



**BGGN 213**

**Hands-on Lab Session**

**Class 03**

**Barry Grant**

**UC San Diego**

<http://thegrantlab.org/bggn213>

# Class 3: Hands-on section

Week 2

<http://thegrantlab.org/bggn213/>

The screenshot shows a web browser window displaying the course page for BGGN 213. The page is titled "UC San Diego BGGN 213" and includes a navigation menu with links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, GDrive, Atmosphere, CloudLaunch, BIMM194, Blink, and News. The main content area features a table with the following schedule entries:

Class	Date	Topic
2	Fri 10/01/21	Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations.
3	Wed 10/06/21	<b>Project: Find a gene project assignment</b> (Part 1) Principles of database searching, due in 2 weeks. (Part 2) Sequence analysis, structure analysis and general data analysis with R due at the end of the quarter.
*	Wed 10/06/21	<b>Optional: Advanced sequence alignment and database searching</b> Detecting remote sequence similarity, Database searching beyond BLAST, Substitution matrices, Using PSI-BLAST, Profiles and HMMs, Protein structure comparisons as a gold standard.
4	Fri 10/08/21	<b>Bioinformatics data analysis with R</b> Why do we use R for bioinformatics? R language basics and the RStudio IDE, Major R data structures and functions, Using R interactively from the RStudio console. Introducing Rmarkdown documents.

Red arrows point to the "Schedule" link in the left sidebar and the Class 3 row in the table.

# Find-a-Gene Project Assignment

- A total of 35% of the course grade will be assigned based on the [“find-a-gene project assignment”](#)

# Find-a-Gene Project Assignment

- A total of 35% 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 35% 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 35% 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 35% 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 its 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 alignment for building a phylogenetic tree that illustrates species divergence.



# UC San Diego

## BGGN 213

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

### (Project:) Find a Gene Assignment Part 1

The [find-a-gene project](#) is a required assignment for BGGN-213. 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 **Wednesday Oct 20th (10/20/21)** at 12pm San Diego time.
- The complete assignment, including responses to all questions, is due **Friday Dec 3rd (12/03/21)** 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.

The screenshot shows a web browser window with the address bar displaying 'bioboot.github.io'. The browser's tab bar shows a tab titled 'Schedule · BGGN 213'. The page content includes a navigation menu with links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, GDrive, Atmosphere, CloudLaunch, BIMM194, Blink, News, and a plus sign. On the left, there is a blue banner for 'UC San Diego BGGN 213' with the text 'A hands-on introduction to'. The main content area features a heading '(Project:) Find a Gene Assignment Part 1' and a paragraph: 'The [find-a-gene project](#) is a required assignment for BGGN-213. 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.'

- 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/bggn213/>

The screenshot shows a web browser window displaying the course page for BGGN 213. The page includes a navigation menu with links to Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, GDrive, Atmosphere, CloudLaunch, BIMM194, Blink, and News. The main content area features a sidebar on the left with the UC San Diego logo and course title 'BGGN 213', followed by a description and a list of navigation links: Overview, Schedule, Computer Setup, and Learning Goals. The 'Schedule' link is highlighted with a red box and an arrow. The main content area contains a table with the following rows:

Week	Date	Topic
2	Fri 10/01/21	Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations.
3	Wed 10/06/21	<b>Project: Find a gene project assignment</b> (Part 1) Principles of database searching, due in 2 weeks. (Part 2) Sequence analysis, structure analysis and general data analysis with R due at the end of the quarter.
*	Wed 10/06/21	<b>Optional: Advanced sequence alignment and database searching</b> Detecting remote sequence similarity, Database searching beyond BLAST, Substitution matrices, Using PSI-BLAST, Profiles and HMMs, Protein structure comparisons as a gold standard.
4	Fri 10/08/21	<b>Bioinformatics data analysis with R</b> Why do we use R for bioinformatics? R language basics and the RStudio IDE, Major R data structures and functions, Using R interactively from the RStudio console. Introducing Rmarkdown documents.

► Details:

Sequence 1

Sequence 2

Match Score Mismatch Score Gap Score

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**  
 $-6 + 1$  (Due to a match between G & G) = -5

**Score from Upper cell**  
 $-8 + -2$  (The Gap score) = -10

**Score from Side cell**  
 $-3 + -2$  (The Gap score) = -5

Winning (max) score is -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.](#)

# 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]

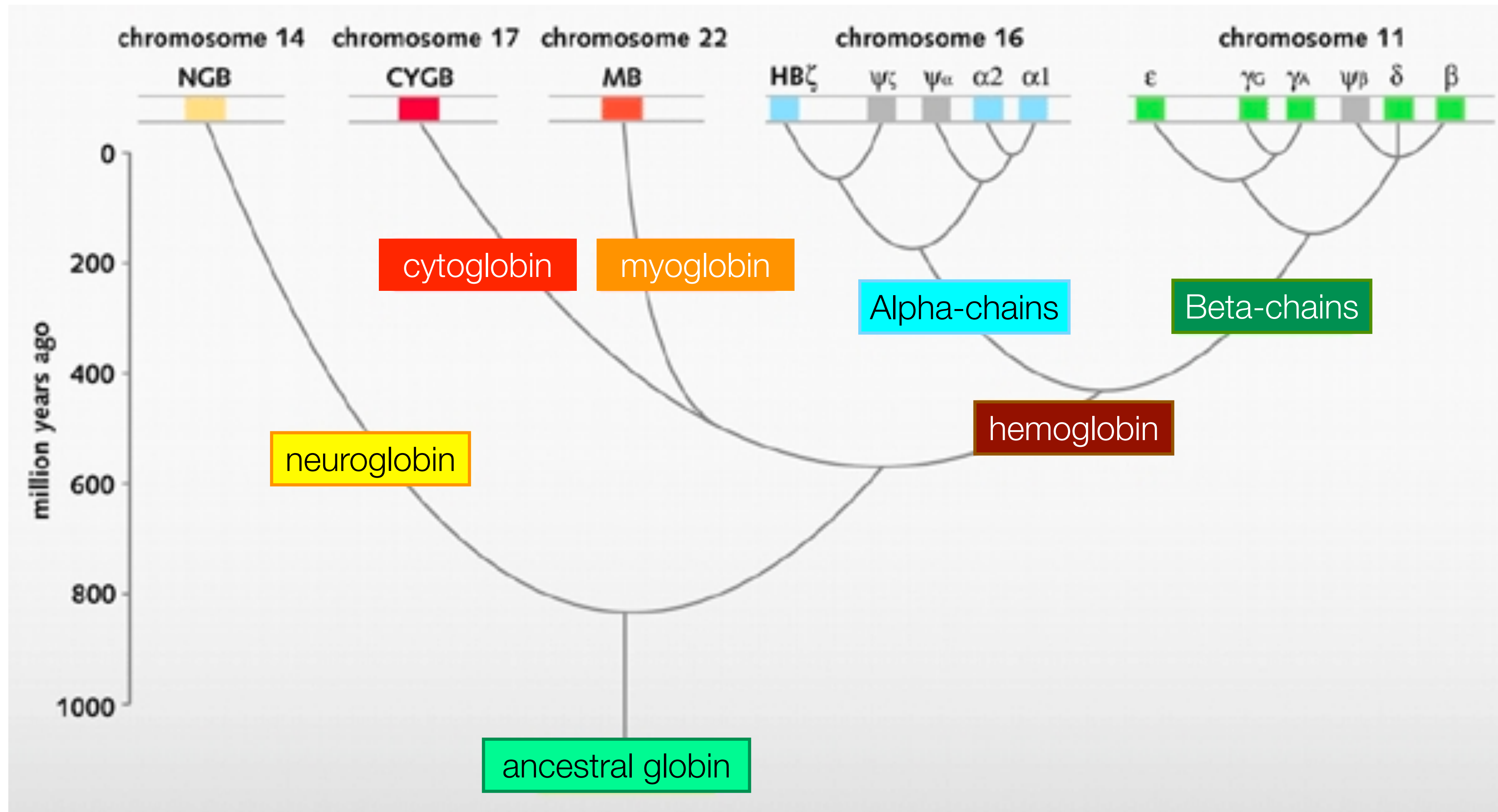
- ▶ Please do answer the last review question (**Q20**).
- ▶ We encourage discussion at your **Table** and on **Piazza!**

# 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]

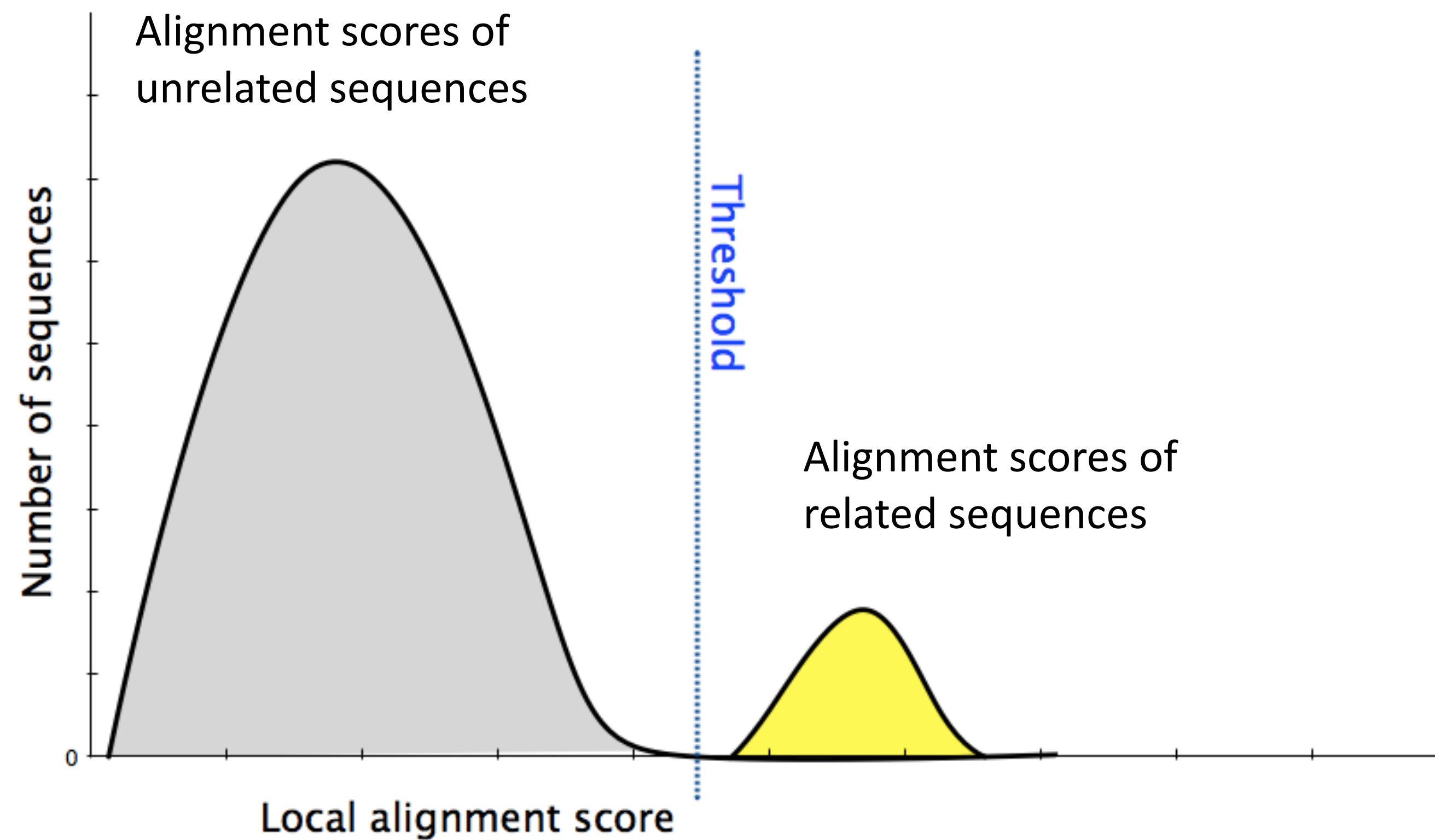
- ▶ Please do answer the last review question (**Q20**).
- ▶ We encourage discussion at your **Table** and on **Piazza!**



**An evolutionary model of human globins.**

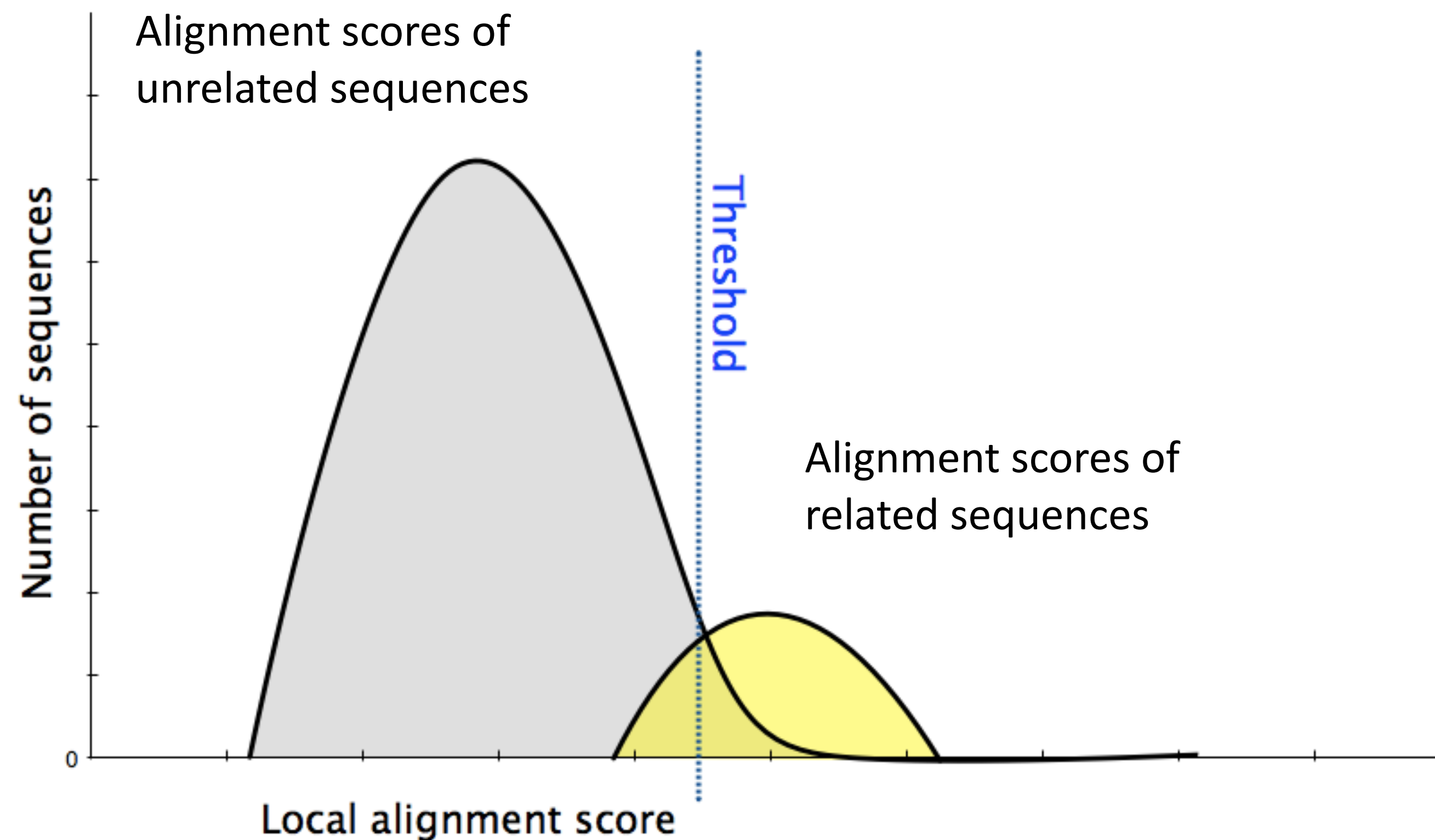
The different locations of globin genes in human chromosomes are reported at the top of the figure, distinguishing between the functional genes (in color) and the pseudogenes (in grey).

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

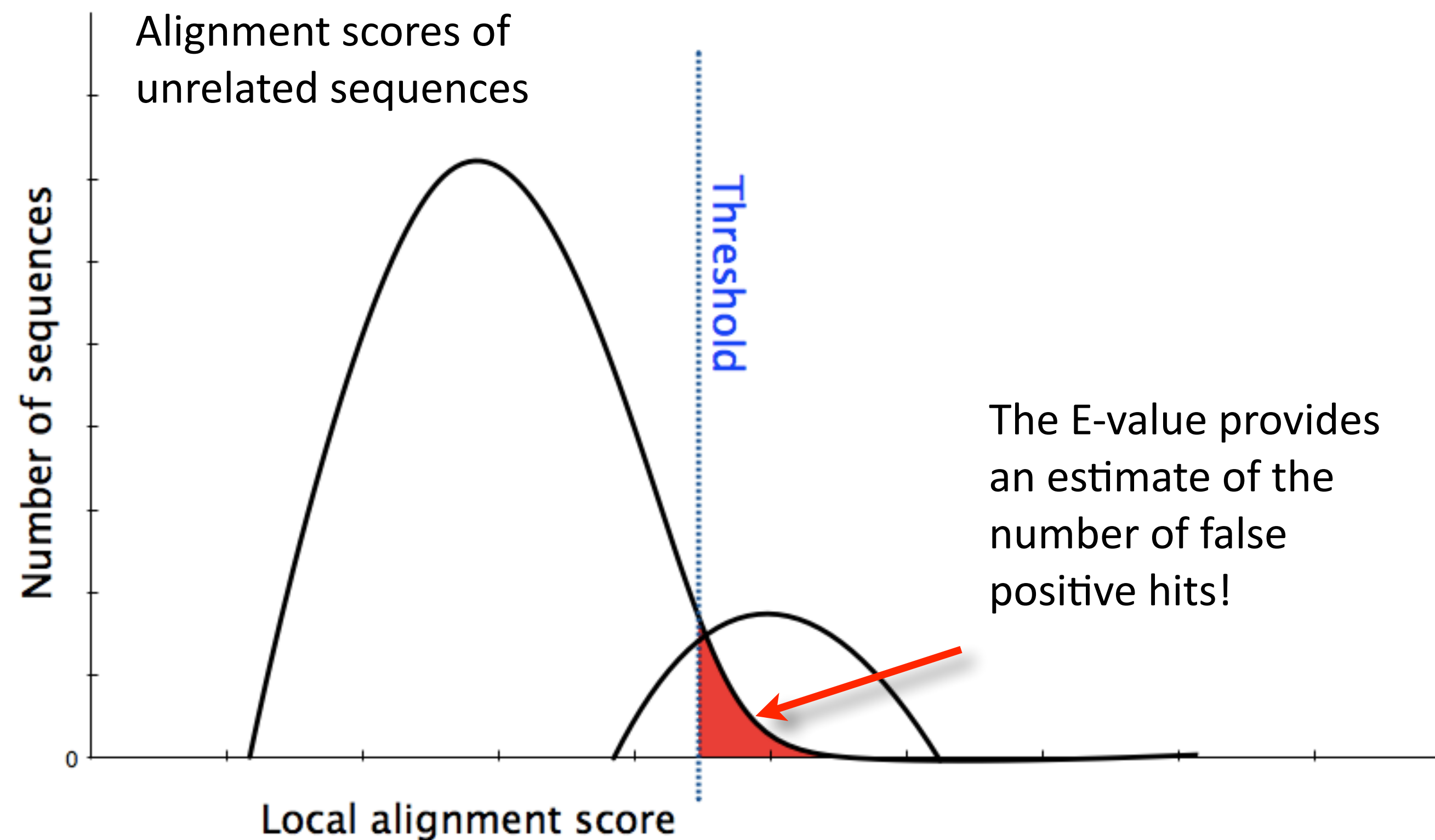




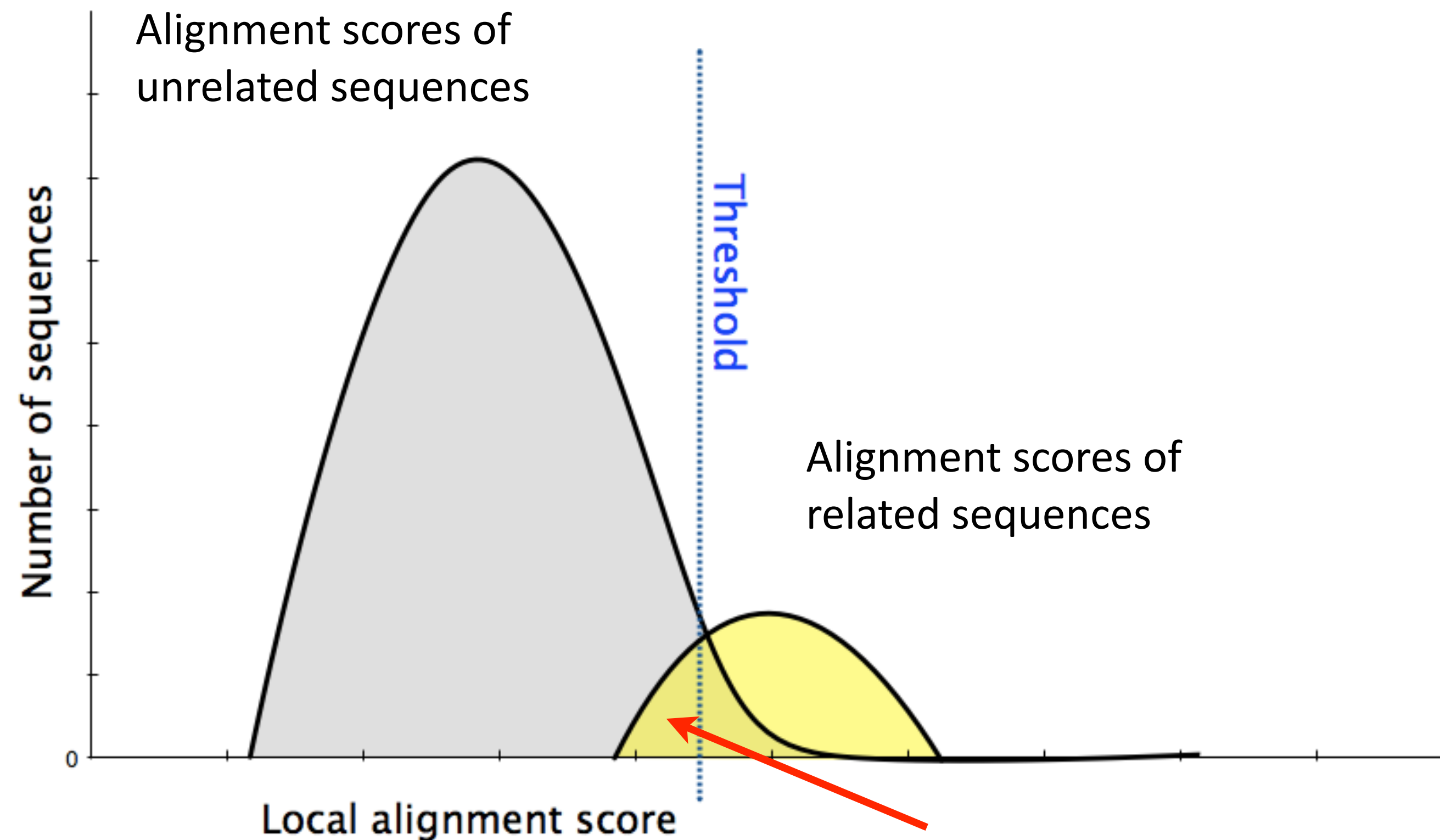
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



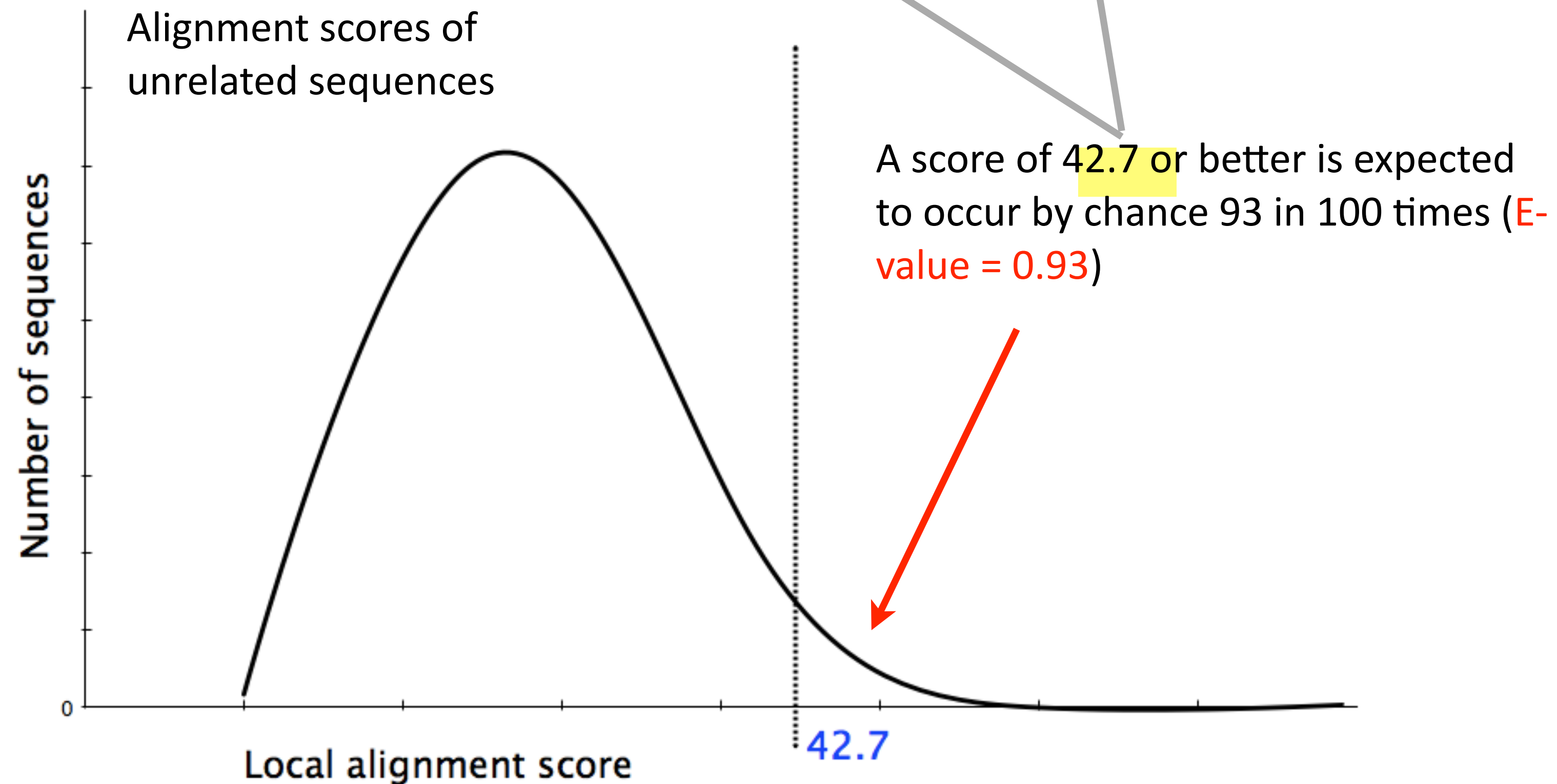
- Unfortunately, often both score distributions overlap
  - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- 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



# 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:** 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).

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 adjustment

## Filters and Masking

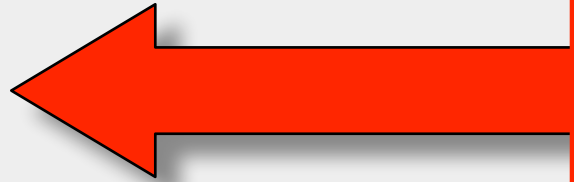
**Filter** |  Low complexity regions

**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

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

**B**locks **S**ubstitution **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	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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

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

# 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

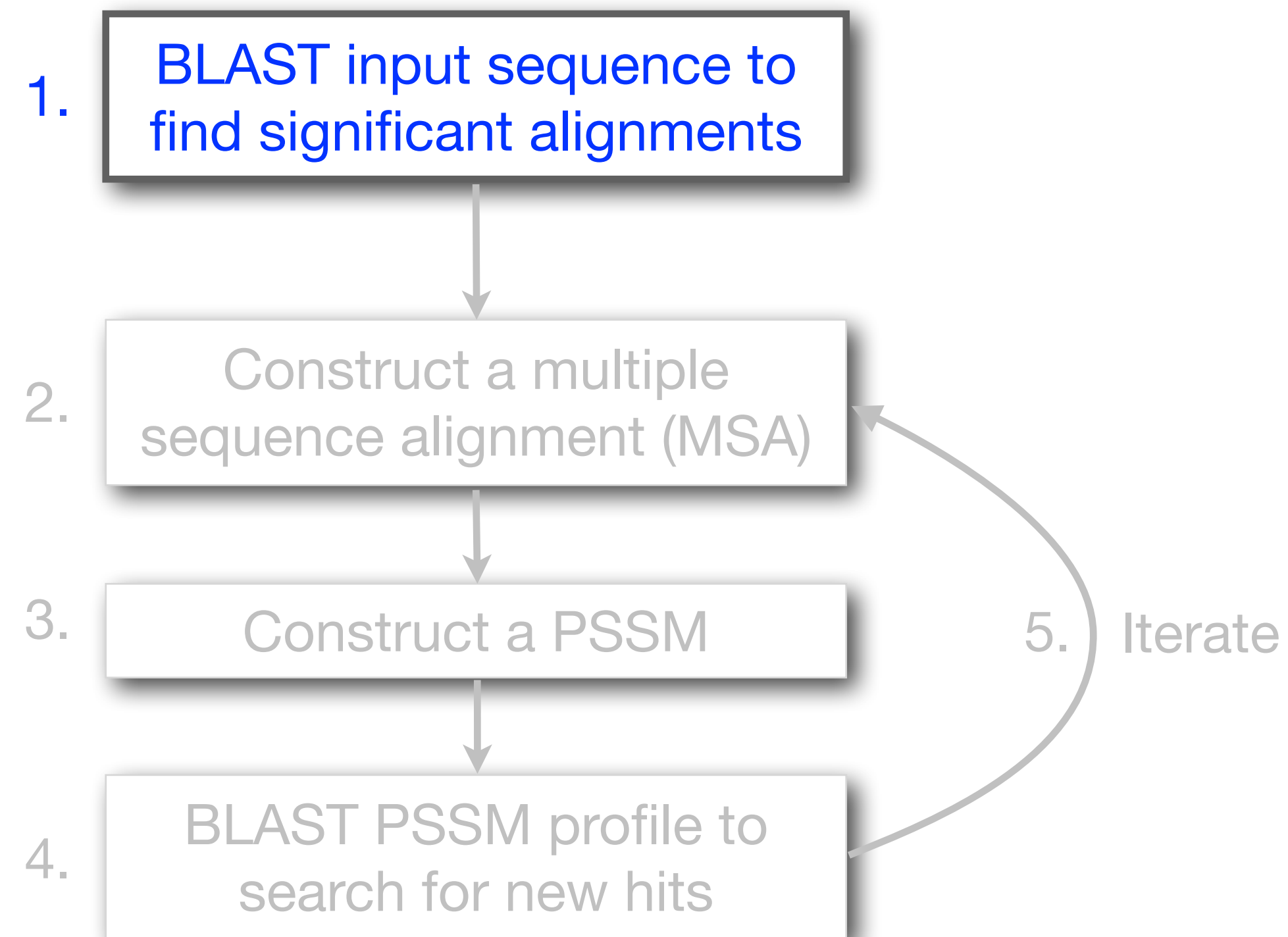
# PSI-BLAST: Position specific iterated BLAST

- 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

# 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

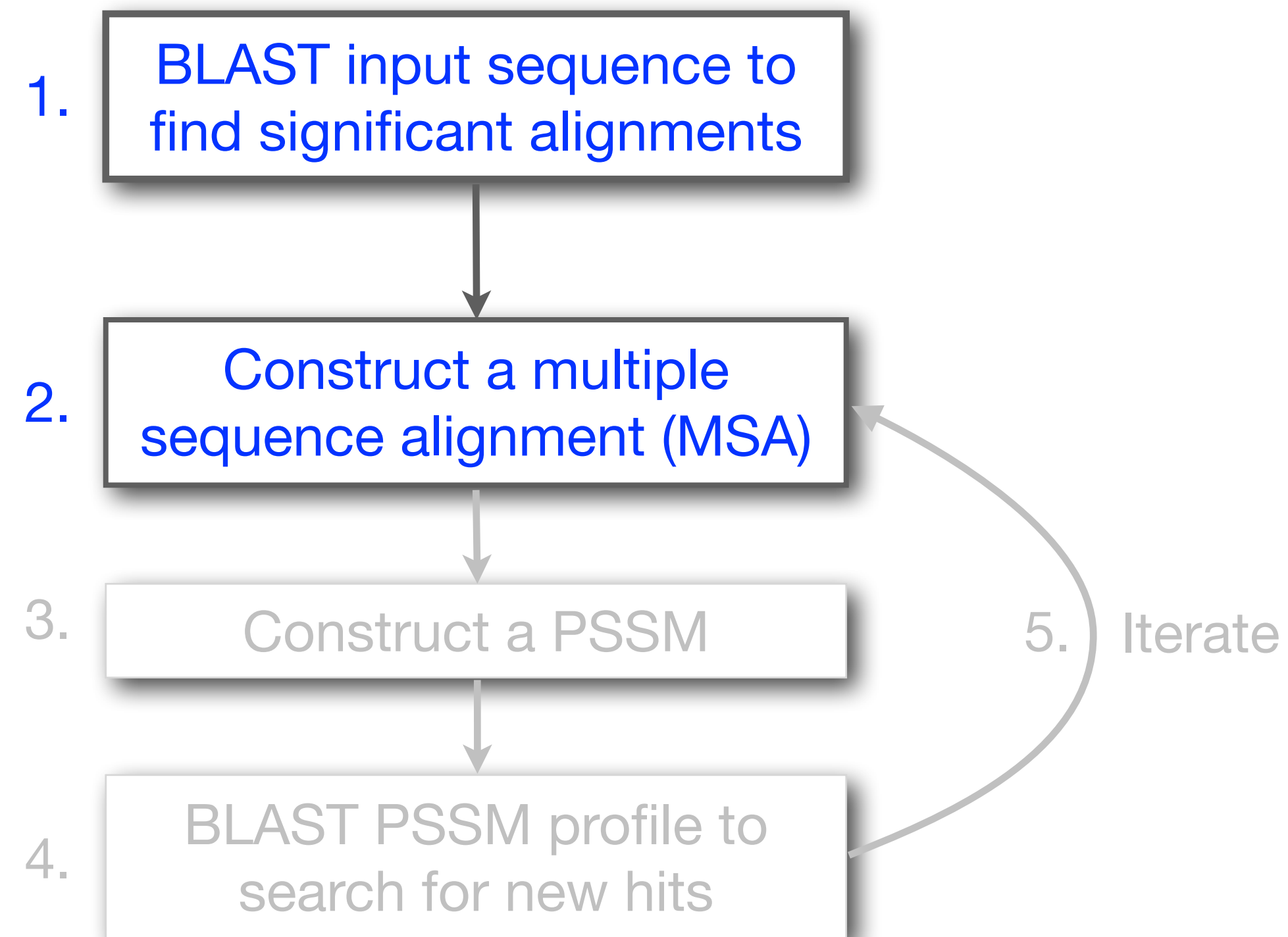


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

# 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

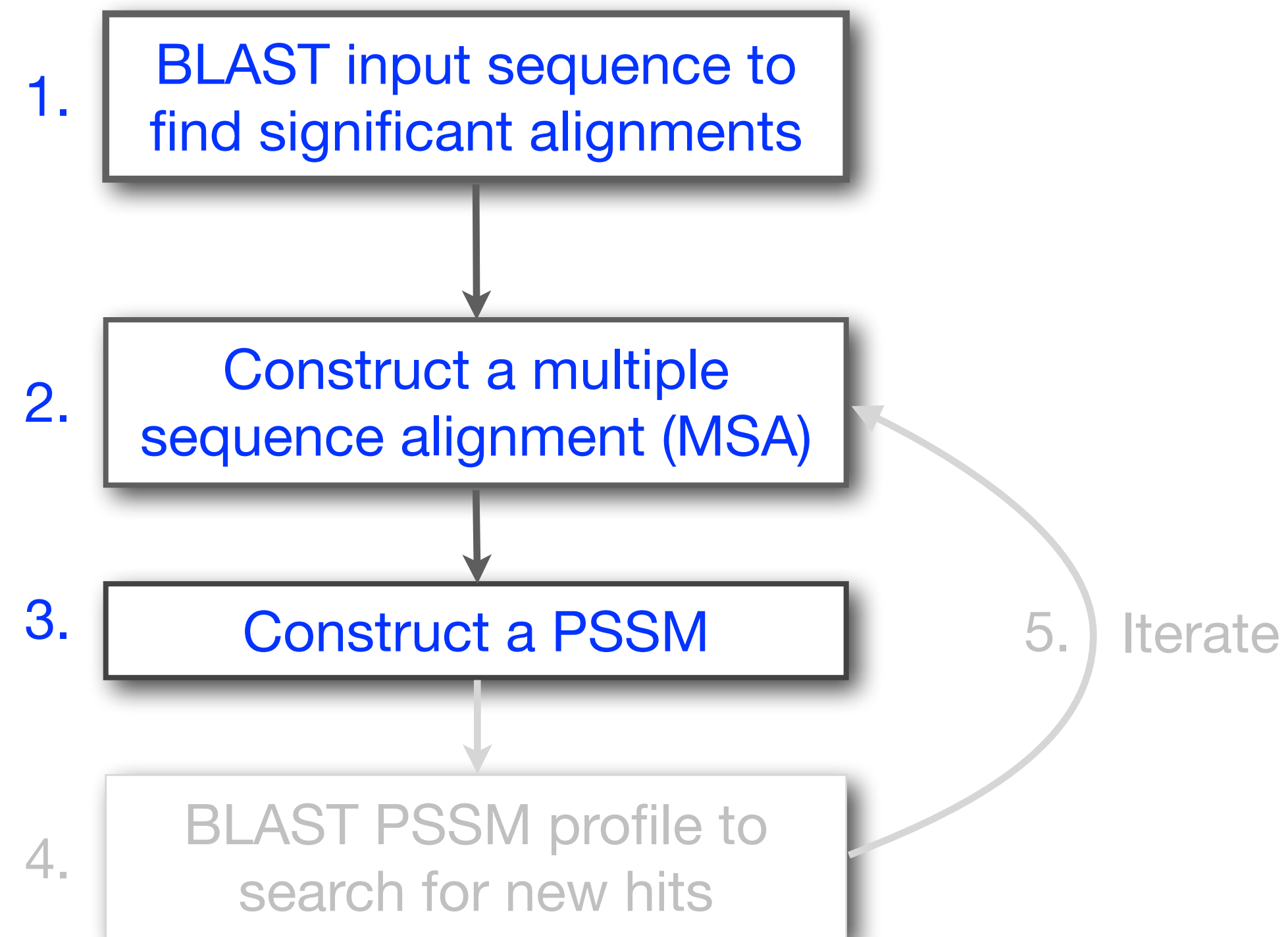


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

# 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)

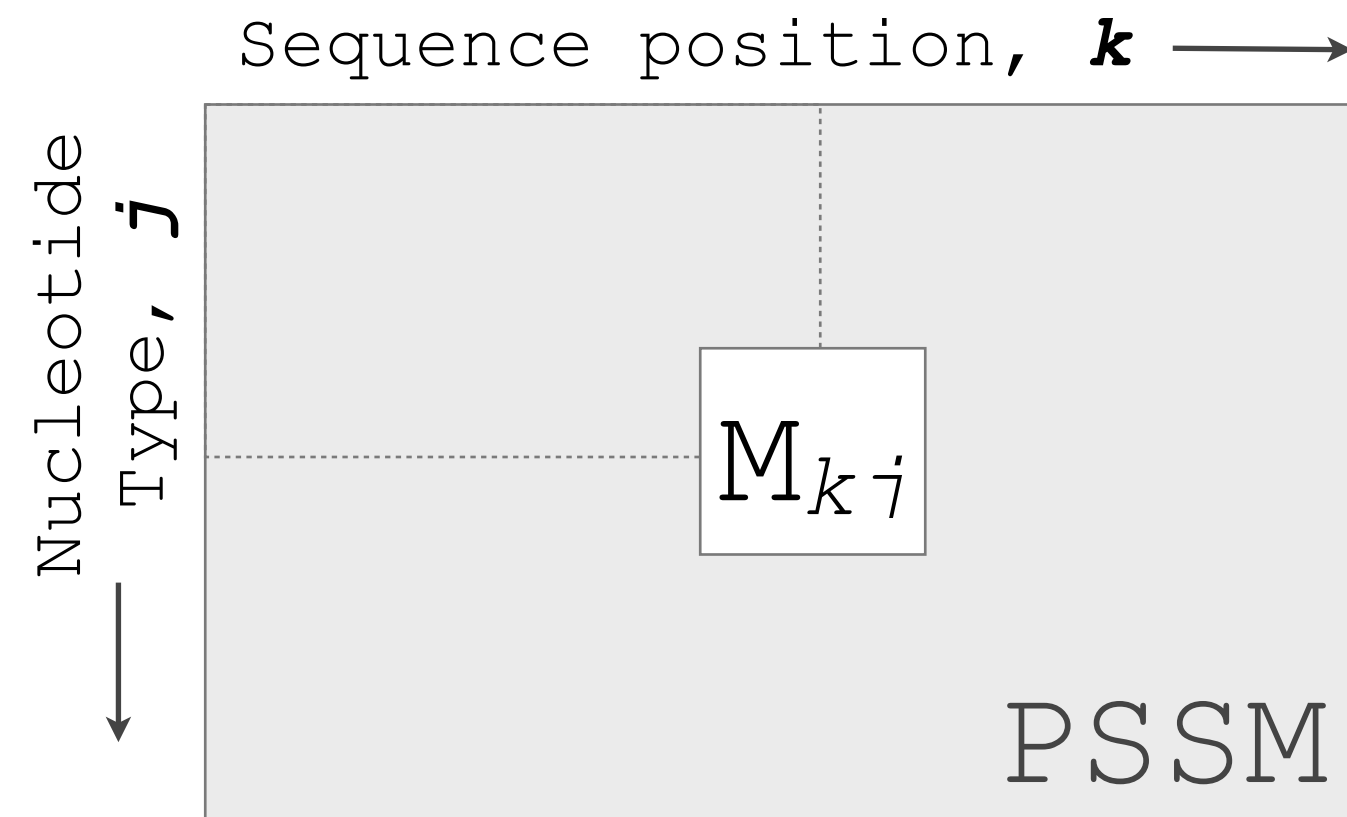
What is a **PSSM**?

# 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  $j$



## Example: Computing a transcription factor bind site PSSM

CCAAATTAGGAAA  
CCTATTAAAGAAAA  
CCAAATTAGGAAA  
CCAAATTCGGATA  
CCCATTTCGAAAA  
CCTATTTAGTATA  
CCAAATTAGGAAA  
CCAAATTGGCAAAA  
TCTATTTTGGAAA  
CCAAATTTCAAAA

The image shows a 10x13 grid of nucleotide sequences. Each row represents a sequence, and each column represents a position. The sequences are: Row 1: CCAAATTAGGAAA; Row 2: CCTATTAAAGAAAA; Row 3: CCAAATTAGGAAA; Row 4: CCAAATTCGGATA; Row 5: CCCATTTCGAAAA; Row 6: CCTATTTAGTATA; Row 7: CCAAATTAGGAAA; Row 8: CCAAATTGGCAAAA; Row 9: TCTATTTTGGAAA; Row 10: CCAAATTTCAAAA. The grid is overlaid with colored blocks: a green vertical bar at column 1; red vertical bars at columns 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, and 13; a blue vertical bar at column 5; and a yellow vertical bar at column 10.

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

CCAAATTAGGAAA  
CCTATTAAGAAAA  
CCAAATTAGGAAA  
CCAAATTCGGATA  
CCCATTTCGAAAA  
CCTATTTAGTATA  
CCAAATTAGGAAA  
CCAAATTGGCAAA  
TCTATTTTGGAAA  
CCAATTTTCAAAA

**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

CCAAATTAGGAAA  
 CCTATTAAGAAA  
 CCAAATTAGGAAA  
 CCAAATTCGGATA  
 CCCATTTTCGAAAA  
 CCTATTTAGTATA  
 CCAAATTAGGAAA  
 CCAAATTGGCAAA  
 TCTATTTTGGAAA  
 CCAATTTTCAAAA

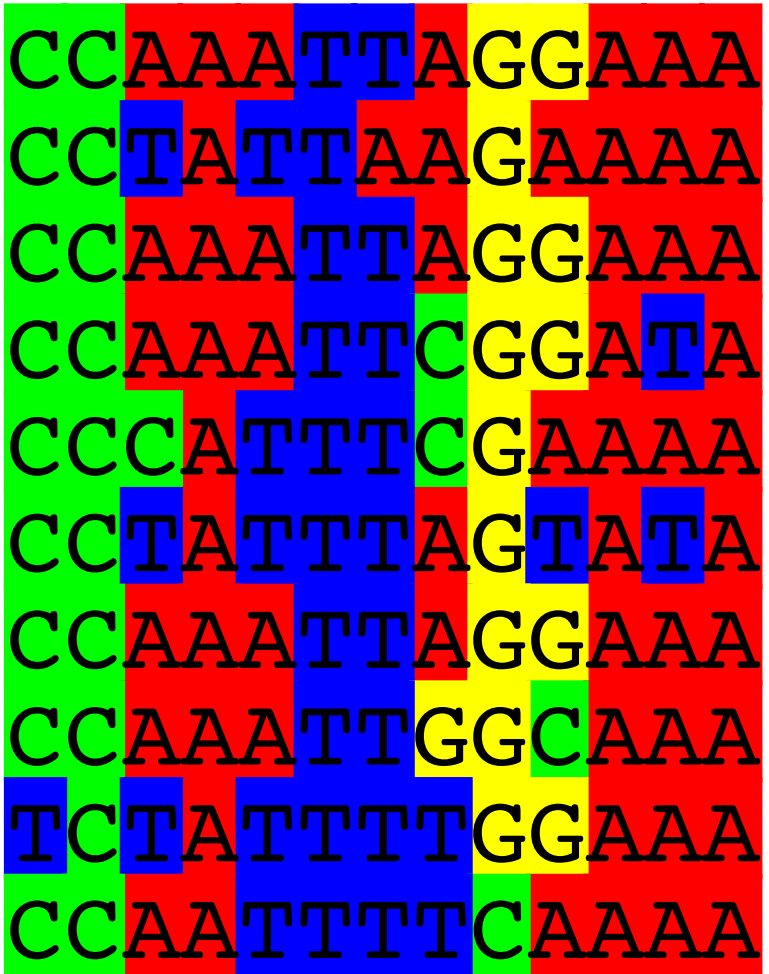
**Alignment Counts matrix:**

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												



# Computing a transcription factor bind site PSSM

CCAAATTAGGAAA  
CCTATTAAGAAA  
CCAAATTAGGAAA  
CCAAATTCGGATA  
CCCATTTCGAAAA  
CCTATTTAGTATA  
CCAAATTAGGAAA  
CCAAATTGGCAAA  
TCTATTTTGGAAA  
CCAATTTTCAAAA

**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

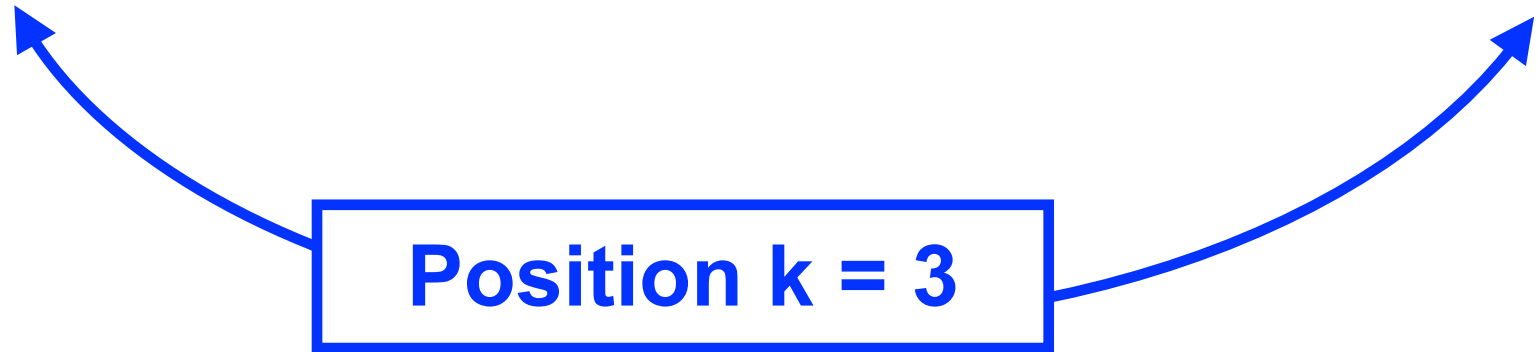
# Computing a transcription factor bind site PSSM

CCAAATTAGGAAA  
 CCTATTAAGAAA  
 CCAAATTAGGAAA  
 CCAAATTCGGATA  
 CCCATTTTCGAAAA  
 CCTATTTAGGTATA  
 CCAAATTAGGAAA  
 CCAAATTGGCAAAA  
 TCTATTTTGGAAA  
 CCAATTTTCAAAA

**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

CCAAATTAGGAAA  
 CCTATTAAGAAA  
 CCAAATTAGGAAA  
 CCAAATTCGGATA  
 CCCATTCGAAAA  
 CCTATTTAGTATA  
 CCAAATTAGGAAA  
 CCAAATTGGCAA  
 TCTATTTTGGAAA  
 CCAATTTCAAAA

**Alignment Counts matrix:**

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



# Computing a transcription factor bind site PSSM

```

CCAAATTAGGAAA
CCATTAAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAAATTTCAAAA
    
```

## 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

## 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

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

# Computing a transcription factor bind site PSSM

```

CCAAATTAGGAAA
CC TATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGGTATA
CCAAATTAGGAAA
CCAAATTGGCAAAA
TCTATTTTGGAAA
CCAATTTTCAAAA
    
```

## 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" = PSSM

$C_{kj}$  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  $k$

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

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

# 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
<b>A:</b>	0	0	6	10	5	0	1	5	0	3	10	8	10
<b>C:</b>	9	10	1	0	0	0	0	2	1	1	0	0	0
<b>G:</b>	0	0	0	0	0	0	0	1	9	5	0	0	0
<b>T:</b>	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
<b>A:</b>	-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
<b>C:</b>	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
<b>G:</b>	-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
<b>T:</b>	-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
<b>A:</b>	-2.4	-2.4	0.8	<b>1.3</b>	0.6	-2.4	-0.8	<b>0.6</b>	-2.4	0.2	<b>1.3</b>	1.1	<b>1.3</b>
<b>C:</b>	<b>1.2</b>	<b>1.3</b>	-0.8	-2.4	-2.4	-2.4	-2.4	-0.2	-0.8	-0.8	-2.4	-2.4	-2.4
<b>G:</b>	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	<b>1.2</b>	<b>0.6</b>	-2.4	-2.4	-2.4
<b>T:</b>	-0.8	-2.4	<b>0.2</b>	-2.4	<b>0.6</b>	<b>1.3</b>	<b>1.2</b>	-0.2	-2.4	-0.8	-2.4	<b>-0.2</b>	-2.4
Test seq:	<b>C</b>	<b>C</b>	<b>T</b>	<b>A</b>	<b>T</b>	<b>T</b>	<b>T</b>	<b>A</b>	<b>G</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>A</b>

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

# Scoring a test sequence

Query Sequence

**CCTATTAGGATA**

PSSM:

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

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

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

# Scoring a test sequence...

Query Sequence

**CCTATTAGGATA**

Best Possible Sequence

**CCAATTAGGAAA**

PSSM:

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

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

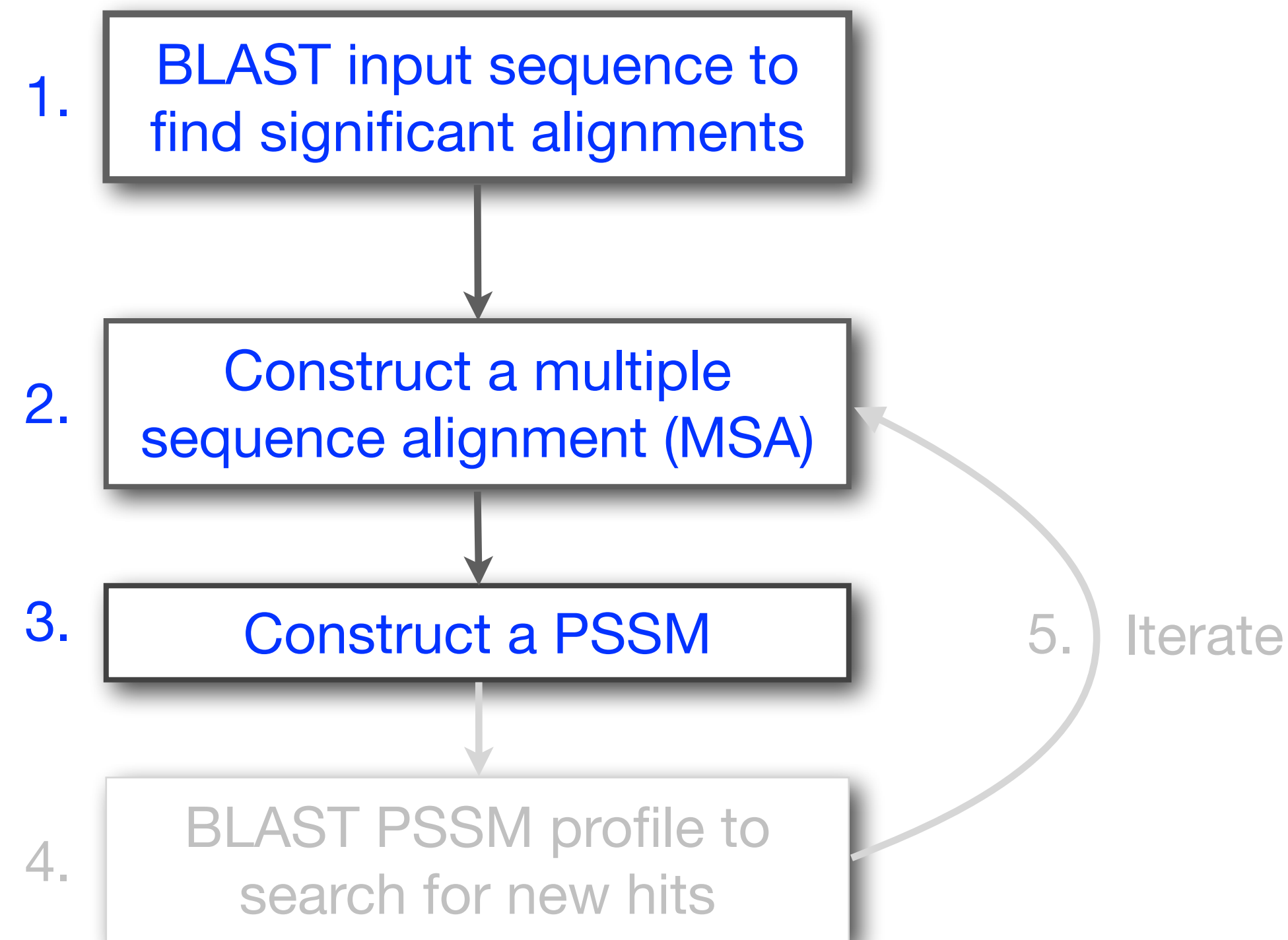
**A.** Following method in Harbison *et al.* (2004) Nature 431:99-104

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!

# 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 href="#">730496</a>	66	FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDEPAKFKMKYWGVASFLQKGNDDH	125
<a href="#">200679</a>	63	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDEPAKFKMKYWGVASFLQKGNDDH	122
<a href="#">206589</a>	34	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDEPAKFKMKYWGVASFLQKGNDDH	93
<a href="#">2136812</a>	2	MSATAKGRVRLLNWDVCADMVGTFTDTEDEPAKFKMKYWGVASFLQKGNDDH	53
<a href="#">132408</a>	65	FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH	124
<a href="#">267584</a>	44	FSVDESGKVTATAHGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQTGNDDH	103
<a href="#">267585</a>	44	FSVDGSGKVTATAQGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH	103
<a href="#">8777608</a>	63	FTIHEDGAMTATAKGRVILNNWEMCADMMATFETTPDPAKFRMRYWGAAAYLQTGNDDH	122
<a href="#">6687453</a>	60	FKVEEDGTMTATAIGRVILNNWEMCANMFGTFEDTEDEPAKFKMKYWGAAAYLQTYDDH	119
<a href="#">10697027</a>	81	FKVQEDGTMTATATGRVILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTYDDH	140
<a href="#">13645517</a>	1	MVGTFTDTEDEPAKFKMKYWGVASFLQKGNDDH	32
<a href="#">13925316</a>	38	FSVDGSGKMTATAQGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH	97
<a href="#">131649</a>	65	YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY	126

M

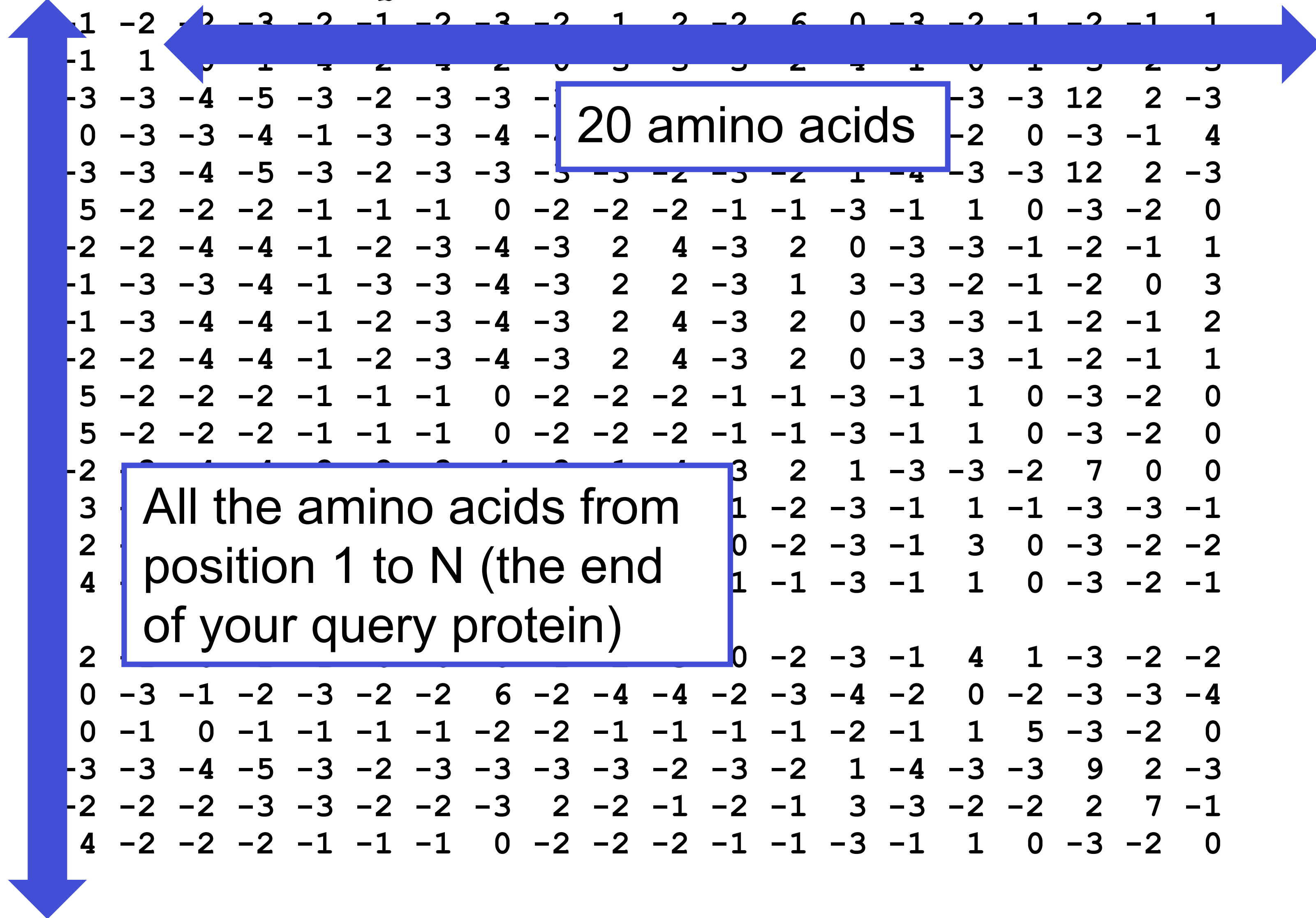
N,M,L,Y,G



		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	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	-2	-1	-2	-2	-2	-2	-2	3	3	3	2	-1	1	0	1	3	2	3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-3	-3	-2	-1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-3	-3	-3	-3	-2	-1	-4	-3	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-3	-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	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-2	-3	-3	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
14	A	3	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
15	A	2	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
16	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
...																					
37	S	2	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
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

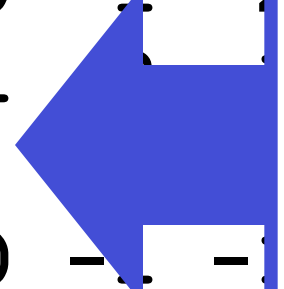
20 amino acids

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



		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

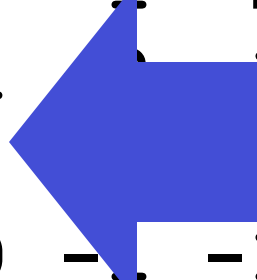
**Note:** A given amino acid (such as alanine) in your query protein can receive different scores for matching alanine depending on the position in the protein (BLOSUM  $S_{AA} = +4$ )



	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.

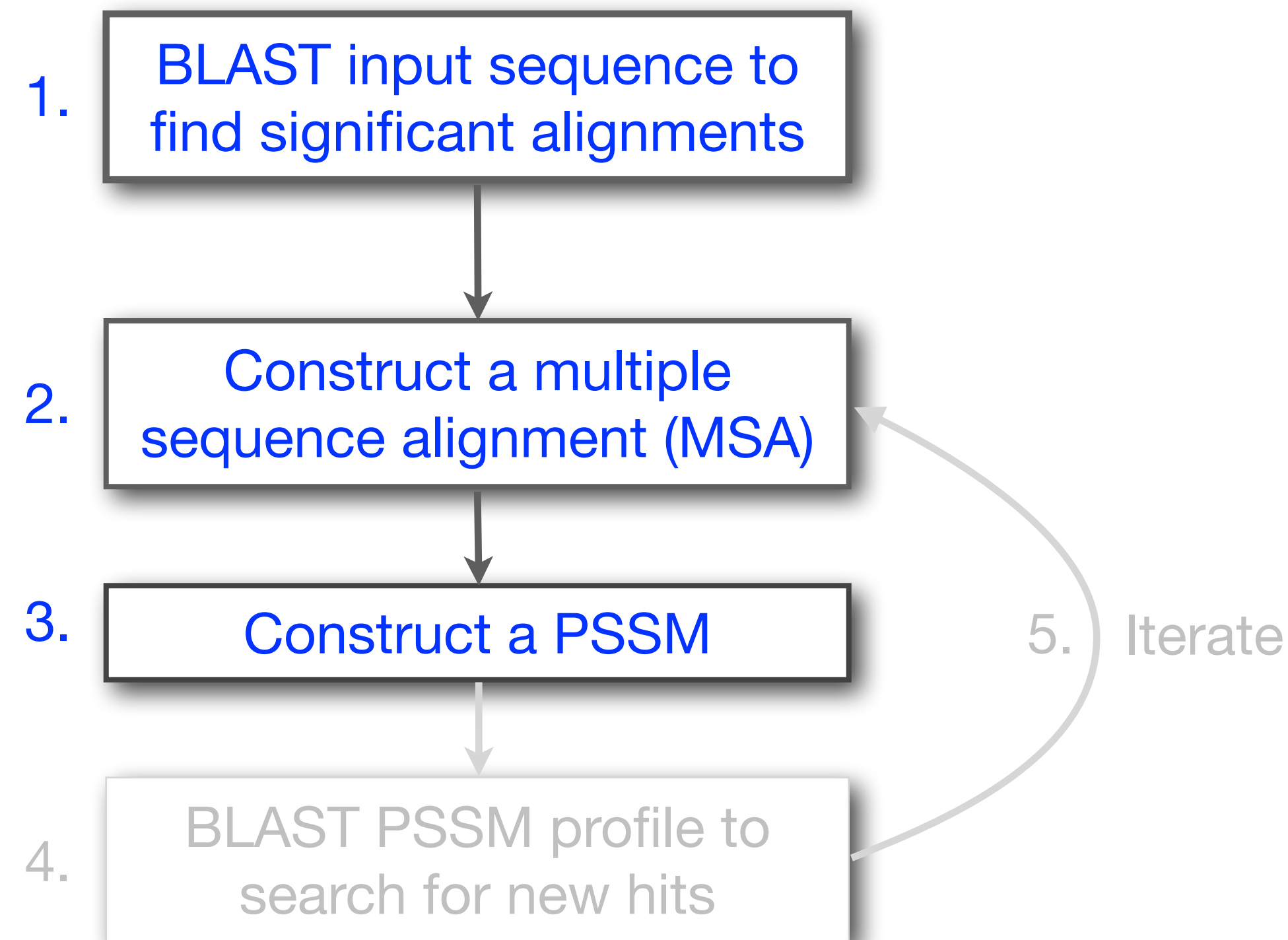
**Note:** A given amino acid (such as alanine) in your query protein can receive different scores for matching alanine depending on the position in the protein (BLOSUM  $S_{AA} = +4$ )



# 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

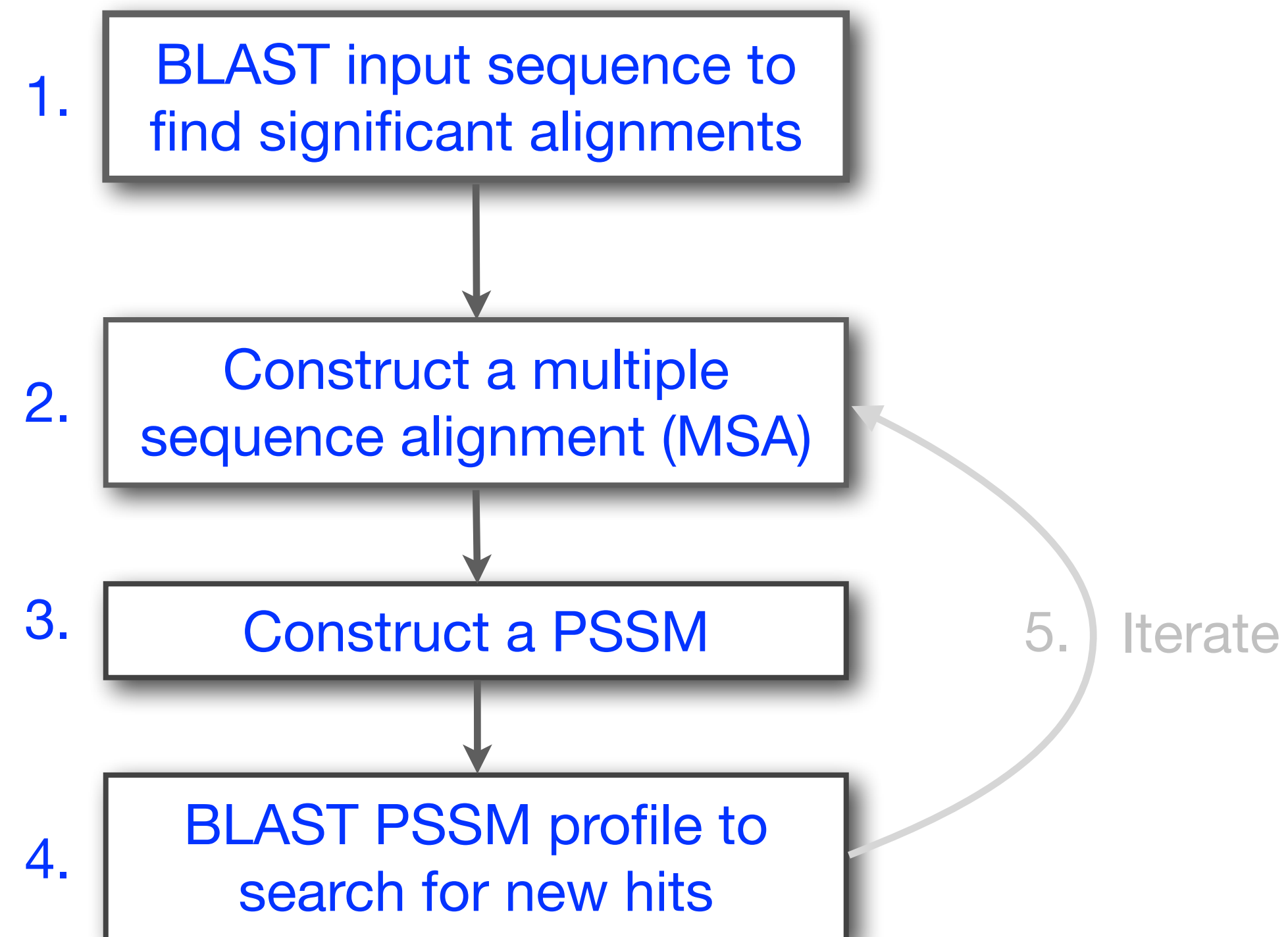


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

# 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

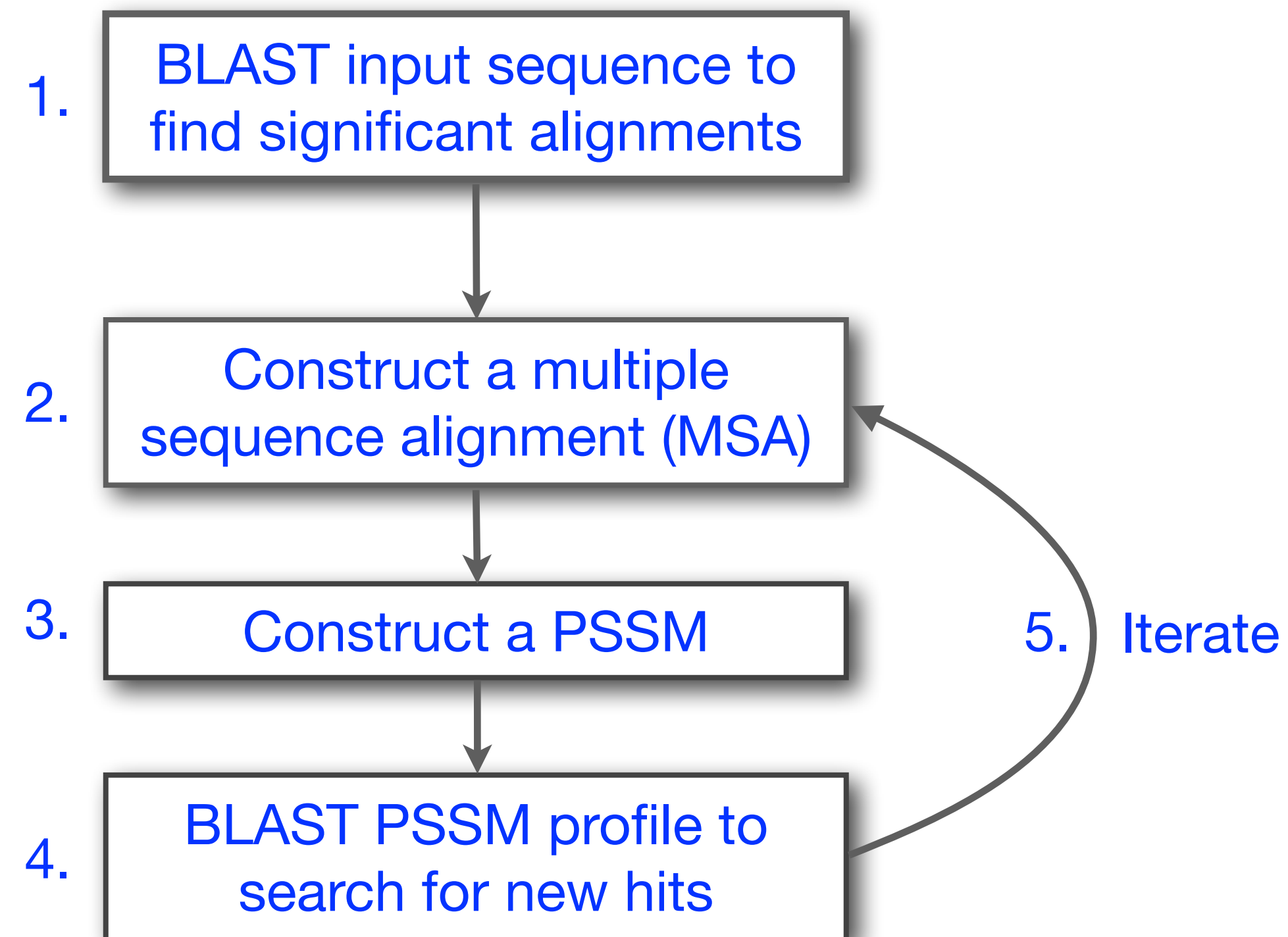


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

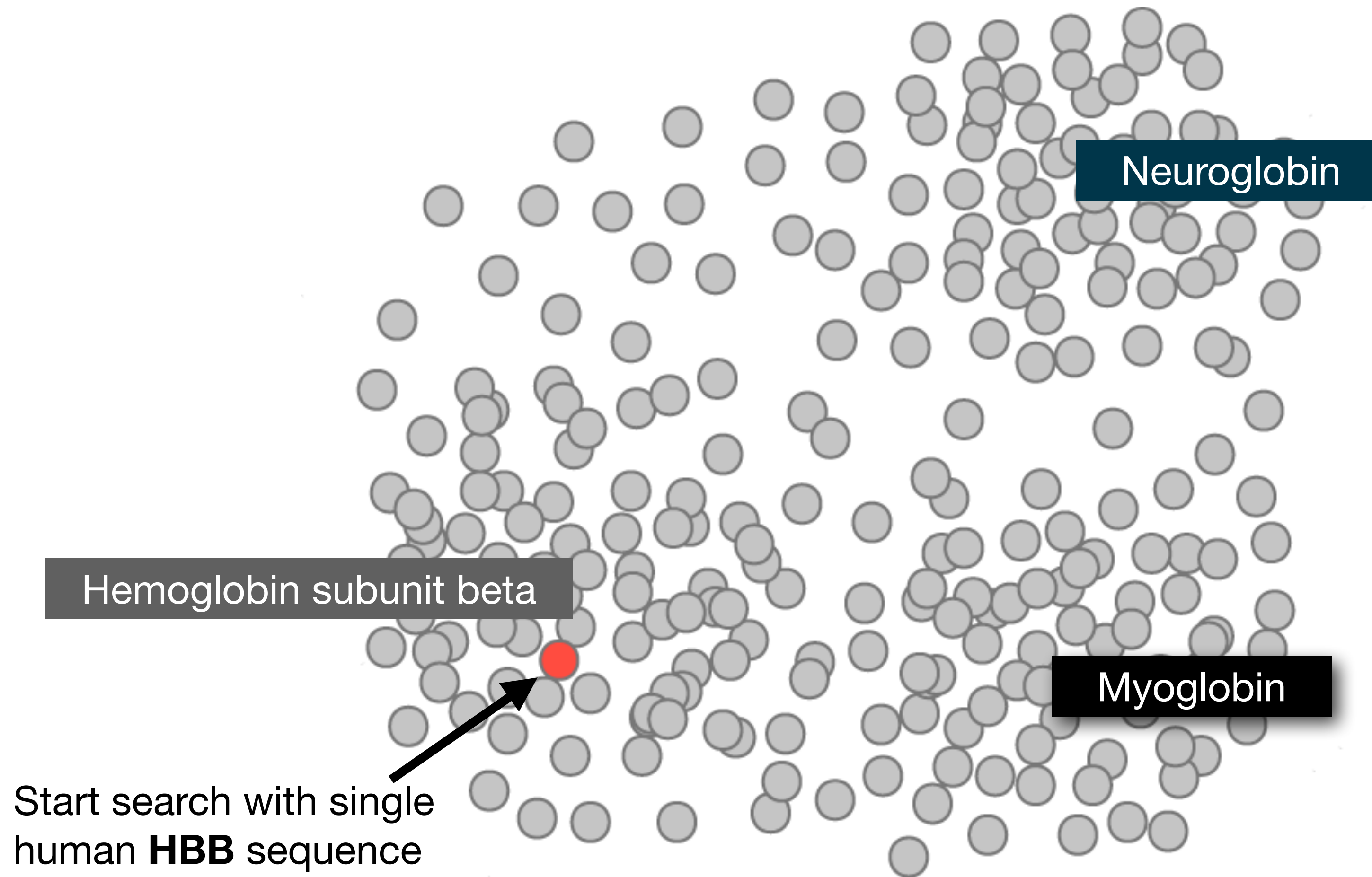
# PSI-BLAST: Position-Specific Iterated BLAST

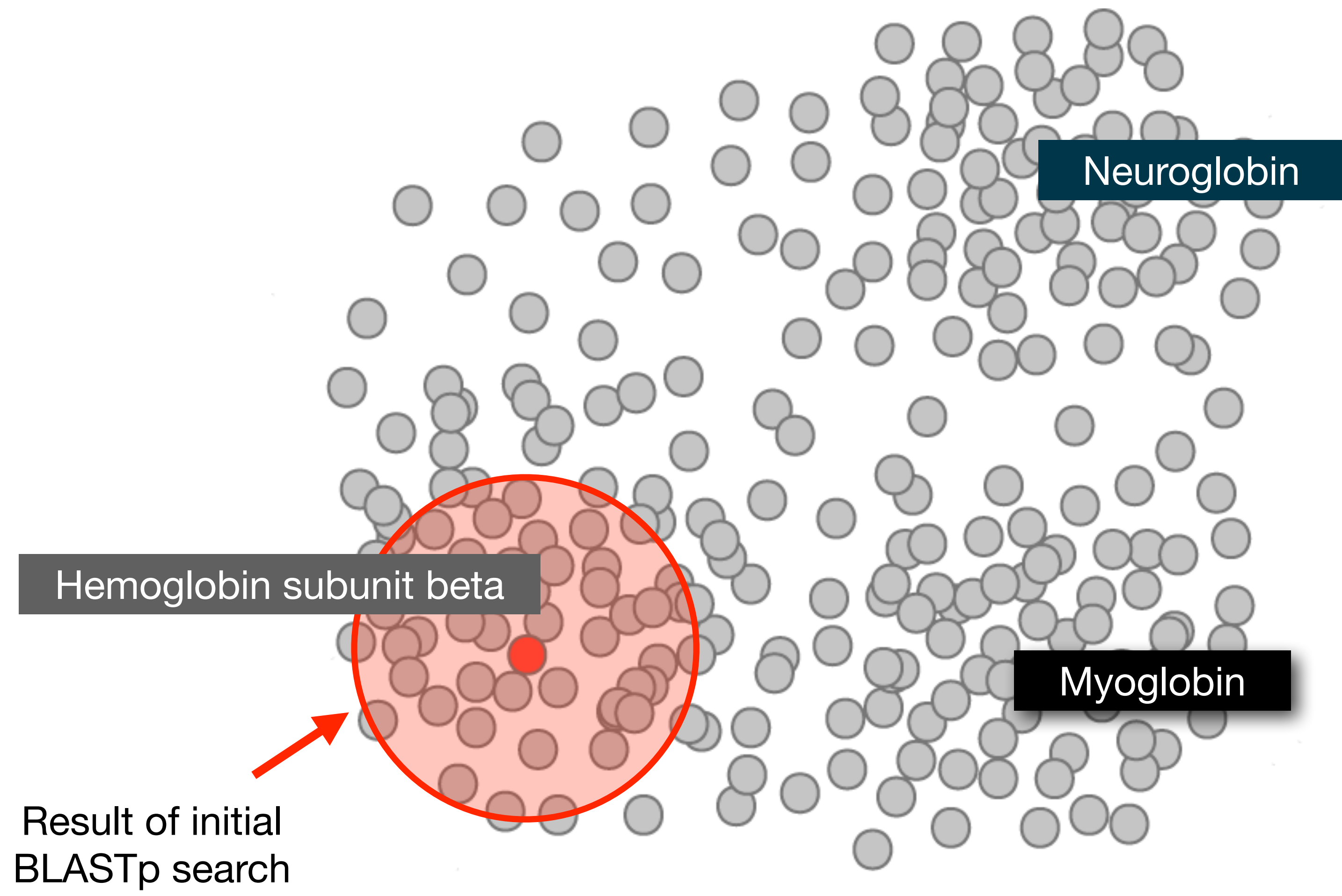
---

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

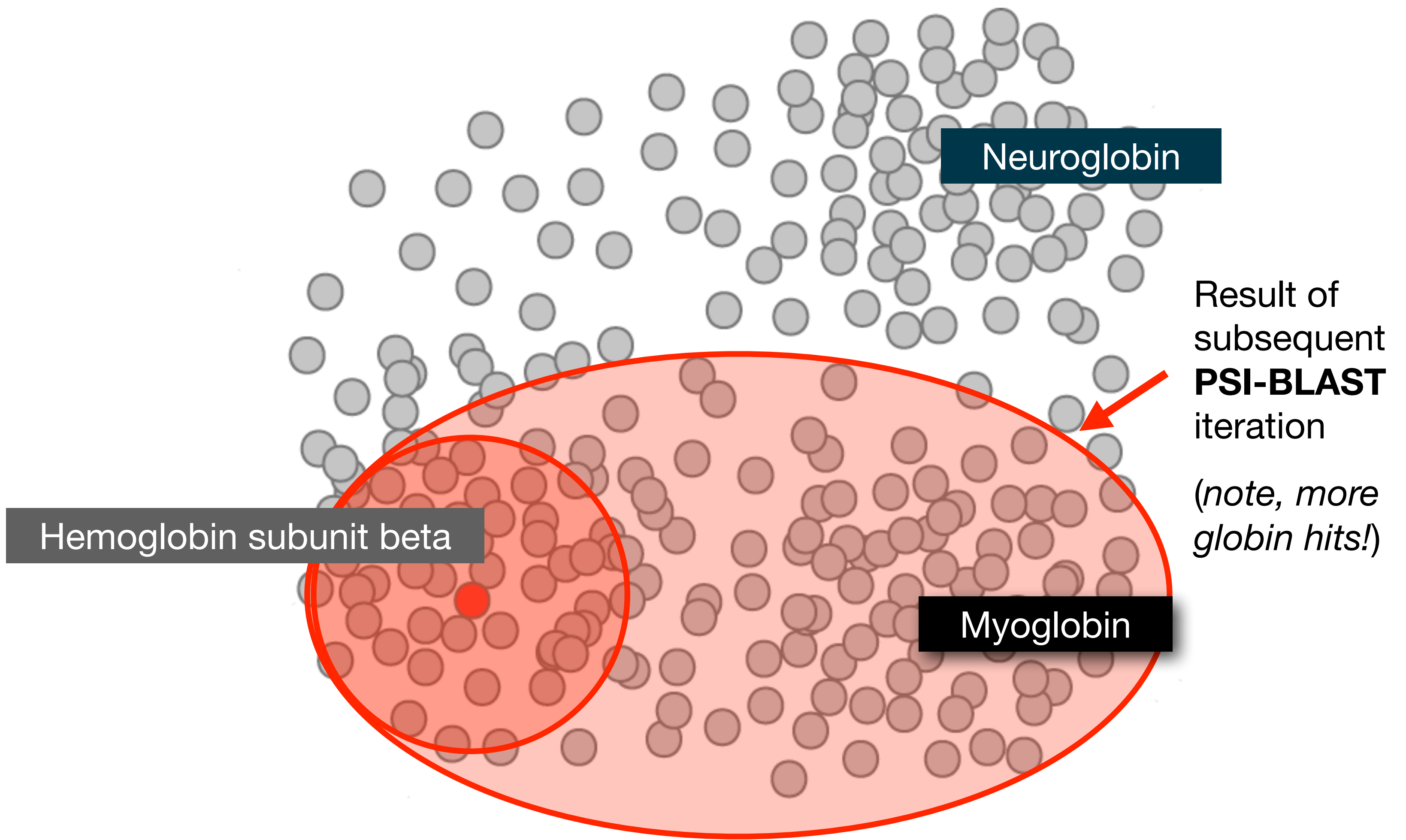


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)







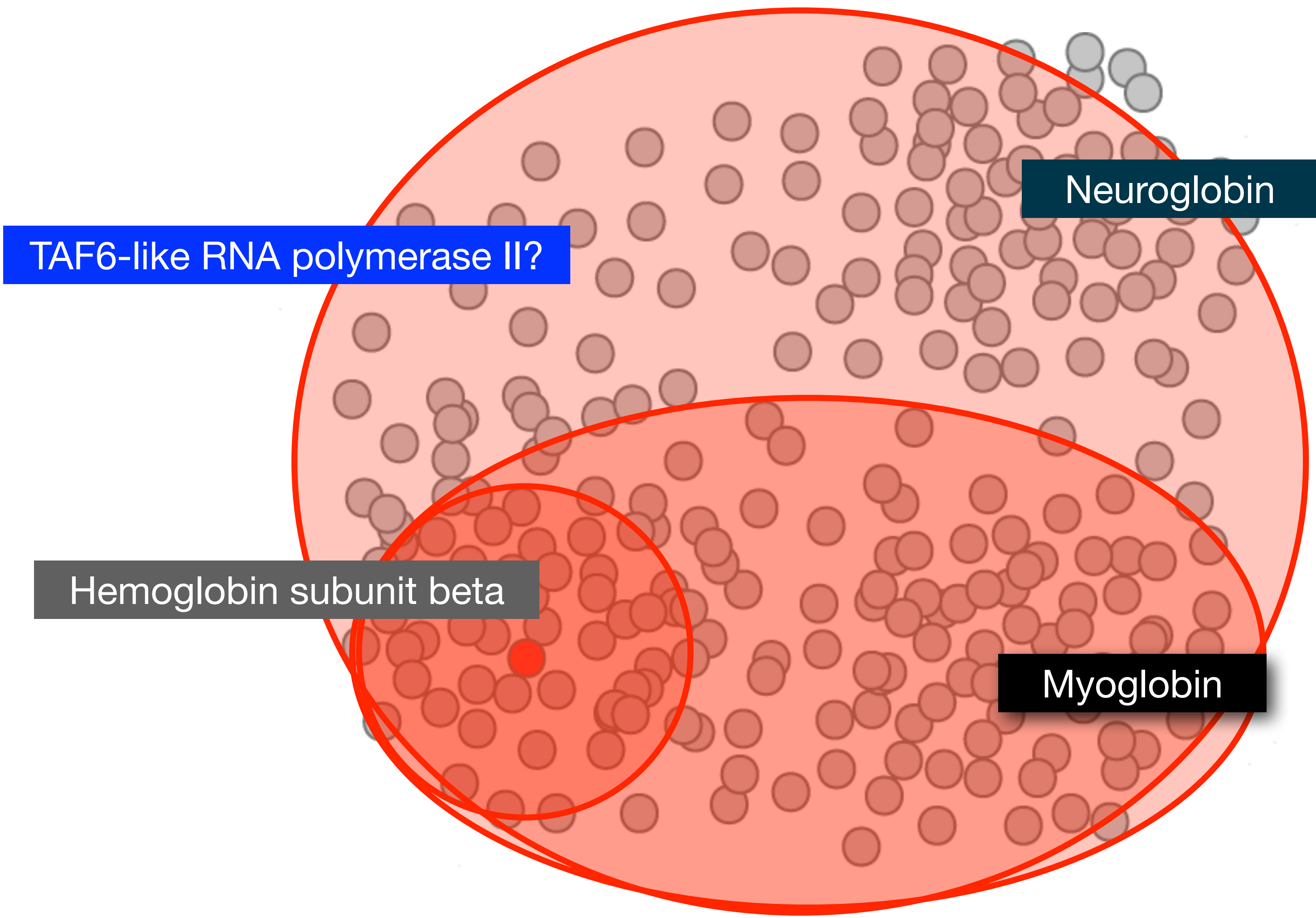


Neuroglobin

Hemoglobin subunit beta

Myoglobin

Result of subsequent **PSI-BLAST** iteration  
*(note, more globin hits!)*



TAF6-like RNA polymerase II?

Neuroglobin

Hemoglobin subunit beta

Myoglobin

Result of later  
**PSI-BLAST**  
iteration  
*(note, potential  
"corruption"!)*

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

Description	Max score	Total score	Query cover	E value	Ident	Accession
<a href="#">hemoglobin subunit beta [Homo sapiens]</a>	301	301	100%	2e-106	100%	<a href="#">NP_000509.1</a>
<a href="#">hemoglobin subunit delta [Homo sapiens]</a>	284	284	100%	7e-100	93%	<a href="#">NP_000510.1</a>
<a href="#">hemoglobin subunit epsilon [Homo sapiens]</a>	240	240	100%	2e-82	76%	<a href="#">NP_005321.1</a>
<a href="#">hemoglobin subunit gamma-2 [Homo sapiens]</a>	235	235	100%	2e-80	73%	<a href="#">NP_000175.1</a>
<a href="#">hemoglobin subunit gamma-1 [Homo sapiens]</a>	232	232	100%	3e-79	73%	<a href="#">NP_000550.2</a>
<a href="#">hemoglobin subunit alpha [Homo sapiens]</a>	114	114	97%	7e-33	43%	<a href="#">NP_000508.1</a>
<a href="#">hemoglobin subunit zeta [Homo sapiens]</a>	100	100	97%	3e-27	36%	<a href="#">NP_005323.1</a>
<a href="#">myoglobin [Homo sapiens]</a>	80.5	80.5	97%	2e-19	26%	<a href="#">NP_005359.1</a>
<a href="#">neuroglobin [Homo sapiens]</a>	54.7	54.7	92%	2e-09	23%	<a href="#">NP_067080.1</a>

1

2

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

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

1

2

3

?

**Inclusion of irrelevant hits can lead to PSSM corruption**

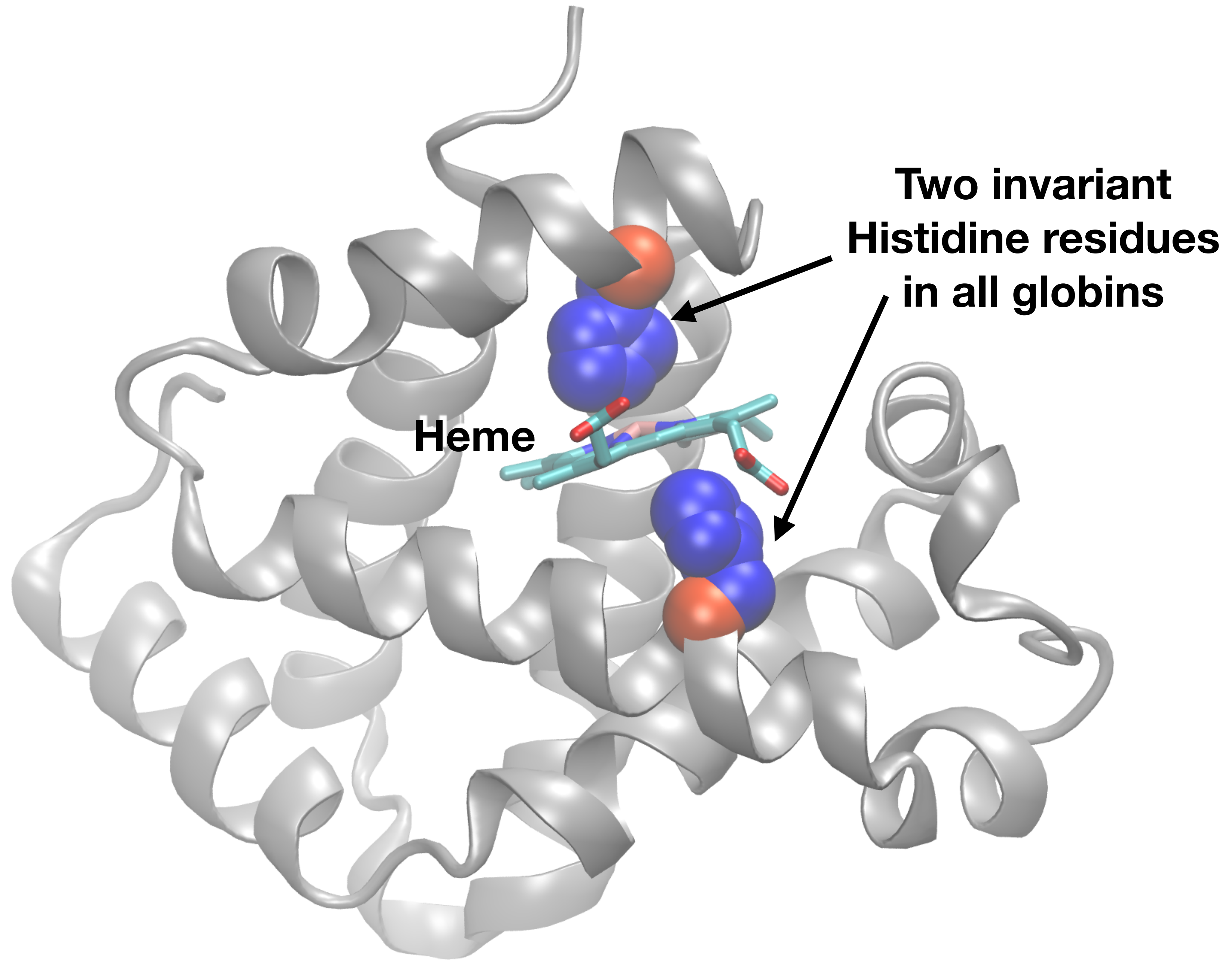
# YOUR TURN!

- 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!**

✓ <a href="#">Query_73613</a>	1	MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF	72
✓ <a href="#">NP_000510.1</a>	1	MVHLTPEEKTAVNALWGKV--NVDAVGGEALGRLLVVYPWTQRFFE-SFGDLSSPDAVM-GNPKVKAHGKKVLGAF	72
✓ <a href="#">NP_000175.1</a>	1	MGHFTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL	72
✓ <a href="#">NP_000509.1</a>	1	MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF	72
✓ <a href="#">NP_005321.1</a>	1	MVHFTAEEKAAVTSLWSKM--NVEEAGGEALGRLLVVYPWTQRFFD-SFGNLSSPSAIL-GNPKVKAHGKKVLTSL	72
✓ <a href="#">NP_000550.2</a>	1	MGHFTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL	72
✓ <a href="#">NP_005323.1</a>	1	-MSLTKTERTIIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFP-HF-----DLHpGSAQLRAHGSKVVA	67
✓ <a href="#">NP_000508.1</a>	1	-MVLSPADKTNVKAAWGKVGGAHAGEYGAEALERMFSLFPTTKTYFP-HF-----DLShGSAQVKGHGKKVADAL	67
✓ <a href="#">XP_005257062.1</a>	1	[ 15 ] SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL	89
✓ <a href="#">NP_001003938.1</a>	1	--MLSAQERAQIAQVWDLIAGHEAQFGAELLRLFTVYPSTKVYFP-HL-----SACQ-DATQLLSHGQRM	66
✓ <a href="#">NP_005322.1</a>	1	-MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFS-H-----LDLSpGSSQVRAHGQKVADAL	67
✓ <a href="#">NP_599030.1</a>	1	[ 15 ] SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL	89
✓ <a href="#">XP_016879605.1</a>	1	-----MEDPLEME-RSPQLRKHACRVMGAL	24
✓ <a href="#">NP_001349775.1</a>	1	-MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPEKLEKFD-KFKHLKSEDEMK-ASEDLKKGATVLTAL	73
✓ <a href="#">NP_067080.1</a>	1	---MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQyNCRQFSSPEDCL-SSPEFLDHIRKVMLVI	72
✓ <a href="#">NP_001369741.1</a>	1	-----MK-ASEDLKKGATVLTAL	18
✓ <a href="#">Query_73613</a>	73	SDGLAHLDNLKGT---FATLSELHCDKLHVDPENFRLGNVLCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147
✓ <a href="#">NP_000510.1</a>	73	SDGLAHLDNLKGT---FSQLSELHCDKLHVDPENFRLGNVLCVLAHNFNGKEFTPQMQAAYQKVVAGVANALAHKYH	147
✓ <a href="#">NP_000175.1</a>	73	GDAIKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQASWQKMTGVASALSSRYH	147
✓ <a href="#">NP_000509.1</a>	73	SDGLAHLDNLKGT---FATLSELHCDKLHVDPENFRLGNVLCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147
✓ <a href="#">NP_005321.1</a>	73	GDAIKNMDNLKPA---FAKLSELHCDKLHVDPENFKLLGNVMVIIILATHFGKEFTPEVQAAWQKLVSAVAIALAHKYH	147
✓ <a href="#">NP_000550.2</a>	73	GDAIKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQASWQKMTAVASALSSRYH	147
✓ <a href="#">NP_005323.1</a>	68	GDAVKSIDDIGGA---LSKLSELHAYILRVDPVNFKLLSHCLLVTLAARFPADFTAEEAHAWDKFLSVVSSVLTEKYR	142
✓ <a href="#">NP_000508.1</a>	68	TNAVAHVDDMPNA---LSALSDLHAHKLVRDPVNFKLLSHCLLVTLAHLPAEFTPAVHASLDKFLASVSTVLT	142
✓ <a href="#">XP_005257062.1</a>	90	NTVVENLHDPDKVssvLALVGKAHALKHKVPEVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK [ 35 ]	202
✓ <a href="#">NP_001003938.1</a>	67	GAAVQHVDNLRAA---LSPLADLHALVLRVDPANFPLLIQCFHVVLASHLQDEFTVQMQAAWDKFLTGVAVVLTEKYR	141
✓ <a href="#">NP_005322.1</a>	68	SLAVERLDDLPHA---LSALSHLHACQLRVDPASFQLLGHCLLVTLARHYPGDFSPALQASLDKFLSHVISALVSEYR	142
✓ <a href="#">NP_599030.1</a>	90	NTVVENLHDPDKVssvLALVGKAHALKHKVPEVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK [ 23 ]	190
✓ <a href="#">XP_016879605.1</a>	25	NTVVENLHDPDKVssvLALVGKAHALKHKVPEVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYK [ 35 ]	137
✓ <a href="#">NP_001349775.1</a>	74	GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK [ 6 ]	154
✓ <a href="#">NP_067080.1</a>	73	DAAVTNVEDLSSLeeyLASLGRKHRA-VGVKLSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWD [ 2 ]	151
✓ <a href="#">NP_001369741.1</a>	19	GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK [ 6 ]	99





# 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				

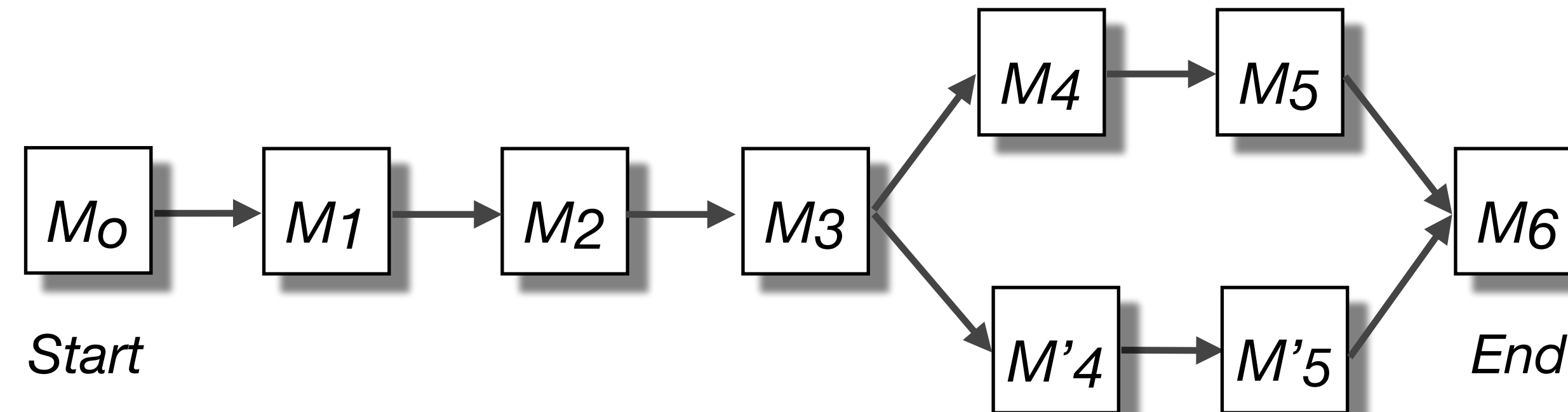
**Note:** We never see **QD** or **RH**, we only see **RD** and **QH**.  
However,  $P(RH)=0.24$ ,  $P(QD)=0.24$ , while  $P(QH)=0.16$

# 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**

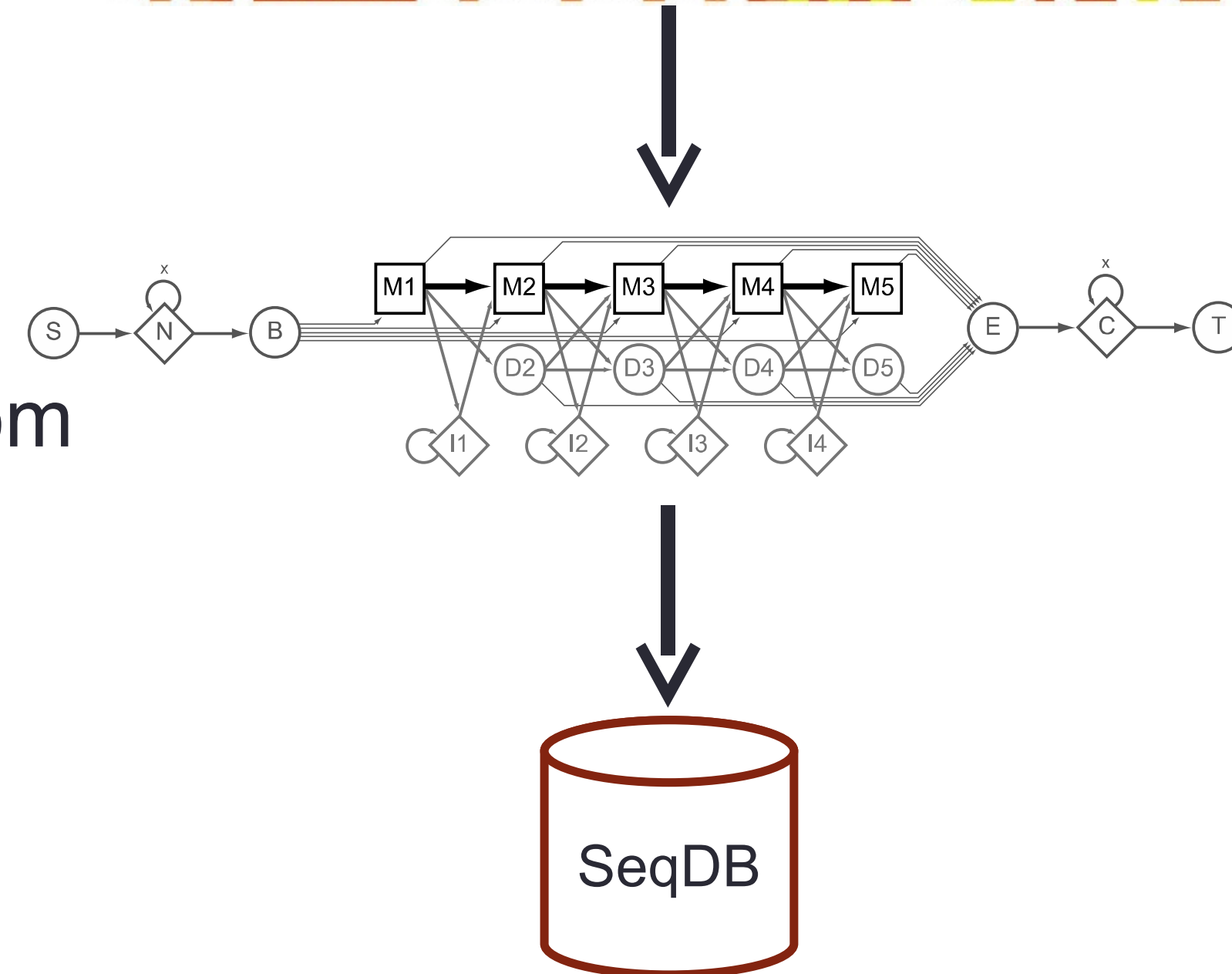


Recall that a PSSM for this motif would give the sequences **WEIRD** and **WEIRH** equally good scores even though the **RH** and **QR** combinations were not observed

# Use of HMMER

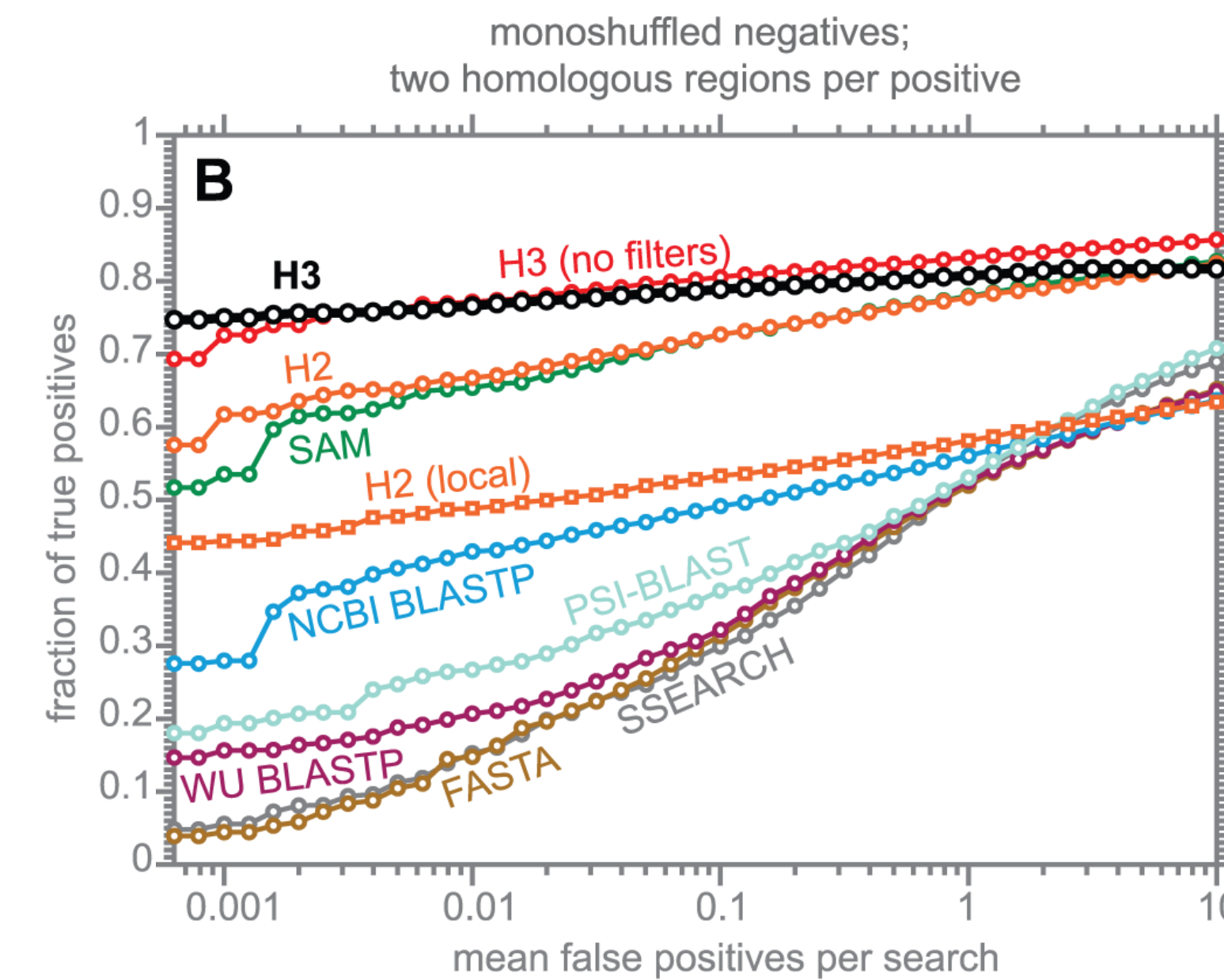
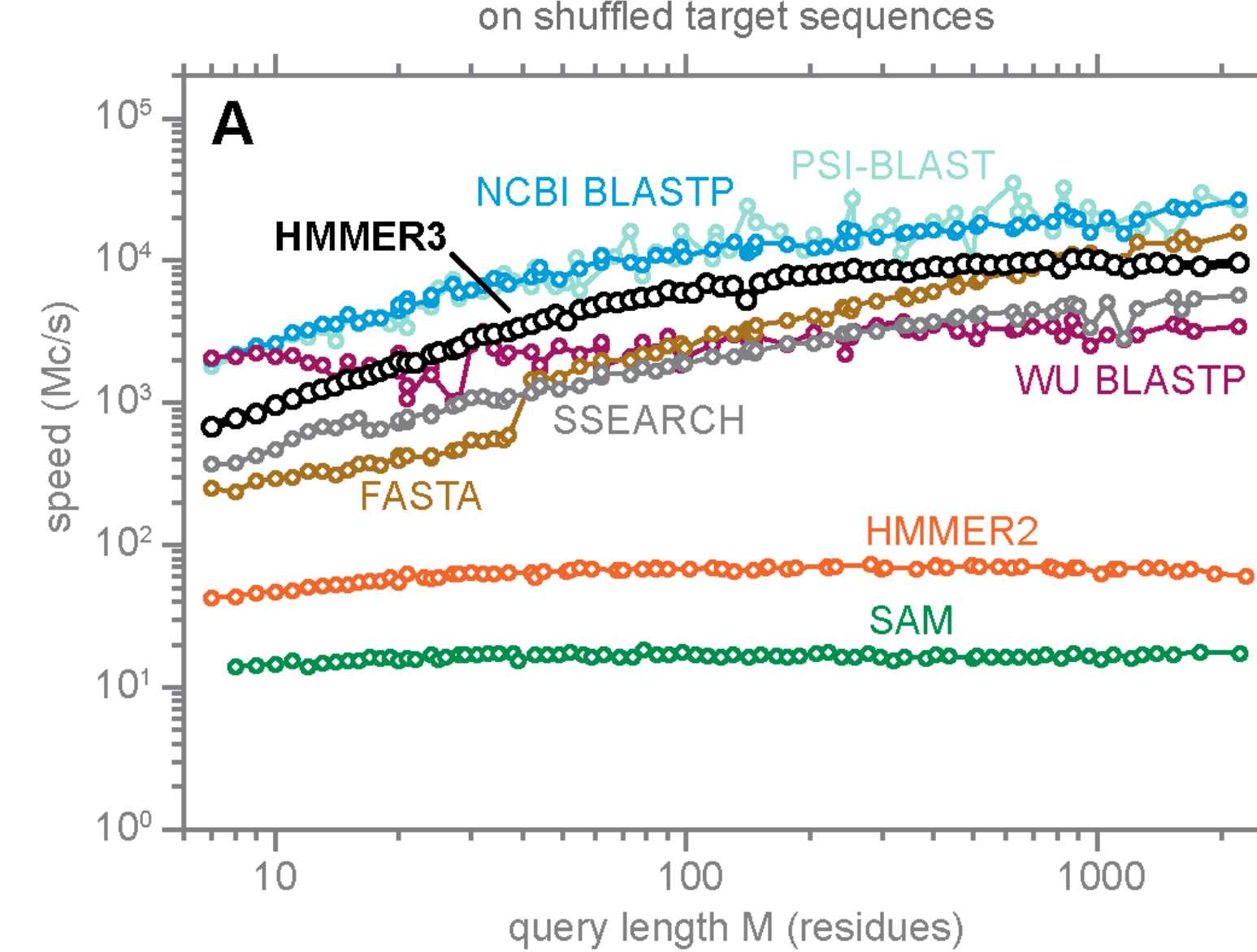
- Widely used by protein family databases
  - Use 'seed' alignments
- Until 2010
  - Computationally expensive
  - Restricted to HMMs constructed from multiple sequence alignments
- Command line application

```
K I I I T G E P G V G K T T L V K K I V E R L - - - G K R A I G F W T E E V R D P E T K K R T G F R I I T T E
K I L I T G R P G V G K T T L I K K L S R L L - - - - Q N A G G F Y T E E M R - - - E G E K R I G F K I I T L D
R F F V S G M P G V G K T T L A K R I A D E V R R E G F K V G G I I T E E I R - - - E G G K R T G F R V I A L D
R I F I T G M P G V G K T T L A L K I A E K L K E L G Y K V G G E I T K E I R - - - D G G K R V G F K I I T L D
R F F V S G M P G V G K T T L A K R I A D E I K R E G F K V G G I I T Q E I R - - - S G A R R S G F R V I A L D
H V F L T G P P G V G K T T L I Q K A I E V L Q S S G L P V D G F Y T Q E V R - - - Q E G K R I G F D V V T L S
H V F L T G V P G V G K T T L V K K V C D A L - - - S G L S V S G F Y T E E V R - - - E H G R R V G F D V V T V S
H V F L T G S P G V G K T T L I Q K A I T V L Q S S G L P V D G F Y T Q E V R - - - Q G G K R I G F D V V T L S
H V F L T G P P G V G K T T L I H K A S E V L K S S G V P V D G F Y T E E V R - - - Q G G R R I G F D V V T L S
```



# 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

**HMMER**  
biosequence analysis using profile hidden Markov models

HHMI  
janelia farm  
research campus

Home Search Results Software Help About

**Search**  
Upload a sequence, HMM or alignment and perform a HMMER search against one of seven sequence databases or the Pfam HMM database.  
**Start**  
hmmsearch jackhmmer hmmscan phmm

**Documentation**  
Download the documentation for the **command line** version of HMMER. (PDF, 392 KB)  
Read the online help for the HMMER **webserver** search service.

**Download HMMER**  
Get the latest version  
**v3.0**  
**Download (MacOSX/Intel)**  
Alternative Download Options

**hmmmer.janelia.org**





# HMMER

Biosequence analysis using profile hidden Markov Models

[Home](#)[Search](#)[Results](#)[Software](#)[Help](#)[About](#)[Contact](#)[phmmer](#)[hmmscan](#)[hmmsearch](#)[jackhmmmer](#)

## 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]  
MVHLTPEEKSAVTALWGKVNVDVEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG  
AFSDGLAHLNLDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN  
ALAHKYH
```

### Sequence Database

Frequently used databases:

Current database selection:

SwissProt










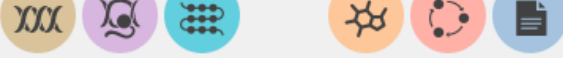



### Restrict by Taxonomy

 Taxon search Pre-defined representatives

Organism:

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

[Customise](#) [Customise](#)

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

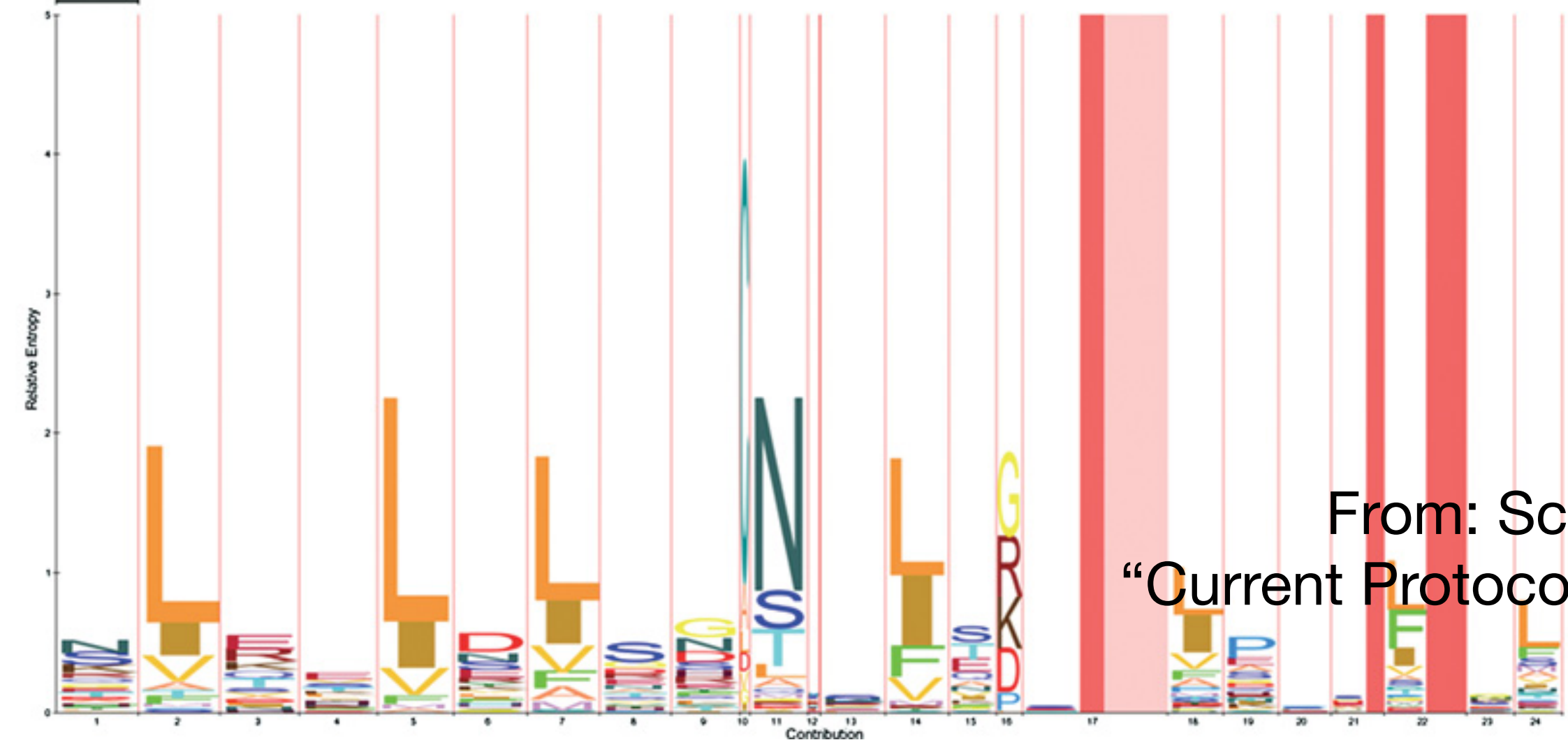
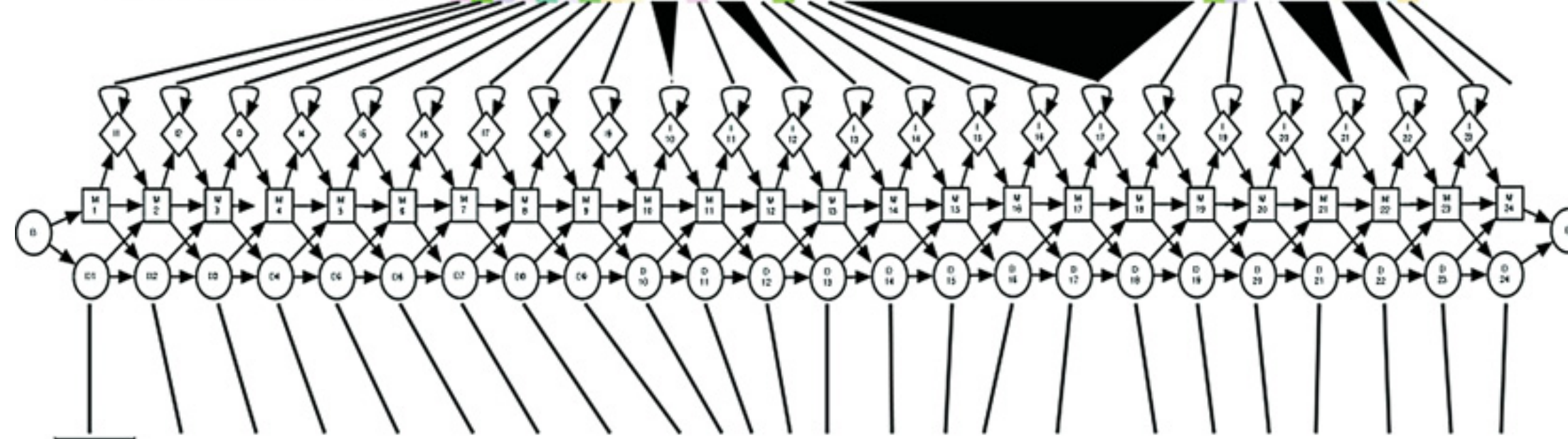
PFAM consists of two databases:

- **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**

```

Q9ARB2_LINUS/823-844 MLEYLDIGRA..P.RIV.H.....LDG...LENL
Q9M8N0_ARATH/320-341 RLTFLNLSFC..S.KLT.G.....LAF...FSII
FLJ_HUMAN/318-339 NLEEFMAAN..N..NLE.L.....VPES..LCRC
Q9VN74_DROME/90-112 ALHSLVIENC...TIV.H.....INDAA.FNQE
Q8L8I7_PNTA/792-814 NLQTIQMYRX..E.SLQ.V.....LPDS..FGNL
Q9FHL8_ARATH/301-324 NLWSLNLSR..N..LFSDP.....LPVVG.ARGF
SLIK6_MOUSE/65-87 RPFHLSLLN..N..GLT.M.....LHTND.FSGL
Q8NIJ8_EMEN/978-1000 TLTSLNIAS..A..KLV.Q.....FRDTL.FDSL
Q9LUQ2_ARATH/92-113 AMKSLDVSF..N..SISE.....LPEQ..IGSA
Q9FH93_ARATH/169-188 RLTSLNLDL..N..RFNGT.....LPS....LN
Q898G0_CLOTE/268-288 YLERINLDK..N..KIKN.....IEE...LEAN
Q8H6V2_MAIZE/678-699 NLRILSIVDC..V.SLQ.K.....LPP...SDSF
Q9AR40_LINUS/692-713 DLKVLINQ..T..EIT.T.....LKGE..VESL
Q9LE82_ARATH/350-377 HLTEIYMSY..L..NLEDEGT.....EALSEAL.LKSA
Q9H5N5_HUMAN/255-278 HLQVLDLHQC...SLT.AD.....DVMSL...TQVI
Q8L4C7_ARATH/185-207 KLEYLDIWG..S..NVT.N.....QGAVS..ILKF
Q9VSA4_DROME/1115-113E QLKALRLQC..N..AIGSH.....GLEAL..LCGQ
TLR1_MOUSE/376-398 RLKTLSLQK..N..QLKN.....LENII.LTSA
Q9TXJ6_LEMA/445-465 GLRDIDLSH..T..KVH.N.....IDA...LQAS
FXL13_MOUSE/409-448 KLIYLDLSGC..T.QVL.VEKCPRISSVVLIGSPHISDSA.FKAL
Q9TXJ6_LEMA/927-948 ALTVVNANSC..V.NLT.S.....IEA...LESA
Q9M4X9_CHLRE/1417-1444 LLAVLHLHD..NP.RLA.ADG.....VAGLAAA..LPGL
Q945S6_LYCPM/656-677 NLRHLDVSN..T..RRL.K.....MPLH..LSRL

```



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!**



# 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!