Representing Protein Domains with PSSMs and HMMs

Biol4230 Tues, February 6, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:

- understand types of domain definitions – folding units, evolutionary (mobile) units; domains vs motifs
- familiarity with InterPro, a "meta"-database of domain databases, and Pfam
- Where do pairwise scoring matrices come from? – the math
- Where do position specific scoring matrices (PSSMs) come from – PSI-BLAST
- What mistakes do Iterative methods (PSI-BLAST) make?

To learn more:

- Domains and InterPro – Pevzner, Part II, Ch. 10
- Scoring Matrices – Pevzner, Part I, Ch. 3
- PSSMs and PSI-BLAST – Pevzner, Part I, Ch. 5, p. 145
- Pick a protein of interest (serine protease, glutathione transferase, your favorite kinase, phosphatase, G-protein)
- Find the protein in interpro. Do the different domain databases find the same domains in the same places?
  - Compare your protein to SwissProt using PSI-BLAST
    - after 3 iterations, look at the domain structure of the five lowest scoring significant (E()<0.001) hits.
    - Are they all homologous (do they have the same domains)?
  - Find the protein in Pfam. What domains are found in the protein?

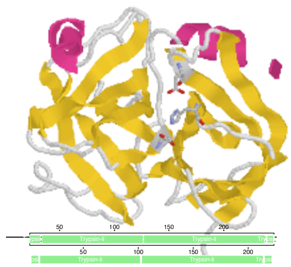
Finding domains with domain models I: from scoring matrices to PSSMs

- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - they are atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM1^{(n)} = PAM(n)$
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
- PSI-BLAST increases sensitivity with PSSMs

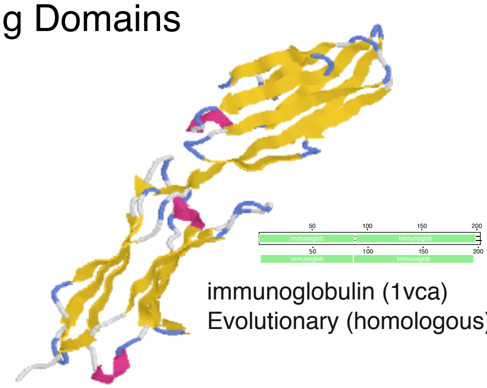
Representing Protein Domains

- Protein domains can be defined structurally, functionally, or based on evolutionary mobility
  - Mobile domains can be identified by duplication: mobile within protein (calmodulin), and alignment context: mobile among proteins
- Multiple-sequence based protein models (PSSMs, HMMs) are extend pair-wise scoring methods to sites on a protein model
- PSI-BLAST and HMMER build sensitive domain models
  - Position-Specific-Scoring Matrix (PSSM) from multiple sequence alignment
- InterPro provides integrated access to most domain annotations on a protein
- PFAM is a high-quality (curated) domain database
- ALL model/domain/sequence methods miss homologs
  - positives are correct, but negatives more ambiguous

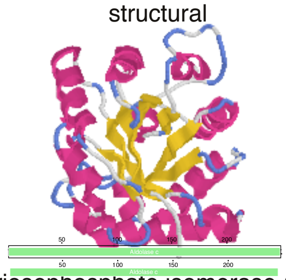
Defining Domains



chymotrypsin (1chg)
structural



immunoglobulin (1vca)
Evolutionary (homologous)



triosephosphate isomerase (1hti)
structural

Domain definitions:
- Structural
  - (molecular scissors)
- Evolutionary
  - (alternate contexts)

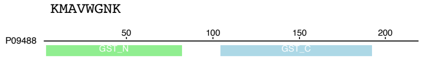
Often repetitive

fasta.bioch.virginia.edu/biol4230 5

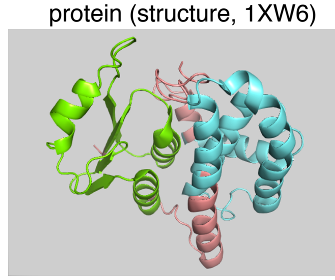
Sequence-based protein domains are evolutionarily mobile

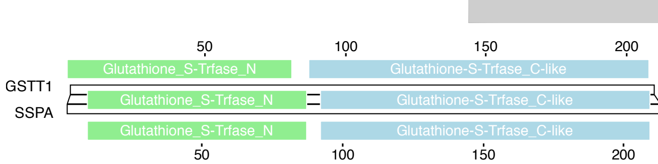
protein (sequence)

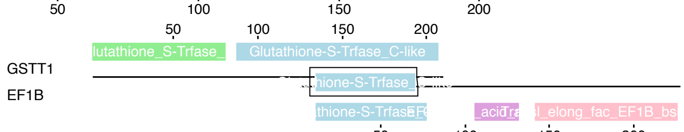
```
>ref|NP_000552.2| GSTM1 (human)
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTM
GDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKI
TQSNAILCYIARKHNLCGETEEEKIRVDILENQTM
DNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYSE
FLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCL
DAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFS
KMAVWGNK
```



protein (structure, 1XW6)





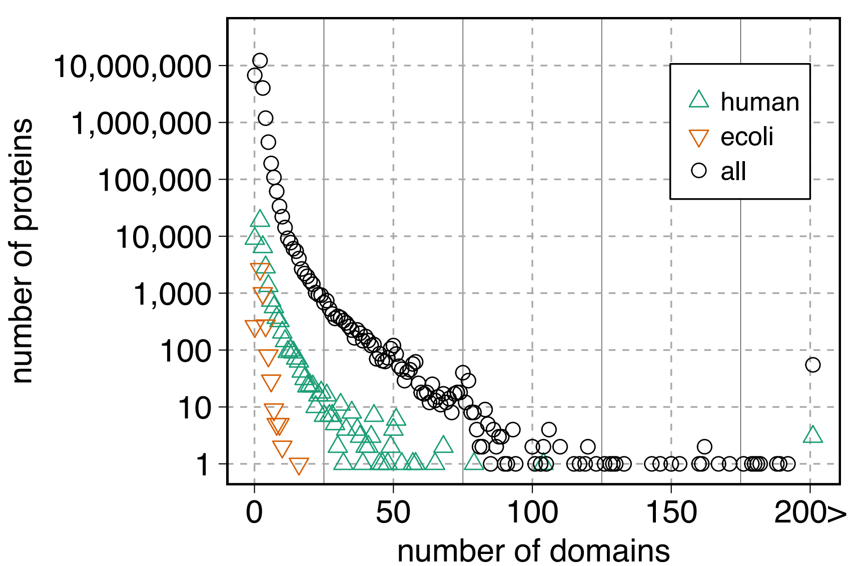


fasta.bioch.virginia.edu/biol4230 6

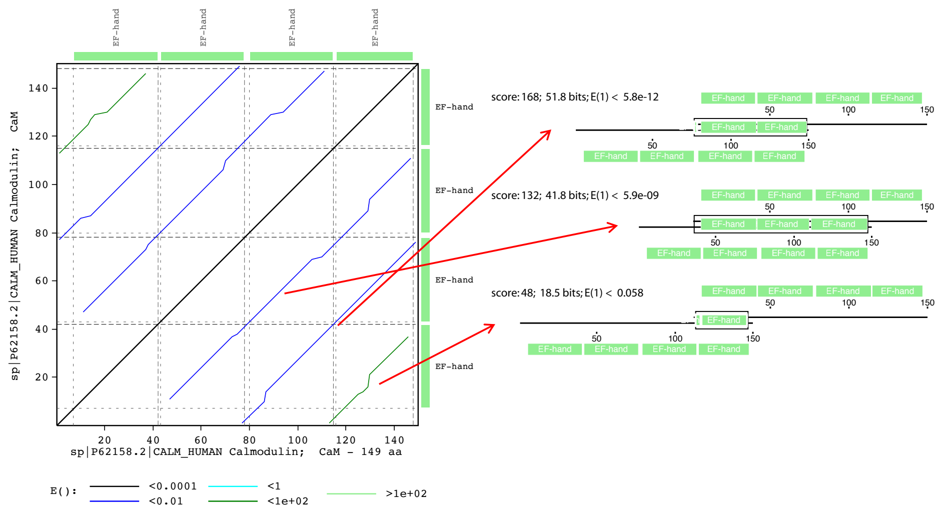
(Evolutionary) Domains vs complete proteins

- Many proteins are made up of multiple domains – structural/sequence units that evolve independently and may fold independently
- For multi-domain proteins, it is the domain, not the protein, that is the “atomic” unit of homology
- For multi-domain proteins, a significant similarity (homology) may apply only to one domain
- Domains are common, >50% of proteins contain more than one domain
- Unlike complete proteins, which have a beginning and end, domain boundaries can be more difficult to determine

Domain abundance (Pfam 31, 2017)



Identifying mobile domains:
mobile (duplicated) domains in local alignments

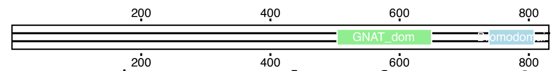


Identifying mobile domains:
homology in different contexts

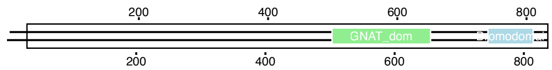
The best scores are:

		s-w	bits	E(454402)	%_id	%_sim	alen
KAT2B_HUMAN Histone acetyltransferase KAT2B	(832)	3820	1456.	0	1.000	1.000	832
KAT2A_HUMAN Histone acetyltransferase KAT2A	(837)	2747	1049.	0	0.721	0.870	813
GCN5_SCHPO Histone acetyltransferase gcn5	(454)	867	334.7	3e-90	0.483	0.768	354
GCN5_YEAST Histone acetyltransferase GCN5	(439)	792	306.2	1.1e-81	0.469	0.760	354
GCN5_ORYSJ Histone acetyltransferase GCN5	(511)	760	294.0	5.9e-78	0.436	0.755	376
GCN5_ARATH Histone acetyltransferase GCN5;	(568)	719	278.4	3.3e-73	0.434	0.740	369
BPTF_HUMAN Nucleosome-remodeling factor sub	(3046)	286	113.6	7.6e-23	0.495	0.804	97
NU301_DROME Nucleosome-remodeling factor su	(2669)	276	109.8	9.1e-22	0.511	0.819	94
CECR2_HUMAN Cat eye syndrome critical regio	(1484)	232	93.2	5e-17	0.371	0.790	105
BRD4_HUMAN Bromodomain-containing protein 4	(1362)	214	86.4	5.2e-15	0.379	0.698	116
BRD4_MOUSE Bromodomain-containing protein 4	(1400)	214	86.4	5.3e-15	0.379	0.698	116
BAZ2A_HUMAN Bromodomain adjacent to zinc fi	(1905)	211	85.2	1.7e-14	0.382	0.683	123
BAZ2A_XENLA Bromodomain adjacent to zinc fi	(1698)	206	83.3	5.5e-14	0.350	0.684	117
FSH_DROME Homeotic protein female sterile;	(2038)	205	82.9	8.8e-14	0.341	0.667	129
BAZ2A_MOUSE Bromodomain adjacent to zinc fi	(1889)	204	82.5	1e-13	0.368	0.680	125
BRDT_MACFA Bromodomain testis-specific prot	(947)	197	80.0	3e-13	0.367	0.697	109
BRD3_HUMAN Bromodomain-containing protein 3	(726)	194	78.9	4.9e-13	0.362	0.664	116

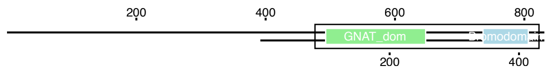
Homology and Domains – Histone acetyltransferase KAT2B



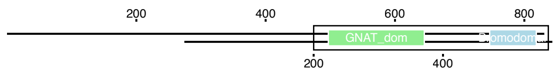
KAT2B_HUMAN Histone acetyltransferase KAT2B E()< 0 832



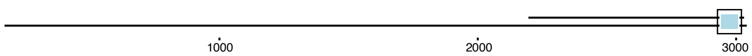
KAT2A_HUMAN Histone acetyltransferase KAT2B E()< 0 813



GCN5_YEAST Histone acetyltransferase GCN5 1.1e-81 354



GCN5_ARATH Histone acetyltransferase GCN5 3.3e-73 369



BPTF_HUMAN Nucleosome-remodeling factor 7.6e-23 97

Identifying mobile domains

Like homologous proteins, homologous domains share statistically significant structural or sequence similarity

- Many domain family members share significant sequence similarity (BLAST), and produce partial sequence alignments
- Internally repeated domains can be identified with lalign
  - Domain boundaries may depend on the scoring matrix
- To find all (or most) domain family members, more sensitive methods are used:
  - PSSMs (Position Specific Scoring Matrices) PSI-BLAST, RPS-BLAST
  - HMMs (Hidden Markov Models) HMMER3 (Pfam)

Protein Motif and Domain Databases

RNA sequence databases
Protein sequence databases
- General sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
  - ASC - Active Sequence Collection
  - Blocks
  - COMe - Co-Ordination
  - CSA - Catalytic Site Atlas
  - eF-site - Electrostatic surface
  - eMOTIF
  - **InterPro**
  - Metalloprotein Database and Browser
  - O-GLYCBASE
  - PhosphoBase
  - PRINTS
  - PROMISE
  - **PROSITE**
- Protein domain databases; protein classification
- Databases of individual protein families

Protein domain databases; protein classification
- BAliBASE
- CDD
- CluSTr - Clusters of Swiss-Prot and TrEMBL
- COG - Clusters of Orthologous Groups
- DomIns - Database of Domain Insertions
- FusionDB
- Hits
- HSSP
- InterDom
- **InterPro**
- iProClass
- MetaFam
- PALI
- **Pfam**
- PIR-ALN
- PIRSF
- ProClass
- ProDom
- ProtoMap
- ProtoNet
- SBASE
- SMART
- SUPFAM
- TIGRFAMs

InterPro, PFAM, and Prosite

InterPro – The database of Protein databases www.ebi.ac.uk/interpro

PFAM – a "domain" database pfam.xfam.org
- Complete domain alignments. Definition of domains.
- Example of searching PFAM on-line; what scores mean.
- Caveats: structural rather than functional classification

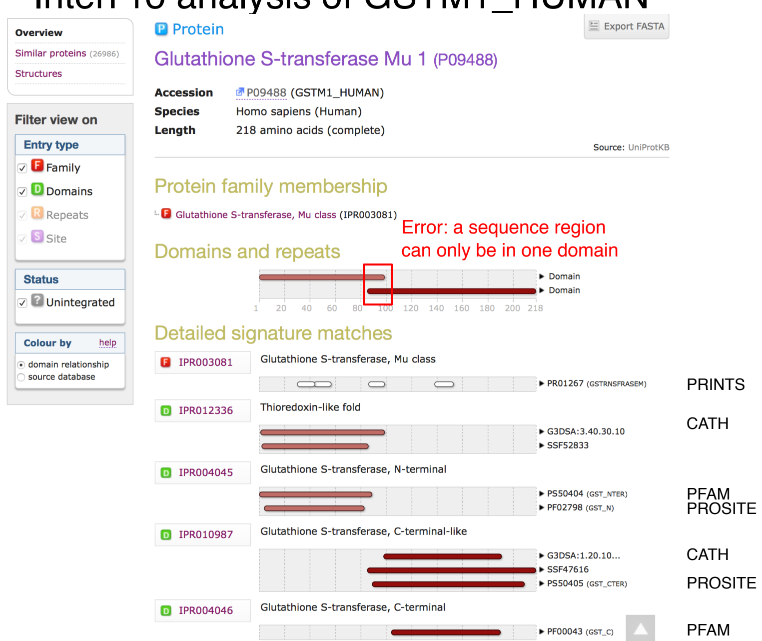
PROSITE – a "motif" database www.expasy.org/prosite
- Patterns and regular expressions
- The information content of a PROSITE pattern
- Examples of searching PROSITE on-line
- Caveats: missing patterns; low-information patterns

Always do control experiments: never trust a server
- Positive controls -- submit sequences for which you know the right answer.
- Negative controls -- random or shuffled sequences.

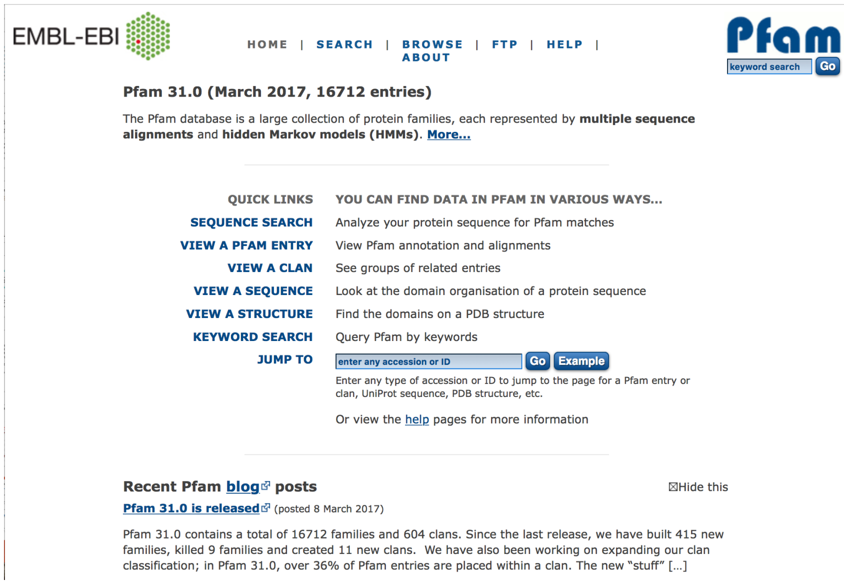
Representations of domains

- Regular expressions – exact match to regular expression (good for absolutely conserved motifs, active sites) – ProSite patterns
- HMM/PSSM/Profile (Hidden Markov Model/Position Specific Scoring matrix/Profile) – HMM most flexible, provides statistical significance estimates
  - Pfam, Tigrfam, SuperFamily, Panther, ProSite profiles, HaMap profiles

InterPro analysis of GSTM1_HUMAN



PFAM – pfam.xfam.org



EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 31.0 (March 2017, 16712 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	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.

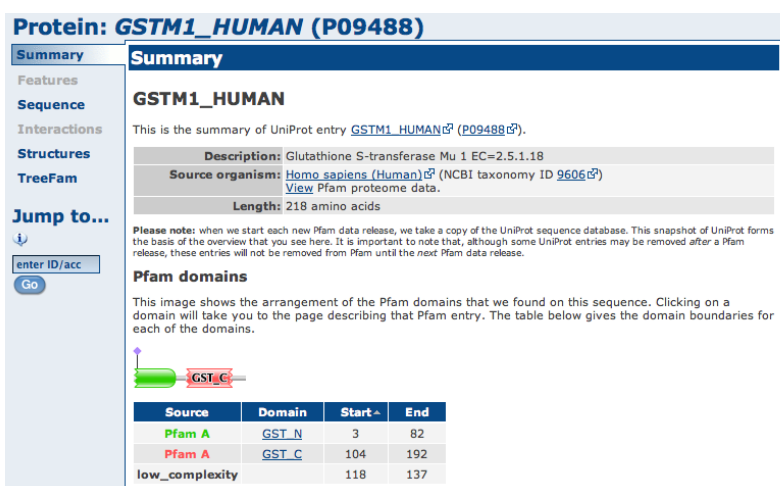
Or view the help pages for more information

Recent Pfam blog posts ☒Hide this

Pfam 31.0 is released (posted 8 March 2017)

Pfam 31.0 contains a total of 16712 families and 604 clans. Since the last release, we have built 415 new families, killed 9 families and created 11 new clans. We have also been working on expanding our clan classification; in Pfam 31.0, over 36% of Pfam entries are placed within a clan. The new "stuff" [...]

Pfam domains on GSTM1_HUMAN



Protein: *GSTM1_HUMAN* (P09488)

Summary
Features
Sequence
Interactions
Structures
TreeFam

Jump to...
enter ID/acc
Go

Summary

GSTM1_HUMAN

This is the summary of UniProt entry GSTM1_HUMAN (P09488).

Description:	Glutathione S-transferase Mu 1 EC=2.5.1.18
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	218 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

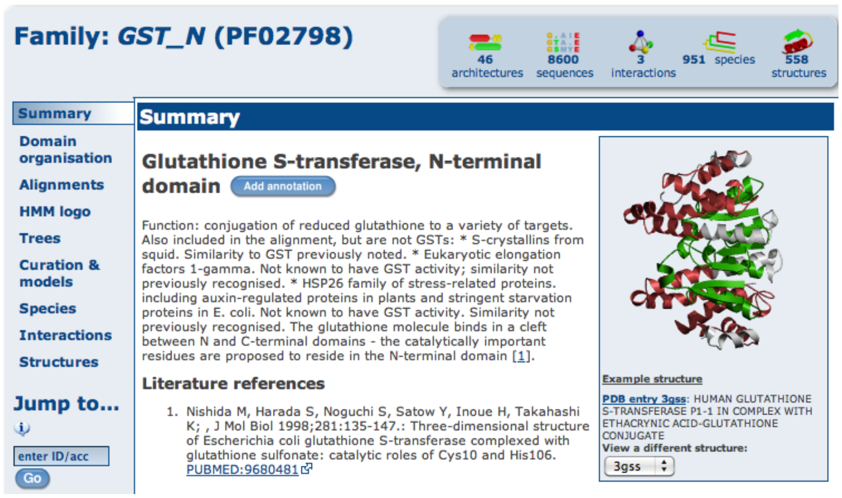
Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

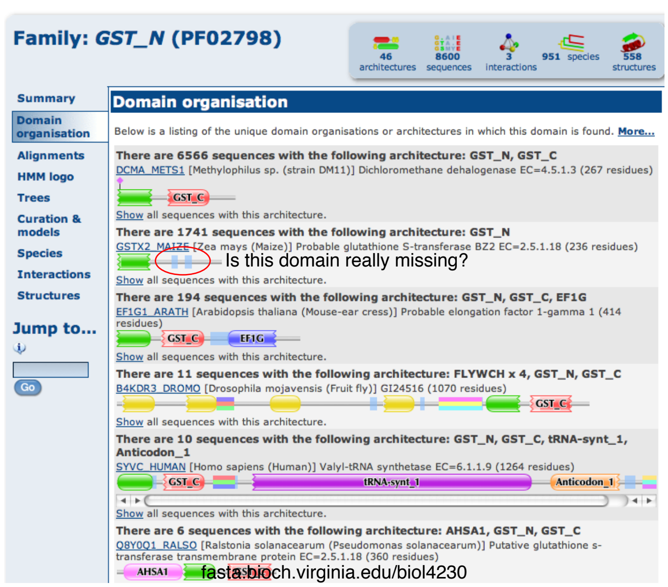
GST_C

Source	Domain	Start	End
Pfam A	GST_N	3	82
Pfam A	GST_C	104	192
low_complexity		118	137

Pfam domain descriptions – GST_N



Pfam GST_N architectures



Finding domains with domain models I: from scoring matrices to PSSMs

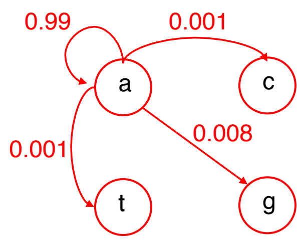
- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - they are atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource

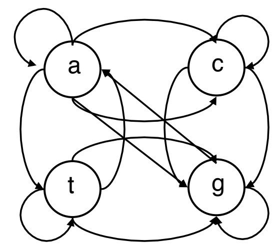
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM1^{(n)} = PAM(n)$
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
- PSI-BLAST increases sensitivity with PSSMs

Improving search sensitivity with Protein family models (PSSMs and HMMs)

- Where do scoring matrices come from
  - Transition probabilities and PAMs
  - Scoring matrices as log-odds values (log(p[related]/p[chance])
- From non-position-specific (PAM250, BLOSUM62) to position-specific – PSI-BLAST

DNA transition probabilities – 1 PAM





	a	c	g	t	
a	0.99	0.001	0.008	0.001	= 1.0
c	0.001	0.99	0.001	0.008	= 1.0
g	0.008	0.001	0.99	0.001	= 1.0
t	0.001	0.008	0.001	0.99	= 1.0

Matrix multiples

can also be calculated from
"instantaneous rate matrix Q"
p(t) = exp(t*Q)

PAM 2
```
M^2={
{0.980, 0.002, 0.016, 0.002},
{0.002, 0.980, 0.002, 0.016},
{0.016, 0.002, 0.980, 0.002},
{0.002, 0.016, 0.002, 0.980}}
```

PAM 5
```
M^5={
{0.952, 0.005, 0.038, 0.005},
{0.005, 0.951, 0.005, 0.038},
{0.038, 0.005, 0.952, 0.005},
{0.005, 0.038, 0.005, 0.952}}
```

PAM 10
```
M^10={
{0.907, 0.010, 0.073, 0.010},
{0.010, 0.907, 0.010, 0.073},
{0.073, 0.010, 0.907, 0.010},
{0.010, 0.073, 0.010, 0.907}}
```

PAM 100
```
M^100={
{0.499, 0.083, 0.336, 0.083},
{0.083, 0.499, 0.083, 0.336},
{0.336, 0.083, 0.499, 0.083},
{0.083, 0.336, 0.083, 0.499}}
```

PAM 1000
```
M^1000={
{0.255, 0.245, 0.255, 0.245},
{0.245, 0.255, 0.245, 0.255},
{0.255, 0.245, 0.255, 0.245},
{0.245, 0.255, 0.245, 0.255}}
```

Where do scoring matrices come from?

$$\lambda S = \log\left(\frac{q_{ij}}{p_j}\right)$$

alignment from homology

probability of alignment by chance

```
qij = M^20=    PAM20(numerator)
{0.828, 0.019, 0.133, 0.019},
{0.019, 0.828, 0.019, 0.133},
{0.133, 0.019, 0.828, 0.019},
{0.019, 0.133, 0.019, 0.828}}
```

```
pi(a,c,g,t)=
         pj=0.25
```

$$\lambda S = 10\log\left(\frac{q_{a,a}}{p_a}\right) = 10\log\left(\frac{0.828}{0.25}\right) = 5.2$$

$$\lambda S = 10\log\left(\frac{q_{a,c}}{p_c}\right) = 10\log\left(\frac{0.019}{0.25}\right) = -11.2$$

$$\lambda_2 = \frac{\log(2)}{10} = 0.33$$

Two expressions for S_{ij}

Transition frequency (probability) - Durbin et al.

```
qij = M^20= PAM20(numerator)
{0.828, 0.019, 0.133, 0.019},
{0.019, 0.828, 0.019, 0.133},
{0.133, 0.019, 0.828, 0.019},
{0.019, 0.133, 0.019, 0.828}}
```

$$\lambda S = \log\left(\frac{q_{ij}^t}{p_j}\right)$$

Alignment frequency (probability) -Altschul

```
qaij = M^20=  PAM20(numerator)
{0.207, 0.005, 0.043, 0.005},
{0.019, 0.207, 0.019, 0.043},
{0.043, 0.005, 0.207, 0.005},
{0.005, 0.043, 0.005, 0.207}}
```

$$\lambda S = \log\left(\frac{q_{ij}^a}{p_i p_j}\right)$$

Altschul $q_{ij}^a = p_i \times$ Durbin q_{ij}^t

$$\lambda S = \log\left(\frac{q_{ij}^a = p_i q_{ij}^t}{p_i p_j}\right)$$

Scoring matrices at DNA PAMs - ratios

blastn (DNA)

```
PAM1={   ratio=1/3.13=+1/-3 H=1.90
{ 1.99, -6.23, -6.23, -6.22},
{-6.23,  1.99, -6.23, -6.23},
{-6.23, -6.23,  1.99, -6.23},
{-6.23, -6.23, -6.23,  1.99}}
```

```
PAM2={   ratio=1/2.65=+2/-5 H=1.82
{ 1.97, -5.24, -5.24, -5.24},
{-5.24,  1.98, -5.24, -5.24},
{-5.24, -5.24,  1.98, -5.24},
{-5.24, -5.24, -5.24, -5.24}}
```

```
PAM10={  ratio=1/1.61=+2/-3 H=1.40
{ 1.86, -3.00, -3.00, -3.00},
{-3.00,  1.86, -3.00, -3.00},
{-3.00, -3.00,  1.86, -3.00},
{-3.00, -3.00, -3.00,  1.86}}
```

```
PAM20={  ratio=1/1.21=+4/-5 H=1.05
{ 1.72, -2.09, -2.09, -2.09},
{-2.09,  1.72, -2.09, -2.09},
{-2.09, -2.09,  1.72, -2.09},
{-2.09, -2.09, -2.09,  1.72}}
```

```
PAM30={  ratio=1/1=+1/-1 H=0.80
{ 1.59, -1.59, -1.59, -1.59},
{-1.59,  1.59, -1.59, -1.59},
{-1.59, -1.59,  1.59, -1.59},
{-1.59, -1.59, -1.59,  1.59}}
```

fasta (DNA)

```
PAM45={  ratio=1.23/1=+5/-4 H=0.54
{ 1.40, -1.14, -1.14, -1.14},
{-1.14,  1.40, -1.14, -1.14},
{-1.14, -1.14,  1.40, -1.14},
{-1.14, -1.14, -1.14,  1.40}}
```


Where do scoring matrices come from?

Pam40

	A	R	N	D	E	I	L
A	8						
R	-9	12					
N	-4	-7	11				
D	-4	-13	3	11			
E	-3	-11	-2	4	11		
I	-6	-7	-7	-10	-7	12	
L	-8	-11	-9	-16	-12	-1	10

Pam250

	A	R	N	D	E	I	L
A	2						
R	-2	6					
N	0	0	2				
D	0	-1	2	4			
E	0	-1	1	3	4		
I	-1	-2	-2	-2	-2	5	
L	-2	-3	-3	-4	-3	2	6

$$\lambda S_{i,j} = \log_b\left(\frac{q_{i,j}}{p_i p_j}\right)$$

q_{ij} : replacement frequency at PAM40, 250

$q_{R:N\,(\ 40)} = 0.000435$ $p_R = 0.051$

$q_{R:N\,(250)} = 0.002193$ $p_N = 0.043$

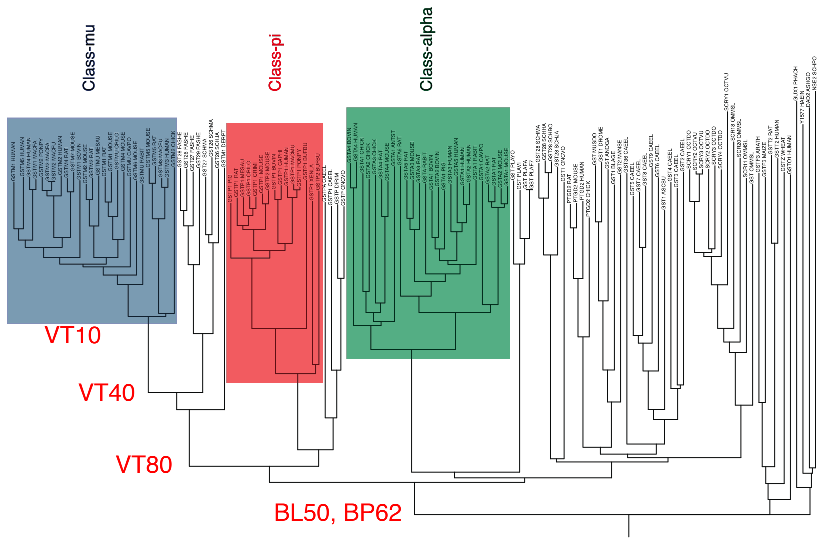
$l_2\ S_{ij} = lg_2\ (q_{ij}/p_ip_j)$ $l_e\ S_{ij} = \ln(q_{ij}/p_ip_j)$ $p_Rp_N = 0.002193$

$l_2\ S_{R:N(\ 40)} = lg_2\ (0.000435/0.00219) = -2.333$

$l_2 = 1/3$; $S_{R:N(\ 40)} = -2.333/l_2 = -7$

$l\ S_{R:N(250)} = lg2\ (0.002193/0.002193) = 0$

Shallow matrices reduce evolutionary look-back
Glutathione Transferases (gstm1_human)



Finding domains with domain models I: from scoring matrices to PSSMs

- Position independent scoring matrices can be built from a simple evolutionary model: $PAM1^{(n)} = PAM(n)$
  - PAM10,20,…,250/VT10,20,...,250 come from evolutionary model
  - BLOSUM50,62,80 do not (and direction is opposite)
  - Shallow (PAM10,20) matrices for short distances
  - Matrices have preferred percent identity/alignment length
  - Shallow matrices for short alignments

- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position

Improving sensitivity with protein/domain family models

- Shallower scoring matrices (dialing back the q_{ij} from the evolutionary model) *reduces* look-back time
  - VT20: 80% identity; VT80: 35% id; BL50: 25% id
  - reduced look-back = reduced sensitivity
- How to *increase* look-back time (more sensitivity)
  - Position Specific Scoring Matrices (PSSMs)
  - Hidden Markov Models (HMMs)

Pairwise Alignment

```
RU1A_HUMAN rrm2      VQAGAAR

PABP_DROME rrm3      EAAEAAV

                      +2 +2 +2
```

score matrices: 20x20, 210 parameters position-*independent*

Cys	12									
Ser	0	2								
Thr	-2	1	3							
Pro	-1	1	0	6						
Ala	-2	1	1	1	2					
Gly	-3	1	0	-1	1	5				
Asn	-4	1	0	-1	0	0	2			
Asp	-5	0	0	-1	0	1	2	4		
Glu	-5	0	0	-1	0	0	1	3	4	
Gln	-5	-1	-1	0	0	-1	1	2	2	4
	C	S	T	P	A	G	N	D	E	Q

Profile Alignment

```
RU1A_HUMAN rrm1        SSATNAL
RU1A_HUMAN rrm2        VQAGAAR      query
SFR1_HUMAN rrm1        RDAEDAV
SXLF_DROME rrm1        MDSQRAI

PABP_DROME rrm3        EAAEAAV      target
                        +3 +4
                          0
```

profile: 20 scores *per column*

position-*dependent*

Where pairwise scores come from –

$$\text{score(AA)=log}\ \frac{P(A|A)}{f(A)}$$

"probability of A given an A"
the observed probability of seeing an A aligned to an A in real alignments

"frequency of A"
the expected frequency of A in any sequence

$$Sc(AA) = \log_2 \frac{0.64}{0.04} = +4$$

$$Sc(AE) = \log_2 \frac{0.01}{0.04} = -2$$

Where profile scores (should) come from

"probability of A at position x"
the observed probability of seeing an A
in the consensus column x

$$\text{score}(A|x) = \log \frac{P(A|\text{position } x)}{f(A)}$$

$$Sc(A|6) = \log_2 \frac{1.00}{0.04} = +4.6 \qquad Sc(A|5) = \log_2 \frac{0.04}{0.04} = 0$$

$$Sc(N|6) = \log_2 \frac{0.00}{0.06} = \text{-inf} \qquad Sc(N|5) = \log_2 \frac{0.06}{0.06} = 0$$

1. what about position-specific gap penalties?
2. how to estimate parameters from small numbers of observations?

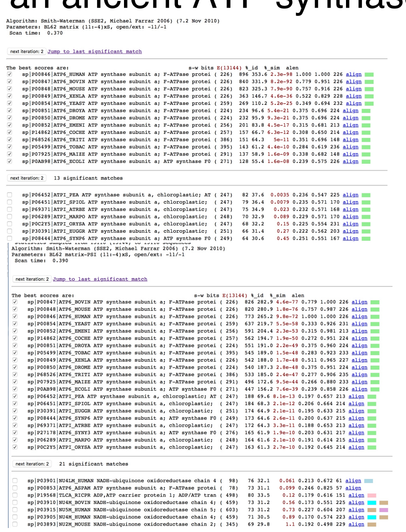
Finding domains with domain models I: from scoring matrices to PSSMs

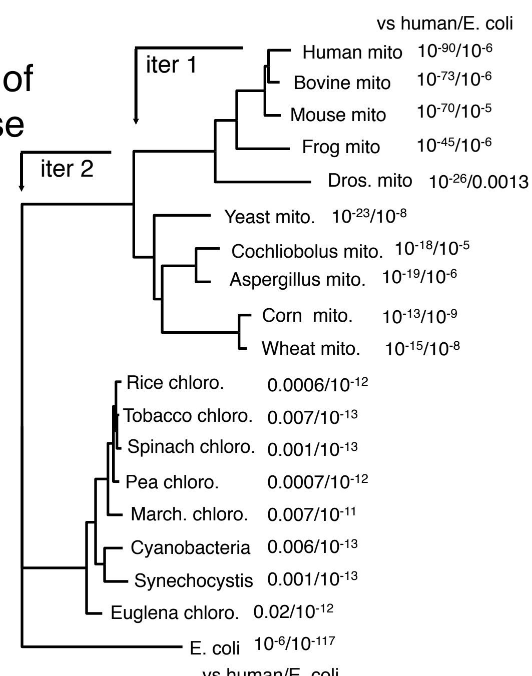
- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - they are atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM1^{(n)} = PAM(n)$
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
- PSI-BLAST increases sensitivity with PSSMs

Improving sensitivity with protein/domain family models

- PSI-BLAST - method
  1. do BLAST search
  2. use query-based implied multiple sequence alignment to build Position Specific Scoring Matrix (PSSM)
  3. repeat steps 1 and 2 with PSSM, for 5 – 10 iterations
- PSI-BLAST – results:
  1. Typically 2X as sensitive as single sequence methods
  2. Over-extension can cause PSSM contamination

PSI-BLAST iteratively builds a model (PSSM) of an ancient ATP synthase





PSI-BLAST ATP6_HUMAN - 4 iterations

Threshold 10^{-20} for demo, use 10^{-3} normally

Results from round:	(1)		(2)		(3)		(4)	
Sequences producing significant alignments:	Score (bits)	E Value	Score (bits)	E Value	Score (bits)	E Value	Score (bits)	E Value
ATP6_HUMAN ATP synthase a chain (ATPase protein 6)	296	3e-81	257	1e-69	241	2e-62	222	5e-59
ATP6_BOVIN ATP synthase a chain (ATPase protein 6)	253	2e-68	257	2e-69	239	8e-65	230	2e-61
ATP6_MOUSE ATP synthase a chain (ATPase protein 6)	245	5e-66	247	3e-66	234	4e-64	225	6e-60
ATP6_XENLA ATP synthase a chain (ATPase protein 6)	142	9e-35	227	1e-60	189	3e-49	177	2e-45
ATP6_DROYA ATP synthase a chain (ATPase protein 6)	101	2e-22	206	3e-54	209	5e-55	196	4e-51
(2)								
ATP6_YEAST ATP synthase a chain precursor (ATPase prot	93	5e-20	97	3e-21	199	4e-52	191	2e-49
ATP6_TRITI ATP synthase a chain (ATPase protein 6)	83	5e-17	96	5e-21	218	1e-57	236	4e-63
(3)								
ATP6_TOBAC ATP synthase a chain (ATPase protein 6)	80	3e-16	90	4e-19	200	2e-52	230	3e-61
ATP6_MAIZE ATP synthase a chain (ATPase protein 6)	76	5e-15	88	1e-18	198	1e-51	219	5e-58
ATP6_COCHE ATP synthase a chain (ATPase protein 6)	75	1e-14	86	9e-18			197	2e-51
ATP6_EMENI ATP synthase a chain precursor (ATPase prot	75	2e-14	84	3e-17	123	5e-29	181	2e-46
(4)								
ATP6_ECOLI ATP synthase a chain (ATPase protein 6)	42	1e-04	40	5e-04	46	8e-06	49	1e-06
ATPI_SPIOL Chloroplast ATP synthase a chain precursor			32	0.12	36	0.006	39	0.001
ATP6_SYNY3 ATP synthase a chain (ATPase protein 6)	28	1.9	32	0.16	44	5e-05	45	1e-05
ATPI_MARPO Chloroplast ATP synthase a chain precursor			31	0.21	44	4e-05	44	3e-05
ATPI_PEA Chloroplast ATP synthase a chain precursor (A			31	0.32	37	0.005		
LAMA2_MOUSE Laminin subunit alpha-2 precursor (Laminin			31	0.34				
ATPI_ATRBE Chloroplast ATP synthase a chain precursor			31	0.39	41	2e-04		
ATP6_SYNP6 ATP synthase a chain (ATPase protein 6)			28	1.7	41	2e-04		
ATPI_EUGGR Chloroplast ATP synthase a chain precursor					39	0.001		
ATPI_ORYSA Chloroplast ATP synthase a chain precursor			28	1.9	36	0.008		
ATPI_ATRBE Chloroplast ATP synthase a chain precursor					36	0.009	38	0.002
ATP6_ASPAM ATP synthase a chain (ATPase protein 6)							36	0.008
POLG_KUNJM Genome polyprotein [Contains: Capsid protei...	27	5.0						
POL_HTL1C Gag-Pro-Pol polyprotein (Pr160Gag-Pro-Pol) [...	27	5.0						
POLG_DEN2J Genome polyprotein [Contains: Capsid protei...	27	5.2	26	7.0				

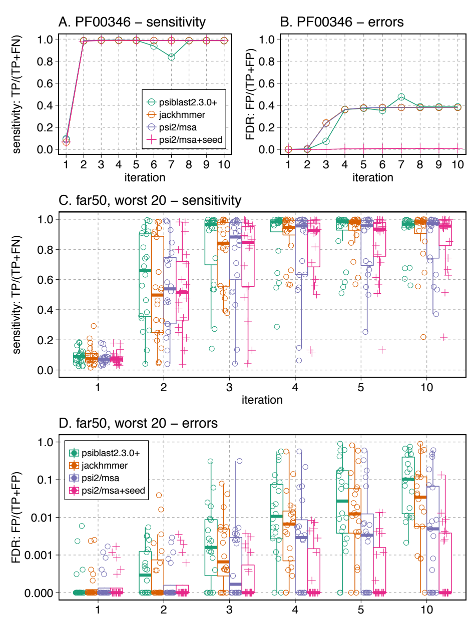
Multiple sequence alignment: Metazoan ATP Synthases

```
CLUSTAL W (1.81) multiple sequence alignment
                                                          46        56
ATP6_BOVIN  MNENLFTSFITPVILGLPLVTLIVLFPSLLF--PTSNRLVSNRFVTLQQWMLQLVSKQMMSIHNSKGQTWT-LML
ATP6_MOUSE  MNENLFASFITPTMMGFPIVVAIIMFPSILF--PSSKRLINNRLHSFQHWLVKLIIKQMMLIHTPKGRTWT-LMI
ATP6_HUMAN  MNENLFASFIAPTILGLPAAVLIILFPPLLI--PTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWS-LML
ATP6_XENLA  MNLSFFDQFMSPVILGIPLIAIAMLDPFTLISWPIQSNGFNNRLITLQSWFLHNFTTIFYQLTSP-GHKWA-LLL
ATP6_DROYA  MMTNLFSVFDPSAIFNLSLNWLSTFLGLLMI--PSIYWLMPSRYNIFWNSILLTLHKEFKTLLGPSGHNGSTFIF
            *  .:*  * ...::.:.     :    ::  *     . .*      ::    . :  :  . *:. : :::
                                          97                               131
ATP6_BOVIN  MSLILFIGSTNLLGLLPHSFTPTTQLSMNLGMAIPLWAGAVITGFRNKTKASLAHFLPQGTPTPLIPMLVIIETI
ATP6_MOUSE  VSLIMFIGSTNLLGLLPHTFTPTTQLSMNLSMAIPLWAGAVITGFRHKLKSSLAHFLPQGTPISLIPMLIIIETI
ATP6_HUMAN  VSLIIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPLIPMLVIIETI
ATP6_XENLA  TSLMLLLMSLNLLGLLPYTFTPTTQLSLNMGLAVPLWLATVIMASKP-TNYALGHLLPEGTPTPLIPVLIIIETI
ATP6_DROYA  ISLFSLILFNNFMGLFPYIFTSTSHLTLTLSLALPLWLCFMLYGWINHTQHMFAHLVPQGTPAILMPFMVCIETI
             **: ::   *::**:*: **.*::*::.:.:*:***   :: .     :  :.*::*:***  *:*.:: ****
   152                                                               210
ATP6_BOVIN  SLFIQPMALAVRLTANITAGHLLIHLIGGATLALMSISTTTALITFTILILLTILEFAVAMIQAYVFTLLVSLYLHDNT
ATP6_MOUSE  SLFIQPMALAVRLTANITAGHLLMHLIGGATLVLMNISPPTATITFIILLLLTILEFAVALIQAYVFTLLVSLYLHDNT
ATP6_HUMAN  SLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
ATP6_XENLA  SLFIRPLALGVRLTANLTAGHLLIQLIATAAFVLLSIMPTVAILTSIVLFLLTLLEIAVAMIQAYVFVLLLSLYLQENV
ATP6_DROYA  SNIIRPGTLAVRLTANMIAGHLLLTLLGNTGPSMSYLLVTFLLVAQIALLVL---ESAVTMIQSYVFAVLSTLYSSEVN
            * :*:* :*.******: *****: *:. :    :  :  .    :    *::*   * **::**:***.:* :** :
```

Position-Specific Scores
ATP Synthase, 4 iterations

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	bits/pos
BL62	Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0.70
46	Q	-2	-1	-2	-2	-4	6	0	1	0	-4	-3	-1	-2	-1	-3	-1	-2	6	4	-3	0.74
	%	0	0	0	0	0	54	0	12	0	0	0	0	0	0	0	0	0	13	20	0	
47	Q	-1	-1	3	3	-3	3	3	-2	3	-4	-4	-1	-3	-4	-2	2	-1	-4	-2	-3	0.51
	%	0	0	13	20	0	16	19	0	8	0	0	0	0	0	0	24	0	0	0	0	
56	Q	-2	-1	-2	-2	-3	5	2	-4	-1	4	-1	-1	-1	-2	-3	-2	-2	-3	-2	0	0.51
	%	0	0	0	0	0	46	13	0	0	41	0	0	0	0	0	0	0	0	0	0	
97	Q	-2	-1	0	-2	-4	4	0	-3	8	-4	-4	-1	-2	-3	-3	-1	-2	-3	0	-4	1.11
	%	0	0	0	0	0	35	0	0	65	0	0	0	0	0	0	0	0	0	0	0	
131	Q	3	-1	-1	-1	-2	5	2	-2	-1	-3	-3	0	-2	-4	-2	1	-1	-3	-3	-2	0.52
	%	44	0	0	0	0	36	11	0	0	0	0	0	0	0	0	9	0	0	0	0	
152	Q	-2	6	-1	-2	-4	4	0	-3	-1	-4	-3	1	-2	-4	-3	-1	-2	-4	-3	-3	1.00
	%	0	77	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
210	Q	-2	0	-1	-1	-4	7	1	-3	0	-4	-3	1	-1	-4	-2	-1	-2	-3	-2	-3	1.13
	%	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

How much improvement with PSSMs/ HMMs?



Pearson (2017) Nuc. Acids Res. 45:e46

Sensitive searches with PSI-BLAST

- PSI-BLAST improves sensitivity by building a Position Specific Scoring Matrix (PSSM)
  - models ancestral sequence (consensus distribution)
  - similar to PFAM HMM (but less sophisticated weights, gaps)
- PSI-BLAST likes larger databases (more data)
- Sensitivity improves with additional iterations
  - model moves to base of tree
- Statistical estimates are difficult
  - once a sequence is in, it is “significant” - validation must be done before a sequence is included
- Very diverse families may not produce a well defined PSSM
  - similar problems with HMMs have led to “clans”

Finding domains with domain models I: from scoring matrices to PSSMs

- Domains are structurally compact, evolutionarily mobile, protein building blocks
  - atomic, they have a characteristic length
  - often repeated, or found in different sequence contexts
  - essential for building detection systems (PSSMs, HMMs), because they focus on the homologous region (a full length protein can be a mixture of domains)
  - Interpro provides large-scale summary
  - Pfam most comprehensive single resource
- Position independent scoring matrices can be built from a simple evolutionary model: $PAM1^{(n)} = PAM(n)$
  - Shallow (low change) for short distances/short alignments
  - Preferred identity/alignment lengtt
- Position Specific Scoring Matrices (PSSMs) generalize frequency data for a single position
  - Improve sensitivity 2 – 10-fold or more
- PSI-BLAST increases sensitivity with PSSMs
  - Also jackhmmer