Using HMMER versus BLAST to find homologs

Gloria I. Giraldo-Calderón
October 2013

VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

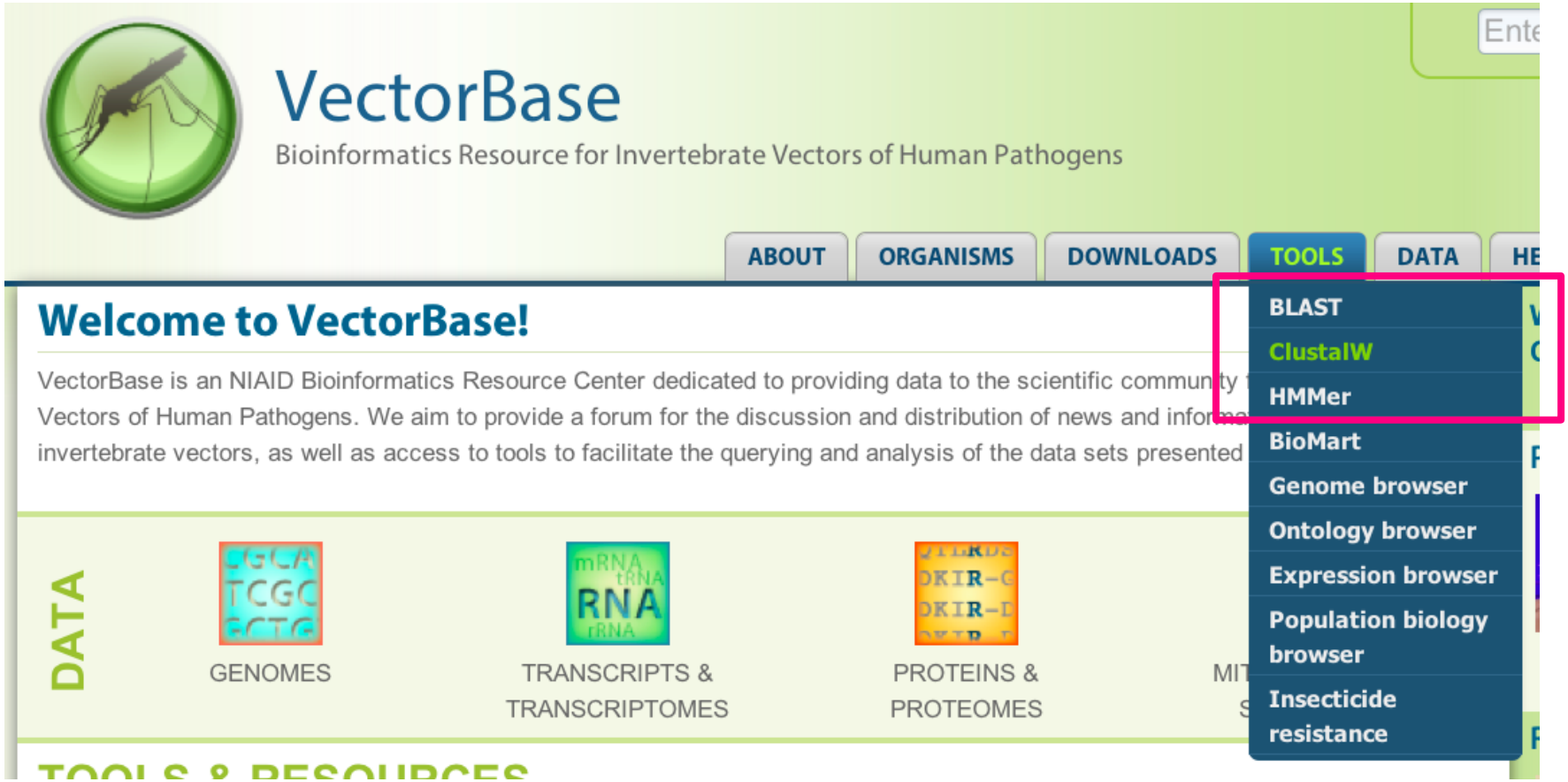
Outline

1. How to find HMMER and BLAST?
2. What can you use HMMER for?
3. How HMMER works?
4. How can you cite HMMER in your paper?

Outline

1. **How to find HMMER and BLAST?**
2. What can you use HMMER for?
3. How HMMER works?
4. How can you cite HMMER in your paper?

1. How to find HMMER and BLAST?



Outline

1. How to find HMMER and BLAST?
2. **What can you use HMMER for?**
3. How HMMER works?
4. How can you cite HMMER in your paper?

2. What can you use HMMER for?

- To search sequence databases for homologs of **protein** sequences using either:
  - multiple sequence alignments (MSA) of a sequence family (very powerful!) or
  - single query sequences (not recommended!)

2. What can you use HMMER for?

Compared to other database search tools (and sequence alignment tools), based on older scoring methodology, HMMER aims to be significantly:

- more accurate
- more able to detect remote homologs (because of the strength of its underlying probability models)
- as fast as BLAST!

2. What can you use HMMER for?

- Nucleotide-nucleotide searches (**blastn**) are <u>not</u> the best method for finding homologous protein coding regions in other organisms.

- That task is better accomplished by performing protein-protein searches (**blastp**) or by translated BLAST searches (**tblastn**, **tblastx** and **blastx**).

- <u>This is because</u> of the codon degeneracy, the greater information available in amino acid sequence, and the more sophisticated algorithm and scoring matrix used in protein-protein BLAST.

NCBI/BLAST/help
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide

Outline

1. How to find HMMER and BLAST?
2. What can you use HMMER for?
3. **How HMMER works?**
4. How can you cite HMMER in your paper?

3. How HMMER works?

HMMER makes a **profile** of the query that assigns a position-specific scoring system for:

- substitutions
- insertions
- deletions

The profiles are probabilistic models called “profile hidden Markov models” (profile HMMs).

3. How HMMER works?

1st step: copy the sample file provided in the front page of this tutorial.

Home » Help » Tutorials » Using HMMER versus BLAST to find homologs

Using HMMER versus BLAST to find homologs

Submitted by ggiraldo on Tue, 2013-02-05 21:10

To follow this tutorial you can use your sequences of interest or download a sample file following this link.

If you want to discuss any issues raised in this tutorial then please contact the help desk.

Download: VectorBase_Using_HMMER_versus_BLAST_to_find_homologs_2013.pdf

Supplementary files: Gene family amino acid sequences Ag.txt

3. How HMMER works?

2nd step: construct a **ClustalW** MSA @ VectorBase.
https://www.vectorbase.org/clustalw (Tools tab).

ClustalW

Paste 2 or more sequences here

Job Control
Load results
Job ID
Add Description
Description
SUBMIT
RESET

Upload ClustalW Input File
Choose File no file selected

Parameters

Basic

Sequence Type
DNA
Protein

Pairwise Alignment
Full
Fast

Full Options

- Click on “Submit”.
- Write down your Job ID
- Click on “Send to HMMER”.

3. How HMMER works?

3rd step: using the MSA build a profile HMM and search with this profile against a sequence database using **hmmsearch**.

HMMER

```
CLUSTAL 2.1 multiple sequence alignment


AgGPRop8        --MGLVQLDNQTAYRPEALIGADQSGLRYLGWN-----------------VPPEELVHIP
AgGPRop9        MFLGNESISEGAMLMPMARTAGEMP--KLLGWN-----------------LPPEEQYLVH
AgGPRop10       --MGRQGSGNAVRISPSSRNQPYFSSAHLSFVVPFPVHSKYVVRSGYVLPVDPLFVAKIN
AgGPRop3        -------MAAFVEP--HFDAWTQSGG-NMSVVDK----------------VPPEMLHMVH
AgGPRop4        -------MAAFVEP--HFDAWTQSGG-NMSVVDK----------------VPPEMLHMVH
AgGPRop1        -------MAAFVEP--HFDAWTQGSG-NMSVVDK----------------VPPEMLHMVH
AgGPRop6        -------MAAFAEP--HFSAWTQTVVSNVTVVDK----------------VPPEMLHMVD
AgGPRop7        -------MPYYGPMQ-QPGLWGQPVA-NLTVVDK----------------VPPEIMHLVD
AgGPRop5        -MMDHRPVGIFGPKSPQALTWTISVA-NLTVVDK----------------VPPEMLHLVD
AgGPRop11       ------------------------------------------------MYDVT-DAAA
AgGPRop12       ------------------------------------------------MNDAPNDVAA
```

Upload HMMer Input File

Choose File no file selected

Job Control

Load results

Job ID

Add Description

Description

SUBMIT

RESET

Parameters

Basic

Sequence Type

Protein

Program

phmmer

hmmsearch

Cut-Offs

E-value

Bit Score

Cut-off values

Significance

Sequence 0.01

Hit 0.03

Report

Sequence 0.01

Hit 0.03

Datasets

- [x] Peptides Aedes aegypti, Liverpool
- [ ] Peptides Anopheles albimanus, ST
- [ ] Peptides Anopheles arabiensis, Dongola strain, AdarD1.0 geneset.
- [ ] Peptides Anopheles christyi, ACHKN1017 strain, AchrA1.0 geneset.
- [ ] Peptides Anopheles darlingi, Coari strain, AdarC1.1 geneset.
- [ ] Peptides Anopheles dirus A, WRAIR2 strain, AdirW1.0 geneset.
- [ ] Peptides Anopheles epiroticus, Epiroticus2 strain, AepiE1.0 geneset.
- [ ] Peptides Anopheles funestus, FUMOZ strain, AfunF1.0 geneset.

- Chose two "Datasets": *Aedes* and *Culex*.
- Click on "Submit".
- Take note of the Job ID

Results

Job [redacted]

Compute Time 2 seconds

Download Raw Results

Jump To Dataset

- ✓ Aedes-aegypti-Liverpool_PEPTIDES_AaegL1.4.fa
- Culex-quinquefasciatus-Johannesburg_PEPTIDES_CpipJ1.3.fa

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.0 (March 2010); http://hmmer.org/
# Copyright (C) 2010 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  hmmer_q2w65g.hmm
# target sequence database:        /vectorbase/dbs/Aedes-aegypti-Liverpool_PEPTIDES_AaegL1.4.fa
# sequence reporting threshold:    E-value <= 0.01
# domain reporting threshold:      E-value <= 0.03
# domain inclusion threshold:      E-value <= 0.03
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       sequence  [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
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
   7.3e-175  581.0  13.7     8e-175  580.9   9.5    1.0  1  AAEL006498-RA long wavelength sensitive opsin|
   1.2e-174  580.3  12.8   1.3e-174  580.1   8.9    1.0  1  AAEL006259-RA long wavelength sensitive opsin|
   2.3e-169  562.9  15.2   2.5e-169  562.8  10.5    1.0  1  AAEL006484-RA long wavelength sensitive opsin|
```

Return to Top

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.0 (March 2010); http://hmmer.org/
# Copyright (C) 2010 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  hmmer_q2w65g.hmm
# target sequence database:        /vectorbase/dbs/Culex-quinquefasciatus-Johannesburg_PEPTIDES_CpipJ1.3.fa
# sequence reporting threshold:    E-value <= 0.01
```

Remember to cite the gene set version that you use in your paper.

3. How HMMER works?

Aedes
output

```
Query:       sequence  [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
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
   7.3e-175  581.0  13.7     8e-175  580.9   9.5    1.0  1  AAEL006498-RA long wavelength sensitive opsin|
   1.2e-174  580.3  12.8   1.3e-174  580.1   8.9    1.0  1  AAEL006259-RA long wavelength sensitive opsin|
   2.3e-169  562.9  15.2   2.5e-169  562.8  10.5    1.0  1  AAEL006484-RA long wavelength sensitive opsin|
   3.1e-167  555.9  15.1   3.5e-167  555.7  10.5    1.0  1  AAEL005625-RA long wavelength sensitive opsin|
   3.8e-167  555.6  15.3   4.3e-167  555.4  10.6    1.0  1  AAEL005621-RA long wavelength sensitive opsin|
   1.4e-157  524.1  10.7   1.8e-157  523.7   7.4    1.0  1  AAEL007389-RA long wavelength sensitive opsin|
   1.7e-146  487.6   7.1    2e-146  487.4   4.9    1.0  1  AAEL009615-RA ultraviolet wavelength sensitive
   2.2e-143  477.4   5.7   2.7e-143  477.1   4.0    1.0  1  AAEL003035-RA short wavelength sensitive opsin
     2e-131  438.0  22.5   3.3e-131  437.3  15.6    1.3  1  AAEL005373-RA pteropsin|protein_coding|superco
   1.9e-110  369.0   6.9   2.4e-110  368.7   4.8    1.1  1  AAEL005322-RA unknown wavelength sensitive ops
    6.3e-40  136.9  14.7    2.6e-32  111.8   6.6    2.1  2  AAEL004396-RA GPCR Octopamine/Tyramine Family|
      1e-38  132.9  11.8    7.4e-30  103.8   5.4    2.2  2  AAEL005834-RA GPCR Dopamine Family|protein_cod
    2.7e-38  131.5   9.7    3.2e-29  101.7   2.0    2.1  2  AAEL017181-RA GPCR Muscarinic Acetylcholine Fa
```

3. How HMMER works?

Culex
output

```
Query:       sequence  [M=379]
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
    ------- ------ -----    ------- ------ -----   ---- --  --------      -----------
     9e-172  571.0  16.0     1e-171  570.9  11.1    1.0  1  CPIJ011571-RA long wavelength sensitive opsin
   3.7e-171  569.0  17.0   4.1e-171  568.9  11.8    1.0  1  CPIJ012052-RA long wavelength sensitive opsin
     6e-169  561.7  17.4   6.6e-169  561.6  12.1    1.0  1  CPIJ011574-RA long wavelength sensitive opsin
     6e-169  561.7  17.4   6.6e-169  561.6  12.1    1.0  1  CPIJ011576-RA long wavelength sensitive opsin
   2.8e-167  556.3  17.0   3.3e-167  556.0  11.8    1.0  1  CPIJ011573-RA long wavelength sensitive opsin
   6.6e-164  545.2  13.4     8e-164  544.9   9.3    1.0  1  CPIJ004067-RA opsin (long wavelength sensitive
   1.6e-163  543.9  21.1   1.7e-163  543.8  14.6    1.0  1  CPIJ020021-RA long wavelength sensitive opsin
     8e-150  498.8   7.3   9.3e-150  498.6   5.1    1.0  1  CPIJ009246-RA ultraviolet wavelength sensitive
   2.4e-149  497.3   8.3   2.8e-149  497.0   5.7    1.0  1  CPIJ013408-RA short wavelength sensitive opsin
     1e-148  495.1   9.2   1.2e-148  494.9   6.4    1.0  1  CPIJ005000-RA short wavelength sensitive opsin
   5.1e-144  479.7  19.1     2e-142  474.4  13.2    2.0  1  CPIJ013056-RA long wavelength sensitive opsin
   3.2e-117  391.5   7.1   4.5e-117  391.0   4.9    1.2  1  CPIJ014334-RA pteropsin|protein_coding|superco
   1.5e-107  359.7   6.0   2.3e-107  359.0   4.1    1.2  1  CPIJ011419-RA unknown wavelength sensitive ops
    6.2e-39  133.8  30.0    4.1e-24   85.0   8.5    3.7  3  CPIJ005574-RA sulfakinin receptor|protein_codi
    1.2e-36  126.3  11.1    1.6e-27   96.3   2.9    2.2  2  CPIJ008330-RA conserved hypothetical protein|p
    2.9e-36  125.0  14.5      4e-36  124.6  10.1    1.1  1  CPIJ018504-RA neuropeptide Y receptor|protein_
```

3. How HMMER works?

You could use other MSA software such as the ones available at EBI http://www.ebi.ac.uk/Tools/msa/

- Clustal Omega: CLUSTAL O(1.1.0) multiple sequence alignment
- Kalign: Kalign (2.0) alignment in ClustalW format
- MAFFT: CLUSTAL format alignment by MAFFT L-INS-1 (v6.850b)
- MUSCLE: MUSCLE (3.8) multiple sequence alignment

Note: You will have to select "Clustal" as the "output format" and replace the alignment output file headers with the "ClustalW2" header:

CLUSTAL 2.1 multiple sequence alignment

3. How HMMER works?

HMMER

```
MUSCLE (3.8) multiple sequence alignment


AgGPRop11      ------------------------------MYDVTDAA-AINSDHQELMAP--------
AgGPRop12      ------------------------------MNDAPNDVAASAVDYEDLMAP--------
AgGPRop10      MGRQGSGNAVRISPSSRNQPYFSSAHLSFVVPFPVHSKY-VVRSGYVLPVDPLFVAKINP
AgGPRop5       --------MMDHRPVGIFGPKSP--------QALTWTIS-VANLTVVDKVPPEMLHLVDT
AgGPRop7       --------------MPYYGPMQ---------QPGLWGQP-VANLTVVDKVPPEIMHLVDP
```

You just need to change the header (as shown by the arrow).

HMMER

```
CLUSTAL 2.1 multiple sequence alignment


AgGPRop11      ------------------------------MYDVTDAA-AINSDHQELMAP--------
AgGPRop12      ------------------------------MNDAPNDVAASAVDYEDLMAP--------
AgGPRop10      MGRQGSGNAVRISPSSRNQPYFSSAHLSFVVPFPVHSKY-VVRSGYVLPVDPLFVAKINP
AgGPRop5       --------MMDHRPVGIFGPKSP--------QALTWTIS-VANLTVVDKVPPEMLHLVDT
AgGPRop7       --------------MPYYGPMQ---------QPGLWGQP-VANLTVVDKVPPEIMHLVDP
```

Outline

1. How to find HMMER and BLAST?
2. What can you use HMMER for?
3. How HMMER works?
4. **How can you cite HMMER in your paper?**

4. How you can cite HMMER in your paper?

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research. Web Server Issue 39:W29-W37.

How to search for more information or help?

E-mail us at

info@vectorbase.org

or go to HMMER home page and download the user manual: http://hmmer.janelia.org