Identification of proteins containing transmembrane domains using Phobius

José Flores

Abstract

Citation: José Flores Identification of proteins containing transmembrane domains using Phobius. **protocols.io** dx.doi.org/10.17504/protocols.io.pdwdi7e
Published: 04 Jun 2018

Guidelines

This protocol was executed via ssh on a Ubuntu machine using the command line.

Before start

Checklist:

1. FASTA file containing the proteins to analyze.

2. FASTX toolkit installed

3. phobius installed

4. pandas installed in python

Protocol

Gathering the sequence from the proteins to analyze

Step 1.

If you already possess a file containing the proteins to screen in FASTA format, skip this step.

Here we will retrieve all the proteins from the Organic Lake phycodnavirus 2 (OLPV2) from the NCBI website.

To download the proteins in the GenBank entry:

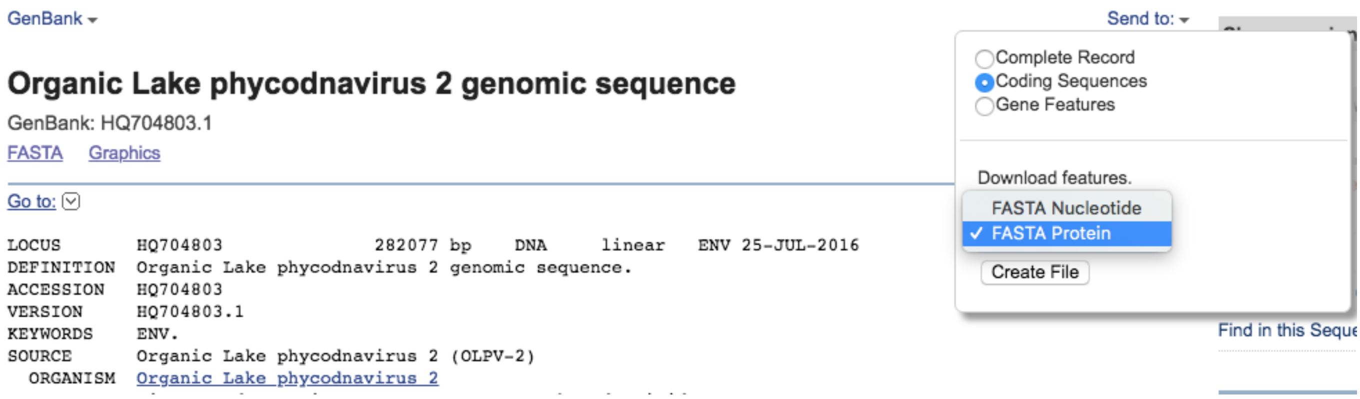
1. Click 'Send to:'

2. Select 'Coding Sequences'

3. Select 'FASTA Protein' as Format.

4. Create File

As shown below:



GenBank ▾

Organic Lake phycodnavirus 2 genomic sequence

GenBank: HQ704803.1

FASTA Graphics

Go to: ⌄

```
LOCUS       HQ704803               282077 bp    DNA     linear   ENV 25-JUL-2016
DEFINITION  Organic Lake phycodnavirus 2 genomic sequence.
ACCESSION   HQ704803
VERSION     HQ704803.1
KEYWORDS    ENV.
SOURCE      Organic Lake phycodnavirus 2 (OLPV-2)
  ORGANISM  Organic Lake phycodnavirus 2
```

Send to: ▾

○ Complete Record
● Coding Sequences
○ Gene Features

Download features.

FASTA Nucleotide
✓ FASTA Protein

Create File

Then rename the file generated to something informative like:

```
OLPV2_prot.txt
```

[Link to the GenBank entry for Organic Lake phycodnavirus 2:](https://www.ncbi.nlm.nih.gov/nuccore/HQ704803.1)

LINK:
https://www.ncbi.nlm.nih.gov/nuccore/HQ704803.1

EXPECTED RESULTS

FASTA format file containing the proteins to analyze.

Reformatting the protein FASTA file.

Step 2.

The FASTA file will be converted into a TSV file with two columns:

1. The sequence header identifier

2. The amino acid sequence

The tool we will use is fasta_formatter from the FASTX toolkit.

SOFTWARE PACKAGE (Ubuntu -)

FASTX Toolkit, 0.0.14

Assaf Gordon

COMMAND (Ubuntu - 14.04.4 LTS)

```
fasta_formatter -t -i OLPV2_prot.txt -o OLPV2_prot.tsv
```

Convert a FASTA format file into a tab separated one.

EXPECTED RESULTS

- A two columns file where the first column contains the sequences header and the second the sequence.

Obtaining the Phobius predictions

Step 3.

Here we will relly on Phobius combined with some command line tools (tail, tr, and awk) to filter the Phobius results and retain only those proteins with at least 1 transmembrane domain.

For an explanation of each part of the command read the steps below, otherwise skip to the next section.

SOFTWARE PACKAGE (Ubuntu -)

Phobius, 1.01

Erik Sonnhammer

COMMAND (Ubuntu - 14.04.4 LTS)

```
phobius.pl -short OLPV2_prot.txt | tail -n+2 | tr -s ' ' | tr ' ' '\t' | awk -
F '\t' '$2 > 0' > OLPV2_prot_TM.tsv
```

Obtains the phobius predictions and filters the results.

EXPECTED RESULTS

A tab-separated file containing the Phobius predictions.

Obtaining the Phobius predictions

Step 4.

cmd COMMAND

```
awk -F '\t' '$2 > 0'  > OLPV_prot_TM.tsv
```

Finally using awk the results are filtered to include only those with TM domains. The $2 > 0 is the parameter indicates awk to retain only lines where the second column ($2 in awk terminology), the one where phobius prints the number of TM, shows presence of TM in the sequence. The output of awk is piped into the tab-separated file OLPV_prot_TM.tsv

Merging the phobius predictions to the FASTA sequences

Step 5.

To merge the tables of sequences and Phobius results we will use the following python script.

Copy paste the following into a file called:

`merge_tables.py`

cmd COMMAND

```
#!/usr/bin/env python

import pandas as pd
import sys

phobius_table = sys.argv[1]
proteins_table = sys.argv[2]
merged_tables_file = 'merged_tables.tsv'

phobius_df = pd.read_table(phobius_table, sep='\t', names=['SEQ_ID', 'TM', 'SP', 'PREDICTION'])
proteins_df = pd.read_table(proteins_table, sep='\t', names=['SEQ_HEADER', 'SEQ'])

proteins_df['SEQ_ID'] = proteins_df['SEQ_HEADER'].apply(lambda x: x.split(' ')[0])
proteins_df['DESCRIPTION'] = proteins_df['SEQ_HEADER'].apply(lambda x: ' '.join(x.split(' ')[1:]))

merged_df = phobius_df.merge(proteins_df, on='SEQ_ID', how='left')

merged_df = merged_df.loc[merged_df['TM'] >= 5]

merged_df.to_csv(merged_tables_file, sep='\t', columns=['SEQ_ID', 'DESCRIPTION', 'TM', 'SP', 'PREDICTION', 'SEQ'])
```

Merging the phobius predictions to the FASTA sequences

Step 6.

cmd COMMAND

```
python merge_tables.py OLPV2_prot_TM.txt OLPV2_prot.tsv
```

The script takes two arguments: 1. The tab-separated file containing the phobius results. 2. The tab-separated file containing the amino acid sequences.

EXPECTED RESULTS

A tab separated file called merged_tables.tsv with six columns:

1. SEQ_ID: Identifier for each sequence.

2. DESCRIPTION: if the analyzed proteins came from GenBank this field contains the annotations.

3. TM: Number of transmembrane domains identified by Phobius.

4. SP: Presence of signal peptide in the protein.

5. PREDICTION: The segments of the protein corresponding to the different transmembrane domains.

6. SEQ: Amino acid sequence of the protein