



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 <u>example report</u> for format and content guidance.
- Your responses to questions Q1-Q4 are due 12pm San Diego time on Monday of week 4 (Apr 21th, 04/21/25).
- The complete assignment, including responses to all questions, is due 12pm Monday of week 10 (Jun 2nd, 06/02/25).

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 (Apr 21th, 04/21/25).
- The complete assignment, including responses to all questions, is due 12pm Monday of week 10 (Jun 2nd, 06/02/25).

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 it's function is known.

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

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

Also include the output of that BLAST search in your document. If appropriate, change the forth Cocratier size in 19 of that the resultance displayed nearly. You can also scene on giture a BLAST output (e.g. alt print screen on a PC or on a MAC press X-shift-1. The pointer becomes a buils eye. Select he area you wish to capture and researe. The image is saved as a fisc 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 liabeled 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 appointed

[03] Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your ELAST results from [02] as FSETs results from [02] as FSETs result post in Recessary or translate your novel DNA sequence using a local called EMBOSS Transeq at the ESI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop code. It may not star with a methionine if you don't have the complete coding region. Make sure the sequence your provide includes a headersubject line and is in traditional FAST for small.

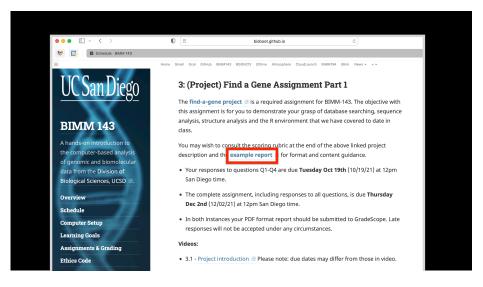
Here, sell me the name of the novel protein, and the species from which it derives. It is very unlikely plus till definitely possible that you will find a novel gene from an organism such as S. cerevisiae, human or mouse, because those genomes have already been throroughly amottated. It is more likely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

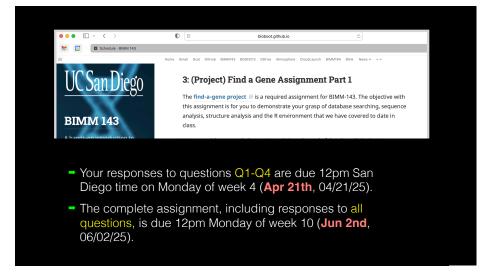
[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

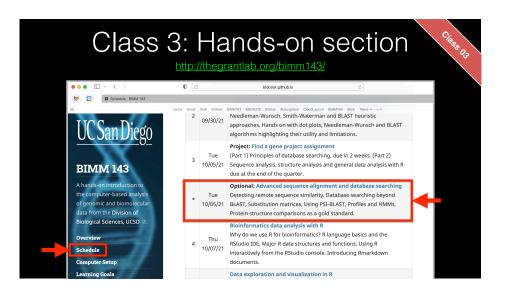
- [Los]), and use in as a query in a loasy securior unit in inductases at involve. If there is a match with 100% amino acid identity to a protein in the database, from the same spacies, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already 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
- 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.

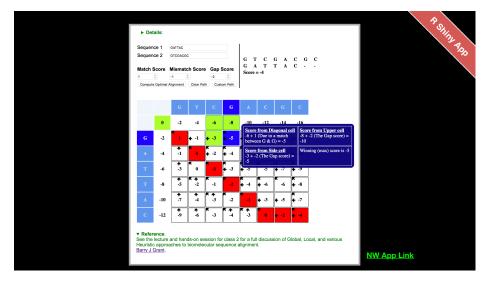
(IOS) Christata a multiple exquence slignment with your post protein, your original lower protein, and group of when remishes at this tentile you offered speakes. A pripal remisher of proving the control of the protein speakes and proving the protein of the resignment purpose is a millimm of S and a maximum of 20 - all founds the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier fort with a size appropriate to fit page width.

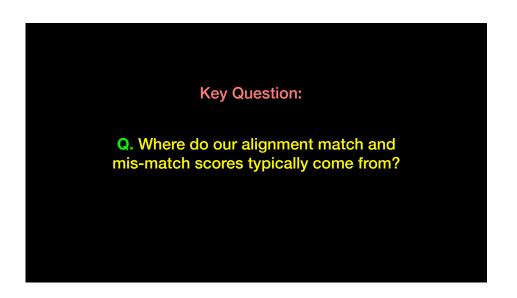
Give-note indicate your sequence in the alignment by choosing an appropriate name or each sequence in the input uniqued sequence life (e. det the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below).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

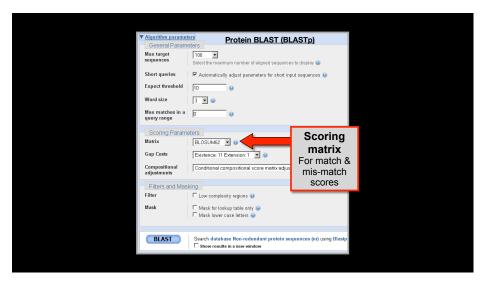




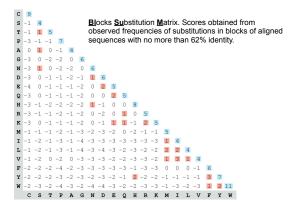




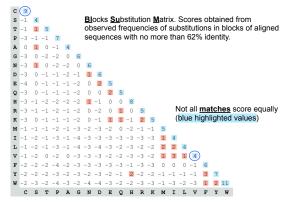




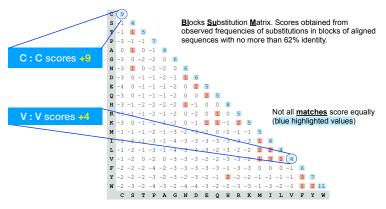
By default BLASTp match scores come from the BLOSUM62 matrix



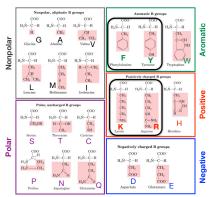
By default BLASTp match scores come from the BLOSUM62 matrix



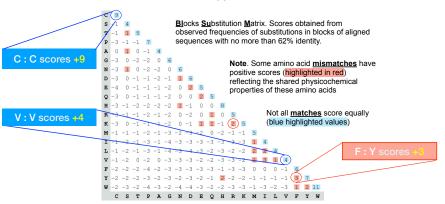
By default BLASTp match scores come from the BLOSUM62 matrix



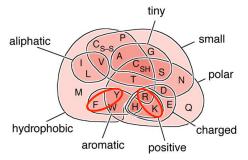
Protein scoring matrices reflect the properties of amino acids



By default BLASTp match scores come from the BLOSUM62 matrix



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]
	DDEAK FIE : 3	

— 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 <u>discussion</u> at your **Table** and on **Piazza**!

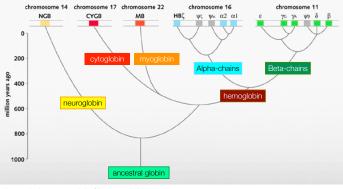
YOUR TURN!

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

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**!

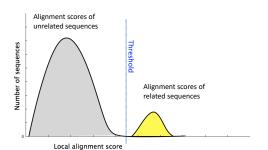


An evolutionary model of human globins.

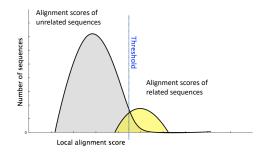
The different locations of globin genes in human chromosomes are reported at the top of the figure, distinguishing between the functional genes (in color) and the pseudogenes (in grey).

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

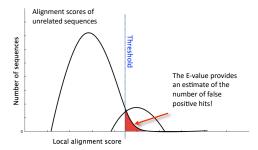
• Ideally, a threshold separates all query related sequences (yellow) from all unrelated sequences (gray)



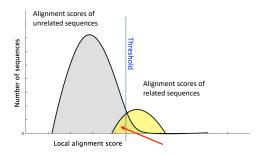
- Unfortunately, often both score distributions overlap
- The E value describes the expected number of hits with a score above the threshold if the query and database are unrelated



- 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…



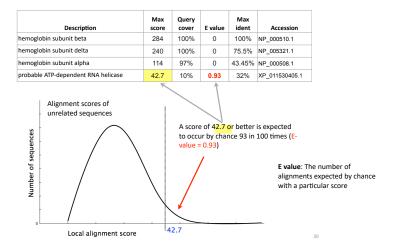
YOUR TURN!

 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 <u>discussion</u> at your **Table** and on **Piazza**!

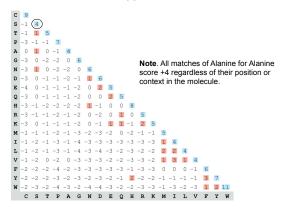


Recall: BLOUSM62 does not take the local context of a particular position into account

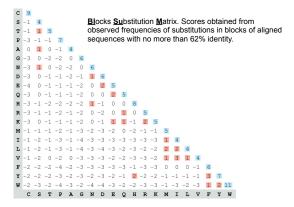
(i.e. all like substitutions are scored the same regardless of their location in the molecules).



By default BLASTp match scores come from the BLOSUM62 matrix



By default BLASTp match scores come from the BLOSUM62 matrix



PSI-BLAST: Position specific iterated BLAST

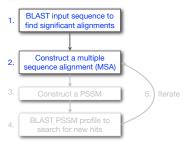
 The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a <u>scoring matrix that is</u> <u>customized to your query</u>

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 <u>scoring matrix that is</u> <u>customized to your query</u>
- 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

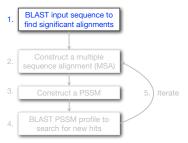
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



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

PSI-BLAST: Position-Specific Iterated BLAST

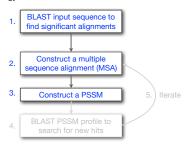
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



(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)



Example: Computing a transcription factor bind site PSSM

CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCGGATA
CCCATTTCGAAAA
CCTATTTAGTATA
CCAAATTAGGAAA
CCAAATTAGGAAA
CCAAATTAGGAAA
CCAAATTTGGCAAA
TCTATTTTGGAAA

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

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

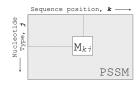
We will build a 13 x 4 **PSSM** (*k*=13, *j*=4).

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).



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

 \mathbf{M}_{kj} score for the *j*th nucleotide at position *k* \mathbf{p}_{kj} probability of nucleotide *j* at position *k*

 $\mathbf{p}_{\mathfrak{I}}$ "background" probability of nucleotide j

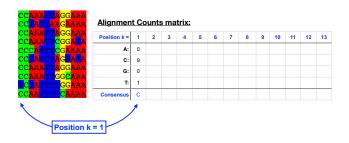
See Gibskov et al. (1987) PNAS 84, 4355

Computing a transcription factor bind site PSSM

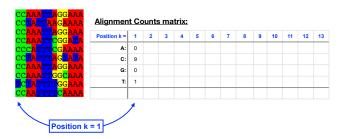
CCAAA <mark>TT</mark> A <mark>GG</mark> AAA CC <mark>T</mark> A <mark>TT</mark> AA <mark>G</mark> AAAA	First we	e wi	ll bui	ld an	alig	nmei	nt Co	ounts	mat	trix				
CCAAATTAGGAAA CCAAATTCGGATA	Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
CCCATTTCGAAAA	A:													
CCTATTTAGTATA	C:													
CCAAATTAGGAAA CCAAATTGGCAAA	G:													
TCTATTTTGGAAA	T:													
CC <mark>AA</mark> TTTT <mark>C</mark> AAAA														

Computing a transcription factor bind site PSSM

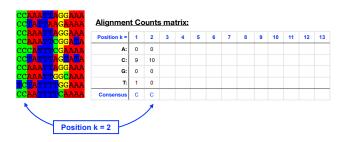
Computing a transcription factor bind site PSSM



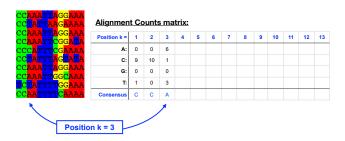
Computing a transcription factor bind site PSSM



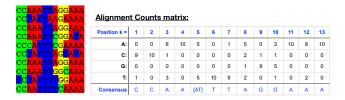
Computing a transcription factor bind site PSSM



Computing a transcription factor bind site PSSM



Computing a transcription factor bind site PSSM

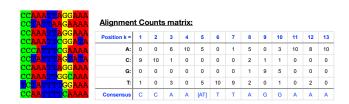


	Average	Profil	<u>e</u> (Fre	quen	cy) ma	atrix:								
Often we will not communicate with	Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
the count matrix	A:	0	0	0.6	1	0.5	0	0.1	0.5	0	0.3	1	0.8	1
but rather the	C:	0.9	1	0.1	0	0	0	0	0.2	0.1	0.1	0	0	0
derived average	G:	0	0	0	0	0	0	0	0.1	0.9	0.5	0	0	0
profile (a.k.a. frequency matrix).	T:	0.1	0	0.3	0	0.5	1	0.9	0.2	0	0.1	0	0.2	0
	Consensus	С	С	Α	Α	[AT]	T	Т	Α	G	G	Α	Α	Α

Computing a transcription factor bind site PSSM

CAAA <mark>TTAGG</mark> AAA C <mark>TATTAAGA</mark> AAA	Alignmen	t Co	ounts	s ma	trix:									
CAAA <mark>TTA</mark> GGAAA CAAATTCGGATA	Position k =	1	2	3	4	5	6	7	8	9	10	11	12	13
CCATTTCGAAAA	A:	0	0	6	10	5	0	1	5	0	3	10	8	10
C <mark>TATTTAGT</mark> ATA	C:	9	10	1	0	0	0	0	2	1	1	0	0	0
CAAA <mark>TTA</mark> GGAAA CAAATTGGCAAA	G:	0	0	0	0	0	0	0	1	9	5	0	0	0
CTATTTTGGAAA	T:	1	0	3	0	5	10	9	2	0	1	0	2	0
CAA <mark>TTTT</mark> CAAAA	Consensus	С	С	Α	Α	[AT]	Т	Т	Α	G	G	Α	Α	Α

Computing a transcription factor bind site PSSM



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

- C_{kj} Number of jth type nucleotide at position k
- Z Total number of aligned sequences
- $\mathbf{p}_{\mathtt{j}}$ "background" probability of nucleotide j
- $\mathbf{p}_{k\uparrow}$ probability of nucleotide j at position k

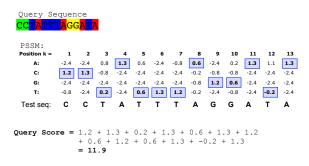
$$M_{kj} = \log\left(\frac{p_{kj}}{p_j}\right) \quad p_{kj} = \frac{C_{kj} + p_j}{Z + 1}$$

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

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

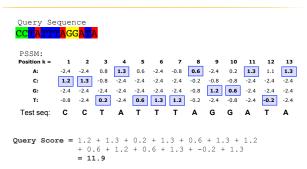
Computing a transcription factor bind site PSSM...

Scoring a test sequence

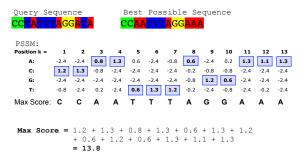


Q. Does the query sequence match the DNA sequence profile?

Scoring a test sequence



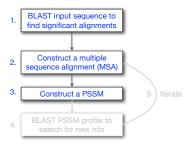
Scoring a test sequence...



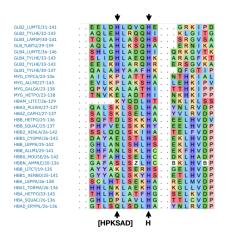
A. Following method in Harbison *et al.* (2004) Nature 431:99-104 Heuristic threshold for match = 60% x Max Score = (0.6 x 13.8 = 8.28); 11.9 > 8.28; Therefore our query is a potential TFBS!

PSI-BLAST: Position-Specific Iterated BLAST

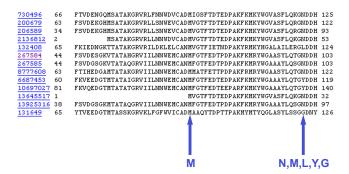
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST

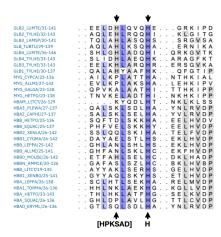


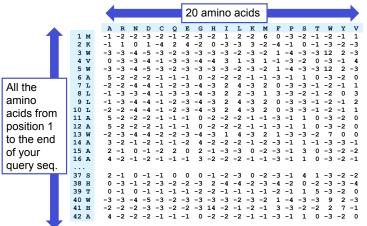
(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)

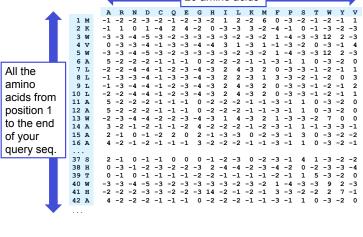


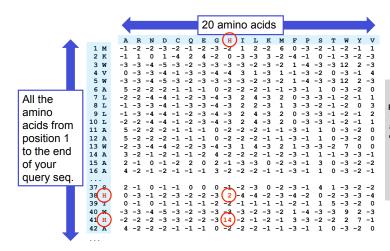
Inspect the blastp output to identify empirical "rules" regarding amino acids tolerated at each position







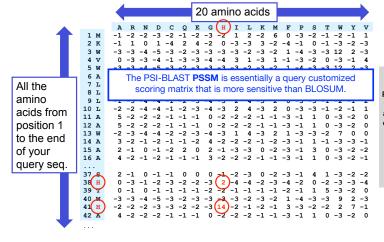




Key Point: PSSM "scores" differ for the same aminoacid type depending on where it is in your

- i.e. sores are POSITION DEPENDENT

protein sequence

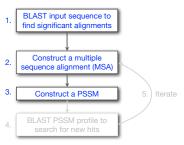


Key Point: PSSM "scores" differ for the same aminoacid type depending on where it is in your protein sequence

- i.e. sores are POSITION DEPENDENT!

PSI-BLAST: Position-Specific Iterated BLAST

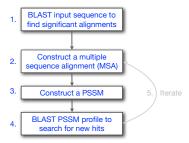
Many proteins in a database are too distantly related to a query to be detected using standard BLAST. In many other cases matches are detected but are so distant that the inference of homology is unclear. Enter the more sensitive PSI-BLAST



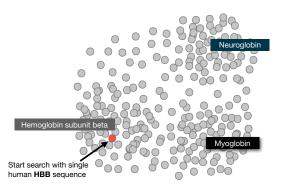
(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)

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

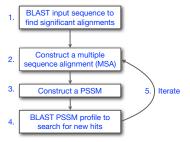


(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)

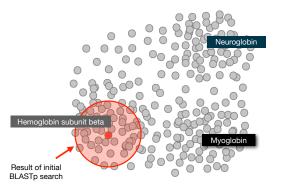


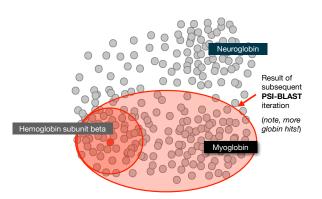
PSI-BLAST: Position-Specific Iterated BLAST

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

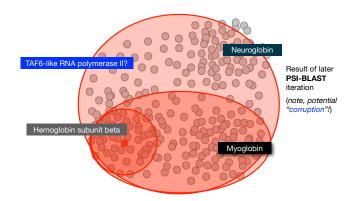


(see Altschul et al., Nuc. Acids Res. (1997) 25:3389-3402)





Description Max Total Cluery E Cover Value Ident Accession	Score Scor							
hemoglobin subunit delta [Homo sapiens] 284 284 100% 7e-100 93% NF_000510.1	hemoglobin subunit delta [Homo sapiens] 284 284 100% 7e-100 93% NF_000510.1 hemoglobin subunit appation [Homo sapiens] 240 240 100% 2e-82 76% NF_000521.1 hemoglobin subunit apmma-2 [Homo sapiens] 235 235 100% 2e-80 73% NF_000175.1 hemoglobin subunit apma-1 [Homo sapiens] 232 232 100% 3e-79 73% NF_00050.2 hemoglobin subunit alpha [Homo sapiens] 114 114 97% 7e-33 43% NF_00050.2	Description					Ident	Accession
hemoglobin subunit agesion (Home sapiens) 240 240 100% 2e-82 76% NE_003231.1 hemoglobin subunit gamma-2 (Home sapiens) 235 235 100% 2e-80 73% NE_000175.1 hemoglobin subunit gamma-1 (Home sapiens) 232 232 100% 3e-79 73% NE_00050.2 hemoglobin subunit alpha (Home sapiens) 114 114 97% 7e-33 43% NE_000508.1	hemoglobin subunit apsilon [Homo sapiens] 240 240 100 2e-82 76% NE_00321_1 hemoglobin subunit gamma-2 [Homo sapiens] 235 235 100 2e-80 73% NE_000175_1 hemoglobin subunit gamma-1 [Homo sapiens] 232 232 100 3e-79 73% NE_00050_2 hemoglobin subunit alpha [Homo sapiens] 114 114 97% 7e-33 43% NE_00050_81	hemoglobin subunit beta [Homo sapiens]	301	301	100%	2e-106	100%	NP_000509.1
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_00058.1	hemoglobin subunit gamma-2 (Homo sapiens) 235 236 100% 2e-80 73% NP_000175.1 hemoglobin subunit gamma-1 (Homo sapiens) 232 232 100% 3e-79 73% NP_000559.2 hemoglobin subunit alpha (Homo sapiens) 114 114 97% 7e-33 43% NP_000581.1	hemoglobin subunit delta [Homo saplens]	284	284	100%	7e-100	93%	NP_000510.1
hemoglobin subunit gamma-[Homo sapiens] 232 232 100% 3e-79 73% NP_000559_2 114 114 97% 7e-33 43% NP_000589_1 114 114 114 116 1	hemoglobin subunit gamma-[Homo sapiens] 232 232 100% 3e-79 73% NP_000559_2 114 114 97% 7e-33 43% NP_000589_1 114 114 114 116 1	hemoglobin subunit epsilon [Homo sapiens]	240	240	100%	2e-82	76%	NP_005321.1
hemoglobin subunit alpha [Homo sapiens] 114 114 97% 7e-33 43% NP_000508.1	hemoglobin subunit alpha [Homo sapiens] 114 114 97% 7e-33 43% NP_000508.1	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 zeta [Homo sapiens] 100 100 97% 3e-27 36% NP_005323.1	hemoglobin subunit zeta [Homo sapiens] 100 100 97% 3e-27 36% NP_005323.1	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



			cover	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

New relevant globins found only by PSI-BLAST

Description	Max	Total score	Query	E value	Ident	Accession	
	score	score	cover	vaiue			
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	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	2
myoglobin [Homo sapiens]	159	159	97%	3e-50	26%	NP 005359.1	
hemoglobin subunit alpha [Homo sapiens]	151	151	97%			NP_000508.1	
hemoglobin subunit mu [Homo sapiens]	147	147	97%			NP 001003938.1	
hemoglobin subunit theta-1 [Homo sapiens]	147	147	97%			NP 005322.1	
neuroglobin [Homo sapiens]	134	134	92%		23%	NP_067080.1	3
PREDICTED: cytoglobin isoform X2 [Homo sapiens]	115	115	66%	3e-33	25%	XP_016879605.1	
PREDICTED: microtubule cross-linking factor 1 isoform X1 [Homo sapir	46.3	46.3	27%	7e-06	39%	XP_011523942.1	٦,
PREDICTED: microtubule cross-linking factor 1 isoform X4 [Homo sapid	46.3	46.3	27%	7e-06	39%	XP_005258156.1	?
Inclusion of irrelevan	t hits	can	lead	to PS	SM c	orruption	
and addition of interest						7.101	

 Query_73613	1	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF
✓NP_000510.1	1	MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFE-SFGDLSSPDAVM-GNPKVKAHGKKVLGAF
✓NP_000175.1	1	MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL
✓NP_000509.1	1	MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDAVM-GNPKVKAHGKKVLGAF
✓NP_005321.1	1	MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFD-SFGNLSSPSAIL-GNPKVKAHGKKVLTSF
✓NP_000550.2	1	MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASAIM-GNPKVKAHGKKVLTSL
✓NP_005323.1	1	-MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFP-HFDLHpGSAQLRAHGSKVVAAV
✓NP_000508.1	1	-MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFP-HFDLShGSAQVKGHGKKVADAL
✓XP_005257062.1	1	[15]SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL
✓NP_001003938.1	1	MLSAQERAQIAQVWDLIAGHEAQFGAELLLRLFTVYPSTKVYFP-HLSACQ-DATQLLSHGQRMLAAV
✓NP_005322.1	1	-MALSAEDRALVRALWKKLGSNVGVYTTEALERTFLAFPATKTYFS-HLDLSpGSSQVRAHGQKVADAL
✓NP_599030.1	1	[15]SEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLEME-RSPQLRKHACRVMGAL
✓XP_016879605.1	1	MEDPLEME-RSPQLRKHACRVMGAL
✓NP_001349775.1	1	-MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFKHLKSEDEMK-ASEDLKKHGATVLTAL
		MERPEPELIROSWRAVSRSPLEHGTVLFARLFALEPDLLPLFOVNCROFSSPEDCL-SSPEFLDHIRKVMLVI
✓NP_067080.1	1	makerelitikgswkkvakarlandivirkkrikterulirirgynckgraafeuch-aaferlunikkvmlvi
✓NP_067080.1 ✓NP_001369741.1	1	MK-ASEDLKKHGATVLTAL
✓NP_001369741.1	-	
	1	MK-ASEDLKKHGATVLTAL
<pre>NP_001369741.1</pre> ✓Query_73613	73	MK-ASEDLKKIGATVLTAL SDGLAHLDNLKGTFATLSELMCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
VNP_001369741.1 VQuery_73613 VNP_000510.1	1 73 73	MK-ASEDLKKIGATVLTAL SDGLAHLDNLKGYFATLSELHICDKLHVDBENFRLLGNVLVCVLAHBFCKEFTPPVQAAYQKVVAGVANALABIKYB SDGLAHLDNLKGYFSQLSELHICDKLHVDPENFRLLGNVLVCVLARNFGKEFTPQMOAAYQKVVAGVANALABIKYB
NP_001369741.1 Query_73613 NP_000510.1 NP_000175.1	1 73 73 73	MK-ASEDLKKIGATVLTAL SDGLAHLDNLKGTFATLSELICCKLINVDPENFALIGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH SDGLAHLDNLKGTFGGSELICCKLINVDPENFALIGNVLVCVLAKHYGKEFTPOQQAAYQKVVAGVANALAHKYH SDGLAHLDNLKGTFGGSELICCKLINVDPENFALIGNVLVVCVLAKHYGKEFTPVQAAYQKVVAGVANALAHKYH SDGLAHLDNLKGTFGGSELICCKLINVDPENFALIGNVVVAVIALHINGKEFTPVQAANQHVNCVAGALAHKYH
VNP_001369741.1 VQuery_73613 VNP_000510.1 VNP_000175.1 VNP_000509.1	73 73 73 73	MK-ASEDLKKIGATVLTAL SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGNVLVCVLAHHFGKEFTPPVQAAYOKVVAGVANALAHKYH SDGLAHLDNLKGTFGGSELHICDKLHVDPENFRLIGNVLVCVLARHFGKEFTPPVQAAYOKVVAGVANALAHKYH SDGLAHLDDLKGTFAQLSELHICDKLHVDPENFRLIGNVLVCVLAHHFGKEFTPFVQAANQHVTGVVAGALSKRYH SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGNVLVCVLAHHFGKEFTPFVQAAYQKVVGVANALAHKYH
<pre>VNP_001369741.1 VQuery_73613 VNP_000510.1 VNP_000175.1 VNP_000509.1 VNP_005321.1</pre>	73 73 73 73 73	SDGLAHLDNLSGT —— FATLSELRCDLINVDENFRILGNYLVCVLANIFGKEFFPVQAAVGOVVAGVAALAINTYN SDGLAHLDNLSGT —— FATLSELRCDLINVDENFRILGNYLVCVLANIFGKEFFPVQAAVGOVVAGVAALAINTYN GDAIRLDNLSGT —— FAGLSELRCDKLINVDENFRILGNYLVCVLANIFGKEFFPVQAAVGOVAGVAALAINTYN SDGLAHLDNLSGT —— FAGLSELRCDKLINVDENFRILGNYLVCVLANIFGKEFFPVQAAVGOVAGVAALAINTYN SDGLAHLDNLSGT —— FATLSELRCDKLINVDENFRILGNYLVCVLANIFGKEFFPVQAAVGOVAGVAALAINTYN SDGLANIFGKEFFPVQAAVGOVAGVAALAINTYN
<pre>NP_001369741.1 Query_73613 NP_000510.1 NP_000175.1 NP_000509.1 NP_005321.1 NP_000550.2</pre>	73 73 73 73 73 73	MK-ASEDLKKIGATVLTAL SDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVARLAHKYH SDGLAHLDNLKGTFQGLSELHCDKLHVDPENFRLIGHVLVCVLARHFGKEFTPCMQAAYQKVVAGVARLAHKYH SDGLAHLDNLKGTFQGLSELHCDKLHVDPENFRLIGHVLVVLALHFIGKEFTPEVQAANQVKVVAGVARLAHKYH SDGLAHLDNLKGTFALSELHCDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAANQVKVAGVARLAHKYH GDAIKHDNLKFAFAKLSELHCDKLHVDPENFRLIGHVVLVVLAHFGKEFTPEVQAANQKLVSAVATLAHKYH GDAIKHDNLKFAFALGSELHCDKLHVDPENFRLIGHVVLVTLAHFGKEFTPEVQAANQKLVSAVATLAHKYH GDAYKHLDDLGKTFAGLSELHCDKLHVDPENFRLIGHVVLVTLAHFGKEFTPEVGAANQKLVSAVATLAHKYH GDAYKHLDGLGKTFAGLSELHCDKLHVDPENFRLIGHVVVTLAHFGKEFTPEVGAANQKLVSAVATLAHKYH
<pre>NP_001369741.1 Query_73613 NP_000510.1 NP_000175.1 NP_000509.1 NP_005321.1 NP_0005323.1</pre>	73 73 73 73 73 73 73 68	SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFAGLSELHICDKLHVDPENFRLIGHVLVVCVLAHHFGKEFTPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVLAHHFGKEFTPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVCVLAHHFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHICDKLHVDPENFRLIGHVLYVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHICDKLHVDPENFRLIGHVLYVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHIGHTLHVDPVNFRLISHCLLVTVLAHAFPADFTAARAHANDFTLSVVSSVLTEKTR THIVANITUODMARLEAKSGLHAHKHVDPVNFRLISHCLLVTLAHAFPADFTAARAHANDFTLSVVSSVLTEKTR
VNP_001369741.1 VQuery_73613 VNP_000510.1 VNP_000175.1 VNP_005321.1 VNP_005323.1 VNP_005323.1	73 73 73 73 73 73 73 68	SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFAGLSELHICDKLHVDPENFRLIGHVLVVCVLAHHFGKEFTPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVLAHHFGKEFTPVQAAYQKVVAGVAHALAHKYB SDGLAHLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVCVLAHHFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDNLKGTFATLSELHICDKLHVDPENFRLIGHVLVVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHICDKLHVDPENFRLIGHVLYVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHICDKLHVDPENFRLIGHVLYVTVLAHTFGKEFTPVQAAYQKUSAAYATALAHKYB GDATHBLDDLKGTFAGLSELHIGHTLHVDPVNFRLISHCLLVTVLAHAFPADFTAARAHANDFTLSVVSSVLTEKTR THIVANITUODMARLEAKSGLHAHKHVDPVNFRLISHCLLVTLAHAFPADFTAARAHANDFTLSVVSSVLTEKTR
ONP_001369741.1 Ouery_73613 ONP_000510.1 ONP_000510.1 ONP_005321.1 ONP_005522.1 ONP_005523.1 ONP_00550.2 ONP_00550.2 ONP_00550.2 ONP_00550.2 ONP_00550.2	73 73 73 73 73 73 73 68 68	MK-ASEDLKKIGATVLTAL SOCIAHLINILSGTFATLSELRCDKLRVDBENFRLIGNIVLVCVLAHIFICREFFPOVAGOVIVAGNAKALAHINTH SOCIAHLINILSGTFOLGSLRCDKLRVDBENFRLIGNIVLVCVLAHIFICREFFFOVAGOVIVAGNAKALAHINTH SOLAHLIDDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVCVLAHIFICREFFFOVAGOVIVAGNAKALAHINTH SOCIAHLIDDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVCVLAHIFIGREFFPOVAGOVIVAGNAKALAHINTH SOCIAHLIDDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVCVLAHIFIGREFFPOVAGOVIVAGNAKALAHINTH GDATRIBUDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVTVLAHIFIGREFFPOVAGOVIVAGNAKALAHINTH SOCIAHLIDDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVTVLAHIFIGREFFPOVAGOVIVAGNAVALAHINTH SOCIAHLIDDLKGTFALGSLRCDKLRVDBENFRLIGNIVLVTVLAHIFIGREFFPOVAGOVIVAGNAVALAHINTH TRIVATUPODHYMALEAKSLBLUK TURDPOVENKLIGNIVLVTVLAHIFIGREFFPOVAGONIVAGNAVUTEGNET TRIVATUPODHYMALEAKSLBLUK TURDPOVENKLIGNIVLVTVLAHIFIGREFFPOVANIALDRIFTANOVIVATUR TRIVATUPODHYMALEAKSLBLUK TURDPOVENKLIGNIVLVTVLAHIFIGREFFPOVANIALDRIFTANOVIVATUR TRIVATUPODHYMALEAKSLBLUK TURDPOVENKLIGNIVLVTVLAHIFIGREFFPOVANIALDRIFTANOVIVATUR TRIVATUPODHYMA
Query_73613 Query_73613 NP_000510.1 NP_000510.1 NP_00509.1 NP_005321.1 NP_00550.2 NP_005323.1 NP_00508.1 XP_00508.1 XP_0050338.1	73 73 73 73 73 73 73 68 68 90 67	MK-ASEDLKKIGATVLTAL SOLIAHLINIKSOT — PATLSELIKONLAVDENFRILGINILVOVIANIFOKEFFPPQAAVGVVAGVARALAHINY SOLIAHLINIKSOT — PROLSELIKONLAVDENFRILGINILVOVIANIFOKEFFPPQAAVGVVAGVARALAHINY SOLIAHLINIKSOT — PROLSELIKONLAVDENFRILGINILVOVIANIFOKEFFPQAAVGVVAGVARALAHINY SOLIAHLINIKSOT — PRALSELIKONLAVDENFRILGINILVOVIANIFOKEFFPQAAVGVVAGVARALAHINY SOLIAHLINIKSOLIAHLINIKSOM — PRALSELIKONLAVPERRILGINILVOVIANIFOKEFFPQAAVGVAGVARALAHINY SOLIAHLINIKSOLIAHLINIKSOM — PRALSELIKONLAVPERRILGINILVOVIANIFOKEFPPQAAVGVAGVARALAHINT SOLIAHLINIKSOLIAHLINIKSOM — PRALSELIKONLAVPERRILGINILVOVIANIFOKEFPPAAVGARALGINITY TIRAVANDOMPAR — LEAKSELIKONLAVEPPARILGINILVOVIANIFOKARPANANDIFIKSIVATIVATI TIRAVANDOMPAR — LEAKSELIKONLAVEPPARILGINILVOVIANIFOKARPATORANIKSISTINITY TIRAVANDOMPAR — LEAKSELIKONLAVEPPARILGINILVOVIANIFOKARPATORANIKATION SINTANIKSI GAAVGUVONIANA — LEAKSELIKONLAVEPPARILGINILVOVIANIFOKARPATORANIKATION SINTANIKSI GAAVGUVONIANA — LEAKSELIKONLAVEPPARILGINILVOVIANIFOKARPATORANIKATION SINTANIKSI TANIAVONIANIKANIKANIKANIKANDOMPARILGINILVOVIANIFOKARPATORANIKATION SINTANIKSI TANIAVONIANIKANIKANIKANIKANIKANIKANIKANIKANIKAN
Query_73613 NP_000510.1 NP_000175.1 NP_00509.1 NP_005321.1 NP_005323.1 NP_005233.1 NP_005257062.1 NP_005322.1 NP_005323.1	73 73 73 73 73 73 73 68 68 90 67 68	SDGLAHLONLROT FATLSELRICDKLRYDPENFILIGNIVLVCVLAHHFGKEFTPPVQAAVGKVVAGVARLAHKYH SDGLAHLONLROT FAGLESELROCKLRYDPENFILIGNIVLVCVLAHHFGKEFTPPVQAAVGKVVAGVARLAHKYH SDGLAHLONLROT FAGLESELROCKLRYDPENFILIGNIVLVCVLAHHFGKEFTPPVQAAVGKVVAGVARLAHKYH SDGLAHLONLROT FAGLESELROCKLRYDPENFILIGNIVLVCVLAHHFGKEFTPPVQAAVGKVVAGVARLAHKYH SDGLAHLONLROT FAGLESELROCKLRYDPENFILIGNIVLVCVLAHHFGKEFTPPVQAAVGKVVAGVARLAHKYH GDATKHLONLROT FAGLESELROCKLRYDPENFILIGNIVLVCVLAHTPGKEFTPEVQAAVGKVAGVARLAHKYH GDATKHLOLKOT FAGLESLROCKLRYDPENFILIGNIVLVTVLATHFGKEFTPEVQAAVGKVAGVARLAHKYH TANAVATUNDERIA LEALSLEJAHKLATUPDVENFILIGNIVLVTVLATHFATEFTEVAANSKOPKTAVASLSKHYH THAVANVONDERIA LEALSLEJAHKLATUPDVENFILIGNIVLVTAARAFDAFFFEVYGAANAKOKTAVUTTEKT THAVANVONDERIA LEALSLEJAHKLATUPDVENFILIGNIVLTVAATRA GARDPATAVATA (3 GAAVQUTOBLAAA LEALSLEJAHKLATUPDVENFILIGNIVLTVAATRA (3 GAAVQUTOBLAAA LEALSLEJAHKLATUPDVENFILIGNIVLTVAATRA (3 GAAVQUTOBLAAA LEALSLEJAHKANKATAVATAVATA (3 GAAVQUTOBLAAA LEALSLEJAHKANKATAVATAVATA (3 GAAVQUTOBLAAA LEALSLEJAHKANKATAVATAVATA (3 GAAVQUTOBLAAA LEALSLEJAKANKATAVATAVATA (3 GAAVGTOBLAAA LEALSLEJAKANKATAVATAVATA (3 GAAVGTOBLAAA LEALSLEJAKANKATAVATAVATATA (3 GAAVGTOBLAAA LEALSLEJAKANKATAVATAVATA (3 GAAVATAVATAVATAVATAVATAVATAVATAVATAVATAV
ONP_001369741.1 Ouery_73613 ONP_000510.1 ONP_000510.1 ONP_00510.1 ONP_00510.1 ONP_00510.1 ONP_005321.1 ONP_005323.1 ONP_005323.1 ONP_005323.1 ONP_005323.1 ONP_005323.1 ONP_001003338.1 ONP_001003338.1 ONP_005322.1 ONP_005322.1 ONP_005322.1	73 73 73 73 73 73 68 68 90 67 68 90	MK-ASEDLKRIGATVLTAL SOCIAHLINILKOT — PATLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOVVAGVARIALBIRVE SOCIAHLINILKOT — PATLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOVVAGVARIALBIRVE SOCIAHLINILKOT — PAGLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOVVAGVARIALBIRVE GDALHILDILKOT — PATLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOVVAGVARIALBIRVE GDALHILDILKOT — PATLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOVVAGVARIALBIRVE GDANKHILDILKOT — PARLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFFPPQAAVGOLIANAVIALBIRVE GDANKHILDILKOT — PAGLSELHCINLHVIDENIFRILGIVILVOLIAHIFOKEFPPPQAAVGOLIANAVIALBIRVE THAVANIVODROPA — LEGALSELHGINLHVIDENIFRILGIVILVOLIAHIFOKEFPPPQAAVGOLIANAVIALBIRVE THAVANIVODROPA — LEGALSELHGINLHVIDENIFRILGIVILVOLIAHIFOKEFPPPRAVIALGILGIVILANIFOKEFP THAVANIVODROPA — LEGALSELHGINLHVINPORVATILGIVILLOVALBAR DORFOTOKANIKAGILLISHVAAKI SI CAAVGUVOLIABAR — LEGALSELHGINLHVINPORVATILGIVILVOLIANAVIA GORFOTAMARIGALISHVAAKI SI CAAVGUVOLIABAR — LEGALSELHGINLHVINPORVATILGIVILVOLIANIFOKEFPPCAMARIGALISH SALVESTE NIVVEHLEIDPONVASILANIVARIALKINFOTOVITETTI SALVESTE NIVVEHLEIDPONVASILANIVARIALKINFOTOVITETTI SALVESTE NIVVEHLEIDPONVASILANIVARIALKINFOTOVITETTI SEVITAKI SI NIVVEHLEIDPONVASILANIVARIALKINKOTOVITETTI SEVITAKI SI NIVVEHLEDPONVASILANIVARIALKINGOTOVITETTI SEVITAKI SI NIVVEHLEDPONVANIN
ONP_001369741.1 Ouery_73613 ONP_000510.1 ONP_000175.1 ONP_000509.1 ONP_005502.2 ONP_005221.1 ONP_005508.1 ONP_005233.1 ONP_005233.1 ONP_005223.1 ONP_005223.1 ONP_00525062.1 ONP_00525062.1	73 73 73 73 73 73 73 68 68 90 67 68 90 25	SDGLAHLDNLRGTFATLSELHCDKLRVDPENFILLGRVLVCVLAHIFGKEFTPFVQAAYQKVVAGVANALAHRYH SDGLAHLDNLRGTFATLSELHCDKLRVDPENFILLGRVLVCVLAHIFGKEFTPFVQAAYQKVVAGVANALAHRYH SDGLAHLDNLRGTFAGLSELHCDKLRVDPENFILLGRVLVVCVLAHIFGKEFTPFVQAAYQKVVAGVANALAHRYH SDGLAHLDNLRGTFAGLSELHCDKLRVDPENFILLGRVLVVLAHIFGKEFTPFVQAAYQKVVAGVANALAHRYH SDGLAHLDNLRGTFALSELHCDKLRVDPENFILLGRVLVVLAHIFGKEFTPFVQAAYQKVAGVANALAHRYH CDATRHLDNLRGTFALSELHCDKLRVDPENFILLGRVLVVLAHIFGKEFTPFVQAAYQKVRAVAALAHRYH SDGLAHLDDLKGTFALSELHCDKLRVDPENFILLGRVLVVLAHIFGKEFTPFVQAARQKURSAVSVUTAKARAHRYH SDATRHLDDLKGTFALSELHCDKLRVDPENFILLGRVLVVLAHIFGKEFTPFVQAARQKURSAVSVUTEKTR THAVAHVDDHPHALGALSDLHAHKRAVDPVHFILLGRVLUTAARFPADFPFPGAARAKHAGHLSVSVUTEKTR THAVAHVDDHPHALGALSDLHAHKRAVDFVTFTKISGVILEVAREFADFPPFTQAARAKHAGHLSVSTVLTETR RVVVEHENDPRVSSPLALWGRAHAKRINGFVTFTKISGVILEVAREFADFPPFTQAARAKHAGHLSVTVATEKTR RVVSHENDPRVSSPLALWGRAHAKRINGFVTFTKISGVILEVAREFADFPPFTQAARAKHAGHLSVAVVILETKR

YOUR TURN!

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

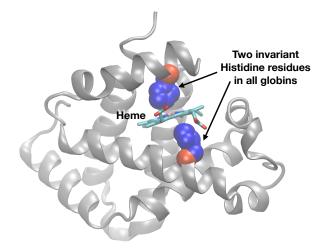
Limits of using BLAST [~10 mins]
 Using PSI-BLAST [~30 mins]
 Examining conservation patterns [~20 mins]

— 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 <u>discussion</u> at your **Table** and on **Piazza**!





Problems with PSSMs: Positional dependencies

Do not capture positional dependencies

WEIRD WEIRD WEIQH WEIRD WEIQH

D					0.6
Е		1			
Н					0.4
Τ			ı		
Q R				0.4	
				0.6	
W	_				

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

YOUR TURN!

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

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

— BREAK [15 mins]—

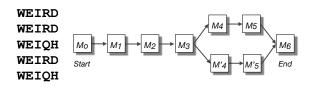
4. [Optional] Using HMMER [~10 mins]
5. Divergence of protein sequence and structure [~25 mins]

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

Markov chains: Positional dependencies



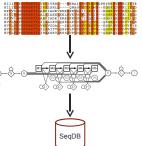
The connectivity or **topology** of a Markov chain can easily be designed to capture dependencies and variable length motifs.



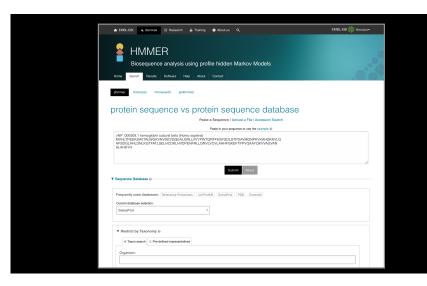
Recall that a PSSM for this motif would give the sequences **WEIRD** and **WEIRH** equally good scores even though the **RH** and **QR** combinations were not observed

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

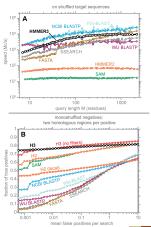






HMMER vs BLAST

	HMMER	BLAST		
Progra m	PHMMER	BIASTP		
Quer y	Single sequence			
Targe t Databas e	Sequenc e databas e			
Progra m	HM M SCA N	RP SB LA S T		
Quer y	Single sequence			
Target Databas e	Profile HMM database, e.g. Pfam	PSSM database e.g. CDD		
Progra m	HM M SEARCH	PSI-BLAST		
Quer y	Profile HMM	PSS M		
Targe t Databas e	Sequenc e databas e			
Progra m	JA CKHM MER	PSI-BLAS T		
Quer y	Single sequence			
Farge t Databas e	Sequence database			



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



Significant Query Matches (12) in swissport (v.2018, 11)							
Sigili	Customate Customate						
	Target	Description	Species	Cross-references	E-value		
>	HBB_HUMANø	Hemoglobin subunit beta	Homo sapienst₽		6.8e-99		
>	HBD_HUMAN⊌	Hemoglobin subunit delta	Homo sapiensı#	m 8 m h 8 0 h	1.6e-91		
>	HBE_HUMANø	Hemoglobin subunit epsilon	Homo sapiensi9	3 (1) (1) (2) (3) (3) (1)	1.5e-74		
>	HBG2_HUMANø	Hemoglobin subunit gamma-2	Homo sapienst9	m (2 (1) (1) (1) (1)	8.8e-73		
>	HBG1_HUMAN®	Hemoglobin subunit gamma-1	Homo sapiensi₽	m 8 m h s 0 h	6.2e-72		
>	HBA_HUMANØ	Hemoglobin subunit alpha	Homo sapiensı#	m (3 (11 (12 (13 (13 (13 (13 (13 (13 (13 (13 (13 (13	3.8e-29		
>	HBAZ_HUMAN₽	Hemoglobin subunit zeta	Homo sapiensr#	m 8 m h % () h	4.5e-23		
>	HBAT_HUMANØ	Hemoglobin subunit theta-1	Homo sapiensi#	m 8 m 8 0 m	5.2e-22		
>	HBM_HUMANø	Hemoglobin subunit mu	Homo sapienst#	m (2) (2) (b)	3.4e-19		
>	CYGB_HUMAN@	Cytoglobin	Homo sapiensı9	3 (B) (B) (B) (C) (B)	3.1e-14		
>	MYG_HUMAN@	Myoglobin	Homo sapiensø	m 8 m b 30 () h	2.3e-06		
>	NGB_HUMAN₽	Neuroglobin	Homo sapiensı#	m 8 m h % () h	0.0017		
(show	(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

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]

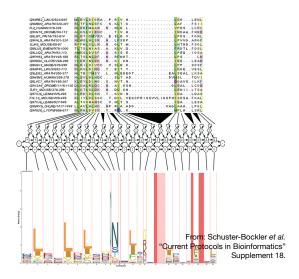
3. Examining conservation patterns [~20 mins]

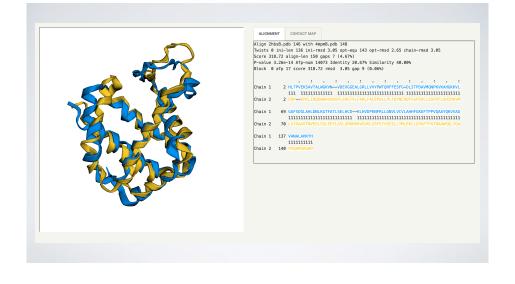
- BREAK [15 mins]-

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 <u>discussion</u> 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!