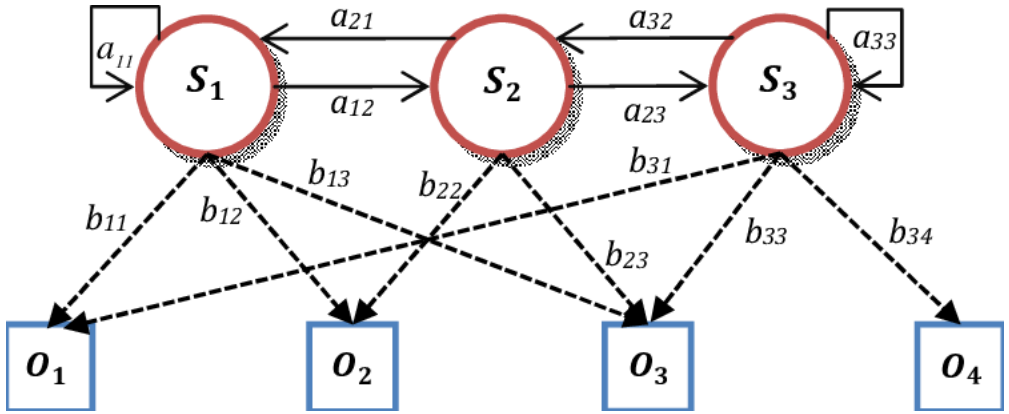
DIPARTIMENTO MATEMATICA
DIPARTIMENTO DI MATEMATICA "TULLIO LEVI-CIVITA"

800 ANNI 1222·2022 UNIVERSITÀ DEGLI STUDI DI PADOVA



HIDDEN MARKOV MODELS (HMM) for detection of intrinsically disordered regions of proteins

Alina Skrylnik

MSc. Data Science (Biological Data Analytics)

2023-2024

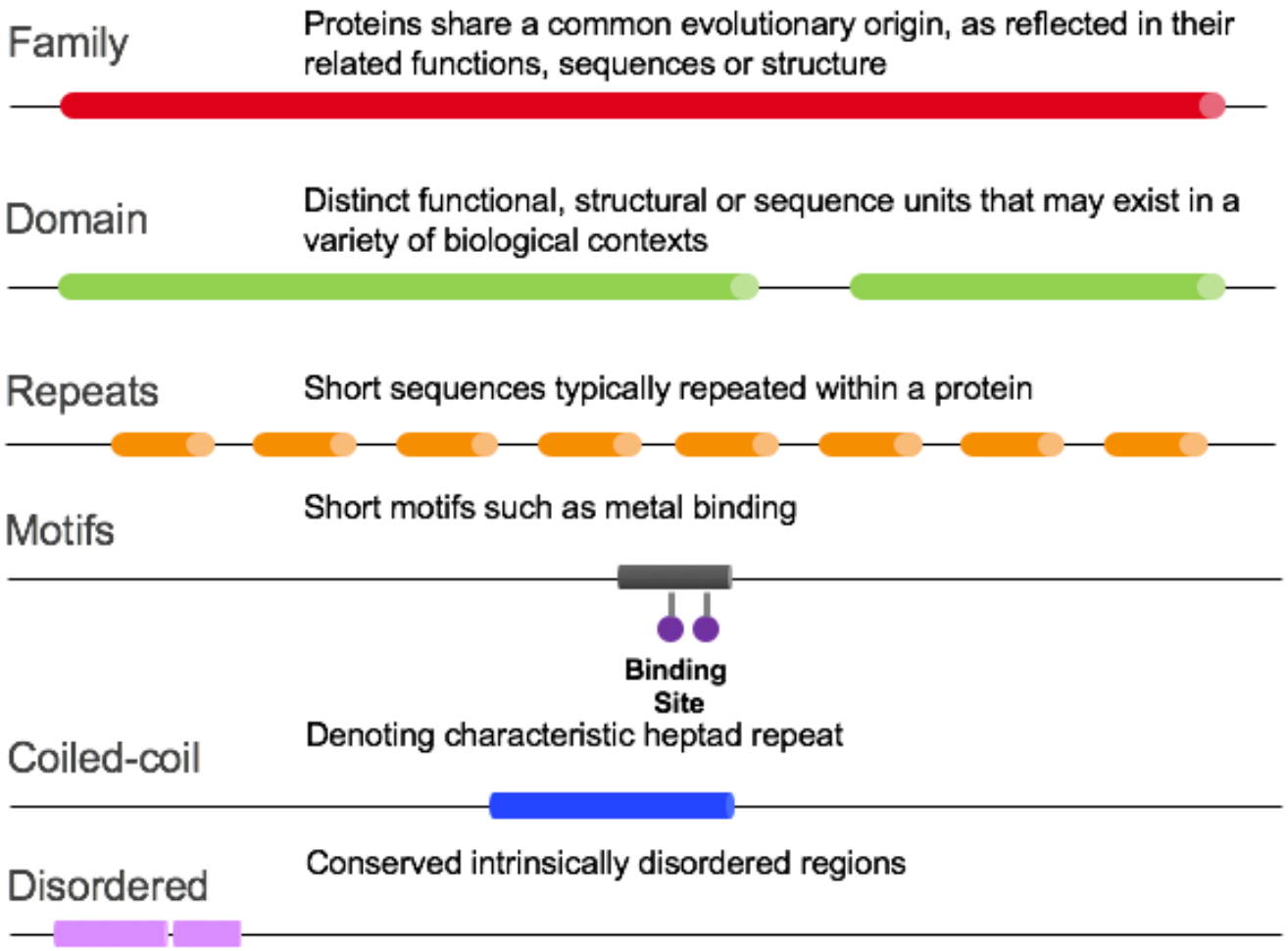
Pfam

Protein families database

Domains are distinct functional and structural units in a protein.

Intrinsically disordered regions (**IDRs**) are considered to be a domain type.

The prediction of Pfam entries is based on the profile HMMs



EMBL-EBI Training **"Pfam: Creating protein families. Pfam entry types"**
https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/pfam-entry-types/

Profile HMM (pHMM):

A probabilistic model that encapsulates the evolutionary changes that have occurred in a set of related sequences.

Start with a multiple sequence alignment

↓

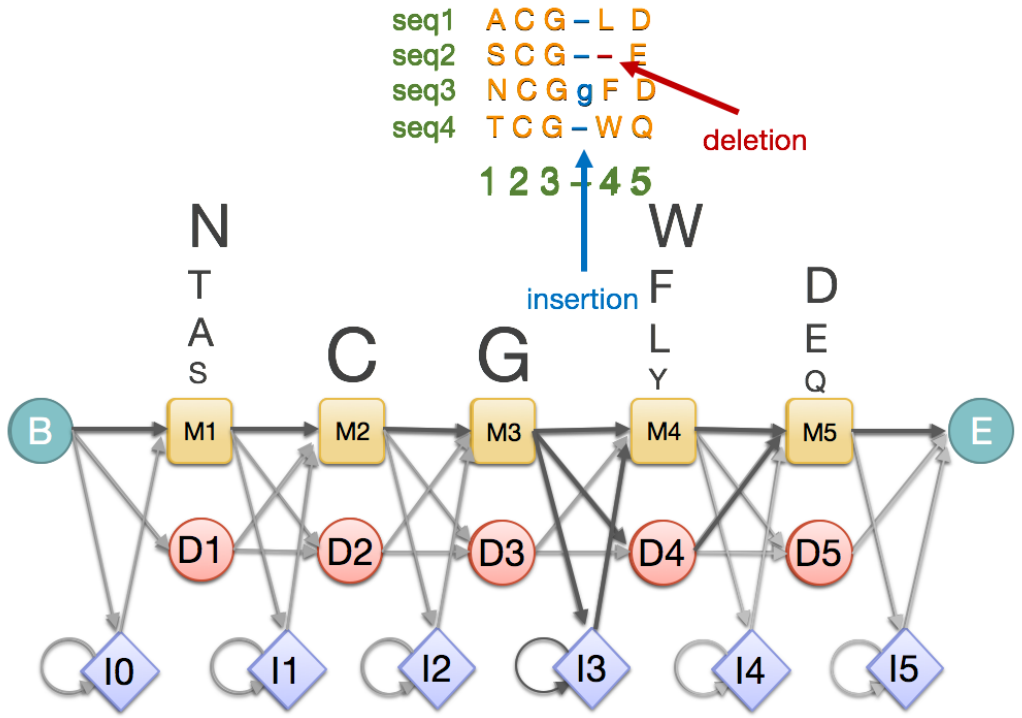
Insertions / deletions can be modelled

↓

Occupancy and amino acid frequency at each position in the alignment are encoded

↓

Profile created



Hidden states: IDRs

Observable states: amino acid sequences

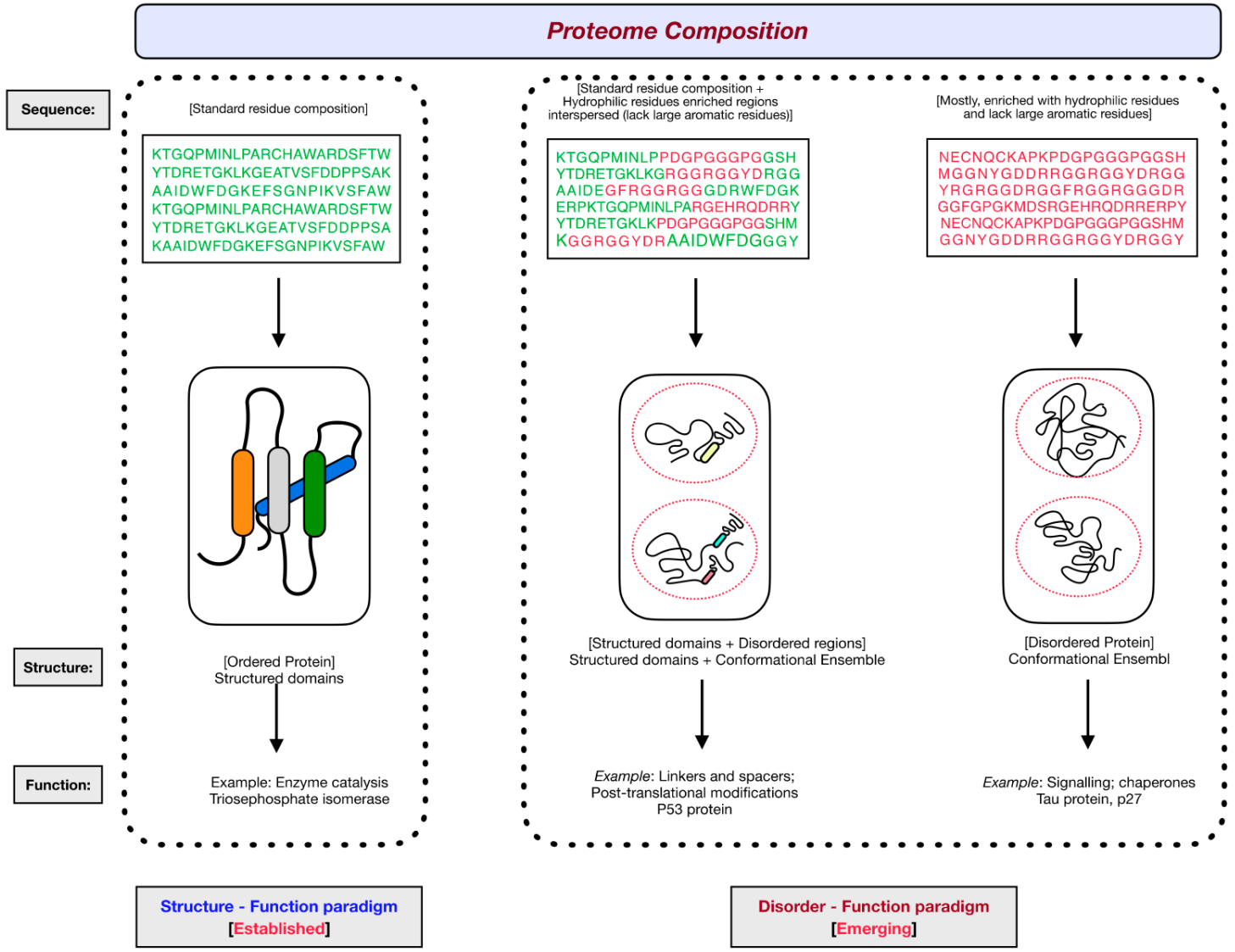
M: match state

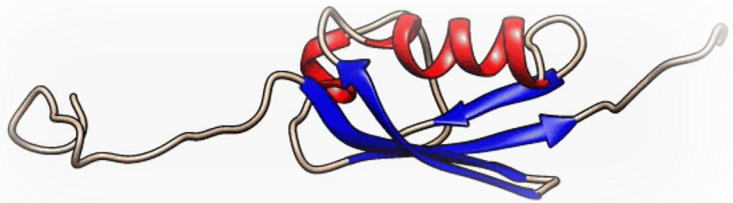
I: insertion state

D: deletion state

EMBL-EBI Training **"Pfam: Creating protein families. What are profile hidden Markov models?"**
https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/

Structure vs disordered paradigms





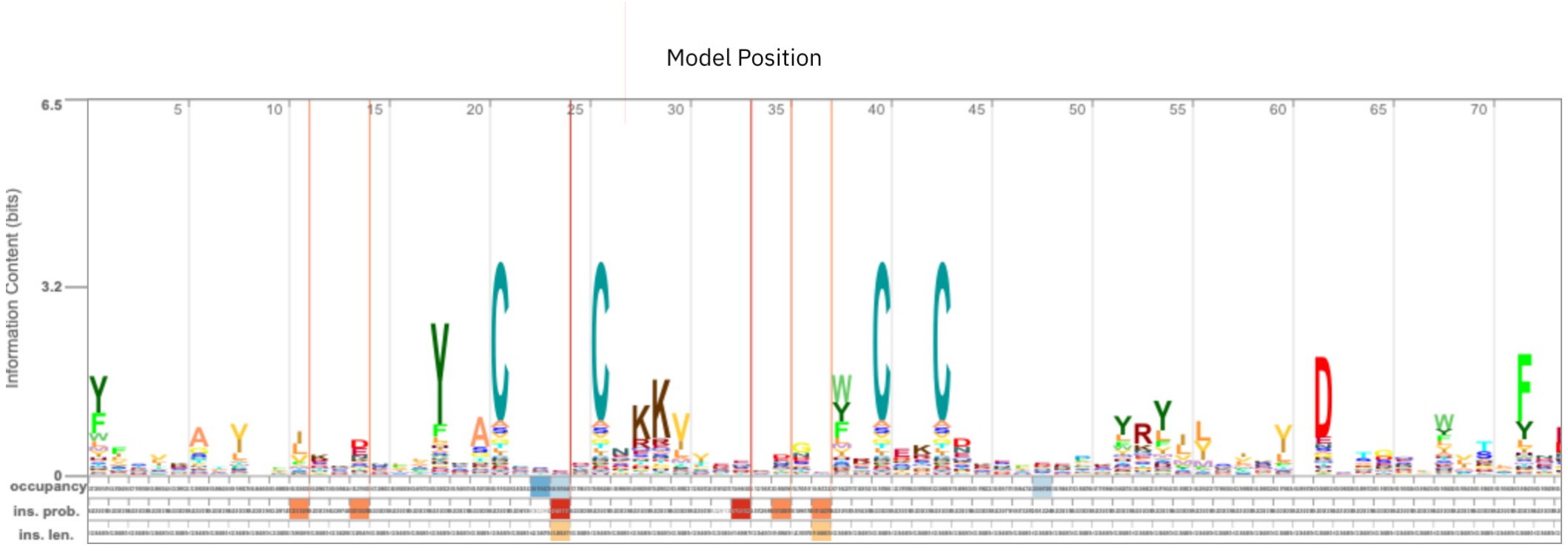
Intrinsically disordered regions (IDRs): polypeptide segments of the protein lacking 3D structure

Main functions: signalling and regulatory

Rakesh Trivedi: Intrinsically Disordered Proteins: An Overview
International Journal of Molecular Sciences (2022)

Goal of the thesis:

analysis of HMMs developed for detecting
intrinsically disordered regions (IDRs) of the proteins

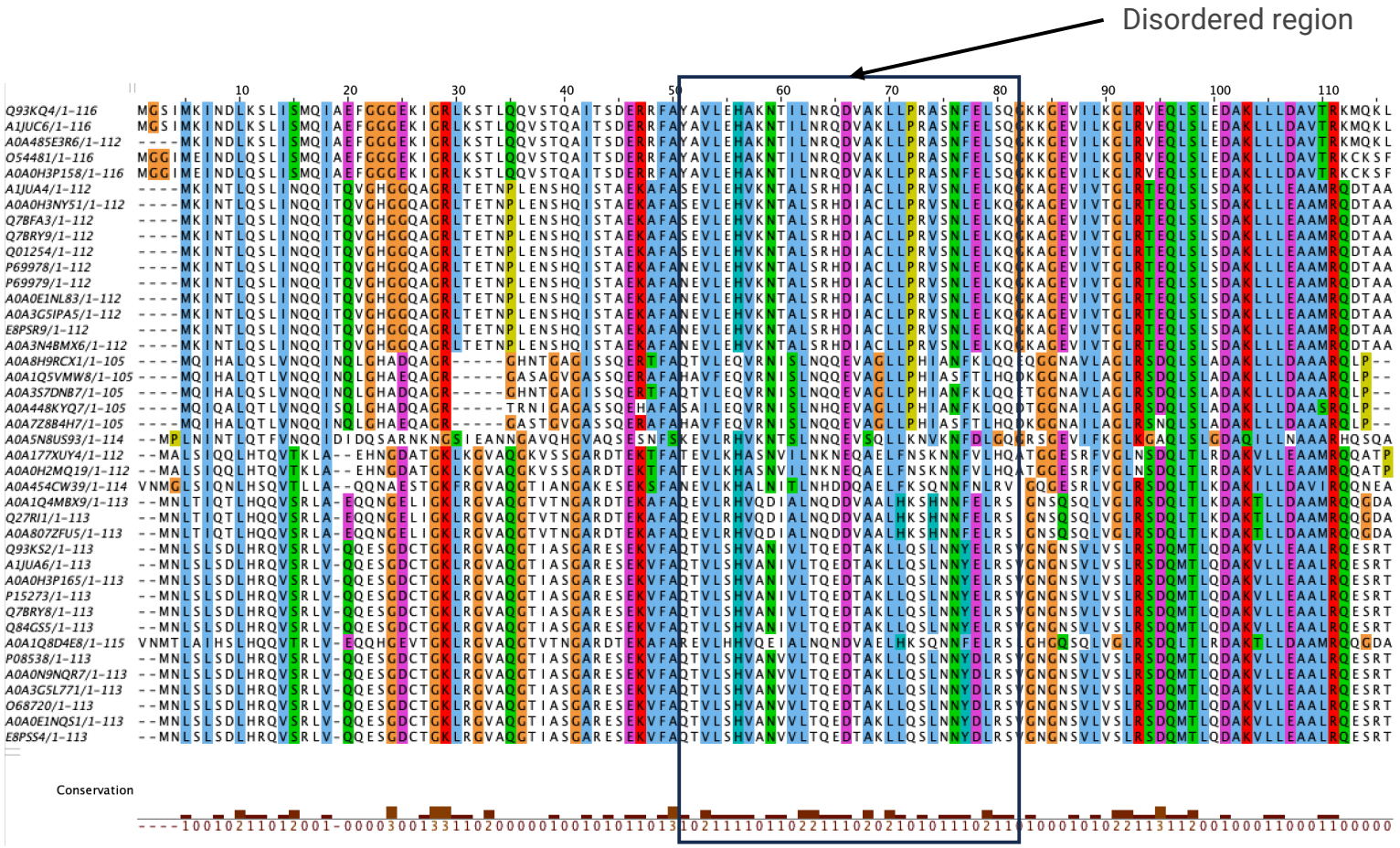


pHMM example: PF08646 - Replication factor-A C terminal domain

Multiple Sequence Alignment (MSA)

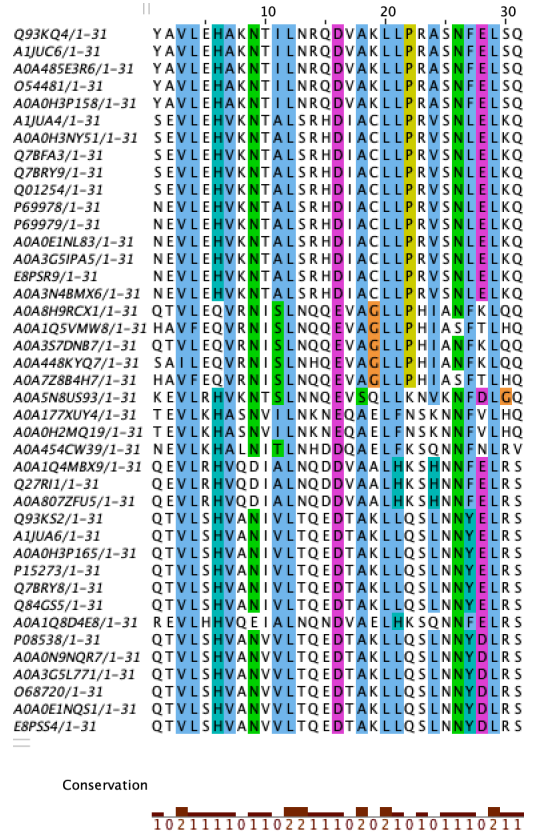
A technique used to arrange **3 or more** protein sequences, in a way that highlights both the similarities and differences among them.

- Helps in identifying **the least conserved** regions



Q93KQ4.fasta - Regulatory protein YscM2

Cut MSA



Q93KQ4_51-81.fasta

Thesis workflow

Databases		
DisProt	Manually curated repository of IDPs and IDRs (**2272** proteins, **3151** regions)	Extract information about IDRs positions
UniProt	Database of protein sequence and functional information	Obtain the seed alignments
Reference Proteome (75%)	Subset of proteomes that have been selected to provide a representative cross-section of the taxonomic diversity	Search for the proteins matching HMMs
Tools		
ClustalOmega	Program to build MSAs to generate alignments between three or more sequences	Build the MSAs
HMMER	Software package for sequence analysis with the various tools to build and analyse HMMs	Build and search HMMs

HMM build and search

- **1 step**: building the HMMs

Input: MSAs of IDRs

Output: HMMs

! Very fast – about **1-2 minutes** for the whole dataset processing

- **2 step**: searching the proteins based on HMMs in a proteome database (RP 75%)

Input: HMMs

Output: table of matched proteins

hmmbuild

```
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
           m->m   m->i   m->d   i->m   i->i   d->m   d->d
  COMPO   4.93836 6.60823 2.78359 1.56765 5.66943 2.59069 6.01310 6.05767 5.13250 5.33880 2.71543 2.63546 5.66889 1.98465 1.98371 2.65961 2.66563 5.72936 6.93832 2.70045
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.00000       *
      1   5.65795 6.94695 6.22410 5.71763 6.76760 5.53020 6.14445 6.68109 4.44313 5.90617 7.00123 5.93877 6.08733 5.45217 0.05374 5.78842 5.91615 6.34709 7.31141 6.63035     1 R
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      2   3.90441 5.75457 5.46612 5.50732 6.22794 4.48908 6.37059 6.06877 5.69397 5.67298 6.53901 5.28158 5.33075 5.87445 5.74200 0.09654 4.48874 5.20650 7.38863 6.39751     2 S
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      3   5.87405 6.77789 6.31463 6.35795 3.73767 5.79855 5.28069 5.84801 6.27452 4.93687 6.35582 5.98148 6.29431 6.18429 6.11330 5.83993 6.17920 5.77203 5.37792 0.07738     3 Y
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      4   5.46200 7.11353 4.52504 0.05997 6.87713 5.21603 6.21336 6.86719 5.44425 6.19136 7.34117 5.16433 5.90552 5.56833 5.83970 5.41278 5.83666 6.44135 7.57335 6.77741     4 E
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      5   5.46108 6.90219 5.34490 5.30534 6.45842 5.37626 6.33943 6.66695 5.13183 5.92175 7.10076 5.70386 6.03006 0.05680 5.28247 5.56877 5.85329 6.29856 7.33411 6.44392     5 Q
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      6   5.23040 6.55133 6.32041 6.24529 5.26901 5.51513 6.68381 4.73863 6.06899 3.84283 0.08323 6.33346 6.12406 6.37380 6.03437 5.60183 5.65175 4.88221 6.97965 5.93746     6 M
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      7   5.46200 7.11353 4.52504 0.05997 6.87713 5.21603 6.21336 6.86719 5.44425 6.19136 7.34117 5.16433 5.90552 5.56833 5.83970 5.41278 5.83666 6.44135 7.57335 6.77741     7 E
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      8   4.23715 6.00464 5.80030 5.81884 6.34938 4.75247 6.56553 5.90568 5.84033 5.59644 6.61614 5.60216 5.56413 6.11511 5.85303 4.46125 0.07963 5.29017 7.42638 6.58977     8 T
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
      9   4.47134 7.25522 0.22347 2.58294 6.54554 4.37002 5.13198 6.24331 4.45865 5.63542 6.64556 3.83040 5.11048 4.34399 5.27949 4.24969 4.82260 5.72546 7.68694 6.05904     9 D
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
     10   5.38193 6.76793 6.21331 6.29051 7.06385 0.03137 7.06456 7.19554 6.57116 6.51064 7.64063 6.39622 6.12961 6.80718 6.47528 5.62143 5.92219 6.52626 7.61280 7.26792    10 G
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
     11   5.46200 7.11353 4.52504 0.05997 6.87713 5.21603 6.21336 6.86719 5.44425 6.19136 7.34117 5.16433 5.90552 5.56833 5.83970 5.41278 5.83666 6.44135 7.57335 6.77741    11 E
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
     12   5.65795 6.94695 6.22410 5.71763 6.76760 5.53020 6.14445 6.68109 4.44313 5.90617 7.00123 5.93877 6.08733 5.45217 0.05374 5.78842 5.91615 6.34709 7.31141 6.63035    12 R
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
     13   5.46108 6.90219 5.34490 5.30534 6.45842 5.37626 6.33943 6.66695 5.13183 5.92175 7.10076 5.70386 6.03006 0.05680 5.28247 5.56877 5.85329 6.29856 7.33411 6.44392    13 Q
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00618 5.48519 6.20753 0.61958 0.77255 0.48576 0.95510
     14   5.07225 6.70828 4.98050 5.04823 6.45418 5.15537 6.37870 6.73284 5.62427 6.12144 7.21838 0.06170 5.88206 5.85265 5.85632 5.18596 5.55846 6.15910 7.39840 6.44416    14 N
          2.68618 4.42225 2.77519 2.73123 3.46354 2.40513 3.72494 3.29354 2.67741 2.69355 4.24690 2.90347 2.73739 3.18146 2.89801 2.37887 2.77519 2.98518 4.58477 3.61503
          0.00416 5.48317       * 0.61958 0.77255 0.00000       *
```

hmmsearch

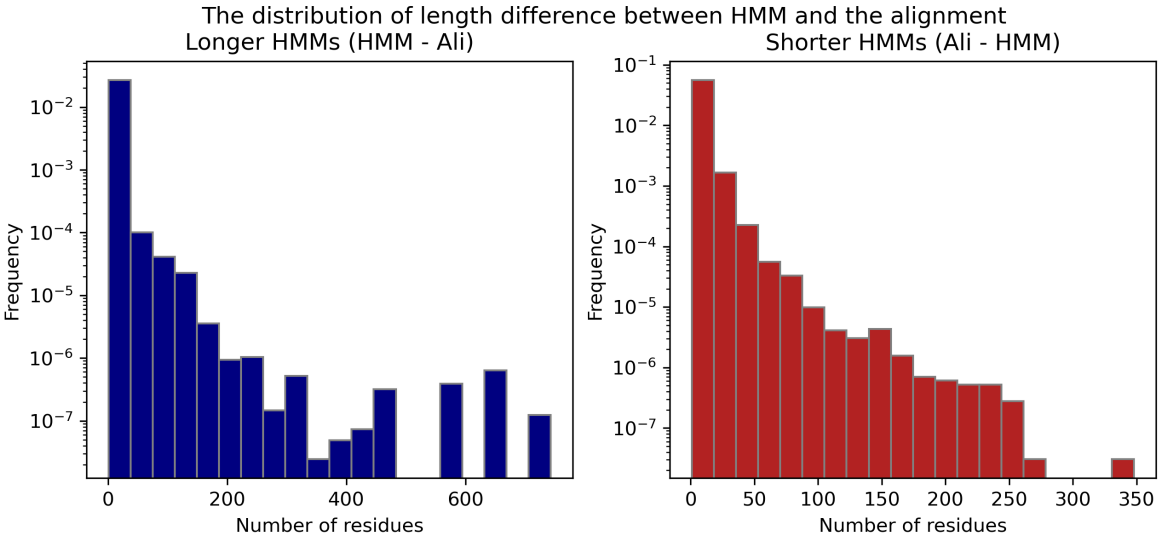
```
Query:       A8AZZ3_146-195  [M=46]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence    Description
    -------  ------ -----    ------- ------ -----    ---- --  --------    -----------
    5.4e-07   42.4  10.3         23   17.9   0.0    5.3  5  R0LE47      R0LE47_STRMT^|^^|^Ornithine carbamoyltransferase
    2.1e-06   40.5  20.6     2.1e-06   40.5  20.6    3.2  3  A8AZZ3      A8AZZ3_STRGC^|^^|^Amylase-binding protein AbpA {E
    5.7e-05   35.9   1.9     2.4e+04    8.3   0.2    8.3  8  A0A371S028  A0A371S028_9BACI^|^^|^Uncharacterized protein {EC
    8.6e-05   35.3   0.1     1.5e+04    8.9   0.0    5.3  4  A0A1S8TMK1  A0A1S8TMK1_9CLOT^|^^|^Autolysin {ECO:0000313|EMBL
     0.0011   31.7  14.9      0.0011   31.7  14.9    1.8  1  T0TW47      T0TW47_9STRE^|^^|^Uncharacterized protein {ECO:00
     0.0022   30.8   2.1     6.6e+04    6.9   0.1    6.5  7  U5MZS1      U5MZS1_CLOSA^|^^|^Surface protein PspC {ECO:00003
     0.0064   29.3   6.6      0.0064   29.3   6.6    3.4  3  E7S8K8      E7S8K8_9STRE^|^^|^Uncharacterized protein {ECO:00
     0.0068   29.2  25.2      0.0068   29.2  25.2    4.6  4  A0A0F2CVN8  A0A0F2CVN8_STRCR^|^^|^Uncharacterized protein {EC
      0.007   29.2  29.4        0.12   25.3   0.0    5.6  4  E1LNB5      E1LNB5_STRMT^|^^|^Cell wall binding repeat family
 ------ inclusion threshold ------

 >> A8AZZ3  A8AZZ3_STRGC^|^^|^Amylase-binding protein AbpA {ECO:0000313|EMBL:ABV10035.1}^|^^|^^|^Streptococcus gordonii s
    #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
  ---  ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
    1 ?   -1.3   0.2       5.7   2.3e+07      18      33 ..      26      40 ..      16      43 .. 0.51
    2 ?    0.3   0.9       1.9   7.5e+06      22      35 ..      64      77 ..      47      80 .. 0.80
    3 !   40.5  20.6   5.3e-13   2.1e-06       2      46 .]     147     195 .]     146     195 .] 0.93

   Alignments for each domain:
   == domain 1  score: -1.3 bits;  conditional E-value: 5.7
   A8AZZ3_146-195 18 eakeeaaaknakagaK 33
                     ea ++a   n +a+++
             A8AZZ3 26 EATDAA-RNNDGAYYL 40
                     222333.556666665 PP

   == domain 2  score: 0.3 bits;  conditional E-value: 1.9
   A8AZZ3_146-195 22 eaaaknakagaKag 35
                     eaaa+ a+++aKa+
             A8AZZ3 64 EAAADPAVVAAKAA 77
                     55588888888887 PP

   == domain 3  score: 40.5 bits;  conditional E-value: 5.3e-13
   A8AZZ3_146-195   2 gaqtgkaakasa..atkpeakeeaa..aknakagaKaggKALPKTsAVK 46
                      ga+++++++a++  at+p++keea+  +++ak++ Ka++KALPKTsAVK
             A8AZZ3 147 GAANNQNPAAEDkgATTPASKEEAKksEAAAKNAGKAAGKALPKTSAVK 195
                      78888888888888**********************************9 PP
```



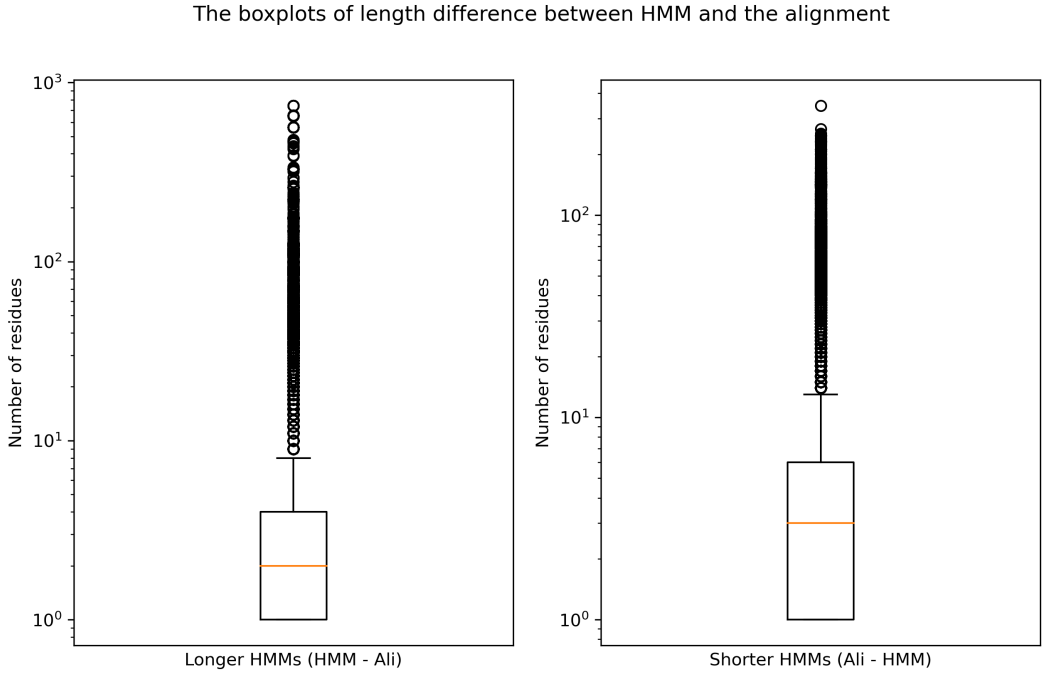
Length comparison: HMM vs alignment

HMM length:

the length of the **query** sequence

Alignment length:

the length of the **target** sequence



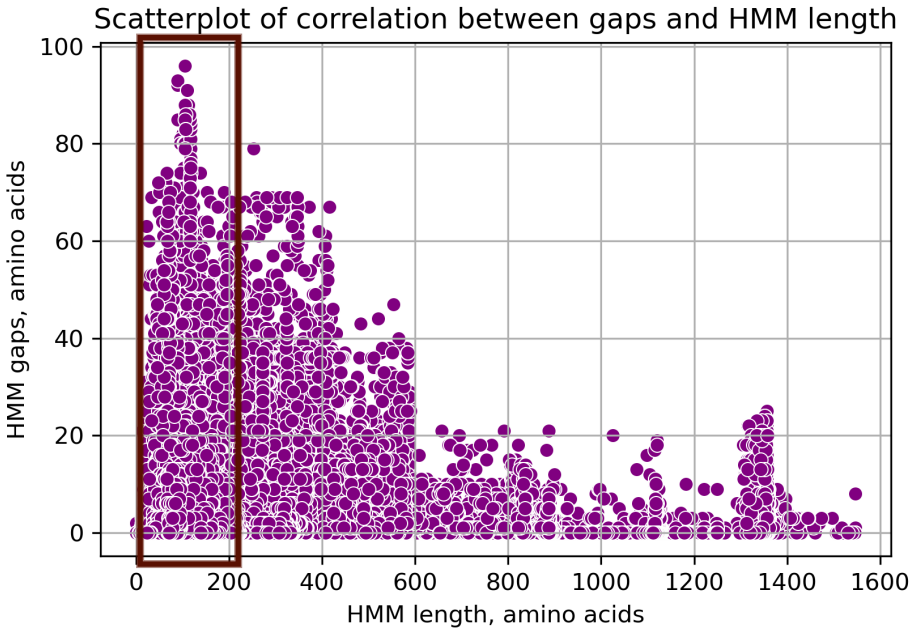
The number of pairwise alignments: **5,878,578**, from which HMMs:

- Match the target sequences length: 2,906,648 (**49.44%**)
- Longer HMMs: 1,097,774 (**36.94%**)
- Shorter HMMs: 1,874,156 (**63.06%**)

- **13.0%** – longer HMMs outliers (8 residues)
- **7.6%** – shorter HMMs outliers (13 residues)

The low proportion of outliers suggests that we can use **ali from/to** for further comparison with Pfam domains

Gaps analysis

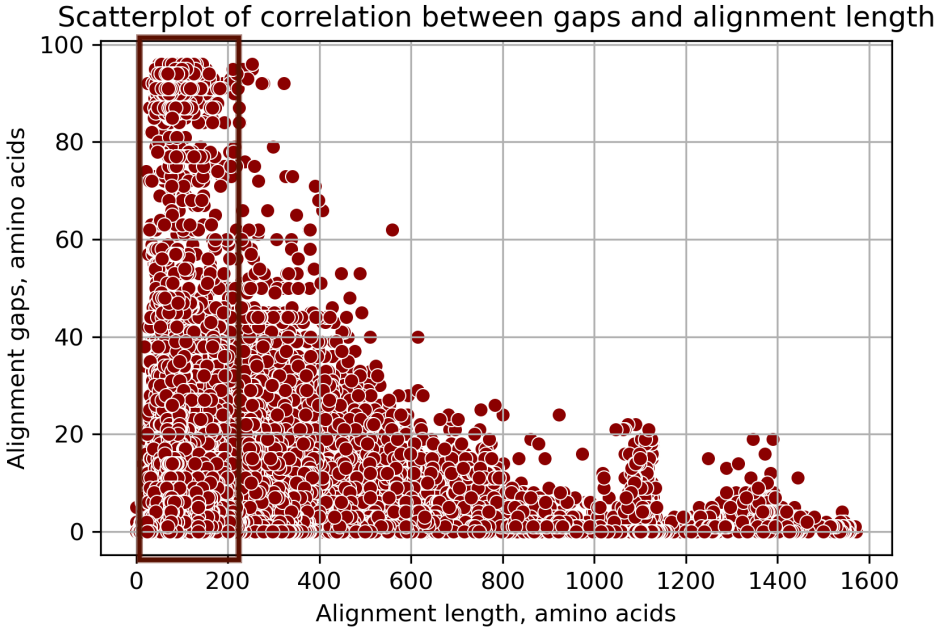



```
A8AZZ3_146-195    1 hgaqtgkaakasa...atkpeakeeaa....aknakagaKagqKALP 40
                    +g++++++ k+++   ++ ++a++++a    ++n++++++ +++ ++
        R0LE47 215 QGQSAKGWQKDAKgqwSYLKDAQGTKAtgwlKDNGTWYYLNAEGVMQ 261
```

Example of the **insertions** for A8AZZ3 protein: HMM

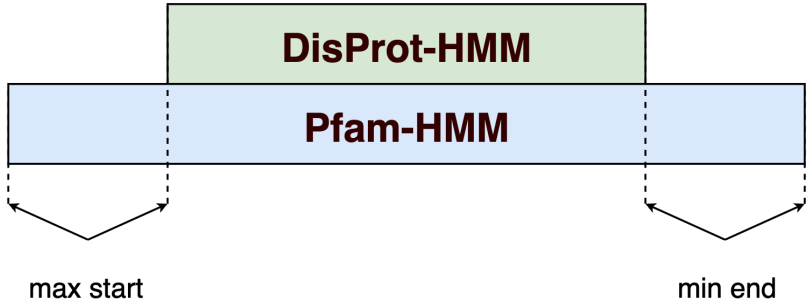
```
   A8AZZ3_146-195    5 tgkaakasaatkpeakeeaaaknakagaKagqKALPKTsAVK 46
                       t++a++a+ + k e++++a a n+ka +++++KALPKTsAVK
          T0TW47 169 TNEATNAA-K-KTEEGAKA-AQNGKASSAQAGKALPKTSAVK 207
```

Example of the **deletions** for A8AZZ3 protein: target sequence

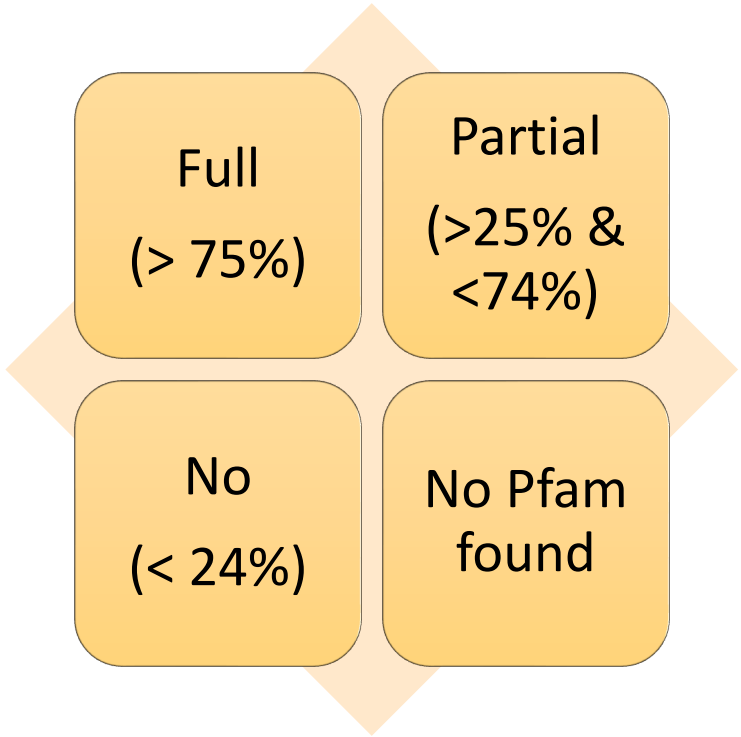


The distribution of gaps across the sequences is similar for both HMM and target sequences. **Shorter sequences** tend to have **more gaps** relative to their length.

Overlap types



Overlap length: end – start
Total length: pfam length + disprot length – overlap
Overlap percentage: overlap/total length (%)



DisProt-HMM distribution:

- Full overlap: 371;
- Partial overlap: 891;
- No overlap: **2196**;

- Full and partial overlaps: 24;
- Full and no overlaps: 341;
- Partial and no overlaps: 766;
- All types of overlaps: 14.

Orphan HMMs

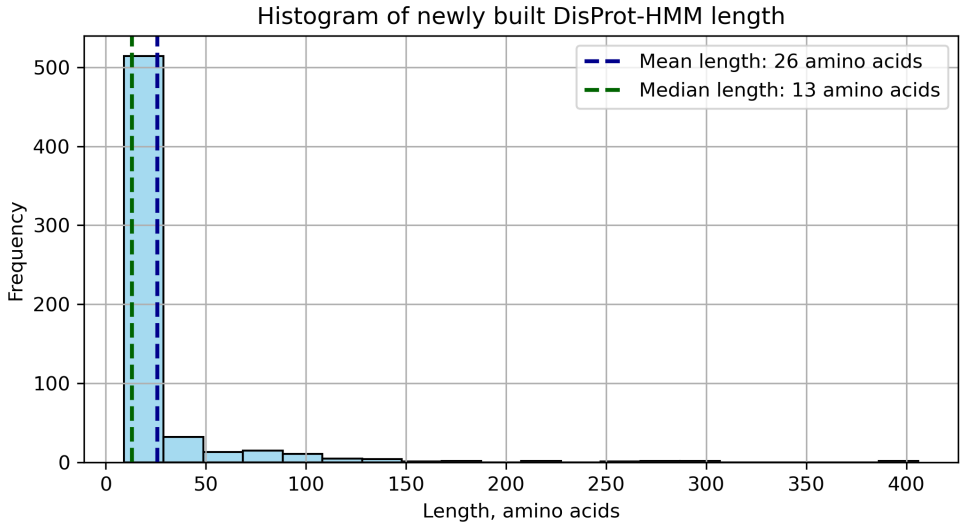
The new HMMs:

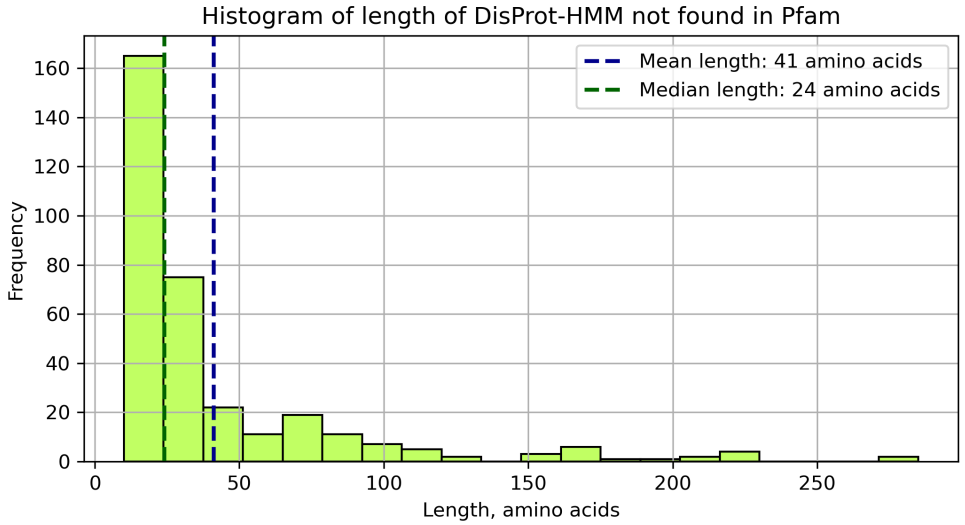
1) generated by hmmbuild but not found in RP 75%;
2) found in RP 75% but not in Pfam.

30.0%
new HMMs
(942 of 3142 generated)

19.3%
without hits in RP 75%
(606 of 3142)

10.7%
without hits in Pfam
(336 of 3142)





Summary statistics

Dataset	# of proteins	# of regions
Curated DisProt	2272	3151
UniProt BLAST	2271	3150
ClustalOmega	2268	3147
hmmbuild	2263	3142
hmmsearch	1971	2536
Pfam *	**1734 (76.32%)**	**2200 (69.82%)**

-1 instance: not found as a query sequence by BLAST

-3 instances: no subject sequences found

-5 instances: failed at cluster splitting in ClustalOmega

-606 instances: not found in RP 75%

-236 instances: not found in Pfam

* - the numbers are dependent on the specified thresholds and may vary accordingly

Results

- **70%** of DisProt-HMMs overlapped Pfam-HMMs, while **30%** of DisProt-HMMs are new;
- Among overlapping instances the larger amount of DisProt-HMMs corresponds to “no overlap” type, comprising **2196** models;
- Most gaps are found in **shorter** sequences, indicating that the HMM considers the IDRs to be shorter than those stated in the Curated DisProt;
- HMM without hits in RP 75% and Pfam are usually short in length.

Possible developments

- Different alignment format: .sto instead of .fasta;
- Other types of MSA: MAFFT, T-Coffee, etc;
- Regions selection: envelope regions instead of alignment;
- Take into account conditional and independent E-values.

Thank you for your attention!