Finding data

HMMER

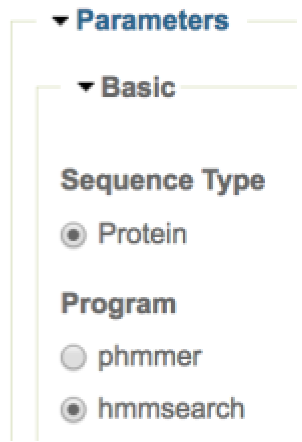
Answer key

HMMER input is prepared using VectorBase ClustalW, which runs a Java application for the graphical representation of the results. If you get an error message that blocks this application, add the URL https://www.vectorbase.org/clustalw to the Exception Site List in the Java Control Pannel to workaround this issue as explained here: https://www.java.com/en/download/faq/exception_sitelist.xml.

Contents:

1. HMMER basics

VectorBase HMMER has two programs implemented, *phmmer* and *hmmsearch*, to search with protein queries against protein databases[1].



▾ Parameters

▾ Basic

Sequence Type

● Protein

Program

○ phmmer

● hmmsearch

[1] HMMER has other functions and programs but these are not implemented in the VectorBase version.

phmmer is used to search with one or more query protein sequences in FASTA format, which makes it a BLASTp-like program. *hmmsearch* is used to search with one or more 'profiles', a multiple sequence alignment (MSA) from ClustalW is the format of the required input. The profiles are probabilistic models called "profile hidden Markov models" or profile HMMs. HMMER main characteristic is that it makes a profile of the query that assigns a position-specific scoring system for substitutions, insertions, and deletions.

Compared to BLAST, which is based on other scoring methodology, HMMER aims to be significantly more accurate and more able to detect remote homologs, because of the strength of its underlying probability models. The currently HMMER version (3.1) is as fast as BLAST for protein search. For more details about HMMER follow this link to its website (http://hmmer.org) and the most current version of its documentation.

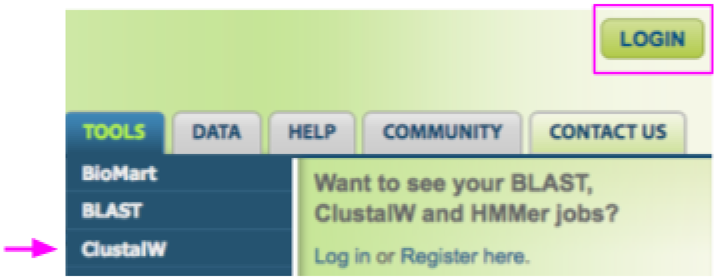
Yet BLAST continued to be the most widely used search program
HMMER User's Guide, v 3.1b2; February 2015

2. How to use these tool and interpret its output?

Quick start

Open Firefox web browser (it is necessary to see ClustalW graphical representation of the results)

Optional: Login to VectorBase. Your ClustalW and Hmmer (and BLAST) jobs will be saved and viewable in your user page.



Paste two or more protein sequences in ClustalW, https://www.vectorbase.org/clustalw

Select protein as sequence type, keep all other parameters as default and click 'Submit'

Click on 'Send to HMMER'

Select *hmmsearch*, the target dataset(s) and click 'Submit'

Output interpretation

The hits with the lowest e-value and highest scores are the best hits.

3. Questions and practice exercises

Question 3.1

What is ClustalW ?

	Answer
A tool for finding remote homologous genes	
A general purpose clustering algorithm	
A tool to align two or more sequences	X

Question 3. 2

True or False:

	True	False
ClustalW only works with protein sequences		X
ClustalW only works with nucleotide sequences		X
ClustalW will align nucleotide and protein sequences to each other		X
ClustalW will align nucleotide or protein sequences but not both at the same time	X	

Question 3.3

Open Firefox web browser (it is necessary to see ClustalW graphical representation of the results)

A sample file with *Anopheles gambiae* sequences for long, short, ultraviolet, Rh7-like and pteropsins is provided in the tutorial page

`VectorBase_HMMER_SampleFile_December2016.txt`

Upload it or copy and paste these sequences into ClustalW (Tools menu)
Select Protein as the sequence type, leave other parameters as the defaults and click 'Submit'

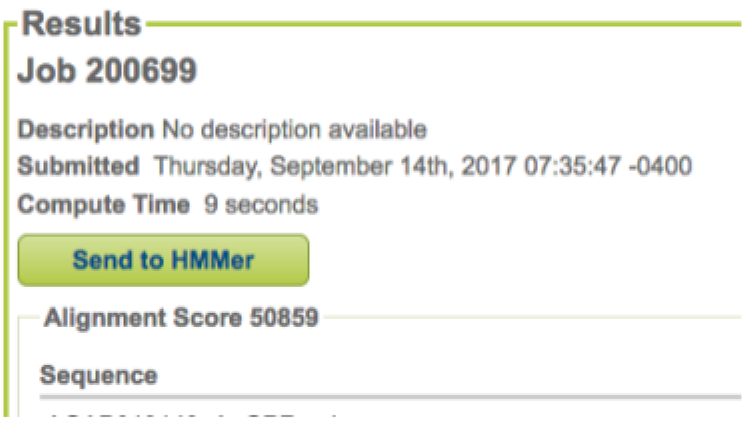
View the result either in VectorBase web-based Java applet (Jalview[2]) or download the alignment file and view on a text editor such as Notepad++ (Windows) or TextWrangler (Mac)

Which opsin sequence(s) has a long "N-terminal extension" according to the alignment?

	Answer
AgGPRop1	
AgGPRop10, 11 and 12	X
AgGPRop12	

You should always be suspicious and critical when you see a deviation from the norm in an alignment like this. Perhaps that gene has an incorrectly predicted gene model? Is that extension present in other species? (These are not questions you have to answer now.)

Click on Send to HMMER. The output file from ClustalW is the input file for HMMER



Results

Job 200699

Description No description available

Submitted Thursday, September 14th, 2017 07:35:47 -0400

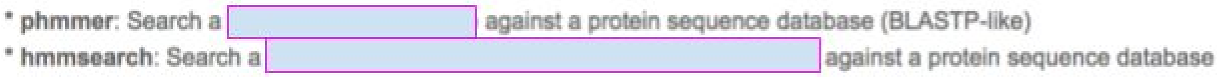
Compute Time 9 seconds

Alignment Score 50859

Sequence

[2] Graphic version of the alignment will work with Firefox, Safari or Explorer, not with Chrome.

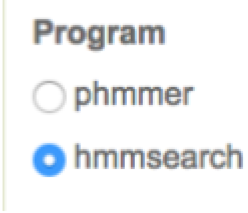
By default the program hmmsearch is selected. Using the information provided in the page complete the sentences of what each program does:



* **phmmer**: Search a against a protein sequence database (BLASTP-like)
* **hmmsearch**: Search a against a protein sequence database

- phmmer:

- hmmsearch:

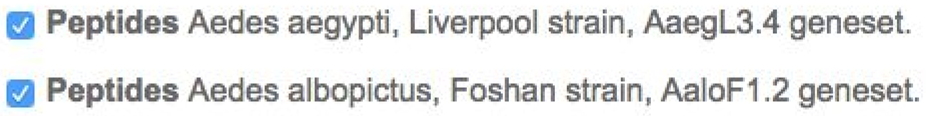


Program

○ phmmer

● hmmsearch

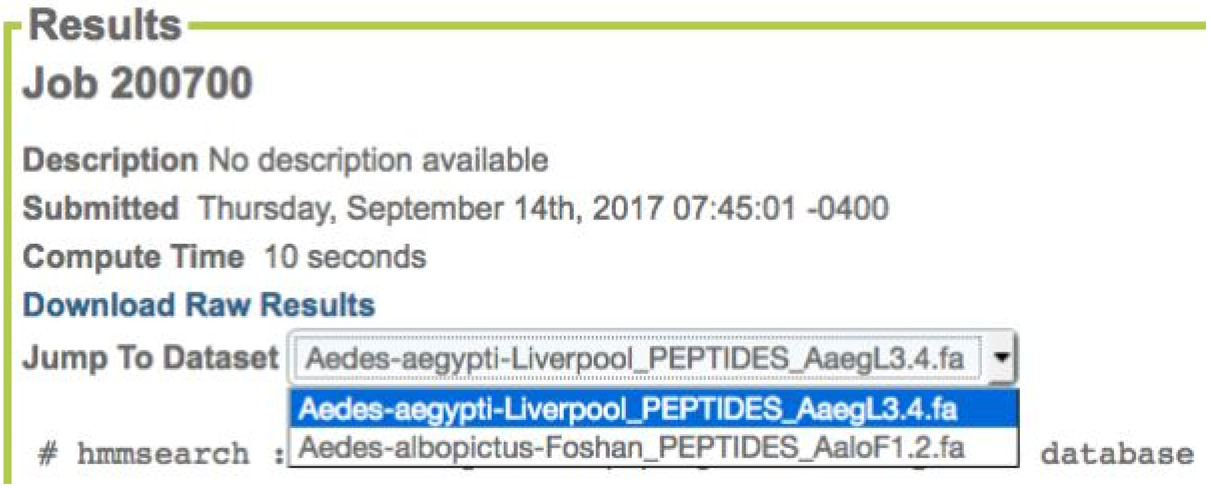
Click on two datasets to be searched against: *Aedes albopictus* and *Ae. aegypti*. Submit the job.



☑ **Peptides** Aedes aegypti, Liverpool strain, AaegL3.4 geneset.

☑ **Peptides** Aedes albopictus, Foshan strain, AaloF1.2 geneset.

You can analyze result directly on VectorBase page or you can click on “Download Raw Results” and open the file in a text editor. Note how to switch between the two sets of results from the two species ("Jump to Dataset" selector).



Results

Job 200700

Description No description available

Submitted Thursday, September 14th, 2017 07:45:01 -0400

Compute Time 10 seconds

Download Raw Results

Jump To Dataset Aedes-aegypti-Liverpool_PEPTIDES_AaegL3.4.fa

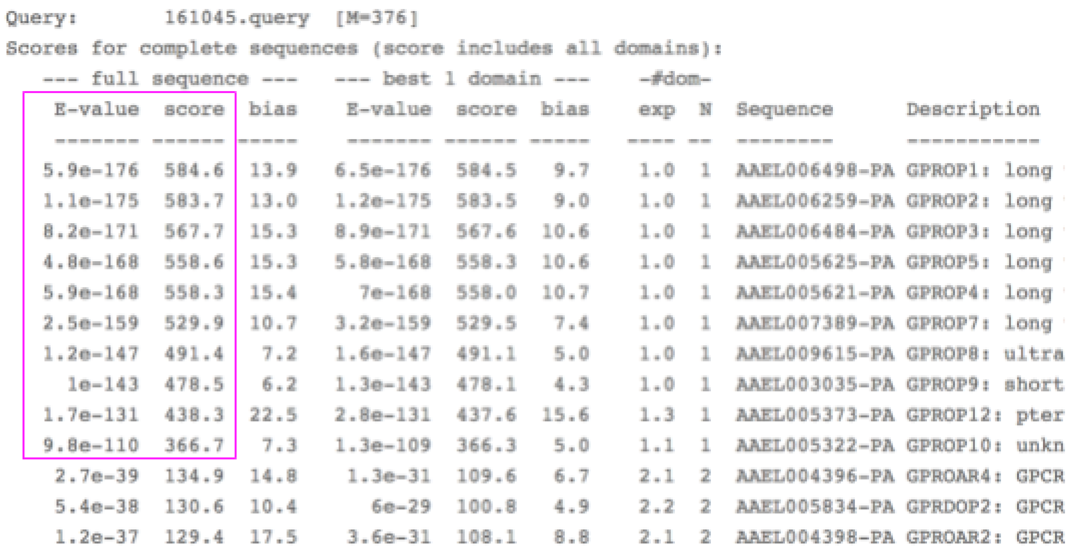
Aedes-aegypti-Liverpool_PEPTIDES_AaegL3.4.fa
Aedes-albopictus-Foshan_PEPTIDES_AaloF1.2.fa

```
# hmmsearch :                                                   database
```

Of the list of genes obtained (see image below) how many counting from the top are true opsin homologous genes?

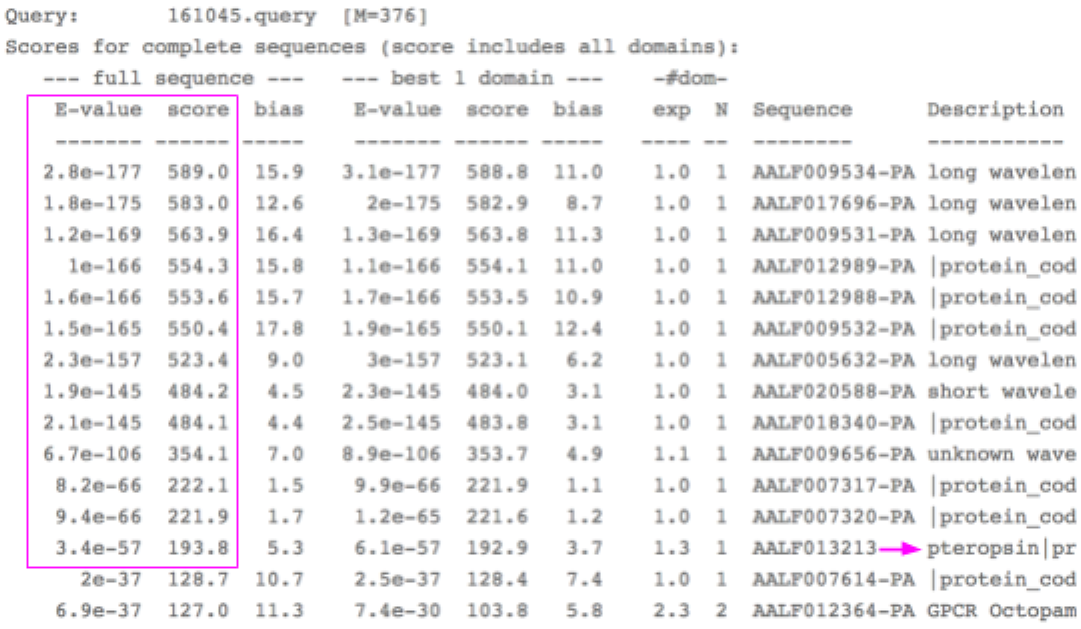
10 for *Aedes aegypti* and 13 for *Ae. albopictus*

Aedes aegypti



```
Query:       161045.query  [M=376]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence      Description
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
   5.9e-176  584.6  13.9   6.5e-176  584.5   9.7    1.0  1  AAEL006498-PA GPROP1: long
   1.1e-175  583.7  13.0   1.2e-175  583.5   9.0    1.0  1  AAEL006259-PA GPROP2: long
   8.2e-171  567.7  15.3   8.9e-171  567.6  10.6    1.0  1  AAEL006484-PA GPROP3: long
   4.8e-168  558.6  15.3   5.8e-168  558.3  10.6    1.0  1  AAEL005625-PA GPROP5: long
   5.9e-168  558.3  15.4     7e-168  558.0  10.7    1.0  1  AAEL005621-PA GPROP4: long
   2.5e-159  529.9  10.7   3.2e-159  529.5   7.4    1.0  1  AAEL007389-PA GPROP7: long
   1.2e-147  491.4   7.2   1.6e-147  491.1   5.0    1.0  1  AAEL009615-PA GPROP8: ultra
     1e-143  478.5   6.2   1.3e-143  478.1   4.3    1.0  1  AAEL003035-PA GPROP9: short
   1.7e-131  438.3  22.5   2.8e-131  437.6  15.6    1.3  1  AAEL005373-PA GPROP12: pter
   9.8e-110  366.7   7.3   1.3e-109  366.3   5.0    1.1  1  AAEL005322-PA GPROP10: unkn
    2.7e-39  134.9  14.8    1.3e-31  109.6   6.7    2.1  2  AAEL004396-PA GPROAR4: GPCR
    5.4e-38  130.6  10.4      6e-29  100.8   4.9    2.2  2  AAEL005834-PA GPRDOP2: GPCR
    1.2e-37  129.4  17.5    3.6e-31  108.1   8.8    2.1  2  AAEL004398-PA GPROAR2: GPCR
```

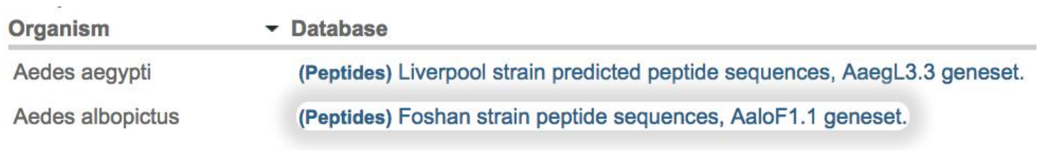
Aedes albopictus



```
Query:       161045.query  [M=376]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence      Description
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
   2.8e-177  589.0  15.9   3.1e-177  588.8  11.0    1.0  1  AALF009534-PA long wavelen
   1.8e-175  583.0  12.6     2e-175  582.9   8.7    1.0  1  AALF017696-PA long wavelen
   1.2e-169  563.9  16.4   1.3e-169  563.8  11.3    1.0  1  AALF009531-PA long wavelen
     1e-166  554.3  15.8   1.1e-166  554.1  11.0    1.0  1  AALF012989-PA |protein_cod
   1.6e-166  553.6  15.7   1.7e-166  553.5  10.9    1.0  1  AALF012988-PA |protein_cod
   1.5e-165  550.4  17.8   1.9e-165  550.1  12.4    1.0  1  AALF009532-PA |protein_cod
   2.3e-157  523.4   9.0     3e-157  523.1   6.2    1.0  1  AALF005632-PA long wavelen
   1.9e-145  484.2   4.5   2.3e-145  484.0   3.1    1.0  1  AALF020588-PA short wavele
   2.1e-145  484.1   4.4   2.5e-145  483.8   3.1    1.0  1  AALF018340-PA |protein_cod
   6.7e-106  354.1   7.0   8.9e-106  353.7   4.9    1.1  1  AALF009656-PA unknown wave
    8.2e-66  222.1   1.5    9.9e-66  221.9   1.1    1.0  1  AALF007317-PA |protein_cod
    9.4e-66  221.9   1.7    1.2e-65  221.6   1.2    1.0  1  AALF007320-PA |protein_cod
    3.4e-57  193.8   5.3    6.1e-57  192.9   3.7    1.3  1  AALF013213→   pteropsin|pr
      2e-37  128.7  10.7    2.5e-37  128.4   7.4    1.0  1  AALF007614-PA |protein_cod
    6.9e-37  127.0  11.3    7.4e-30  103.8   5.8    2.3  2  AALF012364-PA GPCR Octopam
```

Question 3.4

In another web browser tab, perform a VectorBase BLASTp with all the *An. gambiae* genes against *Aedes* peptides using an E-value threshold of 1. Click on the blue (database) link to open the results.



Organism	Database
Aedes aegypti	(Peptides) Liverpool strain predicted peptide sequences, AaegL3.3 geneset.
Aedes albopictus	(Peptides) Foshan strain peptide sequences, AaloF1.1 geneset.

Of the list of genes obtained with BLASTp how many counting from the top are true *Ae. albopictus* opsin homologous genes ?

	True	False
39		X
The list actually has repetitive hits, the redundancy makes it difficult to interpret the results	X	
110		X

Question 3.5

Which of the following statements most accurately reflects what the HMMER results for this query tell you?

	Answer
HMMER finds more homologs than BLAST	
HMMER is missing many of the homologous genes	
BLAST finds more homologs than HMMER	
HMMER data interpretation in easier than BLAST	X

Question 3.6

Some suggested uses for VectorBase's HMMER tool are listed below. Which ones sound accurate, and which are not?

VectorBase HMMER Tool usage scenario	Accurate	Not accurate
Find the closest homologue of gene X (from species Y) in species Z.		X use BLAST!
Starting with a set of protein sequences belonging to a gene family you know well, finding very remote homologues in one or more species.	X	

If you need help with any question and its answer contact us at info@vectorbase.org. Because VectorBase data, tools and resources are updated every two months (6 release cycles per year), answers to these exercises will change too.