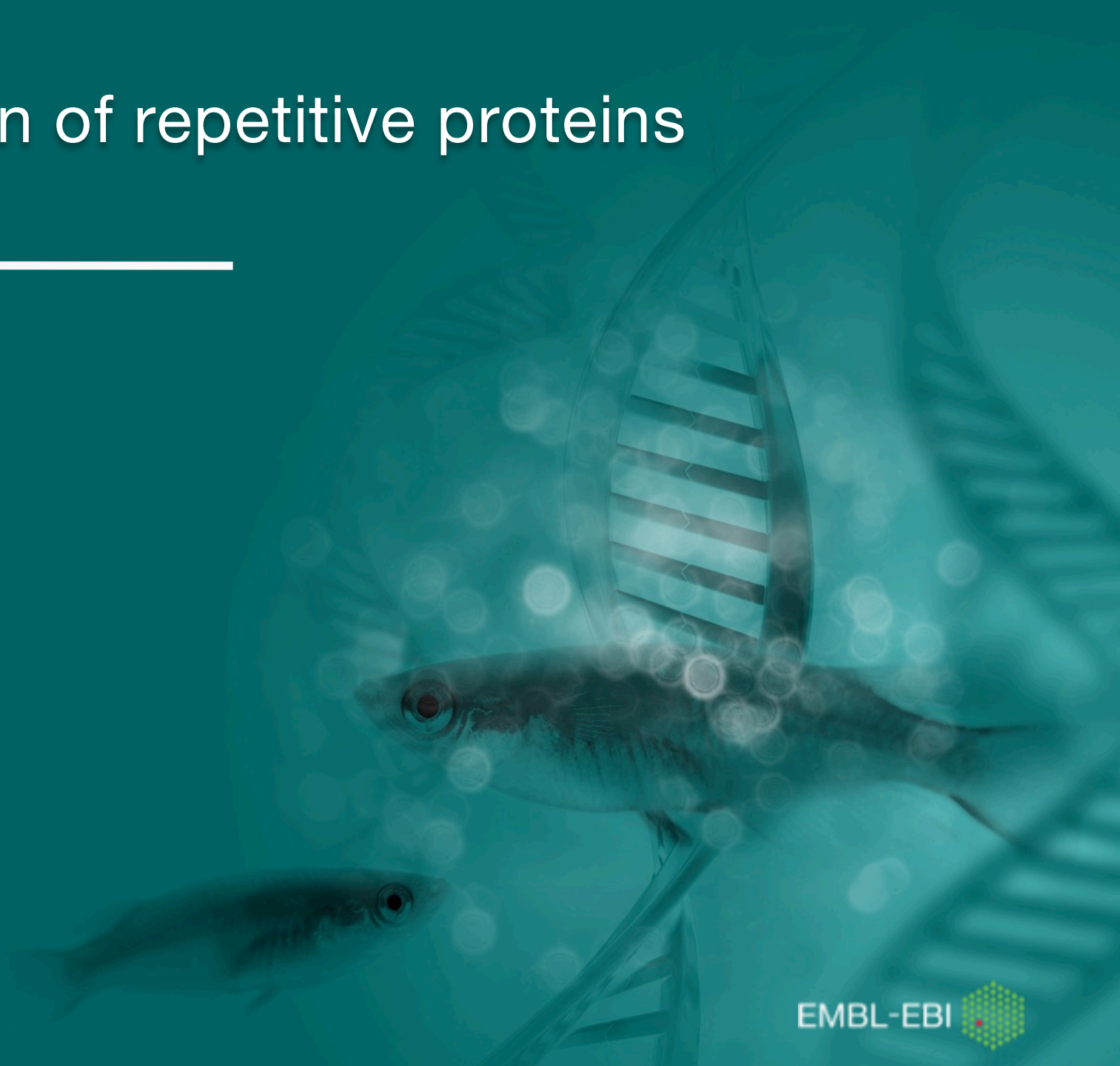
Detection and evolution of repetitive proteins

Aleix Lafita

Bateman Research Group

EMBL-EBI Webinar

27th June 2018



Outline

1. Why should we care about protein repeats?

2. Detection of protein repeats

3. Duplication patterns and evolution

- Example protein throughout the webinar

What can you say about this protein sequence?

MVNRVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIE
EVTNKDTAPQGVEAKSEVTSNKDTIEHEPSVVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQ
EAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHKAEDIVDFLSNK
LNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQG
VEAKSEVTSNKDTIEHSKKEVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIEN
PTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHDTPKE

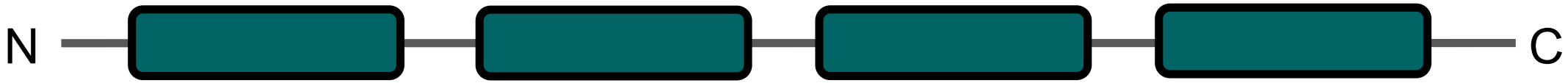
What can you say about this protein sequence?

MVNRVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIE
EVTNKDTAPQGVEAKSEVTSNKDTIEHEPSVVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQ
EAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHKAEDIVDFLSNK
LNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQG
VEAKSEVTSNKDTIEHSKKEVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIEN
PTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHDTPKE

What can you say about this protein sequence?

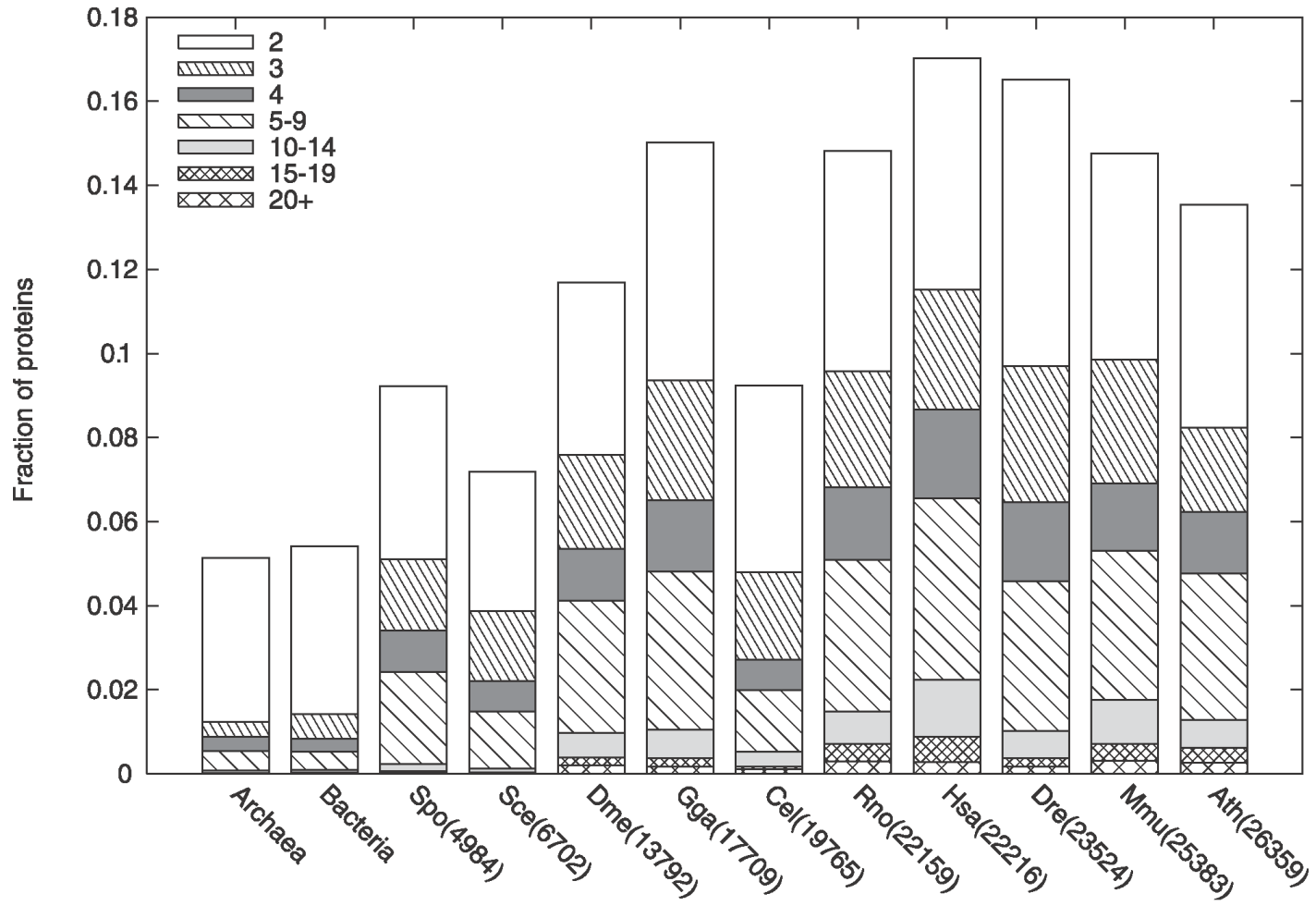
MVNRVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIE
EVTNKDTAPQGVEAKSEVTSNKDTIEHEPSVVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQ
EAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHKAEDIVDFLSNK
LNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQG
VEAKSEVTSNKDTIEHSKKEVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIEN
PTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHDTPKE

What can you say about this protein sequence?



MVNRVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIE
EVTNKDTAPQGVEAKSEVTSNKDTIEHEPSVVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQ
EAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHKAEDIVDFLSNK
LNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIENPTTLKDNVQSKEVKIEEVTNKDTAPQG
VEAKSEVTSNKDTIEHSKKEVDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIEN
PTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEHDTPKE

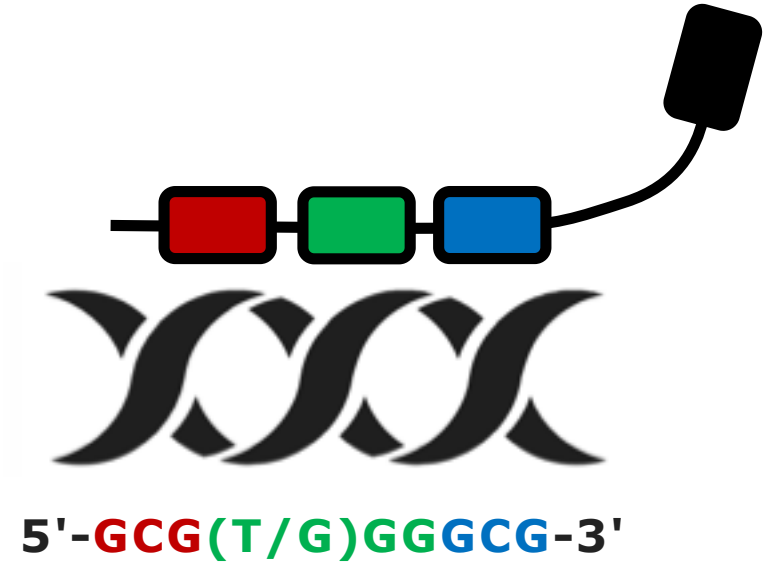
How frequent are tandem repeats in proteins?

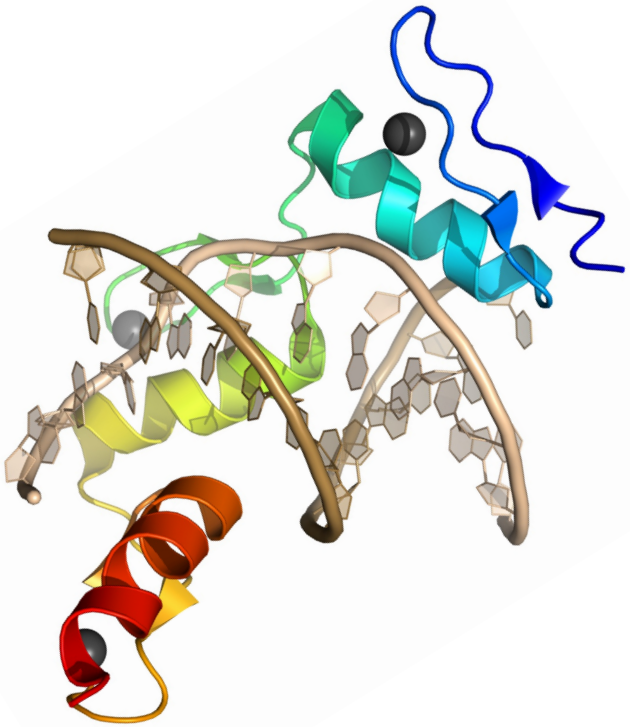


Björklund *et al.* PLOS Comp. Biol. (2006)

Functional relevance of protein tandem repeats

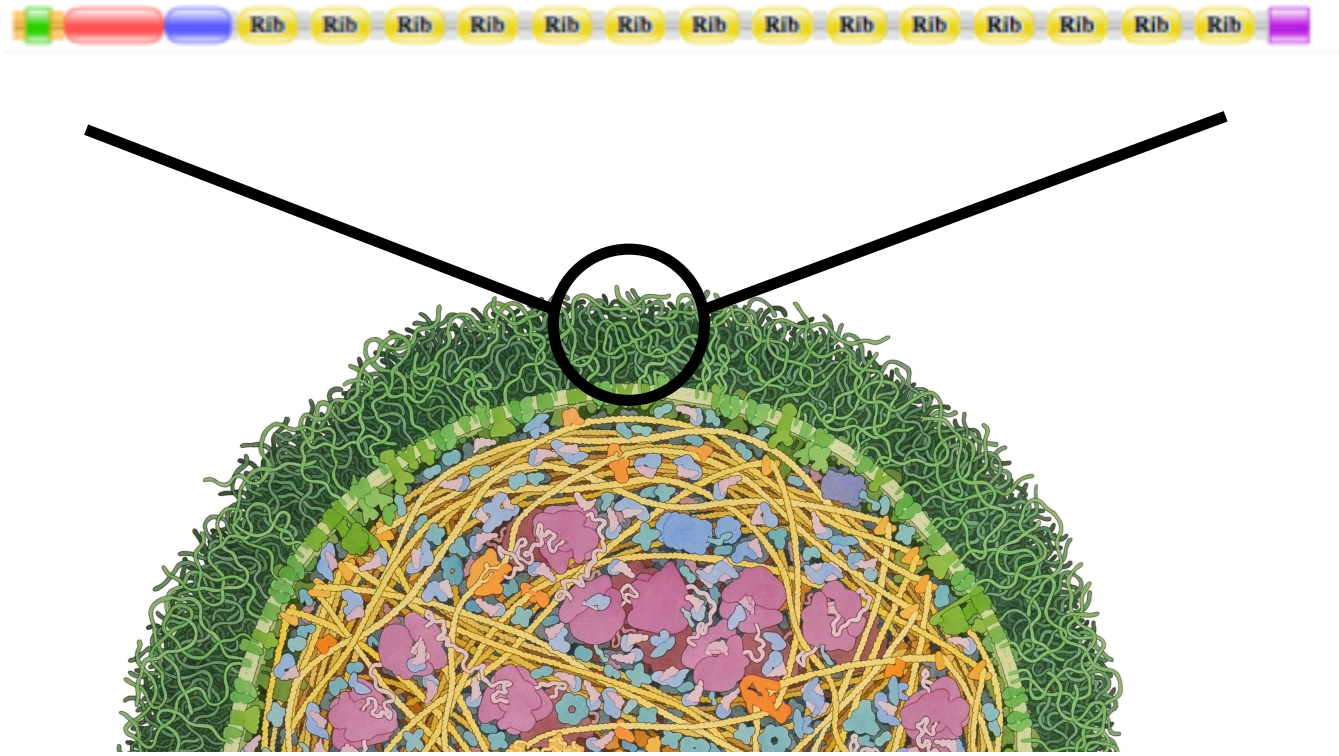
- Binding specificity
  - Transcription factors
  - Protein-protein interactions



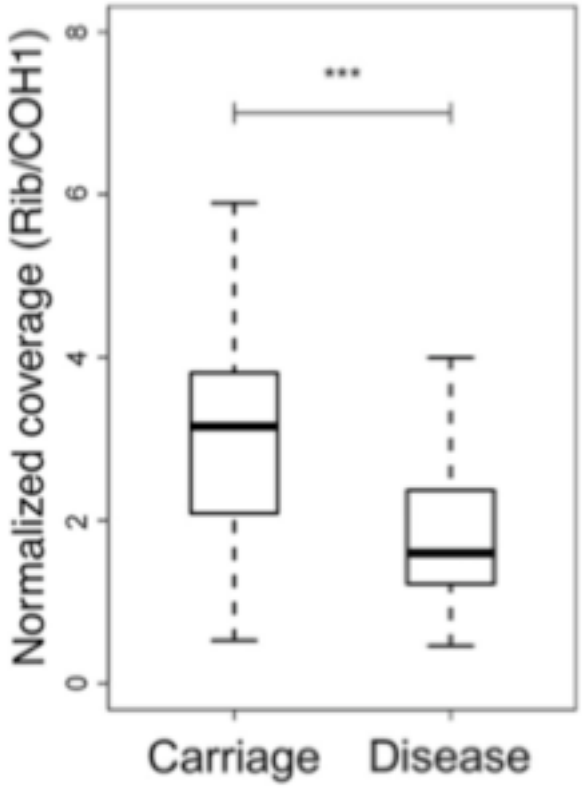


Tandem protein repeats provide an adaptation advantage

Staphylococcus aureus (GBS) cell surface



~ number of tandem domain repeats



Domain of Unknown Function 1220

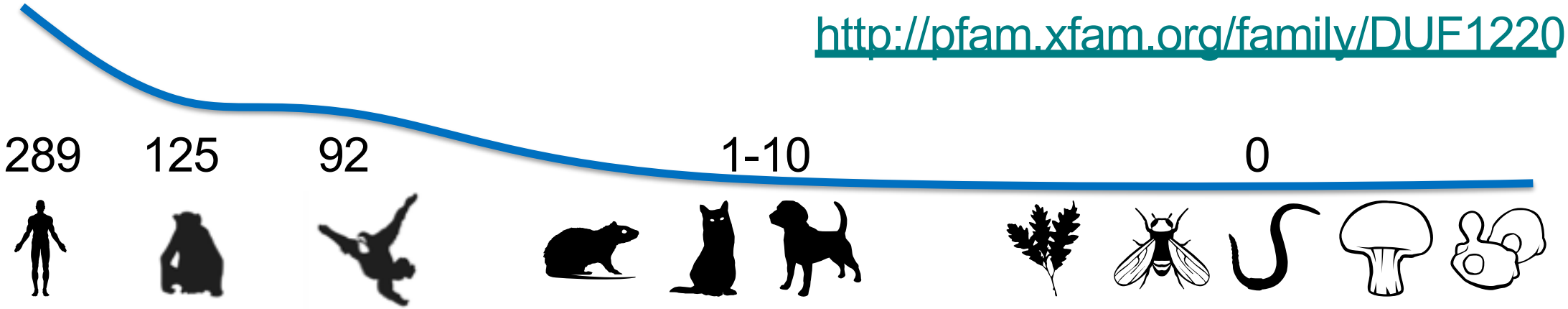
EXAMPLE

DUF1220

From Wikipedia, the free encyclopedia

DUF1220 is a protein domain that shows a striking human lineage-specific (HLS) increase in copy number and may be important to human brain evolution.[1] The DUF1220 domain name has recently been changed to the Olduvai domain based on data obtained since initial discovery of the domain. [2] The copy number of DUF1220 domains increases generally as a function of a species evolutionary proximity to humans. DUF1220 copy number is highest in human (~289, with some person-to-person variations).[3] and shows the largest HLS increase in copy number (an additional 160 copies) of any protein coding region in the human genome. DUF1220 copy number is reduced in African great apes (estimated 125 copies in chimpanzees), further reduced in orangutan (92) and Old World monkeys (35), single- or low-copy in non-primate mammals and absent in non-mammals.[3] DUF1220 domains are approximately 65 amino acids in length and are encoded by a two-exon doublet. In the human genome DUF1220 sequences are located primarily on chromosome 1 in region 1q21.1-q21.2, with several copies also found at 1p36, 1p13.3, and 1p12. Sequences encoding DUF1220 domains show rhythmicity, resonance[4] and signs of positive selection, especially in primates, and are expressed in several human tissues including brain, where their expression is restricted to neurons.[1]

http://pfam.xfam.org/family/DUF1220



Neuroblastoma breakpoint family member 14 (NBPF14)

>tr|A0A087WZJ2|A0A087WZJ2_HUMAN Neuroblastoma breakpoint family member 14 OS=Homo sapiens OX=9606 GN=NBPF14 PE=4 SV=1

MVVSAGPWSSEKAEMNILEINETLRPQLAEKKQQFRNLKEKCFLTQLAGFLANQQKKYKYEECKDLIKFMLRNERQFKEEKLAEQLKQAEELRQYKVLVHSQERELTQLREKLREGRDASRSLY
EHLQALLTPYEPDKSQGQDLQEQLAEGCRLAQHLVQKLSPENDEDEDEDVQVEEAEKVLESSAPREVQKAEESKVPEDSLEECAITCSNSHGPCDSNQPHKNIKITFEEDEVNSTLVVDRESSH
DECQDALNILPVPGPTSSATNVSMVVSAGPLSSEKAEMNILEINEKLRPQLAEKKQQFRNLKEKCFLTQLSGFLANQQKKYKYEECKDLIKFMLRNERQFKEEKLAEQLKQAEELRQYKVLVHAQ
ERELTQLREKLREGRDASRSLNEHLQALLTPDEPDKSQGQDLQEQLAEGCRLAQHLVQKLSPENDNDDDEDVQVEVAEKVQKSSAPREMQKAEEKEVPEDSLEECAITYSNSHGSYDSNQPH
RKTKITFEEDKVDSTLIGSSSHVEWEDAVHIIPENESDDEEEEEKGPVSPRNLQESEEEEVPQESWDEGYSTLSIPPEMLASYQSYSSTFHSLEEQQVCMAVDIGRHRWDQVKKEDQEATGPRLS
RELLDEKGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYQEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLG
LALDVDRIKKDEEEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSREL

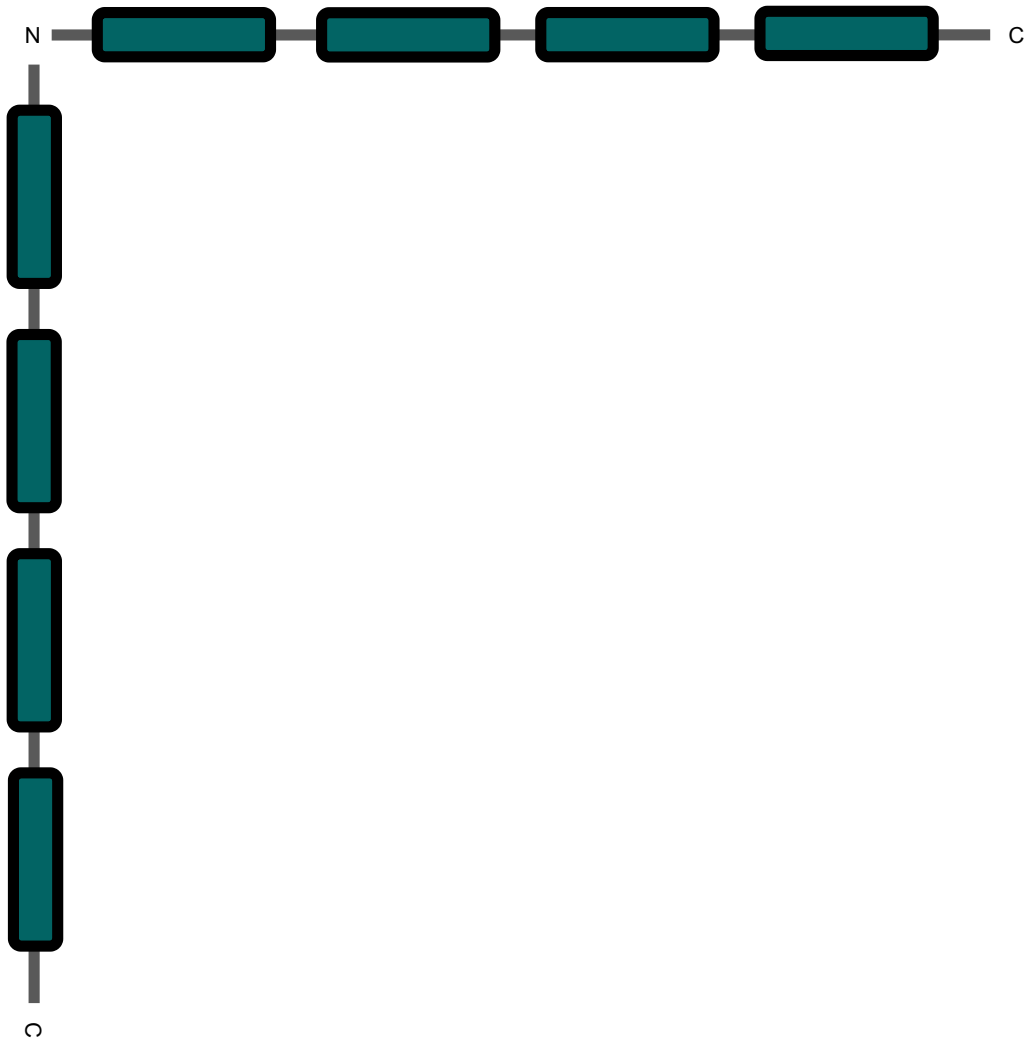
LDEKGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYQEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLAL
DVDRIKKDEEEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDE
KGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYQEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVD
RIKKDEEEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLHEKGP
EVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYILEQQRVGLAVDMDEIEKYKEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDRFKK
DEEEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDEKGPEVLQ
DSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYKEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDRIKKDQE
EEEDQGPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDEKGPEVLQDSL
DRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYKEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDRIKKDQEEEE

DQGPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDEKGPEVLQDSLDR
CYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYKEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDRIKKDQEEEEDQ
GPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDEKGPEVLQDSLDRCY
STPSGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYKEVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDRIKKDQEEEEDQGP
PCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCPRLSRELLDEKGPEVLQDSLDRCYSTP
SGCLELTDSCQPYRSAFYVLEQQRVGLAVDMDEIEKYQEVEEDQDPSCPRLSRELLDEKDPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLALDVDKIEKKGKGKKRRGRRS
KKERRRGRKEGEEDQNPPCPRLNGVLMEVEEREVLQDSLDRCYSTPSMYFELPDSFQHYRSVFYSFEEQHISFALYVDNRFFTLTVTSLHLVFQMGVIFPQ

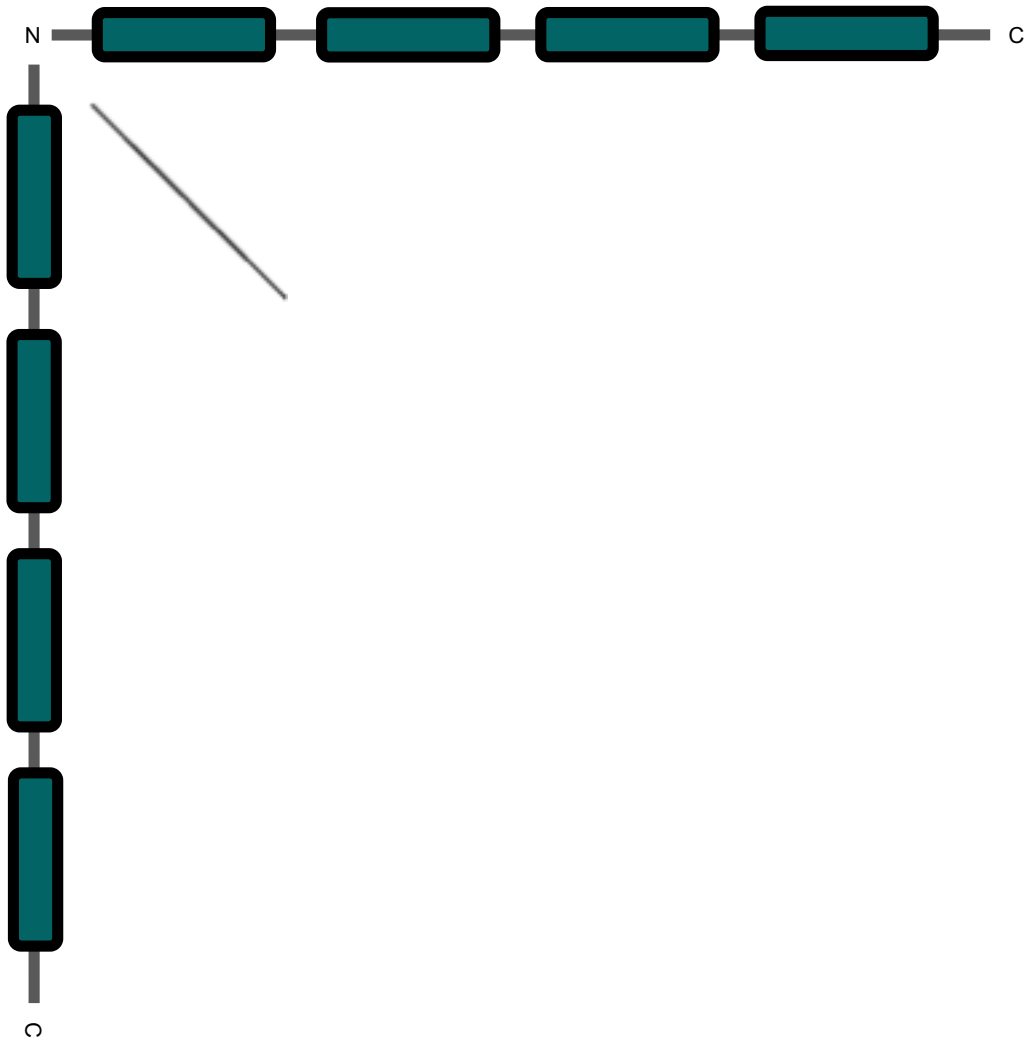
https://www.uniprot.org/uniprot/A0A087WZJ2 http://pfam.xfam.org/protein/A0A087WZJ2

How can we identify the repeats automatically?

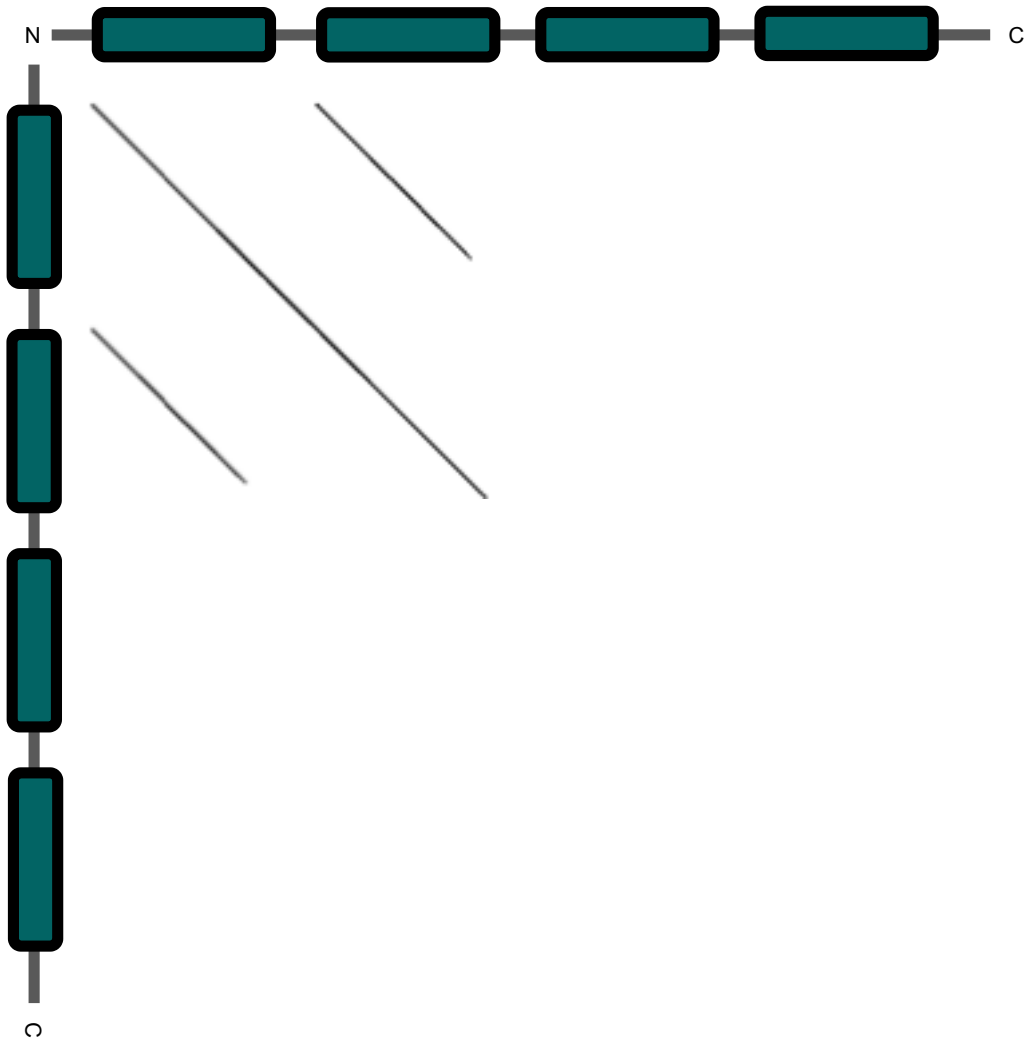
The self-comparison of a protein revels sequence repeats



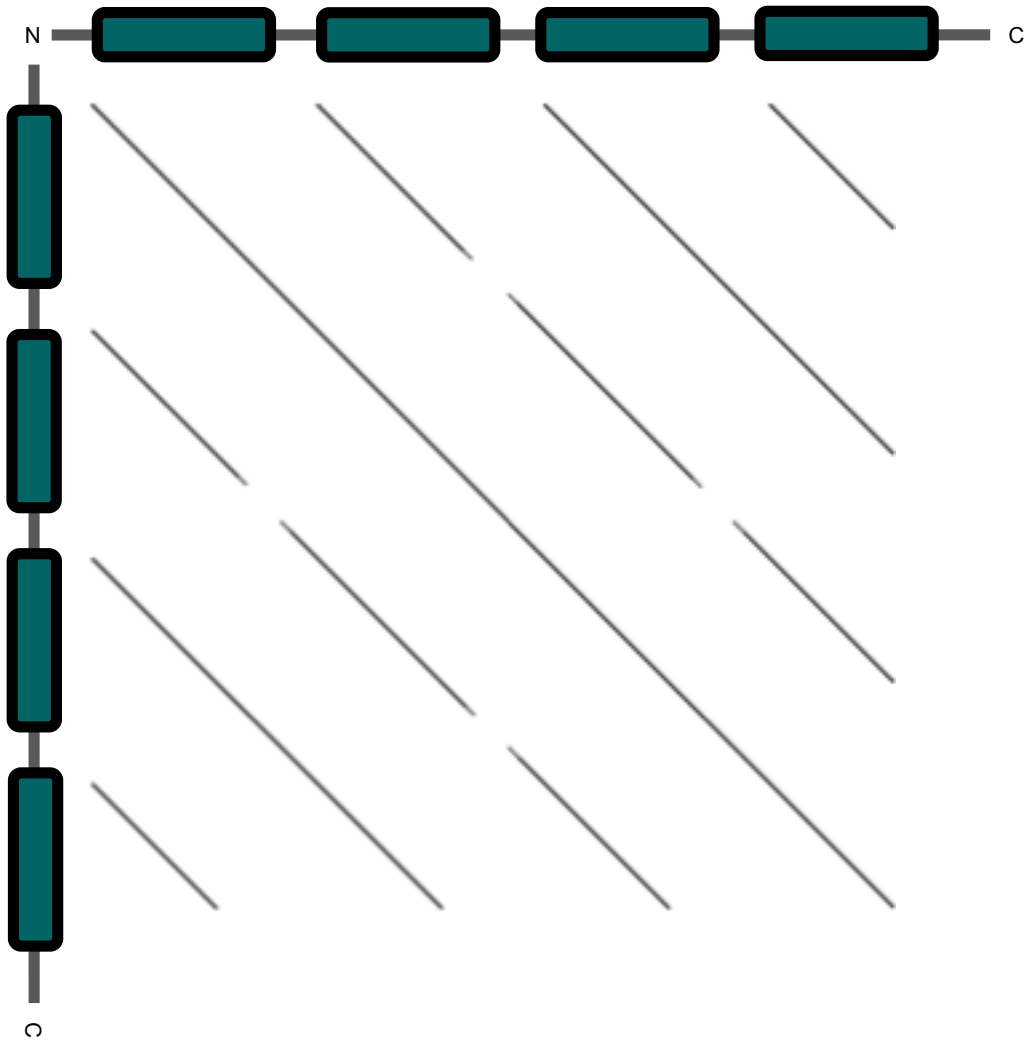
The self-comparison of a protein revels sequence repeats



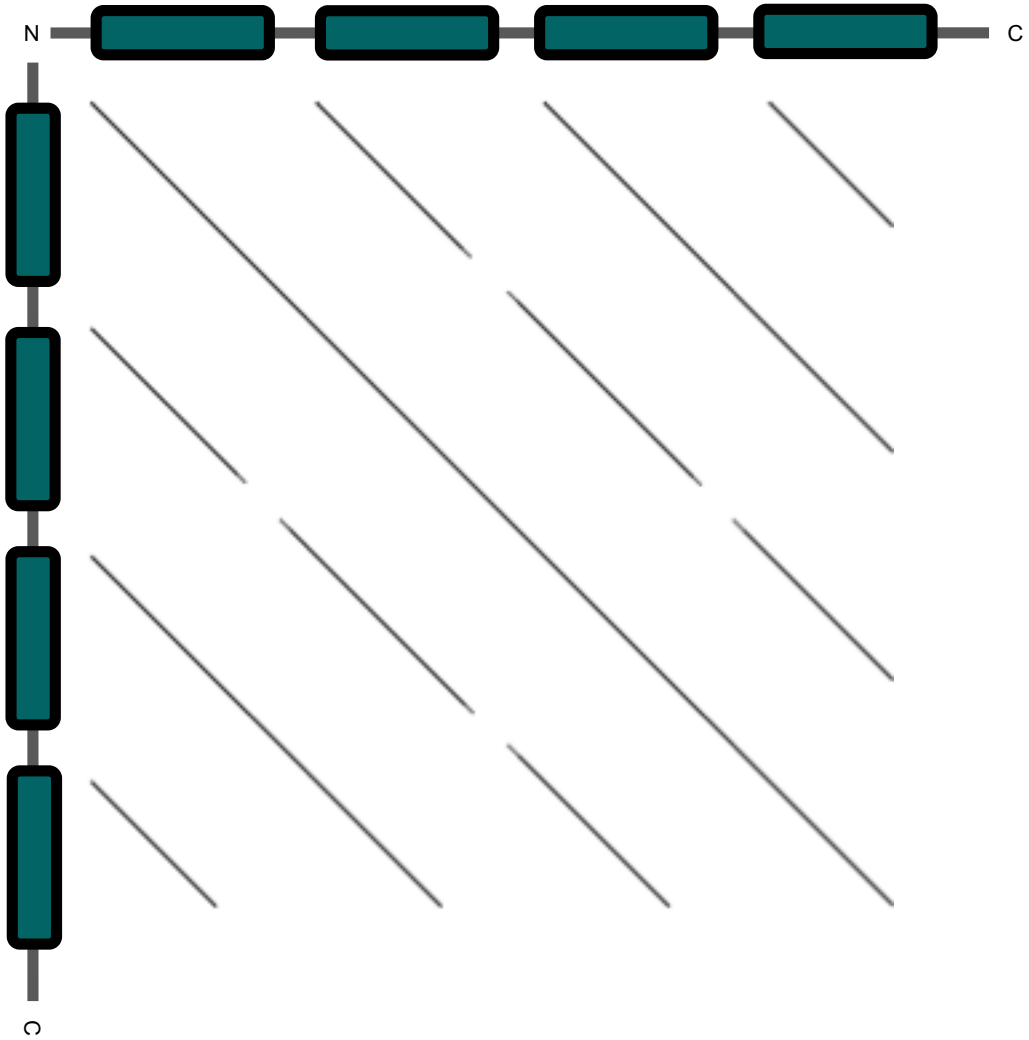
The self-comparison of a protein revels sequence repeats



The self-comparison of a protein revels sequence repeats



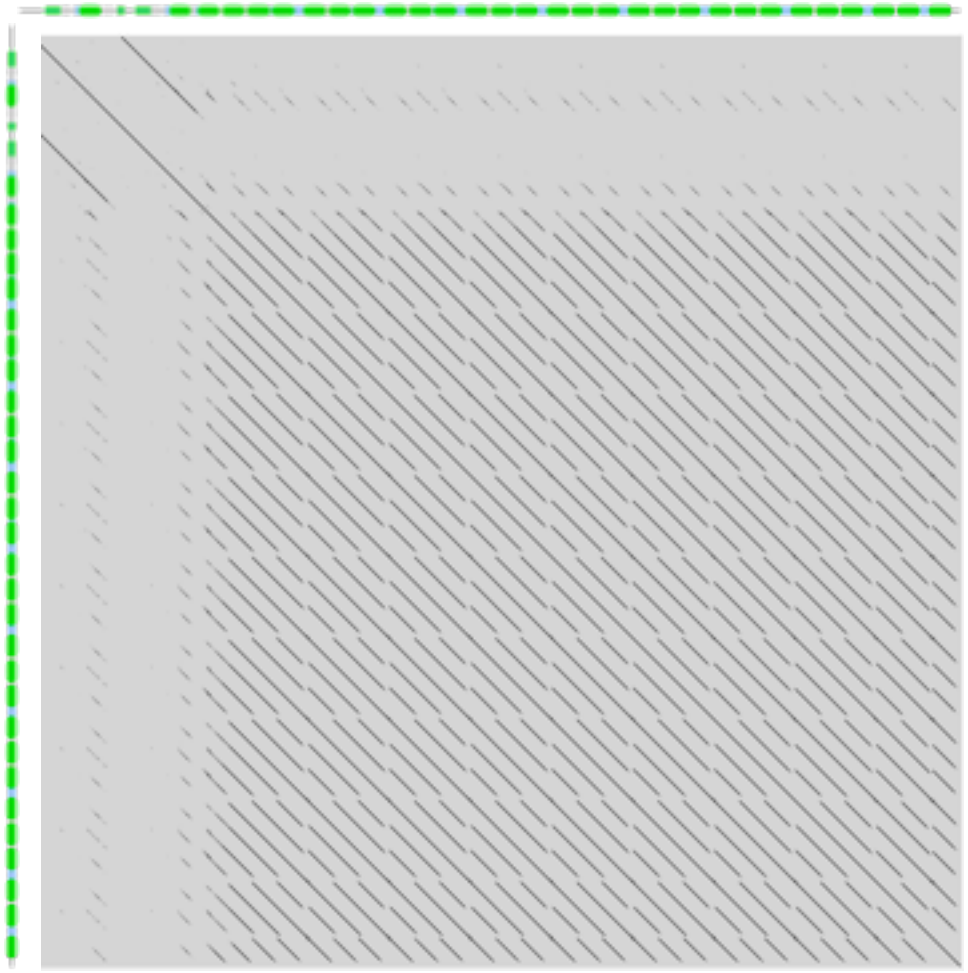
Resources to calculate and visualize sequence dot-plots



Dotter (desktop)
https://www.sanger.ac.uk/science/tools/seqtools

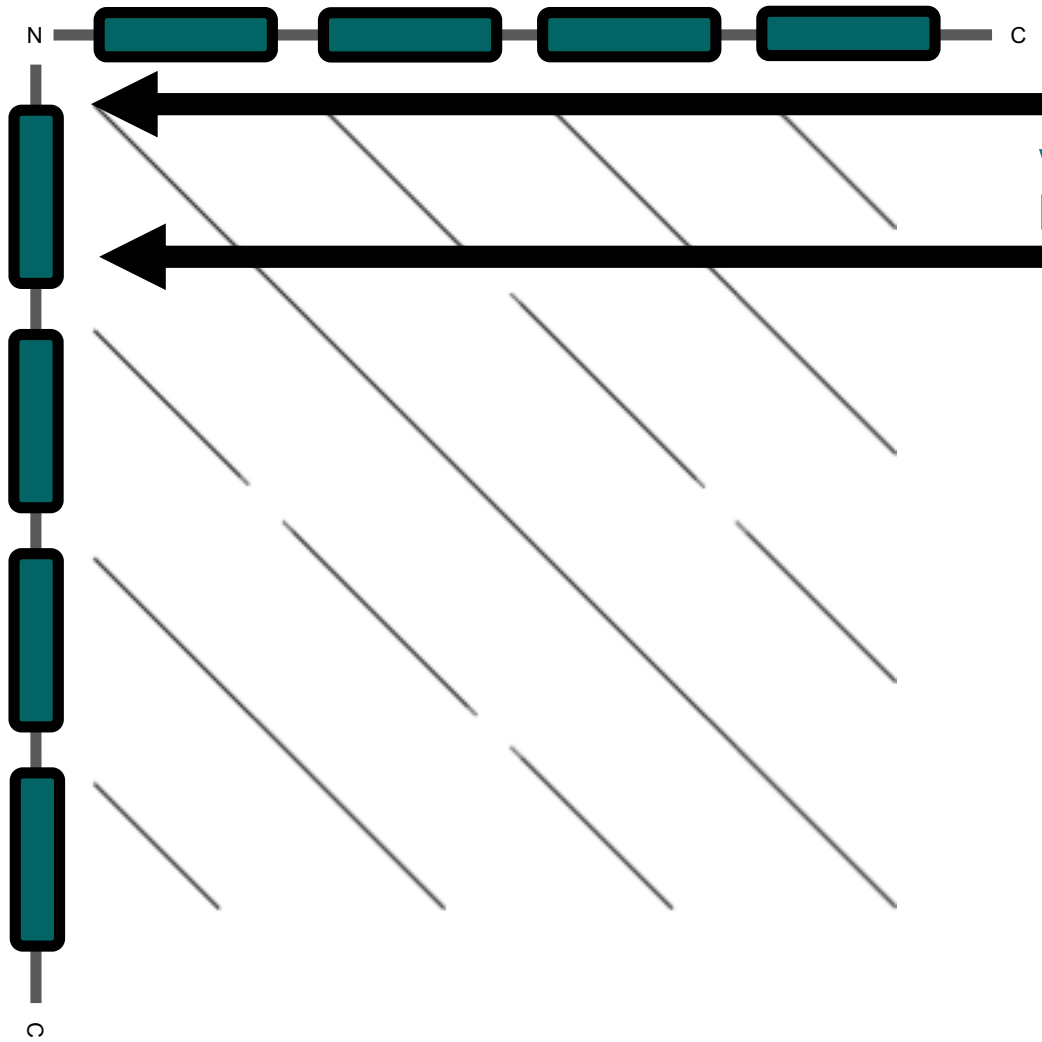
Dotlet JS (web)
https://dotlet.vital-it.ch

Self-dot plot of NBPF14



```
$ dotter A0A087WZJ2.fasta A0A087WZJ2.fasta
```

Methods to extract repeats from protein self-alignments



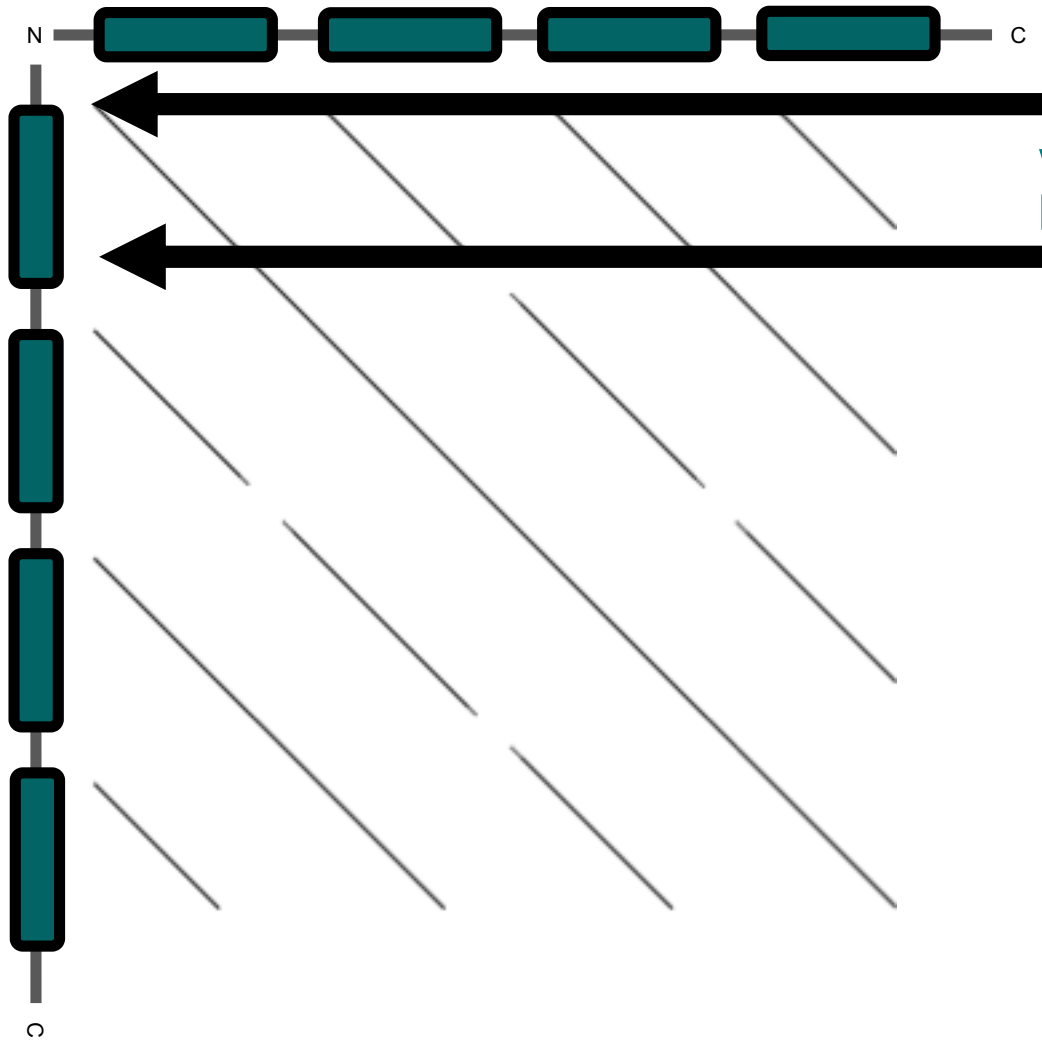
VDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIE
NPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEH

RADAR
https://www.ebi.ac.uk/Tools/pfa/radar/

Swelfe
http://wwwabi.snv.jussieu.fr/public/Swelfe/

T-REKS
http://bioinfo.montp.cnrs.fr/index.php?route=tools&tool=3

Tandem repeats as a multiple sequence alignment



VDFLSNKLNKYSIRKFTVGTASILIGSLMYLGTQQEAEAAENNIE
NPTTLKDNVQSKEVKIEEVTNKDTAPQGVEAKSEVTSNKDTIEH

```
Example/1-100      V-FLSNKLN......TNKDTL-H
Example/101-200    VDF-SNKLN......TNKDTIEH
Example/201-300    V-FLSNKLN......TKKDTL--
Example/301-400    VDF-SNKLN......SKKDTI-H
```

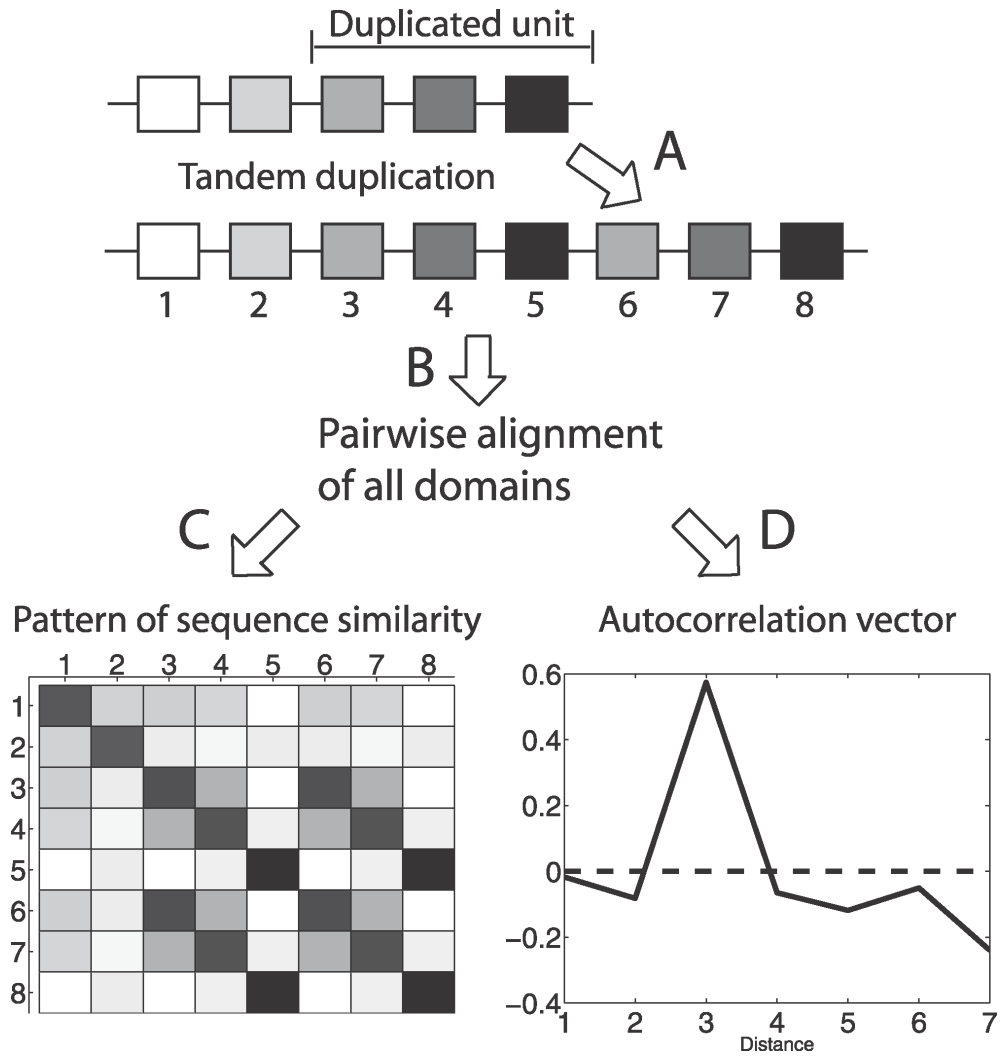
Multiple sequence alignment of NBPF14 domains


```
$ grep A0A087WZJ2 PF06758_full.sto > PF06758_A0A087WZJ2.sto
```

```
A0A087WZJ2_HUMAN/177-240          -EEAEKVLESSAP---REVQKAEESKVPEDSLEECAITCSNSHGPCDSNQPHKNIKITFEEDEVNSTL-
A0A087WZJ2_HUMAN/449-511          VA--EKVQKSSAP---REMQKAEEKEVPEDSLEECAITYSNSHGSYDSNQPHRKTKITFEEDKVDSTL-
A0A087WZJ2_HUMAN/535-597          --EEEEEKGPVSP---RNLQESEEEEVPQESWDEGYSTLSIPPEMLASYQSYSSTFHSLEEQQVCMA-V
A0A087WZJ2_HUMAN/607-672          -V-KKEDQEATGPRLSRELLDEKGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLA-V
A0A087WZJ2_HUMAN/682-747          --EVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLAL-
A0A087WZJ2_HUMAN/757-822          --EEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSL-
A0A087WZJ2_HUMAN/851-916          --EGEEDQNPPCPRLSRELLDEKGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYVLEQQRVGLA-V
A0A087WZJ2_HUMAN/926-991          --EVEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDLGQPYSSAVYSLEEQYLGLAL-
A0A087WZJ2_HUMAN/1001-1066        --EEEEDQDPPCPRLSRELLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKHVGFSL-
...
...
```

http://pfam.xfam.org/family/PF06758/alignment/full/format?format=stockholm&alnType=full&order=a&case=l&gaps=default&download=0

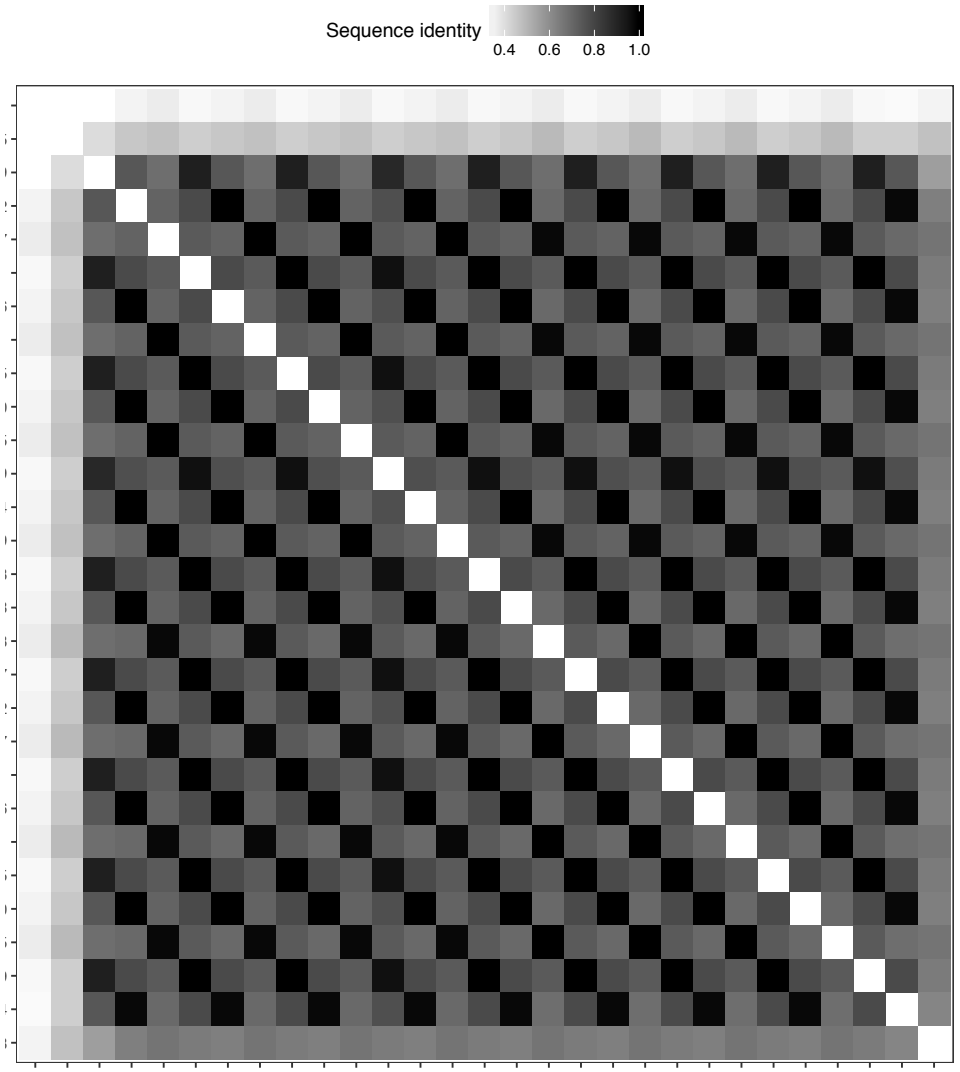
Repeat similarity patterns reveal protein duplication events

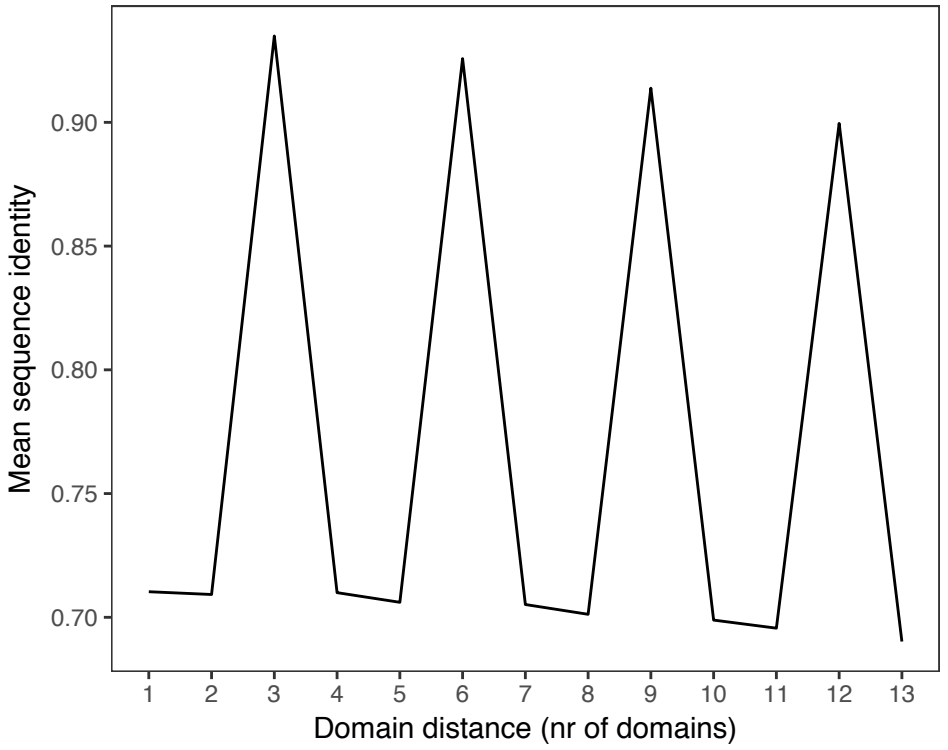


Sequence similarity matrix and duplication patterns

	1	2	3	4	5
1		0.5	0.9	0.4	0.8
2	0.5		0.4	0.9	0.4
3	0.9	0.4		0.5	0.8
4	0.4	0.9	0.5		0.4
5	0.8	0.4	0.8	0.4	

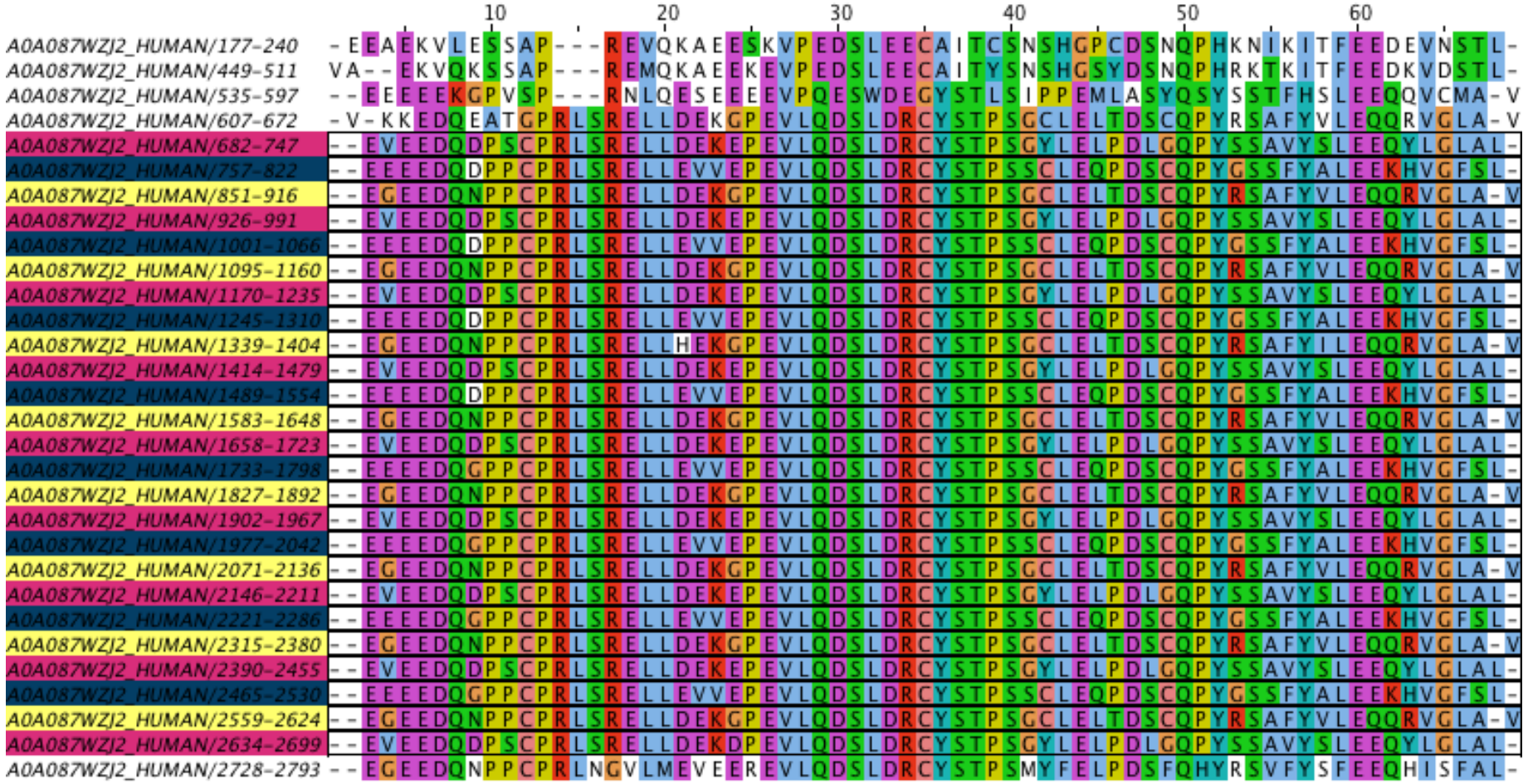
Duplication patterns in NBPF14





The duplication unit in NBPF14

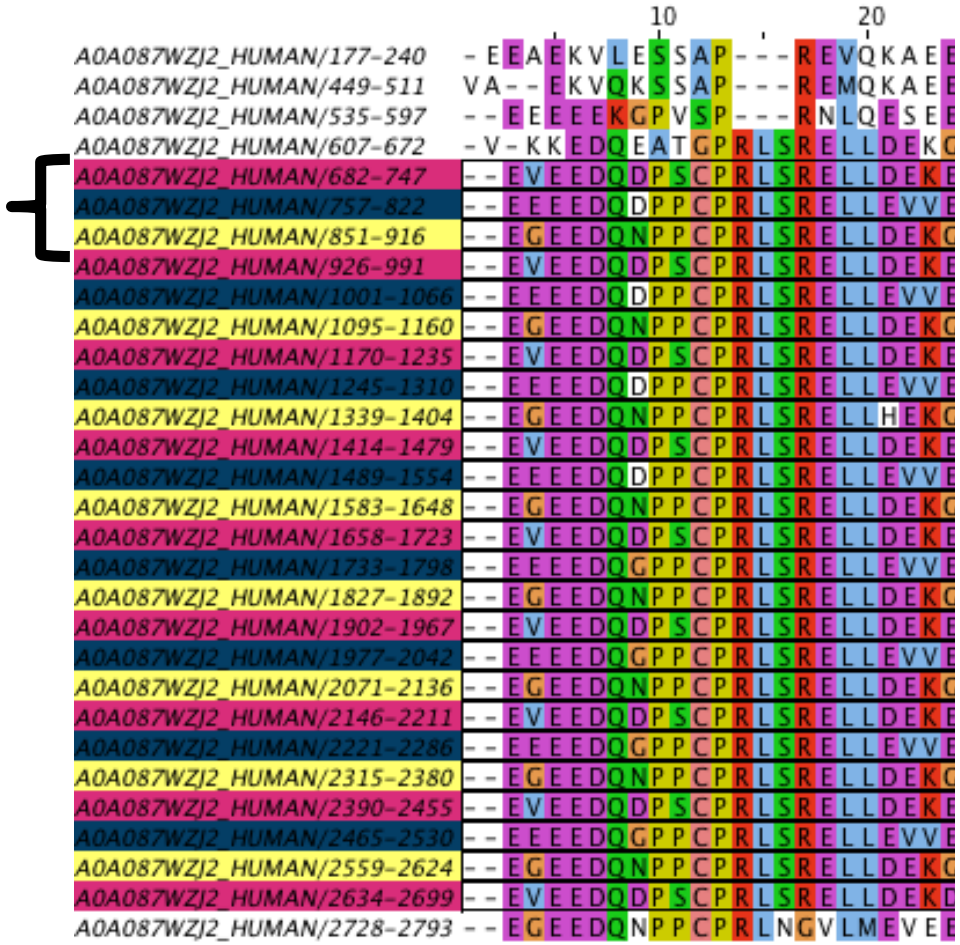
EXAMPLE



The duplication unit in NBPF14

EXAMPLE

>A0A087WZJ2/682-916
VEEDQDPSCPRLSRELLDEKEPEVLQDSLDRCYSTPSGYLELPDL
GQPYSSAVYSLEEQYLGLALDVDRIKKDEEEEEDQDPPCPRLSRE
LLEVVEPEVLQDSLDRCYSTPSSCLEQPDSCQPYGSSFYALEEKH
VGFSLDVGEIEKKGKGKKRRGRRSKKERRRGRKEGEEDQNPPCP
RLSRELLDEKGPEVLQDSLDRCYSTPSGCLELTDSCQPYRSAFYV
LEQQRVGLAVDMDEIEKYQE



https://www.ebi.ac.uk/Tools/hmmer/search/phmmer

Summary

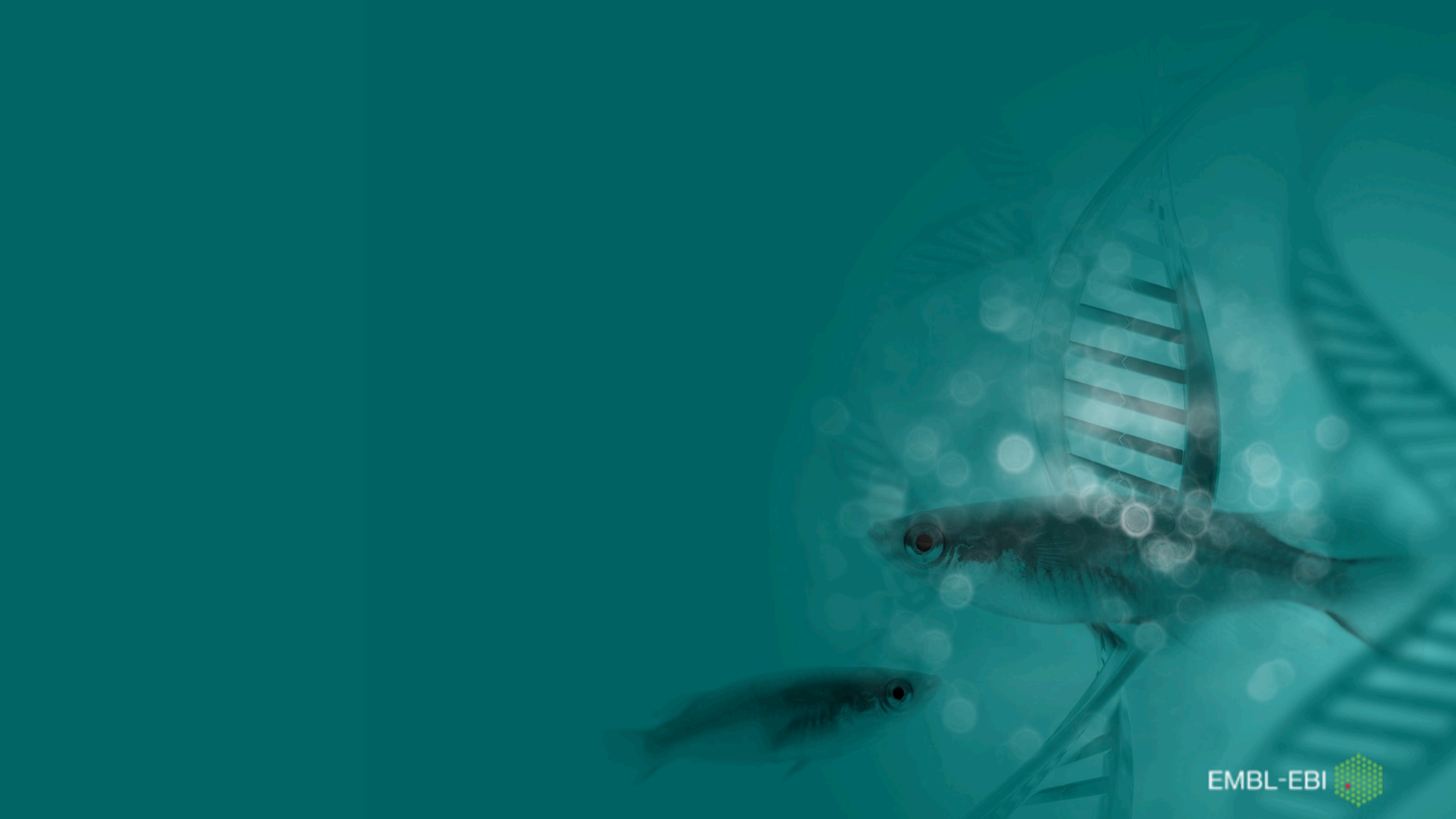
- Tandem repeats are important for protein function and evolution.
- Tools for repeat detection and visualization.
- Evolutionary events can be inferred from sequence similarity patterns.
- Identify recurring protein repeat units.

Aleix Lafita

https://www.ebi.ac.uk/about/people/aleix-lafita-masip

aleixlafita@ebi.ac.uk

@alafitamasip


EMBL-EBI