Mar 26, 2019

Working

SYSB 3036 W08: Multi-Sequence Alignment and Phylogenetics

Version 2

Frank Aylward[1]

[1]Virginia Tech

dx.doi.org/10.17504/protocols.io.ziwf4fe

Frank Aylward
Virginia Tech

ABSTRACT

PROTOCOL STATUS

Working
We use this protocol in our group and it is working

SAFETY WARNINGS

Get the data

1 First we will download some files from GitHub from a repository called "phylogenetics_tutorial":

```
git clone https://github.com/faylward/phylogenetics_tutorial
```

and then we will move into the new folder and see what files are present:

```
cd phylogenetics_tutorial
```

```
ls -lh
```

You should see a file called SSU_seqs.fna. This file contains some Small Subunit (SSU) rRNA genes, also called 16S rRNA genes, that we will use to practice multi-sequence alignment and phylogenetic inference. Every SSU gene here comes from a different microbe, and by examining these genes we can begin to reconstruct the evolutionary relationships between different groups. SSU rRNA genes are often used for this purpose, and in general these genes are considered great "phylogenetic markers" since they are highly conserved and present in all cellular life (they are necessary for functioning of the ribosome).

Look at the sequences

2 Let's take a quick look at the sequences.

Using Unix head:

```
head SSU_seqs.fna
```

And seqkit:

```
seqkit fx2tab -l -g -n SSU_seqs.fna
```

You should find 8 sequences in FASTA format that vary in length between 1,250-1,530 bp. This is pretty typical for full length SSU rRNA genes. Note that the GC content varies between groups- this is also typical.

Look at Clustal Omega options

3 Now that we have an idea of what the file looks like we can start the sequence alignment. The alignment we need to do here is different from the alignment we did in the past with BLAST or any other homology search tool. That's because here we will be using the alignments for phylogenetic inference, so we need to be super careful that each bp is aligned as accurately as possible. With BLAST we were mainly interested in the alignment summary statistics like the e-value, % identity, and bit score. But here we will be interested in each and every location in the resulting alignment, so we must be very careful.

Also, here we also need to align the full length of the sequences. Before we were mainly interested in "local alignment" where some parts of the sequences may remain un-aligned. Here we are performing "global alignment" in that the full sequences will be aligned.

For this we will use a tool called Clustal Omega. Let's take a quick look at the help menu:

```
clustalo --help
```

There are lots of different options that you can play around with, so take a look at them.

Align the sequences

4 Here we will use a simple clustalo command using mainly default parameters:

```
clustalo -i SSU_seqs.fna -o SSU.aln
```

The -i flag specifies the input, and the -o flag specifies the output. The default output format is "FASTA". Let's take a look at what that looks like:

```
head SSU.aln
```

You will notice that the output file looks fairly similar to the input file, but that the sequences have some gaps that are denoted with "-" symbols. These gaps are needed since some sequences may have insertions or deletions ("indels") that make it impossible to align some regions of one sequence to another.

Let's take another look using seqkit:

```
seqkit fx2tab -l -n -g  SSU.aln
```

You will notice here that all of the sequences are the same length now. This should always be the case for aligned sequences, since gaps will be introduced to ensure that every position in one aligned sequence can be related to the same position in another.

Compare different alignment formats

5 The default FASTA alignment we got from clustalo is just one of many alignment formats that we can get. Just to look at another, let's try Phylip format. This format used to be quite commonly used, and many programs still use it.

```
clustalo -i SSU_seqs.fna -o SSU.phy.aln --outfmt=phy
```

Let's take a look using Unix head:

```
head SSU.phy.aln
```

Whereas the FASTA formatted alignment displayed every sequence in full sequentially, here all aligned regions are displayed together. This is called "Phylip interleaved" format, since the aligned sequences are interleaved. Note that names are also truncated to only 10 characters.

Try Clustal format too and see what that looks like:

```
clustalo -i SSU_seqs.fna -o SSU.clu.aln --outfmt=clu
```

```
head SSU.clu.aln
```

Look at the options in FastTree

Look at the options in FastTree

6 Now that we have our multi-sequence alignment file we can create a phylogeny using a program called FastTree.
As the name suggests, this program is quite fast. Early phylogenic estimation programs were very computaionally intense and took a very long time to finish (many still are). But there are some new methods like FastTree that speed things up.

First let's take a look at the options:

```
fasttree
```

Create the phylogenetic tree

7 To run the program on our alignment we can type:

```
fasttree -nt SSU.aln > SSU.nwk
```

The output in this case is a Newick file, with .nwk extension. This is the most commonly used file format to store phylogenetic trees. The format is quite compact and you will see the file is quite small. Let's take a look:

```
ls -la
```

and

```
head SSU.nwk
```

Newick format is a bit hard to read just by looking at it, especially for trees with many species. The general principal is that parentheses are placed to denote evolutionary relationships. So something like A(B,C) would mean that A and B are more closely related than either is to A. Branch lengths and support values are also in there, so it can get a bit dense.

Visualize the tree in iTOL

8 To visualize phylogenetic trees I like to use a handy website called iTOL, which stands for "interactive Tree of Life". You can find it here:

itol.embl.de

If you click on the "upload" tab on the top you should be taken to a page where you can paste your Newick tree and give it a name. For larger trees you may wish to navigate to upload the .nwk file, but here it may be easiest just to cut-and-paste the test from your Newick file into the itol website and give it a name like "SSU_test".

Once you do this you will be taken to the iTOL browser where you should see a visual representation of your tree, with the sequence names at the "leaves". A menu on the right will allow you to manipulate the tree in different ways.

A few things you can do:

1) On the main menu under "basic" you can increase the width of the branch lines. This makes things easier to see.

2) You can go between "circular", "normal", and "unrooted" tree schema in the top of the "basic" menu. Check out these different views. Are some easier to interpret?

3). In the "advanced" meny you can click on "display" next to "bootstraps/metadata". The default behavior here is to show circles on each node which have size proportional to the node support values. You can click on "text" and play with the font size to see the actual values. These numbers tell us the confidence that FastTree has in the branching order that is presented. Here values are between 0 and 1, with higher values indicating higher support.

4) Note that the branch lengths are very long between Nitrososphaera and Sulfolobus and the rest of the tree. That is becuase these two taxa are Archaea, while the rest of the tree contains Bacteria. Archaea and Bacteria are separated by billions of years of evolution, so it is not surprising that their SSU rRNA genes are quite a bit different. In a way it is amazing that we can even align them!
Since we know that Archaea and Bacteria have this split we can "root" the tree by clicking on the branch above Nitrososphaera and Sulfolobus, navigating down to "Tree Structure" in the banner that pops up, and clicking "Reroot tree here". You will see the topology of the tree change (and hopefully it will make more evolutioary sense!).

Overall iTOL is a great tool for visualizing phylogenies, so spend some time playing around with the different options!

9 Now let's say we find a new genome and we want to add it into our phylogeny. How do we go about that?
Let's start by downloading a genome from NCBI. Let's use Rickettsia prowazekii, the causative agent of Typhys. I browsed the NCBI website and found this link for the genome, which we can download using wget:

```
wget -O rickettsia.fna.gz
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/195/735/GCF_000195735.1_ASM19573v1/GCF_000195735.1_ASM19573v1_genomic.fna.gz
```

and then:

```
gunzip rickettsia.fna.gz
```

10 Now we need to get the 16S gene(s) from this organisms. We can predict them with barrnap:

```
barrnap rickettsia.fna > rickettsia.rrna.gff
```

We only want the 16S gene, not all the other rRNAs, so we can extract the lines that correspond to 16S genes with grep:

```
grep "16S" rickettsia.rrna.gff > rick.16S.gff
```

and we can double check everything looks good:

```
more rick.16S.gff
```

Based on this it looks like there is only one 16S gene.

Since barrnap only gives us a gff file, we still need to extract the actual 16S sequences from the genome. We can do this with BedTools:

```
bedtools getfasta -s -fi rickettsia.fna -bed rick.16S.gff -fo rick.16S.fna
```

and to double check the file:

```
more rick.16S.fna
```

11 For the phylogeny we need a combined alignment of this new 16S gene together with our references. So first let's combined the sequences:

```
cat rick.16S.fna SSU_seqs.fna
```

And then we can make the alignment and phylogeny just as we did in the steps above:

```
clustalo -i all_16S.fna -o all_16S.aln
```

```
fasttree -nt all_16S.aln > all_16S.nwk
```

Note- FastTree will not like it if multiple 16S genes with the same name are in the alignment. It truncates the headers at ":" symbols and takes everything before that as the header, so depending on your input file you may need to manually re-name some of the genes to make sure they are not redundant.

And then visualize with iTOL as we did above. Rickettsia is an Alphaproteobacteria, so you should see it clustering with the Pelagibacter in the tree.