

BIMM 143
Hands-on Lab Session

Class 03

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm143>

Homework: DataCamp Signup!

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!

DataCamp

The screenshot shows the DataCamp web interface for an R exercise. The browser address bar displays `campus.datacamp.com/courses/free-introduction-to-r/chapter-1-intro-to-basics-1?ex=1`. The page header includes navigation links like Home, Gmail, Gcal, GitHub, and a course outline. The main content area is divided into three sections: instructions, a code editor, and an R console.

Instructions:

How it works

In the editor on the right you should type R code to solve the exercises. When you hit the 'Submit Answer' button, every line of code is interpreted and executed by R and you get a message whether or not your code was correct. The output of your R code is shown in the console in the lower right corner.

R makes use of the `#` sign to add comments, so that you and others can understand what the R code is about. Just like Twitter! Comments are not run as R code, so they will not influence your result. For example, *Calculate 3 + 4* in the editor on the right is a comment.

You can also execute R commands straight in the console. This is a good way to experiment with R code, as your submission is not checked for correctness.

Instructions 100 XP

- In the editor on the right there is already some sample code. Can you see which lines are actual R code and which are comments?
- Add a line of code that calculates the sum of 6 and 12, and hit the 'Submit Answer' button.

Take Hint (-30 XP)

Code Editor: The editor shows a file named `script.R` with the following content:

```
1 # Calculate 3 + 4
2 3 + 4
3
4 # Calculate 6 + 12
5
```

Buttons for `Run Code` and `Submit Answer` are located at the bottom right of the editor.

R Console: The console is currently empty, showing a prompt `> |`.

DataCamp

campus.datacamp.com/courses/visualization-best-practices-in-r/comparing-distributions

Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News +

Schedule · BIMM 143

A simple boxplot | R

DataCamp

Exercise

A simple boxplot

Let's get started by simply making a box plot similar to the ones we made in the slides.

Modify the `filter()` to look at 'RED' cars instead of blue and then map the x-axis to `gender` and the y-axis to driver `speed`. Add your boxplot geometry and give it a title.

Instructions

100 XP

- Modify `filter()` command to filter to 'RED' vehicles.
- Fill in `aes()` by mapping `x` to `gender` and `y` to `speed`.
- Add a boxplot geometry to the plot.
- Title plot 'Speed of red cars by gender of driver'.

Take Hint (-30 XP)

script.R

```
1 md_speeding %>%
2   filter(vehicle_color == 'BLUE') %>%
3   # Map x and y to gender and speed
   columns respectively
4   ggplot(aes(x=gender, y=speed)) +
5     # add a boxplot geometry
6     geom_boxplot() +
7     # give plot supplied title
8     labs(title = 'My fun subtitle')
```

Run Code

Submit Answer

Plots

My fun subtitle

Previous Plot

1/1

Next Plot

R Console

Slides

```
> md_speeding %>%
  filter(vehicle_color == 'BLUE') %>%
  # Map x and y to gender and speed columns respectively
  ggplot(aes(x=gender, y=speed)) +
  # add a boxplot geometry
  geom_boxplot() +
  # give plot supplied title
  labs(title = 'My fun subtitle')
> |
```

Homework: DataCamp Signup!

Install **R** and **RStudio** on YOUR laptop (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!

Class 3: Hands-on section

<http://thegrantlab.org/bimm143/>

Week2

The screenshot shows the course website for BIMM 143 at UC San Diego. The left sidebar contains navigation links: Overview, Schedule (highlighted with a red arrow), Computer Setup, and Learning Goals. The main content area displays a table of course topics. The table has three columns: a sequence number, a date, and a description. The third row, which includes the 'Project: Find a gene project assignment' and an 'Optional: Advanced sequence alignment and database searching' section, is highlighted with a red border and a red arrow points to it from the right. The fourth row describes 'Bioinformatics data analysis with R'. The table also includes a navigation bar at the top with links to Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, GDrive, Atmosphere, CloudLaunch, BIMM194, Blink, and News.

2	09/30/21	Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations.
3	Tue 10/05/21	Project: Find a gene project assignment (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.
*	Tue 10/05/21	Optional: Advanced sequence alignment and database searching Detecting remote sequence similarity, Database searching beyond BLAST, Substitution matrices, Using PSI-BLAST, Profiles and HMMs, Protein structure comparisons as a gold standard.
4	Thu 10/07/21	Bioinformatics data analysis with R 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.
		Data exploration and visualization in R

BGGN 213
Hands-on Lab Session

Class 03

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

Class 3: Hands-on section

<http://thegrantlab.org/bgggn213/>

Week2

The screenshot shows the BGGN 213 course website. On the left, a sidebar contains the UC San Diego logo, the course title 'BGGN 213', a description, and a navigation menu. The 'Schedule' link in the menu is highlighted with a red box and a red arrow points to it from the left. The main content area displays a table of course topics. The row for 'Project: Find a gene project assignment' (dated 10/06/21) and the row for 'Optional: Advanced sequence alignment and database searching' (dated 10/06/21) are both enclosed in a red rectangular box. A red arrow points from the right side of the image towards this box. The table also includes a row for 'Bioinformatics data analysis with R' (dated 10/08/21).

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		Project: Find a gene project assignment (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		Optional: Advanced sequence alignment and database searching 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		Bioinformatics data analysis with R 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.

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 Monday of week 4 (**Oct 20th**, 10/20/25).
- ➔ The complete assignment, including responses to **all questions**, is due 12pm Monday of week 10 (**Dec 8th**, 12/08/25).

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 Monday of week 4 (**Oct 13th**, 10/13/25).
- ➔ The complete assignment, including responses to **all questions**, is due 12pm Monday of week 10 (**Dec 8th**, 12/08/25).

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

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

HomeGmailGcalGitHubBIMM143BGGN213GDriveAtmosphereCloudLaunchBIMM194BlinkNews+▼

3: (Project) Find a Gene Assignment Part 1

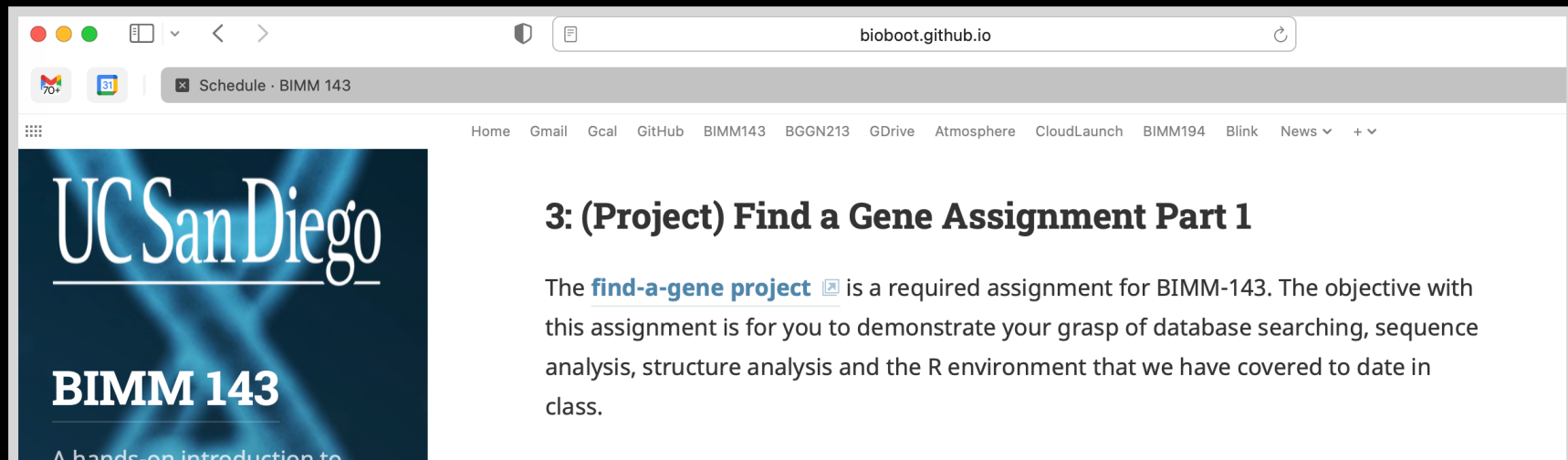
The [find-a-gene project](#)  is a required assignment for BIMM-143. The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.

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 Monday of week 4 (**Oct 20th**, 10/20/25).
- The complete assignment, including responses to **all questions**, is due 12pm Monday of week 10 (**Dec 8th**, 12/08/25).

https://tritongpt.ucsd.edu/

 **TritonGPT**
Powered by Onyx

 New Chat

 My Documents

Assistants

 UC San Diego Assist...

 General AI Assistant

 Fund Manager Coach

Explore Assistants

Chats

▼ Today

BIMM143 Find a Gene

UCSD Project Management

▼ Previous 7 Days

Local vs Global Alignment

Assignment Due Dates Inqui

Lab 2 Sequence Alignment

▼ Over 30 Days

Processing HTML Capability

BIMM143 Molecular Biology

 **UC San Diego Assistant**

UC San Diego-specific guidance or content generation.

As a staff member, what are the most effective,...

As a student, what are the most effective,...

As a faculty member, what are the most effective...

Draft an email to a prospective student highlighti...



Message UC San Diego Assistant...

 File  gpt-oss ▼

☐ Agent 

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

https://tritongpt.ucsd.edu/

General model
for UCSD



 **TritonGPT**
Powered by Onyx

[New Chat](#)

[My Documents](#)

Assistants

[UC San Diego Assist...](#)

[General AI Assistant](#)

[Fund Manager Coach](#)

[Explore Assistants](#)

Chats

▼ Today

BIMM143 Find a Gene

UCSD Project Management

▼ Previous 7 Days

Local vs Global Alignment

Assignment Due Dates Inqui

Lab 2 Sequence Alignment

▼ Over 30 Days

Processing HTML Capability

BIMM143 Molecular Biology

 **UC San Diego Assistant**

UC San Diego-specific guidance or content generation.

As a staff member, what are the most effective,...

As a student, what are the most effective,...

As a faculty member, what are the most effective...

Draft an email to a prospective student highlighti...

Message UC San Diego Assistant...

 File  gpt-oss ▼

☐ Agent 

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

https://tritongpt.ucsd.edu/

Click **Explore Assistants** to find our BIMM143 specific model...

The screenshot shows the TritonGPT web interface. On the left sidebar, under the 'Assistants' section, the 'Explore Assistants' button is highlighted with a red rectangle. The main content area displays the 'UC San Diego Assistant' interface, which includes a header with the UCSD logo and the text 'UC San Diego Assistant'. Below this, a subtitle reads 'UC San Diego-specific guidance or content generation.' There are four example prompts in boxes: 'As a staff member, what are the most effective,...', 'As a student, what are the most effective,...', 'As a faculty member, what are the most effective...', and 'Draft an email to a prospective student highlighti...'. At the bottom, there is a message input field with the placeholder 'Message UC San Diego Assistant...' and a dropdown menu showing 'File' and 'gpt-oss'. A toggle switch for 'Agent' is also visible. At the very bottom, a disclaimer states: 'TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)'.

https://tritongpt.ucsd.edu/

Click **Explore Assistants** to find our BIMM143 specific model...

The screenshot shows the TritonGPT interface with a sidebar on the left and a main content area. The sidebar includes a 'New Chat' button, 'My Documents', and a list of assistants: 'UC San Diego Assistant', 'General AI Assistant', and 'Fund Manager Coach'. Below this is a 'Chats' section with a 'Today' filter and a list of chat titles. The main content area displays a modal titled 'Explore Assistants' with a search bar and filters for 'Pinned', 'Mine', 'Private', and 'Public'. The modal lists six featured assistants in a grid:

- UC San Diego Assistant**: UC San Diego-specific guidance or content generation. 1 Action • Public. Buttons: Start Chat, Unpin.
- General AI Assistant**: AI support for tasks outside UC San Diego — summarizing, paraphrasing, and more. No Actions • Public. Buttons: Start Chat, Unpin.
- Fund Manager Coach**: I provide guidance on budgeting, grant guidelines, financial transactions, payroll, and compliance fo... 1 Action • Public. Buttons: Start Chat, Unpin.
- Email Phishing Analyzer**: Paste the email's sender, subject, and content. This assistant will assess the risk, assign a rating, and... 1 Action • Public. Buttons: Start Chat, Pin.
- Job Description Helper**: Helps you write a Position Overview for jobs in Career Tracks library.
- Expert Notetaker**: Paste a meeting transcript and get a detailed summary with action items.

At the bottom of the modal, a disclaimer states: 'TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)'

https://tritongpt.ucsd.edu/

Filter or scroll to
find “BIMM143
Grant - Student
Assistant”

The screenshot shows the TritonGPT interface with a search modal open. The search bar at the top of the modal contains the text 'BIMM'. Below the search bar, there are filter buttons: 'Pinned', 'Mine', 'Private', and 'Public'. The modal is divided into two main sections: 'Featured Assistants' and 'All Assistants'. The 'Featured Assistants' section displays the message 'No featured assistants match filters'. The 'All Assistants' section lists two assistants: 'BIMM143 Grant - Student Assistant' and 'BIMM143 Grant - Instructor Assistant'. The 'BIMM143 Grant - Student Assistant' entry is highlighted with a red box. It includes a checkmark icon, the name, a description 'An instructional pilot assistant', the creator 'By maholla', a button 'Add to your pinned list', a 'Private' status indicator, and two buttons: 'Start Chat' and 'Pin'. The 'BIMM143 Grant - Instructor Assistant' entry also includes a checkmark icon, the name, a description 'An instructional pilot assistant', the creator 'By maholland@ucsd.edu', '1 Action', a 'Private' status indicator, and two buttons: 'Start Chat' and 'Pin'. The background of the interface shows a sidebar with 'New Chat' and 'My Documents' options, and a list of assistants including 'UC San Diego Assistant', 'General AI Assistant', and 'Fund Manager Coassistant'. The bottom of the interface shows a chat input area with the placeholder text 'Message UC San Diego Assistant...' and a dropdown menu for 'gpt-oss'.

TritonGPT
Powered by Onyx

New Chat

My Documents

Assistants

UC San Diego Assistant

General AI Assistant

Fund Manager Coassistant

Explore Assistants

Chats

Today

BIMM143 Find a Gene

UCSD Project Management

Previous 7 Days

Local vs Global Alignment

Assignment Due Dates In

Lab 2 Sequence Alignment

Over 30 Days

Processing HTML Capability

BIMM143 Molecular Biology

BIMM

Pinned Mine Private Public

Featured Assistants

No featured assistants match filters

All Assistants

BIMM143 Grant - Student Assistant
An instructional pilot assistant
By maholla **Add to your pinned list** Private
[Start Chat](#) [Pin](#)

BIMM143 Grant - Instructor Assistant
An instructional pilot assistant
By maholland@ucsd.edu • 1 Action • Private
[Start Chat](#) [Pin](#)

Message UC San Diego Assistant...

File gpt-oss Agent

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

https://tritongpt.ucsd.edu/

Pin and drag to
top of list...

The screenshot shows the TritonGPT web interface. On the left is a sidebar with the following sections:

- TritonGPT** (Powered by Onyx)
- New Chat**
- My Documents**
- Assistants**
 - BIMM143 Grant - Stu...** (highlighted)
 - BIMM143 Grant - Ins...
 - UC San Diego Assist...
 - General AI Assistant
 - Fund Manager Coach
 - Explore Assistants
- Chats**
 - Today**
 - BIMM143 Find a Gene
 - UCSD Project Management
 - Previous 7 Days**
 - Local vs Global Alignment
 - Assignment Due Dates Inqui
 - Lab 2 Sequence Alignment
 - Over 30 Days**
 - Processing HTML Capability
 - BIMM143 Molecular Biology

The main chat area displays:

- BIMM143 Grant - Student Assistant**
- An instructional pilot assistant**
- A message input field with the placeholder text "Message BIMM143 Grant - Student Assistant..."
- Buttons for "File" and "GPT 4.1" (with a dropdown arrow).
- A toggle switch for "Agent" (currently off) and an upward arrow button.

At the bottom, a disclaimer states: "TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)"

https://tritongpt.ucsd.edu/

Ask away...



Powered by Onyx

New Chat

My Documents

Assistants

BIMM143 Grant - Stu...

BIMM143 Grant - Ins...

UC San Diego Assist...

General AI Assistant

Fund Manager Coach

Explore Assistants

Chats

▼ Today

BIMM143 Find a Gene

UCSD Project Management

▼ Previous 7 Days

Local vs Global Alignment

Assignment Due Dates Inqui

Lab 2 Sequence Alignment

▼ Over 30 Days

Processing HTML Capability

BIMM143 Molecular Biology

what is the project about

Q Searched for: *BIMM143 project description*

layout: projects title: Screen Cast Vide
os show_collection: projects descript...

14 Thu 05/15/25 RNA-Seq mini project
Differential expression analysis proje...

Show All

The main project in BIMM143 is the "Find a Gene" assignment. Its purpose is for you to demonstrate your understanding of database searching, sequence analysis, structure analysis, and exploratory data analysis with R, all covered in the course. You'll use bioinformatics tools to analyze a gene, interpret sequence and structure data, and apply R for data analysis. The project is submitted as a PDF report and counts for 35% of your course grade. There are specific questions to answer, with deadlines for parts and the final report ②, ③.

Do you want to know more about the steps involved, the grading, or the scientific focus?

tell me about how it is graded

Q Searched for: *BIMM143 Find a Gene project grading rubric*

layout: page title: Assignments & Gradi
ng menu: true order: 6 NOTE: Please ...

how effective it w ↓ d tell me you kn
ow do you want to go to zoom instea...

Message BIMM143 Grant - Student Assistant...

File GPT 4.1

Agent

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

Class 3: Hands-on section

<http://thegrantlab.org/bimm143/>

Class 03

The screenshot shows the BIMM 143 course website. On the left, a sidebar contains the UC San Diego logo, the course title 'BIMM 143', a description, and a navigation menu with 'Overview', 'Schedule', 'Computer Setup', and 'Learning Goals'. The 'Schedule' link is highlighted with a red box and an arrow. The main content area displays a table of course sessions. The third session, on Tuesday 10/05/21, is titled 'Optional: Advanced sequence alignment and database searching' and is highlighted with a red box and an arrow. The table also lists sessions 2, 3, and 4.

Session	Date	Topic
2	09/30/21	Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations.
3	Tue 10/05/21	Project: Find a gene project assignment (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.
*	Tue 10/05/21	Optional: Advanced sequence alignment and database searching Detecting remote sequence similarity, Database searching beyond BLAST, Substitution matrices, Using PSI-BLAST, Profiles and HMMs, Protein structure comparisons as a gold standard.
4	Thu 10/07/21	Bioinformatics data analysis with R 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.

Data exploration and visualization in R

► Details:

Sequence 1

Sequence 2

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				
T	-8	-5	-2	-1	-2				
A	-10	-7	-4	-3	-2	-1			
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](#).

[NW App Link](#)

Key Question:

Q. Where do our alignment match and mis-match scores typically come from?

Protein BLAST (BLASTp)

▼ Algorithm parameters

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

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of sequences with no more than 62% identity

By default BLASTp match scores come from the BLOSUM62 matrix

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of sequences with no more than 62% identity

	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

Not all matches score equally
(blue highlighted values)

By default BLASTp match scores come from the BLOSUM62 matrix

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of sequences with no more than 62% identity

Not all **matches** (blue highlighted)

	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

Not all matches score equally
(blue highlighted values)

C : C scores +9

V : V scores +4

By default BLASTp match scores come from the BLOSUM62 matrix

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

Note. Some amino acid mismatches have positive scores (highlighted in red) reflecting the shared physicochemical properties of these amino acids

Not all matches score equally
(blue highlighted values)

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of protein sequences with no more than 62% identity.

Note. Some amino acid matches have positive scores (highlighted in red), reflecting the shared physicochemical properties of these amino acids.

Not all matches are highlighted (blue highlights).

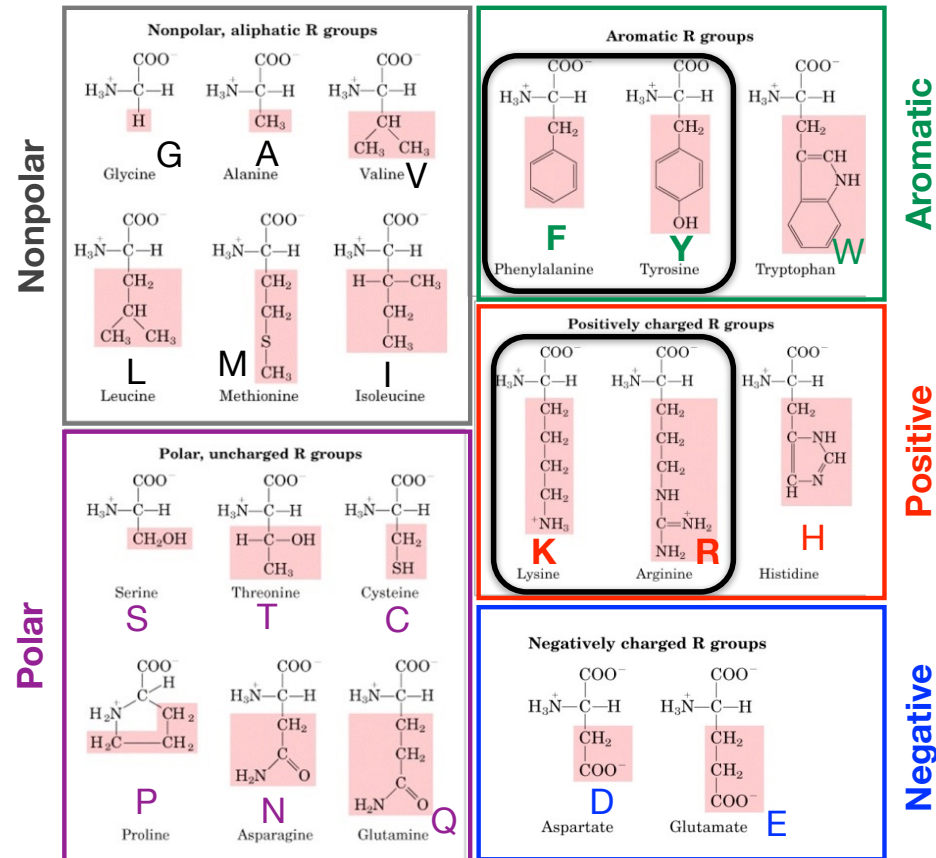
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5							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

C : C scores +9

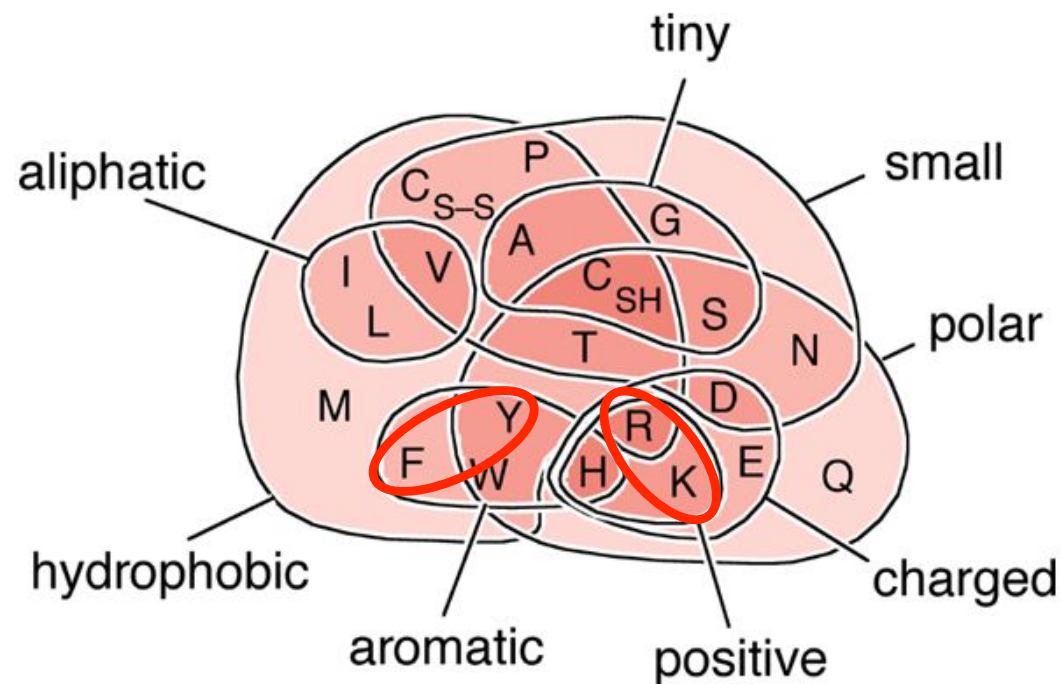
V : V scores +4

F : Y scores +3

Protein scoring matrices reflect the properties of amino acids



Protein scoring matrices reflect the properties of amino acids



Key Trend: High scores for amino acids in the same “biochemical group” and low scores for amino acids from different groups.

N.B. 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).

We will revisit this later...

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]

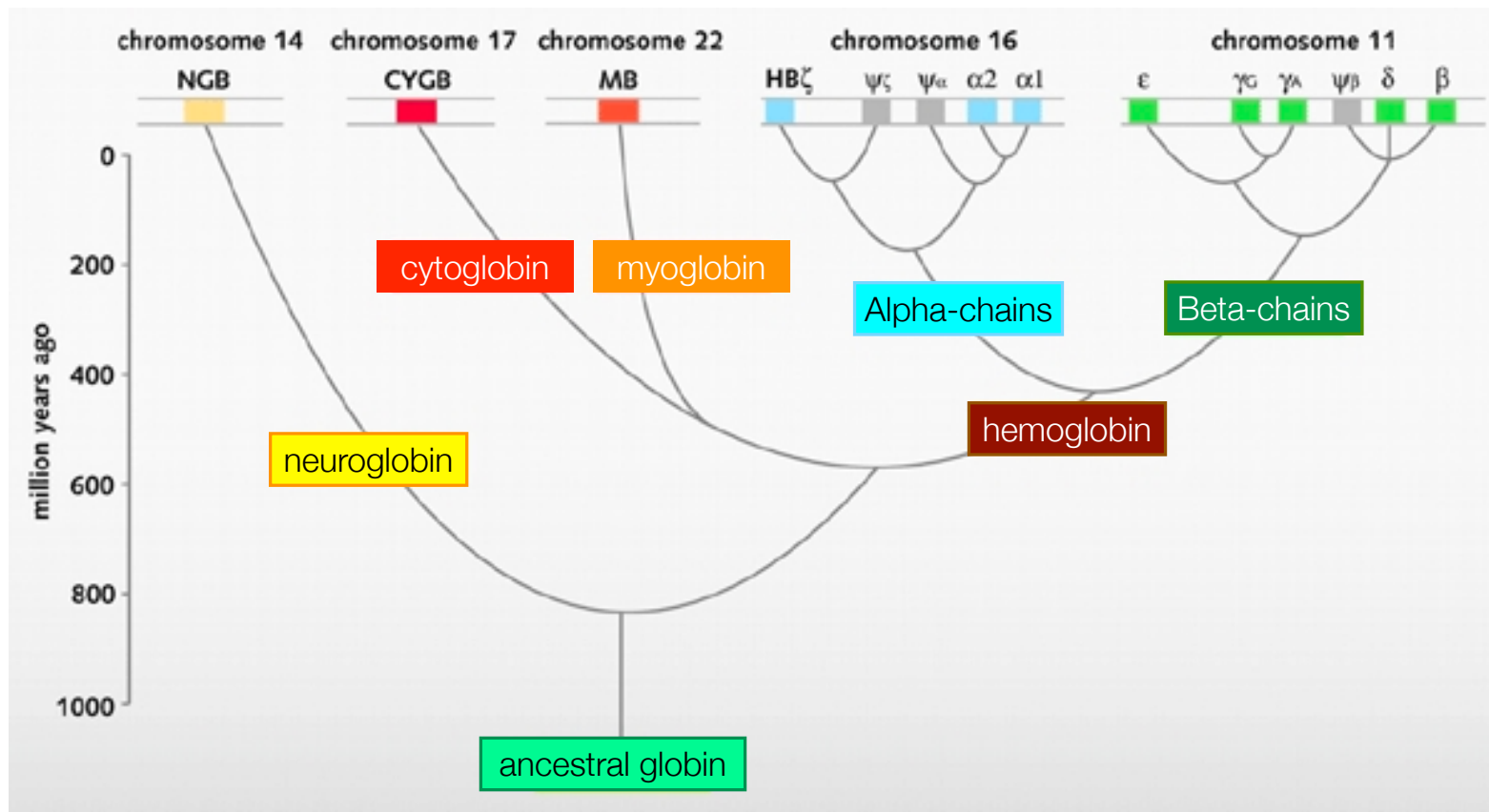
- ▶ Please do answer the last review question (**Q20**).
- ▶ We encourage discussion at your **Table** and on **Piazza**!

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**!



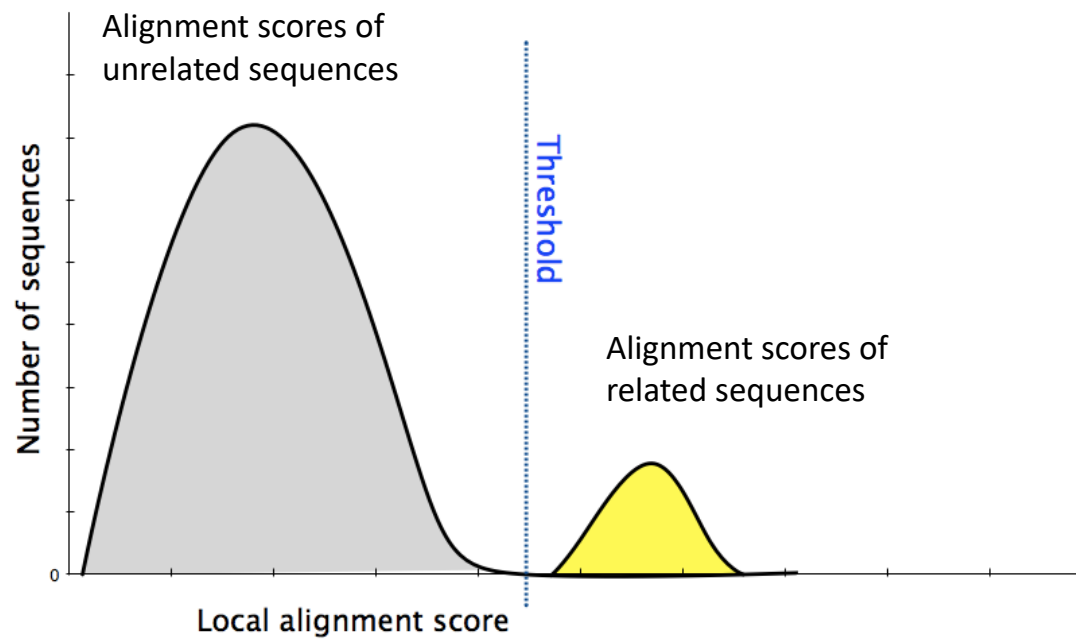
An evolutionary model of human globins.

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).

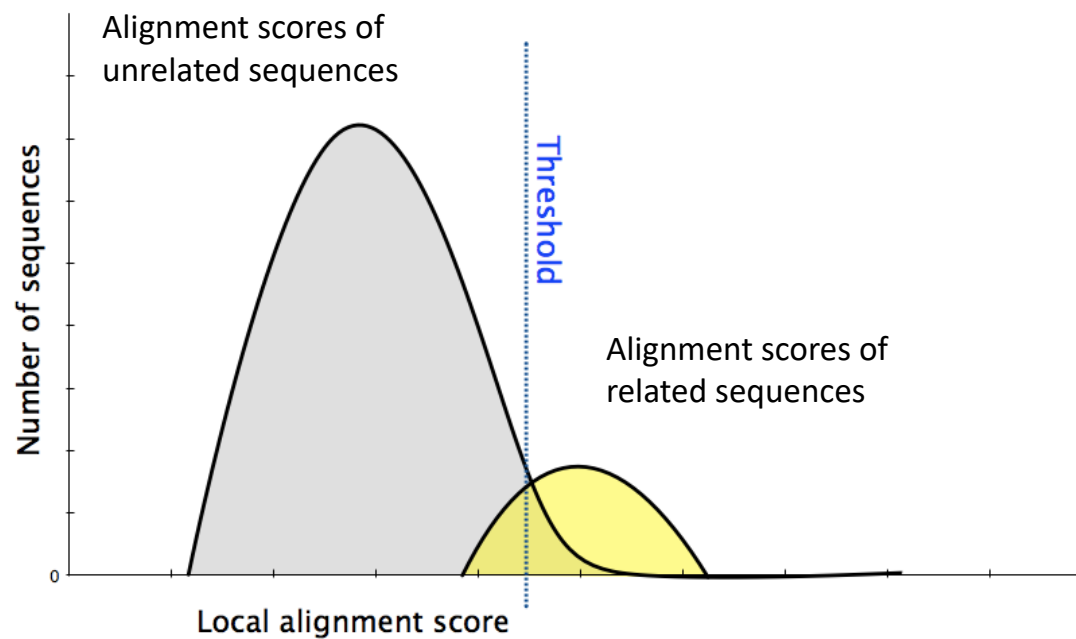
Question:

Q. Can we find and align these homologous globins using SW approaches such as BLAST?

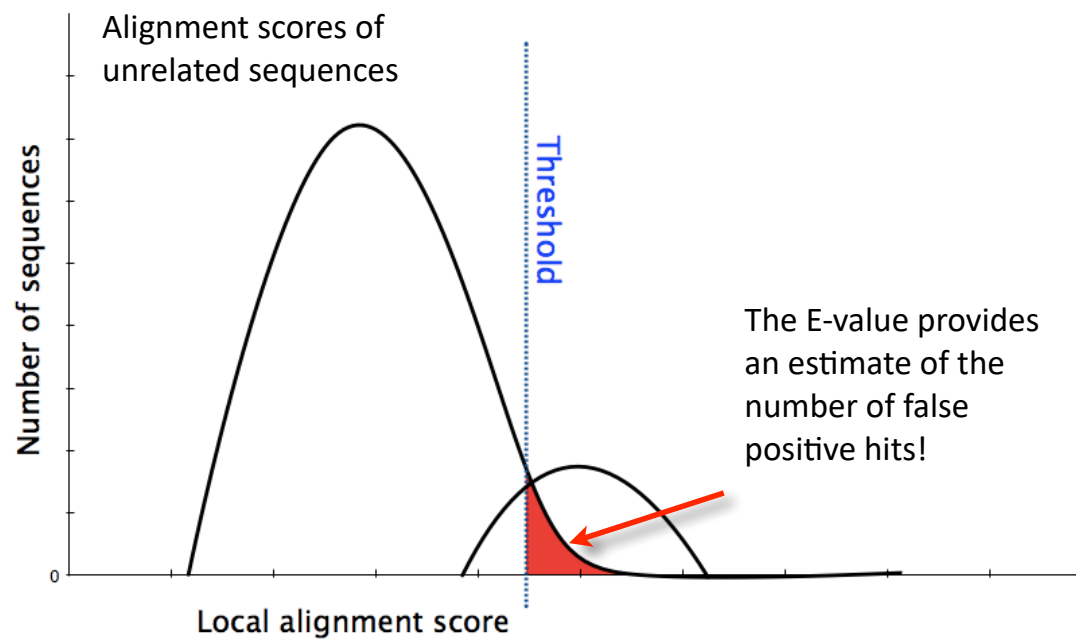
- Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



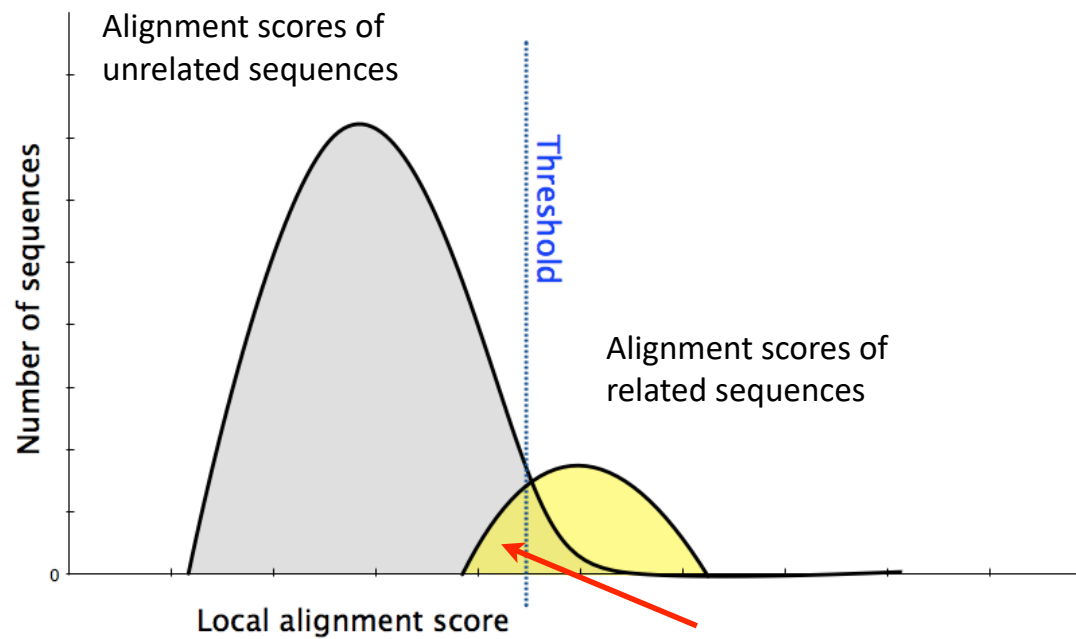
- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



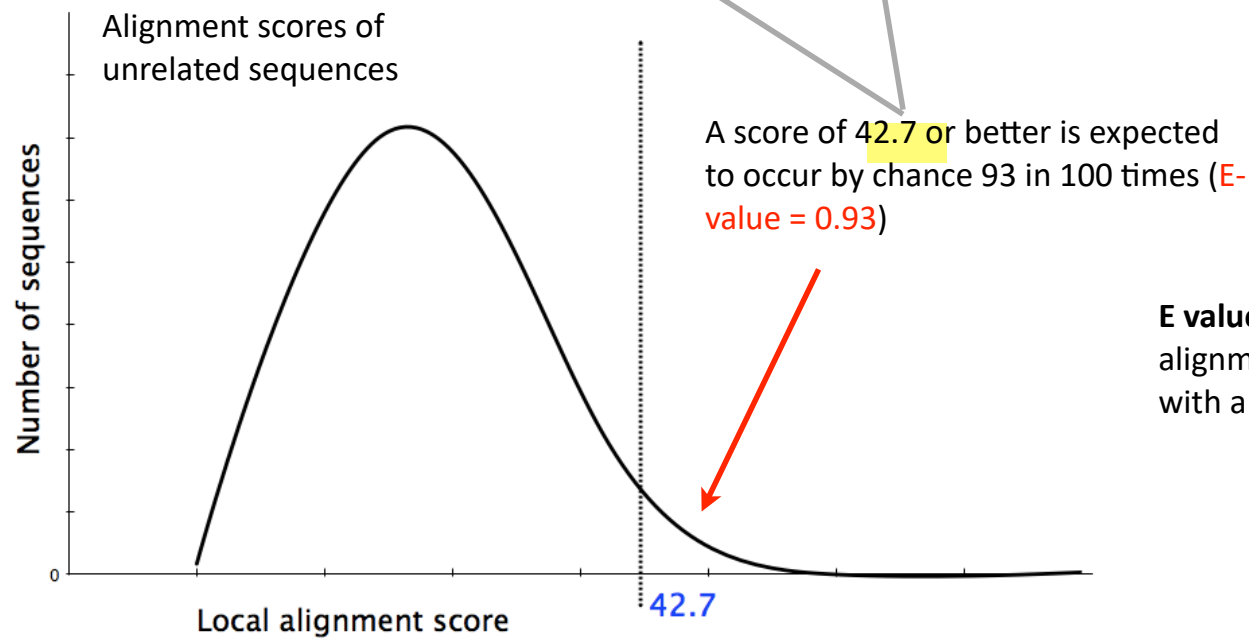
- Unfortunately, often both score distributions overlap
 - The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- 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: The number of alignments expected by chance with a particular score

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).

Protein BLAST (BLASTp)

▼ Algorithm parameters

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

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of aligned sequences with no more than 62% identity.

Blocks Substitution Matrix. Scores obtained from observed frequencies of substitutions in blocks of 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 score +4 regardless of the context in the molecule.

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

PSI-BLAST: Position specific iterated BLAST

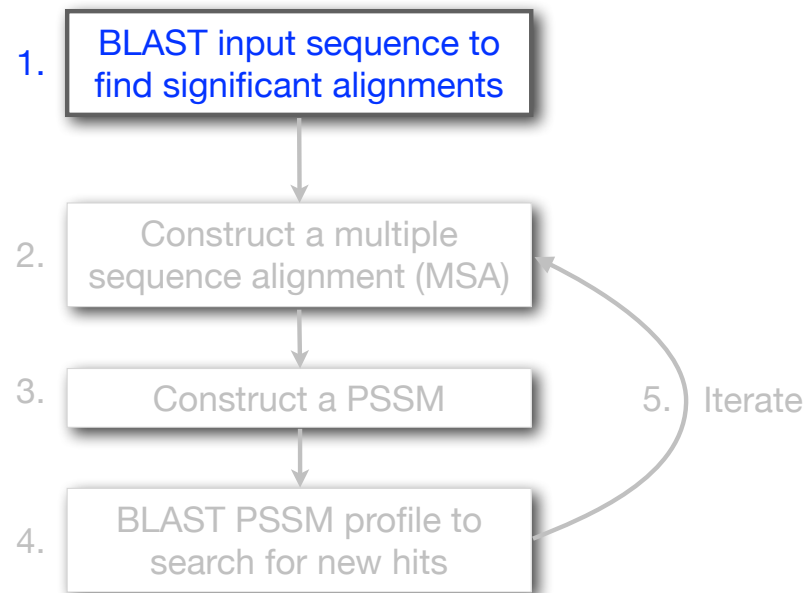
- The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query

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

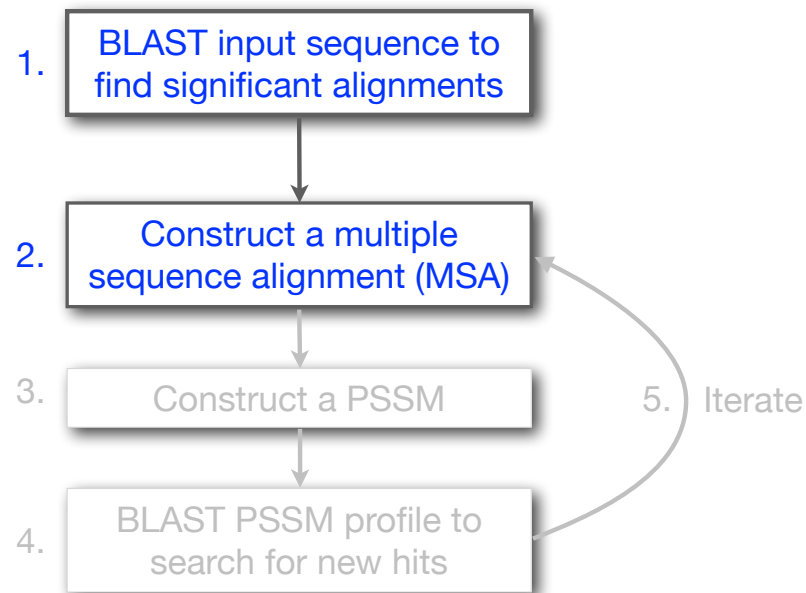
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

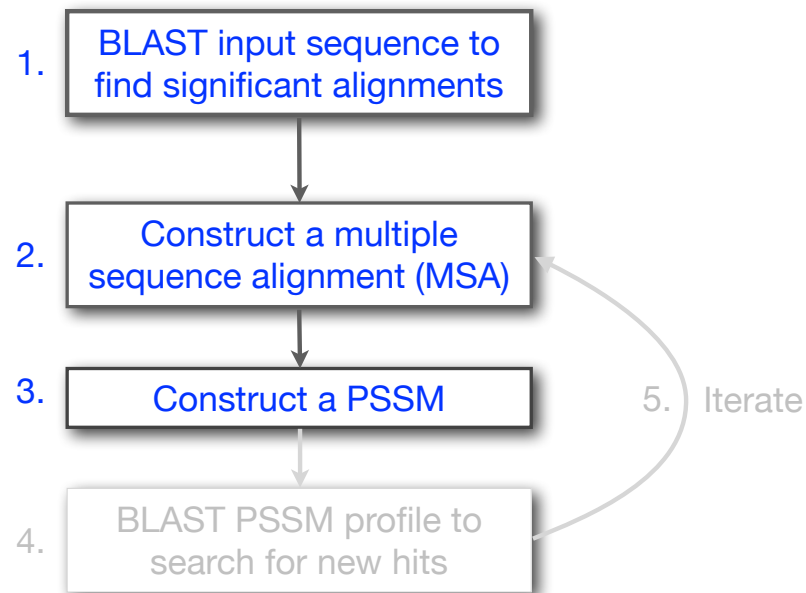
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



(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)

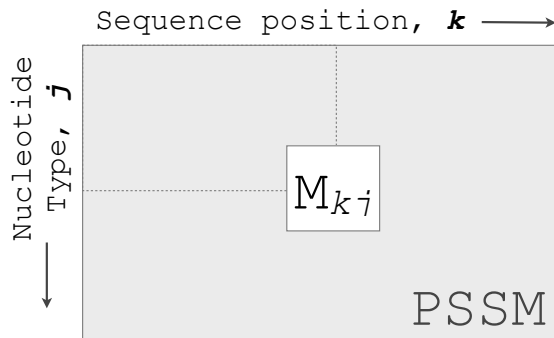
What is a **PSSM**?

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

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

Example: Computing a transcription factor bind site PSSM

```
CCAAATTAGGAAA
CCTATTAAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTTCGAAAA
CCTATTTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA
```

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
CCTATTAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAAATTTTCAAAA

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

[illegible]

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
CCTATTAAGAAAA
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

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:	0												
C:	9												
G:	0												
T:	1												
Consensus	C												

Position k = 1

Computing a transcription factor bind site PSSM

CCAAATTAGGAAA
CCTATTAAGAAAA
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	0											
C:	9	10											
G:	0	0											
T:	1	0											
Consensus	C	C											

Position k = 2

Computing a transcription factor bind site PSSM

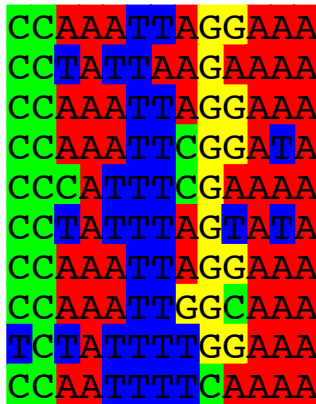
CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA

Alignment Counts matrix:

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

Position k = 3

Computing a transcription factor bind site PSSM



Sequence logo visualization of the alignment counts matrix. The logo shows the conservation of nucleotides at each position. The background is color-coded: red for positions 1, 2, 3, 10, 11, 12, 13; green for positions 4, 5, 6, 7, 8, 9; and blue for positions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. The sequences are aligned to the consensus sequence CCAATTTTCAGAA.

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAGAA

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	A	A	[AT]	T	T	A	G	G	A	A	A

Computing a transcription factor bind site PSSM

CCAAATTAGGAAA
 CCTATTAAGAAAA
 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	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	A	A	[AT]	T	T	A	G	G	A	A	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	A	A	[AT]	T	T	A	G	G	A	A	A

Computing a transcription factor bind site PSSM

CCAAATTAGGAAA
 CCTATTAAGAAAA
 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	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	A	A	[AT]	T	T	A	G	G	A	A	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

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

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

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

Computing a transcription factor bind site PSSM...

Alignment Matrix: C_{kj}

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

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

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

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

PSSM: M_{kj}

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

Scoring a test sequence

Query Sequence

CCTATTAGGATA

PSSM:

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

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

Scoring a test sequence

Query Sequence

CCTATTAGGATA

PSSM:

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

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

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

Scoring a test sequence...

Query Sequence

CCTATTAGGATA

Best Possible Sequence

CCAATTAGGAAA

PSSM:

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

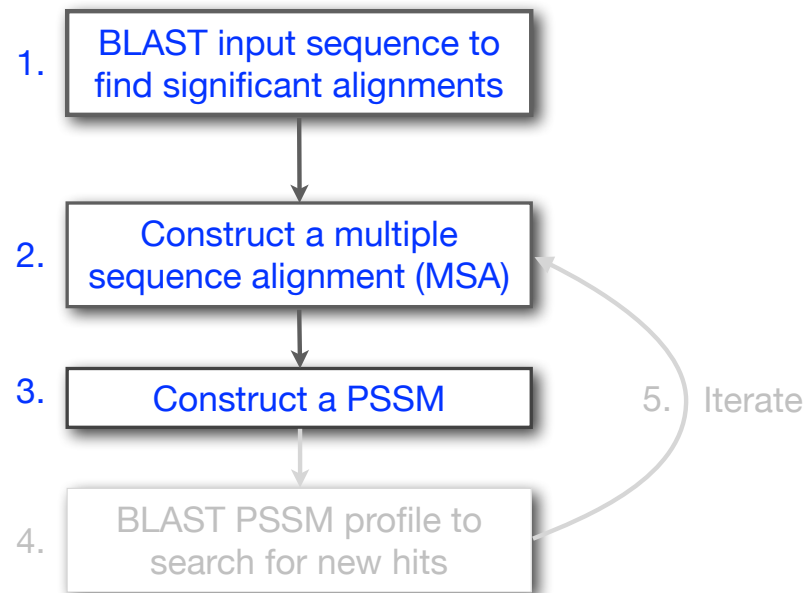
$$\begin{aligned}
 \text{Max Score} &= 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2 \\
 &\quad + 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3 \\
 &= 13.8
 \end{aligned}$$

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

Heuristic threshold for match = 60% x Max Score = (0.6 x 13.8 = 8.28);
 11.9 > 8.28; Therefore our query is a potential TFBS!

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



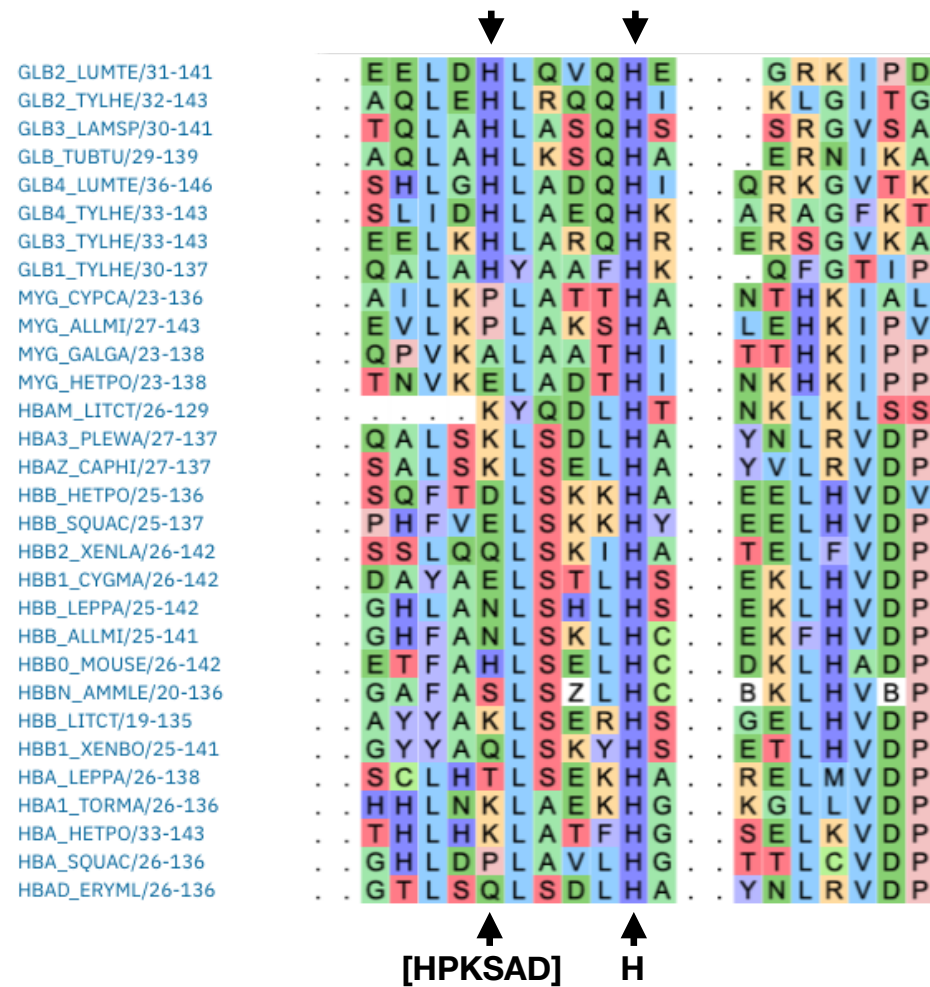
(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

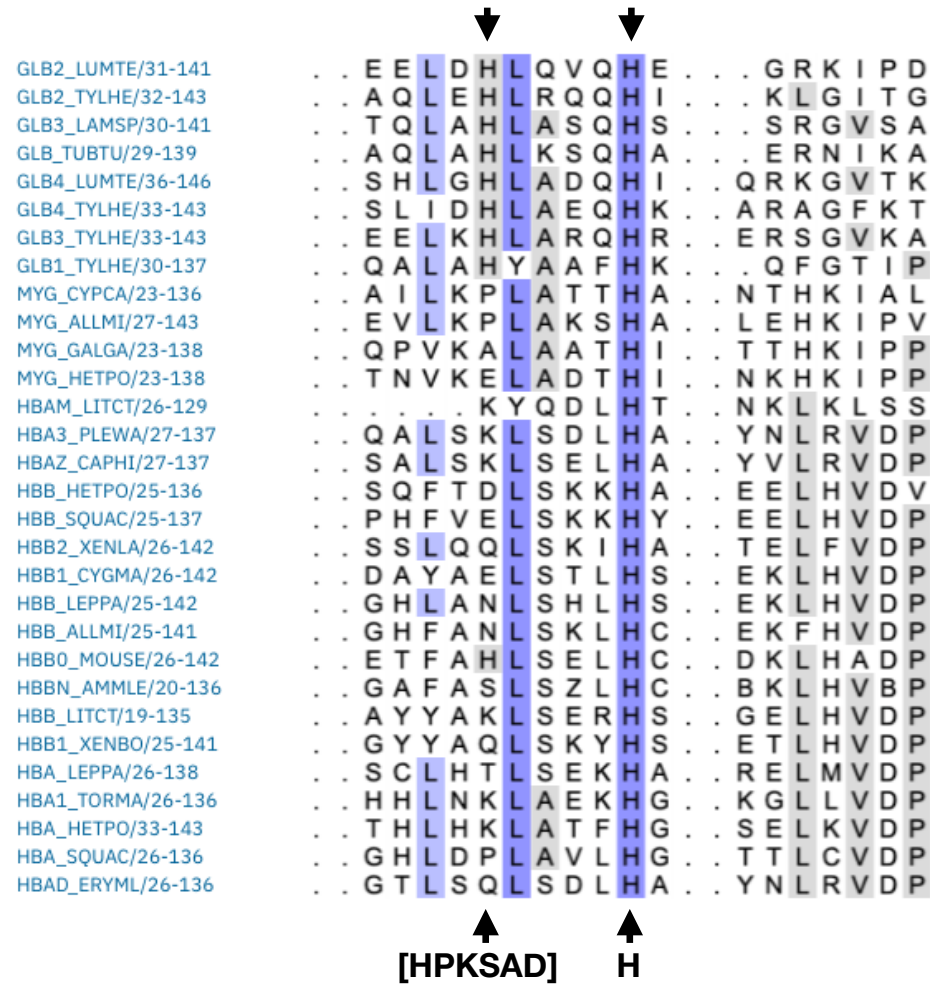
Inspect the blastp output to identify empirical “rules” regarding amino acids tolerated at each position

730496	66	FTVDENGQMSATAKGRVRLFNWVDCADMIGSFTDTEPAKFKMKYWGVAFLQKGNDDH	125
200679	63	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFDTEDPAKFKMKYWGVAFLQKGNDDH	122
206589	34	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFDTEDPAKFKMKYWGVAFLQKGNDDH	93
2136812	2	MSATAKGRVRLLSNWDVDCADMVGTFDTEDPAKFKMKYWGVAFLQKGNDDH	53
132408	65	FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH	124
267584	44	FSVDESGKVTATAHGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQKGNDDH	103
267585	44	FSVDGSGKVTATAQGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQKGNDDH	103
8777608	63	FTIHEDGAMTATAKGRVILNNWEMCADMMATFETTPDPAKFRMRYWGAAAYLQKGNDDH	122
6687453	60	FKVEEDGTMTATAIGRVILNNWEMCANMFGTFEDTPDPAKFKMKYWGAAAYLQKGYDDH	119
10697027	81	FKVQEDGTMTATATGRVILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQKGYDDH	140
13645517	1	MVGTFDTEDPAKFKMKYWGVAFLQKGNDDH	32
13925316	38	FSVDGSGKMTATAQGRVILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQKGNDDH	97
131649	65	YTVEEDGTMTASSKGRVKLFQFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY	126

↑
M

↑
N,M,L,Y,G





All the amino acids from position 1 to the end of your query seq.

20 amino acids

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

■ ■ ■

All the amino acids from position 1 to the end of your query seq.

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

Key Point:
PSSM “scores” differ for the same amino-acid type depending on where it is in your protein sequence

- i.e. scores are POSITION DEPENDENT!

All the amino acids from position 1 to the end of your query seq.

20 amino acids

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

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

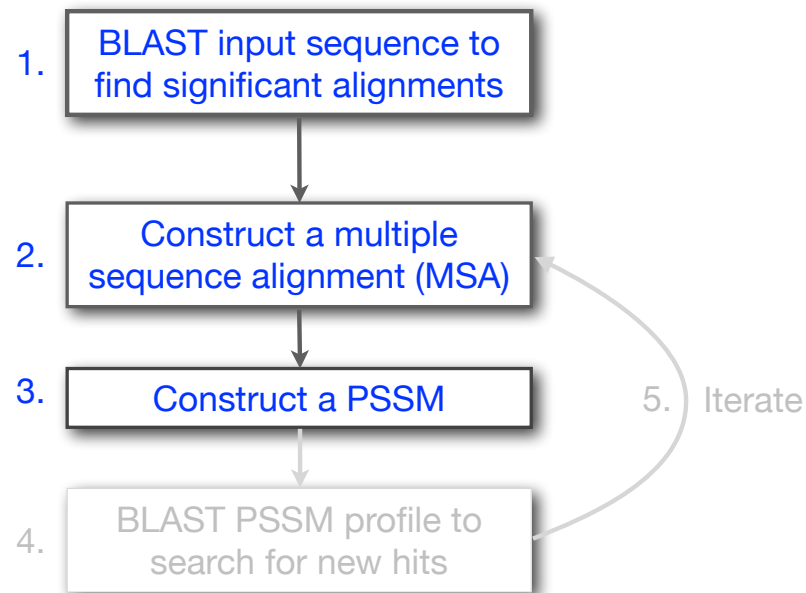
Key Point:

PSSM “scores” differ for the same amino-acid type depending on where it is in your protein sequence

- i.e. scores are POSITION DEPENDENT!

PSI-BLAST: Position-Specific Iterated BLAST

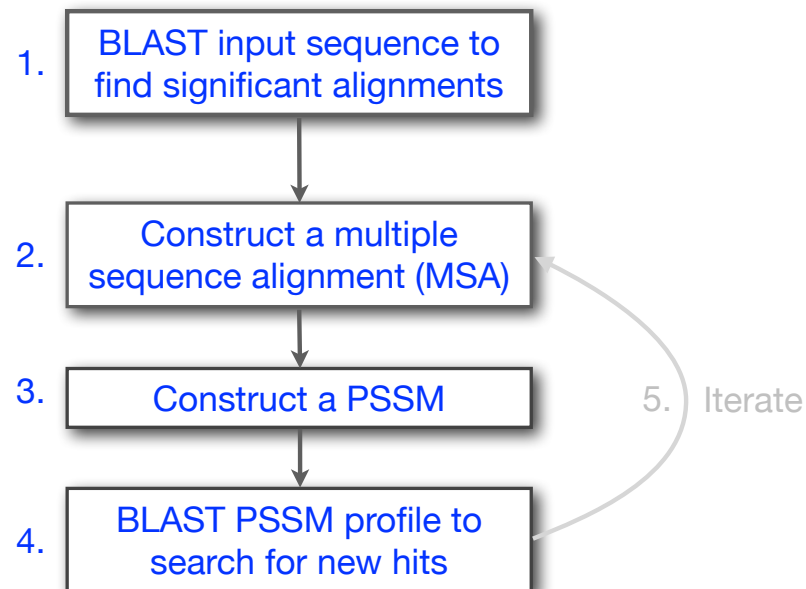
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

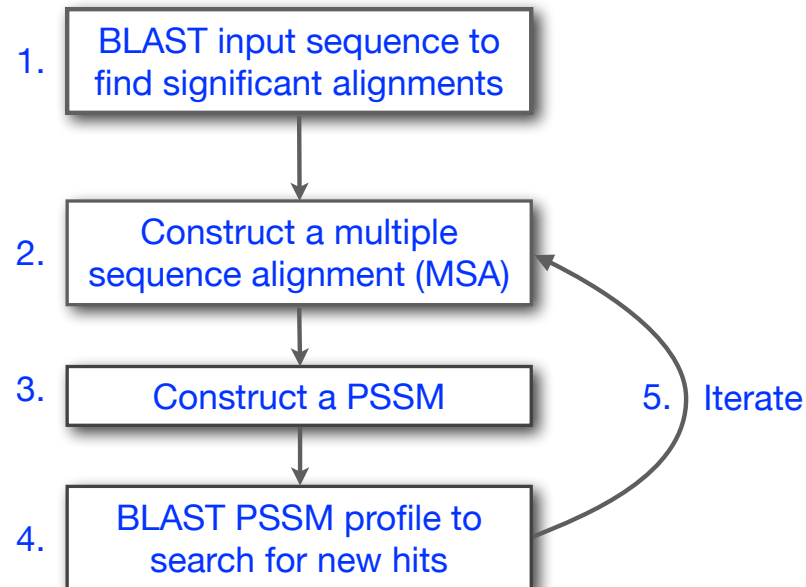
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



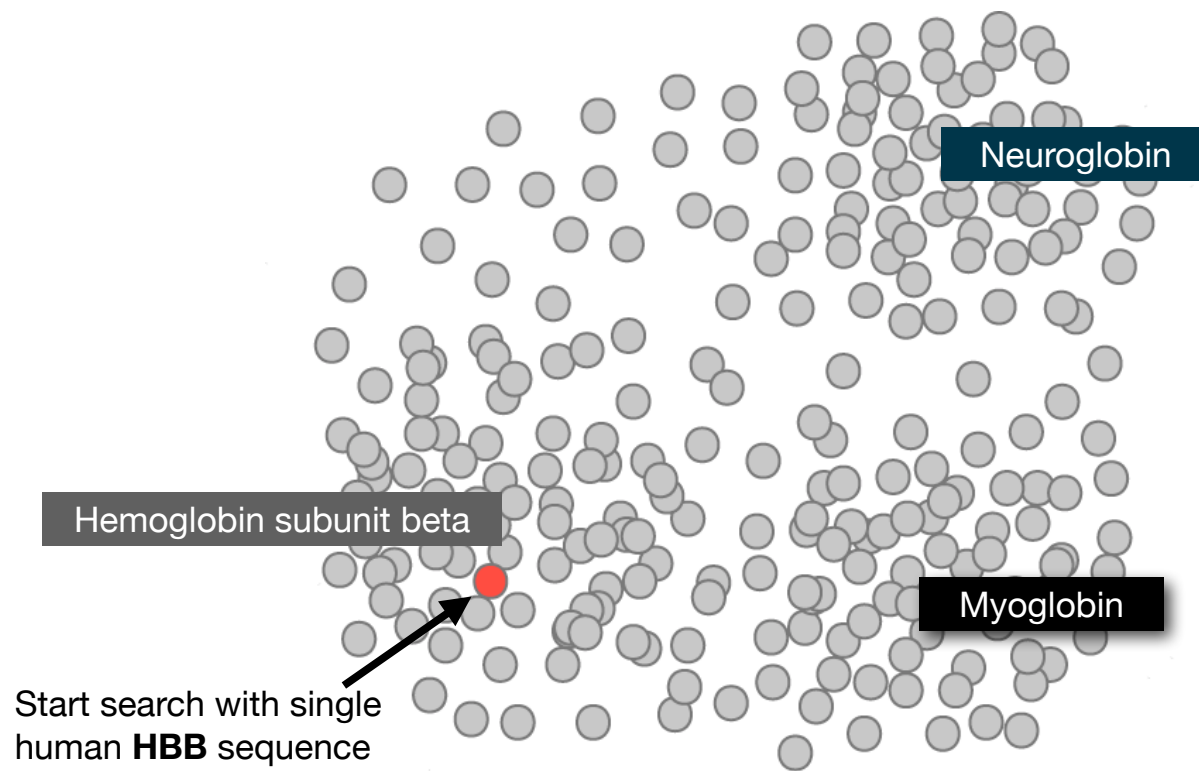
(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)

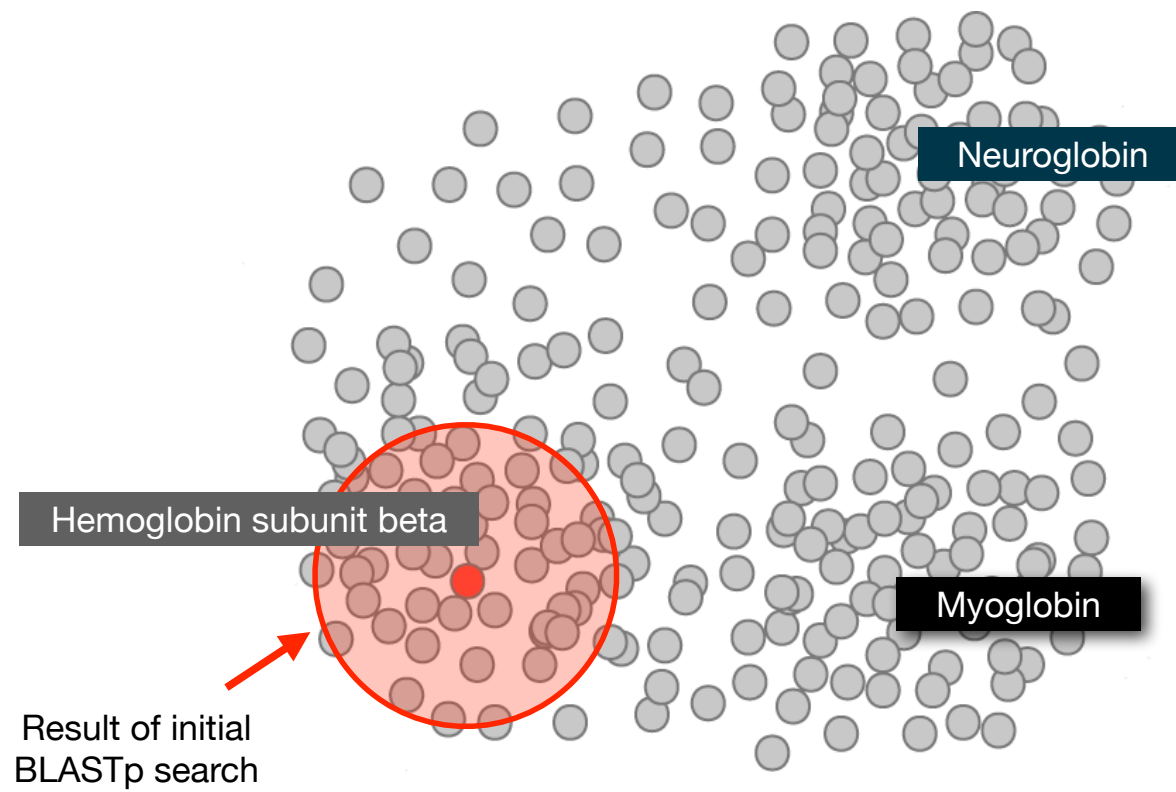
PSI-BLAST: Position-Specific Iterated BLAST

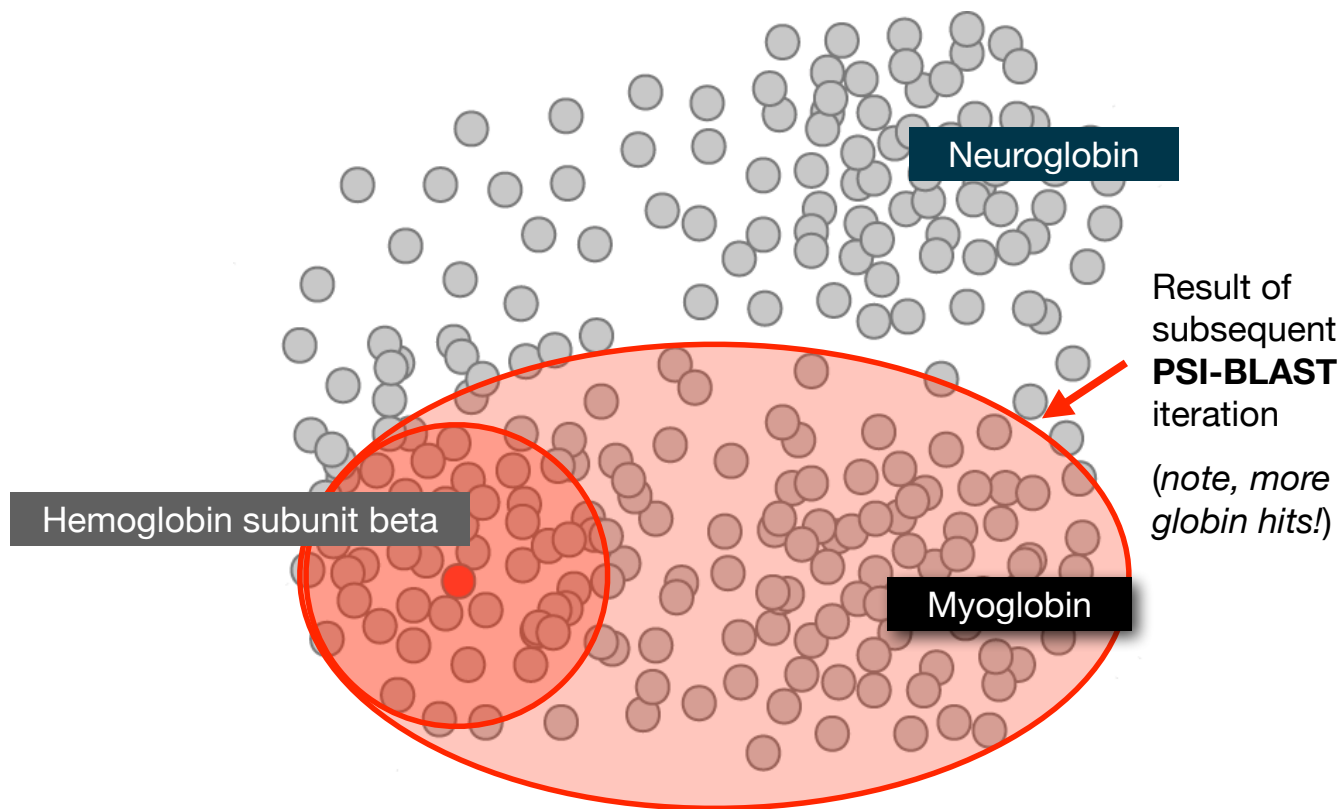
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST

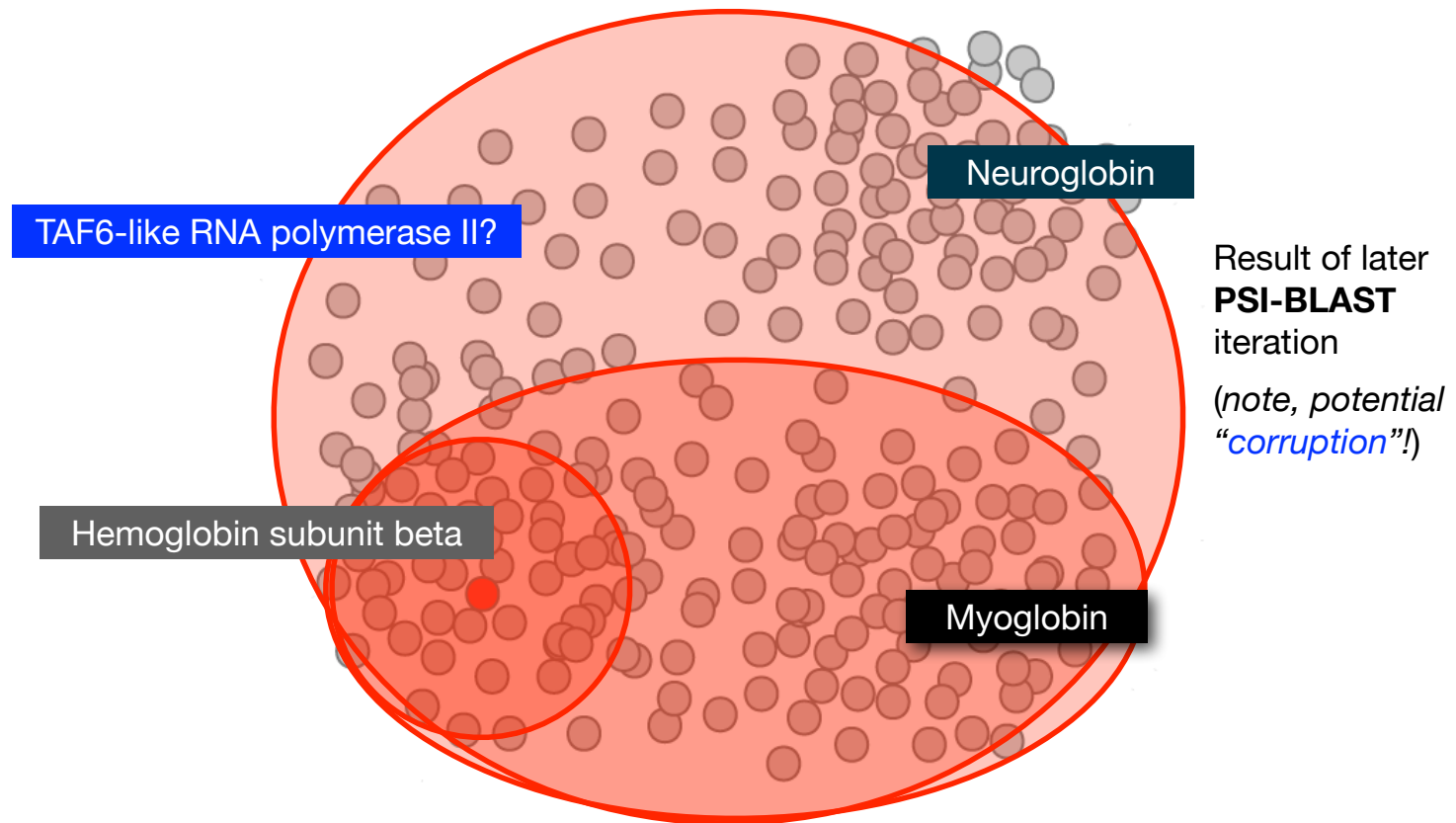


(see Altschul *et al.*, Nuc. Acids Res. (1997) 25:3389-3402)









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
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

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
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73%	NP_000175.1
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	3e-79	73%	NP_000550.2
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43%	NP_000508.1
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	36%	NP_005323.1
myoglobin [Homo sapiens]	80.5	80.5	97%	2e-19	26%	NP_005359.1
neuroglobin [Homo sapiens]	54.7	54.7	92%	2e-09	23%	NP_067080.1
myoglobin [Homo sapiens]	159	159	97%	3e-50	26%	NP_005359.1
hemoglobin subunit alpha [Homo sapiens]	151	151	97%	3e-47	42%	NP_000508.1
hemoglobin subunit mu [Homo sapiens]	147	147	97%	6e-46	35%	NP_001003938.1
hemoglobin subunit theta-1 [Homo sapiens]	147	147	97%	2e-45	37%	NP_005322.1
neuroglobin [Homo sapiens]	134	134	92%	3e-40	23%	NP_067080.1
PREDICTED: cytoglobin isoform X2 [Homo sapiens]	115	115	66%	3e-33	25%	XP_016879605.1
PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_011523942.1
PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapie	46.3	46.3	27%	7e-06	39%	XP_005258156.1

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:

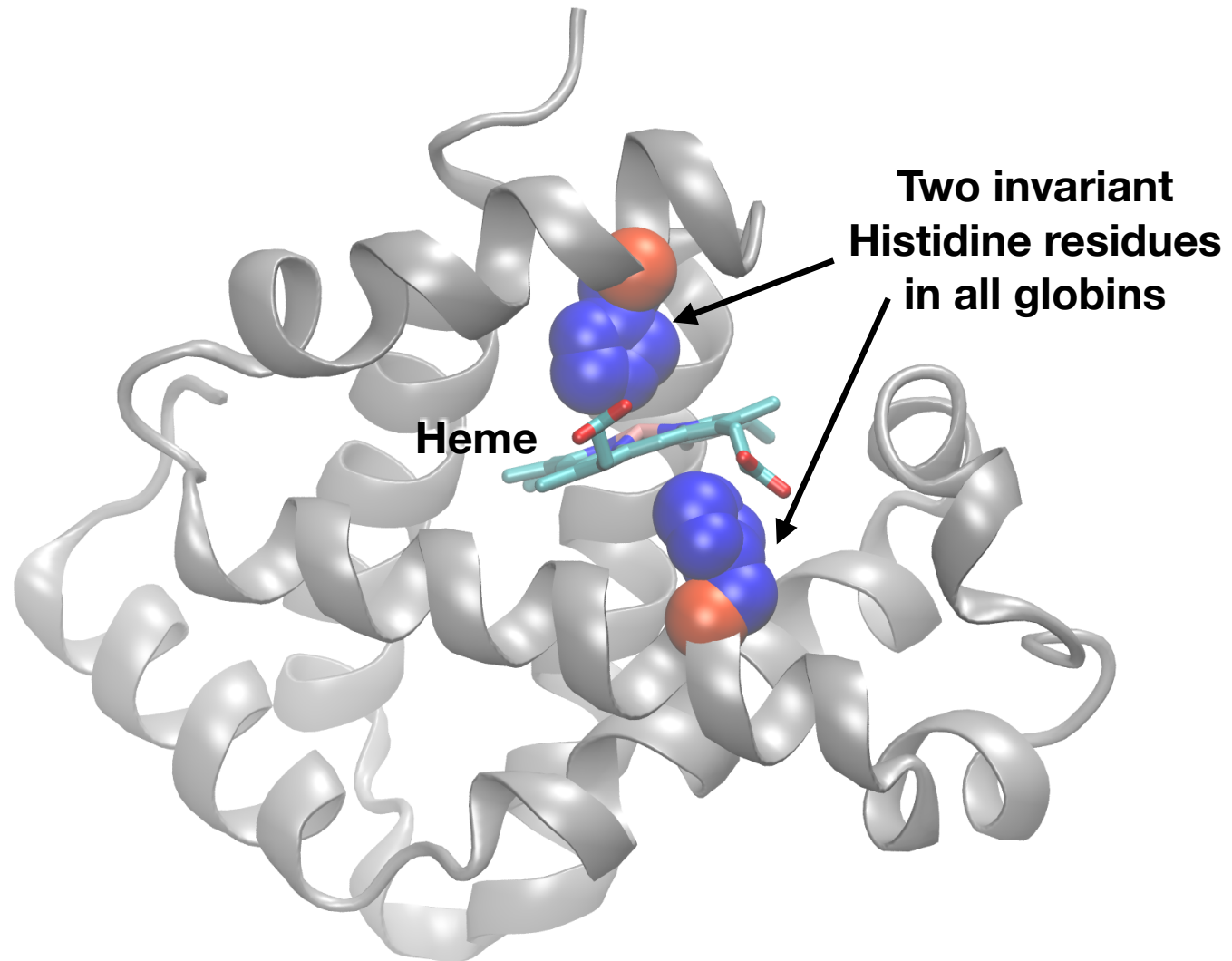
1. Limits of using BLAST [~10 mins]
2. Using PSI-BLAST [~30 mins]
3. **Examining conservation patterns** [~20 mins]

— 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**!

✓Query_73613	1	MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPD	VM-GNPKVKAHGKKVLGAF	72
✓NP_000510.1	1	MVHLTPEEKTAVNALWGKV--NVDAGGGEALGRLLVVYPWTQRFFE-SFGDLSSPD	VM-GNPKVKAHGKKVLGAF	72
✓NP_000175.1	1	MGHFTTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASA	IM-GNPKVKAHGKKVLTSL	72
✓NP_000509.1	1	MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPD	VM-GNPKVKAHGKKVLGAF	72
✓NP_005321.1	1	MVHFTAEEKAAVTSLSWKM--NVEEAGGEALGRLLVVYPWTQRFFD-SFGNLSSPSA	IL-GNPKVKAHGKKVLTSL	72
✓NP_000550.2	1	MGHFTTEEDKATITSLWGKV--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASA	IM-GNPKVKAHGKKVLTSL	72
✓NP_005323.1	1	-MSLTKTERTIIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFP-HF-----	DLHpGSAQLRAHGSKVVA	67
✓NP_000508.1	1	-MVLSPADKTNVKAAWGKVGAGHAGEYGAEALERMFSLFPTTKTYFP-HF-----	DLSHGSAQVKGHGKKVAD	67
✓XP_005257062.1	1	[15]SEELSEAERKAVQAMWARLYANCEDVGVAIVRFFVNFPSAKQYFS-QFKHMEDPLE	ME-RSPQLRKHACRVMG	89
✓NP_001003938.1	1	--MLSAQERAQIAQVWDLIAGHEAQFGAELLRLFTVYPSTKVYFP-HL-----	SACQ-DATQLLSHGQRMLA	66
✓NP_005322.1	1	-MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFS-H-----	LDLSpGSSQVRAHGQKVAD	67
✓NP_599030.1	1	[15]SEELSEAERKAVQAMWARLYANCEDVGVAIVRFFVNFPSAKQYFS-QFKHMEDPLE	ME-RSPQLRKHACRVMG	89
✓XP_016879605.1	1	-----MEDPLE	ME-RSPQLRKHACRVMG	24
✓NP_001349775.1	1	-MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGPETLEKFD-KFKHLKSEDE	MK-ASEDLKKHGATVLT	73
✓NP_067080.1	1	---MERPEPELIRQSWRAVSRSPLHGTVLFARLFALPDLLPLFQyNCRQFSSPED	CL-SSPEFLDHIRKVM	72
✓NP_001369741.1	1	-----MK-ASEDLKKHGATVLT	18	
✓Query_73613	73	SDGLAHLDDLKGT---FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKE	FTPPVQAAYQKVAVAGVANALAHKYH	147
✓NP_000510.1	73	SDGLAHLDDLKGT---FSQLSELHCDKLHVDPENFRLLGNVLVCVLAARNFGKE	FTPQMQAAYQKVAVAGVANALAHKYH	147
✓NP_000175.1	73	GDAIKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKE	FTPEVQASWQKMTGVASALSSRYH	147
✓NP_000509.1	73	SDGLAHLDDLKGT---FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKE	FTPPVQAAYQKVAVAGVANALAHKYH	147
✓NP_005321.1	73	GDAIKNMDNLKPA---FAKLSELHCDKLHVDPENFKLLGNVMVILATHFGKE	FTPEVQAAWQKLVSAVIALAHKYH	147
✓NP_000550.2	73	GDAIKHLDDLKGT---FAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKE	FTPEVQASWQKMTAVASALSSRYH	147
✓NP_005323.1	68	GDAVKSIDDIGGA---LSKLSELHAYILRVDPVNFKLLSHCLLVTLAARFPAD	FTAHAHAWDKFLSVSVLTKYR	142
✓NP_000508.1	68	TNAVAHVDDMPNA---LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAE	FTPAVHASLDKFLASVSTVLT	142
✓XP_005257062.1	90	NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASD	FPPEPQRAWAKLRGLIYSHVTAAYK [35]	202
✓NP_001003938.1	67	GAAVQHVNDLRAA---LSPLADLHALVLRVDPANFPLLIQCFHVVLASHLQDE	FTVQMQAAWDKFLTGVAVVLT	141
✓NP_005322.1	68	SLAVERLDDLPHA---LSALSHLHACQLRVDPASFQLLGHCLLVTLARHPGDF	SPALQASLDKFLSHVISALVSEYR	142
✓NP_599030.1	90	NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASD	FPPEPQRAWAKLRGLIYSHVTAAYK [23]	190
✓XP_016879605.1	25	NTVVENLHDPDKVssvLALVGKAHALKHKVEPVYFKILSGVILEVVAEEFASD	FPPEPQRAWAKLRGLIYSHVTAAYK [35]	137
✓NP_001349775.1	74	GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGD	FGADAQGAMNKALELFRKDMASNYK [6]	154
✓NP_067080.1	73	DAAVTNVEDLSSLeeyLASLGRKHRA-VGVKLSSTFSTVGESLLYMLEKCLGPA	FTPATRAAWSQLYGAVVQAMSRGWD [2]	151
✓NP_001369741.1	19	GGILKKKGHHEAE---IKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGD	FGADAQGAMNKALELFRKDMASNYK [6]	99



HW Q: Make your own PSSM

You can work on it now in-class!

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				

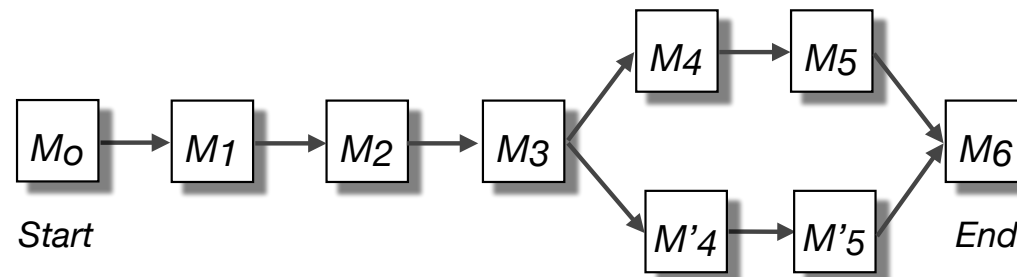
Note: We never see **QD** or **RH**, we only see **RD** and **QH**.
However, $P(RH)=0.24$, $P(QD)=0.24$, while $P(QH)=0.16$

Markov chains: Positional dependencies



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

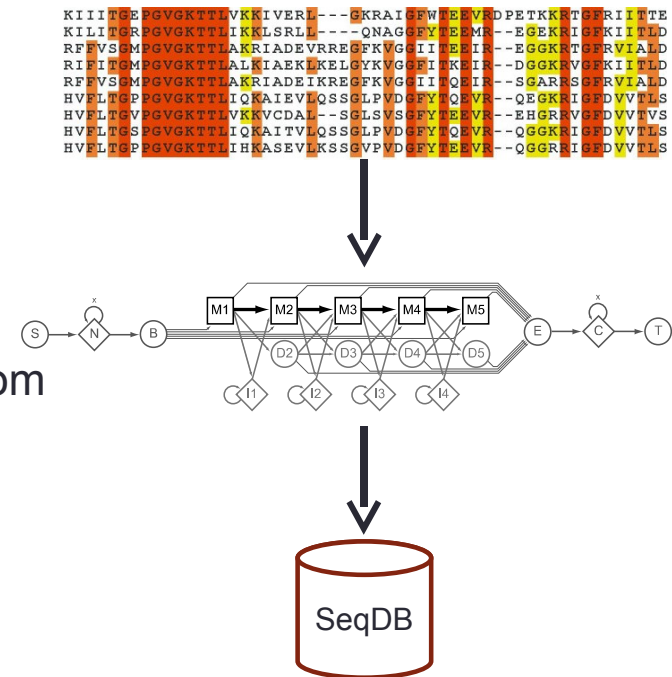
WEIRD
WEIRD
WEIQH
WEIRD
WEIQH



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

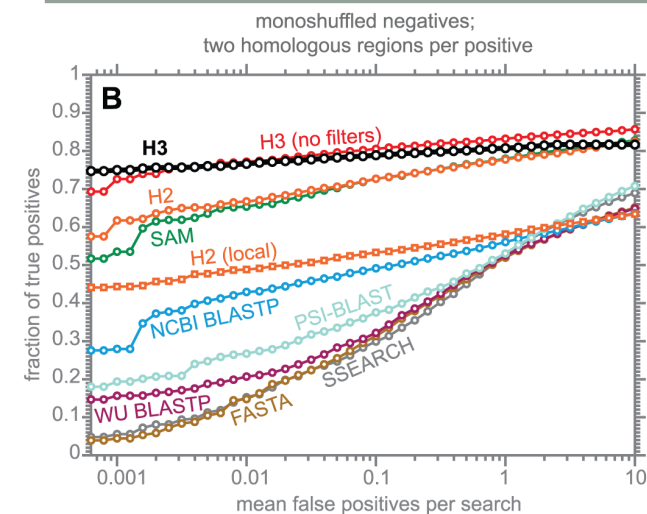
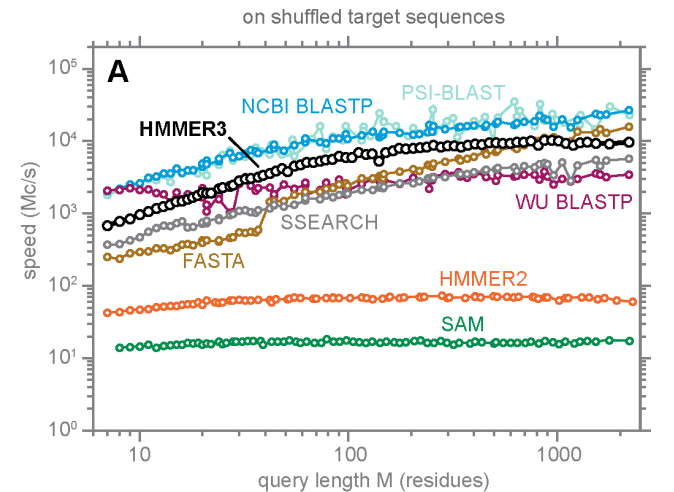
Use of HMMER

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



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.





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]  
MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG  
AFSDGLAHLNLDNLKGTFTLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAAYQKVAGVAN  
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

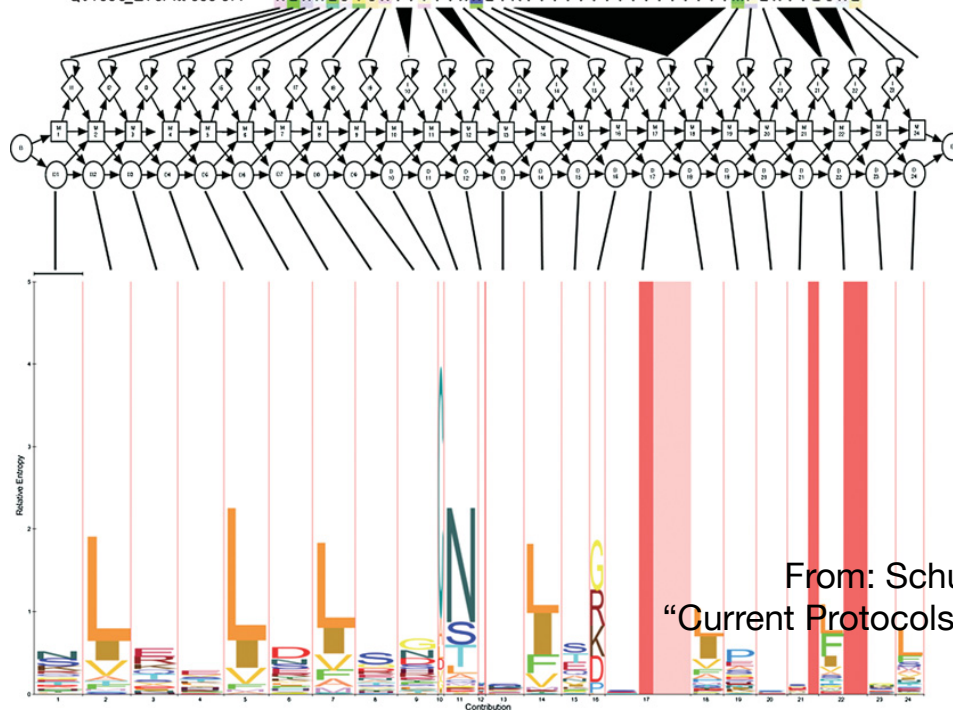
Comprehensive compilation of both multiple sequence alignments and profile HMMs of protein families.

<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**

Q9ARB2_LINUS/823-844 MLEYLDIGRA..P.RIV.H.....LDG...LENL
 Q9MBN0_ARATH/320-341 RLTFNLISFC..S.KLT.G.....LAF...FSII
 FLJ1_HUMAN/318-339 NLEEFMAAN..N.NLE.L.....VPES..LCRC
 Q9VIN74_DROME/90-112 ALHSLVIENC...TIV.H.....INDAA.FNQE
 Q8L8I7_PNTA/792-814 NLQTIQMYRX..E.SLQ.V.....LPDS..FGNL
 Q9FHL8_ARATH/301-324 NLWSNLRSR..N.LFSDP.....LPVVG.ARGF
 SLK6_MOUSE/65-87 RPFHLSLLN..N.GLT.M.....LHTND.FSGL
 Q8NUJ8_EMENI/978-1000 TLTSNLIAS..A.KLV.Q.....FRDTL.FDSL
 Q9LUQ2_ARATH/92-113 AMKSLDVSF..N.SIS.E.....LPEQ..IGSA
 Q9FH93_ARATH/169-188 RLTSNLNDF..N.RFNGT.....LPS...LN
 Q898G0_CLOTE/268-288 YLERINLDK..N.KI.KN.....IEE...LEAN
 Q8H6V2_MAIZE/678-699 NLRIISIVDC..V.SLQ.K.....LPP...SDSF
 Q9AR40_LINUS/692-713 DLKVLIDINQ..T.EIT.T.....LKGE..VESL
 Q9LE82_ARATH/350-377 HLTEIYMSY..L.NLEDEGT.....EALSEAL.LKSA
 Q9H5N5_HUMAN/255-278 HLQVLDLHQC..S.LT.AD.....DVMSL...TQVI
 Q8L4C7_ARATH/185-207 KLEYLDIWG..S.NVT.N.....GAVS..ILKF
 Q9VSA4_DROME/1115-1136 QLKARLQC..N.AI.GSH.....GLEAL..LCGQ
 TLR1_MOUSE/376-398 RLKTLQK..N.QL.KN.....LENII.LTSA
 Q9TXJ6_LEIMA/445-465 GLRDIDLSH..T.KVH.N.....IDA...LQAS
 FXL13_MOUSE/409-448 KLIYDLSGC..T.QVL.VEKCPRISSVVLIGSPHISDSA.FKAL
 Q9TXJ6_LEIMA/927-948 ALTVVNAWSC..V.NLT.S.....IEA...LESA
 Q9M4X9_CHLRE/1417-1444 LLAVLHLHD..NP.R.A.ADG.....VAGLAAA..LPGL
 Q945S6_LYCPIN/656-677 NLRHLDVSN..T.RRL.K.....MPLH..LSRL



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

YOUR TURN!

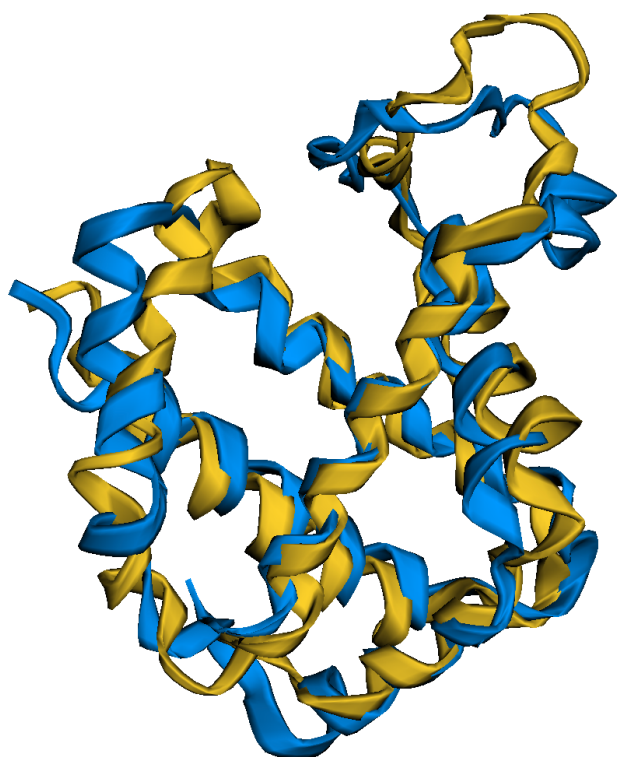
- There are **four required** and **one optional** hands-on sections including:

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**!



ALIGNMENT

CONTACT MAP

Align 2hbsB.pdb 146 with 4mpmB.pdb 148

Twists 0 ini-len 136 ini-rmsd 3.05 opt-eu 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	H	L	T	P	V	E	K	S	A	V	T	A	L	W	G	K	V	N	--	V	D	E	V	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	E	S	F	G	--	D	L	S	T

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!