BIMM 143

Hands-on Lab Session

Class 03

Barry Grant

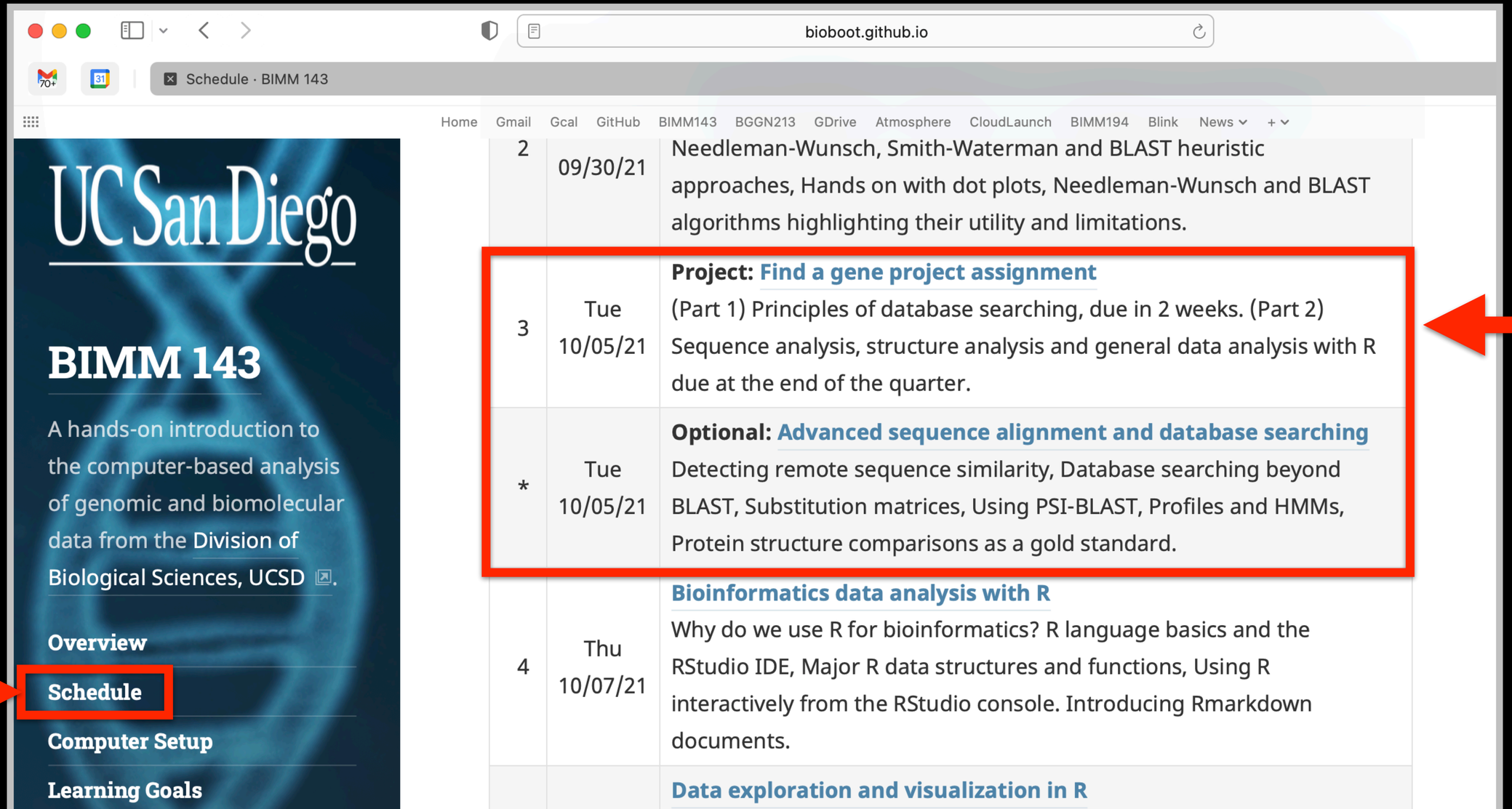
UC San Diego

http://thegrantlab.org/bimm143

Class 3: Hands-on section

Week2

http://thegrantlab.org/bimm143/



Find-a-Gene Project Assignment

- A total of 20% of the course grade will be assigned based on the "find-a-gene project assignment"

Find-a-Gene Project Assignment

- A total of 20% of the course grade will be assigned based on the "find-a-gene project assignment"
- The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.

Find-a-Gene Project Assignment

- A total of 20% of the course grade will be assigned based on the "find-a-gene project assignment"
- The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.
- You may wish to consult the scoring rubric (in the linked project description) and the example report for format and content guidance.

Find-a-Gene Project Assignment

- A total of 20% of the course grade will be assigned based on the "find-a-gene project assignment"
- The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.
- You may wish to consult the scoring rubric (in the linked project description) and the example report for format and content guidance.
  - ➡ Your responses to questions Q1-Q4 are due 12pm San Diego time on Tuesday **Oct 19th** (10/19/21).
  - ➡ The complete assignment, including responses to all questions, is due 12pm San Diego time on **Dec 2nd** (12/02/21).

Find-a-Gene Project Assignment

- A total of 20% of the course grade will be assigned based on the "find-a-gene project assignment"
- The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.
- You may wish to consult the scoring rubric (in the linked project description) and the example report for format and content guidance.

➡ Your responses to questions Q1-Q4 are due 12pm San Diego time on Tuesday **Oct 19th** (10/19/21).

➡ The complete assignment, including responses to all questions, is due 12pm San Diego time on **Dec 2nd** (12/02/21).

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called `Screen Shot [].png` in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.

If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[Q3] Gather information about this “novel” **protein**. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

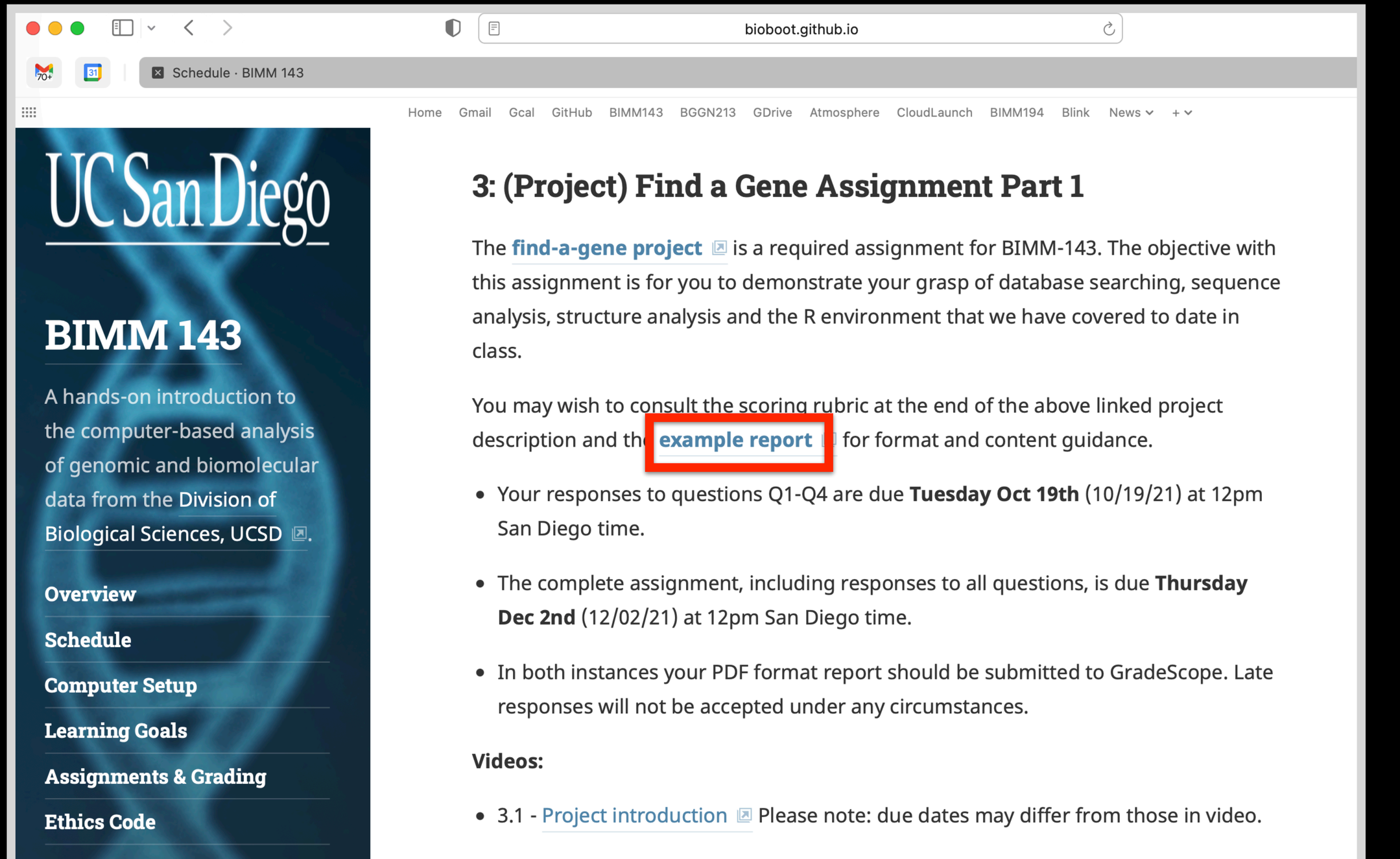
Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

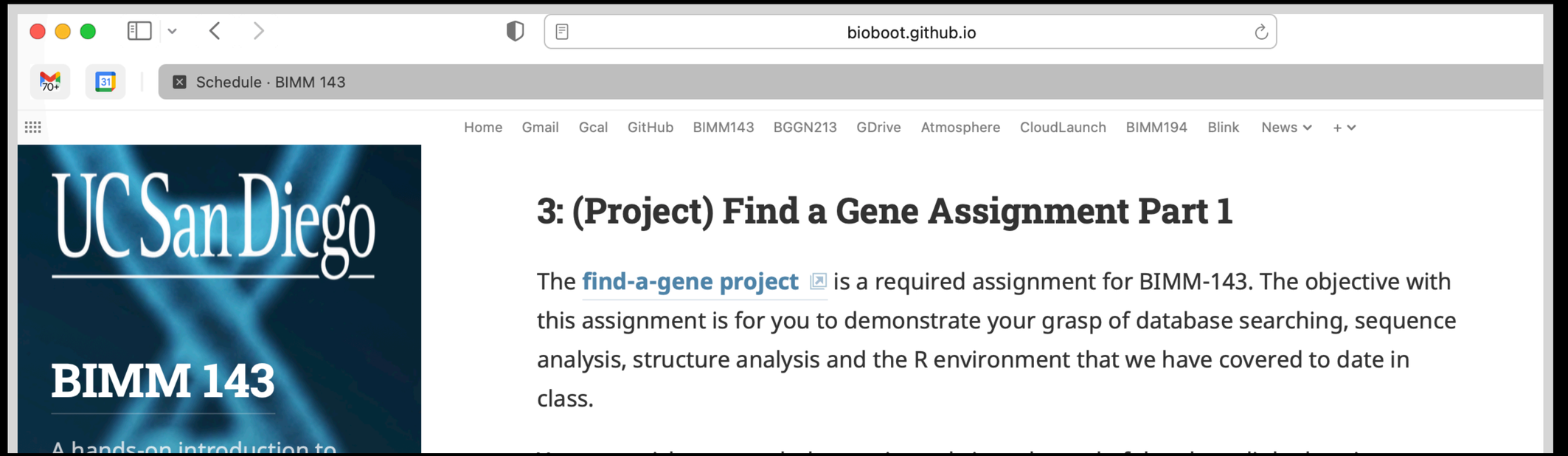
[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.


bioboot.github.io
Schedule · BIMM 143
Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News +
UC San Diego
BIMM 143
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
Overview
Schedule
Computer Setup
Learning Goals
Assignments & Grading
Ethics Code
3: (Project) Find a Gene Assignment Part 1
The find-a-gene project is a required assignment for BIMM-143. The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.
You may wish to consult the scoring rubric at the end of the above linked project description and the example report for format and content guidance.
Your responses to questions Q1-Q4 are due Tuesday Oct 19th (10/19/21) at 12pm San Diego time.
The complete assignment, including responses to all questions, is due Thursday Dec 2nd (12/02/21) at 12pm San Diego time.
In both instances your PDF format report should be submitted to GradeScope. Late responses will not be accepted under any circumstances.
Videos:
3.1 - Project introduction Please note: due dates may differ from those in video.

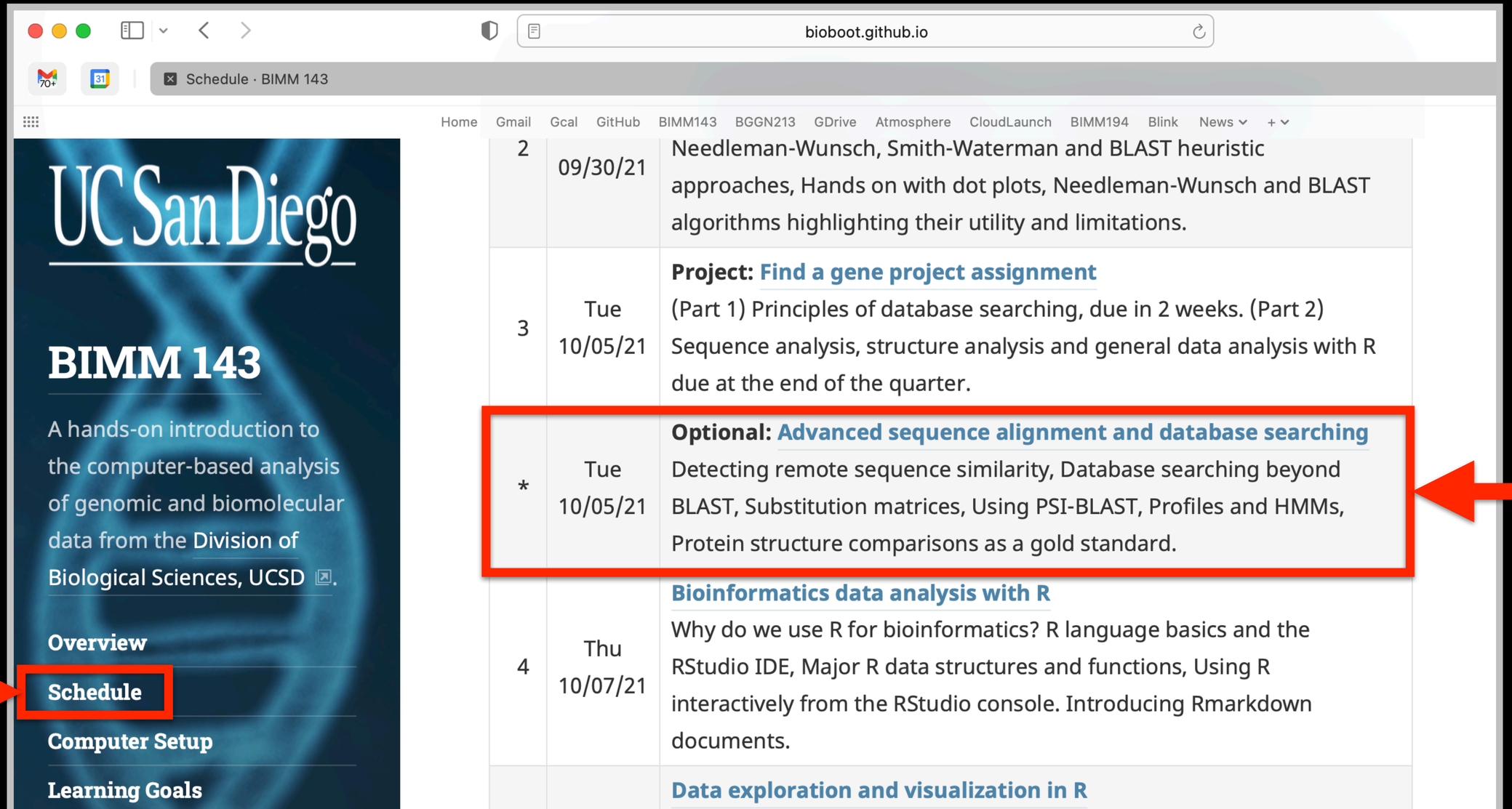


- Your responses to questions Q1-Q4 are due 12pm San Diego time on Tuesday **Oct 19th** (11/19/21).
- The complete assignment, including responses to all questions, is due 12pm San Diego time on Friday **Dec 2nd** (12/02/21).

Class 3: Hands-on section

Class 03

http://thegrantlab.org/bimm143/



R Shiny App

▶ **Details:**

Sequence 1 GATTAC

Sequence 2 GTCGACGC

Match Score	Mismatch Score	Gap Score
1	-1	-2

Compute Optimal Alignment | Clear Path | Custom Path

G T C G A C G C
G A T T A C - -
Score = -4

		G	T	C	G	A	C	G	C
	0	-2	-4	-6	-8	-10	-12	-14	-16
G	-2	1	-1	-3	-5				
A	-4	-1	0	-2	-4				
T	-6	-3	0	-1	-3	-5	-5	-7	-9
T	-8	-5	-2	-1	-2	-4	-6	-6	-8
A	-10	-7	-4	-3	-2	-1	-3	-5	-7
C	-12	-9	-6	-3	-4	-3	0	-2	-4

Score from Diagonal cell	Score from Upper cell
-6 + 1 (Due to a match between G & G) = -5	-8 + -2 (The Gap score) = -10
Score from Side cell	Winning (max) score is -5
-3 + -2 (The Gap score) = -5	

▼ **Reference:**

See the lecture and hands-on session for class 2 for a full discussion of Global, Local, and various Heuristic approaches to biomolecular sequence alignment.

Barry J Grant.

NW App Link

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:

1. Limits of using BLAST [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. Examining conservation patterns [~20 mins]

— BREAK [15 mins]—

4. [Optional] Using HMMER [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!

YOUR TURN!

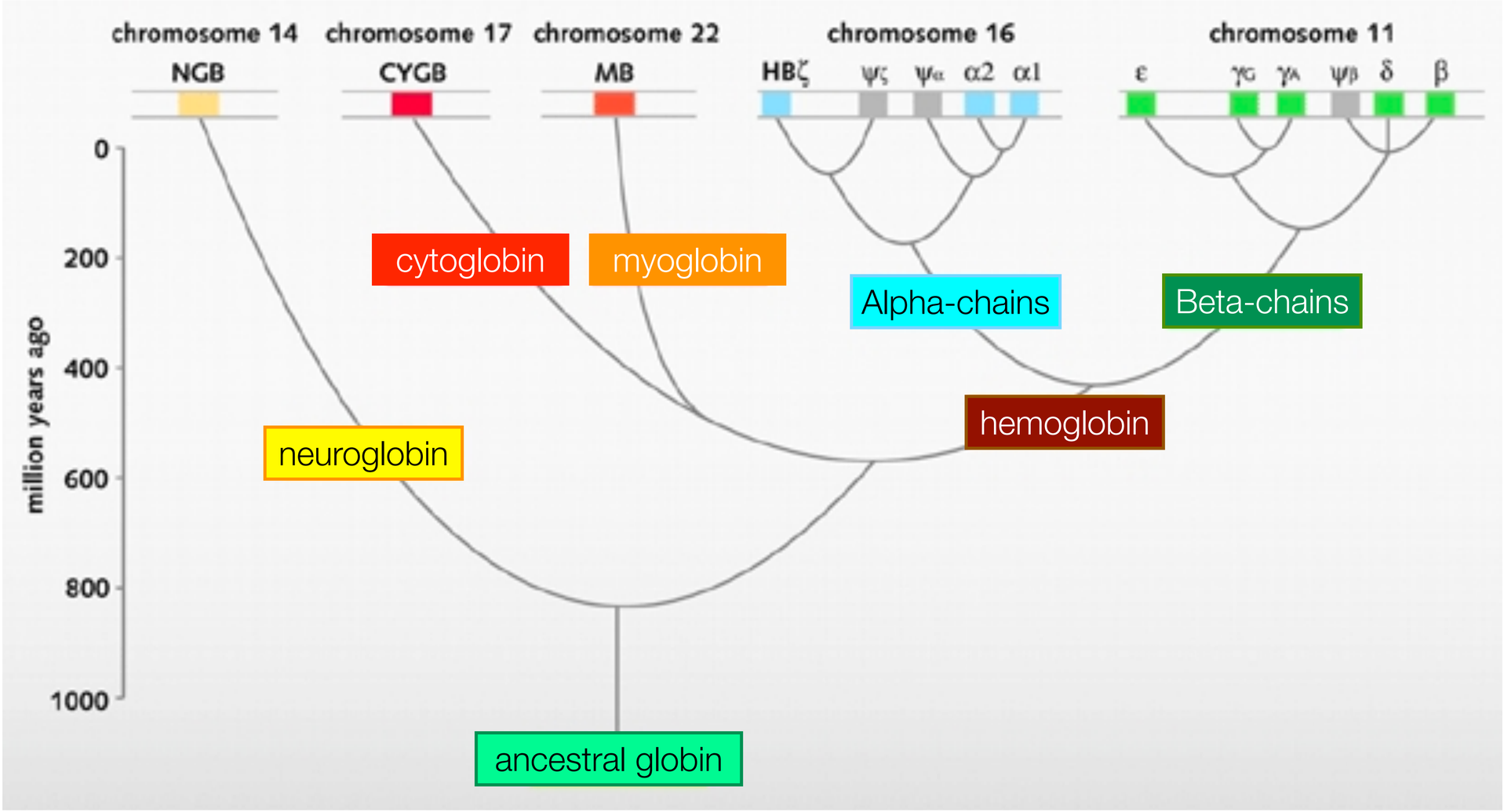
- There are **four required** and **one optional** hands-on sections including:

1. **Limits of using BLAST** [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. Examining conservation patterns [~20 mins]

— BREAK [15 mins]—

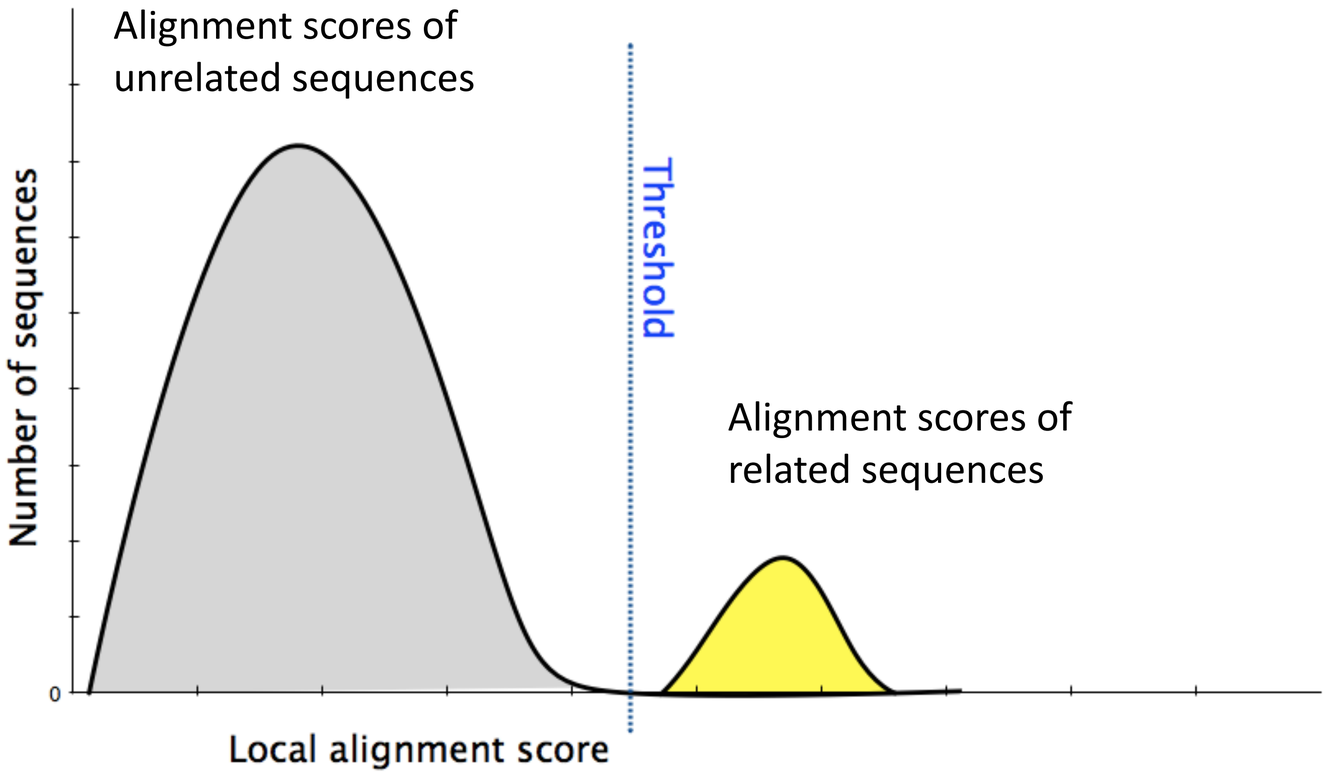
4. [Optional] Using HMMER [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!

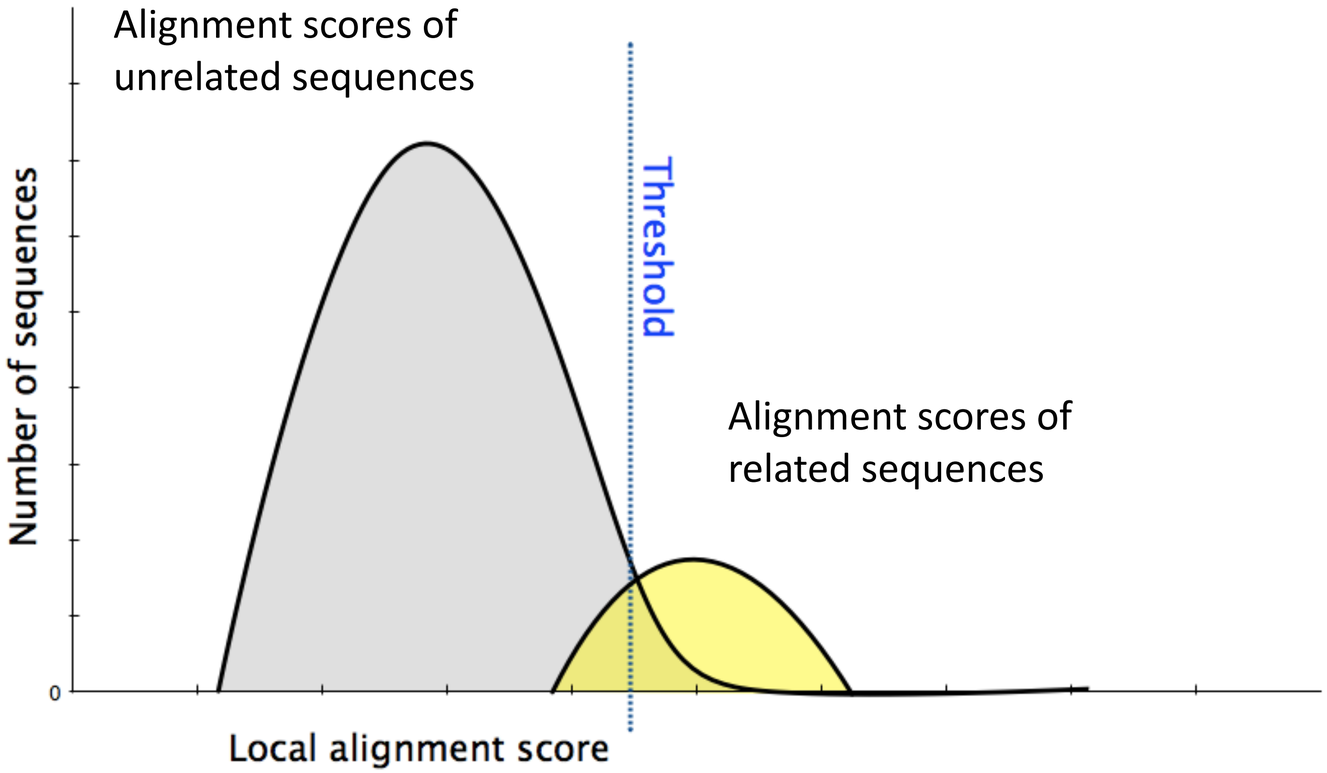


An evolutionary model of human globins.
The different locations of globin genes in human chromosomes are reported at the top of the figure, distinguishing between the functional genes (in color) and the pseudogenes (in grey).

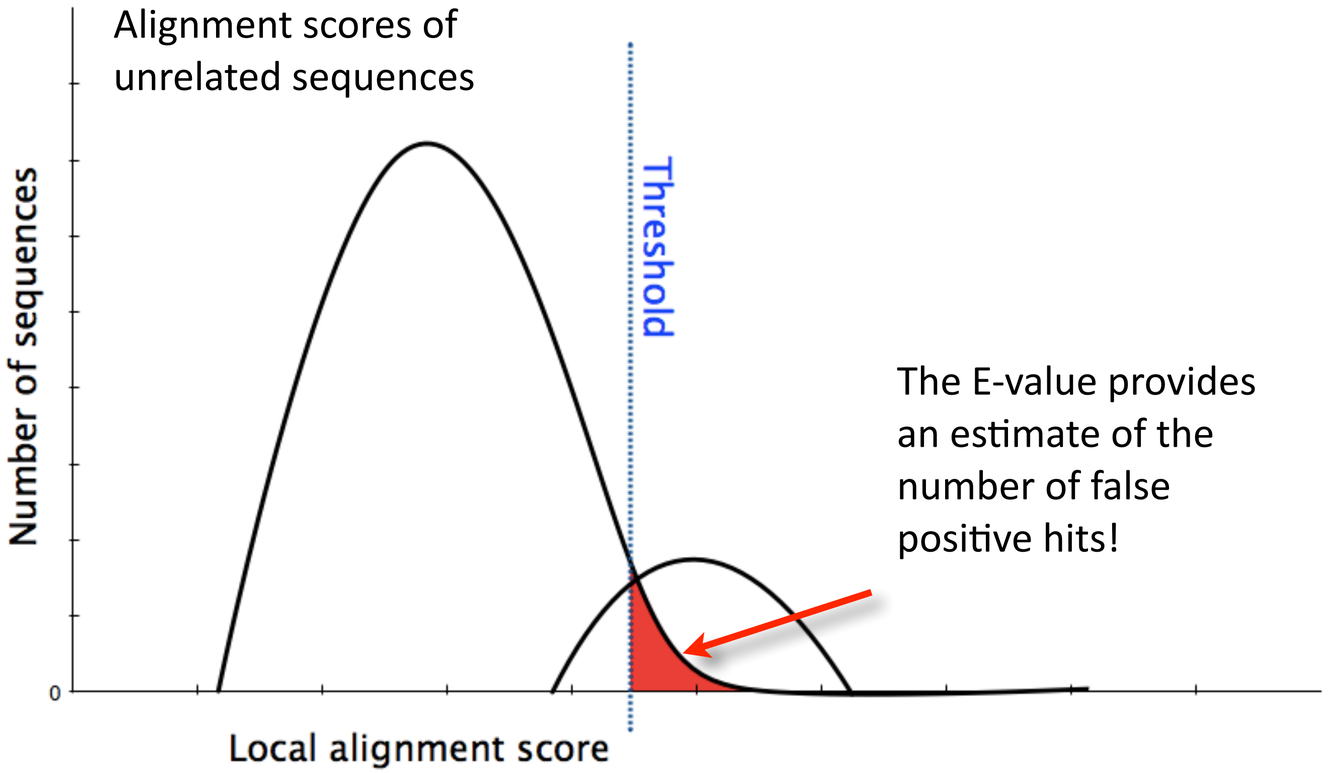
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



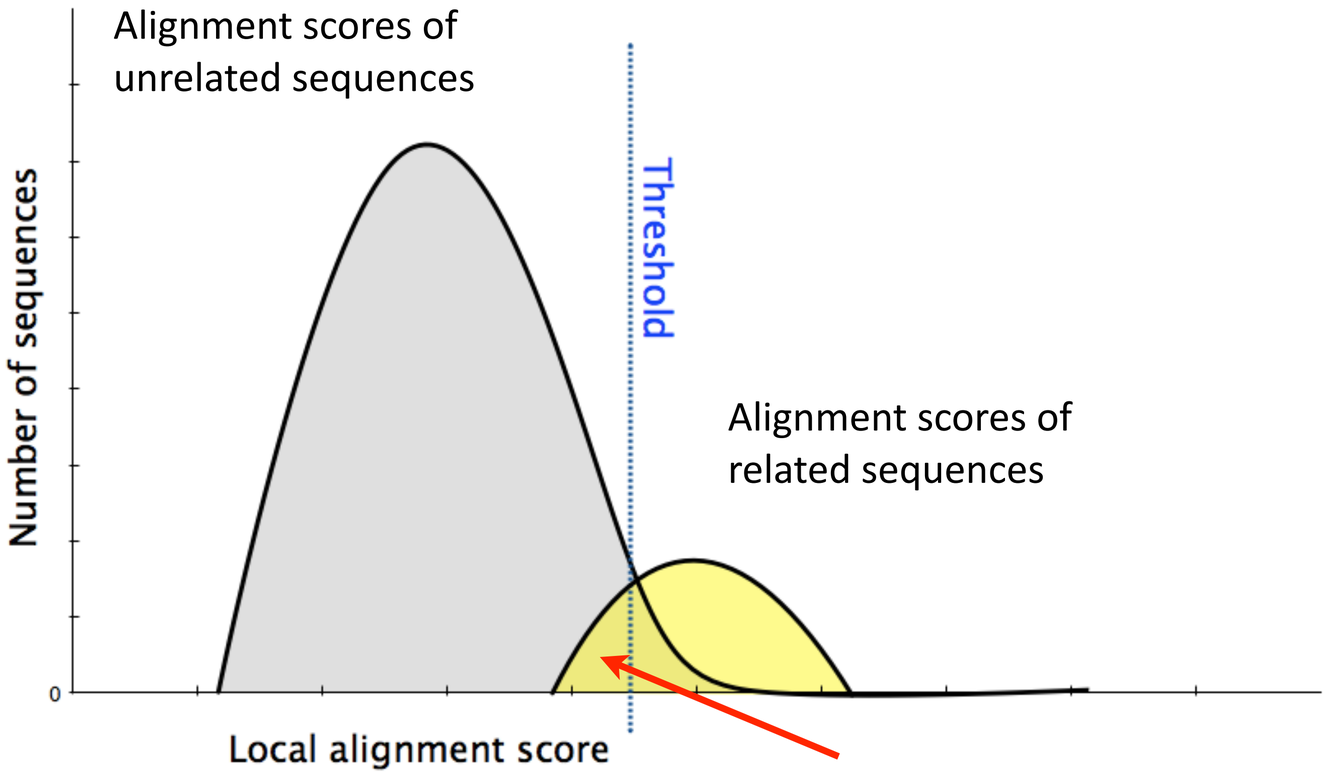
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



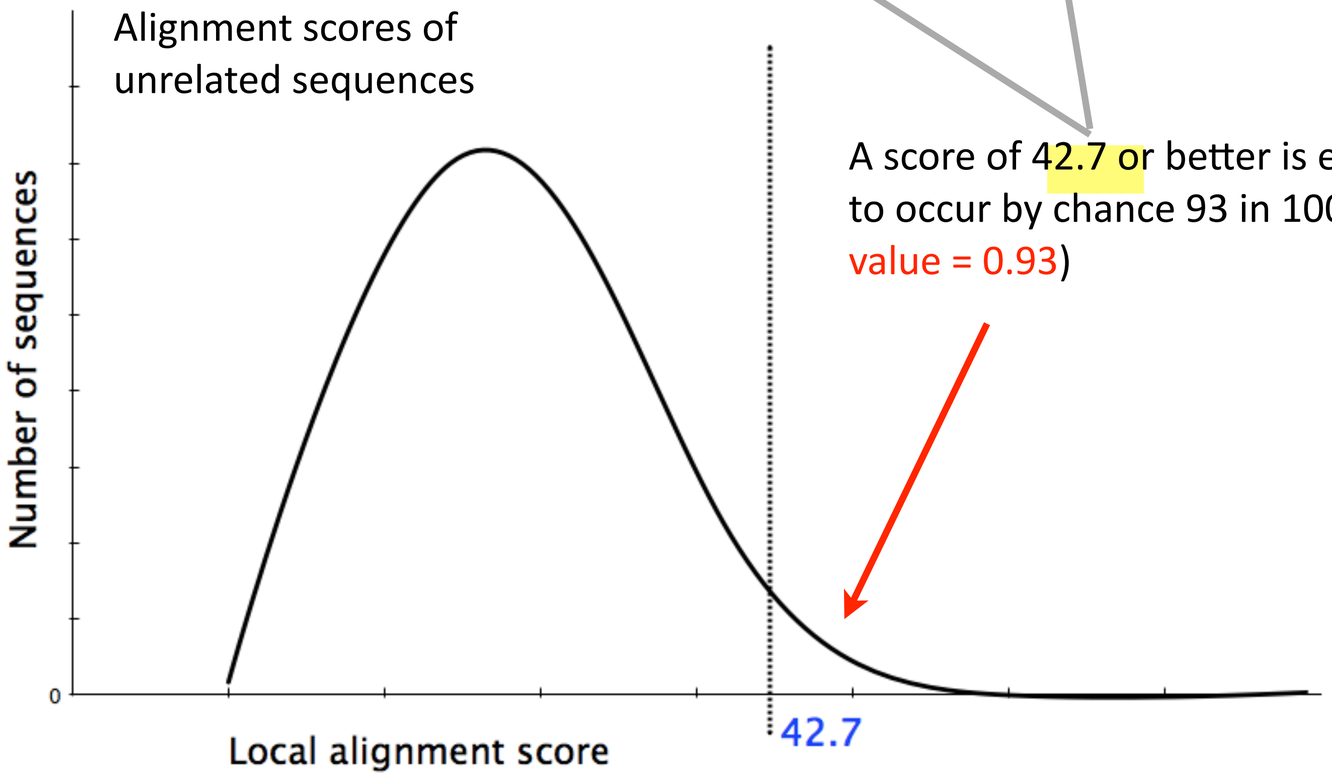
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- Maybe myoglobin, cytoglobin, neuroglobin etc. are found but not reported because of our E-value cutoff?
  - Lets change the cutoff and see...



Description	Max score	Query cover	E value	Max ident	Accession
hemoglobin subunit beta	284	100%	0	100%	NP_000510.1
hemoglobin subunit delta	240	100%	0	75.5%	NP_005321.1
hemoglobin subunit alpha	114	97%	0	43.45%	NP_000508.1
probable ATP-dependent RNA helicase	42.7	10%	**0.93**	32%	XP_011530405.1



E value: Th
alignments
with a part

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:

1. Limits of using BLAST [~10 mins]
2. **Using PSI-BLAST** [~30 mins]
3. Examining conservation patterns [~20 mins]

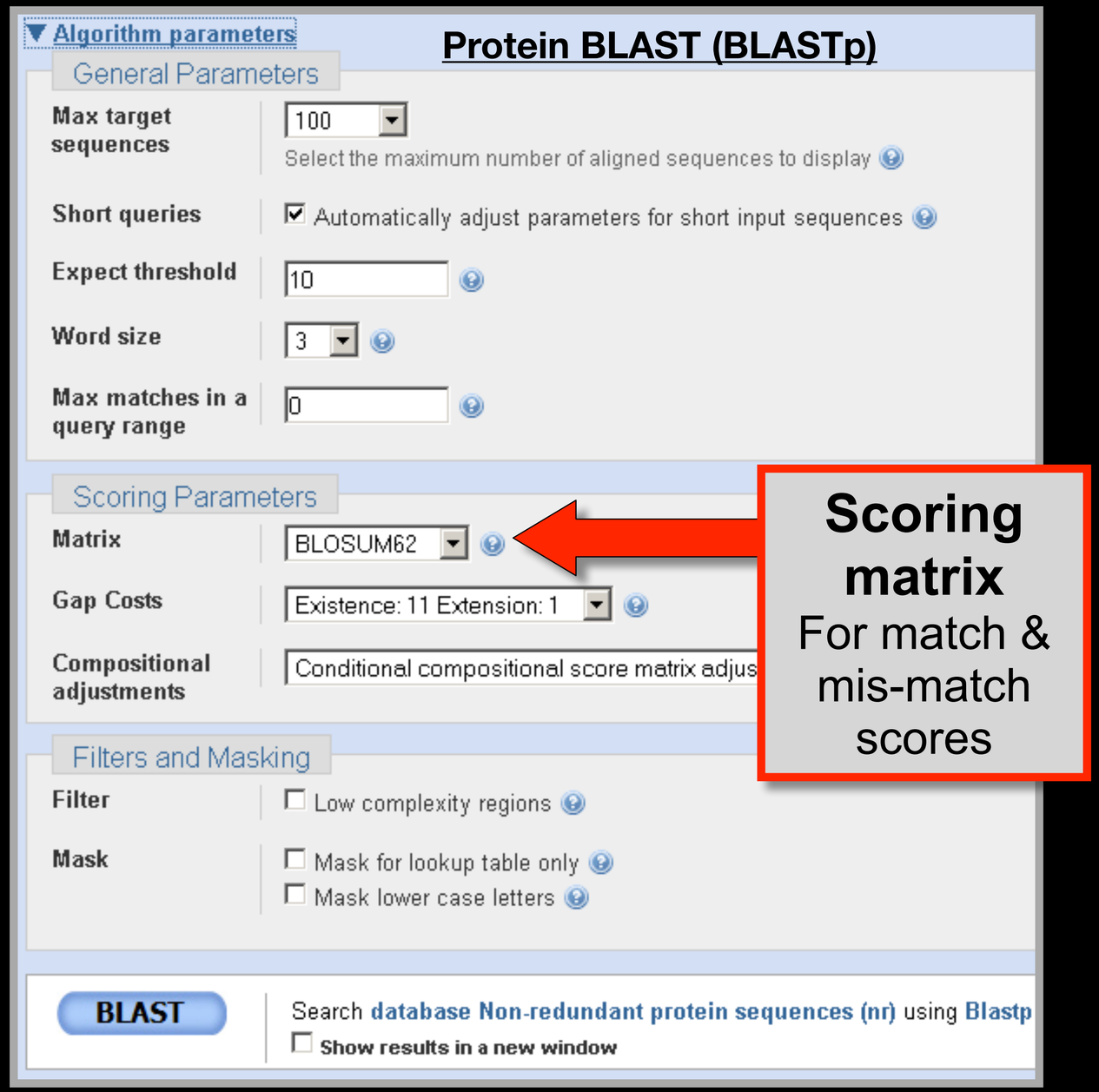
— BREAK [15 mins]—

4. [Optional] Using HMMER [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!

Recall: BLOUSM62 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).


Algorithm parameters
Protein BLAST (BLASTp)
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjus
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Non-redundant protein sequences (nr) using Blastp
Show results in a new window
Scoring matrix
For match & mis-match scores

By default BLASTp match scores come from the BLOSUM62 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

BIlocks **Su**bstitution **M**atrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

By default BLASTp match scores come from the BLOSUM62 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Note. All matches of Alanine for Alanine score +4 regardless of their position or context in the molecule.

PSI-BLAST: Position *s*pecific 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

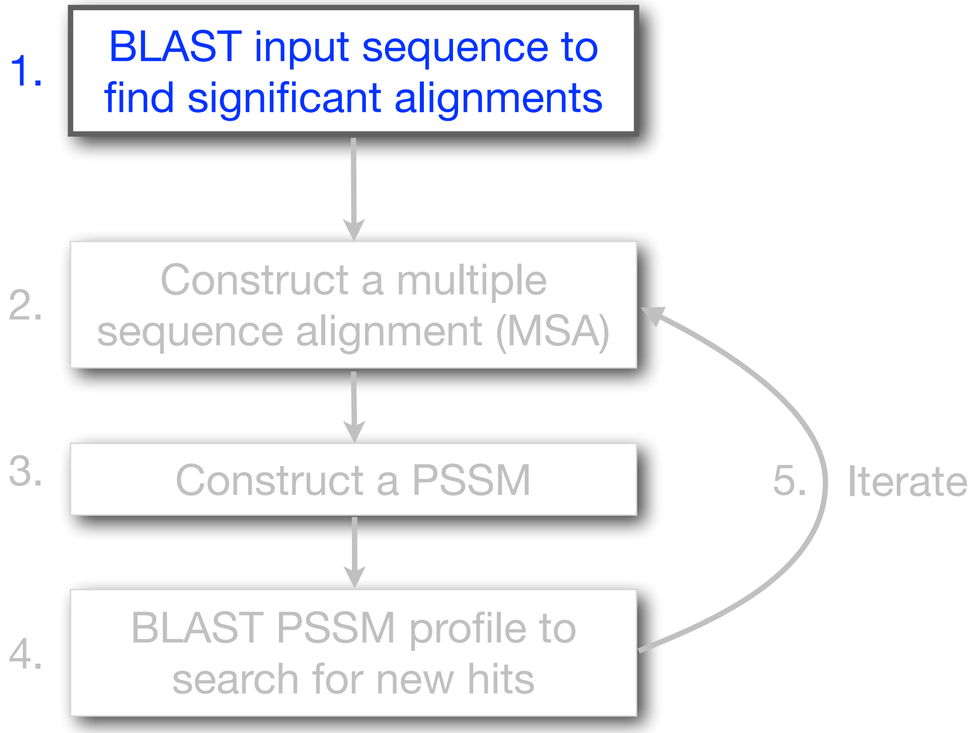
PSI-BLAST: Position *s*pecific iterated BLAST

- 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

PSI-BLAST: Position-Specific Iterated BLAST

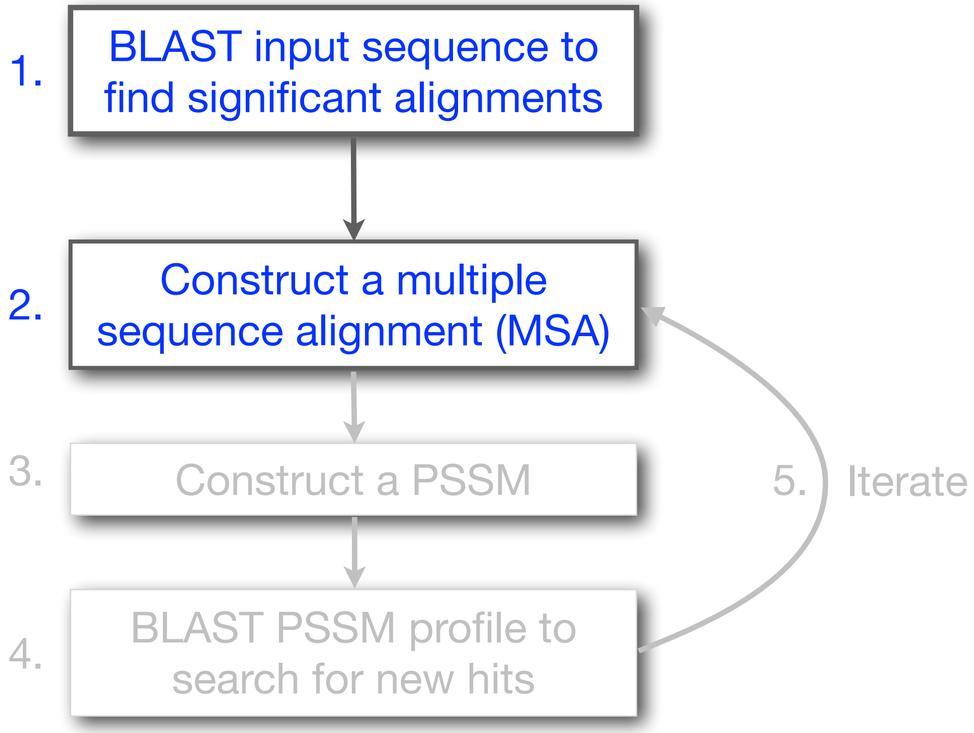
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

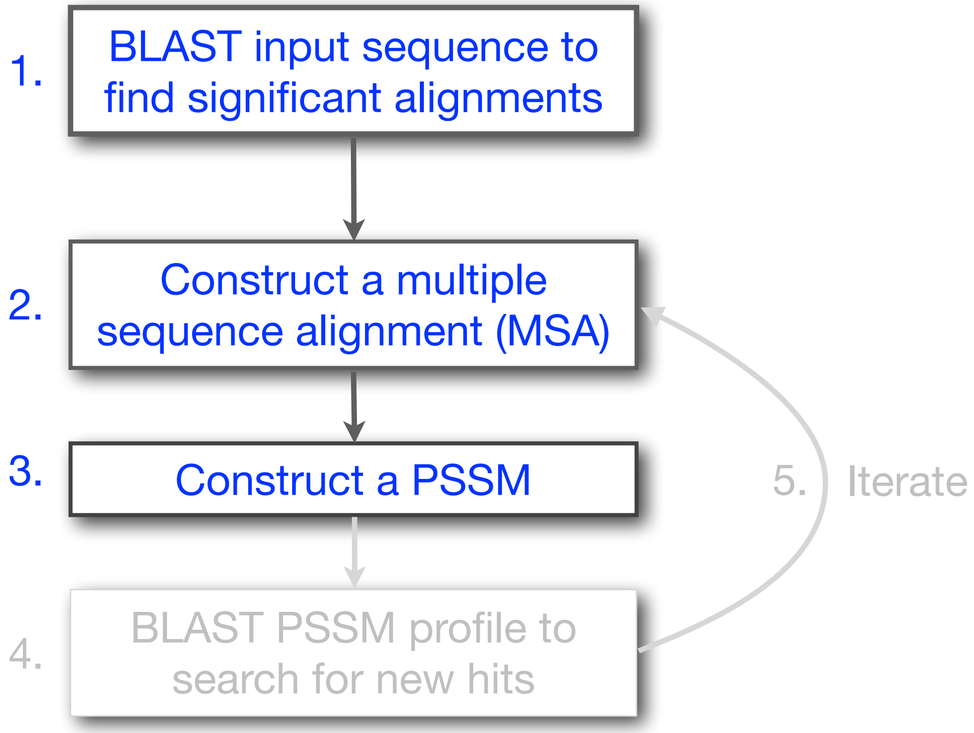
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

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)

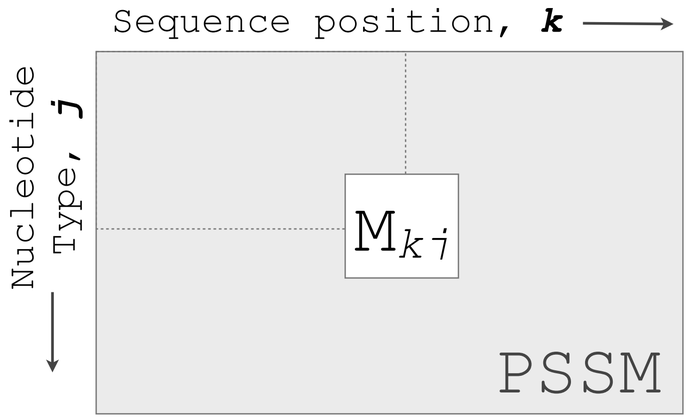
What is a PSSM?

What are PSSM sequence profiles?

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

PSSMs 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).



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

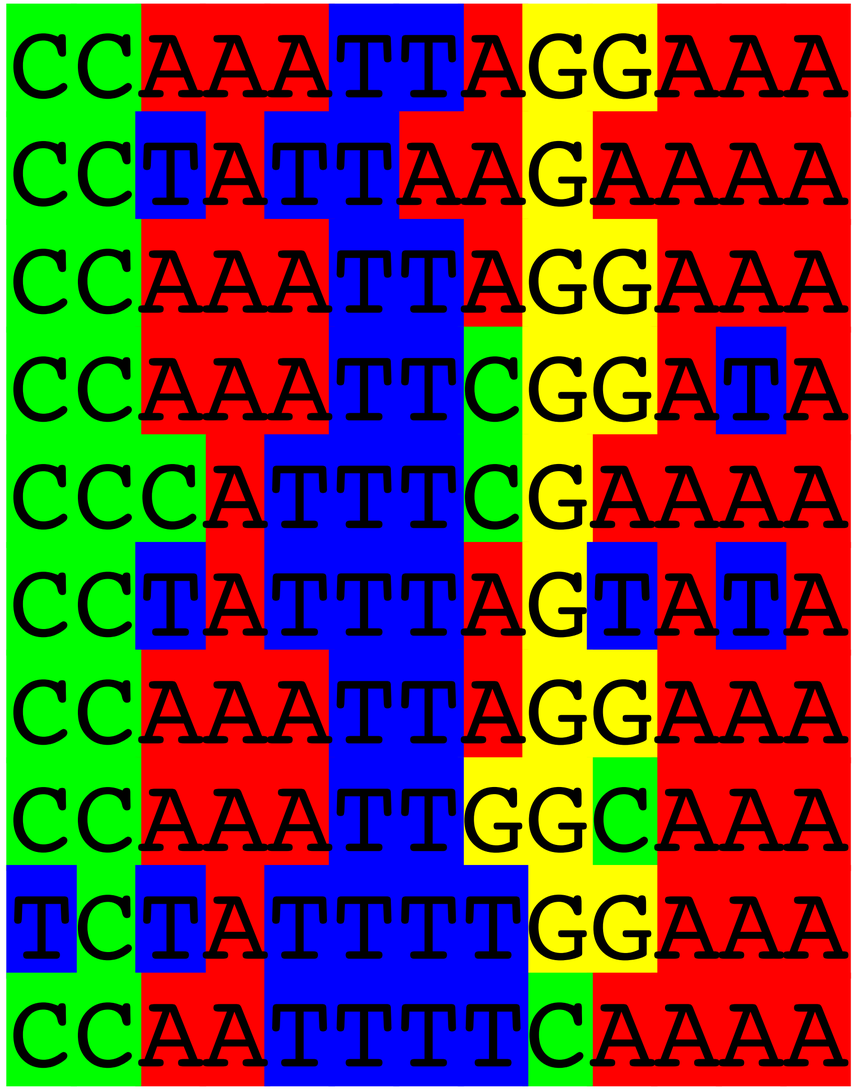
M_{kj} score for the *j*th nucleotide at position *k*

p_{kj} probability of nucleotide *j* at position *k*

p_j "background" probability of nucleotide *j*

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

Example: Computing a transcription factor bind site PSSM

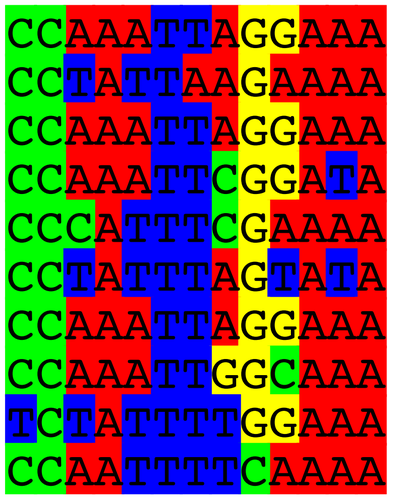


Here we have **10 aligned** transcription factor binding site nucleotide sequences

That span **13 positions** (i.e. columns of nucleotides).

We will build a 13 x 4 **PSSM** (*k*=13, *j*=4).

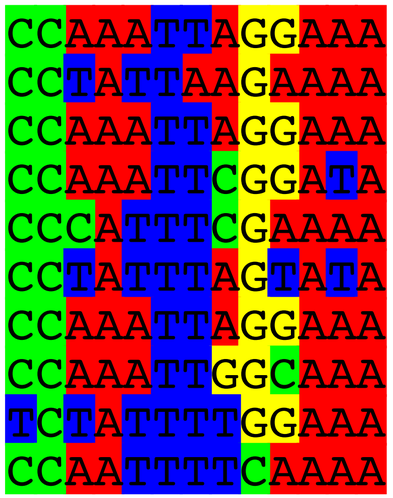
Computing a transcription factor bind site PSSM



First we will build an alignment **Counts matrix**

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

Computing a transcription factor bind site PSSM

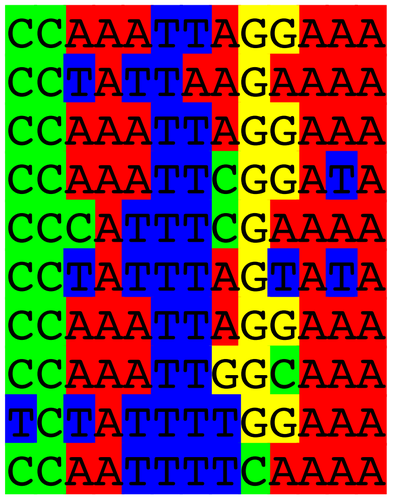


Alignment Counts matrix:

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

Position k = 1

Computing a transcription factor bind site PSSM

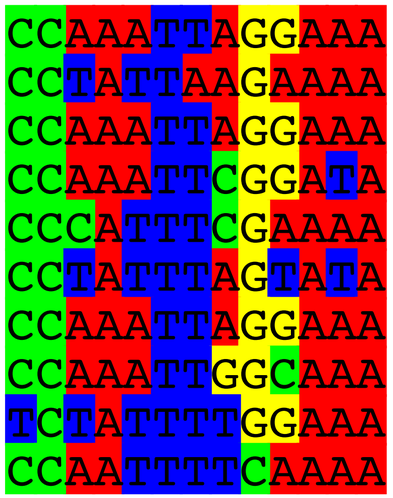


Alignment Counts matrix:

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

Position k = 1

Computing a transcription factor bind site PSSM

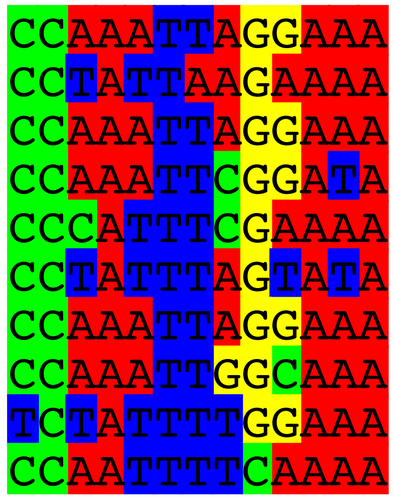


Alignment Counts matrix:

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

Position k = 1

Computing a transcription factor bind site PSSM

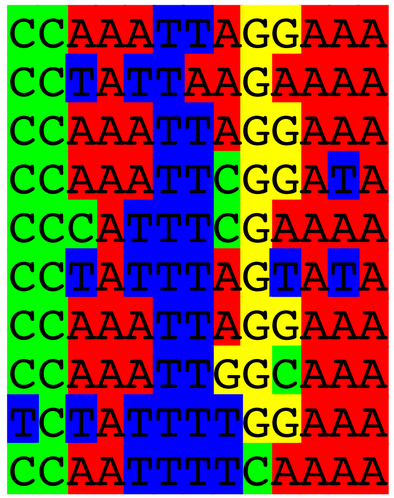


Alignment Counts matrix:

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

Position k = 2

Computing a transcription factor bind site PSSM

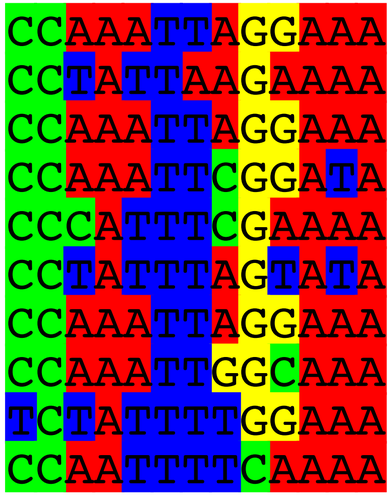


Alignment Counts matrix:

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

Position k = 3

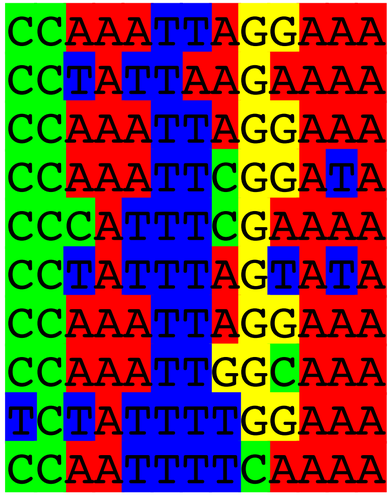
Computing a transcription factor bind site PSSM



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	[AT]	A	[AT]	T	T	[ACT]	G	[GA]	A	[AT]	A

Computing a transcription factor bind site PSSM



Alignment Counts matrix:

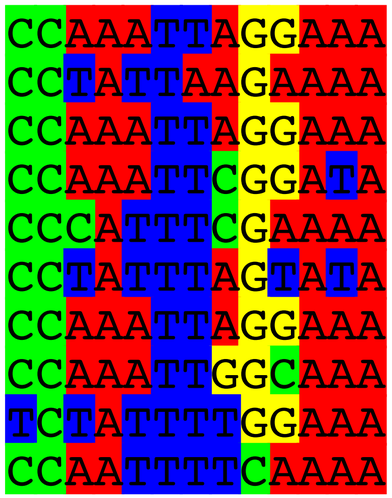
Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C:	9	10	1	0	0	0	0	2	1	1	0	0	0
G:	0	0	0	0	0	0	0	1	9	5	0	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0
Consensus	C	C	[AT]	A	[AT]	T	T	[ACT]	G	[GA]	A	[AT]	A

Often we will not communicate with the count matrix but rather the derived **average profile** (a.k.a. frequency matrix).

Average Profile (Frequency) matrix:

Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:	0	0	0.6	1	0.5	0	0.1	0.5	0	0.3	1	0.8	1
C:	0.9	1	0.1	0	0	0	0	0.2	0.1	0.1	0	0	0
G:	0	0	0	0	0	0	0	0.1	0.9	0.5	0	0	0
T:	0.1	0	0.3	0	0.5	1	0.9	0.2	0	0.1	0	0.2	0
Consensus	C	C	[AT]	A	[AT]	T	T	[ACT]	G	[GA]	A	[AT]	A

Computing a transcription factor bind site PSSM



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	[AT]	A	[AT]	T	T	[ACT]	G	[GA]	A	[AT]	A

Or the "score (M_{kj}) matrix" = PS**S**M

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*

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

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

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

CCTATTTAGGATA

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**

Query Score = 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + −0.2 + 1.3
= 11.9

Scoring a test sequence

Query Sequence

CCTATTTAGGATA

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**

Query Score = 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + −0.2 + 1.3
= 11.9

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

Scoring a test sequence...

Query Sequence: **CCTATTTAGGATA**

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**

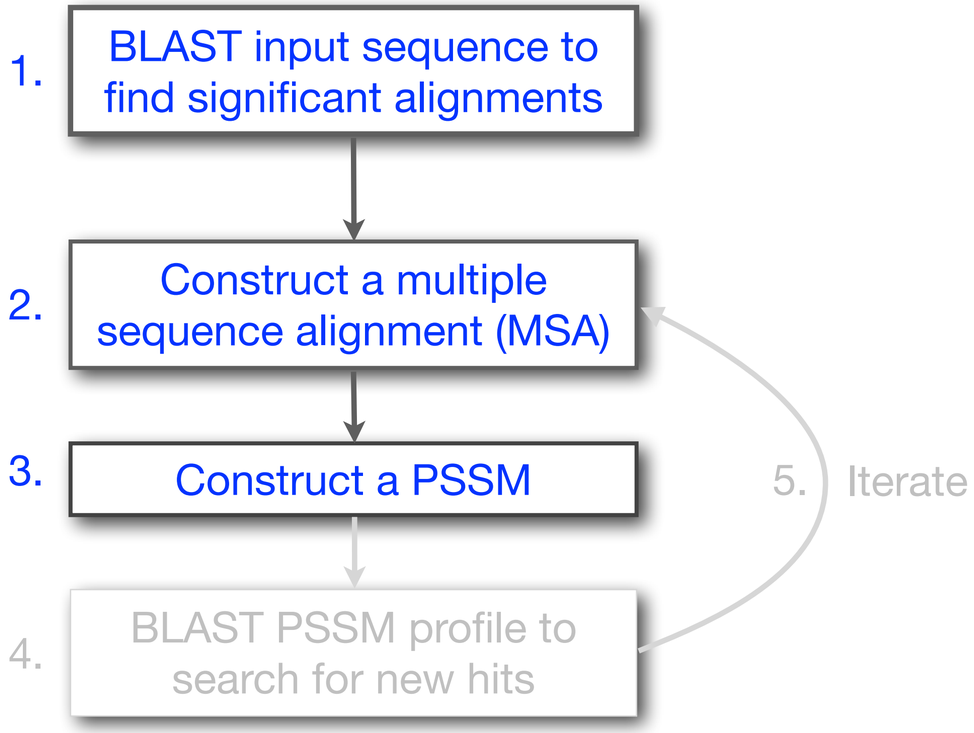
Max Score = 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2
+ 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3
= 13.8

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!

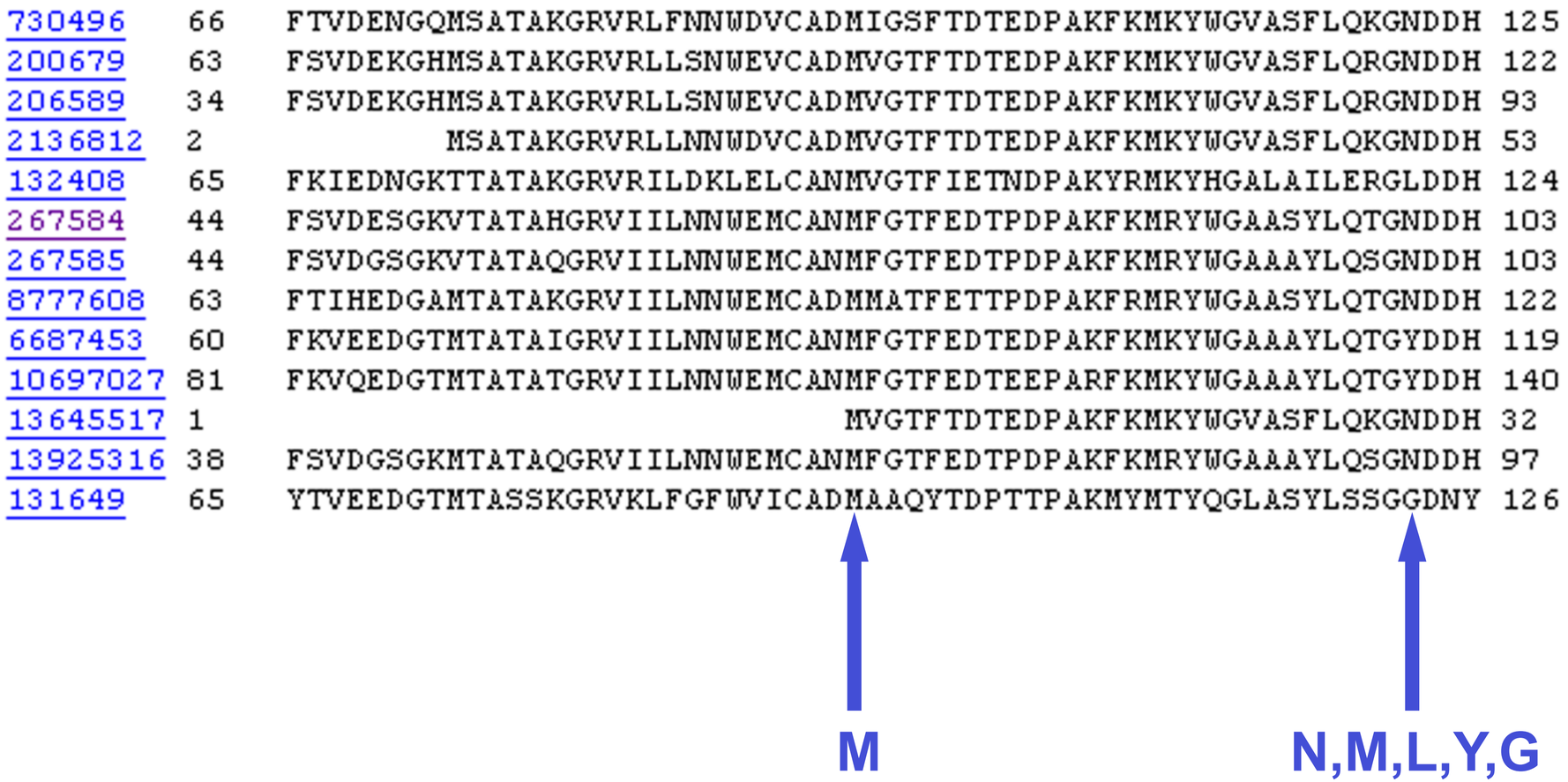
PSI-BLAST: Position-Specific Iterated BLAST

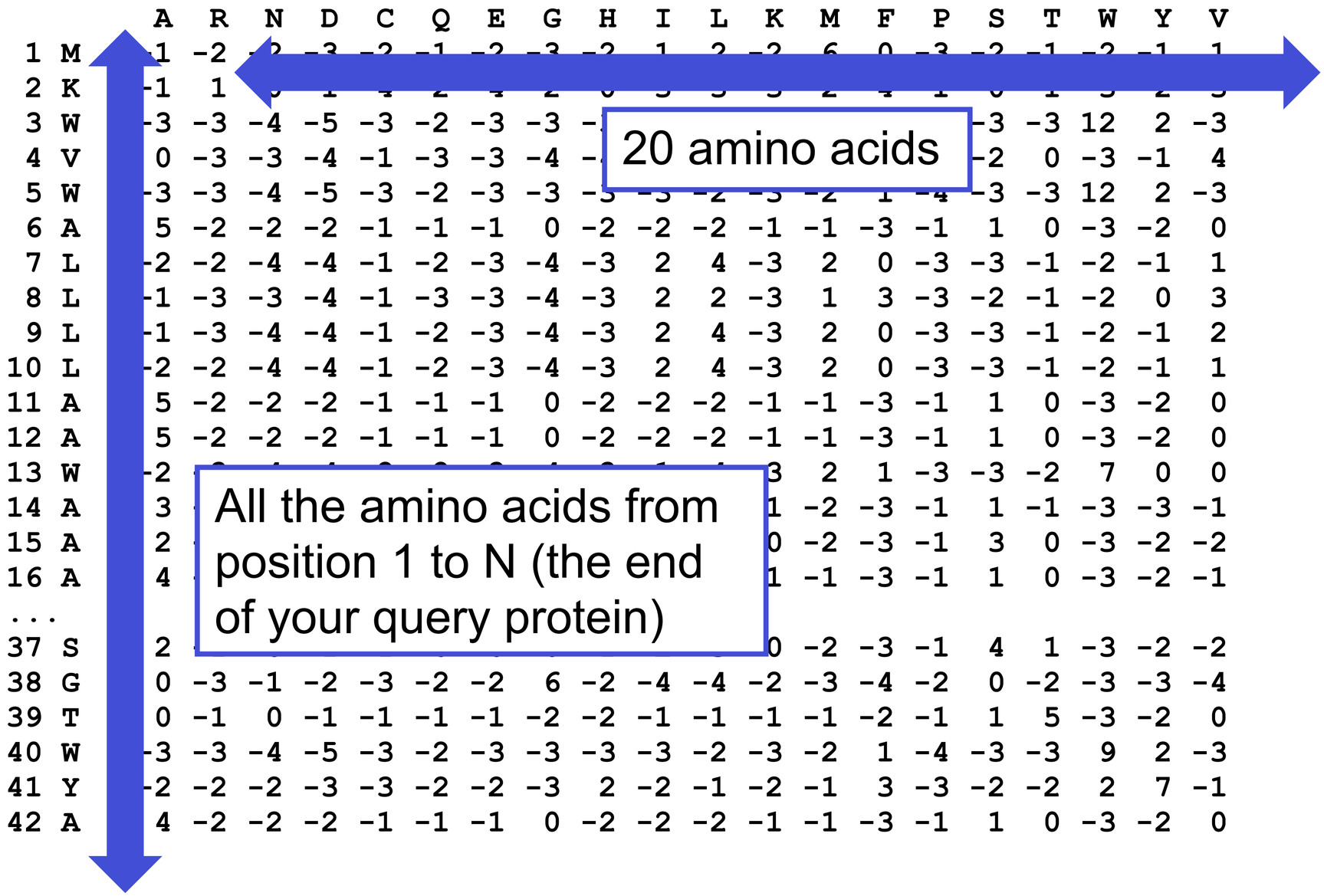
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



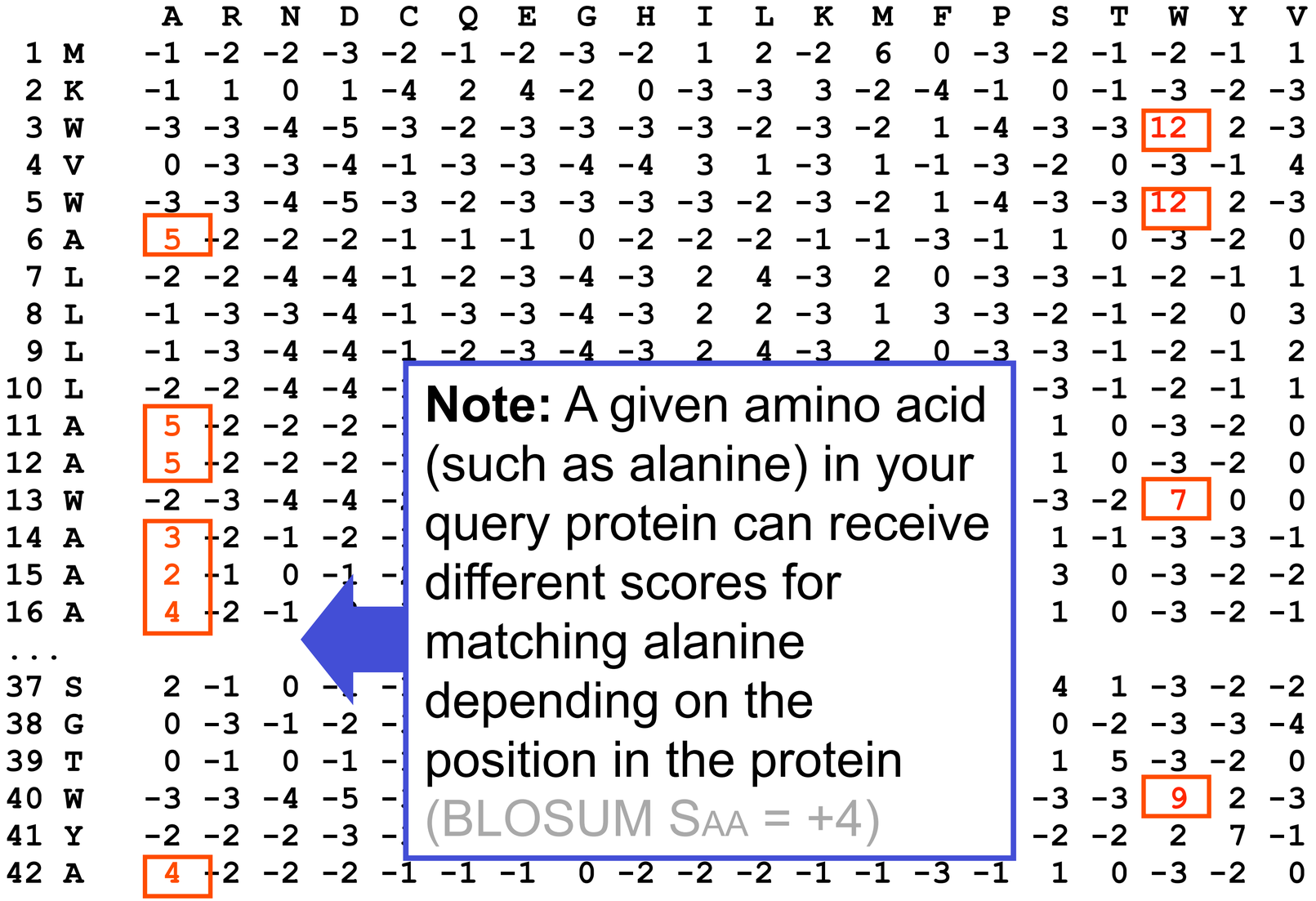
(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

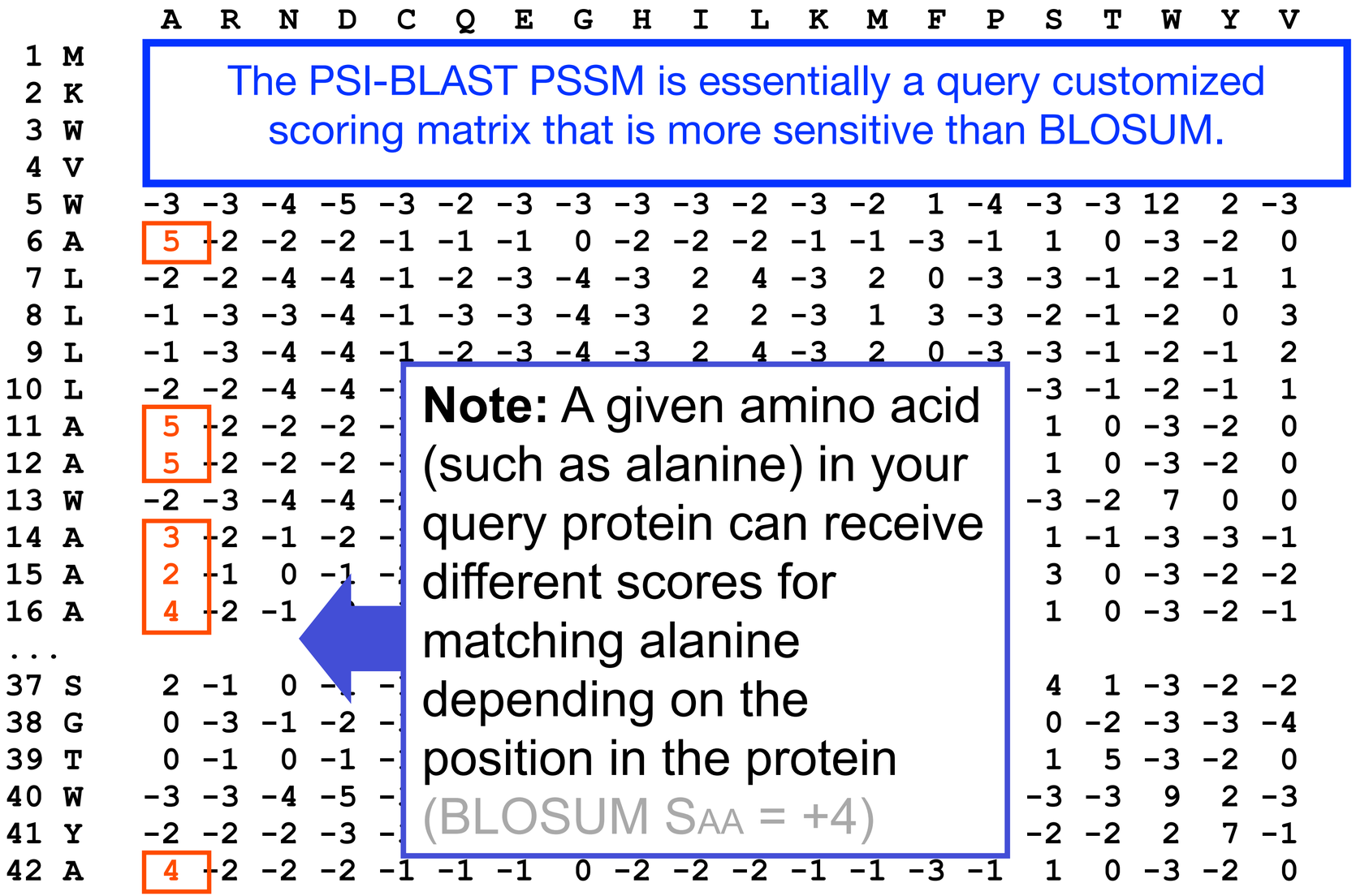
Inspect the blastp output to identify empirical “rules” regarding amino acids tolerated at each position





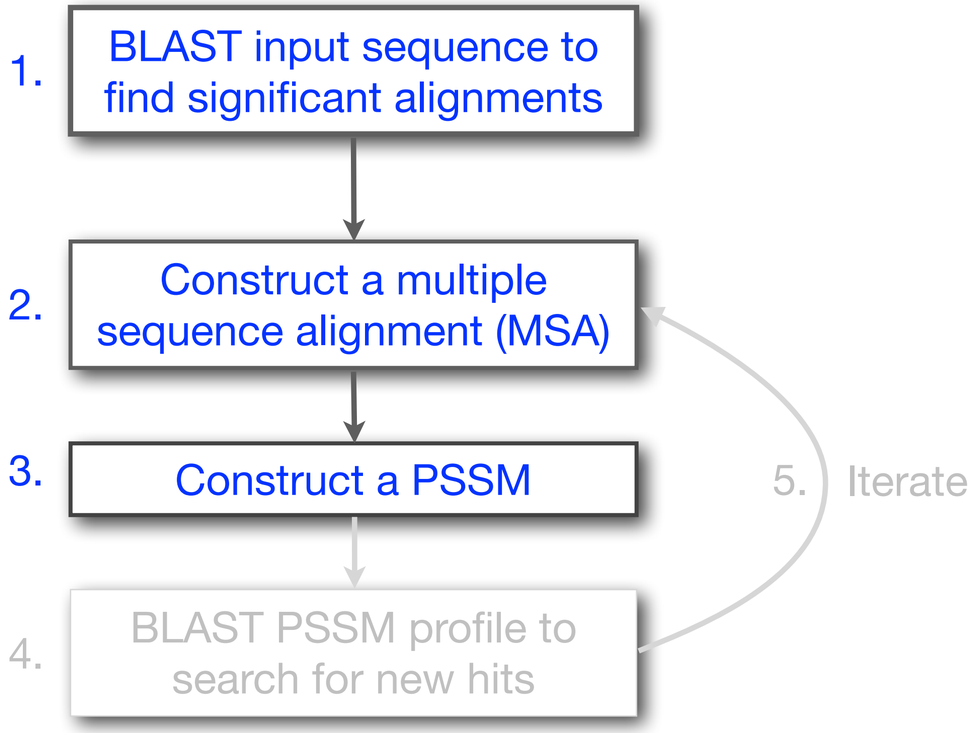
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2																		
2	K	-1	1																		
3	W	-3	-3	-4	-5	-3	-2	-3	-3								-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4								-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3								-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7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


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 -3 -1 -2 -1 1
11 A 5 -2 -2 -2 1 0 -3 -2 0
12 A 5 -2 -2 -2 1 0 -3 -2 0
13 W -2 -3 -4 -4 -3 -2 7 0 0
14 A 3 -2 -1 -2 1 -1 -3 -3 -1
15 A 2 -1 0 -1 3 0 -3 -2 -2
16 A 4 -2 -1 1 0 -3 -2 -1
...
37 S 2 -1 0 4 1 -3 -2 -2
38 G 0 -3 -1 -2 0 -2 -3 -3 -4
39 T 0 -1 0 -1 1 5 -3 -2 0
40 W -3 -3 -4 -5 -3 -3 9 2 -3
41 Y -2 -2 -2 -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 -1 -2 -3 -4 -3 2 4 -3 2 0 -3 -3 -1 -2 -1 2
10 L -2 -2 -4 -4 -3 -1 -2 -1 1
11 A 5 -2 -2 -2 1 0 -3 -2 0
12 A 5 -2 -2 -2 1 0 -3 -2 0
13 W -2 -3 -4 -4 -3 -2 7 0 0
14 A 3 -2 -1 -2 1 -1 -3 -3 -1
15 A 2 -1 0 -1 3 0 -3 -2 -2
16 A 4 -2 -1 1 0 -3 -2 -1
...
37 S 2 -1 0 4 1 -3 -2 -2
38 G 0 -3 -1 -2 0 -2 -3 -3 -4
39 T 0 -1 0 -1 1 5 -3 -2 0
40 W -3 -3 -4 -5 -3 -3 9 2 -3
41 Y -2 -2 -2 -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

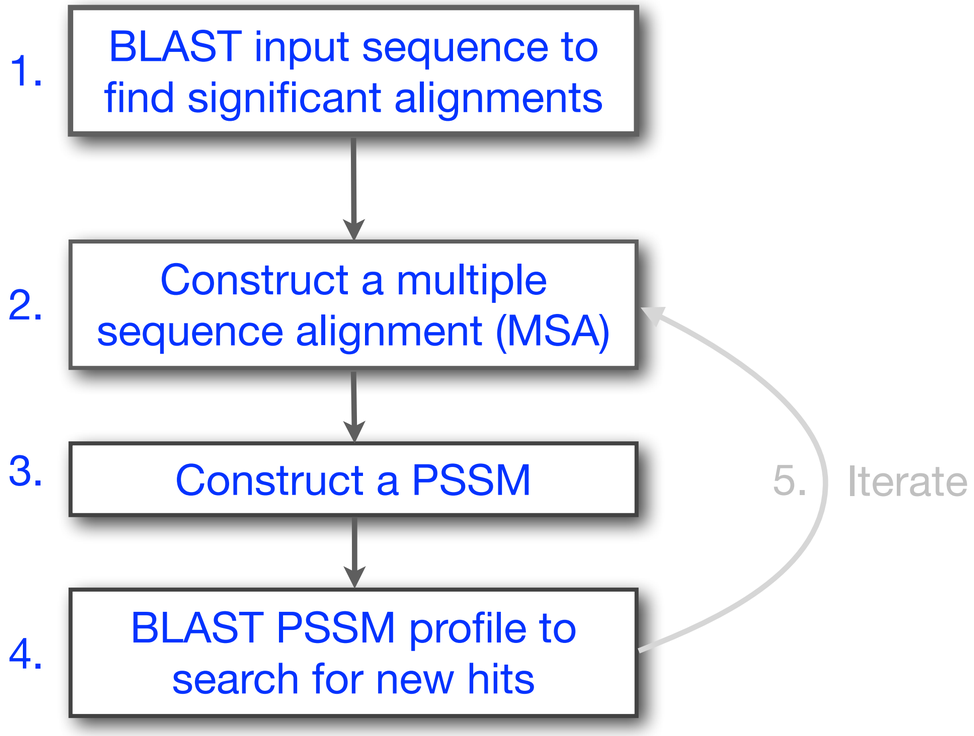
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

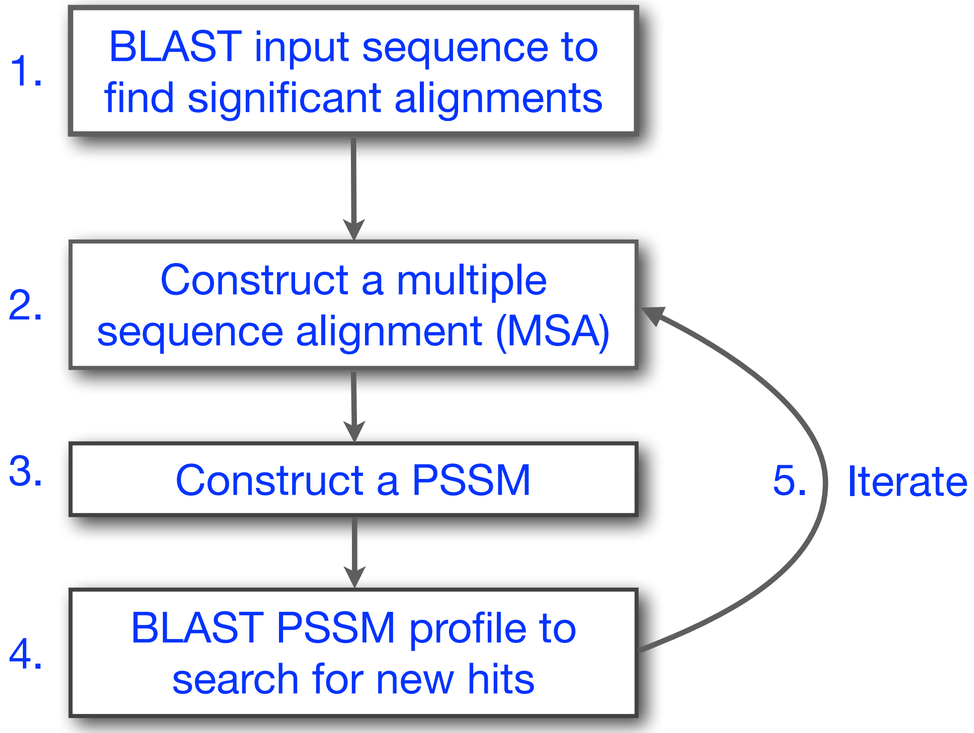
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



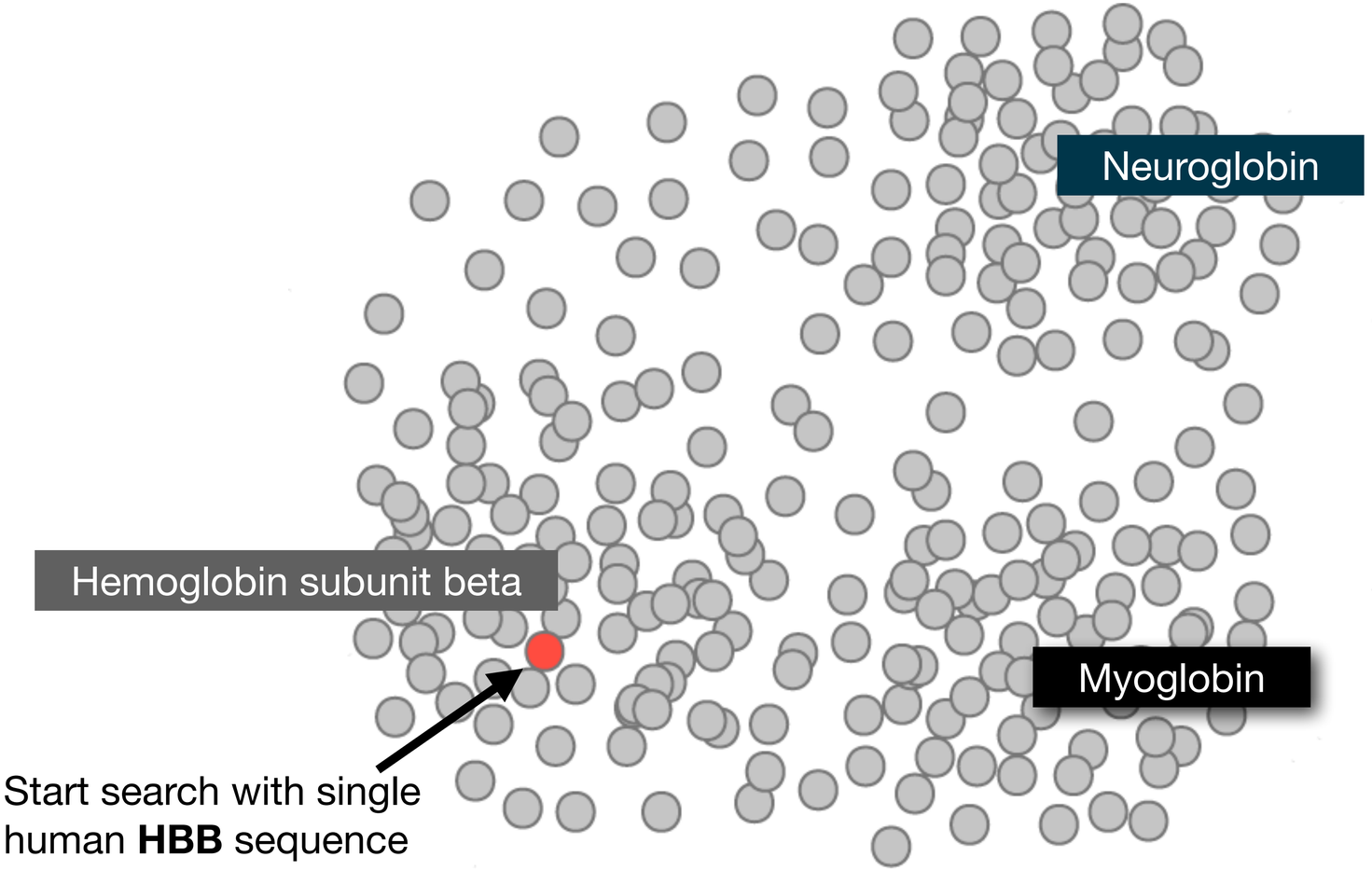
(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

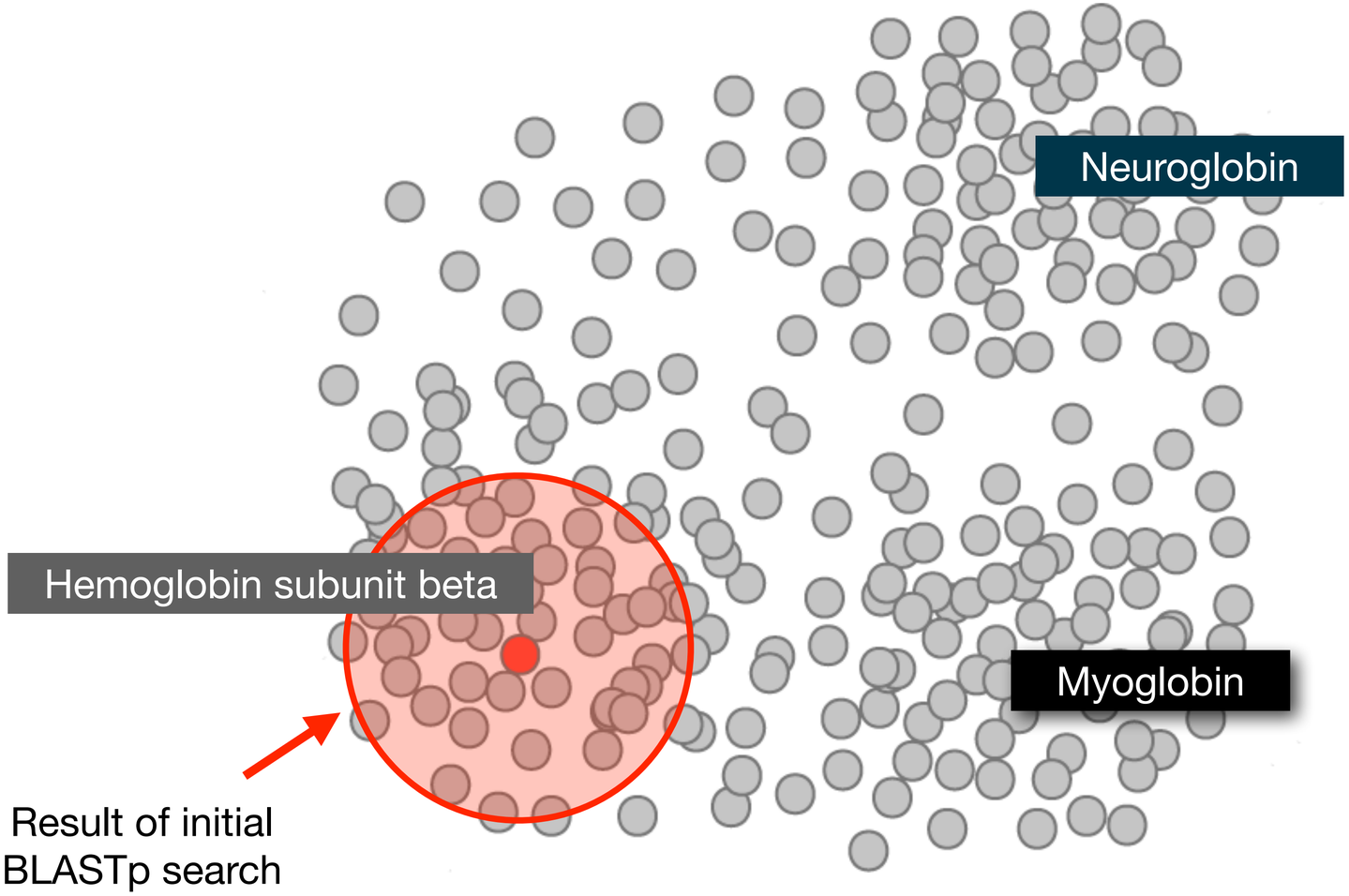
PSI-BLAST: Position-Specific Iterated BLAST

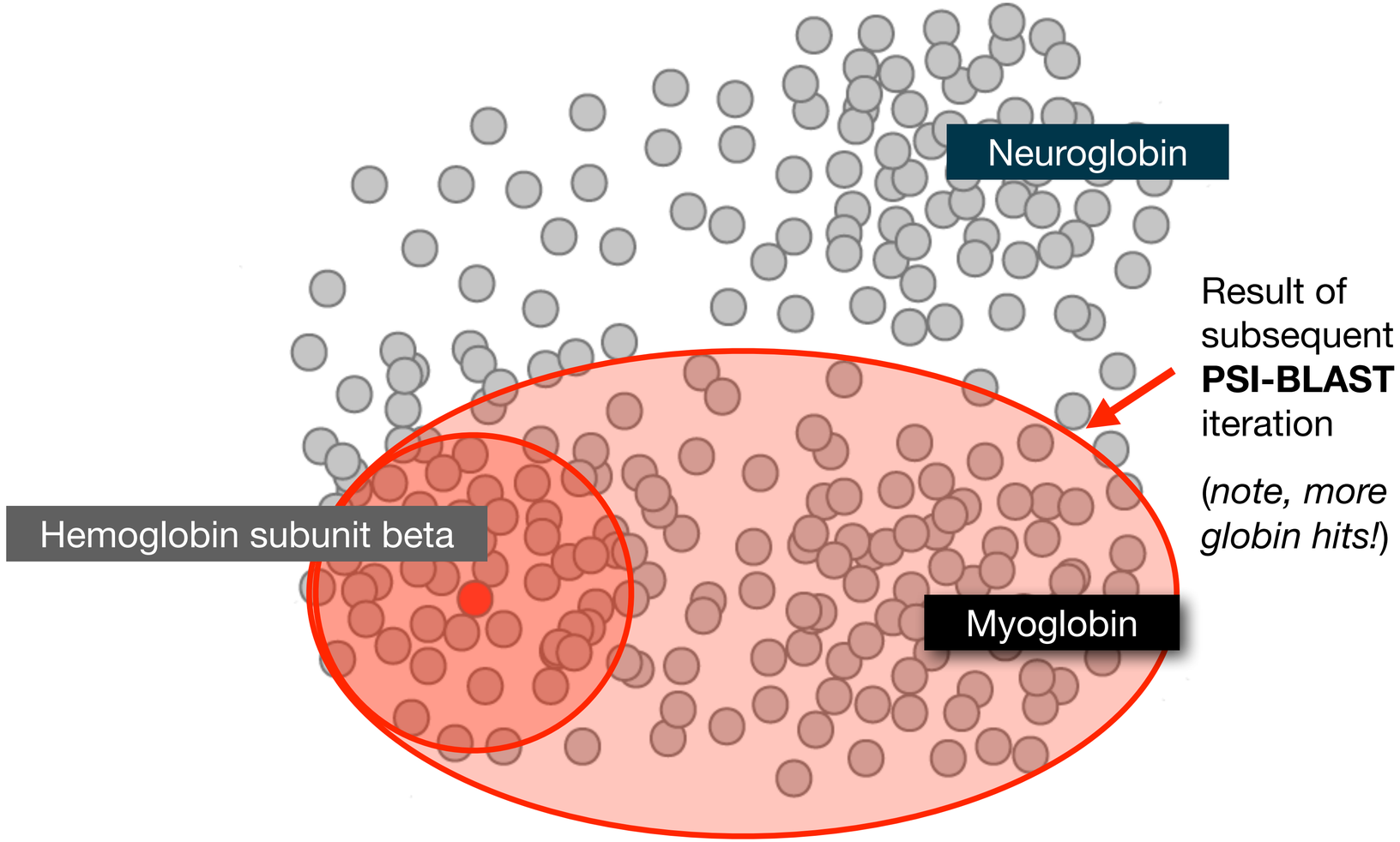
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST

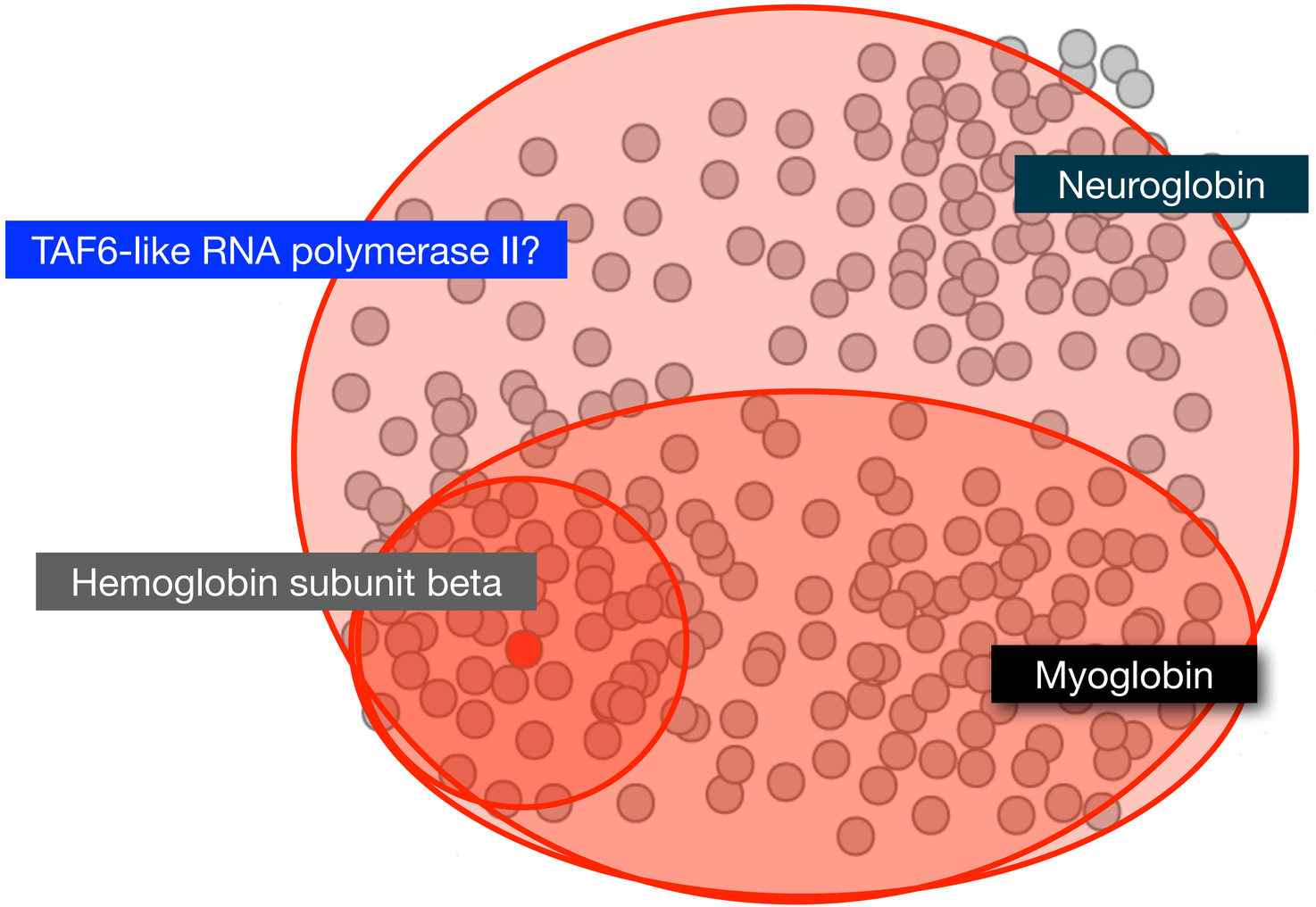


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)


Neuroglobin
Hemoglobin subunit beta
Myoglobin
Start search with single
human **HBB** sequence


Neuroglobin
Hemoglobin subunit beta
Myoglobin
Result of initial
BLASTp search


Neuroglobin
Result of subsequent **PSI-BLAST** iteration
(*note, more globin hits!*)
Hemoglobin subunit beta
Myoglobin



Result of later **PSI-BLAST** iteration

(*note, potential "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

1

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

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

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	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	2
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	3
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

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:

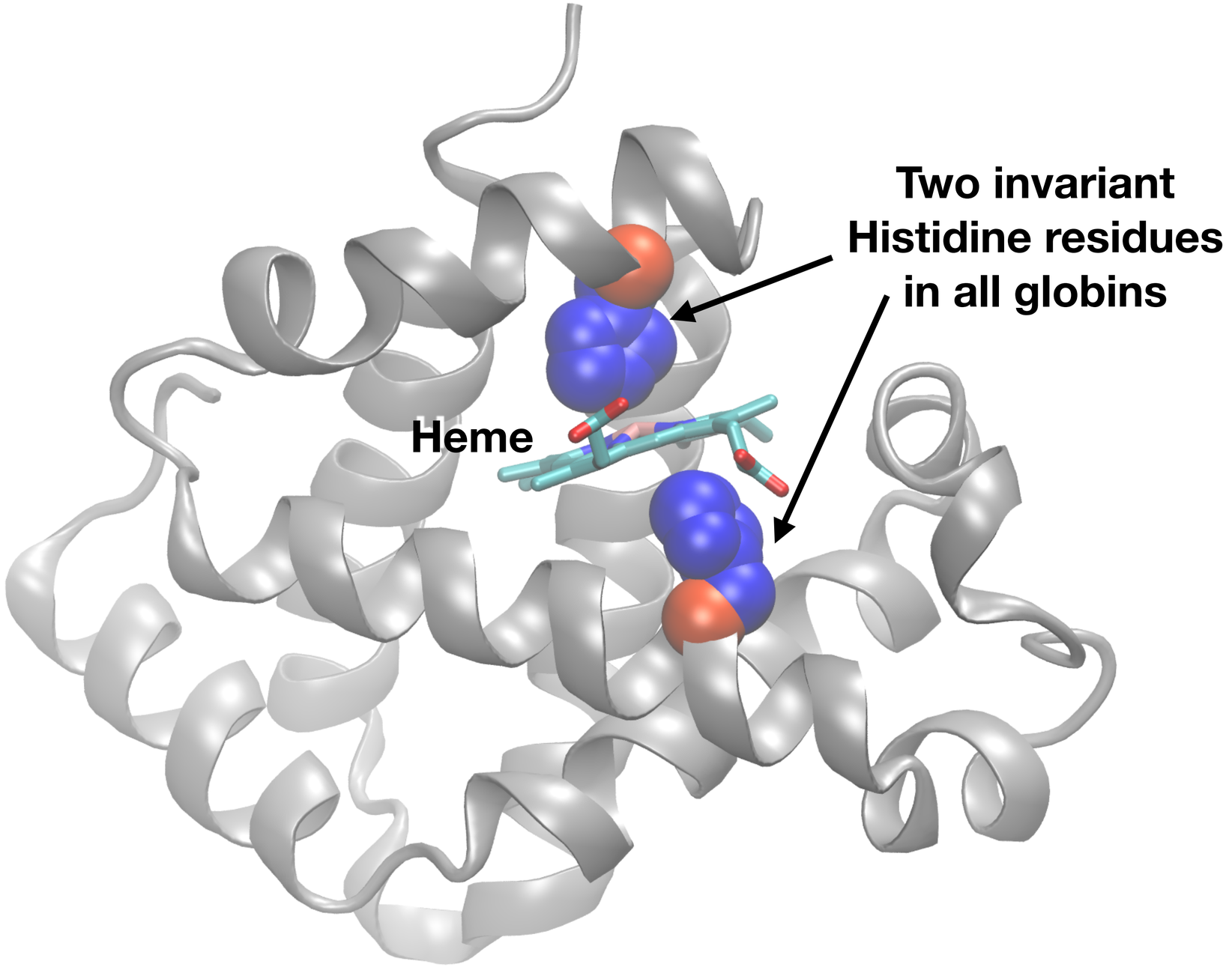
1. Limits of using BLAST [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. **Examining conservation patterns** [~20 mins]

— BREAK [15 mins]—

4. [Optional] Using HMMER [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!


Query_73613 1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF 72
NP_000510.1 1 MVHLTPEEKTAVNALWGKV--NVDAVGGEALGRLLVVYPWTQRFFE-SFGDLSSPDAVM-GNPKVKAHGKKVLGAF 72
NP_000175.1 1 MGHFTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL 72
NP_000509.1 1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF 72
NP_005321.1 1 MVHFTAEEKAAVTSLWSKM--NVEEAGGEALGRLLVVYPWTQRFFD-SFGNLSSPSAIL-GNPKVKAHGKKVLTSF 72
NP_000550.2 1 MGHFTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL 72
NP_005323.1 1 -MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFP-HF-------DLHpGSAQLRAHGSKVVAAV 67
NP_000508.1 1 -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFP-HF-------DLShGSAQVKGHGKKVADAL 67
XP_005257062.1 1 [15]SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL 89
NP_001003938.1 1 --MLSAQERAQIAQVWDLIAGHEAQFGAELLLRLFTVYPSTKVYFP-HL------SACQ-DATQLLSHGQRMLAAV 66
NP_005322.1 1 -MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFS-H-------LDLSpGSSQVRAHGQKVADAL 67
NP_599030.1 1 [15]SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL 89
XP_016879605.1 1 ----------MEDPLEME-RSPQLRKHACRVMGAL 24
NP_001349775.1 1 -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFKHLKSEDEMK-ASEDLKKHGATVLTAL 73
NP_067080.1 1 ---MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQyNCRQFSSPEDCL-SSPEFLDHIRKVMLVI 72
NP_001369741.1 1 ----------MK-ASEDLKKHGATVLTAL 18
Query_73613 73 SDGLAHLDNLKGT---FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
NP_000510.1 73 SDGLAHLDNLKGT---FSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFGKEFTPQMQAAYQKVVAGVANALAHKYH 147
NP_000175.1 73 GDAIKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH 147
NP_000509.1 73 SDGLAHLDNLKGT---FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
NP_005321.1 73 GDAIKNMDNLKPA---FAKLSELHCDKLHVDPENFKLLGNVMVIILATHFGKEFTPEVQAAWQKLVSAVAIALAHKYH 147
NP_000550.2 73 GDATKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTAVASALSSRYH 147
NP_005323.1 68 GDAVKSIDDIGGA---LSKLSELHAYILRVDPVNFKLLSHCLLVTLAARFPADFTAEAHAAWDKFLSVVSSVLTEKYR 142
NP_000508.1 68 TNAVAHVDDMPNA---LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 142
XP_005257062.1 90 NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK[35] 202
NP_001003938.1 67 GAAVQHVDNLRAA---LSPLADLHALVLRVDPANFPLLIQCFHVVLASHLQDEFTVQMQAAWDKFLTGVAVVLTEKYR 141
NP_005322.1 68 SLAVERLDDLPHA---LSALSHLHACQLRVDPASFQLLGHCLLVTLARHYPGDFSPALQASLDKFLSHVISALVSEYR 142
NP_599030.1 90 NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK[23] 190
XP_016879605.1 25 NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK[35] 137
NP_001349775.1 74 GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK[6] 154
NP_067080.1 73 DAAVTNVEDLSSLeeyLASLGRKHRA-VGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWD[2] 151
NP_001369741.1 19 GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK[6] 99


Two invariant
Histidine residues
in all globins
Heme

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:

1. Limits of using BLAST [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. Examining conservation patterns [~20 mins]

— BREAK [15 mins]—

4. **[Optional] Using HMMER** [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!

Problems with PSSMs: Positional dependencies

Do not capture positional dependencies

WEIRD
WEIRD
WEIQH
WEIRD
WEIQH

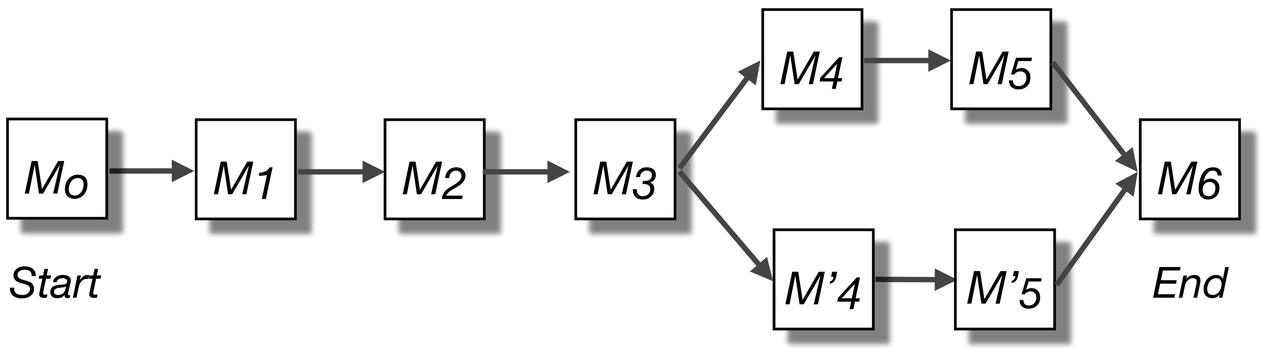
D					0.6
E		I			
H					0.4
I			I		
Q				0.4	
R				0.6	
W	I				

Note: We <u>never</u> 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.

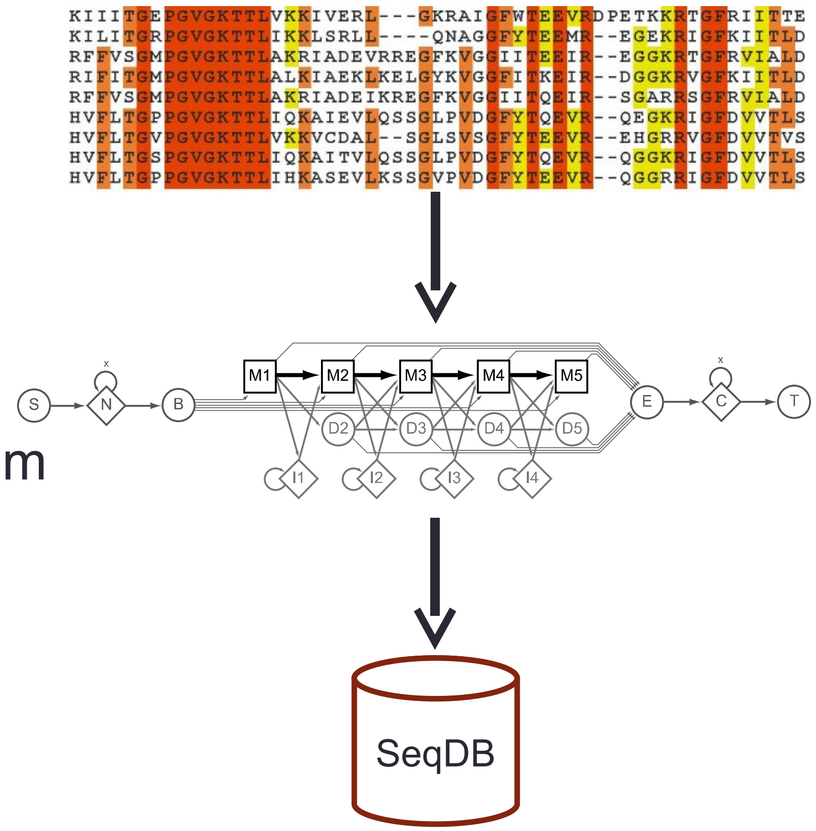
WEIRD
WEIRD
WEIQH
WEIRD
WEIQH



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

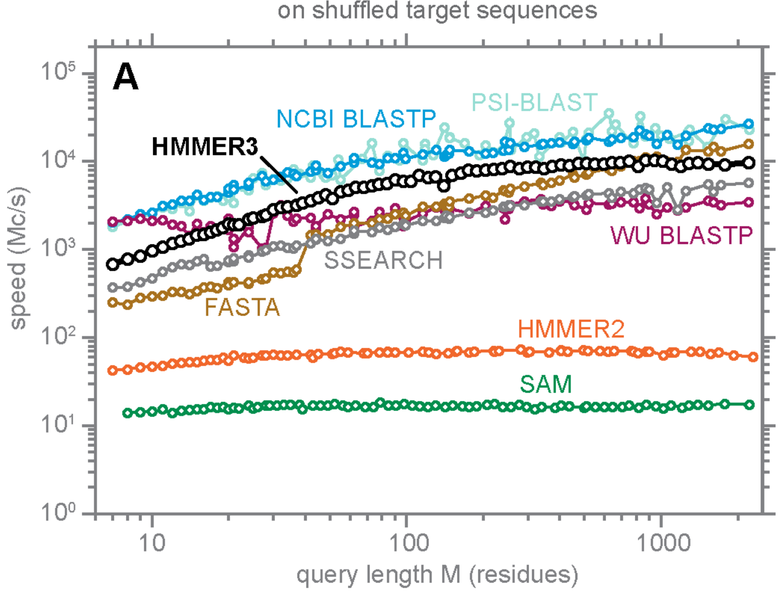
Use of HMMER

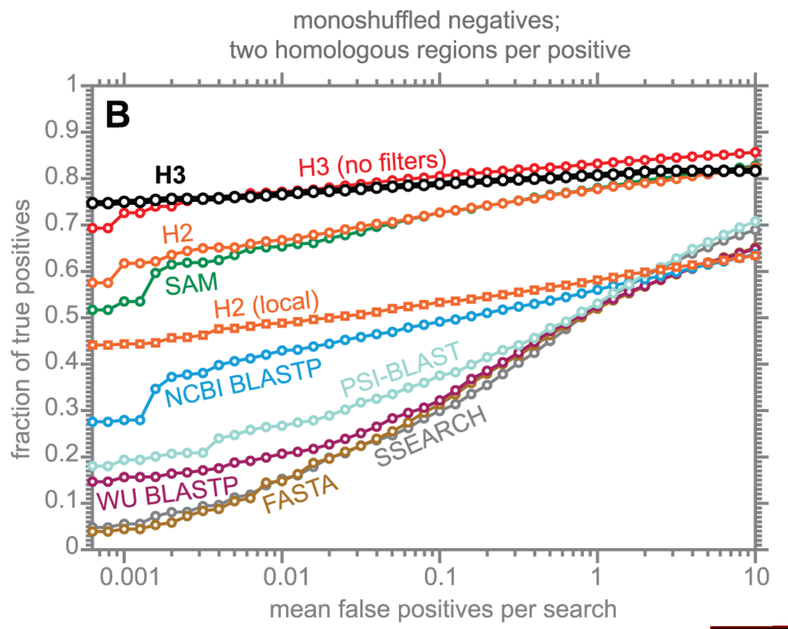
- Widely used by protein family databases
  - Use ‘seed’ alignments
- Until 2010
  - Computationally expensive
  - Restricted to HMMs constructed from multiple sequence alignments
- Command line application



HMMER vs BLAST

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

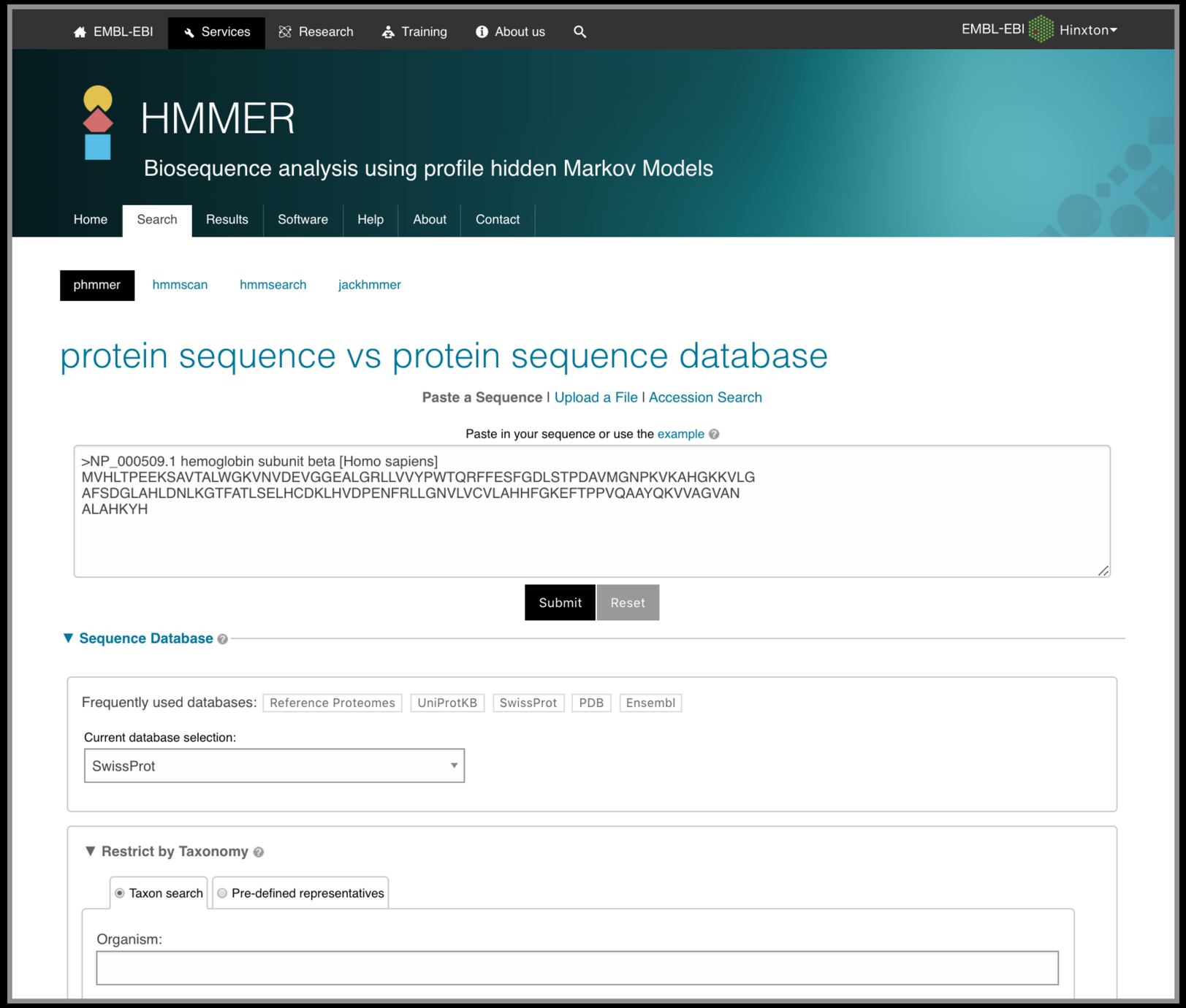




Modified from: S. R. Eddy
PLoS Comp. Biol., 7:e1002195, 2011.

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


EMBL-EBI
Services
Research
Training
About us
EMBL-EBI Hinxton
HMMER
Biosequence analysis using profile hidden Markov Models
Home
Search
Results
Software
Help
About
Contact
phmmer
hmmscan
hmmsearch
jackhmmer
protein sequence vs protein sequence database
Paste a Sequence | Upload a File | Accession Search
Paste in your sequence or use the example
>NP_000509.1 hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
Submit
Reset
▼ Sequence Database
Frequently used databases: Reference Proteomes UniProtKB SwissProt PDB Ensembl
Current database selection:
SwissProt
▼ Restrict by Taxonomy
Taxon search
Pre-defined representatives
Organism:

Significant Query Matches (12) in *swissprot* (v.2018_11)

Customise Customise

	Target	Description	Species	Cross-references	E-value
>	HBB_HUMAN	Hemoglobin subunit beta	Homo sapiens		6.8e-99
>	HBD_HUMAN	Hemoglobin subunit delta	Homo sapiens		1.6e-91
>	HBE_HUMAN	Hemoglobin subunit epsilon	Homo sapiens		1.5e-74
>	HBG2_HUMAN	Hemoglobin subunit gamma-2	Homo sapiens		8.8e-73
>	HBG1_HUMAN	Hemoglobin subunit gamma-1	Homo sapiens		6.2e-72
>	HBA_HUMAN	Hemoglobin subunit alpha	Homo sapiens		3.8e-29
>	HBAZ_HUMAN	Hemoglobin subunit zeta	Homo sapiens		4.5e-23
>	HBAT_HUMAN	Hemoglobin subunit theta-1	Homo sapiens		5.2e-22
>	HBM_HUMAN	Hemoglobin subunit mu	Homo sapiens		3.4e-19
>	CYGB_HUMAN	Cytoglobin	Homo sapiens		3.1e-14
>	MYG_HUMAN	Myoglobin	Homo sapiens		2.3e-06
>	NGB_HUMAN	Neuroglobin	Homo sapiens		0.0017

(show all) alignments | Your search took: 0.06 secs | showing rows 1 - 12 of 12

Local Link

PFAM: Protein Family Database of Profile HMMs

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

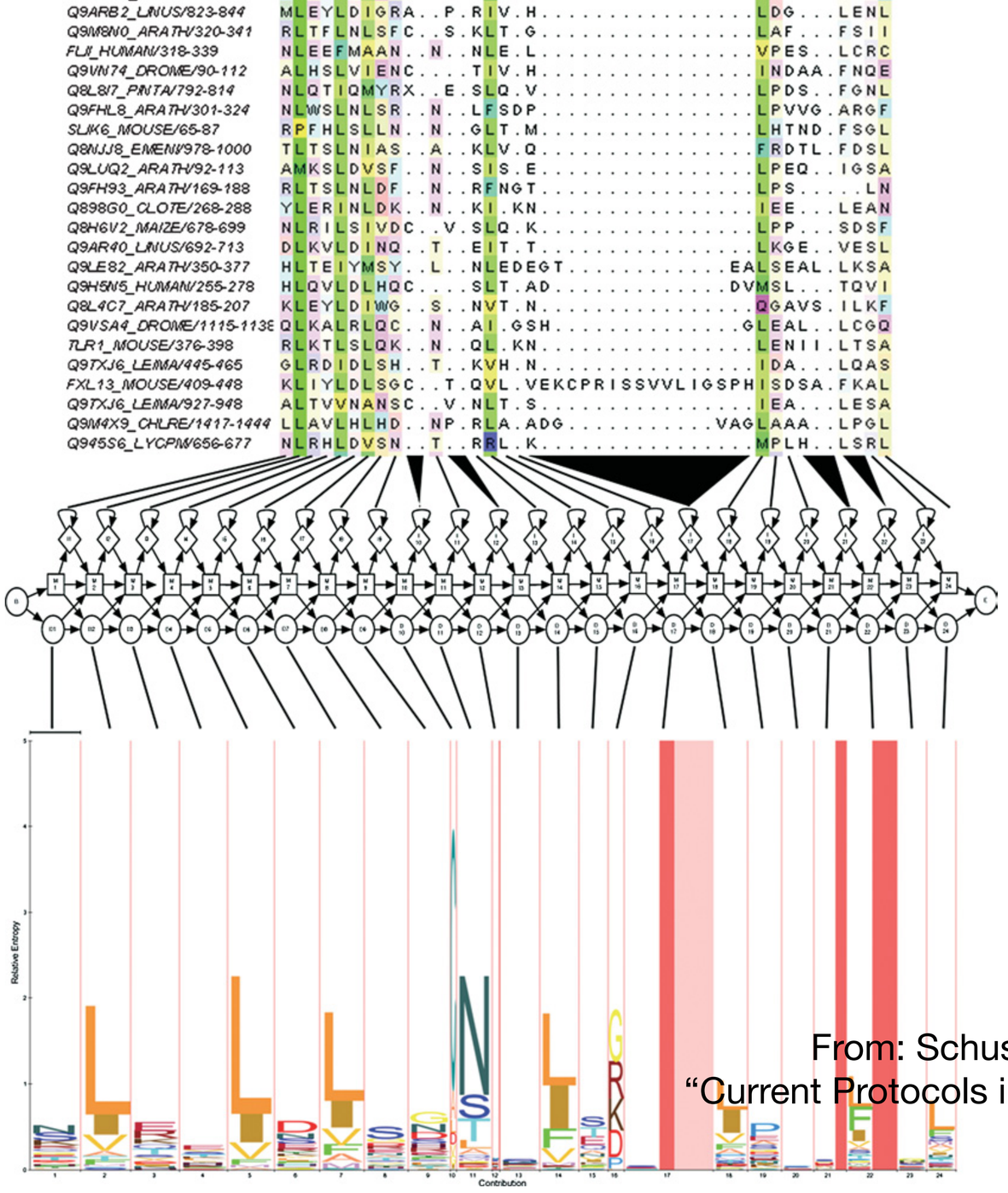
http://pfam.sanger.ac.uk/

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.

YOUR TURN!

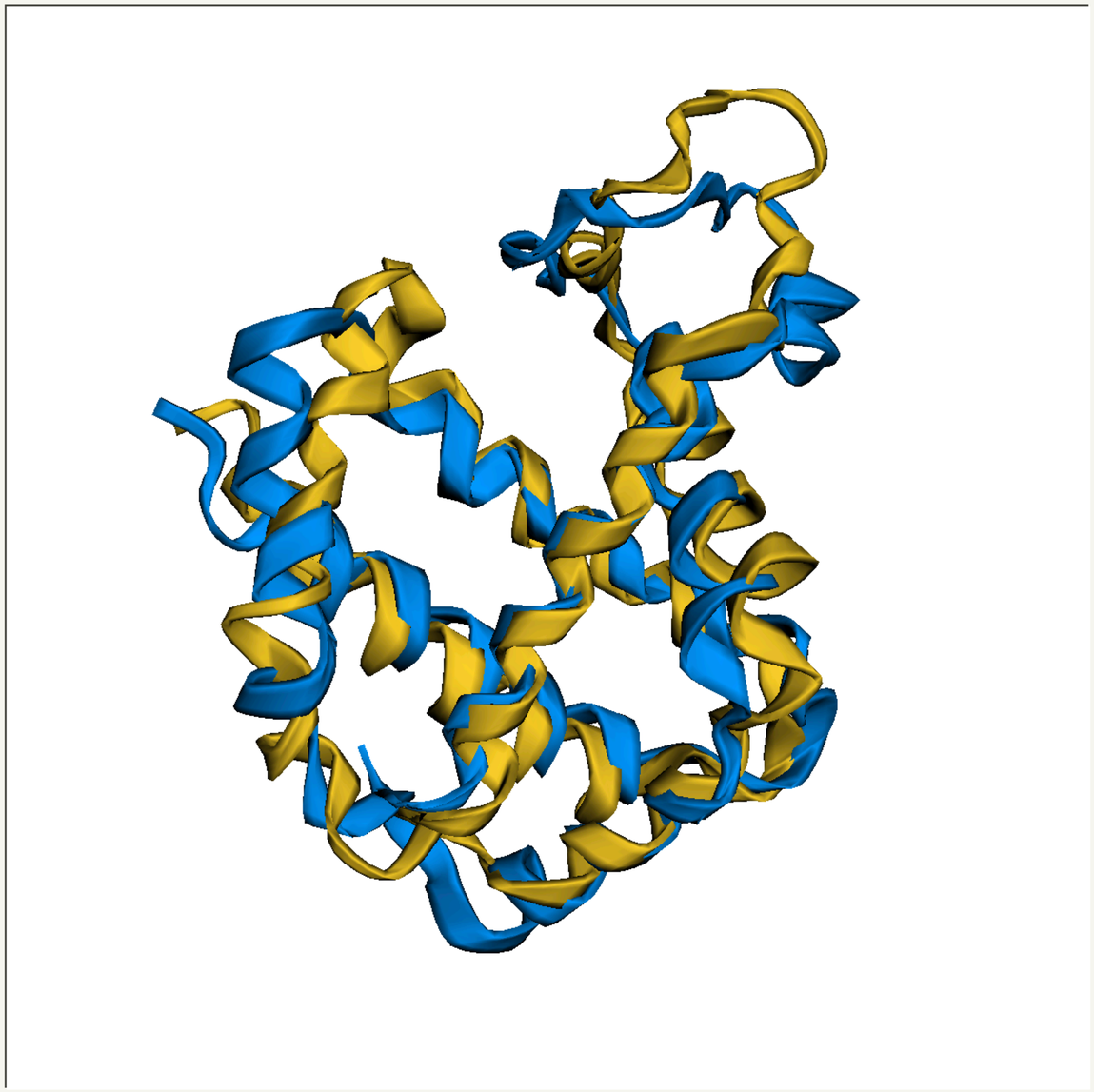
- There are **four required** and **one optional** hands-on sections including:

1. Limits of using BLAST [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. Examining conservation patterns [~20 mins]

— BREAK [15 mins]—

4. [Optional] Using HMMER [~10 mins]
5. **Divergence of protein sequence and structure** [~25 mins]

- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!



ALIGNMENT | CONTACT MAP

```
Align 2hbsB.pdb 146 with 4mpmB.pdb 148
Twists 0 ini-len 136 ini-rmsd 3.05 opt-equ 143 opt-rmsd 2.65 chain-rmsd 3.05
Score 318.72 align-len 150 gaps 7 (4.67%)
P-value 3.26e-14 Afp-num 14073 Identity 20.67% Similarity 40.00%
Block  0 afp 17 score 318.72 rmsd  3.05 gap 9 (0.06%)

                .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1     2 HLTPVEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFG-DLSTPDAVMGNPKVKAHGKKVL
              111  111111111111  11111111111111111111111111 11111111111111111111
Chain 2     2 ERP--EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVM

Chain 1    69 GAFSDGLAHLDNLKGTFATLSELHCD--KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAG
              1111111111111111111111111  1111111111111111111111111111111111111111
Chain 2    70 LVIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGA

Chain 1   137 VANALAHKYH
              1111111111
Chain 2   140 VVQAMSRGWD
```

Summary

- Find a gene project: You can start working on this now. Submit your responses to Q1-Q4 to get feedback.
- PSI-BLAST algorithm: Application of iterative position specific scoring matrices (PSSMs) to improve BLAST sensitivity
- Hidden Markov models (HMMs): More versatile probabilistic model for detection of remote similarities
- Structure comparisons as gold standards: Structure is more conserved than sequence

Homework: DataCamp!

Install R and RStudio (see website)

Complete the Introduction to R course on DataCamp
(Check Piazza for your DataCamp invite and sign up with your UCSD email (i.e. first part of your email address) please.

Let me know NOW if you don't have access to DataCamp!