# Отчет по молекулярной филогенетике № 3

Артем Васильев

22 февраля 2024 года

- 1. Какие программы использовали для анализа? Укажите версии программ.
- 2. Код для запуска 6 возможных алгоритмов выравнивания (clustalw, muscle, mafft, kalign, tcoffee, prank) для 10 последовательностей ДНК (SUP35_10seqs.fa) + вариации параметров, если они были. Если что-то запускали онлайн или в графическом режиме, приложите ссылки на страницы и/или скриншоты.
- 3. Составьте сравнительную таблицу со временем работы* и комментариями по поводу качества выравнивания ДНК для указанных выше алгоритмов**. Какой алгоритм лучше использовать?
- 4. Что не так с выравниванием SUP35_10seqs_strange_aln.fa и как это исправить?
- 5. Команды для запуска 6 возможных вариантов выравнивания (см. п. 2), но для 250 последовательностей ДНК.
- 6. Сравнительная таблица со временем работы и комментариями по поводу качества выравнивания 250 последовательностей ДНК (SUP35_250seqs.fa). Изменился ли наш выбор алгоритма?
- 7. Как получить последовательности аминокислот (транслировать)? Приведите пример команды для перевода в аминокислотные последовательности. Какие проблемы могут возникнуть?
- 8. Команды для запуска 7 возможных вариантов выравнивания для 10 белковых последовательностей + вариации параметров, если они были.
- 9. Сравнительная таблица со временем работы и комментариями по поводу качества выравнивания белков. Какой алгоритм лучше использовать?
- 10. Как добавить к выравниванию 250 нуклеотидных последовательностей ещё две (SUP35_2addseqs.fsa), предварительно
- 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. Что идёт не так при выравнивании последовательностей в файле Parapallasea_18S.fa и с какими параметрами можно получить правильный ответ?
- 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).
- САММАРИ: под конкретный проект и конкретную задачу придется **подбирать** алгос, который лучше всего подойдет.

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

- GUI программа **ugene-49.1** (для визуализации выравниваний + здесь очень много тулов)
- Программы для множественного выравнивания:
  - **clustalw 2.1**
  - **clustalo 1.2.4**
  - **muscle 3.8.31 / 5.1**
  - **mafft 7.520 / 7.149b**
  - **t-coffee 13.46.0.919e8c6b**
  - **prank v.170427**

    - **kalign2 v 2.04**
    - **kalign3 v 3.4.0**
- **emboss toolkit 6.6.0** — transeq
- **ncbi-blast+ 2.15.0**

# 2. Код для запуска 6 возможных алгоритмов выравнивания (clustalw, muscle, mafft, kalign, tcoffee, prank) для 10 последовательностей ДНК (SUP35_10seqs.fa) + вариации параметров, если они были. Если что-то запускали онлайн или в графическом режиме, приложите ссылки на страницы и/или скриншоты.

## - clustalw

Довольно древний алгос, который особо не пытались улучшить. Однако он часто стоит по-умолчанию в разных программах.

```
time clustalw -INFILE=SUP35_10seqs.fa -OUTPUT=FASTA -OUTFILE=02_SUP35_10seqs.clustal
w.fa
```

## - muscle

Чтобы использовать несколько ядер, нужна 5-я версия!

```
time muscle -in SUP35_10seqs.fa -out 02_SUP35_10seqs_muscle.fa
```

## - mafft

Если последовательностей очень много, то он выбирает алгоритм попроще, т.е. как бы меняется цена-качество.

```
time mafft --auto SUP35_10seqs.fa >02_SUP35_10seqs_mafft.fa
```

## - kalign

Мега-быстрый.

```
time kalign <SUP35_10seqs.fa >02_SUP35_10seqs_kalign.fa
```

## - t_coffee

Чай-кофе работает очень долго, но очень хорошо. Если другие алгосы работают плохо, то есть смысл пробовать его. То есть, если у нас последовательности плохо похожи друг на друга.

```
time t_coffee -n_core=8 -infile=SUP35_10seqs.fa -outfile=02_SUP35_10seqs_tcoffee.fa -
output=fasta_aln
```

-n_core=8 # кол-во используемых ядер

## - prank

Тоже работает очень долго (примерно в 2 раза дольше, чем чай-кофе). Автор заметил, что не всегда математически правильное выравнивание — биологически верное. Гораздо лучше справляется с выравниванием последовательностей по кодонам.

```
time prank -d=SUP35_10seqs.fa -o=02_SUP35_10seqs_prank.fa -codon
```

# 3. Составьте сравнительную таблицу со временем работы* и комментариями по поводу качества выравнивания ДНК для указанных выше алгоритмов**. Какой алгоритм лучше использовать?

Качество можно сравнивать по общей длине выравнивания или графическом представлении (например, как в UGENE), а также, оценивая не кратные 3 гэпы. При желании можно воспользоваться TCS: https://tcoffee.crg.eu/apps/tcoffee/do:core (https://tcoffee.crg.eu/apps/tcoffee/do:core)

```
progs_time_10
```

```
##          cpu, % total time, s align. length
## clustalw     99         4.079          2148
## muscle       99         2.314          2275
## mafft        97         0.452          2289
## kalign       97         0.151          2155
## t_coffee    724        52.273          2210
## prank        99       825.960          2373
```

Нет однозначного ответа, какой алгоритм лучше использовать. Многое зависит от 1) кол-ва последовательностей, 2) их длины, 3) отдаленности организмов друг от друга, 4) свободного времени исследователя. Небольшие комментарии оставлены в 1-м вопросе.

- Время выполнения при запуске в терминале можно посмотреть с помощью команды time.
- Если построение выравнивания требует неразумно большого времени, допустима запись вроде “>10 минут, надоело” :)

Если у нас будет не 10, а 250 последовательностей, то время увеличится очень сильно, и будет уже не до тонких настроек выравнивания.

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

Последовательность под названием *SUP35_Spar_A12_Liti_* очень сильно отличается от остальных, хотя это тоже *Saccharomyces*, и это тот же ген.

Дело в том, что она **обратно комплементарна**.

Как исправить:

- ПКМ по названию
- Edit
- Replace selected rows with reverse-complement
- ПКМ по пустому месту в окне выравнивания
- Align with MAFFT…
- ALign

# 5. Команды для запуска 6 возможных вариантов выравнивания (см. п. 2), но для 250 последовательностей ДНК.

- clustalw

```
time clustalw -INFILE=SUP35_250seqs.fa -OUTPUT=FASTA -OUTFILE=05_SUP35_250seqs.clustalw.fa
```

- muscle

```
time muscle -in SUP35_250seqs.fa -out 05_SUP35_250seqs_muscle.fa
```

- mafft

```
time mafft --auto SUP35_250seqs.fa >05_SUP35_250seqs_mafft.fa
```

- kalign

```
time kalign <SUP35_250seqs.fa >05_SUP35_250seqs_kalign.fa
```

- t_coffee

```
time t_coffee -n_core=8 -infile=SUP35_250seqs.fa -outfile=05_SUP35_250seqs_tcoffee.fa -output=fasta_aln
```

-n_core=8 # кол-во используемых ядер

- prank

```
time prank -d=SUP35_250seqs.fa -o=05_SUP35_250seqs_prank.fa
```

# 6. Сравнительная таблица со временем работы и комментариями по поводу качества выравнивания 250 последовательностей ДНК (SUP35_250seqs.fa). Изменился ли наш выбор алгоритма?

```
progs_time_250
```

```
##          cpu, % total time, s align. length
## clustalw     99        867.07          2179
## muscle       99        55.832          2313
## mafft       100        13.691          2322
## kalign       99          8.77          2214
## t_coffee      -             -             -
## prank         -             -             -
```

t-coffee (подождал час) и prank (даже не пытался запускать) работают слишком долго.

kalign выглядит очень приятно.

# 7. Как получить последовательности аминокислот (транслировать)? Приведите пример команды для перевода в аминокислотные последовательности. Какие проблемы могут возникнуть?

## - transeq

Просто делает трансляцию с 1 до последнего нуклеотида. Можно настроить рамку считывания и генетический код, который использует изучаемый организм (митохондрии многих организмов (например, инфузорий) используют нестандартнаый ген. код). Для ген. кода параметр `-table` , который стандартный везде.

```
transeq -sequence SUP35_10seqs.fa -outseq SUP35_10seqs.t.faa
```

## - getorf

Думает, что ему дали последовательность, в которой есть ORF (Met -> STOP, довольно банально), так что тоже нужно настраивать в более сложных случаях.

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

# 8. Команды для запуска 7 возможных вариантов выравнивания для 10 белковых последовательностей + вариации параметров, если они были.

## - clustalw

```
time clustalw -INFILE=SUP35_10seqs.g.faa -OUTFILE=08_SUP35_10seqs.clustalw.faa -OUTPU
T=FASTA -TYPE=protein
```

## - clustalo

Заточен под выравнивание белковых последовательностей.

```
time clustalo --infile=SUP35_10seqs.g.faa --outfile=08_SUP35_10seqs.clustalo.faa --ve
rbose
```

## - muscle

```
time muscle -in SUP35_10seqs.g.faa -out 08_SUP35_10seqs_muscle.faa
```

## - mafft

```
time mafft --auto SUP35_10seqs.g.faa >08_SUP35_250seqs_mafft.faa
```

## - kalign

```
time kalign <SUP35_10seqs.g.faa >08_SUP35_10seqs_kalign.faa
```

## - t_coffee

```
time t_coffee -n_core=8 -infile=SUP35_10seqs.g.faa -outfile=08_SUP35_10seqs_tcoffee.f
aa -output=fasta_aln
```

- prank

```
time prank -d=SUP35_10seqs.g.faa -o=08_SUP35_10seqs_prank.faa
```

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

```
progs_time_10aa
```

```
##          cpu, % total time, s align. length
## clustalw     99         0.224           719
## clustalo    539         0.278           757
## muscle       99         0.148           743
## mafft       105         0.253           759
## kalign      368         0.019           721
## t_coffee    688         6.464           752
## prank       100        65.990           776
```

- В файле с выравниванием после kalign нужно удалить шапку, иначе UGene не распознает файл как фасту.

Все алгоритмы показывают, что N-концевой домен (слева) вариабельный, а C-концевой домен — консервативный.

В данном случае можно прогнать чем угодно, т.к. последовательностей не много. По идее prank выдаст самый биологически правильный результат.

# 10. Как добавить к выравниванию 250 нуклеотидных последовательностей ещё две (SUP35_2addseqs.fsa), предварительно

выровняв их, с помощью mafft и muscle (эти +2 последовательности должны быть похожие в общем и целом)?

- 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

```
mafft --auto SUP35_2addseqs.fa > 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. 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 > Parapallase
a_18.fa
muscle -in Parapallasea_18.fa -out Parapallasea_18.fa.muscle.aln
mafft --auto Parapallasea_18.fa > Parapallasea_18.fa.mafft.aln
```

Проблема в том, что секвенирование кусками. 1 последовательность — полный ген, а 3 последовательности короткие, причем в 1 случае отсеквенирован конец гена, а в 2 других — его начало. В NCBI подобные штуки обычно не обозначают.

**mafft** в данном случае выровнял все, как надо, т.е. он работает с последовательностями, которые не сильно пересекаются, довольно неплохо.

clustalw справляется еще хуже.

t-coffee пытался, но получилось не очень :)

**prank** красава, сделал так же, как и mafft.

# 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?

Извлеките последовательность с лучшим совпадением в отдельный файл.

```
makeblastdb -in Ommatogammarus_flavus_transcriptome_assembly.fa -dbtype nucl -parse_s
eqids
tblastn -query Acanthogammarus_victorii_COI.faa -db Ommatogammarus_flavus_transcripto
me_assembly.fa -outfmt 6 -db_gencode 5
```

fields: qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore

`-parse_seqids` # говорит бласту, чтобы не трогал наши прекрасные названия

`-db_gencode 5` # чтобы учесть митохондриальное происхождение

Мы нашли совпадение на 501 а.к., скорее всего это нужный ген. Для выдергивания из самой БД можно воспользоваться такой командой:

```
blastdbcmd -db Ommatogammarus_flavus_transcriptome_assembly.fa -entry TRINITY_DN8878_
c0_g1_i2 -out Ommatogammarus_flavus_COI.fa
```

САММАРИ: под конкретный проект и конкретную задачу придется **подбирать** алгос, который лучше всего подойдет.