

Week 2

Class 3: Hands-on section

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

The screenshot shows a Google Calendar view for the BIMM 143 course. A red box highlights the entry for Tuesday, October 5, 2021, which is labeled "Project: Find a gene project assignment". This entry includes a detailed description of the assignment, mentioning principles of database searching, sequence analysis, and general data analysis. A red arrow points from the left towards this highlighted box. Another red arrow points from the top right towards the "Project: Find a gene project assignment" text.

Day	Date	Description
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.
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.

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 5 (**Oct 30th**, 10/30/23).
 - The complete assignment, including responses to all questions, is due 12pm Monday of week 10 (**Dec 4th**, 12/03/23).

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 5 (**Oct 30th**, 10/30/23).
 - The complete assignment, including responses to all questions, is due 12pm Monday of week 10 (**Dec 4th**, 12/03/23).

Questions:

[Q1] Take 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 “test” 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 sequences/subject lines from your BLAST result page). You can use UniProt or InterProScan or DNasearch or a tool like the EMBOSS Transl 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 that this protein has been previously reported in a research grant from an organization such as the National Institutes of Health. 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 that 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 sequence, and two other randomly-chosen members of the same protein family. 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 the sequence. Use short common names (rather than accession numbers) display 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.

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.

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.

- Your responses to questions **Q1-Q4** are due 12pm Monday of week 5 (**Oct 30th**, 10/30/23).
- The complete assignment, including responses to **all questions**, is due 12pm Monday of week 10 (**Dec 4th**, 12/03/23).

Class 3: Hands-on section

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

Schedule

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

R Shiny App

Details:

Sequence 1: GATTAC
Sequence 2: GTGAGCG

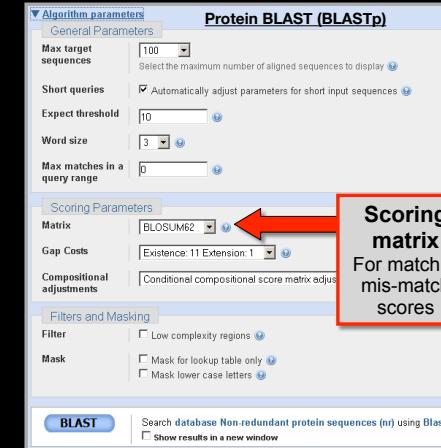
	G	T	C	G	A	C	G	C	
G	0	-2	-4	-6	-8	10	-12	-14	-16
A	-2	1	-1	-3	-5	6 + 1 (Due to a match between G & C) = -5	-8 + -2 (The Gap score) = -10		
T	-4	1	0	-2	-4	Score from Side cell	Score from Upper cell		
C	-6	-3	0	-1	-3	3 + -2 (The Gap score) = -5	Winning (max) score is -5		
G	-8	-5	-2	1	-2	-6	-8		
A	-10	-7	-4	-3	-2	-4	-5	-7	
T	-12	-9	-6	-3	-4	-5	-7	-9	

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?



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 0 -1 4
G	-3 0 -2 -2 0 6
N	-3 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 1 4
L	-1 -2 -1 -3 -1 -4 -3 -4 -3 -2 -3 -2 2 2 4
V	-1 -2 0 -2 0 -3 -3 -3 -2 -2 -3 -3 1 3 1 4
F	-2 -2 -2 -4 -2 -2 -3 -3 -3 -3 -1 -3 -3 0 0 0 -1 6
Y	-2 -2 -2 -3 -2 -3 -2 -1 -2 -2 -1 -1 -1 3 7
W	-2 -3 -2 -4 -3 -2 -4 -3 -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	

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 0 -1 4
G	-3 0 -2 -2 0 6
N	-3 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 1 4
L	-1 -2 -1 -3 -1 -4 -3 -4 -3 -2 -3 -2 2 2 4
V	-1 -2 0 -2 0 -3 -3 -3 -2 -2 -3 -3 1 3 1 4
F	-2 -2 -2 -4 -2 -2 -3 -3 -3 -3 -1 -3 -3 0 0 0 -1 6
Y	-2 -2 -2 -3 -2 -3 -2 -1 -2 -2 -1 -1 -1 3 7
W	-2 -3 -2 -4 -3 -2 -4 -3 -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	

Not all matches score equally
(blue highlighted values)

By default BLASTp match scores come from the BLOSUM62 matrix

C : C scores +9

V : V scores +4

BLOSUM62 Substitution Matrix:

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

Note: Not all matches score equally (blue highlighted values)

By default BLASTp match scores come from the BLOSUM62 matrix

C : C scores +9

V : V scores +4

BLOSUM62 Substitution Matrix:

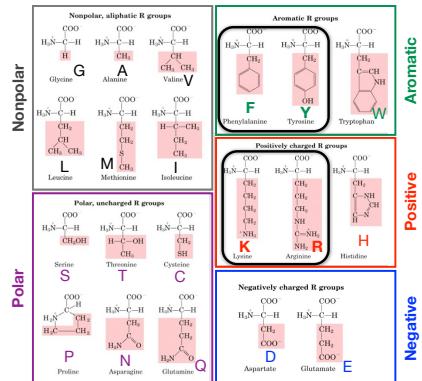
C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
S	-1	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
T	-1	-1	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
P	-1	-1	-1	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	1	0	-1	4	-1	-1	-1	-1	-1	-1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

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

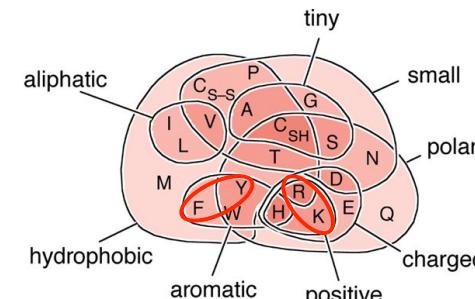
Not all matches score equally (blue highlighted values)

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. BLOSUM62 does not take the local context of a particular position into account
(i.e. all like substitutions are scored the same regardless of their location in the molecules).

We will revisit this later...

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]

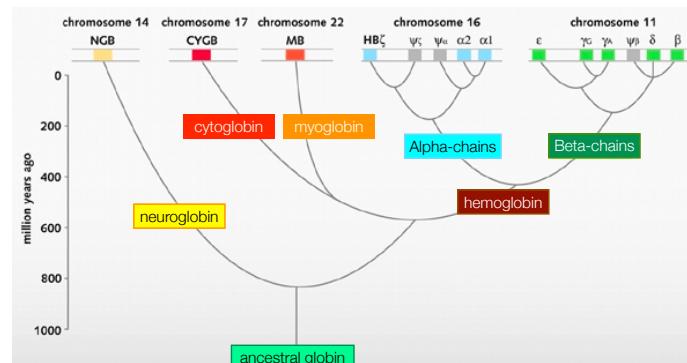
- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!

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]

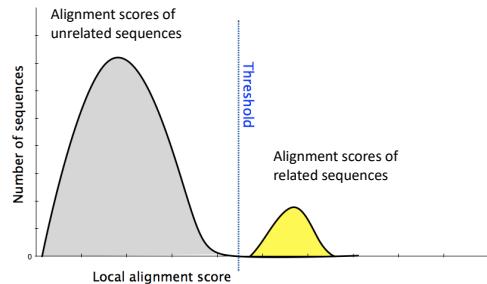
- Please do answer the last review question (**Q20**).
- We encourage discussion at your **Table** and on **Piazza**!



An evolutionary model of human globins.

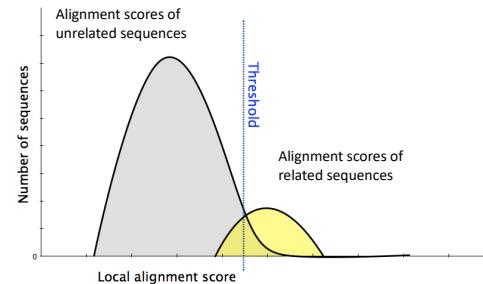
The different locations of globin genes in human chromosomes are reported at the top of the figure, distinguishing between the functional genes (in color) and the pseudogenes (in grey).

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



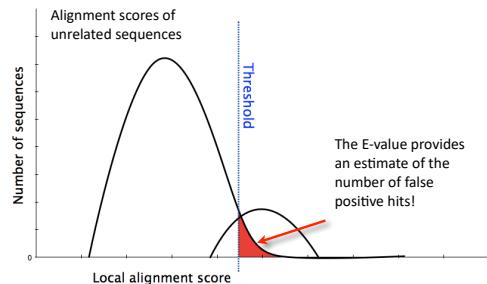
25

- 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



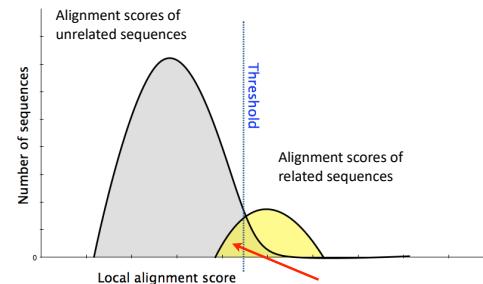
26

- 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



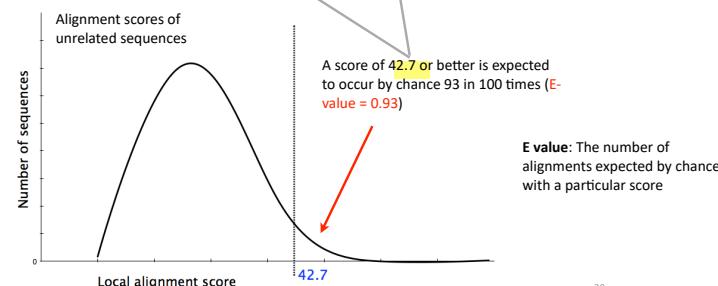
27

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



28

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



Recall: BLOSUM62 does not take the local context of a particular position into account

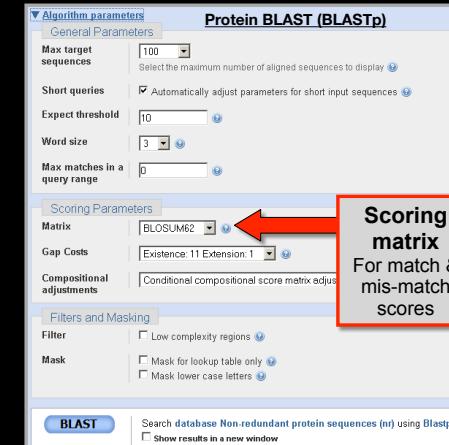
(i.e. all like substitutions are scored the same regardless of their location in the molecules).

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**!



By default BLASTp match scores come from the BLOSUM62 matrix

C	9
S	-1 4
T	-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 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 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 -2 5
M	-1 -1 -1 -2 -1 -3 -2 -3 -2 0 -2 -1 5
I	-1 -2 -1 -3 -1 -4 -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 -2 -2 -3 -2 -1 1 3 1 4
F	-2 -2 -2 -4 -2 -3 -3 -3 -3 -1 -3 -3 0 0 0 -1 6
Y	-2 -2 -2 -3 -2 -3 -2 -1 2 -2 -1 -1 -1 3 7
W	-2 -3 -2 -4 -3 -2 -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	

Blocks Substitution Matrix. 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 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 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 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 5
I	-1 -2 -1 -3 -1 -4 -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 -2 -2 -3 -2 -3 -2 1 3 1 4
F	-2 -2 -2 -4 -2 -3 -3 -3 -3 -1 -3 -3 0 0 0 -1 6
Y	-2 -2 -2 -3 -2 -3 -2 -1 2 -2 -1 -1 -1 3 7
W	-2 -3 -2 -4 -3 -2 -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

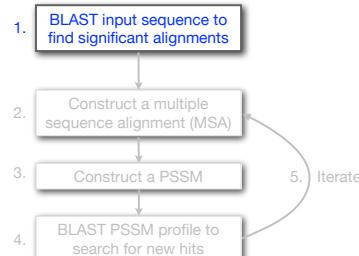
- The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query

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

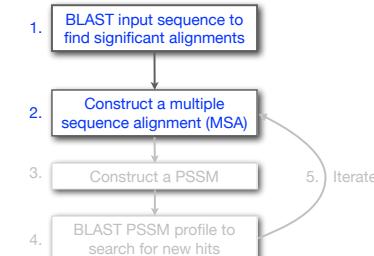
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

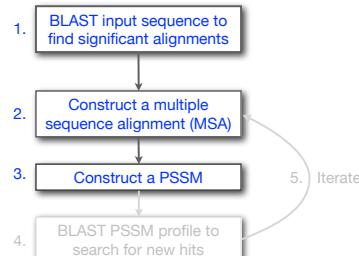
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



(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)

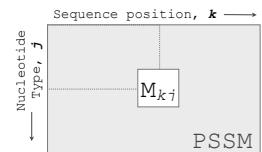
What is a **PSSM**?

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

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

Example: Computing a transcription factor bind site PSSM

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTAGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAGA
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 13×4 **PSSM** ($k=13$, $j=4$).

Computing a transcription factor bind site PSSM

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

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

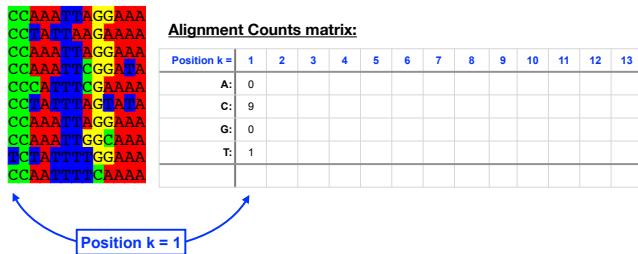
Computing a transcription factor bind site PSSM

Alignment Counts matrix:

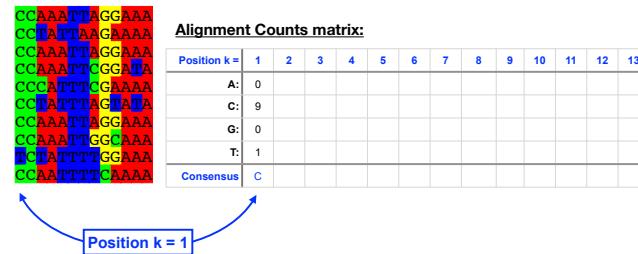
Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
A:													
C:													
G:													
T:													

Position k = 1

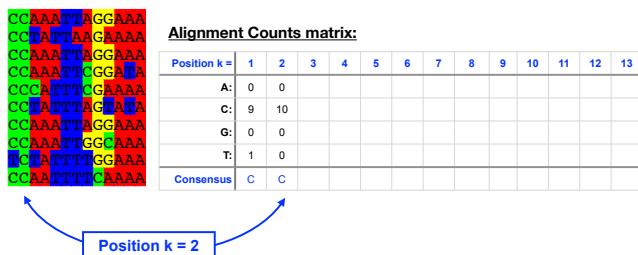
Computing a transcription factor bind site PSSM



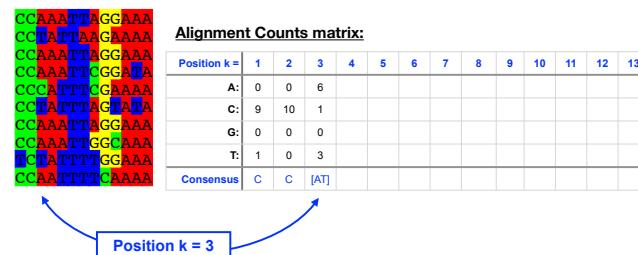
Computing a transcription factor bind site PSSM



Computing a transcription factor bind site PSSM



Computing a transcription factor bind site PSSM



Computing a transcription factor bind site PSSM

CCAAATTTAGGAAA
CCTATTAAAGAAAA
CCAAATTAGGAAA
CCATTTCGGATAA
CCCATTAGGAAA
CCAAATTAGGAAA
CCAAATTGGAAA
CCAAATTGGAAA
CTTATTTCGGAAA
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

Computing a transcription factor bind site PSSM

CCAAATTTAGGAAA
CCTATTAAAGAAAA
CCAAATTAGGAAA
CCATTTCGGATAA
CCCATTAGGAAA
CCAAATTAGGAAA
CCAAATTGGAAA
CCAAATTGGAAA
CTTATTTCGGAAA
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	0	0	0	1	9	5
T:	1	0	3	0	5	10	9	2	0	1	0	2	0
Consensus	C	C	[AT]	A	[AT]	T	T	[ACT]	G	[GA]	A	[AT]	A

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

Average Profile (Frequency) matrix:

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

Computing a transcription factor bind site PSSM

CCAAATTTAGGAAA
CCTATTAAAGAAAA
CCAAATTAGGAAA
CCATTTCGGATAA
CCCATTAGGAAA
CCAAATTAGGAAA
CCAAATTGGAAA
CCAAATTGGAAA
CTTATTTCGGAAA
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

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	0	2	1	1	0	0
G:	0	0	0	0	0	0	0	0	1	9	5	0	0
T:	1	0	3	0	5	10	9	2	0	1	0	2	0

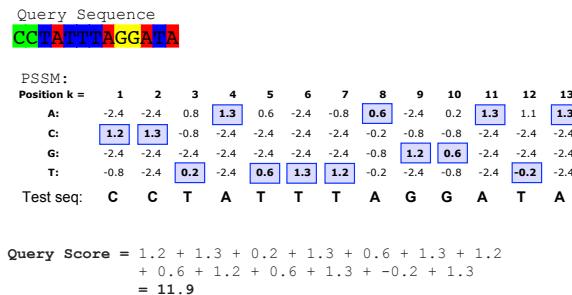
$$k=1, j=A: M_{ij} = \log\left(\frac{C_{ij} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{0 + 0.25 / 10 + 1}{0.25}\right) = -2.4$$

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

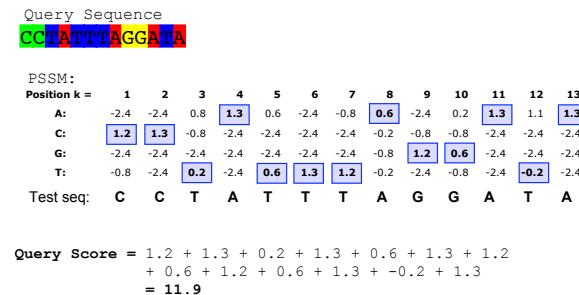
$$k=1, j=T: M_{ij} = \log\left(\frac{C_{ij} + p_j / Z + 1}{p_j}\right) = \log\left(\frac{1 + 0.25 / 10 + 1}{0.25}\right) = -0.8$$

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	-0.2	-0.8	-2.4	-2.4	-2.4	-2.4	-2.4
G:	-2.4	-2.4	-2.4	-2.4	-2.4	-2.4	-0.8	1.2	0.6	-2.4	-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

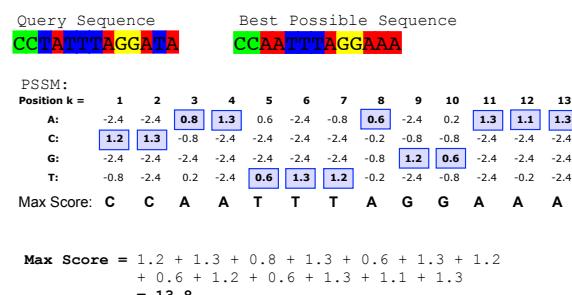


Scoring a test sequence



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

Scoring a test sequence...

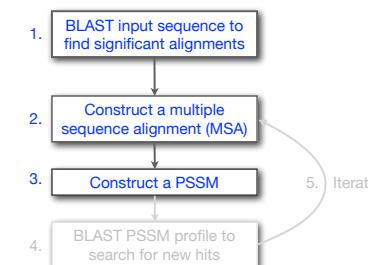


A. Following method in Harbison et al. (2004) Nature 431:99-104

Heuristic threshold for match = $60\% \times \text{Max Score} = (0.6 \times 13.8 = 8.28)$; $11.9 > 8.28$; 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	FTVDENGQMSATAKGRVRLFNNNUDVACDHIGSFTDTEPAKFHKYUWGVASFLQKGNDH 125
200679	63	FSVDEKGHMSATAKGRVRLLSNUEVCADMVGTFDTEDPAFKMKYUWGVASFLQKGNDH 122
206589	34	FSVDEKGHMSATAKGRVRLLSNUEVCADMVGTFDTEDPAFKMKYUWGVASFLQKGNDH 93
2136812	2	HSATAKGRVRLNNNUDVACDMVGTFDTEDPAFKMKYUWGVASFLQKGNDH 53
132408	65	FKIEDNGKTTATAKGRVRLDKLECANMVGTFIETNDPAKFHKYUWGVASFLQKGNDH 124
267584	44	FSVDESGKVTTATAKGRVILNNWEMCANHFGTFEDTPDPARKMKRYUWGAASYLQSGNDH 103
267585	44	FSVDESGKVTTATAKGRVILNNWEMCANHFGTFEDTPDPARKMKRYUWGAASYLQSGNDH 103
8777608	63	FTIHEDGAMTTATAKGRVILNNWEMCADMMAFFTPDPARKMKRYUWGAASYLQSGNDH 122
6687453	60	FKVVEEDGTMATAKGRVILNNWEMCANHFGTFEDTPDPARKMKRYUWGAASYLQSGNDH 119
10697027	81	FKVQEDGTMATATKGRVILNNWEMCANHFGTFEDTPDPARKMKRYUWGAASYLQSGNDH 140
13645517	1	MVGTFTDTEPAFKMKYUWGVASFLQKGNDH 32
13925316	38	FSVDGSGKNTATAQGRVIIILNNWEMCANHFGTFEDTPDPARKMKRYUWGAASYLQSGNDH 97
131649	65	YTVEEDGTMATASSKGRVKLFGFWVICADMAAQYTDPTTPARKMYHTYQGLASYLSSGGDNY 126

M

N,M,L,Y,G

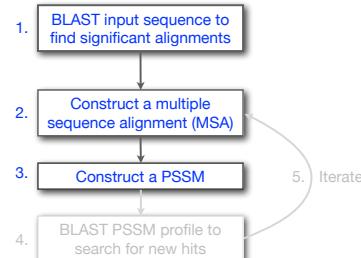
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2 K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	1	0	-1	-3	-2	-3
3 W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3	
4 V	0	-3	-3	-4	1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	-3	-1	-4	
5 W	-3	-3	-4	-5	-3	-2	-3	-3	-3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

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

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

PSI-BLAST: Position-Specific Iterated BLAST

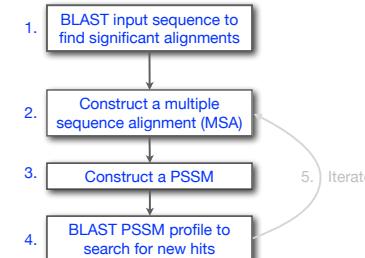
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

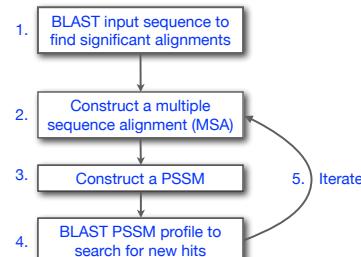
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



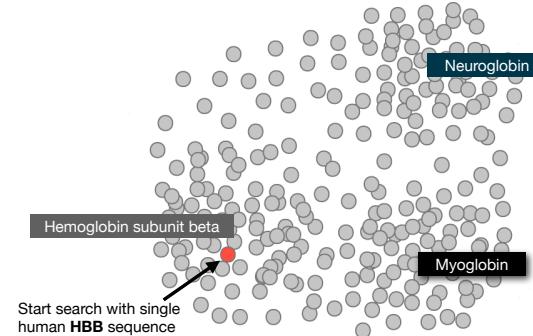
(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)

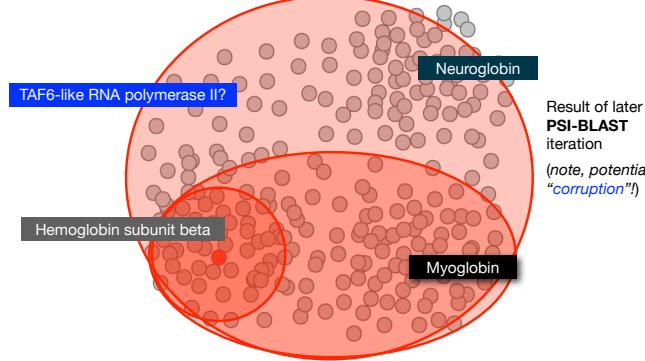
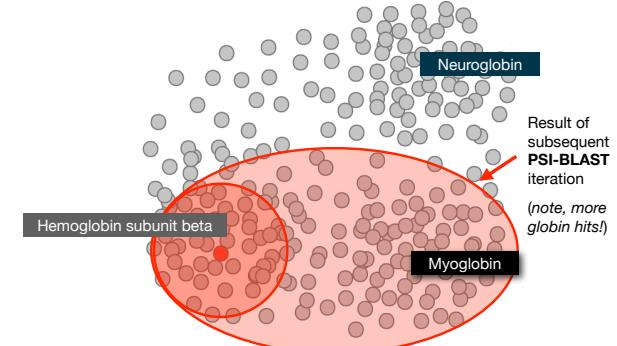
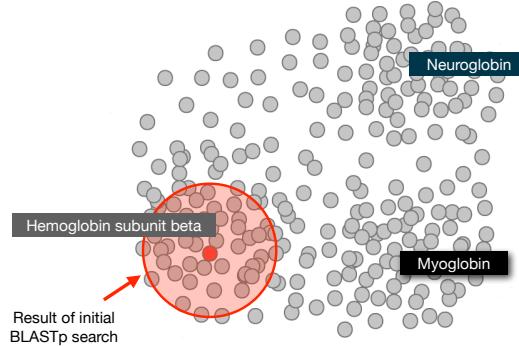
PSI-BLAST: Position-Specific Iterated BLAST

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-100	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

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_000509.2
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

New relevant globins found only by PSI-BLAST

1

2

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_000509.2
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

1

2

?

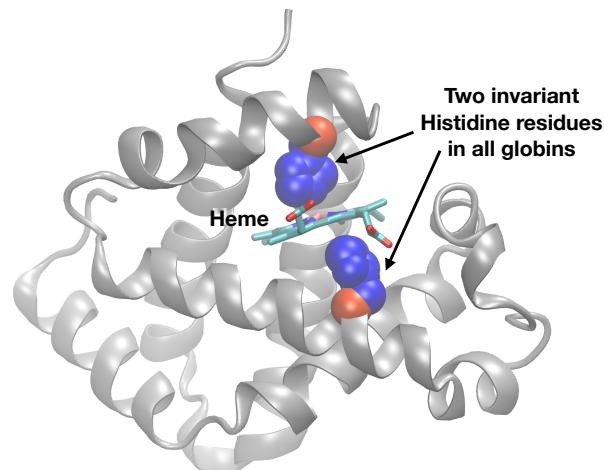
Inclusion of irrelevant hits can lead to PSSM corruption

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:
 - Limits of using BLAST
 - Using PSI-BLAST
 - Examining conservation patterns**
— BREAK [15 mins] —
 - [Optional] Using HMMER
 - Divergence of protein sequence and structure

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

<input checked="" type="checkbox"/> Query_73613	1	MVHLTPEEKSAVTALNKGKV--NVDEVGGEALGRLLVVYWPWTQRFFP--SPGDLSTPDAVM-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_000510.1	1	MVHLTPEEKSAVTALNKGKV--NVDEVGGEALGRLLVVYWPWTQRFFP--SPGDLSTPDAVM-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_000175.1	1	MGHFTTEEEKATITSLNGKV--NVDAGGETLGRLLVVYWPWTQRFFP--SFGNLSSASA1M-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_000509.1	1	MVHLTPEEKSAVTALNKGKV--NVDEVGGEALGRLLVVYWPWTQRFFP--SPGDLSTPDAVM-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_005321.1	1	MVHFTAEKAAVTSLSKMK--NVEAGEGALGRLLVVYWPWTQRFFP--SFGNLSSPSA1L-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_005500.2	1	MGHFTTEEKATITSLNGKV--NVDAGGETLGRLLVVYWPWTQRFFP--SFGNLSSASA1M-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_005359.1	1	MVHLTPEEKSAVTALNKGKV--NVDEVGGEALGRLLVVYWPWTQRFFP--SPGDLSTPDAVM-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_005323.1	1	MVHFTAEKAAVTSLSKMK--NVEAGEGALGRLLVVYWPWTQRFFP--SFGNLSSPSA1L-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_005257062.1	1	MGHFTTEEKATITSLNGKV--NVDAGGETLGRLLVVYWPWTQRFFP--SFGNLSSASA1M-GNPVKVKAHGKVLGAF	72
<input checked="" type="checkbox"/> NP_001003938.1	1	MVLSPAOKTNVKAANGKVAHGEAEGARLMLFSTTKEVYF--HF----DLS-GSAOVRKAGSKVVAAV	67
<input checked="" type="checkbox"/> NP_005322.1	1	MVLSPAOKTNVKAANGKVAHGEAEGARLMLFSTTKEVYF--HF----DLS-GSAOVRKAGSKVVAAV	67
<input checked="" type="checkbox"/> NP_0059030.1	1	MVLSPAOKTNVKAANGKVAHGEAEGARLMLFSTTKEVYF--HF----DLS-GSAOVRKAGSKVVAAV	67
<input checked="" type="checkbox"/> NP_016879605.1	1	MGSLSGEWQLVLNWGKVKAEDIPGHQEVQLRFLKGHPETLEKFD-KFKLILSEDEENK-ASIDLLKKHGTVLW1	24
<input checked="" type="checkbox"/> NP_00134975.1	1	MGSLSGEWQLVLNWGKVKAEDIPGHQEVQLRFLKGHPETLEKFD-KFKLILSEDEENK-ASIDLLKKHGTVLW1	23
<input checked="" type="checkbox"/> NP_067080.1	1	---MERFEPPLLIRQSQAVNSPLEHOTVLFARLFALEPDLLPLFQyNCQGFSPECCL-SSPFELD1IRKVLW1	72
<input checked="" type="checkbox"/> NP_001369741.1	1	-----MK-ASIDLLKKHGTVLW1	18
<input checked="" type="checkbox"/> Query_73613	73	SDGLAHLNDLKGFT---FATLSELKCDKLHVDPENFRLLGNVLVCLVLAHHFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_000510.1	73	SDGLAHLNDLKGFT---FSGLSELKCDKLHVDPENFRLLGNVLVCLVLAHNFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_000175.1	73	SDGLAHLNDLKGFT---FAQLSELKCDKLHVDPENFRLLGNVLVCLVLAHNFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_000509.1	73	SDGLAHLNDLKGFT---FATLSELKCDKLHVDPENFRLLGNVLVCLVLAHNFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_005321.1	73	SDGLAHLNDLKGFT---FAKLSELKCDKLHVDPENFRLLGNVLVCLVLAHNFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_005500.2	73	SDGLAHLNDLKGFT---FAQLSELKCDKLHVDPENFRLLGNVLVCLVLAHNFGKE#TPFPQAAYQKVVAGVANALAHKYH	147
<input checked="" type="checkbox"/> NP_005359.1	68	TNAVAHDDMPMNA---SLSALSDLIAHKLHVPPNPKLSSCLLVTLAHLRPA#TVAHVASLDLKFLASVTVLTKSYR	142
<input checked="" type="checkbox"/> NP_005257062.1	90	NTVVENLHDPOKVSVLALVGKAIALKHKEPVYPTKLSGVILEVAEEFASDFPPEQRAWAKRLGLIYHVTAAYK[35]	202
<input checked="" type="checkbox"/> NP_001003938.1	67	GAAVQHVDNRRAA---SLSPLADLAI[LVRVPDANPPLL]QCPVHVLASHLQD#TVAHVASLDLKFLASVTVLTKSYR	141
<input checked="" type="checkbox"/> NP_005322.1	68	SLAVERLDDLPRA---SLSLHILACQLRVOPASPLQOLLGHCILVTLARHYPGDFSPALQASLDKFLSHVISALSEYR	142
<input checked="" type="checkbox"/> NP_059030.1	90	NTVVENLHDPOKVSVLALVGKAIALKHKEPVYPTKLSGVILEVAEEFASDFPPEQRAWAKRLGLIYHVTAAYK[23]	190
<input checked="" type="checkbox"/> NP_016879605.1	74	GGILKKKGHHEAA---IKPLAQSHATKHKIPVKYLF1SECI1QVQLQSKIPGD#DAQAGMNKALELFRKOMASNYK[6]	154
<input checked="" type="checkbox"/> NP_067080.1	73	DAAVTNVEILSLSeeyASLGLRKRA---VGVLQSLSFSTGESLLYMEKLGFA#TPTRAAMSOLYGAVQAMSRGND[2]	151
<input checked="" type="checkbox"/> NP_001369741.1	19	GGILKKKGHHEAA---IKPLAQSHATKHKIPVKYLF1SECI1QVQLQSKIPGD#GADAQGMNKALELFRKOMASNYK[6]	99



Problems with PSSMs: Positional dependencies

Do not capture positional dependencies

WEIRD	D				0.6
WEIRD	E	I			
WEIQH	H				0.4
WEIRD	I		I		
WEIRD	Q			0.4	
WEIQH	R			0.6	
WEIQH	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 $P(QH)=0.16$

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

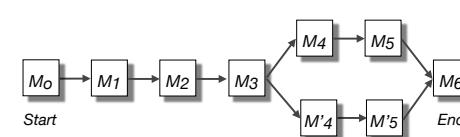
Note: We never see QD or RH, we only see RD and QH.
However, $P(RH)=0.24$, $P(QD)=0.24$, while $P(QH)=0.16$

Markov chains: Positional dependencies



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

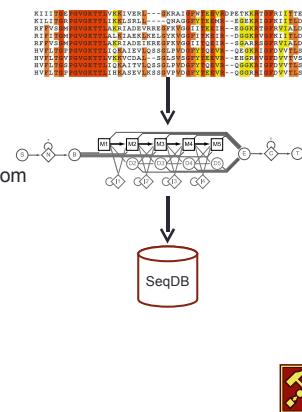
WEIRD
WEIRD
WEIQH
WEIRD
WEIQH



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

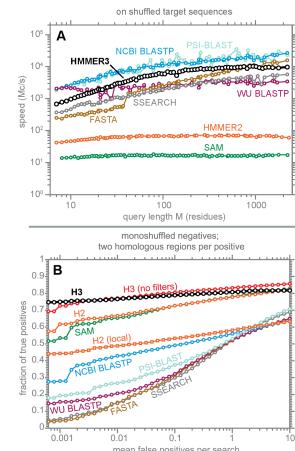
Use of HMMER

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



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



The screenshot shows the HMMER web interface. At the top, there's a navigation bar with links like Home, Search, Results, Software, Help, About, Contact, and a logo for EMBL-EBI. Below the navigation is a header for "HMMER Biosequence analysis using profile hidden Markov Models". The main area has tabs for "HMMER", "HMMER3", "HMMSEARCH", and "JACKHMMER". A form for "protein sequence vs protein sequence database" is shown, with a text input field containing a sequence and a "Submit" button. Below the input is a "Sequence Database" section with dropdown menus for "Frequently used databases" (Reference Proteomes, UniProtKB, SwissProt, PDB, Ensembl) and "Current database selection" (SwissProt). There's also a "Restrict by Taxonomy" section with checkboxes for "Taxon search" and "Pre-defined representatives", and a "Organism" input field.

The screenshot shows the search results page for "swissprot(v.2019_11)". It lists 12 significant query matches, each with a target protein name, description, species, cross-references (represented by colored circles), and E-value. The targets include HBB_HUMAN, HBD_HUMAN, HBE_HUMAN, HBG2_HUMAN, HBG1_HUMAN, HBA_HUMAN, HBAZ_HUMAN, HBAT_HUMAN, HBM_HUMAN, CYGB_HUMAN, MYG_HUMAN, and NGB_HUMAN, all from Homo sapiens. The E-values range from 0.0017 to 6.8e-99. At the bottom, there are buttons for "Local Link" and "show all alignments".

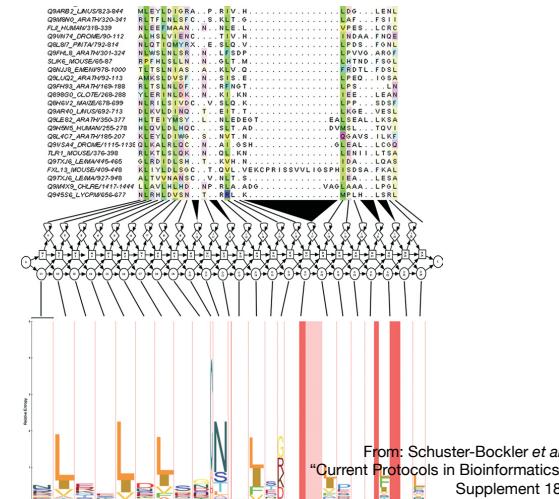
PFAM: Protein Family Database of Profile HMMs

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. HHMER software is used to perform searches.
 - **Pfam-B** contains additional protein sequences that are automatically aligned. Pfam-B serves as a useful supplement that makes the database more comprehensive.
 - Pfam-A also contains higher-level groupings of related families, known as **clans**

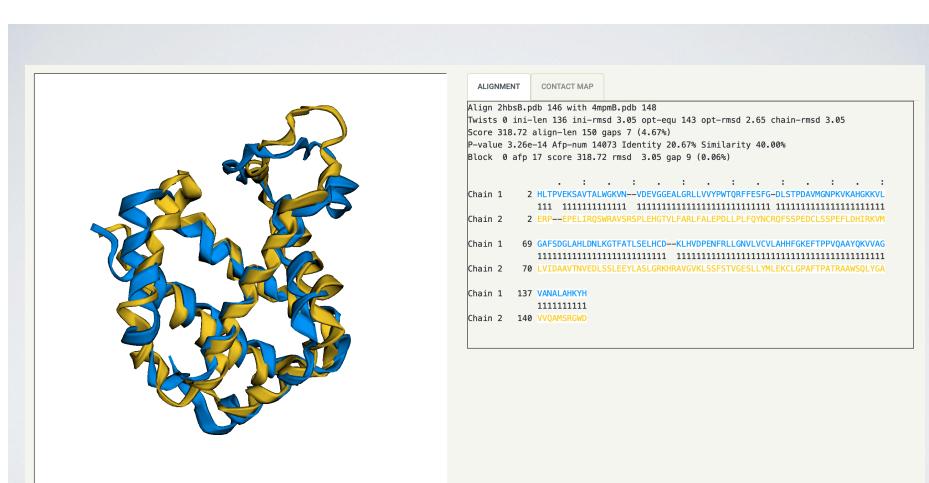


From: Schuster-Bockler et al.
“Current Protocols in Bioinformatics”
Supplement 18

YOUR TURN!

- There are **four required** and **one optional** hands-on sections including:
 1. Limits of using BLAST [~10 mins]
 2. Using PSI-BLAST [~30 mins]
 3. Examining conservation patterns [~20 mins]
— BREAK [15 mins] —
 4. [Optional] Using HMMER [~10 mins]
 5. Divergence of protein sequence and structure [~25 mins]

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



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!