Repeats and composition bias

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

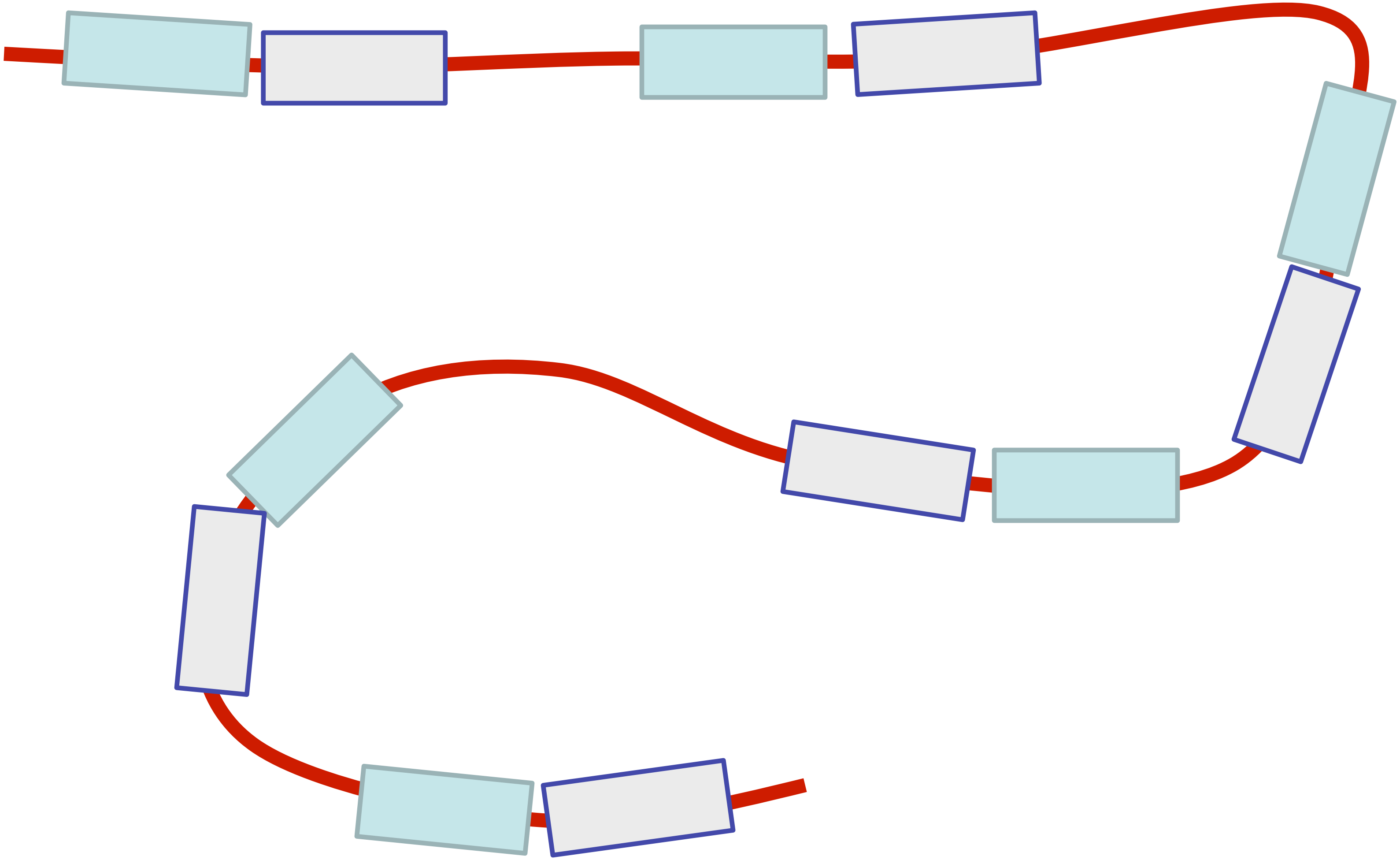
VG**S**PL**SSP**LS S

MK**S**SI**SSP**PS HCS

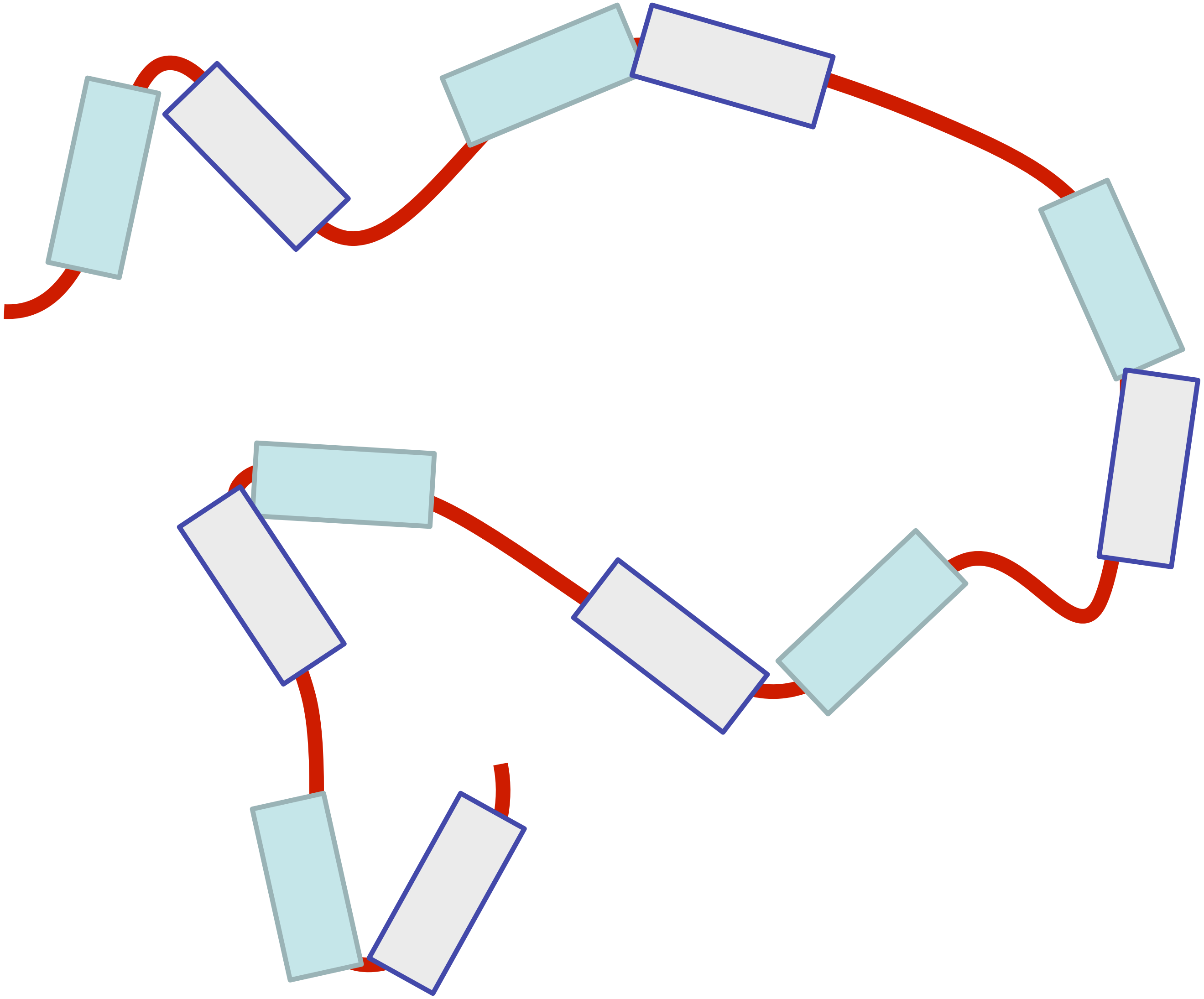
VK**S**P**VSSP**NN VT

LR**S**S**VSSP**AN INN

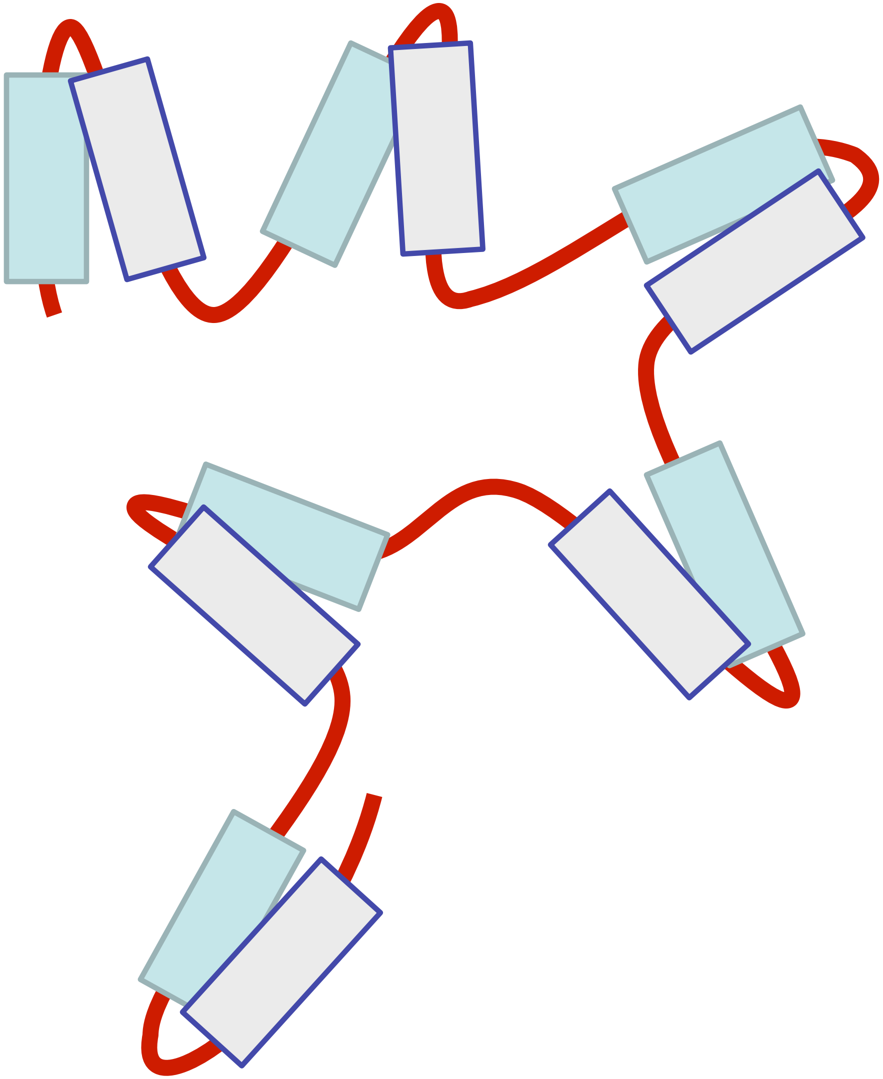
Tandem repeats fold together



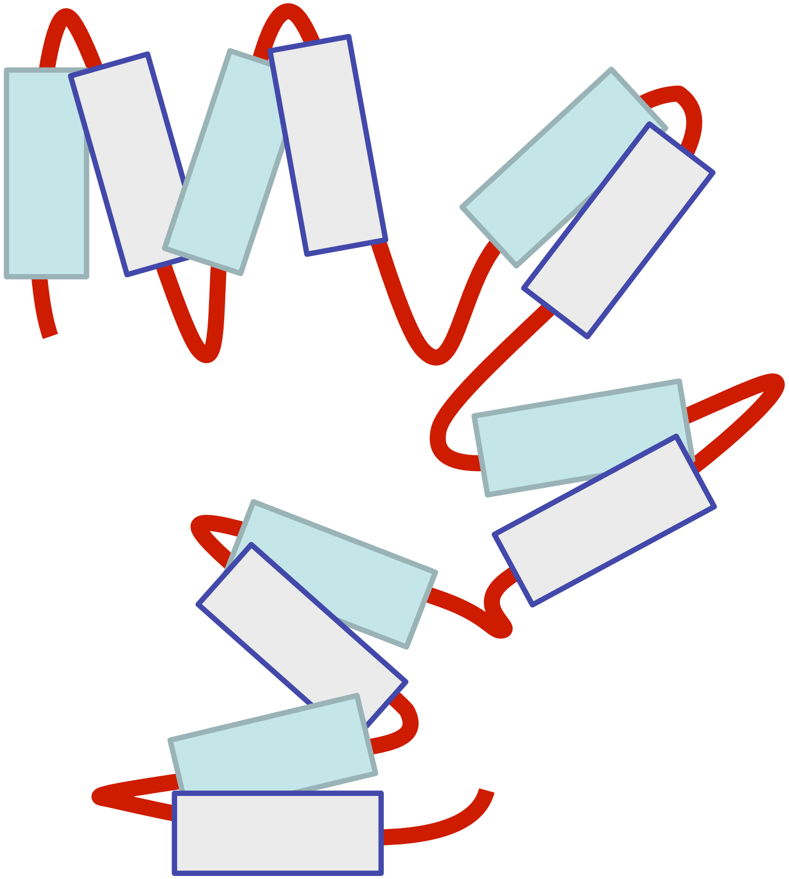
Tandem repeats fold together



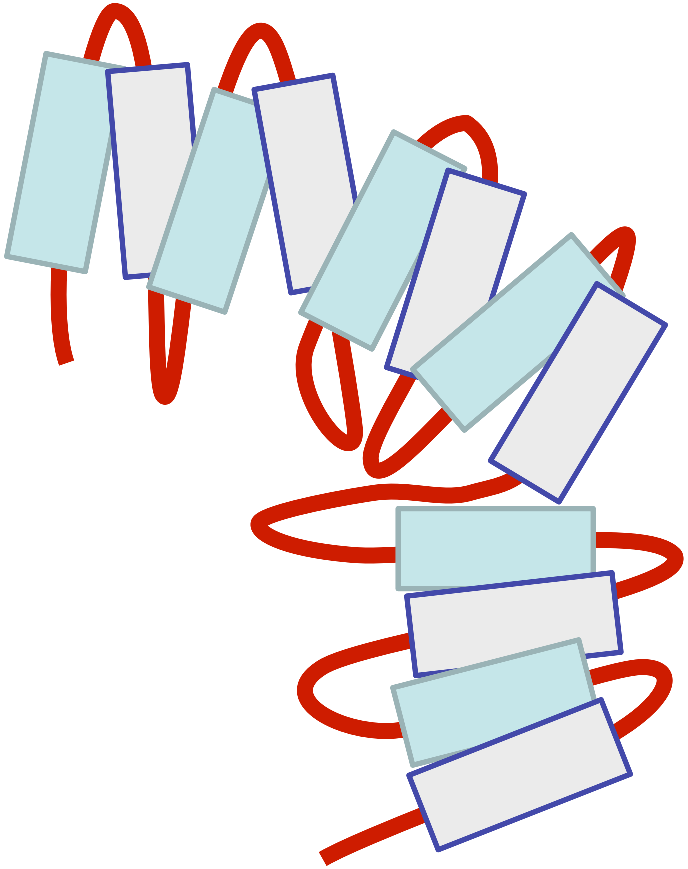
Tandem repeats fold together



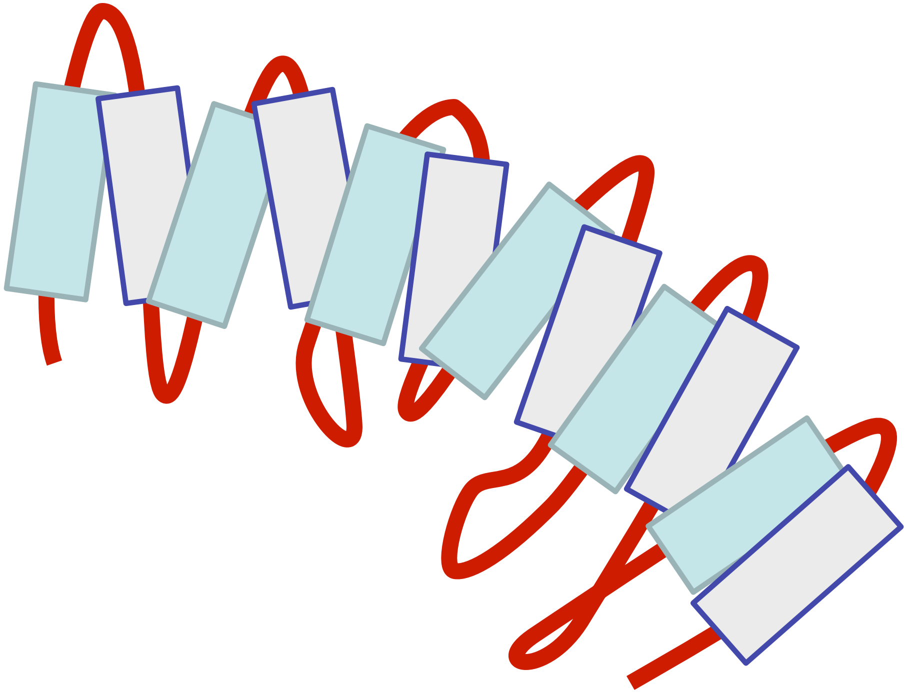
Tandem repeats fold together



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

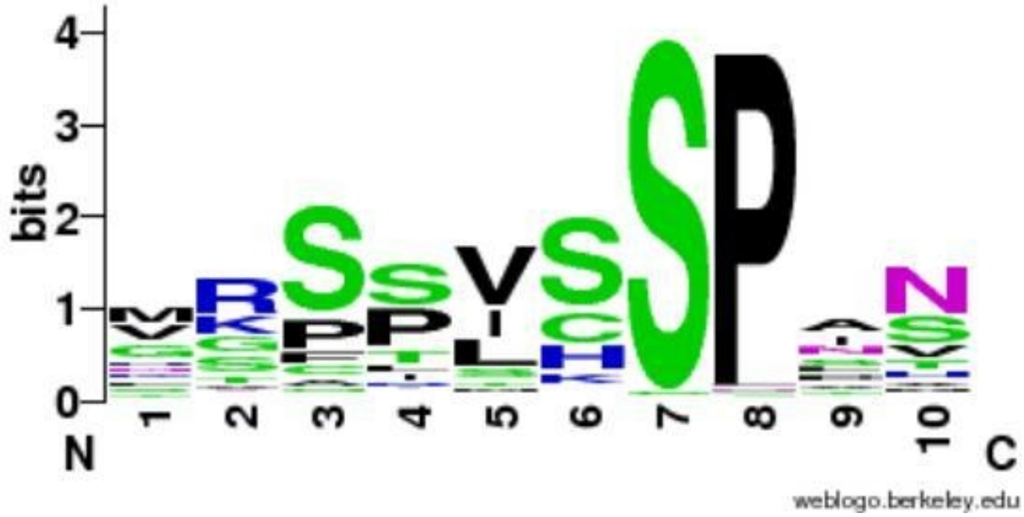
RG**S**RSH**SP**AH ASN

VG**S**PL**SSP**LS S

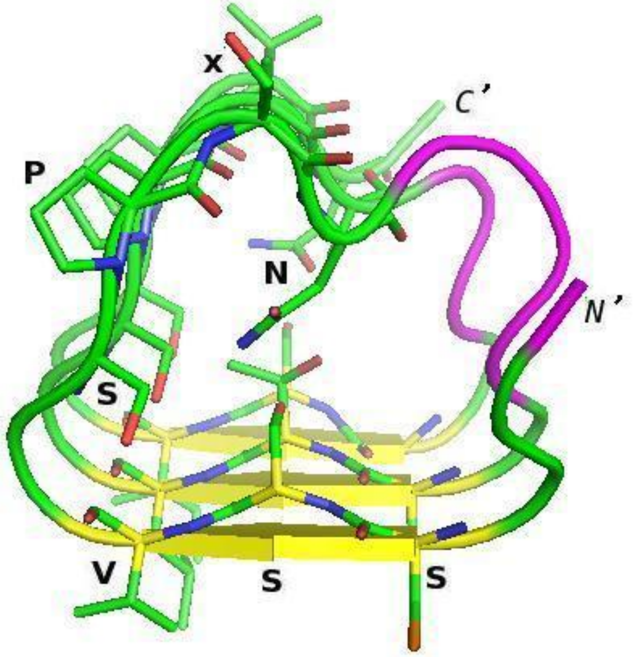
MK**S**SI**SSP**PS HCS

VK**S**P**VSSP**NN VT

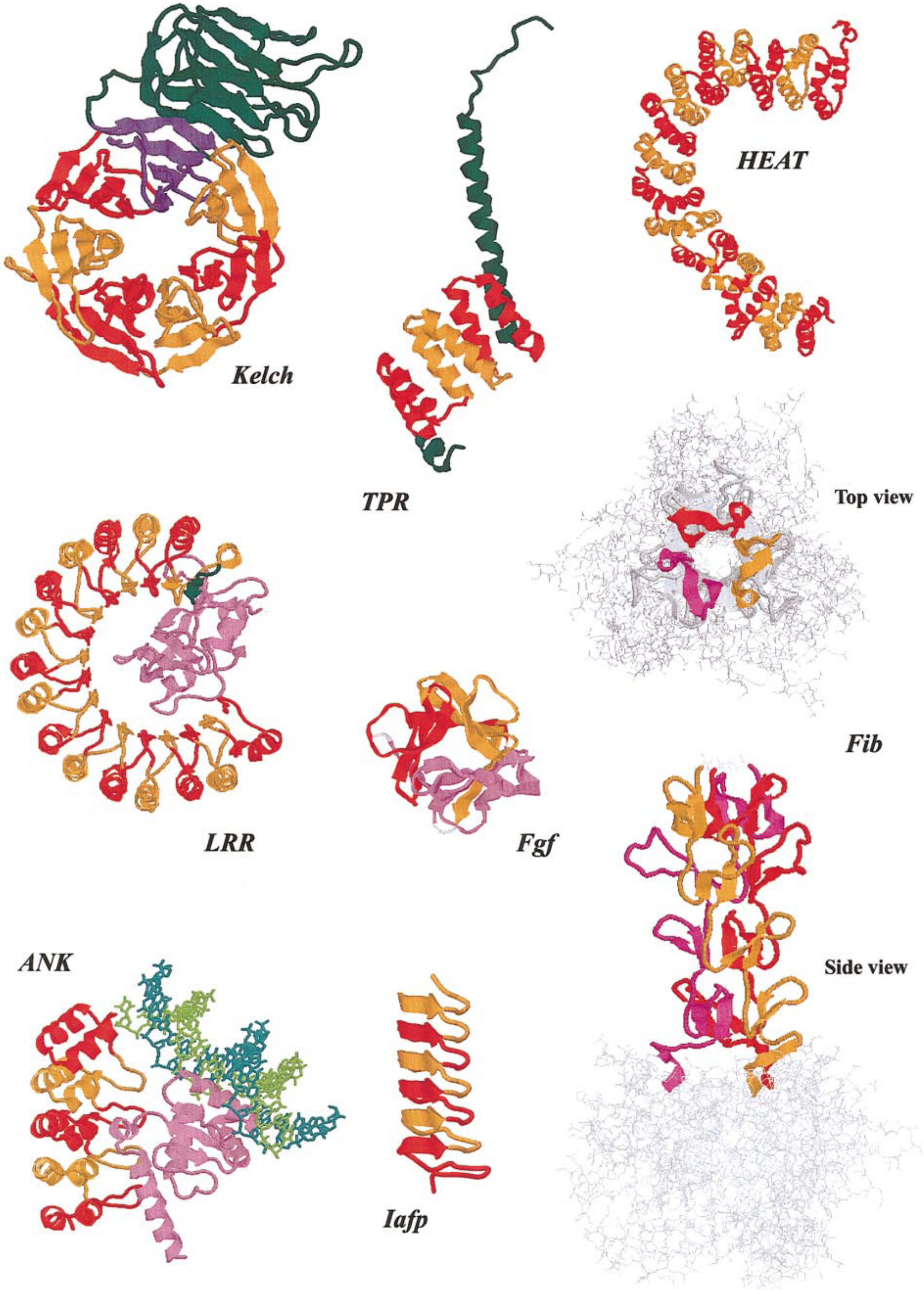
LR**S**S**VSSP**AN INN



http://weblogo.berkeley.edu



(Vlassi et al, 2013)



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
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQP
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**QQQQQQQQQQQQQQQQQQQQQ****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**QQQQQQQQQQQQQQQQQQQQQ****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**
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**QQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQ**L**PQPPPQ**A**QP**
LL**PQPQPPPPPPPPPP**GPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHST**LLILGVLL**T**L**R**YLV**P**LL**QQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
VELIAGGGSSCSPVLSRKQKGKVLLG**EEE**AL**EDD**S**E**SRSDVSSSALTASVKDEISGELAA
SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQ**D**
EDEEATGILPDEASEAFRNSSMALQQAHLLKNMSHCRQPSDSSVDKFVLRDEATEPGDQE
NKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALALSCVG
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSIL

Detection repeats

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

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

Detection repeats

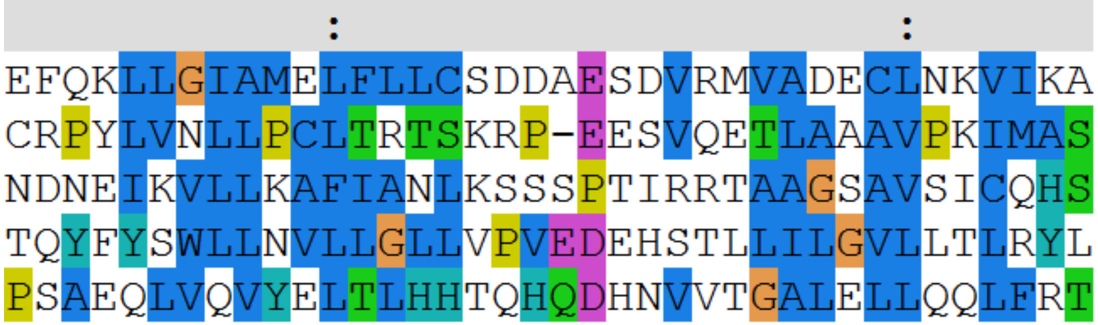
Often NOT straightforward.
N-terminal human Huntingtin.
How many **repeats** can you find?

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

Detection repeats

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

EFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKA
CRPYLVNLLPCLTRTSKRP-EESVQETLAAAVPKIMAS
NDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHS
TQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYL
PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRT



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

Repeats

Frequency repeats

Fraction of proteins annotated with the keyword REPEAT in SwissProt

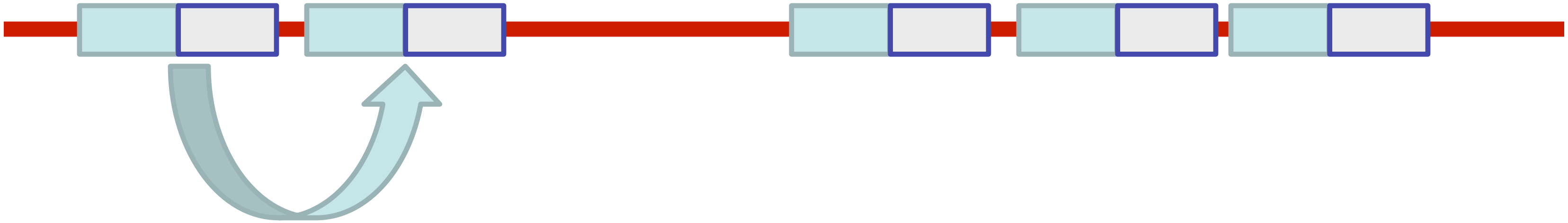
		%
Archaea	27/3428	0.79
Viruses	81/8048	1.00
Bacteria	299/28438	1.05
Fungi	232/8334	2.78
Viridiplantae	153/6963	2.20
Metazoa	1538/28948	5.31
Rest of Eukaryota	92/2434	3.78

(Andrade et al 2001)

Detection of repeats

Dotplots

Comparing a sequence against itself



Detection of repeats

Dotplots

```
TLRSSVSSPANINNS
 NMTSSVCSPANISV
```

Detection of repeats

Dotplots

```
TLRSSVSSPANINNS
    |
 NMTSSVCSPANISV
```

1 match

Detection of repeats

Dotplots

```
TLRSSVSSPANINNS
   ||| |||||
NMTSSVCSPANISV
```

8 matches

Detection of repeats

Dotplots

```
 TLRSSVSSPANINNS
    |  |
NMTSSVCSPANISV
```

2 matches

Detection of repeats

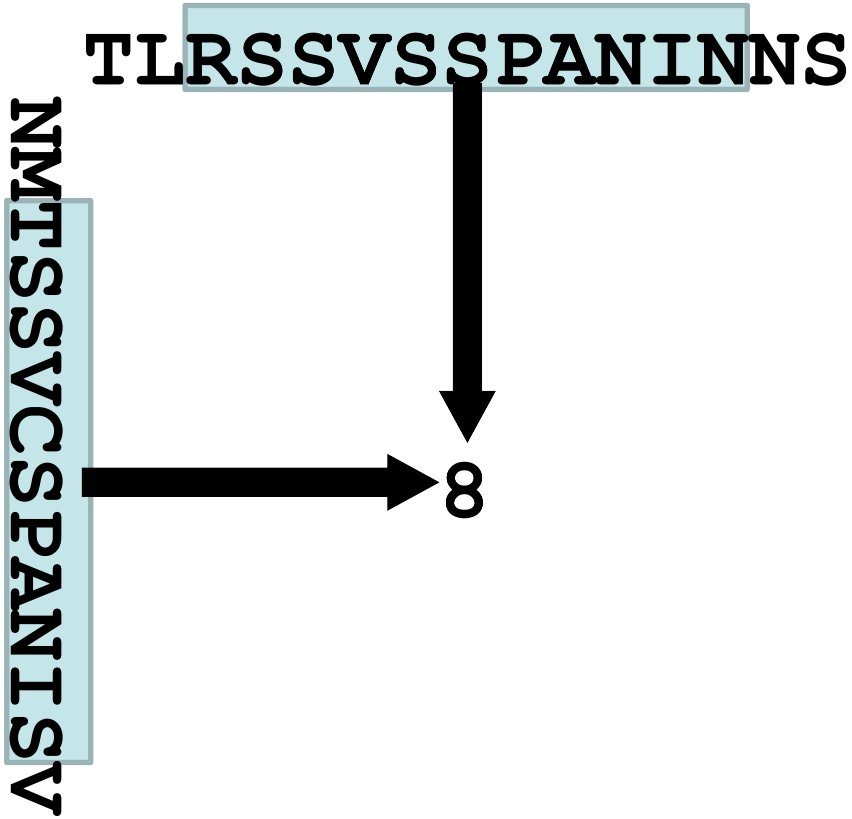
Dotplots

```
  TLRSSVSSPANINNS
  |
NMTSSVCSPANISV
```

1 match

Detection of repeats

Dotplots



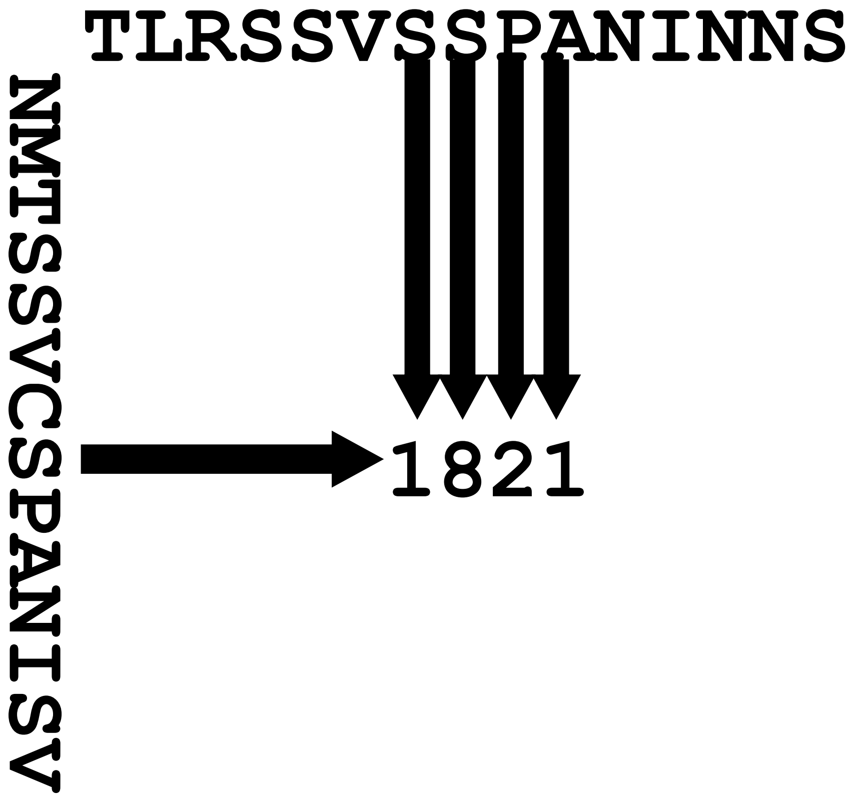
TLRSSVSSPANINNS

NMTSSVCSPANISV

8

Detection of repeats

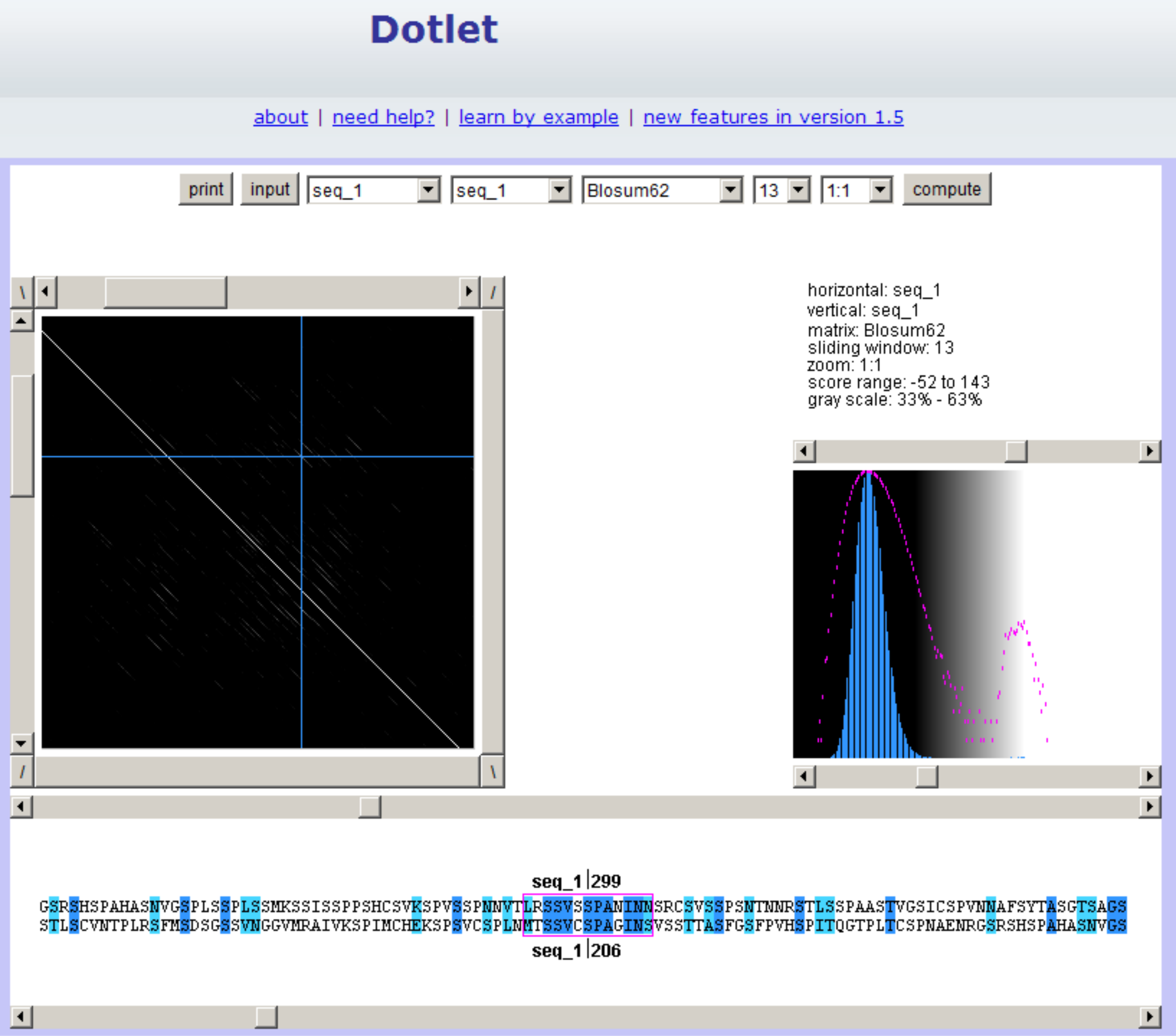
Dotplots



TLRSSVSSPANINNS

NMTSSVCSPANISV

1821



Dotlet

about | need help? | learn by example | new features in version 1.5
print
input
seq_1
seq_1
Blosum62
13
1:1
compute

horizontal: seq_1
vertical: seq_1
matrix: Blosum62
sliding window: 13
zoom: 1:1
score range: -52 to 143
gray scale: 33% - 63%

```
                                          seq_1 |299
GSRSHSPAHASNVGSPLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINNSRCSVSSPSNTNNRSTLSSPAASTVGSICSPVNNAFSYTASGTSAGS
STLSCVNTPLRSFMSDSGSSVNGGVMRAIVKSPIMCHEKSPSVCSPLNMTSSVCSPAGINSVSSTTASFGSFPVHSPITQGTPLTCSPNAENRGSRSHSPAHASNVGS
                                          seq_1 |206
```

Exercise 1/3. Using Dotlet with the human mineralocorticoid receptor (MR)

- Go to the Dotlet web page: http://myhits.isb-sib.ch/cgi-bin/dotlet

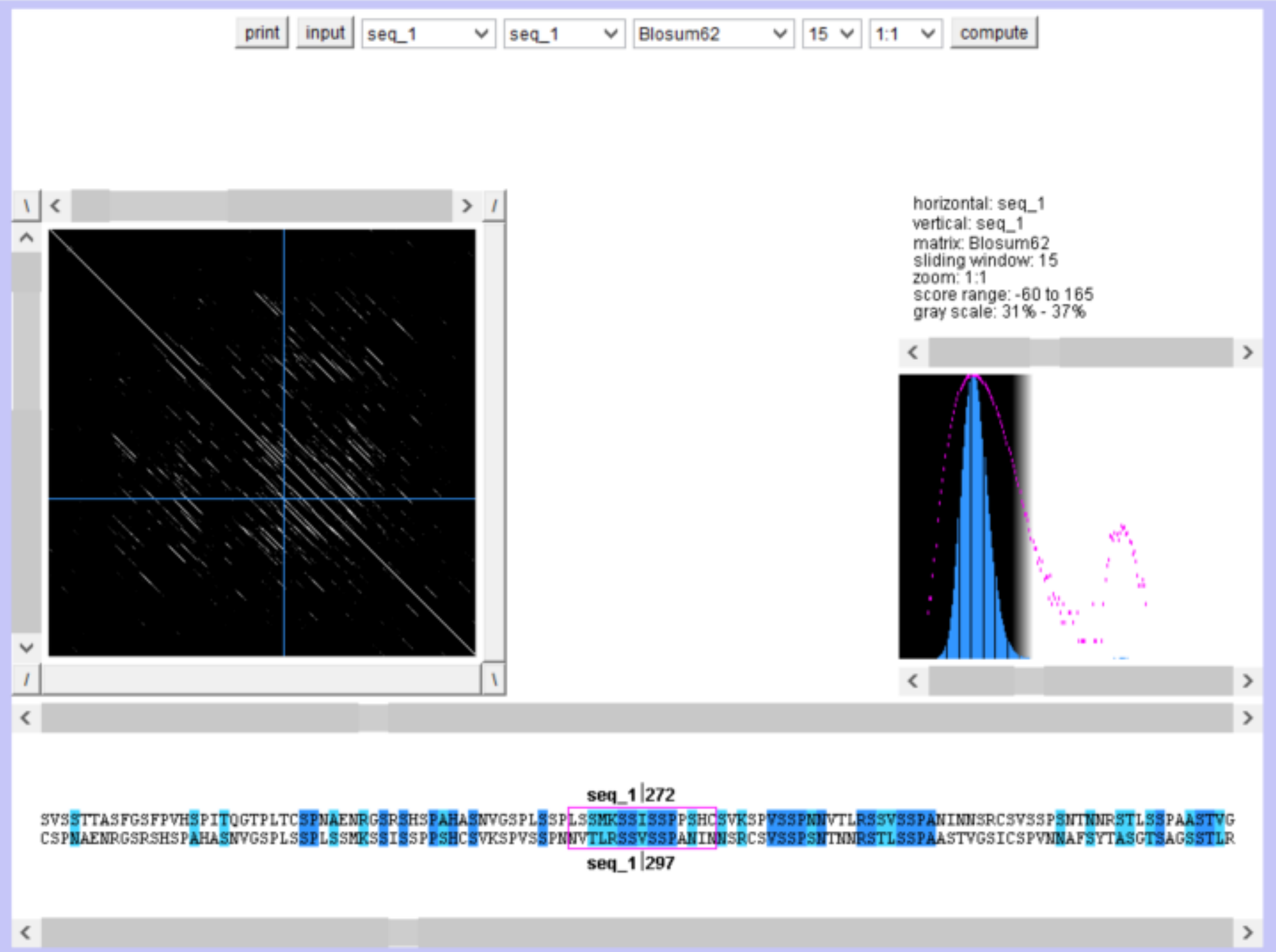
- Click on the input button and paste the sequence of the human mineralocorticoid receptor (UniProt id P08235)

- 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) (Hint2: Range 27%-36% works well)

- Find repetitions clicking in the diagonal patterns: which repeated sequences do you find?

Exercise 1/4. Using Dotlet with the human mineralocorticoid receptor (MR)



horizontal: seq_1
vertical: seq_1
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 165
gray scale: 31% - 37%

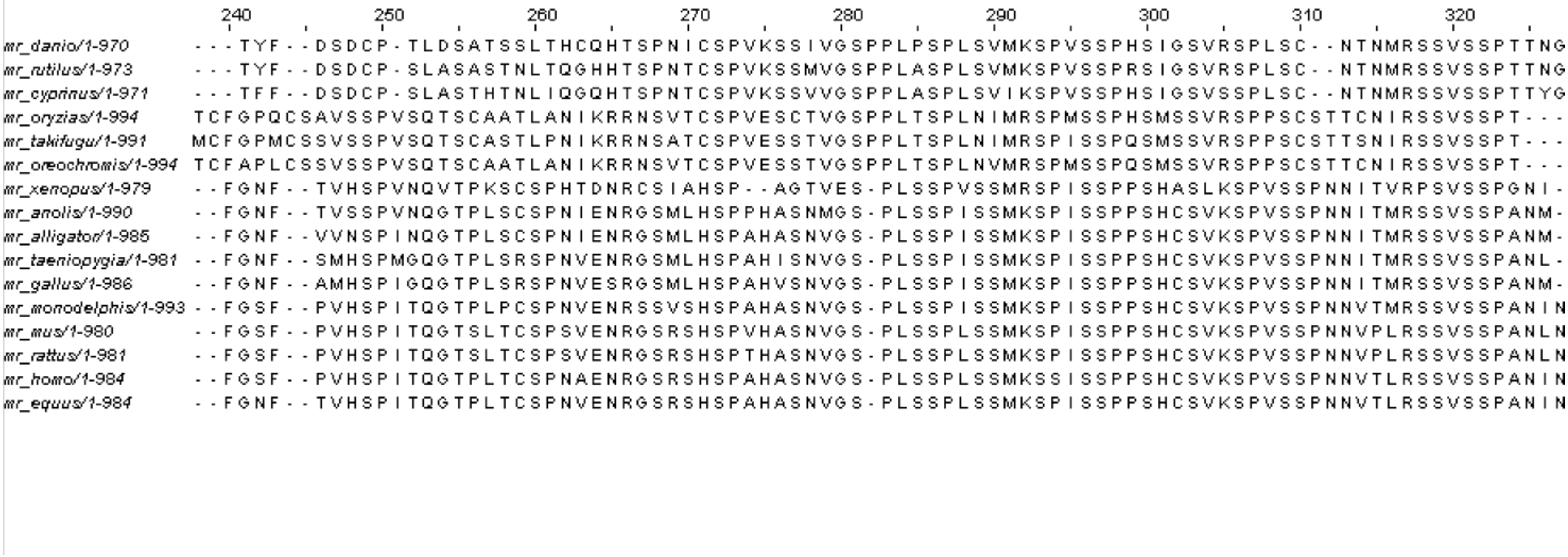
seq_1 |272

```
SVSSTTASFGSFPVHSPITQGTPLTCSPNAENRGSRSHSPAHASNVGSPLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINNSRCSVSSPSNTNNRSTLSSPAASTVG
CSPNAENRGSRSHSPAHASNVGSPLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANINNSRCSVSSPSNTNNRSTLSSPAASTVGSICSPVNNAFSYTASGTSAGSSTLR
```

seq_1 |297

Detection of repeats

Using a multiple sequence alignment helps.
Conserved repeated patterns

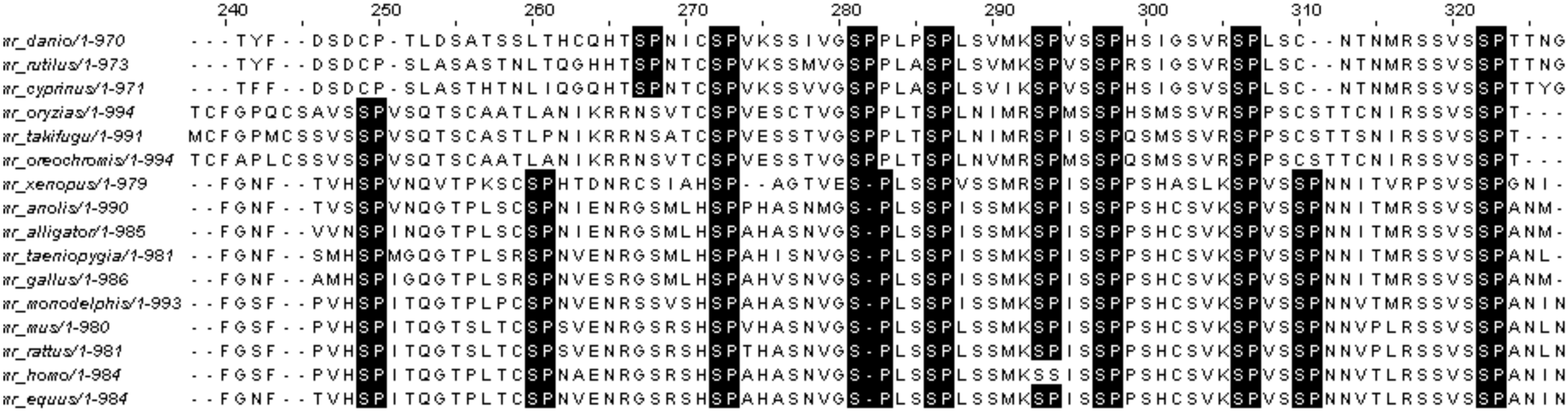


```
                          240       250       260       270       280       290       300       310       320
mr_danio/1-970          ---TYF--DSDCP-TLDSATSSLTHCQHTSPNICSPVKSSIVGSPPLPSPLSVMKSPVSSPHSIGSVRSPLSC--NTNMRSSVSSPTTNG
mr_rutilus/1-973        ---TYF--DSDCP-SLASASTNLTQGHHTSPNTCSPVKSSMVGSPPLASPLSVMKSPVSSPRSIGSVRSPLSC--NTNMRSSVSSPTTNG
mr_cyprinus/1-971       ---TFF--DSDCP-SLASTHTNLIQGQHTSPNTCSPVKSSVVGSPPLASPLSVIKSPVSSPHSIGSVSSPLSC--NTNMRSSVSSPTTYG
mr_oryzias/1-994        TCFGPQCSAVSSPVSQTSCAATLANIKRRNSVTCSPVESCTVGSPPLTSPLNIMRSPMSSPHSMSSVRSPPSCSTTCNIRSSVSSPT---
mr_takifugu/1-991       MCFGPMCSSVSSPVSQTSCASTLPNIKRRNSATCSPVESSTVGSPPLTSPLNIMRSPISSPQSMSSVRSPPSCSTTSNIRSSVSSPT---
mr_oreochromis/1-994    TCFAPLCSSVSSPVSQTSCAATLANIKRRNSVTCSPVESSTVGSPPLTSPLNVMRSPMSSPQSMSSVRSPPSCSTTCNIRSSVSSPT---
mr_xenopus/1-979        --FGNF--TVHSPVNQVTPKSCSPHTDNRCSIAHSP--AGTVES-PLSSPVSSMRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI-
mr_anolis/1-990         --FGNF--TVSSPVNQGTPLSCSPNIENRGSMLHSPPHASNMGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
mr_alligator/1-985      --FGNF--VVNSPINQGTPLSCSPNIENRGSMLHSPAHASNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
mr_taeniopygia/1-981    --FGNF--SMHSPMGQGTPLSRSPNVENRGSMLHSPAHISNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANL-
mr_gallus/1-986         --FGNF--AMHSPIGQGTPLSRSPNVESRGSMLHSPAHVSNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
mr_monodelphis/1-993    --FGSF--PVHSPITQGTPLPCSPNVENRSSVSHSPAHASNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNVTMRSSVSSPANIN
mr_mus/1-980            --FGSF--PVHSPITQGTSLTCSPSVENRGSRSHSPVHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN
mr_rattus/1-981         --FGSF--PVHSPITQGTSLTCSPSVENRGSRSHSPTHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN
mr_homo/1-984           --FGSF--PVHSPITQGTPLTCSPNAENRGSRSHSPAHASNVGS-PLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
mr_equus/1-984          --FGNF--TVHSPITQGTPLTCSPNVENRGSRSHSPAHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
```

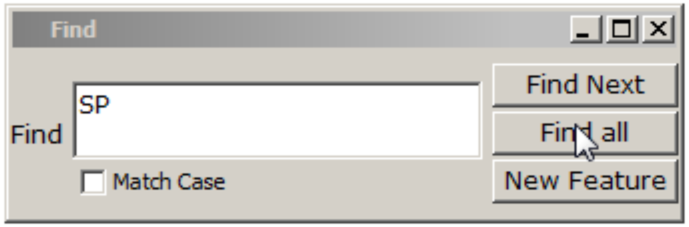
JalView with Regular Expression searches

Detection of repeats

Using a multiple sequence alignment helps
Conserved repeated patterns



```
                     240       250       260       270       280       290       300       310       320
wr_danio/1-970       ---TYF--DSDCP-TLDSATSSLTHCQHTSPNICSPVKSSIVGSPPLPSPLSVMKSPVSSPHSIGSVRSPLSC--NTNMRSSVSSPTTNG
wr_rutilus/1-973     ---TYF--DSDCP-SLASASTNLTQGHHTSPNTCSPVKSSMVGSPPLASPLSVMKSPVSSPRSIGSVRSPLSC--NTNMRSSVSSPTTNG
wr_cyprinus/1-971    ---TFF--DSDCP-SLASTHTNLIQGQHTSPNTCSPVKSSVVGSPPLASPLSVIKSPVSSPHSIGSVSSPLSC--NTNMRSSVSSPTTYG
wr_oryzias/1-994     TCFGPQCSAVSSPVSQTSCAATLANIKRRNSVTCSPVESCTVGSPPLTSPLNIMRSPMSSPHSMSSVRSPPSCSTTCNIRSSVSSPT---
wr_takifugu/1-991    MCFGPMCSSVSSPVSQTSCASTLPNIKRRNSATCSPVESSTVGSPPLTSPLNIMRSPISSPQSMSSVRSPPSCSTTSNIRSSVSSPT---
wr_oreochromis/1-994 TCFAPLCSSVSSPVSQTSCAATLANIKRRNSVTCSPVESSTVGSPPLTSPLNVMRSPMSSPQSMSSVRSPPSCSTTCNIRSSVSSPT---
wr_xenopus/1-979     --FGNF--TVHSPVNQVTPKSCSPHTDNRCSIAHSP--AGTVES-PLSSPVSSMRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI-
wr_anolis/1-990      --FGNF--TVSSPVNQGTPLSCSPNIENRGSMLHSPPHASNMGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
wr_alligator/1-985   --FGNF--VVNSPINQGTPLSCSPNIENRGSMLHSPAHASNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
wr_taeniopygia/1-981 --FGNF--SMHSPMGQGTPLSRSPNVENRGSMLHSPAHISNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANL-
wr_gallus/1-986      --FGNF--AMHSPIGQGTPLSRSPNVESRGSMLHSPAHVSNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-
wr_monodelphis/1-993 --FGSF--PVHSPITQGTPLPCSPNVENRSSVSHSPAHASNVGS-PLSSPISSMKSPISSPPSHCSVKSPVSSPNNVTMRSSVSSPANIN
wr_mus/1-980         --FGSF--PVHSPITQGTSLTCSPSVENRGSRSHSPVHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN
wr_rattus/1-981      --FGSF--PVHSPITQGTSLTCSPSVENRGSRSHSPTHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN
wr_homo/1-984        --FGSF--PVHSPITQGTPLTCSPNAENRGSRSHSPAHASNVGS-PLSSPLSSMKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
wr_equus/1-984       --FGNF--TVHSPITQGTPLTCSPNVENRGSRSHSPAHASNVGS-PLSSPLSSMKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN
```



Find: SP

☐ Match Case

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?

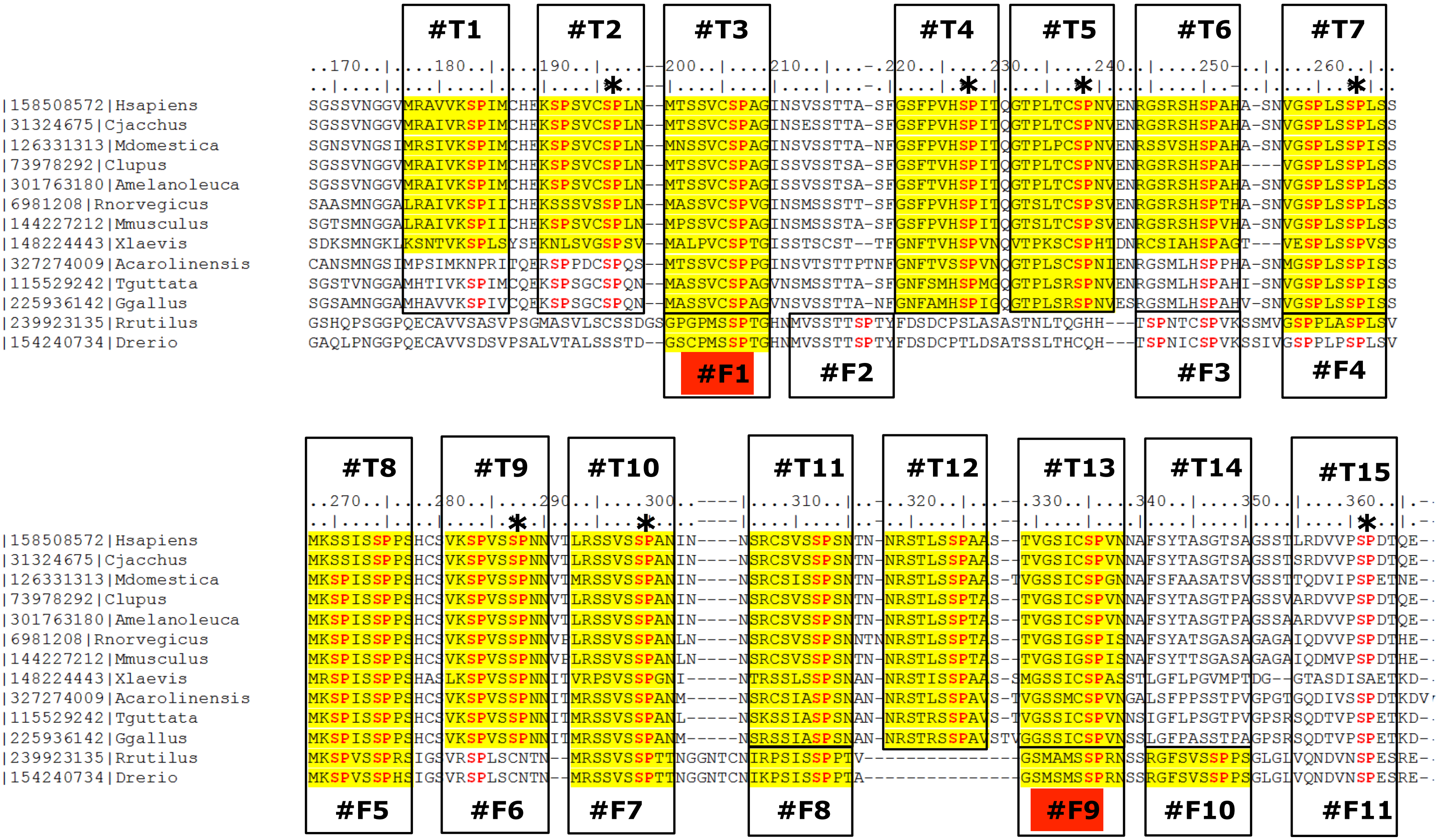
- **`LPTA, SPAA, LPPA, LPAP, SPLA`**

Exercise 2/4. Using JalView with a MSA of the MR with orthologs

- Load the multiple sequence alignment of the MR in JalView: 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?



```
                             ..170..|....180..|....190..|....--200..|....210..|....-.220..|....230..|....240..|....250-..|....260..|..
                             ..|....|....|....|....|..*.--|....|....|....|...-.|....|..*.|....|..*.|....|....|..-..|....|..*|..
|158508572|Hsapiens          SGSSVNGGVMRAVVKSPIMCHEKSPSVCSPLN--MTSSVCSPAGINSVSSTTA-SFGSFPVHSPITQGTPLTCSPNVENRGSRSHSPAHA-SNVGSPLSSPLSS
|31324675|Cjacchus           SGSSVNGGVMRAIVRSPIMCHEKSPSVCSPLN--MTSSVCSPAGINSESSTTA-SFGSFPVHSPITQGTPLTCSPNVENRGSRSHSPAHA-SNVGSPLSSPLSS
|126331313|Mdomestica        SGNSVNGSIMRSIVKSPIMCHEKSPSVCSPLN--MNSSVCSPAGINSVSSTTA-NFGSFPVHSPITQGTPLPCSPNVENRSSVSHSPAHA-SNVGSPLSSPISS
|73978292|Clupus             SGSSVNGGVMRAIVKSPIMCHEKSPSVCSPLN--MTSSVCSPAGISSVSSTSA-SFGSFTVHSPITQGTPLTCSPNVENRGSRSHSPAH----VGSPLSSPLSS
|301763180|Amelanoleuca      SGSSVNGGVMRAIVKSPIMCHEKSPSVCSPLN--MTSSVCSPAGISSVSSTSA-SFGSFTVHSPITQGTPLTCSPNVENRGSRSHSPAHA-SNVGSPLSSPLSS
|6981208|Rnorvegicus        SAASMNGGALRAIVKSPIICHEKSSSVSSPLN--MASSVCSPVGINSMSSSTT-SFGSFPVHSPITQGTSLTCSPSVENRGSRSHSPTHA-SNVGSPLSSPLSS
|144227212|Mmusculus         SGTSMNGGALRAIVKSPIICHEKSPSVCSPLN--MPSSVCSPAGINSMSSSTA-SFGSFPVHSPITQGTSLTCSPSVENRGSRSHSPVHA-SNVGSPLSSPLSS
|148224443|Xlaevis           SDKSMNGKLKSNTVKSPLSYSEKNLSVGSPSV--MALPVCSPTGISSTSCST--TFGNFTVHSPVNQVTPKSCSPHTDNRCSIAHSPAGT---VESPLSSPVSS
|327274009|Acarolinensis     CANSMNGSIMPSIMKNPRITQERSPPDCSPQS--MTSSVCSPPGINSVTSTTPTNFGNFTVSSPVNQGTPLSCSPNIENRGSMLHSPPHA-SNMGSPLSSPISS
|115529242|Tguttata          SGSTVNGGAMHTIVKSPIMCQEKSPSGCSPQN--MASSVCSPAGVNSMSSTTA-SFGNFSMHSPMGQGTPLSRSPNVENRGSMLHSPAHI-SNVGSPLSSPISS
|225936142|Ggallus           SGSAMNGGAMHAVVKSPIVCQEKSPSGCSPQN--MASSVCSPAGVNSVSSTTA-NFGNFAMHSPIGQGTPLSRSPNVESRGSMLHSPAHV-SNVGSPLSSPISS
|239923135|Rrutilus          GSHQPSGGPQECAVVSASVPSGMASVLSCSSDGSGPGPMSSPTGHNMVSSTTSPTYFDSDCPSLASASTNLTQGHH---TSPNTCSPVKSSMVGSPPLASPLSV
|154240734|Drerio            GAQLPNGGPQECAVVSDSVPSALVTALSSSTD--GSCPMSSPTGHNMVSSTTSPTYFDSDCPTLDSATSSLTHCQH---TSPNICSPVKSSIVGSPPLPSPLSV
```

```
                             ..270..|....280..|....290..|....300..----|....310..|..-..320..|..--.330..|....340..|....350..|....360..|.-
                             ..|....|....|....|..*.|....|...*.....----|....|....|..-..|....|..--.|....|....|....|....|....|....|*..|.-
|158508572|Hsapiens          MKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN----NSRCSVSSPSNTN-NRSTLSSPAAS--TVGSICSPVNNAFSYTASGTSAGSSTLRDVVPSPDTQE-
|31324675|Cjacchus           MKSSISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN----NSRCSVSSPSNTN-NRSTLSSPAAS--TVGSICSPVNNAFSYTASGTSAGSSTSRDVVPSPDTQE-
|126331313|Mdomestica        MKSPISSPPSHCSVKSPVSSPNNVTMRSSVSSPANIN----NSRCSISSPSNTN-NRSTLSSPAAS-TVGSSICSPGNNAFSFAASATSVGSSTTQDVIPSPETNE-
|73978292|Clupus             MKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN----NSRCSVSSPSNTN-NRSTLSSPTAS--TVGSICSPVNNAFSYTASGTPAGSSVARDVVPSPDTQE-
|301763180|Amelanoleuca      MKSPISSPPSHCSVKSPVSSPNNVTLRSSVSSPANIN----NSRCSVSSPSNTN-NRSTLSSPTAS--TVGSICSPVNNAFSYTASGTPAGSSAARDVVPSPDTQE-
|6981208|Rnorvegicus        MKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN----NSRCSVSSPSNNTNNRSTLSSPTAS--TVGSIGSPISNAFSYATSGASAGAGAIQDVVPSPDTHE-
|144227212|Mmusculus         MKSPISSPPSHCSVKSPVSSPNNVPLRSSVSSPANLN----NSRCSVSSPSNTN-NRSTLSSPTAS--TVGSIGSPISNAFSYTTSGASAGAGAIQDMVPSPDTHE-
|148224443|Xlaevis           MRSPISSPPSHASLKSPVSSPNNITVRPSVSSPGNI-----NTRSSLSSPSNAN-NRSTISSPAAS-SMGSSICSPASSTLGFLPGVMPTDG--GTASDISAETKD-
|327274009|Acarolinensis     MKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-----NSRCSIASPSNAN-NRSTLSSPAVS-TVGSSMCSPVNGALSFPPSSTPVGPGTGQDIVSSPDTKDV
|115529242|Tguttata          MKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANL-----NSKSSIASPSNAN-NRSTRSSPAVS-TVGSSICSPVNNSIGFLPSGTPVGPSRSQDTVPSPETKD-
|225936142|Ggallus           MKSPISSPPSHCSVKSPVSSPNNITMRSSVSSPANM-----NSRSSIASPSNAN-NRSTRSSPAVSTVGGSSICSPVNSSLGFPASSTPAGPSRSQDTVPSPETKD-
|239923135|Rrutilus          MKSPVSSPRSIGSVRSPLSCNTN--MRSSVSSPTTNGGNTCNIRPSISSPPTV---------------GSMAMSSPRNSSRGFSVSSPPSGLGLVQNDVNSPESRE-
|154240734|Drerio            MKSPVSSPHSIGSVRSPLSCNTN--MRSSVSSPTTNGGNTCNIKPSISSPPTA---------------GSMSMSSPRNSSRGFSVSSPPSGLGLVQNDVNSPESRE-
```

(Vlassi et al, 2013)

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

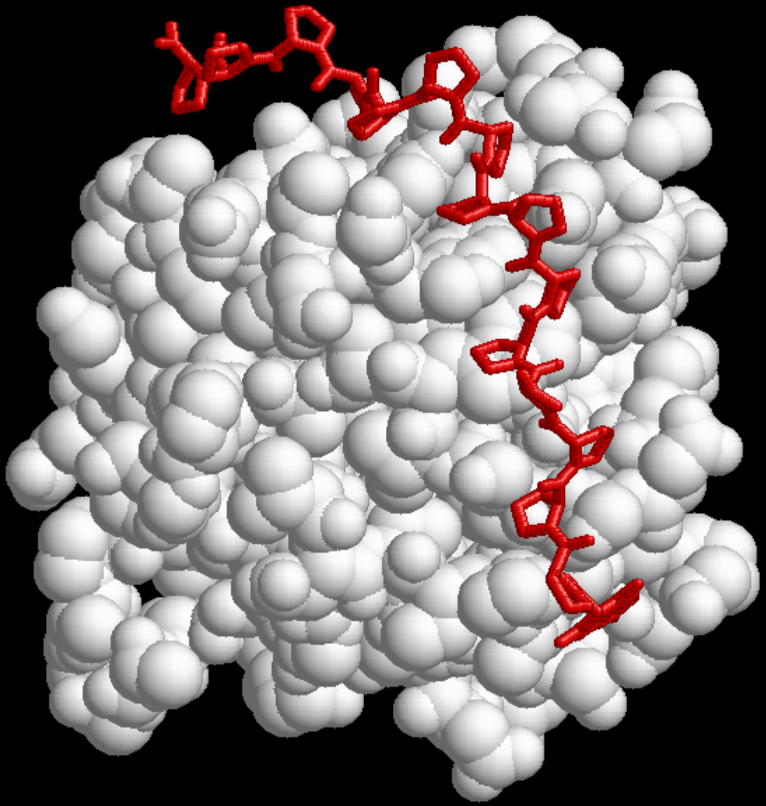
Functions:
Passive: linkers
Active: binding, mediate protein interaction, structural integrity

(Sim and Creamer, 2004)

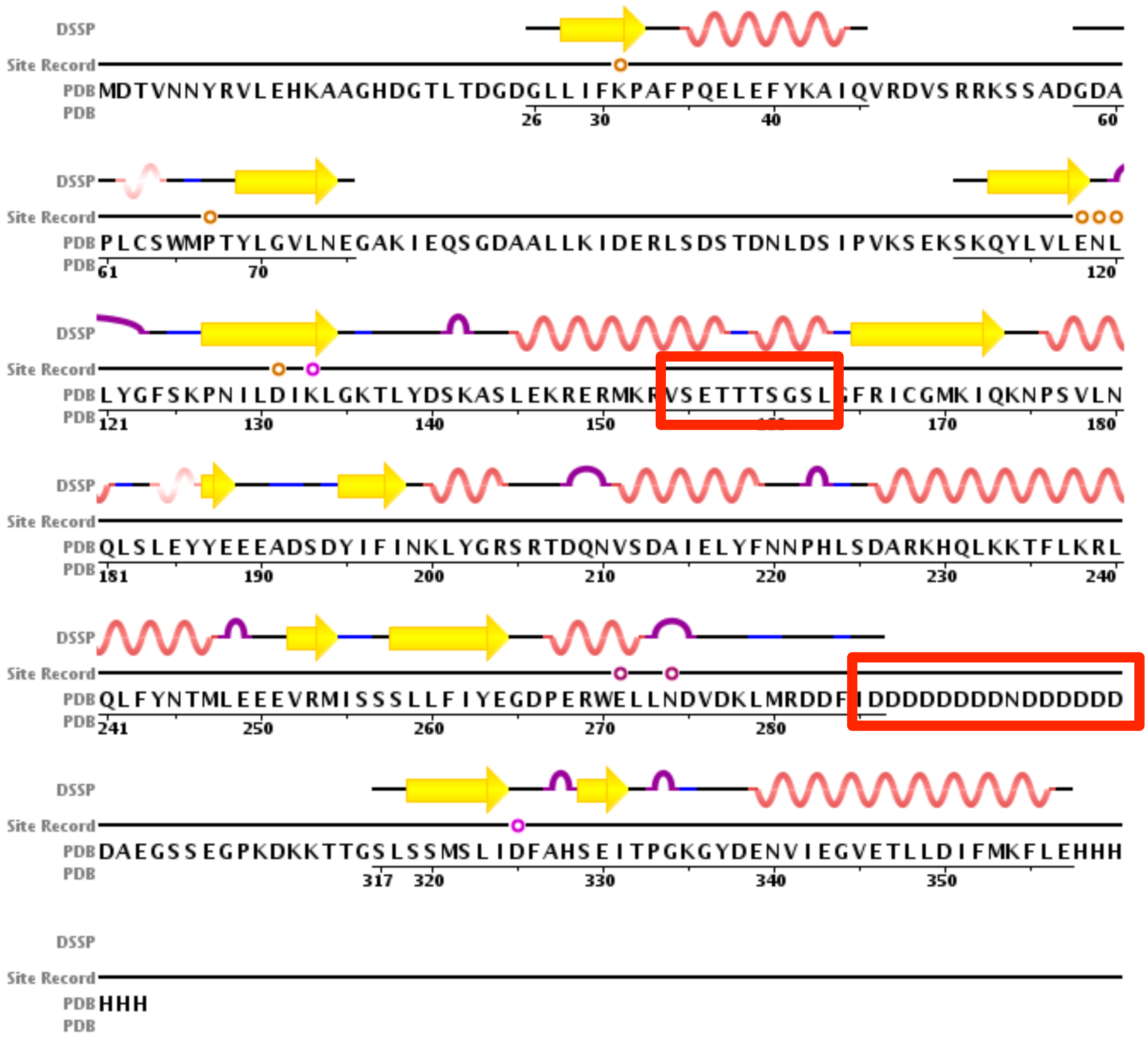
Structure of CBRs

Often variable or flexible: do not easily crystalize

1CJF: profilin bound to polyP



2IF8: Inositol Phosphate Multikinase Ipk2



MDTVNNYRVLEHKAAGHDGTLTDGDGLLIFKPAFPQELEFYKAIQVRDVSRRKSSADGDA

PLCSWMPTYLGVLNEGAKIEQSGDAALLKIDERLSDSTDNLDSIPVKSEKSKQYLVLENL

LYGFSKPNILDIKLGKTLYDSKASLEKRERMKRVSETTTSGSLGFRICGMKIQKNPSVLN

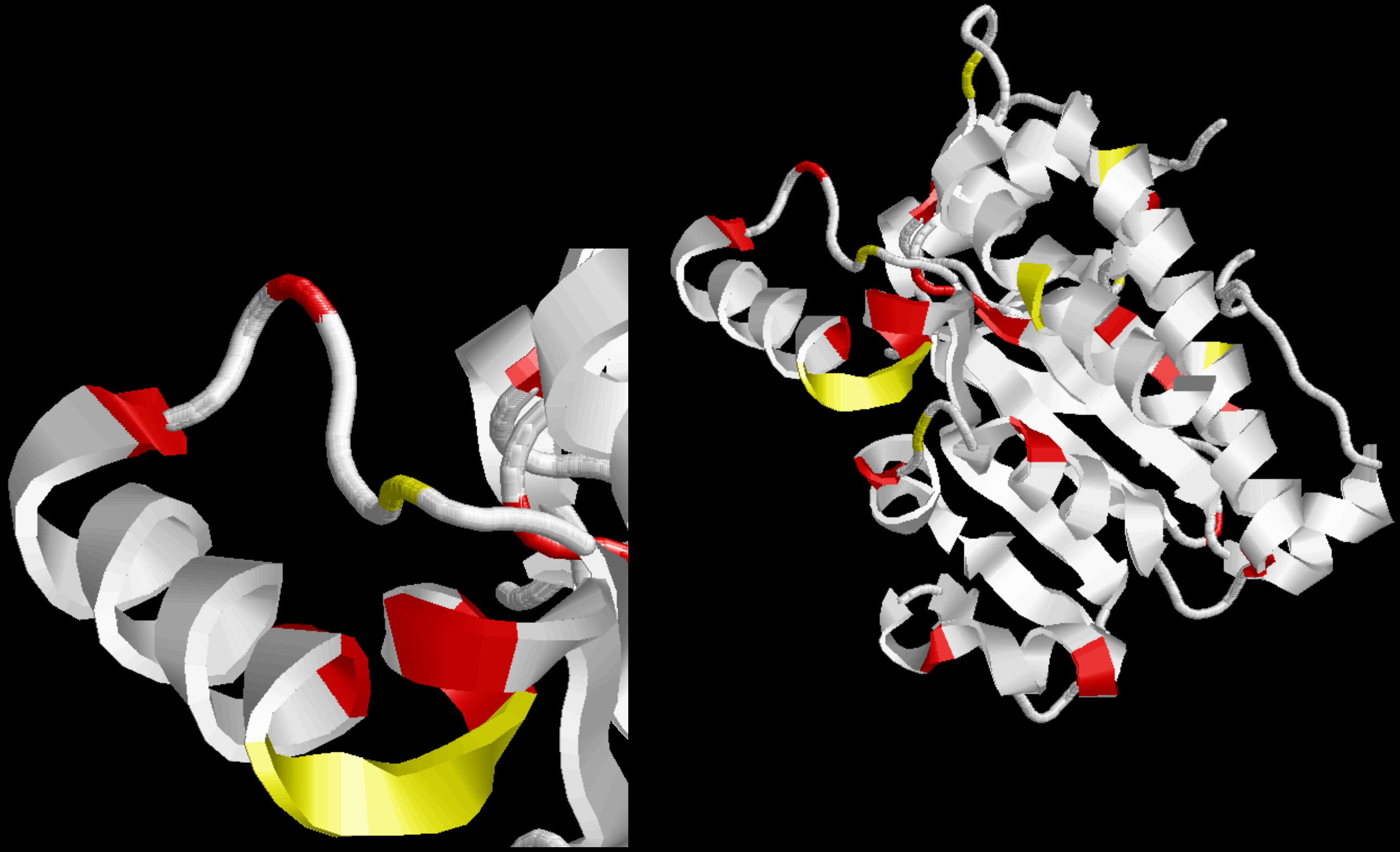
QLSLEYYEEEADSDYIFINKLYGRSRTDQNVSDAIELYFNNPHLSDARKHQLKKTFLKRL

QLFYNTMLEEEVRMISSSLLFIYEGDPERWELLNDVDKLMRDDFIDDDDDDDDNDDDDDD

DAEGSSEGPKDKKTTGSLSSMSLIDFAHSEITPGKGYDENVIEGVETLLDIFMKFLEHHH

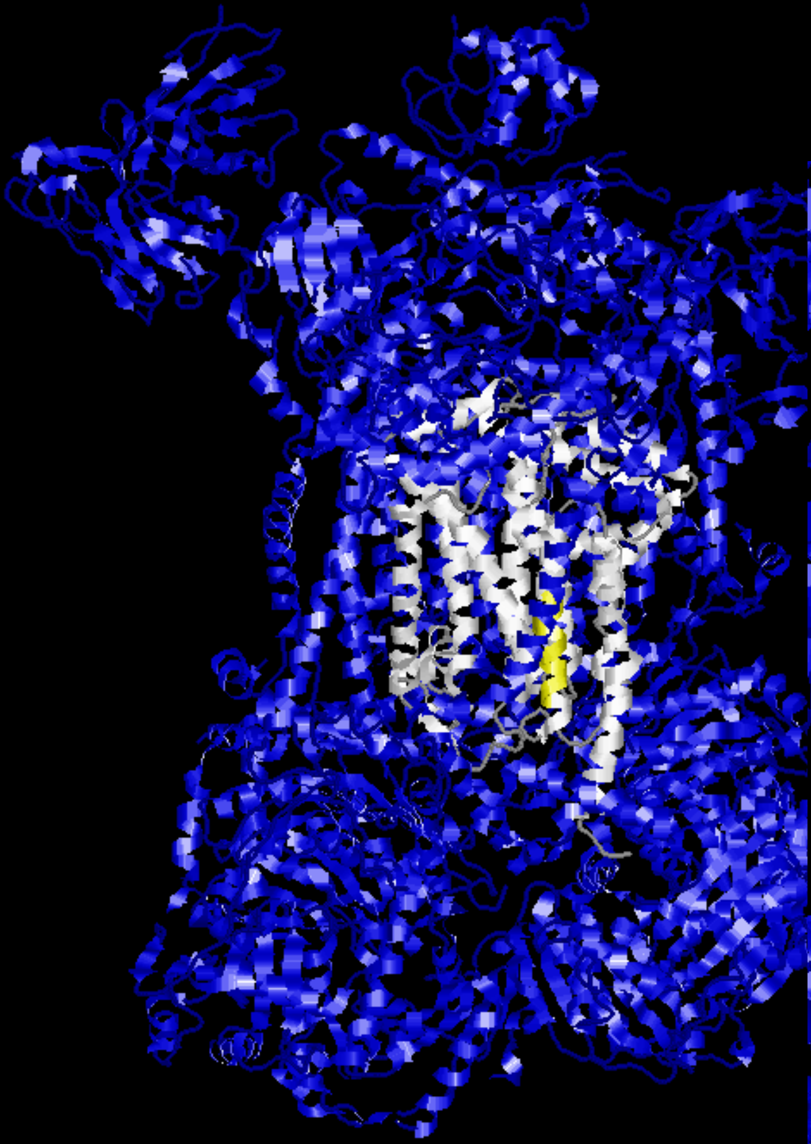
HHH

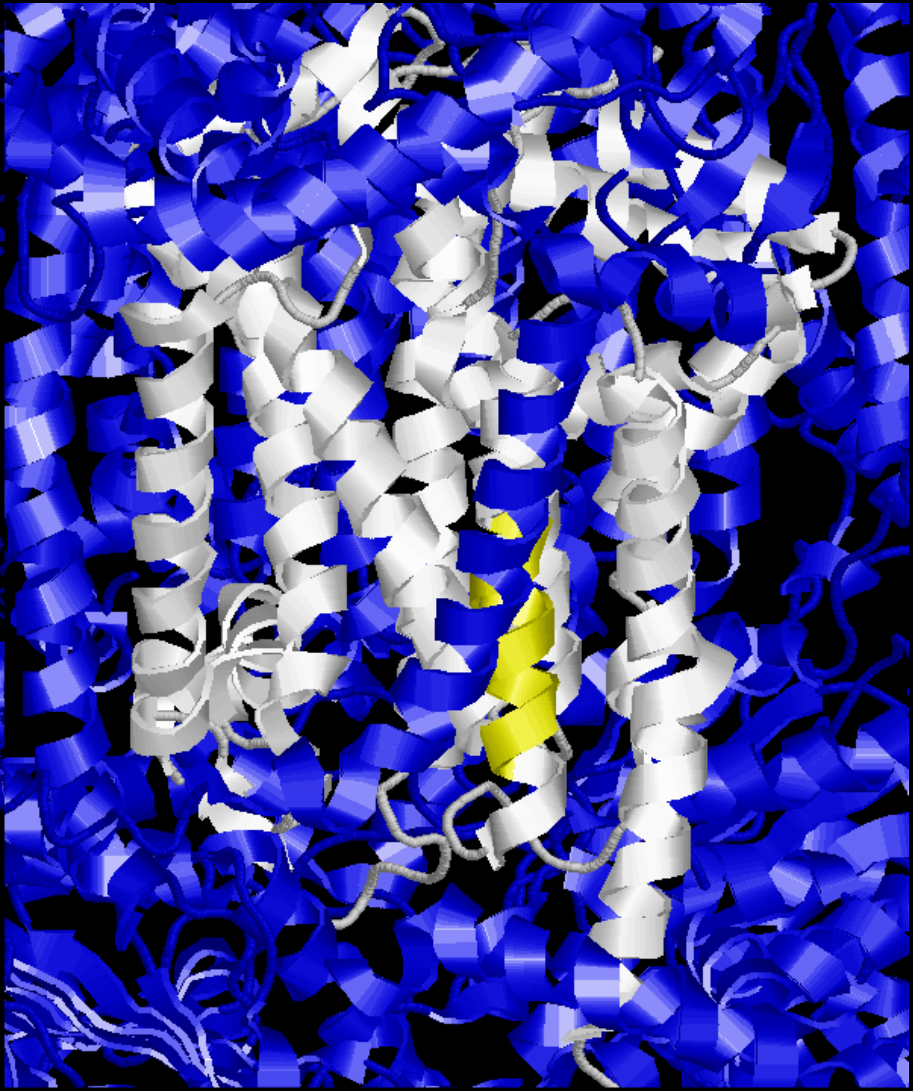
2IF8: Inositol Phosphate Multikinase Ipk2



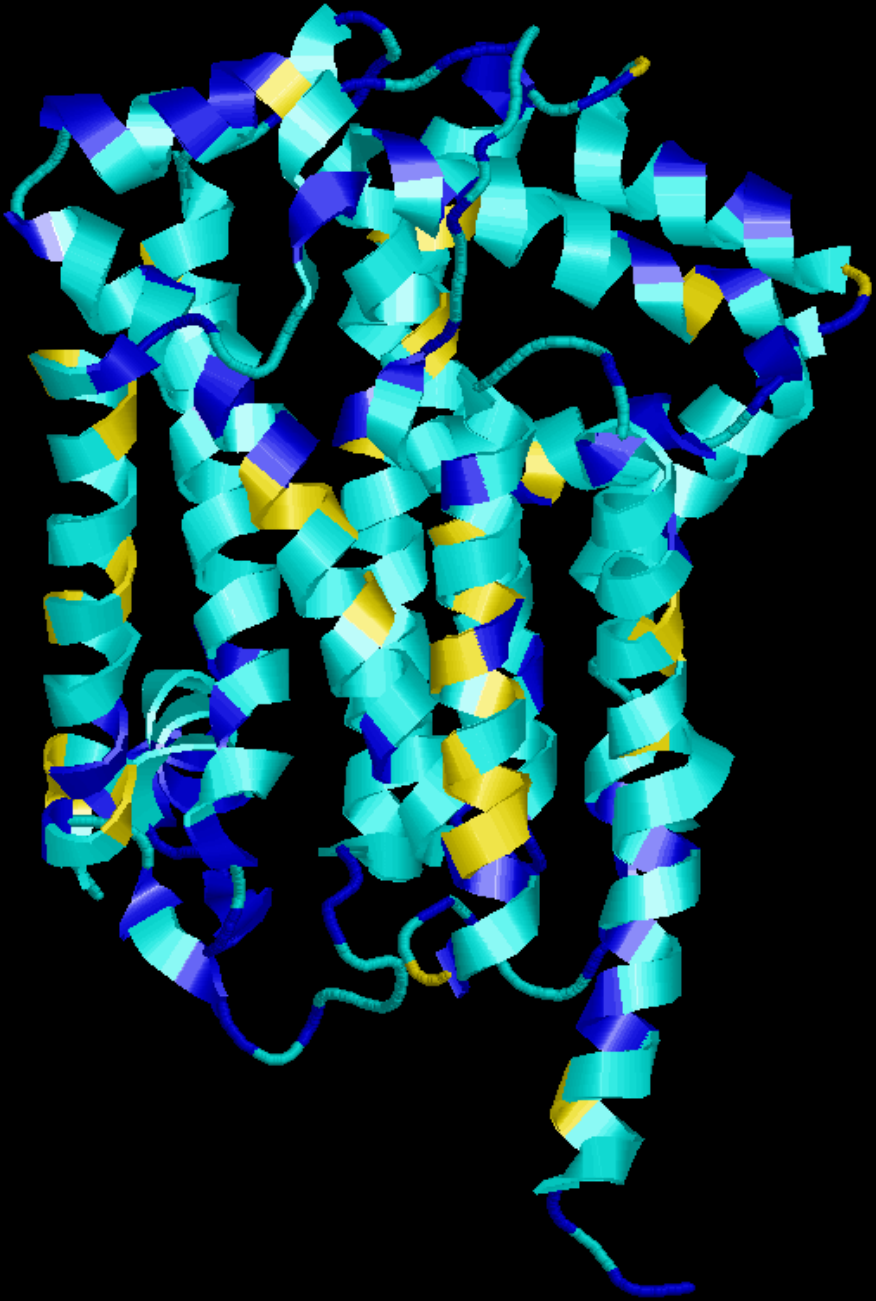
RVSETTTSGSL

2CX5: mitochondrial cytochrome c B subunit N-terminal

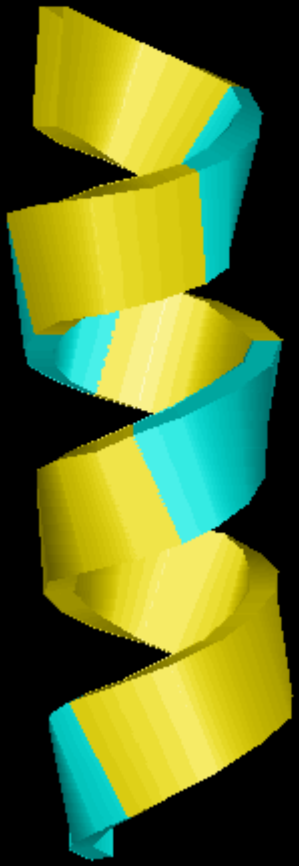




2CX5: mitochondrial cytochrome c B subunit N-terminal



EEEEIEVENE



Types of CBRs

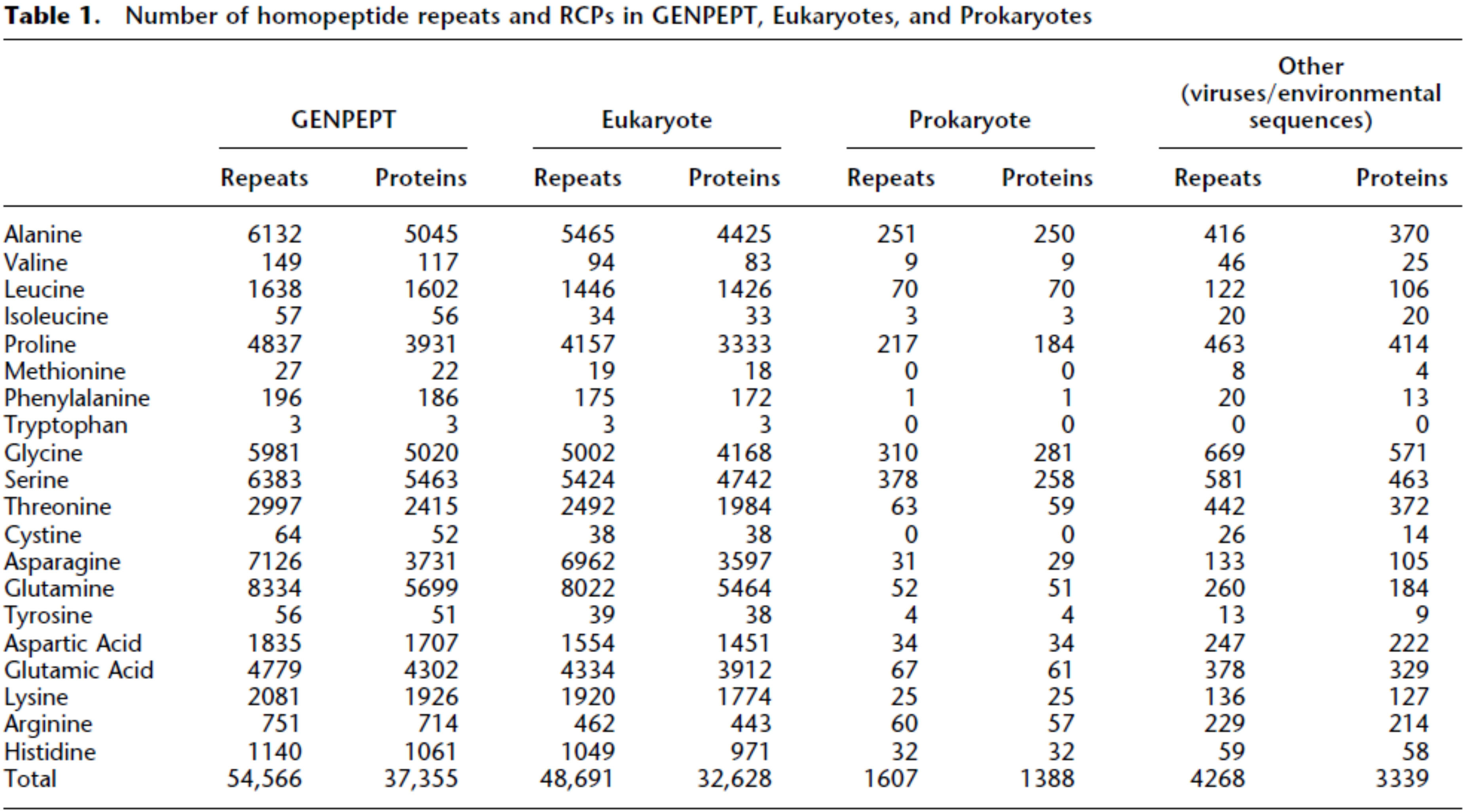


Table 1. Number of homopeptide repeats and RCPs in GENPEPT, Eukaryotes, and Prokaryotes

	GENPEPT		Eukaryote		Prokaryote		Other (viruses/environmental sequences)	
	Repeats	Proteins	Repeats	Proteins	Repeats	Proteins	Repeats	Proteins
Alanine	6132	5045	5465	4425	251	250	416	370
Valine	149	117	94	83	9	9	46	25
Leucine	1638	1602	1446	1426	70	70	122	106
Isoleucine	57	56	34	33	3	3	20	20
Proline	4837	3931	4157	3333	217	184	463	414
Methionine	27	22	19	18	0	0	8	4
Phenylalanine	196	186	175	172	1	1	20	13
Tryptophan	3	3	3	3	0	0	0	0
Glycine	5981	5020	5002	4168	310	281	669	571
Serine	6383	5463	5424	4742	378	258	581	463
Threonine	2997	2415	2492	1984	63	59	442	372
Cystine	64	52	38	38	0	0	26	14
Asparagine	7126	3731	6962	3597	31	29	133	105
Glutamine	8334	5699	8022	5464	52	51	260	184
Tyrosine	56	51	39	38	4	4	13	9
Aspartic Acid	1835	1707	1554	1451	34	34	247	222
Glutamic Acid	4779	4302	4334	3912	67	61	378	329
Lysine	2081	1926	1920	1774	25	25	136	127
Arginine	751	714	462	443	60	57	229	214
Histidine	1140	1061	1049	971	32	32	59	58
Total	54,566	37,355	48,691	32,628	1607	1388	4268	3339

More than 6 aa in length, 1.4% of all, 87% of them in Euk (Faux et al 2005)

Types of CBRs

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)

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

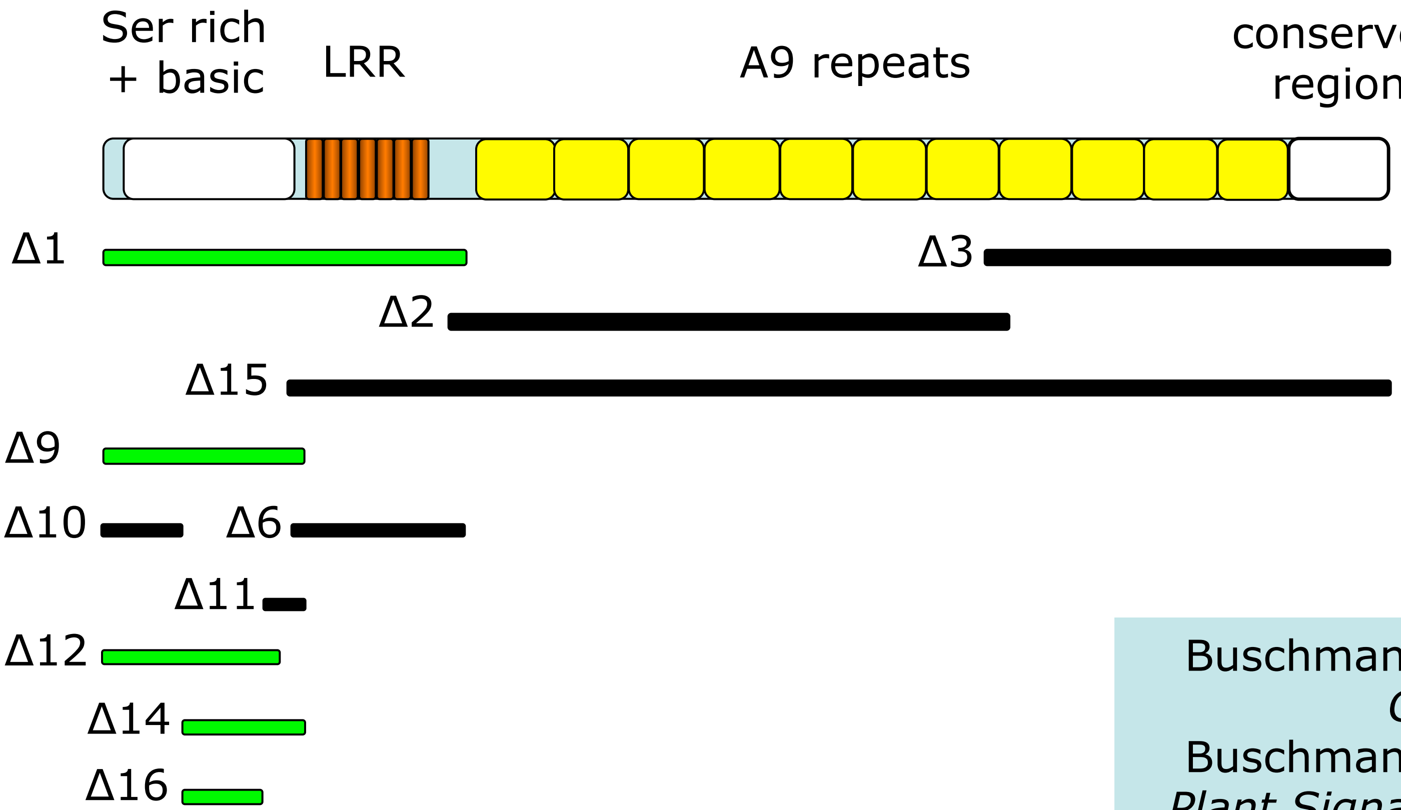
Option for pre-BLAST treatment
SEG algorithm:
1) Identify sequence regions with low information content over a sequence window
2) Merge neighbouring regions

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

(Wootton and Federhen, 1993)

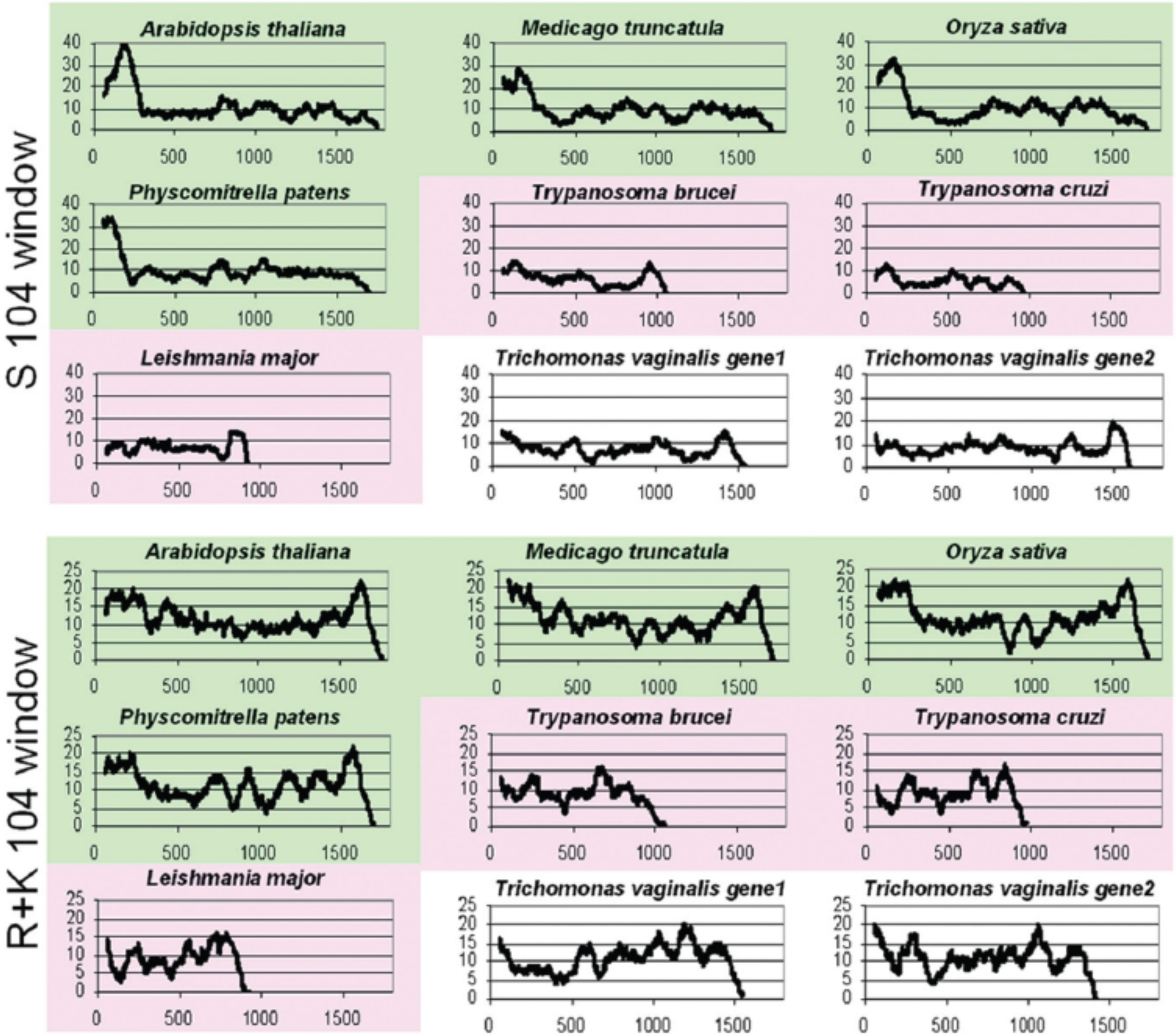
A particular analysis…

AIR9 (1708 aa)



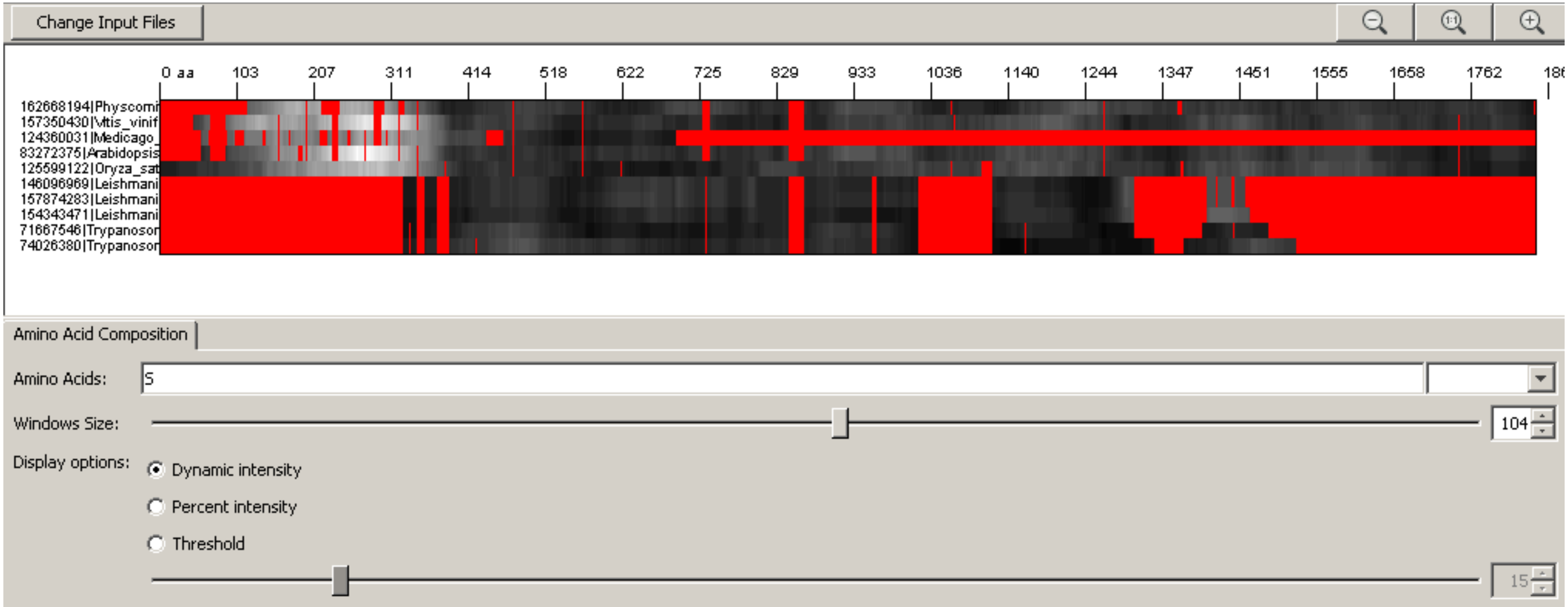
Buschmann, et al (2006). *Current Biology.*
Buschmann, et al (2007). *Plant Signaling & Behavior*

🟩 Microtubule localization of Δx-GFP


S 104 window
Arabidopsis thaliana
Medicago truncatula
Oryza sativa
Physcomitrella patens
Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Trichomonas vaginalis gene1
Trichomonas vaginalis gene2
0
10
20
30
40
0
500
1000
1500
R+K 104 window
Arabidopsis thaliana
Medicago truncatula
Oryza sativa
Physcomitrella patens
Trypanosoma brucei
Trypanosoma cruzi
Leishmania major
Trichomonas vaginalis gene1
Trichomonas vaginalis gene2
0
5
10
15
20
25
0
500
1000
1500

A particular analysis…

…triggers BiasViz



Amino Acids: S

Windows Size: 104

Display options:
● Dynamic intensity
○ Percent intensity
○ Threshold

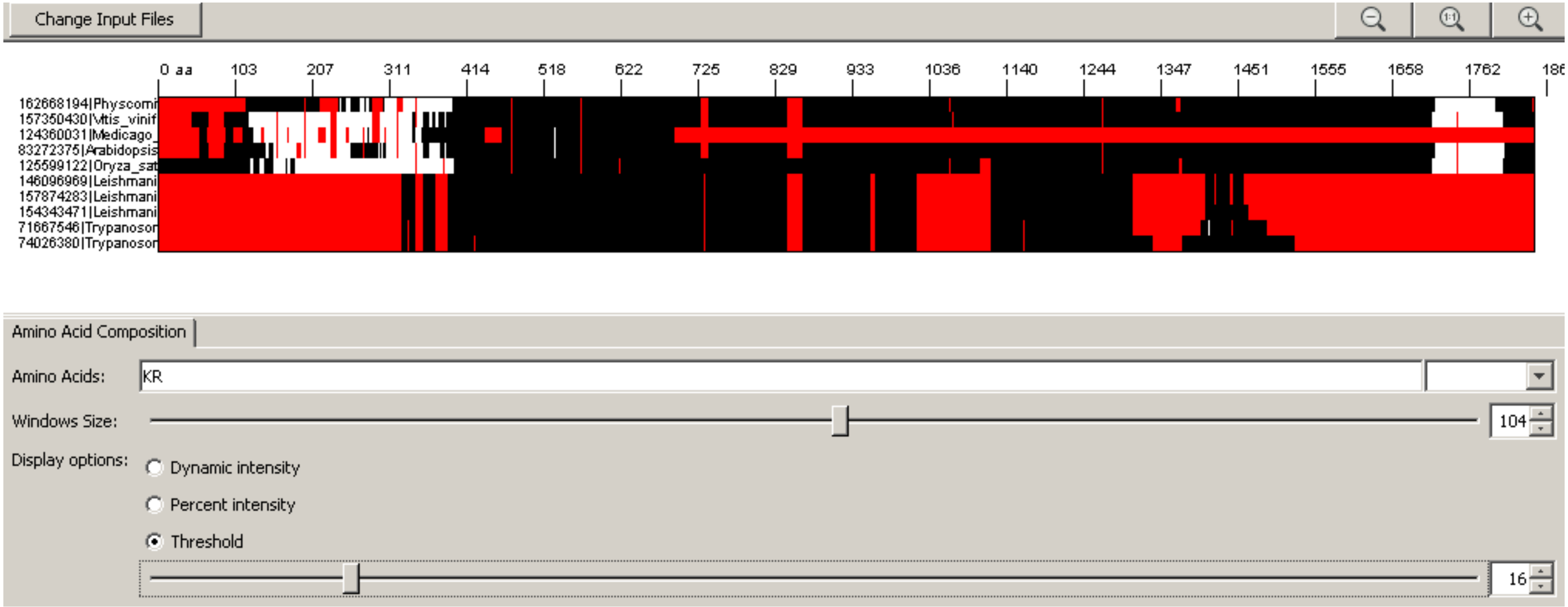
15

http://biasviz.sourceforge.net/

Huska, et al. (2007). *Bioinformatics*

A particular analysis…

…triggers BiasViz



Amino Acids: KR

Windows Size: 104

Display options:

○ Dynamic intensity

○ Percent intensity

● Threshold 16

http://biasviz.sourceforge.net/

Huska, et al. (2007). *Bioinformatics*

Exercise 3/4. Viewing CBRs in an alignment with BiasViz2

- Go to the BiasViz2 web page: http://biasviz.souceforge.net/

- Launch BiasViz2

- Load the alignment little_MSA_fasta.txt on the step 1 section

- Hit the "Go to graphical view" button

- Try to find combinations of parameters that reveal CBRs

- Try hydrophobic residues and window size 10. Remember that this is a transmembrane protein.
  What is this result telling you?

- Can you see other biased regions?

Exercise 4/4. Viewing CBRs in an alignment with BiasViz2

- Exit BiasViz2 and launch it again

- Load the alignment MR1_fasta.txt on the step 1 section

- Hit the "Go to graphical view" button

- Try to find combinations of parameters that reveal CBRs

- Can you find a large (>100aa) Serine rich region? (In Display options, try the threshold option with 25% cut-off)