

Run uproc-DNA using iMicrobe [↗](#)Alise Ponsero¹¹University of Arizona[dx.doi.org/10.17504/protocols.io.u3jeykn](https://doi.org/10.17504/protocols.io.u3jeykn)

iMicrobe

Metafunc course 2018



Alise Ponsero

University of Arizona



ABSTRACT

How to run [uproc v1.2.0](#) ([Meinicke, 2014](#)) through the [iMicrobe](#) platform.

The ultrafast protein classification (UProC) toolbox implements a mosaic matching algorithm for large-scale protein annotation analysis. UProC is up to three orders of magnitude faster than profile-based methods and achieved higher sensitivity on unassembled short reads (100 bp) from simulated metagenomes. UProC does not depend on a multiple alignment of family-specific sequences. Therefore, in addition to the protein domain classification according to the Pfam database, UProC also provide the detection of KEGG Orthologs.

More informations about centrifuge can be found here : <http://uproc.gobics.de/>

TAGS

metagenomics

annotation

EXTERNAL LINK

<https://www.imicrobe.us>

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

GUIDELINES

More informations and details about Uproc can be found here : <https://github.com/gobics/uproc>

Several parameters are available to the user through the iMicrobe app.

PROTEIN CLASSIFICATION OPTIONS:

-P N --pthresh N

Protein threshold level. Allowed values:

0 fixed threshold of 0.0

2 less restrictive

3 more restrictive

Default is 3 .

-l --long

Use long read mode (default): Only accept certain ORFs (see -O below) and report all protein scores above the threshold (see -P above).

-s --short

Use short read mode: Accept all ORFs, report only maximum protein score (if above threshold).

-O N --othresh N

ORF translation threshold level (only relevant in long read mode).

Allowed values:

- 0 accept all ORFs
 - 1 less restrictive
 - 2 more restrictive
- Default is 2.

Note that the iMicrobe app runs automatically the job with the following output format parameters, and therefore produces 6 output files (*.kegg|pfam28.preds, *.kegg|pfam28.stats, *.kegg|pfam28.counts) :

-p --preds

Print all classifications as CSV. The printed fields are in the following order :

- n: sequence number (starting from 1)
- h: sequence header up to the first whitespace
- l: sequence length (this is a lowercase L)
- F: ORF frame number (1-6)
- I: ORF index in the DNA sequence (starting from 1)
- L: ORF length
- f: predicted protein family
- s: classification score

-f --stats

Print "CLASSIFIED,UNCLASSIFIED,TOTAL" sequence counts.

-c --counts

Print "FAMILY,COUNT" where COUNT is the number of classifications for FAMILY

BEFORE STARTING

- You need a working Cyverse account to connect to iMicrobe.
To explore how to log into iMicrobe, read [the dedicated protocol](#).
- Your dataset of interest should be metagenomic reads, in a fasta or fastq format.
- In iMicrobe, there is several ways to run an app on a dataset (from the cart, from your personal datastore and form an URL). If you need more information on how to run an app, [read the protocol associated](#).

Run uproc on metagenomic reads

- 1 This protocol section uses [SRS01342](#), a metagenomic sample from the HMP project. This sample is a posterior formix sample. The app was run on the read pair 1.

NOTES : Uproc doesn't take into account paired-end read technology. To use the tool on a paired-end read dataset, the user can either pair the read using tools like [PEAR or Pandaseq](#) or run the analysis on the two files separately.

In the iMicrobe sample search page, select the mock communities to add them in your cart. In the 'tools' dropdown menu, select 'Apps'. You are presented the list of apps currently available on iMicrobe. Click on [uproc_dna-1.2.0u3](#).

In the page app, provide the input files using the cart or the datastore. The user can tune the following parameters :

- read length (by default "long")
- ORF translation threshold level (by default "More restrictive")
- Protein threshold level (by default "More restrictive")

Note : for more details on the app parameters, please read the "Guidelines" section of this protocol.

After the job is effectively ran, you can access your results using the drop-down menu 'Tools' and selecting 'Jobs'. Select the job corresponding to your centrifuge run, and go to the section 'Outputs'. The uproc output files are now in your cyverse datastore. Click on 'Browse and view output files in the CyVerse Datastore'.

In the job folder created in the CyVerse datastore, the input fasta/fastq files are copied, along with the logs of the job (*.err and *.out). In order to retrieve your results go to the uproc-out folder. This folder should contain the following outputs :

FILE_NAME.kegg.preds and FILE_NAME.pfam28.preds

This output file contains all classifications as comma delimited table. The printed fields are in the following order :

- n: sequence number in the file (starting from 1)
- h: sequence header up to the first whitespace
- l: sequence length (this is a lowercase L)
- F: ORF frame number (1-6)
- l: ORF index in the DNA sequence (starting from 1)
- L: ORF length
- f: predicted protein family - Kegg or pfam family ID.
- s: classification score

FILE_NAME.kegg.stats and FILE_NAME.pfam28.stats

This output contains a unique line stating the classified sequence count, the unclassified sequence count and the total sequence count for the given input file separated by commas.

FILE_NAME.kegg.counts and FILE_NAME.pfam28.counts

This output contains for each protein family the total number of classifications found. The fields are in the following order :

- protein family id (Kegg or Pfam id)
- counts

FILE_NAME.kegg.annotated and FILE_NAME.pfam28.annotated

This output contains protein annotations as a comma delimited table. The printed fields are the following :

- **pfam_id or kegg_id** : predicted protein family (Kegg or pfam28 protein family ID)
- **count** : sequence count for this protein family
- **identifier** (pfam28 only) : short identifier of the protein family
- **name** : complete protein family name
- **pathway module** (kegg only) : pathway of the protein family

EXPECTED RESULT

Results obtained on the exemple set (SRS013542.denovo_duplicates_marked.trimmed.1.fastq), run with the following parameters :

Read length : short

ORF translation threshold level : more restrictive

Protein threshold level : more restrictive

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.kegg.counts

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.kegg.counts.annotated

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.kegg.preds


☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.kegg.stats

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.pfam28.counts

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.pfam28.counts.annotated

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.pfam28.preds

☐ SRS013542.denovo_duplicates_marked.trimmed.1.fastq.pfam28.stats

 This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited