Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт (государственный университет)»

Факультет инноваций и высоких технологий
Кафедра алгоритмов и технологий программирования

**Направление подготовки / специальность**: 01.04.02 Прикладная математика и информатика
**Направленность (профиль) подготовки:** Алгоритмы и технологии программирования

# КЛАССИФИКАЦИЯ ТРИПЛЕКСОВ В СТРУКТУРАХ РНК

магистерская диссертация

**Обучающийся:** Медведева Анастасия Дмитриевна

_____

**Научный руководитель:**
Драль Алексей Александрович,
Старший преподаватель
кафедры АТП ФИВТ МФТИ

_____

Москва 2018

АННОТАЦИЯ

Рибонуклеиновая кислота — один из важнейших классов макромолекул в клетках всех живых организмов, выполняющих множество функций, например, регуляция экспрессии генов. РНК способна принимать множество конфигураций, и от их вида зависит, какие функции выполняет конкретная молекула, поэтому изучение пространственной структуры РНК позволяет делать предсказания о том, в какие взаимодействия способны вступать разные участки молекулы. Одним из важнейших мотивов третичной структуры РНК является триплекс — тройка нуклеотидов с водородными связями между ними.

В данной работе представлена классификация триплексов в молекулах РНК, объединяющая как геометрические, так и структурные свойства участвующих нуклеотидов. Данная классификация позволила выявить новый редкий класс триплексов, который представляет значительный интерес с точки зрения молекулярной биологии.

# ВВЕДЕНИЕ

РНК (рибонуклеиновая кислота) — один из важнейших классов молекул, содержащихся в клетках всех живых организмов. Существует несколько типов РНК, различающихся выполняемыми функциями. На данный момент РНК являются предметом исследований, в частности, изучение структур в этих молекулах позволяет понять, какие функции они выполняют, как они сворачиваются, и какую биологическую роль играют. Функции, выполняемые той или иной молекулой РНК, зависят от её пространственной структуры[1-3]. Молекулы РНК включают три уровня пространственной организации:

1. Первичная структура — последовательность нуклеотидов, задаётся матрицей ДНК
2. Вторичная структура — множество комплементарных спариваний оснований
3. Третичная структура — координаты атомов в пространстве + множество всех типов внутримолекулярных взаимодействий

Цепочки нуклеотидов, составляющие первичную структуру РНК, образуются из четырёх азотистых оснований: A (аденин), C (цитозин), G (гуанин), U (урацил).

Общепринятым методом описания вторичной структуры является “модель ближайших соседей” (Nearest Neighbor Model)[4].

Одним из важнейших  элементов третичной структуры является триплекс. Он представляет из себя тройку нуклеотидов, образующих между собой водородные связи по одному из рёбер. Такие структуры составляют или стабилизируют другие важные третичные мотивы. Можно выделить несколько самых частых:

- Tetraloop-receptor — стабилизируют молекулу РНК, могут исполнять роль сайтов узнавания для белков [5]
- A-minor — стабилизируют взаимодействия между петлями и цепочками РНК [6]

В статье [7] даётся определение триплекса как тройки оснований, расположенных в одной геометрической плоскости. При этом не все нуклеотиды из этой тройки формируют водородные связи с другими частями триплекса. Далее, авторы строят классификацию исходя из геометрических признаков каждого триплекса: того, какие рёбра участвуют в спариваниях, и как основания расположены друг относительно друга.

На основе такого описания строится классификация триплексов по типам спариваний (см. Рис.1), и в дальнейшем описываются все классы, найденные в банке данных Protein Data Bank (PDB, [8]) на момент написания статьи.

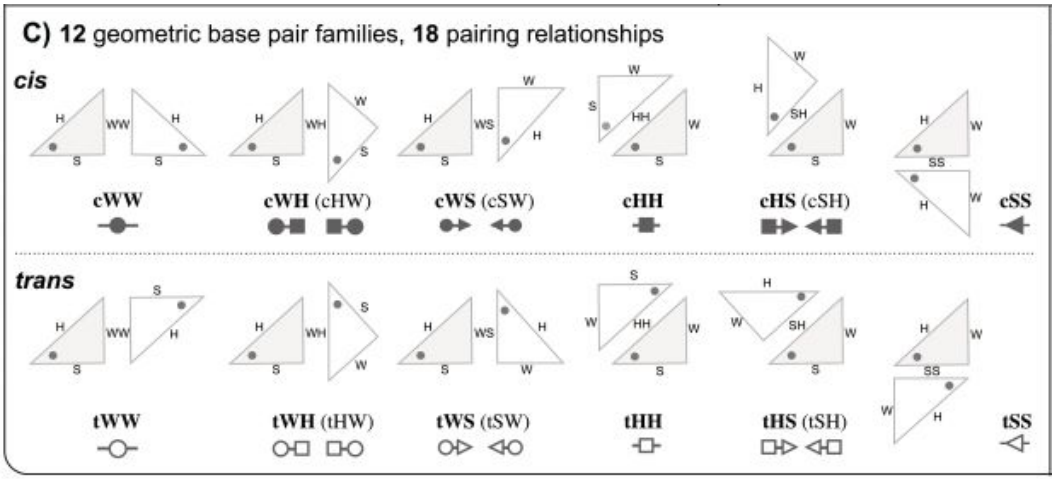


Рис 1. Схематическое обозначение классификации по типам спариваний.

В работе [9] была построена классификация внутренних петель, в том числе по отношению к триплексам, содержащимся в них. Это позволило авторам выделить несколько классов вторичных структур (петель), в зависимости от положения

неспаренного нуклеотида из триплекса по отношению к паре нуклеотидов из стема. Использование такого описания позволяет учитывать расположение триплекса относительно ближайшего стема.

В данной работе построена новая классификация триплексов оснований в РНК, использующая достоинства предыдущих наработок в этой области. Применение описываемой классификации к накопленным в ходе исследований данным позволило выделить новый класс структур, ранее не встречавшийся в печатных источниках. Целью данной работы было построение классификации и исследование различных классов структур, полученных  после применения классификации к данным из PDB . Предметом исследования были данные о структурах РНК из PDB, находящиеся в свободном доступе.

# ОСНОВНАЯ ЧАСТЬ

## Материалы и методы

### Основные определения

Как было указано выше, молекулы РНК обладают различными пространственными структурами, что приводит к их разделению на разные функциональные группы. Точное количество групп и функции, ими выполняемые, до сих пор являются предметами исследований, но среди уже известных можно выделить следующие виды:

1. Матричные РНК — содержат информацию о первичной структуре белков. мРНК синтезируются в ходе транскрипции и выполняют функции переносчиков генетической информации от ДНК к рибосомам [10].
2. Транспортные РНК — переносят аминокислоты для синтеза пептидных связей в рибосоме. Для различных аминокислот существуют свои типы тРНК [11].
3. Рибосомальные РНК — являются составной частью рибосом, участвующих в синтезе белков [12].
4. РНК, участвующие в подавлении экспрессии генов — этот небольшой класс РНК в составе белкового комплекса RISC (RNA-induced silencing complex) разрушает мРНК и препятствует дальнейшему синтезу белков по конкретной матрице [13].
5. Малые ядерные РНК — участвуют в некоторых клеточных процессах, например, в удалении интронов из мРНК (сплайсинге) [14].

Остановимся подробнее на классе рибосомальных РНК, так как третичные мотивы, найденные в этих РНК, представляют особый интерес в рамках данной работы.

Рибосомы — это молекулярные комплексы, выполняющие функцию синтеза протеинов и присутствующие во всех живых клетках, как у прокариотических организмов, так и у эукариотических. В состав рибосом входят упомянутые выше рибосомальные РНК и различные протеины. Структурно рибосомы представляют комплекс из малых и больших субъединиц, различающихся по размерам у прокариот и эукариот. Размеры этих комплексов в единицах СИ — от 200 до 300 нм, размеры в Сведбергах [15] представлены в таблицах 1 и 2.

Таблица 1. Размеры рибосомальных субъединиц в прокариотах.

| Размер рибосомы | Размер субъединицы | Размеры рРНК |
|---|---|---|
| 70S | Большая: 50S | 23S |
| | | 5S |
| | Малая: 30S | 16S |

Таблица 2. Размеры рибосомальных субъединиц в эукариотах.

| Размер рибосомы | Размер субъединицы | Размеры рРНК |
|---|---|---|
| 80S | Большая: 60S | 28S |
| | | 5.8S |
| | | 5S |
| | Малая: 40S | 18S |

Также, существуют отдельные рибосомы в митохондриях животных клеток, рибосомальные РНК малых субъединиц в которых могут иметь вес 12S или 15S.

Как указывалось выше, молекулы РНК имеют первичную, вторичную и третичную структуру. На Рис.2 представлено описание элементов вторичной структуры, формирующихся из пар комплементарных оснований.

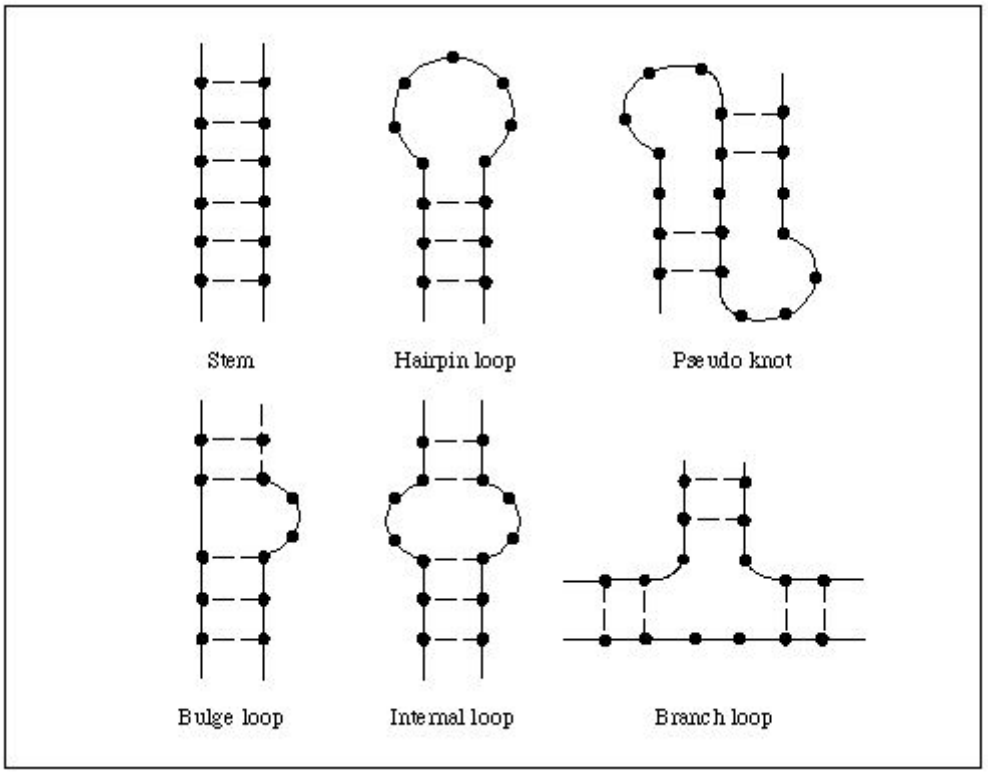


Рис. 2. Схематичное отображение вторичных структур РНК.

Азотистые основания и водородные связи между ними будем рассматривать в виде графа, где основания выполняют роль вершин, пронумерованных от 1 до L — длины цепочки оснований, — в направлении от 5’-конца до 3’-конца. Тогда *спаривание оснований* РНК можно определить следующим образом:

**Спаривание оснований** — это пара нуклеотидов (i, j), i < j, с водородными связями между ними. Существует несколько типов спариваний:

1. Пары комплементарных оснований, или Уотсон-Криковские пары (AU и GC)
2. Неканонические пары оснований:
    a. Wobble-спаривания (GU)
    b. Хугстиновские пары (например, GA)

c. Нетипичные спаривания, содержащие не менее двух водородных связей

Пары оснований (m, n) и (p, q) называются *конфликтующими*, если $m < p < n < q$ или $p < m < q < n$.

Две пары оснований и связи между ними образуют в графе вторичной структуры элементарный цикл. В свою очередь, элементарные циклы, идущие друг за другом, образуют бóльшие циклы, называемые *петлями, стемами и псевдоузлами*. Этим мотивам можно дать строгие определения.

**Стем** — последовательность пар оснований (i, j), (i + 1, j - 1),..., (i + k, j - k), такая, что:

1. $k \geq 1$
2. $i + k < j - k$
3. Все пары оснований (i + x, j - x), x = 0, ..., k, являются Уотсон-Криковскими или Wobble-спариваниями.

Пара (i, j) называется *внешней парой стема*. Пара (i + k, j - k) называется *внутренней парой стема*.

Для любого стема из пар (i, j), (i + 1, j - 1),..., (i + k, j - k) участок цепи РНК с позициями [i, i + k] называется *левым крылом*, участок [j - k, j] называется *правым крылом*.

Каждому стему можно поставить в соответствие отрезок всей цепочки оснований, *внутренний* по отношению к нему — это последовательность нуклеотидов с позициями [i + k, j - k].

В дальнейшем примем следующие обозначения: H — произвольный стем, (i, j) — внутренняя пара этого стема.

Будем говорить, что стем H1 *лежит внутри стема* H, если левое и правое крылья H1 являются внутренними по отношению к стему H,

Основание на позиции t *принадлежит стему* H, если оно является внутренним по отношению к H, и не существует стема H1, лежащего внутри H, т.ч. $x < t < y$, где (x, y) — внешняя пара H1.

**Петля** — совокупность всех оснований, принадлежащих некоторому стему. Существует три вида петель, различающихся их позициями по отношению к стему.

1. Петля называется **шпилькой**, если она не содержит внутренних пар, иными словами, состоит из одной цепочки неспаренных оснований.
2. Петля называется **внутренней**, если она содержит ровно одну внутреннюю пару некоторого стема, иными словами, состоит из двух цепочек неспаренных оснований.
3. Петля называется **выпячиванием**, если она является внутренней, и длина одной из составляющих её цепочек равна 0.
4. Петля называется **мультипетлёй**, если она содержит несколько внутренних пар стемов, такая петля может состоять из нескольких (не меньше трёх) цепочек неспаренных оснований.

Для определения последнего важного мотива вторичной структуры РНК, псевдоузла, следует ввести одно вспомогательное определение.

*Элементарный Закрытый Участок* (ЭЗУ) — это участок цепи [i, j], $i < j$, т.ч.:

1. Не существует пар оснований (k, l), т.ч. ($i \le k \le j$; $l > j$) или ($k < i$; $i \le l \le j$)
2. Не существует позиции l, т.ч. $i < l < j$ и участки [i, l] and [l + 1, j] удовлетворяют условию 1.
3. Пары (i, k) и (l, j) являются парами оснований, при этом, возможно, $k = j$ и $i = l$

**Псевдоузел** — это ЭЗУ, который содержит конфликтующие пары оснований. Иными словами, псевдоузел — это участок РНК, на котором пересекаются как минимум два стема.

Нагляднее всего представлять псевдоузлы при помощи диаграммы дуг (см. Рис.3). На данной диаграмме проводятся рёбра между теми участками РНК, которые являются частями одного стема.

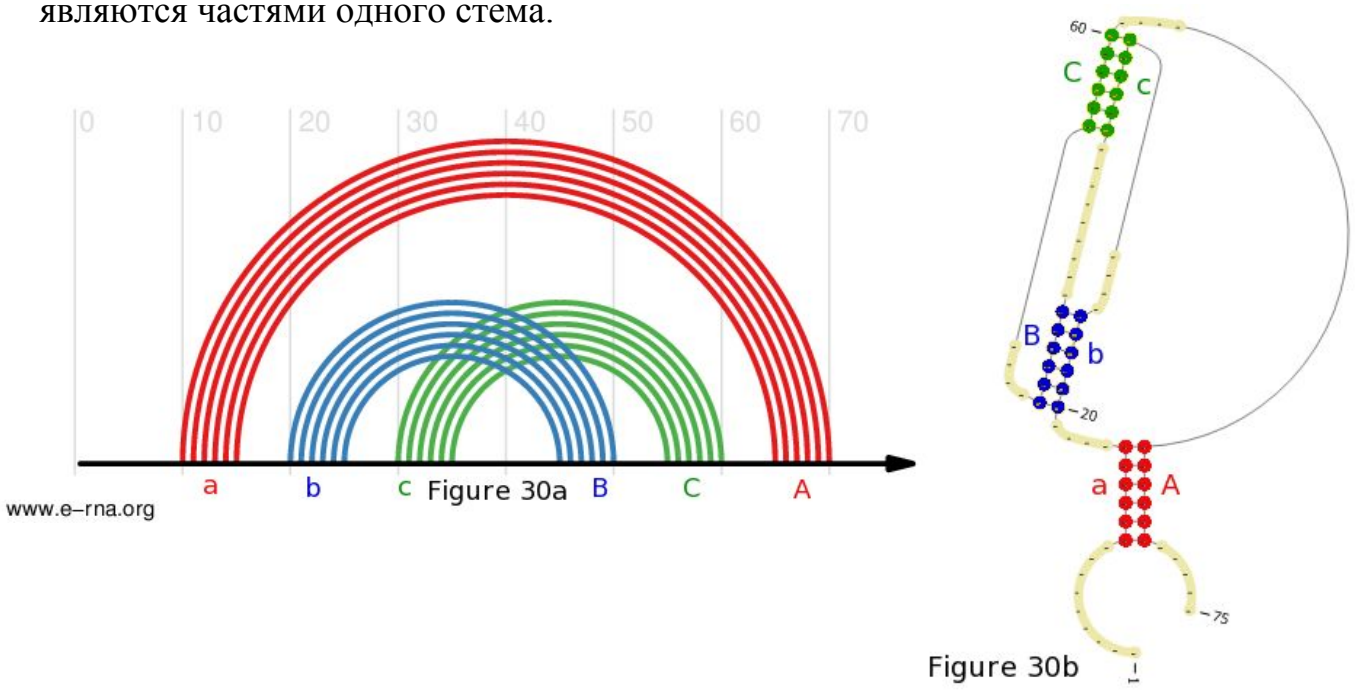


Рис. 3. Дуговая диаграмма и вторичная структура псевдоузла.

Основным элементом третичной структуры является триплекс — кластер из трёх азотистых оснований, рёбра которых участвуют в водородных связях. При этом, не все три нуклеотида должны быть связаны друг с другом, на практике чаще всего встречаются тройки с двумя спариваниями между основаниями. Другим способом определить понятие триплекса можно через плоскости: тройка нуклеотидов может считаться триплексом, если все они лежат в одной плоскости, и находятся на небольшом расстоянии друг от друга (не более 2-4 ангстрем).

Первое определение использовалось в работе [7], в которой вводится классификация триплексов на основании спариваний нуклеотидов, входящих в триплекс. Данная классификация проявила себя очень удобной для описания, хранения и анализа третичных мотивов РНК.

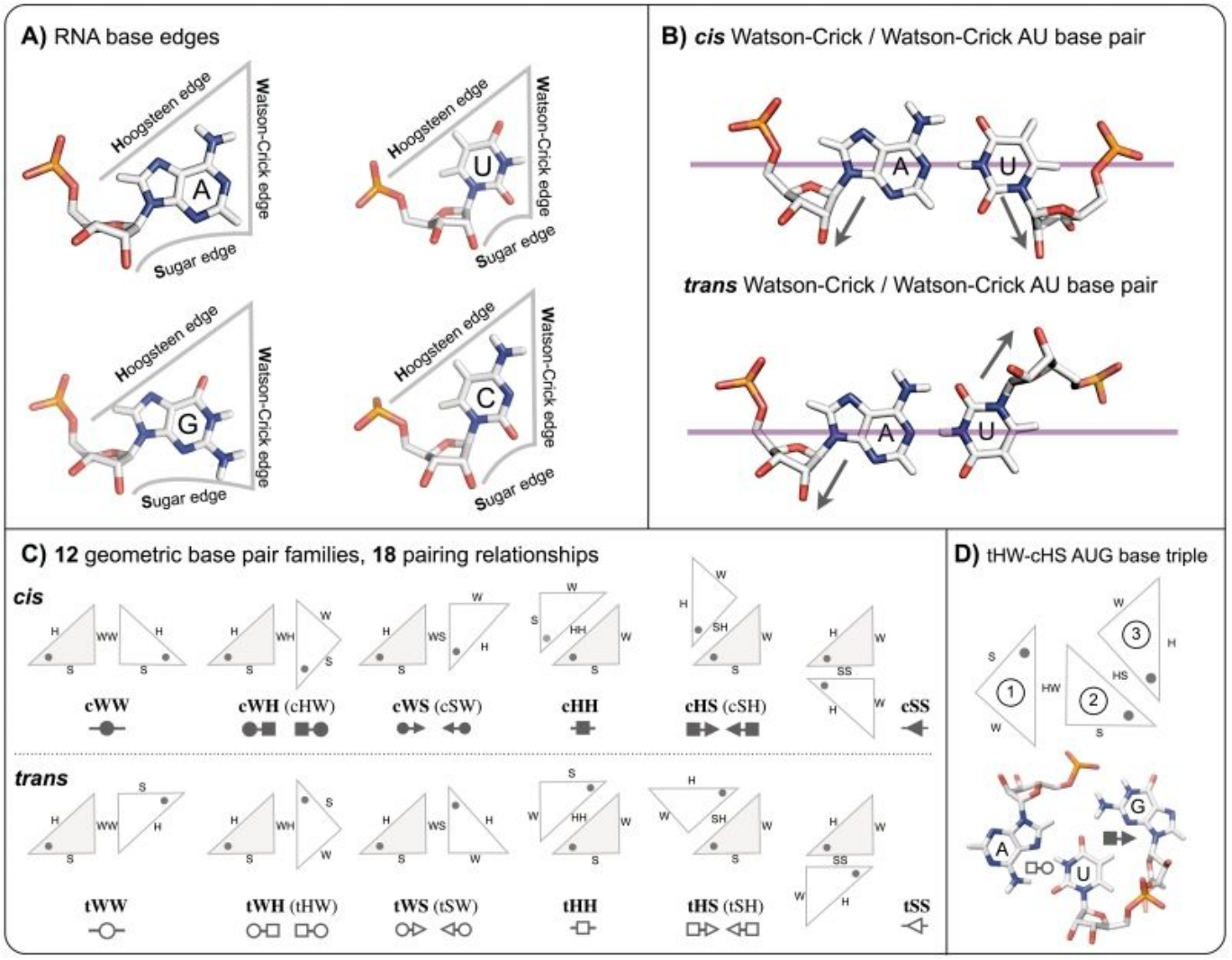


Рис. 4. Визуальное отображение классификации Леонтиса-Вестхофа.

На Рис. 4A определены рёбра нуклеотидов: Уотсон-Криковское (W) ребро, Хугстиновское (H) ребро и сахарное (S) ребро. Таким образом, каждую пару взаимодействующих оснований в триплексе, принимая во внимание их взаимную ориентацию относительно линии, проходящей через центр оснований (цис- или транс-), и тип участвующего в спаривании ребра, можно отнести к одному из восемнадцати семейств. Следует отметить, что некоторые пары оснований переходят друг в друга при зеркальном отражении, поэтому уникальных семейств (без учёта типов оснований) остаётся 12, а не 18.

Два типа ориентации представлены на Рис. 4B. Для краткости названия классов сокращаются до трёх букв: первая указывает на взаимную ориентацию нуклеотидов, t или c, вторая и третья — на рёбра нуклеотидов, участвующие в спаривании.

Для примера можно рассмотреть, как такая классификация накладывается на самые частые виды спариваний. Пара комплементарных оснований, каноническая Уотсон-Криковская пара, будет классифицирована как ‘cWW’, cis Watson-Crick/Watson-Crick, хотя тип cWW не ограничен комплементарными спариваниями

Авторы статьи [7] использовали пиктограммы для более удобного отображения этих классов: круг соответствует W-ребру, квадрат — H-ребру, треугольник — S-ребру. Заливка фигур отражает цис- или транс-ориентацию пары оснований.

Из классификации пар авторы статьи вывели классификацию триплексов. Выбирая одно основание как “центральное” можно описать его взаимодействия с двумя другими основаниями используя нотацию выше. Это позволяет классифицировать все триплексы в РНК на 108 подсемейств. Именно такое описание триплексов используется в данной работе.

### Аннотация триплексов

Одним из распространённых форматов данных о макромолекулах является формат PDB. Он включает в себя информацию о расположении атомов молекулы в кристалле. Однако, измерения производятся с погрешностью, и сами атомы могут быть подвержены небольшим флуктуациям, поэтому поиск и анализ структур всегда связаны с определёнными трудностями: для этих операций необходимо разработать оптимальные алгоритмы поиска в трёхмерном пространстве и определить пороги срабатывания этих алгоритмов.

В статье [7] для поиска триплексов, принадлежащих различным классам, использовалась утилита FR3D (Find RNA 3D) [17]. Данная программа предназначена для поиска любых структур по шаблону, в том числе триплексов.. В основе её работы лежит следующий алгоритм:

1. Утилита предобрабатывает файлы базы данных, собирая из них информацию об основаниях. Для каждого основания высчитывается положение по центральному азоту и строится матрица поворота размера 3x3, описывающая ориентацию плоскости, в которой лежит основание, относительно некоторой основной плоскости.
2. Полученная информация используется для поиска пар оснований, достаточно близких, чтобы они могли вступить во взаимодействие, и их классификации по Леонтису-Вестхофу. Данные о классах сохраняются в памяти для помощи в поиске нужных мотивов.
3. Для поиска мотивов по шаблону, FR3D находит в доступных файлах позиции тяжёлых атомов в азотистом основании, по ним рассчитываются координаты геометрического центра основания и поворотная матрица, как было описано выше.
4. По геометрическим данным о структурах из файлов для поиска, оценивает и ранжирует найденные структуры.

В статье [4] поиск и аннотация триплексов производились при помощи утилиты DSSR (Dissecting the Spatial Structure of RNA) [18]. DSSR предназначена для поиска и классификации пар оснований в файлах PDB. Отметим отличия между алгоритмами поиска этой утилиты и FR3D.

1. Для структур из представленных утилите файлов DSSR считает такие параметры для описания твёрдого тела как сдвиг, сжатие, растяжение, изгиб, кручение, достаточные для определения взаимного положения двух любых оснований.

2. DSSR ищет пары и триплексы в горизонтальных плоскостях с небольшим шагом, используя данные об их взаимном расположении, полученные на шаге 1.

В данной работе строится классификация триплексов, учитывающая геометрические и пространственные признаки оснований. Для поиска, аннотации и классификации спариваний по Леонтису-Вестхофу использовалась утилита DSSR, однако, так как в ходе работы производилось сравнение классифицированных нами данных и результатов работы [7], было важно понять, чем отличались утилиты DSSR и FR3D, чтобы объяснить небольшие расхождения в результатах.

### Данные и реализация

Получение “сырых” данных производилось при помощи SQL-запросов к базе данных URSDB [4] (http://server3.lpm.org.ru/urs/). Эта база данных состоит из 51 таблицы, в которые занесена информация более чем о трёх тысячах файлов с информацией об РНК. База работает с файлами формата mmCIF, поддерживает несколько видов классификации структур, и предоставляет удобный интерфейс для их поиска и просмотра.

Дальнейшая обработка велась на языке программирования Python3 с использованием библиотек pandas, numpy, pymysql, csv. Ознакомиться с программным обеспечением можно по следующей ссылке: https://github.com/anatastasia/RNA-Thesis.

Описание каждого триплекса состоит из следующих элементов:

1. PDB ID - указатель на файл, в котором найдена структура, в PDB
2. RNA Chains - буквенное обозначение цепи РНК, в которой найдена структура
3. Triple - нуклеотидный состав триплекса

4. Nucl1, 2, 3 - порядковые номера нуклеотидов, составляющих триплекс, в цепи
5. BasePair1, 2, 3 - тип спаривания нуклеотидов по Леонтису-Вестхофу
6. SSElem1, 2, 3 - обозначение элемента вторичной структуры, которому принадлежит нуклеотид
7. Respect1-2, 2-3, 3-1 - признак для пары нуклеотидов, указывающий их относительное расположение в контексте вторичной структуры

Значения, которые могут принимать два последних признака:

1. Структурные элементы:
    a. S - стем
    b. H - петля-шпилька
    c. I - внутренняя петля
    d. B - выпячивание
    e. J - мультипетля
2. Также, каждой петле присваивался тип в зависимости от того, где она находится относительно ближайшего псевдоузла:
    a. C - классическая петля, не находящаяся рядом с псевдоузлом
    b. I - изолированная петля, т.е. находящаяся по соседству с псевдоузлом
    c. P - псевдоузловая, т.е. являющаяся частью псевдоузла
3. Пространственное расположение нуклеотидов описывается следующим способом:
    a. Same (SM) - одинаковый, оба нуклеотида принадлежат одному структурному элементу
    b. Local (LC) - локальный, т.е. нуклеотиды находятся в соседних структурных элементах
    c. Long-Range (LR) - удалённый, т.е. нуклеотиды попарно находятся в удалённых друг от друга структурных элементах

Однако, стоит упомянуть одно различие, проявившееся при подготовке данных: если в вышеупомянутой статье [7] все структуры были аннотированы утилитой FR3D, то для данной работы разметка проводилась утилитой DSSR. Различия в подходах к поиску триплексов этих утилит было описано ранее.

Из базы URSDB было выделено 198055 аннотированных триплексов в формате, описанном выше.

Далее были отброшены все триплексы, аннотированные с помощью DSSR, но содержавшие только одно спаривание. Это возможно из-за того, что утилита ищет триплексы по пространственному расположению нуклеотидов, поэтому те структуры, что не могут являться триплексом по такому определению, могли быть извлечены из базы.

Следующим шагом подготовки данных была фильтрация триплексов по списку представителей структур РНК. Неизбыточное подмножество структур (Representative Set) — это список файлов из PDB, где каждому виду РНК каждого организма, называемому классом, сопоставлена одна структура-представитель этой молекулы. Так как данных по одному классу может быть много, фильтрация по этому списку необходима, чтобы избежать дублирования данных.

Всего в неизбыточном подмножестве структур версии 3.11 (по состоянию на март 2018) было выявлено 8240 триплексов.

Каждый триплекс в трёхмерном пространстве может быть описан несколькими способами, так как база данных хранит информацию только об абсолютном положении оснований. Поэтому все строки, полученные после фильтрации, были пропущены через скрипт, генерирующий все возможные перестановки нуклеотидов для заданного триплекса. Триплексы с двумя спариваниями были зеркально отражены (см. Таблицу 3).

Таблица 3. Пример работы генератора перестановок на триплексе с двумя спариваниями.

| | |
|---|---|
| Исходная строка датасета | 1h3e.cif1;B;UGA;B.U.9.;B.G.13.;B.A.22.;tHW;tSH;;JC;HC;HC;LR;SM;LR |
| Полученные строки датасета | 1h3e.cif1;B;UGA;B.U.9.;B.G.13.;B.A.22.;tHW;tSH;;JC;HC;HC;LR;SM;LR<br>1h3e.cif1;B;AGU;B.A.22.;B.G.13.;B.U.9.;tHS;tWH;;HC;HC;JC;LR;SM;LR |

Для триплексов с тремя спариваниями были сгенерированы все перестановки нуклеотидов (см. Таблицу 4).

Таблица 4. Пример работы генератора перестановок на триплексе с тремя спариваниями.

| | |
|---|---|
| Исходная строка датасета | 1exy.cif1;A;GUC;A.G.5.;A.U.26.;A.C.28.;tHW;tSW;cWW;S;BC;S;LC;LC;SM |
| Полученные строки датасета | 1exy.cif1;A;GUC;A.G.5.;A.U.26.;A.C.28.;tHW;tSW;;S;BC;S;LC;LC;SM<br>1exy.cif1;A;CUG;A.C.28.;A.U.26.;A.G.5.;tWS;tWH;;S;BC;S;LC;LC;SM<br>1exy.cif1;A;UCG;A.U.26.;A.C.28.;A.G.5.;tSW;cWW;;BC;S;S;LC;SM;LC<br>1exy.cif1;A;GCU;A.G.5.;A.C.28.;A.U.26.;cWW;tWS;;S;S;BC;SM;LC;LC<br>1exy.cif1;A;CGU;A.C.28.;A.G.5.;A.U.26.;cWW;tHW;;S;S;BC;SM;LC;LC<br>1exy.cif1;A;UGC;A.U.26.;A.G.5.;A.C.28.;tWH;cWW;;BC;S;S;LC;SM;LC |

После генерации получилось 17817 триплексов, которые в дальнейшем были подвергнуты тщательному анализу.

## Анализ триплексов в экспериментально разрешённых структурах

Приведём статистику по семействам пар оснований, самым частым геометрическим и пространственным конфигурациям, полученных после применения нашей классификации к описанному выше датасету.

Для построения таблицы 5 были отфильтрованы все возможные комбинации пары оснований, образующих триплекс. По вертикали указан тип спаривания по Леонтису-Вестхофу первого основания со вторым, по горизонтали — тип спаривания второго основания с третьим. Таблица не симметрична относительно главной оси, так как не была произведена группировка спариваний, переходящих друг в друга при их зеркальном отражении.

Таблица 5. Частоты семейств триплексов в датасете.

| | **cWW** | **tWW** | **cHW** | **tHW** | **cSW** | **tSW** | **cWH** | **tWH** | **cHH** | **tHH** | **cSH** | **tSH** | **cWS** | **tWS** | **cHS** | **tHS** | **cSS** | **tSS** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **cWW** | 249 | 77 | 350 | 343 | 562 | 420 | 40 | 36 | 58 | 94 | 220 | 123 | 31 | 30 | 216 | 70 | 336 | 268 |
| **tWW** | 77 | 7 | 7 | 51 | 12 | 6 | 2 | 33 | 7 | 30 | 4 | 5 | 27 | 3 | 9 | 26 | 16 | 2 |
| **cHW** | 40 | 2 | 3 | 29 | 4 | 3 | 3 | 16 | 2 | 2 | 1 | 4 | 3 | 0 | 4 | 4 | 0 | 1 |
| **tHW** | 36 | 33 | 3 | 17 | 25 | 48 | 16 | 42 | 1 | 2 | 14 | 6 | 7 | 0 | 95 | 11 | 5 | 1 |
| **cSW** | 31 | 27 | 32 | 46 | 1 | 0 | 3 | 7 | 1 | 4 | 0 | 17 | 0 | 7 | 21 | 10 | 0 | 0 |
| **tSW** | 30 | 3 | 3 | 40 | 31 | 2 | 0 | 0 | 1 | 2 | 0 | 3 | 7 | 0 | 3 | 21 | 37 | 0 |
| **cWH** | 350 | 7 | 6 | 15 | 1 | 3 | 3 | 3 | 1 | 1 | 0 | 7 | 32 | 3 | 2 | 2 | 1 | 1 |
| **tWH** | 343 | 51 | 15 | 5 | 7 | 2 | 29 | 17 | 1 | 3 | 0 | 3 | 46 | 40 | 3 | 0 | 3 | 5 |
| **cHH** | 58 | 7 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 5 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| **tHH** | 94 | 30 | 1 | 3 | 12 | 2 | 2 | 2 | 5 | 2 | 1 | 0 | 4 | 2 | 0 | 0 | 1 | 14 |
| **cSH** | 216 | 9 | 2 | 3 | 1 | 0 | 4 | 95 | 0 | 0 | 0 | 1 | 21 | 3 | 1 | 5 | 0 | 1 |
| **tSH** | 70 | 26 | 2 | 0 | 7 | 4 | 4 | 11 | 0 | 0 | 1 | 2 | 10 | 21 | 5 | 0 | 5 | 4 |
| **cWS** | 562 | 12 | 1 | 7 | 0 | 0 | 4 | 25 | 0 | 12 | 5 | 2 | 1 | 31 | 1 | 7 | 0 | 0 |
| **tWS** | 420 | 6 | 3 | 2 | 0 | 0 | 3 | 48 | 0 | 2 | 2 | 1 | 0 | 2 | 0 | 4 | 8 | 0 |
| **cHS** | 220 | 4 | 0 | 0 | 5 | 2 | 1 | 14 | 0 | 1 | 0 | 7 | 0 | 0 | 0 | 1 | 0 | 0 |
| **tHS** | 123 | 5 | 7 | 3 | 2 | 1 | 4 | 6 | 1 | 0 | 7 | 4 | 17 | 3 | 1 | 2 | 0 | 0 |
| **cSS** | 336 | 16 | 1 | 3 | 0 | 8 | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 37 | 0 | 5 | 0 | 2 |
| **tSS** | 268 | 2 | 1 | 5 | 0 | 0 | 1 | 1 | 3 | 14 | 0 | 0 | 0 | 0 | 1 | 4 | 2 | 0 |

Следующая таблица была получена после выделения всех типов пар оснований, представленных в  триплексах. Можно заметить, что самым населённым суперсемейством является cWW — каноническое Уотсон-Криковское спаривание.

Таблица 6. Частоты суперсемейств триплексов в датасете.

| **cHH** | **cHS** | **cHW** | **cSH** | **cSS** | **cSW** |
|---|---|---|---|---|---|
| 382 | 998 | 1045 | 1802 | 1890 | 1641 |
| **cWH** | **cWS** | **cWW** | **tHH** | **tHS** | **tHW** |
| 1459 | 2211 | 14428 | 782 | 838 | 2201 |
| **tSH** | **tSS** | **tSW** | **tWH** | **tWS** | **tWW** |
| 706 | 1228 | 1429 | 2081 | 1705 | 1534 |

Таблица 7 была получена подсчётом количества строк, содержащих триплексы с заданной пространственной конфигурацией. Можно отметить, что чаще всего в датасете встречались триплексы, все три нуклеотида которых принадлежали одному и тому же структурному элементу.

Таблица 7. Частоты различных пространственных конфигураций триплексов.

| **LCLCLC** | **LRLCLC** | **LRLRLC** | **LRLRLR** | **LRLRSM** | **LRSMLC** | **LRSMSM** | **SMLCLC** | **SMSMLC** | **SMSMSM** |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 4 | 144 | 1236 | 16592 | 0 | 0 | 10776 | 0 | 22740 |

В таблице 8 представлены частоты различных структурных конфигураций триплексов. В совокупности с предыдущей таблицей можно отметить, что в нашем датасете одними из самых частотных конфигураций являются те, что уже были описаны в литературе:

1. Триплексы, состоящие из двух нуклеотидов, принадлежащих одному стему, и нуклеотида из прилежащей внутренней петли (IC,S,S,LC,SM,LC). Такая конфигурация с аденином из внутренней петли известна под названием A-minor groove.

2. Триплексы, состоящие из двух нуклеотидов из одного стема и одного нуклеотида из удалённой шпильки (HC,S,S,LR,SM,LR). Такая конфигурация известна как Tetraloop receptor.

Таблица 8 Частоты различных структурных конфигураций триплексов.

| **SSS** | **SSHC** | **SSIC** | **SSBC** | **SSJC** | **SHCHC** |
|---|---|---|---|---|---|
| 1824 | 4628 | 4704 | 2100 | 4704 | 88 |
| **SHCIC** | **SHCBC** | **JCJCJC** | **SICIC** | **SICBC** | **SICJC** |
| 18 | 6 | 8400 | 200 | 4 | 2 |
| **SBCBC** | **SBCJC** | **SJCJC** | **HCHCHC** | **HCHCIC** | **HCHCBC** |
| 44 | 6 | 160 | 4716 | 92 | 28 |
| **HCHCJC** | **HCICIC** | **HCICBC** | **HCBCBC** | **HCJCJC** | **ICICIC** |
| 916 | 472 | 2 | 24 | 548 | 9780 |
| **ICICBC** | **ICICJC** | **ICBCBC** | **ICJCJC** | **BCBCBC** | **BCJCJC** |
| 120 | 392 | 4 | 200 | 24 | 36 |

Преимущества классификации

Основным результатом данной работы является выделение особого класса триплексов, представляющего немалый интерес для дальнейших исследований. Участки РНК, содержащие структуры этого класса, встречаются в рибосомальных РНК десятков различных организмов: Homo Sapiens, E. Coli, Thermus Thermophilus и других. Триплексы, принадлежащие данному классу, характеризуются тем, что все нуклеотиды, входящие в их состав, находятся в разных, удалённых друг от друга, структурных элементах.

Прежде чем описать свойства класса, хотелось бы объяснить, почему его появлению было уделено особое внимание.

В образовании той или иной третичной структуры РНК играет роль то, насколько энергетически выгодно это состояние молекулы. Поэтому гораздо чаще среди всех триплексов встречаются те, что образованы двумя нуклеотидами из одного элемента (это отмечено значением SM ячейки SSElem в нашем датасете), и третьим — из соседнего (значение LC ячейки SSElem). Такое состояние для молекулы является более энергетически выгодным, чем состояние с триплексом, в котором соединённые водородными связями и находящиеся в одной плоскости нуклеотиды находятся в разных элементах.

В датасете из 17817 строк, полученных после фильтрации всех строк по списку представителей, было обнаружено 1236 экземпляров класса LRLRLR.

Среди них 492 экземпляра имели одинаковый нуклеотидный состав — GAA.

Для дальнейшего анализа из списка неизбыточного подмножества структур были отобраны классы структур из малых и больших рибосомальными субъединиц всех организмов. При этом те классы, в которых было недостаточно информации — например, в файле отсутствовали данные по некоторым цепочкам — не были учтены при анализе. Изучение триплексов в этих классах показало, что LRLRLR GAA триплексы встречаются в малых рибосомальных единицах всех организмов. На Рис. 5 показаны 15 представителей данного триплекса, с указанием названия организма, в котором структуры были найдены.

| PDB ID | Chain | Organism | Molecule | Nucleotides | | |
|---|---|---|---|---|---|---|
| 4v4n.cif1 | B2 | Methanocaldococcus jannaschii | 16S ribosomal RNA | B2.G.1276. | B2.A.937. | B2.A.1320. |
| 4v6u.cif1 | A2 | Pyrococcus furiosus | 16S rRNA | A2.G.1276. | A2.A.937. | A2.A.1320. |
| 5jb3.cif1 | 2 | Pyrococcus abyssi | 16S ribosomal RNA | 2.G.1276. | 2.A.937. | 2.A.1320. |
| 1fjg.cif1 | A | Thermus thermophilus | 16S RIBOSOMAL RNA | A.G.1316. | A.A.978. | A.A.1360. |
| 3j2c.cif1 | M | Escherichia coli | 16S rRNA head domain | M.G.1316. | M.A.978. | M.A.1360. |
| 4v4q.cif1 | CA | Escherichia coli | 16S ribosomal RNA | CA.G.1316. | CA.A.978. | CA.A.1360. |
| 4v4b.cif1 | AA | Saccharomyces cerevisiae | 18S ribosomal RNA | AA.G.1316. | AA.A.978. | AA.A.1360. |
| 3j9w.cif1 | AA | Bacillus subtilis | 16S ribosomal RNA | AA.G.1325. | AA.A.988. | AA.A.1369. |
| 5li0.cif1 | a | Staphylococcus aureus | 16S ribosomal RNA | a.G.1327. | a.A.988. | a.A.1371. |
| 5mrc.cif1 | aa | Saccharomyces cerevisiae | 15S RNA | aa.G.1384. | aa.A.1043. | aa.A.1429. |
| 3j9m.cif1 | AA | Homo sapiens | 12S rRNA | AA.G.1401. | AA.A.1232. | AA.A.1444. |
| 3jam.cif1 | 2 | Kluyveromyces lactis | 18S rRNA | 2.G.1551. | 2.A.1202. | 2.A.1595. |
| 4uer.cif1 | A | Lachancea kluyveri | 18S RRNA | A.G.1553. | A.A.1203. | A.A.1597. |
| 3j7p.cif1 | S2 | Sus scrofa | 18S ribosomal RNA | S2.G.1617. | S2.A.1260. | S2.A.1661. |
| 3j7a.cif1 | A | Plasmodium falciparum | 18S ribosomal RNA | A.G.1850. | A.A.1304. | A.A.1894. |

Рис. 5. Примеры консервативного триплекса LRLRLR GAA

Как было сказано выше, к классификации триплексов в соответствии со статьёй Almacarem et al. [7] было решено добавить информацию о структурных и пространственных признаках. Дальнейшим шагом была валидация, то есть проверка соответствия полученной классификации предыдущим наработкам в этой области.

Валидацию было решено проводить через сравнение количества триплексов, принадлежащих различным семействам триплексов согласно классификации Леонтиса. Следует отметить, что авторы оперировали меньшей выборкой структур из PDB, около 30000 триплексов, аннотированных утилитой FR3D. В оригинальной статье частоты всех классов были подсчитаны, результаты представлены таблицей (см. Рис.6).

108 Triple Families

| | cHW | tHW | cHH | tHH | cHS | tHS | cSW | tSW | cSH | tSH | cSS | tSS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **cWW** | 9/36 | 13/36 | 2/17 | 6/45 | 13/54 | 8/39 | 15/60 | 10/53 | 11/52 | 9/37 | 24/60 | 13/30 |
| **tWW** | 0/34 | 3/33 | 1/16 | 2/41 | 1/50 | 6/35 | 2/56 | 1/50 | 0/48 | 1/34 | 9/56 | 1/28 |
| **cHW** | 2/24 | 1/23 | 0/11 | 0/29 | 2/36 | 1/25 | 1/40 | 0/36 | 0/34 | 0/24 | 0/40 | 0/20 |
| **tHW** | 1/24 | 2/23 | 0/11 | 0/29 | 1/36 | 1/25 | 2/40 | 3/36 | 3/34 | 2/24 | 9/40 | 1/20 |
| **cSW** | 2/40 | 4/40 | 0/20 | 0/48 | 4/56 | 0/40 | 0/64 | 0/56 | 0/56 | 6/40 | 3/64 | 1/32 |
| **tSW** | 0/32 | 3/32 | 0/14 | 0/42 | 0/52 | 2/38 | 2/56 | 1/50 | 1/48 | 0/34 | 2/56 | 1/28 |
| **cWH** | | | | | | | 0/40 | 1/34 | 0/36 | 0/26 | 3/40 | 6/24 |
| **tWH** | | | | | | | 1/40 | 2/33 | 2/37 | 0/27 | 9/40 | 9/28 |
| **cHH** | | | | | | | 0/20 | 0/16 | 0/19 | 0/14 | 0/20 | 0/16 |
| **tHH** | | | | | | | 2/48 | 0/41 | 0/43 | 1/31 | 5/48 | 5/28 |
| **cSH** | | | | | | | 0/56 | 0/50 | 0/48 | 1/34 | 4/56 | 1/24 |
| **tSH** | | | | | | | 2/40 | 0/35 | 0/35 | 0/25 | 9/40 | 9/20 |

Рис. 6. Количество триплексов из различных классов, найденных авторами статьи [7]

Цвета в таблице обозначают следующее:

1. Зелёный цвет: в базе данных нашёлся хотя бы один триплекс из соответствующего семейства. Первое число в ячейке — количество различных обнаруженных троек нуклеотидов, второе — максимально возможное количество различных троек.
2. Жёлтый цвет: в базе данных не было обнаружено ни одной тройки нуклеотидов, но подобный триплекс разрешён в их классификации
3. Серый цвет: семейство, которому соответствует ячейка, не может образоваться в их классификации

Проведя аналогичные расчёты мы получили следующие результаты:

Таблица 9. Количество триплексов из различных классов, найденных в данной работе

| | cWW | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cWW | 249 | tWW | | | | | | | | | | | | | | | | |
| tWW | 77 | 7 | cHW | | | | | | | | | | | | | | | |
| cHW | 40 | 2 | 3 | tHW | | | | | | | | | | | | | | |
| tHW | 36 | 33 | 3 | 17 | cSW | | | | | | | | | | | | | |
| cSW | 31 | 27 | 32 | 46 | 1 | tSW | | | | | | | | | | | | |
| tSW | 30 | 3 | 3 | 40 | 31 | 2 | cWH | | | | | | | | | | | |
| cWH | 350 | 7 | 6 | 15 | 1 | 3 | 3 | tWH | | | | | | | | | | |
| tWH | 343 | 51 | 15 | 5 | 7 | 2 | 29 | 17 | | | | | | | | | | |
| cHH | 58 | 7 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | tHH | | | | | | | | |
| tHH | 94 | 30 | 1 | 3 | 12 | 2 | 2 | 2 | 5 | 2 | cSH | | | | | | | |
| cSH | 216 | 9 | 2 | 3 | 1 | 0 | 4 | 95 | 0 | 0 | 0 | tSH | | | | | | |
| tSH | 70 | 26 | 2 | 0 | 7 | 4 | 4 | 11 | 0 | 0 | 1 | 2 | cWS | | | | | |
| cWS | 562 | 12 | 1 | 7 | 0 | 0 | 4 | 25 | 0 | 12 | 5 | 2 | 1 | tWS | | | | |
| tWS | 420 | 6 | 3 | 2 | 0 | 0 | 3 | 48 | 0 | 2 | 2 | 1 | 0 | 2 | cHS | | | |
| cHS | 220 | 4 | 0 | 0 | 5 | 2 | 1 | 14 | 0 | 1 | 0 | 7 | 0 | 0 | 0 | tHS | | |
| tHS | 123 | 5 | 7 | 3 | 2 | 1 | 4 | 6 | 1 | 0 | 7 | 4 | 17 | 3 | 1 | 2 | cSS | |
| cSS | 336 | 16 | 1 | 3 | 0 | 8 | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 37 | 0 | 5 | 0 | tSS |
| tSS | 268 | 2 | 1 | 5 | 0 | 0 | 1 | 1 | 3 | 14 | 0 | 0 | 0 | 0 | 1 | 4 | 2 | 0 |

Здесь используется та же цветовая схема, что на Рис. 6.

В первую очередь, стоит отметить, что большинство обнаруженных авторами семейств было найдено и в нашей классификации. Исключениями являются семейства tSHtHH, cSScWS, tSScWS и tSStWS. Про возможную причину этого (и нескольких других) расхождений будет рассказано ниже.

Во-вторых, ненулевые значения в жёлтых ячейках указывают, что в нашей классификации были обнаружены семейства, ранее не зафиксированные в базе данных. Можно выделить две причины этого:

1. В 2011 года объём данных в PDB существенно увеличился, и тройки нуклеотидов, ранее не присутствовавших в базе, были добавлены позднее
2. В нашей работе аннотация производилась утилитой DSSR, имеющей иной алгоритм работы, нежели утилита FR3D

В третьих, ненулевые значения в серых “невозможных” ячейках, самый заметный пример — cWWcWW, могут также появиться из-за использования иной утилиты. FR3D и DSSR используют различные подходы к выделению триплексов из файлов PDB, подробнее про это было рассказано выше.

Также следует отметить, что пороги точности у утилит различны, и небольшое отклонение позиций атомов в пространстве, вызванное, например, их флуктуациями, могло повлиять на конечный результат.

# ЗАКЛЮЧЕНИЕ

Подводя итоги исследования, можно выделить основные преимущества построенной классификации, и обозначить дальнейшие шаги исследований в этой области.

Во-первых, в ходе работы удалось с высокой точностью валидировать построения, подсчитав частоты различных семейств триплексов и сравнив с результатами статьи [7]. Это сравнение также помогло обнаружить ранее отсутствующие семейства в базе данных PDB. Расхождения, выражающиеся в том, что не все ранее найденные семейства были зафиксированы, могут быть объяснены использованием разных утилит для разметки данных и различными их порогами определения триплексов.

Семейства, недопустимые в классификации Леонтиса-Вестхофа, то есть те, в формировании которых участвуют одинаковые рёбра всех нуклеотидов в триплексе, нуждаются в дальнейшем изучении: следует понять, как они формируются, какую роль в третичной структуре играют, как часто встречаются в структурах РНК.

Во-вторых, работа позволила обнаружить новый класс триплексов, обладающий интересными для исследования свойствами: экземпляры этого класса консервативны и встречаются только в РНК малых субъединиц рибосом. Дальнейшая работа может быть направлена на выяснение биологической роли данной структуры, и на поиск других классов с похожими свойствами.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Montange RK, Batey RT. Riboswitches: emerging themes in RNA structure and function. Annu. Rev. Biophys.. 2008 Jun 9;37:117-33.
2. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nature Reviews Genetics. 2009 Mar;10(3):155.
3. Iwasaki YW, Siomi MC, Siomi H. PIWI-interacting RNA: its biogenesis and functions. Annual review of biochemistry. 2015 Jun 2;84:405-33.
4. Baulin, E., Yacovlev, V., Khachko, D., Spirin, S., & Roytberg, M. (2016). URS DataBase: universe of RNA structures and their motifs. Database: The Journal of Biological Databases and Curation, 2016
5. Thapar, R., Denmon, A. P., & Nikonowicz, E. P. (2014). Recognition Modes of RNA Tetraloops And Tetraloop-Like Motifs By RNA Binding Proteins. Wiley Interdisciplinary Reviews. RNA, 5(1).
6. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B., & Steitz, T. A. (2001). RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. Proceedings of the National Academy of Sciences of the United States of America, 98(9), 4899–4903.
7. Abu Almakarem, A. S., Petrov, A. I., Stombaugh, J., Zirbel, C. L., & Leontis, N. B. (2012). Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic Acids Research, 40(4), 1407–1423.
8. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): the single global macromolecular structure archive. Protein Crystallography: Methods and Protocols. 2017:627-41.
9. Klosterman, P. S., Hendrix, D. K., Tamura, M., Holbrook, S. R., & Brenner, S. E. (2004). Three-dimensional motifs from the SCOR, structural classification of RNA

database: extruded strands, base triples, tetraloops and U-turns. Nucleic Acids Research, 32(8), 2342–2352.

10. Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA. Splicing of messenger RNA precursors. Annual review of biochemistry. 1986 Jul;55(1):1119-50.
11. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research. 1997 Mar 1;25(5):955.
12. Noller HF. Structure of ribosomal RNA. Annual review of biochemistry. 1984 Jul;53(1):119-62.
13. Pratt AJ, MacRae IJ. The RNA-induced silencing complex: a versatile gene-silencing machine. Journal of Biological Chemistry. 2009 Jul 3;284(27):17897-901.
14. Sun JS, Manley JL. A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. Genes & development. 1995 Apr 1;9(7):843-54.
15. https://en.wikipedia.org/wiki/Svedberg
16. Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., & Gralla, J. (1973). Improved estimation of secondary structure in ribonucleic acids. Nature New Biology, 246(150), 40-41.
17. Sarver, M., Zirbel, C. L., Stombaugh, J., Mokdad, A., & Leontis, N. B. (2008). FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. Journal of Mathematical Biology, 56(1-2), 215–252.
18. Lu, Xiang-Jun, Harmen J. Bussemaker, and Wilma K. Olson. “DSSR: An Integrated Software Tool for Dissecting the Spatial Structure of RNA.” Nucleic Acids Research 43.21 (2015): e142. PMC. Web. 7 June 2018. Nucleic Acids Research, Volume 43, Issue 21, 2 December 2015, Pages e142.