Отчет по выравниваниям последовательностей

Жерко Никита

1. Программы для анализа:

- muscle 5.1.osx64
- mafft v7.520 (2023/Mar/22)
- clustalo 1.2.4
- clustalw 2.1
- t-coffee 13.45.0.4846264-7
- prank v.170427
- emboss 6.6.0
- blastp 2.15.0+

2. Запускаем 6 алгоритмов выравнивания для 10 последовательностей.

```
time clustalw -INFILE=SUP35_10seqs.fa -OUTPUT=FASTA -OUTFILE=02_SUP35_10seqs.clustal
w.fa
# ...
# clustalw -INFILE=SUP35_10seqs.fa -OUTPUT=FASTA   4.04s user 0.03s system 99% cpu 4.
093 total
time muscle -in SUP35_10seqs.fa -out 02_SUP35_10seqs_muscle.fa
# ...
# muscle -align SUP35_10seqs.fa -output 02_SUP35_10seqs_muscle.fa  35.24s user 3.59s
system 619% cpu 6.272 total
time mafft --auto SUP35_10seqs.fa >02_SUP35_10seqs_mafft.fa
# ...
# mafft --auto SUP35_10seqs.fa > 02_SUP35_10seqs_mafft.fa  7.63s user 0.58s system 9
5% cpu 8.597 total
time kalign <SUP35_10seqs.fa >02_SUP35_10seqs_kalign.fa
# ...
# kalign < SUP35_10seqs.fa > 02_SUP35_10seqs_kalign.fa  0.01s user 0.00s system 54% c
pu 0.019 total
time t_coffee -infile=SUP35_10seqs.fa -outfile=02_SUP35_10seqs_tcoffee.fa -output=fas
ta_aln
time prank -d=SUP35_10seqs.fa -o=02_SUP35_10seqs_prank.fa -codon
# ...
# prank -d=SUP35_10seqs.fa -o=02_SUP35_10seqs_prank.fa -codon  932.73s user 3.19s sys
tem 76% cpu 20:20.49 total
```

02_SUP35_10seqs.clustalw.fa (https://tcoffee.crg.eu/apps/tcoffee/result?cached=true&rid=5e2286d5)

02_SUP35_10seqs_muscle.fa

02_SUP35_10seqs_mafft.fa

02_SUP35_10seqs_kalign.fa

02_SUP35_10seqs_tcoffee.fa

02_SUP35_10seqs_prank.fa

3. Сравним время работы наших алгоритмов для 10 последовательностей:

clustalw 4.093s

muscle 35.24s

mafft 8.597s

kalign 0.019s ?

t_coffee не получилось, но знаю, что долго)

prank 20min 20s

Prank и t-coffee работают очень долго, но также имеют свое применение.

4. Что не так с выравниванием SUP35_10seqs_strange_aln.fa и как это исправить?

В этом выравнивании одна из последовательность имеет низкий уровень сходства с референсной, мы заключили, что это из-за она обратно комплиментарна. Исправили это путем разворачивания последовательности.

5. Команды для запуска 6 возможных вариантов выравнивания, для 250 последовательностей ДНК.

```
time clustalw -INFILE=SUP35_250seqs.fa OUTPUT=FASTA -OUTFILE=05_SUP35_250seqs.clustal
w.fa
time muscle -in SUP35_250seqs.fa -out 05_SUP35_250seqs_muscle.fa
time mafft --auto SUP35_250seqs.fa >05_SUP35_250seqs_mafft.fa
time kalign <SUP35_250seqs.fa >05_SUP35_250seqs_kalign.fa
time t_coffee -infile=SUP35_250seqs.fa -outfile=05_SUP35_250seqs_tcoffee.fa -output=f
asta_aln
time prank -d=SUP35_250seqs.fa -o=05_SUP35_250seqs_prank.fa
```

6. Сравним время работы наших алгоритмов для 250 последовательностей:

clustalw за 7 минут выравнял 35 последовательностей

muscle больше 10 минут

mafft 33.568s total

kalign не получилось

t-coffee будет очень долго работать

prank будет очень долго работать

7. Команды для перевода в аминокислотные

последовательности:

```
transeq -sequence SUP35_10seqs.fa -outseq SUP35_10seqs.t.faa
getorf -sequence SUP35_10seqs.fa -outseq SUP35_10seqs.g.faa -noreverse -minsize 500
```

Проблемы могут возникнуть, если забыть указать нестандартный генетический код исследуемого организма (митохондриальнвя ДНК, инфузории).

8. Команды для запуска 6 возможных вариантов выравнивания для 10 белковых последовательностей

```
time clustalw -INFILE=SUP35_10seqs.g.faa -OUTFILE=08_SUP35_10seqs.clustalw.faa -OUTPU
T=FASTA -TYPE=protein
time clustalo --infile=SUP35_10seqs.g.faa --outfile=08_SUP35_10seqs.clustalo.faa --ve
rbose
time muscle -in SUP35_10seqs.g.faa -out 08_SUP35_10seqs_muscle.faa
time mafft --auto SUP35_10seqs.g.faa >08_SUP35_250seqs_mafft.fa
time kalign <SUP35_10seqs.faa >08_SUP35_10seqs_kalign.faa
time t_coffee -infile=SUP35_10seqs.g.faa -outfile=08_SUP35_10seqs_tcoffee.faa -output
=fasta_aln
time prank -d=SUP35_10seqs.g.faa -o=08_SUP35_10seqs_prank.faa
```

clustalw 0.387s clustalo 0.326s muscle 0.706s mafft 1.395s t-coffe prank больше минуты

10. Как добавить к выравниванию 250 нуклеотидных последовательностей ещё две (SUP35_2addseqs.fsa), предварительно выровняв их, с помощью mafft и muscle?

```
muscle -in SUP35_2addseqs.fa -out 10_SUP35_2addseqs_muscle.fa
muscle -profile -in1 05_SUP35_250seqs_muscle.fa -in2 SUP35_2addseqs.fa -out 10_SUP35_
252seqs_muscle.fa
mafft --auto SUP35_2addseqs.fa > 10_SUP35_2addseqs_mafft.fa
mafft --add SUP35_2addseqs_mafft.fa 05_SUP35_250seqs_mafft.fa > 10_SUP35_252seqs_maff
t.fa
```

11. Извлеките из NCBI с помощью любой вариации eutils все последовательности по запросу «Parapallasea 18S» (Parapallasea — это таксон, а 18S — это ген) и сохраните в файл fasta.

```
esearch -db nucleotide -query "Parapallasea 18S" | efetch -format fasta >Parapallasea
_18.fa
muscle -in Parapallasea_18.fa -out Parapallasea_18.fa.muscle.aln
mafft --auto Parapallasea_18.fa > Parapallasea_18.fa.mafft.aln
```

12. Команды для того, чтобы сформировать из набора последовательностей Ommatogammarus_flavus_transcriptome_assembly.fa базу для бласта, и для поиска в этой базе белковой последовательности Acanthogammarus_victorii_COI.faa с записью результатов в таблицу (текст с разделением табуляцией). Внимание: происхождение последовательности митохондриальное. Что важно учесть при поиске?

```
makeblastdb -in Ommatogammarus_flavus_transcriptome_assembly.fa -dbtype nucl -parse_seqids
tblastn -query Acanthogammarus_victorii_COI.faa -db Ommatogammarus_flavus_transcriptome_assembly.fa -outfmt 6 -db_gencode 5
# Acanthogammarus_victorii_COI  TRINITY_DN8878_c0_g1_i2  89.621  501  52  0  9  5093
# 1505  0.0  781
# Acanthogammarus_victorii_COI  TRINITY_DN58613_c0_g1_i1  50.000  20  9  1  206  22532  # 88  6.3  20.8
tblastn -query Acanthogammarus_victorii_COI.faa -db Ommatogammarus_flavus_transcriptome_assembly.fa -outfmt 6
# Acanthogammarus_victorii_COI  TRINITY_DN8878_c0_g1_i2  82.834  501  86  0  9  5093   # 1505  0.0  707
# Acanthogammarus_victorii_COI  N_20299_length_1167_cov_109.083184_g14733_i0  26.66745
# 33  0  397  441  458  324  1.7  22.7
# Acanthogammarus_victorii_COI  TRINITY_DN58613_c0_g1_i1  50.000  20  9  1  206  22532  # 88  6.4  20.8
```

Так происхождение митохондриальное, нужно указать `-db_gencode 5` .

Извлечем последовательность с наилучшим совпадением в отдельный файл.

```
blastdbcmd -db Ommatogammarus_flavus_transcriptome_assembly.fa -entry TRINITY_DN8878_c0_g1_i2 -out Ommatogammarus_flavus_COI.fa
```