

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?

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

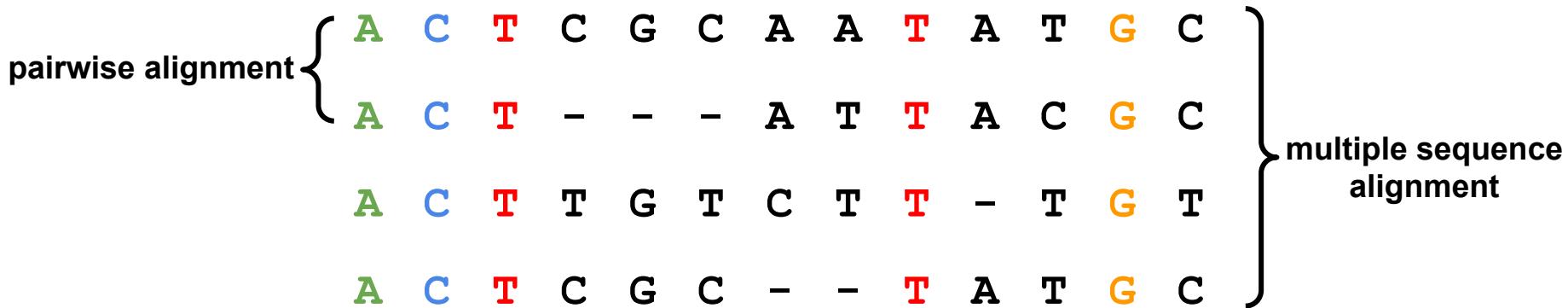
Erik L.L. Sonnhammer,¹ Sean R. Eddy,² and Richard Durbin^{1*}

¹Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

²Department of Genetics, Washington University School of Medicine, St. Louis, Missouri

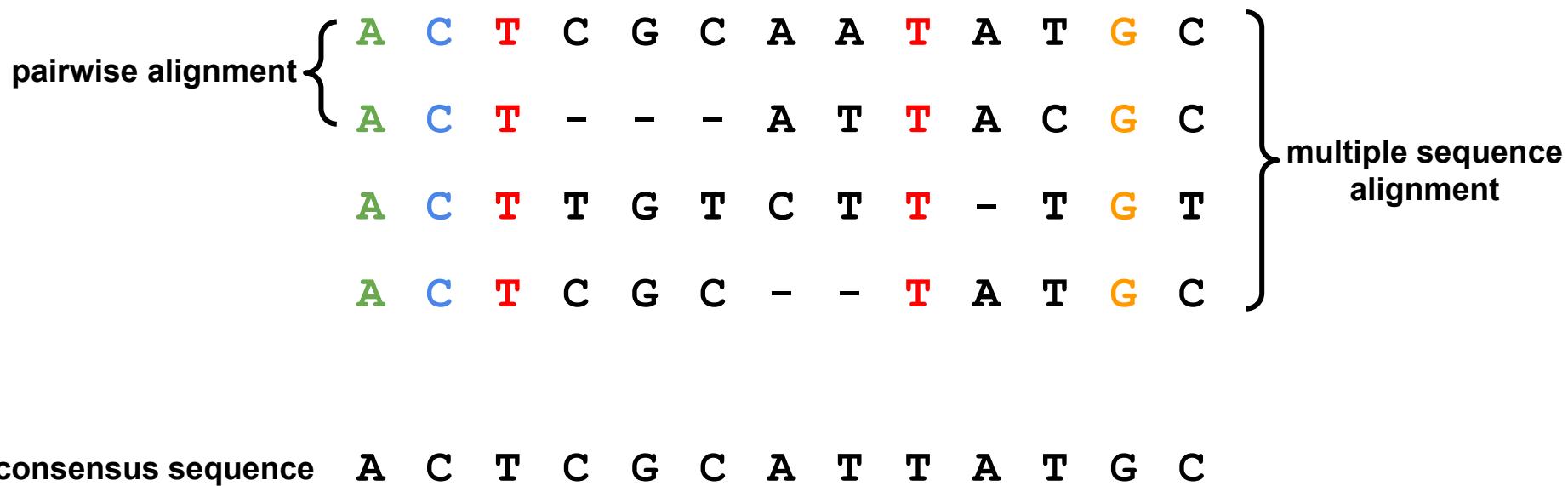
What is an alignment?

- 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**GCATCGCGACGACATCAGCTACGGTCC**
GTCTCA**GCATCGCGACGACATCAGCACGTAGCG**

Local alignment

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

Global alignment

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

'Glocal' alignment



What are alignments used for?

TCAATC**GCATCGCGACGACATCAGCTACGGTCC**
GTCTCA**GCATCGCGACGACATCAGCACGTAGCG**

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

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

Local alignment
find conserved segments

Global alignment
compare similar sequences

'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)

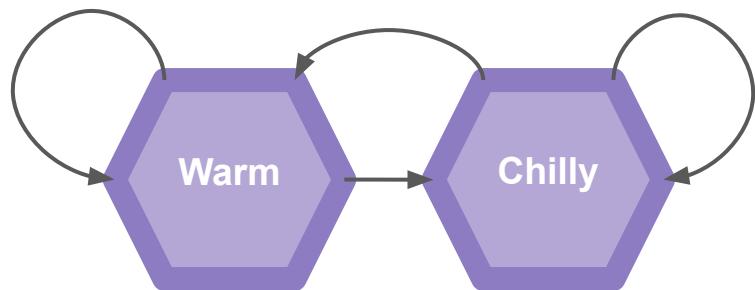
What is a profile Hidden Markov Model (HMM)?

- **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

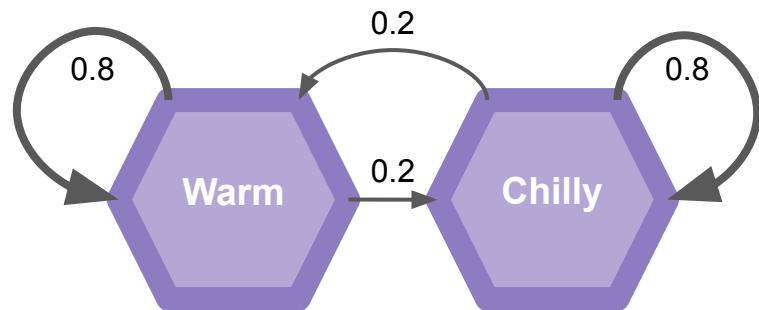
- **Hidden states**
(warm, chilly)



Anatomy of a HMM

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

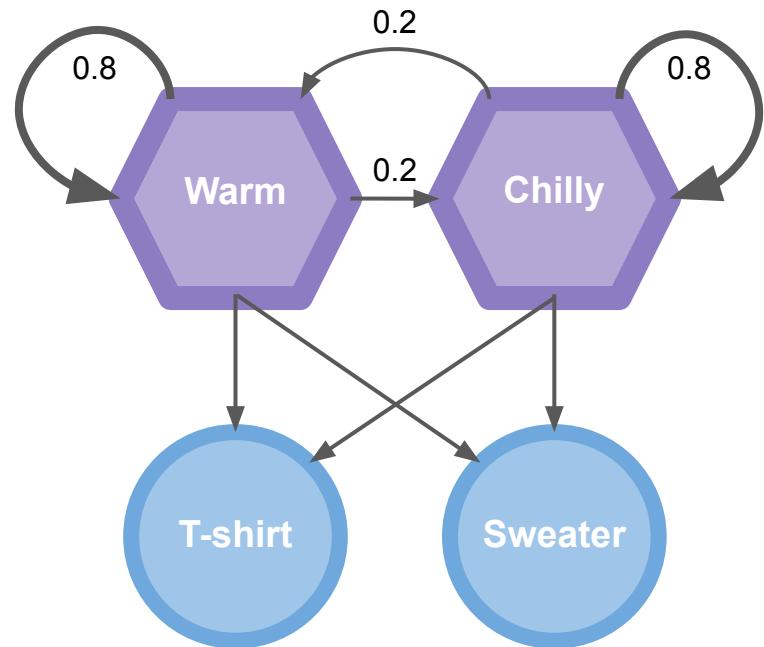
- **Hidden states**
(warm, chilly)
- **Transition probabilities** between states
(likely to have many warm days in a row)



Anatomy of a HMM

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

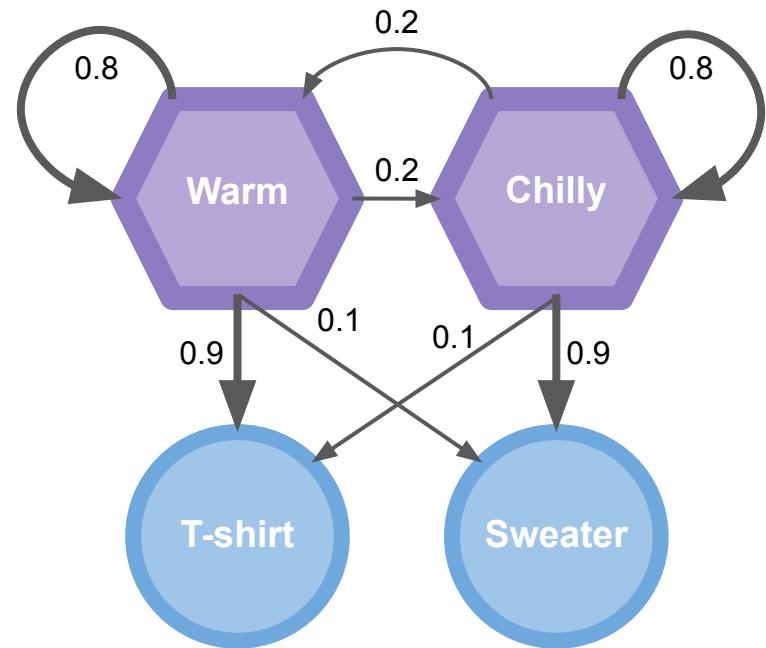
- **Hidden states**
(warm, chilly)
- **Transition probabilities** between states
(likely to have many warm days in a row)
- **Emissions/observable outcomes**
(t-shirt, sweater)



Anatomy of a HMM

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

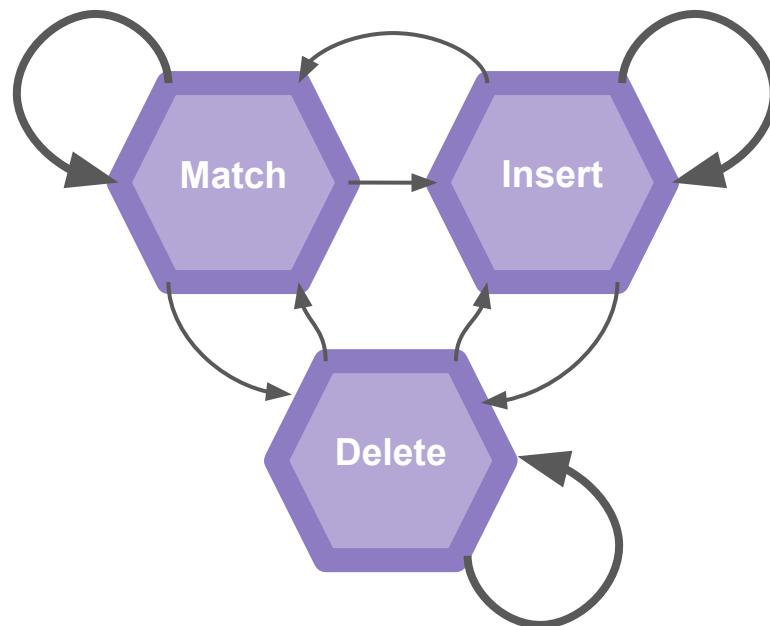
- **Hidden states**
(warm, chilly)
- **Transition probabilities** between states
(likely to have many warm days in a row)
- **Emissions/observable outcomes**
(t-shirt, sweater)
- **Emission probabilities** based on states
(t-shirt more likely during warm day than chilly day)



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

From Profile to HMM

	1	2	3	4	5	6	7	8								
Alignment	A	C	D	E	F	A	A	D								
	A	F	D	A	-	-	C	C								
	A	-	-	E	F	D	F	D								
	A	C	A	E	F	-	A	-								
	A	D	D	E	F	A	A	D								
Alignment*	A	C	D	E	F	A	D	D								
	A	F	D	A	-	C	C	F								
	A	-	-	E	F	F	D	C								
	A	C	A	E	F	A	-	C								
	A	D	D	E	F	A	D	D								
PROFILE(Alignment*)	A	1	0	0	1/5	0	3/5	0								
	C	0	2/4	0	0	0	1/5	1/4								
	D	0	1/4	3/4	0	0	0	3/4								
	E	0	0	0	4/5	0	0	0								
	F	0	1/4	0	0	1	1/5	0								
								3/5								
HMM diagram		M_1	\rightarrow	M_2	\rightarrow	M_3	\rightarrow	M_4	\rightarrow	M_5	\rightarrow	M_6	\rightarrow	M_7	\rightarrow	M_8

- Start with **multiple sequence alignment**

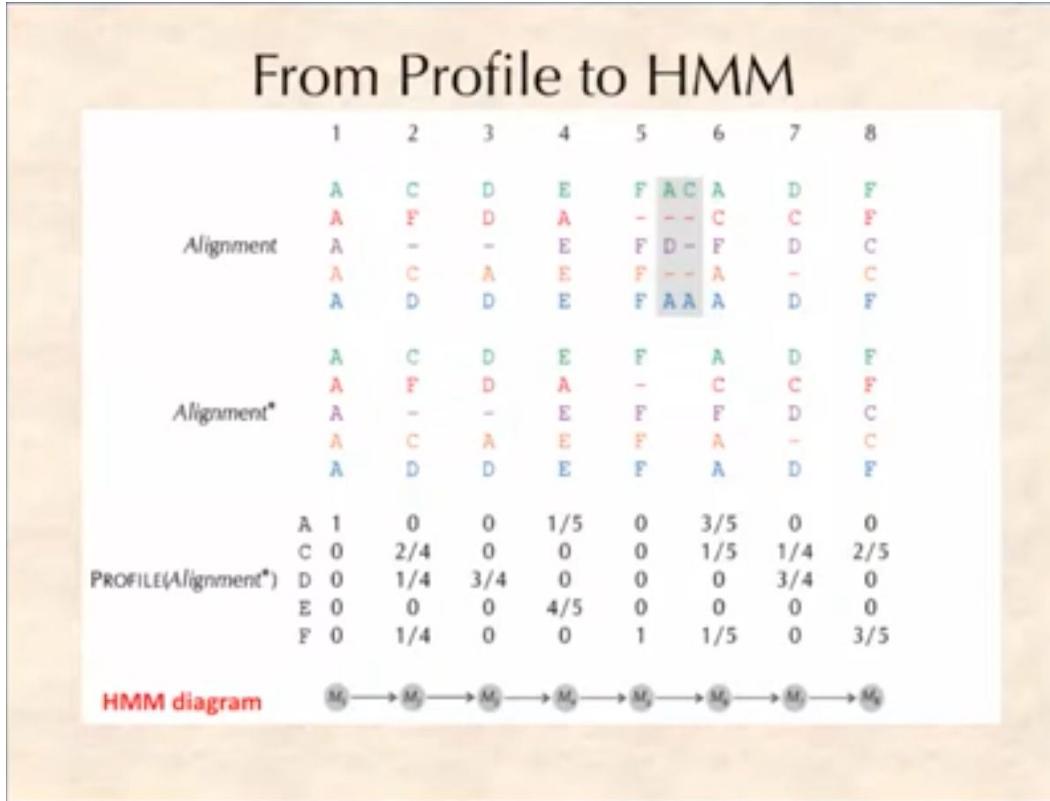
Anatomy of a profile HMM

From Profile to HMM

	1	2	3	4	5	6	7	8
Alignment	A	C	D	E	F	A	A	D
	A	F	D	A	-	-	C	C
	A	-	-	E	F	D	F	D
	A	C	A	E	F	-	A	-
	A	D	D	E	F	A	A	D
Alignment*	A	C	D	E	F	A	D	E
	A	F	D	A	-	C	C	E
	A	-	-	E	F	F	D	C
	A	C	A	E	F	A	-	C
	A	D	D	E	F	A	D	E
PROFILE(Alignment*)	A	1	0	0	1/5	0	3/5	0
	C	0	2/4	0	0	0	1/5	1/4
	D	0	1/4	3/4	0	0	0	3/4
	E	0	0	0	4/5	0	0	0
	F	0	1/4	0	0	1	1/5	0
HMM diagram								
$M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow M_4 \rightarrow M_5 \rightarrow M_6 \rightarrow M_7$								

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

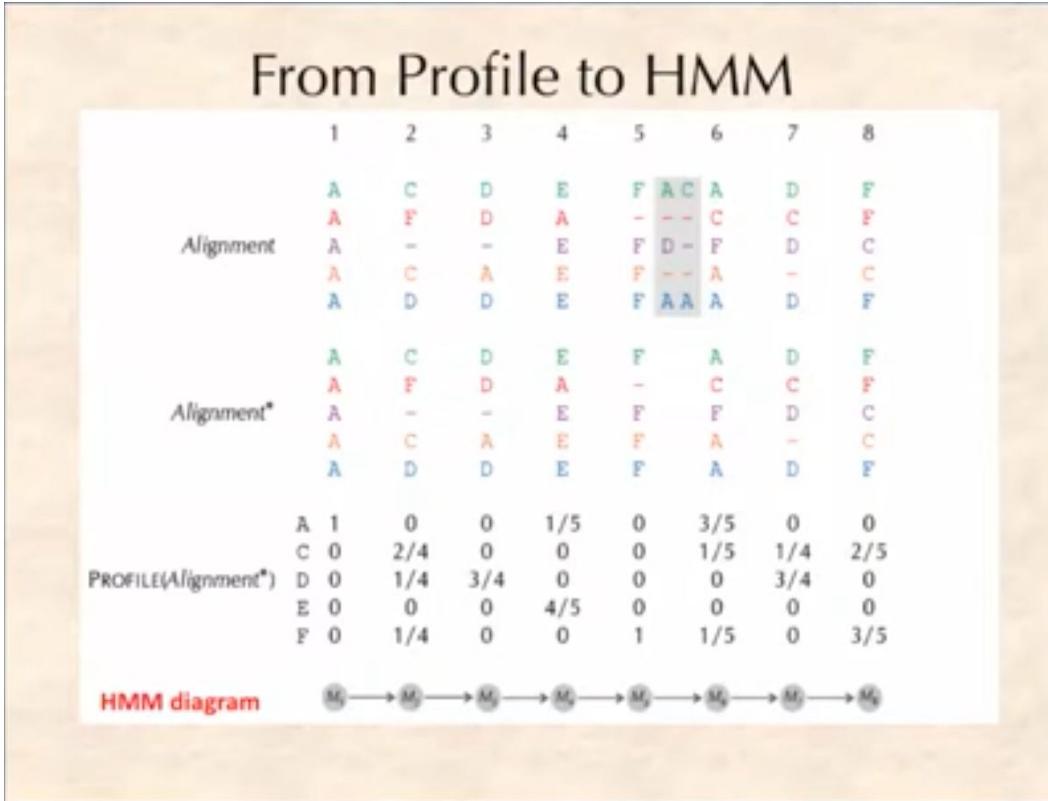
Anatomy of a profile HMM



- 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

Anatomy of a profile HMM

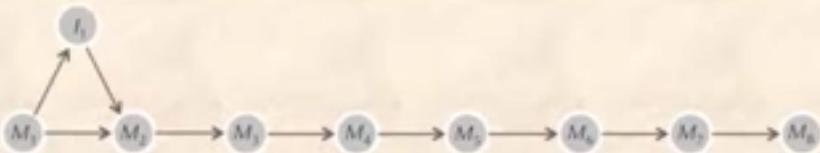


- Start with **multiple sequence alignment**
- Remove columns where the number of sequences with insertions are above a threshold
- Create a **match state*** for each remaining position
- Determine frequency of amino acids at each position
 - This determines the possible **emissions** (and the **emission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

*match is match or mismatch

Anatomy of a profile HMM

Toward a Profile HMM: Insertions

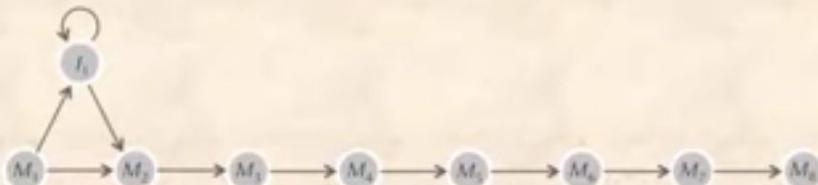


A **F** D D A F F D F

- Add **insertion state** between match states

Anatomy of a profile HMM

Toward a Profile HMM: Insertions

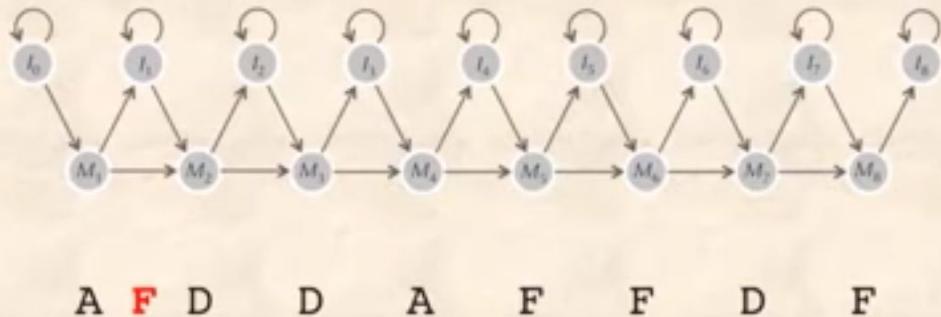


A **F** D D A F F D F

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

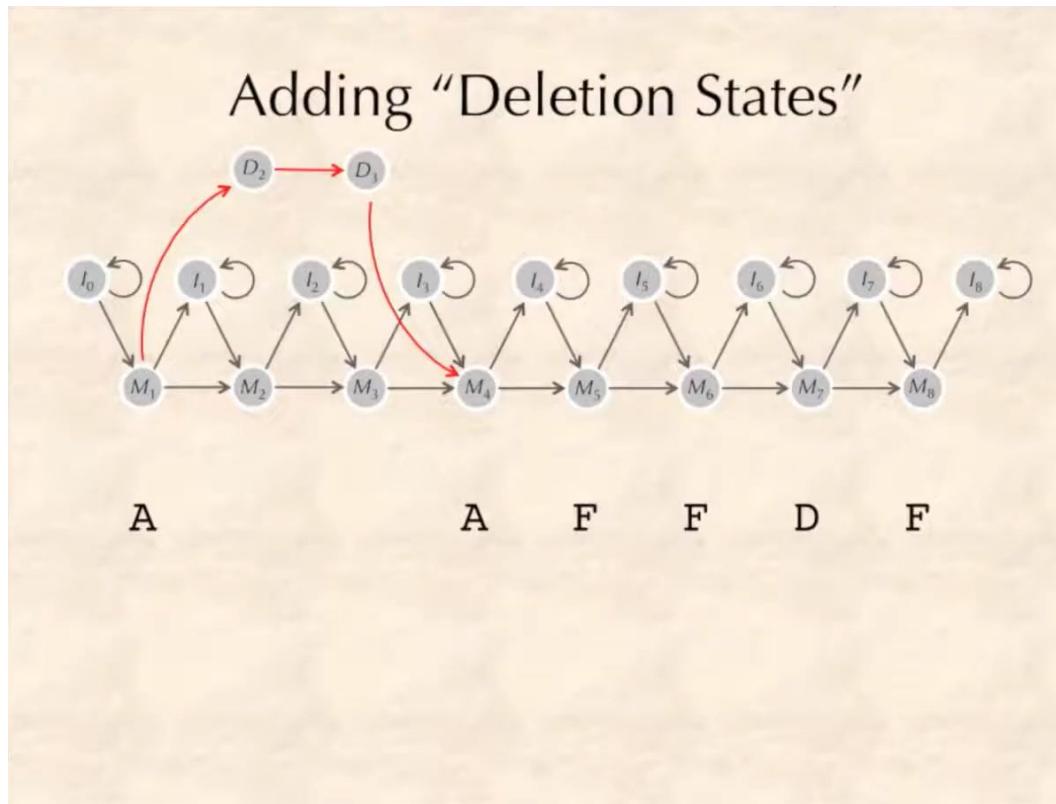
Anatomy of a profile HMM

Toward a Profile HMM: Insertions



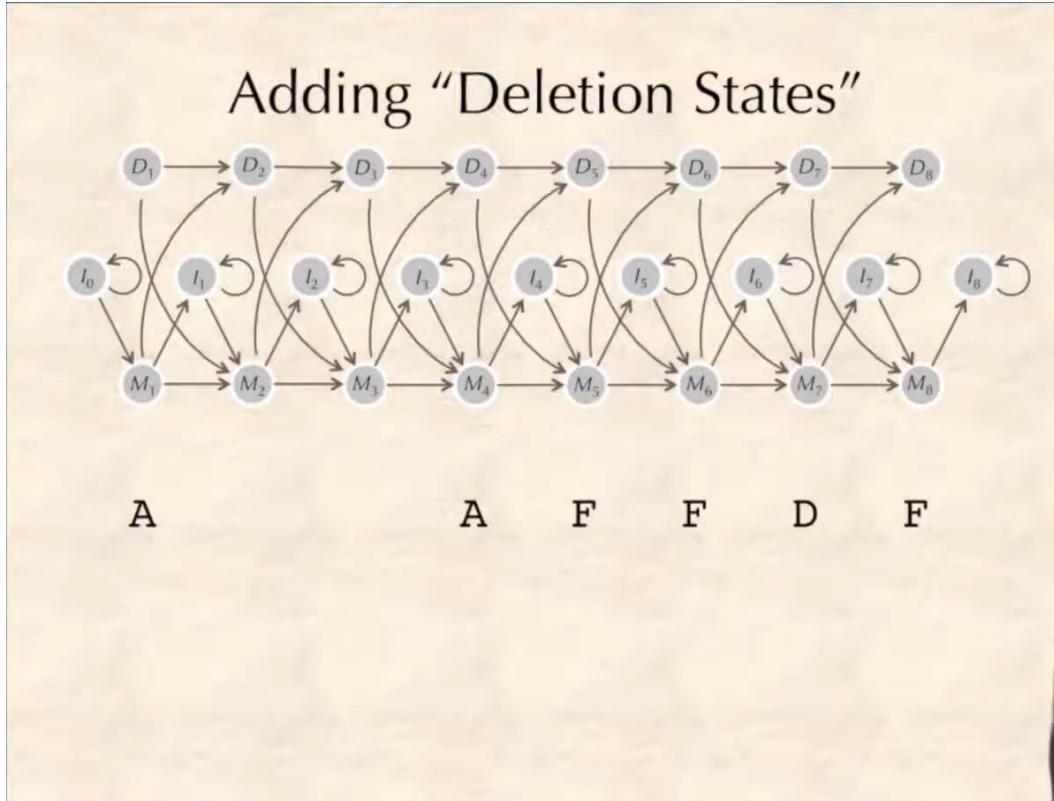
- Add **insertion state** between match states
- Allow for multiple insertions in a row
- Add these insertion states between each match state (and before the first and after the last)
 - Adding them everywhere (instead of only at locations observed in the multiple sequence alignment) allows you to align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

Anatomy of a profile HMM



- Add **deletion states** between match states

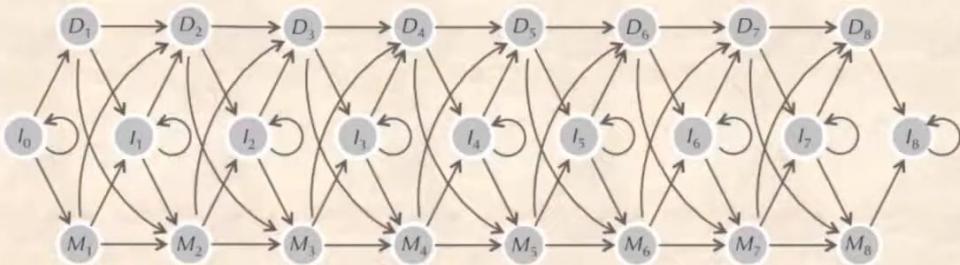
Anatomy of a profile HMM



- 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

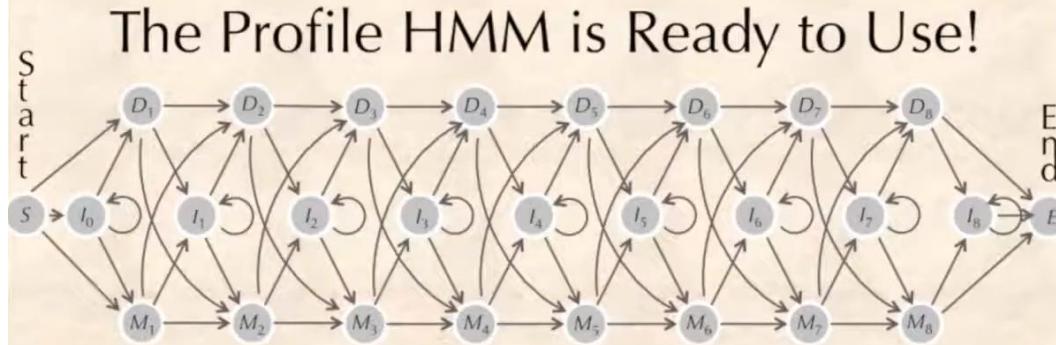
Anatomy of a profile HMM

Adding Edges Between Deletion/Insertion States



- 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

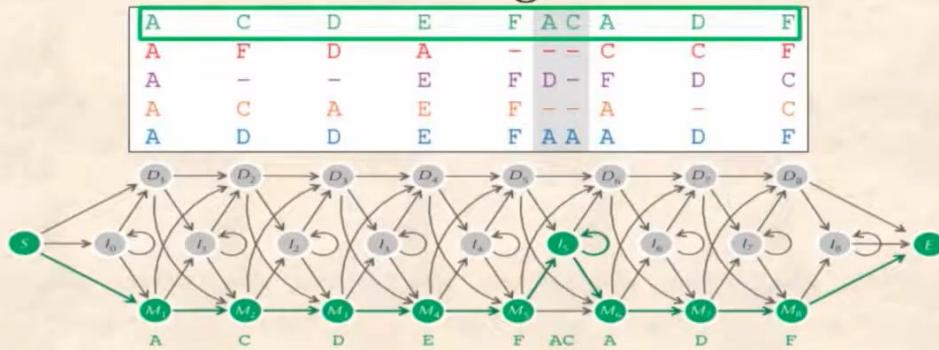
Anatomy of a profile HMM



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

Anatomy of a profile HMM

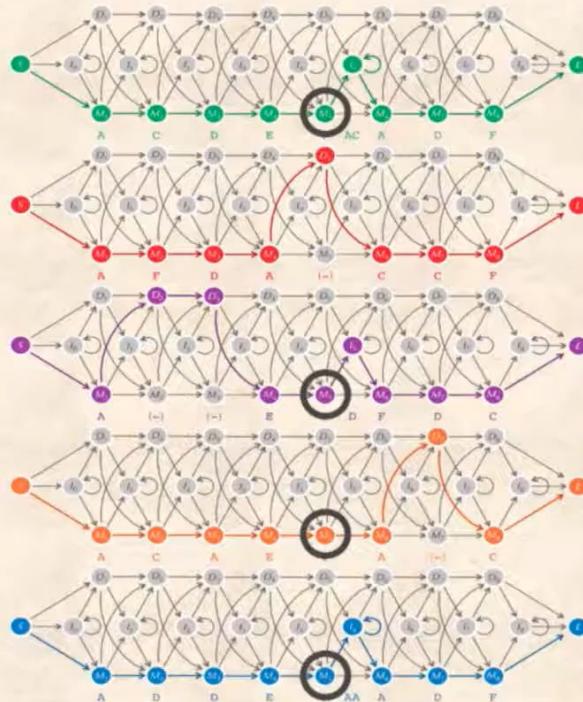
Hidden Paths Through Profile HMM



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

Anatomy of a profile HMM

Transition Probabilities of Profile HMM



4 transitions from M_5 :

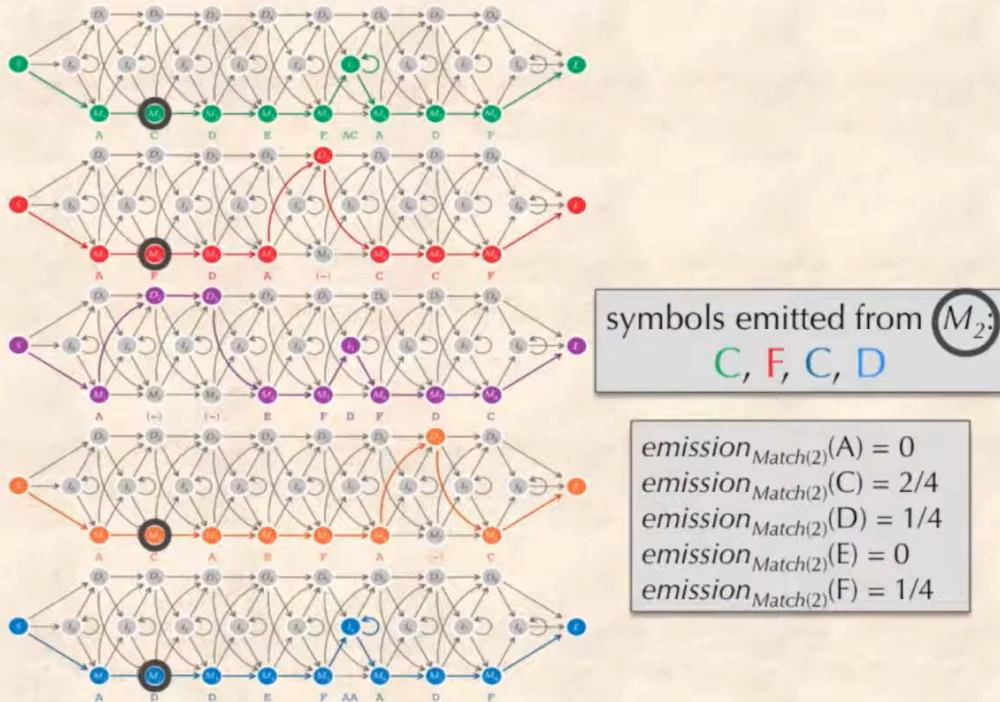
$$1 + 1 + 1 = 3 \text{ into } I_5 \\ 1 \text{ into } M_6 \\ 0 \text{ into } D_6$$

$$\text{transition}_{\text{Match}(5), \text{Insertion}(5)} = 3/4 \\ \text{transition}_{\text{Match}(5), \text{Match}(6)} = 1/4 \\ \text{transition}_{\text{Match}(5), \text{Deletion}(6)} = 0$$

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

Anatomy of a profile HMM

Emission Probabilities of Profile HMM



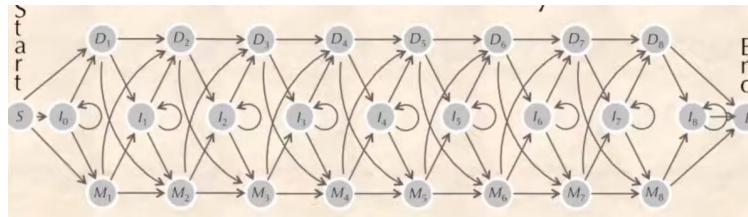
- **Transition probabilities** are determined by looking at the frequency of transitions from the paths of the original sequences
- Likewise, **emission probabilities** are determined by the frequency of amino acids in the original sequences for each state

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 occurrences 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



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*}

¹Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

²Department 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:

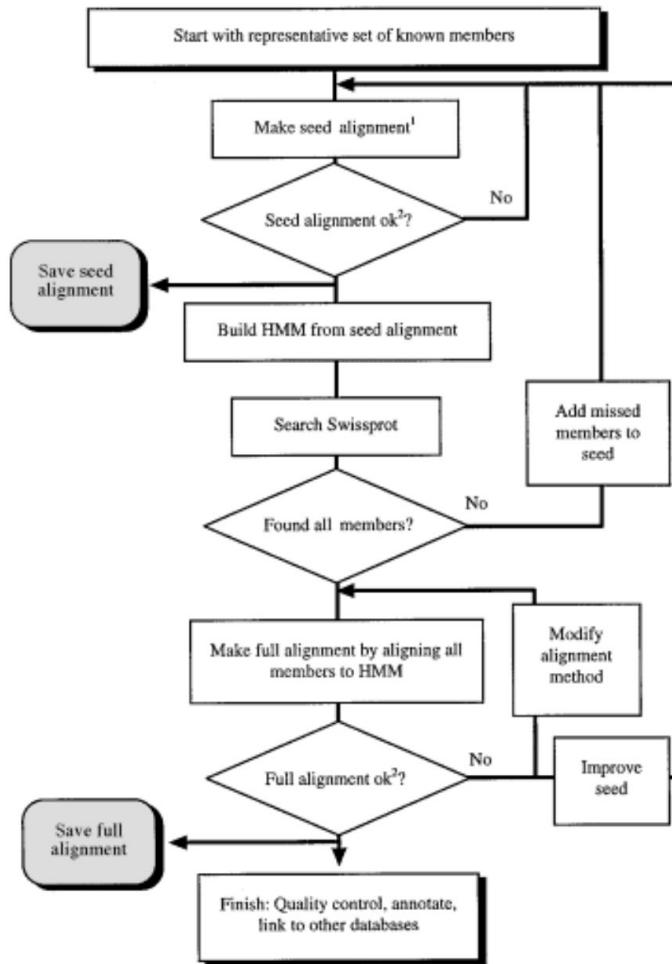
- Cluster all Swissprot sequences not in Pfam-A
- Each cluster given temporary accession numbers

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)

ID response_reg
AC PF00072
DE Response regulator receiver domain
AU Sonnhammer ELL
SE Prodom
AL Clustalw
GA Bic_raw 25 hmmls 25
AM hmmer -qR
RA Pao, G.M., Saier, M.H.
RL J. Mol. Evol. 40:136-154(1995).
DR SCOP; 3ch; 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.

2

PAR 112
TRG 120
LAA 120



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



Figure 3: Construction of Pfam-B by Domainer

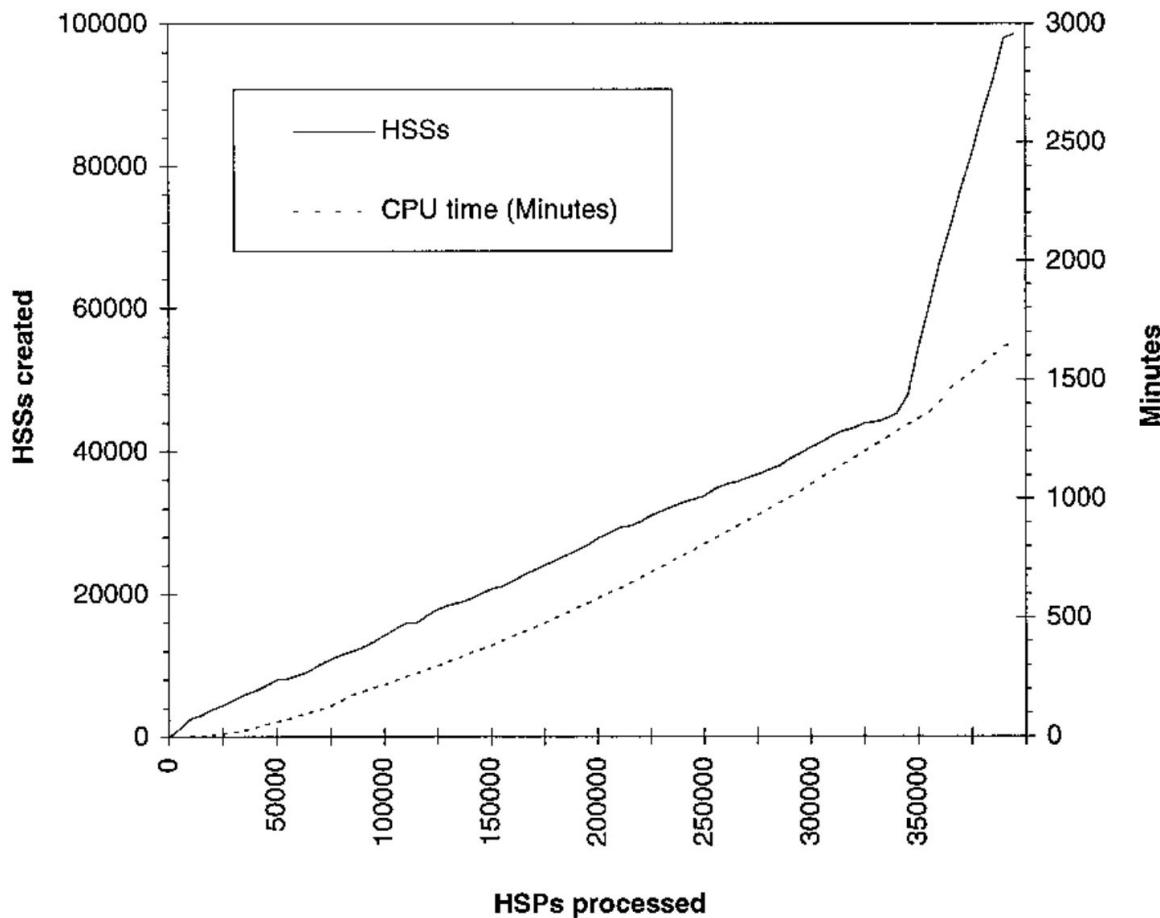
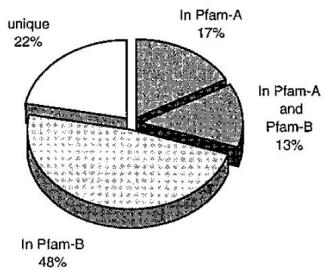


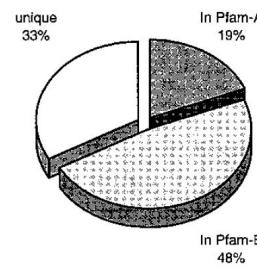


Figure 4: Proportion of Swissprot 33, Wormpep 10 in Pfam

Sequences

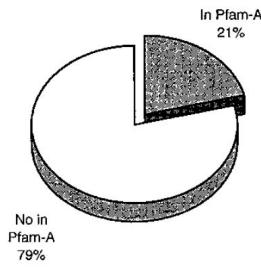


Residues



B. Proportion of Wormpep 10 in Pfam 1.0

Sequences



Residues

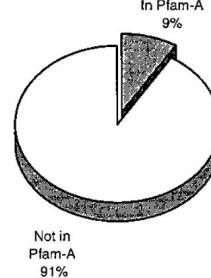


Figure 5: Selected members from Pfam:Cys_knot (PF0007)

APMU_PIG	1062	QKSP...VNIT...	RYNGCT..IKEMARCVGEKKTV..TYDYDIFQLKN...CL...	QEIDYEFRDIVLDPC...STLPYRVRHITAC...SCLD...P...	1145
CE10_CHICK	281	QTKTKKsPS...RF...	TYAGSSVKKYRPKYC...GSCV...DGR...	TPQTRTRVKIRFR...DGETFTKSVM...IQSCRC...N...	354
CGHB_HUMAN	29	CRII...A...AA	MKGCPVCITN...ICAGMCP...MT...RVLQGVLPALP...VV...	CNRDRVR...ESIRL...CPGVN...VVS...A...LSCQC...A...	113
CGHB_PAPAN	29	CRII...A...AA	EKACACPVCVT...N...ICAGMCP...MM...RVLQAVLPPV...P...VV...	CNRDRVR...ESIRL...CPGV...D...MVSV...P...LSCRC...A...	113
CTGF_HUMAN	256	IRTPKISKPK...KF	ELSGCTSMKTYRAKFC...GVCT...DGR...	CPHRTTTLPVEFK...CPGEIMKKNNMF...IKTC...ACHY...N...	329
CTGF_MOUSE	255	IRTPKIAKP...KF	ELSGCTSVKTYRAKFC...GVCT...DGR...	CPHRTTTLPVEFK...CPGEIMKKNNMF...IKTC...ACHY...N...	328
CYR6_MOUSE	284	SKTKKsPEP...RF	TYAGSSVKKYRPKYC...GSCV...DGR...	TPLQTRTVKMRFR...CEGMFSK...NV...IQSCCK...N...	357
FSHB_BOVIN	21	CELT...I...T	EK...ECGFCIS...N...WCAG...CY...RD...LVYKD...PARP...NI...OKT...	CNL...KLV...ETVKV...CAHHADSL...YT...P...TECHC...A...	105
FSHB_HORSE	3	ZL...I...A	EK...GCRFCIS...N...WCAG...CY...RD...LVYKD...PARP...KI...KT...	CNL...KLV...ETVKV...CAHHADSL...YT...P...TZCHC...A...	87
FSHB_HUMAN	21	CELT...I...A	EK...ECR...FCIS...N...WCAG...CY...RD...LVYKD...PARP...KI...KT...	CNL...KLV...ETVRV...CAHHADSL...YT...P...TQCHC...A...	105
FSHB_PIG	21	CELT...I...T	EK...ECNF...CIS...N...WCAG...CY...RD...LVYKD...PARP...NI...KT...	CNL...KLV...ETVKV...CAHHADSL...YT...P...TECHC...A...	105
FSHB_RAT	22	CELT...I...S	EK...ECR...FCIS...N...WCAG...CY...RD...LVYKD...PARP...NI...KT...	CNL...KLV...ETIRL...CARHSDS...LY...P...TECHC...A...	106
FSHB_SHEEP	21	CELT...I...T	EK...EC...FCIS...N...WCAG...CY...RD...LVYKD...PARP...NI...KA...	CNL...KLV...ETVKV...CAHHADSL...YT...P...TECHC...A...	105
GTH1_CORAU	32	CRLN...M...T	ER...DCHG...S...TI...T...G...C...E...TD...LNYQSTWL...PRS...GA...	CNL...K...WS...EEVYLE...CP...P...C...A...N...FFIP...KSCDC...A...	113
GTH1_ONCKE	32	CRLN...M...I	ER...DCHG...S...TI...T...G...C...E...TD...LNYQSTWL...PRS...GV...	CNL...K...WS...EKVYLE...CP...S...E...V...A...N...FFIP...KSCDC...A...	113
GTH1_ONCMA	32	CRLN...M...T	ER...DCHG...S...TI...T...G...C...E...TD...LNYQSTWL...PRS...GV...	CNL...K...WS...EKVYLE...CP...S...E...V...A...N...FFIP...KSCDC...A...	113
GTH1_THUOB	8	CH...K...I...S	ES...G...TITE...L...I...G...E...G...Y...L...ED...P...V...V...I...SH...D...E...K...I...	CNG...E...S...EVVKHIE...CP...V...A...V...T...P...RNCECT...A...	82
GTH2_ONCKE	29	Q...I...Q...MS	EK...G...OPT...CLV...Q...P...I...C...G...H...C...V...KE...PVFKSPF...STV...HV...	CNR...V...R...ETIRL...DC...PP...W...V...D...H...V...T...P...L...SCDC...A...	113
GTH2_ONCMA	29	Q...I...Q...S	EK...G...OPT...CLV...Q...P...I...C...G...H...C...I...KE...PVFRSPF...STV...HV...	CNR...V...R...EMIRL...DC...PP...W...V...D...H...V...T...P...L...SCDC...A...	113
GTHB_MURCI	6	Q...I...E...S	EK...G...CPK...L...VF...P...I...C...G...H...C...I...KD...PSYKSP...L...STV...RV...	CNR...V...R...ETVRV...DC...P...R...D...H...V...T...F...P...L...SCDN...A...	90
GTHB_ONCTS	29	Q...I...Q...S	EK...G...PT...CLV...L...V...P...I...C...G...H...C...V...KE...PVFKSPF...STV...HV...	CNR...V...R...EMIRL...DC...PP...W...S...D...H...V...T...P...L...SCDC...A...	113
LSHB_COTJA	56	CR...I...V...A	EK...G...P...Q...C...M...A...T...I...AC...G...G...C...R...RE...PVYRSP...L...G...P...P...P...SS...	C...GALR...ERWDL...C...PI...E...S...D...E...K...V...I...P...L...L...SCRC...A...	140
LSHB_EQUAS	29	CR...I...A...AA	EK...G...AC...P...I...C...T...F...T...I...C...G...H...C...R...MV...RVMPA...A...L...P...P...I...P...P...PV...	C...R...L...R...GG...S...I...R...L...C...P...P...V...D...M...V...S...P...P...L...SCCHG...A...	113
LSHB_HUMAN	29	CHEI...A...I...A	EK...G...CP...V...C...I...N...I...C...G...H...C...P...M...M...R...V...L...Q...A...V...L...P...P...P...L...P...VV...	C...R...L...V...R...ESIRL...C...P...R...G...V...D...H...V...S...P...P...L...SCRC...A...	113
LSHB_MEGLA	48	CR...I...V...A	EK...G...E...C...P...Q...C...M...A...T...I...AC...G...H...C...R...RE...PVYRSP...L...G...R...P...P...SS...	C...GALR...ERWALW...C...PI...C...S...D...R...V...L...I...P...L...SCRC...A...	132
LSHB_PIG	29	CR...I...A...AA	EK...G...AC...P...V...C...I...T...F...T...I...C...G...H...C...P...M...V...R...V...L...P...A...L...P...P...V...P...V...	C...R...L...S...AS...I...R...V...L...C...P...P...V...D...T...V...S...P...P...L...SCCHG...A...	113
LSHB_SHEEP	29	Q...I...A...AA	EK...G...AC...P...V...C...I...T...F...T...I...C...G...H...C...L...MK...R...V...L...P...V...I...L...P...P...M...P...R...V...	C...H...L...R...AS...V...R...L...C...P...P...V...D...M...V...S...P...P...L...SCCHG...A...	113
MUB1_XENLA	301	QK...VP...A...G...q...g...e...y...d...y...q...	EKTN...CS...A...N...I...MA...K...C...G...C...O...Q...H...K...L...TYD...T...D...N...K...V...V...T...I...C...R...	K...A...R...V...E...P...R...K...A...H...L...V...D...N...K...K...K...I...V...K...K...H...T...S...C...K...C...T...S...	391
MUC2_HUMAN	2170	GSTVP...V...TE	SYAGCT...K...T...L...M...N...H...G...C...G...C...F...V...M...Y...S...A...K...A...Q...A...L...D...H...S...C...S...	K...E...K...T...S...Q...R...E...V...V...L...S...P...C...P...G...G...S...L...T...H...T...H...I...S...C...Q...Q...D...t...V...	2254
MUC5_HUMAN	917	CAVY...R...I	Q...G...G...S...S...S...E...P...R...L...A...Y...R...E...N...G...C...G...D...S...S...M...Y...S...L...E...G...N...T...V...E...H...F...C...Q...	Q...E...L...T...S...L...R...N...V...T...L...H...C...D...S...S...R...A...F...S...T...V...E...E...C...G...C...M...G...R...C...	1004
MUCL_RAT	732	CSAIP...VMKE	SYNG...C...A...K...N...S...M...N...C...G...C...G...C...F...A...M...Y...S...A...Q...A...Q...D...L...D...H...G...C...S...	R...E...R...T...S...V...R...M...V...S...L...D...C...P...D...C...S...K...L...S...H...S...Y...T...H...I...S...C...L...Q...Q...G...t...V...	816
MUCS_BOVIN	471	CRSS...VN...T	NYNG...C...K...K...E...M...A...R...C...G...E...C...K...K...T...I...K...Y...D...D...I...F...Q...L...K...N...S...C...L...	Q...E...N...Y...E...R...E...I...D...L...C...P...D...G...G...T...I...P...Y...R...R...H...I...I...P...S...C...L...D...I...C...	554
NDP_HUMAN	39	QMRH...Y...VD...I...SH	PLYK...C...S...K...M...L...L...A...R...C...G...E...G...C...S...Q...A...S...r...s...E...P...L...V...S...F...S...T...V...L...K...C...	C...R...P...T...S...K...L...K...A...L...R...L...R...S...G...C...M...R...L...T...A...T...R...Y...I...L...S...C...H...C...E...E...C...	131
NDP_MOUSE	37	QMRH...Y...VD...I...SH	PLYK...C...S...K...M...L...L...A...R...C...G...E...G...C...S...Q...A...S...r...s...E...P...L...V...S...F...S...T...V...L...K...C...	C...R...P...T...S...K...L...K...A...L...R...L...R...S...G...C...M...R...L...T...A...T...R...Y...I...L...S...C...H...C...E...E...C...	129
NOV_CHICK	258	CIQT...KKsMKA...RF	HY...N...C...T...S...V...Q...T...Y...K...P...R...Y...C...G...L...C...N...	CP...H...N...T...K...T...I...Q...V...E...F...R...C...P...Q...G...K...F...L...K...K...P...M...M...I...N...T...C...V...C...H...G...	331
NOV_COTJA	260	IRTP...KKsMKA...RF	HY...N...C...T...S...V...Q...T...Y...K...P...R...Y...C...G...L...C...N...	CP...H...N...T...K...T...I...Q...V...E...F...R...C...P...Q...G...K...F...L...K...K...P...M...M...I...N...T...C...V...C...H...G...	333
NOV_HUMAN	264	ILRT...KKsLKA...RF	QF...N...C...T...S...L...H...T...Y...K...P...R...Y...C...G...L...C...N...	CP...H...N...T...K...T...I...Q...A...E...F...Q...C...S...P...Q...I...V...K...K...P...V...M...I...G...T...O...T...C...H...T...N...	337
SLIT_DROME	1409	CRKEQ...V...REYY	TEND...C...R...S...Q...P...K...Y...A...K...C...V...G...G...C...N...	Q...A...A...I...V...R...R...K...V...R...M...V...C...S...N...N...R...K...Y...I...K...N...L...V...R...K...C...G...T...K...	1479
TSHB_BOVIN	22	CI...TE...Y...M...I...H	ER...R...E...C...A...Y...C...L...T...N...I...C...G...M...C...F...V...G...C...M...I...R...D...v...n...G...K...L...F...L...P...K...Y...A...L...S...D...V...	C...R...F...M...I...K...T...A...E...I...C...P...R...H...V...T...Y...F...S...P...P...I...S...O...K...C...G...A...	108
TSHB_HUMAN	22	CI...TE...Y...M...I...H	ER...R...E...C...A...Y...C...L...T...N...I...C...G...M...C...F...V...G...C...M...I...R...D...v...n...G...K...L...F...L...P...K...Y...A...L...S...D...V...	C...R...F...I...I...T...R...V...E...I...C...P...L...H...V...A...Y...F...S...P...P...I...S...O...K...C...G...A...	108
TSHB_ONCMY	22	CV...TE...Y...M...I...E	ER...R...E...C...D...F...C...V...A...N...I...C...G...M...C...F...V...G...C...M...I...R...D...f...n...G...K...L...F...L...P...K...Y...A...L...S...D...V...	C...D...V...E...I...T...R...V...I...L...C...P...L...H...A...N...L...F...T...Y...P...P...I...S...O...K...C...G...A...	108
TSHB_PIG	22	CI...TE...Y...M...I...H	ER...R...E...C...A...Y...C...L...T...N...I...C...G...M...C...F...V...G...C...M...I...R...D...v...n...G...K...L...F...L...P...K...Y...A...L...S...D...V...	C...R...F...M...I...K...T...V...E...I...C...P...H...H...V...T...Y...F...S...P...P...I...S...O...K...C...G...A...	108
TSHB_RAT	22	CI...TE...Y...M...I...Y	ER...R...E...C...A...Y...C...L...T...N...I...C...G...M...C...F...V...G...C...M...I...R...D...v...n...G...K...L...F...L...P...K...Y...A...L...S...D...V...	C...A...F...T...I...T...R...V...E...I...C...P...H...H...V...T...Y...F...S...P...P...I...S...O...K...C...G...A...	108
VWF_HUMAN	2724	QNDIT...AR...QY	IVG...S...U...K...S...E...V...E...V...D...I...H...Y...C...G...K...A...K...A...M...Y...S...I...D...I...N...V...Q...D...C...S...	C...S...P...T...R...T...E...P...M...Q...V...A...L...H...C...T...N...G...S...V...V...Y...H...E...V...L...N...M...E...C...K...Q...S...P...R...K...	2811

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

7LES_DROVI	1917	S.YAE1PPLQLIEL.NAYGMTIA.PGT....PDALSSLTEC.QSLREQ.....LQFN..VAGNH..QMRALAPQPKTRFSCRRA.LAYAATP...API 1997
APU_THETY	1165	P.TAF.V.LQQPGI.ESSRTVN.SPSA..DDVAIFGIEYK.SSSETGPf.....IKIAT..VSDSVY..NYVDTD..VNGNVVYYKV..WDTSYn...RTAS 1248
AXO1_RAT	914	PrRPP.GNISWTF..SSSSLSK.DPVVplrNESTVTGKLL.QNDLHPTptlhltksnWIEIP..VPEDIG..HALVQIRTTGPGGDGIP.AEVHIVRn..EGTS 1009
CHIT_STRLI	142	P.SAP.GTPASASNI.TDTSVKS.SAAT..DDKGKVNMD..LR...DGA.....KVAT..VTCTT..YTNDG..TKGAA.SYSRK..RDTADq..tGPAS 219
CPSF_CHICK	491	P.DPF.QSFRVTSV.GEDWAVS.EAPf.dGGMPITGILER.KKKGSMRW.....mKLNFE..VFPTD..TYESTK.WEGVYMEMRPF..VVNAIV..SQPS 577
CPSF_CHICK	784	P.GPF.QAVRMVEW.WGSNALQ.EPEKd..dgNAEISGTT.QK..ADTRTME.....WFTVL..EHSRPT..RCTVSELVMGNEMRFRG..SENVCF..t..SQEP 869
FAS2_SCHAM	530	P.SAV.LQVKMDVM.TATTVTFKFFGn..dGGLPTKNAQ..Q..KDQSQGW.....EDALN..RTWPVDS..PYILENLKFQ..RINFRFA..QONEVf..GPWS 616
FINC_BOVIN	689	P.VVA.TSESVTEI.TASSFV.SUSA..SDTVSGR..EY..ELSEEVGDe.....PQYLD..LPSTAT..SVNIPDLLPGRK..TVNMY..EISEE.....GEO 768
FINC_BOVIN	780	P.DAF.PDPTDVQV.DDTSVIR.SR..RAPITCR..V..SPSVEGS.....STEYN..LPETAN..SVTLSLDQPGVQVNITID..FVEEN..QES 858
FINC_BOVIN	1511	I.DKE.SQMQTVD.QDNSIS.R.LPS...SSPVTCR..TT..APKNGPGp.....SKTKT..VCPDQT..EMTIEGLQPTVE..VVSY..QON..GES 1590
GUNB_CELFI	651	P.TTP.GTPVATGV.TTVGAS.SAASTD..AGSGVAGHE..LR..VQGTTQ..TLVGT..TTAA..YILRDLI.PGAA..SYVTK..KDVA..n..vSAAS 733
I12B_HUMAN	235	P.DPF.KNLQLKPLkNSRQVE.S.EY..dt..WSTPHSYPS..TP..CVQVQGK..SKREK..KDRVFT..DKTSAT..ICRKNASIS.R..QDRYYs..SSWS 320
IDUA_CANFA	547	P.GPV.TRLRALPL.TRGQVL.V.SDERV..GSKCLWIV.EQ..SADGEV.....YTPIS..RKESTFn..LFVFSPEASV..SGSYR..FDYWAR..pGPFS 632
IDUA_HUMAN	548	P.GQV.TRLRALPL.TQGQVL.V.SDEHV..GSKCLWIV.EQ..SDQGKA.....YTPVS..RKESTFn..LFVFS PDTGAVGSYR..LDYWAR..pGPFS 633
IL7R_HUMAN	129	P.EAF.FDLSVIYRgCANDFVTNTSH1qkKYVVKLMD..IA..YRQEKDENK..WTHVN..LSSSTKL..TLLQRK..QEAAM..EIKER..SIPDHYfkgf..fwSEWS 221
ITB4_HUMAN	1127	L.GAF.QNPNAKAA.GSRKIHFN.LP..SGKPMGTR.KY..WIQGDSEs.....EAHLL..DSKVP..SVELTN..YFYCD..EMKTC..YGAQ..e..SPYS 1208
ITB4_HUMAN	1581	P.DTP.TRLVFSAL.GPTSLR.S.QE.R..CERPLQGNS.EY..QLLNGGE..LHRLN..IPNPAQt..SVVVED..L..PNHSt..VFRFR..HQSQE..w..GRER 1665
KECK_HUMAN	436	Q.TEF.PKVRLEGR.STTSLS.S.SI..pp..QQSRVWKHE..TYR..KKGDS.....NSYN..VRTERGF..SVTLDD..APD..T..LVQ..Q..LTQE..q..PAGS 519
KEK5_CHICK	444	P.SAV.SIMHQVSR.TVDSIT.S.SQ..Dq..PNGVILD..E.Q..YEKNLSE..LNSTA..VKSPTN..TVTVQNLKAG..INVFO..R..RTVA..y..GRYS 528
KMLC_CHICK	60	P.DPFaGTPCASDI.RSSSLT.S.YGSSY..dGGSAVQSMT..EI..WNSVDNK..WTDLT..TCRST..SFNVQD..LQADRE..KFR..R..ANVY..i..SEPS 145
KSEK_MOUSE	441	P.SSI.ALVQAKEV.TRYSVAA..LEDr..PNGVILE..E..K..YEKDQN..ERSYR..IVRTAAr..NTDIKGLNPL..MVFHR..R..RTAA..y..GDFs 525
LAR_DROME	322	P.TAF.TDQVQISEV.TATSVR.E..SYK..GPEDLQY..V..Q..KPKNANQ..AFSEI..SGIITM..YYVVRALSPY..E..EFY..I..VVNNI..y..GPFS 404
MPSF_CHICK	371	P.GAP.MDVKCHDA.NRDYVINT..KPNP..t.SQNPVIGNE..DK..CEVGLEN..WVQCN..DAPFKIC..KYPVTGLEYGRS..IFR..R..VNSA..i..SRPs 457
NCA1_BOVIN	509	P.SSE.SIDQVPE..YSSTAQ..QFDE..ea..tGGVPILK..KKAER..R..AMGEEVw..hSKWYD..AKEASMegIVTIVGLK..F..T..AVRLA..NLNGK..l..GEIS 597
NRCA_CHICK	928	P.SPF.SFLKITNP..TLDLSL..E..GS..Th..PNGVLTSI..K..QPINNThel..gpLVEIR..IPANES..SLILKNLNY..S..R..MKFYFN..QTSV..SGS 1014
PHB_ALCFA	344	G.SAP.TGLAVTAT.TSTSVD.S..NAV..ANASSG..LR..NGS..KVGS..ATATA..YTDSGI..IAGT..T..SYT..T..FDPTAg..eSQPS 418
PTP1_DROME	123	P.DPF.SNLSVQVR.SGKNAII..L..SPIT..QGSYTAHK..KV..LGLSEASSs..yNRTFQ..VNDNTF..QHSVKE..I..P..GAT..QVQAY..TIYDG..KES 205
PTP_HUMAN	554	P.AQV.TDLHVANQgMTSSLFTN..TQA..QGDVEF..Q..LL..IHENVV..IKNES..ISSETS..RYSFHS..KSGSL..SVV..T..TVSGG..ISS 632
PTPK_MOUSE	290	P.PRFIAAPPQLLGV.GPTYLL..QLNANSI..i.GDGPII..K..E..R..MT..SGS..WTETH..AVMAP..TYKLWHL..DFD..E..E..IR..L..I..TRPGE..g..t..OLPG 376
TENA_CHICK	593	V.SPF.TELTVTNV.TDKTVN..E..KHE..NLVNE..L..T..VPTSSGG1..DLQFT..VPGNQT..SATIHE..L..E..PGV..E..F..IR..F..H..ILKN..KKS 671
TIE1_HUMAN	446	P.PVP1AAPRLLTK.QSRQLV..SPLVFSFs..GDGPISTVR..H..RPQDSTM..WSTIV..VDBSE..NVTLMNLRFK..G..SVR..Q..SRPGE..g..eGAWG 533
TIE2_HUMAN	444	L.PKF1NAPNVIDT.GHNFAV..NISSEPY..fGDGPIKSKK..L..I..KPVNHYEA..WQHIQ..VTNEI..V..T..V..L..N..Y..L..P..R..E..M..LCWQ..LVR..R..GEGH 529
TIE2_HUMAN	639	P.PQP.ENIKISNI.THSSAV..S..TILD..GYSISIT..R..KVQGKNE..DQHVDV..K..K..N..T..I..I..Q..Y..Q..L..K..G..L..E..P..E..A..Q..V..D..F..J..ENN..I..s..SNPA 724
UFO_HUMAN	327	L.GPF.ENISATR..NGSQAF..H..Q..E..R..a..pLQGTLG..R..A..Q..Q..Q..Q..D..T..P..E..V..L..V..M..D..I..G..L..R..Q..E..V..T..L..E..L..Q..G..D..G..S..V..N..L..T..V..C..A..Y..T..A..d..GPWS 411



IDUA_HUMAN:

Glycohydrolase

FN3

b

a

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

AGRI_CHICK	154	CVPAS.....	CS.....GVa.ESI VCGSDGKDYRS CDUNKHAC.....DK.....QENVFKKKRDGAC	201
AGRI_RAT	165	CLQPTT.....	CF.....GAp.DGT VCGSDGVDPSECQ LSHAC.....AS.....QEHLFKKENGPC	212
FSA_HUMAN	116	CVCAPD.....	CS.....NItwKGPVCGLDG TYRN ECA LKARC.....KE.....QPELEVQYQGRC	164
FSA_PIG	116	CVCAPD.....	CS.....NItwKGPVCGLDG TYRN ECA LKARC.....KE.....QPELEVQYQGKC	164
FSA_RAT	116	CVCAPD.....	CS.....NItwKGPVCGLDG TYRN ECA LKARC.....KE.....QPELEVQYQGKC	164
FSA_SHEEP	109	CVCAPD.....	CS.....NItwKGPVCGLDG TYRN ECA LKARC.....KE.....QPELEVQYQGKC	157
IAC1_BOVIN	14	CKVYTEA.....	CT.....RE..YNP ICDSAAKTY SNECTF ...CNEKM.NN.....DADLHFNFHFGEC	61
IAC2_BOVIN	7	CAEFKDP.....	KVYCT.....RE..SNPHCG NG TYGNKCAF.....CKAVM. S.....GGK NLKHRGKC	57
IACA_PIG	7	CNVYRSH.....	LFFCT.....RQ..MDPICG NG SYANPCIF.....CSEKG.LR.....NQKFDFGHWGHC	57
IACS_PIG	12	CDVYRSH.....	LFFCT.....RE..MDPICG NG SYANPCIF.....CSEKL.GR.....NEKFDFGHWGHC	62
IAC_MACFA	33	CARYQLPG.....	CRD..FNPVCG DMITYPNECT ...CMKIR. S.....GQN KILRRGPC	81
IOV7_CHICK	94	CSPYLQVVRDGntMVAC	RI..LKPVCG SDSFTYDN E CCAYNA. H.....HTN SKL DGE	150
IOVO_ABUPI	8	CSDHPKP.....	ACL.....QE..QKPLCG SDN TYDNKCSF.....CNAV V.DS.....NGT TLSHFGKC	56
IOVO_ALECH	6	CSEYPKP.....	ACT.....LE..YRPLCG SDS TYGNKCNF.....CNAV V. S.....NGT TLSHFGKC	54
IPSG_VULVU	68	OTEYSDM.....	CT.....MD..YRPLCG SDGKNY SNK CIF.....CNAV V. S.....RGT FLAKHGE	115
IPST_ANGAN	12	CGEMSAMHA.....	CMN..FAPVCG DGNTY PN CCFQRQ. NT.....KTD LITKDDRC	61
IPST_BOVIN	9	CTNEVNG.....	CRI..YNP VCG DGVTY SNECL ...CMENK. R.....QTPVLIQKSGPC	56
IPST_PIG	9	CTSEVSG.....	CKI..YNP VCG DGITY SNECV ...CSENK. R.....QTPVLIQKSGPC	56
IPST_SHEEP	9	CTNEVNG.....	CRI..YNP VCG DGVTY ANECL ...CMENK. R.....QTPVLIQKSGPC	56
OATP_HUMAN	439	CNVDCN.....	CS KI..WDPVCGNNGLSYLSACLA ...GC ..ET.SI.....GTG NMV QNCS	485
OATP_RAT	439	CNTRCS.....	CS.....Tnt..WDPVCGDNGVAYMSACLA ...GCKKFV.GT.....GTN V FQDCSC	486
PE60_PIG	37	CEHMTESPD.....	CS.....RI..YDPVCG DGVTY ESCK ...CLARI. N.....KQD QIVKDGE	86
PGT_RAT	444	CRRDCS.....	CDSf.FHPVCG DNGVEYV SPCHA ...GCSS.....TNTSSEASKEPI	488
PSG1_MOUSE	33	CHDAVAG.....	CRI..YDPVCG DGITY ANECL ...CFENR. R.....IEPV LIRKGGPC	80
QR1_COTJA	466	CICQDPA.....	ACs..tKD..YKRVCG DN TYDGT CQ FGTKCQLEGt KK.....GRQ HLDMMGAC	521
SC1_RAT	424	CVCQDPET.....	Cp..aki..LDQACG DN TYASSCH FATKCMLEGt KK.....GHQ QLDWF GAC	479
SPRC_BOVIN	93	CVCQDP.TS.....	Cap..iGE..FEK VCSNDN TFDSSCHFF FATKCTLEGt KK.....GHK HLDY GPC	149
SPRC_CAEEL	74	CECISK.....	CeldgdP..MDKVCANN N FTSLCD YRERCLCKR KSk ecskafNAK HLEY GEC	135
SPRC_MOUSE	92	CVCQDP.TS.....	Cap..iGE..FEK VCSNDN TFDSSCHFF FATKCTLEGt KK.....GHK HLDY GPC	148
SPRC_XENLA	90	CVCQDPST.....	Cts.vGE..FEK ICG DN TYDSSCHFF FATKCTLEGt KK.....GHK HLDY GPC	146



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?

- $\frac{1}{3}$ of Swissprot proteins ($\frac{1}{5}$ of residues) had at least one domain in Pfam-A
- New members found for some families
- New domains annotated for some proteins
- $\frac{1}{5}$ 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%	

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%

Pfam: Then and now - FN3 domain (PF00041)

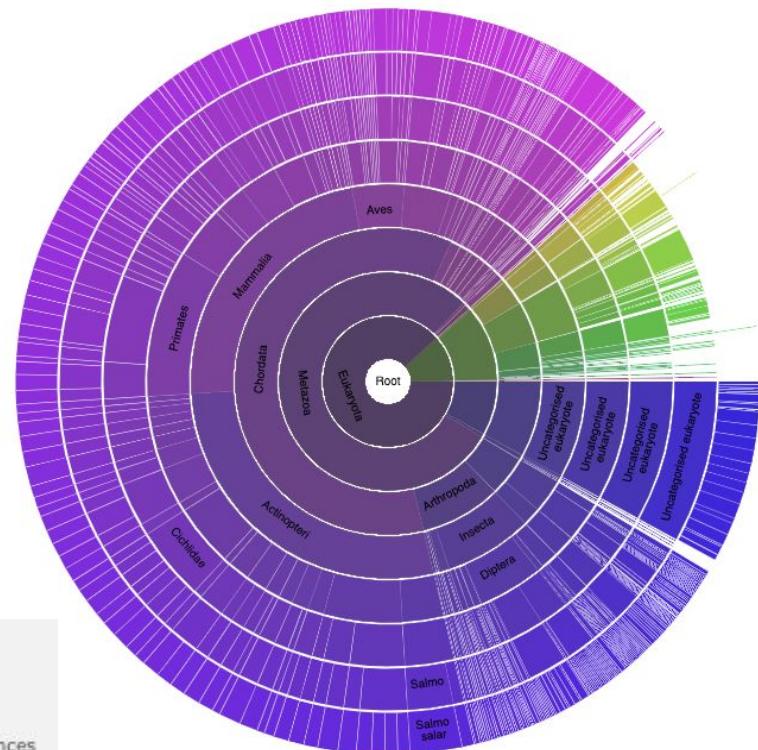
Pfam 1.0 (1997):

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





Assignment

Compare tools for protein domain identification

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

Feel free to work together!