1 Предсказание вторичных структур для отдельных последовательностей

Анализ последовательности CP000627.1/583753-583892

Последовательность РНК:
GCUUGGCCUUAACUCCGAGCUUACCGCGCUAAGUUUA
AACCUUUAAAUAUGCGUUGUAAGCCAGUGACCGCUUG
UCACAAGGGCAGAAUUGGAAAUGAUUUUGCCUCCCGU
AUUUGGAAAGGUGUUCUGUGGCGCAACAA

В процессе вычислений были построены две вторичные структуры с использованием методов минимальной свободной энергии (MFE) и центроидного метода.

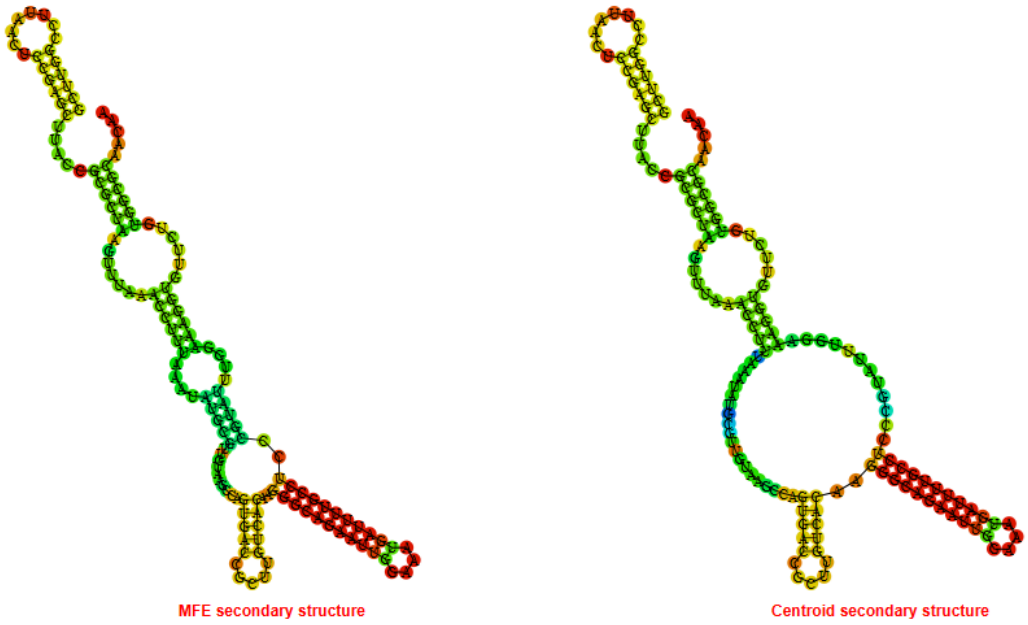


Рис. 1: Вторичные структуры CP000627.1/583753-583892

При анализе обеих структур, сгенерированных RNAfold, результат для CP000627.1/583753-583892 можно посмотреть здесь, Vienna record MFE, Vienna record Centroid, можно сделать следующие наблюдения:
При прогнозировании вторичной структуры РНК широко используются как метод минимальной свободной энергии (MFE), так и метод центроида, но они различаются по целям оптимизации и полученным структурам. Ниже приведен анализ этих методов на основе предоставленных данных:

- Структура MFE (минимальная свободная энергия) соответствует конфигурации, которая имеет минимально возможную свободную энергию. Для этой последовательности структура MFE имеет свободную энергию: −43.90 ккал/моль
((((((.....))))))....((((((..... .((((((....(((((.........(((((.....)
))))..((((((((((....))))))))))..)))))....))))))......))))))....

- Структура центроида представляет собой структуру, которая минимизирует расстояние до всех структур в термодинамическом ансамбле. Обычно ее считают средней структурой. Здесь структура центроида имеет более высокую свободную энергию по

сравнению с MFE −37.50 ккал/моль
((((((.....))))))).....((((((..... .(((((.....(((((.....)
))))..((((((((((....))))))))))..... .....)))))......))))))).....

Обозначения с точками и скобками показывают различия в моделях спаривания оснований. В структуре MFE существует более обширная конфигурация спаривания оснований, что приводит к более отрицательной (более низкой) свободной энергии. Однако структура центроида представляет собой конфигурацию, которая более типична для всего ансамбля, а не является наиболее энергетически стабильной.

Это различие возникает из-за того, что структура MFE направлена исключительно на минимизацию свободной энергии, в то время как структура центроида направлена на представление "средней"структуры в ансамбле, которая может иметь менее обширное спаривание оснований.

Термодинамический ансамбль:
Свободная энергия термодинамического ансамбля, которая является средневзвешенной свободной энергией по всем доступным структурам, рассчитывается как: −45.73 ккал/-моль

Это значение немного ниже энергии MFE, что предполагает, что ансамбль включает некоторые конфигурации с еще более низкой свободной энергией, чем сама структура MFE. Однако эти структуры встречаются с низкой вероятностью или являются частью менее стабильных областей в ансамбле. Кроме того, разнообразие ансамбля 35.27% указывает на значительную изменчивость возможных структур, отражая структурную гетерогенность в молекуле РНК.

Структура инвертированной последовательности:
Вторичная структура РНК зависит от последовательности оснований, поскольку каждое спаривание оснований и взаимодействие стекирования специфично для последовательности. Инвертированная последовательность полностью изменит потенциал спаривания оснований и, вероятно, приведет к другой вторичной структуре. Поэтому инвертированная последовательность вряд ли сформирует ту же самую структуру или будет иметь ту же самую свободную энергию, поскольку взаимодействия спаривания оснований и мотивы вторичной структуры будут перестроены.

Анализ последовательности CP000247.1/832790-832934

Последовательность РНК:
UUAACCACUAAACACUCUAGCCUCUGCACCUGGGUCA
ACUGAUACGGUGCUUUGGCCGUGACAAUGCUCGUAAA
GAUUGCCACCAGGGCGAAGGAAGAAAUGACUUCGCCU
CCCGUAUCUGGAAAGGUGUACAUGGCUUCACAAC

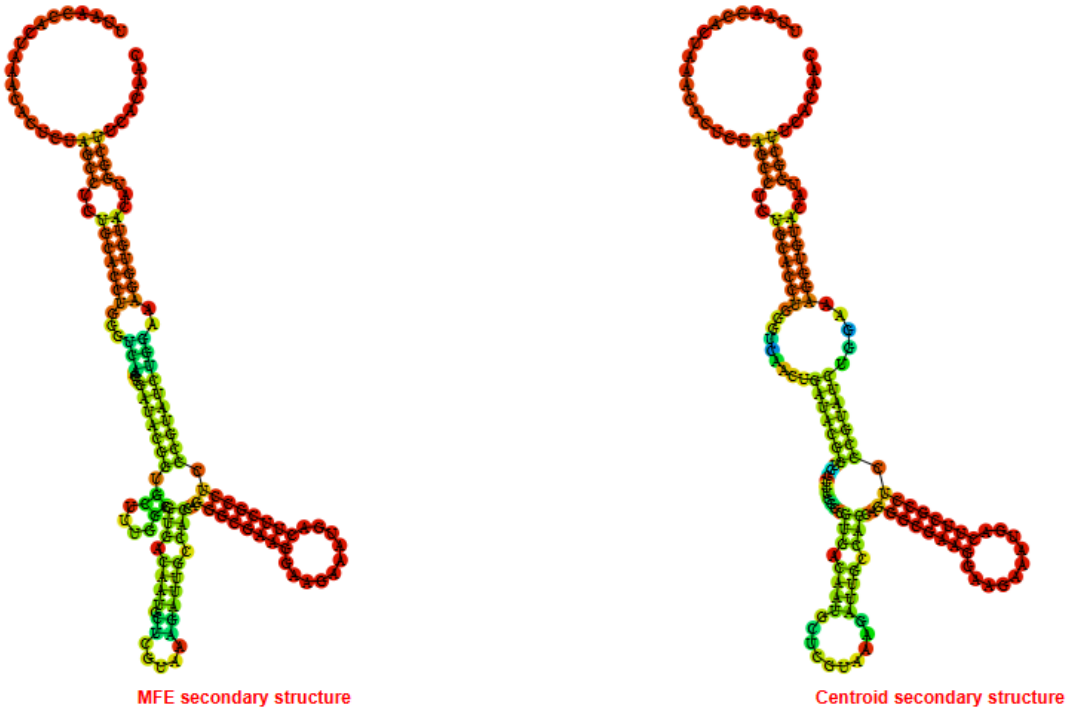


Рис. 2: Вторичные структуры CP000247.1/832790-832934

Результат и Vienna record MFE, Vienna record Centroid
Минимальная свободная энергия (MFE) и центроидные структуры предоставляют различные способы представления оптимальной вторичной структуры последовательности РНК:

- Структура MFE определяется путем нахождения конфигурации, которая дает минимально возможную свободную энергию для данной последовательности, в данном случае −41.50 ккал/моль. Эта структура отдает приоритет минимизации энергии и, как правило, уникальна, с парами оснований, расположенными для достижения минимальной энергии.
  .....((((..(((((((...(((...(((((((.((....)).(((.((((.((.....)
  )))))..)))..(((((((((.........)))))))) .)))))))))..)))))))...))))).....
- Центроидная структура представляет собой структуру с наибольшей вероятностью в термодинамическом ансамбле. Хотя она также стремится к низкой энергии, она учитывает вероятность по распределению потенциальных структур, в результате чего получается структура со свободной энергией −38.50 ккал/моль, что немного выше, чем MFE.
  .....((((..(((((((..... ...(((((((.........(((.((((.........
  .))))..)))..(((((((((.........)))))))) .)))))))).....)))))))...))))).....

Различия между структурами MFE и центроидными структурами возникают из их критериев оптимизации. Структура MFE фокусируется на минимизации энергии независимо от стабильности всего ансамбля, в то время как структура центроида оптимизирует для наиболее вероятной конфигурации в ансамбле, включая стабильность по возможным флуктуациям. Следовательно, структура центроида может демонстрировать вариации в конфигурациях петель или спаривании оснований, которые обеспечивают баланс между энергией и структурной стабильностью, отсюда ее немного более высокая энергия.

Термодинамический ансамбль:
Энергия термодинамического ансамбля, рассчитанная как -43,32 ккал/моль, является средней метрикой, отражающей все вероятные структуры, которые может принять РНК. Эта энергия ансамбля обычно лежит между значениями центроида и MFE и дает представление об общей структурной стабильности в различных конфигурациях, а не о какой-либо одной структуре.

Структура инвертированной последовательности:
Инвертирование последовательности РНК, скорее всего, не приведет к идентичной структуре с теми же энергетическими свойствами. Формирование вторичной структуры в значительной степени зависит от специфичных для последовательности пар оснований и стековых взаимодействий, которые не полностью отражаются при инверсии. Следовательно, инвертированная последовательность может принять отчетливую вторичную структуру с различными характеристиками стабильности и складчатости.

2 Предсказание структуры на основе выравнивания (align&fold)

Метод align&fold позволяет оценить, какие элементы структуры сохраняются, а какие изменяются. Выравнивание позволяет выявить общие черты между последовательностями и предсказать вторичную структуру для их консенсуса.

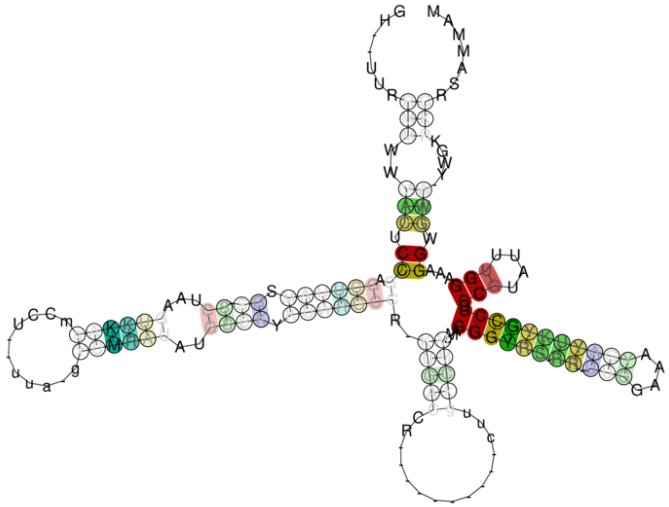


Рис. 3: Реализация метода align&foldдля РНК с вторичными структурами

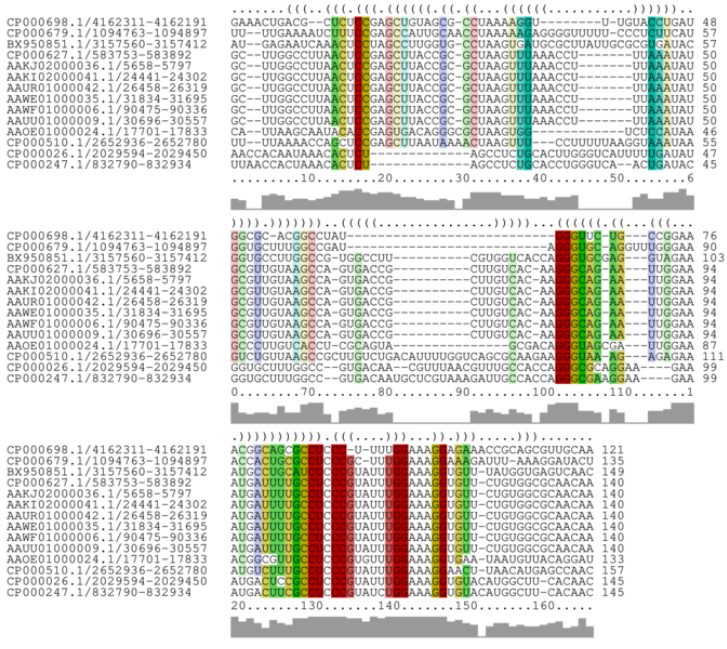


Рис. 4: Выровненные последовательности вторичных структур для РНК

Результаты (также можно посмотреть тут) выравнивания показали:

- Идентичные нуклеотиды, например, локусы GG, обозначены красным.
- Желтым отмечены места, где нуклеотиды различаются.

3 Сравнение структур, построенных двумя методами

Ipynb, Plot, был использован следующий набор последовательностей: combined_sequences.fasta (приложен результат 6го пункта)

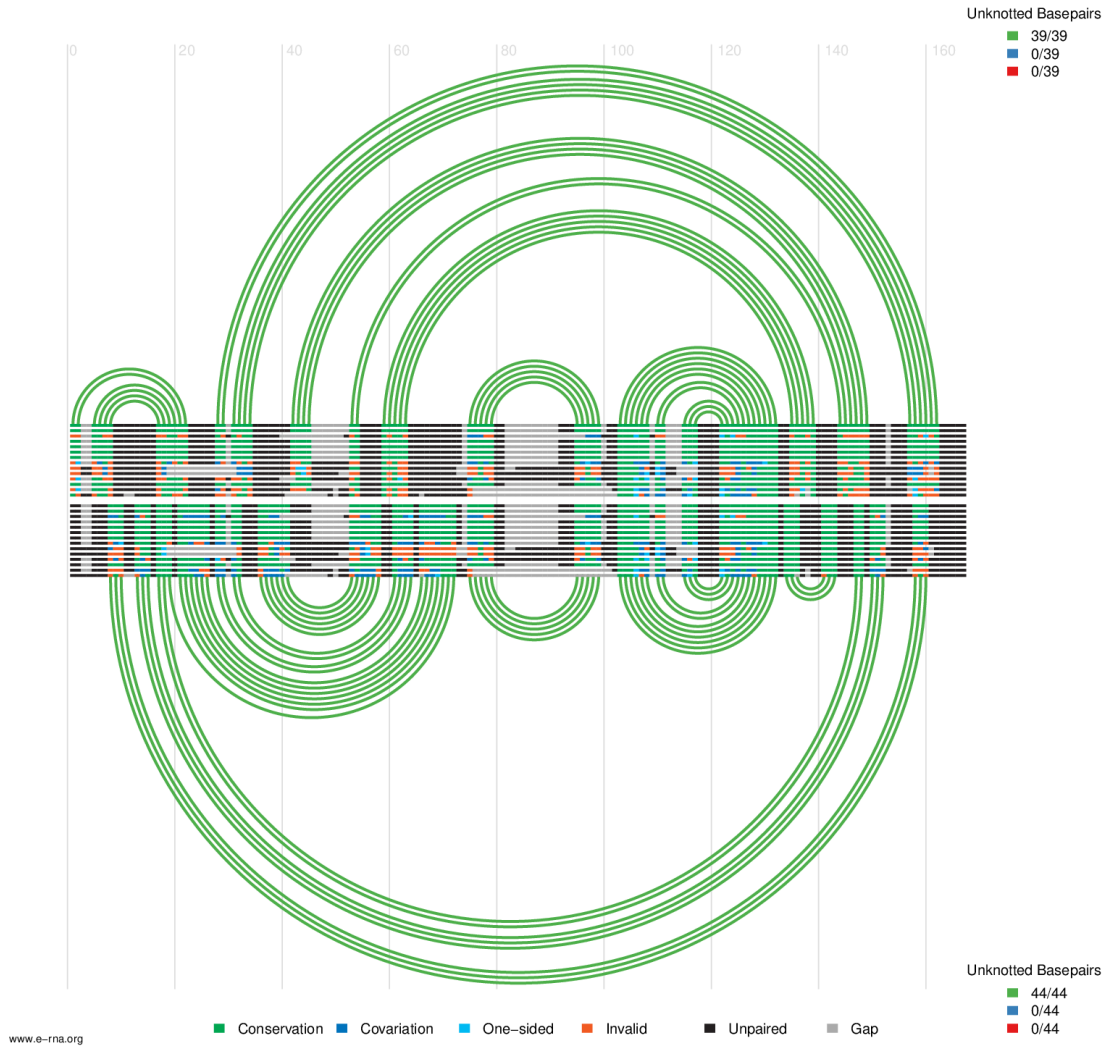
На представленной ниже визуализации видно, что структуры РНК для обеих последовательностей демонстрируют высокий уровень консервативности. Консервативные элементы, обозначенные зелёным цветом, указывают на области, которые сохраняют свою структуру в обеих РНК, что свидетельствует о значимой роли этих участков.

На графике также отображены "неперекрещенные"парные основания, что подтверждает успешное отображение общих черт структуры благодаря совместной обработке файлов, полученных на этапе Align&Fold. Эти пары демонстрируют устойчивые структурные элементы, такие как шпильки, которые сохраняются между последовательностями.

Для улучшения восприятия графического вывода программы было выполнено обрезание изображения, чтобы точнее рассмотреть количество спаренных нуклеотидов и их расположение. На основе визуализации можно отметить следующее:

- Верхняя часть отображает неперекрещенные парные основания, представляющие первую последовательность из предыдущего задания.
- Нижняя часть показывает результаты выравнивания для 14 РНК-последовательностей, что позволяет увидеть общие элементы, характерные для данной группы.

Также отмечены unknotted base pairs, свидетельствуя о том, что совместная загрузка файлов прошлого этапа, отразила общие черты структуры.



4 Оценка структуры

Загрузив "results.stk" в R-scape получим следующий файл (также результаты можно посмотреть тут), в котором будет написано 'no significant pairs'

Для других же наборов результат может быть более информативным, например: Такой

List of covarying basepairs

in given structure	Left base	Right base	Covariation Score	E-value	Substitutions	Power
~	88	102	24.69549	0.00423117	11	0.36

Significantly covarying pairs present in the structure are marked green. Other covarying pairs are marked orange if both residues are unpaired in the structure or there is no structure present. Orange covarying pairs could be an indication of an under annotated structure or pseudoknots. Black covarying pairs could indicate covariation supports for an alternative structure, tertiary interactions, or false positives.

* Base pair in the structure

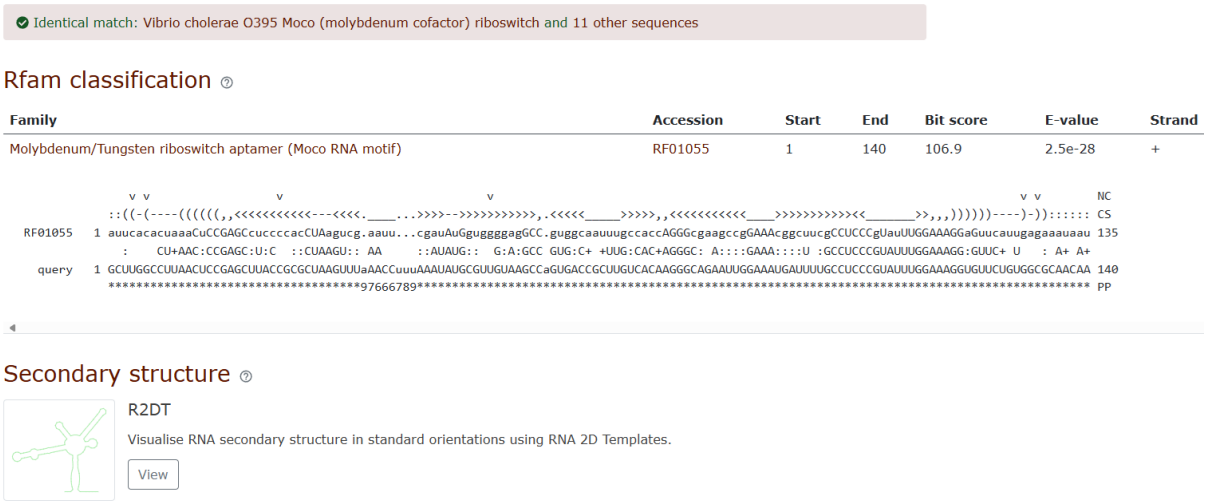
~ Both residues unpaired in the structure, or no structure is present

At least one residue is involved in other pairing in the structure

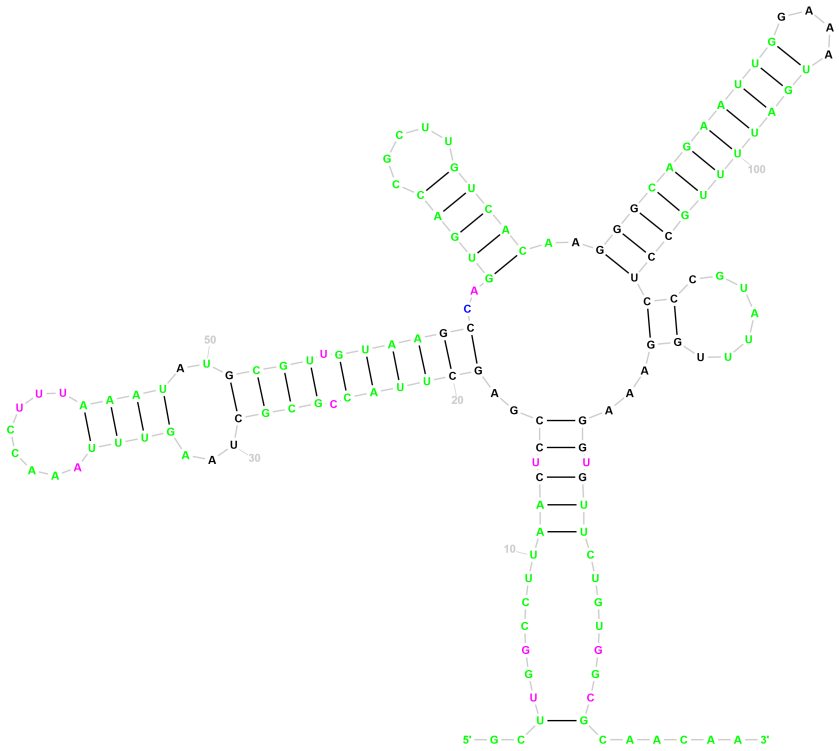
результат говорит о функциональной целостности вторичной структуры РНК, показывая, что пространственная организация остаётся устойчивой, и подтверждая структурную консервативность.

5 Поиск известной структуры по последовательности

Рассмотрим CP000627.1/583753-583892, загрузив эту последовательность в Rfam, получим:
Identical match: Vibrio cholerae O395 Moco (molybdenum cofactor) riboswitch and 11 other sequences



E-value при Identical match равен 2.5×10^{-25}, что указывает на высокую вероятность принадлежности последовательности к данному семейству. Bit score 106.9 подтверждает схожесть структуры со следующей структурой (которую также можно посмотреть тут):



Визуальный и структурный анализ показывает, что структура, полученная с помощью метода минимальной свободной энергии (MFE), и структура, полученная с помощью множественного выравнивания последовательностей, имеет некоторые сходства. Консервативные шпильки и непересекающиеся пары оснований, выявленные во время выравнивания последовательностей, выравниваются по форме и положению с элементами структуры Moco РНК, подтверждая их структурную целостность.

Однако важно отметить, что в нашем случае результаты немного различаются. Это несоответствие можно объяснить значительной изменчивостью внутри ансамбля. Хотя такие методы, как MFE и алгоритмы выравнивания, могут воспроизводить функционально значимые мотивы, их базовые алгоритмы прогнозирования вторичной структуры РНК не всегда могут давать согласованные результаты для разных последовательностей. Таким образом, хотя наши результаты немного различаются для рассмотренных последовательностей, они могут согласовываться при применении к другим последовательностям, или же различаться еще сильнее.

6 Код

Получение fasta файлов на основе "results.stk"

Code

```
# Load the Stockholm file
stk_file_path = '/content/result.stk'
with open(stk_file_path, 'r') as file:
    stockholm_data = file.readlines()

# Extract all IDs from the file
# IDs are typically the first word in lines containing sequences
id_pattern = re.compile(r'^[A-Za-z0-9./-]+')  # Pattern to match the ID format
ids = set()  # Using a set to store unique IDs

for line in stockholm_data:
    match = id_pattern.match(line)
    if match and not line.startswith(("#", "//")):  # Ignore metadata and terminators
        ids.add(match.group())

# Convert the set to a sorted list for consistent processing
ids = sorted(ids)

# Filter sequences for each ID and write to .fasta files
def to_fasta_file(id, sequences, output_folder):
    fasta_filename = f"{output_folder}/{id.replace('/', '_')}.fasta"
    with open(fasta_filename, 'w') as fasta_file:
        fasta_file.write(f">{id}\n")
        for seq in sequences:
            sequence = seq.split(" ", 1)[1].strip()  # Extract the sequence portion
            fasta_file.write(sequence + "\n")
    return fasta_filename

# Create an output folder
output_folder = "/content/fasta_files"
import os
os.makedirs(output_folder, exist_ok=True)

# sequences and write to FASTA files
created_files = []
for id in ids:
    id_sequences = [line.strip() for line in stockholm_data if line.startswith(id)]
    if id_sequences:  # If sequences for the ID exist
        fasta_file = to_fasta_file(id, id_sequences, output_folder)
        created_files.append(fasta_file)

# Output the list of created FASTA files
print(f"FASTA files created for IDs: {', '.join(ids)}")
print(f"Files are saved in {output_folder}")
```

Далее разберемся с просмотром строк через командную строку:

Code

```
# Display lines starting with the first ID
!grep "^CP000627.1/583753-583892" /content/result.stk

# Display lines starting with the second ID
!grep "^CP000247.1/832790-832934" /content/result.stk
```

Получим:

```
CP000627.1/583753-583892 GCUUGGCCUUAACUCCGAGCUUACCGCGCUAAGUUUAAACCU---UUA---AAUAUG---------CGUUGUAAGCCAGUGACCGCUUGUCACAAGGGCAGAAUUGGAAAUGAUUUUGC
CP000627.1/583753-583892 CUCCCGUAUUUGGAAAGGUGUUCUGUGGCGCAACAA
CP000247.1/832790-832934 UUAA---------CCACUAAACACUCUAGCC--UCUGCACCUGGGUCAACUGAUACGGUGCUUUGGCCGUGACAAUGCUCGUAAAGAUUGCCACCAGGGCGAAGGAAGAAAUGACUUCGC
CP000247.1/832790-832934 CUCCCGUAUCUGGAAAGGUGUACAUGGCUUCACAAC
```

Code

```
# Construct the exclusion pattern dynamically from the IDs
!grep -v -e "^CP000698.1/4162311-4162191" \
        -e "^CP000679.1/1094763-1094897" \
        -e "^BX950851.1/3157560-3157412" \
        -e "^CP000627.1/583753-583892" \
        -e "^AAKJ02000036.1/5658-5797" \
        -e "^AAKI02000041.1/24441-24302" \
        -e "^AAUR01000042.1/26458-26319" \
        -e "^AAWE01000035.1/31834-31695" \
        -e "^AAWF01000006.1/90475-90336" \
        -e "^AAUU01000009.1/30696-30557" \
        -e "^AAOE01000024.1/17701-17833" \
        -e "^CP000510.1/2652936-2652780" \
        -e "^CP000026.1/2029594-2029450" \
        -e "^CP000247.1/832790-832934" /content/result.stk | wc -l
```

Получим:

```
6
```

Объединим fasta для удобства:

Code

```
# Concatenate both FASTA files into one
!cat /content/fasta_files/*.fasta > /content/combined_sequences.fasta
```

Получим следующий файл: combined_sequences.fasta