LD12 genomes analysis pipeline

Lucas Sinclair, lucas.sinclair@me.com

Last modified:

Thu Oct 09 2014

All numbers are fake. This document describes a plan, it is not scientific work.

1) Start: with N partial genomes and the FASTA files containing their genes. We also have metadata and annotations for each gene.

#	Name	Source	# Genes	Family	Partial
G1	AAA288-N07	Freshwater (LD12)	560	1	yes
G2	AAA028-D10	Freshwater (LD12)	478	1	yes
G3	AAA023-L09	Freshwater (LD12)	609	1	yes
G4	HIMB140	Marine (SAR11)	1091	2	yes
G5	HTCC9565	Marine (SAR11)	951	3	no
G6	HIMB083	Marine (SAR11)	980	3	no

1B) Extract ribosomal genes: based on the annotations. This will act as the reference tree, these are supposed to be the less likely genes to be exchanged.

#	L29	S17	Rib3	RIb
G1	1	1	1	1
G2	1	1	0	1
G3	1	0	1	1
G4	1	1	1	1
G5	1	1	1	1
G6	1	1	1	1

2) All against all BLAST: Perform a sequence similarity search of all genes against all genes. WIth a e-value cutoff of 0.001

#	g1	g2	g3	g4	g5	g6	...
g1	–	1	89	0	112	1	...
g2	1	–	1	1	1	0	...
g3	45	1	–	1	44	1	...
g4	2	1	1	–	1	1	...
g5	87	0	12	1	–	1	...
g6	1	1	1	1	0	–	...
..	...	...	...	...	...	...	...

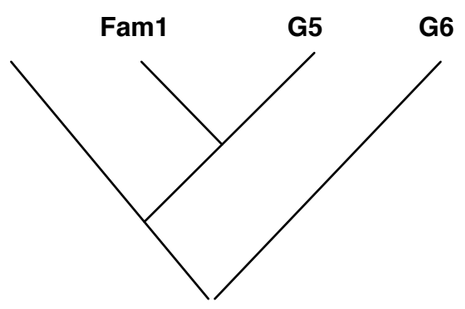
2B) Align each ribosomal protein cluster separately: run gblocks on the result, then concatenate each cluster together into one file, run that into RaxML

Multiple sequence aligement

```
P1| KVFGRCELAAAMKRHGLDNYRGYSLGNWVC----------
P3| --------AAKFESNFNTQATNRNTDGSTDYGILQIN-----
P4| RWWCN--DGRTPGSRNLCN-----IPCSAL------LSSDIT
P5| IWWCAK----------KITPDGNGMNAWVASRNRC-----
G3| PRAA-----AAAERLWSNFF-------EGTFFINKTEIEDAA
G6| -----YVVWQEVFDNKVKVRP-------DTIIQVWREEMPVE
G8| ---------RLWPRAGAVAERLWSSNLTTNIDF----AFKRL
H1| ----------------------------AAAMKRHGLDN--
H2| --KVFGRCELAAAM-------------------------
H4| --EVD--------FTCW------KS-----NP------E---
H5| ----------------------AGFRALLSAPWYL-------
H6| --QKT-----------------NRNTDGSTDY--------
H7| --------RWWCNDGRT----------PGS----------
```

Lorem

3B) Build a seperate tree: the tips can be mixed, if we didn't find a specific ribosomal prtein in one given genome, we can merge it with another genome of the same family and indicate the family on the tree leaf



3) Filter the BLAST hits: Any hit with a minimum identody of 30% and a minima coverage of 50%

#	g1	g2	g3	g4	g5	g6	...
g1	–	1	89	0	112	1	...
g2	1	–	1	1	1	0	...
g3	45	1	–	1	44	1	...
g4	2	1	1	–	1	1	...
g5	87	0	12	1	–	1	...
g6	1	1	1	1	0	–	...
.	...	...	...	...	...	...	...

4) Compte synteny scores: Come up a matrix of synteny distances

#	g1	g2	g3	g4	g5	g6	...
g1	–	1	89	0	112	1	...
g2	1	–	1	1	1	0	...
g3	45	1	–	1	44	1	...
g4	2	1	1	–	1	1	...
g5	87	0	12	1	–	1	...
g6	1	1	1	1	0	–	...
.	...	...	...	...	...	...	...

5) Merge both matrices: Using a elitic-type weighing equation (if the slips is a circle we are just doing 0.5*M1 + 0.5*M2)

#	g1	g2	g3	g4	g5	g6	...
g1	–	1	89	0	112	1	...
g2	1	–	1	1	1	0	...
g3	45	1	–	1	44	1	...
g4	2	1	1	–	1	1	...
g5	87	0	12	1	–	1	...
g6	1	1	1	1	0	–	...
.	...	...	...	...	...	...	...

6) Cluster: Using the MCL clustering, input the distance matrix and get a list of clusters of genes.

Clusters

C1 = g1+g3+g5
C2 = g2+g4
C3 = g6+g7+g8+g9

ETC...

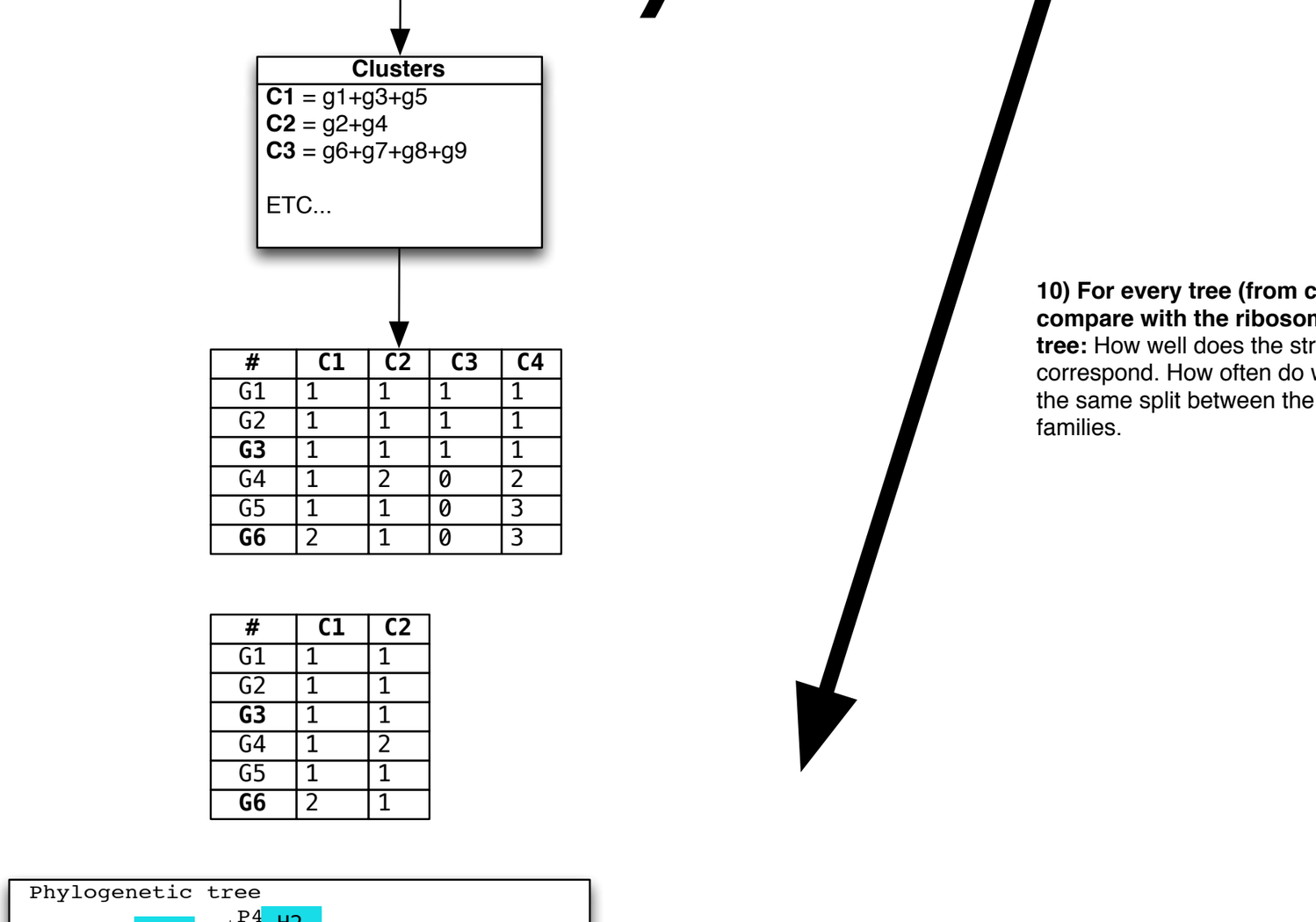
7) Count within genomes: Given each cluster, how many of its members are associated with each genome

#	C1	C2	C3	C4
G1	1	1	1	1
G2	1	1	1	1
G3	1	1	1	1
G4	1	2	0	2
G5	1	1	0	3
G6	2	1	0	3

8) Choose some cluster that we like: optional ?

#	C1	C2
G1	1	1
G2	1	1
G3	1	1
G4	1	2
G5	1	1
G6	2	1

9) Build a tree using the genes in the cluster: Check that the families group together. Report those that don't.



10) For every tree (from clusters) compare with the ribosomal tree: How well does the structure correspond. How often do we find the same split between the families.

0) Taking columns such C3: where a protein family exists in the freshwater, BLAST against "refseq +marine" database. How many hits are to SAR11, alphaproteobactera, proteobacteria, archea.