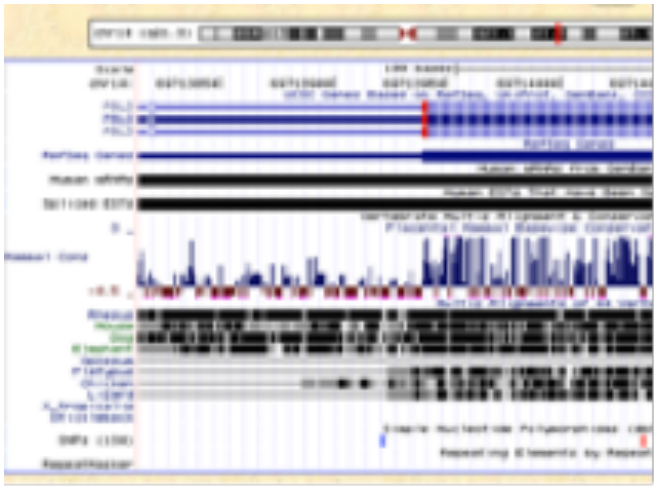
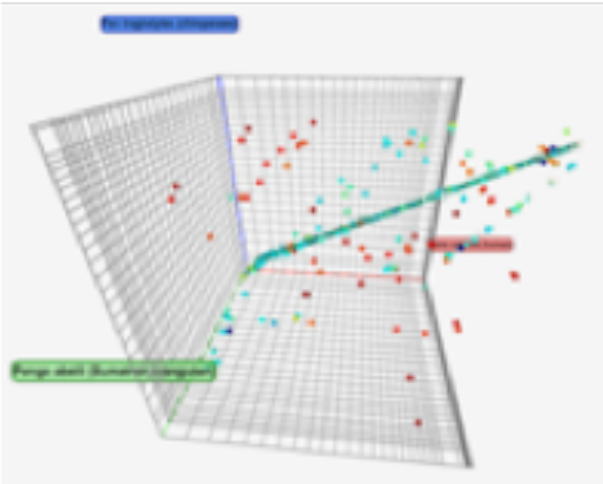
Computational Genomics

Introduction To Databases
Finding:
Genes, and Genomes
Proteins and Proteomes

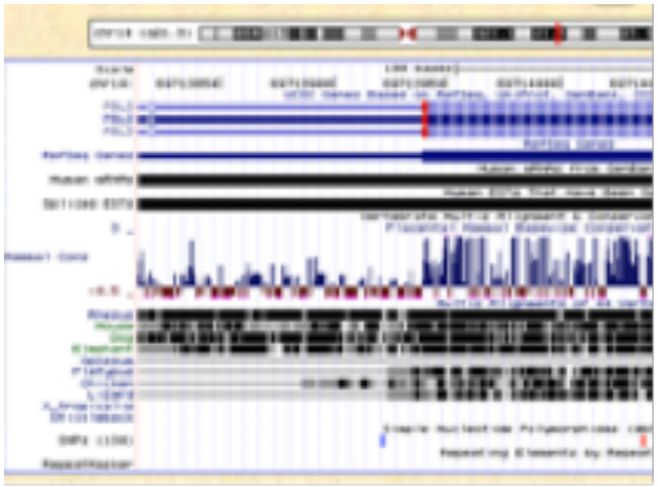


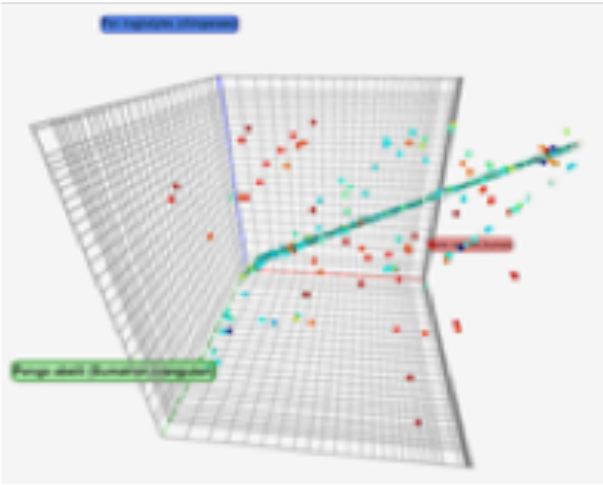


Computational Genomics

Introduction To Databases

Protein Databases





Protein Databases
Pfam

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

QUICK LINKS

SEQUENCE SEARCH Analyze your protein sequence for Pfam matches

VIEW A PFAM ENTRY View Pfam annotation and alignments

VIEW A CLAN See groups of related entries

VIEW A SEQUENCE Look at the domain organisation of a protein sequence

VIEW A STRUCTURE Find the domains on a PDB structure

KEYWORD SEARCH Query Pfam by keywords

JUMP TO enter any accession or ID [Go] [Example]

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Or view the help pages for more information

EMBL-EBI Training

Delivering world-class training in data-driven life sciences.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2016

Week 4: Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

Protein Databases

Pfam

Pfam entry types

- Pfam is a database of conserved evolutionary units. It can be used to explore domain(s) and other functional regions in a protein and give insight into the function of that protein.

- In general, Pfam entries are classified into six different categories, depending on the length and nature of the sequence regions included in the entry: family, domain, repeats, motifs, coiled-coil, and disordered, as shown in Figure 1 below. Each Pfam entry is represented by a set of aligned sequences with their probabilistic representation – called a profile hidden Markov model (HMM). A single profile HMM is often insufficient to model an entire diverse superfamily so evolutionary related entries are combined in sets called Clans.

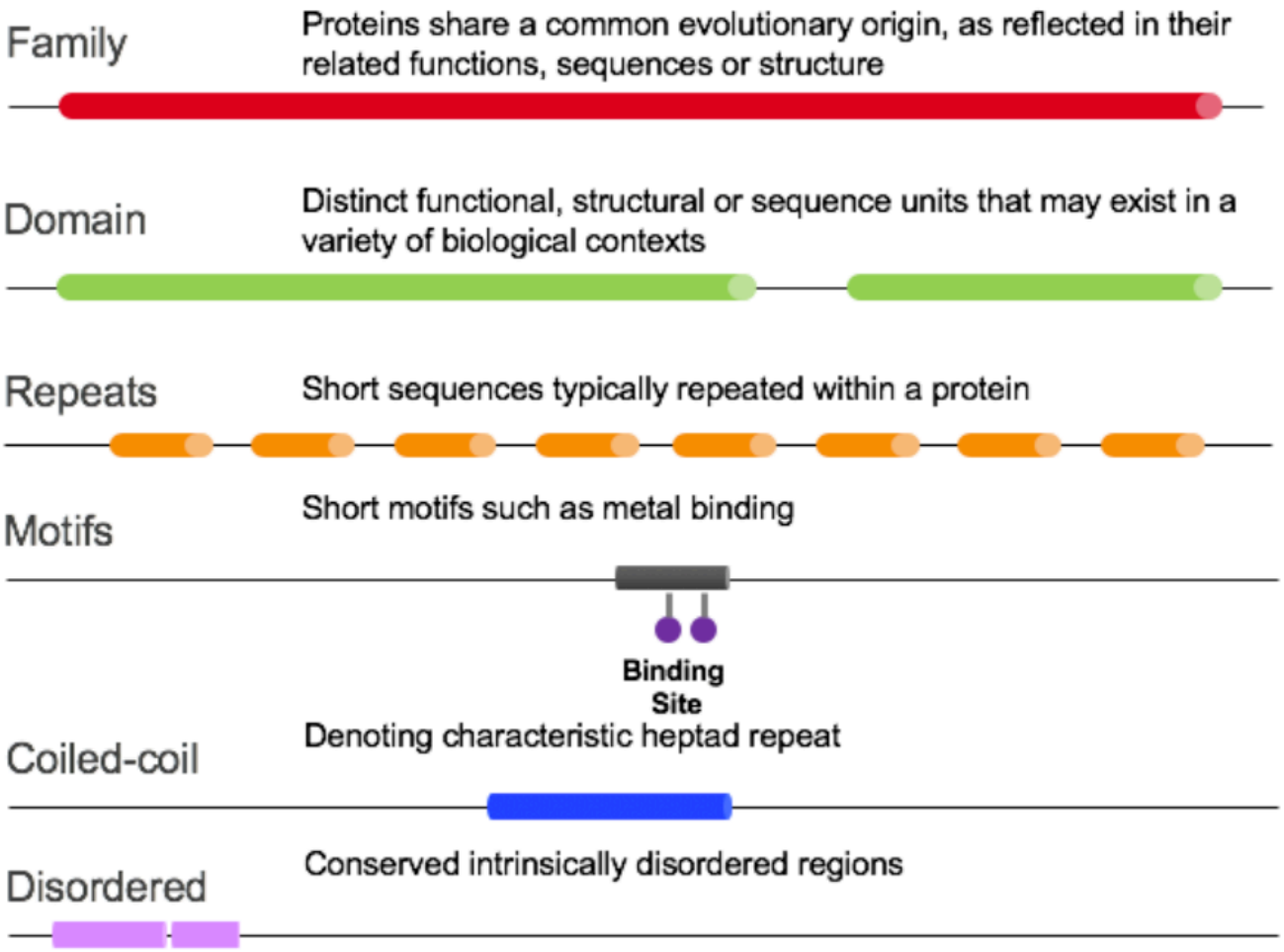


Figure 1 Categories of Pfam entries.

Protein Databases
Pfam
What are profile hidden Markov models?

- They are one of the computational algorithms used for predicting protein structure and function, identifies significant protein sequence similarities allowing the detection of homologs and consequently the transfer of information, i.e. sequence homology-based inference of knowledge. In this section we will describe the algorithm used to create Pfam entries: profile hidden Markov models (HMMs).
- Profile HMMs are probabilistic models that encapsulate the evolutionary changes that have occurred in a set of related sequences (i.e. a multiple sequence alignment). To do so, they capture position-specific information about how conserved each amino acid is in each column of the alignment, see Figure 2.
- The model also captures important information such as the varying degree to which gaps and insertions have occurred. Unlike other sequence homology detection algorithms, profile HMMs use position dependent gap penalties and substitution probabilities which better reflect biological reality.

The boxes in yellow are the match states (M). In the M state the probability distribution is the frequency of the amino acids in that position. The row of diamond shaped states are insert states (I) which are used to model highly variable regions in the alignment. The circular states are delete states (D). These are called silent states since they do not match any residues, and they are there merely to make it possible to jump over one or more columns in the alignment.The final probabilistic model conveys the estimation of the observed frequencies of the amino acids in each position, as well as the transitions between the amino acids derived from the observed occupancy of each position in a multiple sequence alignment.

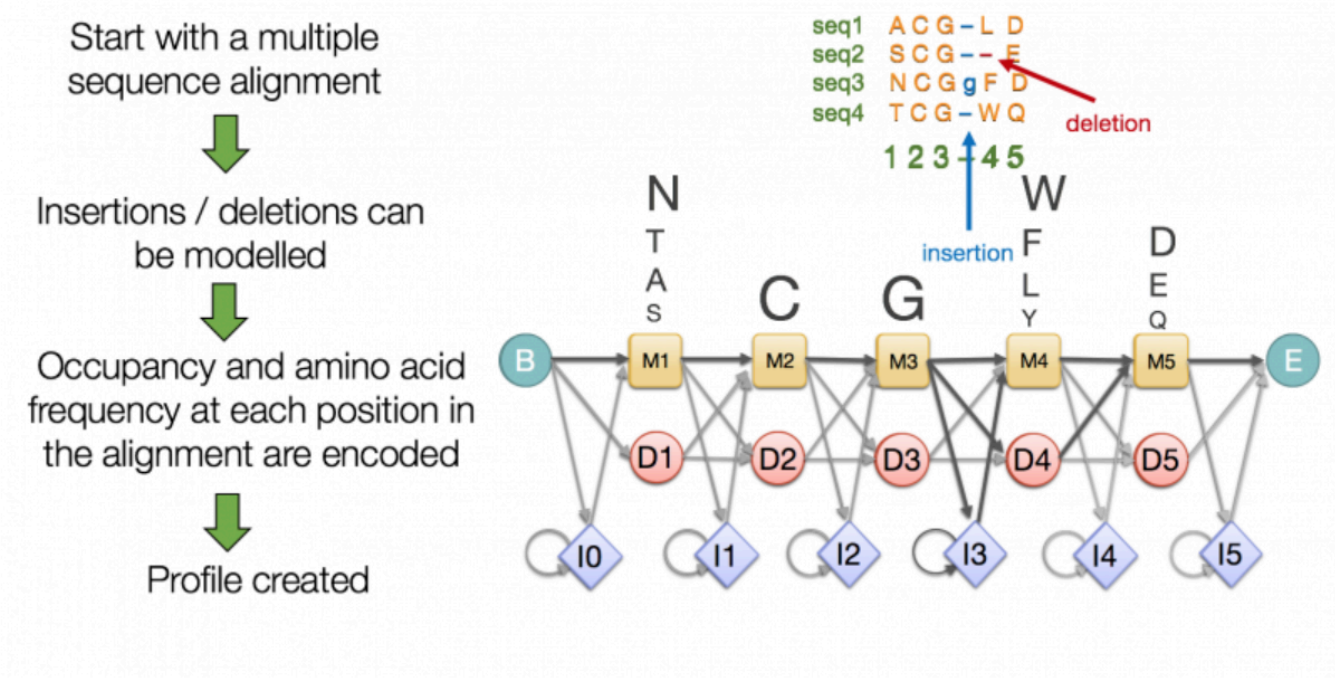


Figure 2 A profile HMM modelling a multiple sequence alignment.

Protein Databases
Pfam
Modelling in Pfam

- Creating a Pfam model is an iterative process. The starting point is the selection and alignment of curated example sequences (i.e. seed alignment), which is used to calculate a profile hidden Markov model (HMM), see panels 1 and 2 in Figure 3. This profile HMM is then used to search against a reference proteomes database to find additional matching members that pass the inclusion thresholds. Thresholds are adjusted to avoid inclusion of false positives.

- The information in this new set of sequences (full alignment) is used to improve the probabilities in the model which may then lead to a slightly different alignment, see panel 3 in Figure 3. This adjusted full alignment is refined (through the determination of boundaries and minimisation of redundancy) to produce a new seed, and a new profile HMM is generated. This iterative process is repeated until no more homologs are detected in the sequence database search (i.e. the search has converged).

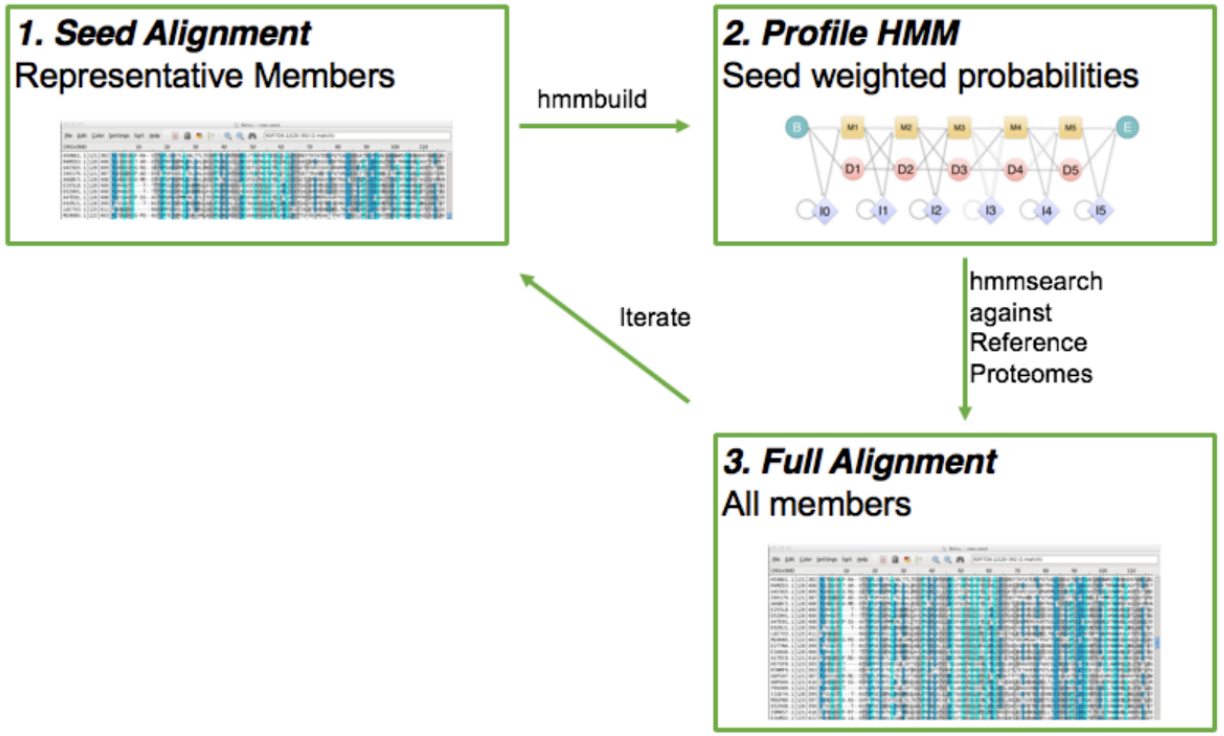


Figure 3 Creating a Pfam entry using HMM is an iterative process involving seed alignment, construction of profile HMM models and full alignment of sequences against a reference database.

Protein Databases
Pfam
Boundaries determination

- Part of the iterative process of building an entry involves trimming the full alignment to produce a new seed alignment for each round, a process which includes boundary determination. Figure 4 shows the difference between the full alignment and the resulting seed alignment after manually assigning boundaries. In Pfam, we adopt a number of different approaches to define the correct boundaries within the alignment:
    1. Comparing sequences to known protein structures
    2. Building models and altering the boundaries until an optimal solution is found
    3. Based on information derived from external resources such as literature and/or experts in the field

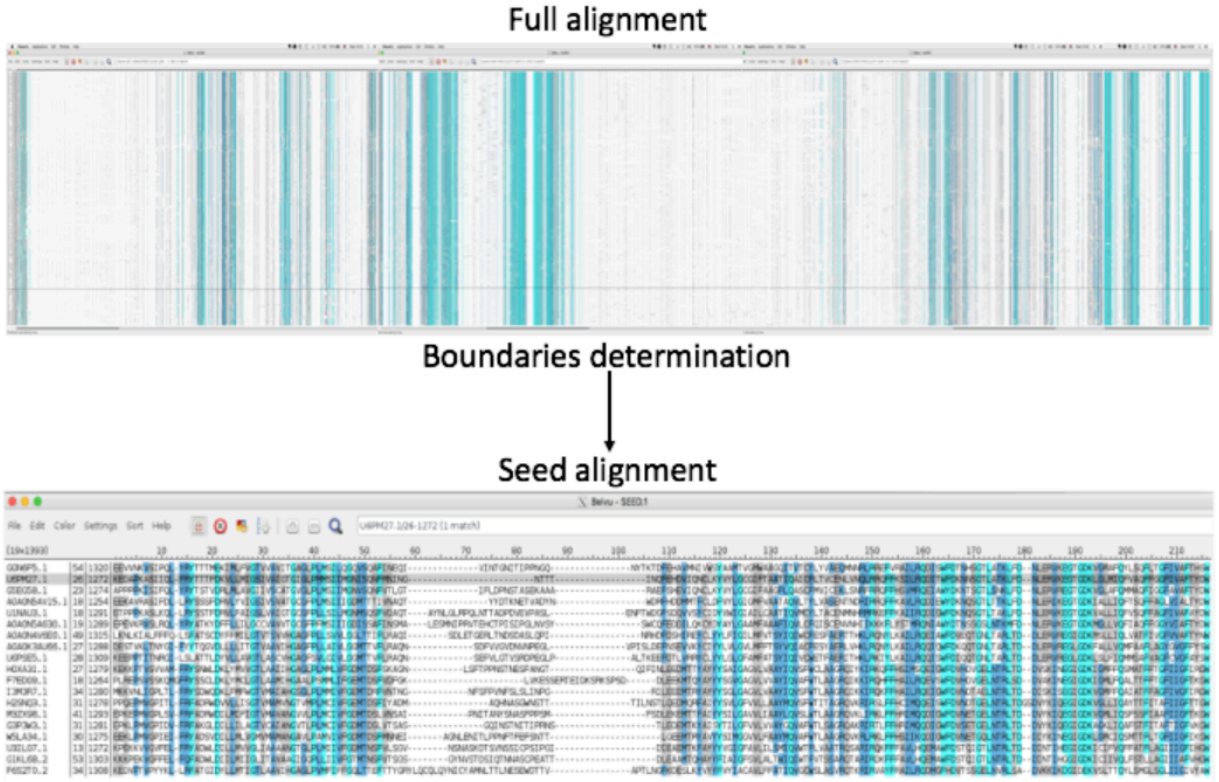
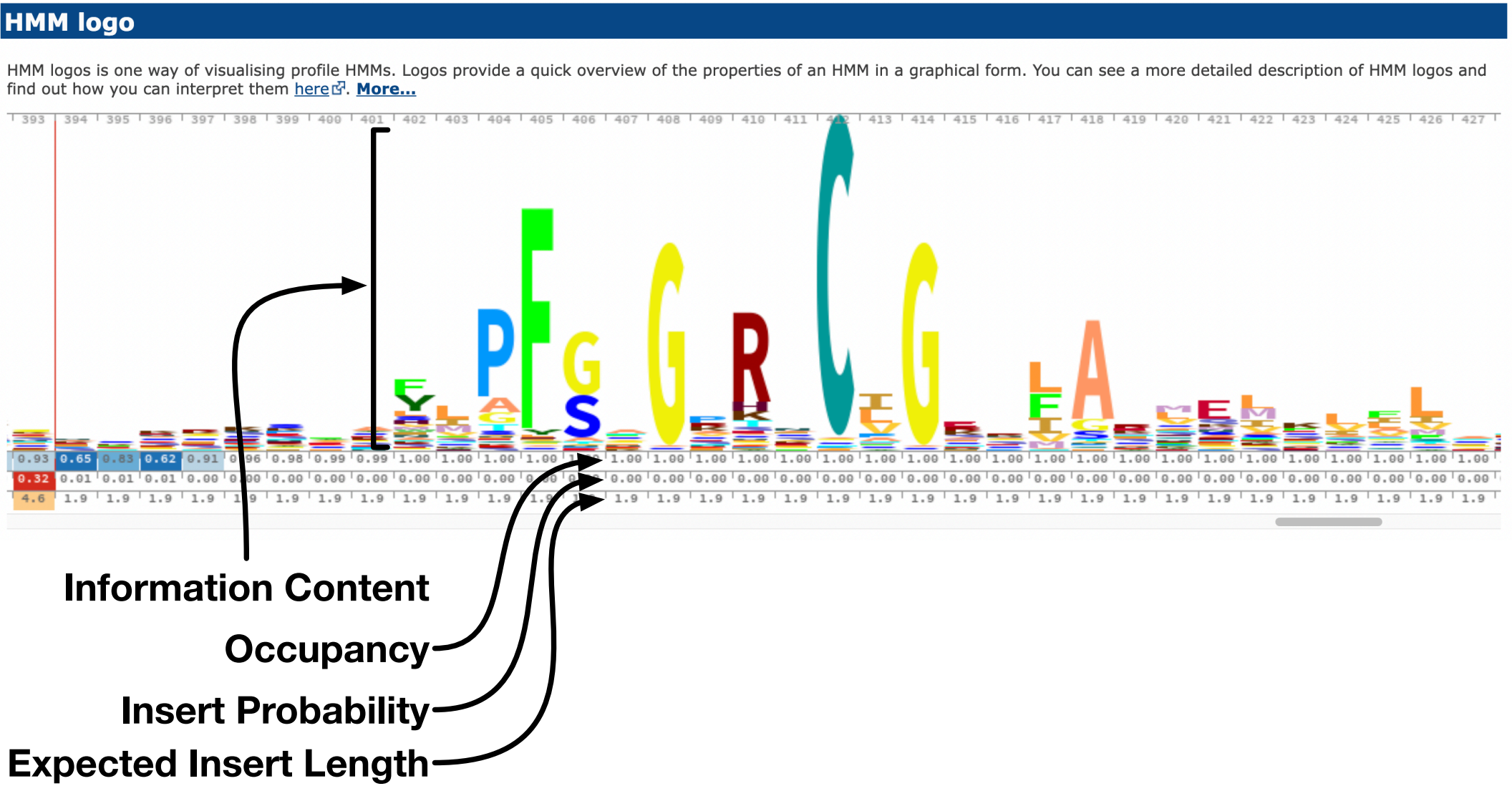


Figure 4 Full alignment analysis and boundaries determination, creating a new seed alignment.

Protein Databases

Pfam

HMM Logo



Protein Databases
Pfam
Threshold adjustment

- The software suite used to construct and search the models, called HMMER, reports two non-independent values that can be used to determine significance of a match; bit-scores and E-values.
- The bit-score is a measure of how well a sequence matches the profile HMM (i.e. if the sequence is homologous to the model) and is independent of the size of the database. E-values provide an indication of how likely it is that the given bit-score would have been achieved given the size of the database (i.e. it measures how statistically significant the bit-score is). The larger the database, the greater the chance that a match with the same bit-score could have occurred by chance and hence be a false positive. Therefore, the E-value changes with the size of the database.
- Two scores are reported for every match, one for the individual hit (domain score) and one for the sequence (sequence score). The sequence bit-score is the sum of all hit bit-scores found between the sequence and the profile HMM. Sequence scores (Figure 5) are calculated for every round of iteration and analysed manually. For most Pfam entries of type domain,
the sequence and hit thresholds defined are very similar (around 25.0 bit-score) as you do not typically find domains repeated within a sequence.

```
# ===========
#
# Sequence scores
# ---------------
#
# name           description                                    bits      evalue

C1ECX3.1         C1ECX3_MICCC UNCHARACTERIZED PROTEIN {ECO:     548.1     7e-163
K0RDT8.1         K0RDT8_THAOC CARBONIC ANHYDRASE {ECO:00003     269.1    2.9e-77
B8CG97.1         B8CG97_THAPS UNCHARACTERIZED PROTEIN {ECO:     259.8    1.9e-74
A6FY58.1         A6FY58_9DELT UNCHARACTERIZED PROTEIN {ECO:     259.0    3.5e-74
C1N5U2.1         C1N5U2_MICPC PREDICTED PROTEIN {ECO:000031     258.3    5.9e-74
F9ZEW7.1         F9ZEW7_9PROT CARBONIC ANHYDRASE, CADMIUM-B     216.0    5.3e-61
A0A0G1CI10.1     A0A0G1CI10_9BACT UNCHARACTERIZED PROTEIN {     186.7    5.2e-52
A0A0F7KI59.1     A0A0F7KI59_9PROT UNCHARACTERIZED PROTEIN {     132.7      2e-35
A0A0G1S0Q0.1     A0A0G1S0Q0_9BACT UNCHARACTERIZED PROTEIN {      93.4    2.3e-23
R1CFN9.1         R1CFN9_EMIHU UNCHARACTERIZED PROTEIN {ECO:      63.0    4.9e-14
A5KSD2.1         A5KSD2_9BACT UNCHARACTERIZED PROTEIN {ECO:      54.1    2.6e-11
R1EGY0.1         R1EGY0_EMIHU UNCHARACTERIZED PROTEIN {ECO:      51.6    1.5e-10
A0A0G1V756.1     A0A0G1V756_9BACT UNCHARACTERIZED PROTEIN {      46.8    4.4e-09
R4PXW2.1         R4PXW2_9BACT UNCHARACTERIZED PROTEIN {ECO:      38.6    1.4e-06
R1FN91.1         R1FN91_EMIHU UNCHARACTERIZED PROTEIN {ECO:      27.6     0.0035
A0A0C4E3I8.1     A0A0C4E3I8_MAGP6 UNCHARACTERIZED PROTEIN {      23.0      0.092
R7Z4F5.1         R7Z4F5_CONA1 UNCHARACTERIZED PROTEIN {ECO:      22.9      0.095
A0A0D5YUA3.1     A0A0D5YUA3_9FLAO XAA-PRO DIPEPTIDASE FAMIL      22.1       0.17
A0A0A3I6H2.1     A0A0A3I6H2_9BACI PHOSPHORIBOSYLFORMYLGLYCI      21.6       0.23
A0A094EWJ3.1     A0A094EWJ3_9PEZI UNCHARACTERIZED PROTEIN {      21.6       0.24
L8LR03.1         L8LR03_9CHRO CALX-BETA DOMAIN-CONTAINING P      21.5       0.25
J3P2K8.1         J3P2K8_GAGT3 UNCHARACTERIZED PROTEIN {ECO:      20.7       0.45
A0A0X8CQN5.1     A0A0X8CQN5_9BACL PHOSPHORIBOSYLFORMYLGLYCI      20.5       0.52
A0A150YJU0.1     A0A150YJU0_9BACI PHOSPHORIBOSYLFORMYLGLYCI      20.0       0.76
W4XGA1.1         W4XGA1_STRPU UNCHARACTERIZED PROTEIN {ECO:      19.9       0.77
A0A0L0QN47.1     A0A0L0QN47_VIRPA UNCHARACTERIZED PROTEIN {      19.8       0.85
A0A135T7T0.1     A0A135T7T0_9PEZI HEAVY METAL TRANSLOCATING      19.5        1.1
K0CCT8.1         K0CCT8_ALCDB PUTATIVE FERRODOXIN {ECO:0000      19.3        1.2
A0A0P0E6Q6.1     A0A0P0E6Q6_9MICO UNCHARACTERIZED PROTEIN {      19.1        1.4
A0A177C037.1     A0A177C037_9PLEO HEAVY METAL TRANSLOCATIN       18.8        1.7
A0A100YS74.1     A0A100YS74_9FIRM UNCHARACTERIZED PROTEIN {      18.4        2.2
A0A0L0BUT7.1     A0A0L0BUT7_LUCCU UNCHARACTERIZED PROTEIN {      18.3        2.4
E1VKJ6.1         E1VKJ6_9GAMM FERREDOXIN {ECO:0000313|EMBL:      18.1        2.9
D5GKS6.1         D5GKS6_TUBMM UNCHARACTERIZED PROTEIN {ECO:      18.1        2.8
Q5CA10.1         Q5CA10_ALCBS FERREDOXIN, 2FE-2S {ECO:00003      18.1        2.8
A0A094E209.1     A0A094E209_9PEZI UNCHARACTERIZED PROTEIN {      17.9        3.3
```

Figure 5 Example Pfam sequence scores output including Uniprot accession, Uniprot description, Bit Score, E-value.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2016

Week 1: Biological Sequence Analysis I

Andy Baxevanis, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

Sequence Comparisons

- Homology searches
  - Usually 'one-against-one': *BLAST, FASTA*
  - Allows for comparison of individual sequences against databases comprised of individual sequences
- Profile searches
  - Uses collective characteristics of a family of proteins
  - Search can be 'one-against-many': *Pfam, CDD*
    or 'many-against-one': *PSI-BLAST, DELTA-BLAST*

Profiles, Patterns, Motifs, and Domains

Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly related proteins

Profile Construction

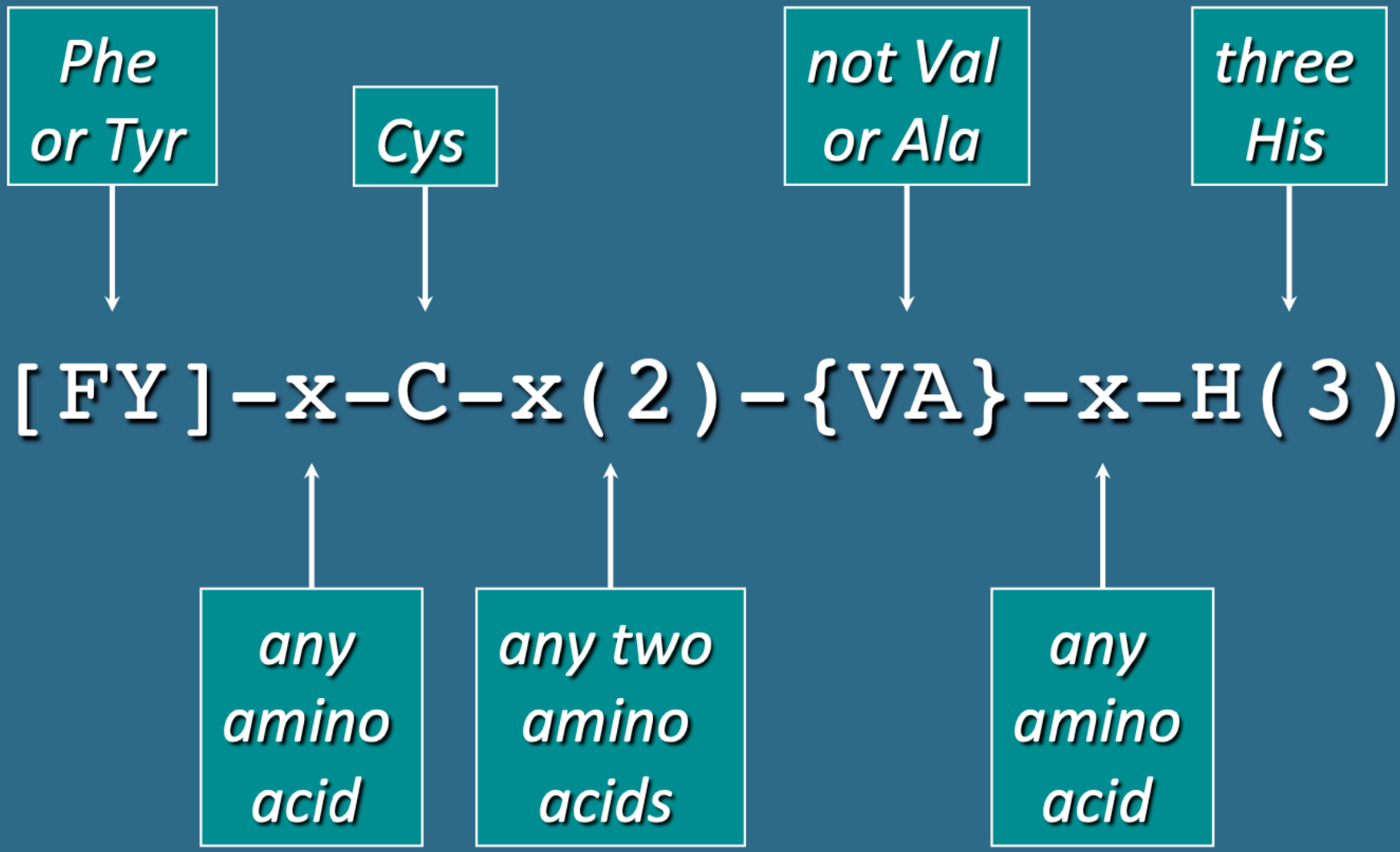
```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- *Which residues are seen at each position?*
- *What is the frequency of observed residues?*
- *Which positions are conserved?*
- *Where can gaps be introduced?*

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	41	19	11	9	6	16	11	9	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Patterns


Phe or Tyr
Cys
not Val or Ala
three His
[FY]-x-C-x(2)-{VA}-x-H(3)
any amino acid
any two amino acids
any amino acid

Pfam

- Collection of multiple alignments of protein domains and conserved protein regions that probably have structural, functional, or evolutionary importance
- Each Pfam entry contains:
  - Multiple sequence alignment of family members
  - Protein domain architectures
  - Species distribution of family members
  - Information on known protein structures
  - Links to other protein family databases

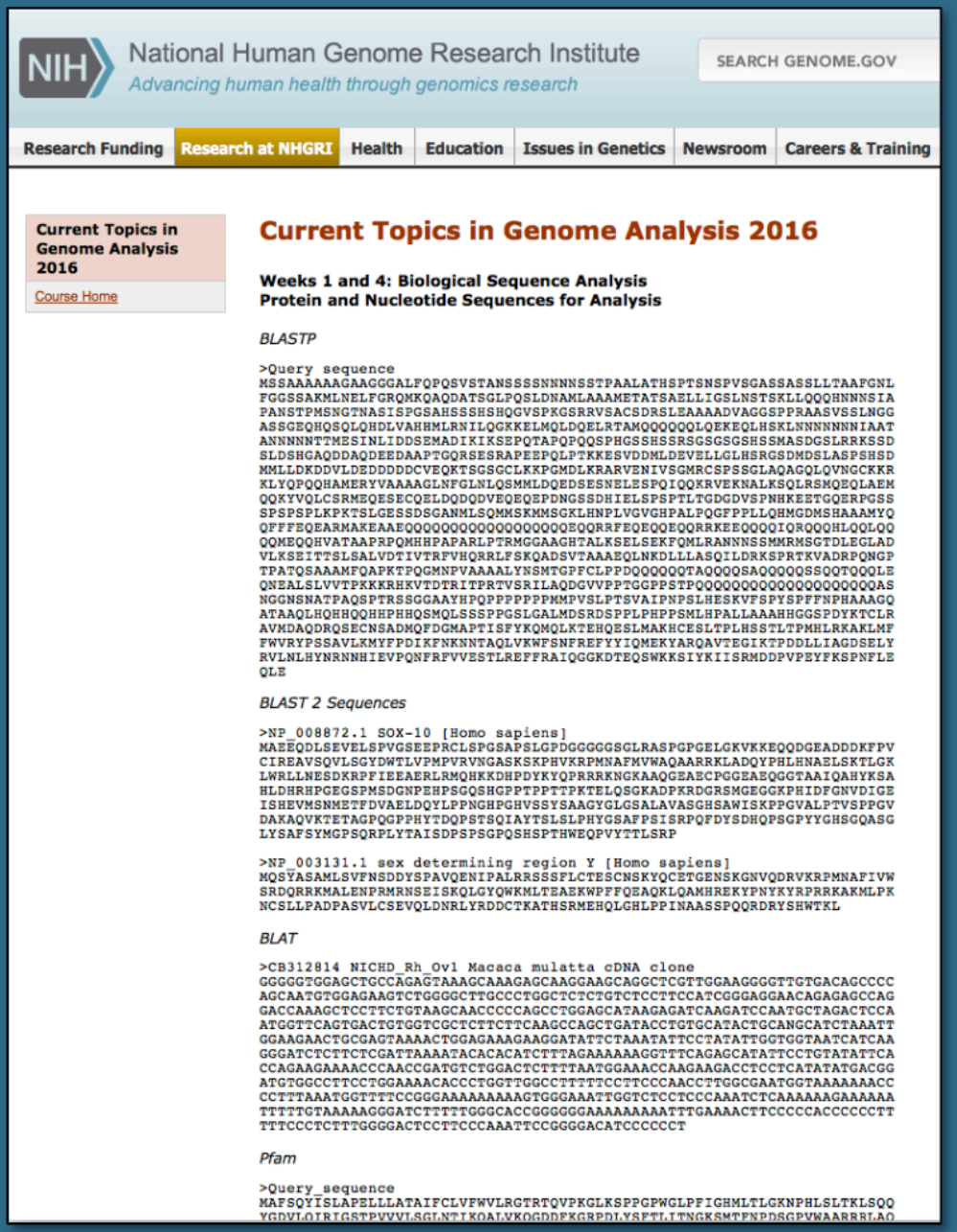
Finn et al., Nucleic Acids Res. 44: D279-D285, 2016

Pfam A

- Based on *curated* multiple alignments of known members of a protein family ('seed alignment')
  - *Pfam definition of 'family': a collection of related protein regions*
  - *Based on reference proteomes (UniProtKB)*
- HMMER used to find all detectable protein sequences belonging to the family
- New 'true members' of the family are then used to generate the 'full alignment' for the protein family
- Given the method used to construct the alignments, hits are highly likely to be true positives

Sequences Used in Examples

http://research.nhgri.nih.gov/teaching/seq_analysis.shtml



National Human Genome Research Institute
Advancing human health through genomics research
SEARCH GENOME.GOV

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training

Current Topics in Genome Analysis 2016
Course Home

Current Topics in Genome Analysis 2016

Weeks 1 and 4: Biological Sequence Analysis
Protein and Nucleotide Sequences for Analysis

BLASTP

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAAT
ANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSD
SLDSHGAQDDAQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKR
KLYQPQQHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQLAEM
QQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSPNHKEETGQERPGSS
SPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHPALPQGFPPLLQHMGDMSHAAAMYQ
QFFFEQEARMAKEAAEQQQQQQQQQQQQQQQQQQQEQQRRFEQEQQEQQQRRKEEQQQQIQRQQQHLQQLQQ
QQMEQQHVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLEGLAD
VLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRKSPRTKVADRPQNGP
TPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSAQQQQQSSQQTQQQLE
QNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQAS
NGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQ
ATAAQLHQHHQQHHPHHQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLR
AVMDAQDRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMF
FWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAGDSELY
RVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSPNFLE
QLE
```

BLAST 2 Sequences

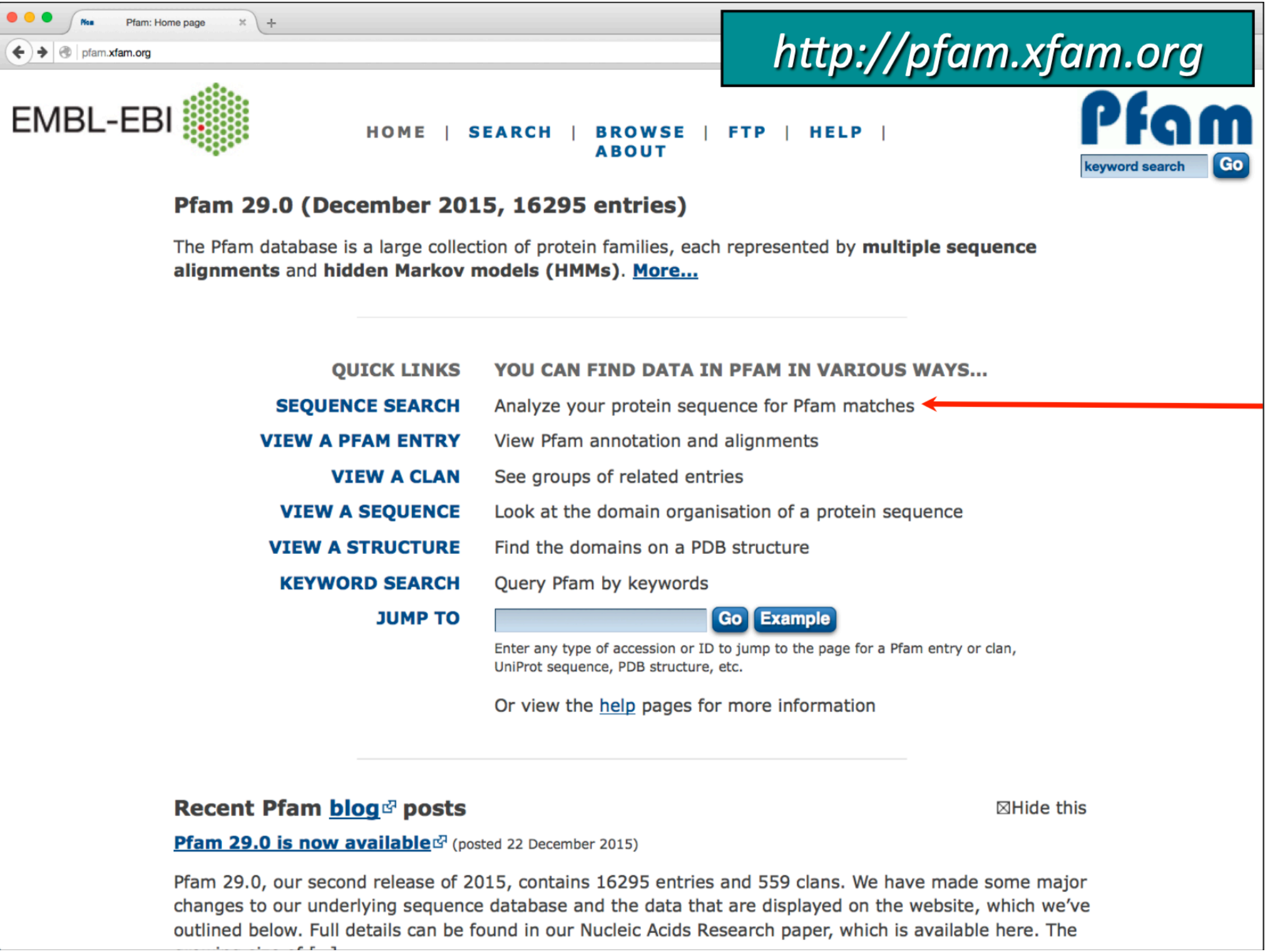
```
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVALPTVSPPGV
DAKAQVKTETAGPQGPPHYTDQPSTSQIAYTSLSLPHYGSAFPSISRPQFDYSDHQPSGPYYGHSGQASG
LYSAFSYMGPSQRPLYTAISDPSPSGPQSHSPTHWEQPVYTTLSRP

>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
```

BLAT

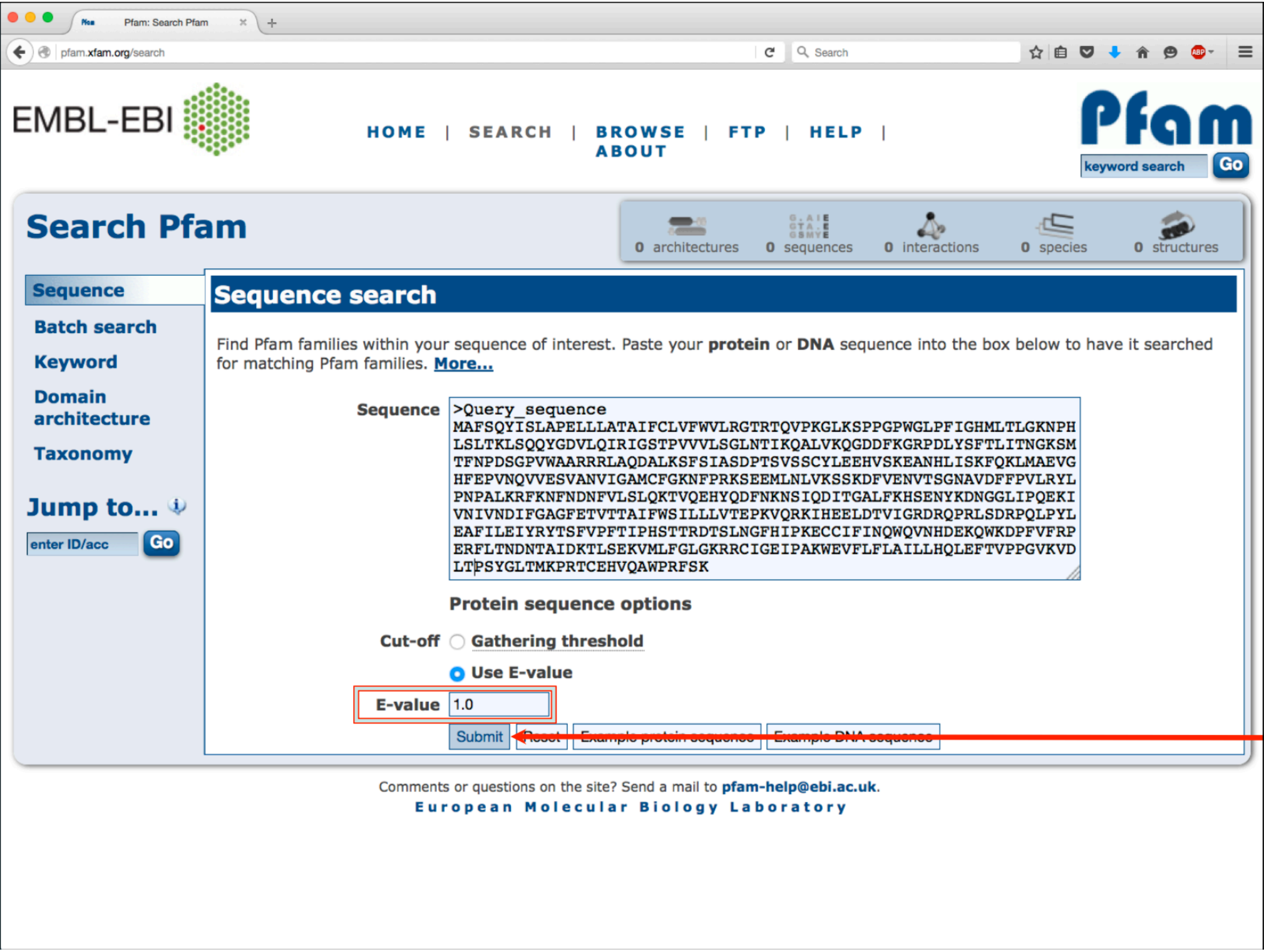
```
>CB312814 NICHD_Rh_Ov1 Macaca mulatta cDNA clone
GGGGGTGGAGCTGCCAGAGTAAAGCAAAGAGCAAGGAAGCAGGCTCGTTGGAAGGGGTTGTGACAGCCCC
AGCAATGTGGAGAAGTCTGGGGCTTGCCCTGGCTCTCTGTCTCCTTCCATCGGGAGGAACAGAGAGCCAG
GACCAAAGCTCCTTCTGTAAGCAACCCCCAGCCTGGAGCATAAGAGATCAAGATCCAATGCTAGACTCCA
ATGGTTCAGTGACTGTGGTCGCTCTTCTTCAAGCCAGCTGATACCTGTGCATACTGCANGCATCTAAATT
GGAAGAACTGCGAGTAAAACTGGAGAAAGAAGGATATTCTAAATATTCCTATATTGGTGGTAATCATCAA
GGGATCTCTTCTCGATTAAAATACACACATCTTTAGAAAAAAGGTTTCAGAGCATATTCCTGTATATTCA
CCAGAAGAAAACCCAACCGATGTCTGGACTCTTTTAATGGAAACCAAGAAGACCTCCTCATATATGACGG
ATGTGGCCTTCCTGGAAAACACCCTGGTTGGCCTTTTCCTTCCCAACCTTGGCGAATGGTAAAAAAACC
CCTTTAAATGGTTTTCCGGGAAAAAAAAAGTGGGAAATTGGTCTCCTCCCAAATCTCAAAAAAGAAAAAA
TTTTTGTAAAAAGGGATCTTTTTGGGCACCGGGGGGAAAAAAAAATTTGAAAACTTCCCCCACCCCCCTT
TTTCCCTCTTTGGGGACTCCTTCCCAAATTCCGGGGACATCCCCCCT
```

Pfam

```
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPHLSLTKLSQQ
YGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTFNPDSGPVWAARRRLAQ
```


Pfam: Home page
pfam.xfam.org
http://pfam.xfam.org
EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
Pfam 29.0 (December 2015, 16295 entries)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH
Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY
View Pfam annotation and alignments
VIEW A CLAN
See groups of related entries
VIEW A SEQUENCE
Look at the domain organisation of a protein sequence
VIEW A STRUCTURE
Find the domains on a PDB structure
KEYWORD SEARCH
Query Pfam by keywords
JUMP TO
Go
Example
Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.
Or view the help pages for more information
Recent Pfam blog posts
Hide this
Pfam 29.0 is now available (posted 22 December 2015)
Pfam 29.0, our second release of 2015, contains 16295 entries and 559 clans. We have made some major changes to our underlying sequence database and the data that are displayed on the website, which we've outlined below. Full details can be found in our Nucleic Acids Research paper, which is available here. The


Pfam: Home page
pfam.xfam.org
Search
EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
Pfam 29.0 (December 2015, 16295 entries)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
SEQUENCE SEARCH
VIEW A PFAM ENTRY
VIEW A CLAN
VIEW A SEQUENCE
VIEW A STRUCTURE
KEYWORD SEARCH
JUMP TO
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES
Paste your protein sequence here to find matching Pfam entries.
Go
Example
This search will use and an E-value of 1.0. You can set your own search parameters and perform a range of other searches here.
Recent Pfam blog posts
Hide this
Pfam 29.0 is now available (posted 22 December 2015)
Pfam 29.0, our second release of 2015, contains 16295 entries and 559 clans. We have made some major changes to our underlying sequence database and the data that are displayed on the website, which we've outlined below. Full details can be found in our Nucleic Acids Research paper, which is available here. The growing size of [...]
Moving to xfam.org (posted 1 May 2014)


Pfam: Search Pfam
pfam.xfam.org/search
Search
EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
Search Pfam
0 architectures
0 sequences
0 interactions
0 species
0 structures
Sequence
Batch search
Keyword
Domain architecture
Taxonomy
Jump to...
enter ID/acc
Go
Sequence search
Find Pfam families within your sequence of interest. Paste your protein or DNA sequence into the box below to have it searched for matching Pfam families. More...
Sequence
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPH
LSLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSM
TFNPDSGPVWAARRRLAQDALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVG
HFEPVNQVVESVANVIGAMCFGKNFPRKSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYL
PNPALKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQDITGALFKHSENYKDNGGLIPQEKI
VNIVNDIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIGRDRQPRLSDRPQLPYL
EAFILEIYRYTSFVPFTIPHSTTRDTSLNGFHIPKECCIFINQWQVNHDEKQWKDPFVFRP
ERFLTNDNTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFTVPPGVKVD
LTPSYGLTMKPRTCEHVQAWPRFSK
Protein sequence options
Cut-off
Gathering threshold
Use E-value
E-value
1.0
Submit
Reset
Example protein sequence
Example DNA sequence
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

EMBL-EBI

Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant)

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
p450	Cytochrome P450	Domain	n/a	41	505	41	500	1	**457**	463	344.0	1.1e-102	n/a	Show

Sequence search results

Hide the detailed description of this results page.

Below are the details of the matches that were found. We separate Pfam-A matches into two tables, containing the significant and insignificant matches. A significant match is one where the bits score is greater than or equal to the gathering threshhold for the Pfam domain. Hits which do not start and end at the end points of the matching HMM are **highlighted**.

The Pfam graphic below shows only the **significant** matches to your sequence. Clicking on any of the domains in the image will take you to a page of information about that domain.

Pfam does not allow any amino-acid to match more than one Pfam-A family, unless the overlapping families are part of the same clan. In cases where two members of the same clan match the same region of a sequence, only one match is show, that with the lowest E-value.

A small proportion of sequences within the enzymatic Pfam families have had their active sites experimentally determined. Using a strict set of rules, chosen to reduce the rate of false positives, we transfer experimentally determined active site residue data from a sequence within the same Pfam family to your query sequence. These are shown as "Predicted active sites". Full details of Pfam active site prediction process can be found in the accompanying paper.

For Pfam-A hits we show the alignments between your search sequence and the matching HMM. You can show individual alignments by clicking on the "Show" button in each row of the result table, or you can show all alignments using the links above each table.

This alignment row for each hit shows the alignment between your sequence and the matching HMM. The alignment fragment includes the following rows:

#HMM:	consensus of the HMM. Capital letters indicate the most conserved positions
#MATCH:	the match between the query sequence and the HMM. A '+' indicates a positive score which can be interpreted as a conservative substitution
#PP:	posterior probability. The degree of confidence in each individual aligned residue. 0 means 0-5%, 1 means 5-15% and so on; 9 means 85-95% and a '*' means 95-100% posterior probability
#SEQ:	query sequence. A '-' indicate deletions in the query sequence with respect to the HMM. Columns are coloured according to the posterior probability

0% 100%

You can bookmark this page and return to it later, but please use the URL that you can find in the "Search options" section below. Please note that old results may be removed after **one week**.

We found **1** Pfam-A match to your search sequence (**all** significant)

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope Start
p450	Cytochrome P450	Domain	n/a	41

```
#HMM      Ppgptplplvgnllqlgrkeelhevlrklgkkygpifrlklgskpvvvlsgpeavkevlikkgeefsgrpdeallatsrkpfkgkgvlfang..ekwkklRrftptltsf.......gkl.sleelveeeaedlveklrkkagesseldite
#MATCH    Ppgp  lp++g++l lg  +++h  l+kl+++yg+++++++gs+pvvvlsg + +k++l+k+g++f+grpd   ++     ++gk++ f+    + w   Rr+   +l sf        +   lee v +ea+ l+ k++k   e + +++++
#PP       89999***********..9**************************************97755555...588999988875556***************99999886779***************************
#SEQ      PPGPWGLPFIGHMLTLG-KNPHLSLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFT--LITNGKSMTFNPDsgPVWAARRRLAQDALKSFsiasdptsVSSCYLEEHVSKEANHLISKFQKLMAEVGHFEPVNQVV
```

Family: *p450* (PF00067)

455 architectures | 41973 sequences | 4 interactions | 929 species | 1275 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Cytochrome P450 | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Cytochrome P450

Provide feedback

Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

Example structure

PDB entry 4C9P: Structure of camphor bound T260A mutant of CYP101D1

View a different structure: 4C9P

Literature references

1. Graham-Lorence S, Amarneh B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. PUBMED:7549871 EPMC:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 EPMC:8405421
3. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; , DNA Cell Biol 1993;12:1-51.: The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. PUBMED:7678494 EPMC:7678494
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 EPMC:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 EPMC:3304150
6. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272 EPMC:11178272

External database links

HOMSTRAD:	p450
PRINTS:	PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE:	PDOC00081
SCOP:	2cpp

Family: *p450* (PF00067)

455 architectures | 41973 sequences | 4 interactions | 929 species | 1275 structures

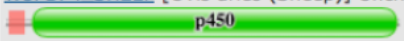
Jump to... enter ID/acc Go

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. More...

There are 34971 sequences with the following architecture: p450

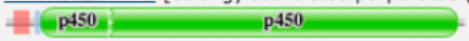
W5PB74_SHEEP [Ovis aries (Sheep)] Uncharacterized protein {ECO:0000313|Ensembl:ENSOARP00000007685} (494 residues)



Show all sequences with this architecture.

There are 2629 sequences with the following architecture: p450 x 2

W4ZE85_STRPU [Strongylocentrotus purpuratus (Purple sea urchin)] Uncharacterized protein {ECO:0000313|EnsemblMetazoa:SPU_026477-tr} (575 residues)



Show all sequences with this architecture.

There are 252 sequences with the following architecture: p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1

Q89R90_BRADU [Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)] Blr2882 protein {ECO:0000313|EMBL:BAC48147.1} (1078 residues)

Show all sequences with this architecture.

There are 152 sequences with the following architecture: p450 x 3

M4FE06_BRARP [Brassica rapa subsp. pekinensis (Chinese cabbage) (Brassica pekinensis)] Uncharacterized protein {ECO:0000313|EnsemblPlants:Bra039327.1-P} (983 residues)

Show all sequences with this architecture.

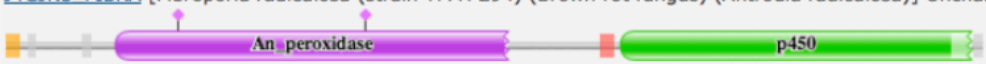
There are 93 sequences with the following architecture: An_peroxidase x 2, p450

W7MZF6_GIBM7 [Gibberella moniliformis (strain M3125 / FGSC 7600) (Maize ear and stalk rot fungus) (Fusarium verticillioides)] Prostaglandin-endoperoxide synthase 1 {ECO:0000313|EMBL:EWG53184.1} (1101 residues)

Show all sequences with this architecture.

There are 68 sequences with the following architecture: An_peroxidase, p450

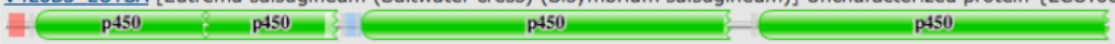
J4GJN3_FIBRA [Fibroporia radiculosa (strain TFFH 294) (Brown rot fungus) (Antrodia radiculosa)] Uncharacterized protein {ECO:0000313|EMBL:CCL99225.1} (1228 residues)



Show all sequences with this architecture.

There are 28 sequences with the following architecture: p450 x 4

V4L6B3_EUTSA [Eutrema salsugineum (Saltwater cress) (Sisymbrium salsugineum)] Uncharacterized protein {ECO:0000313|EMBL:ESQ39189.1} (1387 residues)



Show all sequences with this architecture.

There are 15 sequences with the following architecture: p450, Fer2

X5EDG3_9CORY [Corynebacterium glyciniphilum AJ 3170] Cytochrome P450 {ECO:0000313|EMBL:AHW64656.1} (774 residues)

Show all sequences with this architecture.

Family: *p450* (PF00067)

455 architectures | 41973 sequences | 4 interactions | 929 species | 1275 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database (reference proteomes) using the family HMM. We also generate alignments using four representative proteomes (RP) sets, the UniProtKB sequence database, the NCBI sequence database, and our metagenomics sequence database. **More...**

View options

We make a range of alignments for each Pfam-A family. You can see a description of each **above**. You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (50)	Full (41973)	Representative proteomes				UniProt (105935)	NCBI (141176)	Meta (2644)
			RP15 (9588)	RP35 (24353)	RP55 (37142)	RP75 (44573)			
Jalview	✓	✓	✓	✓	✓	✓	✓	✓	✓
HTML	View	–	×	×	×	×	×	×	×
PP/heatmap	×[1]	–	×	×	×	×	×	×	×

[1]Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, × not generated, – not available.

Format an alignment

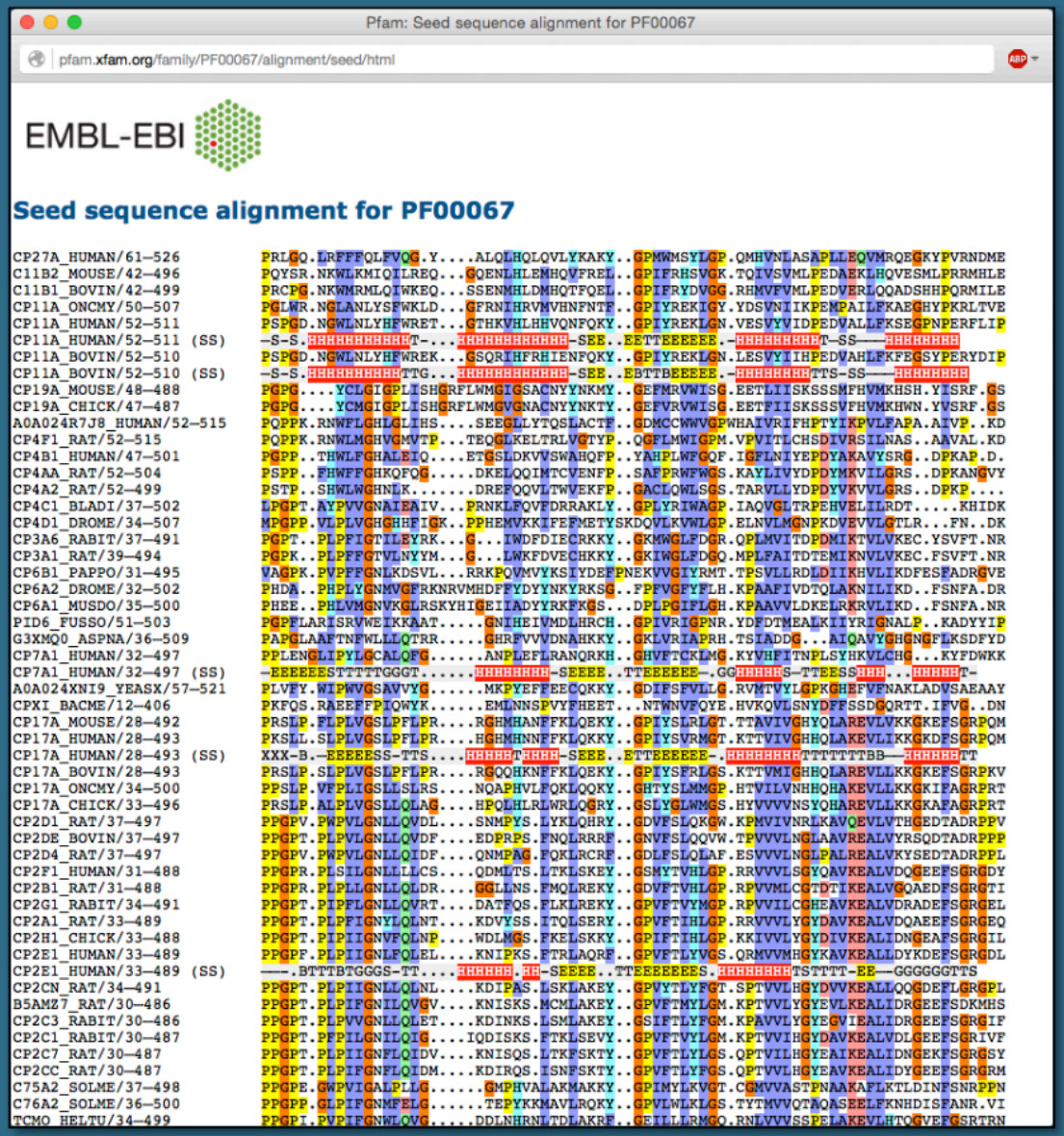
	Seed (50)	Full (41973)	Representative proteomes				UniProt (105935)	NCBI (141176)	Meta (2644)
			RP15 (9588)	RP35 (24353)	RP55 (37142)	RP75 (44573)			
Alignment:	◉	○	○	○	○	○	○	○	○
Format:	Selex								
Order:	◉ Tree	○ Alphabetical							
Sequence:	◉ Inserts lower case	○ All upper case							
Gaps:	Gaps as "." or "-" (mixed)								
Download/view:	◉ Download	○ View							

Generate

Download options

We make all of our alignments available in Stockholm format. You can download them here as raw, plain text files or as gzip-compressed files.

	Seed (50)	Full (41973)	Representative proteomes				UniProt (105935)	NCBI (141176)	Meta (2644)
			RP15 (9588)	RP35 (24353)	RP55 (37142)	RP75 (44573)			
Raw Stockholm	✓	✓	✓	✓	✓	✓	–	–	✓



- Glycine (G)
- Proline (P)
- Small or hydrophobic (A,V,L,I,M,F,W)
- Hydroxyl or amine amino acids (S,T,N,Q)
- Charged amino-acids (D,E,R,K)
- Histidine or tyrosine (H,Y)

- C Random coil
- H Alpha-helix
- G 3(10) helix
- I Pi-helix
- E Hydrogen bonded beta-strand (extended strand)
- B Residue in isolated beta-bridge
- T H-bonded turn (3-turn, 4-turn, or 5-turn)
- S Bend (five-residue bend centered at residue i)

Family: *p450* (PF00067)

455 architectures | 41973 sequences | 4 interactions | 929 species | 1275 structures

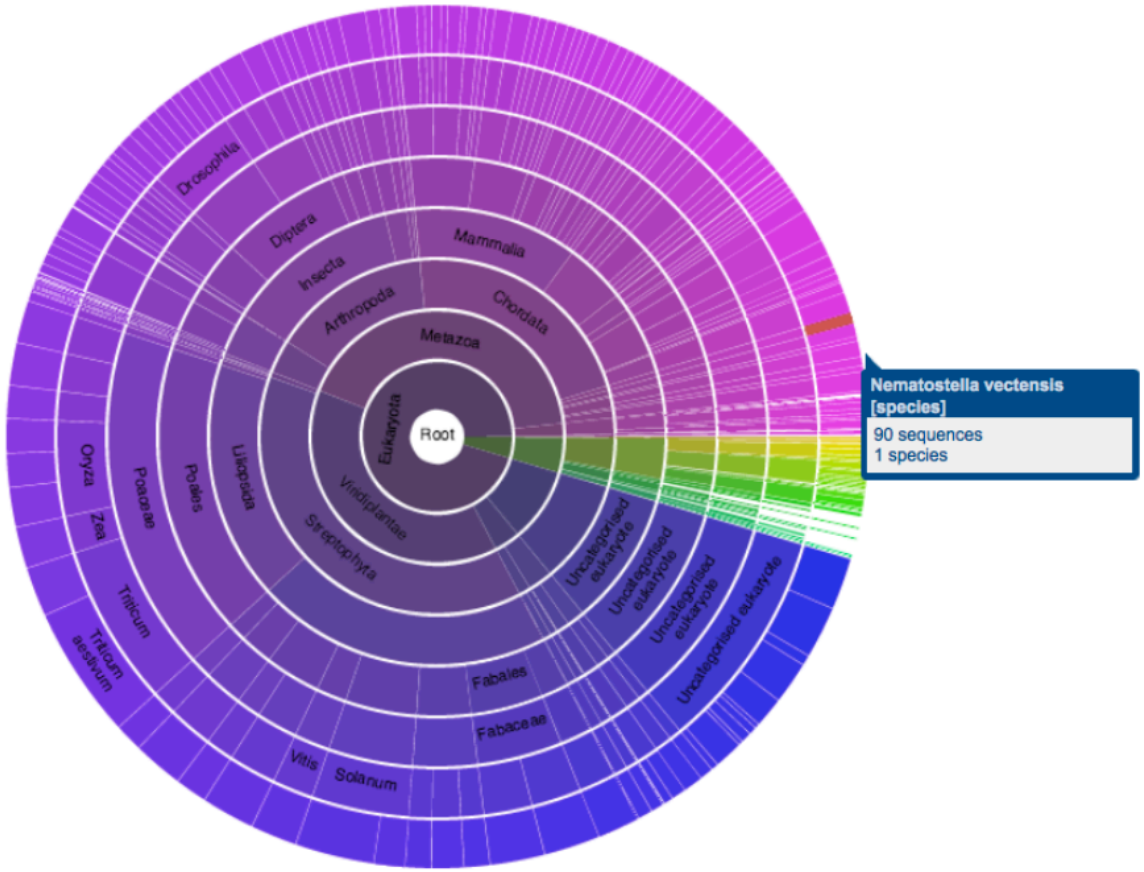
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Species distribution

Sunburst | Tree

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab. **More...**



Sunburst controls Hide

Nematostella vectensis

Root
Eukaryota
Metazoa
Cnidaria
Anthozoa
Actiniaria
Edwardsiidae
Nematostella
Nematostella vectensis

Weight segments by...
- (•) number of sequences
- ( ) number of species

Change the size of the sunburst
Small — Large

Colour assignments
Archea, Eukaryota, Bacteria, Other sequences, Viruses, Unclassified, Viroids, Unclassified sequence

Selections
Align selected sequences to HMM
Generate a FASTA-format file
Clear selection
Currently selected:
- 90 sequences
- 1 species

Note: selection tools show results in pop-up windows. Please disable pop-up blockers.

Family: *p450* (PF00067)

455 architectures | 41973 sequences | 4 interactions | 929 species | 1275 structures

Jump to... enter ID/acc Go

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Cytochrome P450 | Pfam | InterPro

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Cytochrome P450

Provide feedback

Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

Example structure

PDB entry 4C9P: Structure of camphor bound T260A mutant of CYP101D1

View a different structure: 4C9P

Literature references

1. Graham-Lorence S, Amarneh B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. PUBMED:7549871 EPMC:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 EPMC:8405421
3. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; , DNA Cell Biol 1993;12:1-51.: The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. PUBMED:7678494 EPMC:7678494
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 EPMC:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 EPMC:3304150
6. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272 EPMC:11178272

External database links

HOMSTRAD:	p450
PRINTS:	PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE:	PDOC00081
SCOP:	2cpp

Protein Databases
Pfam
Grouping Pfam entries into Clans

- Structural properties are often more conserved than the underlying sequence. Therefore, a single profile HMM is often insufficient to model an entire, diverse, structural superfamily. In Pfam there is a hierarchical level of classification which integrates evolutionary related entries in to sets, termed Clans, see Figure 6.
- The relationship between entries in a Clan may be defined by:
  - sequence similarity (whilst still originating from a common ancestor)
  - similarity of known three-dimensional structures
  - functional similarity
  - and/or similarity between their profile HMMs (as determined by algorithms such as HHsearch)
- The majority of Pfam Clans are groupings of domains and families.

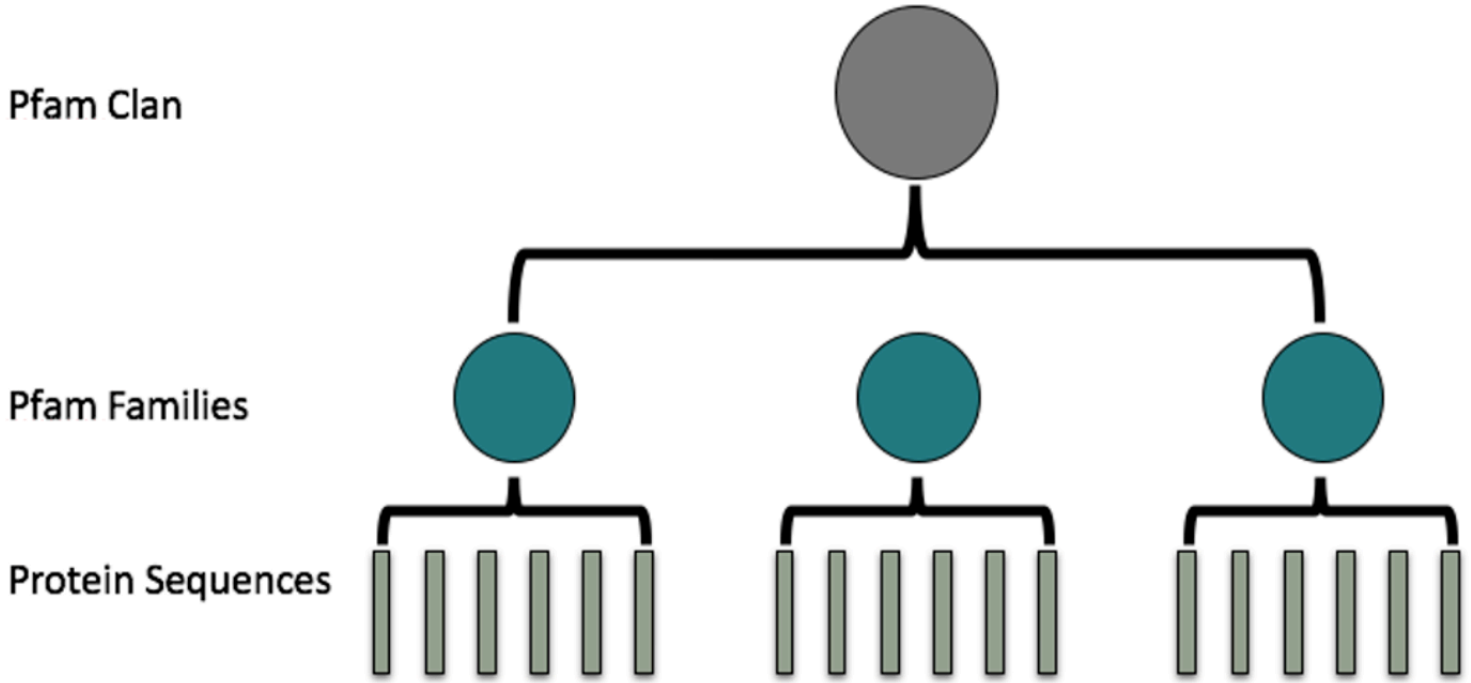


Figure 6 Grouping of protein sequences and Pfam families.

Protein Databases
Pfam
Summary

- Pfam is a database of conserved evolutionary units, each represented by a multiple sequence alignment and a profile hidden Markov model (HMM)
- The profile HMMs are created through an iterative process of building a model, finding new matching members and improving the model
- The inclusion of new matching members is dependant on manually curated cut-off thresholds
- The alignment is refined by defining boundaries to create a new seed, and from that a new model
- The process improves the model with each iteration until no new matches are found and the model is mature
- Pfam entries that have been identified as being related are grouped into sets called ‘clans’

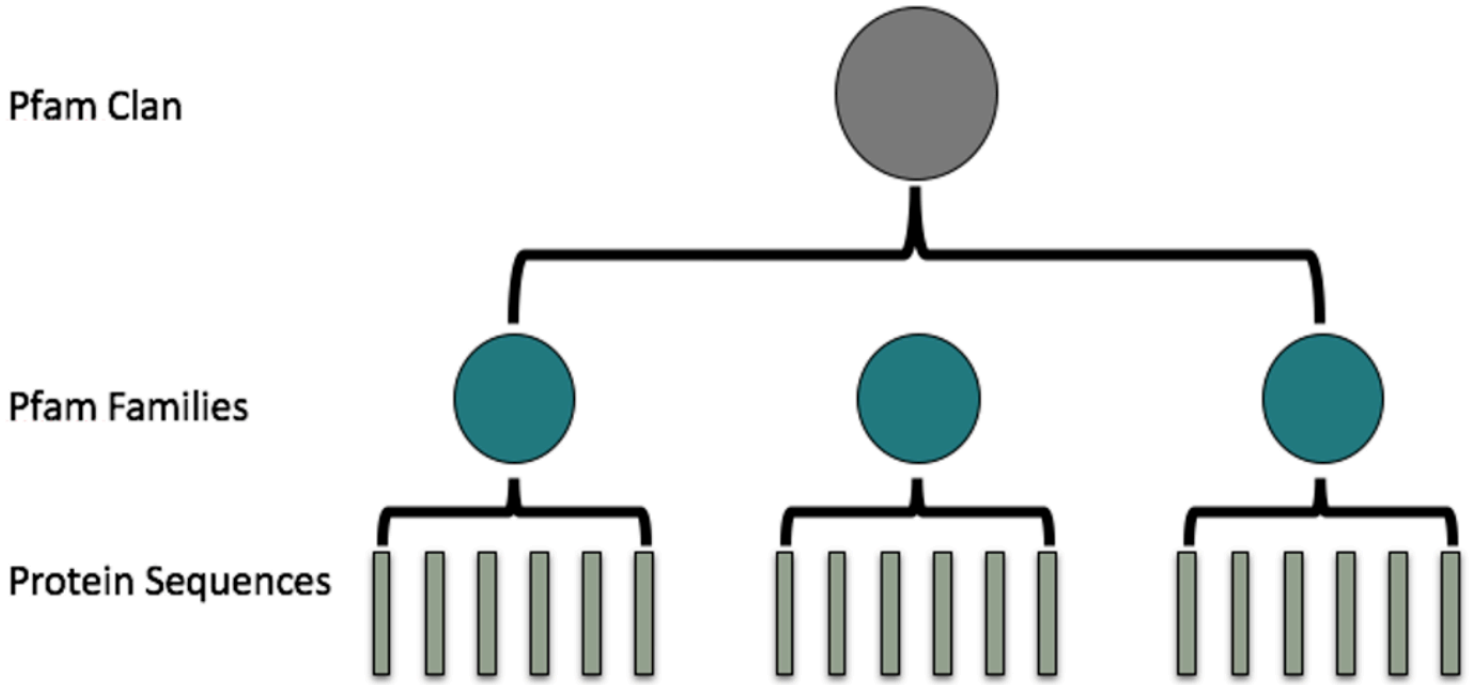


Figure 6 Grouping of protein sequences and Pfam families.