OrthoFinder on the Command Line

OrthoFinder is a program created by Emms and Kelly that provides information about orthogroups, among many other things (Emms and Kelly 2015). Orthogroups are groups of genes that descended from a single gene in a common ancestor – or homologs. Orthogroups can therefore represent gene families, which can consist of orthologs and paralogs. Orthologs are genes in different species that come from a common ancestral gene whereas paralogs are genes that came from a duplication event within a single species. Homologs are any genes that share an origin and thus include both orthologs and paralogs. Orthofinder allows you to analyze the relationships between proteins from two or more datasets. Additionally, analysis with OrthoFinder yields a lot of other useful information. It can create gene trees based on the orthogroups, and species trees for the fasta files provided, assuming each fasta file represents an individual species. OrthoFinder provides additional information such as duplication events and specific sequences that are putative orthologs. When it comes to trying to find similarities or differences between proteomes, OrthoFinder is an extremely useful program.

In this tutorial, I will be looking at different aspects of the proteomes of SARS, MERS, and SARS-Cov2 such as the relationships between the viruses and relationships between the proteins that the viruses express. These proteomes were downloaded from NCBI in fasta format and then transferred to my personal UC Riverside Cluster account. The links to these proteomes can be found below in the step by step tutorial. Using these proteomes, we can determine genetic similarities and differences among the viruses. Pinpointing these differences is the key to understanding how the virus takes control of host cells and expresses its proteins. It is an important step in trying to figure out how to make an effective treatment or vaccine. Before I jump into this example however, I am first going to go over some of the basics of OrthoFinder to show how to run it and how to interpret the results.

Accompanying videos:

Tutorial Part 1: https://www.youtube.com/watch?v=Vo2sWCdyIng
Tutorial Part 2: https://www.youtube.com/watch?v=O1u41vkmu9g

Manual

About

OrthoFinder is a useful program for comparative genomics. It is capable of finding orthogroups, gene trees, species trees, and other useful information for comparing proteomes. OrthoFinder can also be paired with other programs to get additional information such as alignments. OrthoFinder can be used on the command allowing for easy setup and analysis of output.

Setup

In order to use OrthoFinder, you must first have two or more proteomes that you would like to analyze. OrthoFinder is only able to recognize files with the extensions ".fa", ".fasta", ".faa", ".fas", and ".pep" so it is important that the files you use are in proper format. The files you are using must also all be in the same directory. This directory must only contain the files you are analyzing because the directory must be specified as the input for OrthoFinder. You must either have the OrthoFinder program downloaded or loaded as a module in order to use it. Information on how to download the OrthoFinder program can be found at the following website: https://github.com/davidemms/OrthoFinder .

Running OrthoFinder

In order to run OrthoFinder, specify the program pathway and use the parameter "-f" to denote the data you are using. See example below where "Directory" represents the directory containing your input files:

```
OrthoFinder -f /rhome/dblumsack/shared/Directory
```

There are numerous additional parameters that can be added on to modify the results that OrthoFinder returns. The parameters that you can use are listed below and come directly from the OrthoFinder manual which can be accessed using the "—help" command.

```
-t <int>      Number of parallel sequence search threads [Default = 32]
-a <int>      Number of parallel analysis threads [Default = 1]
-M <txt>      Method for gene tree inference. Options 'dendroblast' & 'msa'
              [Default = dendroblast]
-S <txt>      Sequence search program [Default = diamond]
              Options: blast, diamond, blast_gz, mmseqs
-A <txt>      MSA program, requires '-M msa' [Default = mafft]
              Options: mafft, muscle
-T <txt>      Tree inference method, requires '-M msa' [Default = fasttree]
              Options: fasttree, raxml, raxml-ng, iqtree
-s <file>     User-specified rooted species tree
```

```
-I <int>        MCL inflation parameter [Default = 1.5]
-x <file>       Info for outputting results in OrthoXML format
-p <dir>        Write the temporary pickle files to <dir>
-1              Only perform one-way sequence search
-X              Don't add species names to sequence IDs
-n <txt>        Name to append to the results directory
-o <txt>        Non-default results directory
-h              Print this help text

WORKFLOW STOPPING OPTIONS:
-op            Stop after preparing input files for BLAST
-og            Stop after inferring orthogroups
-os            Stop after writing sequence files for orthogroups
               (requires '-M msa')
-oa            Stop after inferring alignments for orthogroups
               (requires '-M msa')
-ot            Stop after inferring gene trees for orthogroups

WORKFLOW RESTART COMMANDS:
-b  <dir>       Start OrthoFinder from pre-computed BLAST results in <dir>
-fg <dir>       Start OrthoFinder from pre-computed orthogroups in <dir>
-ft <dir>       Start OrthoFinder from pre-computed gene trees in <dir>
```

As previously mentioned, the parameters listed above are optional and not required in order to run OrthoFinder. The only information that must be specified is the pathway to the directory specified by “-f”.

Output

The output for OrthoFinder is put in a directory within the directory containing your input. This directory is labelled “OrthoFinder”. Within this directory is yet another directory named “Results_[date]” where [date] refers to the date that the OrthoFinder analysis was ran on. In the Results directory, there are numerous other directories named according to the information that they contain. Below is a description of these directories and what they contain. Additional information can be found in the github repository for OrthoFinder.

Orthogroups: Contains files pertaining to the orthogroups created from your input. Orthogroups are defined as groups of genes thought to have originated from a single gene. There is information regarding the number of sequences in each orthogroup, the exact sequences in each orthogroup, the orthogroups with only one sequence, and the genes unassigned to any orthogroup.

Phylogenetic_Hierarchical_Orthogroups: Orthogroups found using a method different from the gene similarity method that produced the results in the Orthogroups directory described above. These orthogroups are more accurate than the orthogroup inference method based on gene similarity.

Resolved_Gene_Trees: Gives resolved gene trees from your input. The files are in newick format, and can be viewed in tree viewers such as ETE Toolkit which as a web gui that's easy to use. This is a representation of how the genes in your input are related to one another. These gene trees are resolved using the OrthoFinder hybrid species-overlap/duplication-loss coalescent model (Emms and Kelly 2015).

Species_Tree: Gives a species tree from your input. Assuming each input file you provided represents proteins from an individual species, the species tree reports how each of these species are related to one another based on the orthogroups derived from the provided input.

Gene_Duplication_Events: Gives information regarding possible duplications events that could have occurred with the gene families/orthogroups derived from your input. This information comes from the orthogroup gene tree created by OrthoFinder.

Gene_Trees: Gives resolved gene trees from you input. The files can be viewed in a tree viewing software as they are in newick format. These gene trees are not resolved using OrthoFinder's unique model.

Orthogroup_Sequences: Contains fasta files for every orthogroup found by OrthoFinder. Each of the fasta files contains all the sequences in the specified orthogroup.

Orthologues: Contains TSV files describing possible orthologues found by OrthoFinder among the proteomes given as input. Orthologues are genes found in different species that have evolved from a common ancestral gene through a speciation event.

Putative_Xenologs: Contains TSV files describing possible xenologs predicted by OrthoFinder among the proteomes given as input. Xenologs are orthologues where homologous sequences are found in different species because of horizontal gene transfer.

Single_Copy_Orthologue_Sequences: Contains all sequences that belong to orthogroups which contain only a single sequence from your input.

Comparative_Genomics_Statistics: Gives general information regarding the analysis performed by OrthoFinder. Provides overall statistics and statistics per species as well as summaries for the various directories of information produced as output from OrthoFinder.

WorkingDirectory: Provides a working directory of what OrthoFinder has done to arrive at the results. This is useful if OrthoFinder has returned an error and you do not know what caused the error.

Useful Additional Information
Website: https://github.com/davidemms/OrthoFinder
Manual on Command Line: OrthoFinder --help

Example
Now that we have gone through the basics of how to run OrthoFinder and what the results mean, I am going to go through an example of how one might apply the OrthoFinder program to real-world research. See the accompanying videos, links to the videos can be found at the end of the tutorial.

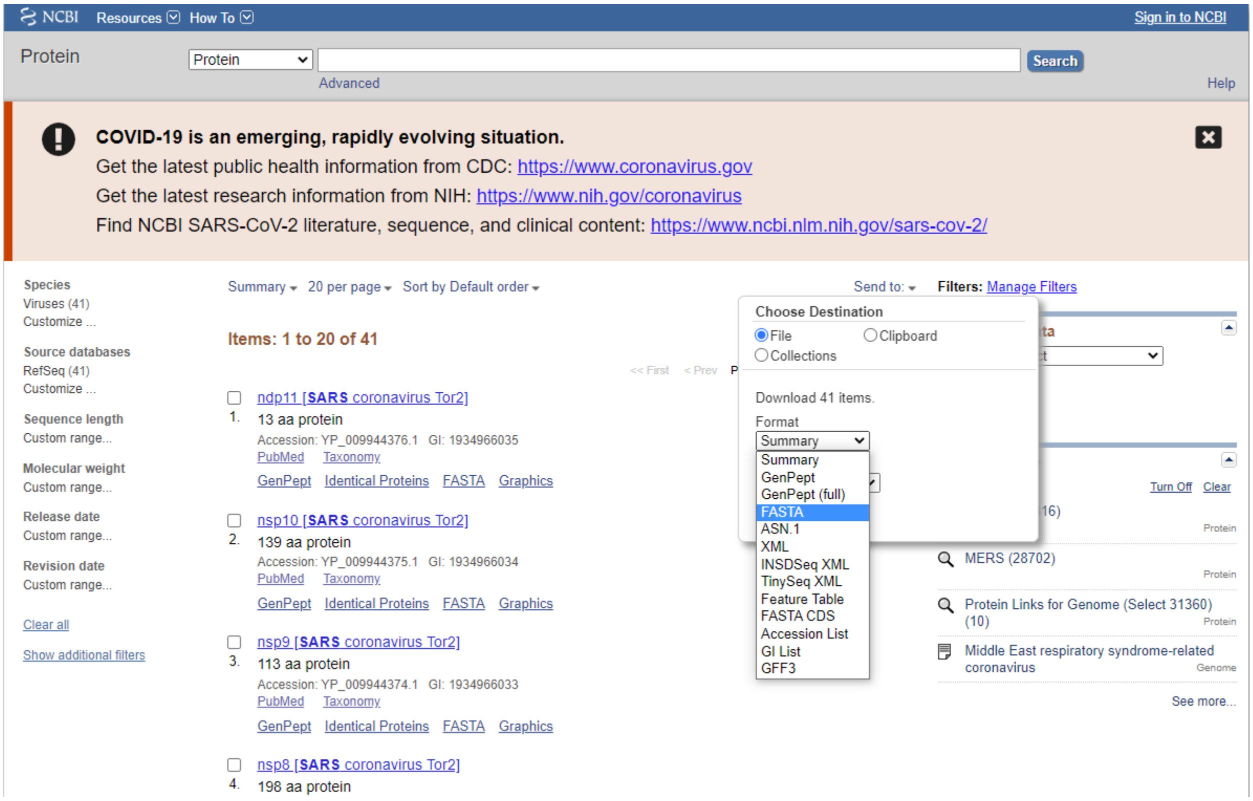
I focus on comparing the proteomes of SARS, MERS, and SARS-Cov2. These are three viruses that share ancestry, which means that they should share similarities among their proteins. If you would like to follow along with the tutorial, I have provided links to the NCBI pages for each virus dataset. It is on these pages that you can download the proteomes in fasta format.

SARS Proteome (41 proteins) :
https://www.ncbi.nlm.nih.gov/protein/1934966035,1934966034,1934966033,1934966032,1934966031,1934966030,1934966029,1934966028,1934966027,1934966026,1934966025,1934966024,1904811885,1873624195,1845982731,1845982730,1845982729,1845982728,1845982727,1845982726,1845982725,1845982724,1845982723,1845982722,1845982721,1845982720,1845982719,34555778,34555776,34555775,34555774,30133975,29837507,29837506,29837505,29837503,29837502,29837501,29837500,29837499,29837498

MERS Proteome (37 proteins) :
https://www.ncbi.nlm.nih.gov/protein/667489425,667489424,667489423,667489422,667489421,667489420,667489419,667489418,667489417,667489416,667489415,667489414,667489413,667489412,667489411,667489410,667489409,667489408,667489407,667489406,667489405,667489404,667489403,667489402,667489401,667489400,667489399,667489398,667489397,667489396,667489395,667489394,667489393,667489392,667489391,667489390,667489389

SARS-Cov2 Proteome (38 proteins) :
https://www.ncbi.nlm.nih.gov/protein/1826688927,1826688926,1826688925,1826688924,1826688923,1826688922,1826688921,1826688920,1826688919,1826688918,1820616061,1802476820,1802476819,1802476818,1802476817,1802476816,1802476815,1802476814,1802476813,1802476812,1802476811,1802476810,1802476809,1802476808,1802476807,1802476806,1802476805,1802476803,1798174256,1798174255,1796318604,1796318603,1796318602,1796318601,1796318600,1796318599,1796318598,1796318597

Screenshot of How To Download Proteins from NCBI



COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: https://www.coronavirus.gov

Get the latest research information from NIH: https://www.nih.gov/coronavirus

Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/

Items: 1 to 20 of 41

Choose Destination

● File ○ Clipboard

○ Collections

Download 41 items.

Format

Summary

1. ndp11 [**SARS** coronavirus Tor2]
13 aa protein
Accession: YP_009944376.1 GI: 1934966035
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

2. nsp10 [**SARS** coronavirus Tor2]
139 aa protein
Accession: YP_009944375.1 GI: 1934966034
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

3. nsp9 [**SARS** coronavirus Tor2]
113 aa protein
Accession: YP_009944374.1 GI: 1934966033
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

4. nsp8 [**SARS** coronavirus Tor2]
198 aa protein

16)
Protein

MERS (28702)
Protein

Protein Links for Genome (Select 31360) (10)
Protein

Middle East respiratory syndrome-related coronavirus
Genome

As seen in the screenshot above, the proteins can be downloaded by pressing the drop down menu labelled "Send to". You will then have to check off "file" and specify that you want to download the proteins in FASTA format.

If you are working on a high performance computing cluster, the next step is to move the files from your local computer to your directory on the cluster. If you are not working on a cluster, please disregard this step. In order to move the proteomes onto your cluster, you must first open your terminal. From there, use the **scp** command to move the fasta files onto your cluster. An example of this command is shown below. Where the command says username, it should be your username.

```
scp /c/Users/[username]/Desktop/refseq_SARS_proteins.fa
[username]@cluster.hpcc.ucr.edu:/rhome/[username]/shared/Bioinformatics
```

The next step once the files are in the correct location, is to create a directory specific for the protein fasta files. To do this, use the command **mkdir** and then specify the name of your directory after. Once again, an example command follows.

```
mkdir Virus_orthofinder_analysis
```

The files can then be moved to this new directory using the command **mv**. This command must be specified along with both the name of the file you want to move and the destination path.

```
mv refseq_SARS_proteins.fa /Virus_orthofinder_analysis
```

After moving the files to their own directory, the fun begins. Call the OrthoFinder program if you have it downloaded by specifying the path to the program on your local. If you are working on a cluster where it is installed as a module, load the OrthoFinder module and call the program by specifying the program's name. I am working on a cluster so I will show you how to load the module and call the program.

```
module load orthofinder
orthofinder
```

Calling the program will give you information on how to use the function and what parameters can be used with the program. This information is covered briefly in the first part of the tutorial and also in the Orthofinder manual. In order to run OrthoFinder on the three proteomes that you have retrieved, all you have to do is once again call OrthoFinder, and then specify the directory that holds the three proteomes.

```
orthofinder /Virus_orthofinder_analysis
```

This command should tell OrthoFinder to analyze the three proteomes located in the specified directory. After OrthoFinder completes its analysis, you will be left with a subdirectory located within the directory that you have created for the proteomes. That subdirectory will be labelled "OrthoFinder" and will contain another subdirectory labelled "Results_Oct05" where the date will be the date you ran OrthoFinder. In the results directory, you will find the many directories that contain the results as described above in the first part of the tutorial.

Please check my videos on Youtube for further clarification on how to run OrthoFinder and how to interpret the results.

Tutorial Part 1: https://www.youtube.com/watch?v=Vo2sWCdyIng
Tutorial Part 2: https://www.youtube.com/watch?v=O1u41vkmu9g

References

Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (1): 157. https://doi.org/10.1186/s13059-015-0721-2.