MMG1001 Genomics
Week 1 Tutorial

Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments

TA: Heather Gibling

How this course works

Monday
Lecture and paper #1 discussion with Tim

Wednesday
Paper #2 discussion with TA (~40 min)
Short break (~5-10min)
Assignment overview & work through (~60-70min)

Friday
Tim presents paper #2 and asks students to answer questions
TAs review assignments

Purpose of the Wednesday session

- Experience reading and dissecting papers
- Practical hands-on experience with some common bioinformaticsy activities
- Get participation marks
- Get to know your cohort

Things to think about when reading a paper

What are the main **goals** of the paper?

- *See abstract, last paragraph of introduction*

How did they do it?

- *See methods*

What are the major **findings**?

- *See abstract, results, discussion*

Why does it **matter**?

- *See discussion*

Things to remember

- Please be respectful
- Please participate!  
  - Combination of verbal questions and polls
  - **I will state if a poll is anonymous--most will not be so I can track participation (no one else can see your name)**
- It’s ok not to know something (I don’t know a lot!)
- I won’t require you to turn your video on
  - ...but if you don’t mind, it makes it less awkward!
  - Cats/dogs/pets welcome on camera
- Questions/concerns about this course/grad school? Ask me!

Tell us about yourself!

- Who are you?
- Where and what did you study before?
- What lab did you join/are you currently rotating in?
- What's one *boring* fact about yourself?

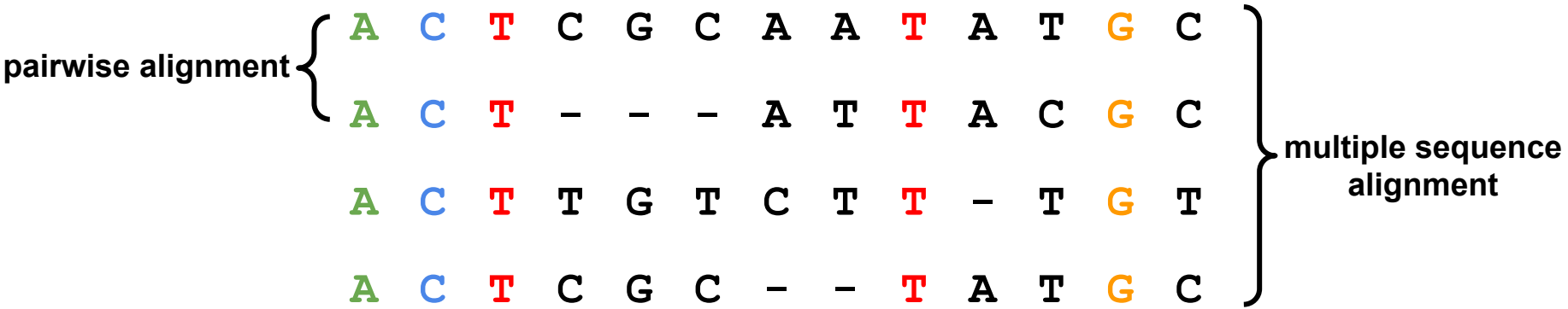
PROTEINS: Structure, Function, and Genetics 28:405–420 (1997)

Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments

Erik L.L. Sonnhammer,[1] Sean R. Eddy,[2] and Richard Durbin[1]*
[1]*Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*
[2]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri*

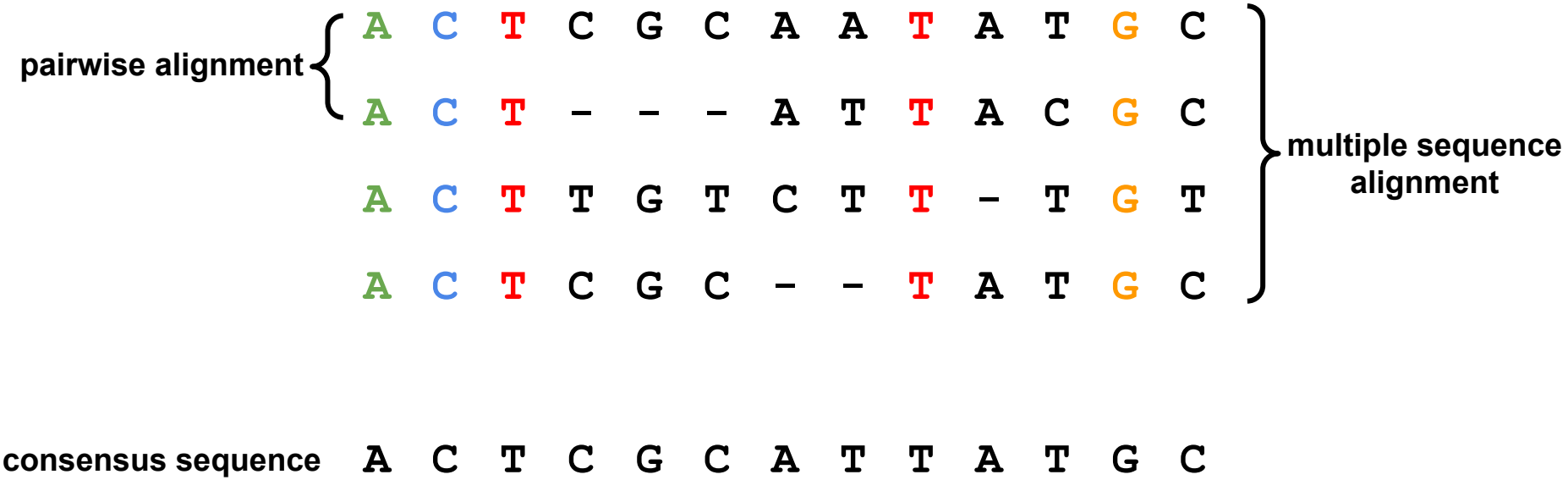
What is an alignment?

- Arrangement of sequences to identify similarities
- Can be nucleotides or amino acids



What is an alignment?

- Arrangement of sequences to identify similarities
- Can be nucleotides or amino acids



What are alignments used for?

TCAATC**GCATCGCGACGACATCAGC**TACGGTCC
GTCTCA**GCATCGCGACGACATCAGC**ACGTAGCG

Local alignment

-TCAATC-GCATCGCGACGACATCAGCTACGGT-CC-
GTC--TCAGCATCGCGACGACATCAGC-ACG-TAGCG

Global alignment

...TCAATC**GCATCGCGACGACATCAGC**TACGGTCC...
GCATCGCGACGACATCAGC

‘Glocal’ alignment

What are alignments used for?

TCAATC**GCATCGCGACGACATCAGC**TACGGTCC
GTCTCA**GCATCGCGACGACATCAGC**ACGTAGCG

Local alignment
find conserved segments

-TCAATC-GCATCGCGACGACATCAGCTACGGT-CC-
GTC--TCAGCATCGCGACGACATCAGC-ACG-TAGCG

Global alignment
compare similar sequences

...TCAATC**GCATCGCGACGACATCAGC**TACGGTCC...
GCATCGCGACGACATCAGC

'Glocal' alignment
align to larger genome

What is a profile Hidden Markov Model (HMM)?

What is a profile Hidden Markov Model (HMM)?

- **Markov property**: future condition depends on the current condition, but no past conditions

What is a profile Hidden Markov Model (HMM)?

- **Markov property**: future condition depends on the current condition, but no past conditions
- **Markov model**: probabilistic model with a Markovian assumption

What is a profile Hidden Markov Model (HMM)?

- **Markov property**: future condition depends on the current condition, but no past conditions
- **Markov model**: probabilistic model with a Markovian assumption
- **Hidden Markov model**: Markov model with unobservable/hidden states (can learn about states by observing the outcome)

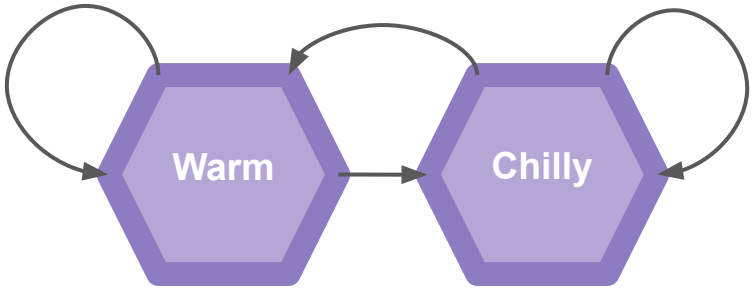
What is a profile Hidden Markov Model (HMM)?

- **Markov property**: future condition depends on the current condition, but no past conditions
- **Markov model**: probabilistic model with a Markovian assumption
- **Hidden Markov model**: Markov model with unobservable/hidden states (can learn about states by observing the outcome)
- **Profile HMM**: HMM that describes sequence conservation
    - (basically, a special type of HMM used by biologists)

Anatomy of a HMM

Example: Predict the temperature outside by looking at pedestrian clothing while stuck inside during quarantine

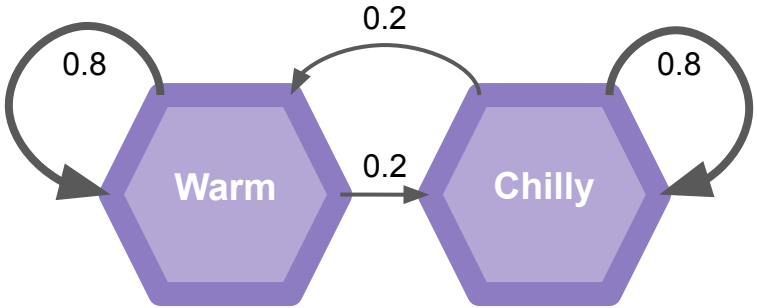
- **Hidden states**
  (warm, chilly)



Anatomy of a HMM

Example: Predict the temperature outside by looking at pedestrian clothing while stuck inside during quarantine

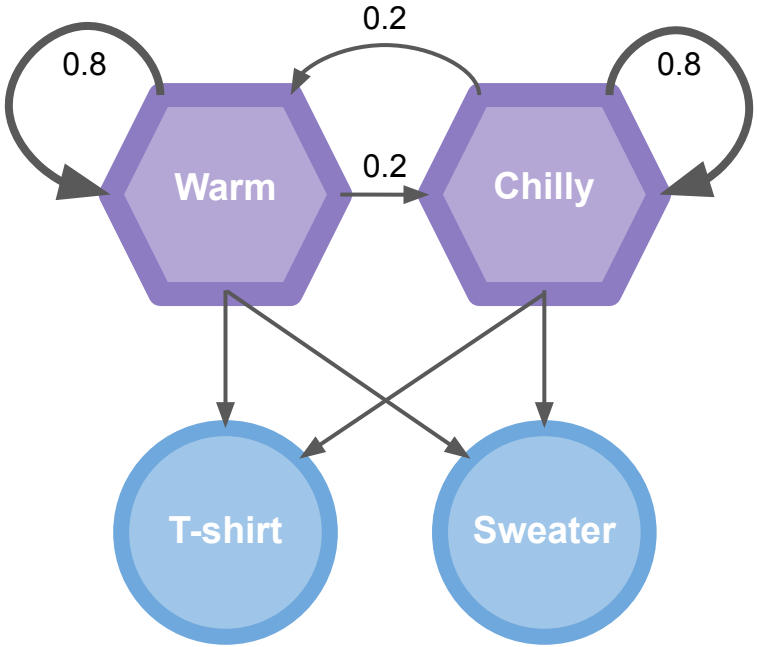
- **Hidden states**
  (warm, chilly)
- **Transition probabilities** between states
  (likely to have many warm days in a row)



Anatomy of a HMM

Example: Predict the temperature outside by looking at pedestrian clothing while stuck inside during quarantine

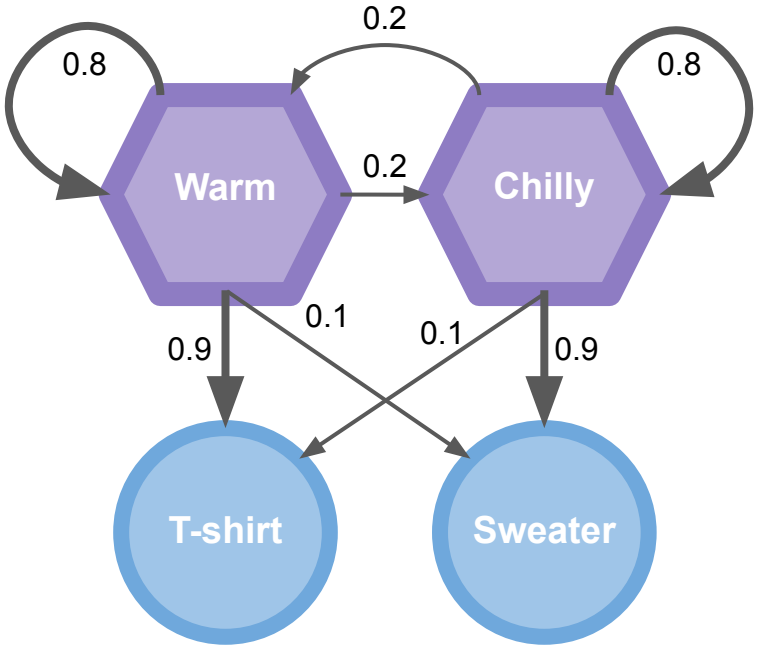
- **Hidden states**
  (warm, chilly)
- **Transition probabilities** between states
  (likely to have many warm days in a row)
- **Emissions**/observable outcomes
  (t-shirt, sweater)



Anatomy of a HMM

Example: Predict the temperature outside by looking at pedestrian clothing while stuck inside during quarantine

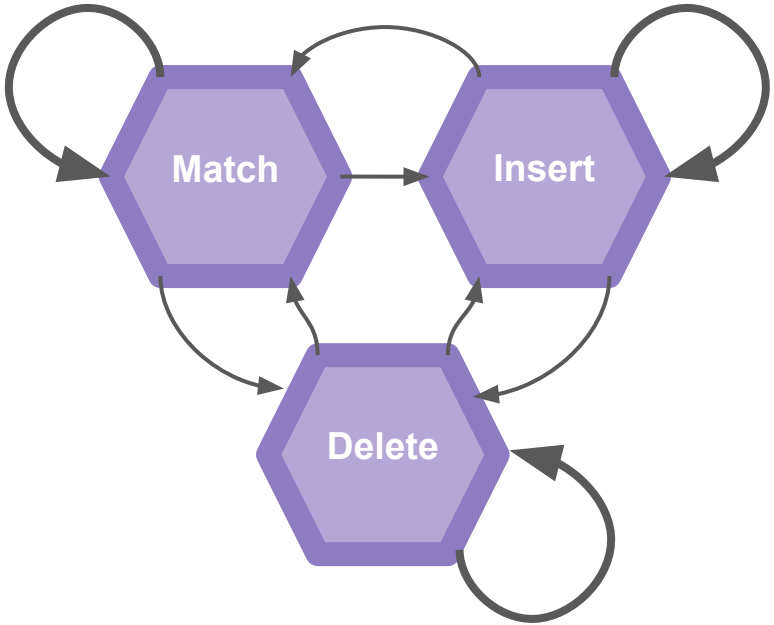
- **Hidden states** (warm, chilly)
- **Transition probabilities** between states (likely to have many warm days in a row)
- **Emissions**/observable outcomes (t-shirt, sweater)
- **Emission probabilities** based on states (t-shirt more likely during warm day than chilly day)



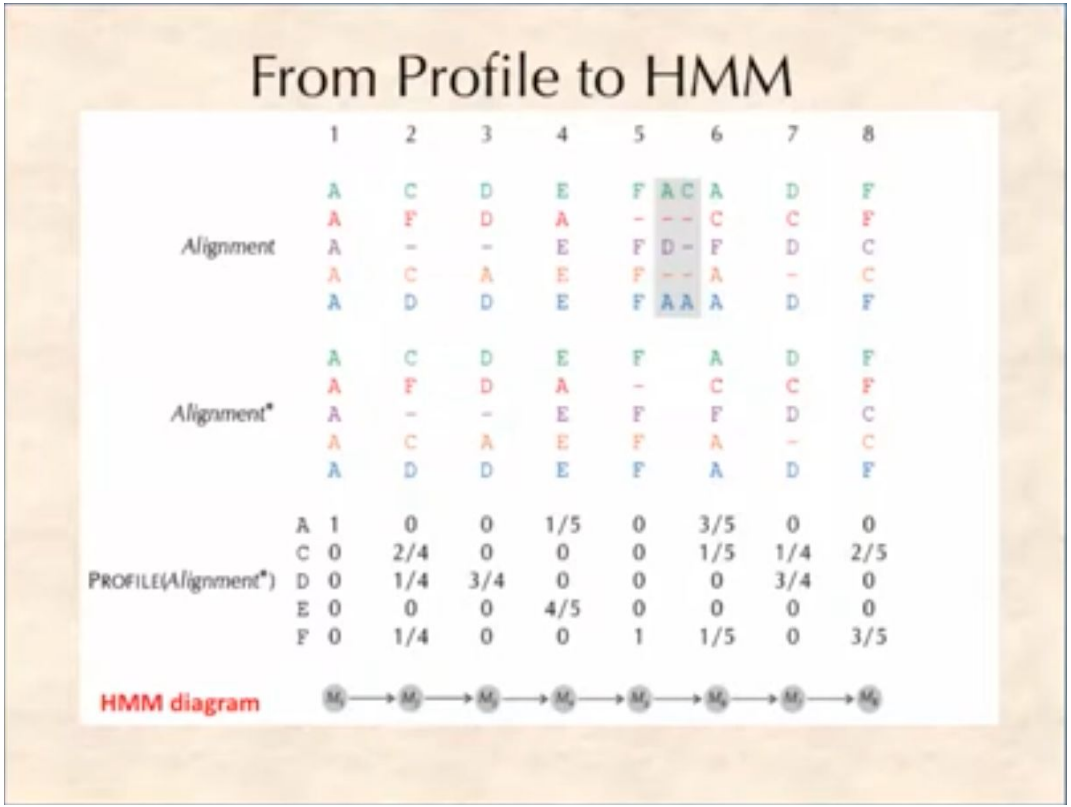
Anatomy of a HMM

Example: Aligning two DNA sequences

- **Hidden states**
  (match, insertion, deletion)
- **Transition probabilities** between states
  (based on alignment penalties)
- **Emissions**/observable outcomes
  ([x,y], [x,-], [-,y])
- **Emission probabilities** based on states
  (fixed)



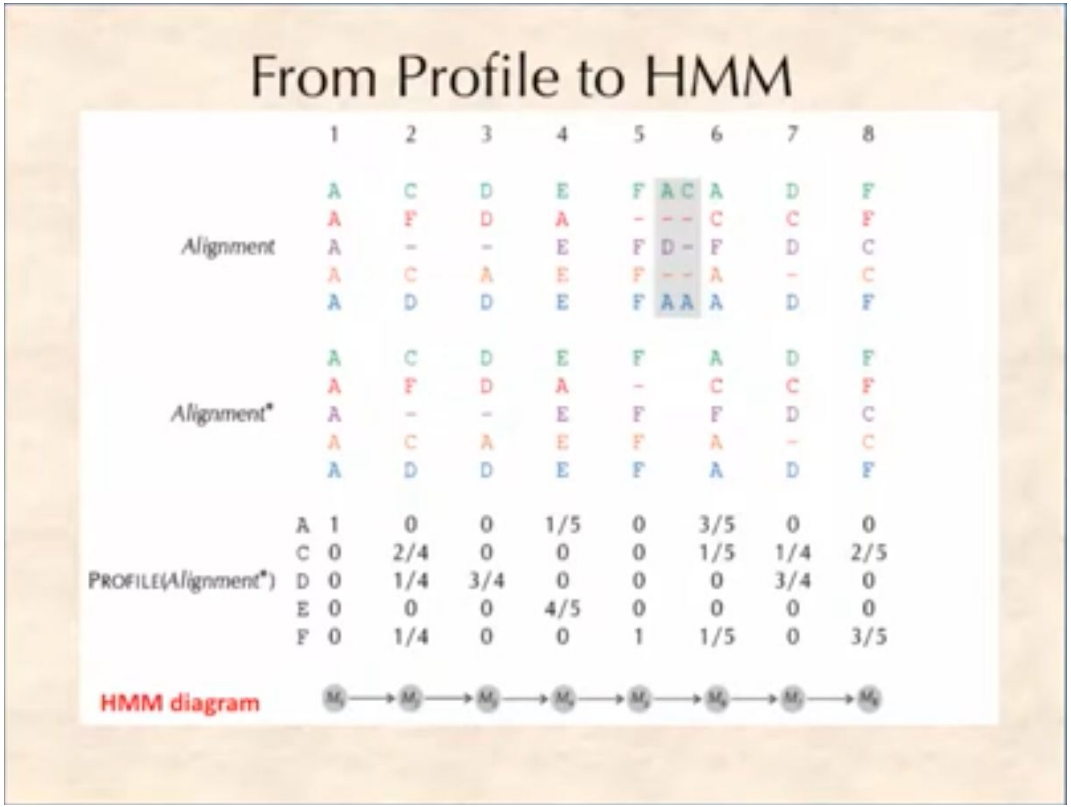
Anatomy of a profile HMM



- Start with **multiple sequence alignment**

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

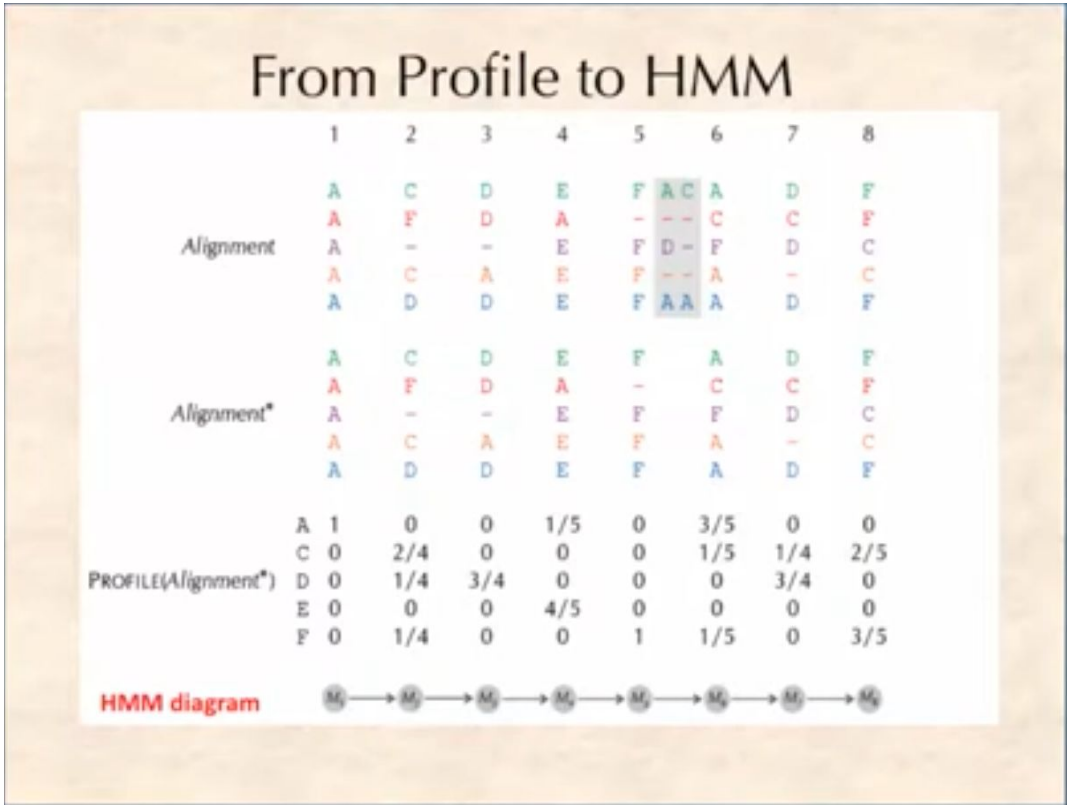
Anatomy of a profile HMM



- Start with **multiple sequence alignment**
- Remove columns where the number of sequences with insertions are above a threshold

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

Anatomy of a profile HMM

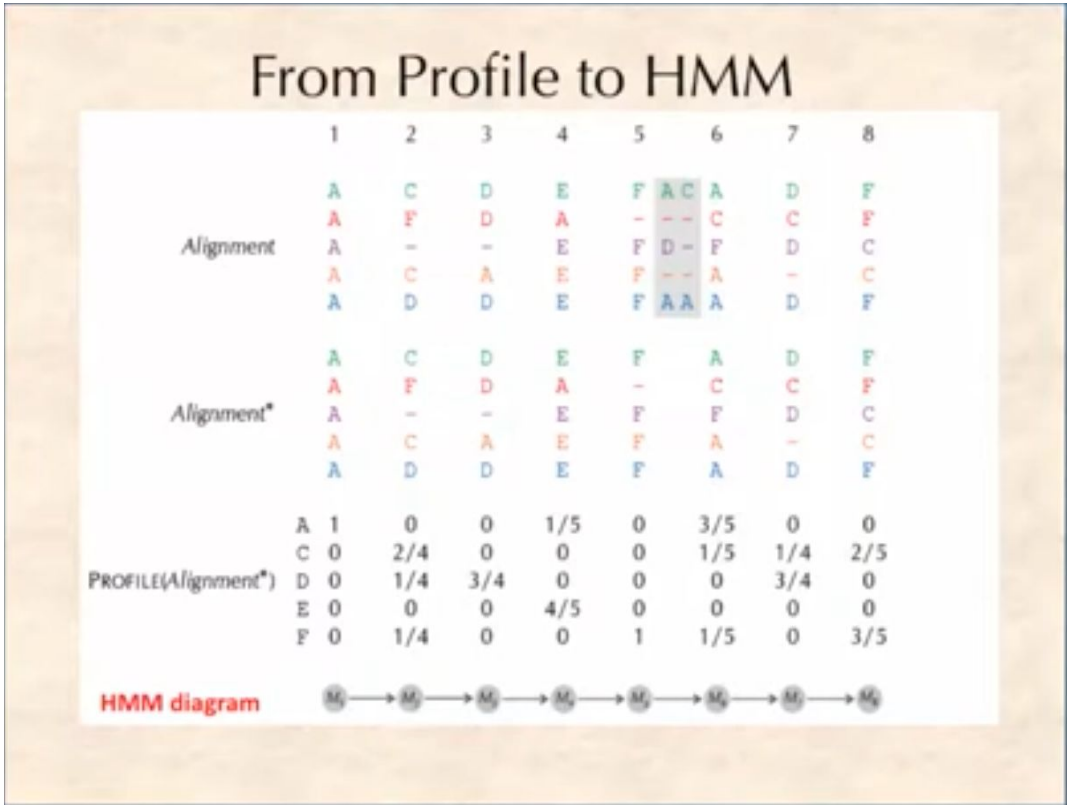


- Start with **multiple sequence alignment**
- Remove columns where the number of sequences with insertions are above a threshold
- Create a **match state*** for each remaining position

**match is match or mismatch*

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

Anatomy of a profile HMM

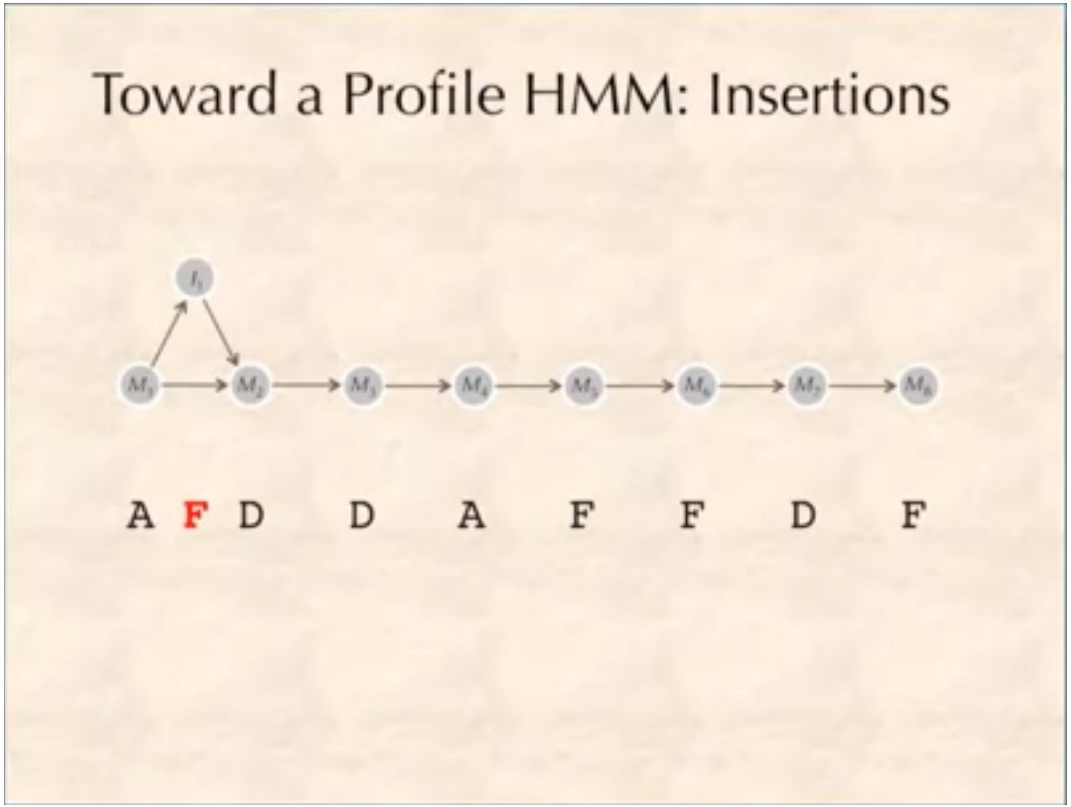


- Start with **multiple sequence alignment**
- Remove columns where the number of sequences with insertions are above a threshold
- Create a **match state*** for each remaining position
- Determine frequency of amino acids at each position
  - This determines the possible **emissions** (and the **emission probabilities**) at each match state

**match is match or mismatch*

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

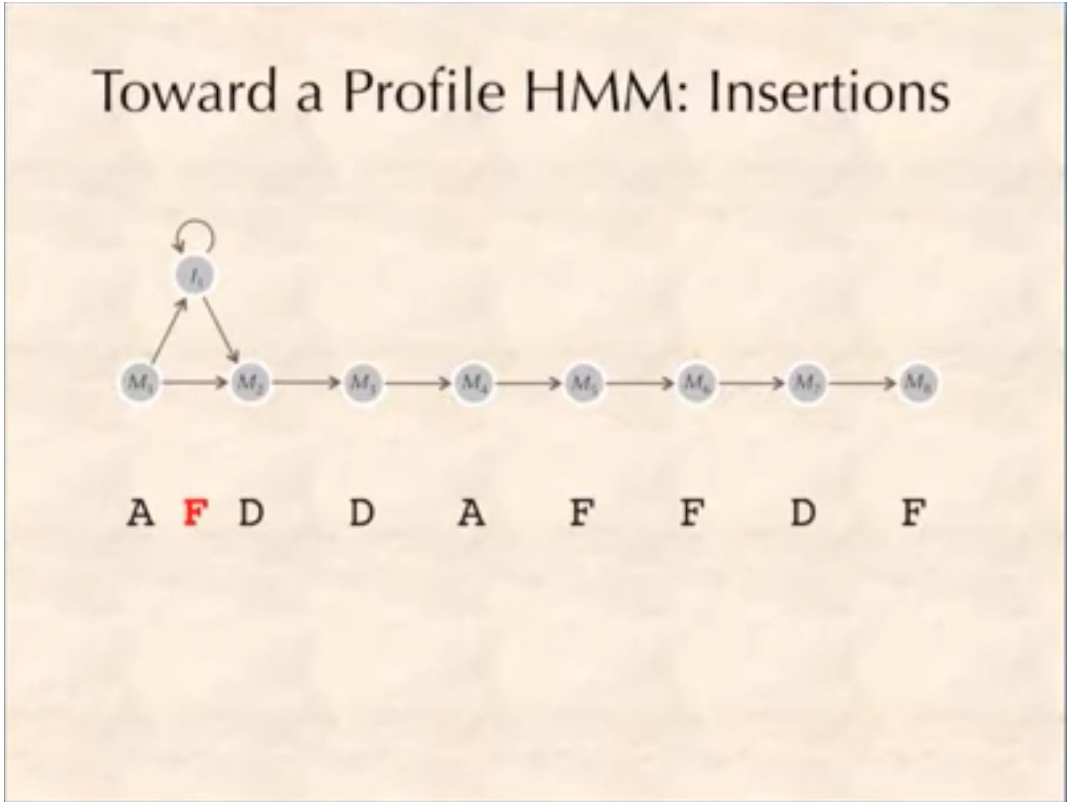
Anatomy of a profile HMM



- Add **insertion state** between match states

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

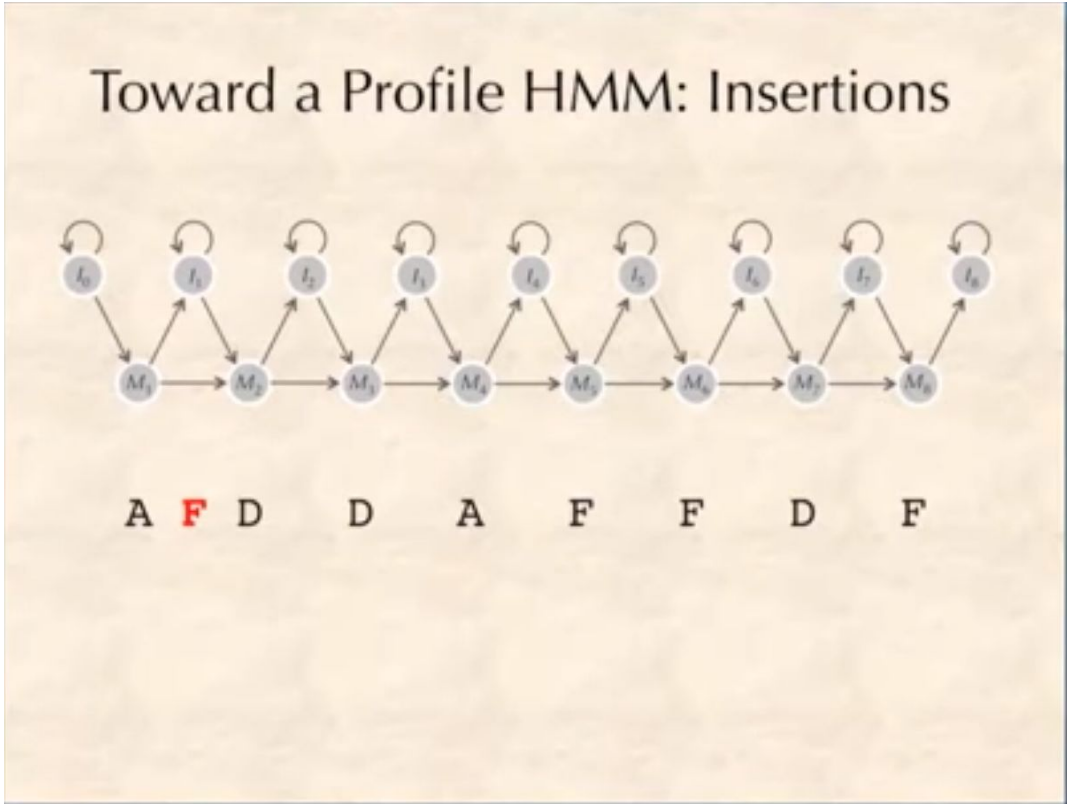
Anatomy of a profile HMM



- Add **insertion state** between match states
- Allow for multiple insertions in a row

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

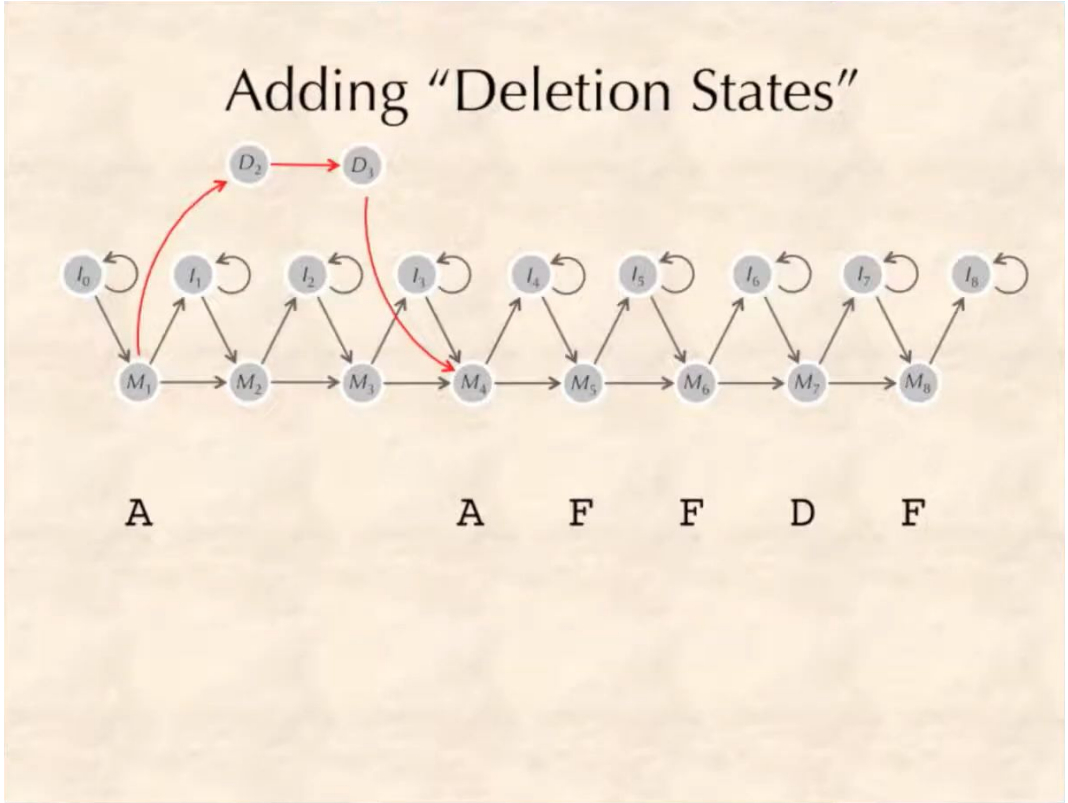
Anatomy of a profile HMM



- Add **insertion state** between match states
- Allow for multiple insertions in a row
- Add these insertion states between each match state (and before the first and after the last)
    - Adding them everywhere (instead of only at locations observed in the multiple sequence alignment) allows you to align new sequences to the HMM

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

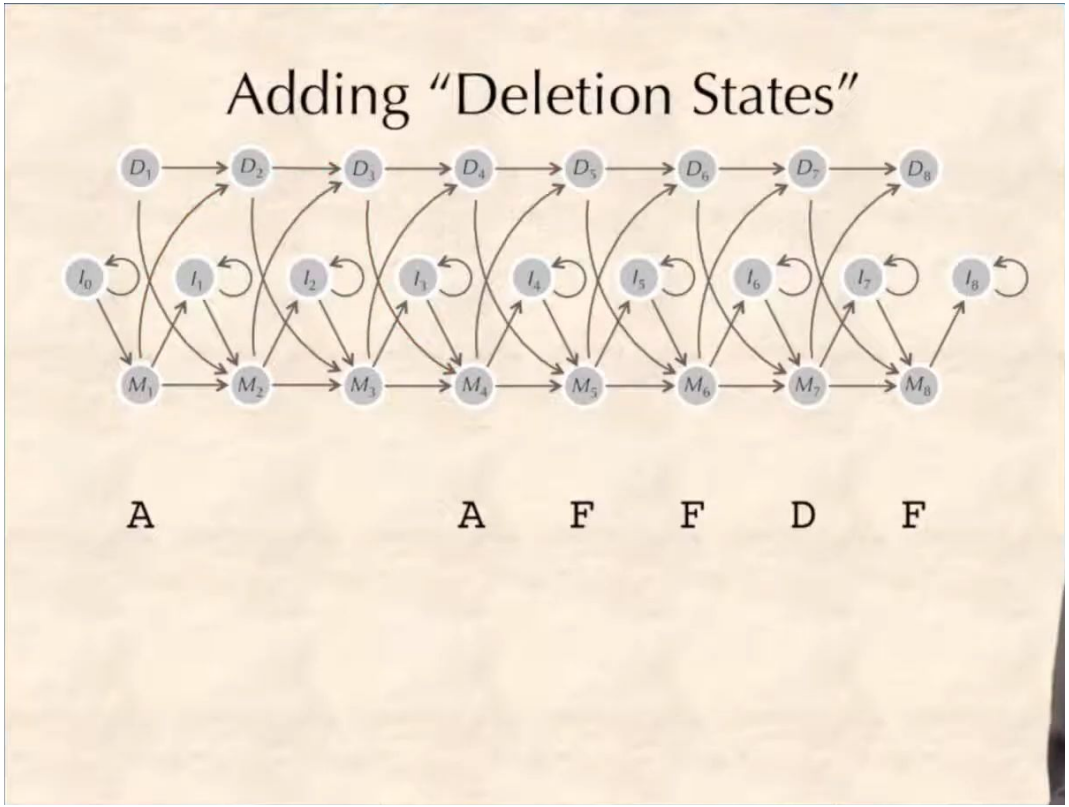
Anatomy of a profile HMM



- Add **deletion states** between match states

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

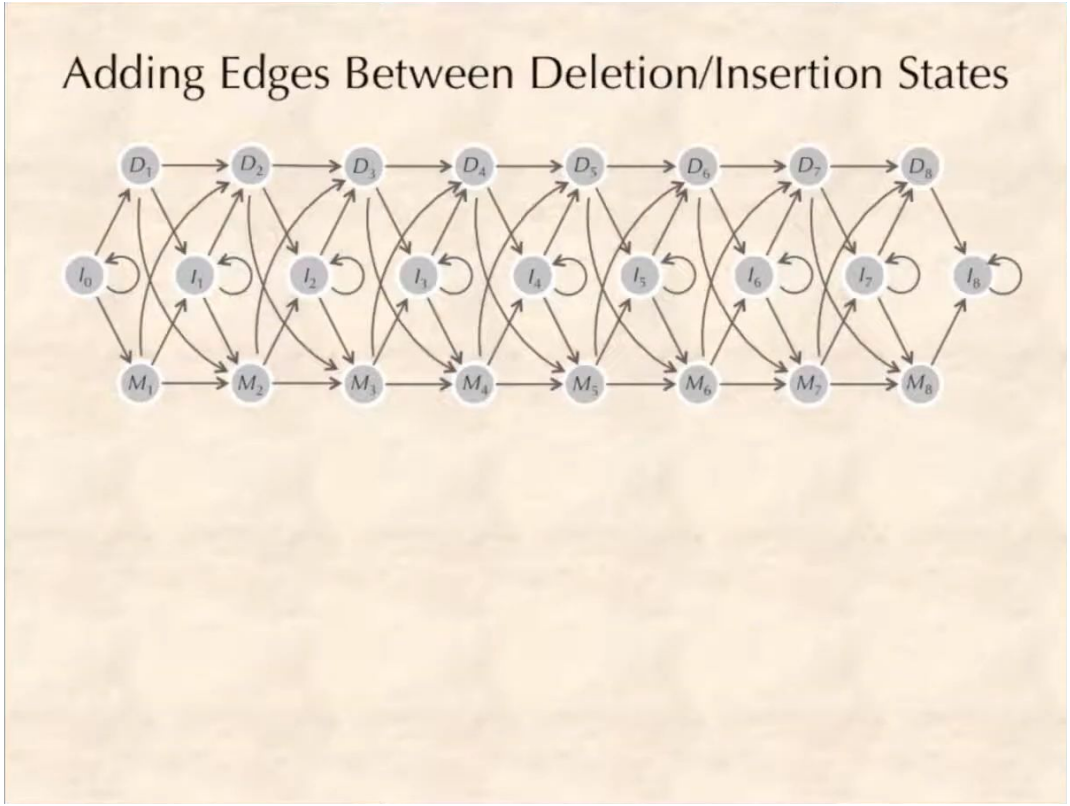
Anatomy of a profile HMM



- Add **deletion states** between match states… all of them
  - Adding them everywhere (instead of only at locations observed in the multiple sequence alignment) allows you to align new sequences to the HMM

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

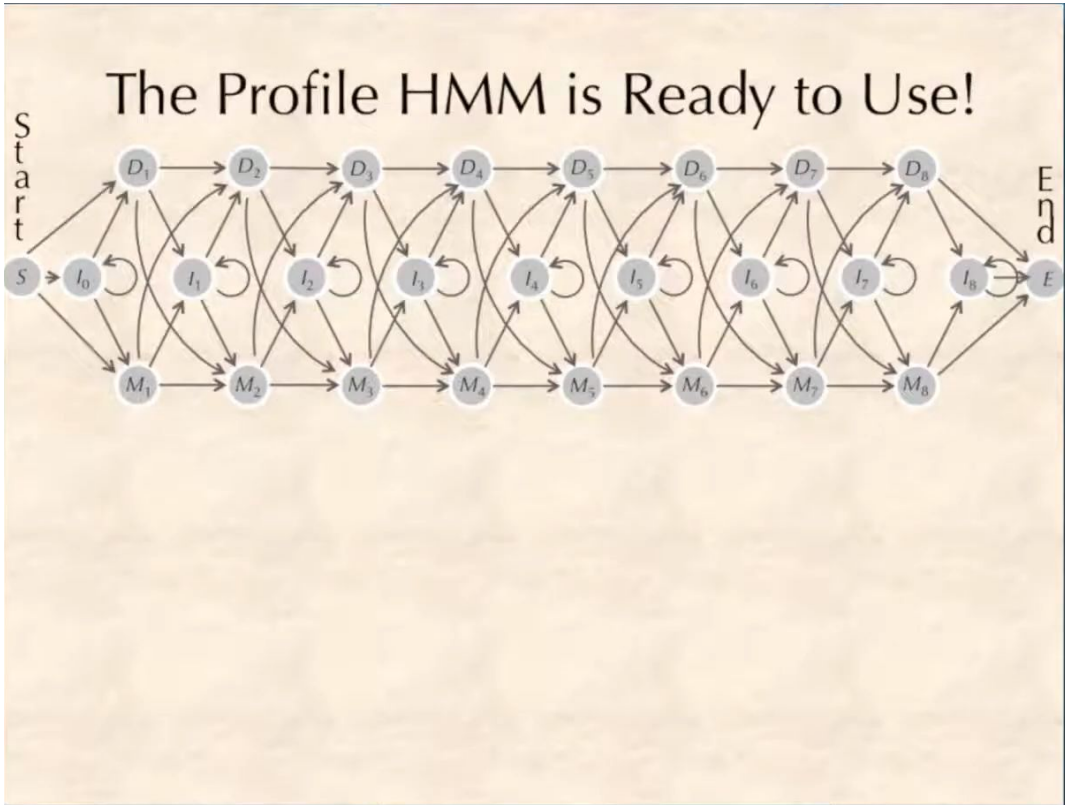
Anatomy of a profile HMM



- Add **deletion states** between match states… all of them
  - Adding them everywhere (instead of only at locations observed in the multiple sequence alignment) allows you to align new sequences to the HMM
- Allow for transition between deletion and insertion states

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

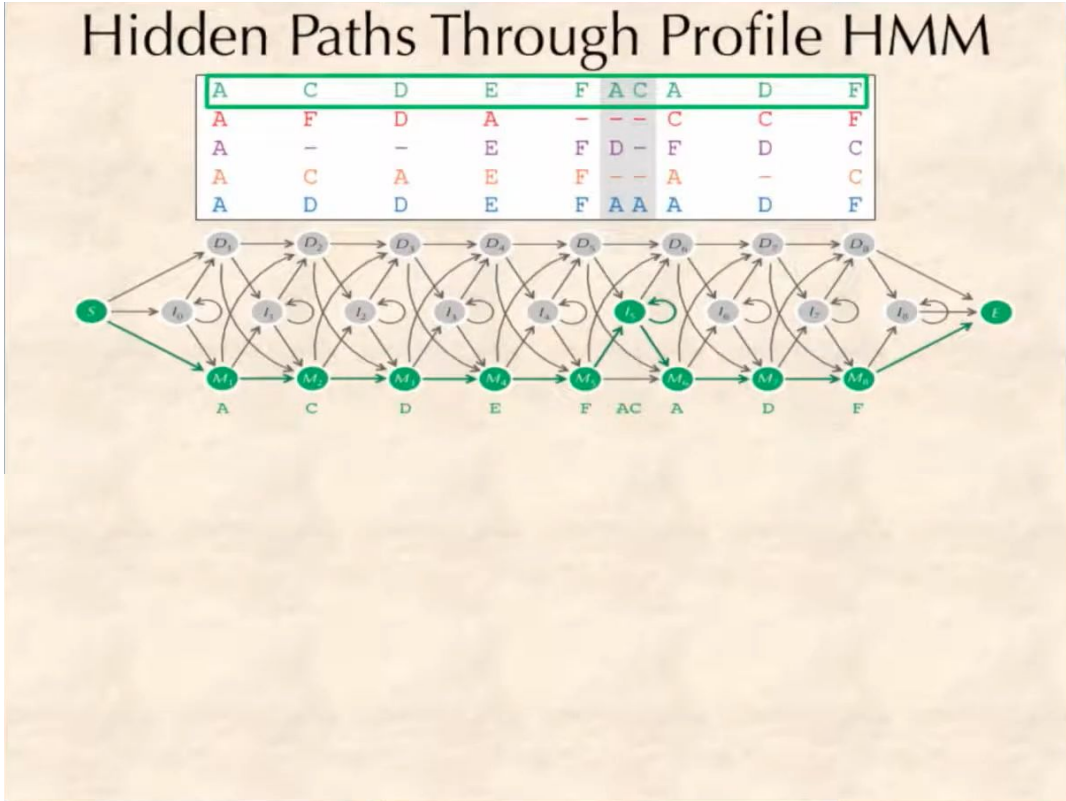
Anatomy of a profile HMM



- Add **start** and **end states** (no emissions)

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

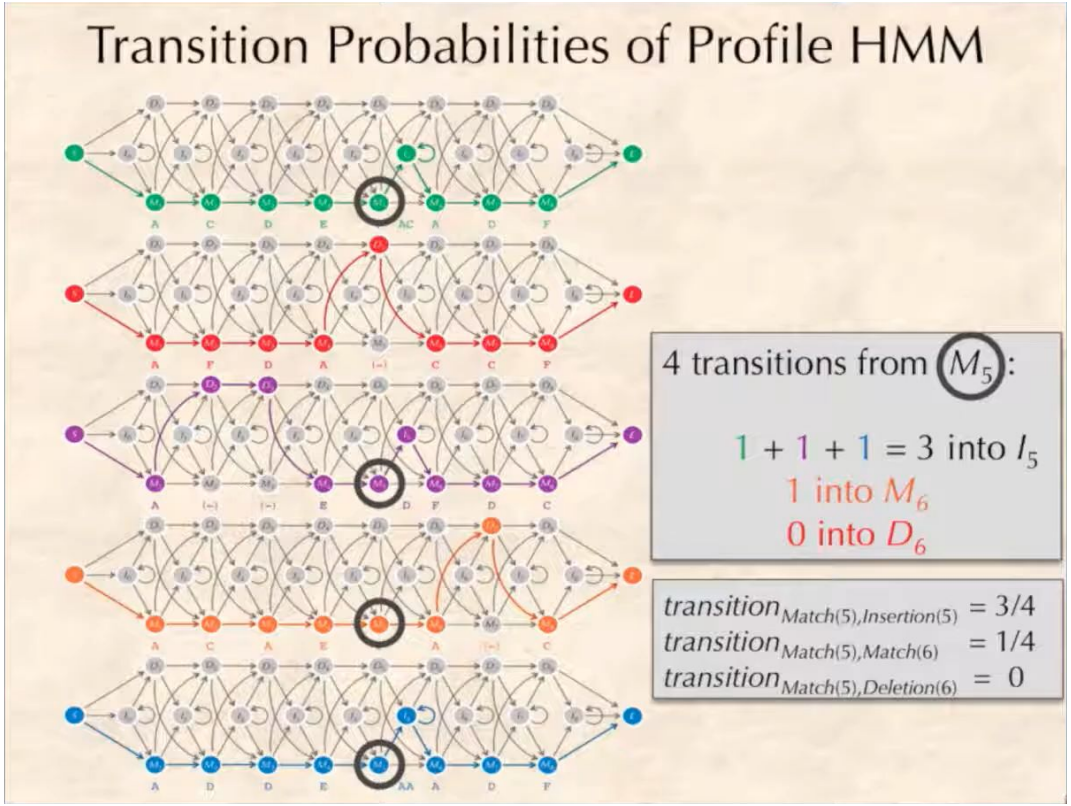
Anatomy of a profile HMM



- Example of the **path** taken through the profile HMM by the first sequence in the original multiple sequence alignment

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

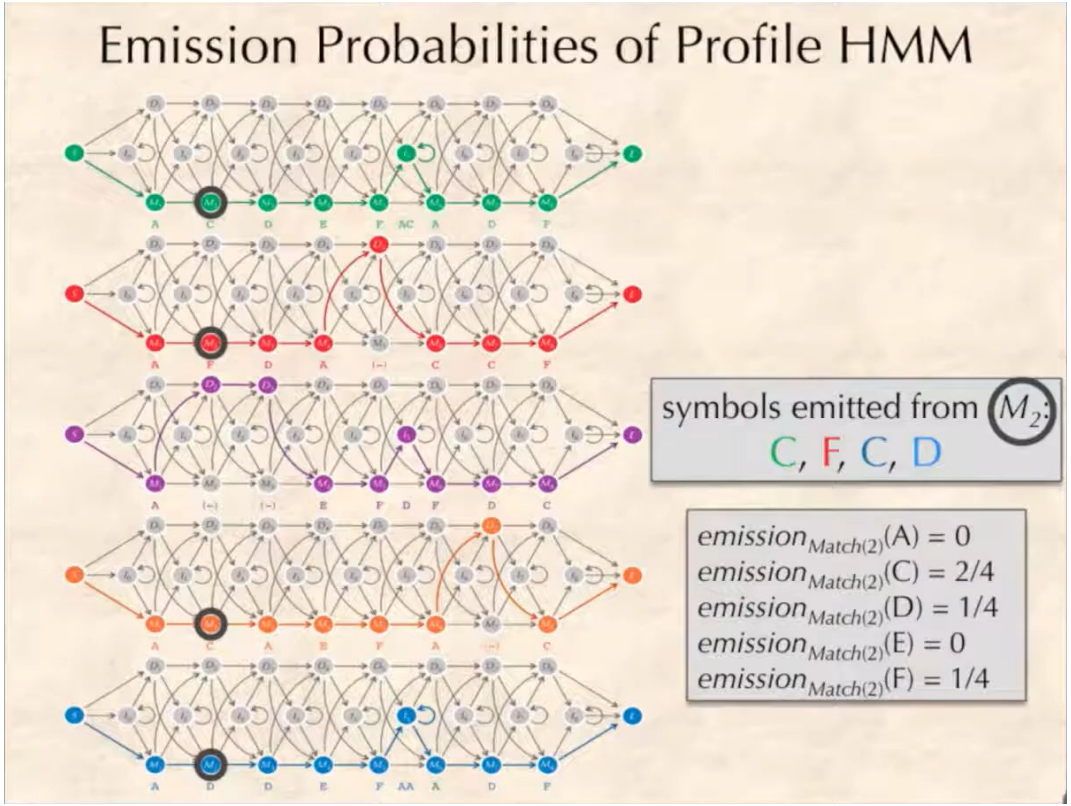
Anatomy of a profile HMM



- **Transition probabilities** are determined by looking at the frequency of transitions from the paths of the original sequences in the multiple sequence alignment

awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

Anatomy of a profile HMM



- **Transition probabilities** are determined by looking at the frequency of transitions from the paths of the original sequences
- Likewise, **emission probabilities** are determined by the frequency of amino acids in the original sequences for each state

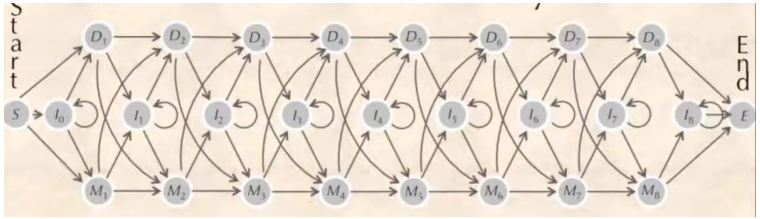
awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao

Anatomy of a profile HMM

A profile HMM is an HMM with **match**, **deletion**, and **insertion states** for each position in a **multiple sequence alignment**.

The **probabilities** of **transitioning** from one state to another and of **emitting** specific amino acids at each state depend on the occurences in the original multiple sequence alignment.

When aligning new sequences to the HMM, individual scoring parameters for each edge **capture subtle similarities** that evade traditional alignments



awkwardly cropped from Pavel Pevzner: https://www.youtube.com/watch?v=vO_6xfLwGao; https://www.youtube.com/watch?v=HbA0odlLuZs

PROTEINS: Structure, Function, and Genetics 28:405–420 (1997)

Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments

Erik L.L. Sonnhammer,[1] Sean R. Eddy,[2] and Richard Durbin[1]*
[1]*Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*
[2]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri*

What are the main goals of the paper?

What are the main goals of the paper?

- Create database with quality alignments to describe as many protein families as possible
  - Stable accession numbers, easily updatable data
- Show how quality database can be used to classify unannotated proteins and define new family memberships for known proteins

How did they do it?

How did they do it?

Pfam-A:

- Constructed **seed alignment** (starter set) for each known family
  - Using known nonredundant set of full length domain sequences
  - Clustalw software
- Then built profile HMM for seed alignment
  - HMMer software
  - Transition probabilities estimated from multiple sequence alignment or determined iteratively
- Then aligned other sequences to HMM to find additional family members
  - Searched all sequences in Swissprot database

- Ad hoc approach to choosing families (went for ones with many known members)

How did they do it?

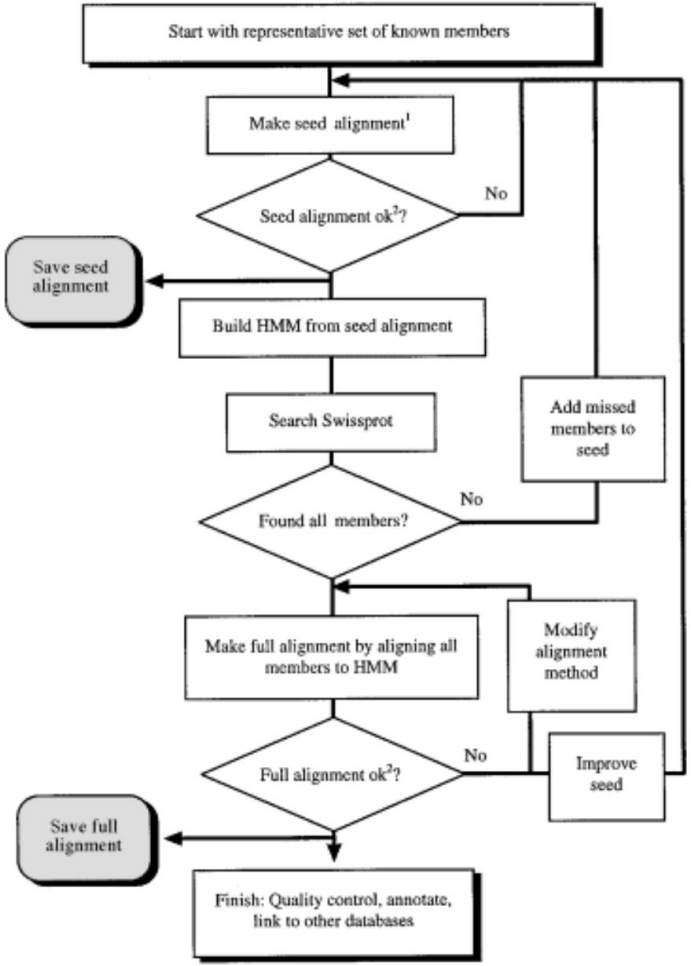
Pfam-B:

- Cluster all Swissprot sequences not in Pfam-A
- Each cluster given temporary accession numbers

Wormpep:

- Aligned Wormpep proteins (4874) to HMMs

Figure 1: The procedure to construct the alignments and HMM for a Pfam-A family



What were the major findings?

Figure 2: Example of the Pfam-A family response_reg (PF00072)

a

```
ID   response_reg
AC   PF00072
DE   Response regulator receiver domain
AU   Sonnhammer ELL
SE   Prodom
AL   Clustalw
GA   Bic_raw 25 hmmls 25
AM   hmma -qR
RA   Pao, G.M., Saier, M.H.
RL   J. Mol. Evol. 40:136-154(1995).
DR   SCOP; 3chy; fa;
CC   This domain receives the signal from the sensor partner in
CC   bacterial two-component systems.  It is usually found N-terminal
CC   to a DNA binding effector domain.
```

b

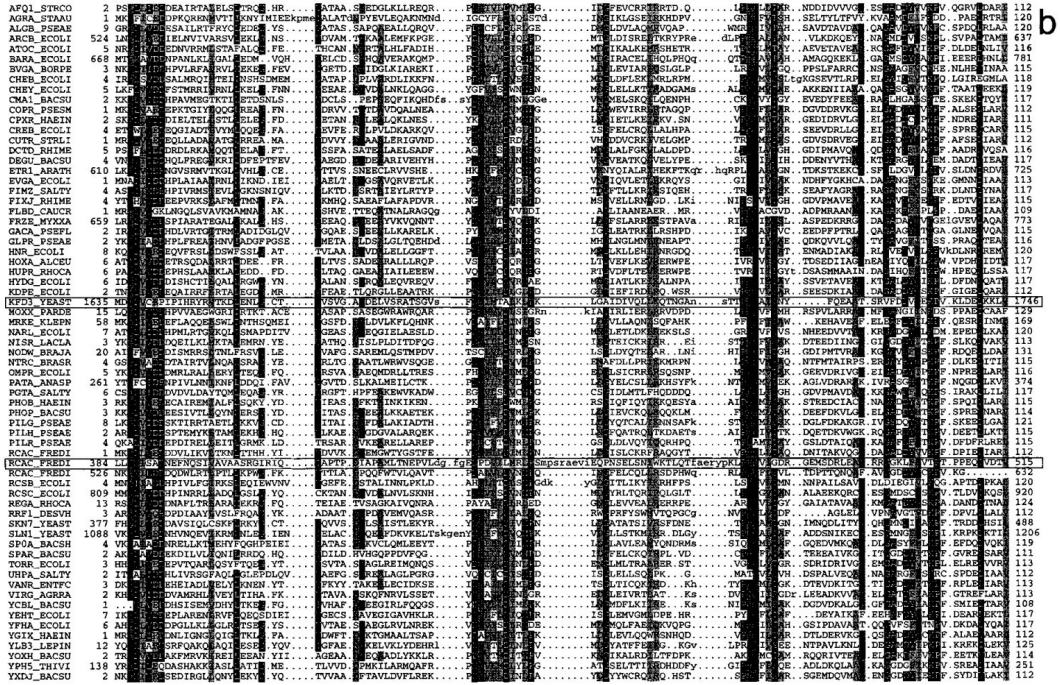


Figure 2: Example of the Pfam-A family response_reg (PF00072)

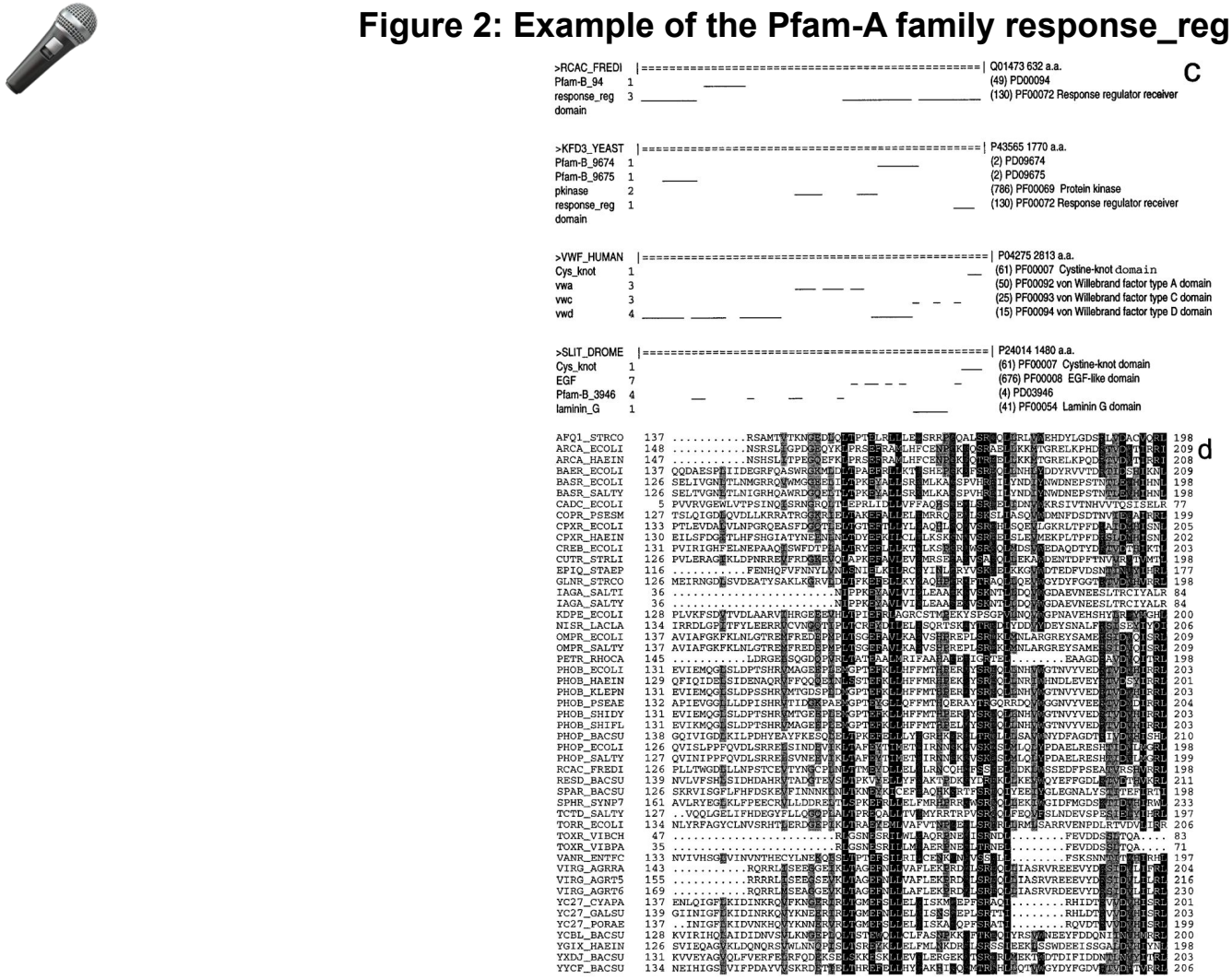


Figure 3: Construction of Pfam-B by Domainer

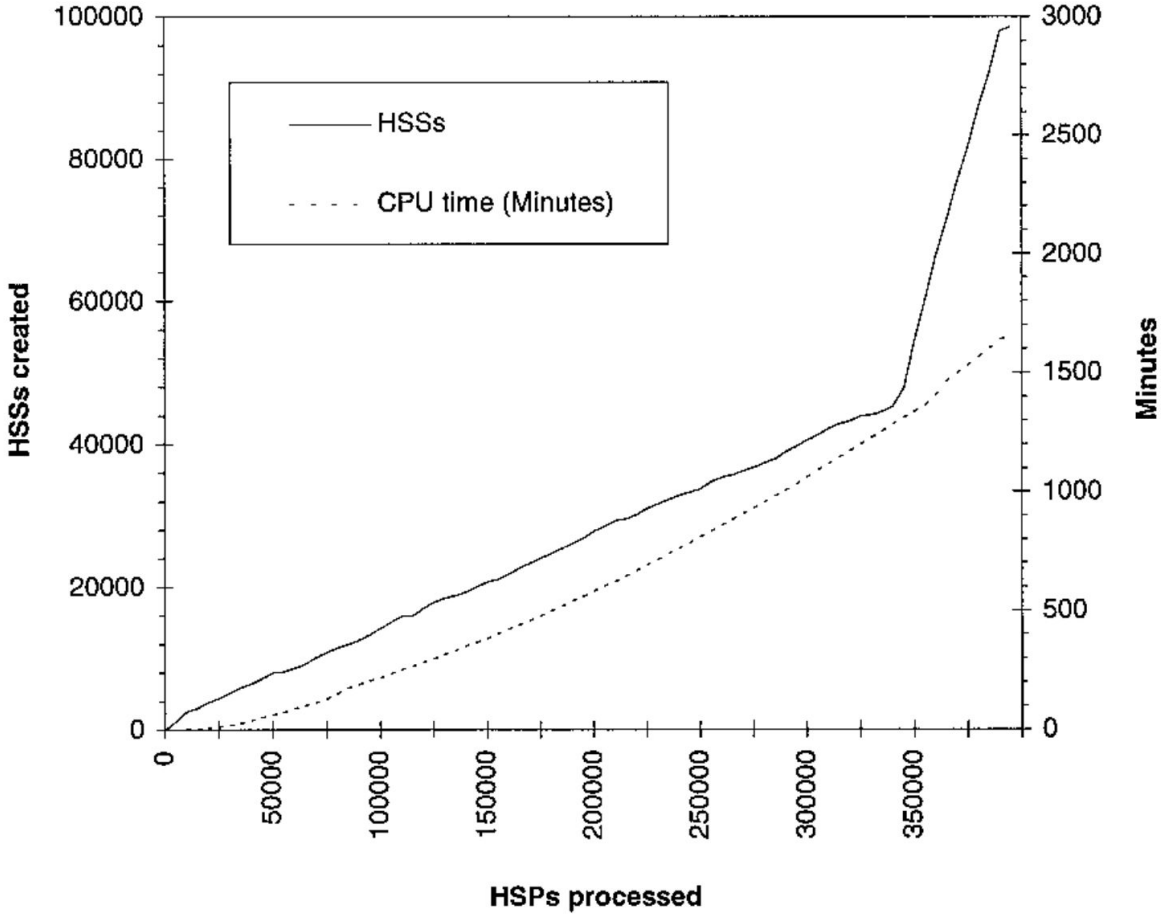
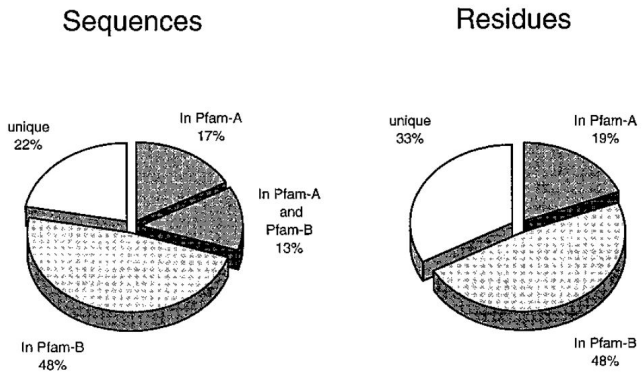


Figure 4: Proportion of Swissprot 33, Wormpep 10 in Pfam



B. Proportion of Wormpep 10 in Pfam 1.0

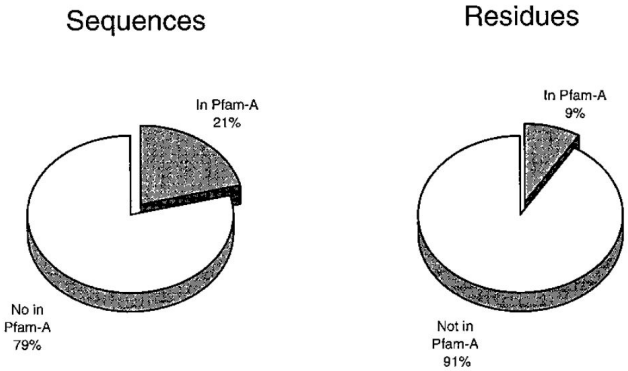


Figure 5: Selected members from Pfam:Cys_knot (PF0007)

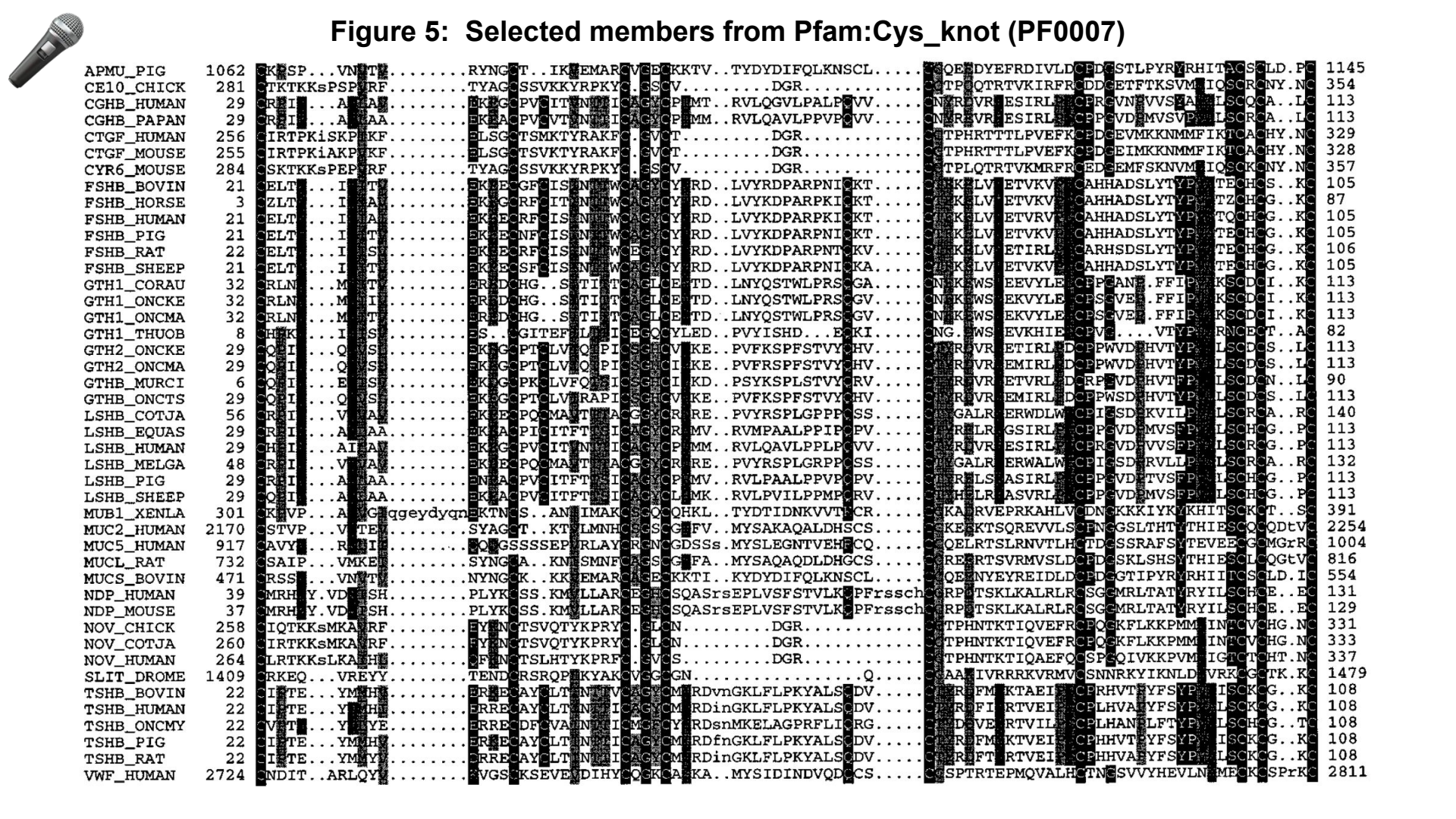


Figure 6: Selected members from Pfam:fn3 (PF00041)

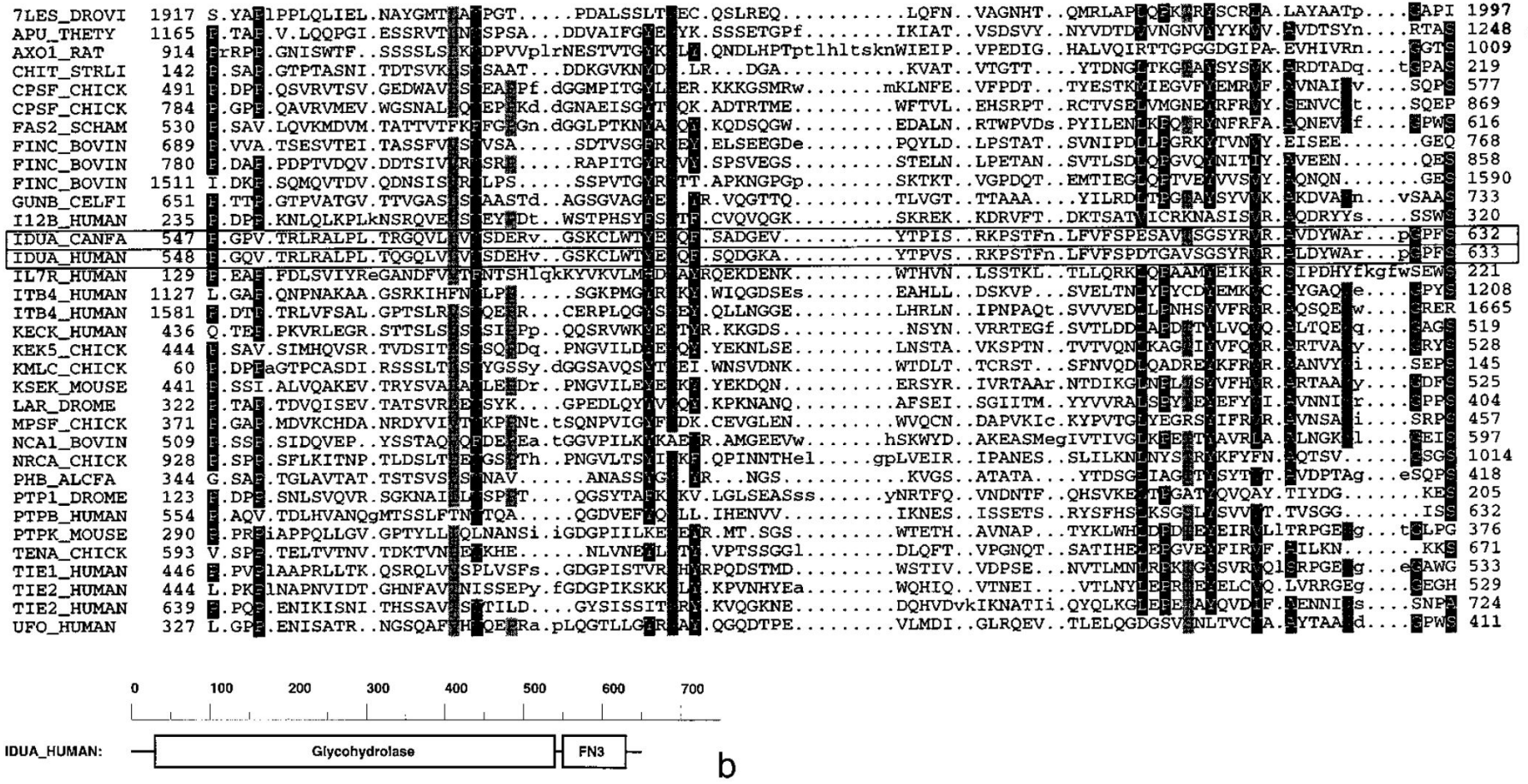


Figure 7: Selected members from Pfam:kazal (PF00050) showing the novel members

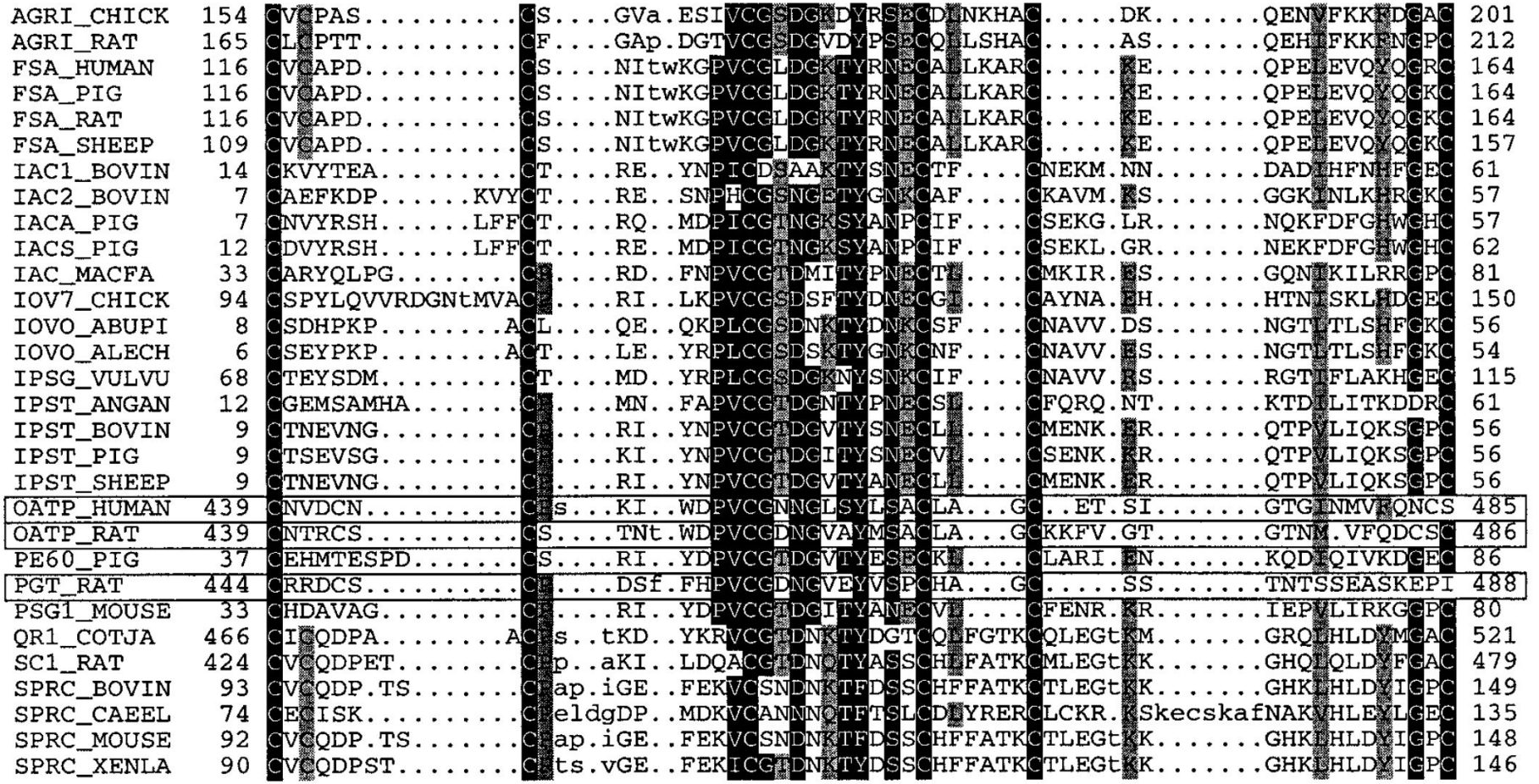


Table I: All families in Pfam-A and their sizes

Table II: Novel Pfam annotations for unclassified *C. elegans* proteins

TABLE III. Comparison of Databases That Contain Protein Family Clusters and Multiple Alignments

	Pfam-A 1.0	Pfam-B 1.0	ProDom 28.0	PIRALN 11.0	BLOCKS 13.0	PRINTS 10.0
Alignment construction	Manual, clustal, HMM	Domainer	Domainer	Pileup	Motif	SOPMA
Source database	Swissprot 33	Swissprot 33	Swissprot 28	PIR 48	Swissprot 32	OWL 26
Clusters	175	11,929	8,031	2,059	872	500
Sequences	15,604	31,931	23,048	11,367	18,593	16,231
Average alignment width (including gaps)	297	180	154	354	32	18
Average cluster size	127	5.7	3.3	6.5	19	37

What were the major findings?

- ⅓ of Swissprot proteins (⅕ of residues) had at least one domain in Pfam-A
- New members found for some families
- New domains annotated for some proteins
- ⅕ of WormPep proteins(1/10 of residues) had at least one domain in Pfam-A
  - Includes some missed by BLAST
- Larger number of members per cluster means Pfam families retain lots of evolutionary information about domains

Why does it matter?

Why does it matter?

- Better alignments - better evolution information
- Quality database with reliable information and cross references
- Publicly available
- Profile HMMs - fast, good at finding related sequences
- Semi-automated method of annotation - saves time

Pfam: Then and now

	Pfam 1.0 (1997)	**Pfam 32.0 / 33.1* (2019-2020)**
Number of families	175	
Number of sequences	15,604	
% SwissProt/UniProt sequences with match in Pfam	30%	
% SwissProt/UniProt residues with match in Pfam	19%	

El-Gebali et al. 2018 Nucleic Acids Research

Pfam: Then and now

	Pfam 1.0 (1997)	**Pfam 32.0 / 33.1* (2019-2020)**
Number of families	175	18,259*
Number of sequences	15,604	~151 million (?)**
% SwissProt/UniProt sequences with match in Pfam	30%	77.2%
% SwissProt/UniProt residues with match in Pfam	19%	53.2%

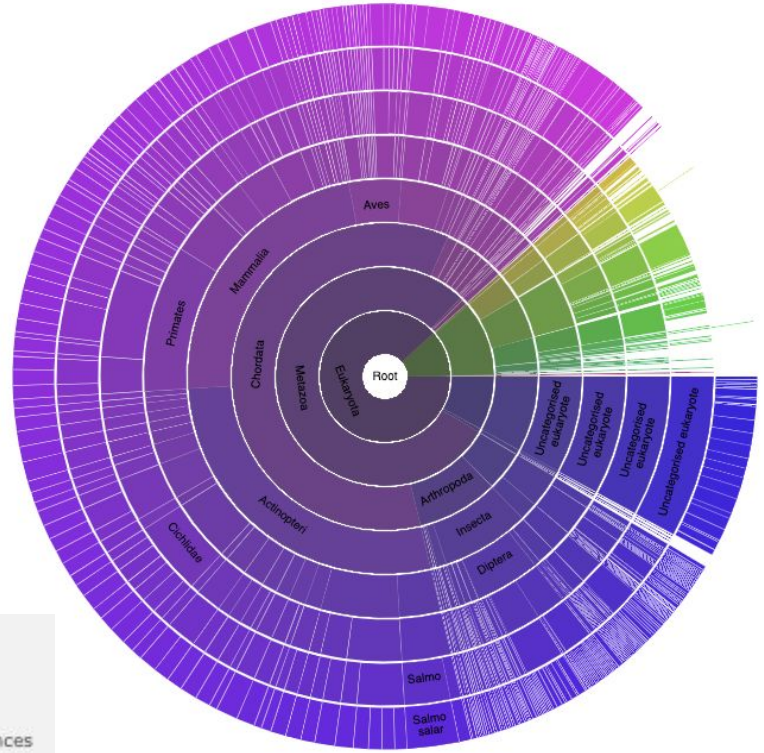
El-Gebali et al. 2018 Nucleic Acids Research

** # UniProt seqs (in 2020) x 0.772

Pfam: Then and now - FN3 domain (PF00041)

Pfam 1.0 (1997):
Thought to be only bacterial, but found 3 eukaryote examples

Pfam 33.1 (2020):
Prokaryotes: 6,439 seqs & 2,013 species
Eukaryotes: 50,108 seqs & 708 species



https://pfam.xfam.org/family/PF00041#tabview=tab7

Assignment

Compare tools for protein domain identification

- Look up information about known protein
- Predict domains of unknown protein

Feel free to work together!