JOHANNES GUTENBERG UNIVERSITÄT MAINZ

# Repeats and homorepeats

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

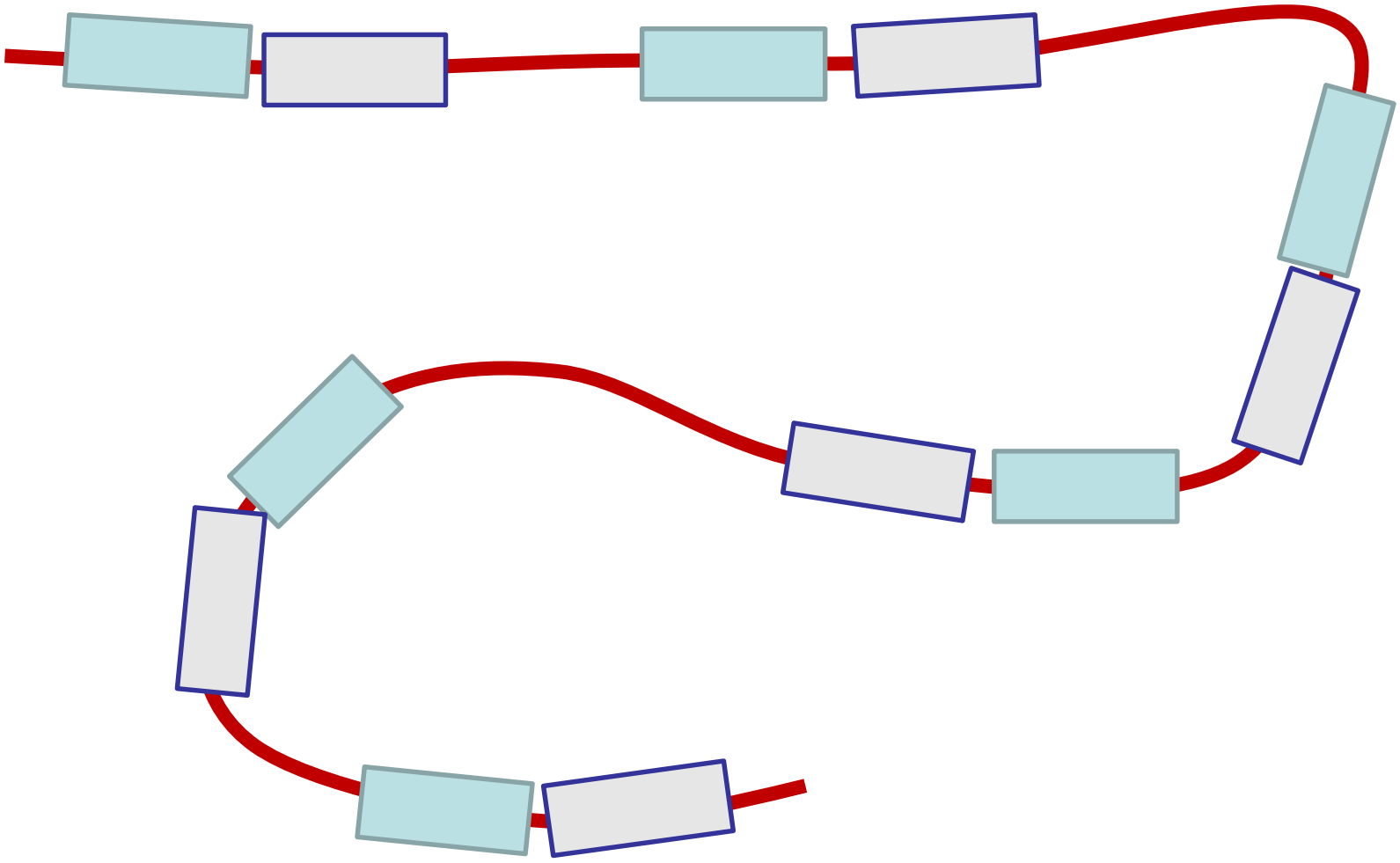
MRAVVK**SP**IM CHE
KSPSVC**SP**LN
MTSSVC**SP**AG INSVSSTTASF
GSFPVH**SP**IT Q
GTPLTC**SP**NV EN
RGSRSH**SP**AH ASN
VGSPLS**SP**LS S
MKSSIS**SP**PS HCS
VKSPVS**SP**NN VT
LRSSVS**SP**AN INN

# Definition repeats

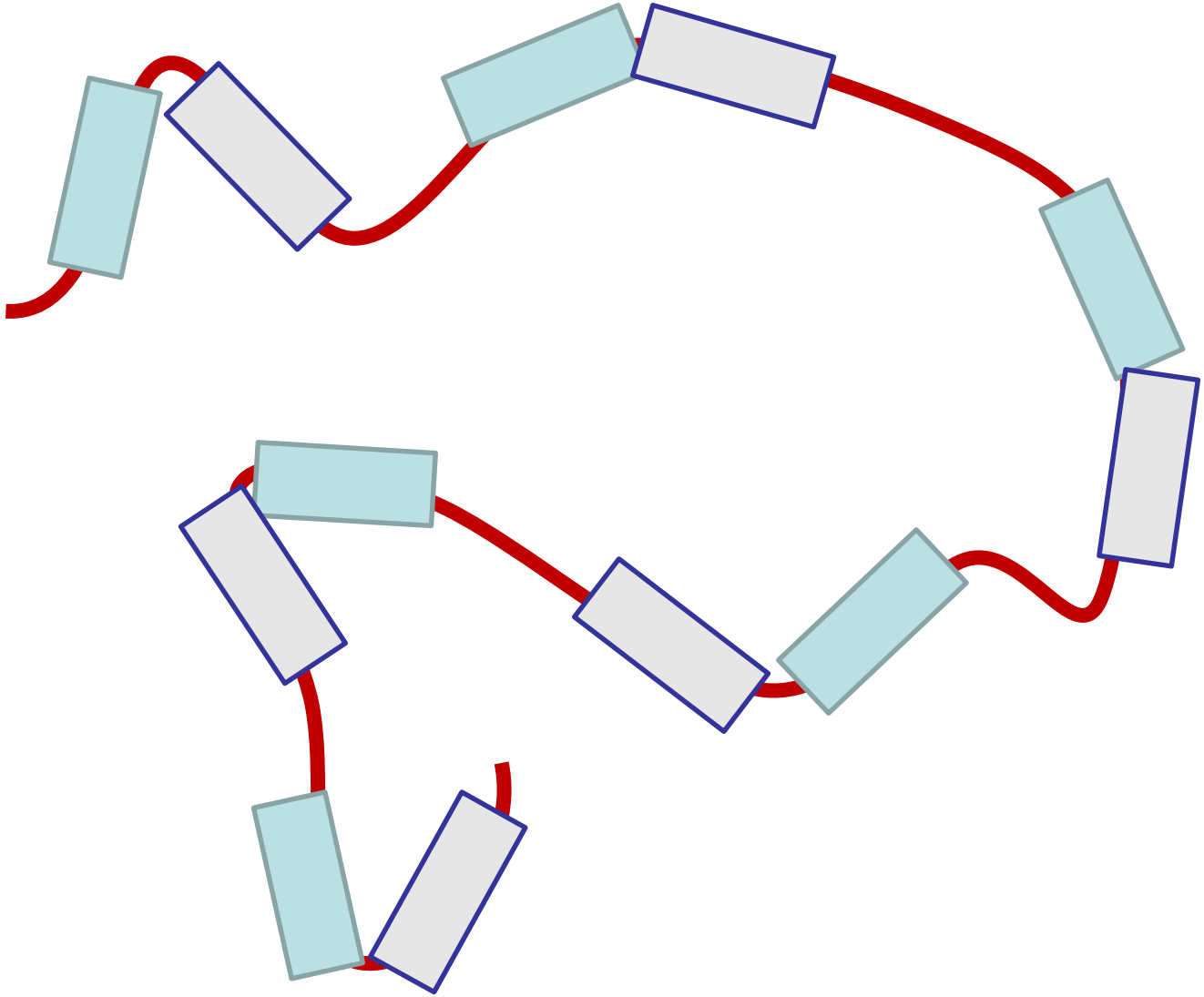
Sequence, long, imperfect, tandem

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**SSP**LS S
MK**S**SI**SSP**PS HCS
VK**S**P**VSSP**NN VT
LR**S**S**VSSP**AN INN

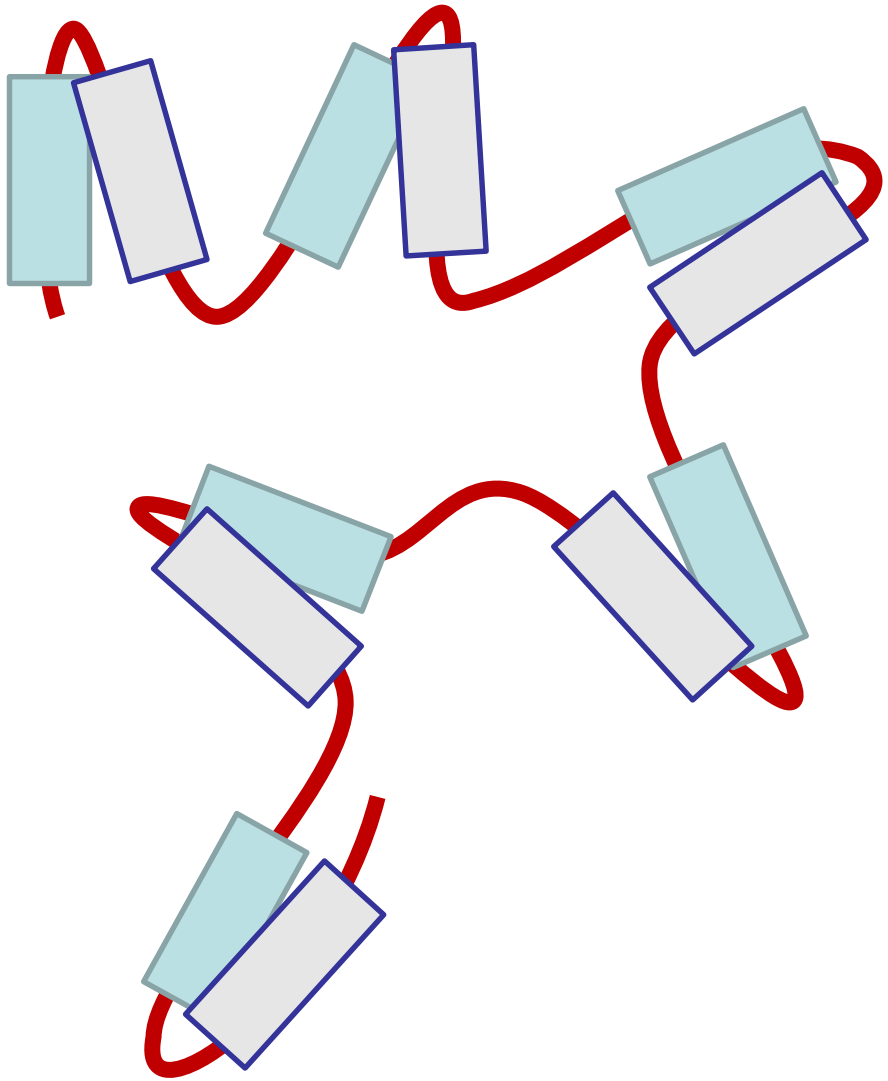
# Tandem repeats fold together



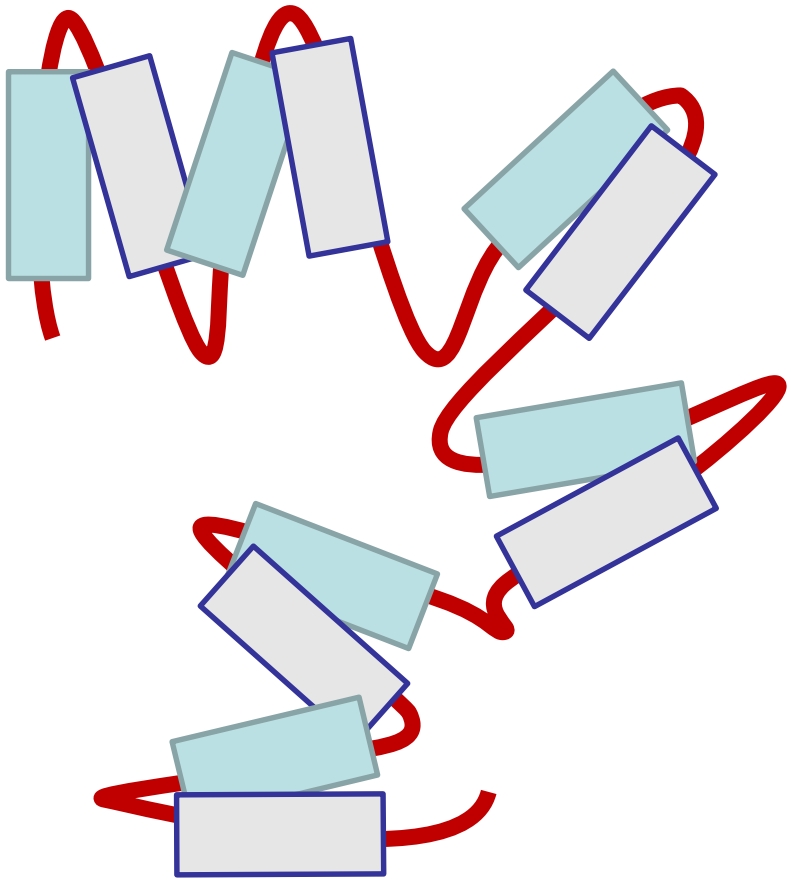
# Tandem repeats fold together



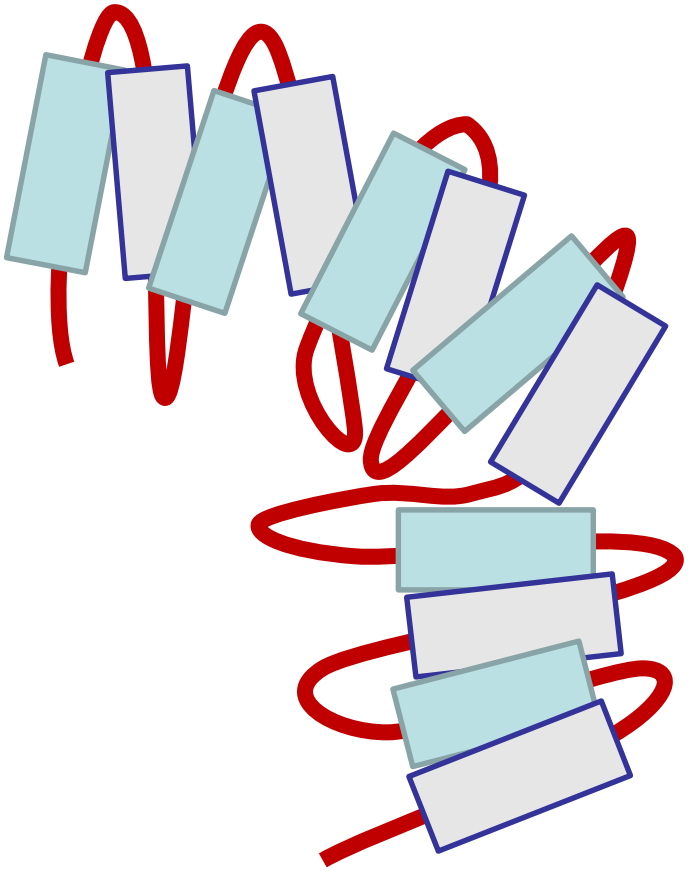
# Tandem repeats fold together



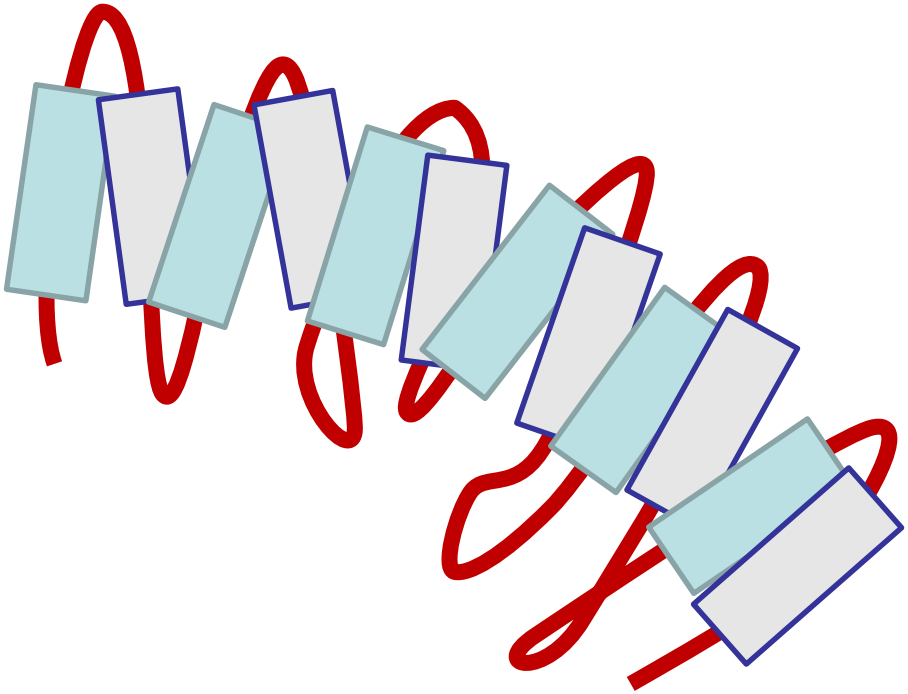
# Tandem repeats fold together



# Tandem repeats fold together



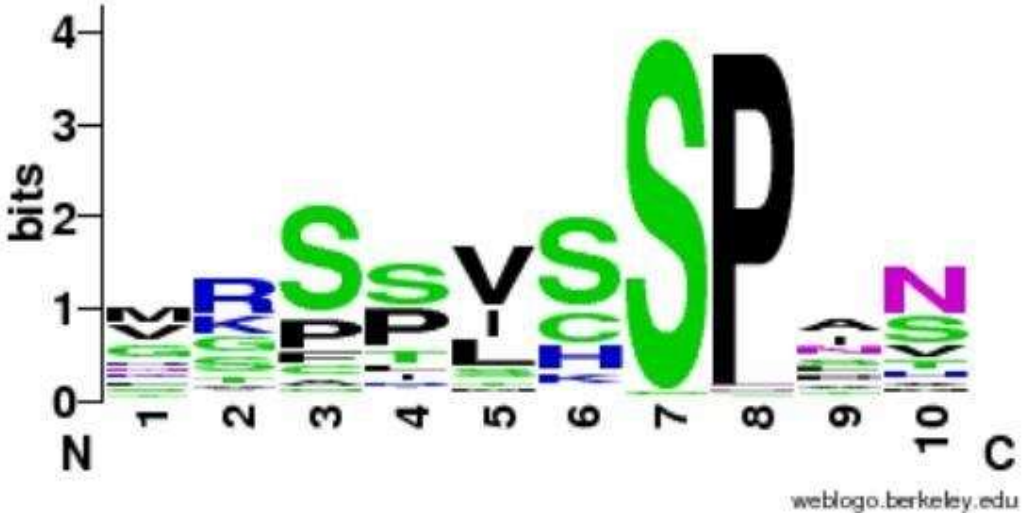
# Tandem repeats fold together



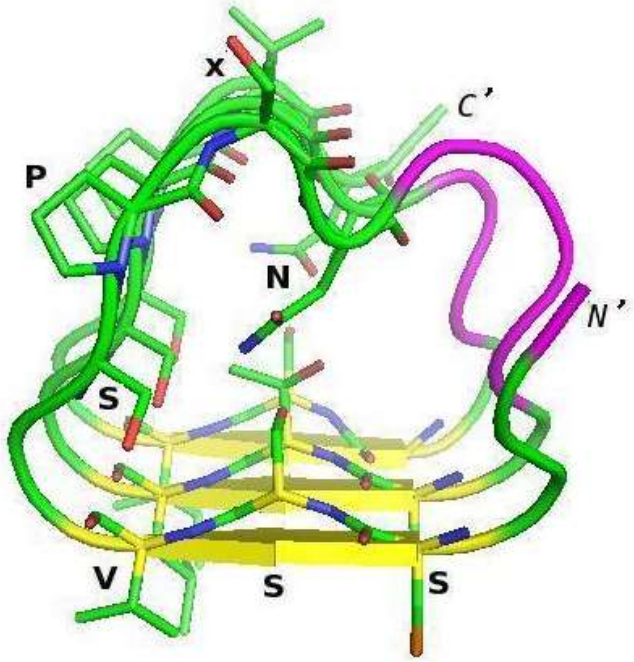
# Definition repeats

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**SSP**LS S
MK**S**SI**SSP**PS HCS
VK**S**P**VSSP**NN VT
LR**S**S**VSSP**AN INN




http://weblogo.berkeley.edu

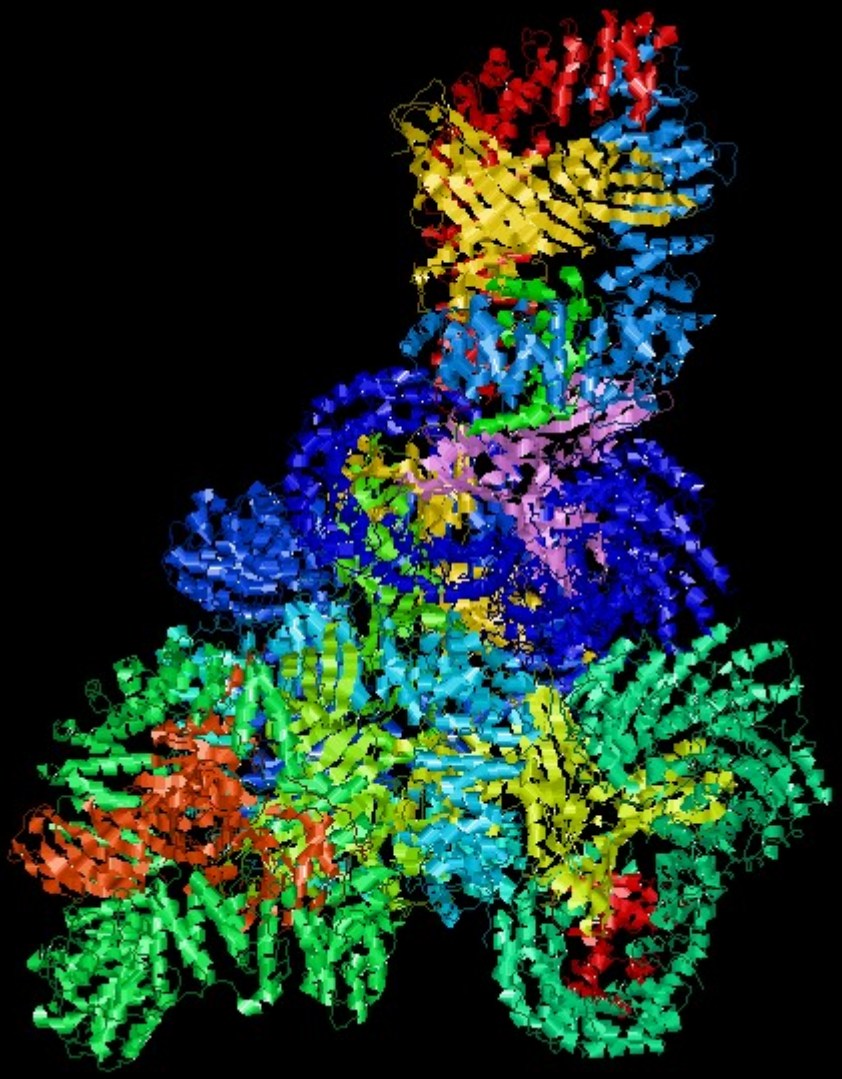



(Vlassi et al, 2013)

# A subunit PP2A structure

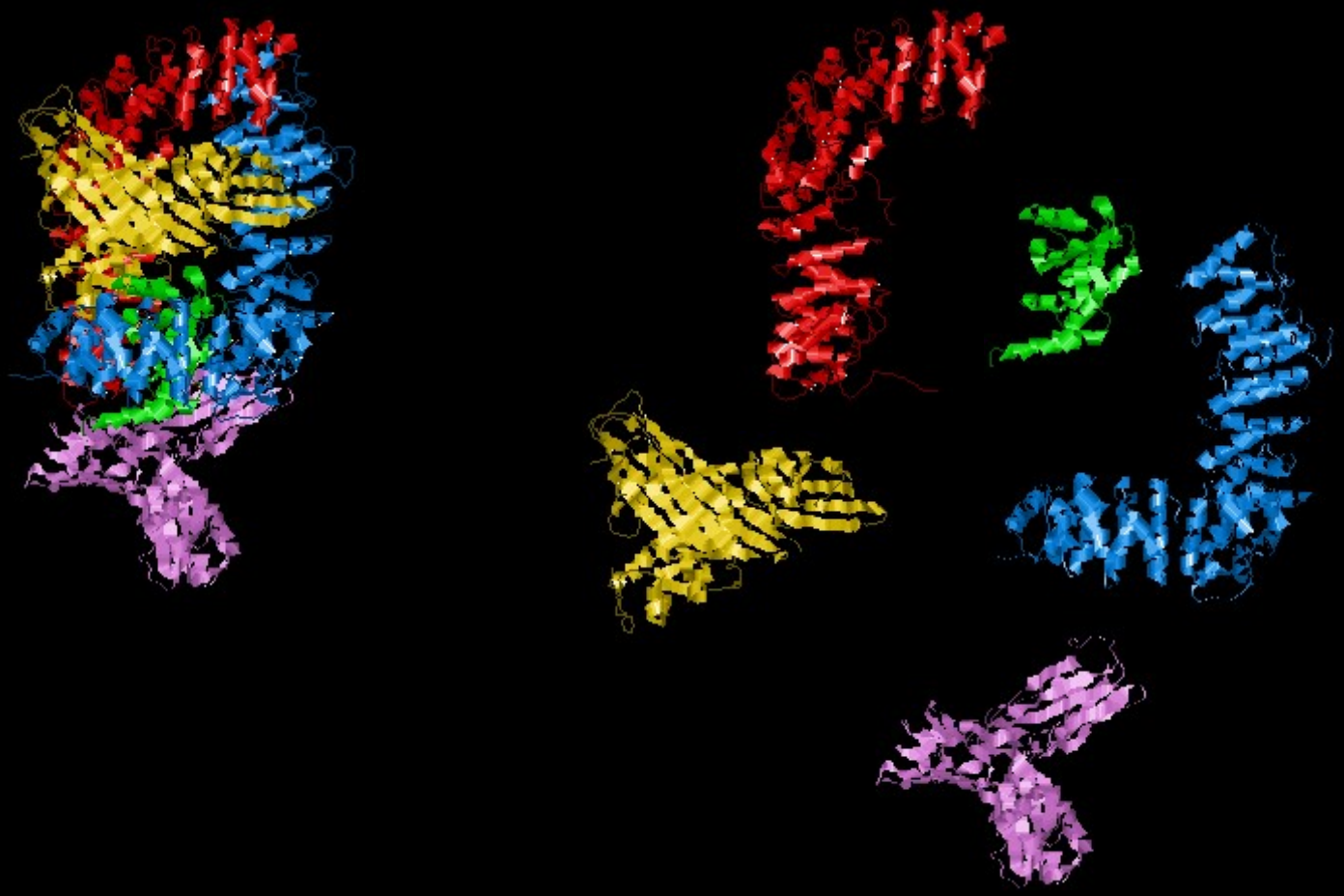
PDB:1b3u
Groves et al. (1999) *Cell*

# Ap1 Clathrin Adaptor Core

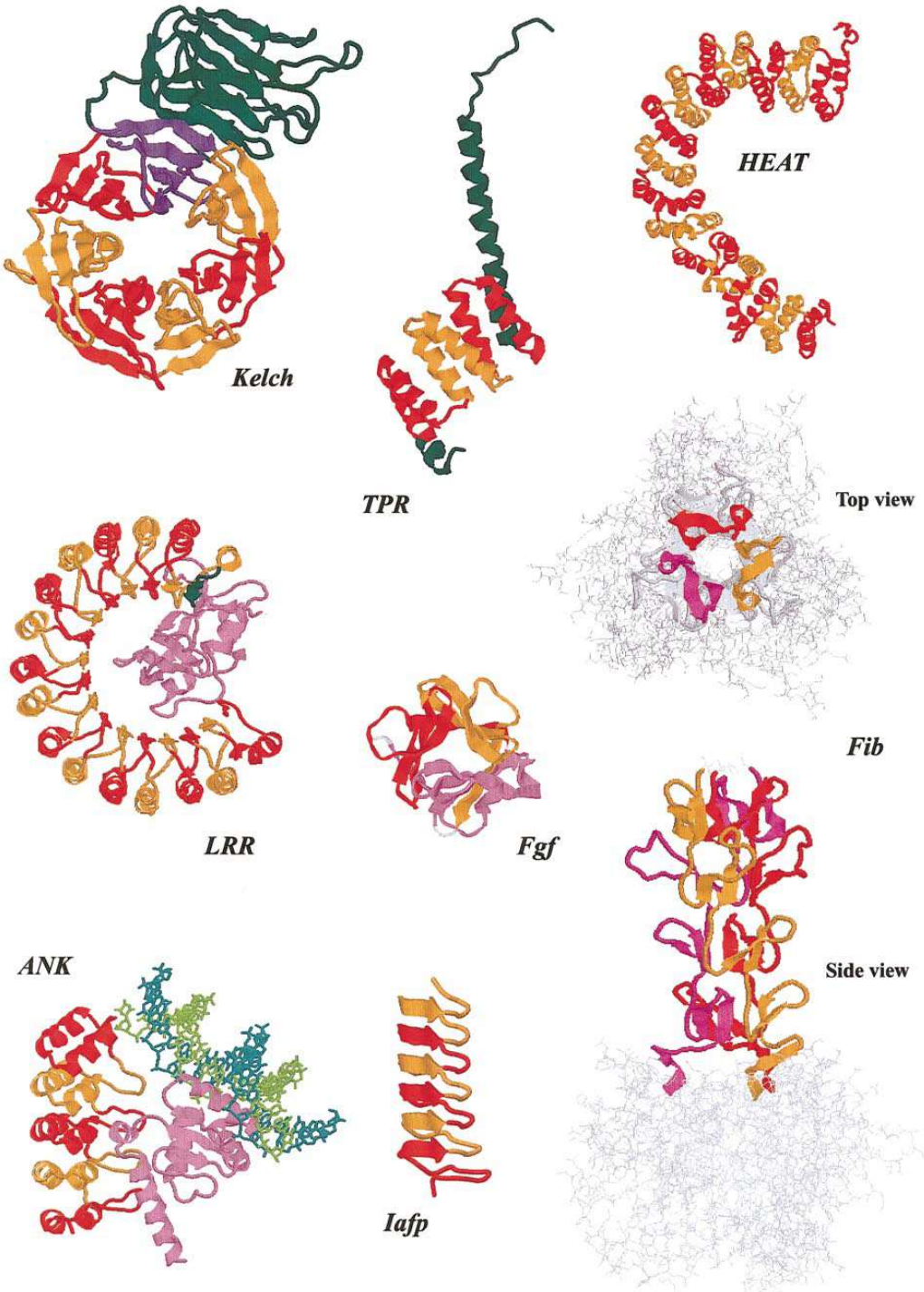


PDB:1w63
Heldwein et al. (2004) *PNAS*

# Ap1 Clathrin Adaptor Core



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




Andrade et al. (2001)
*J Struct Biol*

# Definition CBRs

Perfect repeat: 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**QQQQQQQQQQQQQQQQQQQQQQ****PPPPPPPPPPP**QLPQPPPQAQP
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**QQQQQQQQQQQQQQQQQQQQQQ****PPPPPPPPPPP**QLPQPPPQAQP
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**
**EDEE**ATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

# 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 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**
**FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA**LMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQK**CRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMAS**FG
NFA**NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS**RR**TQYFYSWLLNVLLGLLV**
**PVEDEHSTLLILGVLLTLRYL**VPLLQQQVKDTSLKGSFGVTRKEMEVS**PSAEQLVQVYEL**
**TLHHTQHQDHNVVTGALELLQQLFRT**PPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

# Detection repeats

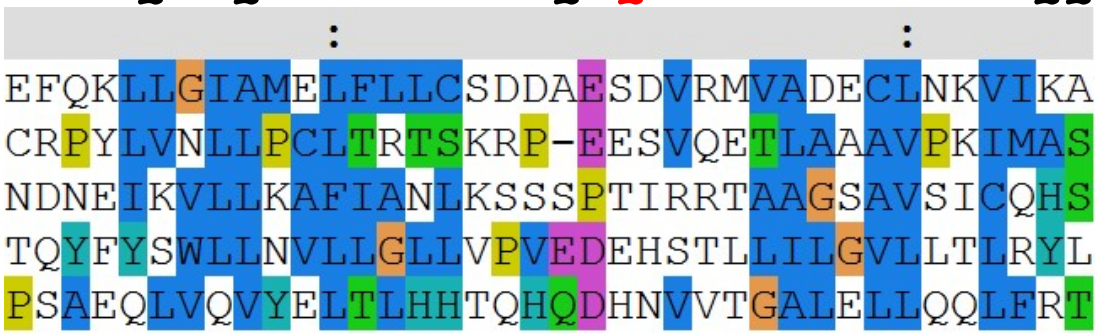
Often NOT straightforward.
N-terminal human Huntingtin.
How many **repeats** can you find?

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

# Detection repeats

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

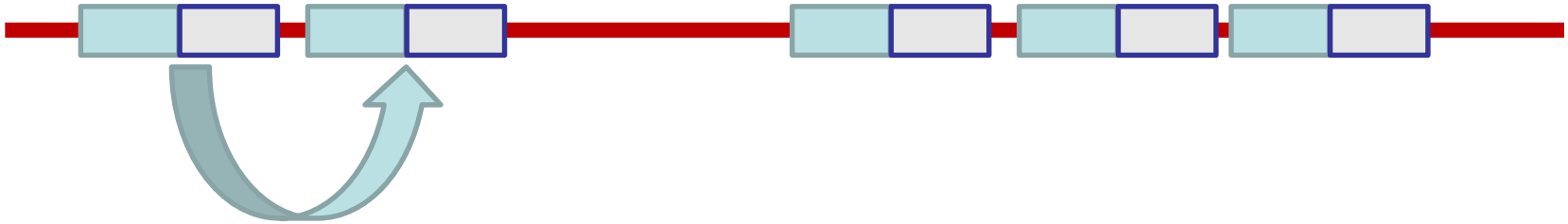
**EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA**
**CRPYLVNLLPCLTRTSKRP-EESVQETLAAAVPKIMAS**
**NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS**
**TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL**
**PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRT**



# Detection of repeats

## Dotplots

## Comparing a sequence against itself



# Detection of repeats

## Dotplots

# Detection of repeats

## Dotplots


1 match

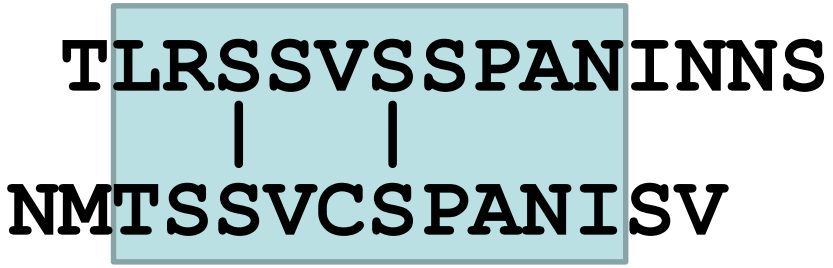
# Detection of repeats

## Dotplots


8 matches

# Detection of repeats

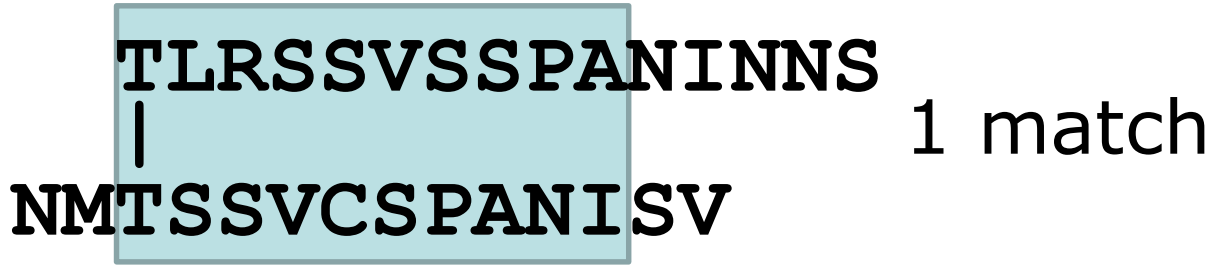
## Dotplots




2 matches

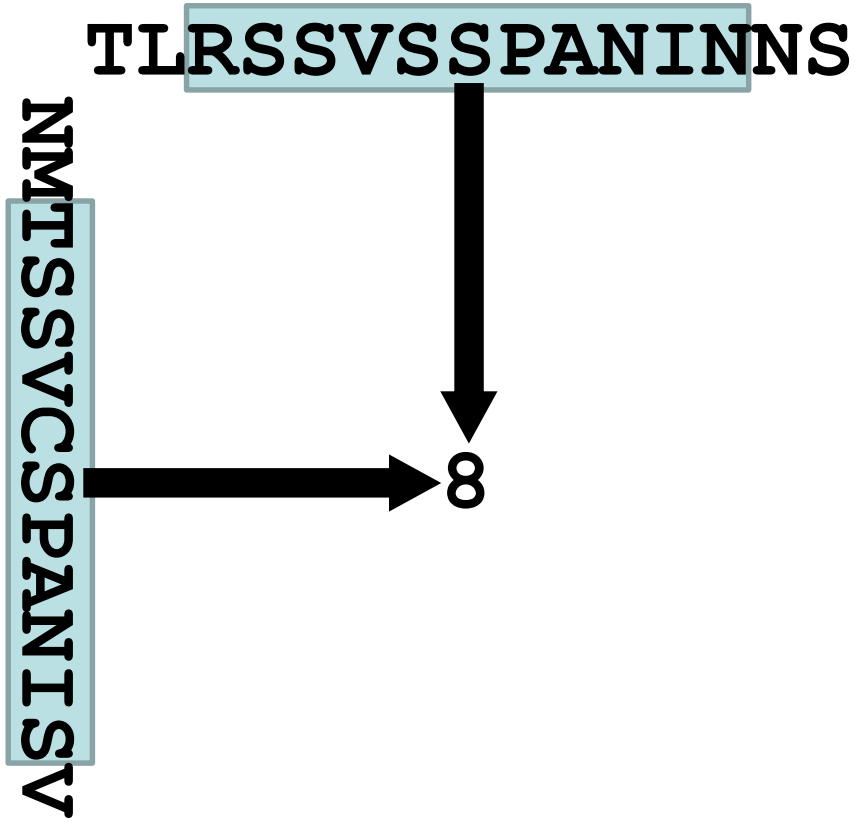
# Detection of repeats

## Dotplots



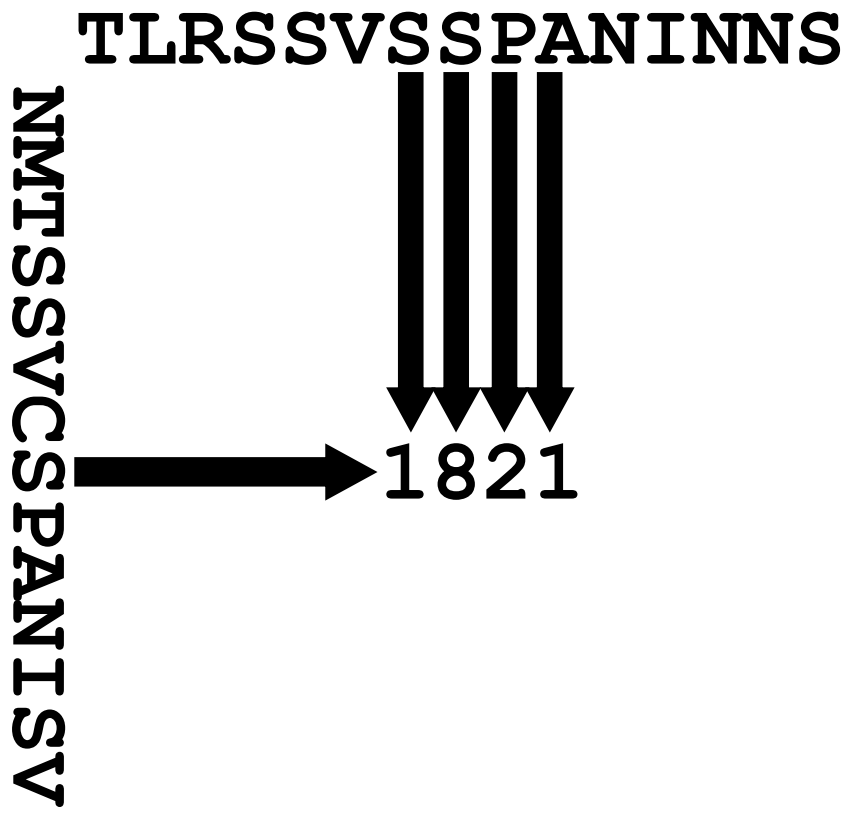
# Detection of repeats

## Dotplots

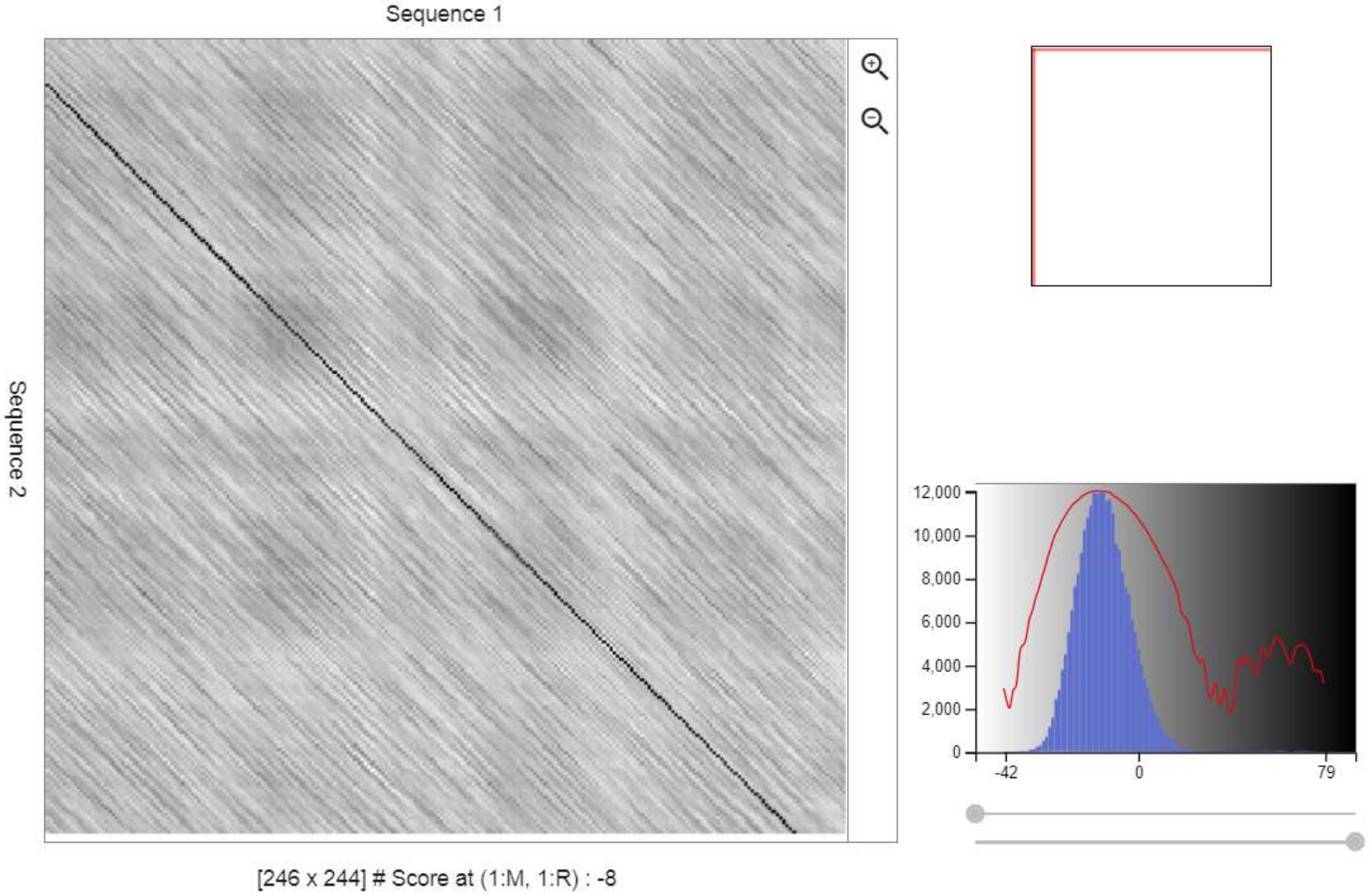


# Detection of repeats

## Dotplots



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




```
Seq1:1
|
MTMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQGHE
RKPLQTPTPIRRLWTMDTSELVQKAKLAEQAERYDDM
|
```

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

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

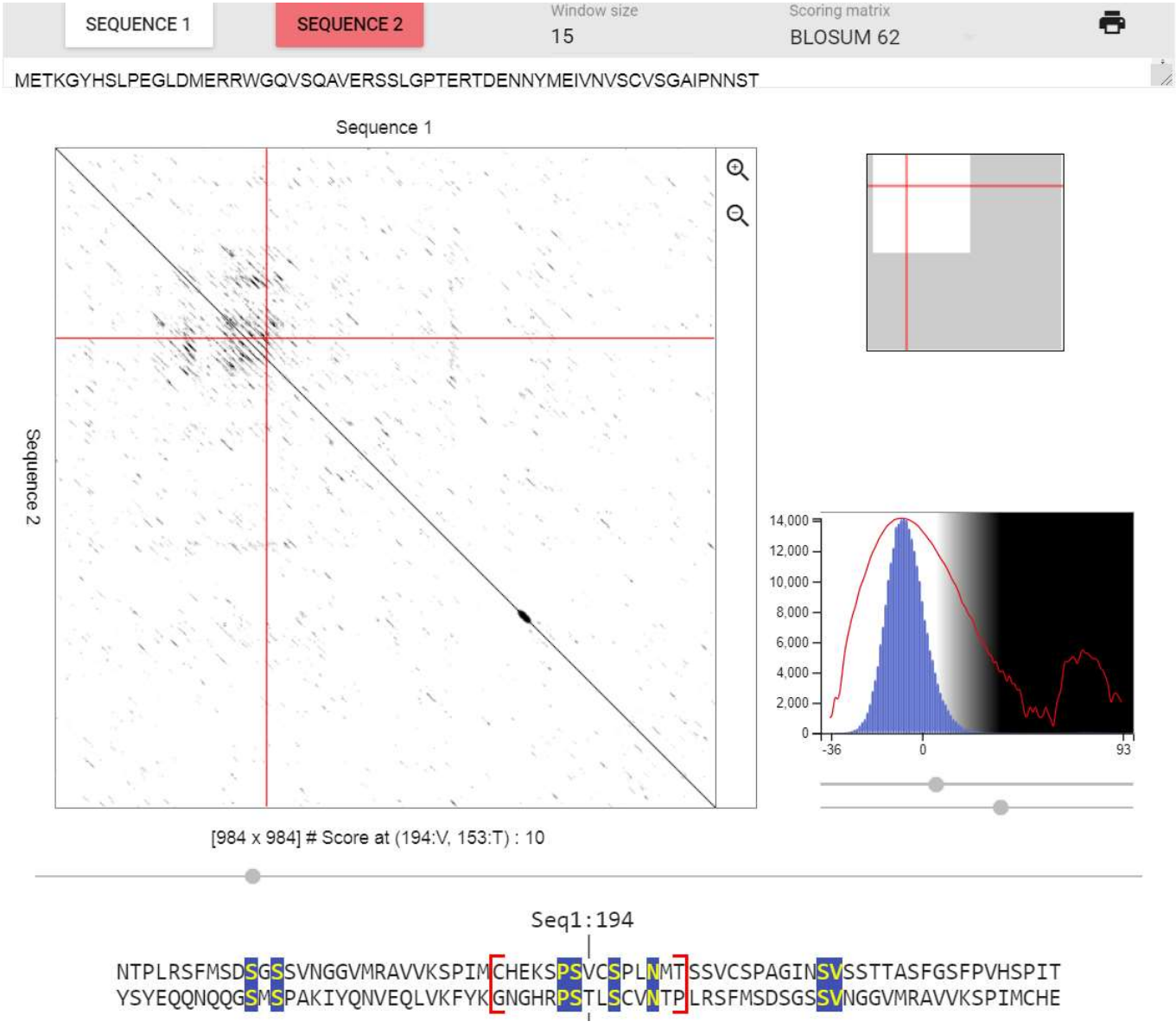
•Click on the input button and paste the sequence of the human mineralocorticoid receptor (UniProt id P08235)

•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)

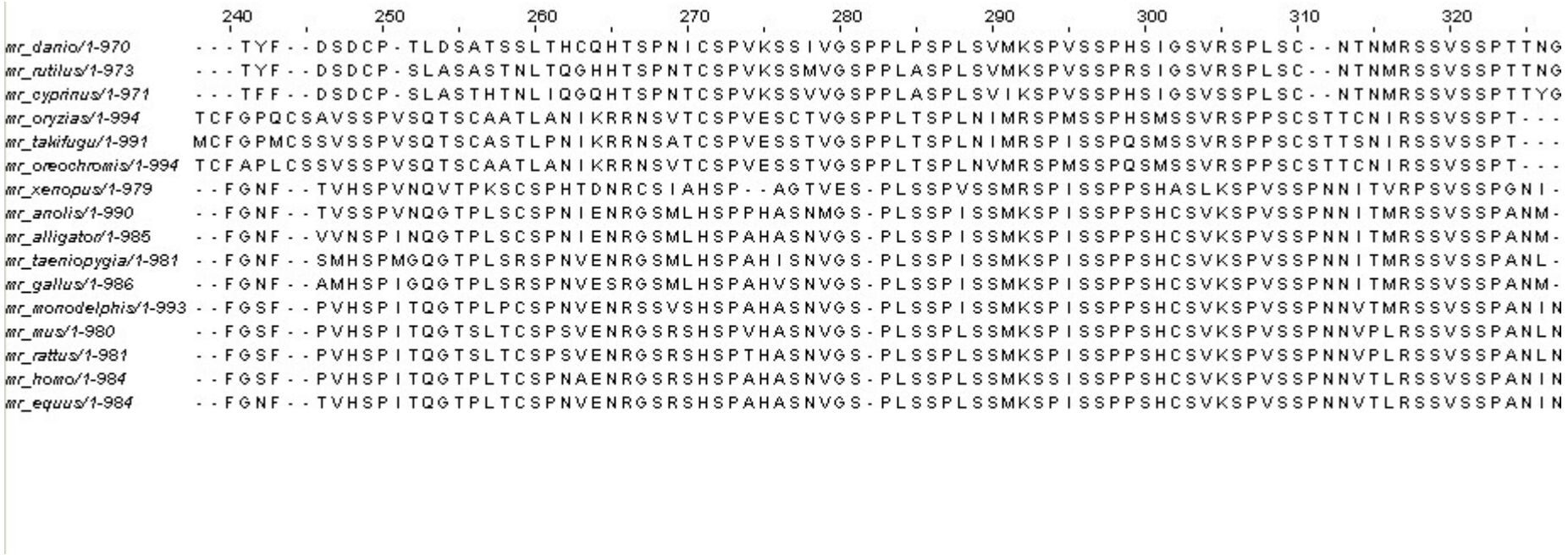
•Find repetitions clicking in the diagonal patterns

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



# Detection of repeats

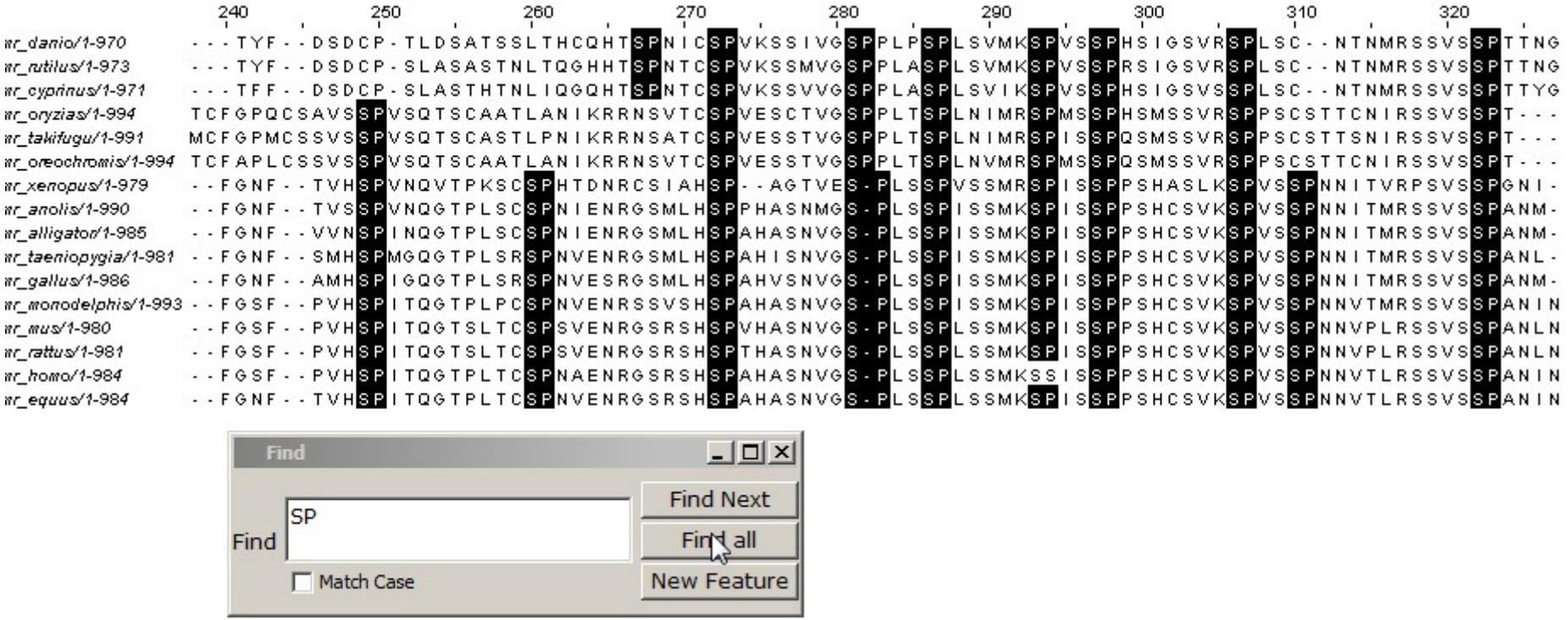
Using a multiple sequence alignment helps.
Conserved repeated patterns




**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?

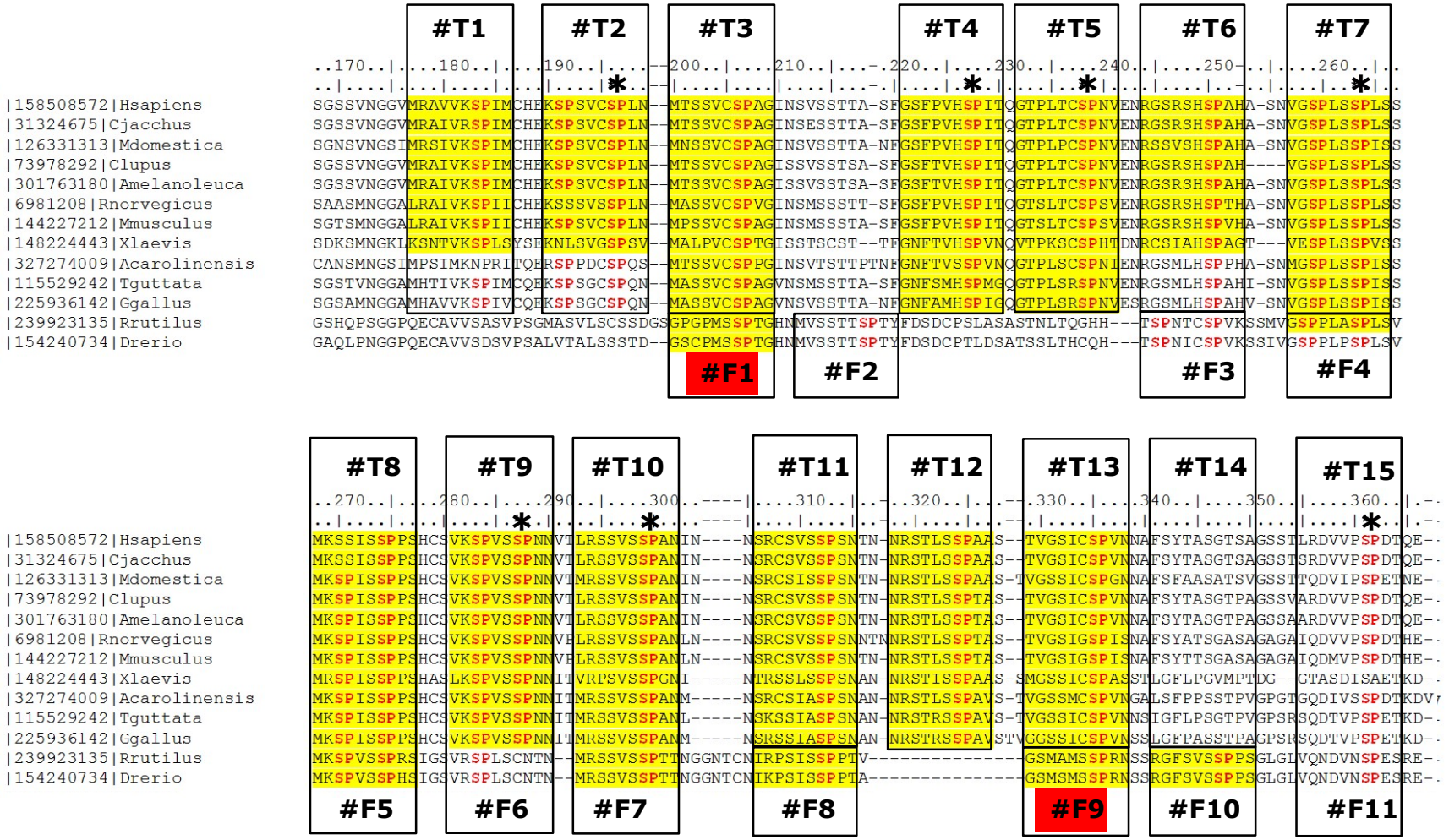
- **`LPTA, SPAA, LPPA, LPAP, SPLA`**

# 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)

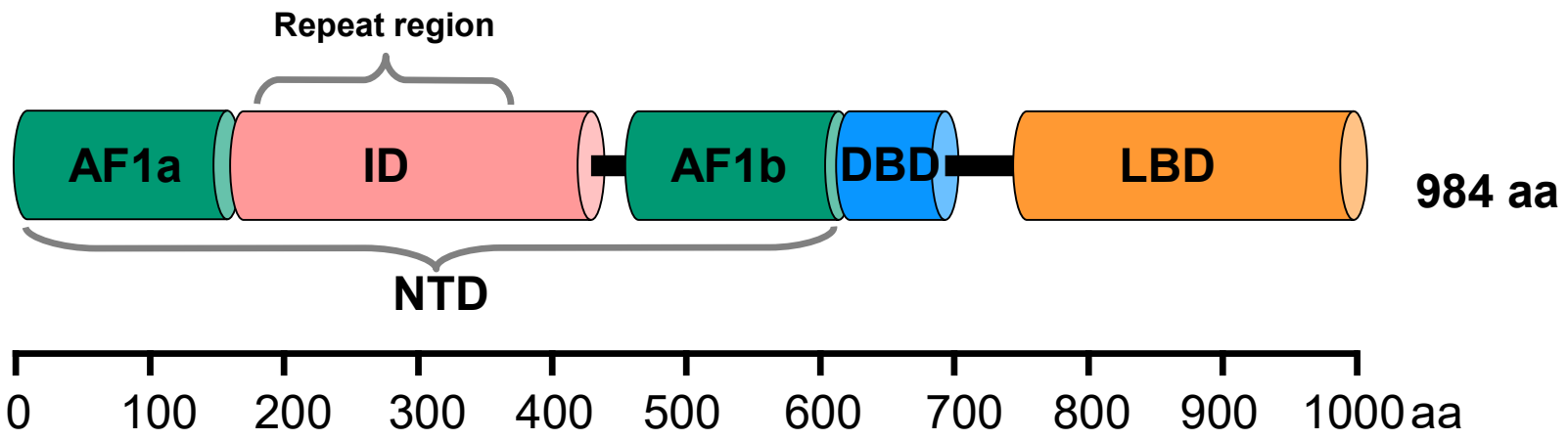
•Use the "Select > find" (of Ctrl+F) option with a regular expression and mark all matches (**click the "Find all" option!**)

•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?

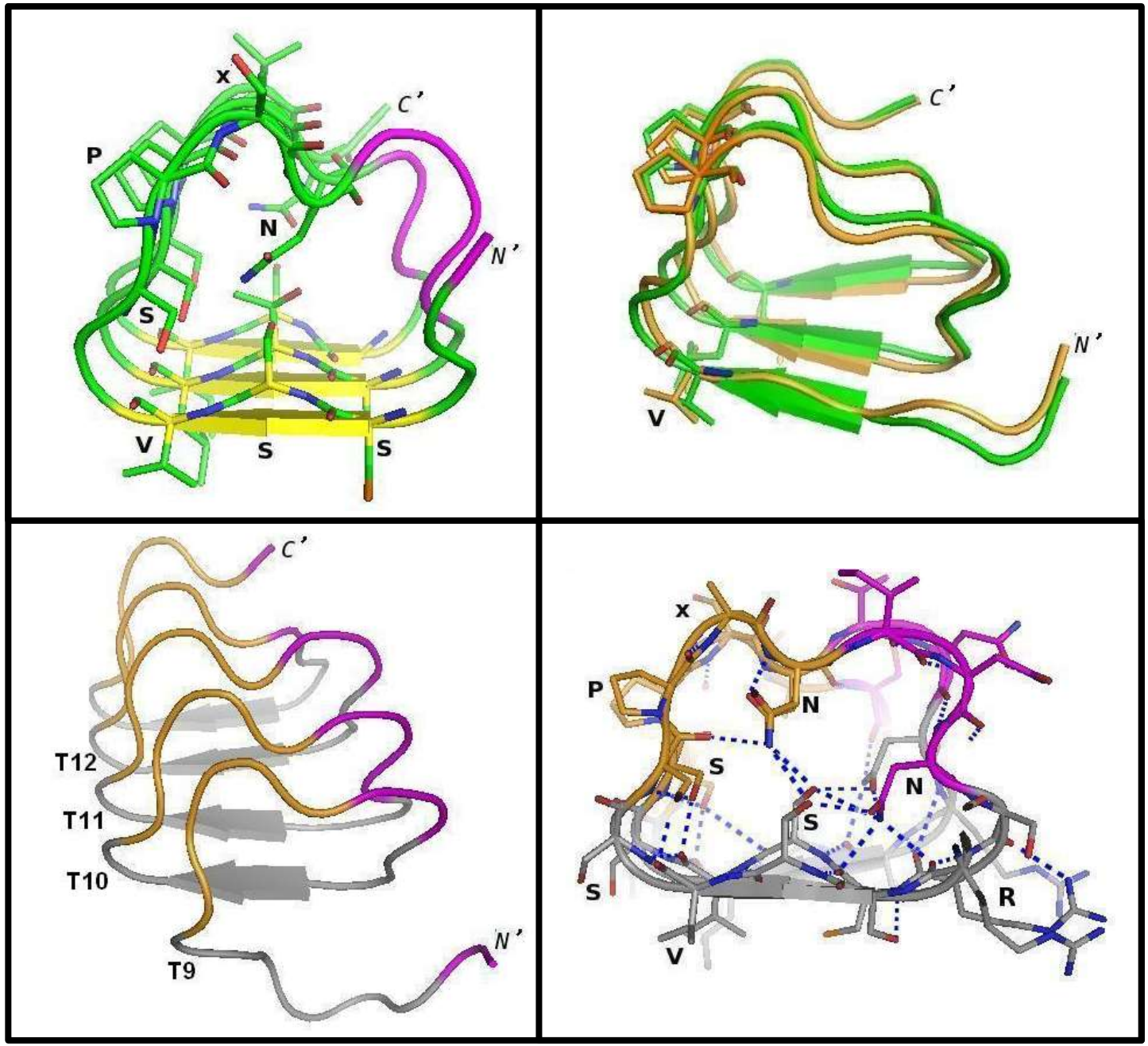



Vlassi *et al.* (2013) *BMC Struct. Biol.*

# 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

# 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

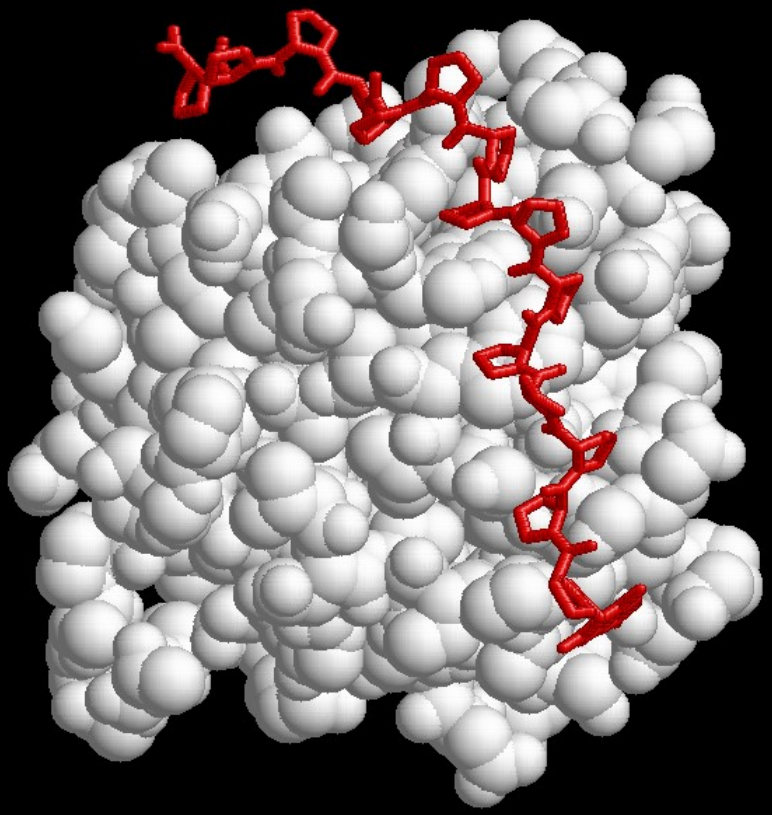
Functions:
Passive: linkers
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

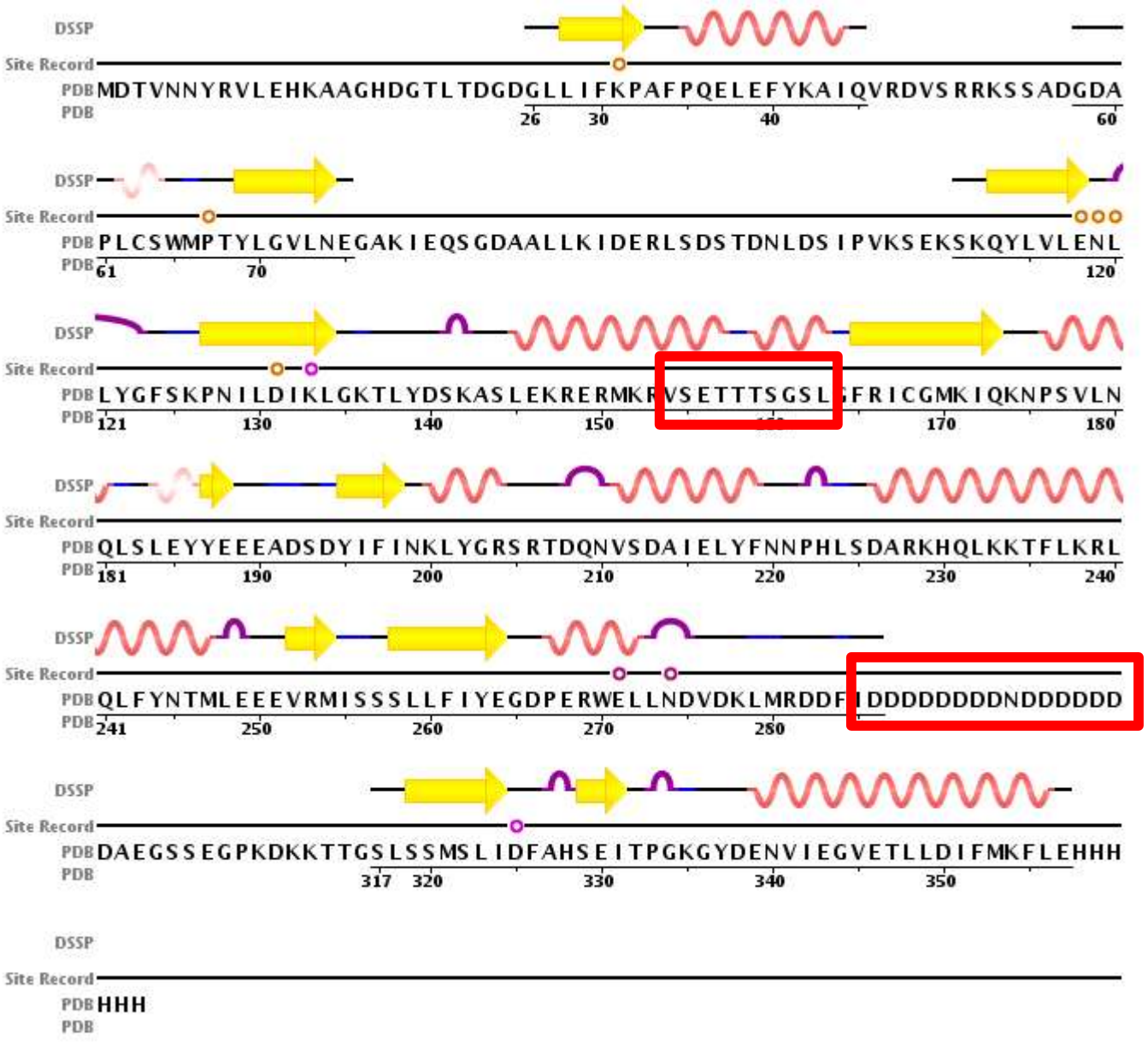
# Structure of CBRs

Often variable or flexible: do not easily crystalize

1CJF: profilin bound to polyP



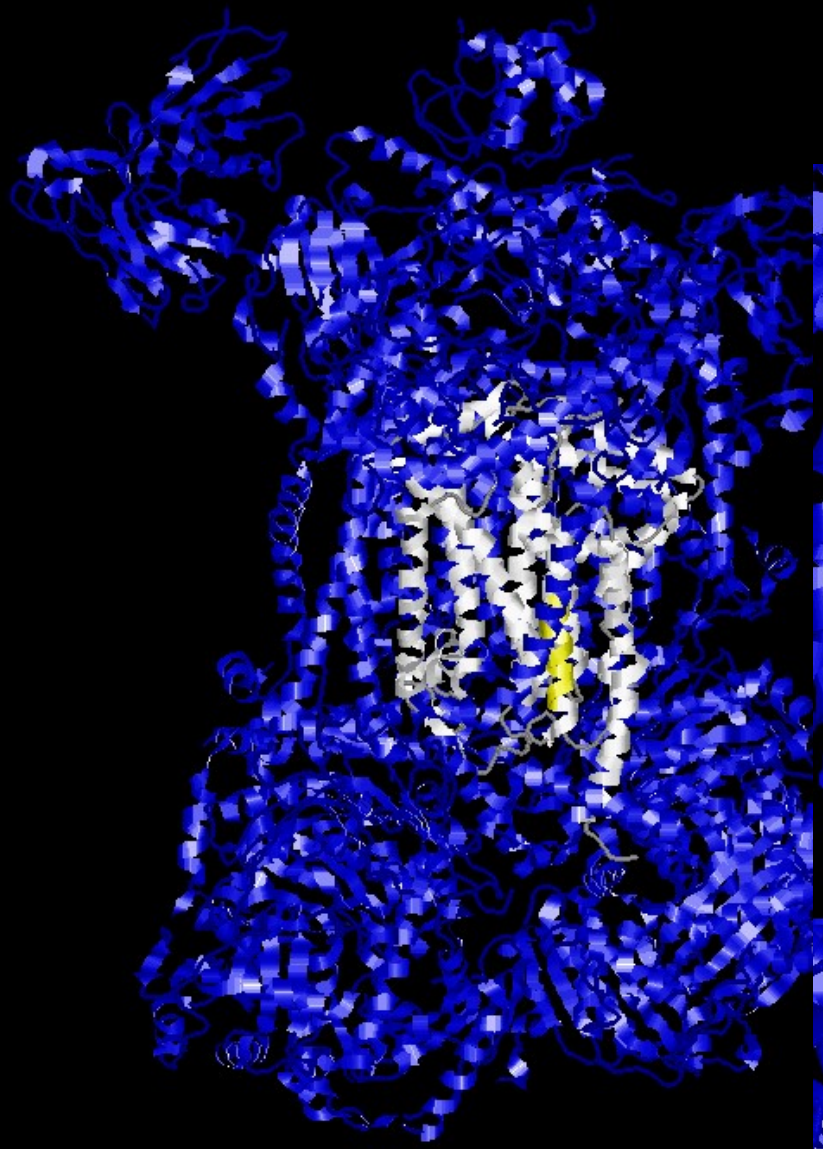
# 2IF8: Inositol Phosphate Multikinase Ipk2

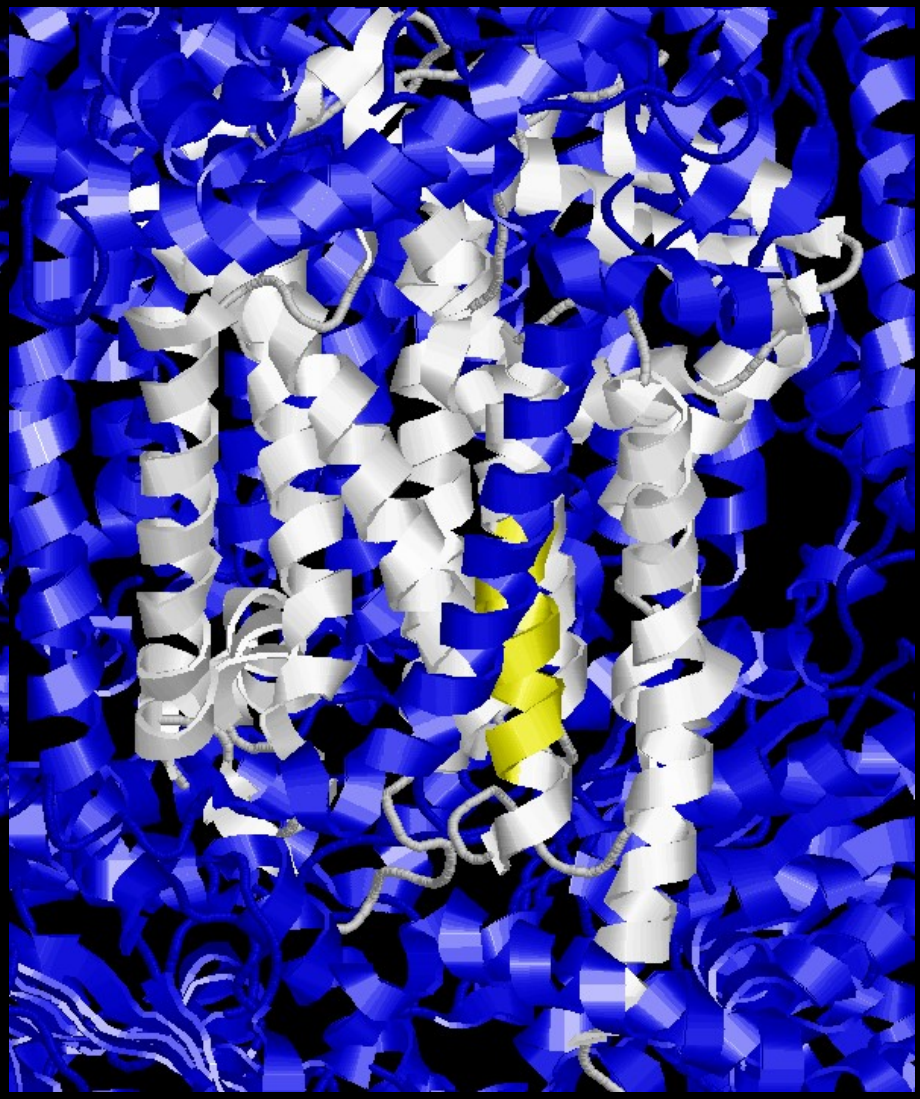


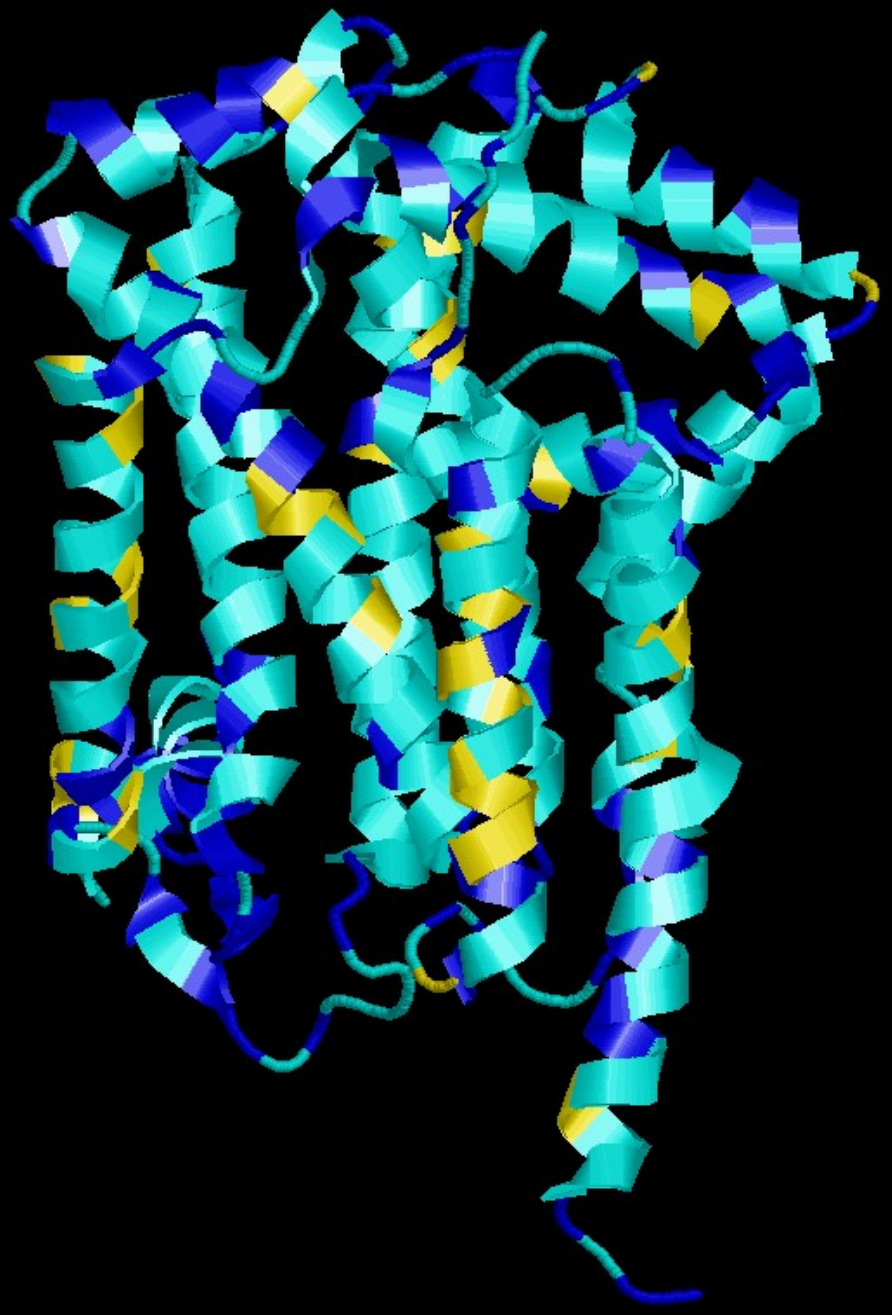
# 2IF8: Inositol Phosphate Multikinase Ipk2

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

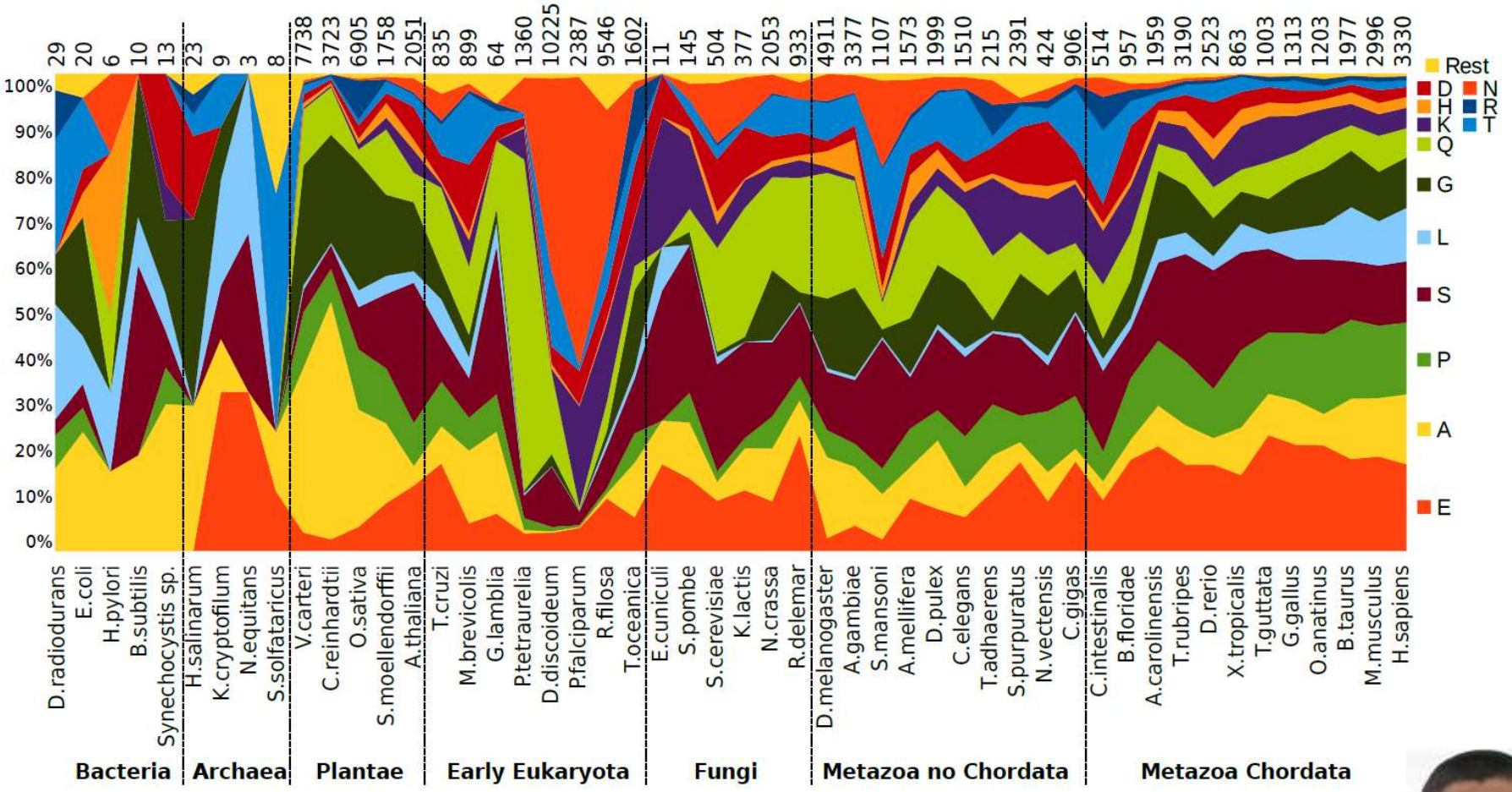
Prokaryota:
Most common: poly-S, poly-G, poly-A, poly-P
Relatively rare: poly-Q, poly-N

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

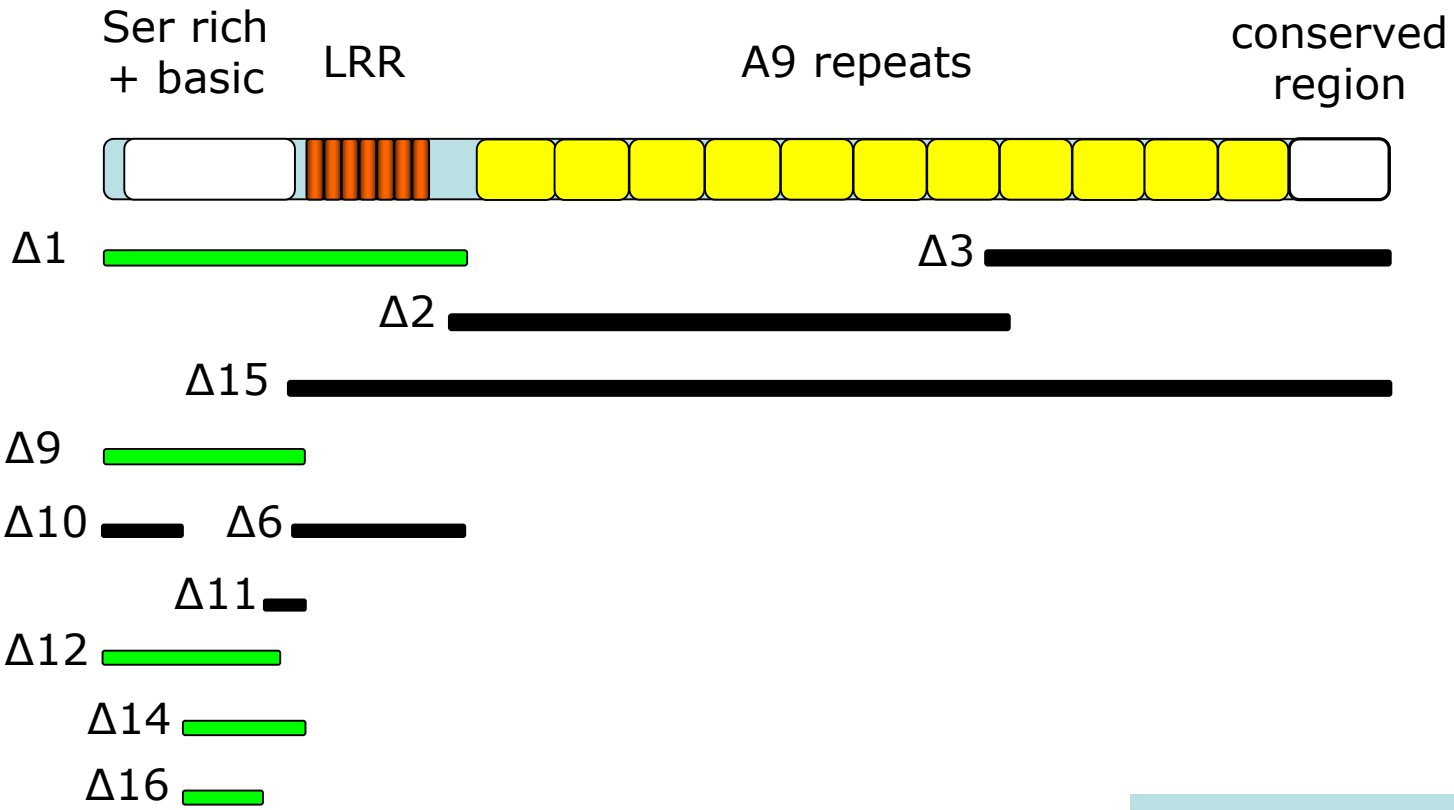
# Filtering out CBRs

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

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:

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQP

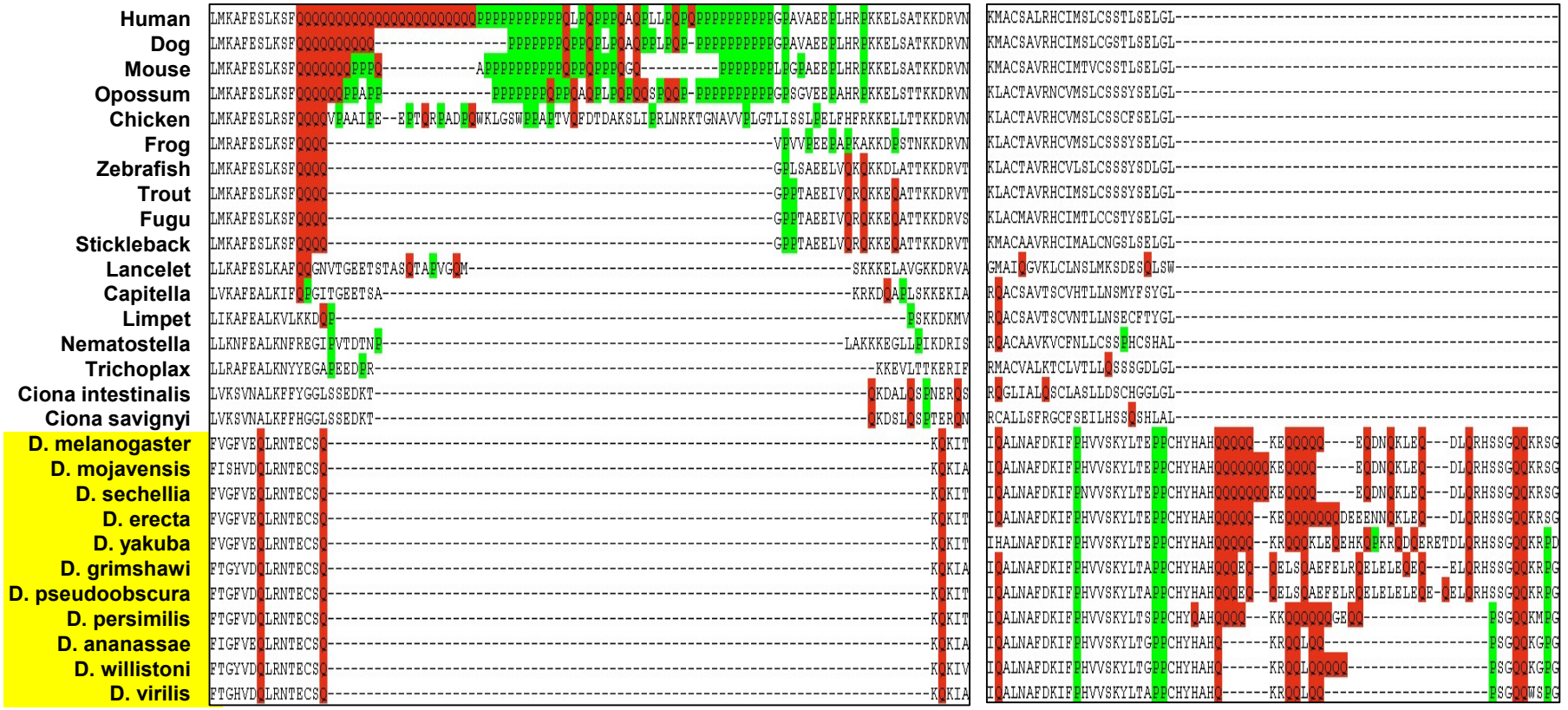
- 10% of human proteins have homorepeats
- lack sequence conservation
- not possible to predict function by homology

Homorepeats need to be studied in context

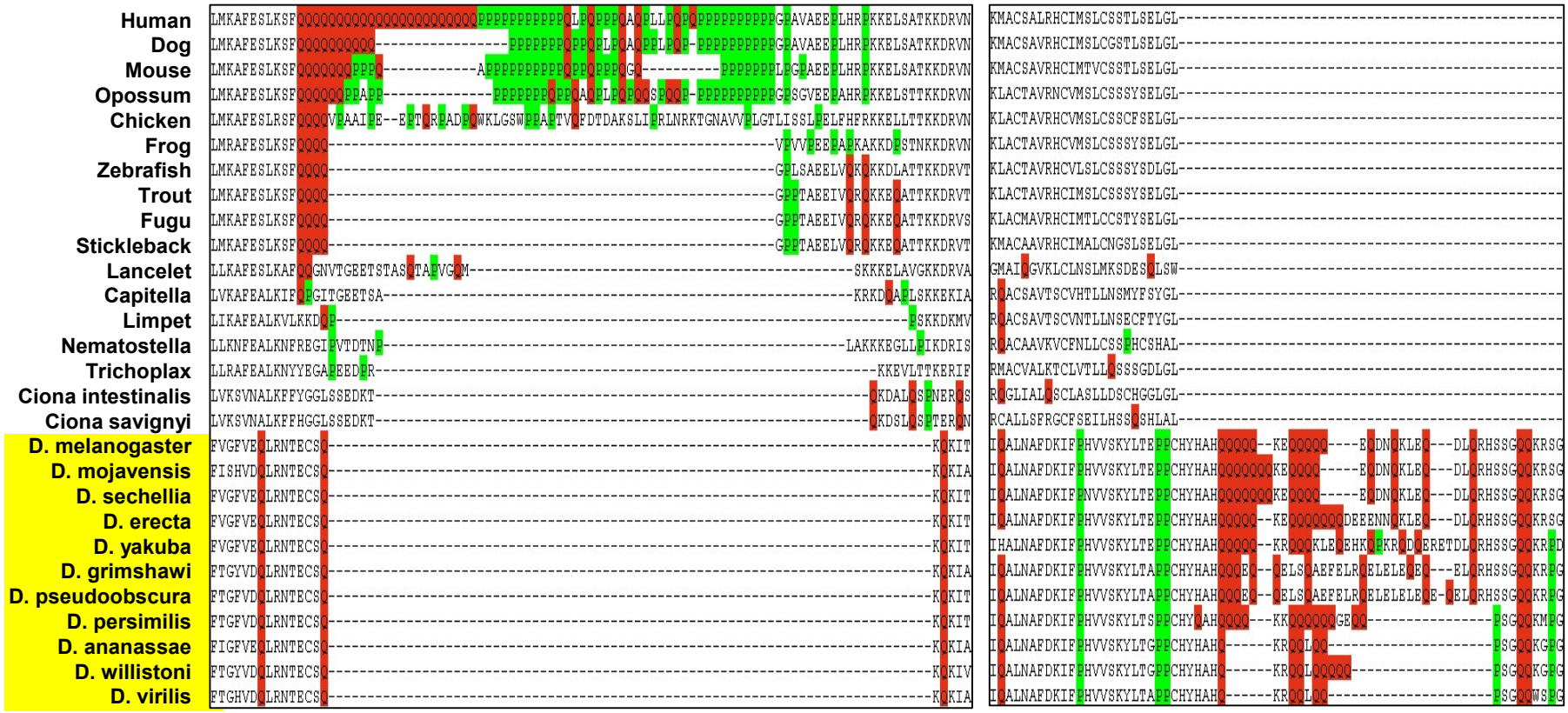
# Function of polyQ

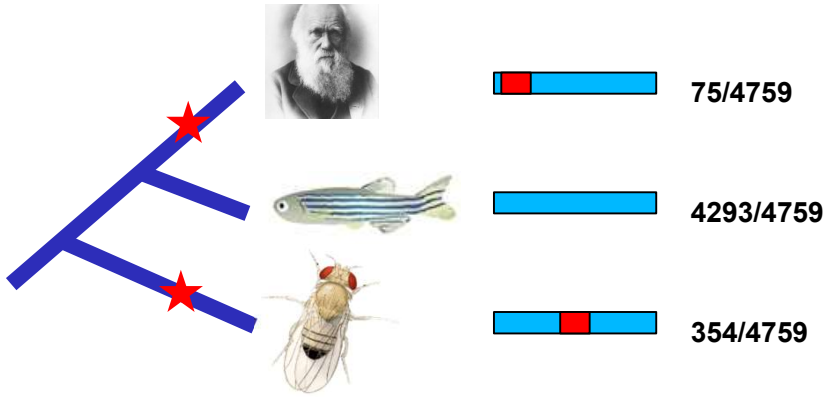
Martin Schaefer

## polyQ in Huntingtin

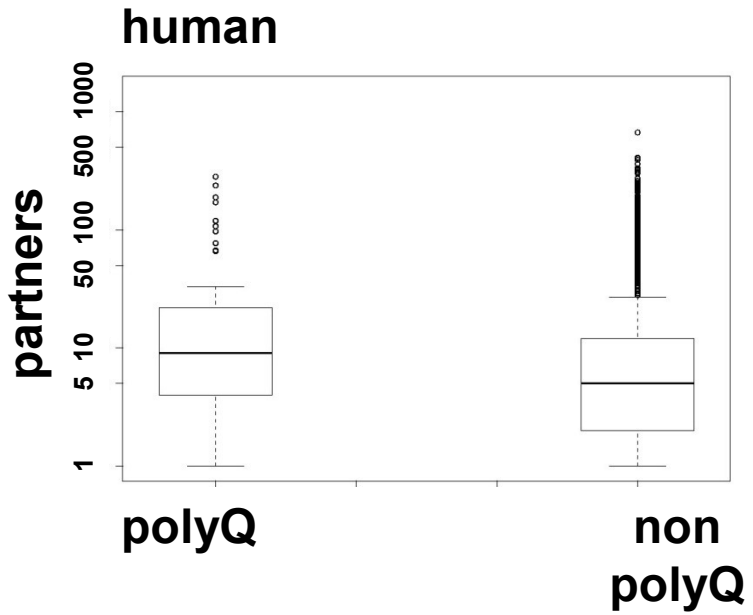


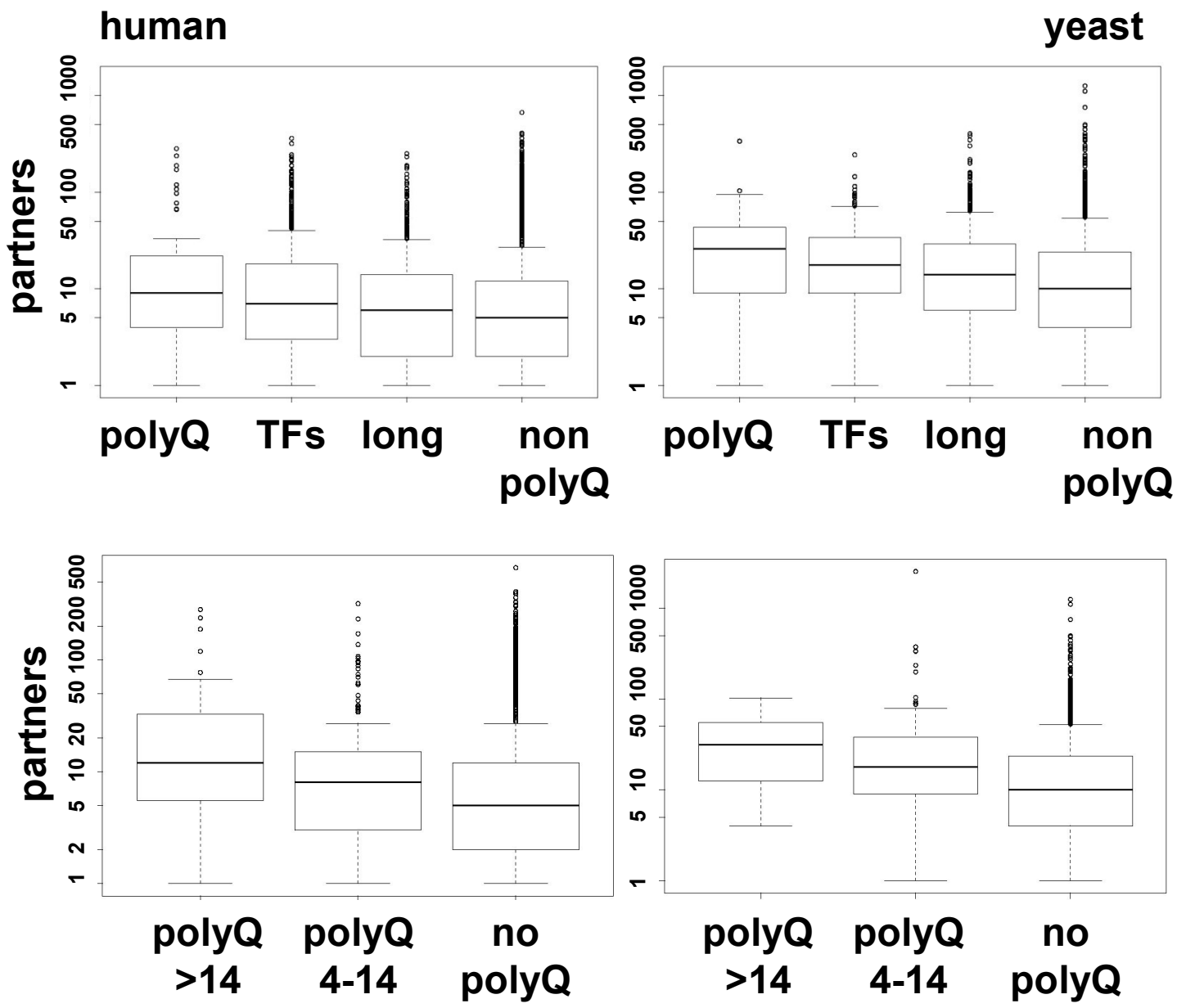

**Schaefer et al (2012) *Nucleic Acids Res.***



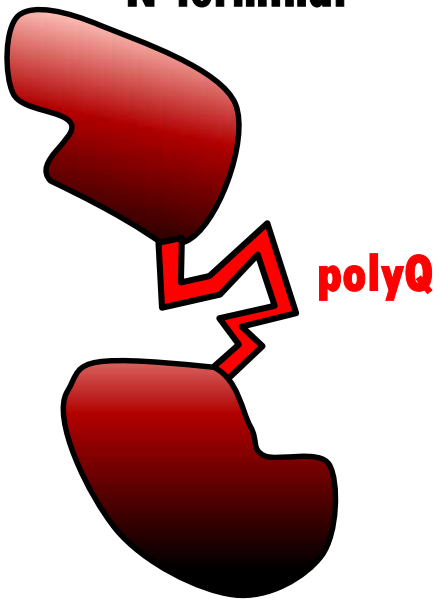

14 trios
P-val < 0.05


human
partners
1
5
10
50
100
500
1000
polyQ
non
polyQ


human
yeast
partners
1
5
10
50
100
500
1000
polyQ
TFs
long
non polyQ
1
2
5
10
20
50
100
200
500
polyQ >14
polyQ 4-14
no polyQ

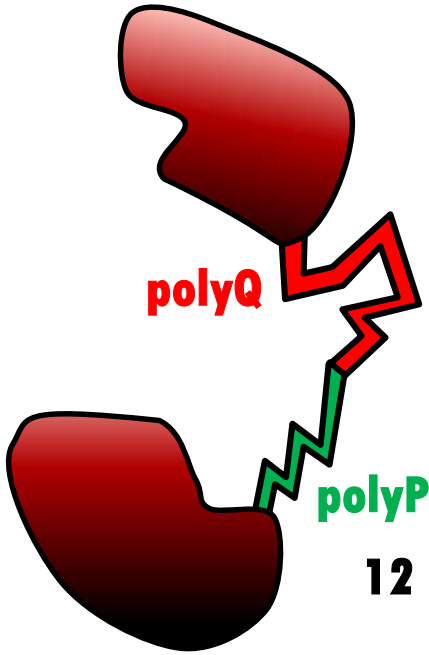
86 human
polyQ
proteins

N-terminal




C-terminal

13 polyQ
proteins with
near polyP




12 C-term

109 polyQ regions
54 overlap/near
coiled coil

86 human
polyQ
proteins

N-terminal

polyQ

C-terminal

13 polyQ
proteins with
near polyP

polyQ

polyP

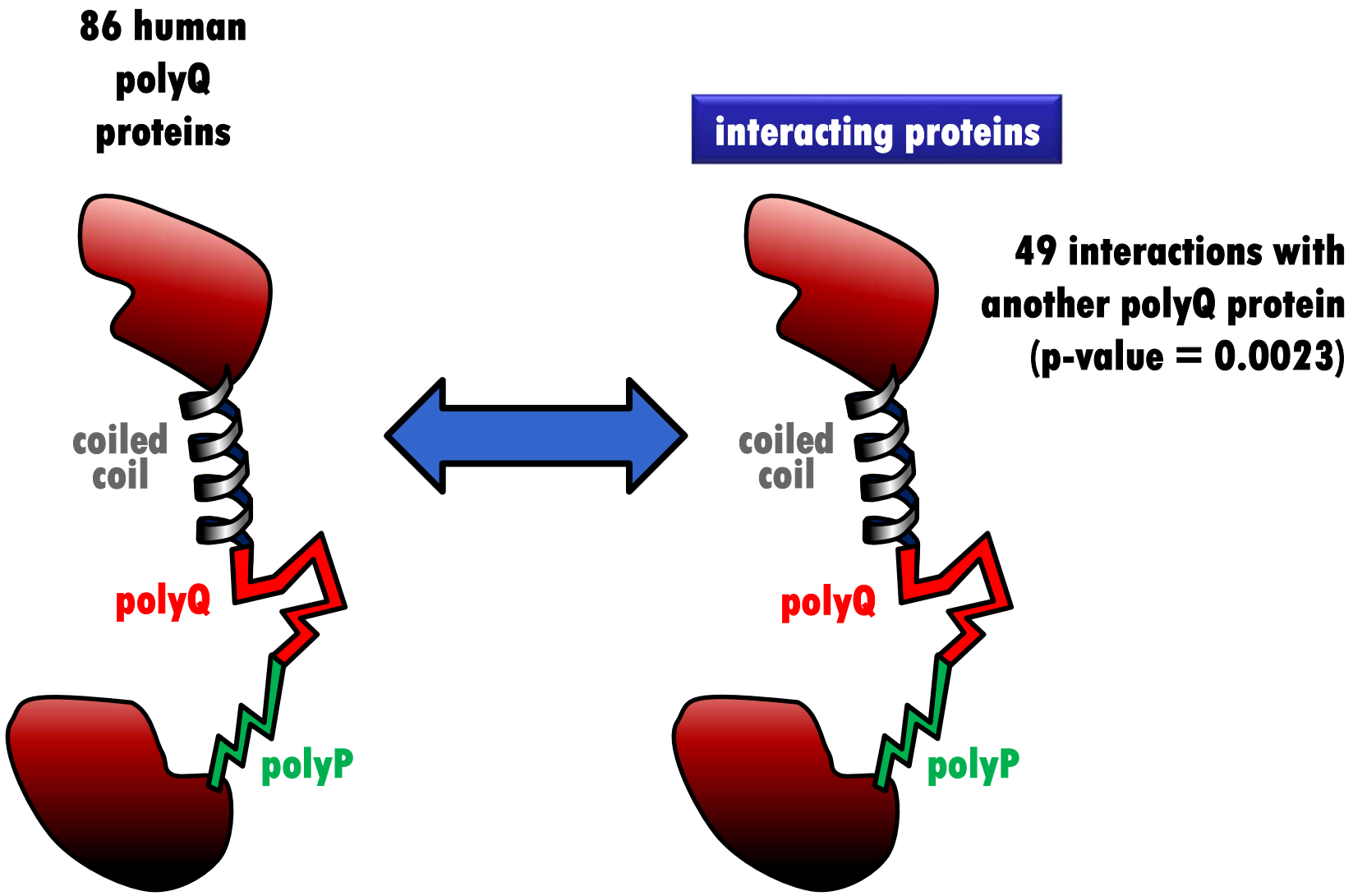
12 C-term

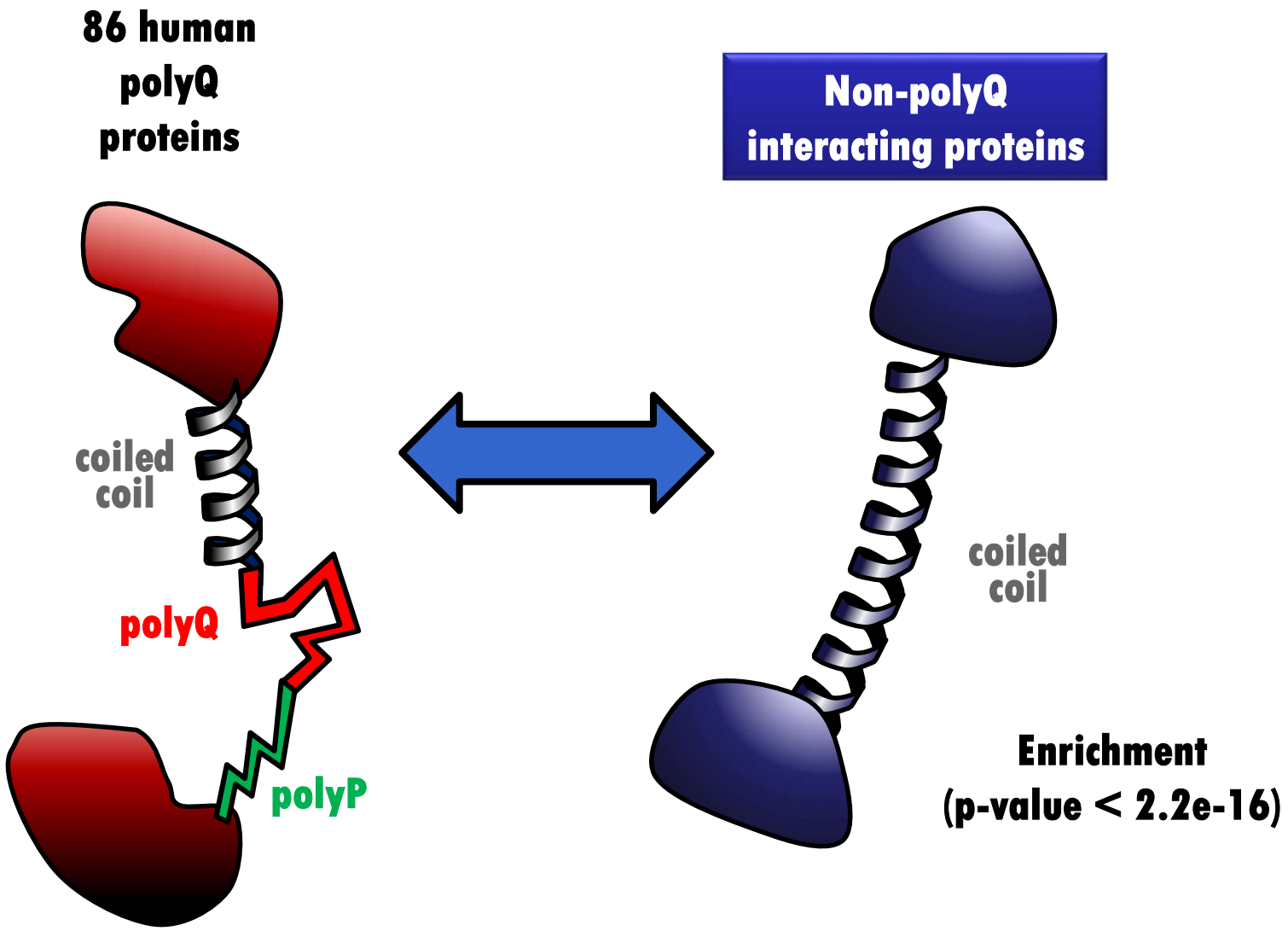
40 human
polyQ/coiled-coil
proteins
(no polyP)

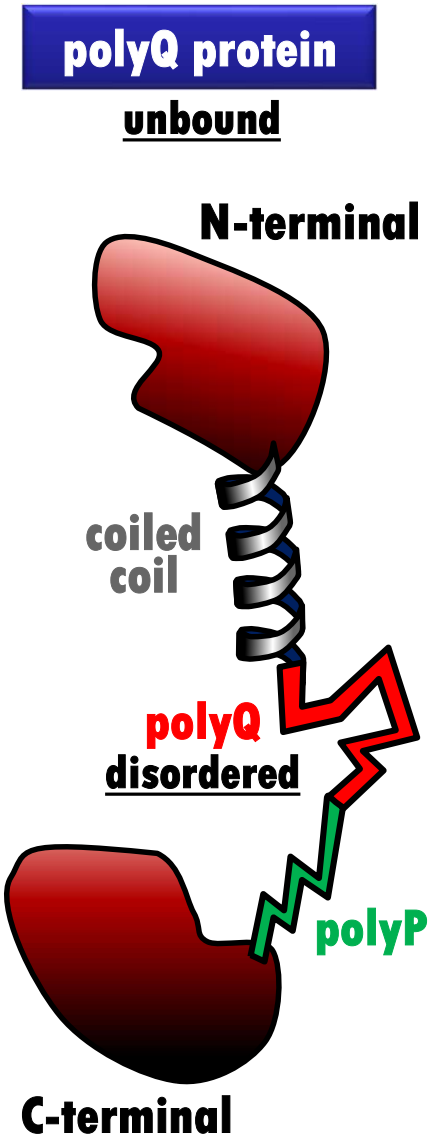
coiled
coil

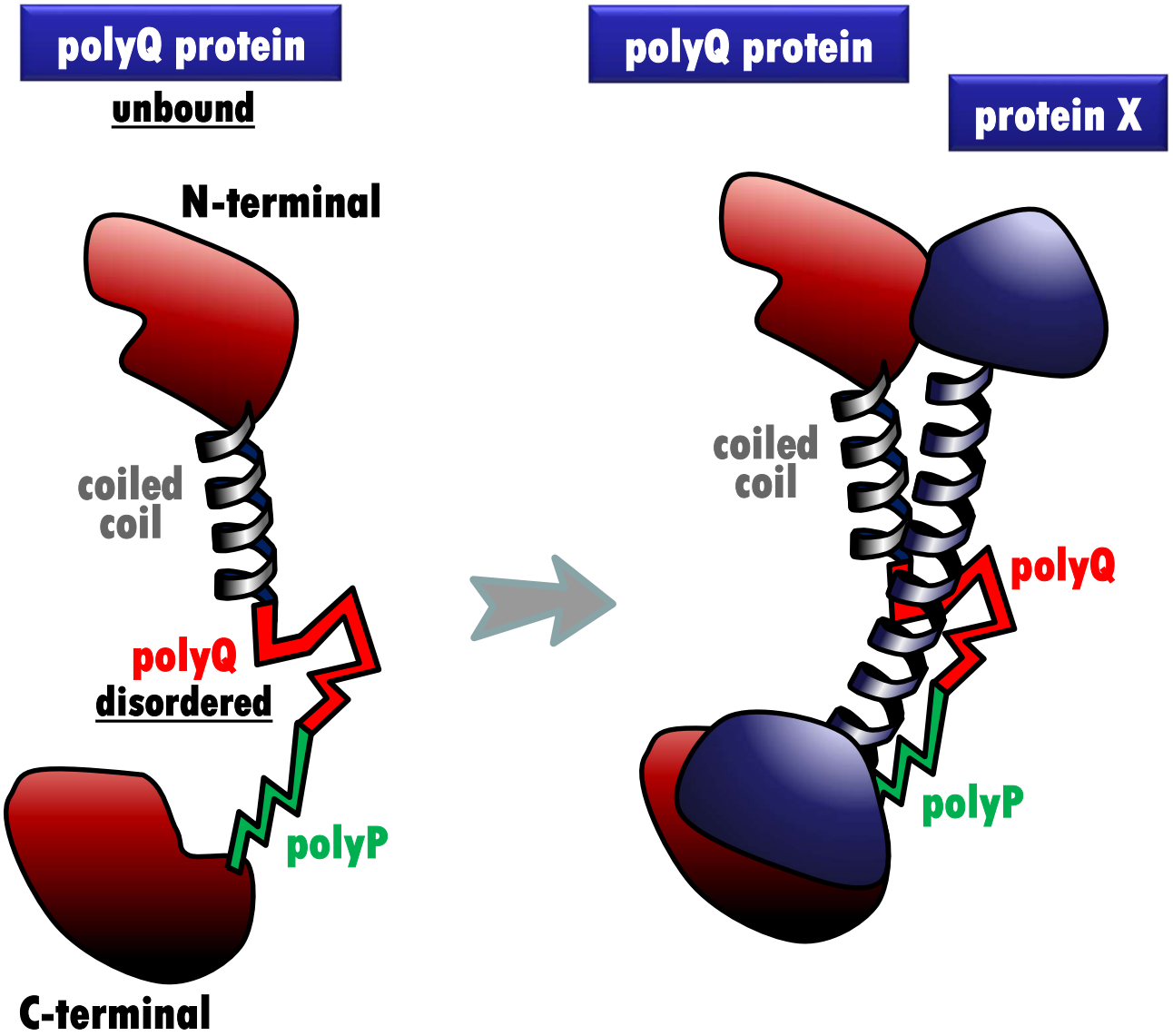
36 N-term

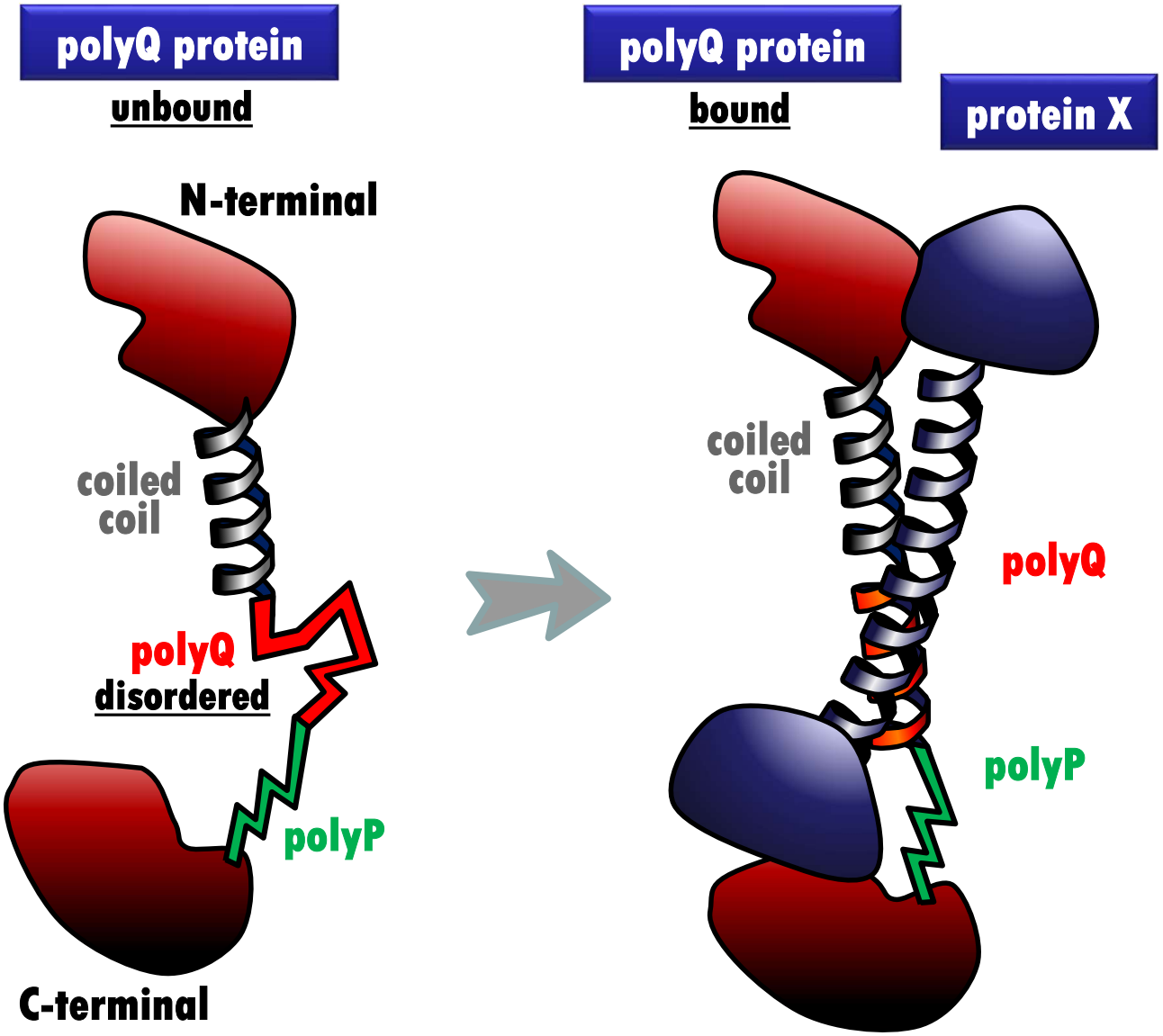
polyQ


86 human
polyQ
proteins
interacting proteins
49 interactions with
another polyQ protein
(p-value = 0.0023)
coiled
coil
polyQ
polyP
coiled
coil
polyQ
polyP

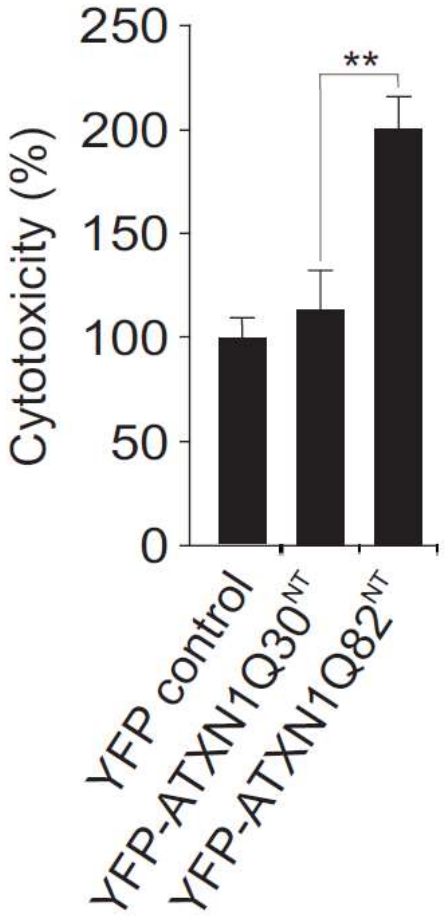

86 human
polyQ
proteins
coiled
coil
polyQ
polyP
Non-polyQ
interacting proteins
coiled
coil
Enrichment
(p-value < 2.2e-16)


polyQ protein
unbound
N-terminal
coiled
coil
polyQ
disordered
polyP
C-terminal

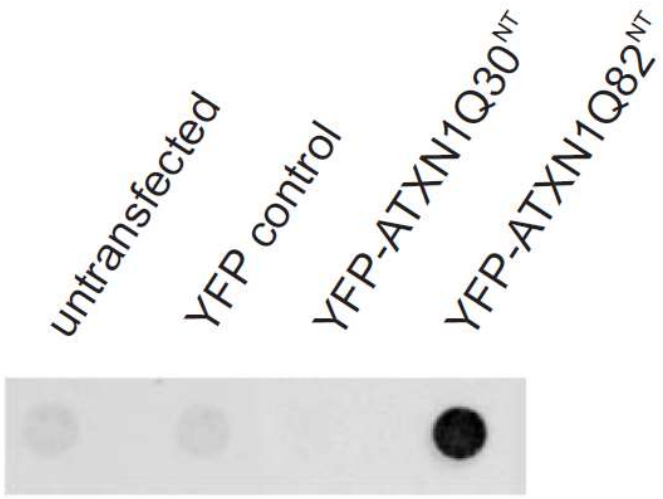

polyQ protein
unbound
N-terminal
coiled
coil
polyQ
disordered
polyP
C-terminal
polyQ protein
protein X
coiled
coil
polyQ
polyP


polyQ protein
unbound
N-terminal
coiled
coil
polyQ
disordered
polyP
C-terminal
polyQ protein
bound
protein X
coiled
coil
polyQ
polyP

## ATXN1Q82$^{NT}$ is toxic




## ATXN1Q82$^{NT}$ aggregates

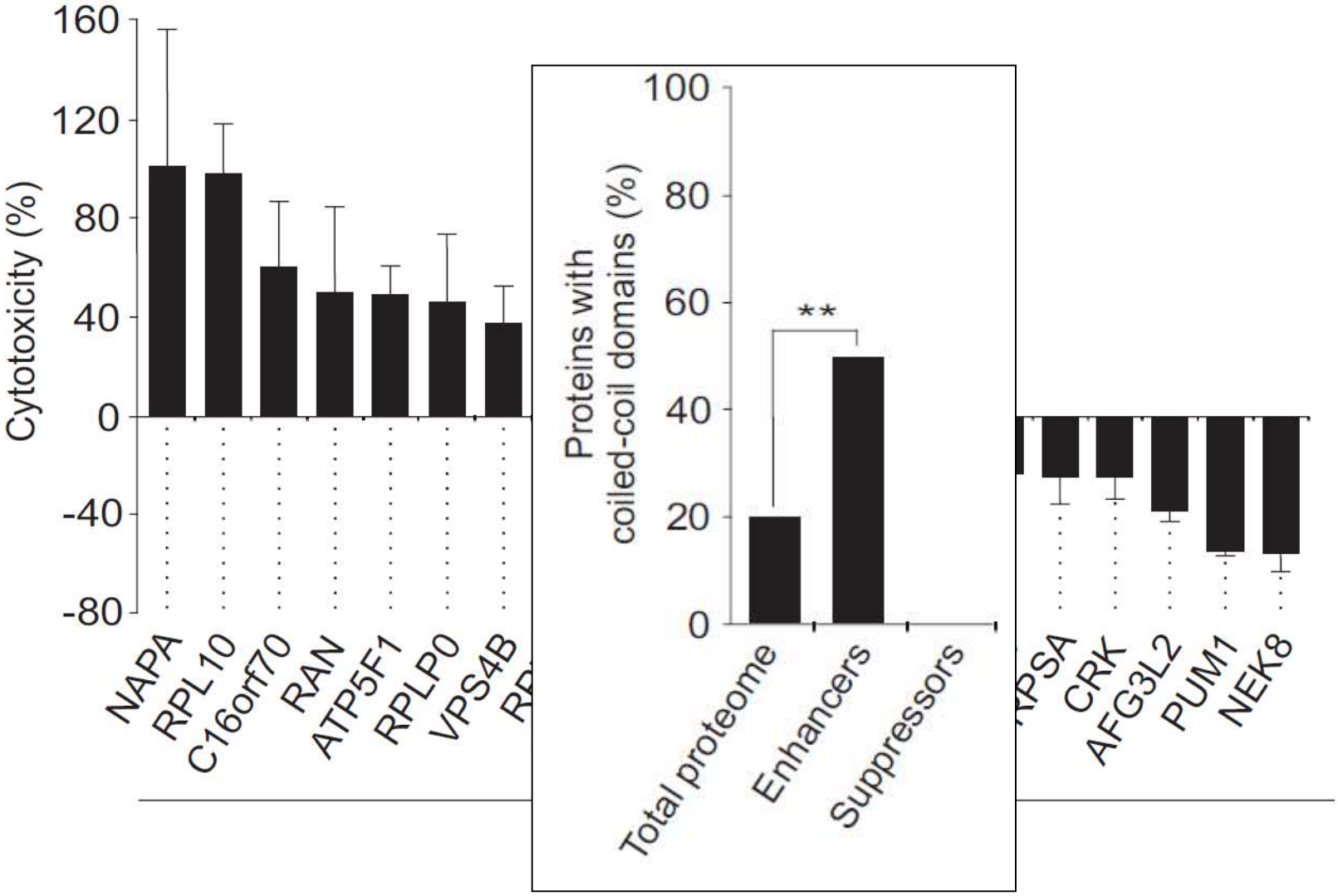


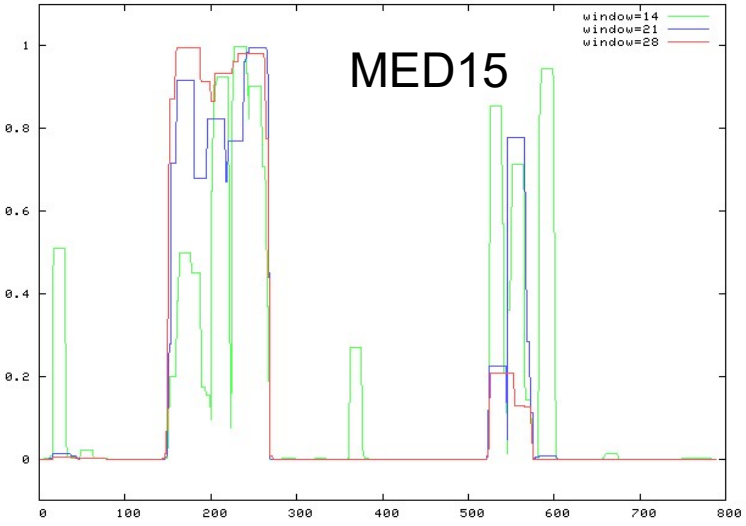

Spyros Petrakis

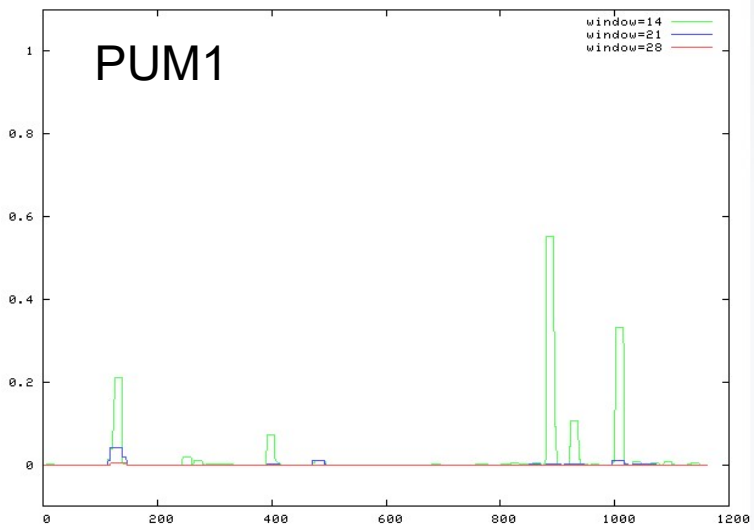
Erich Wanker

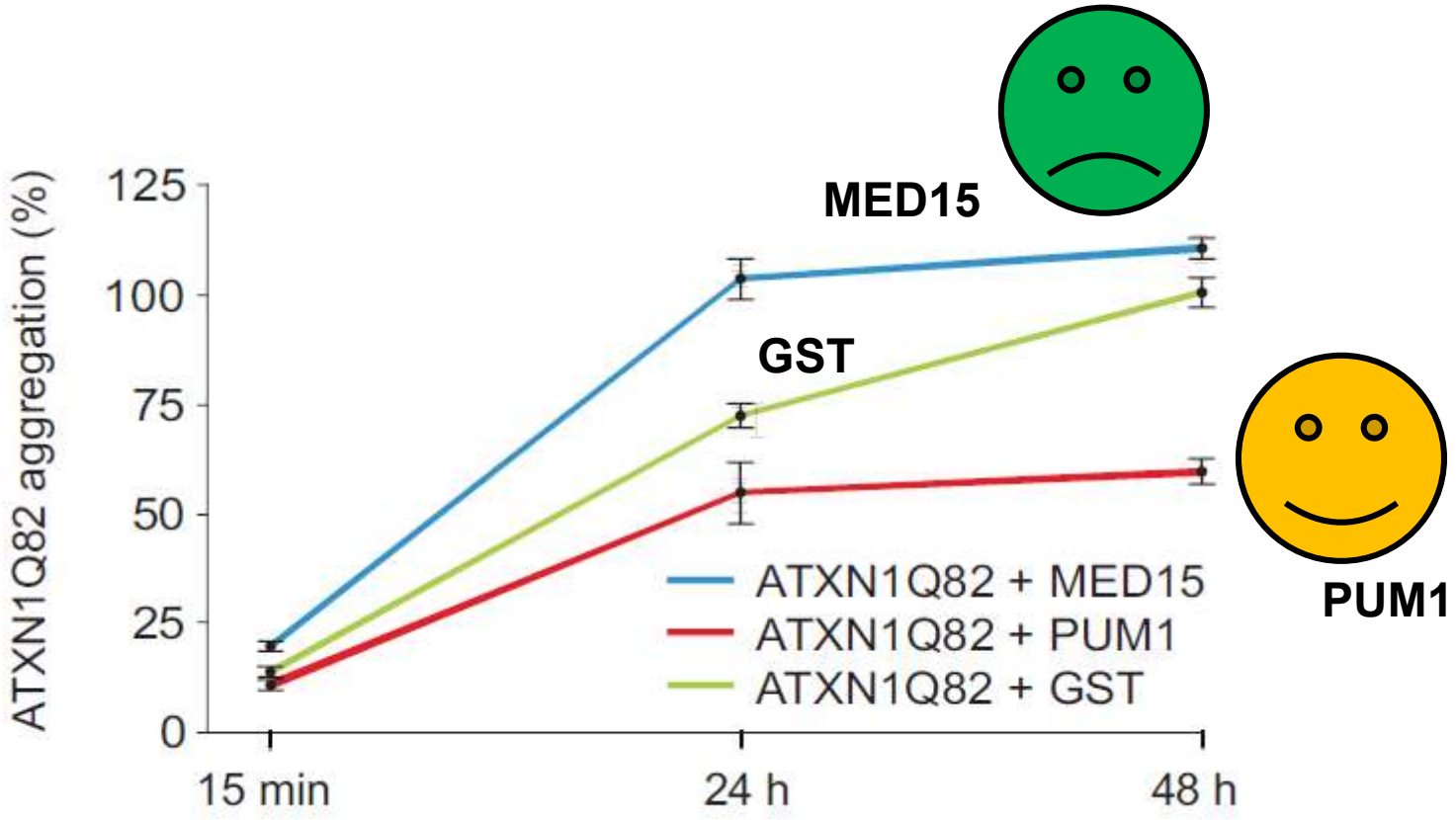
Petrakis *et al.* (2012) *PLoS Genetics*

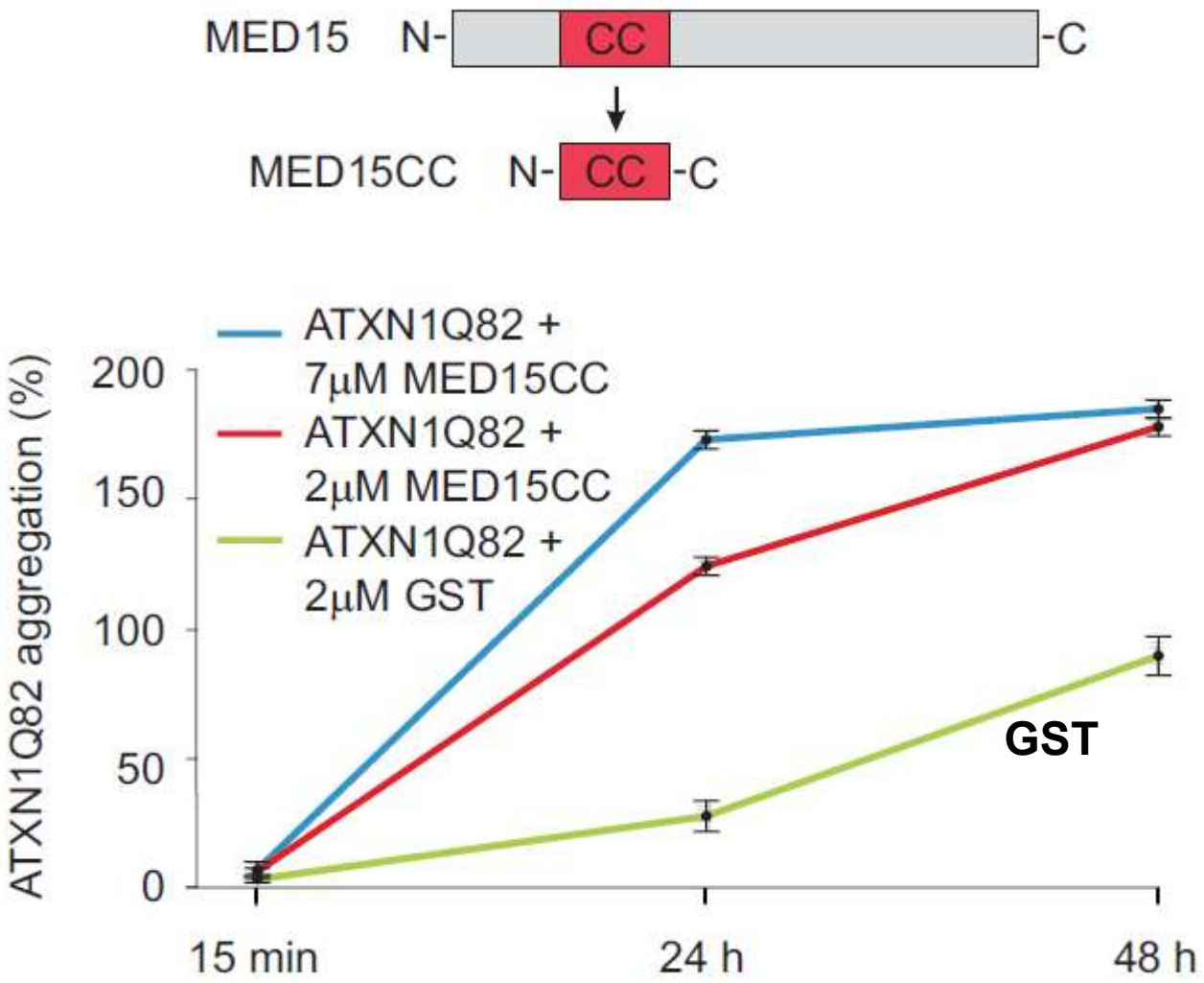
# interactors that change ATXN1Q82$^{NT}$ toxicity

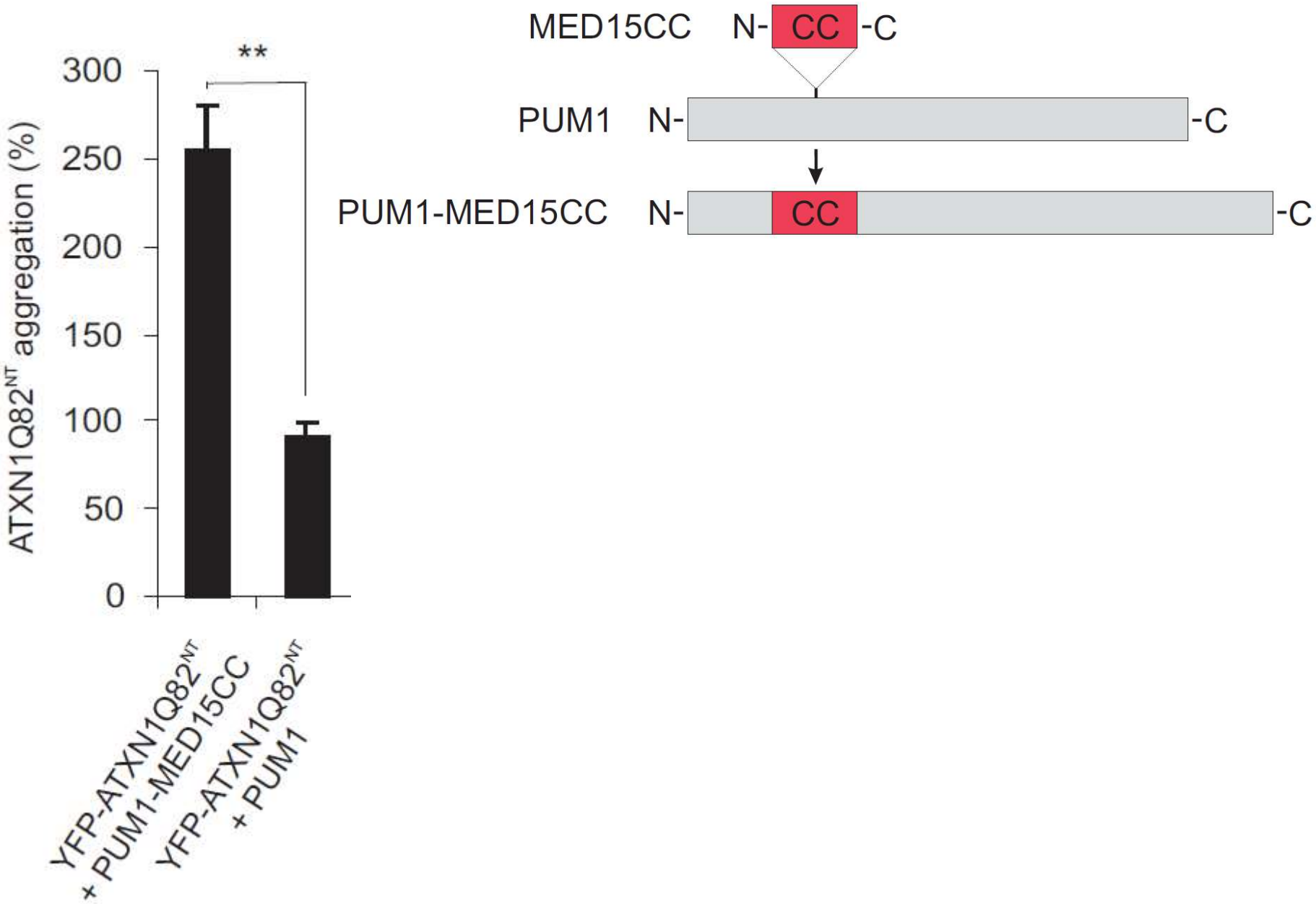


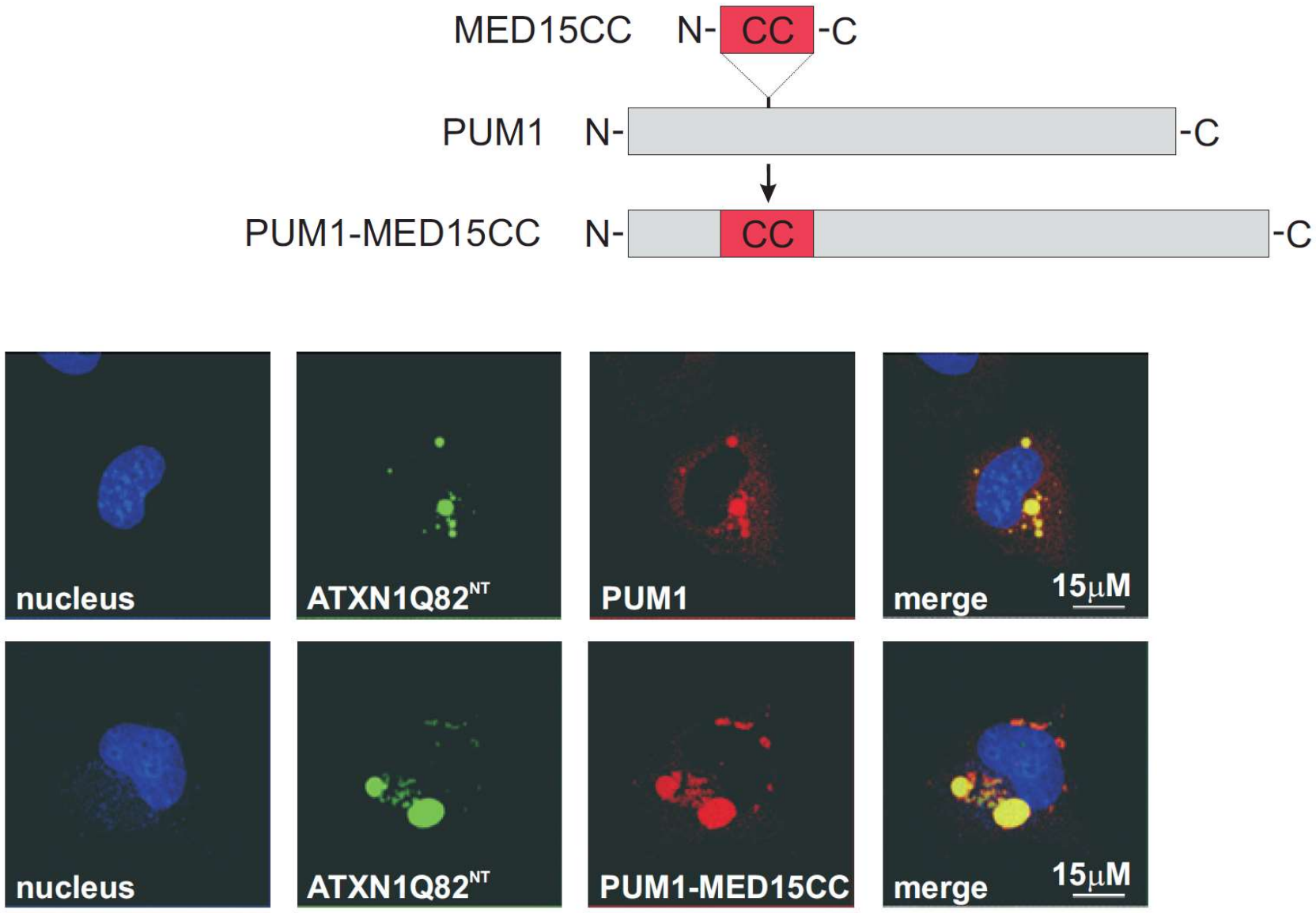

MED15
window=14
window=21
window=28
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
700
800

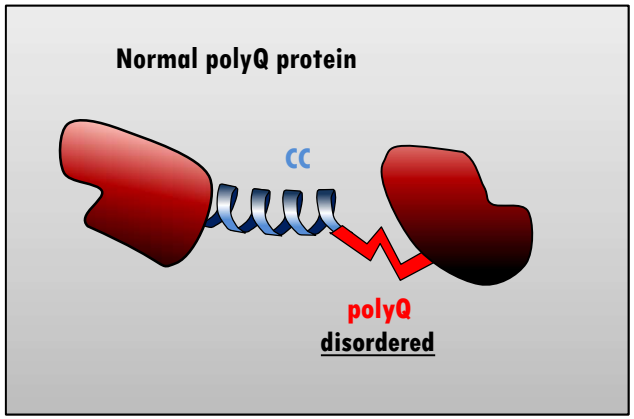

PUM1
window=14
window=21
window=28
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200

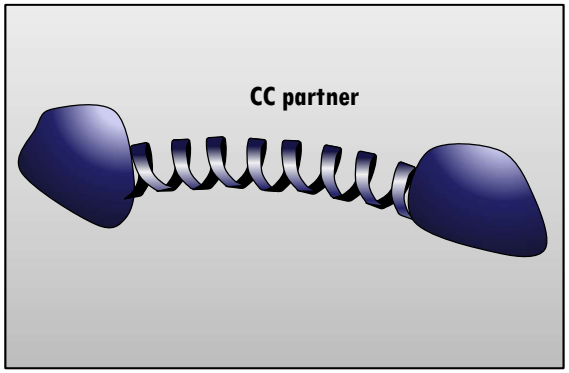

MED15
GST
PUM1
ATXN1Q82 aggregation (%)
125
100
75
50
25
0
15 min
24 h
48 h
ATXN1Q82 + MED15
ATXN1Q82 + PUM1
ATXN1Q82 + GST

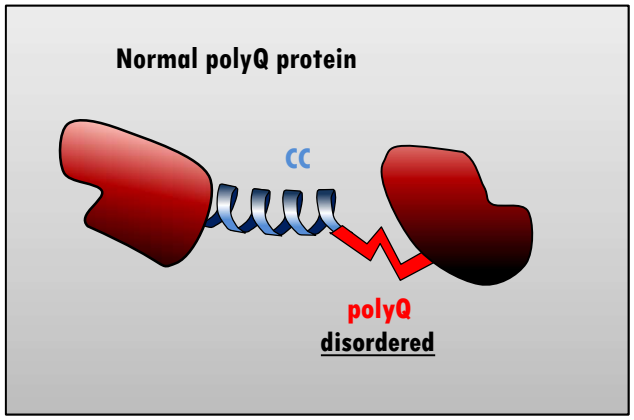

MED15
N-
CC
-C
MED15CC
N-
CC
-C
ATXN1Q82 +
7μM MED15CC
ATXN1Q82 +
2μM MED15CC
ATXN1Q82 +
2μM GST
ATXN1Q82 aggregation (%)
200
150
100
50
0
15 min
24 h
48 h
GST

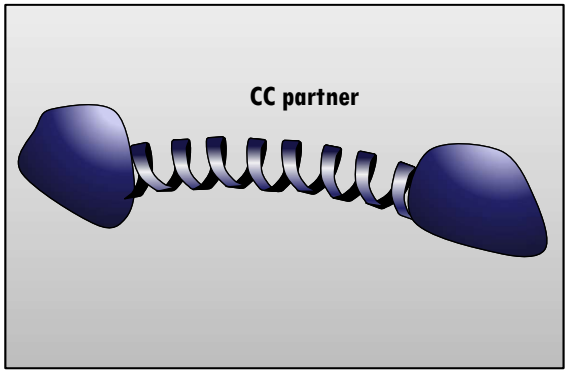

MED15CC
N-
CC
-C
PUM1
N-
-C
PUM1-MED15CC
N-
CC
-C
**
ATXN1Q82NT aggregation (%)
300
250
200
150
100
50
0
YFP-ATXN1Q82NT + PUM1-MED15CC
YFP-ATXN1Q82NT + PUM1

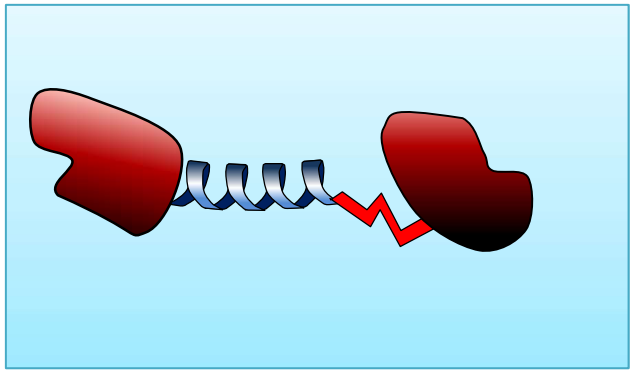

MED15CC
N-
CC
-C
PUM1
N-
-C
PUM1-MED15CC
N-
CC
-C
nucleus
ATXN1Q82NT
PUM1
merge
15μM
nucleus
ATXN1Q82NT
PUM1-MED15CC
merge
15μM

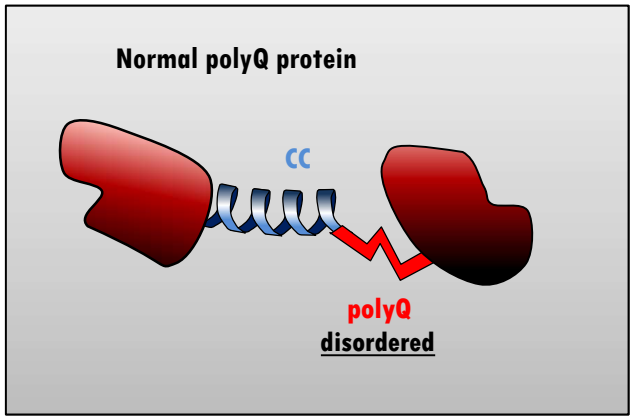

Normal polyQ protein
CC
polyQ
disordered

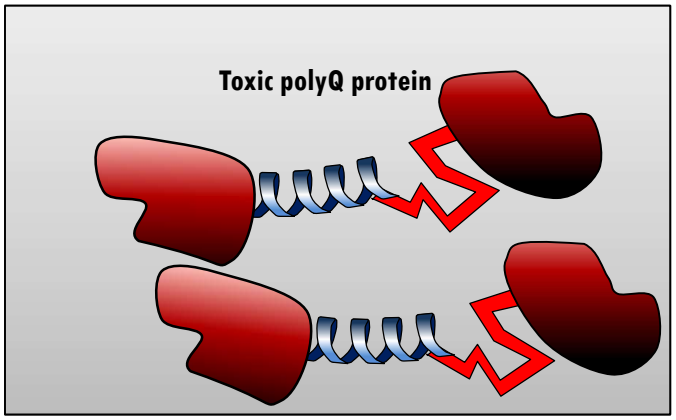

CC partner

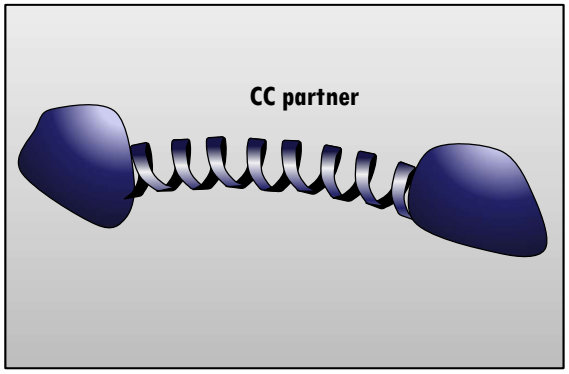

Normal polyQ protein
CC
polyQ
disordered

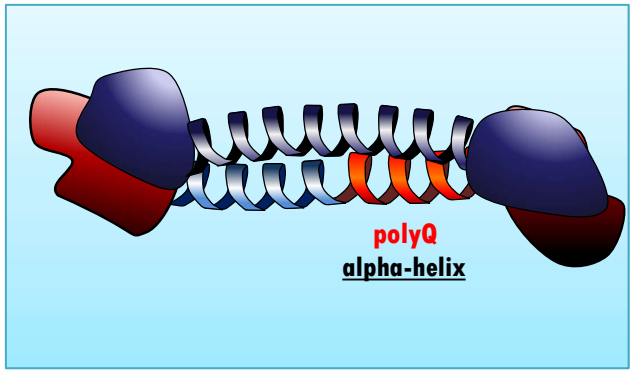

CC partner

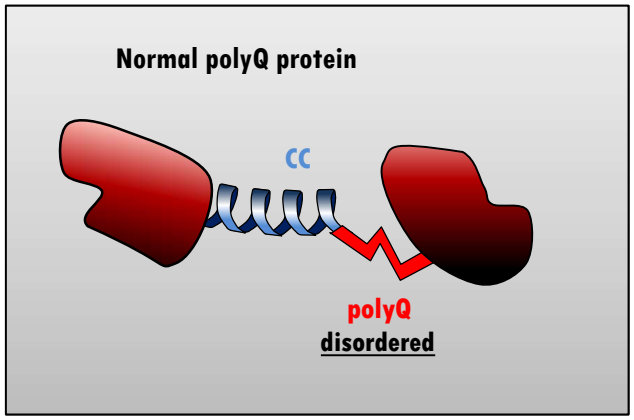


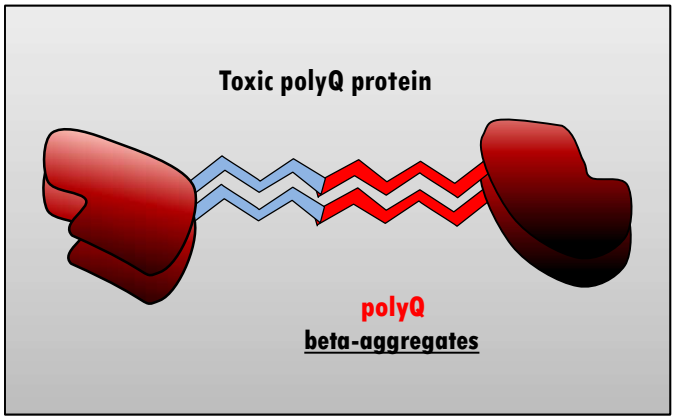

Normal polyQ protein
CC
polyQ
disordered

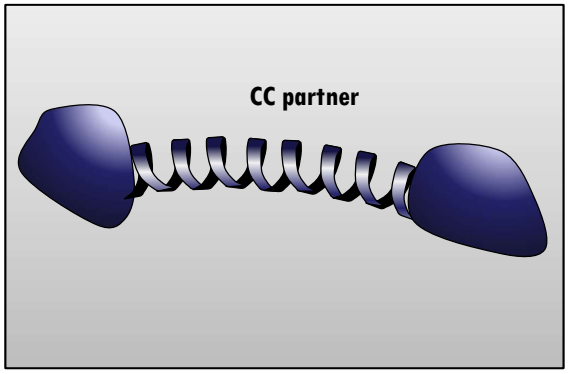

Toxic polyQ protein

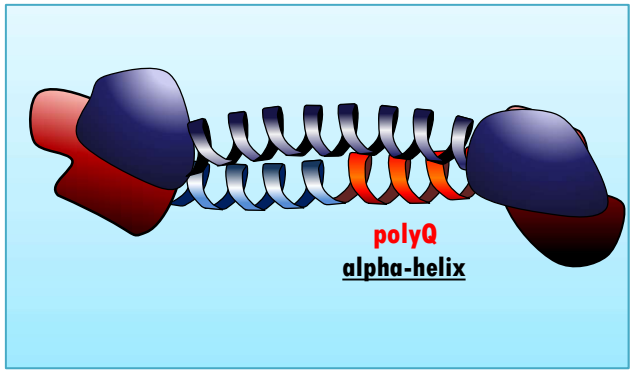

CC partner

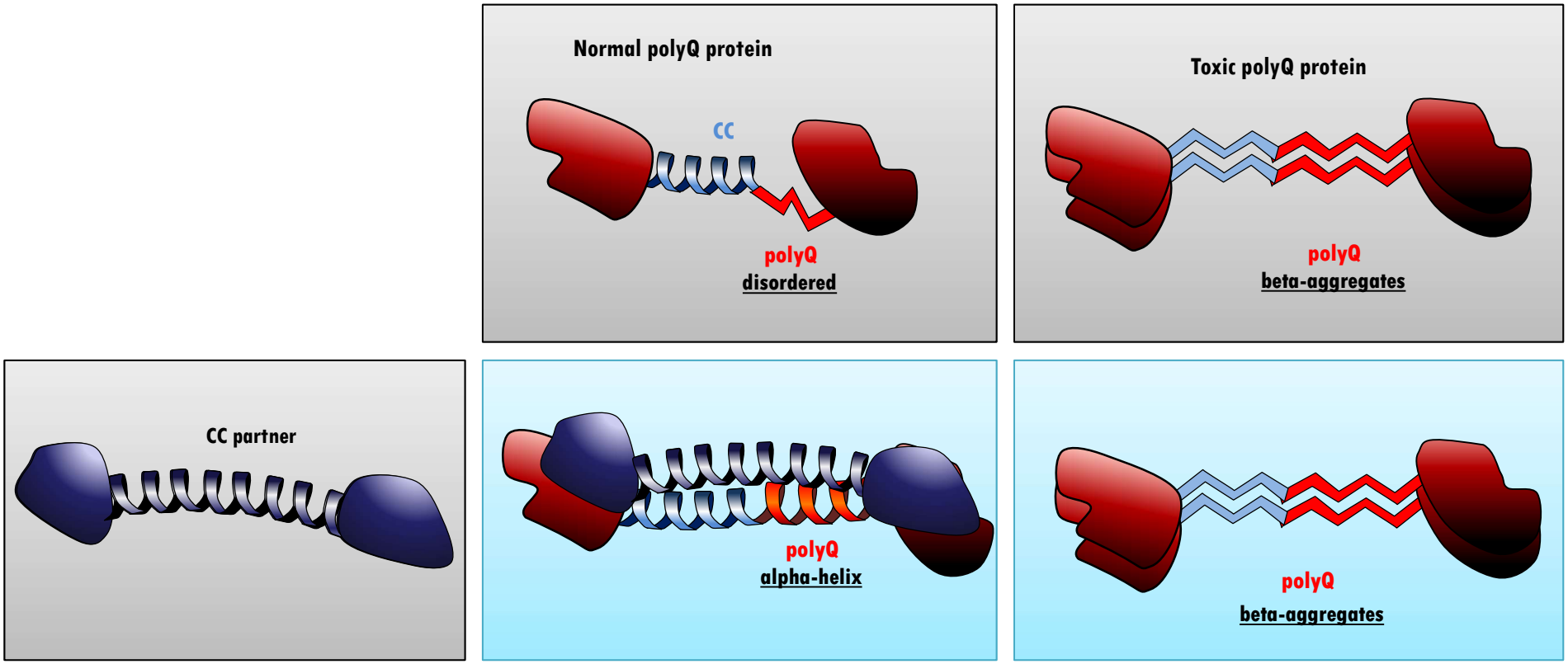

polyQ
alpha-helix

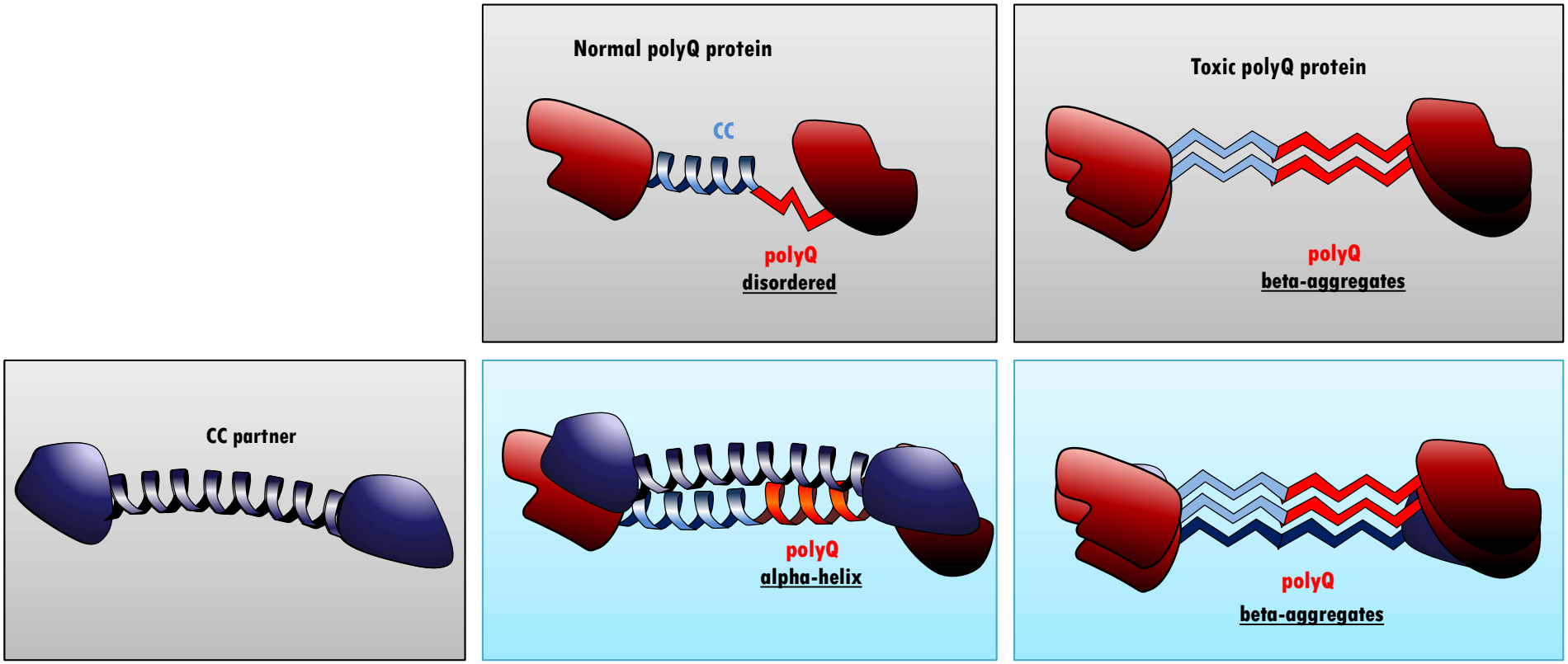

Normal polyQ protein
CC
polyQ
disordered

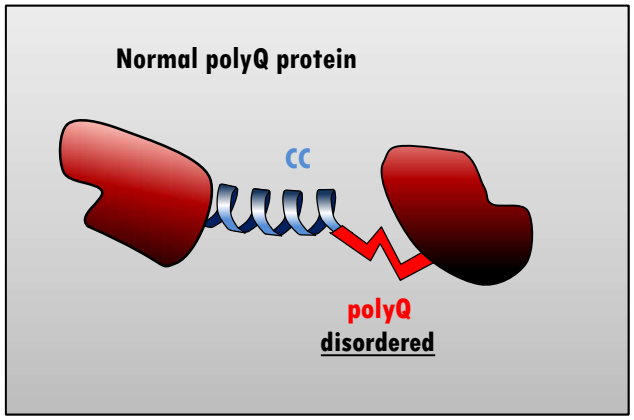

Toxic polyQ protein
polyQ
beta-aggregates

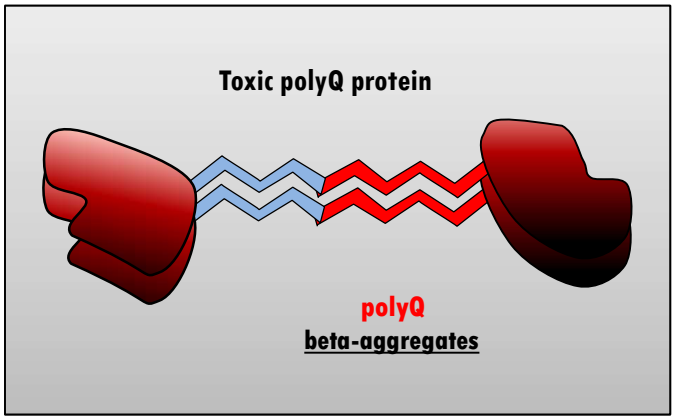

CC partner

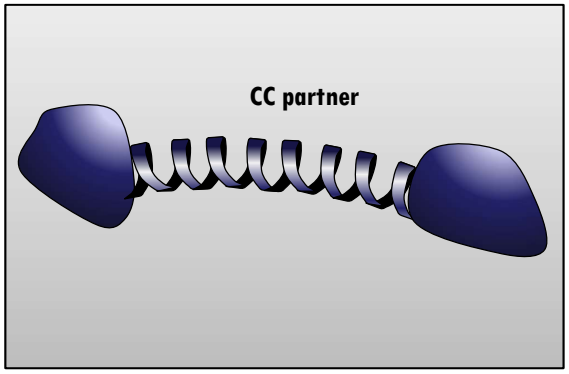

polyQ
alpha-helix

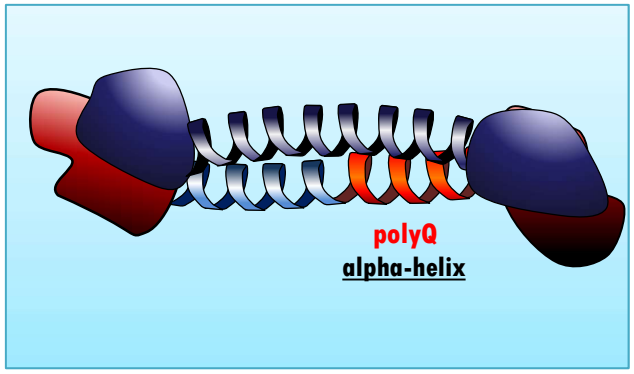

Normal polyQ protein
CC
polyQ
disordered
Toxic polyQ protein
polyQ
beta-aggregates
CC partner
polyQ
alpha-helix
polyQ
beta-aggregates

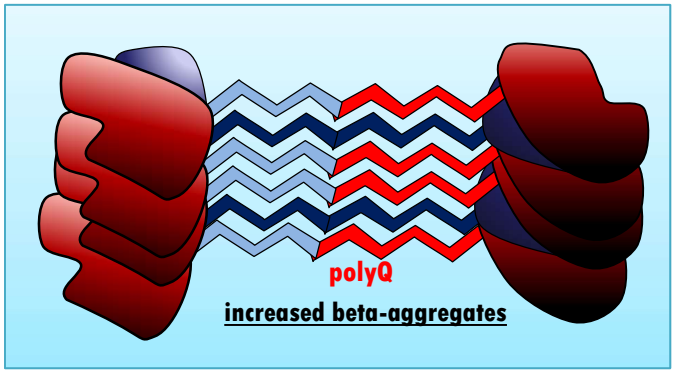

Normal polyQ protein
CC
polyQ
disordered
Toxic polyQ protein
polyQ
beta-aggregates
CC partner
polyQ
alpha-helix
polyQ
beta-aggregates

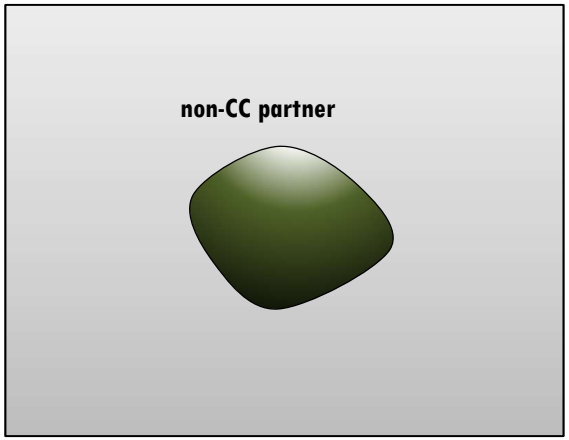

Normal polyQ protein
CC
polyQ
disordered

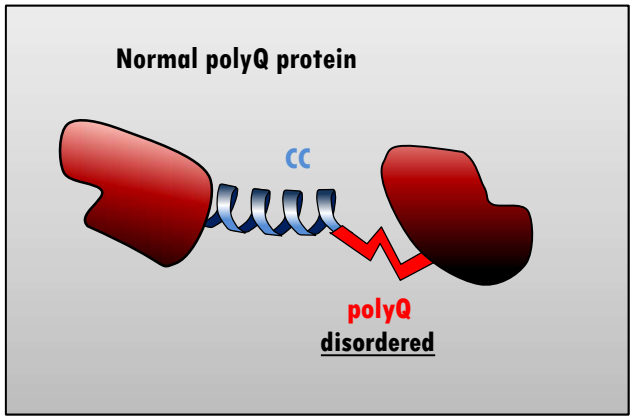

Toxic polyQ protein
polyQ
beta-aggregates

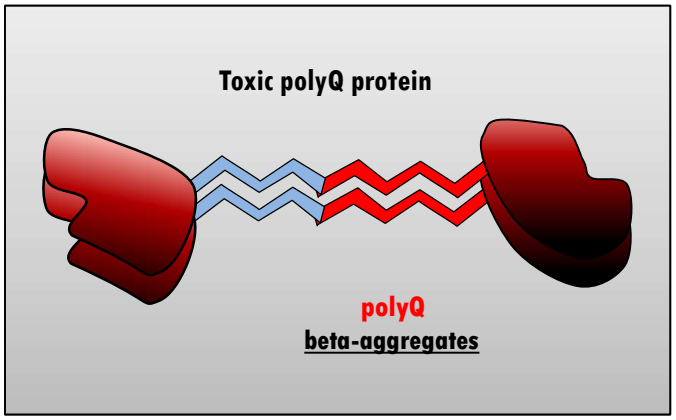

CC partner

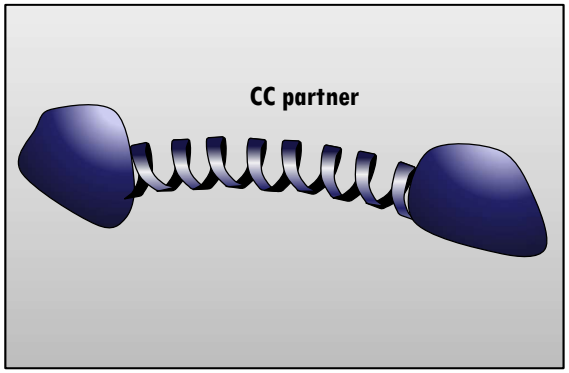

polyQ
alpha-helix

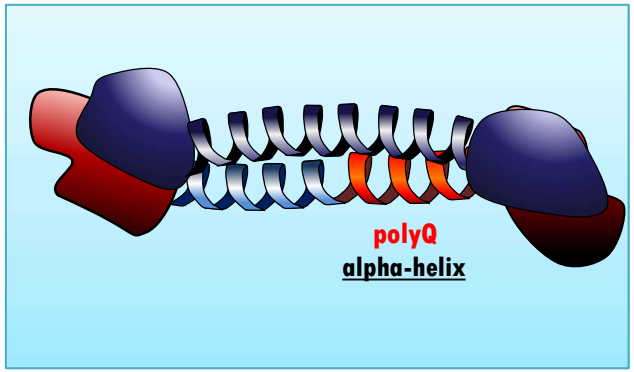

polyQ
increased beta-aggregates

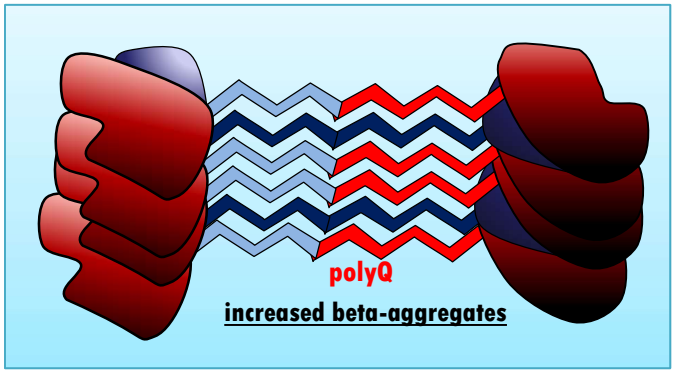

non-CC partner

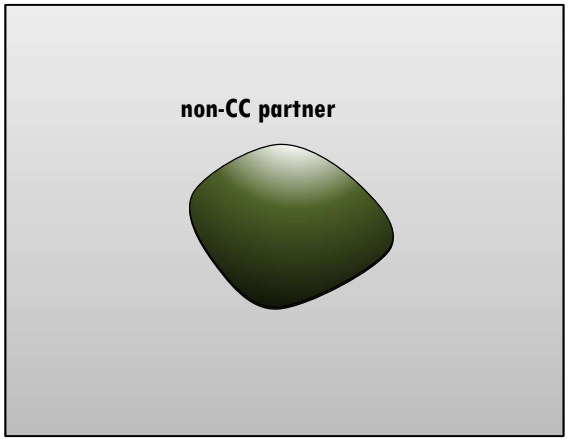

Normal polyQ protein
CC
polyQ
disordered

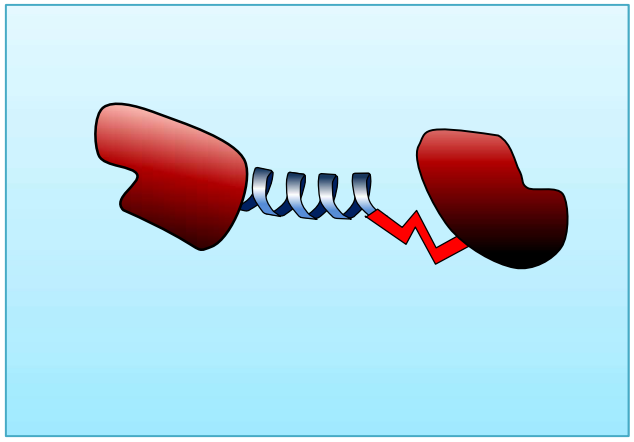

Toxic polyQ protein
polyQ
beta-aggregates

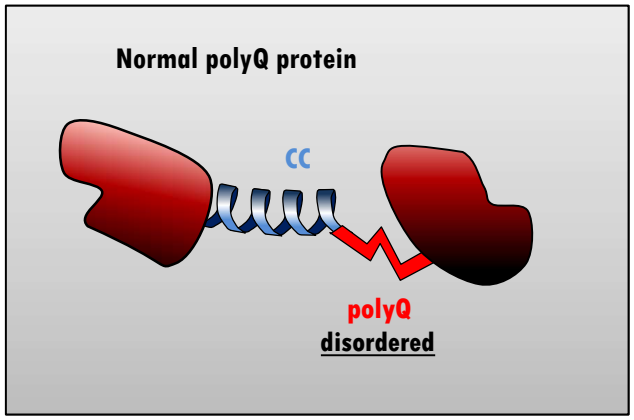

CC partner

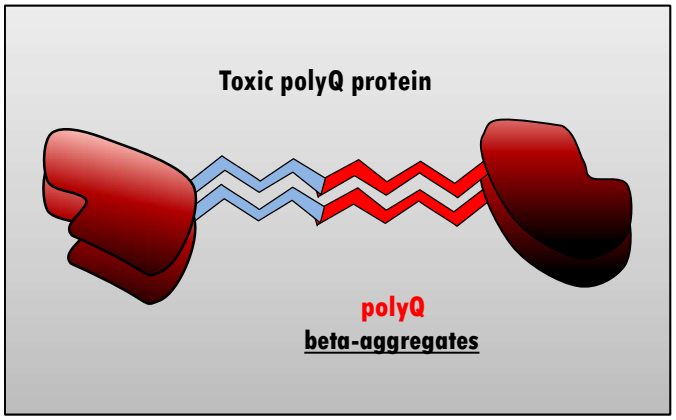

polyQ
alpha-helix

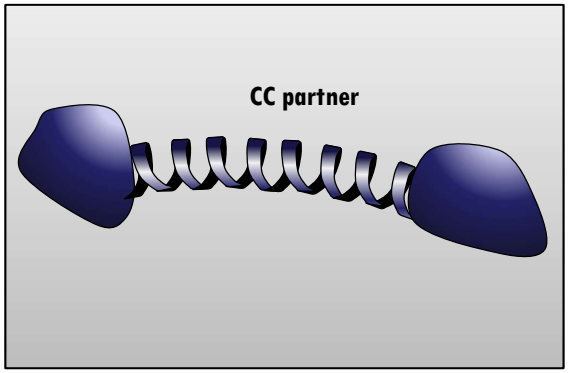

polyQ
increased beta-aggregates

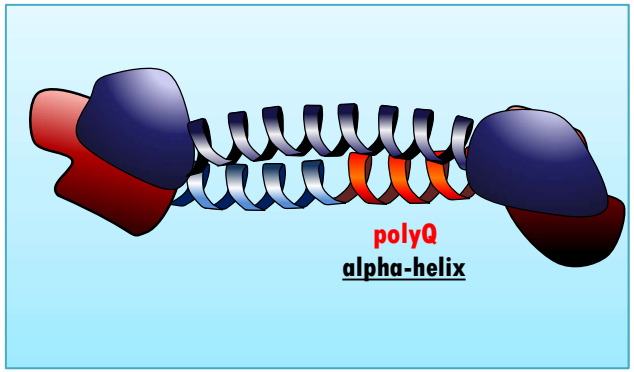

non-CC partner

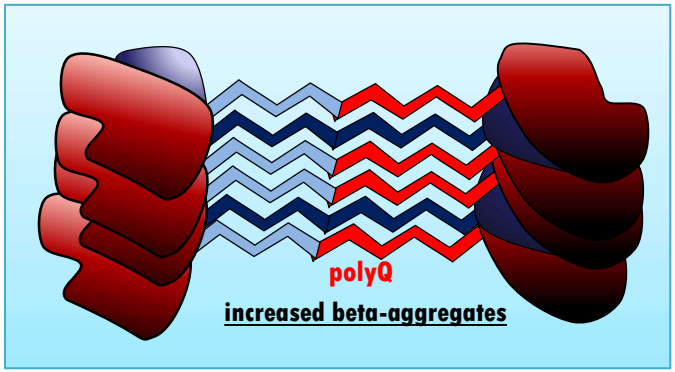


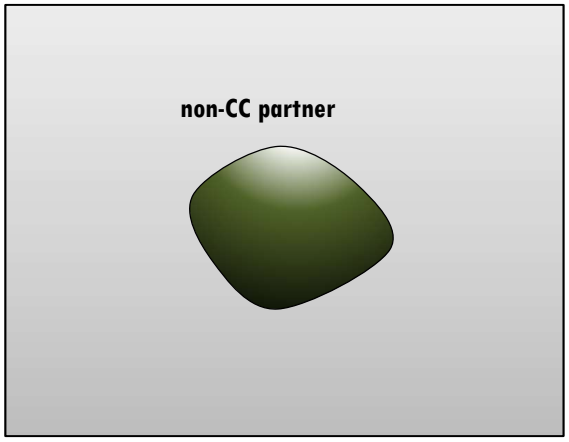

Normal polyQ protein
CC
polyQ
disordered

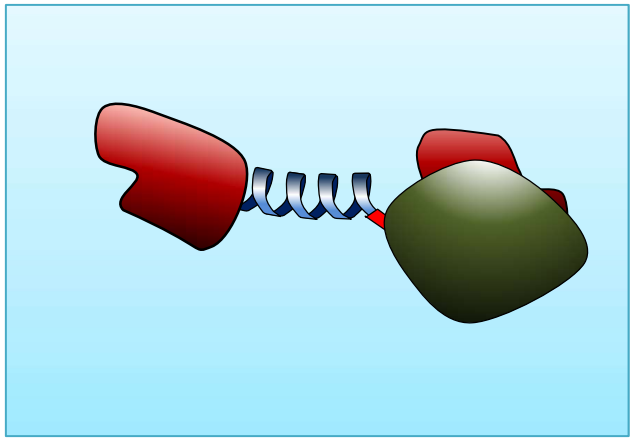

Toxic polyQ protein
polyQ
beta-aggregates

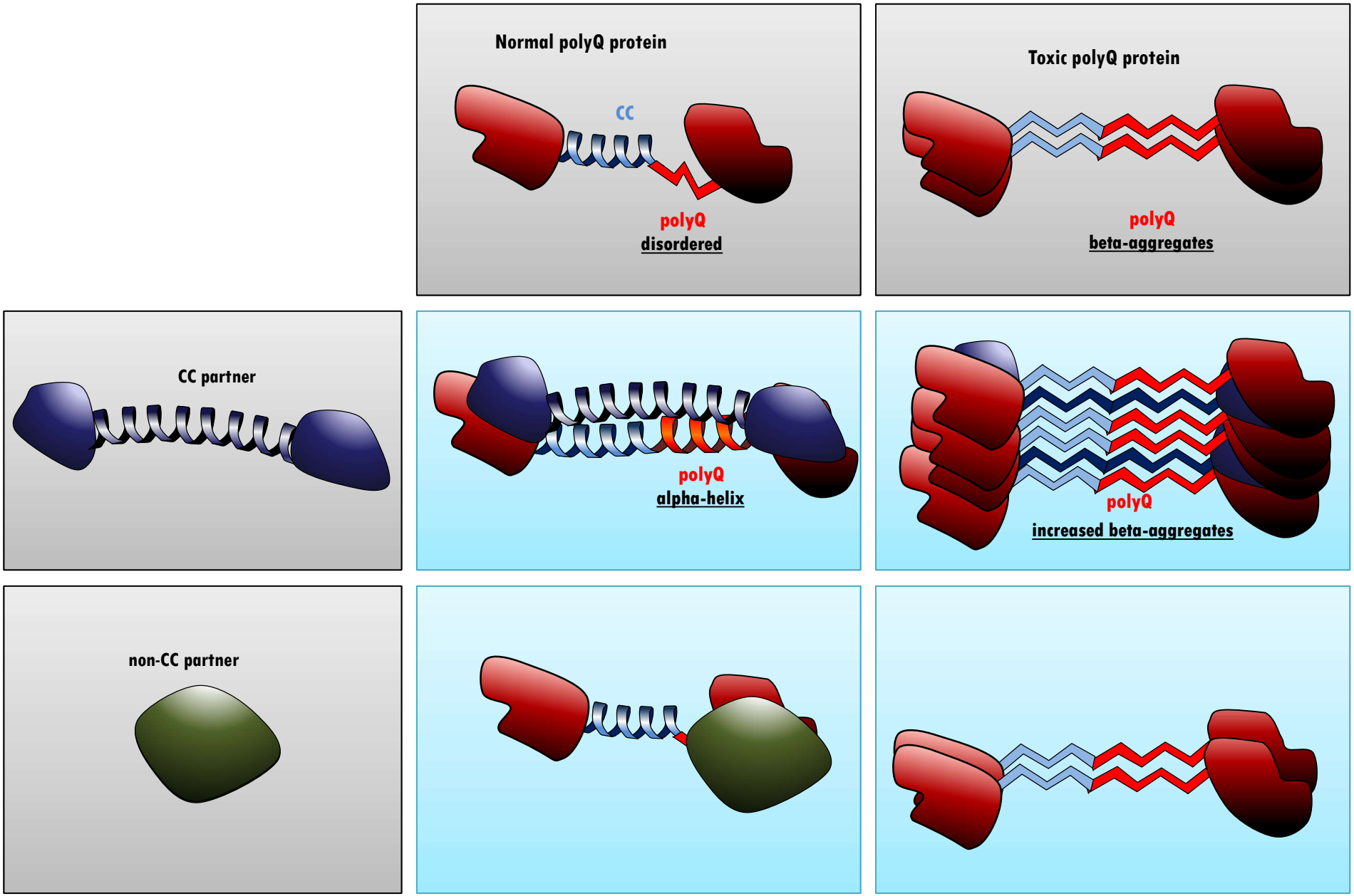

CC partner

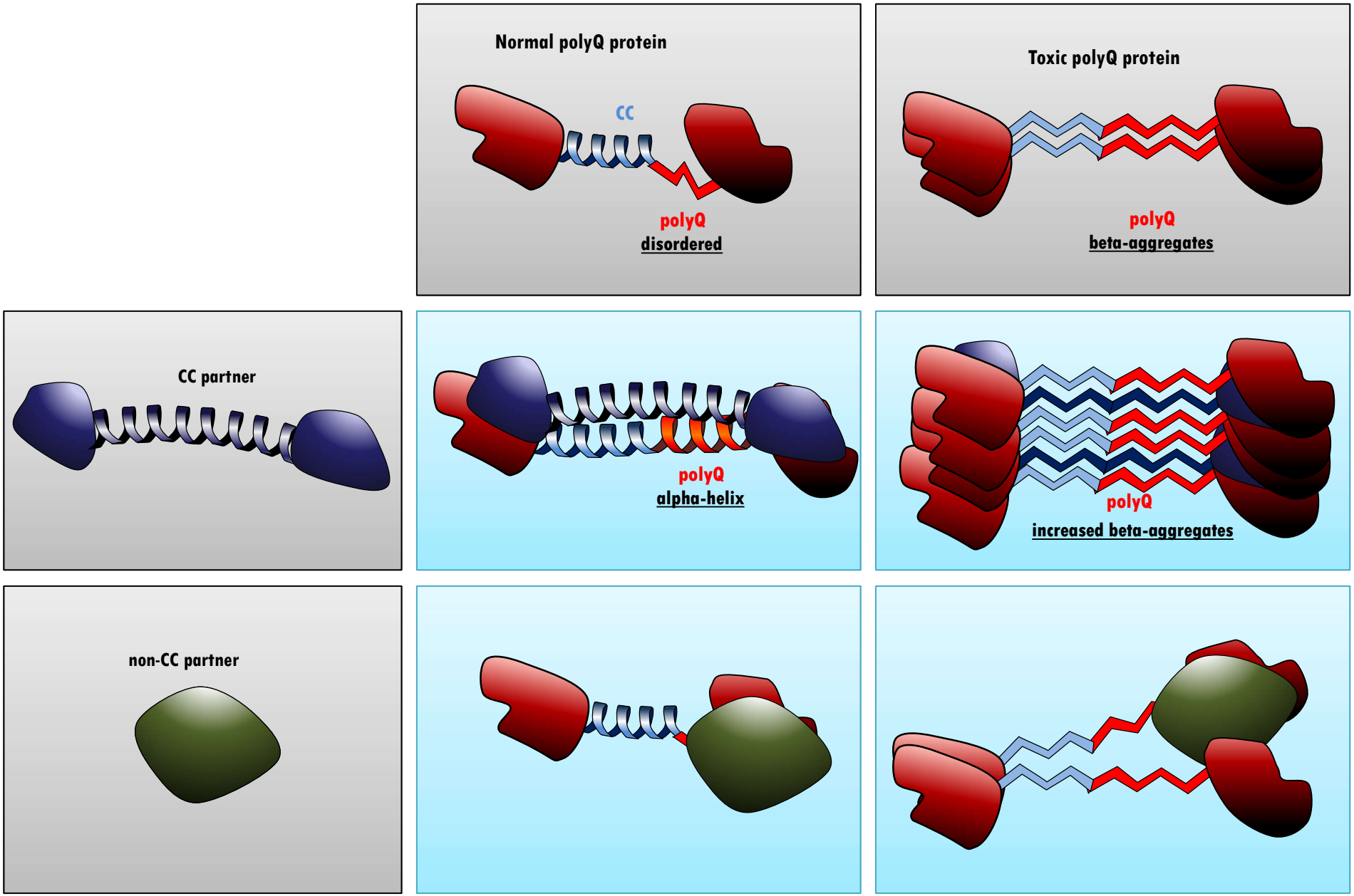

polyQ
alpha-helix

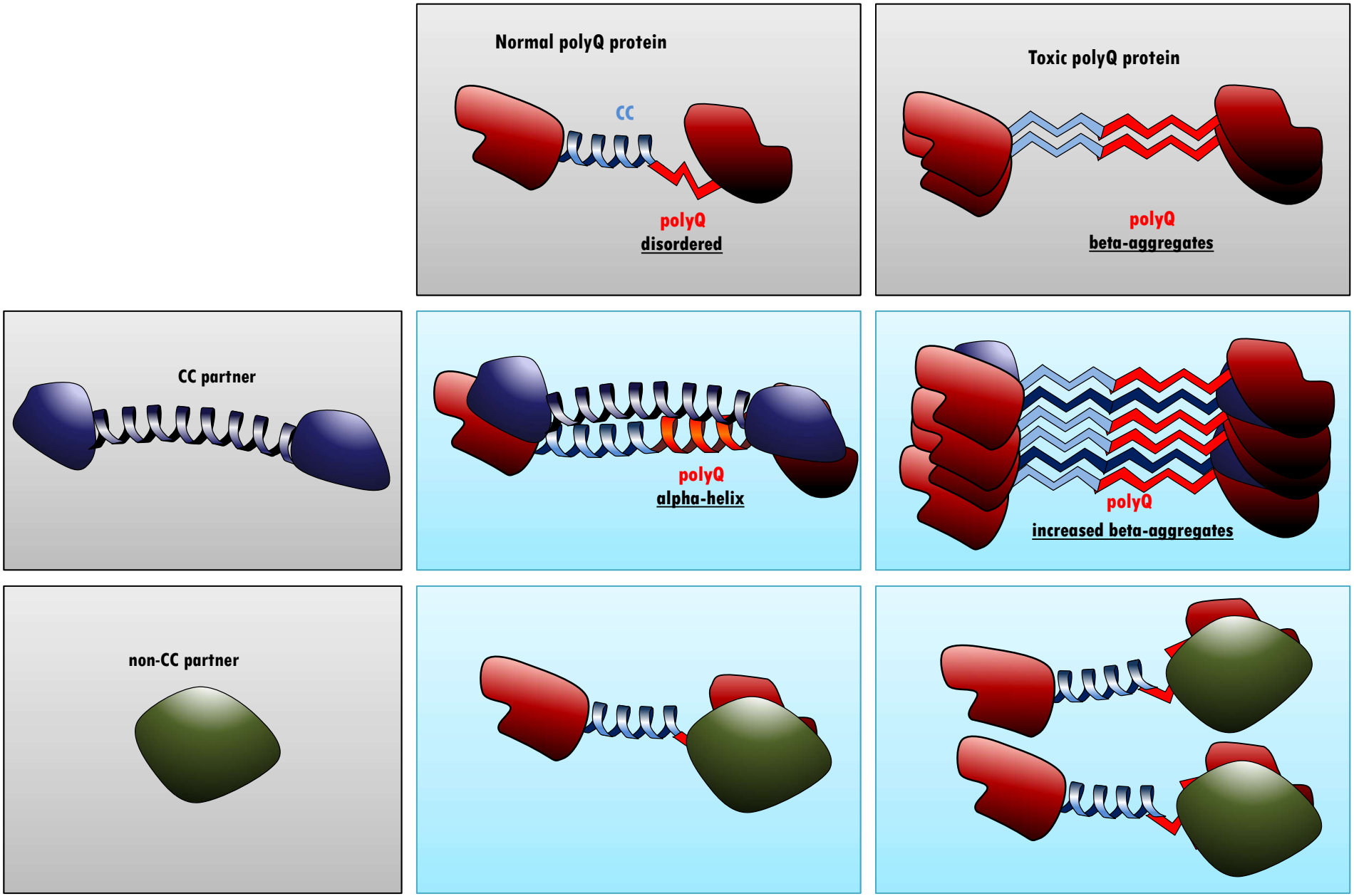

polyQ
increased beta-aggregates


non-CC partner




Normal polyQ protein
CC
polyQ
disordered
Toxic polyQ protein
polyQ
beta-aggregates
CC partner
polyQ
alpha-helix
polyQ
increased beta-aggregates
non-CC partner


Normal polyQ protein
CC
polyQ
disordered
Toxic polyQ protein
polyQ
beta-aggregates
CC partner
polyQ
alpha-helix
polyQ
increased beta-aggregates
non-CC partner


Normal polyQ protein
CC
polyQ
disordered
Toxic polyQ protein
polyQ
beta-aggregates
CC partner
polyQ
alpha-helix
polyQ
increased beta-aggregates
non-CC partner

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

- Open in jalview the alignment of the mineralocorticoid receptor: MR1_fasta.txt

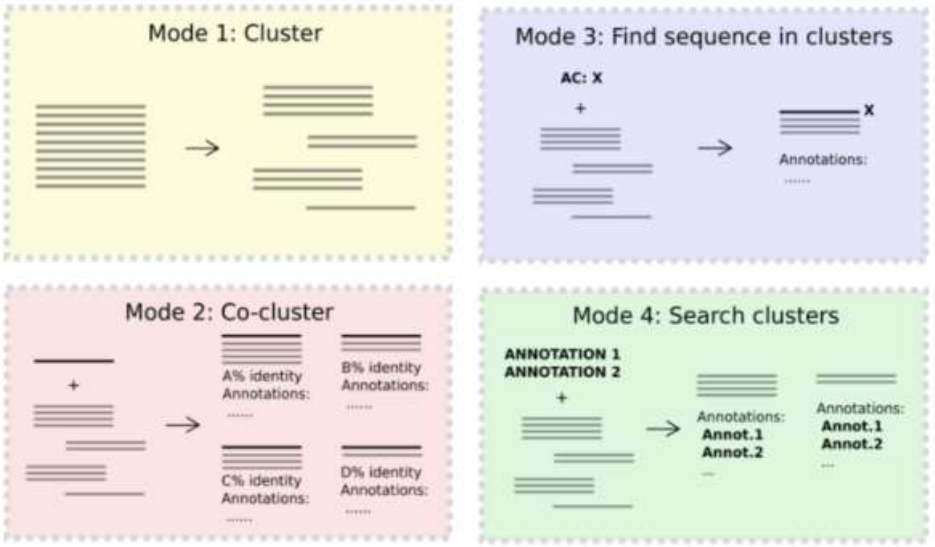
- Find a polyQ insertion.

Do you see any other biased region nearby?

# Clustering proteins

## FastaHerder2

Pablo Mier




Mier and Andrade-Navarro (2016) *J. Comp. Biol.*

# Clustering proteins

Pablo Mier

Results overview

Search settings

| The cluster MUST have... | The cluster MUST NOT have... |
|---|---|
| Organism/s = escherichia | - |
| PolyQ | |

Number of clusters found 2

Download link file .txt 718580686328099.txt

**Click on the leader to display its annotations**

Leader: (1:)sp|A0K4S8|DNAK_BURCH
Leader: (1:)sp|Q83S00|FTSK_SHIFL

Time elapsed: 3 seconds

**polyQ**

...polyS regions? DM ...polyQ regions? DM
...polyG regions? DM ...polyA regions? DM
...polyL regions? DM ...polyM regions? DM
...polyW regions? DM ...polyY regions? DM

es separated by "+")

separated by "+")

In the cluster, ...

***Escherichia***

there **MUST** be at least one sequence from the following organism/s: (taxonomic id from an organism, e.g. **9606** for *H.sapiens*, or taxon name, e.g. ***Homo***)*

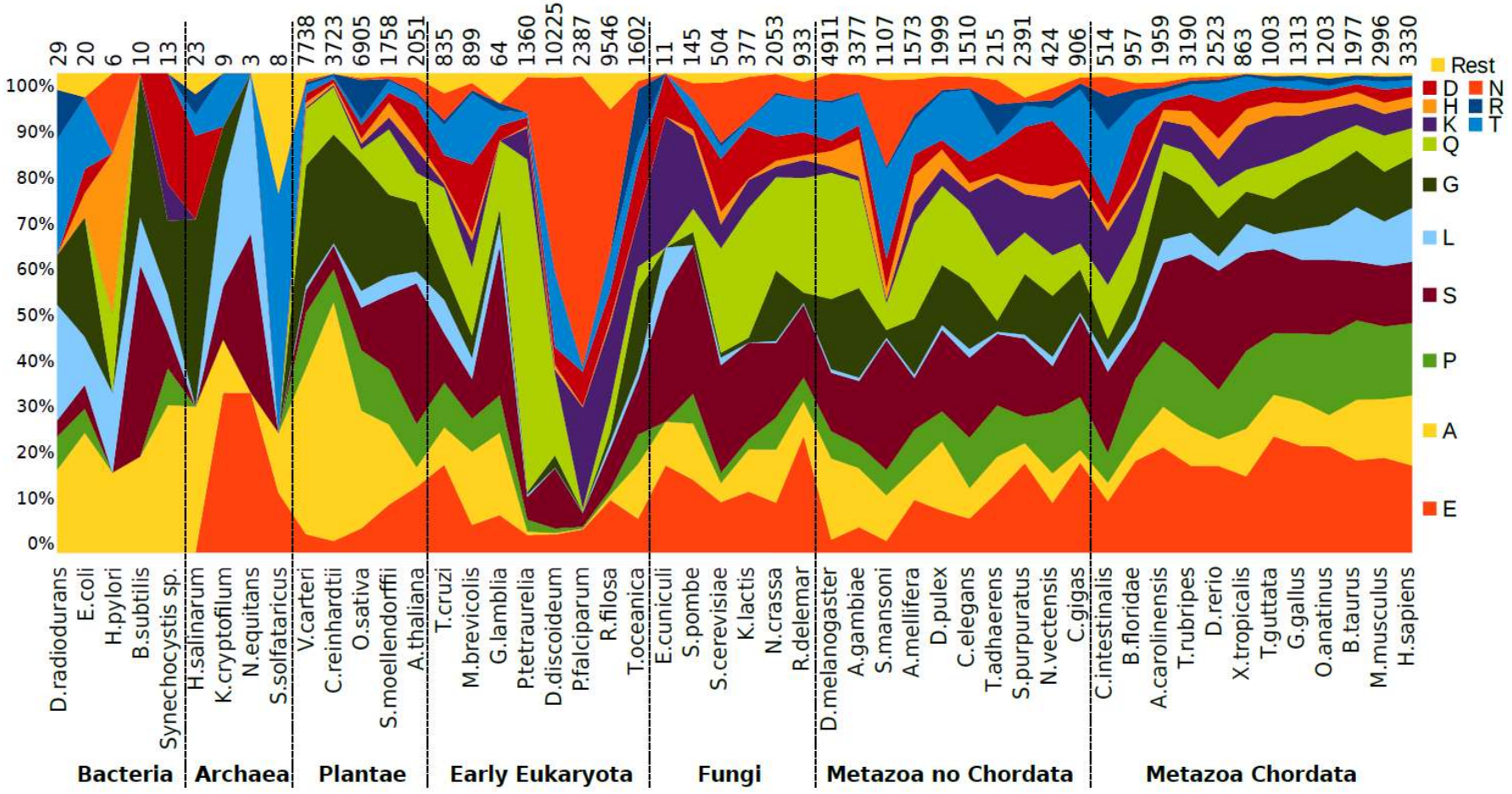
there **MUST NOT** be any sequence from the following organism/s: (taxonomic id from an organism, e.g. **9606** for *H.sapiens*, or taxon name, e.g. ***Homo***)*

**if more than one, separate them by "+", e.g. **9606+Homo***

SUBMIT GO MODE 4! | What's this? | e.g. example 1, example 2

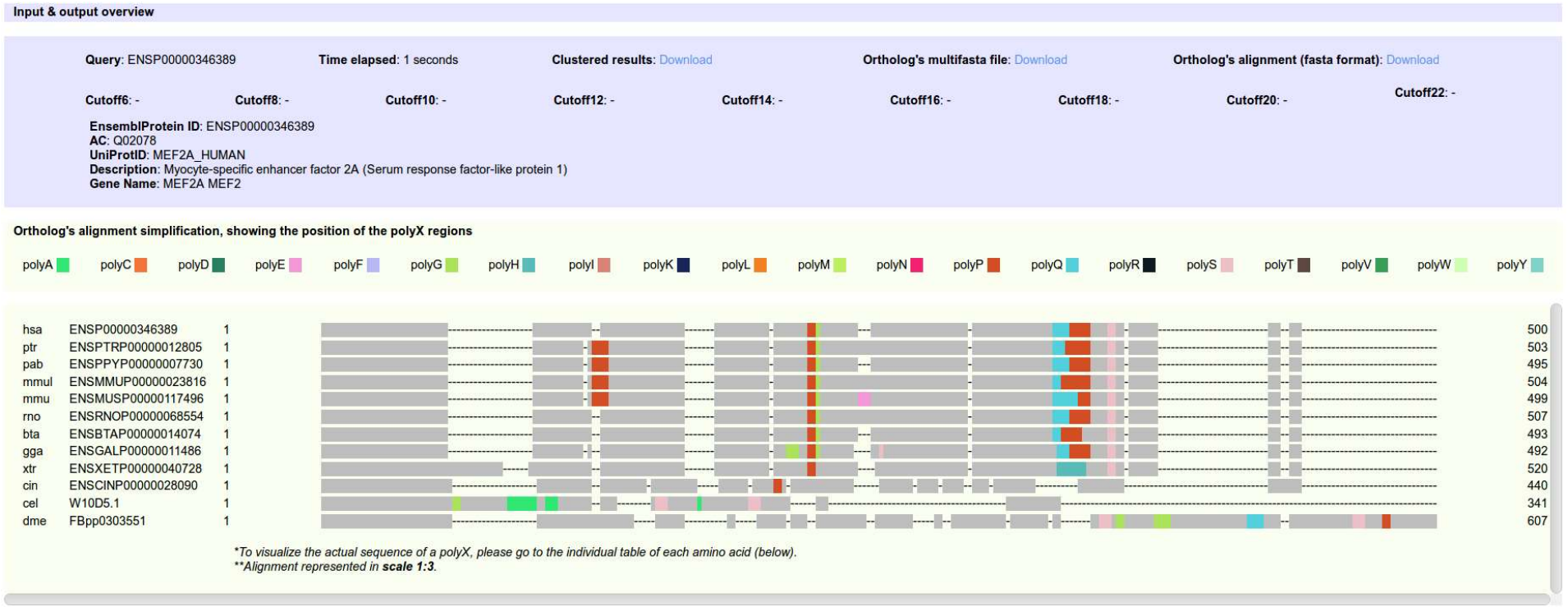
Mier and Andrade-Navarro (2016) *J. Comp. Biol.*

# Frequency of homorepeats in 50 species




Mier *et al.* (2017) Proteins

# Context and evolution of homorepeats

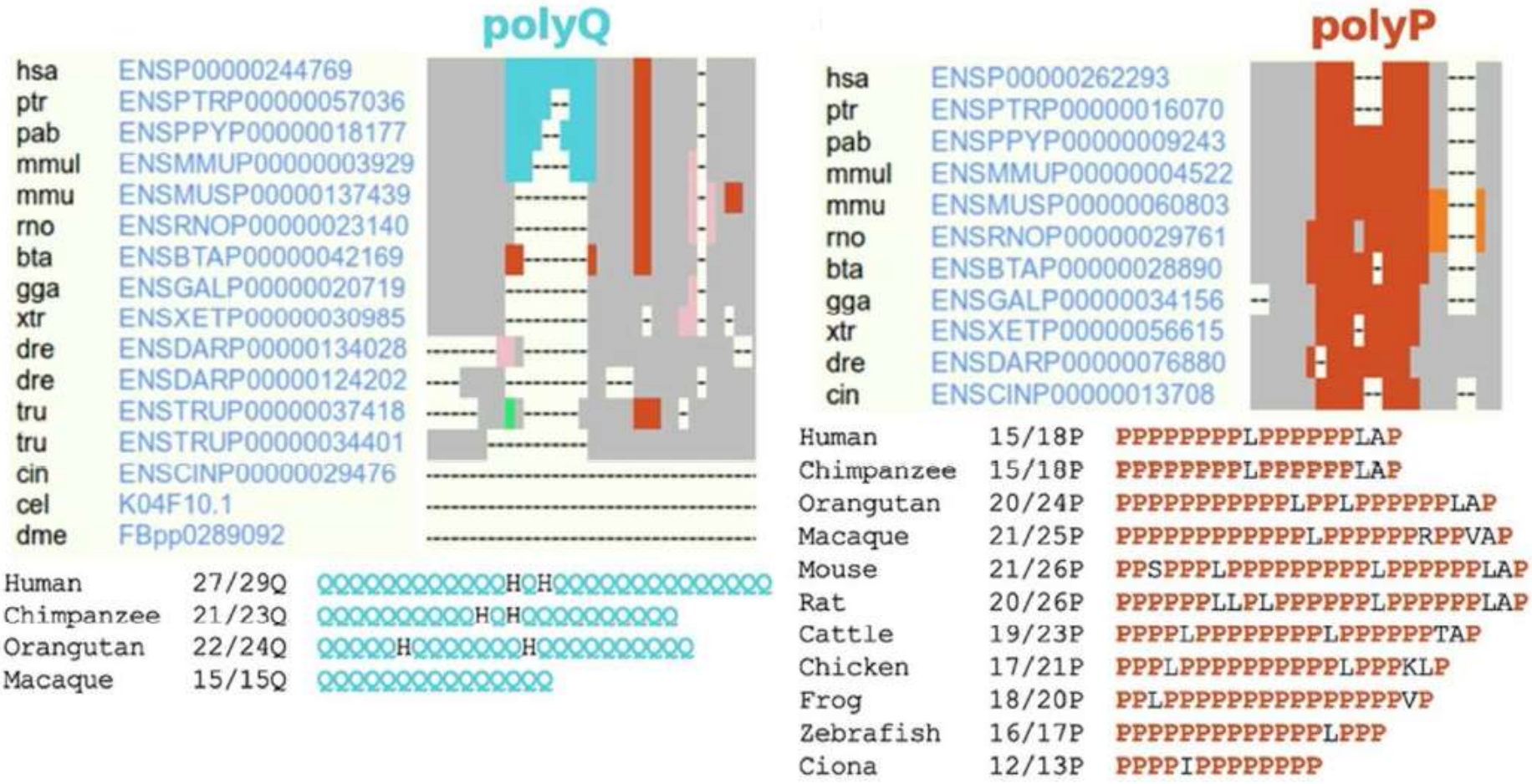




**dA**tabase of **P**olyX **E**volution

Mier *et al.* (2016) Bioinformatics

# Context and evolution of homorepeats





**dA**tabase of **P**olyX **E**volution

Mier *et al.* (2016) Bioinformatics