Projekt z Genomiki Porównawczej

Cel projektu

Jako cel założyłem jak najlepsze odtworzenie ewolucji cyjanobakterii na podstawie tej pracy. (https://www.frontiersin.org/articles/10.3389/fmicb.2019.01612/full#h9)

Wybrałem 18 gatunków:

- Acaryochloris marina MBIC11017
- Prochlorococcus marinus MIT9313
- Rivularia sp. PCC 7116
- Prochlorococcus sp. MIT 0801
- Gloeomargarita lithophora Alchichica-D10
- Anabaena cylindrica PCC 7122
- Gloeobacter violaceus PCC 7421
- Synechococcus elongatus PCC 6301
- Trichodesmium erythraeum IMS101
- Chroococcidiopsis thermalis PCC 7203
- Moorea producens 3L Ga0081465_101
- Gloeobacter kilaueensis JS1
- Synechococcus sp. JA-2-3B'a(2-13)
- Pleurocapsa sp. PCC 7327
- Thermosynechococcus elongatus BP-1
- Cyanothece sp. PCC 7822
- Cyanobium gracile PCC 6307
- Nostoc sp. PCC 7120

Pobieranie danych

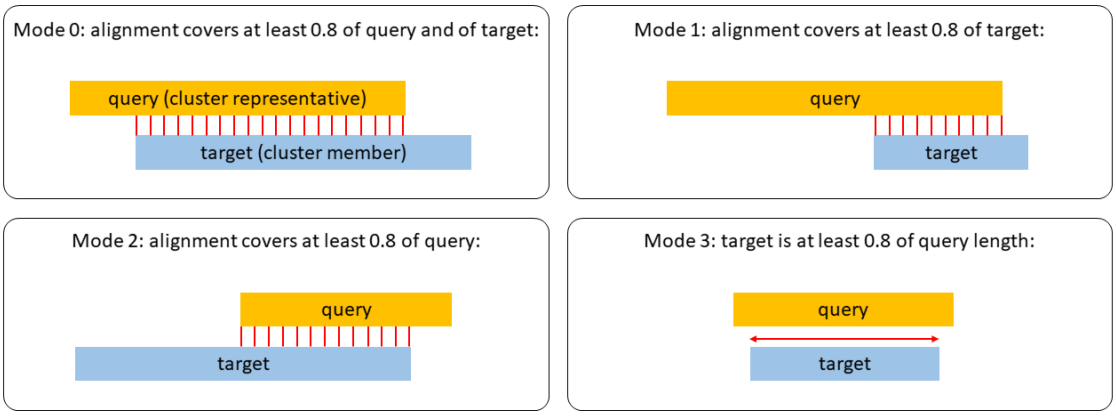
Sekwencje pobrałem korzystając z `BioPythona`.

Z plików `.gb` wyciągnąłem regiony kodujące.

Klastrowanie

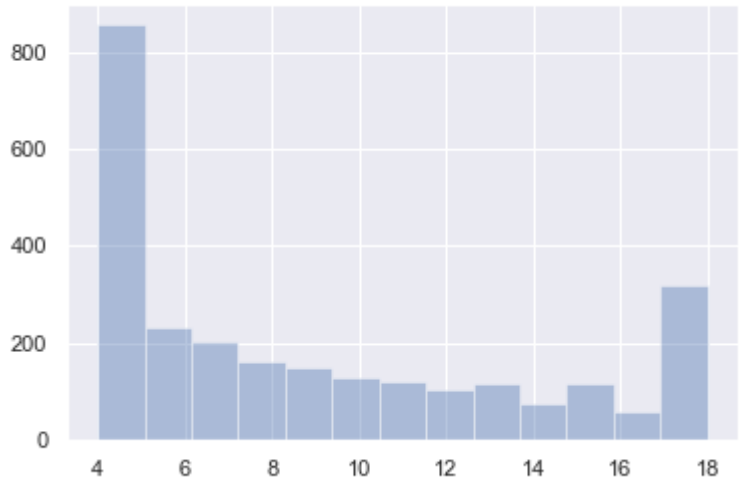
Następnie dokonałem klastrowania korzystając z programu `MMseqs2`.

Użyłem opcji `--cov-mode 1`, oznaczającej porównywanie pokrycia jak na rysunku:



W wyniku dostałem *35727* klastrów, z których usunąłem te, które miały mniej niż cztery sekwencje i takie, w których powtarzały się sekwencje z jednego gatunku.

Ostatecznie zostały *2633* klastry. Oto histogram liczności klastrów:



Przyrównywanie sekwencji

Do przyrównywania sekwencji w ramach klastrów użyłem programu `MAFFT` z opcją `--auto`.

Tworzenie drzew

Do inferencji drzew użyłem programu `RAxML` w wersji ósmej. Dla każdego klastra zrobiłem 100 drzew bootstrapowych z wykorzystaniem metody Maximum Likelihood.

Następnie korzystając z tego samego programu stworzyłem dla każdego klastra drzewo konsensusowe z użyciem konsensusu większościowego.

Filtrowanie drzew

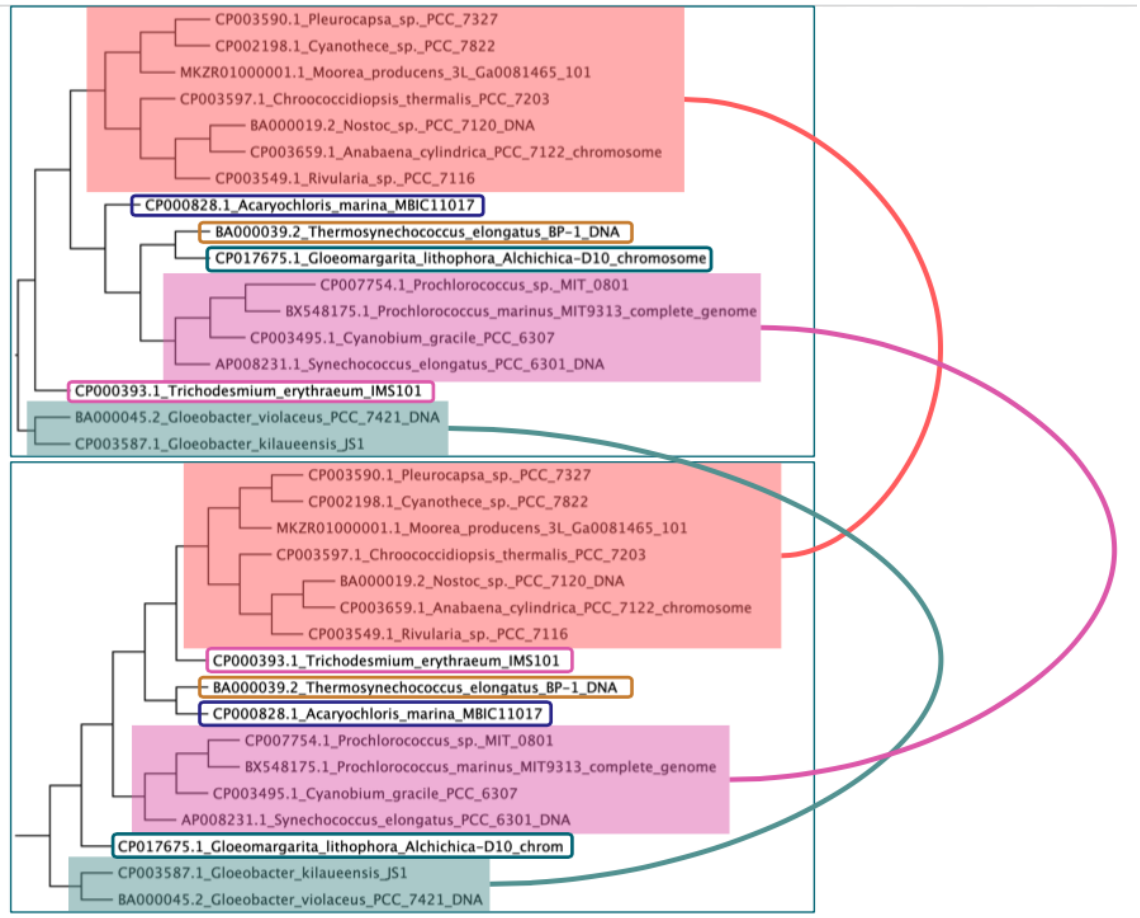
Na tym etapie usunąłem drzewa, które nie były binarne. Przez to niestety z analizy „znikł” jeden gatunek: Synechococcus sp. JA-2-3B a 2-13.

Inferencja drzewa gatunków

Do tworzenia drzewa genów użyłem `fasturec2`. Program uruchomiłem 100 razy i wybrałem drzewo z najmniejszym kosztem DL.

Wynik

Porównanie drzewa wynikowego i faktycznego jest tutaj:



Tutaj widać drzewa ukorzenione grupą zewnętrzną złożoną z *Gloeobacter kilauneensis* i *Gloeobacter violaceus*.

Całość analizy zajmuje około 7 godzin na 16 wątkach.

Można zauważyć, że trzy klady zostały dobrze zrekonstruowane, oznaczone są one kolorem turkusowym, różowym i ~pomarańczowym.

Wystąpiły jednak dość znaczące różnice w przypadku czterech z gatunków. Może być to spowodowane wieloma czynnikami:

- horyzontalnym transferem genów
- zbyt dużym zbiorem genów

Praca, na której się opierałem, wykorzystywała do liczenia drzew sekwencje rybosomalne, jako dobre markery filogenetyczne, ewoluujace w mniej więcej stałym tempie.

W przypadku blisko spokrewnionych gatunków (jak jest w tym przypadku), analiza całych genomów może być skomplikowana.