

Phylogenetics Tutorial 1:

Making Phylogenies

Finlay Maguire

Faculty of Computer Science

Table of contents

1. Overview
2. Installation
3. Data
4. Multiple Sequence Alignment
5. Trimming
6. Approximate ML Tree
7. Maximum-Likelihood Tree
8. Phylogenomics

Overview

Protein Phylogeny Aims

- Get a protein
- Using pairwise alignment to find potential homologs
- Perform a multiple sequence alignment
- Trim the alignment
- Infer a NJ distance phylogeny
- Infer an approximate Maximum Likelihood phylogeny
- Infer an accurate Maximum Likelihood phylogeny
- Compare the trees

Core Genome Phylogeny

- Get genomes
- Find core genome
- Extract SNPs
- Infer a Maximum Likelihood phylogeny
- Visualise Phylogeny

Requirements

- mafft
- trimal
- aliview
- FastTree2
- iqtree
- FigTree
- prokka
- roary
- snp-sites

Installation

If you don't have miniconda

<https://docs.conda.io/en/latest/miniconda.html>

```
conda create -n phylo -c bioconda mafft trimal prokka
```

```
fasttree iqtree roary snp-sites
```

```
conda activate phylo
```

or if older miniconda version:

```
source activate phylo
```


Unfortunately, not everything is in bioconda:

- AliView
<https://github.com/AliView/AliView/releases>
- FigTree
<https://github.com/rambaut/figtree/releases>

Data

Starting Sequence

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase
Swiss-Prot (559,228)
S Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.
TrEMBL (146,106,279)
Automatically annotated and not reviewed.

UniRef
y

UniParc
D

Proteomes
FHL

Supporting data

Literature citations l	Taxonomy h	Subcellular locations c
Cross-ref. databases x	Diseases d	Keywords o

News

Forthcoming changes
There are currently no changes planned

UniProt release 2019_02
Let's twist again with Myo1D | Removal of the cross-references to CleanEx | Change of URIs for Orphanet

UniProt release 2019_01
Engaging and disengaging: CRISPR rings | Cross-references to JPOST

[News archive](#)

Figure 1: High-quality protein reference database: swiss-prot
<http://www.uniprot.org>

Starting Sequence

The screenshot shows a search interface with a 'View by' menu on the left and search results on the right. The 'View by' menu includes options like 'Results table', 'Taxonomy', 'Keywords', 'Gene Ontology' (which is highlighted), 'Enzyme class', and 'Pathway'. The search results list several Gene Ontology terms with their respective result counts and a 'Z' icon.

View by

Results table

Taxonomy

Keywords

Gene Ontology

Enzyme class

Pathway

Search:

- molecular_function (471698 results) Z
- cellular_component (411411 results) Z
- biological_process (440181 results) Z
- reproduction (749 results) Z
- cell killing (1960 results) Z
- immune system process (8229 results) Z
- sulfur utilization (4 results)

Figure 2: Choose 'Gene Ontology' and 'biological process'

Starting Sequence

- [-] detoxification (798 results) Z
 - [+] toxin catabolic process (143 results) Z
 - mycothiol-dependent detoxification (4 results)
 - detoxification of zinc ion (5 results)
 - [+] detoxification of nitrogen compound (20 results) Z
 - [+] detoxification of inorganic compound (99 results) Z
 - detoxification of arsenic-containing substance (8 results)
 - [+] cellular detoxification (560 results) Z

Figure 3: Go down to 'detoxification' and expand

Starting Sequence

Filter by¹

Reviewed (8)
Taxa: Prot

Popular organisms

- A. thaliana (1)
- C. elegans (1)
- ECOLX (2)
- HALSA (1)
- ACIMA (1)

BLAST	Align	Download	Add to basket	Columns	d	1 to 8 of 8
Entry	Entry name	Protein names	Gene names	Organism		
1 result(s) selected. (Clear Selection)						
<input type="checkbox"/>	P30632 ASNA_CAEEL	S ATPase asna-1	asna-1 tag-205, ZK637.5	Caenorhabditis elegans		
<input checked="" type="checkbox"/>	P08690 ARSA1_ECOLX	S Arsenical pump-driving ATPase	arsA	Escherichia coli		
<input type="checkbox"/>	O52027 ARSA_HALSA	S Putative arsenical pump-driving ATP...	arsA arsA2, VNG_5180G	Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)		
<input type="checkbox"/>	O50593 ARSA_ACIMA	S Arsenical pump-driving ATPase	arsA	Acidiphilium multivorum (strain DSM 11245 / JCM 8867 / NBRC 100883 / AJU301)		
<input type="checkbox"/>	A8Q372 ASNA_BRUMA	S ATPase ASNA1 homolog	Bm1_42140	Brugia malayi (Filarial nematode worm)		

Figure 4: Select '8 results' next to 'detoxification of arsenic'

Using BLAST to find related sequences

Filter by!

BLAST | Align | Download | Add to basket | Columns | 1 to 8 of 8 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length	
1 result(s) selected (Clear Selection)						
<input checked="" type="checkbox"/>	P39632	ASNA_CAEEL	ATPase asna-1	asna-1 tag-205, ZK637.5	Caenorhabditis elegans	342
<input type="checkbox"/>	P06690	ARSA1_ECOLX	Arsenical pump-driving ATPase	arsA	Escherichia coli	583
<input type="checkbox"/>	O52027	ARSA_HALSA	Putative arsenical pump-driving ATR...	arsA arsA2_VNG_5180G	Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)	644
<input type="checkbox"/>	O50593	ARSA_ACIMA	Arsenical pump-driving ATPase	arsA	Acidiphilium multivorum (strain DSM 11245 / JCM 8867 / NBRC 100883 / AJJ301)	583

Figure 5: Select the *C. elegans* sequence and BLAST

Using BLAST to find related sequences

BLAST

Job status: RUNNING

Running **blastp** job against **UNIPROTKB** for 15s 

[job information](#)

Query sequence ¹	<pre>>sp P38632 ASNA_CAEEL ATPase asna-1 OS=Caenorhabditis elegans OX=6239 GN=asna-1 PE=1 SV=1 MSDQLEASIKNILEQKTLKWI FVGKGGVGTTCSCSLAAQLSKVRERVLLISTDPAHNI SDAFSQKFTXPTLVVGFKNLFAMEIDSNPWGEGVEMNIEMLQNAAQNEGSGGFSMG</pre>
-----------------------------	--

Figure 6: Wait...

Using BLAST to find related sequences

The screenshot shows a BLAST search results page titled "Alignments". A modal dialog box is open, allowing the user to download selected sequences. The dialog has two options: "Download selected (10)" (which is selected) and "Download all (250)". Below these options, there is a "Format:" dropdown menu set to "FASTA", and radio buttons for "Compressed" and "Uncompressed" (the latter is selected). A "Preview first 10" link and a "Go" button are also present. The background shows a table of alignment results with columns for "Info", "E-value", "Score", and "Ident.". Three results are visible, each with a corresponding sequence alignment visualization.

Info	E-value	Score	Ident.
64326E5E763500BDED0012FB9	0.0	1,759	100.0%
A0A2G5UM79_9PELO - ATPase Cni-asna-1 - Caenorhabditis n...	0.0	1,622	92.1%
ASNA_CAEBR - ATPase asna-1 - Caenorhabditis b...	0.0	1,621	91.8%

Figure 7: Download 10 sequences across a range of similarity

Multiple Sequence Alignment

```
mafft-linsi arsenic.faa > arsenic.afa
```

Trimming

```
java -jar aliview.jar
```

```
trimal -nogaps -in arsenic.afa -out arsenic_nogaps.mask
```

```
trimal -automated1 -in arsenic.afa -out arsenic_auto.mask
```

```
java -jar aliview.jar
```


Approximate ML Tree

```
FastTree -lg arsenic_auto.mask > arsenic_dist.tree
```

FigTree

Maximum-Likelihood Tree

- Generate 100 parsimony trees

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood
- Perform hill-climbing NNI (stochastic followed by hill-climbing) on each and optimise

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood
- Perform hill-climbing NNI (stochastic followed by hill-climbing) on each and optimise
- Retain top 5 topologies as candidate trees

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood
- Perform hill-climbing NNI (stochastic followed by hill-climbing) on each and optimise
- Retain top 5 topologies as candidate trees
- Randomly perturb candidates (stochastic NNI) and optimise (hill-climbing)

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood
- Perform hill-climbing NNI (stochastic followed by hill-climbing) on each and optimise
- Retain top 5 topologies as candidate trees
- Randomly perturb candidates (stochastic NNI) and optimise (hill-climbing)
- If new tree is better than top candidate, replace

- Generate 100 parsimony trees
- Optimise all 100 with lazy SPR moves
- Collect resulting unique topologies and optimise branch lengths
- Select top 20 by likelihood
- Perform hill-climbing NNI (stochastic followed by hill-climbing) on each and optimise
- Retain top 5 topologies as candidate trees
- Randomly perturb candidates (stochastic NNI) and optimise (hill-climbing)
- If new tree is better than top candidate, replace
- If top candidate doesn't change after 100 random perturbations then output.

```
iqtree -mset LG,JTT,WAG -s arsenic_auto.mask
```

Note: IQTree does output a neighbour joining distance tree too (.bionj).

FigTree

Phylogenomics

Download the 6 listeria genomes

```
wget finlaymagui.re/assets/listeria_genomes.tar.gz  
tar xvf listeria_genomes.tar.gz
```

Annotate genomes

For genome GCA000008258:

```
prokka --kingdom Bacteria --outdir prokka_GCA_000008285  
--genus Listeria --locustag GCA_000008285  
GCA_000008285.1_ASM828v1_genomic.fna
```

Repeat for all genomes

Find shared parts

```
mkdir annotations  
cp */*.gff annotations  
roary -f core_genome -e -n -v annotations/*.gff
```

```
snp-sites -o listeria_snps.fna  
    core_genome/core_gene_alignment.aln
```

```
iqtree -mset GTR -s listeria_snps.fna
```

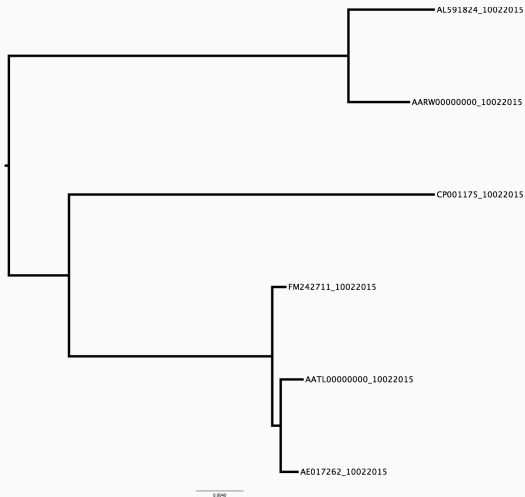


Figure 8: Roary Tutorial

Questions?