



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
LSSMKSSISSLPPSHCSVKSPVSSPNNVTLRSSVSSPANINN

Definition repeats

Sequence, long, imperfect, tandem

MRAVVK**SP**IMCHEKSPSVC**SP**LNMTSSVC**SP**AGINSVSSTTASF
GSFPVH**SP**ITQGTPLTC**SP**NVENRGSRSH**SP**AHASNVGSPLS**SP**
LSSMKSSIS**SP**PSHCSVKSPV**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**VKSP**IM CHE

KSPSVC**SPLN**

MT**SSVCSP**AG INSVSSTTASF

GSFP**VHSP**IT Q

GTP LTC**SPNV** EN

RG**SRSHSPAH** ASN

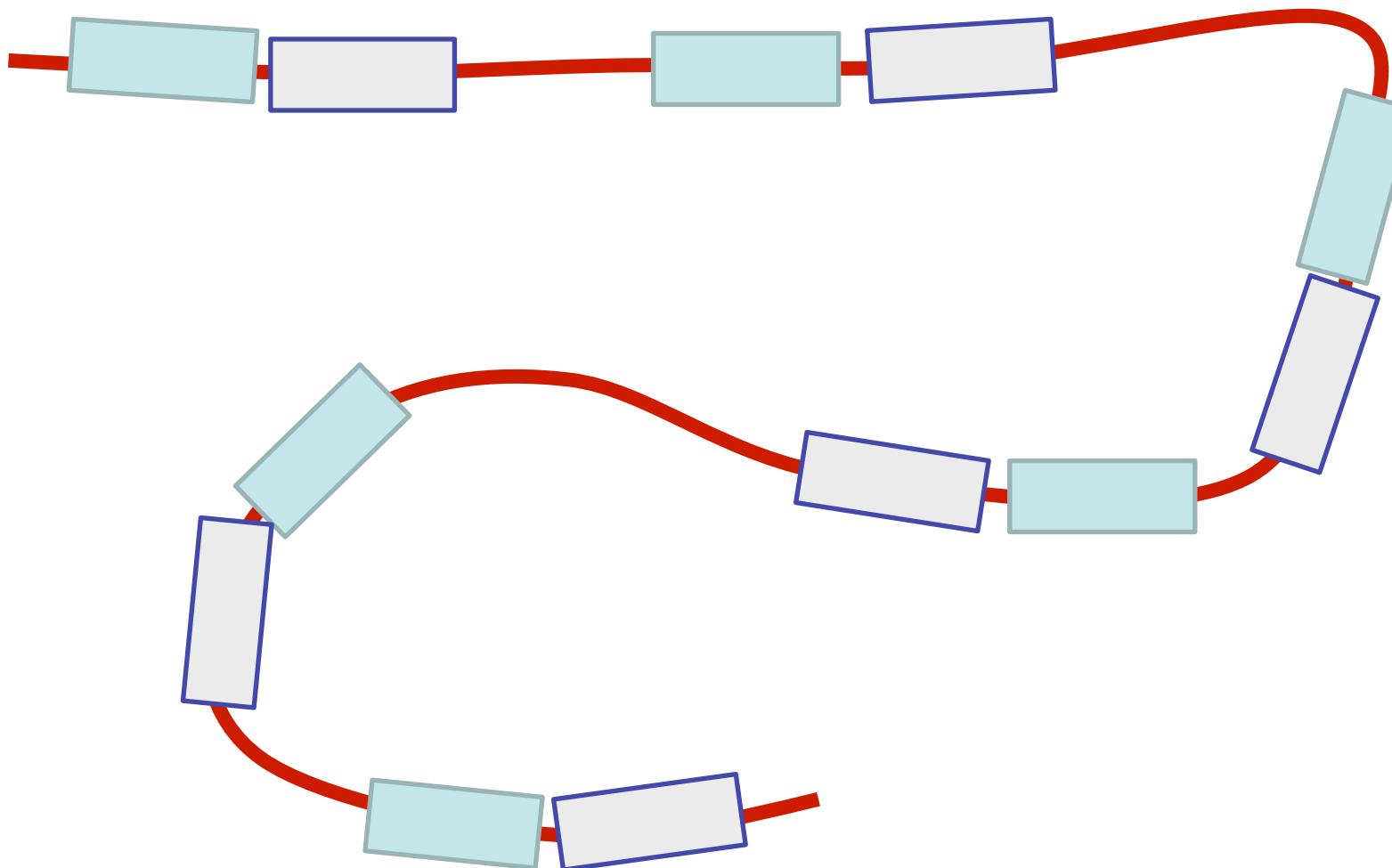
VG**SPLSSP**LS S

MK**SSISSP**PS HCS

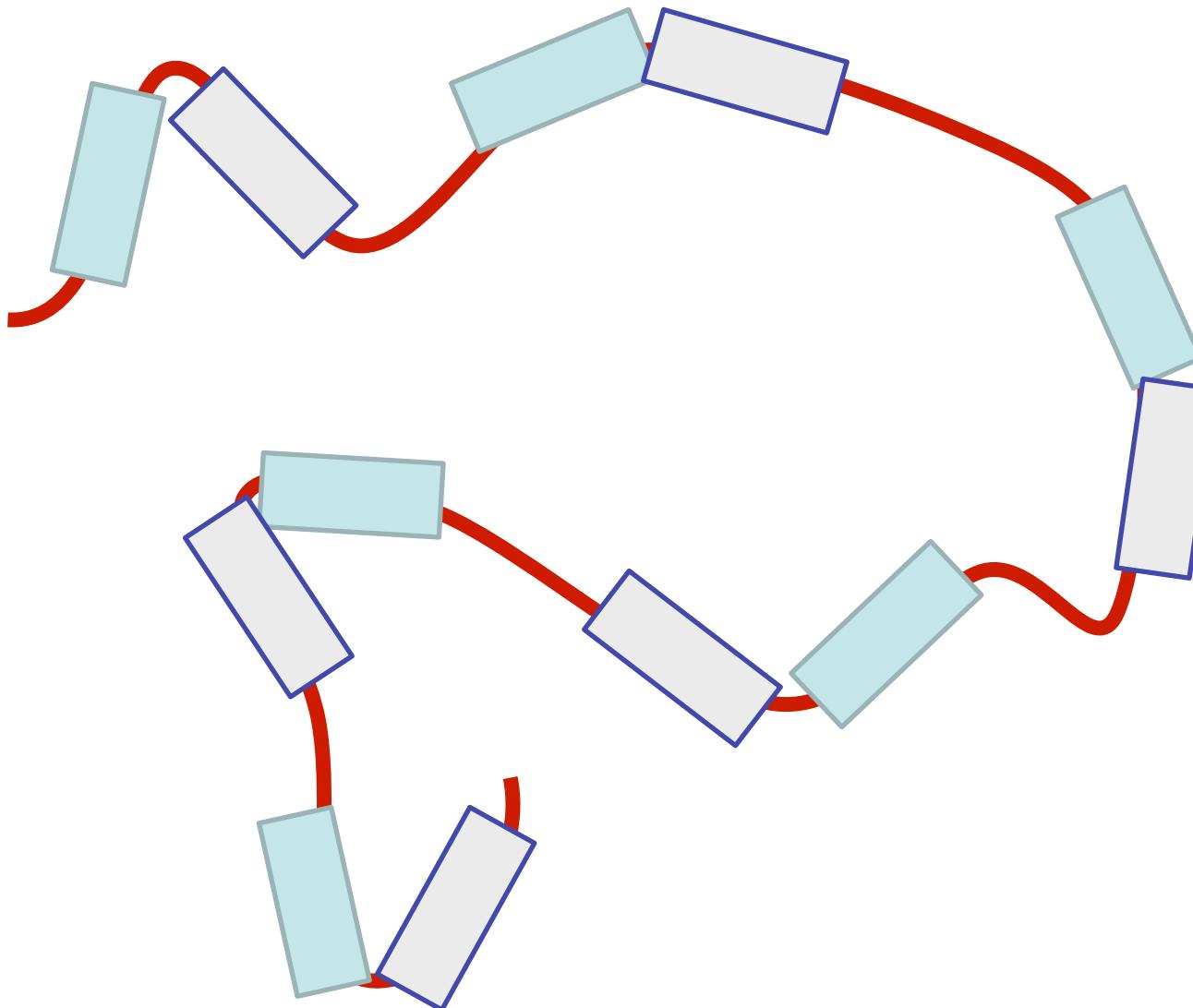
VK**SPVSSP**NN VT

LR**SSVSSP**AN INN

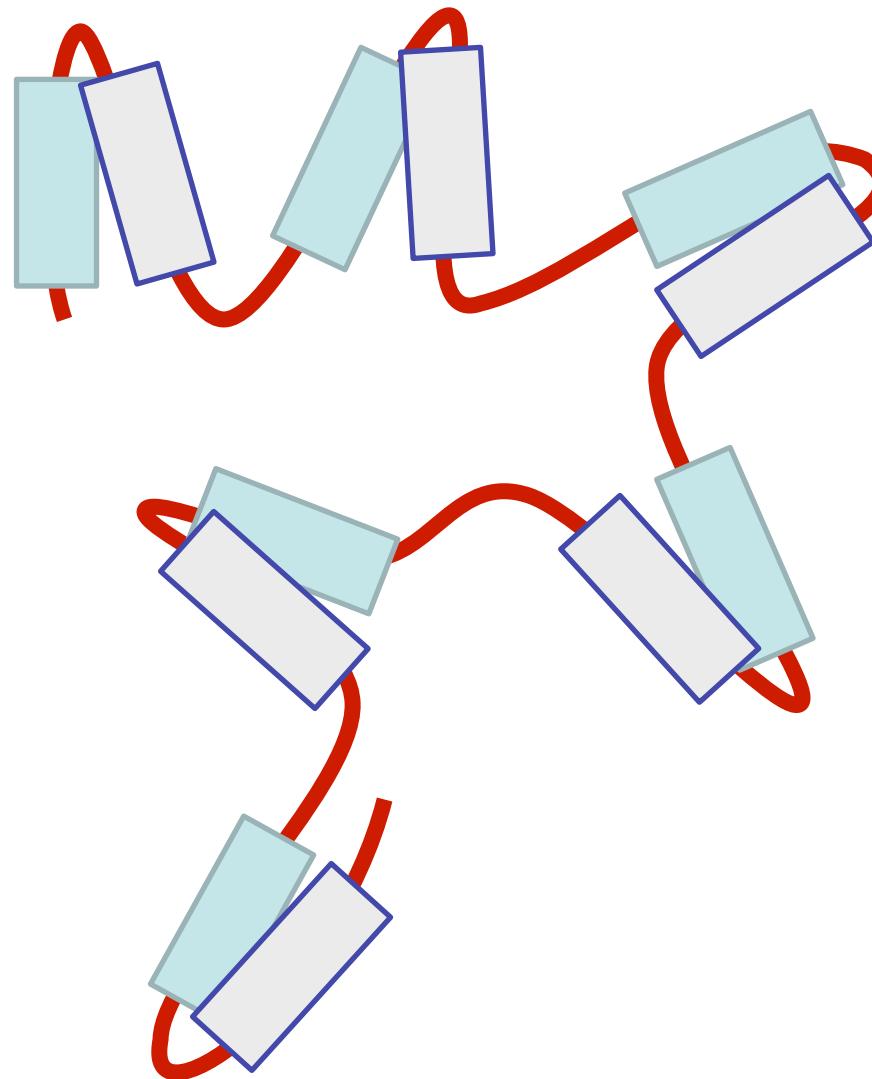
Tandem repeats fold together



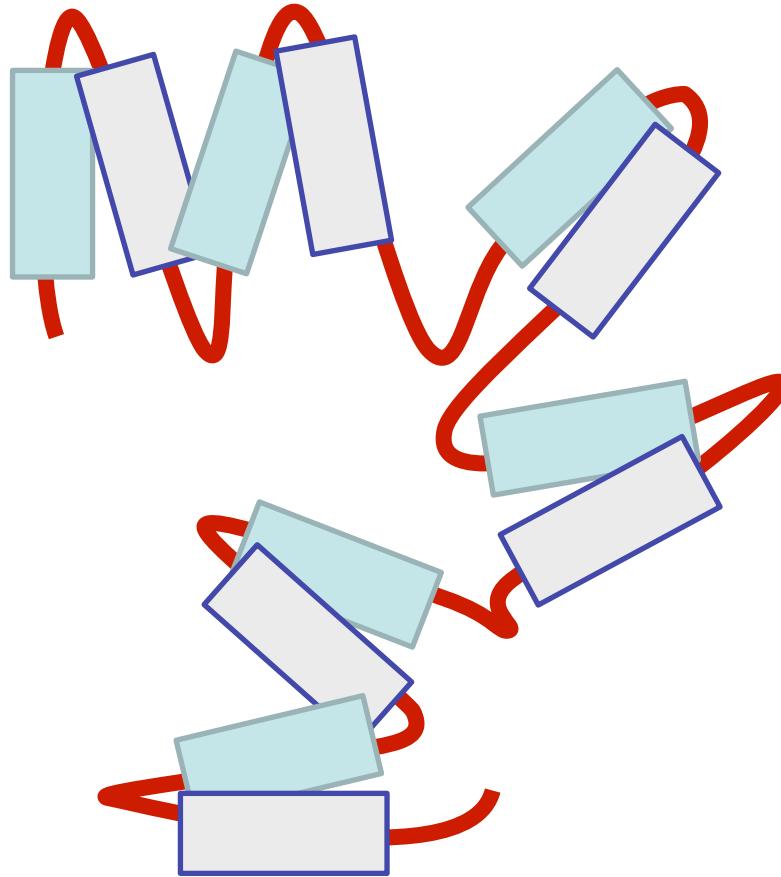
Tandem repeats fold together



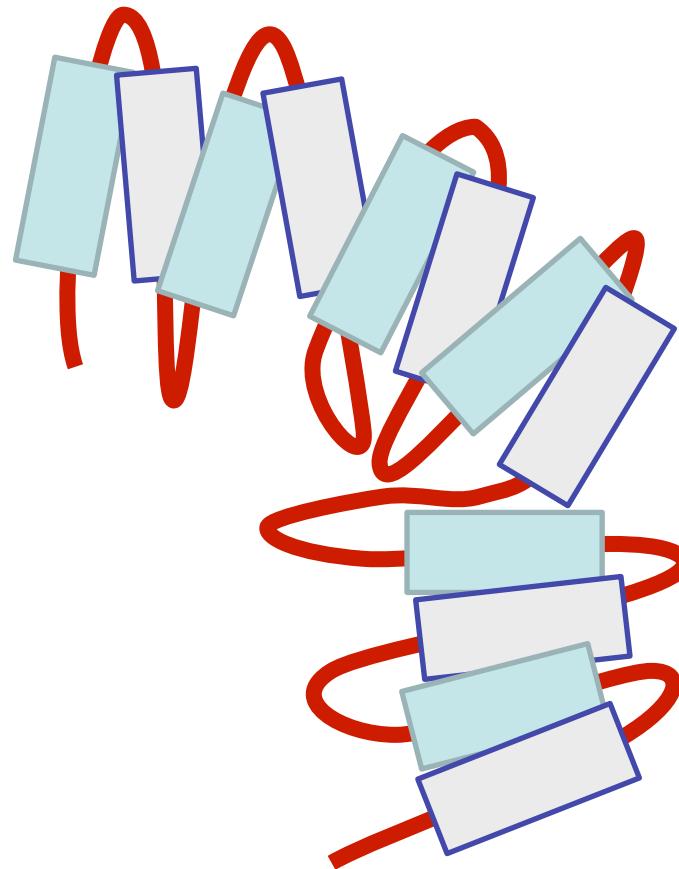
Tandem repeats fold together



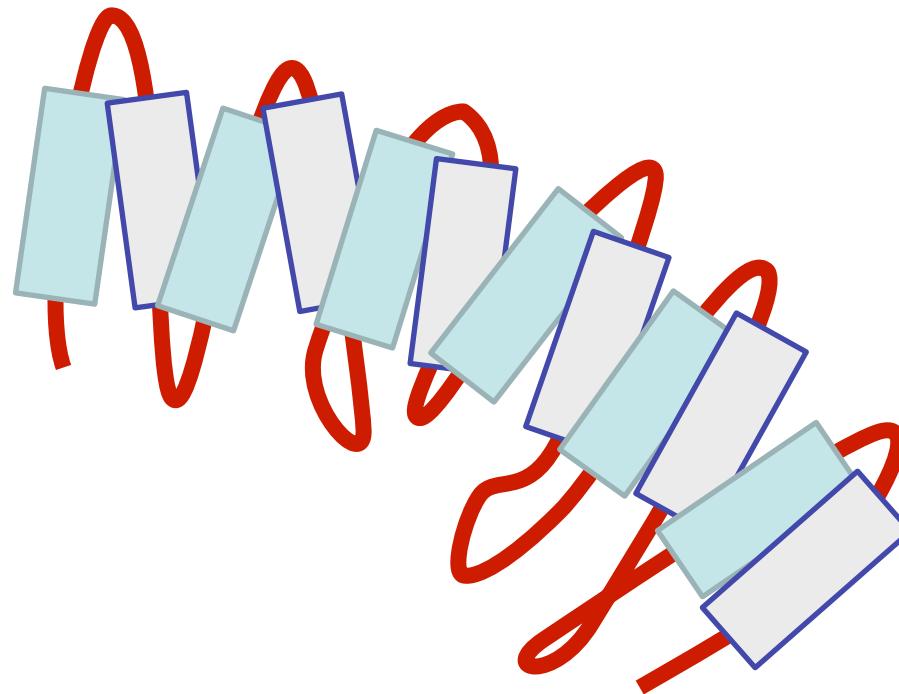
Tandem repeats fold together



Tandem repeats fold together



Tandem repeats fold together



Definition repeats

Sequence, long, imperfect, tandem

MRAV**VKSP**IM CHE

KSPSVC**SPLN**

MT**SSVCSP**AG INSVSSTTASF

GSFP**VHSP**IT Q

GTP LTC**SPNV** EN

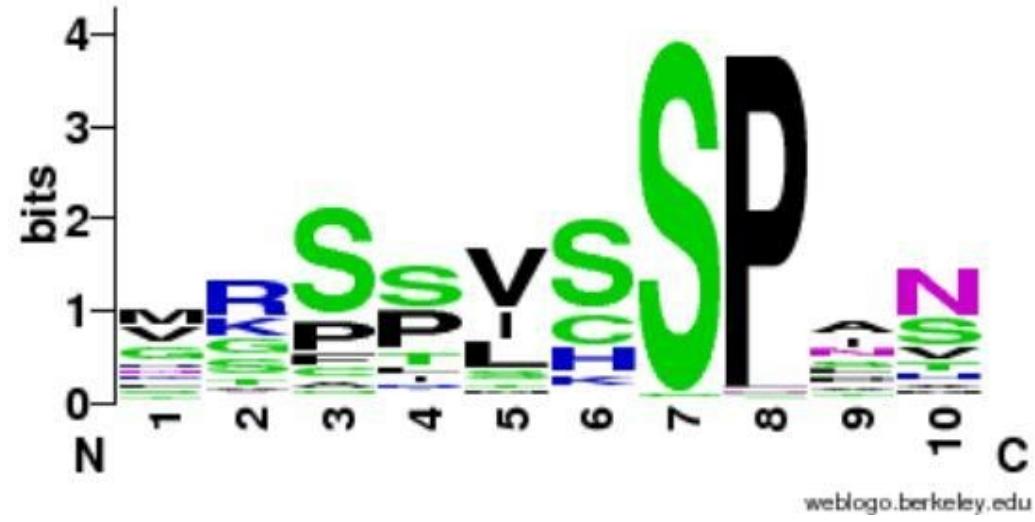
RG**SRSHSP**AH ASN

VG**SPLSSP**LS S

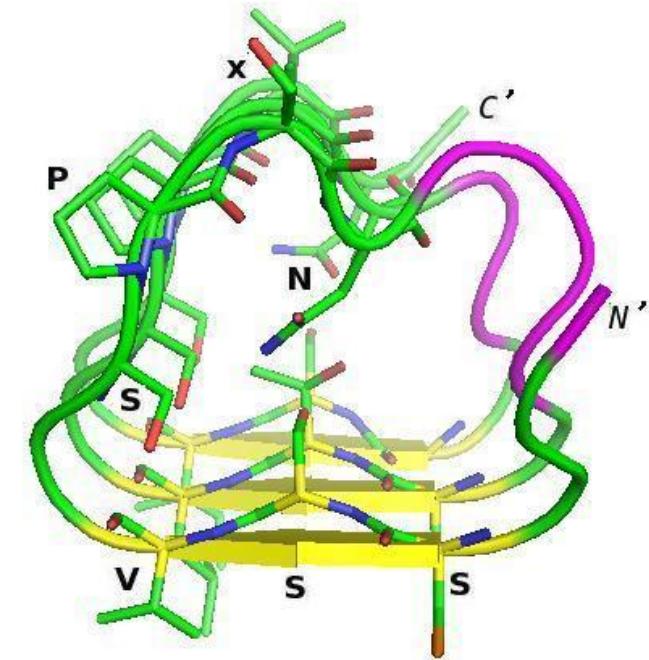
MK**SSISSP**PS HCS

VK**SPVSSP**NN VT

LR**SSVSSP**AN INN



<http://weblogo.berkeley.edu>



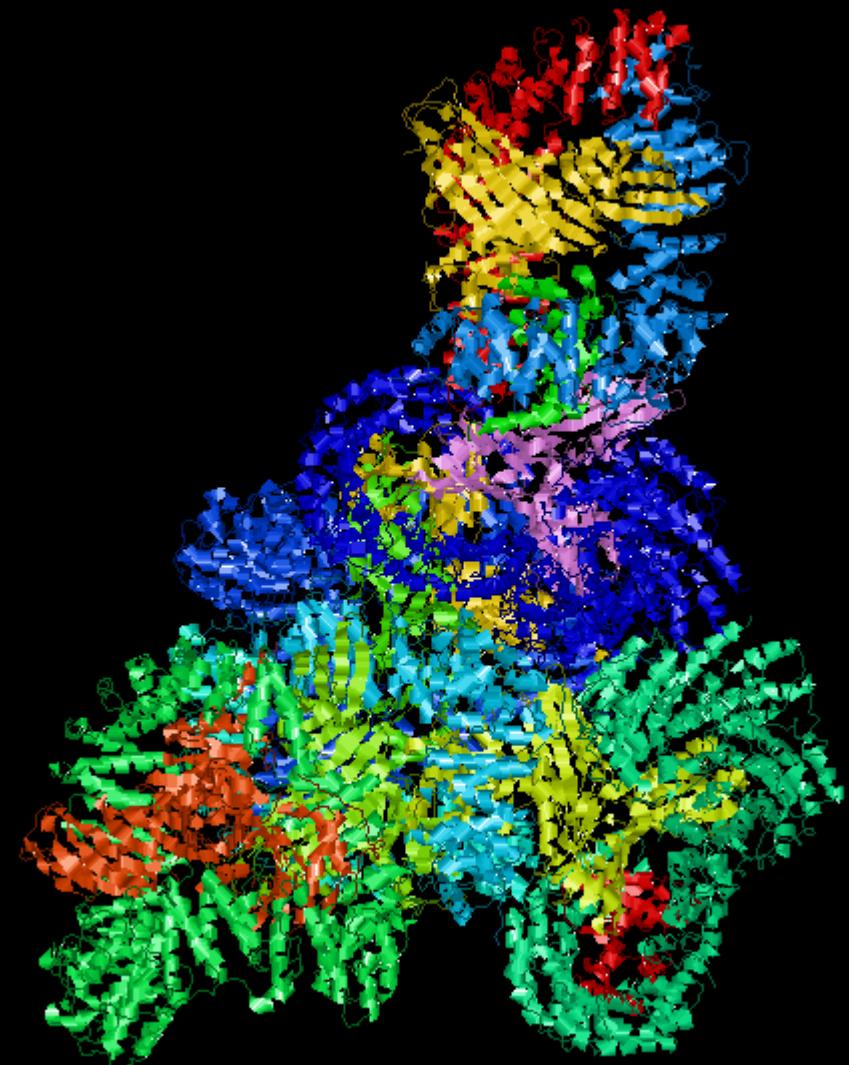
(Vlassi et al, 2013)

A subunit PP2A structure



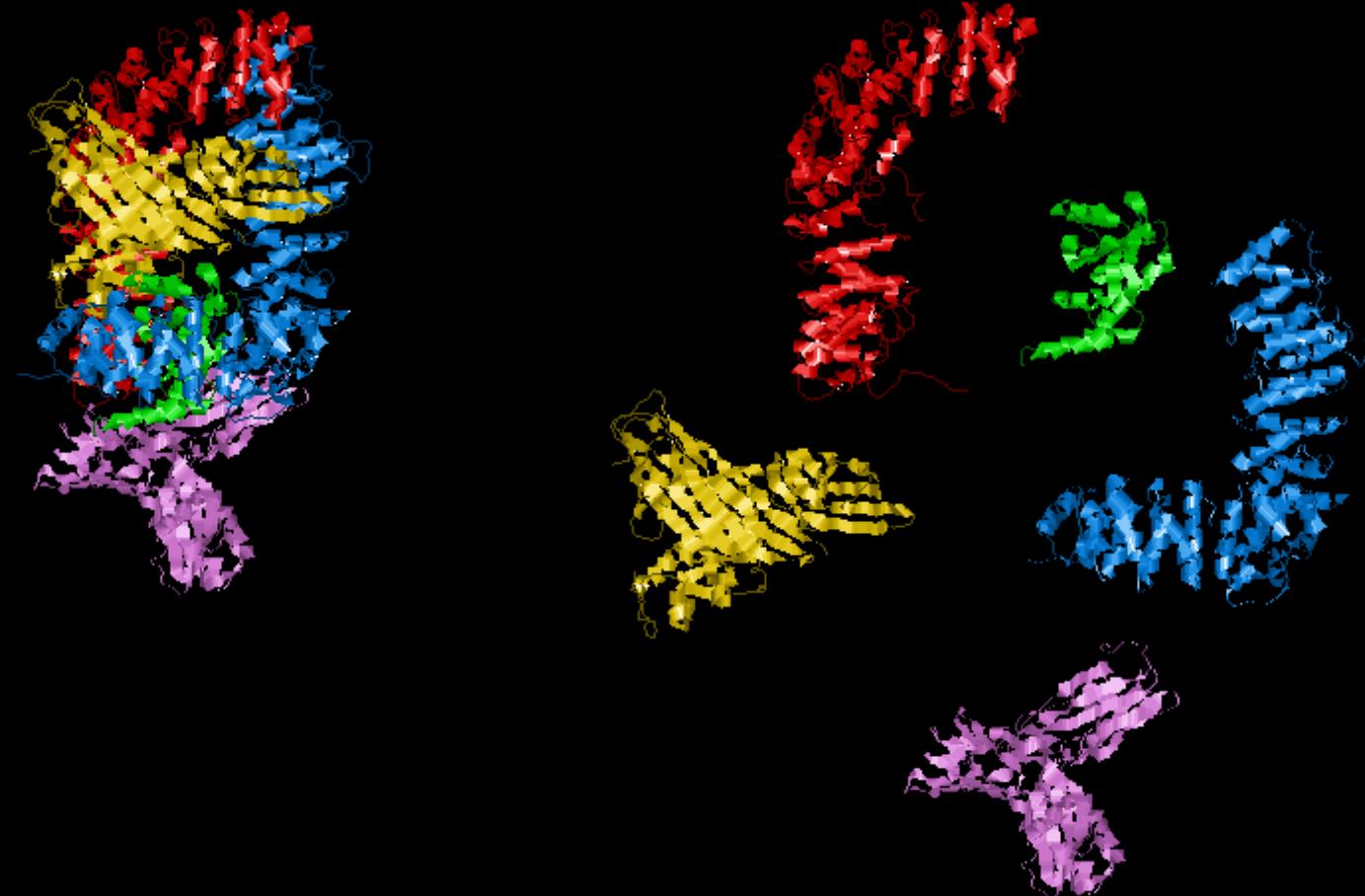
PDB:1b3u
Groves et al. (1999) *Cell*

Ap1 Clathrin Adaptor Core



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

Ap1 Clathrin Adaptor Core



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

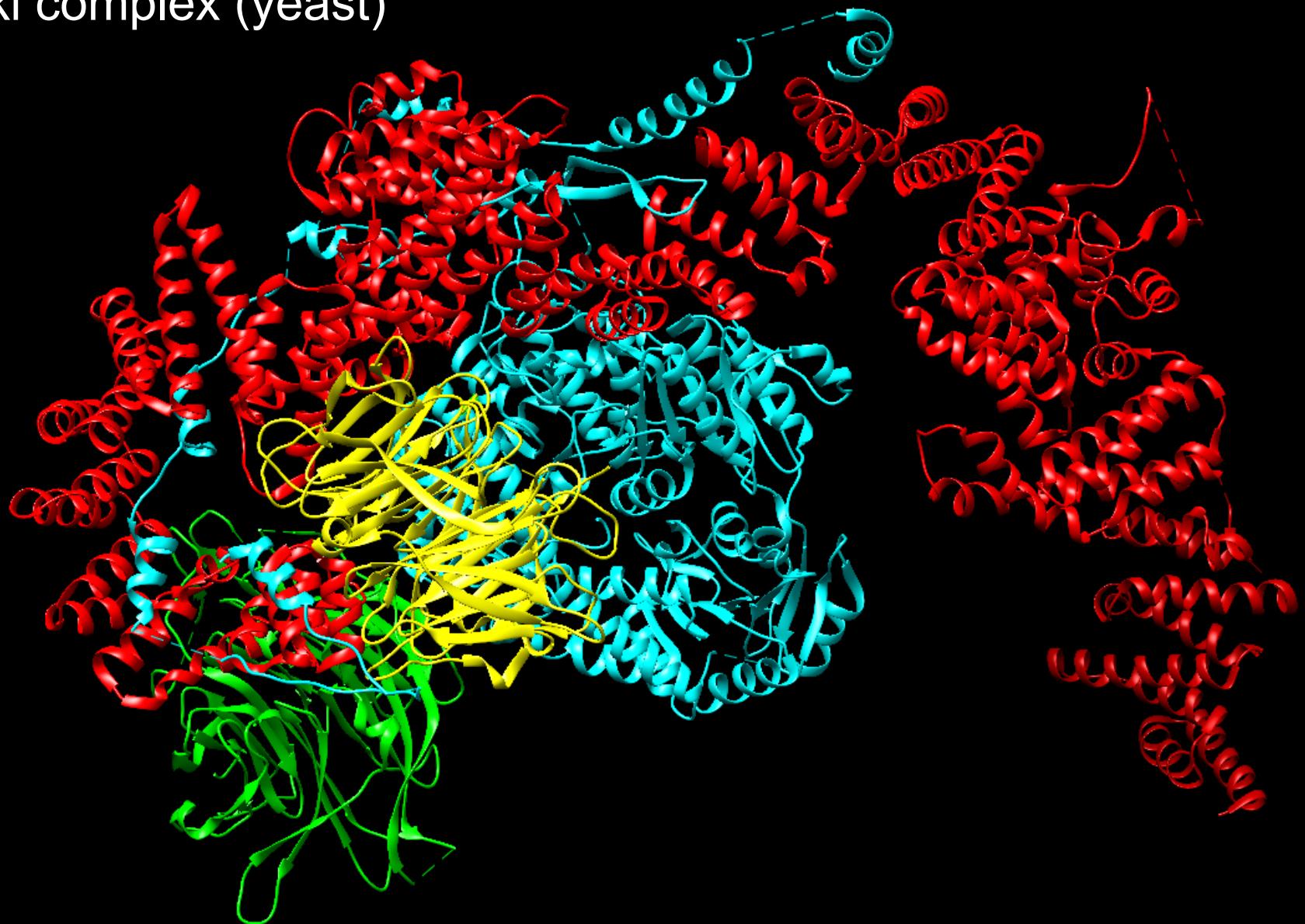
i-TASSER model of *D. melanogaster* thr protein

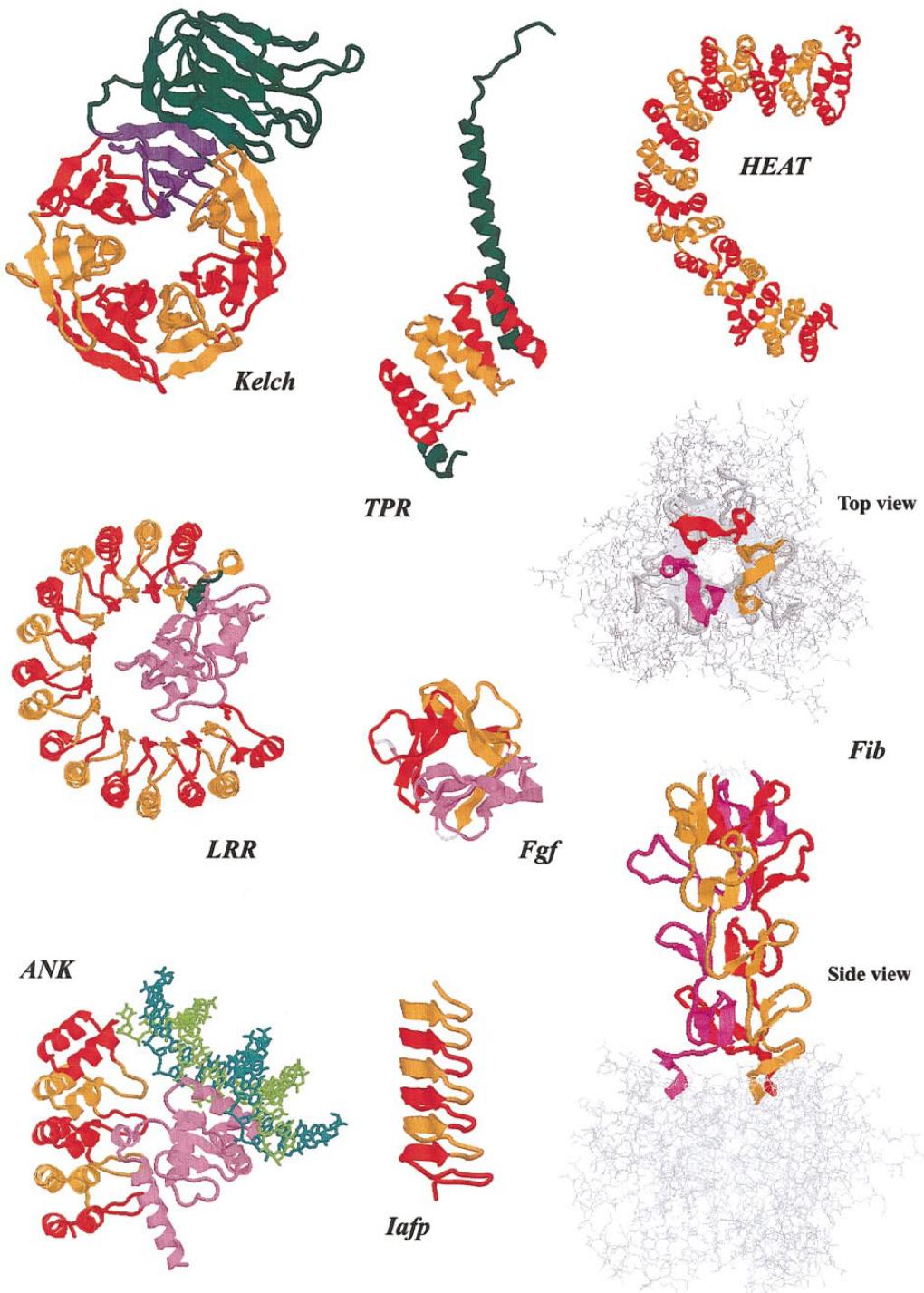


Based on PDB 4BUJ chain B

PDB 4BUJ

Ski complex (yeast)





Andrade et al. (2001)
J Struct Biol

Definition CBRs

Perfect repeat: QQQQQQQQQQQQQQ

Imperfect: QQQQPQQQQQQQ

Amino acid type: DDDDDDEEEDEED

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 Huntington OS=Homo sapiens
MATLEKLMKAFESIKSFQQQQQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECINKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGCALELLQQQFRTPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSVDSSALTASVKDEISGELAA
SSGVSTPGSAGHDIIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDFVLRDEATEPGDQE
NKPCRIKGDIGQSTDSSAPLVHCVRLLSASFLITGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSL
```

Detection CBRs

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

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKILLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRQLELYKEIKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGILV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSGFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSVDVSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAP
PSDPAMDLNQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLKNMSHCRQPSDSSVDFVLRDEATEPGDQE
NKPCRIKGDIGQSTDSSAPlVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSL

Detection CBRs

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

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKILLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRQLELYKEIKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGILV
PVEDEHSTLLILGVLLTLRYLVLLQQQVKDTSLKGSGFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLEEEAEDDSESRSDVSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSQVSAP
PSDPAMDLNQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPD
EDEEATGILPDEASEAFRNSSMALQQAHLKNMSHCRQPSDSSVDFVLRDEATEPGDQE
NKPCRIKGDIGQSTDSSAPlVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSI

Detection CBRs

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

```
>sp|P42858|HD_HUMAN Huntington OS=Homo sapiens
MATLEKLMKAFESIKSFQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQLPQPPPQAQP
LLPQPQPPPPPPPPPQPAQPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECINKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPPLLQQQVKDTSLKGSGFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGAEELLQQQLFRTPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDSSAPLVHCVRLLSASFLITGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSL
```

Detection repeats

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

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESIKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECINKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGCALELLQQLFRTPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSVDSSALTASVKDEISGELAA
SSGVSTPGSAGHDIIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDSSAPLVHCVRLLSASFLITGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSL
```

Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

```
>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens
MATLEKLMKAFESIKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWVNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLGEEEEALEDDSESRSDVSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL
```

Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

EFQKLLGIAMELFLLCSD**DA**ESDVRMVADECLNKVIKA
CRPYLVNLLPCLRTSKR**P-E**ESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKS**SSPTIRRTAAGSAVSICQHS**
TQYFYSWLLNVLLGLLVP**VEDEHSTLLILGVLLTLRYL**
PSAEQLVQVYELTLHHTQ**HQDHNVVTGALELLQQLFRT**

Detection repeats

Often NOT straightforward.

N-terminal human Huntingtin.

How many **repeats** can you find?

EFQKLLGIAMELFLLCSD**DA**ESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKR**P**-EESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLK**S**SPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLG**L**VP**VE**DEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQ**HQ**DHNVVTGALELLQQLFRT

: :
EFQKLLGIAMELFLLCSD**DA**ESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKR**P**-EESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLK**S**SPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLG**L**VP**VE**DEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQ**HQ**DHNVVTGALELLQQLFRT

Repeats

Frequency repeats

Fraction of proteins annotated with the keyword REPEAT in SwissProt

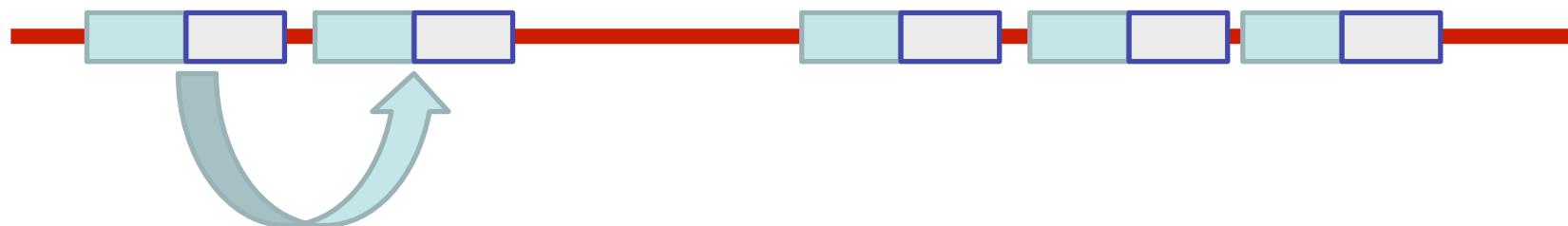
| | | % |
|-------------------|------------|------|
| Archaea | 27/3428 | 0.79 |
| Viruses | 81/8048 | 1.00 |
| Bacteria | 299/28438 | 1.05 |
| Fungi | 232/8334 | 2.78 |
| Viridiplantae | 153/6963 | 2.20 |
| Metazoa | 1538/28948 | 5.31 |
| Rest of Eukaryota | 92/2434 | 3.78 |

(Andrade et al 2001)

Detection of repeats

Dotplots

Comparing a sequence against itself



Detection of repeats

Dotplots

TLRSSVSSSPANINNS
NMTSSVCSPANISV

Detection of repeats

Dotplots



TLRSSVSSSPANINNS
|
NMTSSVCSPANISV

1 match

Detection of repeats

Dotplots



TLRSSVSSPANINNS
| | | | | | | |
NMTSSVCSPANISV

8 matches

Detection of repeats

Dotplots



TLRSSVSSSPANINNS
| |
NMTSSVCSPANISV

2 matches

Detection of repeats

Dotplots

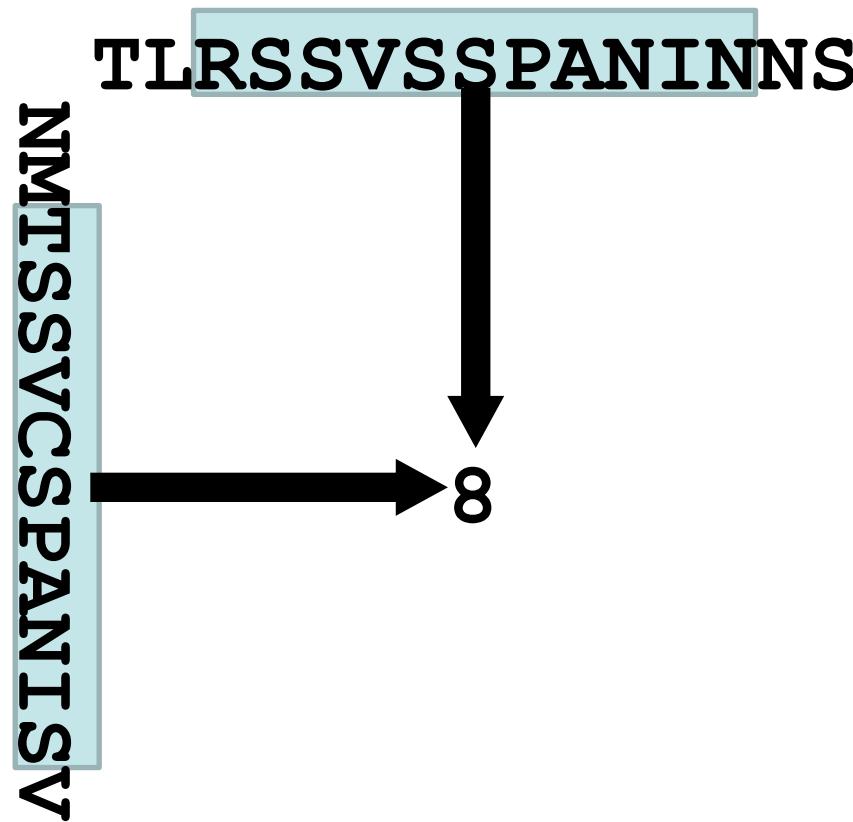


TLRSSVSSPANINNS
|
NMTSSVCSPANISV

1 match

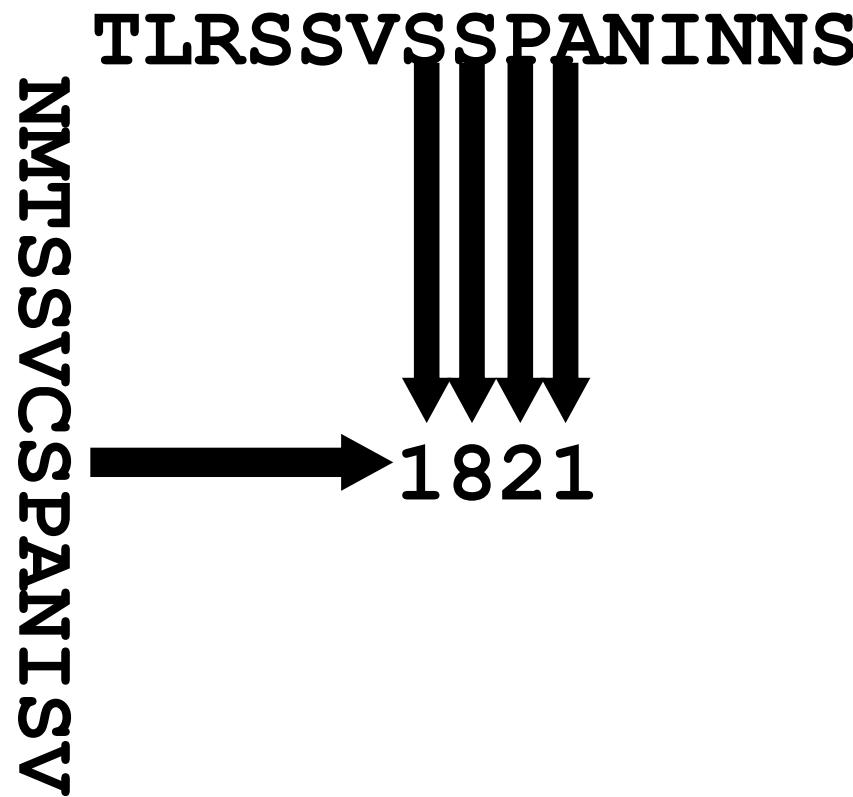
Detection of repeats

Dotplots



Detection of repeats

Dotplots



SEQUENCE 1

SEQUENCE 2

Window size

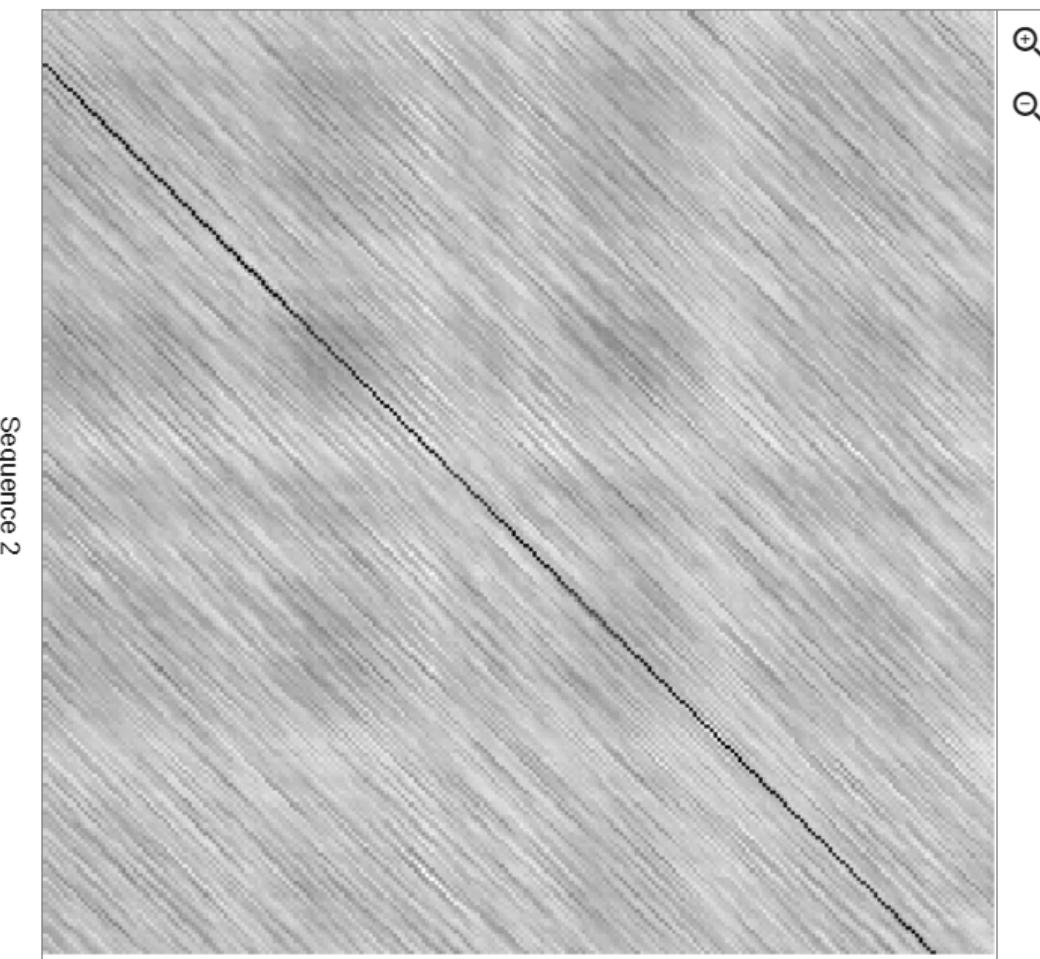
15

Scoring matrix

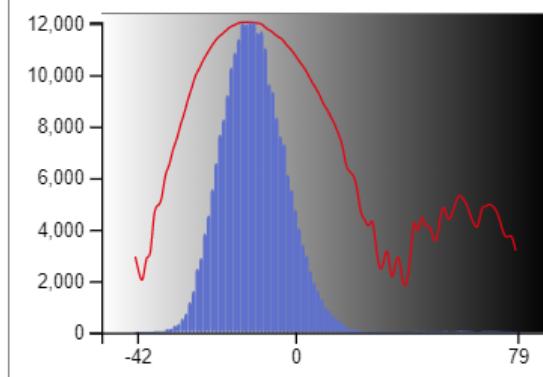
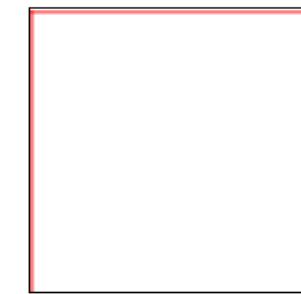
BLOSUM 62



Sequence 1



Sequence 2



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

Seq1:1

[

MTMDKSEL[VQKAKLAEQAERYDDM[AAMKAVTEQGHE
RKPLQTPTPIRRLWTMDTSELVQKA[KLAEQAERYDDM

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

- Go to the Dotlet web page:

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

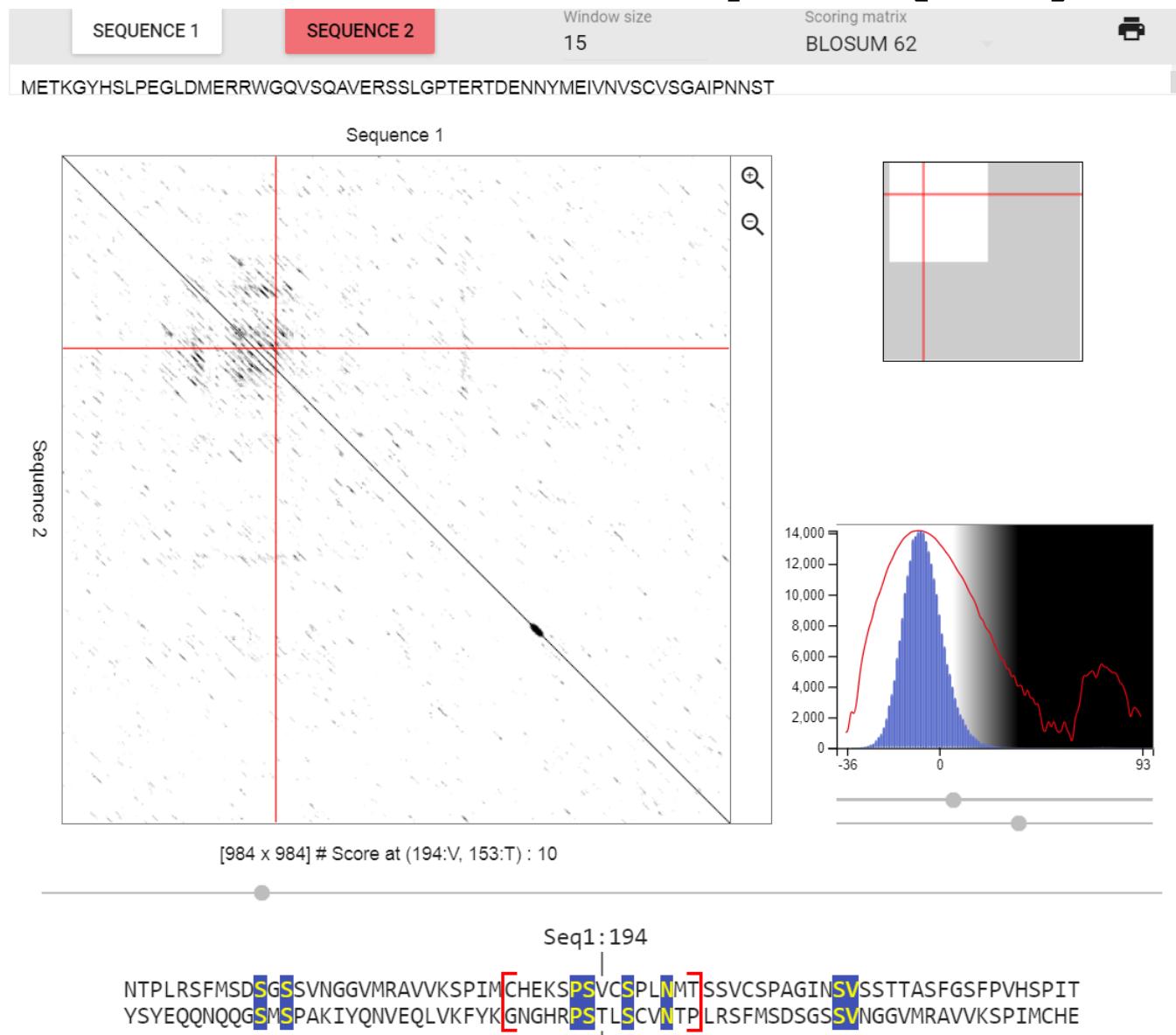
- Click on the input button and paste the sequence of the human mineralocorticoid receptor (UniProt id P08235)

- 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

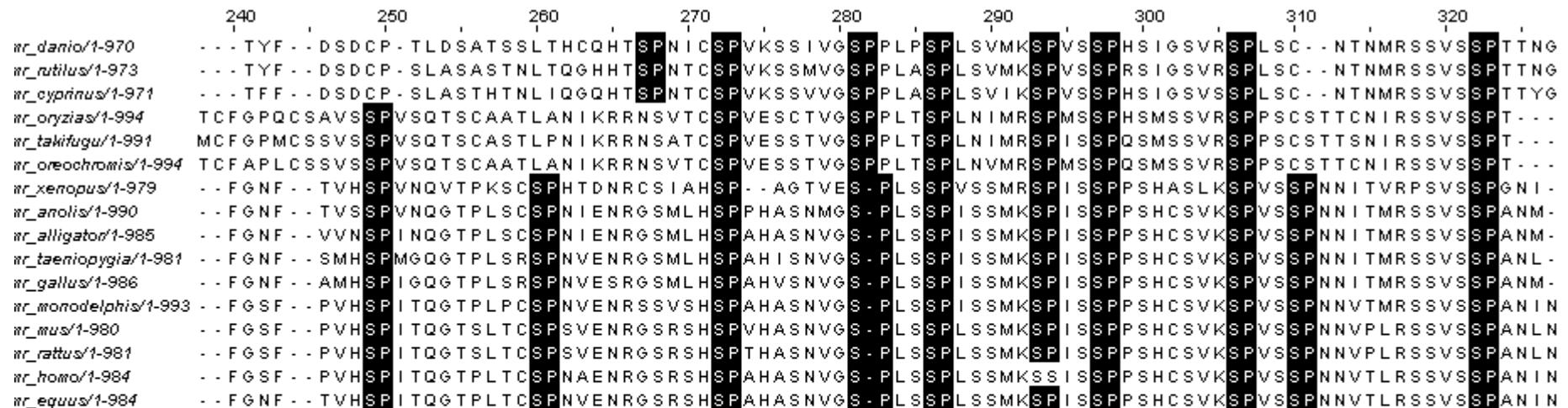
Using a multiple sequence alignment helps.
Conserved repeated patterns

| | 240 | 250 | 260 | 270 | 280 | 290 | 300 | 310 | 320 |
|-----------------------------|--|-----|-----|-----|-----|-----|-----|-----|-----|
| <i>mr_danio/1-970</i> | - - - TYF - - DSDCP - TLD SAT SSLTHCQHTSPN I CSPV KSSIV GSPPL P S L SVMK SPV SS PHS IGS VRSPLSC - - N TNM RSSV S SPTT NG | | | | | | | | |
| <i>mr_nutilus/1-973</i> | - - - TYF - - DSDCP - SLAS A STNL T QGH HTSPN T C SPV KSSMV GSPPL ASPLS VMK SPV SS PHS IGS VRSPLSC - - N TNM RSSV S SPTT NG | | | | | | | | |
| <i>mr_cyprinus/1-971</i> | - - - TFF - - DSDCP - SLAS H TNL I QGQHTSPN T C SPV KSSVV GSPPL ASPLS VIK SPV SS PHS IGS VSSPLSC - - N TNM RSSV S SPTT YG | | | | | | | | |
| <i>mr_oryzias/1-994</i> | T C FGP QCSA VSSP VSQT SCA AT LAN I KRRN S VTC SPV E SCT V GSPPL T S P L N I M RSP M S S P H S M S S V R S P P S C S T T C N I R S S V S S P T - - - | | | | | | | | |
| <i>mr_takifugu/1-991</i> | M C FGP MC S S V S S P VSQT SCA STLP N I KRRN S A T C SPV E S S T V GSPPL T S P L N I M RSP I S S P Q S M S S V R S P P S C S T T S N I R S S V S S P T - - - | | | | | | | | |
| <i>mr_oreochromis/1-994</i> | T C FAPLC S S V S S P VSQT SCA AT LAN I KRRN S VTC SPV E S S T V GSPPL T S P L N V M RSP M S S P Q S M S S V R S P P S C S T T C N I R S S V S S P T - - - | | | | | | | | |
| <i>mr_xenopus/1-979</i> | - - FGNF - - TVHSPVNQVTPKSCSPHTDNRC SIAHSP - - AGTVES - PLSSPVSSMRSP I S S P P S H A S L KSPVSSPNN I TVRPSVSSPGNI - - | | | | | | | | |
| <i>mr_anolis/1-990</i> | - - FGNF - - TVSSPVNQG TPLSCSPN I ENRG SML HSPPH AS NM G S - PLSSP I S S M K S P I S S P P S H C S V K S P V S S P N N I TMRSSVSSPANM - - | | | | | | | | |
| <i>mr_alligator/1-985</i> | - - FGNF - - VVNSP I NQG TPLSCSPN I ENRG SML HSPA H AS NV G S - PLSSP I S S M K S P I S S P P S H C S V K S P V S S P N N I TMRSSVSSPANM - - | | | | | | | | |
| <i>mr_taeniopygia/1-981</i> | - - FGNF - - SMHSPMGQGTPLSRSPN VENRG SML HSPA H I S NV G S - PLSSP I S S M K S P I S S P P S H C S V K S P V S S P N N I TMRSSVSSPANL - - | | | | | | | | |
| <i>mr_gallus/1-986</i> | - - FGNF - - AMHSP I GQG TPLSRSPN VESRG SML HSPA H V S NV G S - PLSSP I S S M K S P I S S P P S H C S V K S P V S S P N N I TMRSSVSSPANM - - | | | | | | | | |
| <i>mr_monodelphis/1-993</i> | - - FGSF - - PVHSP I T Q G TPLPCSPN VENRG S RSHSPV H AS NV G S - PLSSP I S S M K S P I S S P P S H C S V K S P V S S P N N V T M R S S V S S P A N I N - - | | | | | | | | |
| <i>mr_mus/1-980</i> | - - FGSF - - PVHSP I T Q G T S L T C S P S VENRG S RSHSPV H AS NV G S - PLSSPL L S S M K S P I S S P P S H C S V K S P V S S P N N V P L R S S V S S P A N L N - - | | | | | | | | |
| <i>mr_rattus/1-981</i> | - - FGSF - - PVHSP I T Q G T S L T C S P S VENRG S RSHSPV H AS NV G S - PLSSPL L S S M K S P I S S P P S H C S V K S P V S S P N N V P L R S S V S S P A N L N - - | | | | | | | | |
| <i>mr_homo/1-984</i> | - - FGSF - - PVHSP I T Q G TPLTCSPNAENRG S RSHSPV H AS NV G S - PLSSPL L S S M K S S I S S P P S H C S V K S P V S S P N N V T L R S S V S S P A N I N - - | | | | | | | | |
| <i>mr_equus/1-984</i> | - - FGNF - - TVHSP I T Q G TPLTCSPN VENRG S RSHSPV H AS NV G S - PLSSPL L S S M K S P I S S P P S H C S V K S P V S S P N N V T L R S S V S S P A N I N - - | | | | | | | | |

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:

[L S] 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:

[L S] P . A

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

Which one is not matched?

- LPTA, SPAA, LPAA, LPAP, SPLA**

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns

JalView with Regular Expression searches

- Regular Expressions:

[L S] P . A

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

Which one is not matched?

- LPTA, SPAA, LPAA, LPAP, SPLA**

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

- Load the multiple sequence alignment of the MR in JalView: MR1_fasta.txt (from URL: https://cbdm.uni-mainz.de/files/2015/02/MR1_fasta.txt)
- Use the “Select > find” (or Ctrl+F) option with a regular expression and mark all matches (**click the “Find all” option!**)
- 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?

|158508572|Hsapie ns
 |31324675|Cjacchus
 |126331313|Mdomestica
 |73978292|Clupus
 |301763180|Amelanoleuca
 |6981208|Rnorvegicus
 |144227212|Mmusculus
 |148224443|Xlaevis
 |327274009|Acarolinensis
 |115529242|Tguttata
 |225936142|Ggallus
 |239923135|Rrutilus
 |154240734|Drerio

| #T1 | #T2 | #T3 | #T4 | #T5 | #T6 | #T7 |
|-----------------------------------|--------------|----------------------|---|---|---|---|
| ..170... ...180... ... | 190... ... * | 200... | 210... ... 220... ... 230... ... 240... ... 250... ... 260... ... | * ... * ... * ... * ... * ... * ... * ... | * ... * ... * ... * ... * ... * ... * ... | * ... * ... * ... * ... * ... * ... * ... |
| SGSSVNGGVRAVVKSPIMCHE | KSPSVCSPLN | MTSSVCSPAG | INVSSTTA-SF | GSFPVHSPI | GTPLTCPNVENRGSRSHSPAHA-SNVGSPPLSSPLLS | SGSSVNGGVRAIVKSPIMCHE |
| SGSSVNGGVRAIVRSPIMCHE | KSPSVCSPLN | MTSSVCSPAG | INSESSTTA-SF | GSFPVHSPI | GTPLTCPNVENRGSRSHSPAHA-SNVGSPPLSSPLLS | SGNSVNGSIMRSIVKSPIMCHE |
| SGNSVNGSIMRSIVKSPIMCHE | KSPSVCSPLN | MNSSVCSPAG | INVSSTTA-NE | GSFPVHSPI | GTPLPCSPNVENRSSVSHSPAHA-SNVGSPPLSSPISS | SGSSVNGGVRAIVKSPIMCHE |
| SGSSVNGGVRAIVKSPIMCHE | KSPSVCSPLN | MTSSVCSPAG | INSSSSTA-SF | GSFTVHSPI | GTPLTCPNVENRGSRSHSPAHA-SNVGSPPLSSPLLS | SGSSVNGGVRAIVKSPIMCHE |
| SGSSVNGGVRAIVKSPIMCHE | KSPSVCSPLN | MTSSVCSPAG | INSSSSTA-SF | GSFTVHSPI | GTPLTCPNVENRGSRSHSPAHA-SNVGSPPLSSPLLS | SGASMNNGALRAIVKSPII |
| SAASMNNGALRAIVKSPII | CHE | INSSVSPLN | INMSSTTA-SF | GSFPVHSPI | GTPSLTCPNVENRGSRSHSPTHA-SNVGSPPLSSPLLS | SGTSMNNGALRAIVKSPII |
| SGTSMNNGALRAIVKSPII | CHE | INSSVSPLN | INMSSTTA-SF | GSFPVHSPI | GTSLTCSPNVENRGSRSHSPVHA-SNVGSPPLSSPLLS | SDKSMNGKIKSNTVKSPLSYSE |
| SDKSMNGKIKSNTVKSPLSYSE | CHE | KNLSVGSPSV | MALPVCSPTG | ISSTSCST--T | GNFTVHSPIQVTPKSCSPHTDNRCIAHSPACT--VE | SPLSSPVSS |
| CANSMNGSIMPSIMKNPRITQER | SPPDCCPQS | MTSSVCSPPG | INSVTSTTPTNF | GNFTVHSPIQVTPKSCSPHTDNRCIAHSPACT--VE | SPLSSPVSS | CANSMNGSIMPSIMKNPRITQER |
| SGSTVNGGAMHTIVKSPIMCQE | KSPSGCSPQN | MASSVCSPAG | VNSMSSTTA-SF | GNFSMHSPMG | GTPLSRSPNVENRGSMMLHSPAHI-SNVGSPPLSSPISS | SGSTVNGGAMHTIVKSPIMCQE |
| SGSAMNGGAMHAVVKSPIVCQE | KSPSGCSPQN | MASSVCSPAG | VNSVSSTTA-NE | GNFAMHSPIC | GTPLSRSPNVESRGSMMLHSPAHV-SNVGSPPLSSPISS | GSQPSGGPQECAVVSAVASVPGMASVLSCSSDG |
| GSQPSGGPQECAVVSAVASVPGMASVLSCSSDG | SP | GPGFMSPPTGHNMVSSTTSP | T | FDSDCFLASASTNLTQGH | TSPNTCSPVKSSMV | GSPFLASPLSV |
| GAQLPNGGPQECAVVSDSVPSALVTALSSSTD- | GSPCMSPPTQHN | MVSSTTSP | T | FDSDCPTLDSATSSLTHCQH | TSPNICSPVKSSIV | GSPPLPSPLSV |

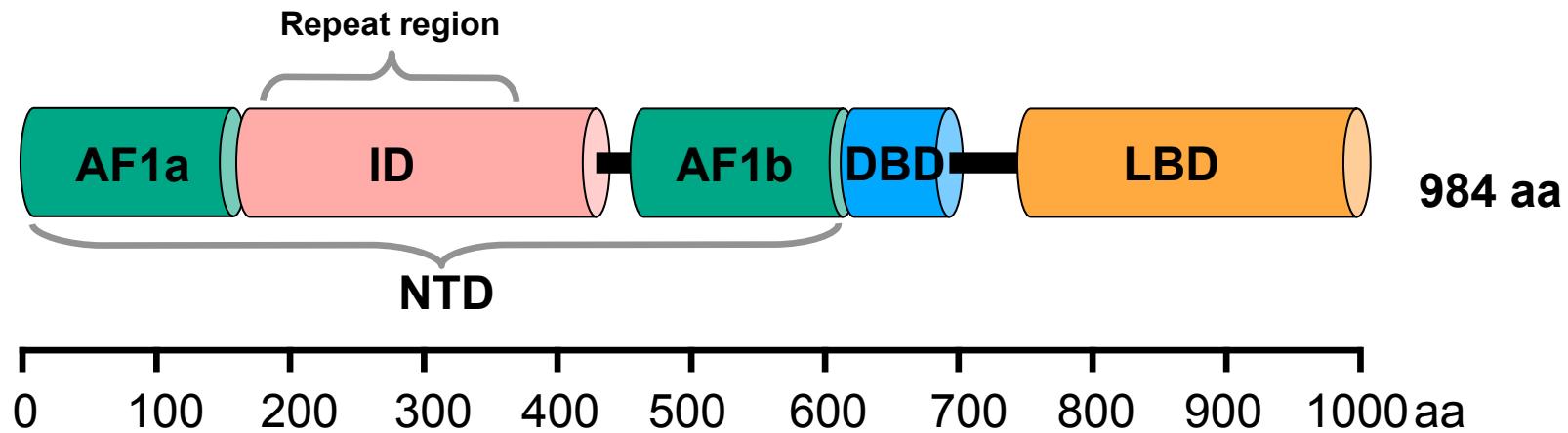
#F1 #F2 #F3 #F4

|158508572|Hsapie ns
 |31324675|Cjacchus
 |126331313|Mdomestica
 |73978292|Clupus
 |301763180|Amelanoleuca
 |6981208|Rnorvegicus
 |144227212|Mmusculus
 |148224443|Xlaevis
 |327274009|Acarolinensis
 |115529242|Tguttata
 |225936142|Ggallus
 |239923135|Rrutilus
 |154240734|Drerio

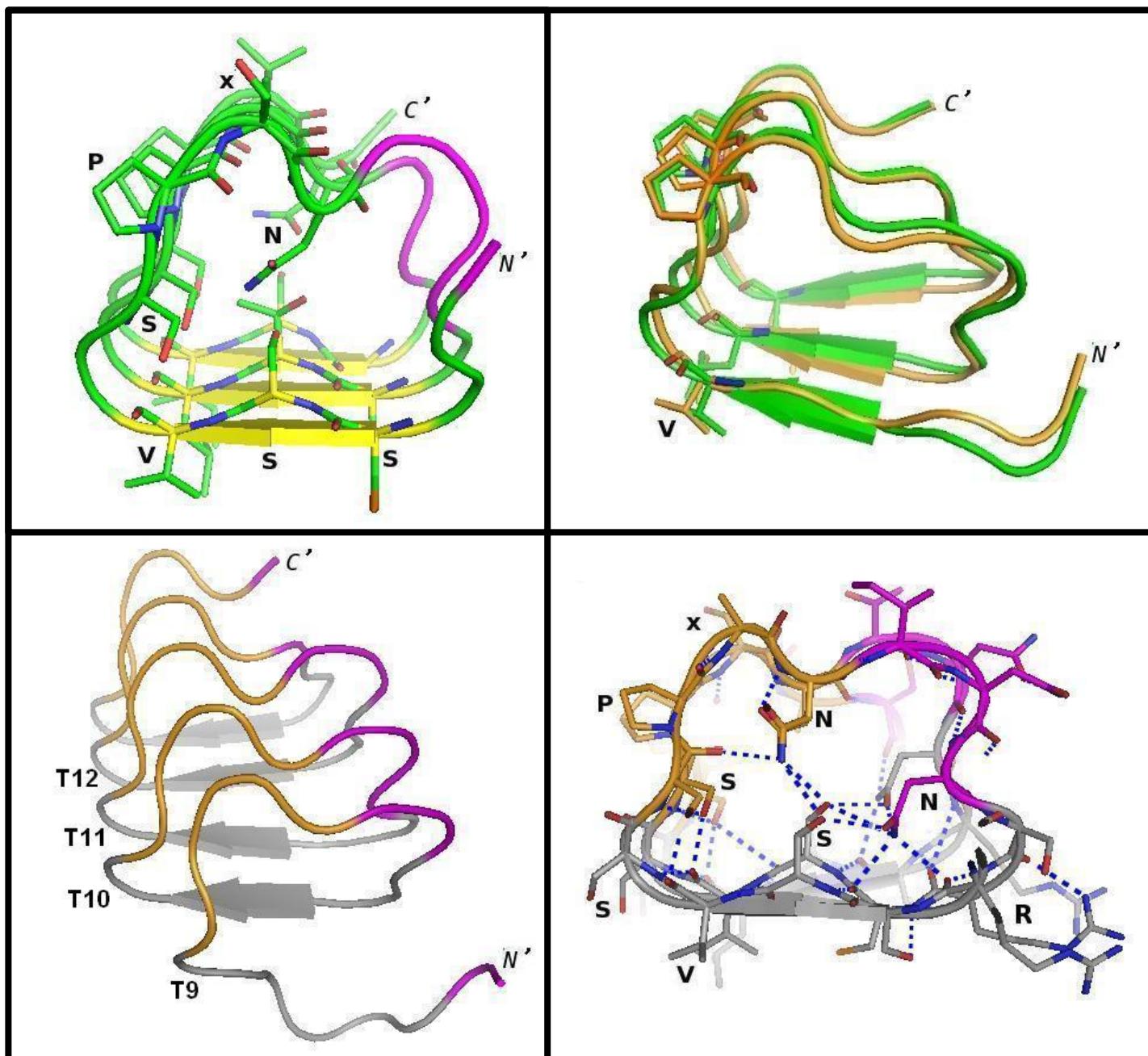
| #T8 | #T9 | #T10 | #T11 | #T12 | #T13 | #T14 | #T15 |
|------------------------|--------------|--------------|----------------|----------------|----------------|---|----------------------|
| ..270... ...280... ... | 290... ... * | 300... ... * | 310... | 320... | 330... | 340... | 350... * ... |
| MKSSISSPSPHCS | VKSPVSSPNV | LRSSVSSPAN | IN--- | NRCSVSSPNTN | NRSTLSSPAAS | TVGSICSPVNNAFSYTASGTSAGSSTL | RDVVPSPDTQE-- |
| MKSSISSPSPHCS | VKSPVSSPNV | LRSSVSSPAN | IN--- | NRCSVSSPNTN | NRSTLSSPAAS | TVGSICSPVNNAFSYTASGTSAGSSTISRDVVPSPDTQE-- | MRSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPNV | MRSSVSSPAN | IN--- | NRCSVSSPNTN | NRSTLSSPAAS | TVGSICSPGNNAFSAASATSVGSSTTQDVIPSPETNE-- | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPNV | MRSSVSSPAN | IN--- | NRCSVSSPNTN | NRSTLSSPNTA | TVGSICSPVNNAFSYTASGTPAGSSWARDVVPSPD | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPNV | MRSSVSSPAN | IN--- | NRCSVSSPNTN | NRSTLSSPNTA | TVGSICSPVNNAFSYTASGTPAGSSAARDVVPSPD | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPNV | MRSSVSSPAN | LN--- | NRCSVSSPNTN | NRSTLSSPNTA | TVGSIGSPISNAFSYATSGASAAGAGAIQDVVPSPDTHE-- | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPNV | MRSSVSSPAN | LN--- | NRCSVSSPNTN | NRSTLSSPNTA | TVGSIGSPISNAFSYTTSGASAAGAGAIQDMVPSPDTHE-- | MKSPISSPPHCS |
| MKSPISSPPHCS | LKSPVSSPN | VRFSVSSPGNI | I--- | NRSSLSSPNTN | NRSTISSPAA | MGSSICSPASSTLGFLPGVMPTDG-GTASDISAE | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPN | MRSSVSSPAN | I--- | NRSSLSSPNTN | NRSTISSPAA | MGSSICSPASSTLGFLPGVMPTDG-GTASDISAE | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPN | MRSSVSSPAN | I--- | NRSSLSSPNTN | NRSTLSSPNTA | VGSSMCSPVNNGALSFPPSSTPVGPGTGQDIVSPDTKD | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPN | MRSSVSSPAN | I--- | NRSSLSSPNTN | NRSTLSSPNTA | VGSSMCSPVNNGALSFPPSSTPVGPGTGQDIVSPDTKD | MKSPISSPPHCS |
| MKSPISSPPHCS | VKSPVSSPN | MRSSVSSPAN | M--- | NRSSLSSPNTN | NRSTRSSPNTA | GGSSICSPVNSIGFLPSGTVPGPSRSQDTPVSPETKD | MKSPISSPPHCS |
| MKSPVSSPRIGS | VRSPLCNTN | MRSSVSSPTT | NGNTCN | IRPSISSPP | ST | GSMAMSSPRNRSRGFSVSSPPSGLGI | MKSPVSSPRIGS |
| MKSPVSSPRIGS | VRSPLCNTN | MRSSVSSPTT | NGNTCN | IRPSISSPP | ST | GSMAMSSPRNRSRGFSVSSPPSGLGI | MKSPVSSPRIGS |

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

Mineralocorticoid receptor



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



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: QQQQQQQQQQQQQQ

Imperfect: QQQQPQQQQQQQ

Amino acid type: DDDDDDEEEDEED

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

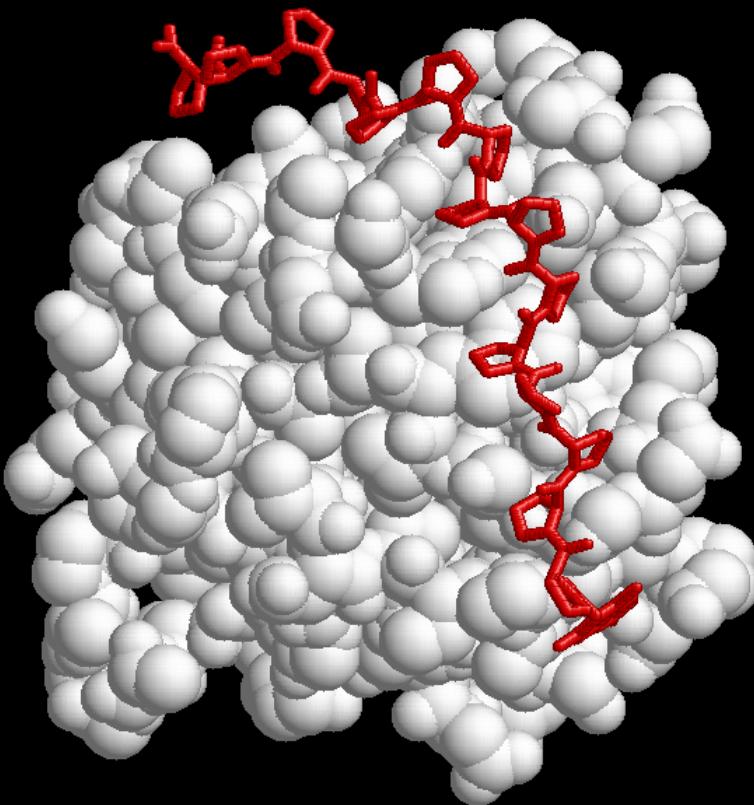
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

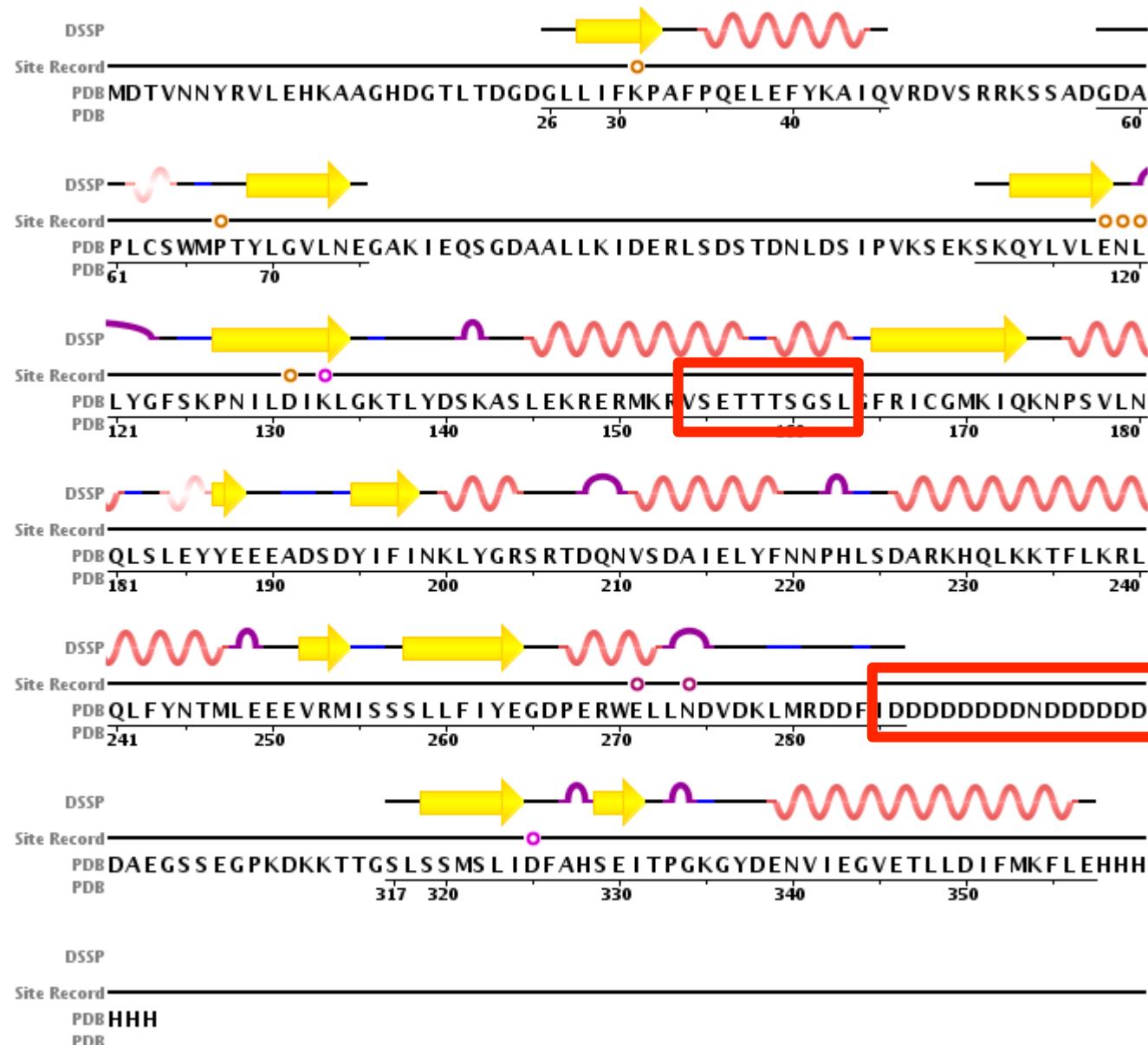
Structure of CBRs

Often variable or flexible: do not easily
crystallize

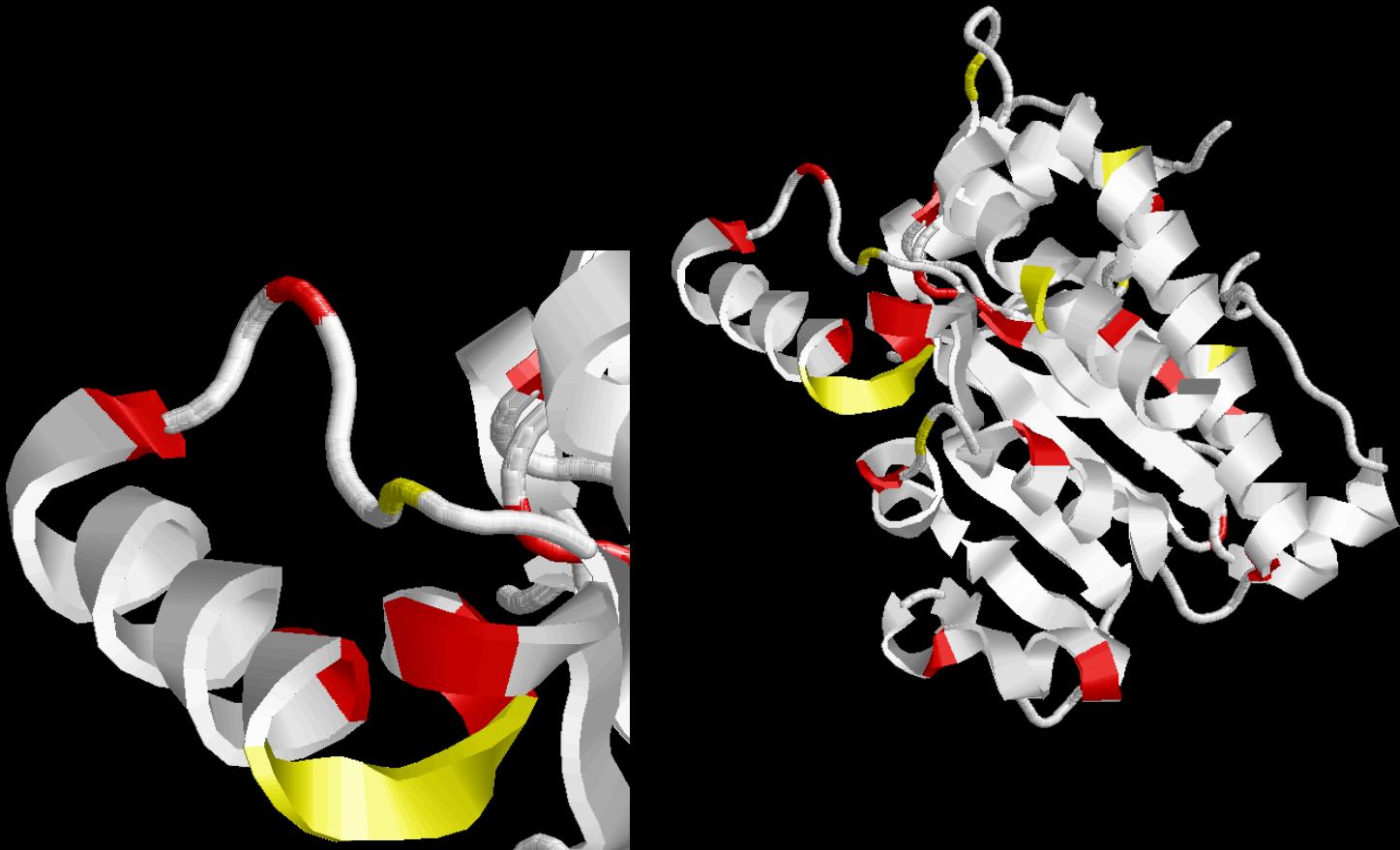
1CJF: profilin bound to polyP



2IF8: Inositol Phosphate Multikinase Ipk2

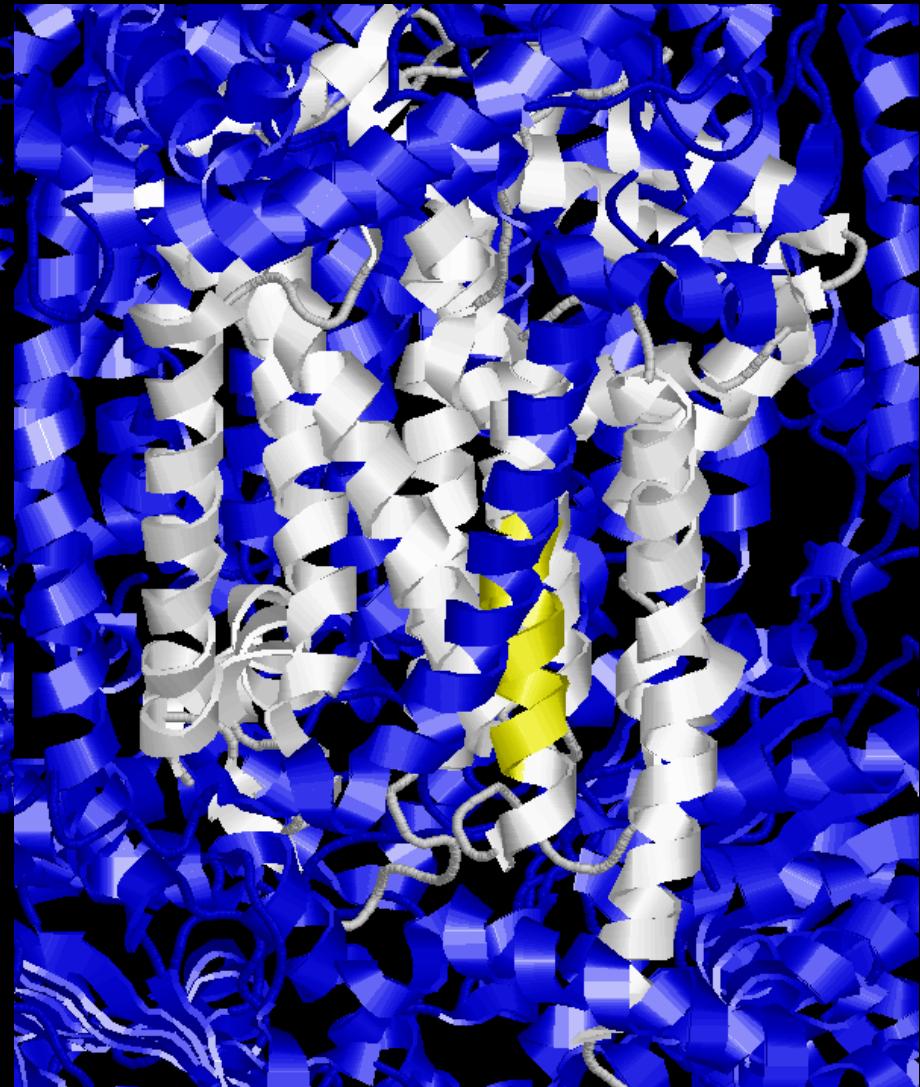
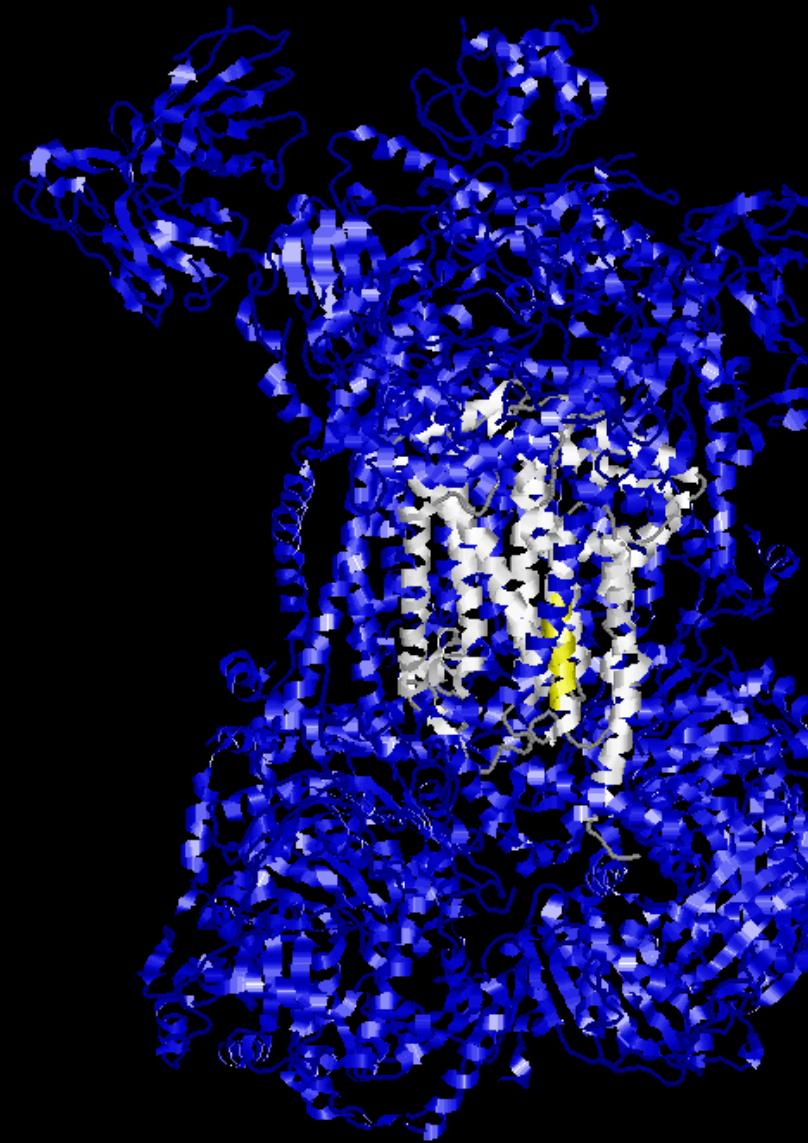


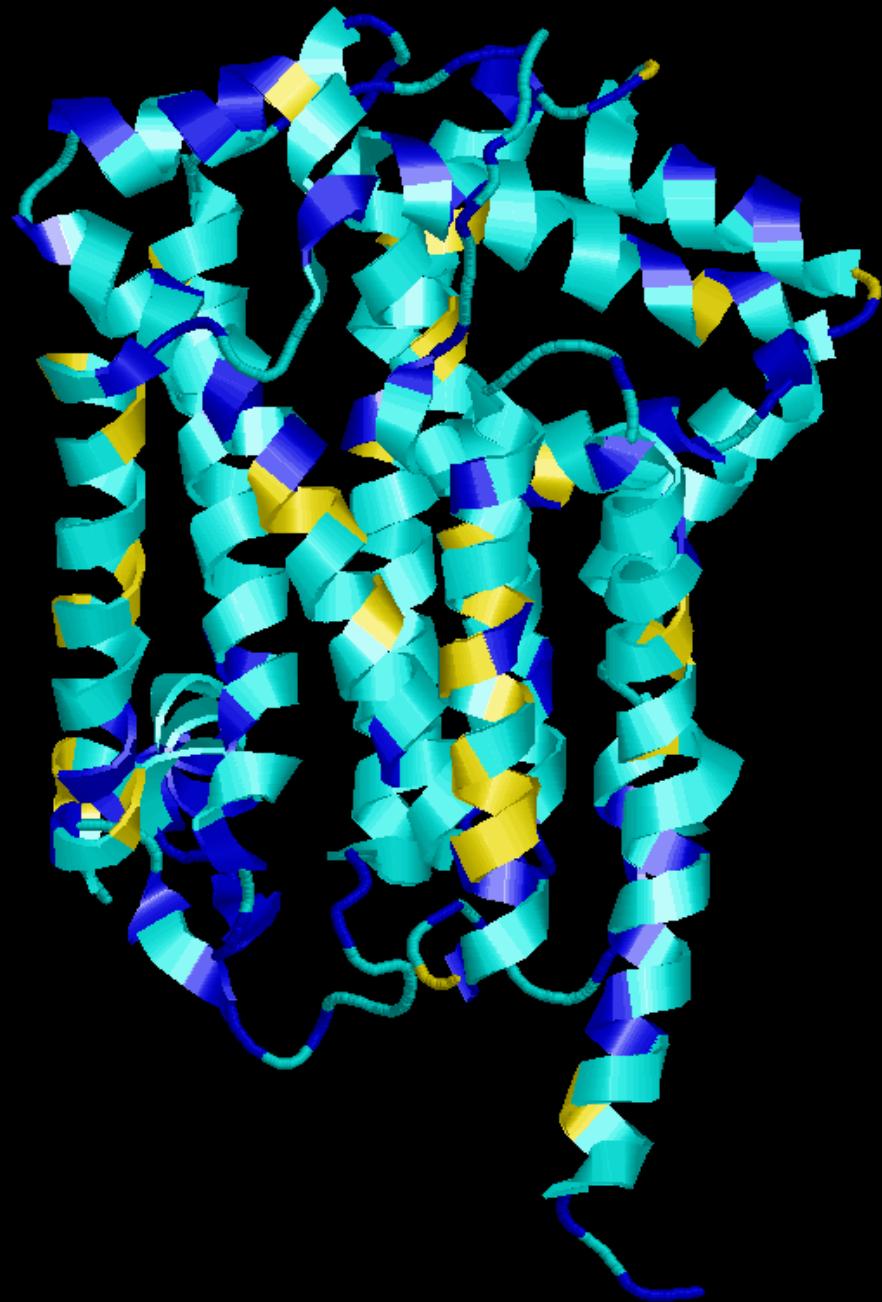
2IF8: Inositol Phosphate Multikinase Ipk2



RV**S**ETTT**S**GSL

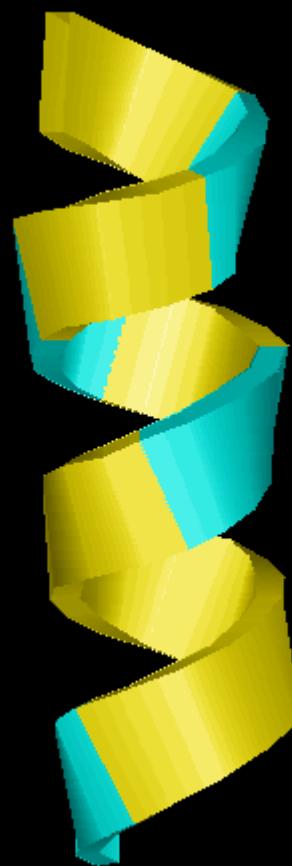
2CX5: mitochondrial
cytochrome c
B subunit N-terminal





2CX5: mitochondrial
cytochrome c
B subunit N-terminal

EFFEFFVVFNE



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

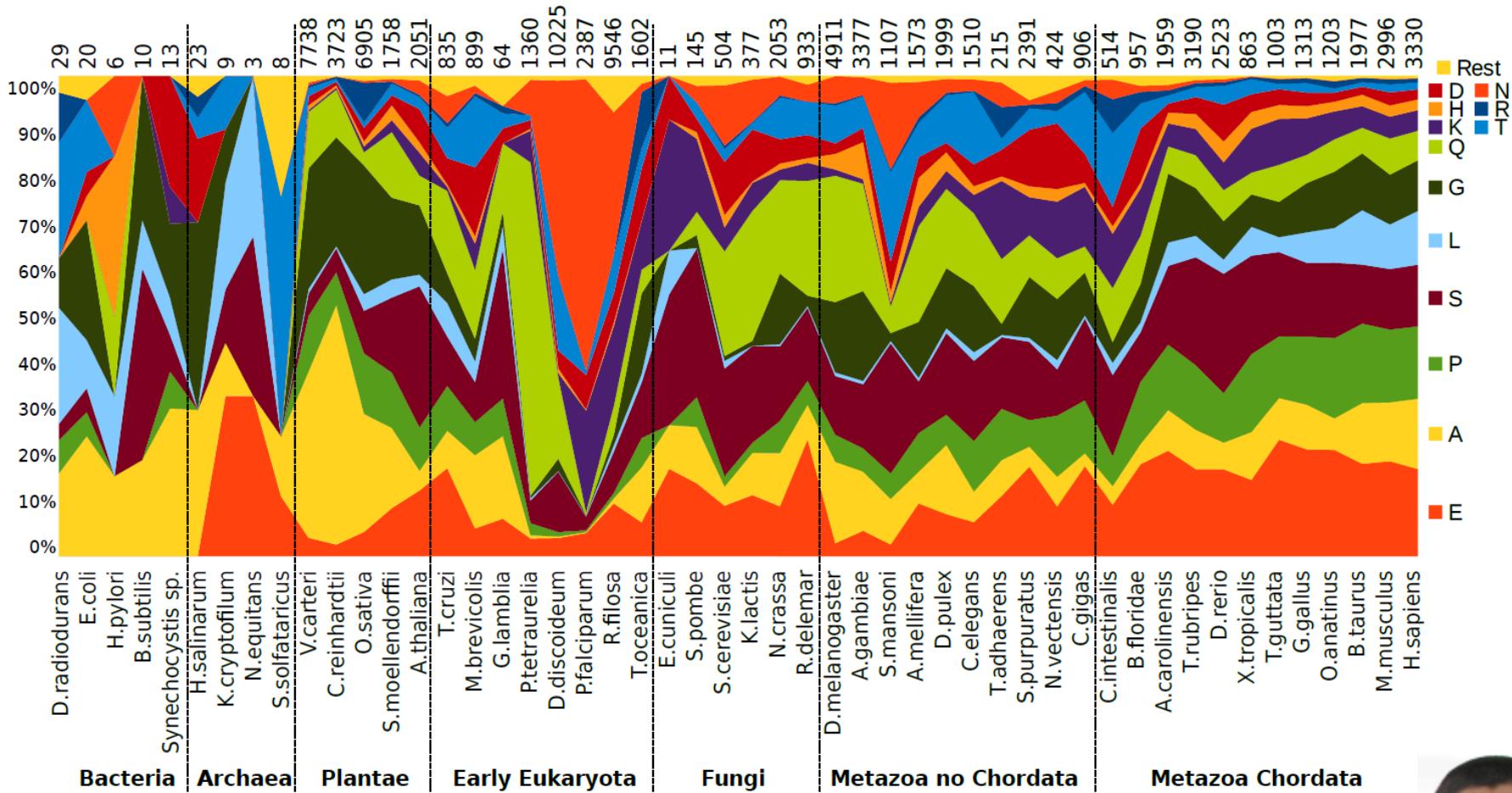
Very rare or absent in both eukaryota and prokaryota:

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:

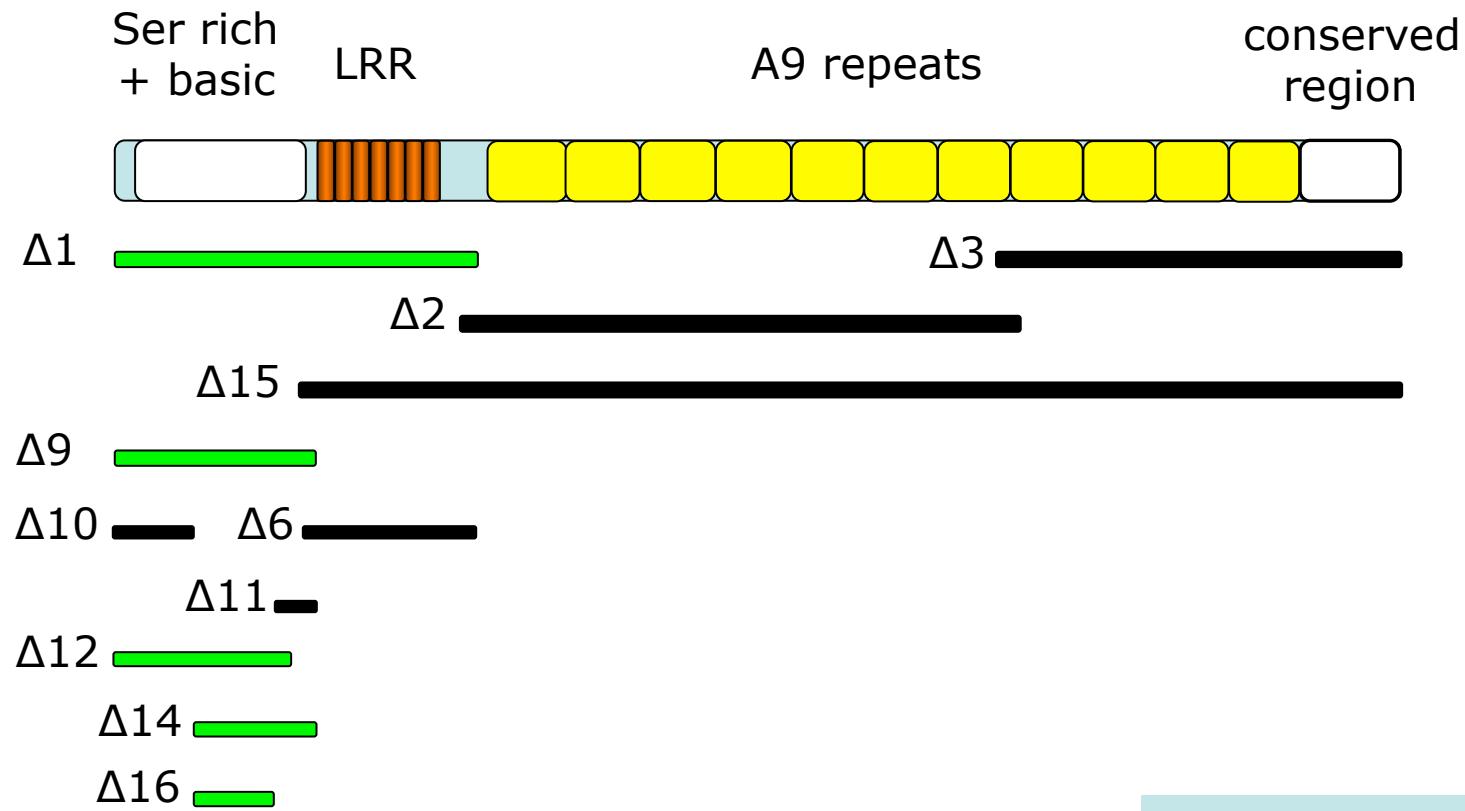
- 1) Identify sequence regions with low information content over a sequence window
- 2) Merge neighbouring regions

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:

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPPQLPQP

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

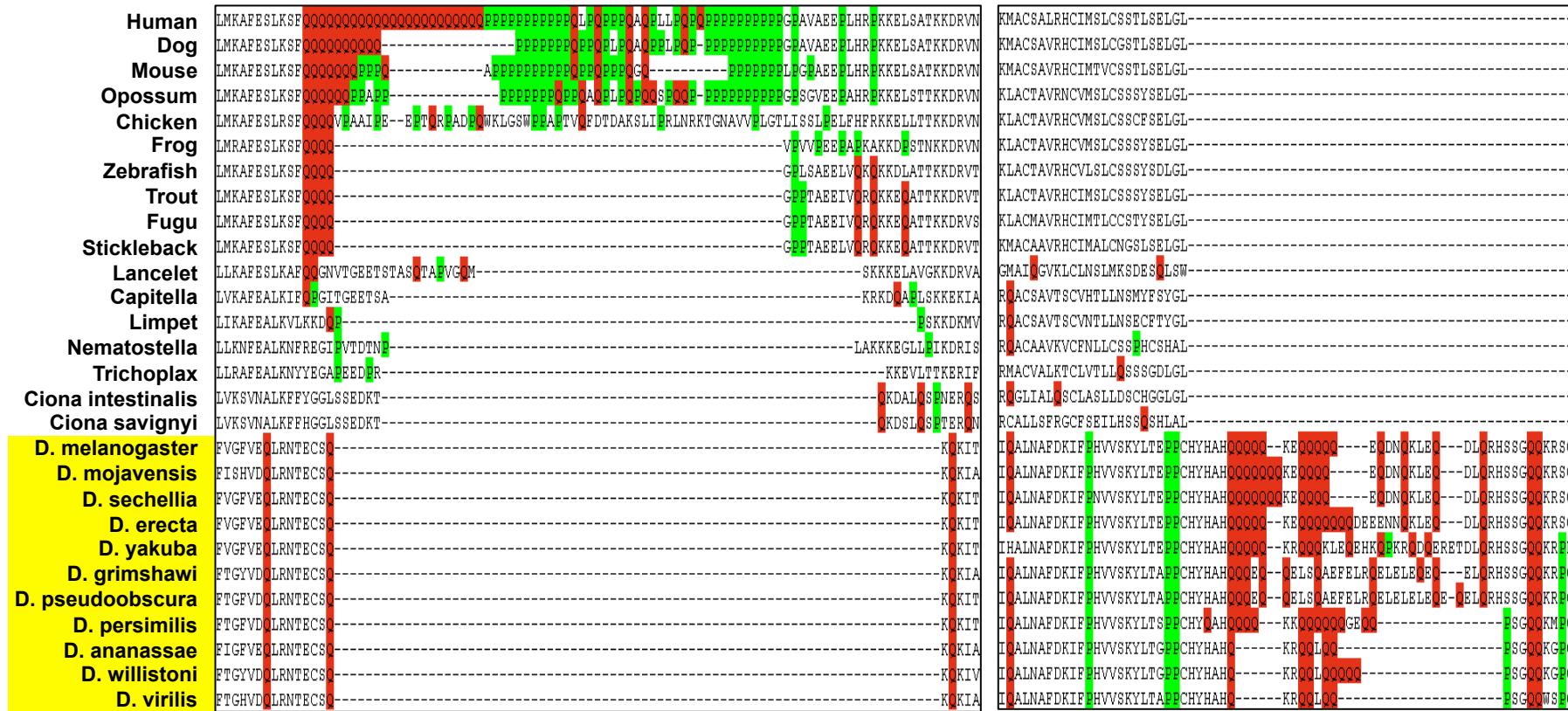
Homorepeats need to be studied in context

Function of polyQ

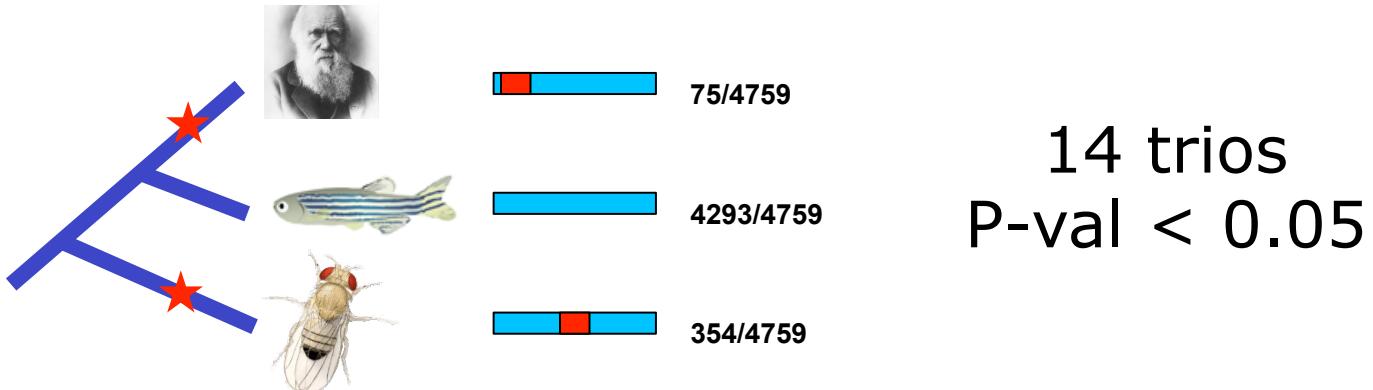
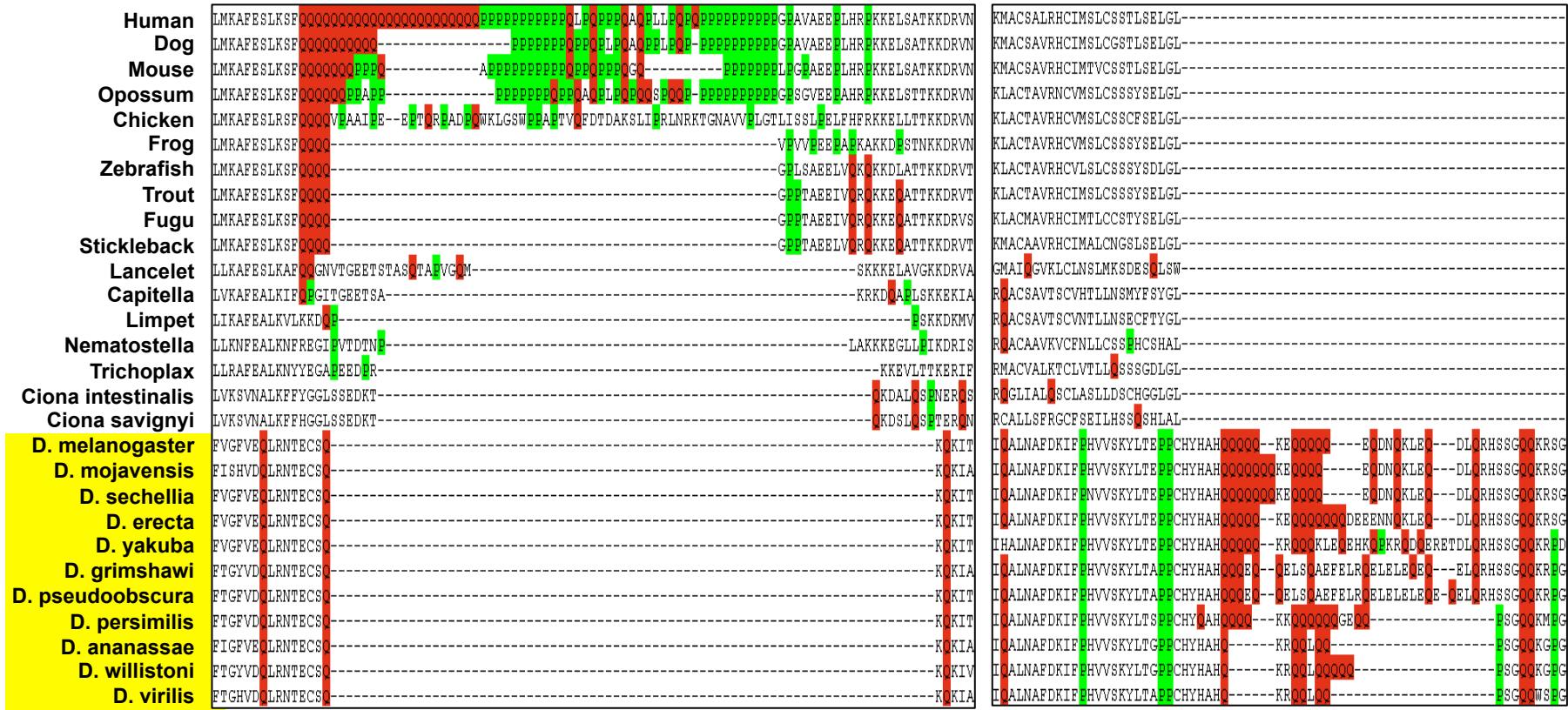
Martin
Schaefer

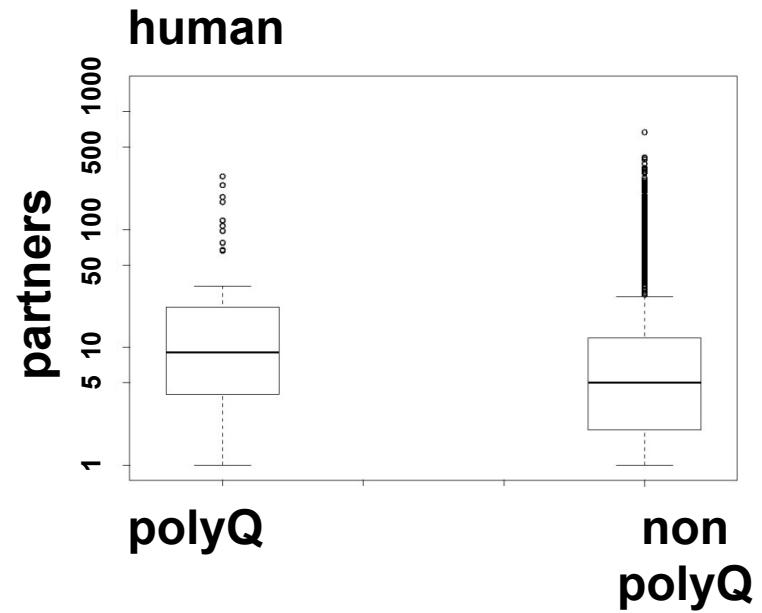


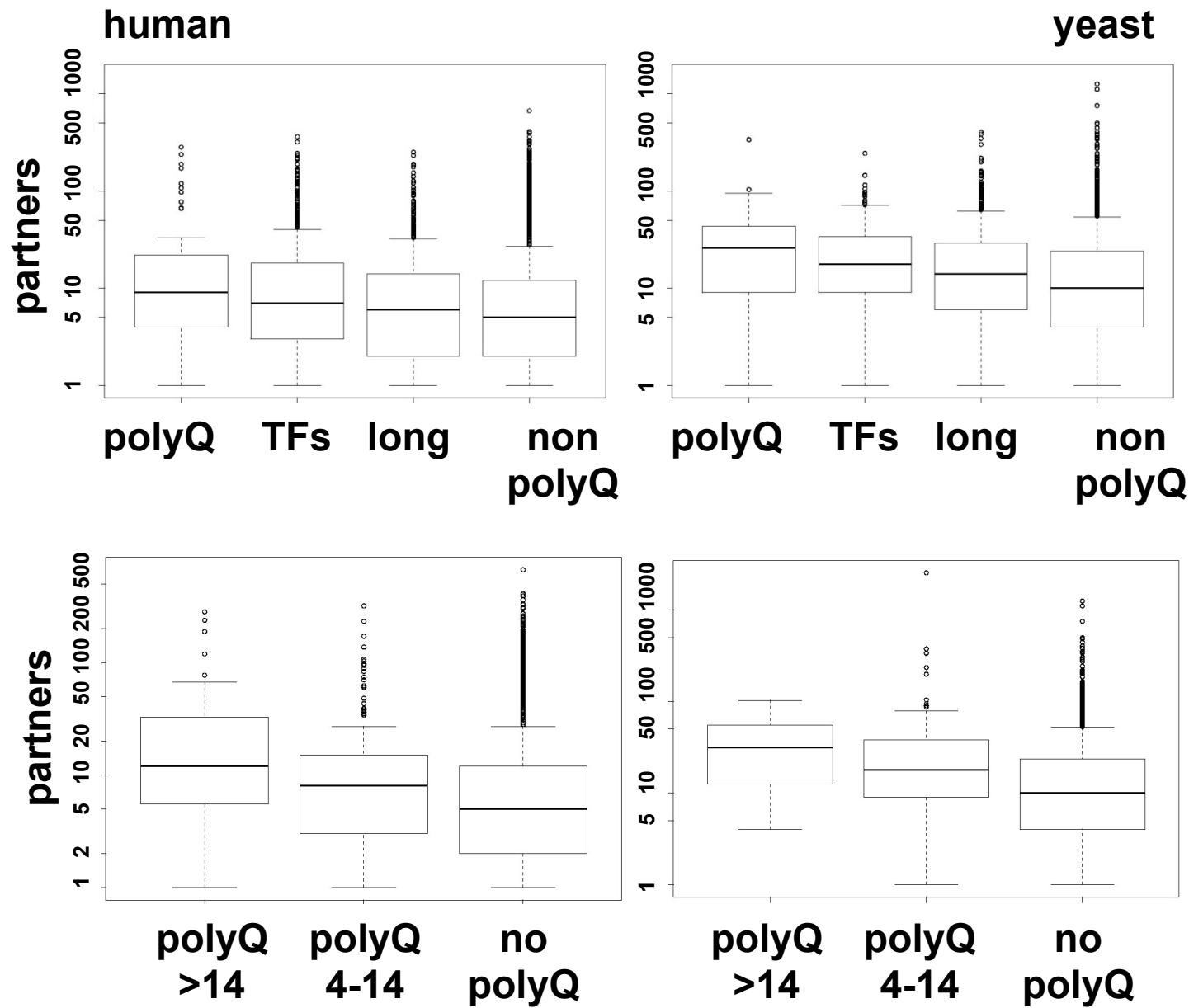
polyQ in Huntington



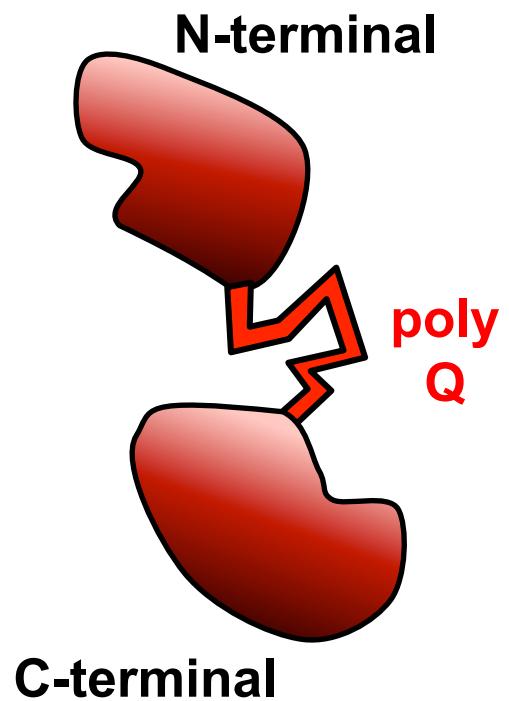
Schaefer et al (2012) *Nucleic Acids Res.*



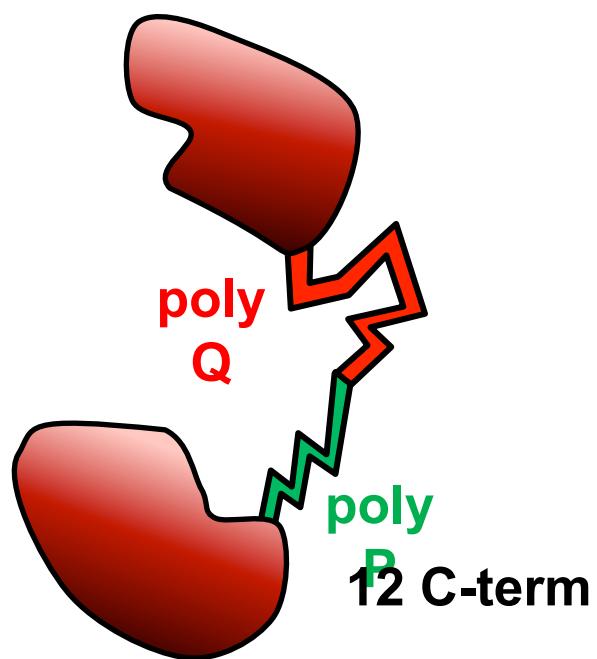




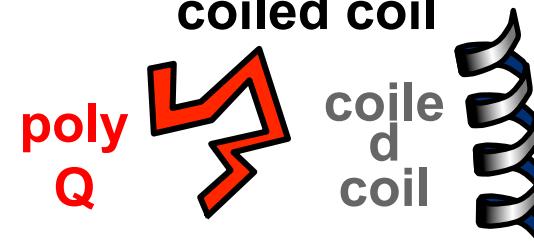
86 human
polyQ
proteins



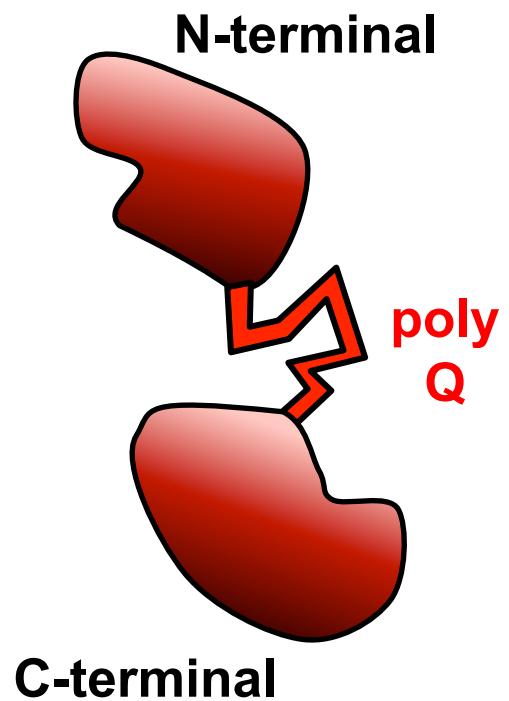
13 polyQ
proteins
with near
polyP



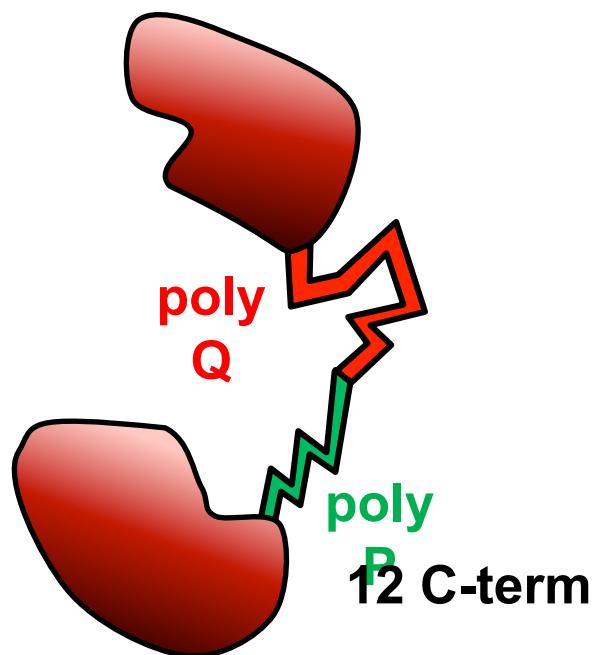
109 polyQ
regions
54 overlap/near
coiled coil



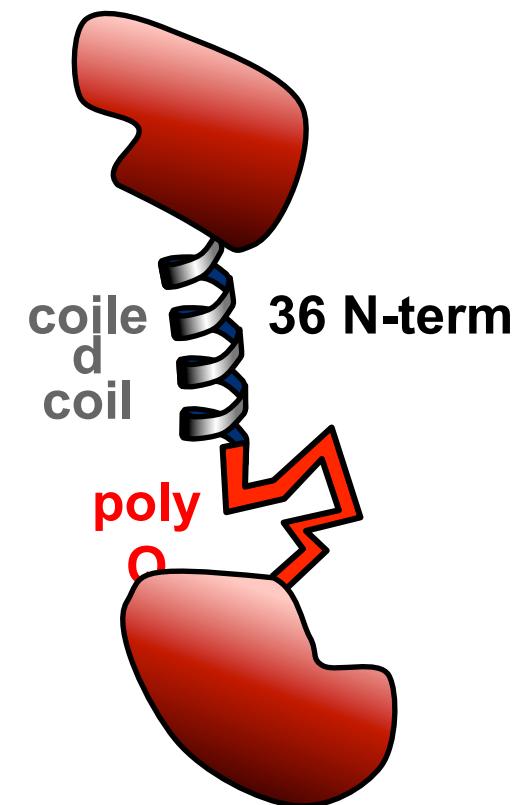
86 human
polyQ
proteins



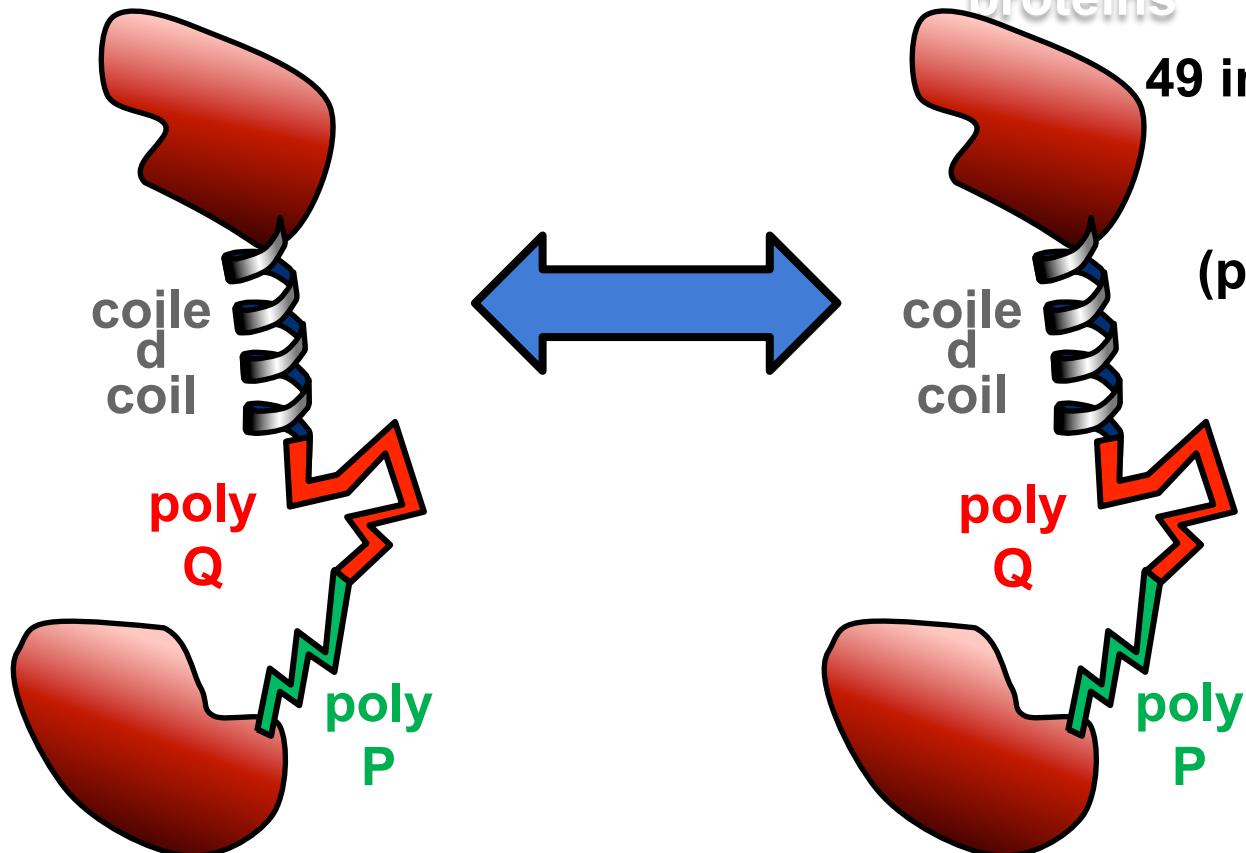
13 polyQ
proteins
with near
polyP



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

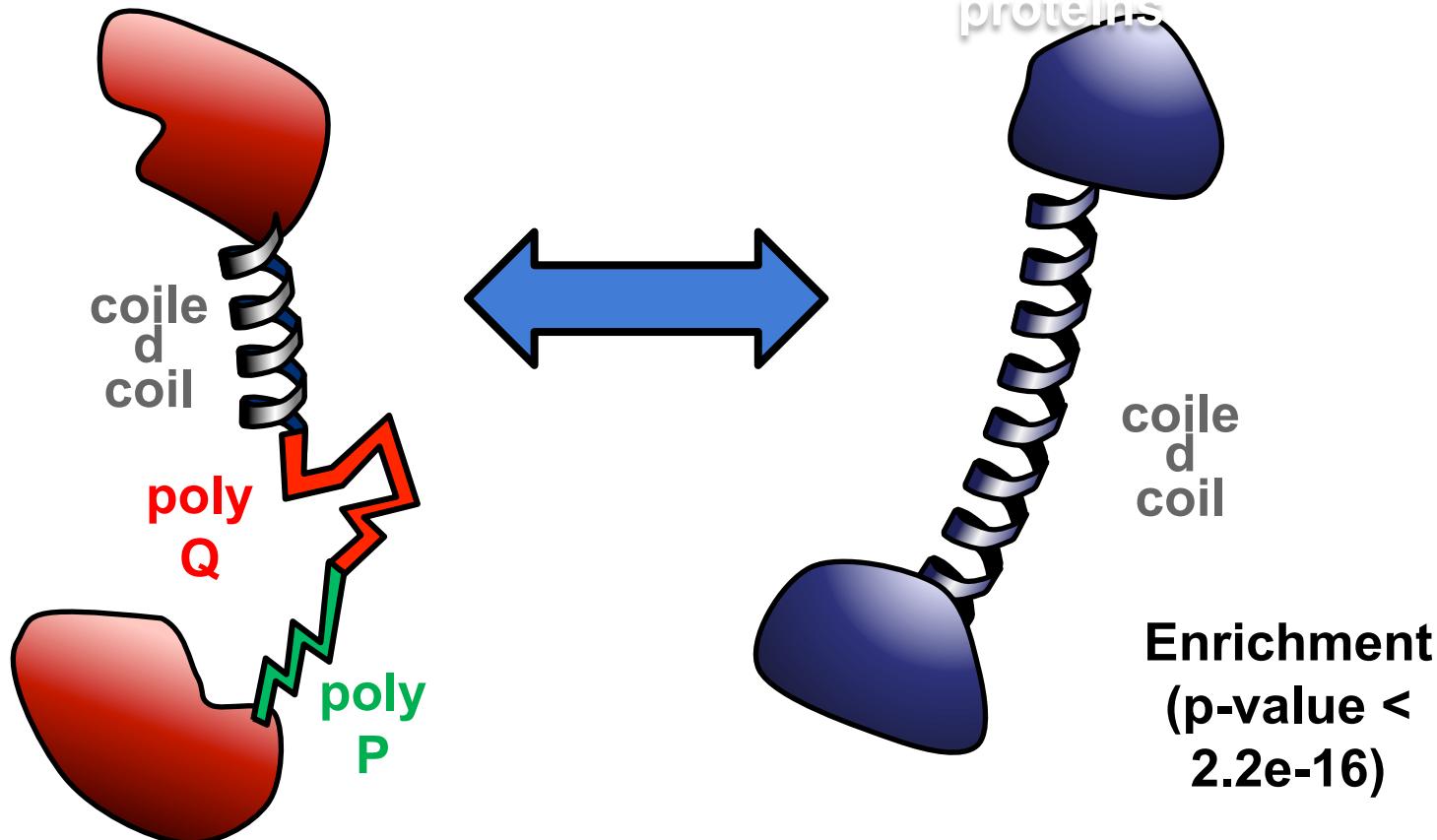


86 human
polyQ
proteins



86 human
polyQ
proteins

Non-polyQ
interacting



polyQ

unbound

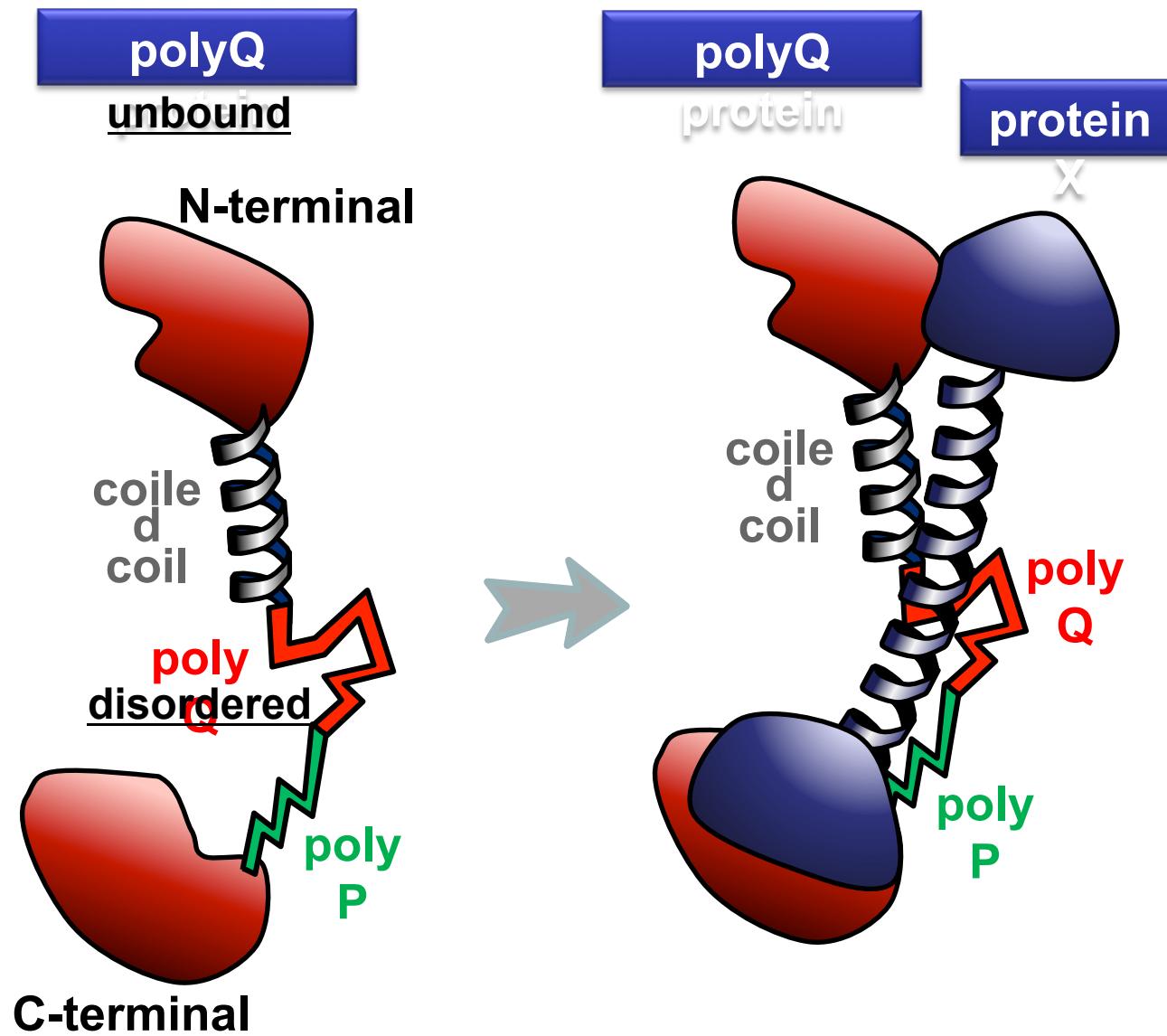
N-terminal

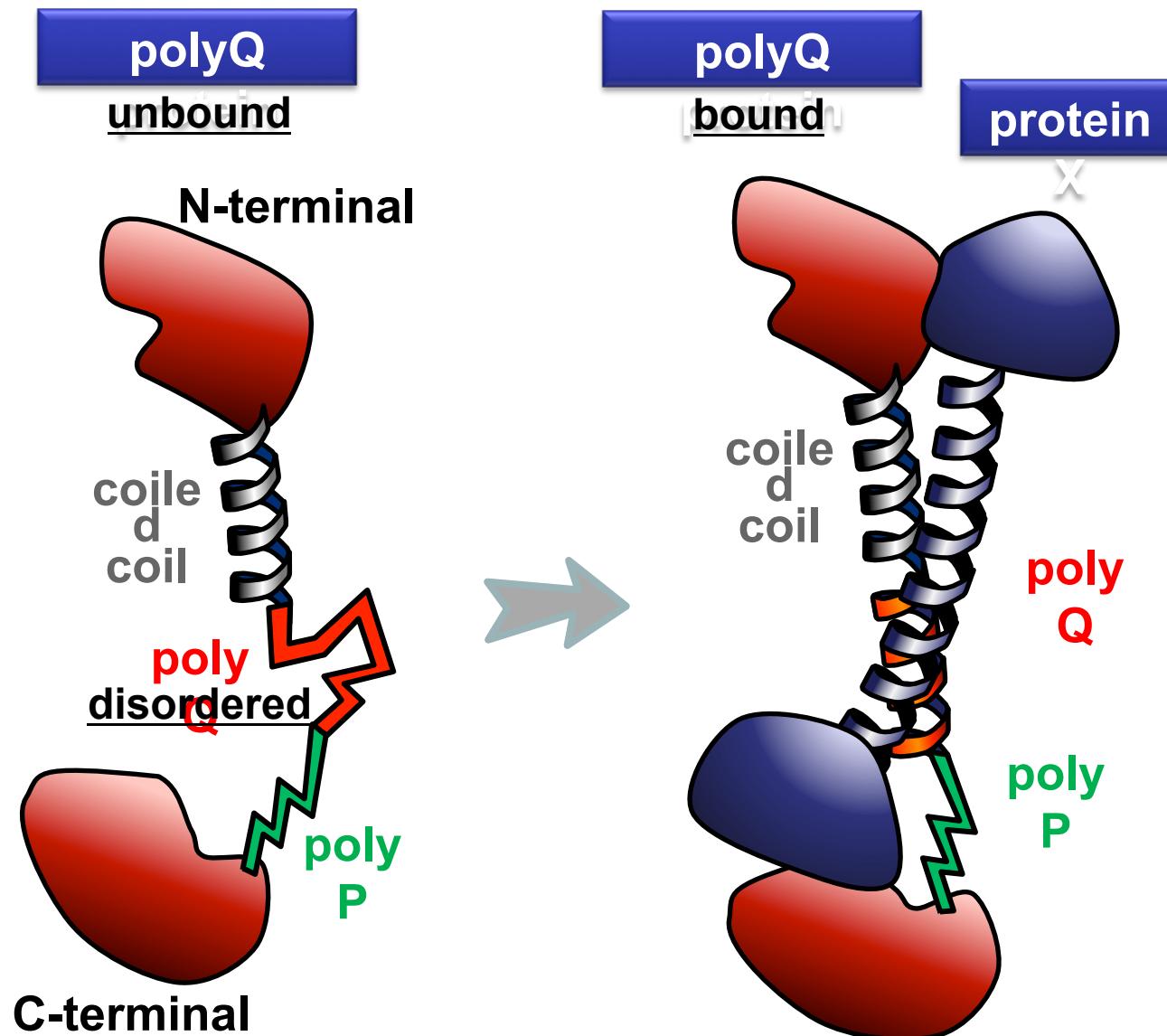
coiled
coil

poly
disordered

poly
P

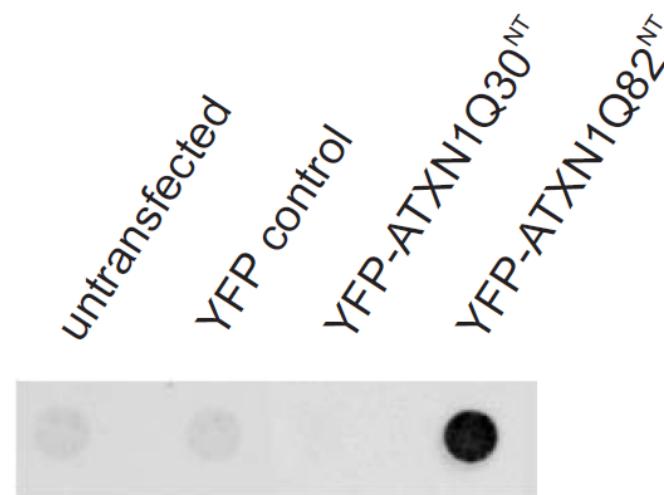
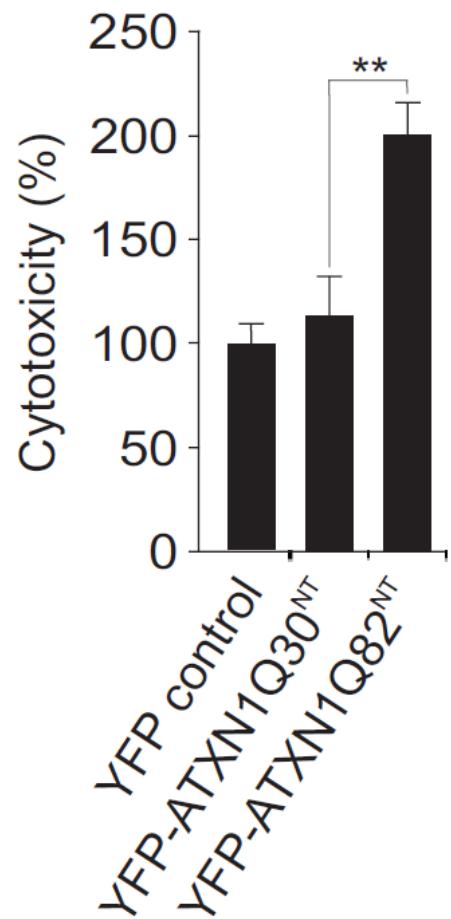
C-terminal





ATXN1Q82^{NT} is toxic

ATXN1Q82^{NT} aggregates



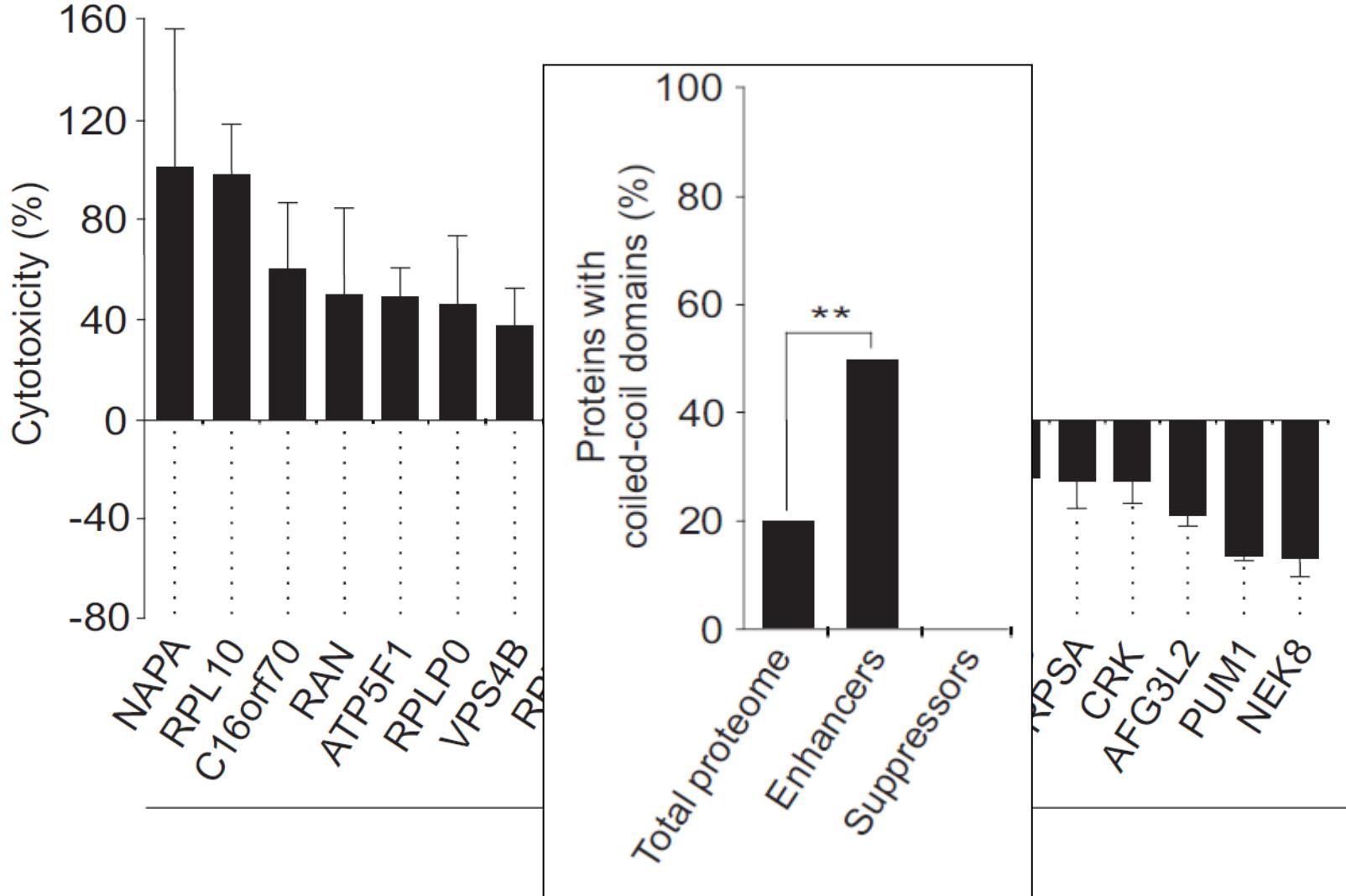
Erich
Wanker

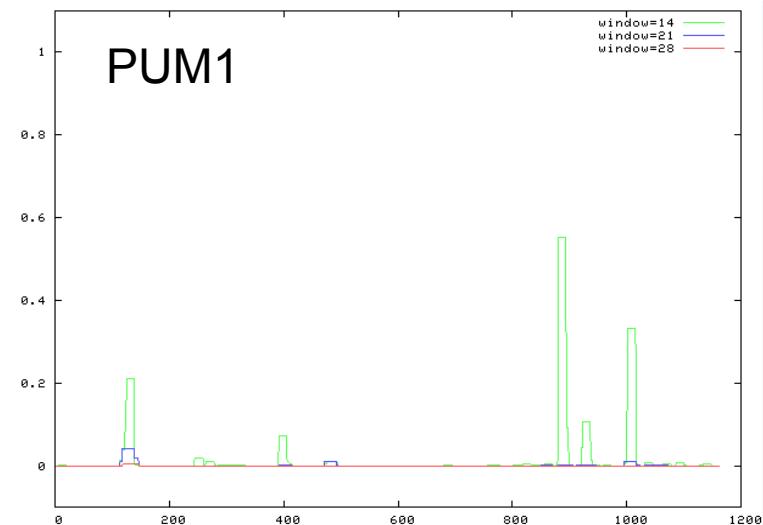
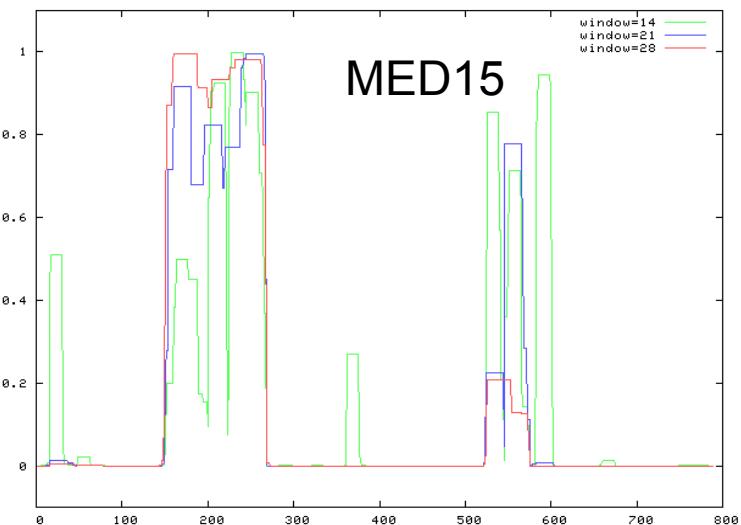


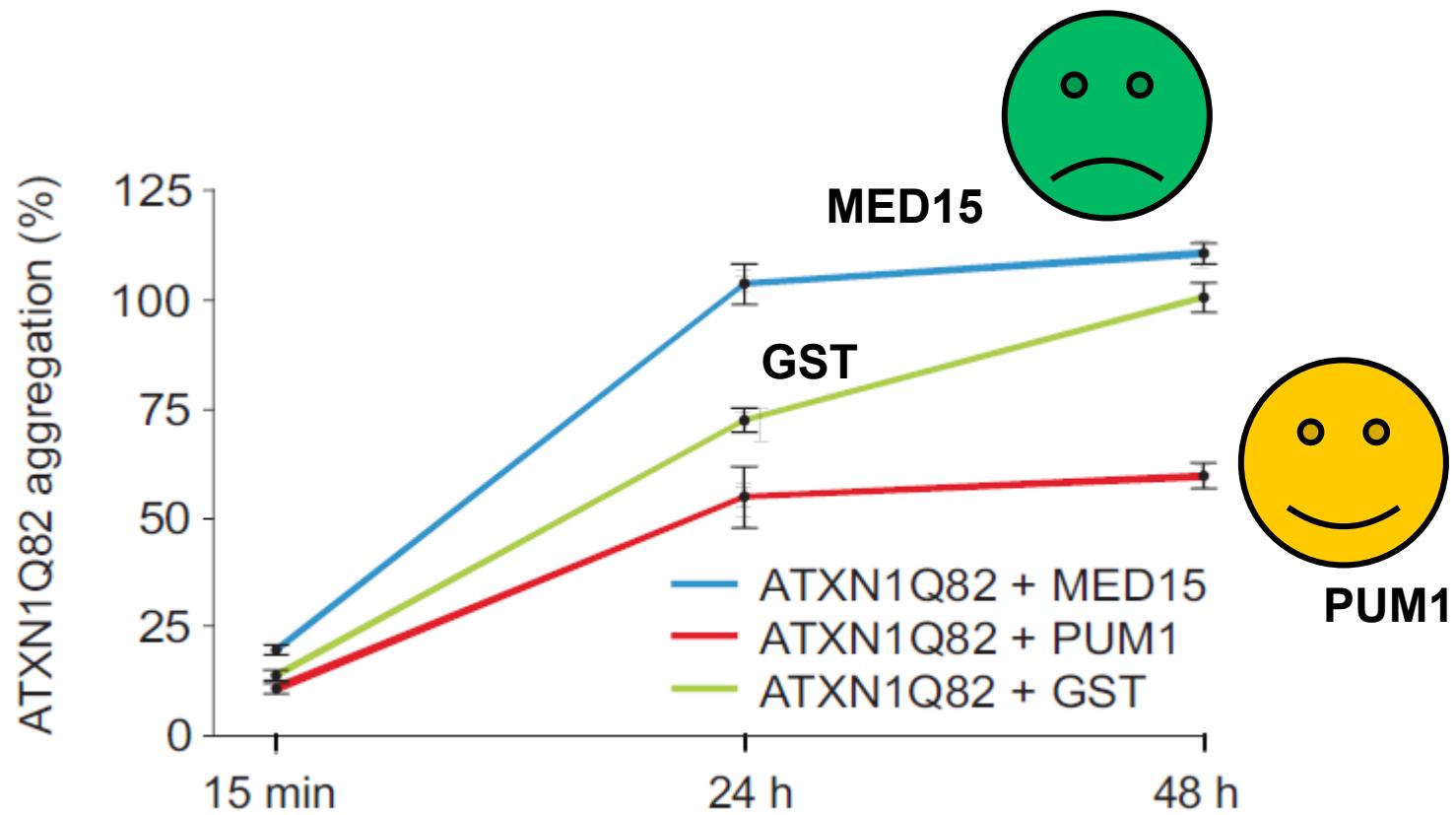
Spyros
Petrakis

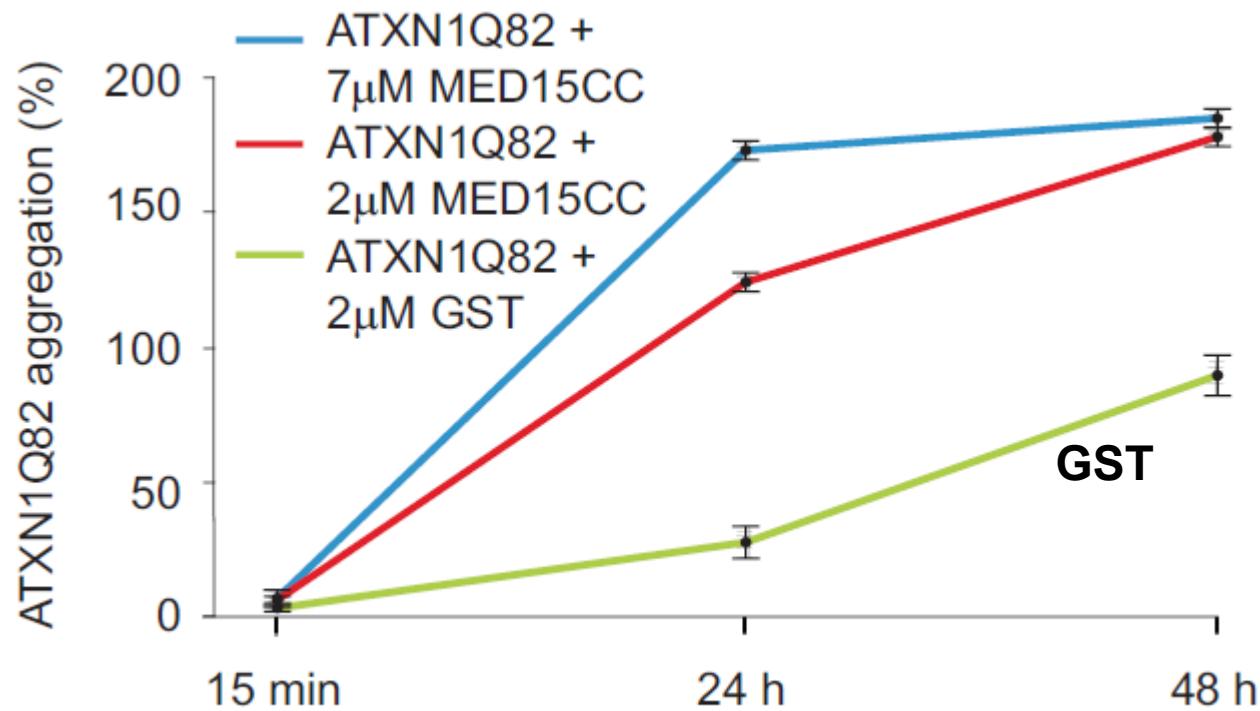
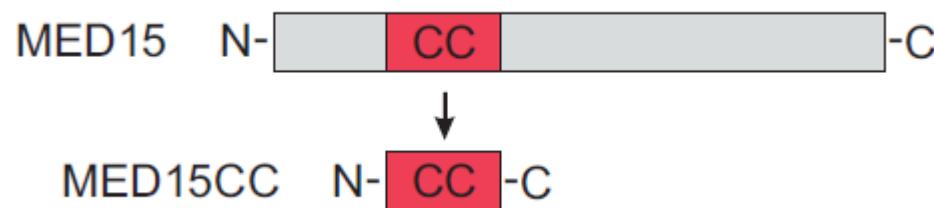


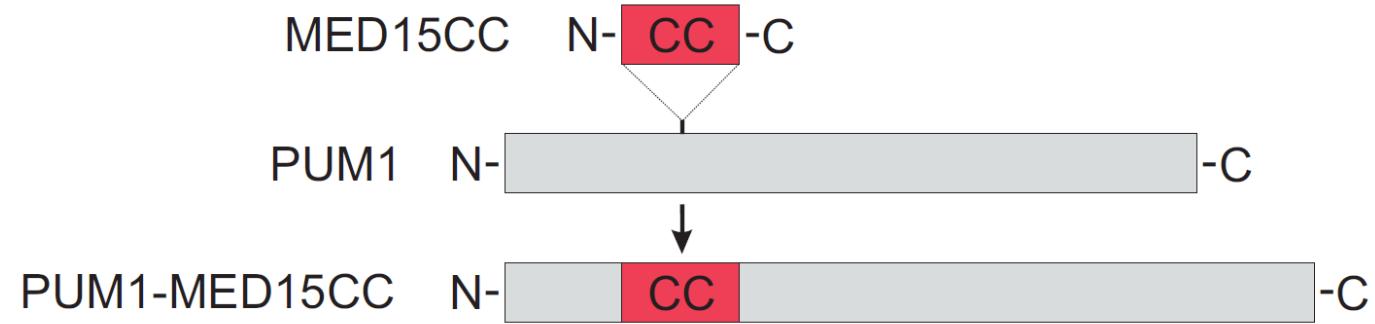
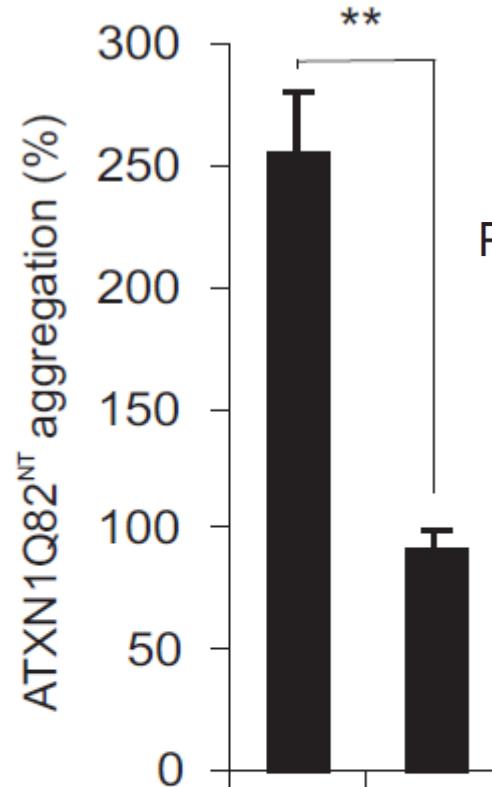
interactors that change ATXN1Q82^{NT} toxicity

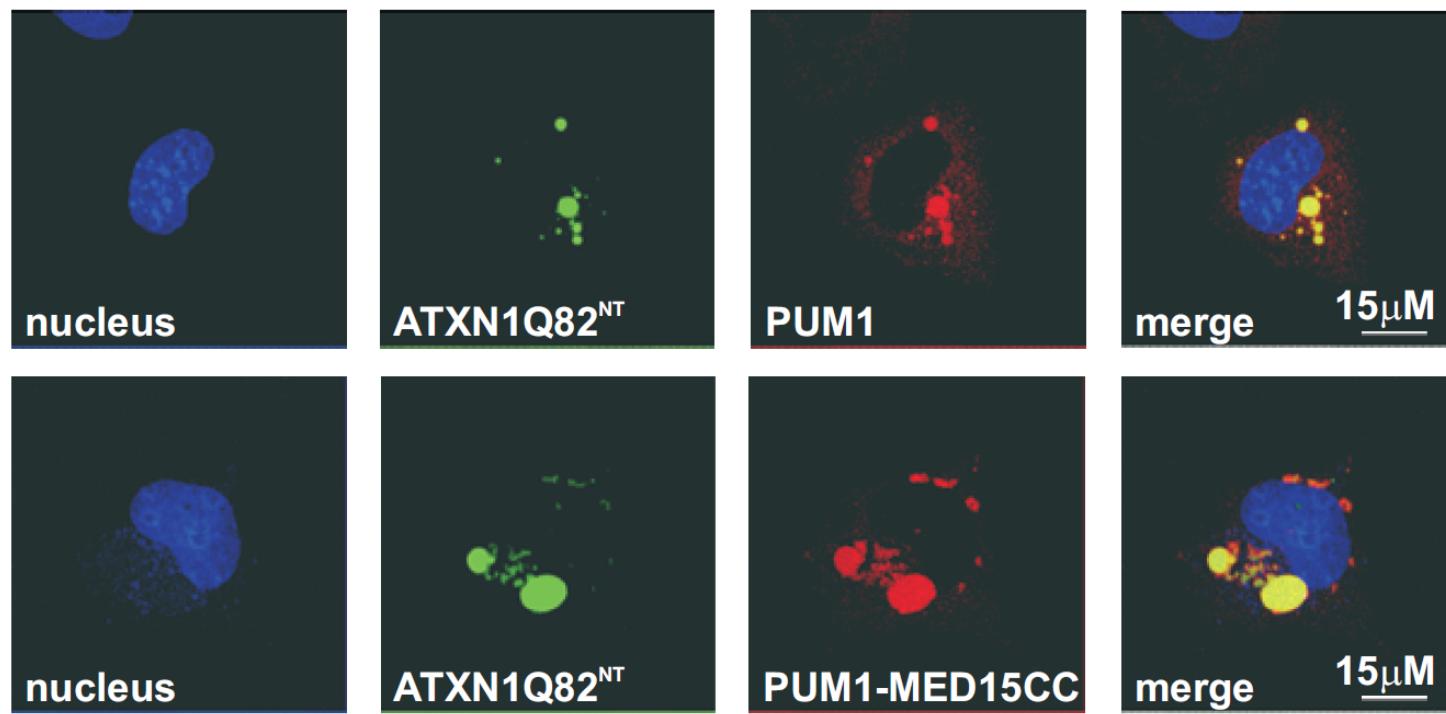
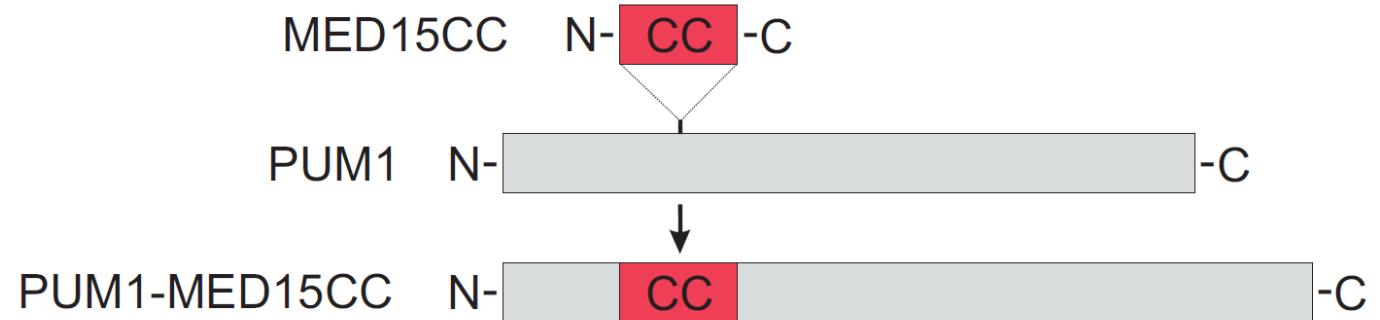




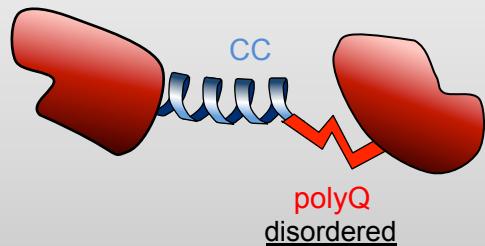




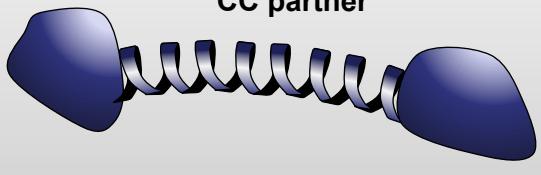




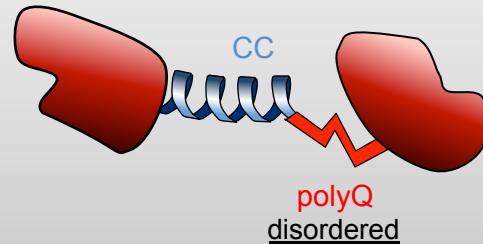
Normal polyQ protein



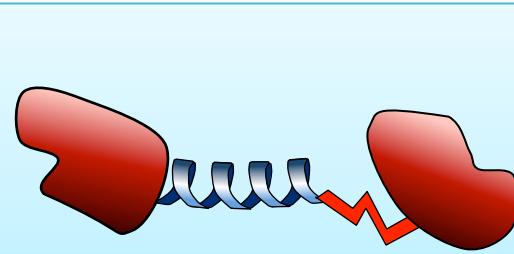
CC partner



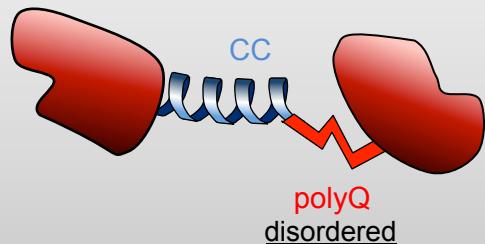
Normal polyQ protein



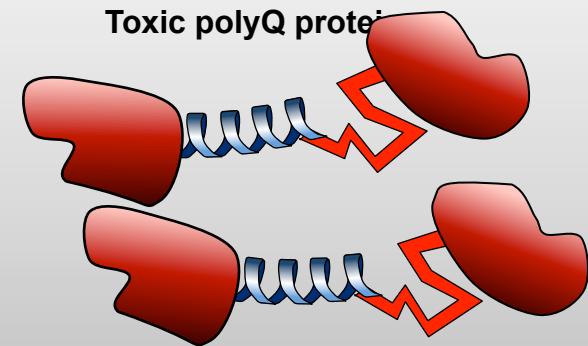
CC partner



Normal polyQ protein



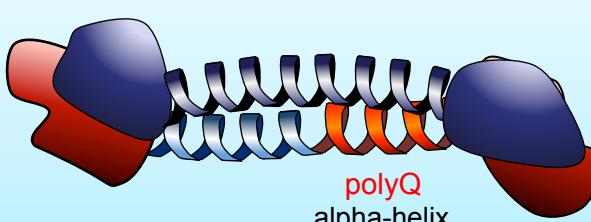
Toxic polyQ protein

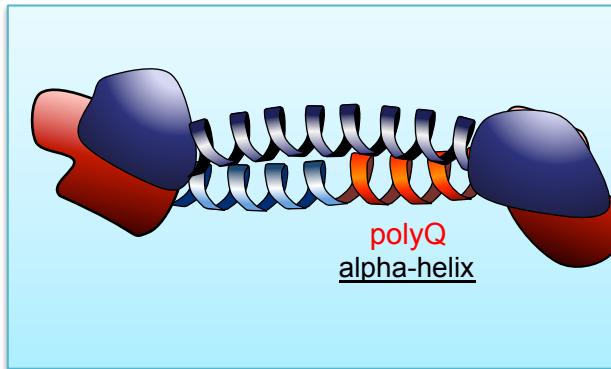
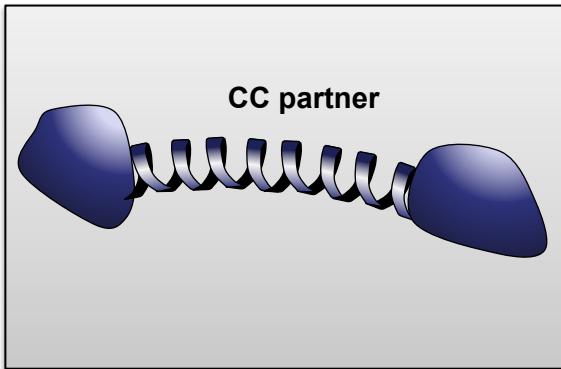
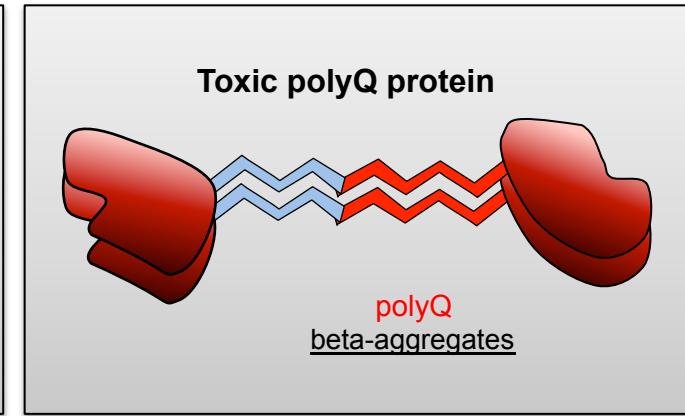
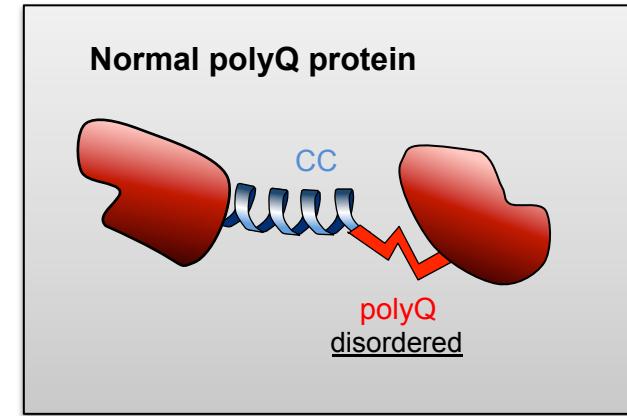


CC partner

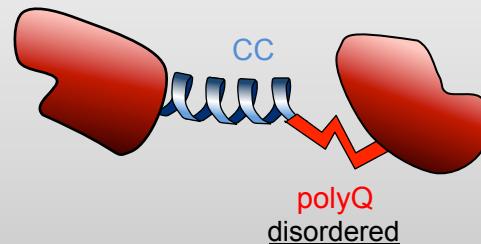


polyQ
alpha-helix

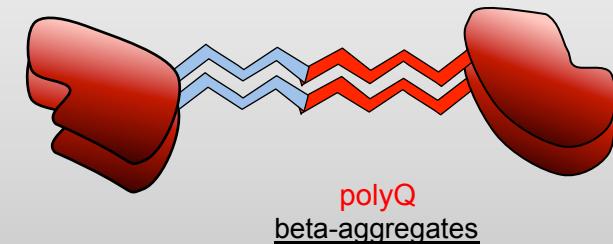




Normal polyQ protein



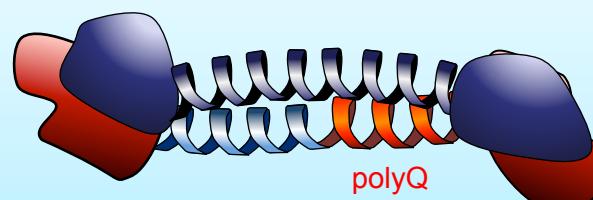
Toxic polyQ protein



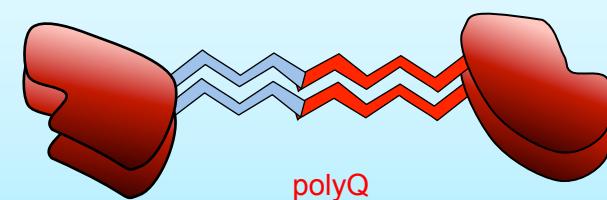
CC partner



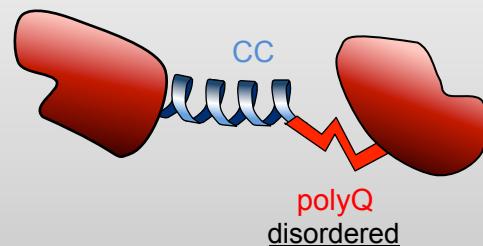
**polyQ
alpha-helix**



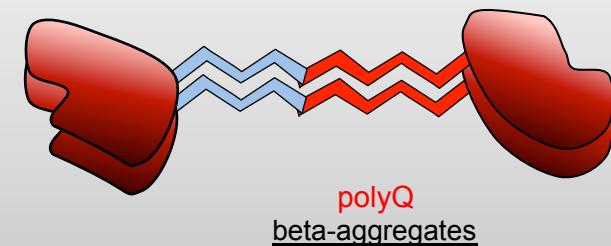
**polyQ
beta-aggregates**



Normal polyQ protein



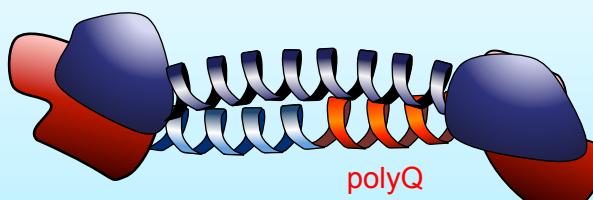
Toxic polyQ protein



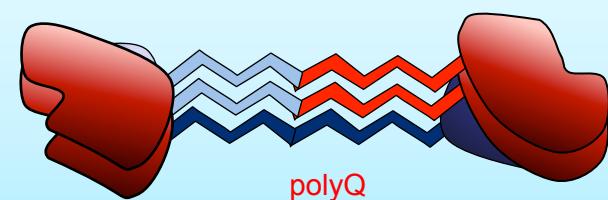
CC partner



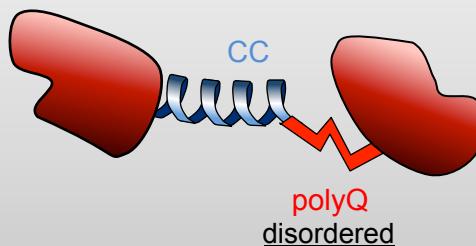
**polyQ
alpha-helix**



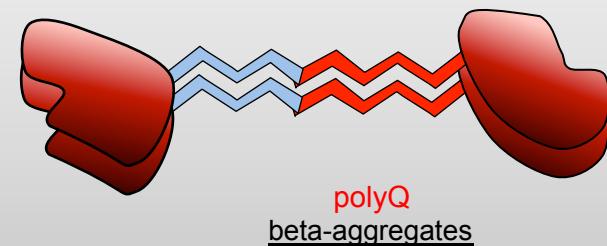
**polyQ
beta-aggregates**



Normal polyQ protein



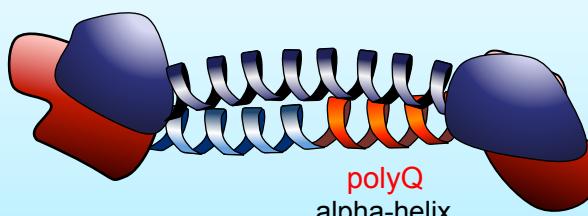
Toxic polyQ protein



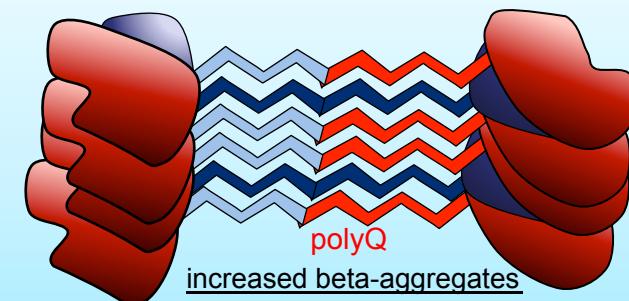
CC partner



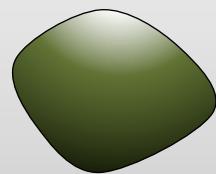
**polyQ
alpha-helix**



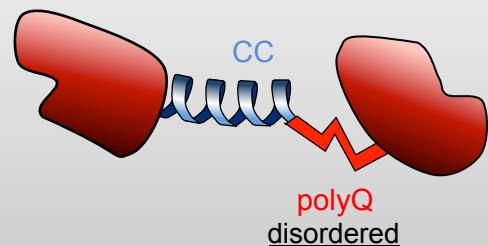
increased beta-aggregates



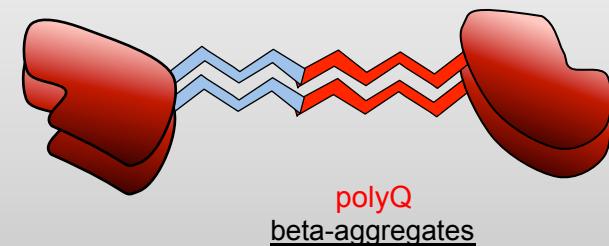
non-CC partner



Normal polyQ protein



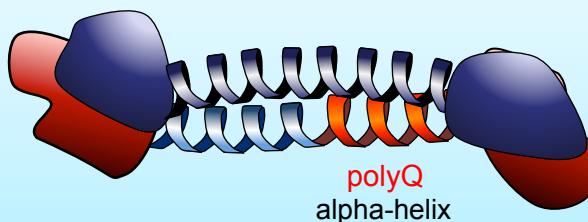
Toxic polyQ protein



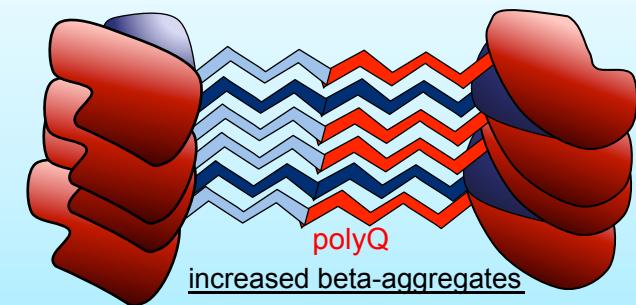
CC partner



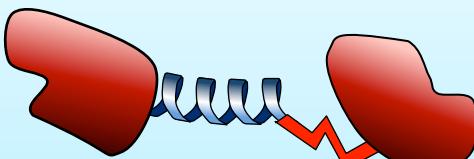
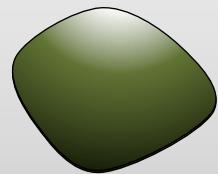
**polyQ
alpha-helix**



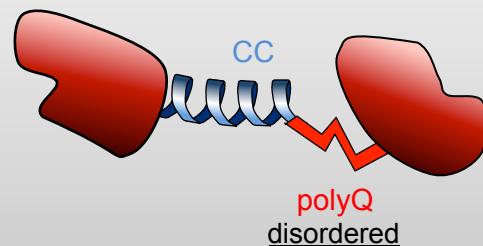
**polyQ
increased beta-aggregates**



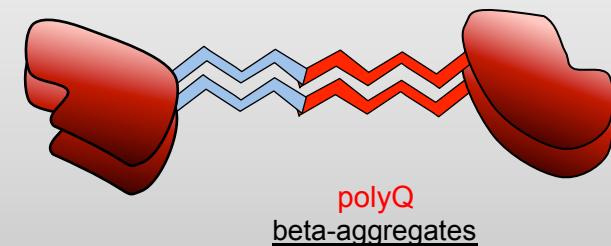
non-CC partner



Normal polyQ protein



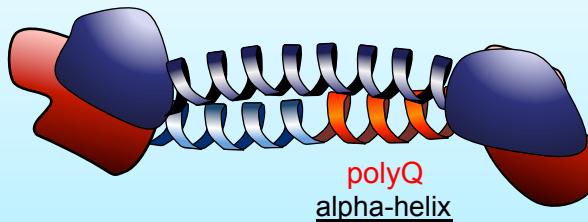
Toxic polyQ protein



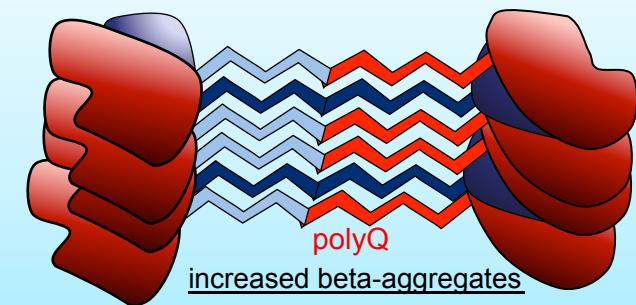
CC partner



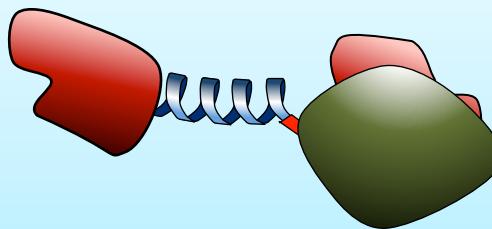
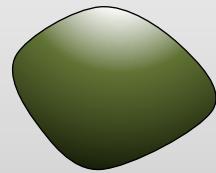
**polyQ
alpha-helix**



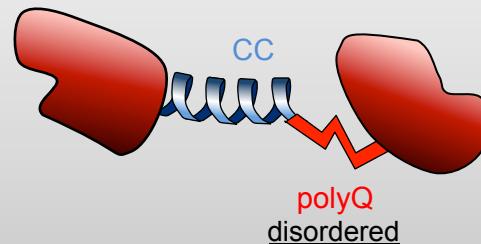
**polyQ
increased beta-aggregates**



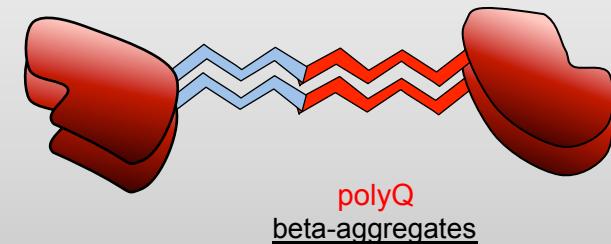
non-CC partner



Normal polyQ protein



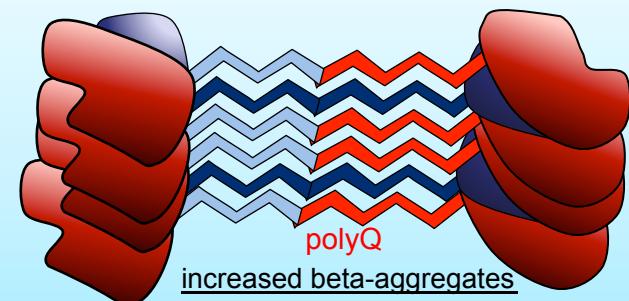
Toxic polyQ protein



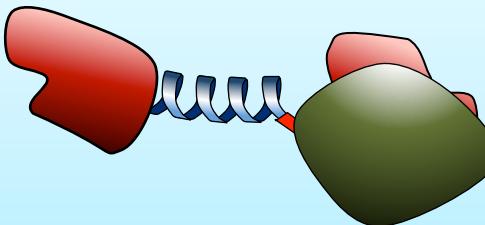
CC partner



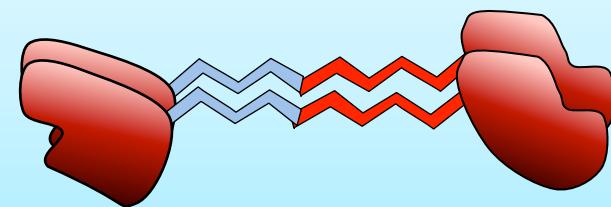
**polyQ
alpha-helix**



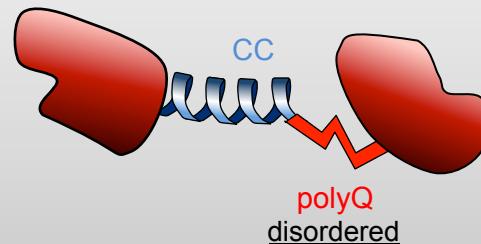
non-CC partner



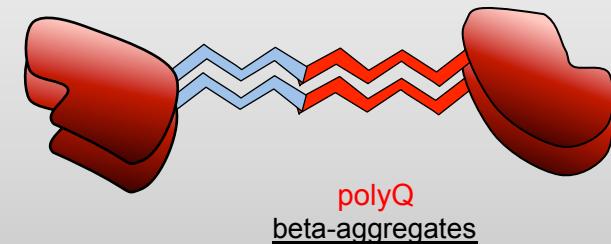
increased beta-aggregates



Normal polyQ protein



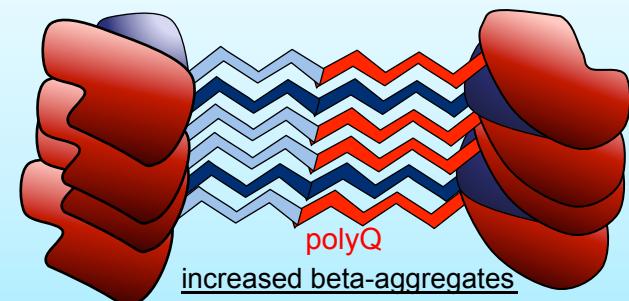
Toxic polyQ protein



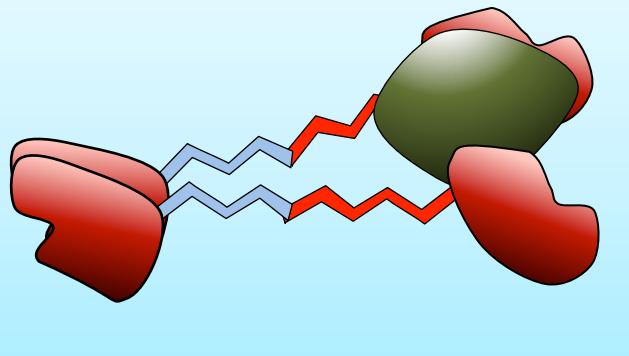
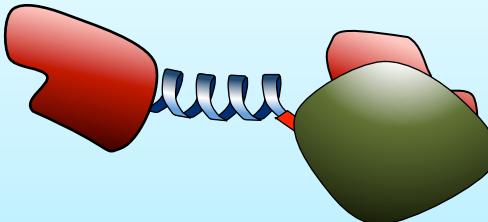
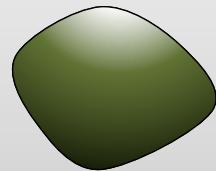
CC partner



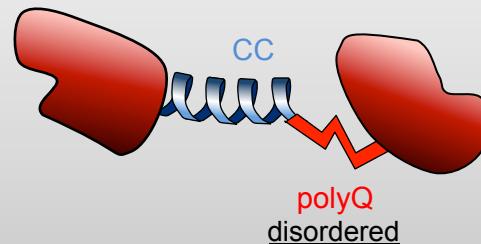
**polyQ
alpha-helix**



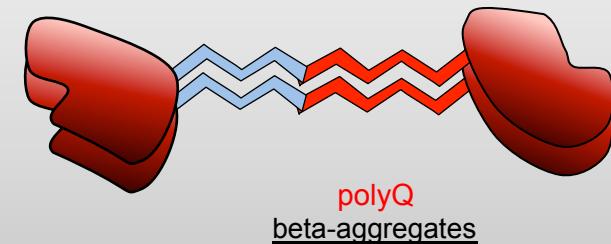
non-CC partner



Normal polyQ protein



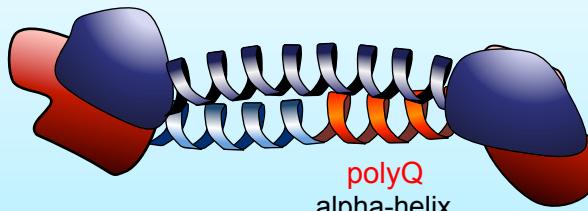
Toxic polyQ protein



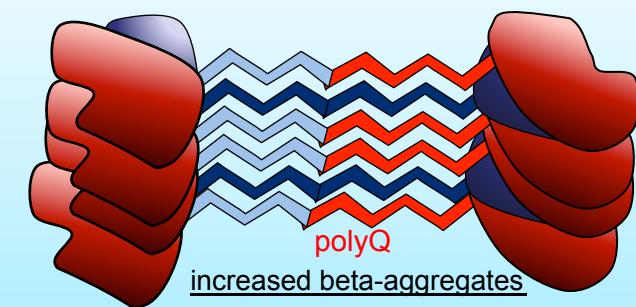
CC partner



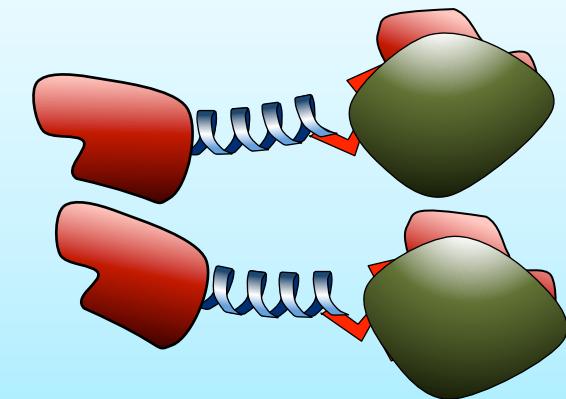
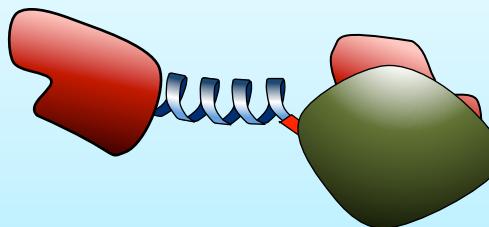
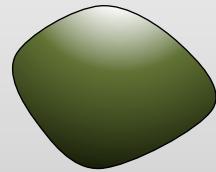
**polyQ
alpha-helix**



**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.

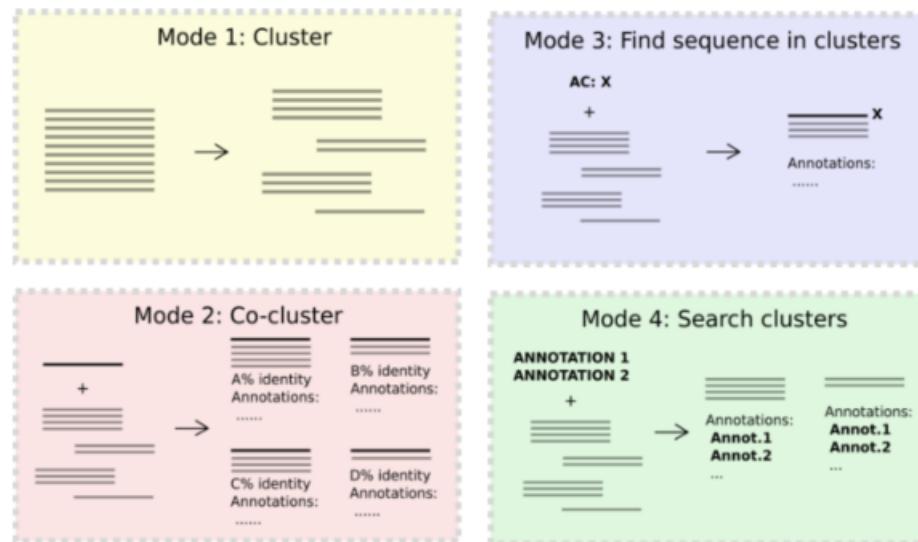
Do you see any other biased region nearby?

Clustering proteins

Pablo Mier



FastaHerder2



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*)*

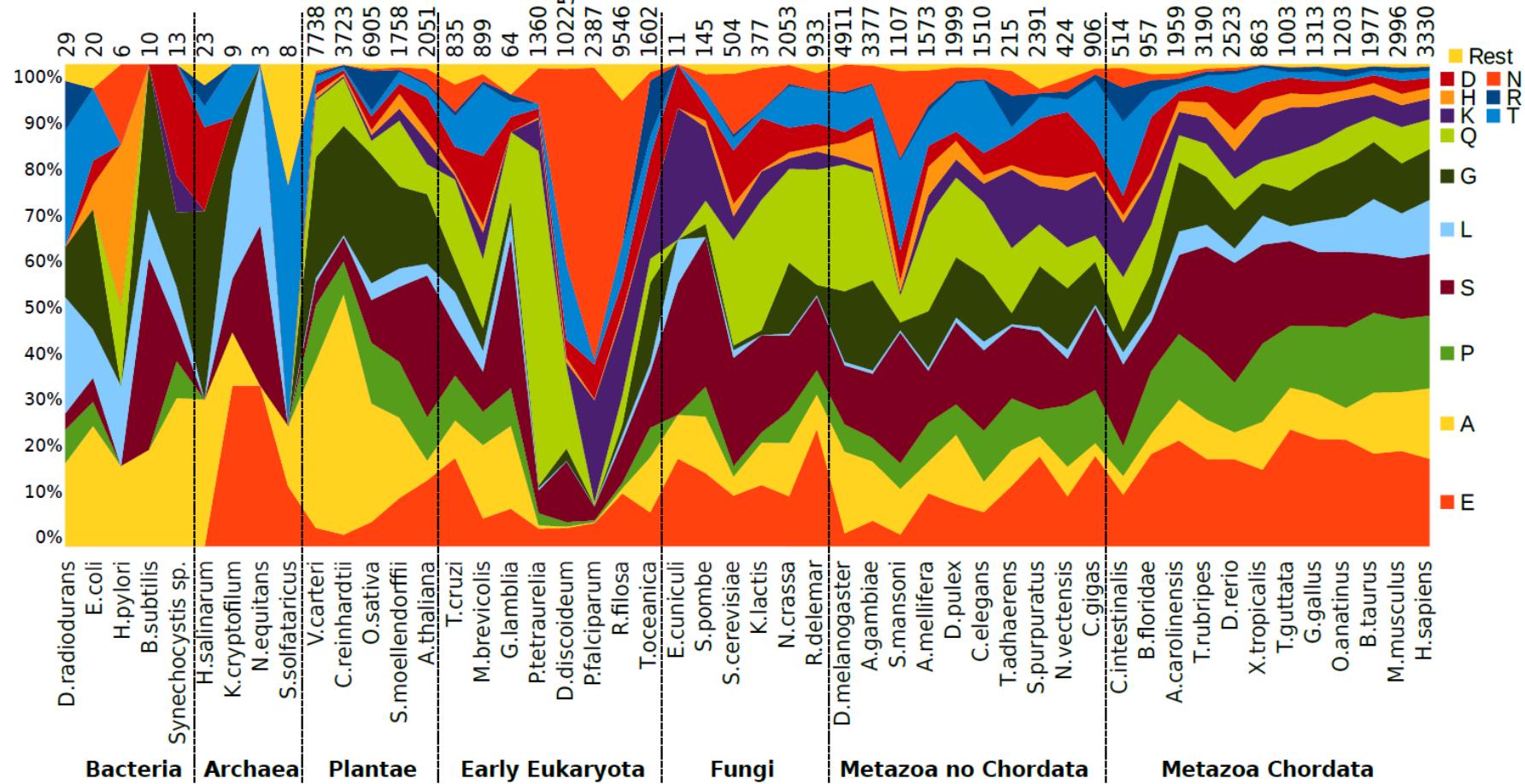
*if more than one, separate them by "+", e.g. 9606+Homo

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

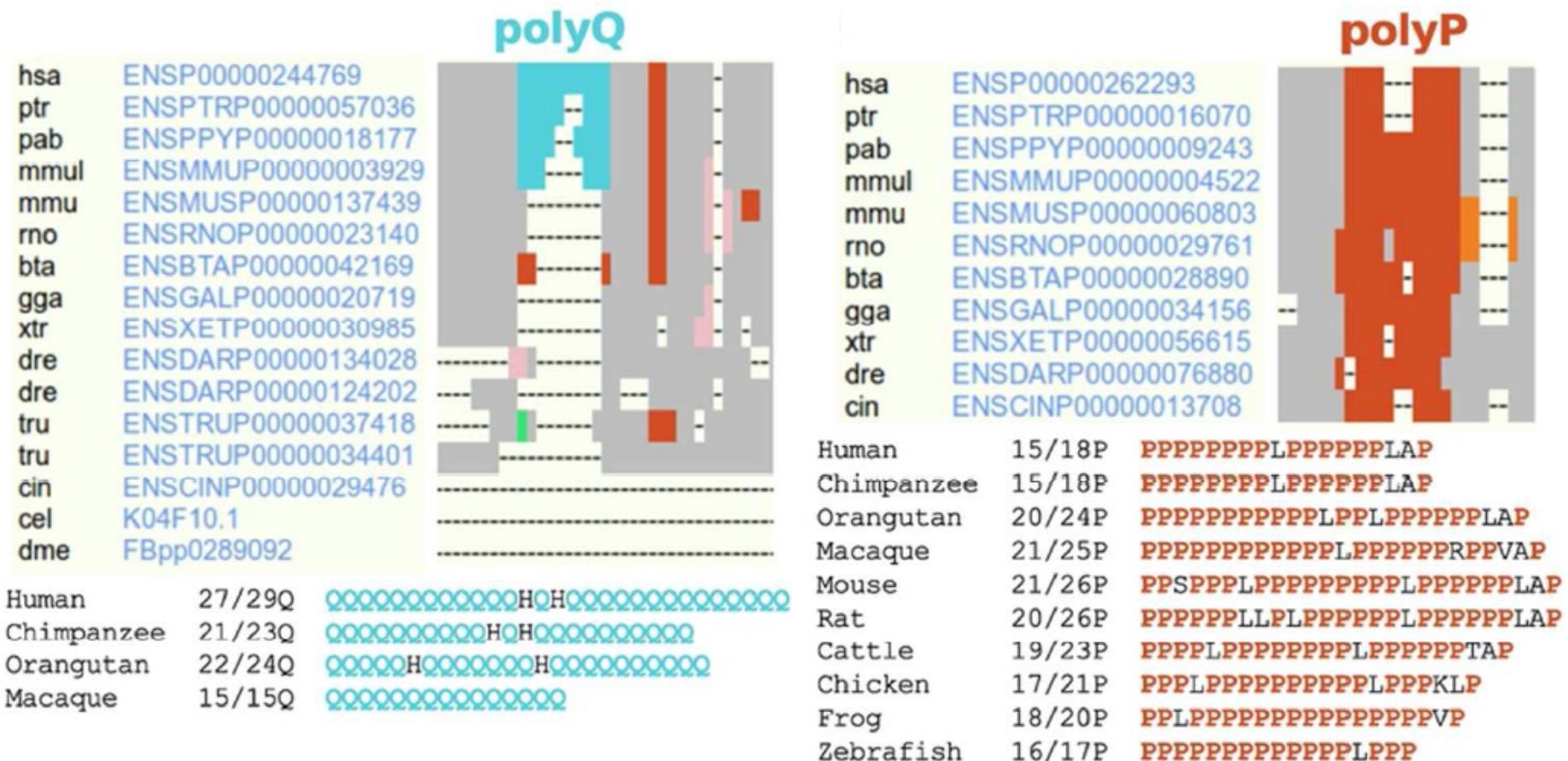
Context and evolution of homorepeats



dAtabase of PolyX Evolution

Mier *et al.* (2016) Bioinformatics

Context and evolution of homorepeats



dAtabase of PolyX Evolution

Mier et al. (2016) Bioinformatics