Федеральное государственное учреждение
«Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»

Направление подготовки: 09.06.01 – Информатика и Вычислительная техника
Направленность (профиль) подготовки: 05.13.18 – Математическое моделирование, численные методы и комплексы программ

Допущен к ГИА
Заведующий аспирантурой

_____/_____
(подпись) расшифровка подписи

«___» _____2019 г.

Выпускная квалификационная работа

«Структурные мотивы РНК: классификация, поиск, база данных»

Аспирант: Баулин Евгений Федорович

_____
(подпись аспиранта)

Научный руководитель: Яковлев В.В.,
к.ф.-м.н., и.о. зав. ЛПМ ИМПБ РАН –
филиала ИПМ им. М.В. Келдыша РАН

_____
(подпись научного руководителя)

Москва 2019

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Исследование пространственной структуры рибонуклеиновых кислот (РНК) - одно из важнейших направлений современной биоинформатики и молекулярной биологии [1]. За последние 25-30 лет было открыто множество типов молекул некодирующих РНК, выполняющих ключевые функции в клетках живых организмов [2-4]. Помимо матричных, транспортных и рибосомальных РНК, задействованных в процессе синтеза белков, были выявлены классы РНК, участвующие в процессах репликации ДНК, регуляции экспрессии генов, сплайсинга генов и других [5-7]. Функция, выполняемая РНК, напрямую зависит от пространственной структуры молекулы, что определяет актуальность изучения принципов ее структурной организации. Исследование структуры РНК представляется важным и с точки зрения теории (теория эволюции, понимание внутриклеточных процессов) и с точки зрения практики (создание лекарств) [8-10].

В структуре РНК выделяют, как правило, четыре уровня организации [11]: первичная структура (последовательность нуклеотидов), вторичная структура (множество канонических спариваний оснований), третичная структура (расположение атомов молекулы в пространстве), четвертичная структура (комплекс, образованный двумя и более молекулами).

На данный момент для описания вторичной структуры РНК используется общепринятая модель Цукера-Мэтьюза-Тернера (Nearest Neighbor Model, NNM, [12, 13]), в рамках которой вторичная структура РНК разбивается на петли (loops) и стемы (stems). Данная модель используется как для описания вторичной структуры РНК, так и для решения задачи предсказания вторичной структуры путем минимизации свободной энергии структуры. Модель NNM неприменима к структурам РНК, содержащим псевдоузлы, т.н. псевдоузловым структурам. Задача предсказания псевдоузловых структур РНК изучена в меньшей степени, чем задача

предсказания «классических» структур РНК, которые не содержат псевдоузлов [14]. При этом в настоящий момент не существует общепринятой классификации псевдоузлов. Описаны простейшие типы псевдоузлов (kissing hairpins, H-structures и др., см. [15]), псевдоузлы, играющие важную роль в клеточных процессах, см. например, [16]; в [17] проведен анализ типов псевдоузлов, введенных в различных работах по предсказанию вторичной структуры РНК по ее последовательности. В базе данных [18, 19] собраны примеры псевдоузлов в экспериментально определенных структурах РНК. Однако единого метода описания вторичной структуры при наличии псевдоузлов, подобного тому, который был предложен в [12, 13] для описания вторичных структур без псевдоузлов, на данный момент нет.

Третичная структура РНК обусловлена относительно слабыми взаимодействиями, образованными «поверх» вторичной структуры. Данные взаимодействия образуют рекуррентные третичные мотивы. Такие мотивы в РНК часто выступают в роли функциональных единиц, участвуя либо в стабилизации пространственной структуры [20], либо в узнавании других молекул [21]. Выделяют мотивы, содержащие неканонические спаривания оснований, изгибы сахаро-фосфатного остова, стекинг участков двойной спирали и др. [22-24]. В настоящее время отсутствует единый подход к описанию третичных мотивов РНК. Взаимосвязь третичных мотивов и элементов вторичной структуры остается слабо изученной.

Степень научной разработанности темы. Основными алгоритмическими задачами современной биоинформатики РНК являются задачи предсказания вторичной структуры РНК, поиска РНК-генов и предсказания пространственной структуры РНК.

Задача предсказания вторичной структуры РНК приобрела популярность во второй половине XX века, благодаря работам Р. Нуссинов и Д. Санкоффа. Существенное влияние на решение задачи предсказания вторичных структур, не содержащих псевдоузлы, оказали работы Д.Х.

Мэтьюза, М. Цукера и Д.Х. Тернера. В последние годы данной задаче уделяли внимание П.Ф. Стадлер и И. Хофакер. Изучению псевдоузловых структур в значительной мере способствовали работы А. Кондон, А.П. Гультяева и С.М. Рейдиса.

В работах А.А. Миронова, Р. Бекофена и Ш.Р. Эдди алгоритмы предсказания вторичной структуры РНК применялись к задаче поиска РНК-генов.

Существенный вклад в развитие многих алгоритмов, применяемых в биоинформатике РНК, внесли работы Е. Ривас и Ш.Р. Эдди.

Привлечению внимания к изучению пространственной структуры РНК способствовали базы данных экспериментально определенных структур, отраженные в работах Х.М. Бермана и Е. Ривас. Значительную роль в сборе данных и систематизации третичных мотивов сыграли работы С.Е. Бреннера, С.Р. Холбрука и К.Д. Лю. Анализу пространственной структуры РНК методами молекулярной динамики уделял внимание Я. Шпунер. Значительную роль в решении задачи предсказания вторичной структуры РНК сыграла экспериментальная методика SHAPE, изложенная в работах К. Викса. Методика SHAPE была использована для предсказания пространственной структуры РНК в работах Н.В. Дохоляна. В последнее время анализу третичных мотивов и предсказанию пространственной структуры РНК уделяли внимание К.Л. Зирбель, Н. Леонтис, Я. Буйницкий, А.А. Богданов, М. Попенда, Т. Шлик и Х.Я. Вольфсон.

Исследования вторичных структур РНК в значительной части охватывают только классические структуры, не содержащие псевдоузлов, т.к. задача предсказания вторичной структуры РНК при наличии псевдоузлов является NP-трудной. Как следствие, отсутствует единый язык описания произвольных вторичных структур РНК, в том числе содержащих псевдоузлы, что затрудняет изучение взаимосвязи между вторичной структурой и третичными мотивами. В работах по изучению третичных мотивов либо не рассматривается контекст вторичной структуры, либо

рассматриваются только локальные (внутри одной петли) третичные мотивы РНК. Данная работа направлена на восполнение указанных пробелов.

Задачи исследования. Основной целью исследования является поиск закономерностей строения структур РНК, которые могут быть использованы для улучшения качества предсказания вторичных и пространственных структур РНК. Вспомогательная цель исследования состоит в разработке единого языка описания произвольной вторичной структуры РНК, который может быть использован для систематизации и анализа структурных мотивов РНК.

Исходя из поставленных целей, были сформулированы и решены следующие задачи:

1. Разработать модель описания вторичной структуры РНК при наличии псевдоузлов.

2. Создать базу данных пространственных структур РНК и структурных мотивов на основе предложенной модели.

3. Разработать классификацию третичных мотивов РНК на основе предложенной модели.

4. Исследовать возможность использования разработанной классификации для предсказания третичных мотивов РНК по данным о последовательности и вторичной структуре.

5. Исследовать строение псевдоузлов в экспериментально определенных пространственных структурах РНК.

Научная новизна исследования. Научная новизна работы состоит в следующих новых результатах:

1. Разработана модель описания произвольной вторичной структуры РНК, учитывающая псевдоузловые структуры.

2. Показано, что все псевдоузлы в экспериментально определенных пространственных структурах функциональных РНК образованы с участием энергетически неустойчивых коротких участков двойной спирали, состоящих из 2-3 спариваний оснований.

3. Показано, что в экспериментально определенных структурах РНК более половины мотивов типа А-минор существуют в кластерах. Описан новый тип мотива A-cluster.

4. Разработана классификация третичных мотивов РНК, отражающая связь со вторичной структурой и позволяющая выявлять функциональную роль отдельных представителей мотивов типа триплекс и А-минор.

5. Сформулирована и решена задача предсказания триплексов рибонуклеотидов по данным о последовательности и вторичной структуре РНК. Показано, что разработанная классификация третичных мотивов значимо повышает результаты предсказания.

Положения, выносимые на защиту.

1. Предложенная модель описания вторичной структуры РНК позволяет использовать для анализа мотивов произвольные структуры РНК, в том числе содержащие псевдоузлы.

2. Многообразие псевдоузловых структур обеспечивается энергетически неустойчивыми короткими участками двойной спирали, состоящими из 2-3 спариваний оснований.

3. Предложенная классификация третичных мотивов РНК позволяет выявлять функциональные особенности отдельных представителей различных классов мотивов, а также может быть использована для предсказания третичных мотивов по данным о последовательности и вторичной структуре РНК.

Научная значимость и практическая ценность исследования. Научная значимость исследования заключается в постановке нового варианта задачи предсказания третичных мотивов РНК, а также в разработанной методике решения этой задачи. Практическая ценность состоит в подготовленной базе данных пространственных структур РНК, которая может быть использована как источник исходных данных о структурных мотивах РНК. Результаты работы использовались при выполнении проектов РФФИ 14-01-93106-НЦНИЛ_а «Алгоритмы на графах в задачах

молекулярной биологии» и 16-04-01640-А «Структурные мотивы РНК: классификация, поиск, база данных, алгоритмы». Результаты работы можно рекомендовать к использованию для улучшения качества решения в задачах предсказания вторичных и пространственных структур РНК, а также в задаче поиска генов некодирующих РНК.

Апробация результатов исследования. Результаты работы были представлены на международных конференциях MCCMB'11 (Москва, 2011 г.), MCCMB'13 (Москва, 2013 г.), BGRS'14 (Новосибирск, 2014 г.), ICMBB'14 (Пущино, 2014 г.), MCCMB'15 (Москва, 2015 г.), MCCMB'17 (Москва, 2017 г.), ICMBB'18 (Пущино, 2018 г.), ECCB'18 (Афины, 2018 г.), на семинарах в Институте математических проблем биологии РАН - филиале Института прикладной математики им. М.В. Келдыша РАН, Институте белка РАН.

Публикации по теме исследования. По результатам работы опубликовано 14 печатных работ, в том числе 5 статей в журналах и 1 статья в сборнике.

ГЛАВА 1. Обзор литературы

1.1 Базы данных пространственных структур РНК

В настоящее время существует около 15 баз данных, которые содержат сведения о пространственных структурах РНК [19, 25-37]. Стоит отметить, что кроме баз структур большое значение при изучении РНК имеют базы, основанные на множественном выравнивании семейств РНК, например, Rfam [38].

Все базы пространственных структур РНК основаны на данных PDB [39] (в случае RNAStrand [35] – и на других источниках) и различаются возможностями поиска (например, поиск только по структурам в целом или по отдельным элементам), полнотой охвата известных структур (например, включены ли в базу сведения о РНК-белковых комплексах), полнотой описания отдельных структур (некоторые базы содержат сведения только о петлях, в большинстве баз нет сведений о псевдоузлах). К сожалению, ни одна из рассматриваемых баз данных не обеспечивает полный набор возможностей, необходимых исследователю.

Во всех изученных работах, так или иначе, используются инструменты разметки водородных связей по пространственным координатам атомов. Для некоторых баз с этой целью были созданы собственные программы, однако большинство баз используют один из трех наиболее популярных инструментов: FR3D, RNAView и MC-Annotate [40-42].

Все существующие базы данных можно условно разделить на три типа: вспомогательные базы для структурного моделирования; базы, посвященные отдельным элементам вторичной структуры РНК и универсальные базы пространственных структур РНК. Остановимся подробнее на последнем классе баз. Наиболее интересными с точки зрения полноты имеющихся данных и функциональности веб-интерфейса представляются базы RnaFRABASE [33] и RNAStrand. Можно отметить такие специфичные достоинства указанных баз, как наличие гибкого поиска структурных

элементов (RnaFRABASE) и обширность и полноту представленных данных (RNAStrand). Однако данные базы имеют и важные недостатки, так, например, в RnaFRABASE исключены псевдоузловые структуры, а в RNAStrand отсутствует возможность поиска отдельных структурных элементов. Более того, в RNAStrand представлены не все содержащие РНК структуры из PDB. К универсальным можно отнести и базу RNA3DHub [31]. Однако данная база не рассматривает такой важный элемент вторичной структуры РНК, как спирали, а, значит, не вполне соответствует требованиям универсальности. Тем не менее, с точки зрения выделения мотивов, которые встречаются в различных пространственных структурах, база RNA3DHub, как и другие разработки группы Н. Леонтиса и К. Зирбеля, представляют большой интерес. С точки зрения удобства пользования и полноты охвата известных структур интерес представляет и база NPIDB [28], однако эта база специализирована для анализа нуклеиново-белковых контактов и не содержит разметки вторичной структуры РНК.

Следует отметить, что, в отличие от многих биоинформатических баз данных, в частности, баз данных по первичным структурам РНК и белков, базы данных пространственных структур РНК используются, в основном, для гомологического моделирования, подбора knowledge-based параметров и тестирования программ предсказания и сравнения структур РНК [43-49] (в дополнение к упоминавшимся базам стоит указать специализированную базу данных BRASERO [50]). Их потенциал для собственно биологических исследований используется недостаточно.

1.2 Псевдоузлы

Традиционно [51] псевдоузел определяется, как элемент вторичной структуры РНК, в котором нарушено следующее условие вложенности: если в цепи РНК спарены нуклеотиды *(i, j)* и *(m, n)*, то отрезки *[i, j]* и *[m, n]* либо не пересекаются, либо один из них лежит строго внутри другого.

Отметим, что это определение не указывает точно, какой именно фрагмент РНК относится к псевдоузлу, общепринятого определения сейчас нет. Обзор различных определений дан в работе [17]. Однако в этой работе лишь перечислены классы структур, которые охватываются тем или иным алгоритмом предсказания вторичных структур РНК. Для анализа всех встречающихся в природе псевдоузловых структур представляется важным разработать классификацию, основанную не на внешних обстоятельствах (свойствах алгоритмов), а на свойствах самих псевдоузлов. Классификация А. Кондон была расширена в работе [52] за счет учета алгоритмов, опубликованных после выхода работы А. Кондон [17].

Псевдоузловые структуры встречаются в различных молекулах РНК, играющих важную роль в жизнедеятельности клетки. В качестве примеров можно указать молекулы большой рибосомальной РНК, РНК рибонуклеазы Р и других. Однако традиционно основное внимание при изучении РНК уделяется структурам, не содержащим псевдоузлов. В экспериментальных исследованиях это связано с тем, что структуры, содержащие псевдоузлы, значительно менее распространены (~20%). А с другой стороны определение энергетических параметров для них более затруднительно. С алгоритмической точки зрения наличие псевдоузлов делает невозможным использование алгоритмов динамического программирования [53-56] и вынуждает использовать другие, менее эффективные методы [57-60]. Таким образом, алгоритмы предсказания структур, содержащих псевдоузлы, с одной стороны имеют большую временную сложность, а с другой стороны некоторые из них допускают наличие чрезмерно общего класса псевдоузлов, возможность которых в реальных РНК вызывает сомнения, см. обзор в [17]. Еще одной причиной недостаточного качества программ, предсказывающих по последовательности структуры с возможным наличием псевдоузлов, является то, что физические параметры, используемые в программах предсказания, определены хуже, чем аналогичные параметры для предсказания классических структур РНК, см. [61].

Единственной базой, которая включает в себя более или менее универсальный инструмент анализа псевдоузловых структур является база RNAStrand. Однако, и этот инструмент не лишен недостатков. Так, он не позволяет анализировать петли, находящиеся внутри псевдоузлового участка (это можно увидеть, например, при изучении структур 1FFK и 1BJ2). Помимо RNAStrand информация о псевдоузлах представлена в таких базах, как PseudoBase++ [19] и RNAJunctions [34]. Однако, эти базы организованы как набор разрозненных классов псевдоузловых структур, выбор этих классов носит эмпирический характер и, например, обнаруженная нами на предварительном этапе исследований структура «тройной узел» [62] в этих базах не описана. Кроме того, эти базы не универсальные; они описывают только отдельные элементы вторичных структур (псевдоузлы в PseudoBase++; петли и псевдоузлы типа “kissing hairpins” в RNAJunctions).

Возможно, одной из причин отсутствия баз по структурам, содержащим псевдоузлы, является то, что в настоящее время недостаточно разработан язык для описания псевдоузлов, Описаны лишь наиболее простые (но наиболее распространенные) виды таких структур, например, kissing hairpins и узлы H-типа. В работе [63] была предложена топологическая классификация псевдоузлов. Этот подход представляется наиболее перспективным, однако он до настоящего времени не использовался при построении баз данных.

1.3 Алгоритмы предсказания классических вторичных структур РНК

Предсказание вторичной структуры РНК – одна из классических задач вычислительной молекулярной биологии. Знание оптимальной, т.е. имеющей минимально возможную свободную энергию, вторичной структуры молекулы РНК является решающим для понимания функции РНК [54, 64-67]. В основном, методы компьютерного предсказания оптимальных структур

РНК рассматривали структуры, не содержащие псевдоузлов, мы здесь также рассмотрим только такие алгоритмы.

Методы, описанные в пионерских работах [68, 69] впоследствии совершенствовались в трех направлениях. Первое направление – использование все более реалистичных энергетических функций. В ранних работах (например, [53]) свободная энергия оценивалась просто как число пар нуклеотидов, образующих водородные связи. В настоящее время используется значительно более сложная и точная модель NNM (Nearest Neighbour Model, см. [70]). В такой модели вторичная структура РНК рассматривается как составленная из петель (loop) различных типов, таких как стекинг-пары, выпячивания (bulges), шпильки (hairpins), внутренние петли (internal loops) и мульти-петли (multi-branch loops). NNM включает правила, которые присваивают энергию петлям каждого из указанных типов; энергия полной структуры при этом будет суммой энергий составляющих ее петель. Параметры NNM уточнялись в серии экспериментальных работ (см. [12, 14] и обзор [13]).

Другое направление – рассмотрение различных объектов, связанных со вторичной структурой РНК. Среди таких объектов отметим множество пар оснований, входящих в субоптимальную структуру [71], множество субоптимальных структур [71, 72], статистическая сумма и вероятность наличия в структуре заданных спариваний нуклеотидов [54, 73], неветвящаяся оптимальная структура [74].

Алгоритмы для поиска данных объектов базируются на соответствующих вариантах метода динамического программирования. Удивительно, что наиболее сложной проблемой оказался анализ внутренних петель, т.е. петель, содержащих только две пары оснований, образующих водородные связи, и два региона с неспаренными нуклеотидами между ними. Алгоритм, вычисляющий все внутренние петли молекулы РНК длиной L за время $O(L^3)$ был предложен только в 1999 г [67]. Поиск оптимальной неветвящейся вторичной структуры (НВС) тесно соотносится с

определением энергий всех возможных внутренних петель. В работе [74] предложен алгоритм, использующий метод динамического программирования для разреженных матриц (SDP) , который находит оптимальную НВС. Время работы алгоритма имеет порядок $O(M*log^2(L))$, а требуемый объем памяти - $O(M)$, где M – число допустимых пар нуклеотидов и L – длина молекулы РНК. Очевидно, $M < L^2$; это означает, что время работы алгоритма имеет порядок $O(L^2*log^2(L))$.

Однако, этот интересный алгоритм не используется в программах предсказания вторичной структуры РНК. Причина в том, что он требует чтобы энергия внутренних циклов была выпуклой или вогнутой функцией от суммы длин двух неспаренных участков, которые образуют цикл. Между тем энергетические функции используемые в наиболее популярной в настоящее время модели NNM зависят как от суммы так и от разности двух этих длин. Также часто необходимо найти не только оптимальную НВС, но и множество всех «разумных» НВС, чего не позволяет метод SDP.

Третье направление связано с построением консенсусной структуры, общей для нескольких родственных РНК, см., например, [75, 76] и обзоры в этих работах. К этому направлению примыкают работы по выравниванию последовательностей РНК с учетом знаний о вторичной структуре, см., например [77]. Общим недостатком работ по построению консенсусной вторичной структуры является их невысокое быстродействие.

1.4 Анализ пространственных структур РНК

Вопрос о свойствах множества возможных пространственных структур РНК представляет интерес, как с теоретической (пример: эволюционная теория, см. [78]), так и с практической (пример: дизайн РНК, см. обзор [79]) точки зрения.

Среди существующих работ можно выделить два направления. Первое направление – это классификация структурных мотивов определенного типа и сбор сведений о вариациях таких структур, их встречаемости и т.п., т.е.

создание «атласов» в терминологии RNA 3D Motif Atlas [22]. В цитированной работе изучаются «мотивы» - устойчивые образования взаимодействующих нуклеотидов (“well-defined geometric arrangements of interacting nucleotides”). Авторы этой интересной работы предлагают автоматизированную методику выделения и кластеризации локальных мотивов (мотивов, которые являются разновидностями шпилек и внутренних петель), описывают возможные виды подобных мотивов. В частности, в этой работе описаны такие распространенные локальные мотивы, как Kink-turn [23], C-loop [80], Sarcin-Ricin [81], T-loop [82], GNRA [83] и др. Другой важной работой по описанию локальных мотивов является база SCOR [36]. Мотивы в указанной базе, кроме кластеризации, были также проклассифицированы в виде ациклического направленного графа мотивов и функционально аннотированы. К сожалению, база SCOR в настоящее время больше не существует.

Другим важным классом мотивов являются мотивы с дальнодействием (long-range motifs). Такие мотивы изучены значительно хуже, в силу сложностей в их идентификации. На данный момент не существует ни одной базы данных, содержащей такие мотивы, даже в виде коллекции отдельных их видов, которые являются наиболее распространенными. В работе [84] проведен поиск по неизбыточному множеству структур таких мотивов, как Coaxial helix [24], A-minor [85], ribose zipper [86], loop-loop receptor [87], tRNA D-loop/T-loop [88] и др. Результаты этой работы не были внедрены ни в одну базу данных, хотя доступны онлайн в виде неинтерактивных карт вторичной структуры с наложением третичных взаимодействий. В данной работе для поиска мотивов использовался инструмент FR3D [40]. Помимо FR3D наиболее популярными инструментами являются NASSAM [89] (поиск элементарных мотивов, таких как триплексы и третичные взаимодействия) и DSSR [90] (поиск, в том числе, A-minor, Kink-turn, U-turn [91] и ribose zipper). Стоит также отметить, что с 2015 года основным форматом банка PDB стал формат mmCIF [92]. На данный момент программа NASSAM не может

обрабатывать формат mmCIF; программа DSSR начала обрабатывать mmCIF с февраля 2015 года.

Анализу видов псевдоузлов посвящены работы [93] и [94]. В данных работах предлагается топологическая классификация псевдоузлов (подобная классификация предложена также в [63]) и анализируется ряд представленных в PDB структур с точки зрения этой классификации. Однако, представленные в статьях данные неполны с точки зрения современного состояния базы PDB.

Другое направление работ связано с анализом распределений различных характеристик мотивов пространственной структуры. Интересным примером таких работ является работа [95]. В этой работе показана асимметрия в длинах петель и спиралей в псевдоузлах H-типа и дается объяснение этого эффекта с точки зрения свойств малой и большой бороздок спиралей РНК; см. также обзор [96]. Работы этого направления важны и с точки зрения эволюционной теории. Так, в цитированной работе [78] исследуется компенсаторная эволюция митохондриальных РНК. Методика этой работы основана на анализе статистики спариваний нуклеотидов в сочетании с методами сравнительной геномики. Работа выполнена, используя только данные о корреляциях в последовательностях РНК, что ограничивает исследование Уотсон-Криковскими спариваниями (канонические спаривания).

1.5 Выводы и результаты по главе

Существующие в настоящее время базы данных пространственных структур РНК имеют ряд недостатков, среди них: ограниченные возможности поиска структурных элементов, исключение РНК-белковых комплексов, исключение псевдоузловых структур РНК и др.

Отсутствие модели описания произвольной вторичной структуры РНК обусловлено проблемой предсказания псевдоузловых структур. Задача

предсказания вторичной структуры РНК при наличии псевдоузлов является NP-полной, поэтому основное внимание уделяется классическим структурам.

Анализ мотивов пространственной структуры РНК не носит систематический характер. Изученные работы ограничиваются либо локальными мотивами, либо рассматривают отдельные классы мотивов с дальнодействием.

Представляется необходимой попытка разработать единый язык описания структурных мотивов (как локальных, так и мотивов с дальнодействием), что позволит не только внедрить разметку всех мотивов в аннотированную базу данных, но и открыть новые, более редкие виды повторяющихся мотивов с дальнодействием, обладающих, возможно, важными функциями.

ГЛАВА 2. Модель описания вторичной структуры РНК

Описанные в данной главе результаты опубликованы в работах [97, 98].

2.1 Основные определения

Молекулу РНК мы будем представлять, как последовательность нуклеотидов, иначе говоря, как символьную последовательность в алфавите *{A, C, G, U}*. Каждый нуклеотид в молекуле имеет свой номер от *1* до *L*, где *L* – длина последовательности.

Связь (Спаривание) – это пара нуклеотидов *(i, j),* где $i < j$, которая образует водородные связи. При этом допускаются не только связи между комплементарными нуклеотидами (A-U и G-C Watson-Crick pairs) и G-U связи (Wobble pairs), но и неканонические связи, см. [90, 99].

Спираль (Стем) – это последовательность пар нуклеотидов вида *(i, j), (i+1, j-1),…, (i+k, j-k)* такая, что

1) $i < j,\ i+k < j-k,\ k > 1;$

2) все пары вида *(i+x, j-x),* где $x = 0, \ldots, k,$ образуют Уотсон-Криковские связи (WC-связи), т.е. связи между комплементарными нуклеотидами, или G-U связи.

Участок цепи *[i, i+k]* будем называть *левым крылом* стема, соответственно участок *[j-k, j]* будем называть *правым крылом* стема.

Пару *(i, j)* будем называть *внешней парой* стема или *торцом* стема, пару *(i+k, j-k)* будем называть *внутренней парой* стема.

Нить – это такой участок цепи *[i, j],* где $i < j$, что

1) не существует такой WC-связи или G-U связи *(k, t),* что $i <= k <= j$ или $i <= t <= j$ и связь *(k, t)* является парой стема;

2) существуют пары, являющиеся частью стемов, в которые входят нуклеотиды *i-1* и *j+1*.

Замечание. Допускаются нити «нулевой длины», для их обозначения используется запись *[i+1, i]*, где *i* – номер последнего нуклеотида

предшествующего крыла.

Вторичная структура РНК – это такое множество WC и G-U связей, что

1) каждый нуклеотид входит не более чем в одну связь;

2) каждая пара входит в некоторую спираль.

Отметим, что в экспериментально определенных пространственных структурах РНК есть значительное число водородных связей, не входящих в стемы [100], роль таких связей в настоящее время изучена слабо. Мы исходим из предположения, что полезно отдельно рассматривать «базовую» вторичную структуру, образованную стемами, и (над этой структурой) – одиночные спаривания, называемые *линками*.

Линк – одиночное спаривание *(i, j),* не являющееся частью стема.

Будем говорить, что два стема (стем и линк, два линка) находятся *в конфликте,* если между крыльями одного стема (линка) находится одно, и только одно крыло другого стема (линка).

Псевдоузел – участок вторичной структуры, содержащий хотя бы одну пару стемов, находящихся в конфликте друг с другом (см. Рис. 2.1).

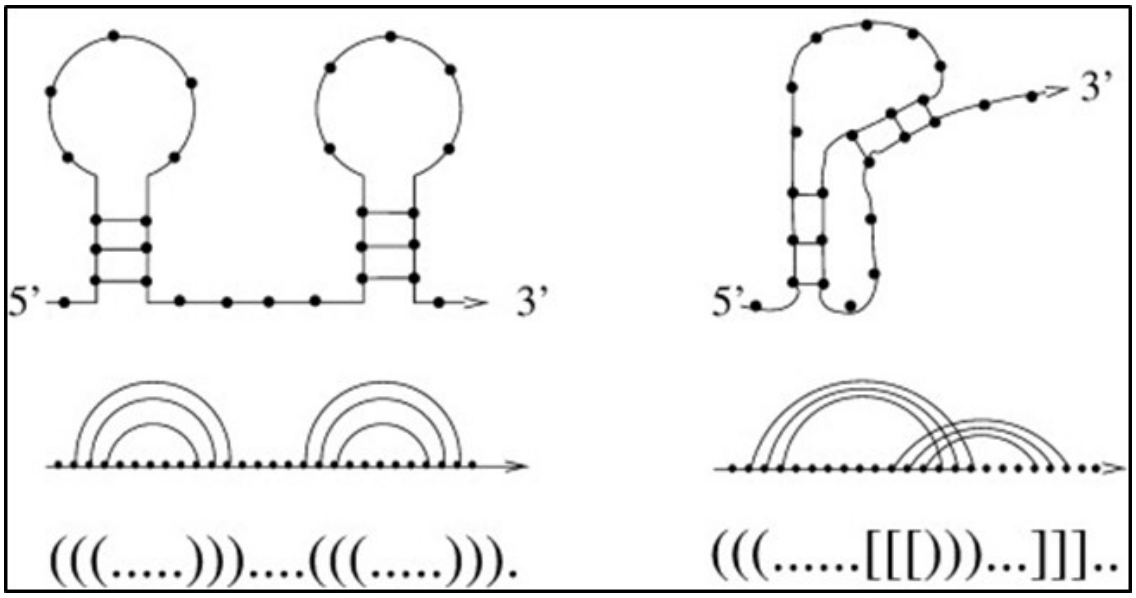


Рисунок 2.1. Пример структуры без псевдоузлов (слева) и с псевдоузлами (справа)

По наличию конфликтов линки делятся на три типа. Линк называется *внутренним*, если он не конфликтует с другими стемами (линками); *связанным*, если он конфликтует с линками, но не со стемами; *свободным*, если он конфликтует со стемами.

2.2 Стемы и петли

Здесь и далее будем считать фиксированной цепь РНК с заданной на ней вторичной структурой. Эту цепь можно рассматривать как чередующуюся последовательность нитей и крыльев. Для удобства изложения мы будем считать, что перед первым и после последнего нуклеотида цепи добавлены крылья «внешнего стема» (ср. с [101]).

С каждым стемом связан *внутренний* по отношению к нему участок цепи – участок между концом левого крыла и началом правого крыла, иначе говоря, – между нуклеотидами, образующими внутреннюю пару стема. Для фиктивного внешнего стема внутренним участком является вся исходная последовательность РНК.

Пусть *H* – стем и *(i, j)* – его внутренняя пара.

Определение 1. Позиция цепи *t* – *внутренняя* для стема *H* (синоним: *лежит внутри H*) , если $i < t < j$. Фрагмент цепи - *внутренний* для стема *H* (синоним: *лежит внутри H*) , если все его позиции – внутренние для стема *H*. Стем *H1 лежит внутри* стема *H* (является *внутренним* для *H*), если все позиции его крыльев – внутренние для *H*.

Определение 2. Позиция цепи *t принадлежит* стему *H*, если она внутренняя для *H* и не существует стема *H1*, лежащего внутри *H*, такого, что $x < t < y$, где *(x, y)* – внешняя пара (торец) *H1*.

Определение 3. *Петля* стема *H* – это множество всех позиций, принадлежащих стему *H*.

Очевидно, каждая позиция, не входящая в связь, принадлежит хотя бы одной петле – обычной или внешней. При этом если какая-то позиция нити

(крыла) принадлежит некоторой петле, то и вся нить (все крыло) принадлежит этой петле.

Если в структуре нет псевдоузлов, то каждая петля в смысле определения 3 является петлей согласно модели NNM и наоборот. При этом каждая нить принадлежит ровно одной петле (возможно, внешней), а ни одно крыло не принадлежит какой-либо петле. Для структур с псевдоузлами оба эти свойства нарушаются.

2.3 Структура петель

Определение 4. Пусть *H* – стем и *(u, v)* – его внутренняя пара. Участок, *[i, j]* называется *элементарным замкнутым участком* относительно *H* (в общем случае *элементарный замкнутый относительно стема участок*, *stem-related elementary closed region, S-ECR*)**,** если

1) *[i, j]* лежит внутри *H*;

2) не существует таких связей *(k, t),* что *(i <= k <= j < t < v)* или *(u < k < i <= t <= j);*

3) существуют связи *(i, k)* и *(t, j),* где *k <= j; i <= t.*

4) не существует отличного от *[i, j]* участка *[i', j']* такого, что *i' <= i < j <= j'* и участок *[i', j']* удовлетворяет условиям 1) - 3).

Пара нуклеотидов *(i, j)* называется *торцом* замкнутого относительно *H* участка.

Утверждение 1. Пусть *Z = [f, g]* – участок, замкнутый относительно стема *H*; *(u, v)* – внутренняя пара стема *H*. Тогда:

1) Участок *Z* целиком лежит внутри спирали *H*.

2) Крыло либо целиком лежит в *Z,* либо целиком лежит вне *Z.*

3) Замкнутый относительно *H* участок начинается левым крылом некоторого стема *H1*, лежащего внутри *H*, и заканчивается правым крылом некоторого стема *H2*, лежащего внутри *H*.

4) Если *H1 = H2* – это один и тот же стем, то торец *(s, t)* участка *Z* – это торец данного стема. В противном случае *s* – это начало левого крыла стема

H1, *t* – это конец правого крыла стема *H2*.

Доказательство – следует из определения 4 и того, что крылья не пересекаются.

Определение 5. Пусть *Z* – участок, замкнутый относительно стема *H*. Участок *Z* называется *простым*, если его торец – это торец некоторого стема и *сложным* в противном случае. Сложные участки для краткости будем называть *блоками* (см. Рис. 2.2).

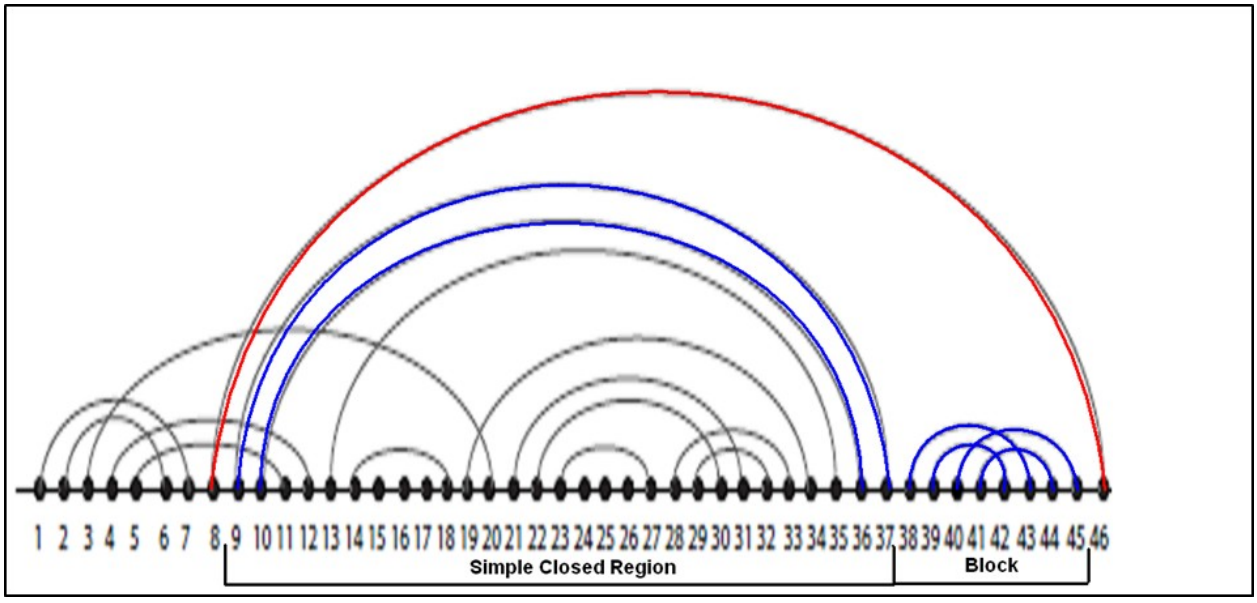


Рисунок 2.2. Примеры замкнутых участков

Утверждение 2. Пусть *H* – стем; *(u, v)* – его внутренняя пара. Тогда

1) Никакие два участка, замкнутых относительно *H*, не пересекаются.

2) Пусть позиция *t* лежит внутри спирали *H*. Позиция *t* НЕ принадлежит спирали *H* тогда и только тогда, когда *t* лежит внутри некоторого участка *Z,* замкнутого относительно *H* (т.е. лежит в *Z,* но не входит в его торец).

Доказательство – следует из определений 1, 2 и 4.

Определение 6. Пусть *H* – стем и *(u, v)* – его внутренняя пара. Пусть $(s_1, t_1), \ldots, (s_n, t_n)$ – торцы всех участков, замкнутых относительно *H*; $s_1 < t_1 < \ldots < s_n < t_n$. Для удобства пусть $t_0 = u$; $s_{n+1} = v$. Пусть *k* – целое; $1 <= k <= n + 1$. Тогда *k*-я *грань* петли *H* – это фрагмент $[t_{k-1}+1, s_k-1]$.

Замечание. Если $s_k = t_{k-1}+1$, то *k*-я грань петли *H* – пустой отрезок.

Утверждение 3. Пусть H – стем и (u, v) – его внутренняя пара. Пусть $(s_1, t_1), \ldots, (s_n, t_n)$ – торцы всех участков, замкнутых относительно H; $s_1 < t_1 < \ldots < s_n < t_n$.. Для удобства пусть $t_0 = u$; $s_{n+1} = v$. Тогда петля стема H – это объединение торцов всех участков, замкнутых относительно H, и расположенных между ними граней.

Доказательство – следует из утверждения 2.

Утверждение 4. Пусть H – стем и (u, v) – его внутренняя пара и позиция x принадлежит грани (t, s) петли стема H. Тогда

1) Позиция x либо не участвует в связи, либо принадлежит крылу стема H', другое крыло которого лежит вне стема H.

2) Если x принадлежит нити (крылу стема), то все позиции этой нити (этого крыла) принадлежат той же грани петли стема H.

Доказательство – следует из определения граней и того, что крылья не пересекаются.

Утверждения 3 и 4 описывают возможные структуры петель. Отметим, что в случае структур, которые не содержат псевдоузлов, все замкнутые участки – простые и каждая грань состоит из единственного однонитевого участка. Поэтому можно дать такое определение.

Определение 7. Петля называется *классической* (classical), если она не содержит крыльев и торцов блоков. Петля называется *изолированной* (isolated), если она не содержит крыльев. и *узловой* (pseudoknotted), если она содержит крылья.

Стем называется *узловым*, если его петля – узловая.

Применим классификацию петель модели NNM к введенному нами обобщению, основываясь на количестве торцов, входящих в петлю. Отметим, что в нашем случае торцы могут быть как торцами стемов (иными словами – простых замкнутых участков), так и торцами блоков (сложных замкнутых участков).

Определение 8. Петля называется *шпилькой* (hairpin loop), если она не содержит торцов и, соответственно имеет одну грань. Петля называется

внутренней (internal loop), если она содержит ровно один торец, и, соответственно, имеет две грани. Петля называется *мульти-петлей* (multiple junction, multiple loop), если она содержит более одного торца, и, соответственно, более двух граней (см. Рис. 2.3).

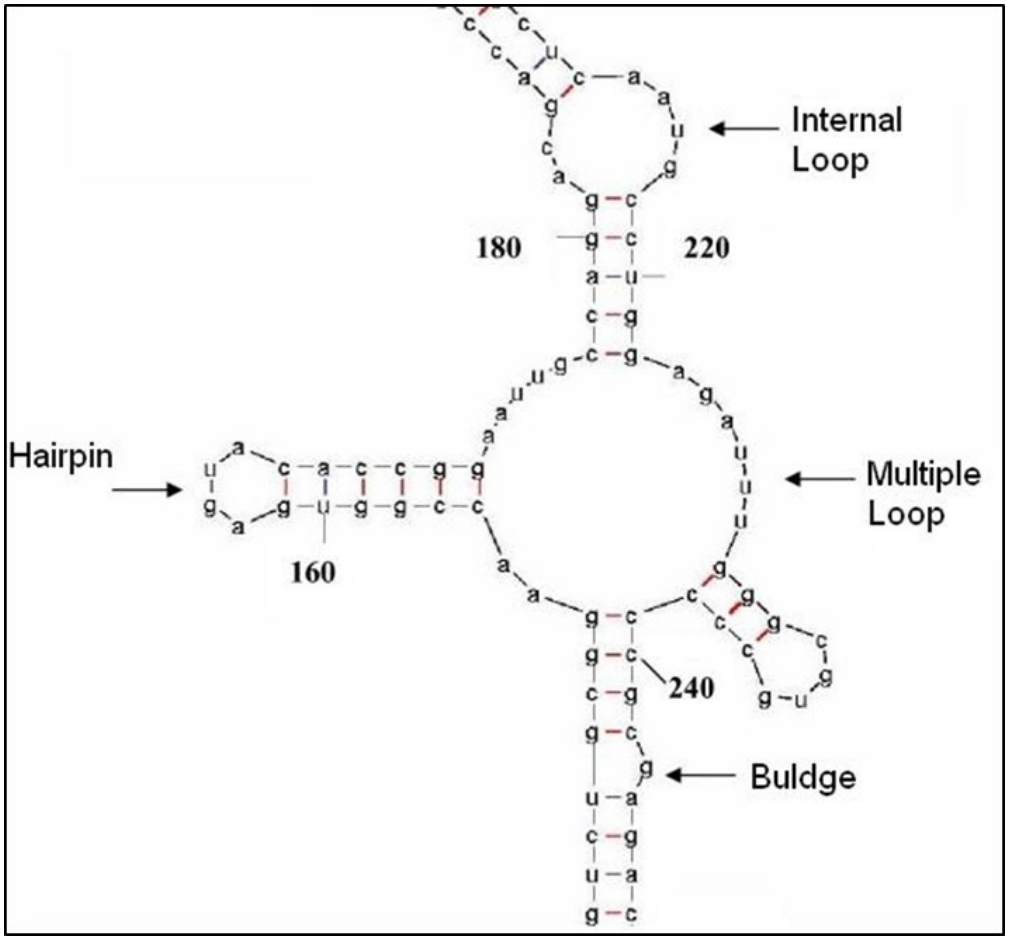


Рисунок 2.3. Различные типы петель

Замечание 1. Будем называть выпячиванием (Bulge) такую внутреннюю петлю, одна из граней которой является нитью нулевой длины.

Замечание 2. Данная классификация распространяется как на обычные, так и на внешние петли (принадлежащие «внешним» стемам).

2.4 Псевдоузлы

Элементарный замкнутый участок (ЭЗУ, elementary closed region, ECR) – минимальный участок *[i, j],* такой что:

1) Не существует связи *(k, l)*, такой что *(i <= k <= j; l > j)* или *(k < i; i <= l <= j)*;

2) Не существует позиции *l*, такой что $i < l < j$ и оба участка *[i, ..., l]* и *[l +1, ..., j]* удовлетворяют условию 1);

3) Существуют связи *(i, k)* и *(l, j)*; допускаются равенства $k = j$ и $i = l$.

Пара *(i, j)* называется *торцом* ЭЗУ *[i, j]*. Отметим, что если пара *(i, j)* является связью и принадлежит некоторому стему, то торец ЭЗУ совпадает с торцом данного стема.

ЭЗУ *[k, l]* является *суб-ЭЗУ* (sub-ECR) относительно ЭЗУ *[i, j],* если $i < k < l < j$ и не существует такого ЭЗУ *[m, n]*, что $i < m < k < l < n < j$.

ЭЗУ называется *псевдоузлом* (синоним: *псевдоузловой ЭЗУ*) если принадлежащие ему стемы находятся в конфликте. В противном случае ЭЗУ называется свободным от псевдоузлов или *классическим*.

Классификация псевдоузлов, реализованная в разработанной базе данных URSDB, основана на понятии *сигнатуры*. Классификация схожа с топологической классификацией псевдоузлов, предложенной в работе [102]. Основным отличием нашей классификации является исключение из рассмотрения одиночных спариваний (линков).

Рассмотрим все спирали ЭЗУ и обозначим их буквами латинского алфавита в соответствии с позициями их крыльев от 5`- к 3`-концу. Левое крыло будем обозначать строчными буквами, например, **a**, а правое крыло – заглавными буквами, например, **A**. Таким образом, каждый стем будет обозначен двумя буквами, например, **aA**.

Полной сигнатурой ЭЗУ называется последовательность его крыльев в соответствии с их позициями от 5`- к 3`-концу.

Пример 1. Пусть ЭЗУ *[10, 70]* содержит три стема, *([10, 15]; [65, 70]), ([20, 25]; [45, 50]), ([30, 35]; [55, 60])*, здесь *[10, 15]* и *[65, 70]* – крылья

стема *([10, 15]; [65, 70])*, и т.д. Тогда стем *([10, 15]; [65, 70])* обозначим как **aA**, стем *([20, 25]; [45, 50])* – **bB**, а стем *([30, 35]; [55, 60])* – **cC**. Полной сигнатурой данного ЭЗУ будет последовательность **abcBCA**. Участок *[20, 60]* является суб-ЭЗУ относительно исходного ЭЗУ (см. Рис. 2.4, 2.5).

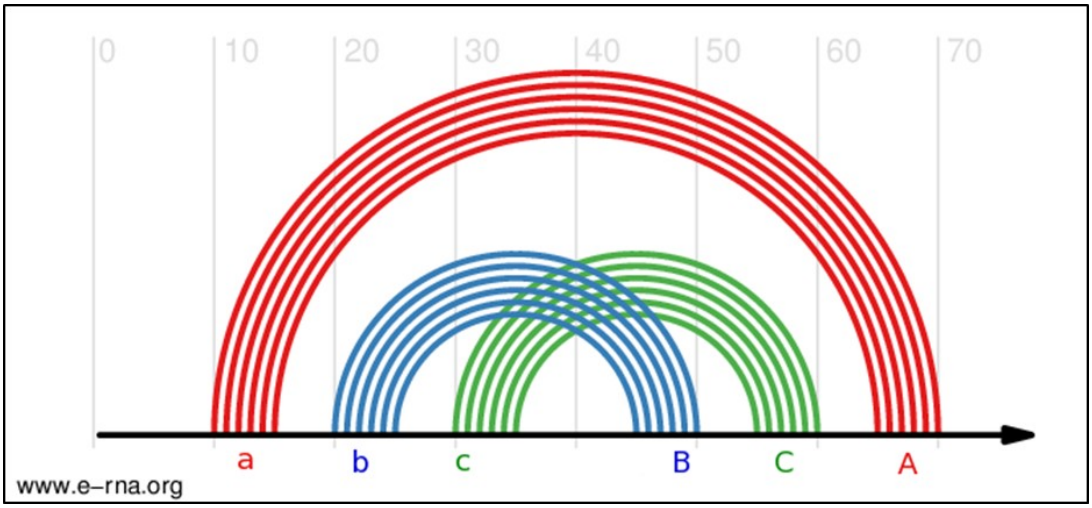


Рисунок 2.4. Дуговая диаграмма крыльев ЭЗУ из примера 1.

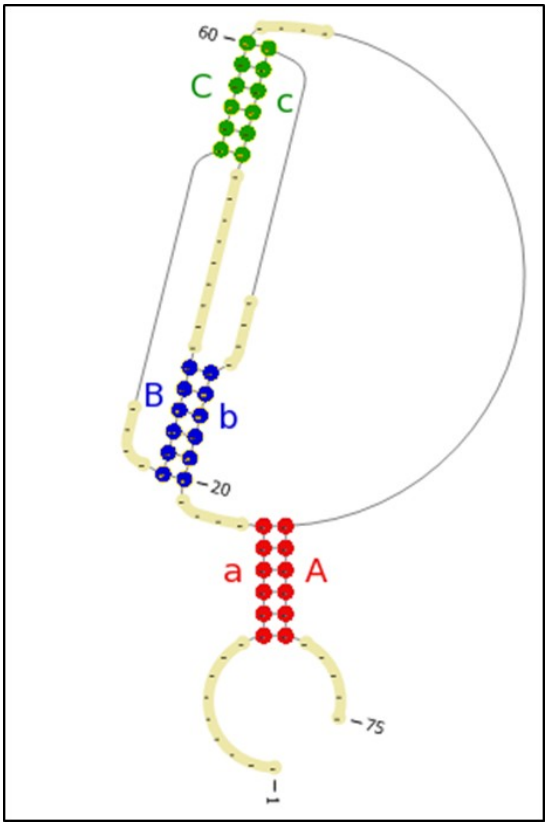


Рисунок 2.5. Схема вторичной структуры ЭЗУ из примера 1.

Пример 2. Пусть ЭЗУ содержит четыре спирали, *([10, 15]; [70, 75]), ([20, 25]; [50, 55]), ([30, 35]; [40, 45]), ([60, 65]; [80, 85])*, здесь *[10, 15]* и *[70, 75]* – крылья стема *([10, 15]; [70, 75])*, и т.д. Тогда стем *([10, 15]; [70,*

75]) обозначим как **aA**, стем *([20, 25]; [50, 55])* – **bB**, стем *([30, 35]; [40, 45])* – **cC**, а стем *([60, 65]; [80, 85])* – **dD**. Полной сигнатурой данного ЭЗУ будет последовательность **abcCBdAD**. Участок *[20, 55]* является суб-ЭЗУ относительно исходного ЭЗУ, а участок *[30, 45]* является суб-ЭЗУ относительно участка *[20, 55]* (см. Рис. 2.6, 2.7).

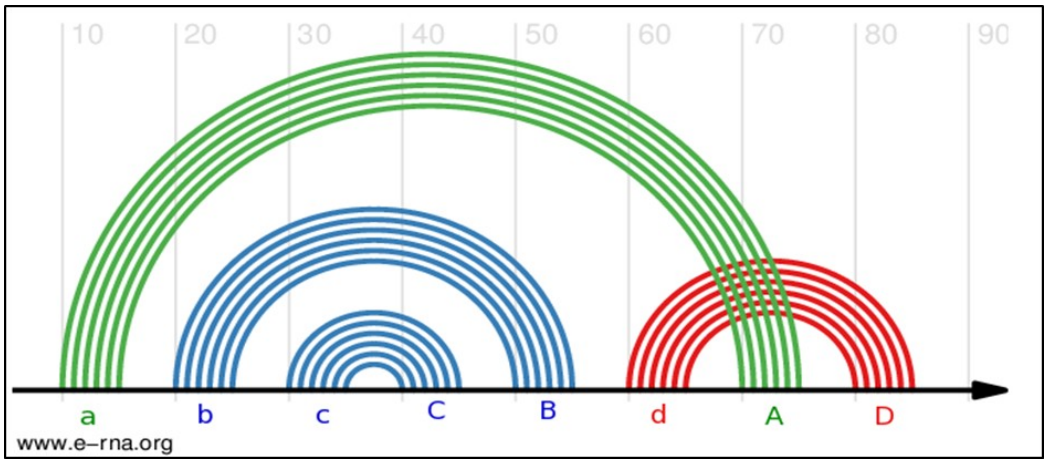


Рисунок 2.6. Дуговая диаграмма крыльев ЭЗУ из примера 2.

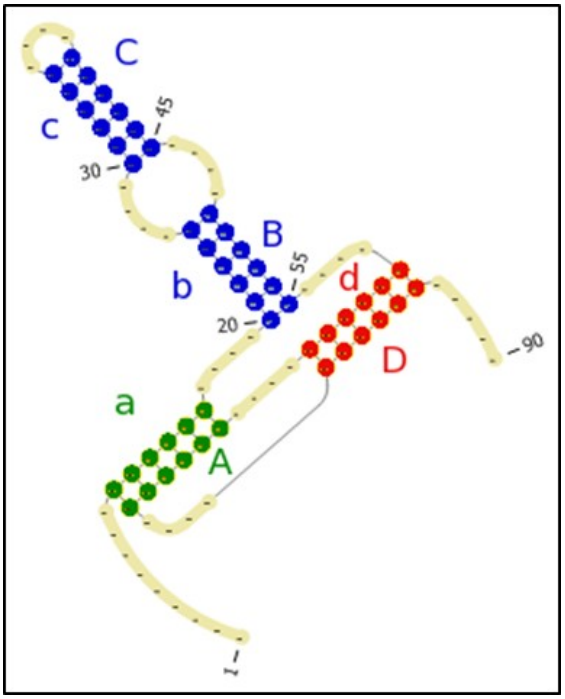


Рисунок 2.7. Схема вторичной структуры ЭЗУ из примера 2.

Пример 3. Пусть ЭЗУ содержит шесть стемов, *([2, 7]; [90, 95]), ([10, 15]; [80, 85]), ([20, 25]; [50, 55]), ([30, 35]; [40, 45]), ([60, 65]; [120, 125]), ([70, 75]; [110, 115])*, здесь *[2, 7]* и *[90, 95]* – крылья стема *([2, 7]; [90, 95])*, и т.д. Тогда стем *([2, 7]; [90, 95])* обозначим как **aA**, стем *([10, 15]; [80, 85])* – **bB**, стем *([20, 25]; [50, 55])* – **cC**, стем *([30, 35]; [40, 45])* – **dD**, стем *([60,*

65]; [120, 125]) – **eE**, а стем *([70, 75]; [110, 115])* – **fF**. Полной сигнатурой данного ЭЗУ будет последовательность **abcdDCefBAFE**. Участок *[20, 55]* является суб-ЭЗУ относительно исходного ЭЗУ, а участок *[30, 45]* является суб-ЭЗУ относительно участка *[20, 55]* (см. Рис. 2.8, 2.9).

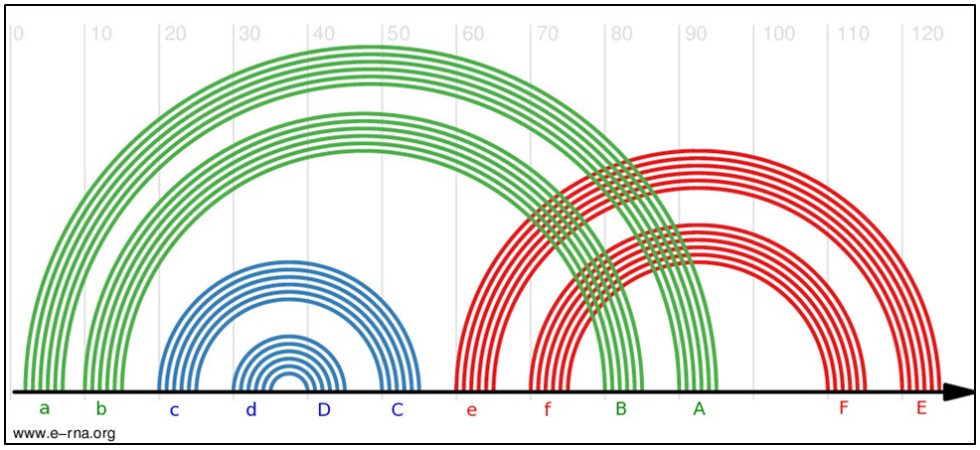


Рисунок 2.8. Дуговая диаграмма крыльев ЭЗУ из примера 3.

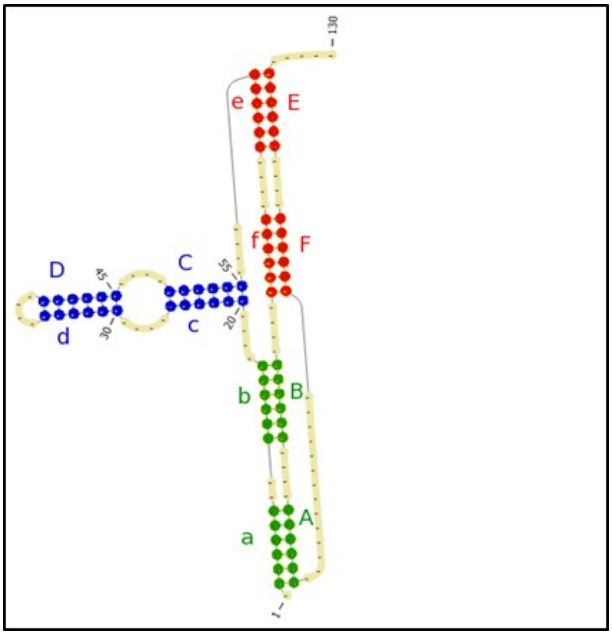


Рисунок 2.9. Схема вторичной структуры ЭЗУ из примера 3.

Верхняя сигнатура ЭЗУ получается из полной сигнатуры в результате:

1) Удаление крыльев, соответствующих всем суб-ЭЗУ;

2) Переименования стемов в порядке использования последовательных букв латинского алфавита (сохраняя порядок крыльев).

Верхней сигнатурой ЭЗУ из примера 1 будет последовательность **aA**; участок **bcBC**, отвечающий суб-ЭЗУ *[20, 60]*, был удален из полной сигнатуры **abcBCA**.

Верхней сигнатурой ЭЗУ из примера 2 будет последовательность **abAB**. Сначала, участок **bcCB**, отвечающий суб-ЭЗУ *[20, 55]*, был удален из полной сигнатуры **abcCBdAD**. Затем мы, заменяя **d** и **D** на **b** и **B**, получаем **abAB**.

Аналогично, верхней сигнатурой ЭЗУ из примера 3 будет последовательность **abcdBADC**.

Стемы **xX**, **yY**, **...** называются *связанными в верхней сигнатуре*, если оба слова **xy...** и **...YX** являются подстроками верхней сигнатуры.

Сигнатура ЭЗУ (*усеченная сигнатура* ЭЗУ) – это последовательность, полученная из верхней сигнатуры в результате:

1) удаления всех букв, кроме **x** и **X** (первая буква левой части и последняя буква правой части), соответствующих цепочкам связанных стемов.

2) Переименования стемов в порядке использования последовательных букв латинского алфавита (сохраняя порядок крыльев).

Сигнатуры ЭЗУ из примеров 1 и 2 совпадают со своими верхними сигнатурами. Сигнатурой ЭЗУ из примера 3 будет последовательность **abAB**, что совпадает с сигнатурой ЭЗУ из примера 2.

Примеры типичных сигнатур:

a) H-узел (H-knot): **abAB**;

b) «Целующиеся петли» (Kissing Loops): **abAcBC**;

c) Тройной узел (Triple knot): **abcABC**.

2.5 Выводы и результаты по главе

В данной главе дается описание предлагаемой нами модели описания вторичной структуры РНК, а именно: вводится вся необходимая терминология, включая основное для предлагаемого подхода понятие петли, обобщающее понятие петли модели NNM, а также доказываются утверждения, позволяющие установить общий вид петель и согласовать новую модель с моделью NNM.

В главе вводятся принципиально новые понятия, такие как линк, грань, блок и др., которые позволяют обобщить модель NNM на случай псевдоузловых структур. Также, наглядно доказано, что при отсутствии псевдоузлов представленные модели полностью совпадают.

ГЛАВА 3. База данных

Описанные в данной главе результаты опубликованы в работе [98].

3.1 Общие сведения

В качестве исходных данных были выбраны все РНК-содержащие экспериментально определенные пространственные структуры из банка данных PDB (Protein Data Bank, www.rcsb.org, [39]). По состоянию на апрель 2019 года база данных включала структуры из более чем 4300 документов банка PDB в .mmCIF формате. В связи с тем, что с начала 2015 года основным форматом PDB стал формат mmCIF [92], разработанная нами база данных также перешла на использование документов в данном формате. Преимуществом формата mmCIF (см. Рис. 3.1) является отсутствие ограничений на размер документа, что позволило объединить в один файл многие разделенные из-за ограничений формата .pdb структуры.

```
_struct_ref_seq.pdbx_auth_seq_align_beg
_struct_ref_seq.pdbx_auth_seq_align_end
1 1 2KMJ A 1 ? 28 ? 2KMJ 16 43 16 46
2 2 2KMJ B 1 ? 4  ? 2KMJ 1  4  1  4
3 2 2KMJ C 1 ? 4  ? 2KMJ 1  4  1  4
#
loop_
_chem_comp.id
_chem_comp.type
_chem_comp.mon_nstd_flag
_chem_comp.name
_chem_comp.pdbx_synonyms
_chem_comp.formula
_chem_comp.formula_weight
G   'RNA linking'       y "GUANOSINE-5'-MONOPHOSPHATE" ? 'C10 H14 N5 O8 P' 363.223
C   'RNA linking'       y "CYTIDINE-5'-MONOPHOSPHATE"  ? 'C9 H14 N3 O8 P'  323.199
A   'RNA linking'       y "ADENOSINE-5'-MONOPHOSPHATE" ? 'C10 H14 N5 O7 P' 347.224
U   'RNA linking'       y "URIDINE-5'-MONOPHOSPHATE"   ? 'C9 H13 N2 O9 P'  324.183
DAR 'D-peptide linking' . D-ARGININE                   ? 'C6 H15 N4 O2 1'  175.210
ZUK 'D-peptide linking' . 5-PYRIMIDIN-2-YL-D-NORVALINE ? 'C9 H13 N3 O2'    195.221
NH2 NON-POLYMER         . 'AMINO GROUP'                ? 'H2 N'            16.022
#
```

Рисунок 3.1. Фрагмент документа 2KMJ из банка PDB в формате mmCIF.

По состоянию на апрель 2019 года база URSDB включает 4303 документа банка PDB, из которых 2485 документов содержат структуры РНК-белковых комплексов, 1380 документов содержат структуры РНК, 387 документов содержат структуры РНК в комплексе с ДНК и белками и 51 документ содержит структуры РНК в комплексе с ДНК. Стоит отметить, что

документ банка PDB может содержать несколько моделей одной и той же структуры. Это связано с тем, что в ходе экспериментального определения структуры не всегда удается однозначно восстановить координаты всех ее атомов. Описанные 4303 документа в формате mmCIF содержат в общей сложности 12014 моделей структур. Всего представлено 11686 цепей РНК (без учета представления одной цепи РНК в нескольких моделях).

Для разметки водородных связей, образующих вторичную структуру РНК, была использована программа DSSR [90]. Данная программа была выбрана среди более распространенных аналогов [40-42] как самая свежая и обладающая наиболее богатым функционалом. Более того, мы принимали активное участие в ее тестировании. Выходные данные программы DSSR содержат 7 файлов с аннотациями спариваний, элементов вторичной структуры и некоторых третичных мотивов. Основной выходной файл имеет формат .out (см. Рис. 3.2) и содержит подробное описание спариваний между основаниями нуклеотидов, а также других водородных связей и стекинг-взаимодействий.

```
List of 20 base pairs
      nt1              nt2             bp  name         Saenger    LW DSSR
   1 .A.GB.1.         .A.CC.18.        g-c WC           19-XIX    cWW cW-W
       [-175.5(anti) C2'-exo lambda=53.1] [-158.3(anti) C3'-endo lambda=49.9]
       d(C1'-C1')=10.87 d(N1-N9)=9.03 d(C6-C8)=9.78 tor(C1'-N1-N9-C1')=-8.7
       H-bonds[3]: "O6(carbonyl)-N4(amino)[2.69],N1(imino)-N3[2.94],N2(amino)-O2(carbonyl)[3.13]"
       bp-pars: [0.08    -0.07   -0.35   -21.13  -17.16  -6.56]
   2 .A.G.2.          .A.A.16.         G-A Imino        08-VIII   cWW cW-W
       [-168.8(anti) C3'-endo lambda=33.4] [-156.8(anti) C3'-endo lambda=31.3]
       d(C1'-C1')=13.31 d(N1-N9)=10.82 d(C6-C8)=10.44 tor(C1'-N1-N9-C1')=9.3
       H-bonds[1]: "O6(carbonyl)-N6(amino)[2.47]"
       bp-pars: [0.34    1.25    -2.48   -9.11   -27.21  -44.75]
   3 .A.G.2.          .A.C.17.         G-C WC           19-XIX    cWW cW-W
       [-168.8(anti) C3'-endo lambda=52.2] [-154.9(anti) C3'-endo lambda=47.4]
       d(C1'-C1')=10.96 d(N1-N9)=9.06 d(C6-C8)=9.79 tor(C1'-N1-N9-C1')=-25.3
       H-bonds[3]: "O6(carbonyl)-N4(amino)[2.80],N1(imino)-N3[2.92],N2(amino)-O2(carbonyl)[3.19]"
       bp-pars: [0.03    -0.12   0.09    -16.39  -27.61  -8.11]
   4 .A.U.3.          .A.A.16.         U-A WC           20-XX     cWW cW-W
       [-157.4(anti) C3'-endo lambda=49.8] [-156.8(anti) C3'-endo lambda=50.9]
       d(C1'-C1')=10.97 d(N1-N9)=9.09 d(C6-C8)=9.87 tor(C1'-N1-N9-C1')=-24.4
       H-bonds[2]: "N3(imino)-N1[2.94],O4(carbonyl)-N6(amino)[2.88]"
       bp-pars: [-0.11   -0.07   0.16    -3.57   -27.58  -8.11]
```

Рисунок 3.2. Фрагмент основного выходного файла программы DSSR в формате .out.

Для каждой структуры из банка PDB в качестве исходных данных использовались координаты атомов и описание цепей биополимеров из документа в формате mmCIF и разметка спариваний оснований из выходного файла программы DSSR в формате .out.

Для обработки исходных данных был разработан программный комплекс, реализованный на языке программирования Python 3 [103]. Программа выполнена в виде библиотеки, состоящей из 27 модулей. Такая реализация позволяет добавлять новые модули, изменять или удалять существующие независимо от остального кода. Общий объем кода составляет 9391 строку и занимает 324 килобайта.

Описанная библиотека содержит все необходимые функции для парсинга исходных данных, их обработки, построения базы данных и анализа, как полученных элементов вторичной структуры, так и взаимодействий РНК с другими молекулами. По выполняемым действиям программа делится на три части:

1) Подготовка исходных данных (разделение документов на модели, прогонка моделей через DSSR);

2) Конструктор текстовых файлов для наполнения базы данных (парсинг моделей и out-файлов, разметка элементов вторичной структуры, разметка РНК-белковых и других взаимодействий, разметка сигнатур псевдоузлов, разметка третичных мотивов).

3) Конструктор скрипта в формате SQL для создания базы данных на SQL сервере и ее наполнения.

Алгоритм разметки водородных связей между РНК и белком был взят из [28] с согласия авторов. Данный алгоритм использует численный инвариант, который отражает достоверность водородных связей – от 0.1 (недостоверная связь) до 1.0 (достоверная связь).

Начиная с марта 2015 года для работы с разработанным программным комплексом используется система контроля версий Git [104]. Также, с

помощью веб-ресурса GitLab (аналога GitHub), запущенного в нашей лаборатории, создан открытый репозиторий с исходным кодом, доступный по адресу http://git.lpm.org.ru/baulin/ursdb.

Для анализа экспериментально полученных структур РНК была разработана схема базы данных, основанная на новом способе описания петель. Данная разработка направлена на углубленное изучение структуры РНК и сбор статистики для последующего применения в рамках предсказания реальных последовательностей РНК, учитывая наличие псевдоузлов, а также для составления атласа третичных мотивов и их систематизации.

Предлагаемая схема состоит из 51 таблицы и содержит исчерпывающий набор данных, необходимых для дальнейших исследований, включая таблицы стемов, нитей, петель, граней, псевдоузлов и др.

Таблицы базы данных URSDB можно условно разделить на 4 блока данных:

1) Блок данных, полученных в результате парсинга mmCIF документов (таблицы моделей, молекул, цепей биополимеров, мономеров, атомов);

2) Блок данных, полученных в результате переработки данных о вторичной структуре РНК (таблицы спариваний, стемов, петель, крыльев, линков и пр.);

3) Блок данных, полученных непосредственно из выходного файла программы DSSR, предназначенный для верификации разметки (таблицы некоторых третичных мотивов, мультиплетов, и пр.)

4) Блок данных, полученных в процессе исследования структуры РНК (таблицы псевдоузлов, стемовых мотивов, РНК-белковых контактов и пр.)

Полный объем текстовых данных составляет 12,1 ГБ. По состоянию на апрель 2019 года база данных содержит описания 11686 цепей РНК, около 2,3 млн. спариваний и 7422 псевдоузла. Данные регулярно обновляются (не реже, чем раз в месяц).

3.2 Веб-интерфейс

3.2.1 Детали реализации

Для взаимодействия с базой данных URSDB в режиме онлайн разработан веб-интерфейс URS. Веб-интерфейс доступен по адресу http://urs.lpm.org.ru.

Веб-интерфейс реализован в виде набора CGI скриптов [105], выполненных на языке программирования Python 2 [103]. В процессе работы использовалась система контроля версий Git [104]. Также, в процессе работы были освоены языки разметки HTML и CSS [106, 107], а также скриптовый язык программирования JavaScript [108].

В рамках реализации веб-интерфейса использовались следующие технологии:

1) Local Storage – возможность HTML5 [106], позволяющая сохранять данные на стороне клиента. Была использована для сохранения выдачи результатов поиска в течение сессии пользователя.

2) jQuery [109] – сторонняя библиотека языка JavaScript. Была задействована для выполнения асинхронных запросов к серверу (Ajax) в процессе выдачи информации об отдельной структуре.

3) Ajax [110] – технология выполнения асинхронных запросов к серверу. Использовалась для выдачи информации об отдельной структуре и ее визуализации.

4) Jmol [111] – Java-апплет визуализации пространственной структуры. Был интегрирован в окно отдельной структуры для ее визуализации.

5) JSmol [112] – HTML5-версия апплета Jmol. Был интегрирован в окно отдельной структуры для ее визуализации.

6) ExtJS [113] – сторонняя библиотека CSS. Использовалась для оформления окна отдельной структуры.

Общий объем кода HTML, CSS и JavaScript составляет 4455 строк (без учета сторонних библиотек). Общий объем CGI скриптов составляет 5267 строк.

3.2.2 Возможности использования

Разработанный интерфейс позволяет пользователю формировать выборку интересующих его РНК-содержащих структур, после чего собирать статистику по имеющимся в них структурным элементам. Кроме того пользователю доступны индивидуальные структуры и элементы с подробными данными для детального анализа.

Формирование выборки структур происходит согласно запросу к базе данных; запрос состоит из конъюнкции нескольких дизъюнкций элементарных условий. Элементарные условия, доступные пользователю на данный момент, состоят из 4 групп: (1) общая информация о PDB документе, (2) информация о содержащихся макромолекулах, (3) шаблоны структурных элементов, а также (4) информация о содержащихся спариваниях между основаниями в РНК и между атомами РНК и белка. Результатом запроса является список структур, удовлетворяющих введенным параметрам (см. Рис. 3.3).

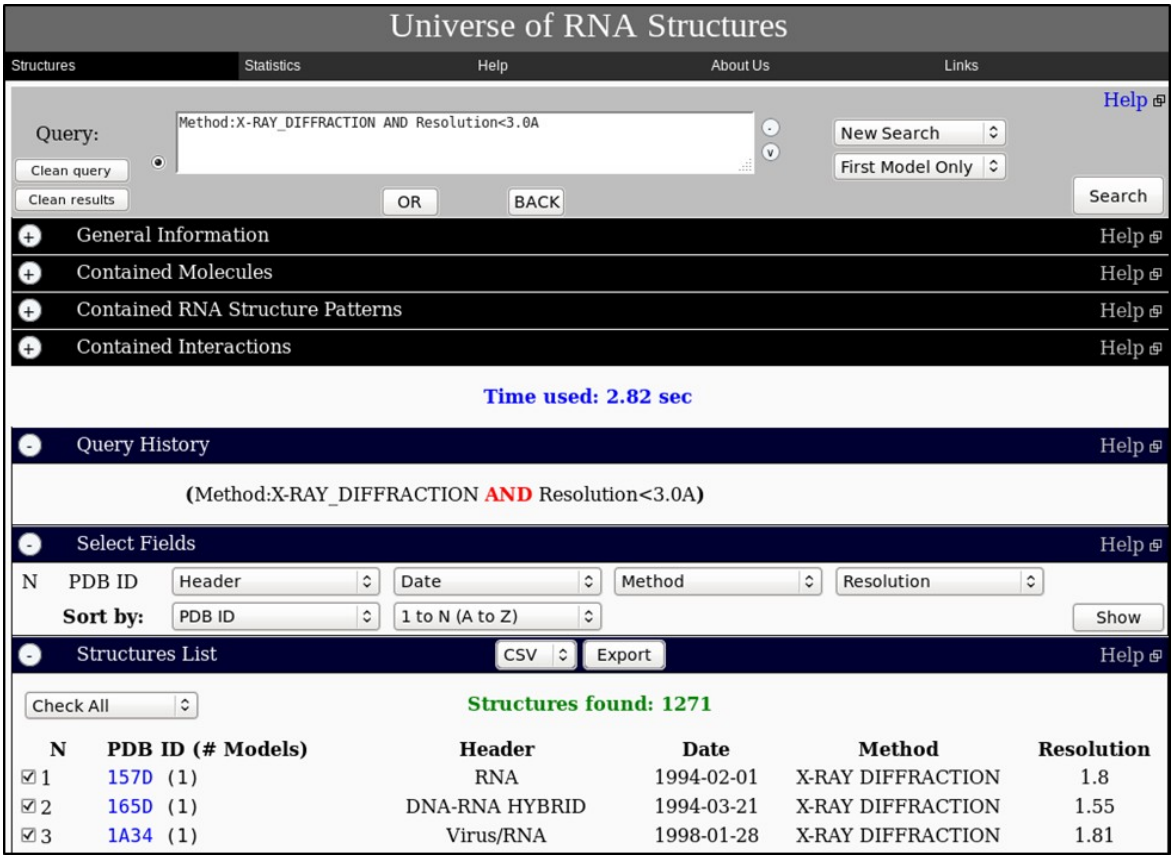


Рисунок 3.3. Страница поиска структур.

Для индивидуального анализа интересующей пользователя структуры имеется окно структуры, доступное по клику на ID структуры в списке результатов поиска структур. Данное окно содержит детальную информацию об имеющихся в структуре цепях, спариваниях, петлях, спиралях и псевдоузлах. В будущем этот список будет пополняться. Также в окно встроен 3D-визуализатор молекул JSmol [112] (см. Рис. 3.4).

Рисунок 3.4. Окно отдельной структуры

После формирования выборки документов пользователь имеет возможность собирать статистику структурных элементов, содержащихся в выбранных структурах. На данный момент в качестве таких элементов доступны цепи, спаривания, петли, спирали, псевдоузлы, мультиплеты, а также РНК-белковые взаимодействия. По каждому виду элемента доступны фильтры, позволяющие ограничивать элементы необходимыми параметрами, а также полные списки найденных элементов с возможностью вывода детальной информации об отдельном элементе по клику (см. Рис. 3.5, 3.6).

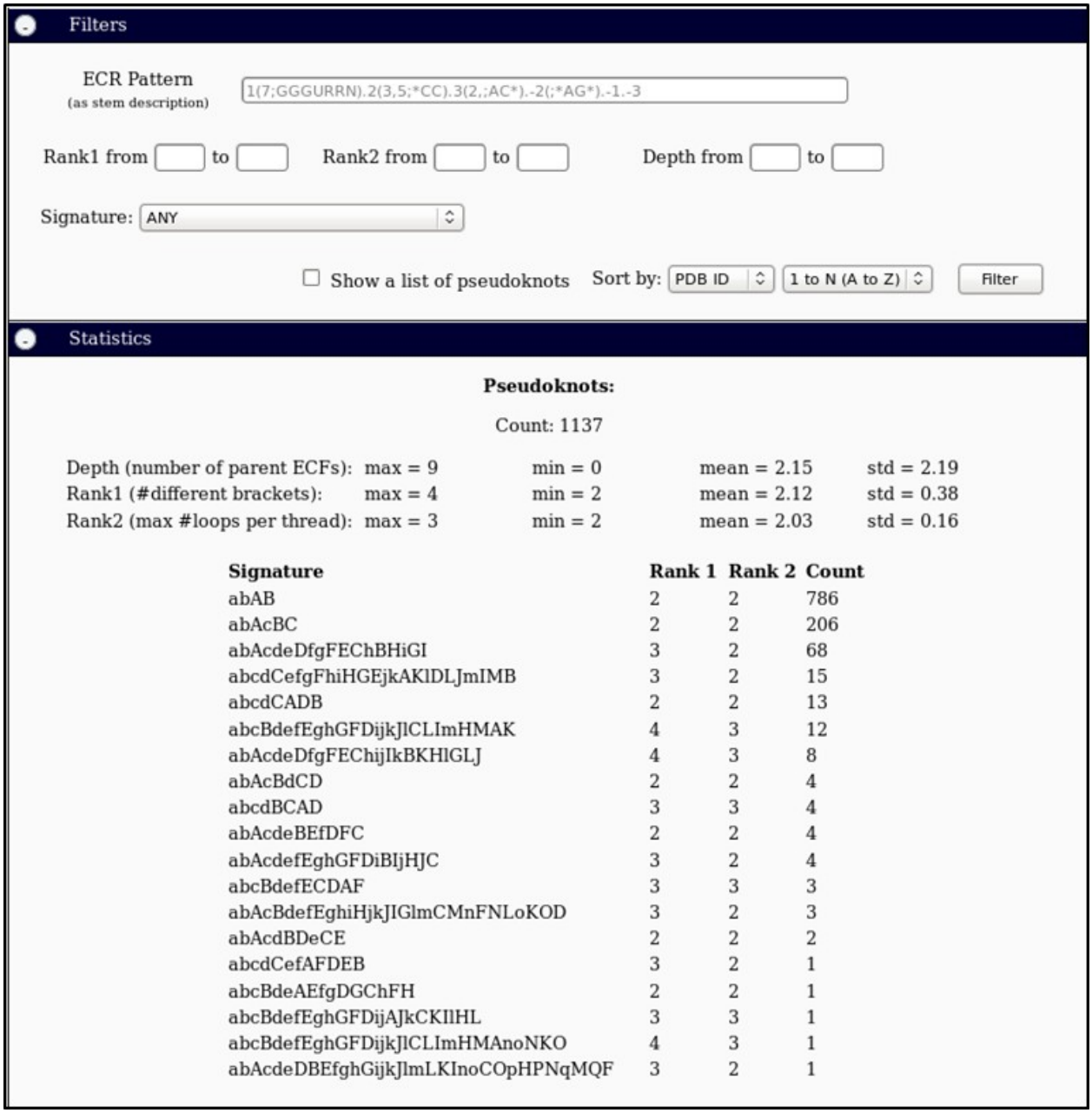


Рисунок 3.5. Результат запроса статистики псевдоузлов

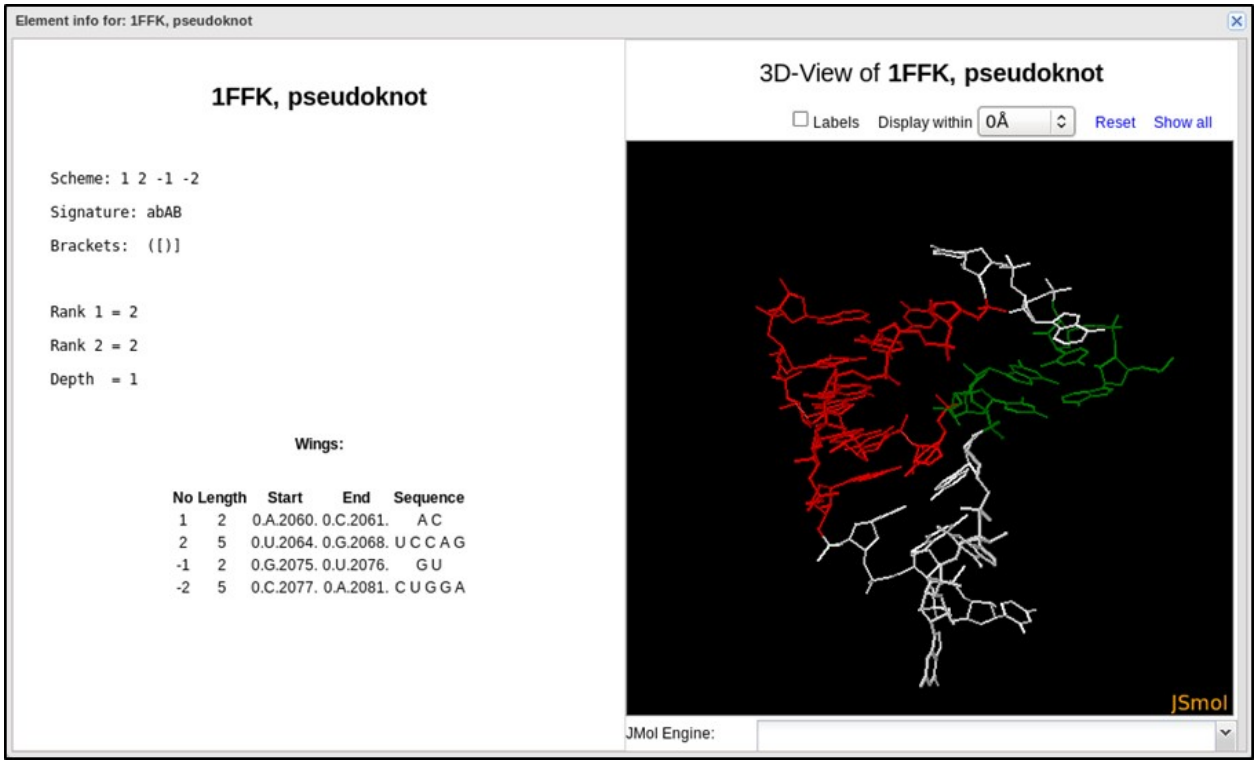


Рисунок 3.6. Окно отдельного псевдоузла

3.3 Выводы и результаты по главе

Была спроектирована и реализована база данных экспериментально определенных пространственных структур РНК. Все цепи РНК и их структурные элементы проаннотированы. В частности, размечены все элементы вторичной структуры, в том числе, связанные с псевдоузлами, и нестандартные мотивы, образованные с участием неканонических водородных связей. Подобная база разработана впервые. Она существенно расширит возможности исследования пространственных структур РНК, в том числе, – их неканонических элементов.

Для взаимодействия конечных пользователей с базой разработан веб-интерфейс, обеспечивающий поиск нужных структур в базе, сбор статистик по этим структурам, визуализацию и хранение результатов поиска. Веб-интерфейс позволяет работать не только с целыми структурами РНК, но и с различными типами их структурных элементов, таких как, спаривания, петли, псевдоузлы, РНК-белковые контакты и др. Также, веб-интерфейс содержит подробную страницу помощи.

ГЛАВА 4. Анализ структурных мотивов РНК

4.1 Короткие стемы в псевдоузловых структурах РНК

Описанные в данном параграфе результаты опубликованы в работе [114].

В процессе работы было замечено, что псевдоузлы в экспериментально разрешенных структурах РНК более чем наполовину состоят из стемов, образованных двумя или тремя парами оснований. Согласно эмпирическим термодинамическим параметрам образование стема вносит отрицательный вклад в свободную энергию структуры только при длине в четыре спаривания и более [115]. Таким образом, короткие стемы могут быть энергетически нестабильны и не должны вносить существенного вклада в свободную энергию структуры.

Среди всех РНК-содержащих документов PDB были отобраны структуры из неизбыточного подмножества структур РНК ([116], версия 2.156_all от 24 ноября 2017 г.). В подмножество вошло 2300 структур, образованных 3974 молекулами РНК из 1987 документов PDB. Т.к. база данных URSDB учитывает межцепочечные стемы, все сигнатуры псевдоузлов были пересчитаны после исключения таких стемов. Далее сигнатуры были пересчитаны в трех различных вариантах: после удаления стемов длины 2; после удаления стемов длины 3 и после удаления стемов длины 2 и 3. Все процедуры были также выполнены на всем множестве структур РНК из банка PDB.

В рассматриваемое подмножество структур вошло 12793 стема различной длины. Среди всех стемов, входящих в псевдоузлы, коротких стемов оказалось более половины – 1171 из 2303. В процессе работы была проверена гипотеза о том, что короткие стемы предпочитают находиться рядом с внутренними петлями и избегать шпилек. Данная гипотеза была вызвана тем, что внутренние петли в реальных структурах РНК представляют собой набор неканонических спариваний нуклеотидов, а не

последовательность неспаренных нуклеотидов, как это следует из графического представления элементов вторичной структуры РНК. Таким образом, в пространственных структурах последовательные элементы "стем – внутренняя петля – стем" образуют один "неканонический стем", что способствует дополнительной стабилизации структуры. Однако в случаях, когда короткий стем замыкает шпильку, какие-либо дополнительные стабилизирующие факторы отсутствуют. Гипотеза была подтверждена – точный тест Фишера показал p-value < $2.2 \cdot 10^{-16}$ для шпилек и p-value = $6.29 \cdot 10^{-9}$ для внутренних петель.

Таблица 4.1. Список исходных сигнатур псевдоузлов.

Сигнатура	**Кол-во**
abAB	289
abAcBC	67
abcdBCAD	2
abcdCADB	2
abAcdBDeCE	3
abcdCABeDE	1
abAcdCeBEfDF	1
abAcdeBEfDFC	9
abcdCeAEfDFB	1
abcdCeBEAfDF	1
abAcdCefDFgBGE	1
abAcdeDfgFEChBHiGI	17
abAcdeDfghiHFECjkGKlBLJmIM	2
Всего	**396**

Таблица 4.2. Список сигнатур псевдоузлов, пересчитанных без учета коротких стемов

Сигнатура	**Кол-во**
abAB	74
abAcBC	43
Всего	**117**

Всего в рассматриваемом подмножестве структур РНК было обнаружено 396 псевдоузлов, принадлежащих 13 различным типам, согласно топологической классификации псевдоузлов (см. таблицу 4.1). Наиболее распространенными типами являются H-узлы (сигнатура abAB, 289 псевдоузлов) и kissing loops (сигнатура abAcBC, 67 псевдоузлов). После исключения коротких стемов и пересчета сигнатур было выявлено 117 псевдоузлов (см. таблицу 4.2), принадлежащих к двум простейшим типам (abAB и abAcBC).

Подсчет сигнатур был также произведен на всем множестве структур РНК из банка PDB. Было обнаружено 6986 псевдоузлов. После исключения коротких стемов было выявлено 1276 псевдоузлов, 1270 из которых принадлежат двум простейшим типам, а оставшиеся 6 псевдоузлов имеют сигнатуру abcdCefAFDEB.

Все 6 псевдоузлов с сигнатурой abcdCefAFDEB были обнаружены в различных экземплярах молекулы интрона группы II бактерии *Oceanobacillus iheyensis*. Всего в PDB представлено 26 структур данной молекулы, которые образуют класс эквивалентности NR_all_35054.3. Анализ данных структур показал, что псевдоузел с сигнатурой abcdCefAFDEB образуется только в присутствии фрагмента экзона под названием Intron Binding Site 1 (IBS1, последовательность AUAA). Среди 26 структур данного класса 6 структур (идентификаторы 4ds6, 4faq, 4fau, 4y1n и два экземпляра 4y1o) содержат лигированный IBS1 и образуют указанный псевдоузел. Еще 3 структуры (4far, 3eoh, 3eog) содержат IBS1 в виде изолированной молекулы и

также образуют указанный псевдоузел (однако данные экземпляры содержат межцепочечные стемы, поэтому при подсчетах не учитывались). Таким образом, следует полагать, что данный псевдоузел существует только в момент сплайсинга данного интрона. Представителем данного класса эквивалентности в неизбыточном подмножестве структур РНК является структура с идентификатором 5j01, в которой IBS1 отсутствует, что объясняет отсутствие псевдоузла с сигнатурой abcdCefAFDEB в таблицах 4.1 и 4.2.

Интроны группы II представлены в PDB тремя организмами: *Oceanobacillus iheyensis* (26 структур, класс NR_all_35054.3), *Lactococcus lactis* (2 структуры – 5g2x, 5g2y, класс NR_all_28269.1) и *Pylaiella littoralis* (1 структура – 4r0d, класс NR_all_05993.1). Структуры 5g2x и 4r0d содержат IBS1 и также образуют сложные псевдоузлы, однако при исключении коротких стемов их сигнатуры сводятся к сигнатуре abAcBC. Таким образом, во всем PDB существует всего 1 уникальный псевдоузел, который не сводится к простейшему типу при исключении коротких стемов.

4.2 Предсказание сайтов связывания ионов Mg2+ с РНК

Описанные в данном параграфе результаты опубликованы в работе [117].

Используя оригинальную модель описания вторичной структуры РНК был разработан метод предсказания сайтов связывания ионов магния с РНК, основанный на алгоритме машинного обучения “случайный лес” [118]. Проведен сравнительный анализ работы представленной модели и сервисов FEATURE [119] и MetalIonRNA [120]. Насколько нам известно, классические методы машинного обучения к решению данной задачи не применялись.

Сформулируем задачу классификации фрагментов РНК на два типа – связанные с ионом магния (класс 1) и несвязанные (класс 0). Фрагмент принадлежит к классу 1, если в радиусе X A° находится хотя бы один ион магния. По результатам предварительных экспериментов в качестве

фрагментов РНК были выбраны O/N атомы РНК, значение радиуса $X = 7$. Составленная выборка данных содержала около 330 структур РНК. Доля элементов класса 1 составляла около 10%. Каждый элемент выборки содержал 383 значения признаков.

Параметры алгоритма были определены в ходе тестовых запусков. Итоговая модель включала следующие значения основных параметров: *Max_depth = 26, Min_samples_leaf = 20, Max_features = 0.7*.

Результаты работы описанной модели были сопоставлены с результатами сервисов FEATURE и MetalIonRNA. Поскольку онлайн-версия сервиса FEATURE в настоящее время недоступна, сравнения проводились на 12 структурах, описанных в работе [119]. Для 8 структур из 12 результаты нашей модели незначительно превосходили результаты существующих сервисов, однако разница составляла 5-10% и находилась в рамках погрешности. На рис 4.1. представлены результаты предсказаний трех алгоритмов на примере структуры 1hc8. Визуализация результата FEATURE взята из работы [119].

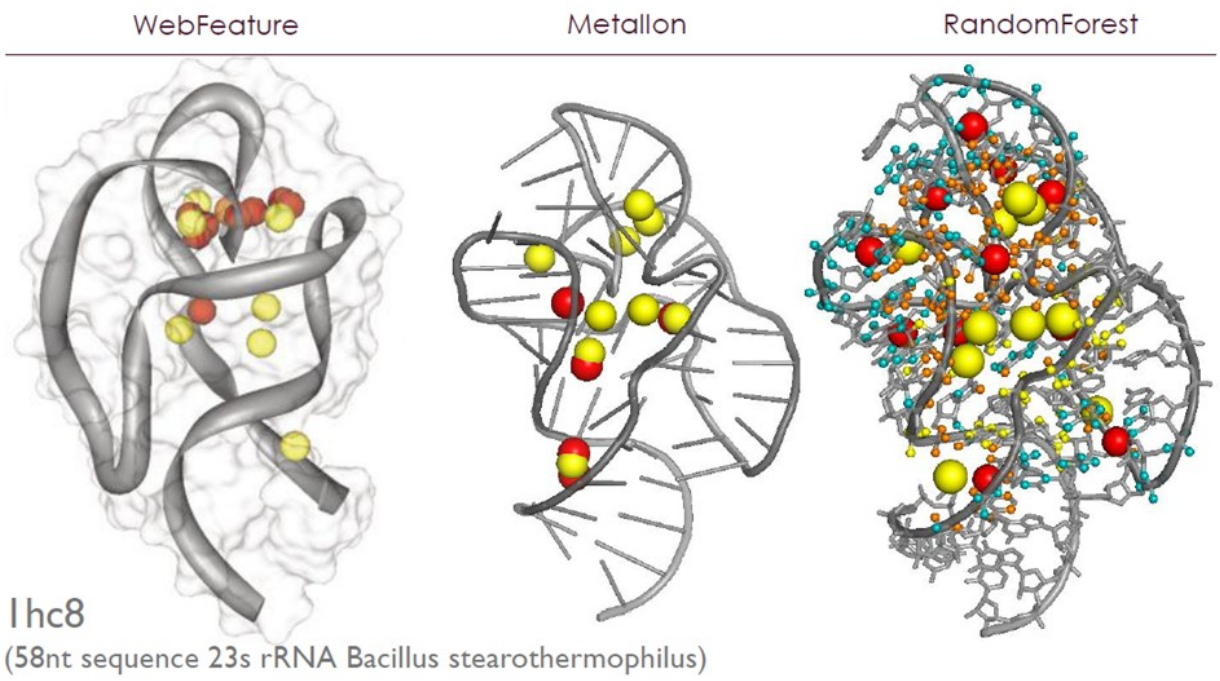


Рисунок 4.1. Результаты работы алгоритмов WebFeature, MetalIon и RandomForest. Реальные ионы магния представлены большими желтыми шарами, предсказанные – большими красными. Маленькими шариками выделены результаты классификации атомов РНК алгоритмом RandomForest – верно отнесенные к классу 1 (рыжий цвет), ошибочно отнесенные к классу 1 (голубой цвет), ошибочно отнесенные к классу 0 (желтый цвет).

Стоит отметить, что несмотря на достаточное качество полученных результатов для выбранных 12 структур, средние результаты работы описанной модели на всей выборке не позволяют сделать вывод об успешном решении поставленной задачи. Так, значение F-меры для всей выборки не превышает значения 0.3. Также, сервис MetalIonRNA для произвольных структур РНК показывал точность не более 60% по количеству предсказанных ионов магния от числа ионов магния, содержащихся в структуре. Таким образом, на данный момент не существует сервиса, способного с приемлемой точностью предсказывать сайты связывания ионов магния для произвольных пространственных структур РНК.

4.3 Классификация третичных мотивов РНК

На основе предложенной модели описания вторичной структуры РНК была разработана классификация третичных мотивов РНК. Согласно классификации каждому нуклеотиду мотива приписывается тип соответствующего элемента вторичной структуры. Так, нуклеотиду соответствует метка *S*, если нуклеотид принадлежит стему, и метка $T_1C_1T_2C_2...T_nC_n$, если нуклеотид принадлежит *N* петлям, где T_i - тип петли (*H* - шпилька, *B* - выпячивание, *I* - внутренняя петля, *J* - мульти-петля), C_i - класс петли (*C* - классическая петля, *I* - изолированная петля, *P* - псевдоузловая петля). Каждой паре нуклеотидов мотива ставится в соответствие взаимное расположение их элементов вторичной структуры. Так, паре нуклеотидов соответствует метка *SM* (same), если нуклеотиды принадлежат одному элементу вторичной структуры, метка *LC* (local), если нуклеотиды принадлежат соседним элементам, и метка *LR* (long-range), если нуклеотиды принадлежат взаимно удаленным элементам.

Описанная классификация была применена к двум наиболее распространенным типам третичных мотивов РНК - триплексам оснований и мотивам А-минор. Триплексы оснований представляют собой тройки оснований, расположенных в одной плоскости и образующих спаривания. А-минор - мотив, образованный как правило каноническим спариванием оснований и аденином, образующим водородные связи с нуклеотидами спаривания со стороны малой бороздки. Строгие определения данных типов третичных мотивов РНК см. в работе [90]. На Рис. 4.2 показан пример триплекса класса IC-IC-S SM-LR-LR. Два нуклеотида триплекса принадлежат одной классической внутренней петле, а третий нуклеотид принадлежит стему, удаленному относительно данной петли.

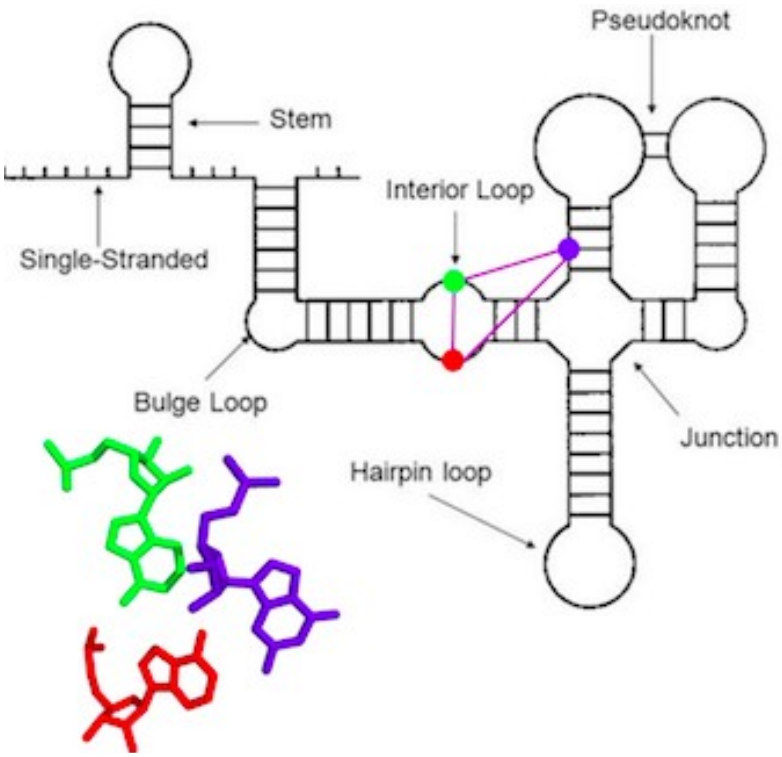


Рисунок 4.2. Пример триплекса класса IC-IC-S SM-LR-LR

Применение разработанной классификации к описанию триплексов и A-миноров, содержащихся в экспериментально определенных структурах РНК из банка PDB, позволило выявить функциональные подтипы данных мотивов. Так, анализ структур показал, что A-миноры класса IC-IC-IC SM-SM-SM являются частью более сложного мотива Kink-turn и стабилизируют его структуру (резкий изгиб сахаро-фосфатного остова); A-миноры класса JC-S-S LC-LC-SM стабилизируют коаксиальный стекинг участков двойной спирали. Кроме того, среди 10000 триплексов было выявлено около 100 триплексов класса LR-LR-LR. Такие триплексы образованы тремя попарно удаленными по вторичной структуре нуклеотидами и представляют особый интерес, т.к. образование подобной структуры энергетически невыгодно, а значит, носит важный функциональный характер.

В процессе работы было замечено, что более половины всех A-миноров в реальных структурах РНК существуют не поодиночке, а образуют кластеры. Помимо описанного в работе [85] вида кластера A-patch, в котором аденины образуют стекинг, были найдены кластеры, в которых стекинг

образуют только канонические спаривания A-миноров. Мы назвали такие кластеры PAM (Pile of A-minors).

Была сформулирована задача предсказания триплексов оснований по данным о последовательности и вторичной структуре РНК. Классификация третичных мотивов была использована для описания триплексов при составлении выборки данных. Для решения задачи использовалась методика, описанная в нашей работе [117], основанная на алгоритме машинного обучения “случайный лес”.

Решалась задача классификации троек нуклеотидов на два класса - образующие триплекс (класс 1) и не образующие триплекс (класс 0). В выборку вошли триплексы и случайно отобранные тройки нуклеотидов, не образующие триплексы, из более чем 300 структур РНК неизбыточного подмножества РНК-содержащих документов банка PDB. Соотношение классов было выбрано равным.

Результаты кросс-валидации показали значения метрик precision и recall на уровне 92% и 95% соответственно (см. Рис. 4.3). Признаки, соответствующие классам триплексов согласно разработанной классификации, дали более 50% значимости согласно анализу важности признаков (feature importance), что демонстрирует практическую ценность данной классификации.

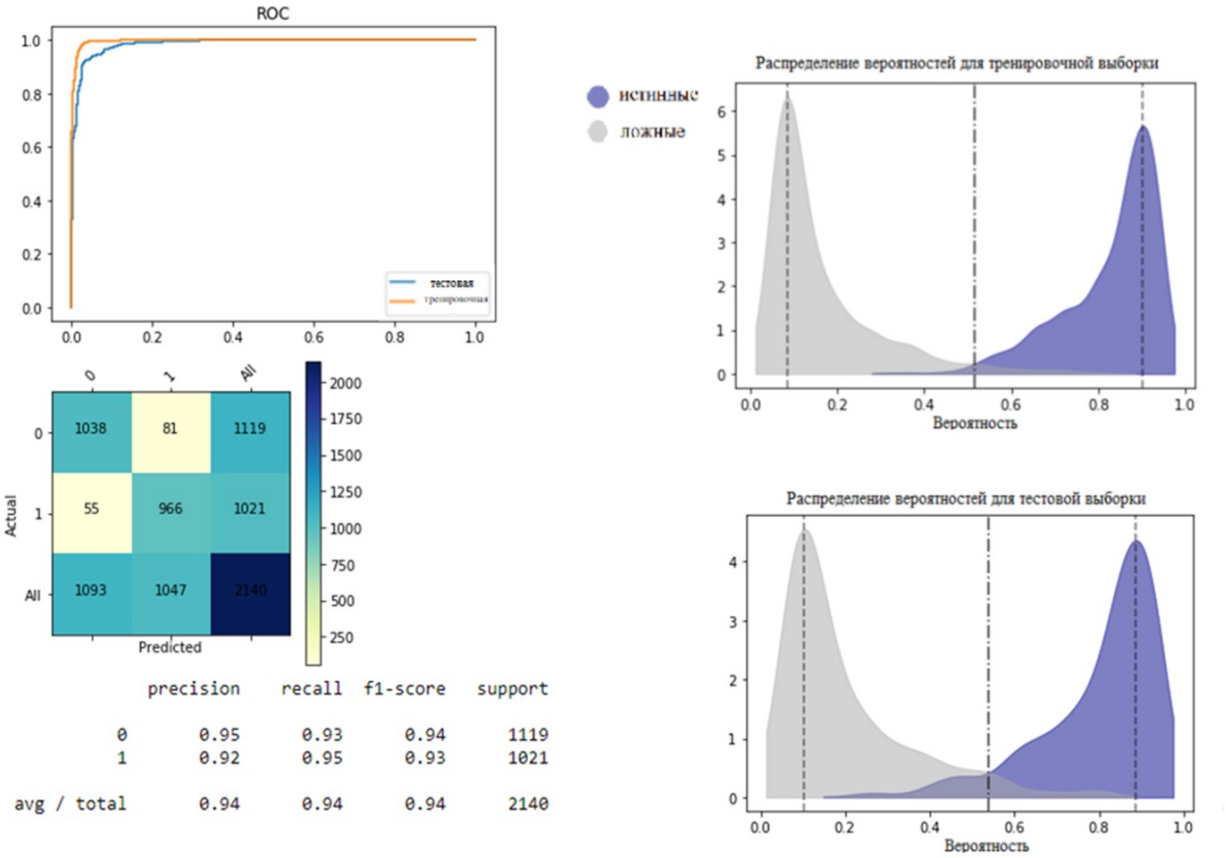
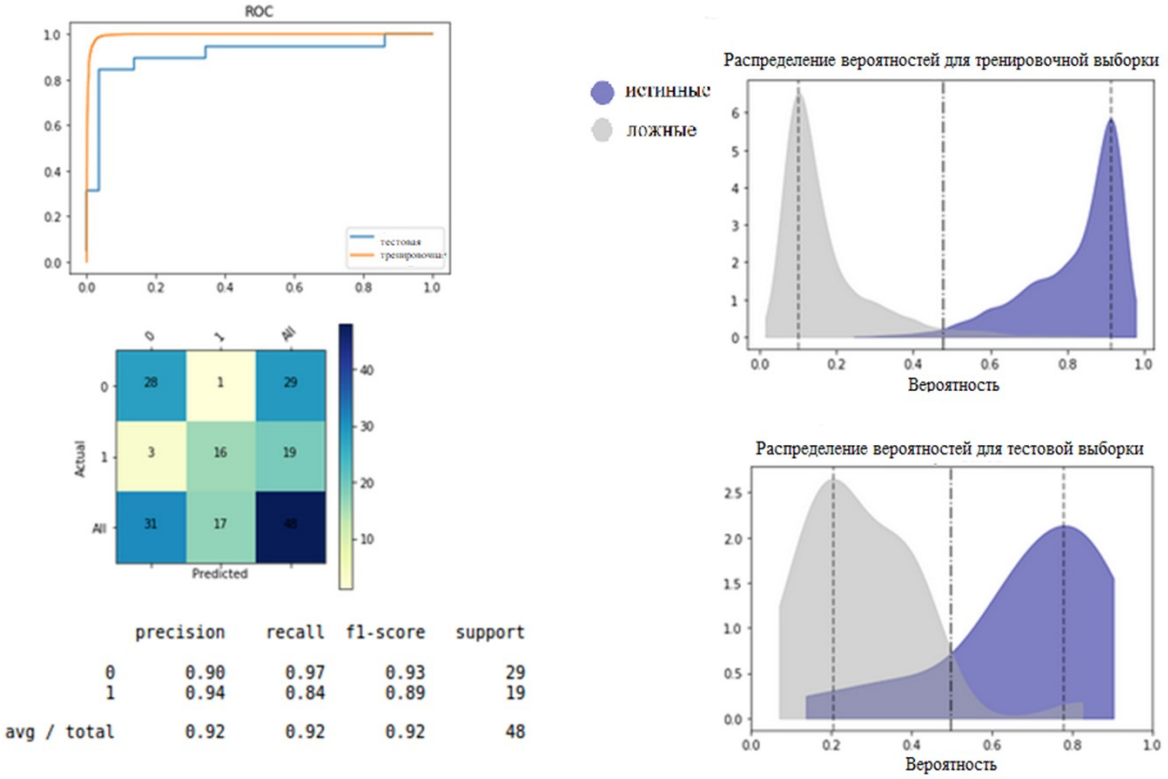


Рисунок 4.3 Результаты кросс-валидации на сбалансированной выборке

Для исследования различий свойств триплексов в разных типах молекул РНК и возможности предсказания триплексов в новых типах молекул был проведен эксперимент, в котором в качестве тестовой выборки использовались 3 структуры SAM-рибосвитчей. При этом в обучающую выборку структуры SAM-рибосвитчей не входили. Результаты показали значения метрик precision и recall на уровне 94% и 84% соответственно (см. Рис. 4.4).



	precision	recall	f1-score	support
0	0.90	0.97	0.93	29
1	0.94	0.84	0.89	19
avg / total	0.92	0.92	0.92	48

Рисунок 4.4 Результаты тестирования на структурах SAM-рибосвитчей модели, обученной на других типах молекул РНК

4.4 Выводы и результаты по главе

Был проведен анализ роли коротких стемов в образовании псевдоузловых структур РНК. Показано, что при исключении коротких стемов из рассмотрения, псевдоузлы в экспериментально разрешенных пространственных структурах РНК сводятся к двум простейшим типам. Кроме того, на основе экспериментальных данных были получены косвенные подтверждения энергетической нестабильности коротких стемов.

Было показано, что в настоящий момент не существует сервиса, способного с приемлемой точностью предсказывать сайты связывания ионов магния с РНК. Данная проблема связана с низким качеством данных о координатах ионов магния в документах банка PDB.

Была разработана классификация третичных мотивов РНК. Классификация была апробирована на примере триплексов и мотивов A-минор, ее применение позволило выявить функциональные классы данных мотивов. Также, предложенная классификация была успешно использована для решения задачи предсказания триплексов оснований РНК.

ЗАКЛЮЧЕНИЕ

В процессе работы впервые была предложена модель описания произвольной вторичной структуры РНК, обобщающая модель NNM на случай псевдоузловых структур. На основе новой модели была спроектирована и реализована универсальная база данных пространственных структур и структурных мотивов РНК URSDB. Также, предложенная модель была использована при разработке единой классификации третичных мотивов РНК. Разработанные методики открывают новые возможности для проведения детального анализа пространственных структур РНК.

Проведен анализ роли коротких стемов в образовании псевдоузловых структур РНК. В ходе работы были получены косвенные подтверждения энергетической нестабильности коротких стемов. Анализ сигнатур псевдоузлов показал, что все псевдоузлы из неизбыточного подмножества структур РНК сводятся к двум простейшим типам (abAB и abAcBC) при исключении коротких стемов из рассмотрения и последующего пересчета сигнатур. Полученные результаты могут быть использованы при разработке нового алгоритма предсказания вторичной структуры РНК, допускающего только псевдоузлы указанных типов. Такой алгоритм может, например, состоять из двух итераций: 1) предсказание вторичной структуры, не допускающей коротких стемов в составе псевдоузлов; 2) последующее наложение на полученную структуру коротких стемов, вносящих отрицательный вклад в свободную энергию структуры.

Была сформулирована и решена для случая сбалансированной выборки задача предсказания триплексов оснований по данным о последовательности нуклеотидов и вторичной структуре РНК. Результаты таких предсказаний могут быть использованы в качестве дополнительных ограничений при моделировании пространственной структуры РНК *de novo*. Однако, для успешного решения задачи предсказания триплексов необходимо решить проблему дисбаланса классов в случае полной выборки (число троек

нуклеотидов пропорционально кубу длины последовательности, число триплексов пропорционально длине последовательности).

В дальнейшем представляются перспективными два направления работ: (1) использование полученных результатов для разработки новых алгоритмов решения вычислительных задач биоинформатики РНК и (2) применение разработанных методик для анализа новых типов третичных мотивов РНК.

СПИСОК ЛИТЕРАТУРЫ

1. Gorodkin J., Ruzzo W. L. (ed.). RNA sequence, structure, and function: computational and bioinformatic methods. – New York, NY : Humana Press, 2014.

2. Brosnan C. A., Voinnet O. The long and the short of noncoding RNAs //Current opinion in cell biology. – 2009. – Т. 21. – №. 3. – С. 416-425.

3. Waters L. S., Storz G. Regulatory RNAs in bacteria //Cell. – 2009. – Т. 136. – №. 4. – С. 615-628.

4. Lafontaine D. L. J. Noncoding RNAs in eukaryotic ribosome biogenesis and function //Nature structural & molecular biology. – 2015. – Т. 22. – №. 1. – С. 11.

5. Christov C. P. et al. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication //Molecular and cellular biology. – 2006. – Т. 26. – №. 18. – С. 6993-7004.

6. Morris K. V., Mattick J. S. The rise of regulatory RNA //Nature Reviews Genetics. – 2014. – Т. 15. – №. 6. – С. 423.

7. Cech T. R., Steitz J. A. The noncoding RNA revolution—trashing old rules to forge new ones //Cell. – 2014. – Т. 157. – №. 1. – С. 77-94.

8. Ponting C. P., Oliver P. L., Reik W. Evolution and functions of long noncoding RNAs //Cell. – 2009. – Т. 136. – №. 4. – С. 629-641.

9. Mercer T. R., Mattick J. S. Structure and function of long noncoding RNAs in epigenetic regulation //Nature structural & molecular biology. – 2013. – Т. 20. – №. 3. – С. 300.

10. Ling H., Fabbri M., Calin G. A. MicroRNAs and other non-coding RNAs as targets for anticancer drug development //Nature reviews Drug discovery. – 2013. – Т. 12. – №. 11. – С. 847.

11. Saenger W. Principles of nucleic acid structure. – Springer Science & Business Media, 2013.

12. Xia T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson− Crick base pairs //Biochemistry. – 1998. – Т. 37. – №. 42. – С. 14719-14735.

13. Zuker M., Mathews D. H., Turner D. H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide //RNA biochemistry and biotechnology. – Springer, Dordrecht, 1999. – С. 11-43.

14. Mathews D. H., Turner D. H. Prediction of RNA secondary structure by free energy minimization //Current opinion in structural biology. – 2006. – Т. 16. – №. 3. – С. 270-278.

15. Pleij C. W. A. RNA pseudoknots //Current Opinion in Structural Biology. – 1994. – Т. 4. – №. 3. – С. 337-344.

16. Gultyaev A. P., Olsthoorn R. C. L. A family of non-classical pseudoknots in influenza A and B viruses //RNA biology. – 2010. – Т. 7. – №. 2. – С. 125-129.

17. Condon A. et al. Classifying RNA pseudoknotted structures //Theoretical Computer Science. – 2004. – Т. 320. – №. 1. – С. 35-50.

18. Van Batenburg F. H. D. et al. PseudoBase: a database with RNA pseudoknots //Nucleic Acids Research. – 2000. – Т. 28. – №. 1. – С. 201-204.

19. Taufer M. et al. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots //Nucleic acids research. – 2008. – Т. 37. – №. suppl_1. – С. D127-D135.

20. Butcher S. E., Pyle A. M. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks //Accounts of chemical research. – 2011. – Т. 44. – №. 12. – С. 1302-1311.

21. Schwalbe H. et al. Structures of RNA switches: insight into molecular recognition and tertiary structure //Angewandte Chemie International Edition. – 2007. – Т. 46. – №. 8. – С. 1212-1219.

22. Parlea L. G. et al. The RNA 3D Motif Atlas: Computational methods for extraction, organization and evaluation of RNA motifs //Methods. – 2016. – Т. 103. – С. 99-119.

23. Matsumura S., Ikawa Y., Inoue T. Biochemical characterization of the kink-turn RNA motif //Nucleic acids research. – 2003. – Т. 31. – №. 19. – С. 5544-5551.

24. Tyagi R., Mathews D. H. Predicting helical coaxial stacking in RNA multibranch loops //Rna. – 2007. – Т. 13. – №. 7. – С. 939-951.

25. Vanegas P. L. et al. RNA CoSSMos: characterization of secondary structure motifs—a searchable database of secondary structure motifs in RNA three-dimensional structures //Nucleic acids research. – 2011. – Т. 40. – №. D1. – С. D439-D444.

26. Abraham M. et al. Analysis and classification of RNA tertiary structures //RNA. – 2008. – Т. 14. – №. 11. – С. 2274-2289.

27. Schnabl J., Suter P., Sigel R. K. O. MINAS—a database of Metal Ions in Nucleic AcidS //Nucleic acids research. – 2011. – Т. 40. – №. D1. – С. D434-D438.

28. Kirsanov D. D. et al. NPIDB: nucleic acid—protein interaction database //Nucleic acids research. – 2012. – Т. 41. – №. D1. – С. D517-D523.

29. Lewis B. A. et al. PRIDB: a protein–RNA interface database //Nucleic acids research. – 2010. – Т. 39. – №. suppl_1. – С. D277-D282.

30. Schudoma C., May P., Walther D. Modeling RNA loops using sequence homology and geometric constraints //Bioinformatics. – 2010. – Т. 26. – №. 13. – С. 1671-1672.

31. Petrov A. I., Zirbel C. L., Leontis N. B. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas //Rna. – 2013. – Т. 19. – №. 10. – С. 1327-1340.

32. Chojnowski G., Waleń T., Bujnicki J. M. RNA Bricks—a database of RNA 3D motifs and their interactions //Nucleic acids research. – 2014. – Т. 42. – №. D1. – С. D123-D131.

33. Popenda M. et al. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures //Bmc Bioinformatics. – 2010. – Т. 11. – №. 1. – С. 231.

34. Bindewald E. et al. RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign //Nucleic acids research. – 2007. – Т. 36. – №. suppl_1. – С. D392-D397.

35. Andronescu M. et al. RNA STRAND: the RNA secondary structure and statistical analysis database //BMC bioinformatics. – 2008. – Т. 9. – №. 1. – С. 340.

36. Tamura M. et al. SCOR: Structural Classification of RNA, version 2.0 // Nucleic acids research. – 2004. – Т. 32. – №. suppl_1. – С. D182-D184.

37. Coimbatore Narayanan B. et al. The Nucleic Acid Database: new features and capabilities //Nucleic acids research. – 2013. – Т. 42. – №. D1. – С. D114-D122.

38. Kalvari I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families //Nucleic acids research. – 2017. – Т. 46. – №. D1. – С. D335-D342.

39. Burley S. K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive //Protein Crystallography. – Humana Press, New York, NY, 2017. – С. 627-641.

40. Sarver M. et al. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures //Journal of mathematical biology. – 2008. – Т. 56. – №. 1-2. – С. 215-252.

41. Yang H. et al. Tools for the automatic identification and classification of RNA base pairs //Nucleic acids research. – 2003. – Т. 31. – №. 13. – С. 3450-3460.

42. Gendron P., Lemieux S., Major F. Quantitative analysis of nucleic acid three-dimensional structures //Journal of molecular biology. – 2001. – Т. 308. – №. 5. – С. 919-936.

43. Kirillova S., Tosatto S. C. E., Carugo O. FRASS: the web-server for RNA structural comparison //BMC bioinformatics. – 2010. – Т. 11. – №. 1. – С. 327.

44. Wang C. W., Chen K. T., Lu C. L. iPARTS: an improved tool of pairwise alignment of RNA tertiary structures //Nucleic acids research. – 2010. – Т. 38. – №. suppl_2. – С. W340-W347.

45. Muppirala U. K., Honavar V. G., Dobbs D. Predicting RNA-protein interactions using only sequence information //BMC bioinformatics. – 2011. – Т. 12. – №. 1. – С. 489.

46. Nebel M. E., Weinberg F. Algebraic and combinatorial properties of common RNA pseudoknot classes with applications //Journal of Computational Biology. – 2012. – Т. 19. – №. 10. – С. 1134-1150.

47. Rother M. et al. ModeRNA: a tool for comparative modeling of RNA 3D structure //Nucleic acids research. – 2011. – Т. 39. – №. 10. – С. 4007-4022.

48. Zhao Y. et al. Automated and fast building of three-dimensional RNA structures //Scientific reports. – 2012. – Т. 2. – С. 734.

49. Puton T. et al. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction //Nucleic acids research. – 2013. – Т. 41. – №. 7. – С. 4307-4323.

50. Allali J. et al. BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms //Advances in bioinformatics. – 2012. – Т. 2012.

51. Rivas E., Eddy S. R. A dynamic programming algorithm for RNA structure prediction including pseudoknots //Journal of molecular biology. – 1999. – Т. 285. – №. 5. – С. 2053-2068.

52. Saule C. et al. Counting RNA pseudoknotted structures //Journal of Computational Biology. – 2011. – Т. 18. – №. 10. – С. 1339-1351.

53. Nussinov R., Jacobson A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA //Proceedings of the National Academy of Sciences. – 1980. – Т. 77. – №. 11. – С. 6309-6313.

54. McCaskill J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure //Biopolymers: Original Research on Biomolecules. – 1990. – Т. 29. – №. 67. – С. 1105-1119.

55. Finkelstein A. V., Roytberg M. A. Computation of biopolymers: a general approach to different problems //BioSystems. – 1993. – Т. 30. – №. 1-3. – С. 1-19.

56. Backofen R. et al. Locality and gaps in RNA comparison //Journal of Computational Biology. – 2007. – Т. 14. – №. 8. – С. 1074-1087.

57. Bon M., Micheletti C., Orland H. McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots //Nucleic acids research. – 2012. – Т. 41. – №. 3. – С. 1895-1900.

58. Kato Y. et al. Rtips: fast and accurate tools for RNA 2D structure prediction using integer programming //Nucleic acids research. – 2012. – Т. 40. – №. W1. – С. W29-W34.

59. Bon M., Orland H. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots //Nucleic acids research. – 2011. – Т. 39. – №. 14. – С. e93-e93.

60. Jabbari H., Condon A. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures //BMC bioinformatics. – 2014. – Т. 15. – №. 1. – С. 147.

61. Andronescu M. S., Pop C., Condon A. E. Improved free energy parameters for RNA pseudoknotted secondary structure prediction //RNA. – 2010. – Т. 16. – №. 1. – С. 26-42.

62. Baulin E.F., Roytberg M.A. The database of RNA secondary structure elements // Proceedings of 6-th Moscow Conference on Computational Molecular Biology (MCCMB 13) July 25–28, 2013.

63. Reidys C. M. et al. Topology and prediction of RNA pseudoknots //Bioinformatics. – 2011. – Т. 27. – №. 8. – С. 1076-1085.

64. Zuker M., Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information //Nucleic acids research. – 1981. – Т. 9. – №. 1. – С. 133-148.

65. Zuker M., Sankoff D. RNA secondary structures and their prediction //Bulletin of mathematical biology. – 1984. – Т. 46. – №. 4. – С. 591-621.

66. Hofacker I. L. et al. Fast folding and comparison of RNA secondary structures //Monatshefte für Chemie/Chemical Monthly. – 1994. – Т. 125. – №. 2. – С. 167-188.

67. Lyngso R. B., Zuker M., Pedersen C. N. Fast evaluation of internal loops in RNA secondary structure prediction //Bioinformatics (Oxford, England). – 1999. – Т. 15. – №. 6. – С. 440-445.

68. Tinoco I., Uhlenbeck O. C., Levine M. D. Estimation of secondary structure in ribonucleic acids //Nature. – 1971. – Т. 230. – №. 5293. – С. 362.

69. Tinoco I. et al. Improved estimation of secondary structure in ribonucleic acids //Nature New Biology. – 1973. – Т. 246. – №. 150. – С. 40.

70. Jaeger J. A., Turner D. H., Zuker M. Improved predictions of secondary structures for RNA //Proceedings of the National Academy of Sciences. – 1989. – Т. 86. – №. 20. – С. 7706-7710.

71. Zuker M. On finding all suboptimal foldings of an RNA molecule //Science. – 1989. – Т. 244. – №. 4900. – С. 48-52.

72. Wuchty S. et al. Complete suboptimal folding of RNA and the stability of secondary structures //Biopolymers: Original Research on Biomolecules. – 1999. – Т. 49. – №. 2. – С. 145-165.

73. Hofacker I. L., Stadler P. F., Stocsits R. R. Conserved RNA secondary structures in viral genomes: a survey //Bioinformatics. – 2004. – Т. 20. – №. 10. – С. 1495-1499.

74. Eppstein D. et al. Sparse dynamic programming II: convex and concave cost functions //Journal of the ACM (JACM). – 1992. – Т. 39. – №. 3. – С. 546-567.

75. Will S. et al. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs //Rna. – 2012. – Т. 18. – №. 5. – С. 900-914.

76. Sorescu D. A. et al. CARNA—alignment of RNA structure ensembles // Nucleic acids research. – 2012. – Т. 40. – №. W1. – С. W49-W53.

77. Chiu J. K. H., Chen Y. P. P. Pairwise RNA secondary structure alignment with conserved stem pattern //Bioinformatics. – 2015. – Т. 31. – №. 24.

– C. 3914-3921.

78. Meer M. V. et al. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness //Nature. – 2010. – T. 464. – №. 7286. – C. 279.

79. Laing C., Schlick T. Computational approaches to RNA structure prediction, analysis, and design //Current opinion in structural biology. – 2011. – T. 21. – №. 3. – C. 306-318.

80. Groebe D. R., Uhlenbeck O. C. Characterization of RNA hairpin loop stability //Nucleic acids research. – 1988. – T. 16. – №. 24. – C. 11725-11735.

81. Szewczak A. A. et al. The conformation of the sarcin/ricin loop from 28S ribosomal RNA //Proceedings of the National Academy of Sciences. – 1993. – T. 90. – №. 20. – C. 9581-9585.

82. Krasilnikov A. S., Mondragón A. On the occurrence of the T-loop RNA folding motif in large RNA molecules //Rna. – 2003. – T. 9. – №. 6. – C. 640-643.

83. Heus H. A., Pardi A. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops //Science. – 1991. – T. 253. – №. 5016. – C. 191-194.

84. Xin Y. et al. Annotation of tertiary interactions in RNA structures reveals variations and correlations //Rna. – 2008. – T. 14. – №. 12. – C. 2465-2477.

85. Nissen P. et al. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif //Proceedings of the National Academy of Sciences. – 2001. – T. 98. – №. 9. – C. 4899-4903.

86. Tamura M., Holbrook S. R. Sequence and structural conservation in RNA ribose zippers //Journal of molecular biology. – 2002. – T. 320. – №. 3. – C. 455-474.

87. Davis J. H. et al. RNA helical packing in solution: NMR structure of a 30 kDa GAAA tetraloop–receptor complex //Journal of molecular biology. – 2005. – T. 351. – №. 2. – C. 371-382.

88. Ueda T. et al. The T-loop region of animal mitochondrial tRNASer (AGY) is a main recognition site for homologous seryl-tRNA synthetase //Nucleic

acids research. – 1992. – Т. 20. – №. 9. – С. 2217-2222.

89. Hamdani H. Y. et al. NASSAM: a server to search for and annotate tertiary interactions and motifs in three-dimensional structures of complex RNA molecules //Nucleic acids research. – 2012. – Т. 40. – №. W1. – С. W35-W41.

90. Lu X. J., Bussemaker H. J., Olson W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA //Nucleic acids research. – 2015. – Т. 43. – №. 21. – С. e142-e142.

91. Ashraf S. S. et al. The uridine in “U-turn”: contributions to tRNA-ribosomal binding //Rna. – 1999. – Т. 5. – №. 4. – С. 503-511.

92. Westbrook J. D., Fitzgerald P. M. D. The PDB format, mmCIF formats, and other data formats //Structural bioinformatics. – 2009. – Т. 44.

93. Bon M. et al. Topological classification of RNA structures //Journal of molecular biology. – 2008. – Т. 379. – №. 4. – С. 900-911.

94. Chiu J. K. H., Chen Y. P. P. Conformational features of topologically classified RNA secondary structures //PloS one. – 2012. – Т. 7. – №. 7. – С. e39907.

95. Aalberts D. P., Hodas N. O. Asymmetry in RNA pseudoknots: observation and theory //Nucleic acids research. – 2005. – Т. 33. – №. 7. – С. 2210-2214.

96. Peselis A., Serganov A. Structure and function of pseudoknots involved in gene expression control //Wiley Interdisciplinary Reviews: RNA. – 2014. – Т. 5. – №. 6. – С. 803-822.

97. Баулин Е. Ф., Астахова Т. В., Ройтберг М. А. Классификация элементов вторичной структуры РНК //Математическая биология и биоинформатика. – 2012. – Т. 7. – №. 2. – С. 567-571.

98. Baulin E. et al. URS DataBase: universe of RNA structures and their motifs //Database. – 2016. – Т. 2016.

99. Leontis N. B., Westhof E. Geometric nomenclature and classification of RNA base pairs //Rna. – 2001. – Т. 7. – №. 4. – С. 499-512.

100. Baulin E., Ivankov D., Roytberg M. Statistics of RNA structures //Proceedings of Moscow Conference on Computational Molecular Biology (July 21-24, 2011. Moscow). – 2011. – С. 325-326.

101. Exterior loop in RNA secondary structure: [Электронный ресурс]. URL: http://x3dna.org/articles/exterior-loop-in-rna-secondary-structure. (Дата обращения: 01.04.2019)

102. Andersen J. E. et al. Topological classification and enumeration of RNA structures by genus //Journal of mathematical biology. – 2013. – Т. 67. – №. 5. – С. 1261-1278.

103. Oliphant T. E. Python for scientific computing //Computing in Science & Engineering. – 2007. – Т. 9. – №. 3. – С. 10-20.

104. Loeliger J., McCullough M. Version Control with Git: Powerful tools and techniques for collaborative software development. – " O'Reilly Media, Inc.", 2012.

105. Gundavaram S. CGI programming on the World Wide Web. – O'Reilly & Associates, 1996.

106. Pilgrim M. HTML5: up and running: dive into the future of web development. – " O'Reilly Media, Inc.", 2010.

107. Sikos L. Web standards: mastering HTML5, CSS3, and XML. – Apress, 2014.

108. Goodman D. JavaScript bible. – John Wiley & Sons, 2004.

109. Bibeault B., Kats Y. jQuery in Action. – Dreamtech Press, 2008.

110. Garrett J. J. et al. Ajax: A new approach to web applications. – 2005.

111. Herraez A. Biomolecules in the computer: Jmol to the rescue //Biochemistry and Molecular Biology Education. – 2006. – Т. 34. – №. 4. – С. 255-261.

112. Hanson R. M. et al. JSmol and the next generation web based representation of 3D molecular structure as applied to proteopedia //Israel Journal of Chemistry. – 2013. – Т. 53. – №. 34. – С. 207-216.

113. Sunderaraman P. Overview of Ext JS 4 //Practical Ext JS 4. – Apress, Berkeley, CA, 2013. – С. 9-15.

114. Баулин Е. Ф. и др. Короткие стемы в псевдоузловых структурах РНК //Математическая биология и биоинформатика. – 2018. – Т. 13. – №. 2. – С. 526-533.

115. Tan Z. et al. RNA folding: structure prediction, folding kinetics and ion electrostatics //Advance in Structural Bioinformatics. – Springer, Dordrecht, 2015. – С. 143-183.

116. Leontis N. B., Zirbel C. L. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking //RNA 3D structure analysis and prediction. – Springer, Berlin, Heidelberg, 2012. – С. 281-298.

117. Баулин Е. Ф., Тихонова П. О., Ройтберг М. А. Предсказание сайтов связывания ионов магния с РНК методами машинного обучения // Доклады Международной конференции "Математическая биология и биоинформатика". — Т. 7. — ИМПБ РАН Пущино, 2018.

118. Liaw A. et al. Classification and regression by randomForest //R news. – 2002. – Т. 2. – №. 3. – С. 18-22.

119. Banatao D. R., Altman R. B., Klein T. E. Microenvironment analysis and identification of magnesium binding sites in RNA //Nucleic acids research. – 2003. – Т. 31. – №. 15. – С. 4450-4460.

120. Philips A. et al. MetalionRNA: computational predictor of metal-binding sites in RNA structures //Bioinformatics. – 2011. – Т. 28. – №. 2. – С. 198-205.