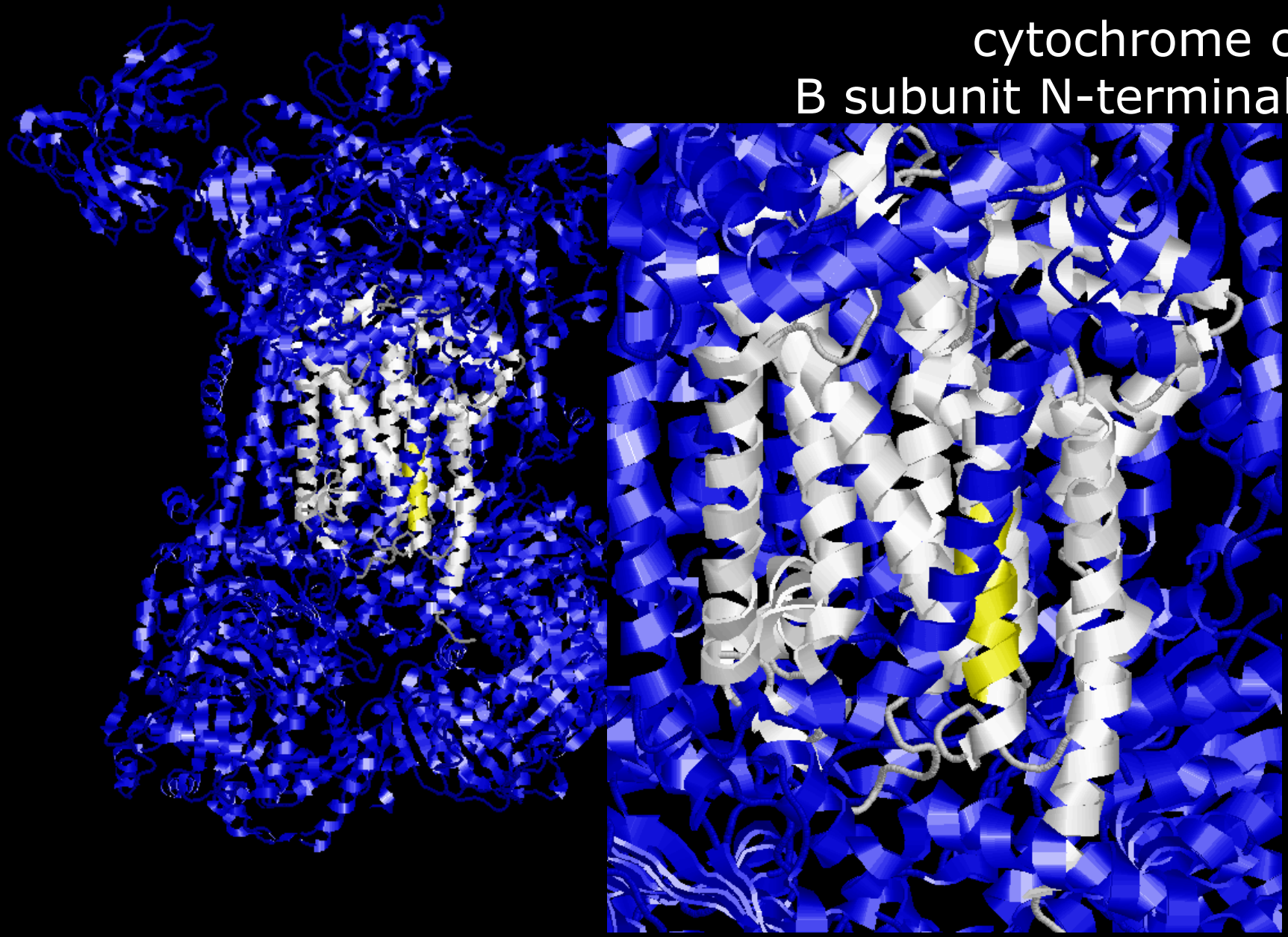


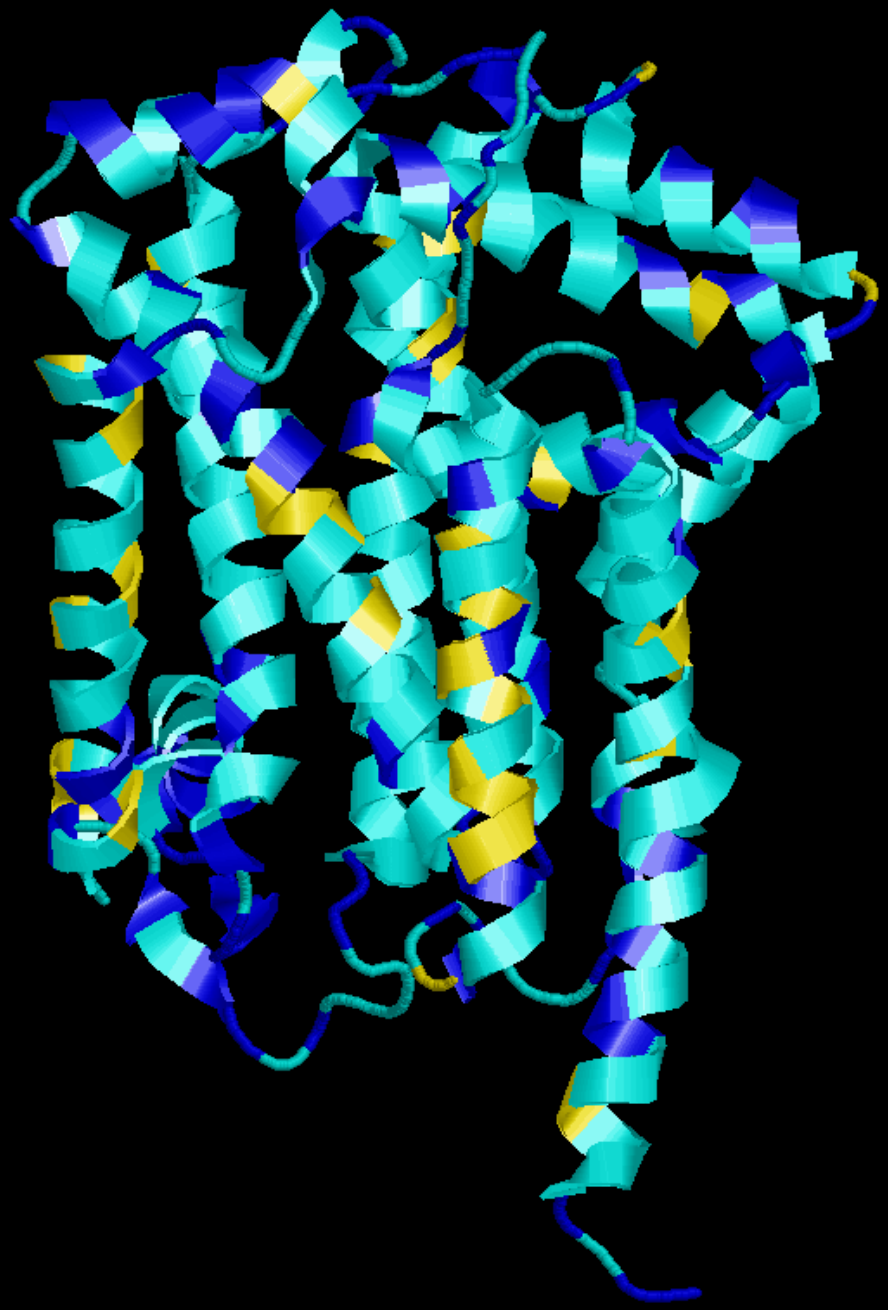
2IF8: Inositol Phosphate Multikinase Ipk2



RVSETTTSGSL

2CX5: mitochondrial
cytochrome c
B subunit N-terminal





2CX5: mitochondrial cytochrome c B subunit N-terminal

Amino acid repeats

Distribution is not random:

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

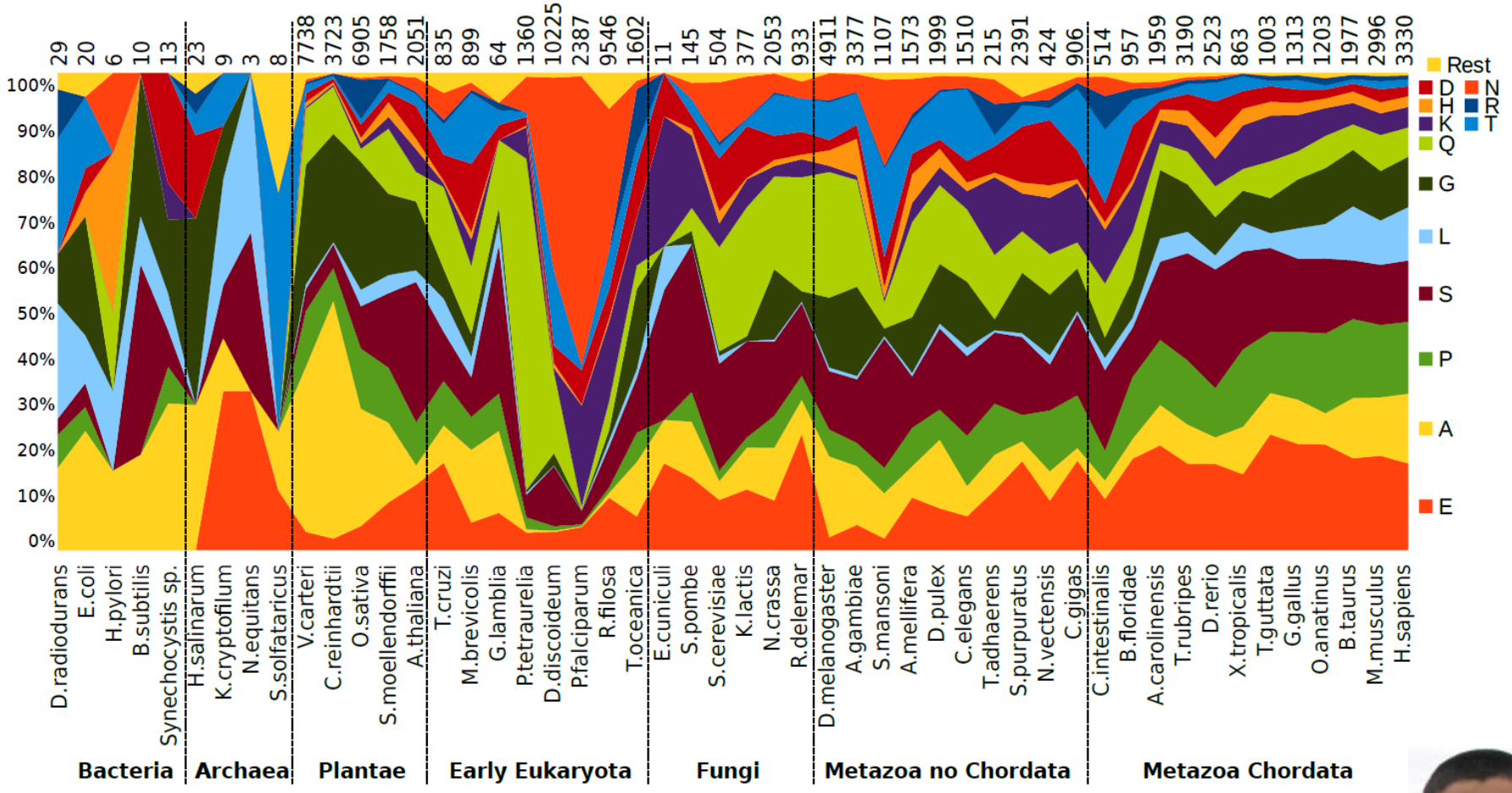
Prokaryota:
Most common: poly-S, poly-G, poly-A, poly-P
Relatively rare: poly-Q, poly-N

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



Mier *et al.* (2017) Proteins

Pablo Mier

Filtering out CBRs

Normally filtered out as low complexity region: they give spurious BLAST hits

```
QQQQQQQQQQ
||||||||||
QQQQQQQQQQ   10/10 id

IDENTITIES
||||||||||
IDENTITIES   10/10 id
```

Filtering out CBRs

Normally filtered out as low complexity region: they give spurious BLAST hits

```
QQQQQQQQQQ
||||||||||
QQQQQQQQQQ
```
Shuffle: 10/10 id

```
IDENTITIES
||||||||||
IDENTITIES
```
10/10 id

Filtering out CBRs

Normally filtered out as low complexity region: they give spurious BLAST hits

```
QQQQQQQQQQ
||||||||||
QQQQQQQQQQ Shuffle: 10/10 id

IDENTITIES
   | |
SIINDIETTE Shuffle: 2/10 id
```

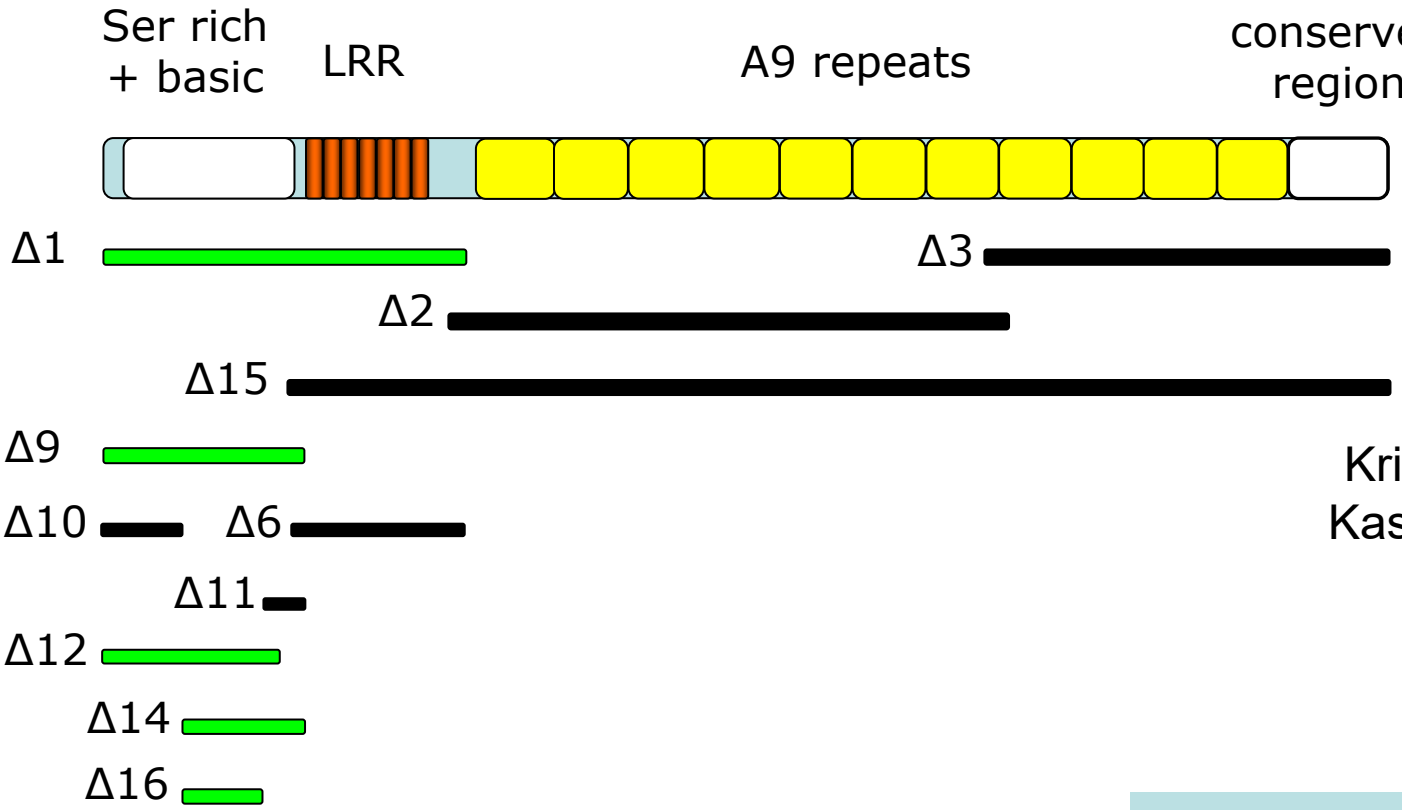
Filtering out CBRs

Option for pre-BLAST treatment
SEG algorithm:
1) Identify sequence regions with low information content over a sequence window
2) Merge neighbouring regions

Eliminates hits against common acidic-, basic- or proline-rich regions

(Wootton and Federhen, 1993)

AIR9 (1708 aa)



Microtubule localization of Δx-GFP

Kristina
Kastano

Buschmann, et al (2006). *Current Biology.*
Buschmann, et al (2007). *Plant Signaling & Behavior*

Homorepeats are frequent but difficult to characterize

Pablo Mier

e.g. polyQ:

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQP

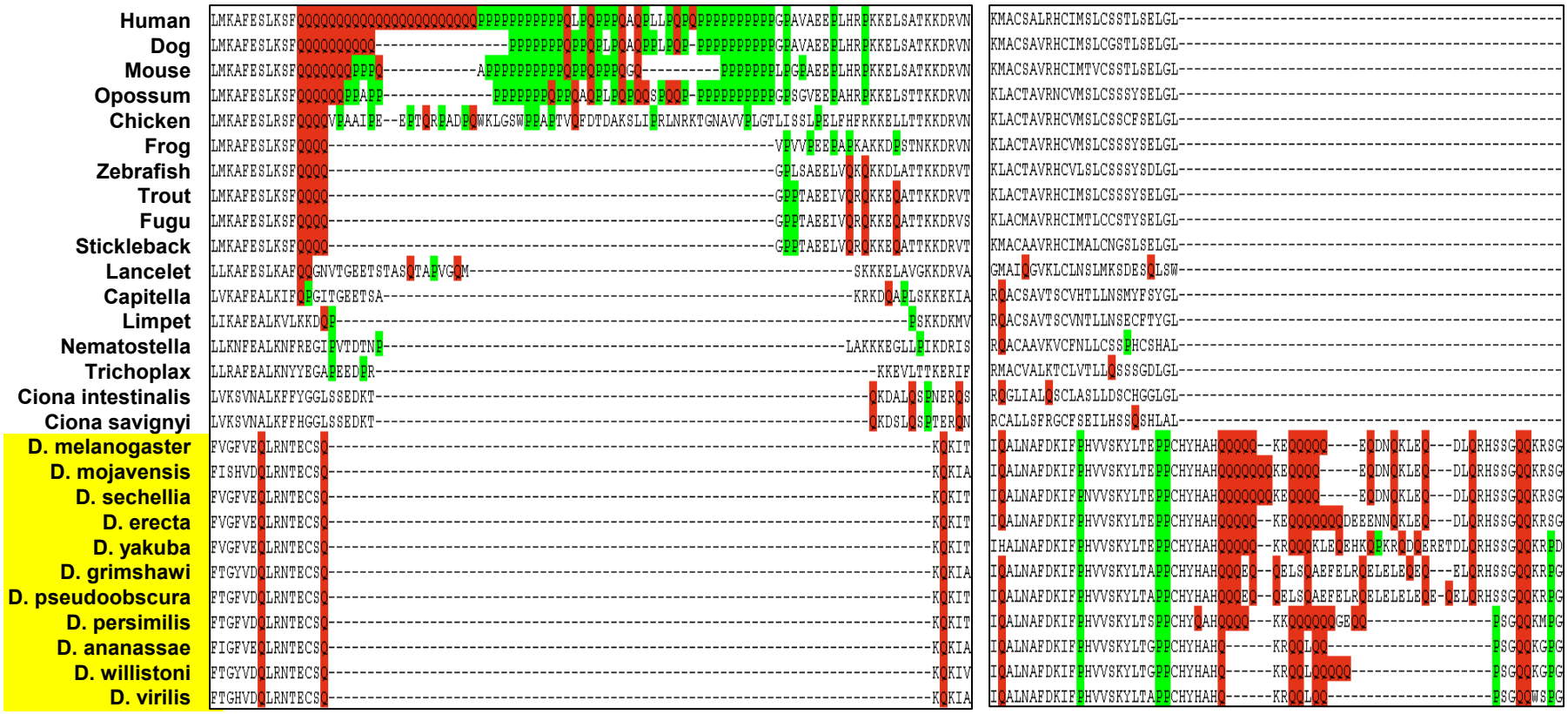
- 10% of human proteins have homorepeats
- lack sequence conservation
- not possible to predict function by homology

Homorepeats need to be studied in context

Function of polyQ

Martin Schaefer

polyQ in Huntingtin



Schaefer et al (2012) *Nucleic Acids Res.*

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?