Concatenated Tree Construction Demo: Identify Marker Genes

Nina Dombrowski and Kiley Seitz

Abstract

Demo for identifying marker genes using RiboDB.

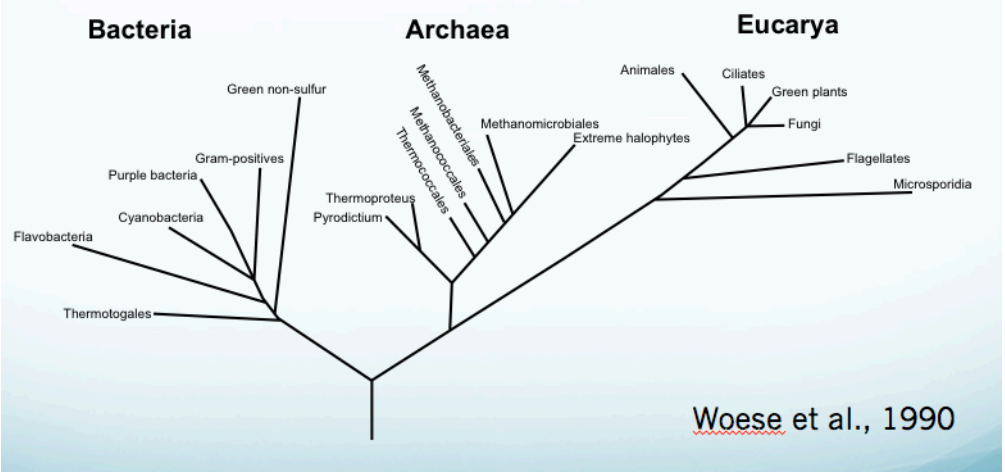
Citation: Nina Dombrowski and Kiley Seitz Concatenated Tree Construction Demo: Identify Marker Genes. **protocols.io** dx.doi.org/10.17504/protocols.io.fa7bihn
Published: 25 Jul 2016

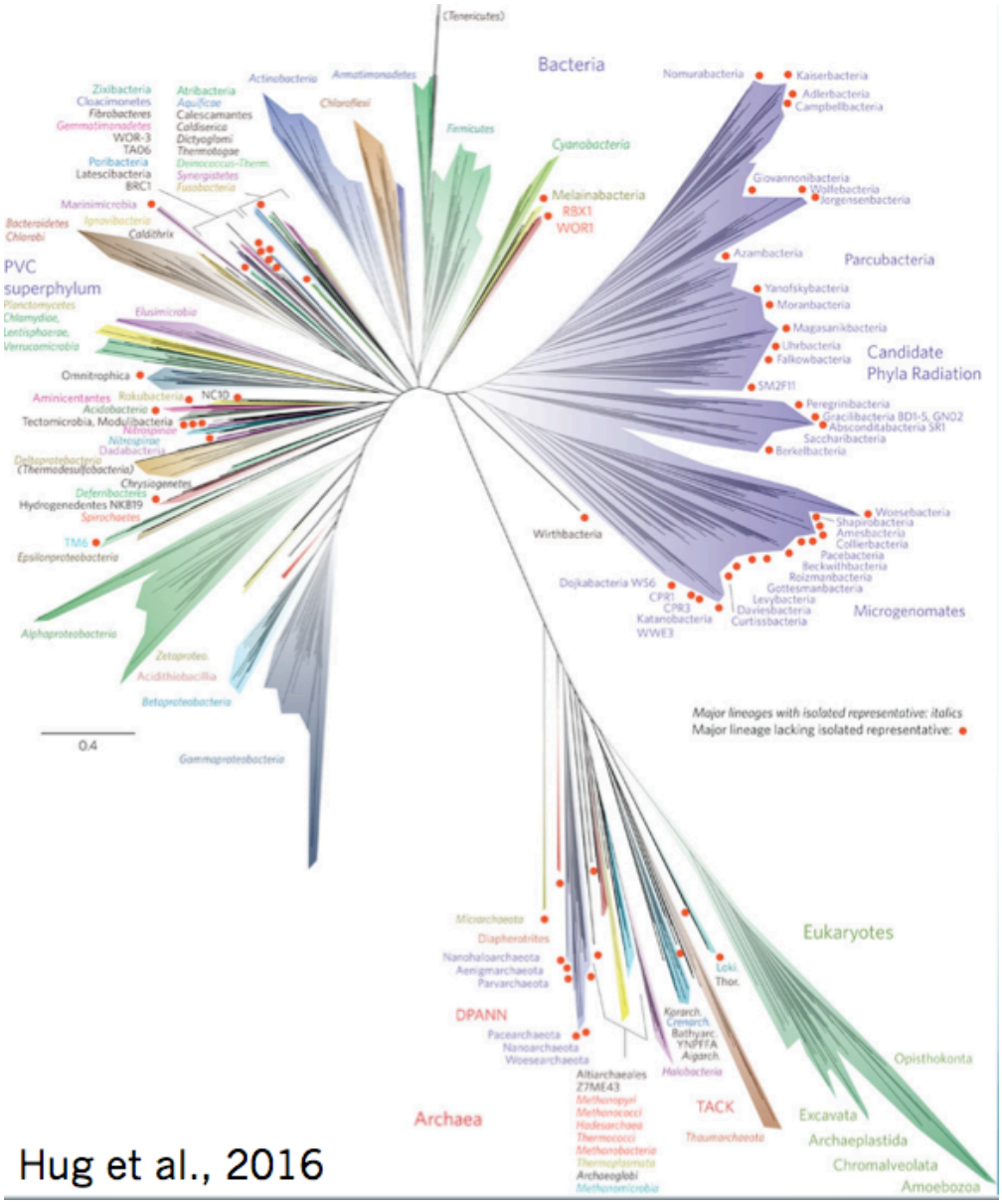
Guidelines

Why important?

1. Identify novel taxonomic lineages
2. Understanding evolutionary processes (Speciation, Geography, Age of taxa, Endosymbiosis, Tree of Life…)
3. Predict function of novel genes (Dsr genes)

An example:





What to use?

1. DNA
2. RNA
3. Protein

Non-coding **DNA** regions: higher mutation rate

→ Rapid evolving sequences for close relatives

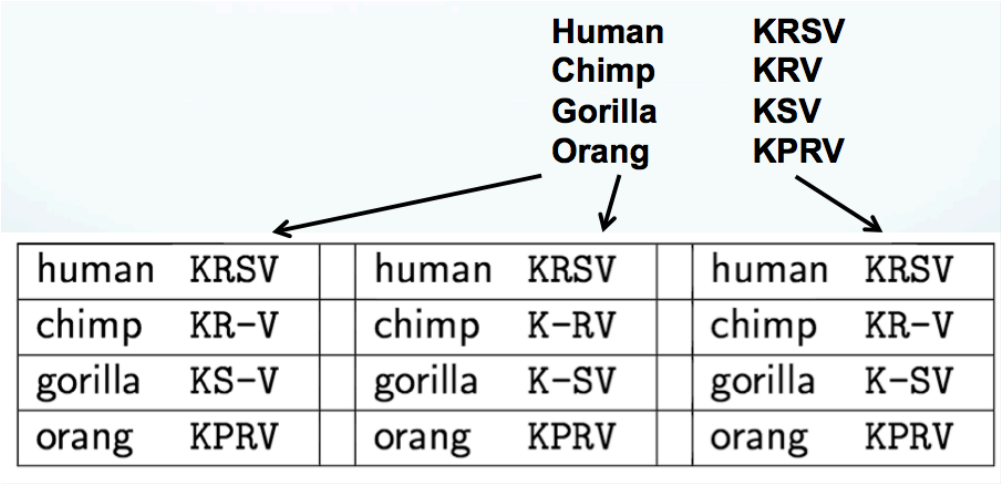
Proteins: mutate slower (must maintain function)

→ Good for distantly related species

Types of alignments

1. Pair-wise alignment
2. Multiple alignment
3. Local alignment (Identify sub-sequences sharing high similarity)
4. Global alignment (Align entire sequences, up to both ends of each sequence)
5. Structure-guided alignments (ie. 16s, Takes secondary structure into account)

Problems: possible to align two sequences by different combination of gaps



human	KRSV		human	KRSV		human	KRSV
chimp	KR-V		chimp	K-RV		chimp	KR-V
gorilla	KS-V		gorilla	K-SV		gorilla	K-SV
orang	KPRV		orang	KPRV		orang	KPRV

Programs for alignments

- ARB (2nd structure)
- MUSCLE
- Geneious
- MEGAN

Tree construction methods

- Distance (Neighbour-joining)
- Maximum Parsimony
- Maximum Likelyhood (ML)
- Bayesian

Distance:

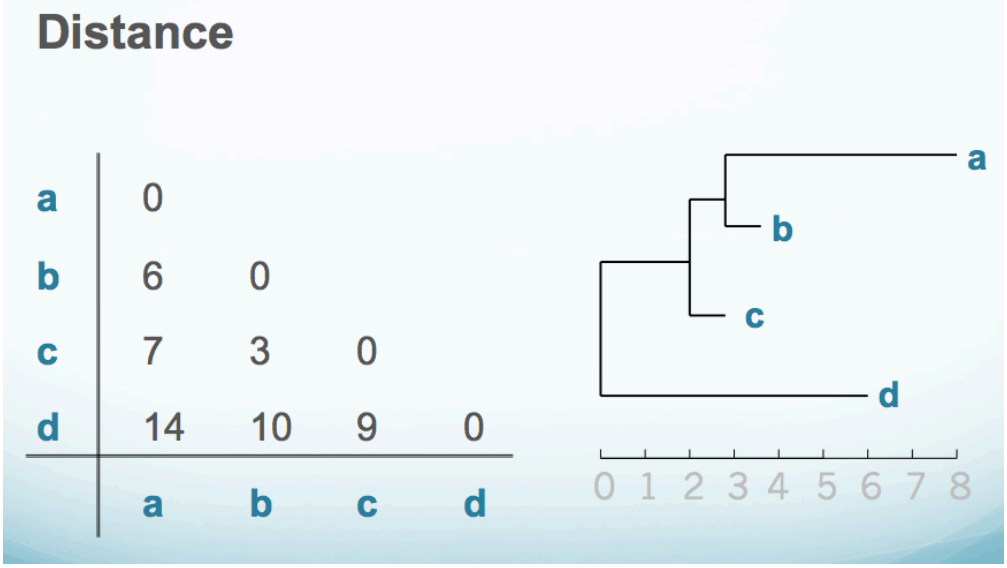
Calculate pairwise distances on alignment

Count number of differences between seqs

Phylogenetic tree calculated on distance matrix

+ Fast (many seqs), good for first tree, ok when low distance between sequences

- Sequence info based on 1 parameter, does not account for multiple mutations at one site



Maximum parsimony:

All possible trees determined from each position of the sequence alignment

Each tree gets a score based on number of steps to generate tree

Chosen tree = min. nr. of mutations that could produce the data

+ Uses all data, Good for close evol. distances

- Slow, assumes equal rate of mutations, many scenarios possible

Maximum likelihood:

ML employs a model of evolution, i.e. different rates of transitions/transversion

Probability calculated how each position reflects sequencing data (for all 4 nucleotide sites)

Tree generated that is most likely to have produced the observed data is generated

+ Model of evolution, uses all sequence info, corrects for multiple mutations → good for large evol. distances

- Very slow, depends on model of evolution used

Bayesian:

Incorporation of prior information about a parameter

Then calculate the likelihood of a given site after observing some data (posterior probability)

Chose tree with highest posterior probability

+ relatively fast, allows more complex models, gives both tree estimate and measure of uncertainty

- needs to specify prior information, very different run times for different trees

Evolutionary models:

Interpreters of phylogenetic info in a sequence. Each model makes different basic assumptions.

→ Which model for my data?

→ Did my model create reliable data?

1. Aminoacid substitution model
2. Codon based models
3. Models depending on secondary structure

Model selection:

jModelTest2:

http://code.google.com/p/jmodeltest2/

http://jmodeltest.org/

- Several selection criteria employing maximum likelihood (ML) scores
- Free
- Not ideal for Bayesian approaches

Protocol

Step 1.

Commonly → 16S rRNA gene

+ 100,000s of environmental sequences

+ Well developed trees

- 16S rRNA gene tends to break assemblers

- Not always present in MAGs

Step 2.

Concatenated ribosomal markers

+ Increased resolution of phylogeny

- Phylogenetic placements among genomes

- Genes need same evolutionary history

Step 3.

Collect marker genes from reference genomes Hug et al (2016).

→ 16 Ribosomal proteins:

L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, S19

Step 4.

One resource for ribosomal markers: RiboDB

Flavobacteria and Prochlorococcaceae →

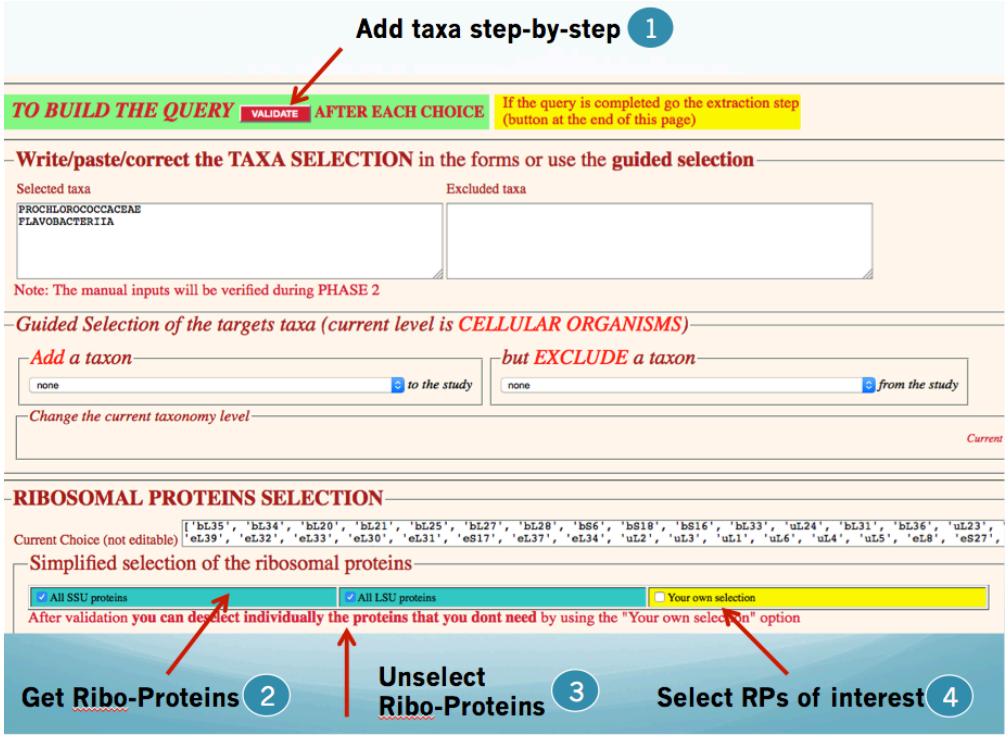
RpL14 (uL14), RpL22 (uL22), RpS8 (bS8), RpS18 (uS18)

LINK:
https://ribodb.univ-lyon1.fr/ribodb/ribodb-in.cgi

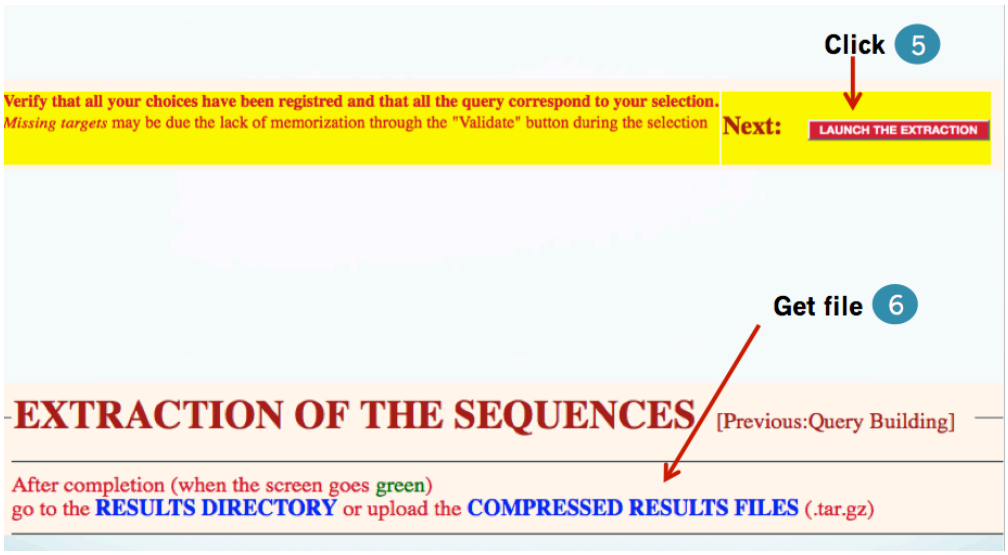
ANNOTATIONS
Elisha Wood-Charlson 05 Aug 2016

Paste in Flavobacteriia (yes, with 2 i's)

Step 5.



Step 6.



Step 7.

Move compressed file to /home/c-debi/ecogeo/phylogenetics

cmd COMMAND

```
$ mv R-PROTS.tar.gz /home/c-debi/ecogeo/phylogenetics
```

Step 8.

Decompress file:

cmd COMMAND

```
$ tar zxvf R-PROTS.tar.gz
```

NOTES
Elisha Wood-Charlson 13 Jul 2016

Contains 4 folders – bS18, uL14, uL22, uS8

Each contains a number of files – 3 nucleotide and 3 protein FASTA files

Step 9.

Copy the *_prot.fst files up two directories to /home/ecogeo/phylogenetics

cmd COMMAND
```
$ cp *_prot.fst ../../home/ecogeo/phylogenetics
```

ANNOTATIONS
Elisha Wood-Charlson 05 Aug 2016

Navigate to each folder and use:

$ cp *_prot.fst /home/c-debi/ecogeo/phylogenetics

Step 10.

Clean up header names & repeat:

cmd COMMAND
```
$ cut -f1 -d "~" bS18_prot.fst | sed 's/\.//' | sed 's/|//' > temp1_bS18
$ rm bS18_prot.fst
$ mv temp1_bS18 bS18_prot.fst
```

NOTES
Elisha Wood-Charlson 13 Jul 2016

When renaming:

Check that you get unique names!