Adverb 1.0 Instructions

Your download bundle should contain:

* This ReadMe document
* A jar file
* Source code
* A sample order-level fasta file of BOLD records (Order Amphipoda)

To identify a COI nucleotide query sequence, you need a fasta file of all BOLD sequences of the query’s order or class. If neither of these is known in advance, blast the query against BOLD and use the class of the best hit.

Identification proceeds in 3 steps which are invoked from the Linux command line:

STEP 1: Extract per-genus training sets from the fasta file.

STEP 2: Align all the training sets using Clustal-Omega.

STEP 3: Compute an HMM from each training set, and compute the query sequence’s log-Viterbi probability on each HMM. The family of the HMM with the highest probability is the Adverb prediction for the family of the query.

**To prepare for Step 1**, install Clustal-Omega locally. The download is available at <http://www.clustal.org/omega/#Download>. Also create a nucleotide fasta file of all BOLD sequences of the query’s order or class. Each defline must specify the phylum, class, order, family, genus, and species of the sequence in the format shown below:

>P\_Arthropoda\_\_C\_Malacostraca\_\_O\_Amphipoda\_\_F\_Epimeriidae\_\_G\_Epimeriella\_\_S\_Epimeriella macronyx

Copy the fasta and Adverb\_1.0.jar into a new directory. cd into the directory and give yourself write and execute permission by typing

chmod +wx .

**To execute Step 1**, cd into the directory containing the fasta and Adverb\_1.0.jar and type

java -cp Adverb\_1.0.jar adverb.CollectTrainingSets *fasta* *clustal-path*

For *fasta*, use the name of your fasta file. For *clustal-path*, provide the path to the Clustal-Omega executable (it probably ends with “clustalo”). Execution should take up to a few minutes.

2 subdirectories and 1 bash script will be created:

* Subdirectory full\_genus\_fastas contains 1 fasta file for every genus in the input fasta, unless the genus has no sequences in the length range 480-782. It is provided for reference, and may be deleted.
* Subdirectory unaligned\_training\_fastas contains 1 training set for each genus fasta in full\_genus\_fastas. Up to 25 unique sequences are chosen, all with length in the range 480-782, representing the species diversity of the genus.
* Bash script align\_all.sh invokes Clustal-Omega to align the training fastas.

**To execute Step 2**, type

./align\_all.sh

The script creates subdirectory aligned\_training\_fastas. Execution should take up to a few minutes. After execution, unaligned\_training\_fastas is no longer needed and may be deleted.

**To execute Step 3**, type

java -cp Adverb\_1.0.jar adverb.BuildAndExecuteHmms *query-seq* *parallel*

For *query-seq*, use your nucleotide query sequence; note that all characters other than acgtACGT will be deleted from the query. For *parallel*, type “true” or “false” to compute using parallel threads or a single serial thread. Parallel computation provides efficient acceleration but has high memory requirements, and is only recommended for execution on a high-performance cluster with abundant memory.

This step is time-consuming (hours or days on a single core). Computation of a single log-Viterbi probability takes 1-2 minutes on a 2.7 GHz Intel i7 core, and a large class may contain thousands of genera.

**Practice Session:**

This practice session can be completed in under 10 minutes. It takes you through the steps of classifying a sequence from order Amphipoda. It has been tested on MacOS and should work on any Linux system.

1. Create a new directory. Copy Adverb\_1.0.jar and O\_Amphipoda.fa into the new directory.
2. Open a terminal window and cd into the new directory.
3. Execute Step 1 (create training fastas):

java -cp Adverb\_1.0.jar adverb.CollectTrainingSets O\_Amphipoda.fa *clustal-path*

1. Execute Step 2 (align):

./align\_all.sh

1. Begin Step 3 (compute):

java -cp Adverb\_1.0.jar adverb.BuildAndExecuteHmms *query-seq* false

For *query-seq*, type tail -1 O\_Amphipoda.fa and copy/paste the output.

**License:**

License is hereby granted for any non-commercial use. You may freely distribute the source code and the jar file. All source code is Copyright © 2021 Philip Heller.