Molecular Phylogeny III

1. Какие программы использовали для анализа? Укажите версии программ.

- clustalw - CLUSTAL 2.1
- muscle - MUSCLE v3.8.1551
- mafft - v7.490 (2021/Oct/30)
- kalign - kalign 3.3.1
- t_coffee - –
- prank - prank v.170427
- UGENE - UGENE 46.0
- EMBOSS - EMBOSS:6.6.0.0
- BLAST - BLAST 2.13.0+

2. Код для запуска 6 возможных алгоритмов выравнивания (clustalw, muscle, mafft,288,63 kalign, tcoffee, prank) для 10 последовательностей ДНК (SUP35_10seqs.fa) + вариации параметров, если они были. Если всё запускали онлайн — ссылки на страницы, можно приложить скриншоты, если это кажется осмысленным.

ClustalW:

```
clustalw -INFILE=SUP35_10seqs.fa -OUTPUT=FASTA -
OUTFILE=SUP35_10seqs.clustalw.fa
```

Формат вывода по умолчанию - очевидно, clustal, поэтому нужно сменить на FASTA.
MUSCLE:

```
muscle -in SUP35_10seqs.fa -out SUP35_10seqs_muscle.fa
```

The output is in .fasta format by default.
MAFFT:

```
mafft --auto SUP35_10seqs.fa > SUP35_10seqs_mafft.fa
```

The output is in .fasta format by default.
kalign

```
kalign < SUP35_10seqs.fa > SUP35_10seqs_kalign.fa
```

Формат вывода по умолчанию - .fasta
T-Coffee

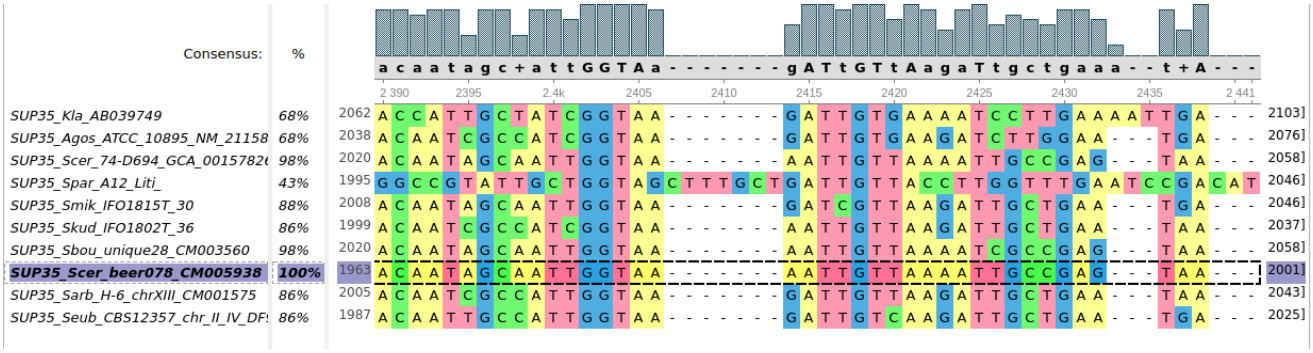
```
t_coffee -infile=SUP35_10seqs.fa -outfile=SUP35_10seqs_tcoffee.fa
```

Таблица со временем работы и качеством

	clustalw	muscle	mafft	kalign	tcoffee	prank
Time, s	5.90	4.89	4.12	0.28863	--	170.93
Quality	????	?????	??????		--	????
Length					--	

T-Coffe установить не удалось. Не ясно, что именно подразумевалось под качеством, если на семинаре говорилось, что общепринятых метрик качества множественных выравниваний на реальных данных не существует.

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



Последовательность Spar*A12_Liti* имеет очень плохое качество выравнивания. Причина, очевидно, в том, что это обратная комплиментарная последовательность. В UGENE имеется опция, позволяющая заменить последовательность на обратную комплиментарную. Сделав это и заново выровняв последовательности при помощи встроенного в UGENE MUSCLE, получаем гораздо лучшее качество выравнивания:

4. Команды / скриншоты для запуска 6 возможных вариантов выравнивания (см. п. 2), но для 250 последовательностей ДНК. Сравнительная таблица со временем работы и комментариями по поводу качества выравнивания 250 последовательностей ДНК (SUP35_250seqs.fa). Изменился ли наш выбор алгоритма?

```
muscle -in ./MolPhylo2023-3_data/SUP35_250seqs.fa -out
./250_seqs_together/SUP35_250seqs_muscle.fa

time kalign < ./MolPhylo2023-3_data/SUP35_250seqs.fa >
./250_seqs_together/SUP35_250seqs_kalign.fa

time clustalw -INFILE=./MolPhylo2023-3_data/SUP35_250seqs.fa -
OUTPUT=FASTA -OUTFILE=SUP35_250seqs_clustalw.fa

time mafft --auto ./MolPhylo2023-3_data/SUP35_250seqs.fa >
./250_seqs_together/SUP35_250seqs_mafft.fa
```

	clustalw	muscle	mafft	kalign	tcoffee	prank
Time, mm:ss:msms	Запускать его было ошибкой 41:80:??	2:45:??	0:40:96	00:05:51	--	Вы знаете, я даже не стал пробовать
Quality	????	?????	??????		--	????

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

```
muscle -in SUP35_2addseqs.fa -out SUP35_2addseqs_muscle.fa
muscle -profile -in1 SUP35_250seqs_muscle.fa -in2 SUP35_2addseqs.fa -out
SUP35_252seqs_muscle.fa
```

6. Как получить последовательности аминокислот (транслировать)? Пример команды для перевода в аминокислотные последовательности. Какие проблемы возникают.

Можно воспользоваться опцией ПКМ по последовательности → Экспорт → Экспортировать транслированное выравнивание в UGENE. Возможные проблемы

- неверная рамка считывания, альтернативный генетический код.
Также можно воспользоваться transeq из пакета EMBOSS:

```
transeq ./MolPhylo2023-3_data/SUP35_10seqs.fa
./10_seqs_together/SUP35_10seqs_to_protein.fa -frame 1
```

Здесь `-frame 1` означает, что рамка считывания начинается с первого нуклеотида первого кодона.

7. Команды / скриншоты для запуска 6 возможных вариантов выравнивания для 10 белковых последовательностей + вариации параметров, если они были.

Вместо clustalw следует использовать clustalo (ClustalΩ). В случае, если требуется использовать именно clustalw, нужно использовать специальный параметр `-TYPE=PROTEIN`

```
time clustalw -INFILE=SUP35_10seqs.g.faa -
OUTFILE=SUP35_10seqs.clustalw.faa -OUTPUT=FASTA -TYPE=protein
time clustalo --infile=SUP35_10seqs.g.faa --
outfile=SUP35_10seqs.clustalo.faa --verbose
time muscle -in SUP35_10seqs.g.faa -out SUP35_10seqs_muscle.faa
time mafft --auto SUP35_10seqs.g.faa > SUP35_10seqs_mafft.faa
time kalign <SUP35_10seqs.g.faa >SUP35_10seqs_kalign.faa
time t_coffee -infile=SUP35_10seqs.g.faa -
outfile=SUP35_10seqs_tcoffee.faa
time prank -d=SUP35_10seqs.g.faa -o=SUP35_10seqs_prank.faa
```

Сравнительная таблица со временем работы и комментариями по поводу качества выравнивания белков. Какой алгоритм лучше использовать?

	clustalw	muscle	mafft	kalign	tcoffee	prank
Time	732 ms	290 ms	374 ms	41 ms	--	128 s
Quality	????	?????	??????		--	????

Не знаю, но не prank

8. Извлеките из NCBI (с помощью любой вариации eutils или скриншоты, если делали в браузере) все последовательности по запросу «Parapallasea 18S» (Parapallasea — это таксон, а 18S — это ген) и сохраните в

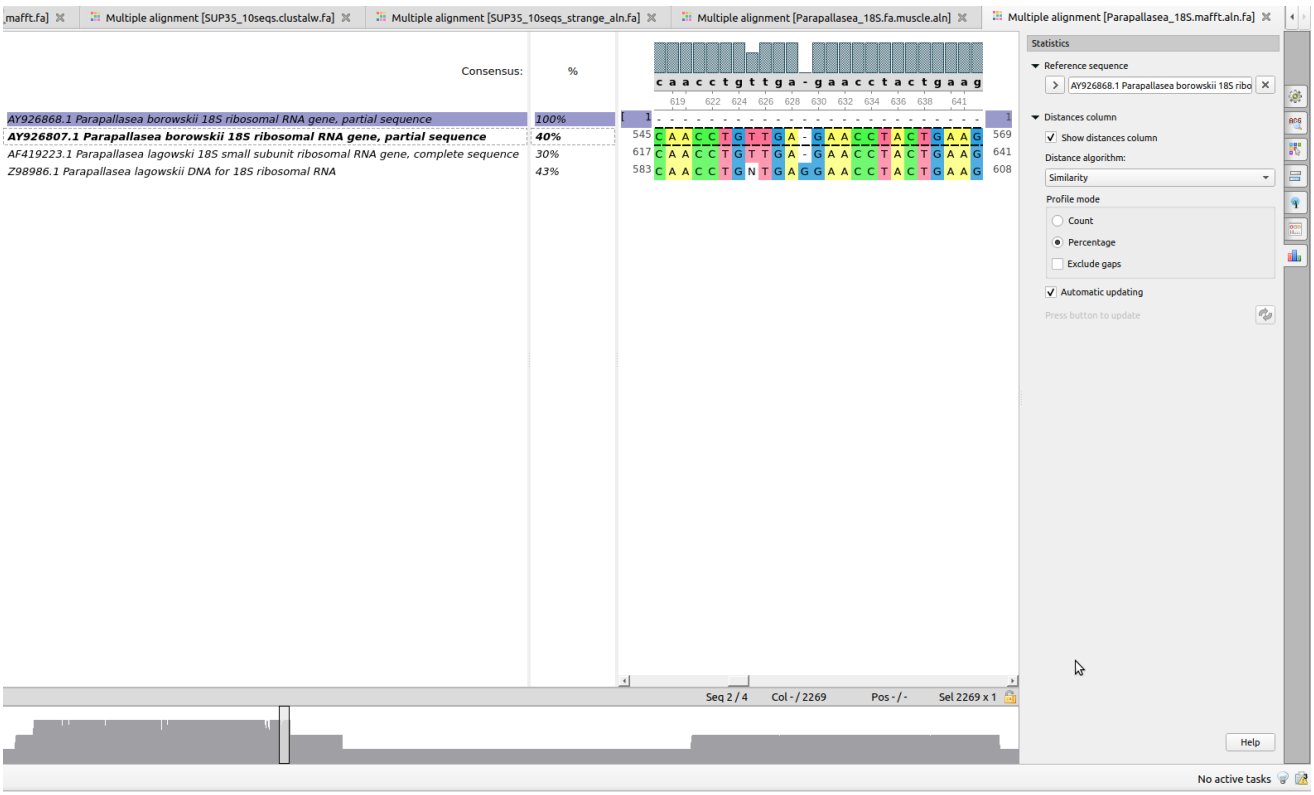
файл fasta. Что идёт не так при выравнивании последовательностей в файле Parapallasea_18S.fa и с какими параметрами можно получить правильный ответ?

```
esearch -db nucleotide -query "Parapallasea 18S" | efetch -format fasta > Parapallasea_18S.fa
```

Вывровняем последовательности при помощи muscle

```
muscle -in ./Parapallasea_18S.fa -out ./Parapallasea_18S.fa.muscle.aln
```

и посмотрим на выравнивание при помощи UGENE.



Вообще говоря, кажется, что всё идёт так. В файле присутствует как полная последовательность 18S ДНК, так и две частичные, одна из которых состоит из двух кусков, близеих к 5' и 3' концам. Кажется, что качество выравнивания, осуществуляемого mafft, заметно выше.

9. Приведите 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). Извлеките последовательность с лучшим совпадением в отдельный файл.

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

10. Внимание: происхождение последовательности митохондриальное. Что важно учесть при поиске?

Альтернативный генетический код. По умолчанию при поиске в базе данных BLAST используется стандартная таблица генетического кода. Однако первая

субъединица цитохром-оксидазы, с которой мы работам, кодируется частью митохондриального генома, с альтернативным генетическим кодом.
В BLAST генетический код, используемый при поиске, настраивается опцией `-db_gencode`, в нашем случае нужно использовать опцию `-db_gencode 5`.