

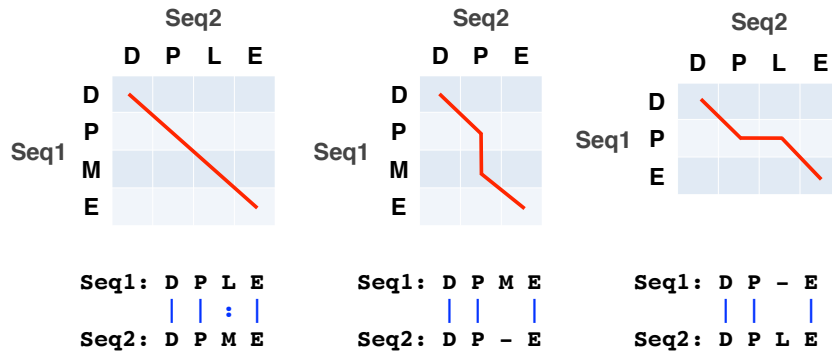
**BIMM 143**  
**Advanced Database Searching**  
**Lecture 3**  
**Barry Grant**  
**UC San Diego**  
<http://thegrantlab.org/bimm143>

## Recap From Last Time:

- Sequence alignment is a fundamental operation underlying much of bioinformatics.
- Introduced **dot matrices**, **dynamic programming** and the **BLAST heuristic** approaches.
  - *Key point:* Even when optimal solutions can be obtained they are not necessarily unique or reflective of the biologically correct alignment.
- Introduced classic **global** and **local alignment** algorithms (Needleman–Wunsch and Smith–Waterman) and their major application areas.
- **Heuristic approaches** are necessary for large database searches and many genomic applications.

[Feedback](#)

## Muddy Point: Different paths represent different alignments



(Mis)matches are represented by diagonal paths & Indels with horizontal or vertical path segments

## Today's Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns
- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations
- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity
- **Hidden Markov models (HMMs):** More versatile probabilistic model for detection of remote similarities



**N.B.** BLOSUM62 does not take the local context of a particular position into account

(i.e. all like substitutions are scored the same regardless of their location in the molecules).

We will revisit this later...

# Today's Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns
- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations
- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity
- **Hidden Markov models (HMMs):** More versatile probabilistic model for detection of remote similarities

## Functional cues from conservation patterns

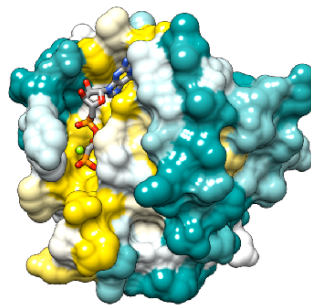
Within a protein or nucleic acid sequence there may be a small number of characteristic residues that occur consistently. These conserved "sequence fingerprints" (or **motifs**) usually contain functionally important elements

- E.g., the amino acids that are consistently found at enzyme active sites or the nucleotides that are associated with transcription factor binding sites.

ATP/GTP-binding proteins: G-x(4)-G-K-T

```

*      ***
FYGPPGLGKTSNIGG
LYGPPGLGKTANMGV
LFGPPGLGKTAHLGV
LIGPPGLGKTAALGV
LSGPPGLGKTAFMNA
ISGPIGTGKSAGIGI
LHGNPFTGKTASFSA
VCGLPGMGKTVETGF
VAGTPGVGKTVKLRP
IAGTPGVGKTVKMKF
LHGVPGTGKTMKKGY
G      GKT
  
```

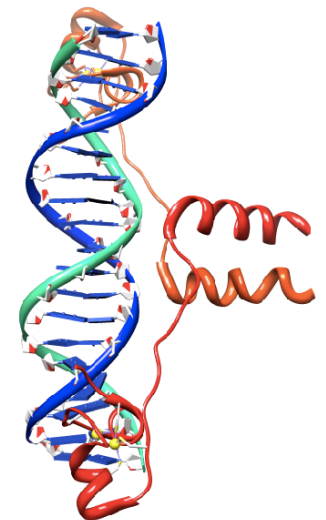
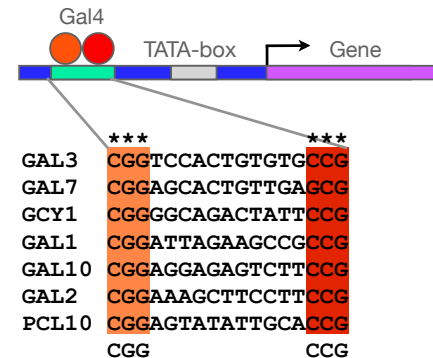


Conservation

## Functional cues from conservation patterns...

Many DNA patterns are binding sites for Transcription Factors.

- E.g., The Gal4 binding sequence  
C-G-G-N(11)-C-C-G



## Representing recurrent sequence patterns

Beyond knowledge of invariant residues we can define **position-based** representations that highlight the range of permissible residues per position.

- **Pattern:** Describes a motif using a qualitative consensus sequence (e.g., IUPAC or regular expression). N.B. Mismatches are not tolerated!

[LFI]-x-G-[PT]-P-G-x-G-K-[TS]-[AGSI]

- **Profile:** Describes a motif using quantitative information captured in a position specific scoring matrix (weight matrix). Profiles quantify similarity and often span larger stretches of sequence.
- **Logos:** A useful visual representation of sequence motifs.



Image generated by:  
[weblogo.berkeley.edu](http://weblogo.berkeley.edu)

## PROSITE is a protein pattern and profile database

Currently contains > 1790 patterns and profiles: <http://prosite.expasy.org/>  
Example PROSITE patterns:

PS00087; SOD\_CU\_ZN\_1  
[GA]-[IMFAT]-H-[LIVF]-H-{S}-x-[GP]-[SDG]-x-[STAGDE]  
The two Histidines are copper ligands

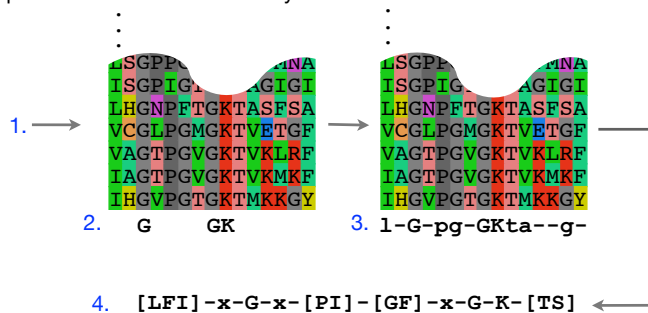
- Each position in the pattern is separated with a hyphen
- x can match any residue
- [ ] are used to indicate ambiguous positions in the pattern  
• e.g., [SDG] means the pattern can match S, D, or G at this position
- { } are used to indicate residues that are not allowed at this position  
• e.g., {S} means NOT S (not Serine)
- ( ) surround repeated residues, e.g., A(3) means AAA

Information from <http://ca.expasy.org/prosite/prosuser.html>

## Defining sequence patterns

There are four basic steps involved in defining a new PROSITE style pattern:

1. Construct a multiple sequence alignment (MSA)
2. Identify conserved residues
3. Create a core sequence pattern (i.e. *consensus sequence*)
4. Expand the pattern to improve **sensitivity** and **specificity** for detecting desired sequences - more on this shortly...



## Pattern advantages and disadvantages

### Advantages:

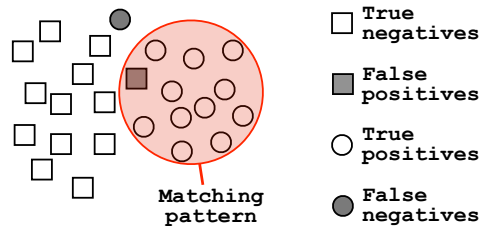
- Relatively straightforward to identify (exact pattern matching is fast)
- Patterns are intuitive to read and understand
- Databases with large numbers of protein (e.g., PROSITE) and DNA sequence (e.g., JASPER and TRANSFAC) patterns are available.

### Disadvantages:

- Patterns are qualitative and *deterministic* (i.e., either matching or not!)
- We lose information about relative frequency of each residue at a position  
E.g., [GAC] vs 0.6 G, 0.28 A, and 0.12 C
- Can be difficult to write complex motifs using regular expression notation
- Cannot represent subtle sequence motifs

## Side note: pattern sensitivity, specificity, and PPV

In practice it is not always possible to define one single regular expression type pattern which matches all family sequences (*true positives*) while avoiding matches in unrelated sequences (*true negatives*).



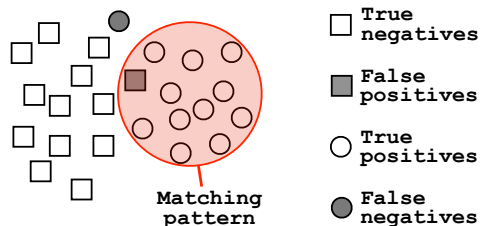
$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{PPV} = \frac{TP}{TP+FP}$$

The positive predictive value (or PPV) assesses how big a proportion of the sequences matching the pattern are actually in the family of interest. (i.e., the probability that a positive result is truly positive!)

## Side note: pattern sensitivity, specificity, and PPV

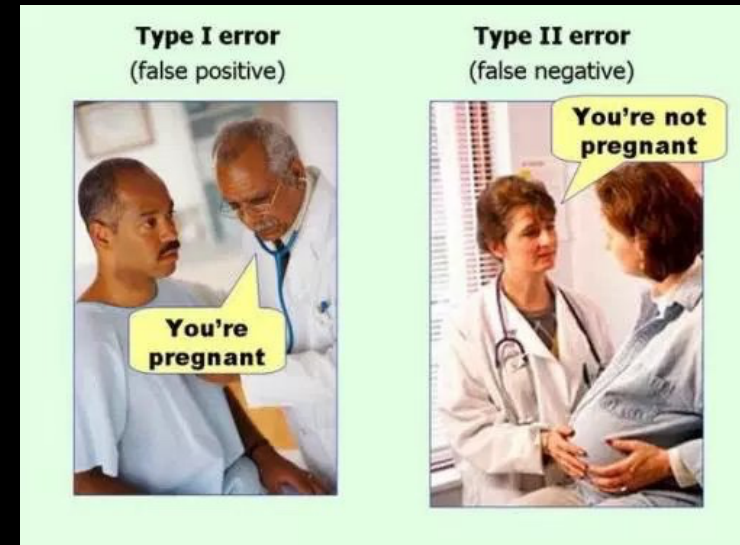


$$\text{Sensitivity} = \frac{TP}{TP+FN} = \text{Fraction of total circles we found (i.e. things we want!)}$$

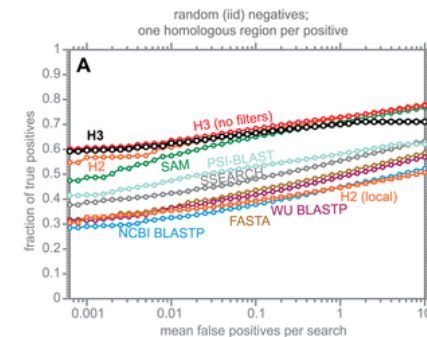
$$\text{Specificity} = \frac{TN}{TN+FP} = \text{Fraction of total squares we missed (i.e. things we don't want!)}$$

$$\text{PPV} = \frac{TP}{TP+FP} = \text{Fraction of our highlighted matches that are actually circles (i.e. proportion of the things we found that are what we want!)}$$

ROC plot example



## ROC plot of sequence searching performance...



H3 (HMMER3) has a much higher search sensitivity and specificity than BLASTP

In each benchmark, true positive subsequences have been selected to be no more than 25% identical to any sequence in the query alignment ... (see paper for details).

See: Eddy (2011) PLoS Comp Biol 7(10): e1002195



# Today's Menu

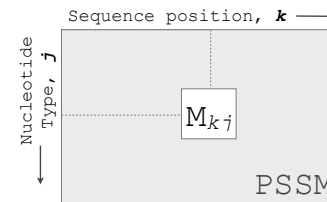
- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns
- **Sequence profiles and position specific scoring matrices (PSSMs):** Building and searching with profiles, Their advantages and limitations
- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity
- **Hidden Markov models (HMMs):** More versatile probabilistic model for detection of remote similarities

## Sequence profiles

A sequence profile is a **position-specific scoring matrix** (or **PSSM**, often pronounced 'possum') that gives a *quantitative* description of a sequence motif.

Unlike deterministic patterns, profiles assign a score to a query sequence and are widely used for database searching.

A simple PSSM has as many columns as there are positions in the alignment, and either 4 rows (one for each DNA nucleotide) or 20 rows (one for each amino acid).



$$M_{kj} = \log \left( \frac{P_{kj}}{p_j} \right)$$

- $M_{kj}$  score for the  $j$ th nucleotide at position  $k$
- $P_{kj}$  probability of nucleotide  $j$  at position  $k$
- $p_j$  "background" probability of nucleotide  $j$

See Gibskov et al. (1987) PNAS 84, 4355

## Computing a transcription factor bind site PSSM

```

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTCGGAAA
TCTATTTTGGAAA
CCAAATTTTCAAAA
    
```

Alignment Counts Matrix:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0
Consensus:	C	C	[ACT]	A	[AT]	T	T	N	G	N	A	[AT]	A

$$M_{kj} = \log \left( \frac{P_{kj}}{p_j} \right) \quad p_{kj} = \frac{C_{kj} + p_j}{Z + 1}$$

- $C_{kj}$  Number of  $j$ th type nucleotide at position  $k$
- $Z$  Total number of aligned sequences
- $p_j$  "background" probability of nucleotide  $j$
- $P_{kj}$  probability of nucleotide  $j$  at position  $k$

$$M_{kj} = \log \left( \frac{C_{kj} + p_j / Z + 1}{p_j} \right)$$

Adapted from Hertz and Stormo, Bioinformatics 15:563-577

## Computing a transcription factor bind site PSSM...

Alignment Matrix:  $C_{kj}$

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0

$$k=1, j=A: M_{kj} = \log \left( \frac{C_{kj} + p_j / Z + 1}{p_j} \right) = \log \left( \frac{0 + 0.25 / 10 + 1}{0.25} \right) = -2.4$$

$$k=1, j=C: M_{kj} = \log \left( \frac{C_{kj} + p_j / Z + 1}{p_j} \right) = \log \left( \frac{9 + 0.25 / 10 + 1}{0.25} \right) = 1.2$$

$$k=1, j=T: M_{kj} = \log \left( \frac{C_{kj} + p_j / Z + 1}{p_j} \right) = \log \left( \frac{1 + 0.25 / 10 + 1}{0.25} \right) = -0.8$$

PSSM:  $M_{kj}$

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	-2.4	-2.4	0.8	1.3	0.6	-2.4	-0.8	0.6	-2.4	0.2	1.3	1.1	1.3
C:	1.2	1.3	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	0.6	1.3	1.2	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

## Scoring a test sequence

Query Sequence

**CCTATTAGGATA**

PSSM:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	-2.4	-2.4	0.8	1.3	0.6	-2.4	-0.8	0.6	-2.4	0.2	1.3	1.1	1.3
C:	1.2	1.3	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	0.6	1.3	1.2	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

Test seq: C C T A T T T A G G A T A

$$\begin{aligned} \text{Query Score} &= 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2 \\ &\quad + 0.6 + 1.2 + 0.6 + 1.3 + -0.2 + 1.3 \\ &= 11.9 \end{aligned}$$

## Scoring a test sequence

Query Sequence

**CCTATTAGGATA**

PSSM:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	-2.4	-2.4	0.8	1.3	0.6	-2.4	-0.8	0.6	-2.4	0.2	1.3	1.1	1.3
C:	1.2	1.3	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	0.6	1.3	1.2	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

Test seq: C C T A T T T A G G A T A

$$\begin{aligned} \text{Query Score} &= 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2 \\ &\quad + 0.6 + 1.2 + 0.6 + 1.3 + -0.2 + 1.3 \\ &= 11.9 \end{aligned}$$

Q. Does the query sequence match the DNA sequence profile?

## Scoring a test sequence...

Query Sequence

**CCTATTAGGATA**

Best Possible Sequence

**CCAATTTAGGAAA**

PSSM:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	-2.4	-2.4	0.8	1.3	0.6	-2.4	-0.8	0.6	-2.4	0.2	1.3	1.1	1.3
C:	1.2	1.3	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-2.4	-2.4
T:	-0.8	-2.4	0.2	-2.4	0.6	1.3	1.2	-0.2	-2.4	-0.8	-2.4	-0.2	-2.4

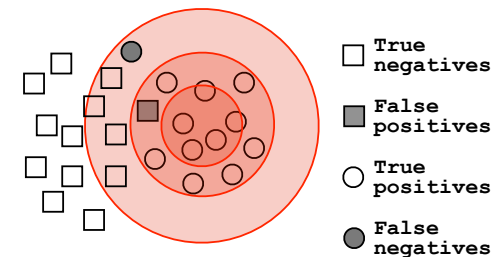
Max Score: C C A A T T T A G G A A A

$$\begin{aligned} \text{Max Score} &= 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2 \\ &\quad + 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3 \\ &= 13.8 \end{aligned}$$

A. Following method in Harbison *et al.* (2004) Nature 431:99-104  
 Heuristic threshold for match = 60% x Max Score = (0.6 x 13.8 = 8.28);  
 11.9 > 8.28; Therefore our query is a potential TFBS!

## Picking a threshold for PSSM matching

Again, you want to select a threshold that **minimizes FPs** (e.g., how many shuffled or random sequences does the PSSM match with that score) and **minimizes FNs** (e.g., how many of the 'real' sequences are missed with that score).



FP=0, FN=7, TP=5  
 FP=1, FN=1, TP=11  
 FP=5, FN=0, TP=12

$5/(5+0) = 1$   
 $11/(11+1) = 0.92$   
 $12/(12+5) = 0.71$

Q. Which threshold has the best PPV (TP/(TP+FP)) ?

## Searching for PSSM matches

If we do not allow gaps (i.e., no insertions or deletions):

- Perform a linear scan, scoring the match to the PSSM at each position in the sequence - the “sliding window” method



If we allow gaps:

- Can use dynamic programming to align the profile to the protein sequence(s) (with gap penalties)  
We will discuss PSI-BLAST shortly...  
see Mount, Bioinformatics: sequence and genome analysis (2004)
- Can use hidden Markov Model-based methods  
We will cover HMMs in the next lecture...  
see Durbin et al., Biological Sequence Analysis (1998)

## Side note: Building PSSMs from unaligned sequences

Patterns and profiles are most often built on the basis of known site equivalences (i.e. from a pre-calculated MSA).

However, a number of programs have been developed that employ local multiple alignments to search for common sequence elements in unaligned sequences.



Gibbs *sampling* methods:

Motif Sampler - <http://bayesweb.wadsworth.org/gibbs/gibbs.html>

AlignAce - <http://atlas.med.harvard.edu/cgi-bin/alignace.pl>

*Expectation maximization* method:

MEME - <http://meme.sdsc.edu/>

See: Lawrence et al. (1993) Science. 262, 208-14

## Side note: Profiles software and databases

**Pftools** is a package to build and search with profiles,  
<http://www.isrec.isb-sib.ch/ftp-server/pftools/>

The package contains (among other programs):

- ▶ **pfmake** for building a profile starting from multiple alignments
- ▶ **pfsearch** to search a protein database with a profile
- ▶ **pfscan** to search a profile database with a protein

**PRINTS** database of PSSMs

<http://bioinf.man.ac.uk/dbbrowser/PRINTS>

Collection of conserved motifs used to characterize a protein

- ▶ Uses fingerprints (conserved motif groups).
- ▶ Very good to describe sub-families.

**BLOCKS** is another PSSMs database similar to prints

<http://www.blocks.fhcrc.org>

**ProDom** is collection of protein motifs obtained automatically using PSI-BLAST

<http://prodes.toulouse.inra.fr/prodom/doc/prodom.html>

## Side note: Profiles software and databases...

**InterPro** is an attempt to group a number of protein domain databases.  
<http://www.ebi.ac.uk/interpro>

It currently includes:

- ▶ Pfam
- ▶ PROSITE
- ▶ PRINTS
- ▶ ProDom
- ▶ SMART
- ▶ TIGRFAMs

- InterPro tries to have and maintain a high quality of annotation
- The database and a stand-alone package (**iprscan**) are available for UNIX platforms, see:  
<ftp://ftp.ebi.ac.uk/pub/databases/interpro>





# PSI-BLAST: Position specific iterated BLAST

- The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query
- PSI-BLAST constructs a multiple sequence alignment from the results of a first round BLAST search and then creates a “profile” or specialized **position-specific scoring matrix (PSSM)** for subsequent search rounds

37

## Inspect the blastp output to identify empirical “rules” regarding amino acids tolerated at each position

```

730496 66 FTVDENGQMSATAKGRVRLFNWVDCADMIGSFDTEDPAKFKMKYUGVASFLQKGNDDH 125
200679 63 FSVDEKGHMSATAKGRVRLSNWEVCADMVGTFTDTEDEPAKFKMKYUGVASFLQKGNDDH 122
206589 34 FSVDEKGHMSATAKGRVRLSNWEVCADMVGTFTDTEDEPAKFKMKYUGVASFLQKGNDDH 93
2136812 2 MSATAKGRVRLSNWVDCADMVGTFTDTEDEPAKFKMKYUGVASFLQKGNDDH 53
132408 65 FKIEDNGKTTATAKGRVRLDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584 44 FSVDESGKVTATAHGRVILNWNWEMCANMFGTFEDTPDPAKFKMRYUGAASYLQKGNDDH 103
267585 44 FSVDESGKVTATAQGRVILNWNWEMCANMFGTFEDTPDPAKFKMRYUGAASYLQKGNDDH 103
8777608 63 FTIHEDGAMTATAKGRVILNWNWEMCADMMATFETTPDPAKFKMRYUGAASYLQKGNDDH 122
6687453 60 FKVEEDGTMNTATAIGRVIILNWNWEMCANMFGTFEDTEDEPAKFKMKYUGAASYLQKGYDDH 119
10697027 81 FKVQEDGTMNTATATGRVILNWNWEMCANMFGTFEDTEEPARFKMKYUGAASYLQKGYDDH 140
13645517 1 MVGTFTDTEDEPAKFKMKYUGVASFLQKGNDDH 32
13925316 38 FSVDESGKMTATAQGRVILNWNWEMCANMFGTFEDTPDPAKFKMRYUGAASYLQKGNDDH 97
131649 65 YTVDEEDGTMNTASSKGRVKLFGFWVICADMAAQYDPTTPAKMYNTYQGLASVLSGGDNY 126
    
```

R,I,K
C
D,E,T
K,R,T
N,L,Y,G

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2 K	-1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
4 V	0	-3	-3	-4	-1	-3	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3
6 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9 L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0
13 W	-2	-2	-3	-4	-4	-4	-4	-4	-4	-3	2	1	-3	-3	-2	7	0	0	0	0
14 A	3	-2	-2	-2	-2	-2	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1	-1	-1
15 A	2	-2	-2	-2	-2	-2	-2	-2	-2	0	-2	-3	-1	3	0	-3	-2	-2	-2	-2
16 A	4	-2	-2	-2	-1	-1	0	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1	-1	-1
...																				
37 S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38 G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39 T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	9	2	-3
41 Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42 A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

All the amino acids from position 1 to N (the end of your query protein)

20 amino acids

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4 V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9 L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0
13 W	-2	-3	-4	-4	-4	-4	-4	-4	-4	-3	2	1	-3	-3	-2	7	0	0	0	0
14 A	3	-2	-2	-2	-2	-2	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1	-1	-1
15 A	2	-2	-2	-2	-2	-2	-2	-2	-2	0	-2	-3	-1	3	0	-3	-2	-2	-2	-2
16 A	4	-2	-2	-2	-1	-1	0	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1	-1	-1
...																				
37 S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38 G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39 T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	9	2	-3
41 Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42 A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

Note: A given amino acid (such as alanine) in your query protein can receive different scores for matching alanine depending on the position in the protein (BLOSUM SAA = +4)

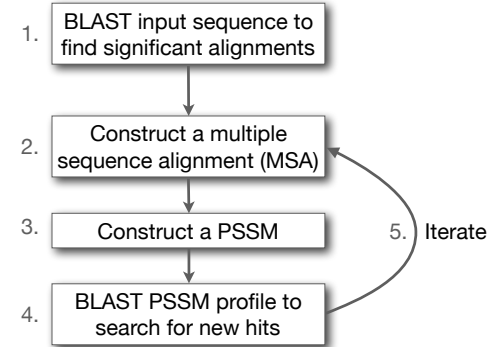
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M																				
2 K																				
3 W																				
4 V																				
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6 A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9 L	-1	-3	-4	-4											3	-3	-1	-2	-1	2
10 L	-2	-2	-4	-4											3	-3	-1	-2	-1	1
11 A	5	-2	-2	-2											-1	1	0	-3	-2	0
12 A	5	-2	-2	-2											-1	1	0	-3	-2	0
13 W	-2	-3	-4	-4											-3	-3	-2	7	0	0
14 A	3	-2	-1	-2											-1	1	-1	-3	-3	-1
15 A	2	-1	0	1											-1	3	0	-3	-2	-2
16 A	4	-2	-1												-1	1	0	-3	-2	-1
...																				
37 S	2	-1	0	-1											-1	4	1	-3	-2	-2
38 G	0	-3	-1	-2											-2	0	-2	-3	-3	-4
39 T	0	-1	0	-1											-1	1	5	-3	-2	0
40 W	-3	-3	-4	-5											-4	-3	-3	9	2	-3
41 Y	-2	-2	-2	-3											3	-2	-2	2	7	-1
42 A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

The PSI-BLAST PSSM is essentially a query customized scoring matrix that is more sensitive than BLOSUM.

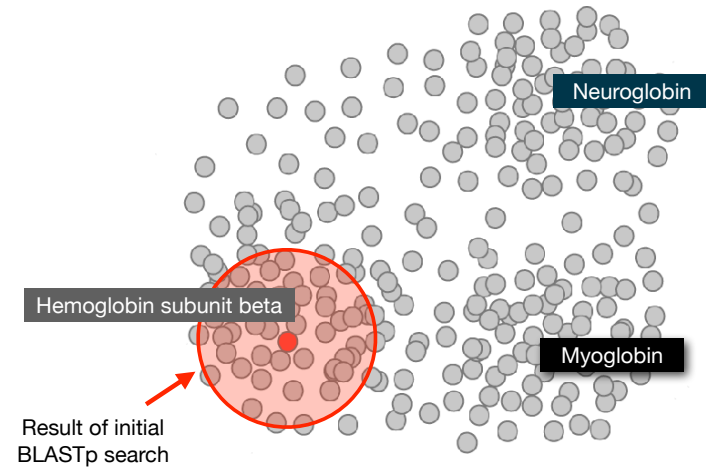
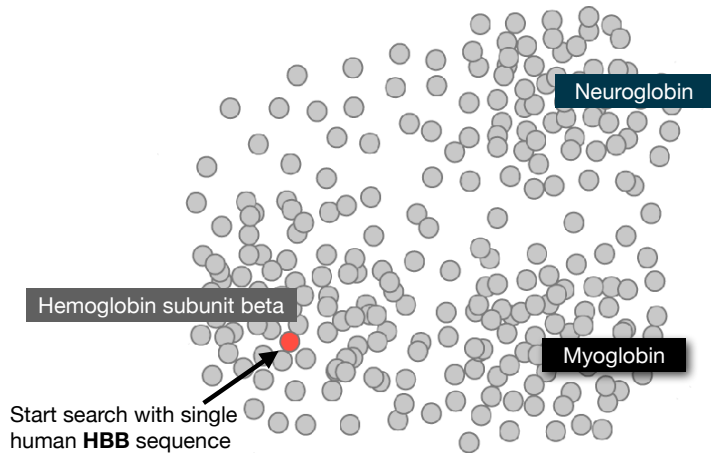
Note: A given amino acid (such as alanine) in your query protein can receive different scores for matching alanine depending on the position in the protein (BLOSUM SAA = +4)

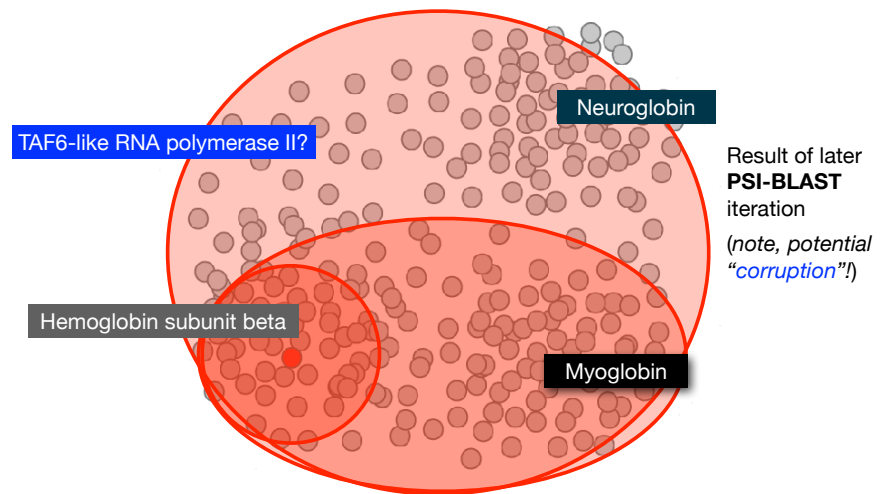
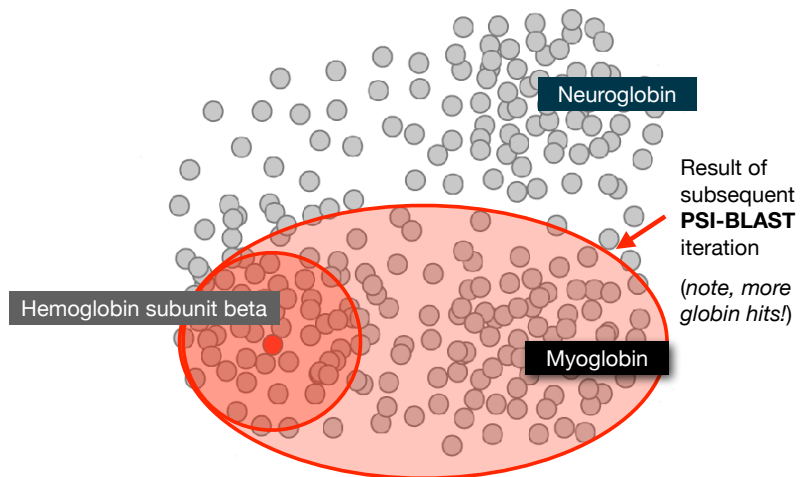
## PSI-BLAST: Position-Specific Iterated BLAST

Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)





Description	Max score	Total score	Query cover	E value	Ident	Accession
hemoglobin subunit beta [Homo sapiens]	301	301	100%	2e-106	100%	<a href="#">NP_000509.1</a>
hemoglobin subunit delta [Homo sapiens]	284	284	100%	7e-100	93%	<a href="#">NP_000510.1</a>
hemoglobin subunit epsilon [Homo sapiens]	240	240	100%	2e-82	76%	<a href="#">NP_005321.1</a>
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73%	<a href="#">NP_000175.1</a>
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	3e-79	73%	<a href="#">NP_000550.2</a>
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43%	<a href="#">NP_000508.1</a>
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	36%	<a href="#">NP_005323.1</a>

1

Description	Max score	Total score	Query cover	E value	Ident	Accession
hemoglobin subunit beta [Homo sapiens]	301	301	100%	2e-106	100%	<a href="#">NP_000509.1</a>
hemoglobin subunit delta [Homo sapiens]	284	284	100%	7e-100	93%	<a href="#">NP_000510.1</a>
hemoglobin subunit epsilon [Homo sapiens]	240	240	100%	2e-82	76%	<a href="#">NP_005321.1</a>
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73%	<a href="#">NP_000175.1</a>
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	3e-79	73%	<a href="#">NP_000550.2</a>
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43%	<a href="#">NP_000508.1</a>
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	36%	<a href="#">NP_005323.1</a>

1

myoglobin [Homo sapiens]	80.5	80.5	97%	2e-19	26%	<a href="#">NP_005359.1</a>
neuroglobin [Homo sapiens]	54.7	54.7	92%	2e-09	23%	<a href="#">NP_067080.1</a>

2

**New relevant globins found only by PSI-BLAST**

Description	Max score	Total score	Query cover	E value	Ident	Accession
hemoglobin subunit beta [Homo sapiens]	301	301	100%	2e-106	100%	NP_000509.1
hemoglobin subunit delta [Homo sapiens]	284	284	100%	7e-100	93%	NP_000510.1
hemoglobin subunit epsilon [Homo sapiens]	240	240	100%	2e-82	76%	NP_005321.1
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73%	NP_000175.1
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	3e-79	73%	NP_000550.2
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43%	NP_000508.1
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	36%	NP_005323.1
myoglobin [Homo sapiens]	80.5	80.5	97%	2e-19	26%	NP_005359.1
neuroglobin [Homo sapiens]	54.7	54.7	92%	2e-09	23%	NP_067080.1
myoglobin [Homo sapiens]	159	159	97%	3e-50	26%	NP_005359.1
hemoglobin subunit alpha [Homo sapiens]	151	151	97%	3e-47	42%	NP_000508.1
hemoglobin subunit mu [Homo sapiens]	147	147	97%	6e-46	35%	NP_001003938.1
hemoglobin subunit theta-1 [Homo sapiens]	147	147	97%	2e-45	37%	NP_005322.1
neuroglobin [Homo sapiens]	134	134	92%	3e-40	23%	NP_067080.1
PREDICTED: cytoglobin isoform X2 [Homo sapiens]	115	115	66%	3e-33	25%	XP_016879605.1
PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_011523942.1
PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_005258156.1

Inclusion of irrelevant hits can lead to PSSM corruption

Description	Max score	Total score	Query cover	E value	Ident	Accession
hemoglobin subunit beta [Homo sapiens]	301	301	100%	2e-106	100%	NP_000509.1
hemoglobin subunit delta [Homo sapiens]	284	284	100%	7e-100	93%	NP_000510.1
hemoglobin subunit epsilon [Homo sapiens]	240	240	100%	2e-82	76%	NP_005321.1
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73%	NP_000175.1
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	3e-79	73%	NP_000550.2
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43%	NP_000508.1
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	36%	NP_005323.1
myoglobin [Homo sapiens]	80.5	80.5	97%	2e-19	26%	NP_005359.1
neuroglobin [Homo sapiens]	54.7	54.7	92%	2e-09	23%	NP_067080.1
myoglobin [Homo sapiens]	159	159	97%	3e-50	26%	NP_005359.1
hemoglobin subunit alpha [Homo sapiens]	151	151	97%	3e-47	42%	NP_000508.1
hemoglobin subunit mu [Homo sapiens]	147	147	97%	6e-46	35%	NP_001003938.1
hemoglobin subunit theta-1 [Homo sapiens]	147	147	97%	2e-45	37%	NP_005322.1
neuroglobin [Homo sapiens]	134	134	92%	3e-40	23%	NP_067080.1
PREDICTED: cytoglobin isoform X2 [Homo sapiens]	115	115	66%	3e-33	25%	XP_016879605.1
PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_011523942.1
PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_005258156.1

Score and E value depends on PSSM

## PSI-BLAST is performed in five steps

- A normal blastp search uses a scoring matrix (e.g., BLOSUM62) to perform pairwise alignments of your query sequence (such as RBP) against the database. PSI-BLAST also begins with a protein query that is searched against a database of choice.
- PSI-BLAST constructs a multiple sequence alignment (MSA) from an initial blastp-like search. It then creates a **PSSM** based on that multiple alignment.
- This **PSSM** is then used as a query to search the database again.
- PSI-BLAST estimates the statistical significance of the database matches, essentially using the parameters we described for gapped alignments.
- The search process is continued iteratively, typically 3 to 5 times. At each step a new PSSM is built.

## PSI-BLAST returns dramatically more hits

You must decide how many iterations to perform and which sequences to include!

You can stop the search process at any point - typically whenever few new results are returned or when no new "sensible" results are found.

Iteration	Hits with E < 0.005	Hits with E > 0.005
1	34	61
2	314	79
3	416	57
4	432	50
5	432	50

Human retinol-binding protein 4 (RBP4; P02753) was used as a query in a PSI-BLAST search of the RefSeq database.

## Example PSI-BLAST PSSM at iteration 3

The PSI-BLAST PSSM is essentially a query customized scoring matrix that is more sensitive than BLOSUM (e.g. BLOSUM  $S_{AA} = +4$ )

		20 amino acids types																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Query residues/positions	1 M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
	2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
	3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	1.2	2	-3
	4 V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
	5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	1.2	2	-3
	6 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0
	7 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
	8 L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
	9 L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
	10 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
	11 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0
	12 A	5	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0
	13 W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
	14 A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
	15 A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
	16 A	4	-2	-1	-2	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1	1
...																					
37 S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2	
38 G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	
39 T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0	
40 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	1.2	2	-3	
41 Y	-2	-2	-2	-3	-3	-2	-2	-3	2	2	-1	-2	-1	3	-3	-2	-2	2	7	-1	
42 A	4	-2	-2	-2	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0	0	
...																					

## PSI-BLAST errors: the corruption problem

The main source of error in PSI-BLAST searches is the spurious amplification of sequences that are unrelated to the query.

There are three main approaches to stopping corruption of PSI-BLAST queries:

- Perform multi-domain splitting of your query sequence
  - If a query protein has several different domains PSI-BLAST may find database matches related to both individually. One should not conclude that these hits with different domains are related.
  - Often best to search using just one domain of interest.
- Inspect each PSI-BLAST iteration removing suspicious hits.
  - E.g., your query protein may have a generic coiled-coil domain, and this may cause other proteins sharing this motif (such as myosin) to score better than the inclusion threshold even though they are not related.
  - Use your biological knowledge!
- Lower the default expect level (e.g.,  $E = 0.005$  to  $E = 0.0001$ ).
  - This may suppress appearance of FPs (but also TPs)

## Profile advantages and disadvantages

### Advantages:

- Quantitate with a good scoring system
- Weights sequences according to observed diversity
  - Profile is specific to input sequence set
- Very sensitive
  - Can detect weak similarity
- Relatively easy to compute
  - Automatic profile building tools available

### Disadvantages:

- If a mistake enters the profile, you may end up with irrelevant data
  - The corruption problem!
- Ignores higher order dependencies between positions
  - i.e., correlations between the residue found at a given position and those found at other positions (e.g. salt-bridges, structural constraints on RNA etc...)
- Requires some expertise and oversight to use proficiently

# Today's Menu

- **Sequence motifs and patterns:** Simple approaches for finding functional cues from conservation patterns
- **Sequence profiles** and position specific scoring matrices (PSSMs): Building and searching with profiles, Their advantages and limitations
- **PSI-BLAST algorithm:** Application of iterative PSSM searching to improve BLAST sensitivity
- **Hidden Markov models (HMMs):** More versatile probabilistic model for detection of remote similarities



Your Turn!

**Hands-on sections 3 & 4:**  
**Comparing methods and the trade-off**  
**between sensitivity, selectivity and**  
**performance**

~30 mins

Problems with PSSMs: Positional dependencies

Do not capture positional dependencies

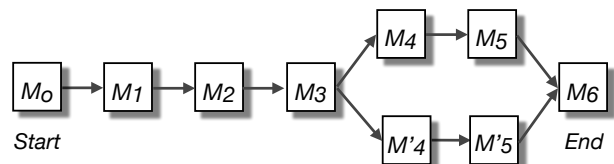
WEIRD	D				0.6
WEIRD	E	I			
WEIQH	H				0.4
WEIRD	I		I		
WEIQH	Q			0.4	
	R			0.6	
	W	I			

**Note:** We never see QD or RH, we only see RD and QH.  
 However, P(RH)=0.24, P(QD)=0.24, while P(QH)=0.16

Markov chains: Positional dependencies ✓

The connectivity or **topology** of a Markov chain can easily be designed to capture dependencies and variable length motifs.

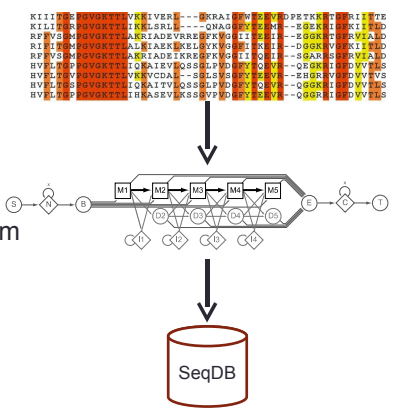
WEIRD  
 WEIRD  
 WEIQH  
 WEIRD  
 WEIQH



Recall that a PSSM for this motif would give the sequences **WEIRD** and **WEIRH** equally good scores even though the **RH** and **QR** combinations were not observed

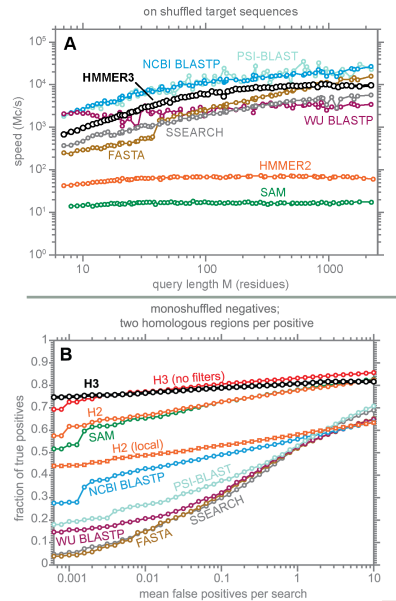
Use of HMMER

- Widely used by protein family databases
  - Use 'seed' alignments
- Until 2010
  - Computationally expensive
  - Restricted to HMMs constructed from multiple sequence alignments
- Command line application



# HMMER vs BLAST

	HMMER	BLAST
Program	<i>PHMMER</i>	<i>BLASTP</i>
Query	Single sequence	
Target Database	Sequence database	
Program	<i>HMMSCAN</i>	<i>RPSBLAST</i>
Query	Single sequence	
Target Database	Profile HMM database, e.g. Pfam	PSSM database, e.g. CDD
Program	<i>HMMSEARCH</i>	<i>PSI-BLAST</i>
Query	Profile HMM	PSSM
Target Database	Sequence database	
Program	<i>JACKHMMER</i>	<i>PSI-BLAST</i>
Query	Single sequence	
Target Database	Sequence database	



Modified from: S. R. Eddy  
PLoS Comp. Biol., 7:e1002195, 2011.



# Fast Web Searches

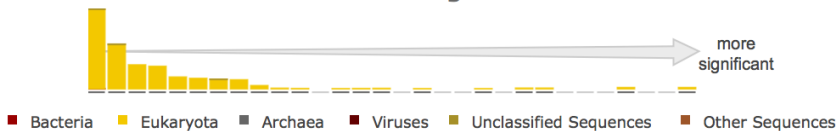
- Parallelized searches across compute farm
  - Average query returns ~1 sec
- Range of sequence databases
  - Large Comprehensive
  - Curated / Structure
  - Metagenomics
  - Representative Proteomes
- Family Annotations
  - Pfam
- Batch and RESTful API
  - Automatic and Human interface



# Visualization of Results – By Score

## Pfam Domains

### Distribution of Significant Hits



Target	Description	Species	Known Structure	Hit Positions	E-value
G3WV2S_SARHA2	Uncharacterized protein	Sarcophilus harrisi			8.0e-107



# Visualization of Results – By Score



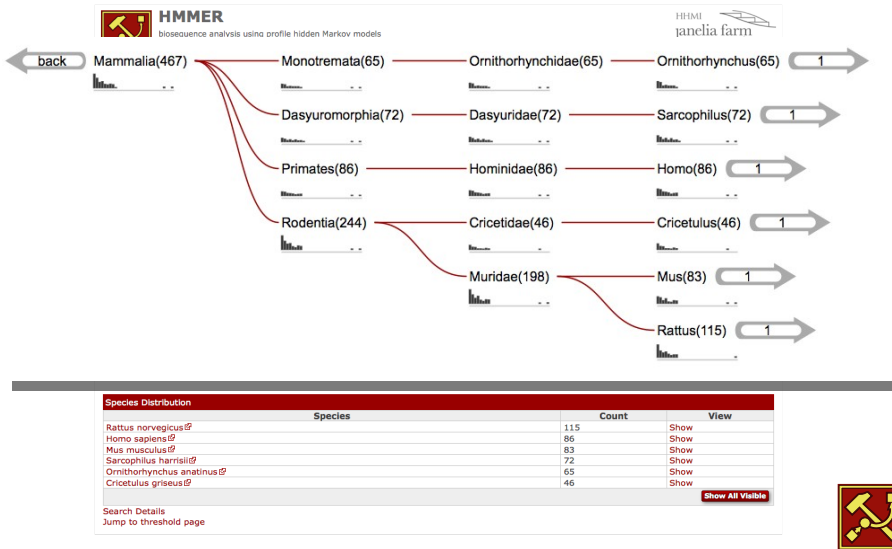
Query	Target Envelope	Target Alignment	Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value
start end	start end	start end						Ind. Cond.
7 62	4 81	9 63	0.02	0.81	36.4 (20)	50.9 (28)	19.5	0.2 0.00011

```

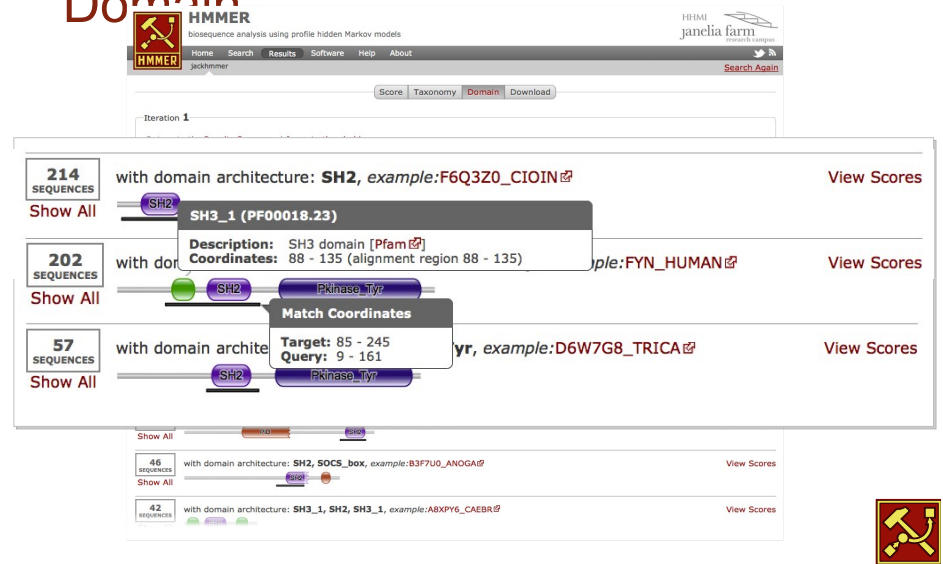
Query      7  .p.n.l.f.v.a.l.y.d.f.v.a.s.g.d.n.t.l.s.i.t.k.g.e.k.l.r.v.l.g.y.n.h.g.e.w.c.e.a.g.t.k.g.g.g.w.v.p.s.n.y.i.t 62
          d  v.a.y.d.+ a g l + k.e.+ +l + w q n g+v.p.s.n.y.+
Target    9  D.V.C.Y.V.V.A.K.Y.D.Y.A.A.Q.G.A.Q.E.L.D.L.R.K.N.E.R.Y.L.L.L.D.--D.S.K.H.W.R.V.Q.N.S.H.N.Q.S.G.V.P.S.N.Y.V.K 63
          PP  5566799*****987775..455677766516777*****96
    
```



# Visualization of Results – By Taxonomy



# Visualization of Results – By Domain



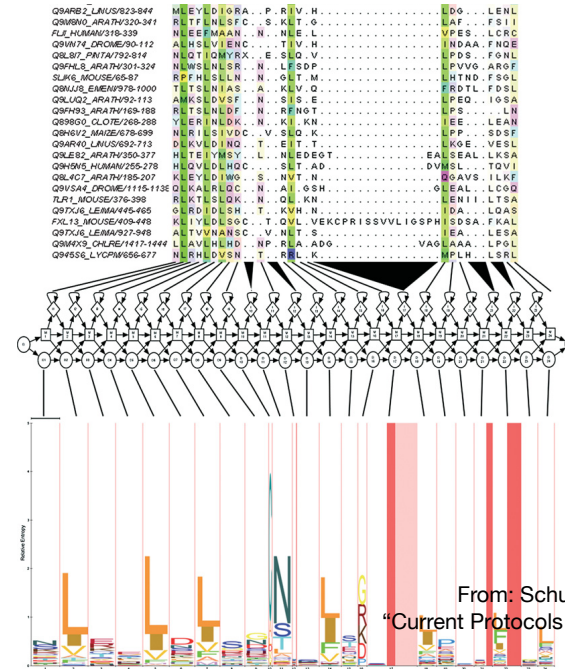
## PFAM: Protein Family Database of Profile HMMs

Comprehensive compilation of both multiple sequence alignments and profile HMMs of protein families.

<http://pfam.sanger.ac.uk/>

PFAM consists of two databases:

- **Pfam-A** is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. HMMER software is used to perform searches.
- **Pfam-B** contains additional protein sequences that are automatically aligned. Pfam-B serves as a useful supplement that makes the database more comprehensive.
- Pfam-A also contains higher-level groupings of related families, known as **clans**



From: Schuster-Bockler *et al.*  
"Current Protocols in Bioinformatics"  
Supplement 18.

## HMM limitations

HMMs are linear models and are thus **unable to capture higher order correlations** among positions (e.g. distant cysteins in a disulfide bridge, RNA secondary structure pairs, etc).

Another flaw of HMMs lies at the very heart of the mathematical theory behind these models. Namely, that the probability of a sequence can be found from the product of the probabilities of its individual residues.

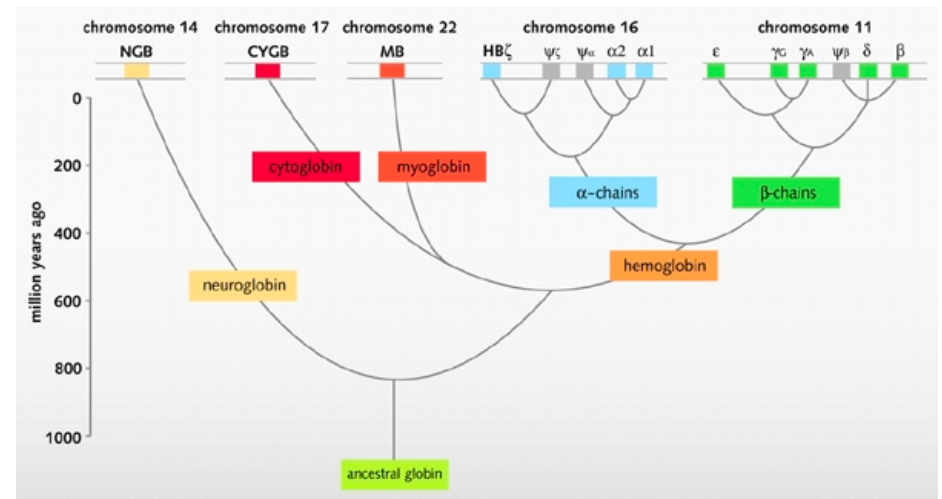
This claim is only valid if the probability of a residue is independent of the probabilities of its neighbors. In biology, there are frequently **strong dependencies between these probabilities** (e.g. hydrophobic residues clustering at the core of protein domains).

These biological realities have motivated research into new kinds of statistical models. These include hybrids of HMMs and neural nets, dynamic Bayesian nets, factorial HMMs, Boltzmann trees and stochastic context-free grammars.

See: Durbin et al. "Biological Sequence Analysis"



## Side Note: Human Globins

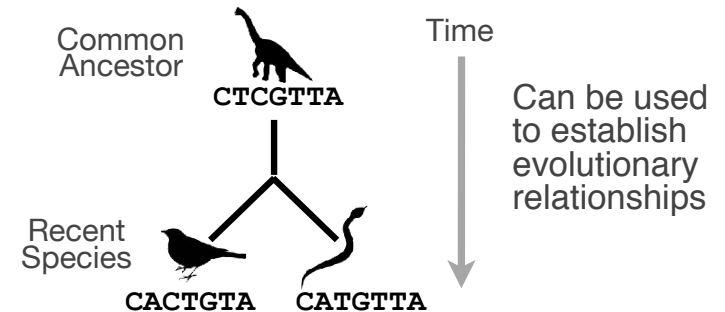


## Side Note: Orthologs vs Paralogs

Sequence comparison is most informative when it detects **homologs**

**Homologs** are sequences that have common origins *i.e.* they share a **common ancestor**

- They may or may not have common activity



74

## Key terms

When we talk about related sequences we use specific terminology.

*Homologous sequences* may be either:

- **Orthologs** or **Paralogs**

(Note. these are all or nothing relationships!)

*Any pair of sequences* may share a certain level of:

- **Identity** and/or **Similarity**

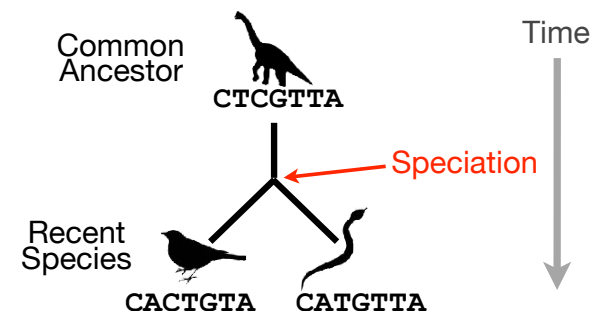
(Note. if these metrics are above a certain level we often infer homology)

75

## Orthologs tend to have similar function

**Orthologs**: are homologs produced by speciation that have diverged due to divergence of the organisms they are associated with.

- Ortho = [greek: straight] ... implies direct descent

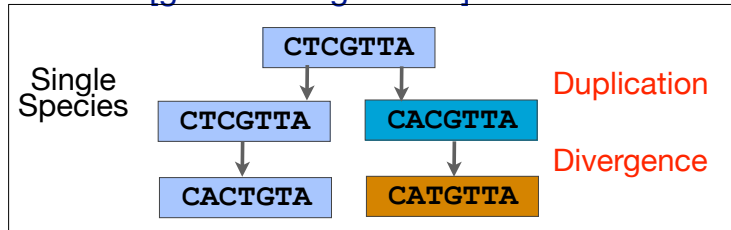


76

## Paralogs tend to have slightly different functions

**Paralogs:** are homologs produced by **gene duplication**. They represent genes derived from a common ancestral gene that  *duplicated within an organism*  and then subsequently  *diverged by accumulated mutation* .

– Para = [greek: along side of]



77

## Orthologs vs Paralogs

- In practice, determining ortholog vs paralog can be a complex problem:
  - gene loss after duplication,
  - lack of knowledge of evolutionary history,
  - weak similarity because of evolutionary distance
- **Homology does not necessarily imply exact same function**
  - may have similar function at very crude level but play a different physiological role

78