Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича

Российской академии наук

На правах рукописи

Виноградова Светлана Владимировна

Предсказание структурных элементов РНК с использованием экспериментальных данных

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:

кандидат физико-математических наук,

доктор биологических наук, профессор

Андрей Александрович Миронов

Москва, 2016

Оглавление

Введение

Актуальность темы исследования и степень ее разработанности

Рибонуклеиновая кислота (РНК) – одна из основных универсальных макромолекул, присутствующая в всех живых клетках и выполняющая самые разнообразные функции. Существует большое количество классов РНК, регулирующих самые разные клеточные процессы, от транскрипции до сплайсинга и модификации хроматина. Последние двадцать лет мы наблюдаем прорыв в области биологии РНК, сопровождающийся открытием десятков новых классов некодирующих РНК.

Многие функциональные некодирующие РНК обладают вторичной структурой, и именно вторичная структура часто играет ключевую роль для функциональности данных РНК. Наличие консервативной и функциональной вторичной структуры молекулы РНК чаще всего говорит об её участии в биологических процессах клетки [61], поэтому анализ структур и поиск новых структур в масштабе целого генома является фундаментальной и актуальной задачей, решение которой поможет глубже понять клеточные процессы.

Открытие новых классов некодирующих РНК со стабильной вторичной структурой ставит задачу *de novo* поиска структурированных элементов в длинных последовательностях РНК [31]. Методы сравнительной геномики позволяют находить функциональные структурные элементы РНК [65, 83], детектируя давление эволюционного отбора на структуру [67] или опираясь на ковариационные модели [63]. Однако часто набор гомологичных последовательностей для анализа с подходящим уровнем дивергенции и компенсаторных замен может быть недоступен. В этом случае приходится работать с отдельными последовательностями РНК. Высокая стабильность вторичной структуры РНК является её важной особенностью, и функциональные

РНК элементы действительно имеют более низкую свободную энергию структуры, чем случайные последовательности той же длины и нуклеотидного состава. Программа RNASurface позволяет сканировать длинные последовательности РНК, выделяя функциональные локально структурированные элементы РНК [84], реконструируя полный ландшафт структурованности последовательности.

Экспериментальные методы анализа вторичной структуры РНК предоставляют альтернативный источник информации. Данные методы позволяют детектировать позиции РНК, более доступные для химических реагентов или ферментов и, тем самым, в зависимости от эксперимента, более или менее склонные к образованию вторичной структуры. В настоящее время такие эксперименты проводят *in vitro* и *in vivo*, на уровне отдельных молекул и в масштабах целого транскриптома, что позволяет получить информацию о структурах РНК в различных условиях.

Сложность использования экспериментальных данных для анализа вторичной структуры заключается в том, что эксперимент предоставляет только вероятностную информацию о статусе конкретного нуклеотида. Одной этой информации недостаточно для того, чтобы определить вторичную структуру РНК. Традиционным подходом является использование информации о спаренности нуклеотидов в качестве ограничения при процедуре минимизации энергии [20, 104]. Было показано, что использование экспериментальных данных при поиске гомологичных РНК позволяет повысить точность и эффективность поиска [29]. Несмотря на большое количество экспериментальных данных для различных организмов, на данный момент не существует универсального алгоритма, позволяющего проводить полногеномный поиск структурированных РНК с использованием экспериментальных данных по определению структуры.

Цели и задачи исследования

Целью данного исследования было разработать метод, позволяющий использовать различные экспериментальные данные по определению принадлежности отдельных нуклеотидов к вторичной структуре РНК при сканировании генома для распознавания локальных стабильных структур РНК.

Для достижения цели были поставлены следующие задачи:

1) Анализ существующих экспериментальных методик определения принадлежности отдельных нуклеотидов к вторичной структуре РНК; разработка единого представления экспериментальных данных, полученных с помощью различных методик, для последующего использования в качестве дополнительного источника информации при поиске структурированных РНК.
2) Разработка алгоритма, позволяющего учитывать экспериментальные данные при поиске структурированных РНК, и оценка эффективности данного алгоритма на примере транскриптома человека.

Научная новизна и теоретическая и практическая значимость работы

Традиционным подходом к *de novo* поиску структурированных РНК в длинных последовательностях является сканирование последовательностей с целью поиска сегментов с низкой энергией и стабильной вторичной структурой. Данные подходы не используют доступные данные по экспериментальному определению вторичной структуры РНК, которые являются важным альтернативным источником структурной информации.

В настоящей работе мы проанализировали несколько типов экспериментальных данных по определению вторичной структуры РНК. Мы

разработали теоретический поход для преобразования экспериментальных с целью дальнейшего включения их в энергетическую модель программы RNASurface. Поиск структурированных элементов РНК в масштабах транскриптома человека с помощью модифицированной версии программы RNASurface, использующей экспериментальные данные, показал, что включение данных эксперимента позволяет увеличить эффективность поиска и находить функциональные структурированные РНК элементы.

Положения, выносимые на защиту

1) Разработана методика анализа и преобразования экспериментальных данных, касающихся принадлежности отдельных нуклеотидов к вторичной структуре РНК. Методика основана на сопоставлении каждому нуклеотиду количественной характеристики, отражающей его склонность быть включенным во вторичную структуру.
2) Проведено сравнение данных эксперимента ДМС *in vivo* и *in vitro*. Показано, что кодирующие области мРНК являются менее структурированными в клетке по сравнению с состоянием *in vitro*.
3) Алгоритм RNASurface расширен на случай использования в оценке степени структурированности фрагмента экспериментальных данных. Построена фоновая модель для оценки структурированности РНК, учитывающая как энергетические параметры, так и экспериментальные данные.
4) Разработан и запущен веб-сервис RNASurface (http://bioinf.fbb.msu.ru/RNASurface/), позволяющий визуализировать результаты работы алгоритма по предсказанию структурированных элементов РНК с использованием экспериментальных данных.
5) На основании данных эксперимента PARS проведен анализ структурированности РНК элементов в масштабах транскриптома человека,

показавший, что использование экспериментальных данных при поиске структурированных элементов РНК позволяет улучшить качество предсказания.

Степень достоверности и апробация результатов

По материалам диссертации опубликовано 2 статьи в рецензируемых научных журналах. Результаты работы были представлены на международных конференциях: Интеллектуальные системы молекулярной биологии (Intelligent Systems for Molecular Biology – ISMB'14), Европейская конференция по вычислительной биологии (European Conference on Computational Bilology – ECCB'14), Седьмая Московская конференция по вычислительной молекулярной биологии (Moscow Conference on Computational Molecular Biology – MCCMB’15), Симпозиум Европейской Организации Молекулярной Биологии и Европейского Института Биоинформатики (EMBO/EMBL Symposium'15), а также на конференциях Информационные Технологии и Системы (ИТИС'14, ИТИС'15).

Глава 1. Обзор Литературы

1.1 Разнообразие мира РНК

Рибонуклеиновая кислота (РНК) – одна из трёх основных типов макромолекул, которые содержатся во всех живых клетках. Молекулы РНК выполняют самые разнообразные функции в клетке. Длины молекул РНК также весьма разнообразны от нескольких десятков нуклеотидов в случае малых РНК до нескольких тысяч нуклеотидов в случае сложных длинных молекул РНК.

Молекулы мРНК (англ. messenger RNA, или информационной РНК) принимают участие в трансляции: последовательность нуклеотидов, из которой состоит РНК, позволяет кодировать генетическую информацию и служит промежуточным звеном между ДНК и белком. Кроме мРНК, существует также огромное количество некодирующих РНК (нкРНК), которые также участвуют в трансляции и во многих других клеточных процессах. Многие высокоструктурированные РНК принимают участие в синтезе белков, например, транспортные РНК (тРНК) служат для узнавания кодонов и доставки соответствующих аминокислот к месту синтеза белка, а рибосомные РНК (рРНК) являются структурной и каталитической основой рибосом [103].

Кроме участия в трансляции, молекулы РНК выполняют самые разнообразные функции в клетке и участвуют практически во всех биологических процессах. Например, малые ядерные РНК (мяРНК) принимают участие в сплайсинге эукариотических матричных РНК и других процессах [53]. Малые ядрышковые РНК (мякРНК) – класс малых РНК, которые участвуют в химических модификациях (метилировании и псевдоуридилировании) рРНК, а также тРНК и мяРНК [38]. МикроРНК принимают участие в транскрипционной и посттранскипционной регуляции экспрессии генов [5]. piРНК (англ. Piwi-

interacting RNA, piRNA) представляют собой самый большой класс малых РНК, экспрессирующихся в клетках животных. PiРНК образуют комплексы с PIWI-белками, участвуя в эпигенетической и пост-транскрипционной регуляции экспрессии ретротраспозонов и других генетических элементов в зародышевых линиях [80].

Существует также отдельный класс РНК, характеризующийся длинной последовательностью, более 200 нуклеотидов, – длинные некодирующие РНК (англ. lncRNA). РНК данного класса также выполняют самые разнообразные функции, от регуляции транскрипции [26] до участия в эпигенетических процессах [102]. А для многих длинных некодирующих РНК функция до сих пор остается неясной.

1.2 Структура РНК

РНК – это полимер, состоящий из нуклеотидов четырёх видов: аденина (обозначается как A), цитозина (C), гуанина (G) и урацила (U). Последовательности нуклеотидов молекулы РНК представляет собой первичную структуру РНК. Кроме первичной, различают вторичную, третичную (пространственную) и четвертичную структуру РНК. Вторичная структура характеризуется образованием Уотсон-Криковских пар нуклеотидов, которые приводят к формированию структуры двойной спирали различной длины. Пространственная структура РНК – структура, характеризующаяся взаимодействием элементов вторичной структуры. Так, возможно образование дополнительных водородных связей между нуклеотидами или связей между OH-группами остатков рибозы и основаниями. Третичная структура РНК часто стабилизирована ионами двухвалентных металлов, например ионами Mg^{2+}, связывающимися не только с фосфатными группами, но и с основаниями. Четвертичная структура РНК характеризуется взаимодействием отдельных молекул РНК между собой и с белками.

В настоящее время методы исследования пространственной структуры РНК – как экспериментальные, так и вычислительные – весьма ограничены. Однако изучение вторичной структуры позволяет приблизиться к пониманию функциональности структуры молекулы РНК.

1.2.1 Вторичная структура РНК

Каждый нуклеотид РНК состоит из сахаро-фосфатного остова (рибоза-5-фосфат), к которому в положении 1’ присоединено одно из азотистых оснований (рис. 1.1). При образовании вторичной структуры азотистые основания образуют водородные связи: цитозин и гуанин образуют три водородные связи, а аденин и урацил – две водородные связи. Такие пары оснований называются комплементарными. Кроме того, гуанин и урацил также могут образовывать две водородные связи; более того, в некоторых случаях другие неканонические пары оснований также образуют связи [44].

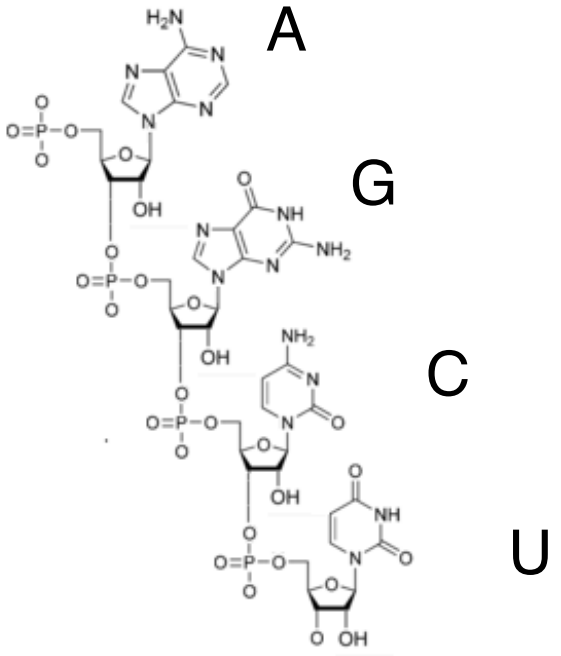


Рисунок 1.1. Химическое строение цепи РНК.

Одиночные пары оснований термодинамически не стабильны, однако формирование нескольких последовательных пар приводит к существенному увеличению стабильности: образуется стебель. На рис. 1.2 приведены мотивы, наиболее часто встречающихся во вторичных структурах. Однонитевые участки РНК, ограниченные спаренными основаниями, называются петлями. Петли могут быть как на конце стебля, так и в самом стебле, в последнем случае такая петля называется внутренней петлей. Группа неспаренных оснований только в одной из нитей РНК называется выпетливанием. Кроме того, различают разветвленные петли – петли, от которых отходит три и более стеблей.

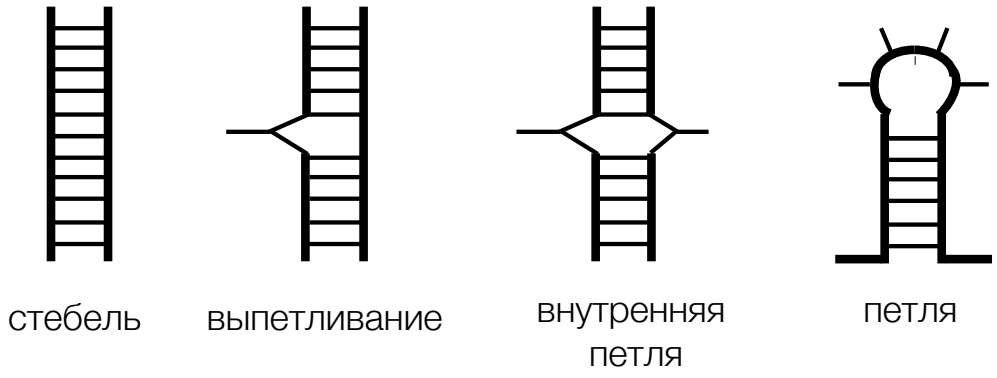


Рисунок 1.2. Мотивы вторичной структуры РНК.

1.2.2 Пространственная структура РНК

Третичная структура образуется на основе элементов вторичной структуры с помощью водородных связей внутри молекулы. В естественных условиях РНК укладывается в 3D структуру, при этом спирали и неспаренные участки строго определенно располагаются друг относительно друга в пространстве, образуя так называемые третичные взаимодействия [10].

Коаксиальный или спиральный стэкинг является основой третичных взаимодействий в структуре РНК. Такой стэкинг возникает между двумя двухцепочечными фрагментами РНК (спиралями), определенным образом расположенными друг относительно друга в пространстве. Если два двухцепочечных фрагмента РНК располагаются друг за другом, то есть разделены только фосфодиэфирной связью, то крайние нуклеотиды этих фрагментов образуют стэкинг-взаимодействие, а сами фрагменты оказываются параллельными (коаксиальными). Впервые коаксиальный стэкинг был описан для молекулы фенилаланин-тРНК [70]. Позже было показано, что коаксиальный стэкинг присутствует также в молекулах других РНК, например, рибосомальных РНК [95] и интронов РНК 1 и 2 групп [93]. В целом, именно связующие элементы

между отдельными стеблями и являются критичными для образования правильной пространственной структуры РНК, определяя взаимное расположение этих стеблей. Важно также учитывать, что концентрации ионов в клетке оказывают большое влияние на связующие элементы и их расположение и, таким образом, на всю пространственную структуру РНК в целом [45]. Также в пространственной структуре часто возникают взаимодействия между отдельными петлями. Такие взаимодействия включают в себя образование псевдоузлов и kissing-петель (Рис. 1.3).

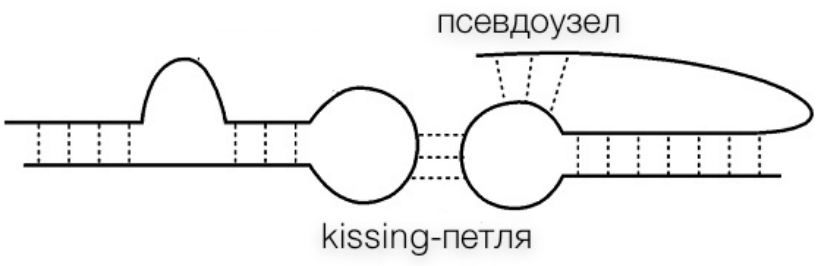


Рисунок 1.3. Примеры третичных взаимодействий

1.3 Вычислительные подходы предсказания вторичной структуры РНК

Принимая во внимание существующие трудности в экспериментальных методах определения вторничной структуры РНК, было разработано и применено множество алгоритмов для предсказания структуры РНК по её последовательности. Безусловно, данные подходы имеют большие преимущества: они могут предсказать структуру для абсолютно любой последовательности, не требуя наличия экспериментальных данных и даже более того – физического существования самой последовательности РНК. Подходы, основанные на вычислительном предсказании структур, позволили сделать многие биологические открытия и наблюдения. Например, для определенных классов

некодирующих РНК, чья структура важна для функциональности, применение таких подходов позволило выявить новых членов класса: это верно для тРНК, мякРНК и микроРНК. А комбинирование подхода предсказания вторичной структуры со сравнительным геномным анализом позволяет решать более общую задачу поиска новых классов нкРНК [67, 73].

1.3.1 Свободная энергия вторичной структуры

При вычислительном подходе предсказания вторичной структуры РНК необходимо найти такую структуру, в которую РНК сворачивается с большей вероятностью, относительно других возможных структур. Для оценки того, насколько вероятна та или иная структура, используется метод оценки изменения свободной энергии при температуре 37 °C ΔG_{37}. Для данной молекулы РНК в равновесии существует равновесие между последовательностью, свернутой в структуру S_1 и развернутой структурой RC:

$$RC \rightleftharpoons S_1,$$

где константа равновесия K1 вычисляется как:

$$K_1 = \frac{[S_1]}{[RC]}$$

Через изменение свободной энергии для структуры S_1 $\Delta G_{37}(1)$ можно вычислить стабильность данной структуры:

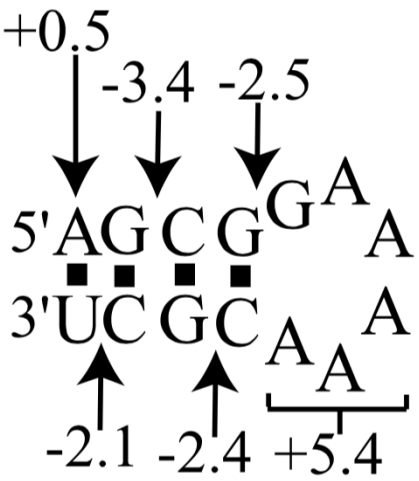
$$K_1 = e^{-\Delta G_{37}(1)/RT}$$

где R – газовая постоянная, а T – абсолютная температура. Таким образом, можно выразить отношение стабильностей и концентраций двух структур через их свободную энергию:

$$\frac{K_1}{K_2} = \frac{[S_1]}{[S_2]} = e^{(\Delta G_{37}(2) - \Delta G_{37}(1))/RT}$$

Таким образом, структура с минимальной энергией является наиболее представленной в равновесии в растворе.

Для расчета свободной энергии вторичной структуры используется эмпирическая модель «ближайшего соседа» [91]. Метод носит название «ближайшего соседа», так как при расчете свободной энергии основания учитывается вид основания и контекст, а именно ближайшие соседние основания. На рис. 1.4 представлен расчет свободной энергии для структуры типа «стебель-петля». Методы предсказания структуры с наименьшей свободной энергией называются методами минимизации свободной энергии (МСЭ).



$\Delta G_{37\ total}$ = 0.5-2.1-3.4-2.4-2.5+5.4=
=-4.5 ккал/моль

Рисунок 1.4. Пример расчета свободной энергии структуры, рисунок адаптирован из [54]**.**

1.3.2 Алгоритмы предсказания вторичной структуры

Если структура с наименьшей энергией является самой стабильной, то для того, чтобы предсказать самую стабильную структуру, достаточно найти

структуру с минимальной энергией. Простейший метод – перебрать все возможные структуры для данной последовательности, но было показано, что число возможных структур растет экспоненциально с длиной последовательности [106]. Для последовательности длиной 100 нуклеотидов существует около 10^{25} возможных структур. Первое эффективное и наиболее популярное решение данной проблемы – использование метода динамического программирования.

Тиноко и соавторы в 1971 предложили простой метод для оценки вторичной структуры РНК, основанный на разделении последовательности на отдельные блоки, а именно петли (внутренние и внешние), выпетливания и стебли [92]. Каждому блоку присваивается вес, в зависимости от того, стабилизирует он структуру или дестабилизирует. Петлям присваивается отрицательный вес, что ведет к увеличению свободной энергии, а стеблям присваивается положительный вес, что в свою очередь ведет к уменьшению свободной энергии. Кроме того, вводится понятие матрицы спаривания нуклеотидов: каждой паре нуклеотидов присваивается некоторый положительный вес, если они могут образовать пару. Далее на основании матрицы спаривания нуклеотидов и информации о положительном и отрицательном вкладе в энергию стеблей и петель можно рассчитать вес структуры. Данная работа явилась первым шагом к разработке алгоритмов эффективного поиска стабильных вторичных структур РНК.

В 1978 году Нуссинов предложила использовать подход динамического программирования для поиска структуры с наибольшим количеством спаренных нуклеотидов [64]. В данном подходе была использована идея динамического программирования, заключающаяся в том, что для оценки какого-либо свойства (например, числа спаренных нуклеотидов или свободной энергии) более длинной последовательности можно опираться на результат, полученной для более короткой последовательности, а это значительно облегчает расчеты. Алгоритм Нуссинов позволяет находить структуры с максимальным количеством спаренных нуклеотидов, однако, найденная структура далеко не всегда является структурой с наименьшей энергией.

В 1981 году был разработан алгоритм Зукера – алгоритм динамического программирования, в котором свободная энергия вторичной структуры оценивается как сумма свободных энергий отдельных элементов [107]. Важное преимущество по сравнению с алгоритмом Нуссинов состоит в том, что при расчете учитывается стэкинг – энергия взаимодействия соседних пар, образующих стебель. Параметры для данного алгоритма была рассчитаны из результатов экспериментальных термодинамических исследований малых РНК [27]. Эти параметры учитывают стэкинг, длины петель и выпетливаний, одиночных нуклеотидов и неспаренных концов стеблей. Данный алгоритм работает за время $O(N^3)$, где N – длина последовательности.

1.3.3 Субоптимальные структуры РНК

Сравнение вторичных структур РНК, имеющих минимальную свободную энергию, с экспериментально расшифрованными структурами показывает, что далеко не всегда структура с минимальной энергией реализуется в клетке и выполняет биологическую функцию [23]. Отчасти это объясняется тем, что в клетке структуры стабилизированы третичными и четвертичными взаимодействиями. Но кроме того, энергетический спектр субоптимальных структур может быть достаточно плотным: клетка содержит целый ансамбль структур для одной последовательности РНК. Более того, некоторые РНК, например, рибопереключатели, существуют сразу в нескольких функциональных конформациях [15, 33, 97]. Таким образом, часто бывает важным получить не только структуру с минимальной энергией, но и оценить субоптимальные с точки зрения энергии структуры.

Можно рассчитать статистическую сумму Q как сумму всех констант равновесия K_i для всех возможных структур:

$$Q = \sum_i K_i = \sum_i e^{-\Delta G_i/RT}$$

Таким образом, молекула РНК пребывает в данной конформации *i* с частотой, описываемой распределением Больцмана:

$$P_i = \frac{e^{-\Delta G_i/RT}}{Q}$$

Для расчета статистических сумм применяется алгоритм МакКаскила, реализованный с помощью эффективного динамического программирования [57]. Данный алгоритм позволяет вычислить вероятности спаривания нуклеотидов последовательности, на основании полного спектра структур. Кроме того, на основании алгоритма МакКаскила был разработан метод, позволяющий семплировать структуры из ансамбля структур, согласно их вероятностям [21].

1.3.4 Эволюционный подход

Рассмотренные выше методы предсказания вторичной структуры опираются на одну последовательность. Но в некоторых случаях может быть доступен набор гомологичных последовательностей с подходящем уровнем дивергенции. В этом случае применимы методы сравнительной геномики. Методы, основанные на сравнительном анализе последовательностей, опираются на тот факт, что многие известные функциональные РНК структуры сохраняются в процессе эволюции. Примером служат тРНК, рибосомные РНК (рРНК) и рибозимы (интроны I и II групп). Ковариационный метод определяет вторичную структуру путем исследования паттернов консервативности пар нуклеотидов в ортологичных или паралогичных генах [35] [100].

Основным ограничением эволюционного подхода является условие наличия ортологичных последовательностей, а для многих подходов требуется и их

выравнивание. Однако построение точного выравнивания нкРНК до сих пор остается нерешенной задачей по двум причинам. Во-первых, в отличие от белков, последовательности нкРНК составлены из четырехбуквенного алфавита, что делает сложным использование сходства последовательностей в качестве меры оценки биологической осмысленности выравнивания. Как следствие, появляется так называемая «сумеречная зона», в районе 70% сходства, за пределами которой теряется информативность парных выравниваний. Вторая причина сложности выравнивания нкРНК заключается в том, как проходит их эволюция: большинство функциональных нкРНК обладает структурой, поддерживаемой компенсаторными мутациями. Это приводит к тому, что многие родственные нкРНК могут обладать сходной структурой, но сильно разошедшимися последовательностями, а это, в свою очередь, затрудняет выравнивание нкРНК на основании только их последовательностей. Таким образом, при выравнивании нкРНК необходимо принимать во внимание также их структуру.

Алгоритм одновременного сворачивания и выравнивания нкРНК был предложен Санковом еще в 1985 году [79], однако этот алгоритм требует огромных затрат как по времени, так и по памяти. Существующие на данный момент упрощения данного алгоритма в основном вносят ограничения по длине нкРНК или типам рассматриваемых структур. Таким образом, эволюционный подход, требующий наличие выравнивания, применим далеко не всегда.

1.4 Сканирование генома для поиска структурированных участков РНК

Для поиска локально структурированных элементов длиной L в длинной последовательности длиной N можно использовать подход скользящего окна: рассчитать свободную энергию всех возможных окон длины L с помощью одной из программ, реализующих алгоритм Зукера [107]. Алгоритм Зукера требует $O(L^3)$ времени, всего окон $N - L$, таким образом, сканирование всей последовательности займет время $O(NL^3)$. В 2004 году был разработан элегантный подход,

позволяющий сократить время работы до $O(NL^2)$ [36]. Основанием для ускорения работы алгоритма является наблюдение о том, что при расчете матрицы динамического программирования размером N на N на самом деле необходимо рассчитать значения только для диагональной части этой матрицы шириной L. Программа RNAslider реализует дальнейшее ускорение работы алгоритма до $O(NL)$ [37].

Однако задача *de novo* поиска функциональных структурированных РНК отличается от предсказания структуры по последовательности. Программы, реализующие алгоритм Зукера (например, mfold [105] и RNAfold [50]) или его модификации могут свернуть в структуру любую последовательность, однако это не значит, что данная структура является функциональной. Чтобы отличить структурированную последовательность от фона, нужна мера, соответствующая уровню структурированности, позволяющая эффективно разделять фон и функциональные структуры. Несмотря на то, что многие функциональные РНК являются более стабильными и обладают меньшей свободной энергией, чем случайные последовательности с тем же динуклеотидным составом [14], значение свободной энергии не является статистически значимым сигналом для выделения структурированных РНК [73]. Несмотря на это, свободная энергия РНК может быть использована для поиска новых членов отдельных семейств функциональных РНК, например, микро РНК [8], а в сочетании со сравнительной геномикой является хорошим инструментом для поиска нкРНК в целом [100].

Другие меры, оценивающие структурированность РНК, такие как нормализованная энергия, Z-значение, p-значение минимальной энергии, энтропия Шеннона и другие могут также быть использованы для эффективного предсказания функциональных РНК [28]. Например, Z-значение рассчитывается по формуле

$$Z\text{–значение} = \frac{E - \mu}{\sigma}$$

где E – значение свободной энергии последовательности, μ – среднее и σ – стандартное отклонение энергий перемешанных последовательностей той же длины и динуклеотидного состава.

1.4.1 Программа RNASurface

Программа RNASurface, разработанная на основе быстрого алгоритма RNAslider [37], позволяет искать локально структурированные участки в длинных последовательностях РНК [84] и использует Z-значение в качестве меры структурированности подпоследовательности длинной РНК. RNASurface реконструирует полный ландшафт структурированности последовательности, рассчитывая Z-значения для всех подпоследовательностей, длиной меньше пороговой (рис. 1.5). Далее на основании полученных Z-значений выбираются локально оптимальные структурированные участки РНК.

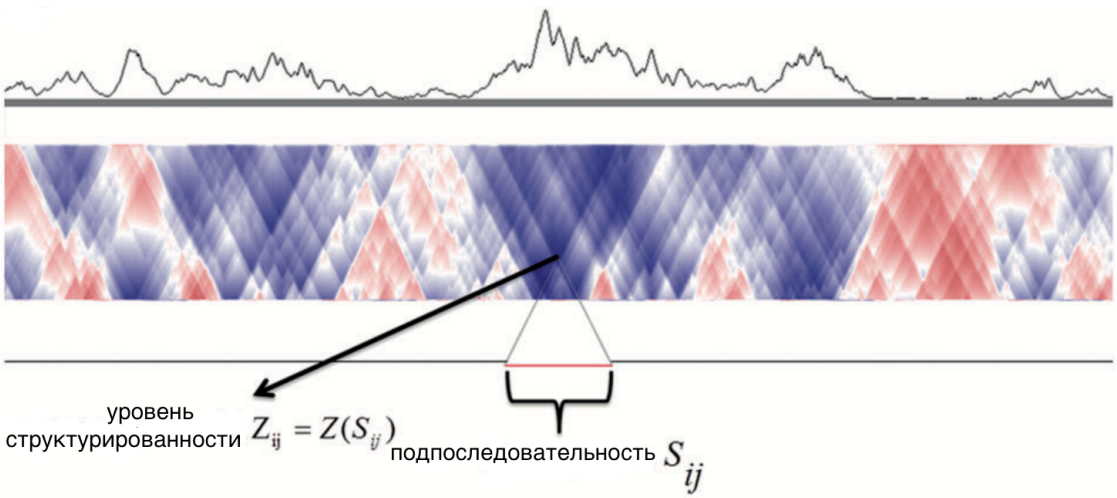


Рисунок 1.5. Ландшафт структур РНК, рассчитанный программой RNASurface.

1.4.2 Эволюционный подход

Использование эволюционного подхода для поиска локально структурированных РНК позволяет значительно улучшить эффективность алгоритмов поиска. Дополнительный сигнал, который может быть использован при поиске функциональных структур, основан на том, что замены нуклеотидов в стеблях РНК происходят с сохранением спаренных нуклеотидов, то есть компенсаторно. То есть происходит сразу две замены, таким образом, чтобы пара спаренных нуклеотидов сохранилась, например, если один нуклеотид из пары заменился с C на A, то другой заменится с G на U [74]. Комбинация филогенетических моделей и контекстно-свободных грамматик позволяет оценить, насколько паттерн замен для каждой колонки выравнивания соответствует вторичной структуре РНК [25]. Программа EvoFold использует очень общую модель структуры РНК, которая позволяет моделировать любую подструктуру от коротких шпилек до больших разветвленных структур, включая новые подструктуры, не содержащиеся в обучающей выборке. В программе заложена модель, отражающая ко-эволюцию спаренных нуклеотидов на филогенетическом дереве [67]. На данный момент программа EvoFold может считаться «золотым стандартом» для поиска локально структурированных участков РНК, с которым можно сравнивать другие программы.

1.5 Экспериментальные методы определения структуры РНК

Экспериментальные подходы к определению структуры РНК включают в себя рентгеноструктурный анализ [32], ЯМР-спектроскопию [43], крио-электронную микроскопию [62] и футпринтинг [2, 9, 19, 59, 75]. В основе футпринтинга, или пробинга РНК (англ. probing), лежит исследование молекулы РНК с помощью химических веществ или ферментов. Такие эксперименты проводятся как *in vitro*, так и *in vivo*, в присутствии или отсутствии белков и

других лигандов, при различных температурах и других условиях. Как в случае исследования с помощью химических реагентов, так и в случае ферментов, доступность молекулы РНК по большой части определяет ее реакциоспособность. В случае химических реагентов определенное химическое вещество, способное взаимодействовать с определенным основанием или сахаро-фосфатным остовом, реагирует с молекулой РНК. Если в реакции участвует фермент, то он реагирует или с одноцепочечными участками РНК, или наоборот, с двухцепочечными, чаще всего внося разрывы в цепи РНК. Положение модифицированных нуклеотидов или разрывов в РНК определяют с помощью обратной транскрипции с использованием меченого праймера. Синтез растущей цепи обрывается, когда обратная транскриптаза доходит до модифицированного основания в РНК, которое не может образовывать «нормальную» комплементарную пару или до места разрыва в РНК. Синтезированные фрагменты комплементарной ДНК определяют с помощью электрофореза. В случае электрофореза в полиакриламидном геле положение модифицированных нуклеотидов или разрывов определяется по паттерну движения полос, и интенсивность полос оценивается с помощью программ анализа изображений, например, программы для полу-автоматического анализа футпринтов [18]. Применение капиллярного электрофореза для анализа футпринтов стало важным шагом к высокопроизодительному анализу структур РНК. Несмотря на простоту процедуры с случае коротких РНК, исследование структуры более длинных РНК является непростой задачей. Электрофорез в полиакриламидном геле обычно позволяет обрабатывать одновременно только РНК длиной до 100 нуклеотидов, в то время как исследование РНК длиной несколько тысяч нуклеотидов приводит к необходимости одновременного использования десятков и сотен гелей. Капиллярный электрофорез позволяет увеличить разрешение до 300-650 нуклеотидов, тем самым позволяя исследовать более длинные РНК, особенно при условии параллельных экспериментов [59].

РНК футпринтинг также может быть проведен *in vivo*. Так как в некоторых случаях пространственная структура РНК *in vitro* может сильно отличаться от

структуры *in vivo* [76], проведение экспериментов *in vivo* может представлять интерес с точки зрения регуляции сворачивания РНК в клетке [78]. РНК футпринтинг может проводиться внутри живых клеток с помощью химических веществ, способных проникать через клеточную мембрану (например, Pb^{2+} или DMS) или с помощью рентгеновского излучения [1, 49].

1.5.1 Методы SHAPE, DMS, PARS

В 2005 году в лаборатории доктора Викса был разработан метод SHAPE (англ. selective 2-hydroxyl acylation analysed by primer extension – выборочное ацилирование 2'-гидроксильной группы, анализируемое методом удлинения праймера) для высокочувствительного анализа структур РНК [58]. Было показано, что реакциоспособность гидроксила рибозы РНК во многом определяется гибкостью участка, в котором находится соответственный нуклеотид, ввиду конформационных ограничений [12]. Данная особенность легла в основу метода SHAPE, который использует ангидрид N-метилизатоиновой кислоты (NMIA) и его производные, чтобы исследовать гибкие регионы во вторичной структуре РНК [58]. Реакция с NMIA ведет к образованию 2-О аддукта, что позволяет оценивать локальную структурированность участков РНК. Такие 2-О аддукты могут быть определены с помощью обратной транскрипции с последующим капиллярным электрофорезом. Далее требуется обработка данных, полученных с электрофореграм, для преобразования их в сигнал: вводится понятие «реактивности» SHAPE для каждого нуклеотида, представляющей склонность данного нуклеотида к образованию 2-О аддукта. Было показано, что SHAPE реактивность коррелирует с локальной пространственной неструктурированностью и таким образом может являться мерой динамики структуры [30]. В целом, нуклеотиды с высокой реактивностью является в среднем неструктурированными, в то время как нуклеотиды с низкой

реактивностью ограничены в своей динамике канонической или неканонической вторичной или пространственной структурой.

Было продемонстрировано, что метод SHAPE может определить структуру длинных РНК, например, 16S рРНК и РНК генома ВИЧ [20, 101]. Реконструкция вторичной структуры генома ВИЧ с помощью метода SHAPE стало доказательством того, что экспериментальные методы действительно могут быть очень полезными при анализе структур РНК. Геном ВИЧ состоит из однонитевой РНК длиной 9 тысяч нуклеотидов, в нем закодировано 9 открытых рамок считывания, которые транслируются в 15 белков, важных для заражения и репликации вируса. Начальное исследование первых 900 нуклеотидов при различных условиях показало высокое сходство структур внутри и вне вириона. Регуляторные регионы внутри этих 900 нуклеотидов оказались более структурированы, чем белок-кодирующие области. Структурированные РНК домены улучшили понимание процессов сдвига рамки считывания Gag-Pol и траслокации белка Env. Интересно, что нуклеотиды между независимо сворачивающимися доменами белка обладают большей структурированностью и способны спариваться, образуя структуры, тормозящие продвижение рисобомы, тем самым способствуя котрансляционному сворачиванию доменов [101].

Таким образом, сочетание РНК футпринтинга с капиллярным секвенированием открыло путь к изучению длинных РНК, а также целых РНК-геномов.

Альтернативным реагентом, также являющимся чувствительным к структуре РНК является диметилсульфат (ДМС) [90]. Из химических реагентов, используемых для исследования структуры РНК, ДМС является одним из старейших и универсальных. Он был использован для определения структуры РНК в 1980 году, когда Питти и Гилберт адаптировали методы, которые ранее были использованы для секвенирования ДНК и РНК [66]. В данном методе реагент ДМС метилирует аденозин и цитидин, которые находятся в неспаренном состоянии и доступны для модификации. Далее места модификаций можно

обнаружить с помощью обратной транскрипции: обратная транскриптаза блокируется при встрече с модифицированными сайтами. Метод ДМС-пробинга является мощным инструментом для определения структуры РНК и был применен к большому количеству молекул. Первые опыты были проведены на рибосомальных РНК и РНК-белковых комплексах [60, 87]. В настоящее время большое количество функциональных РНК было исследовано в ДМС экспериментах: интроны I и II группы [11, 16], сплайсосомные мяРНК [34, 42], рибонуклеаза P [94], структурированные участки мРНК [77].

Одним из больших преимуществ ДМС метода является возможность проведения эксперимента *in vivo*: ДМС легко проникает в клетки, таким образом, существует возможность исследовать структуру РНК внутри живых клеток и получать информацию о функциональных структурах, принимающих участие в клеточных процессах. Рускин и коллеги в 2013 году провели серию экспериментов на дрожжах *Saccharomyces cerevisiae* по определению вторичной структуры РНК *in vivo* и *in vitro* [76]. В качестве контроля они использовали образцы РНК, нагретые до 95°C, предполагая, что при данной температуре структура денатурирует и что данный контроль покажет шум в самом методе. Данные ДМС-пробинга *in vivo* и *in vitro* очень хорошо согласуется с известными структурами для нескольких РНК. Однако в случае длинных РНК, например, рибосомальных, картина более сложная. В случае *in vivo* эксперимента точность предсказания неспаренных нуклеотидов составила 94%, а чувствительность 90%. Однако в случае *in vitro* эксперимента предсказание оказалось значительно хуже, что говорит о том, что структура рибосомальных РНК в отсутствии белков сильно отличается от структуры в клетке.

Далее авторы сравнивали структурированность мРНК *in vivo* и *in vitro*, введя две метрики: коэффициент корреляции Пирсона (r значение), отражающий уровень сходства паттерна модификации между экспериментом и контролем, и коэффициент Джини, отражающий неравномерность распределения реактивностей между петлей и стеблем РНК. Далее авторы применили эти метрики к окнам, содержащим по 50 A/C нуклеотидов. Оказалось, что мРНК *in vitro* гораздо более структурированы, чем мРНК *in vivo*: авторы наблюдали сильный сдвиг r значения вниз и коэффициента Джини вверх в случае *in vitro*, в случае *in vivo* эффект был значительно менее сильным (рис. 1.6).

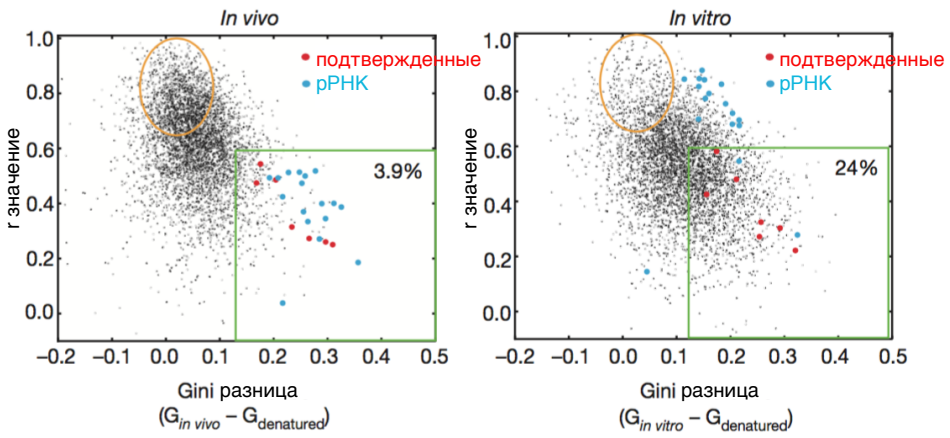


Рисунок 1.6. Точечный график разницы коэффициентов Джини против r значения для *in vivo* и *in vitro* против контроля, для не пересекающихся 5000 случайно выбранных окон мРНК, содержащих по 50 нуклеотидов A/C, данные для клеточной линии K562. Красные точки – регионы, соответствующие подтвержденным структурам мРНК, синие точки – регионы из рибосомальных РНК. Адаптировано из [76]**.**

Одним из центральных вопросов статьи Рускина [76] был вопрос о том, какой механизм приводит к столь сильным различиям в экспериментах *in vivo* и *in vitro*. Авторы предполагают, что несмотря на то, что трансляция играет роль в раскручивании структур, трансляция не является главной причиной различий в структурированности мРНК в двух состояниях. Авторы предполагают, что и активные механизмы (например, РНК-хеликазы), и пассивные механизмы (например, связывание белков с неспаренными участками РНК) вносят вклад с различие в раскручивание РНК *in vivo*. Для изучения энергозависимых процессов, вносящих вклад в раскручивание мРНК *in vivo*, авторы дополнительно провели ДМС-пробинг в дрожжах с выключенной работой АТФ. В этом случае наблюдалось увеличение структурированности мРНК *in vivo*, причем изменения в структурированности наблюдались в тех же сайтах, что и в случае *in vitro*. Таким образом, авторы подтвердили гипотезу о том, что активные, АТФ-зависимые механизмы вносят вклад в раскручивание мРНК *in vivo*.

Другим методом экспериментального определения структуры РНК является метод PARS (англ. parallel analysis of RNAstructure – параллельный анализ структуры РНК) [39]. Метод PARS состоит в применении сразу двух нуклеаз: S1 и V1. Первая нуклеаза расщепляет одноцепочечную ДНК или РНК, вторая – двухцепочечную. Таким образом, при проведении двух экспериментов параллельно можно получить информацию сразу и о неспаренных, и о спаренных нуклеотидах. Посчитав число чтений, начинающихся или заканчивающихся на определенной позиции последовательности, возможно оценить, насколько часто данный нуклеотид разрезается той или иной нуклеазой, и, соответственно, насколько вероятно то, что он находится в неспаренном состоянии.

1.5.2 Полногеномные карты структур РНК

Благодаря достижениями методов секвенирования нового поколения стало возможным исследовать структуры не только отдельных длинных РНК или РНК-

геномов, но и полных транскриптов геномов прокариот и эукариот. Разрывы или модификации в одно- или двухцепочечных структурах РНК могут быть детектированы путем перевода РНК в библиотеки кДНК, которые можно легко секвенировать. Полученные в результате секвенирования чтения картируются обратно на геном или транскриптом для определения местонахождения разрыва или модификации. Частота разрыва или модификации может быть определена из плотности чтений, картированных на данный участок. Такая стратегия позволяет одновременно детектировать сразу тысячи событий в одном эксперименте.

В методе PARS проводится сравнение чтений, полученных высокопроизводительным секвенированием одно- и двухцепочечных фрагментов РНК, которые были подвержены действию нуклеаз V1 и S1, соответственно [39]. Другой метод Frag-seq оценивает количество чтений, полученных в результате работы РНКазы P1, разрезающей одноцепочечные участки [96].

Метод SHAPE также был усовершенствован с помощью применения высокопроизводительного секвенирования [4][51]. Стало возможным анализировать не одну молекулу РНК, а целую смесь молекул сразу. Для того, чтобы можно было различать эти молекулы между собой, используют технику уникальных бар-кодов рядом с 3’ нетранслируемой областью (3' НТО) РНК. Далее образцы разделяются на (+) и (-) и обрабатываются SHAPE реагентом и контрольным растворителем, соответственно. Далее проводится процедура высокопроизводительного секвенирования с использованием платформы Illumina в режиме парных чтений. Необходимо прочитать только 50 нуклеотидов с каждого конца: с одной стороны чтений содержится бар-код, а с другой стороны - информация о позиции модификации SHAPE реагентом (Рис. 1.7). Чтобы от плотностей чтений, полученных в результате секвенирования, перейти к реактивностям нуклеотидов, необходимо провести нормализацию и шкалирование данных. Лукс и коллеги [4] разработали гибкую математическую модель для поиска оптимальных значений реактивностей, соответствующих наблюдаемым распределениям плотностей чтений в (+) и (-) экспериментах.

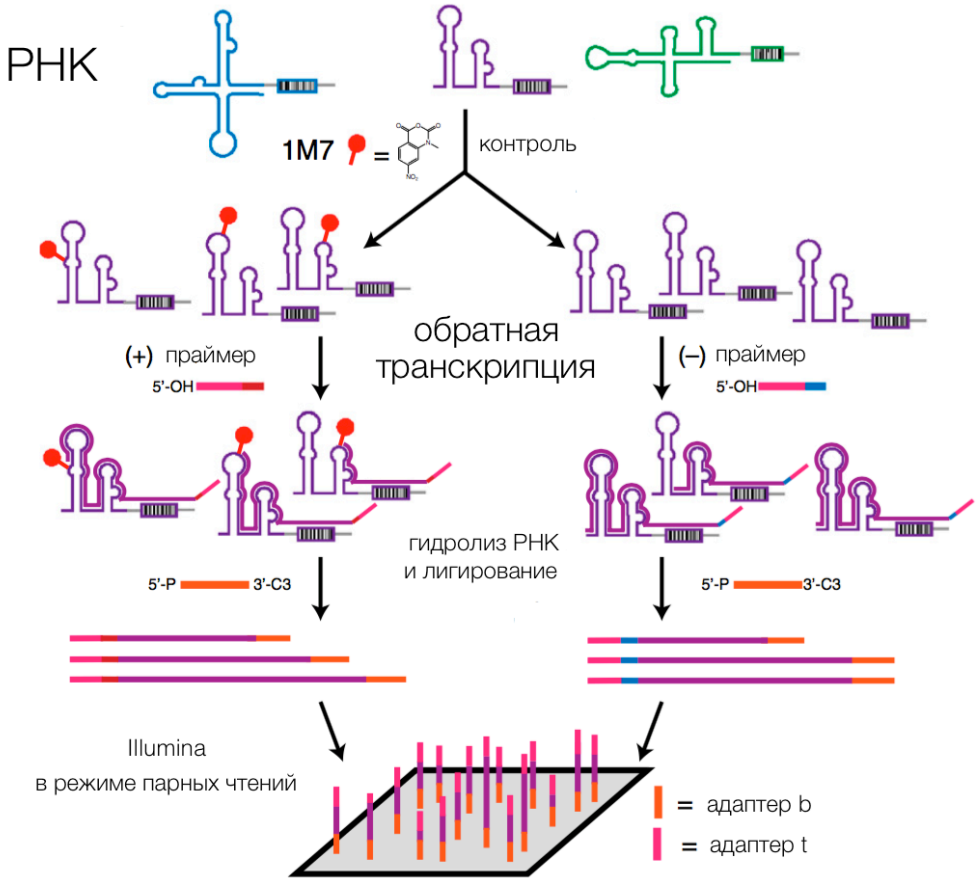


Рисунок 1.7. Схема работы метода SHAPE-seq, адаптировано из [4]**.**

Метод SHAPE продолжает развиваться. В настоящее время самыми многообещающими подходами являются методы дифференциальный SHAPE [72, 86], SHAPE-MaP [82] и SHAPE-selection [68]. Первый метод использует два реагента (NMIA и 1M6) и позволяет детектировать нуклеотиды, участвующие в неканонических и пространственных взаимодействиях. Метод SHAPE-MaP уникален тем, что положения нуклеотидов, модифицированных реагентом, определяются следующим образом: обратная транскриптаза совершает ошибки в данных позициях, таким образом, продуцируя цепь, не комплементарную исходной. Далее места этих “ошибок” легко вычисляются с помощью

высокопроизводительного секвенирования. Метод SHAPE-selection является развитием технологии SHAPE-seq: реагент ангидрид пропановой кислоты (англ. N-propanone isatoic anhydride - NPIA), взаимодействующий с гидроксильной группой рибоз, также обладает способностью связываться с биотином. Далее с помощью стрептавидина рассматриваемые РНК можно изолировать и, таким образом, избавиться от фонового шума в эксперименте. Кроме того, в 2013 году был разработан метод SHAPE *in vivo*, позволяющий анализировать структуры *in vivo*: метод использует электрофил [85], способный проникать в живые клетки, хорошо растворимый в воде. С помощью данного метода была проанализирована с высокой точностью 5S рибосомальная РНК эмбриональной клеточной линии мыши.

ДМС метод также получил развитие с помощью высокопроизодительного секвенирования [22, 76, 89]. В настоящее время получены полногеномные карты РНК структур для дрожжей [76, 89], человека [76], *Arabidopsis thaliana* [22].

1.5.3 Использование экспериментальных данных при вычислительных подходах к определению структур РНК

Быстро прогрессирующие экспериментальные подходы к определению структур РНК позволяют проверять и совершенствовать существующие вычислительные методы определения структур. Например, программа RNAStructure [55], первоначально разработанная для предсказания вторичных структур РНК на основе термодинамического подхода, использует данные из SHAPE экспериментов при процедуре минимизации энергии. Было продемонстрировано, что использование информации о структуре, полученной в ходе эксперимента, в дополнение к МСЭ, позволяет значительно улучшить предсказание [20].

Тем не менее, высокопроизводительное определение структур РНК накладывает определенные ограничения. Во-первых, существующие

аналитические методы предполагают доступность экспериментальных данных с очень высоким покрытием (должна быть информация о каждом нуклеотиде), однако полногеномные карты РНК (степень структурированности РНК для всех нуклеотидов генома) далеко не всегда обладают подобным разрешением. С другой стороны, применимость экспериментальных данных ограничена масштабированием на весь транскриптом.

Как и в случае с другими экспериментальными методами, результаты пробинга РНК содержат в себе шум. Возникает этот шум по разным причинам. Во-первых, на данный момент химические реагенты не позволяет прямо оценить, спарен конкретный нуклеотид с другим нуклеотидом или нет, а позволяют получить лишь косвенные оценки. На эту оценку, естественно, влияет большое количество факторов, например, при исследовании РНК в живой клетке взаимодействие РНК с белками затрудняет оценку структуры в месте контакта. Кроме того, способность нуклеотида реагировать с тем или иным химическим веществом может зависеть от окружения этого нуклеотида и его взаимного расположения относительно элементов вторичной и третичной структуры РНК. Во-вторых, в результате пробинг-эксперимента мы получаем информацию не об одной молекуле РНК, а взвешенную информацию сразу обо всем ансамбле молекул РНК в клетке или в растворе. Эти факторы необходимо учитывать при анализе экспериментальных данных.

Кроме того, основная сложность при инкорпорировании данных заключается в том, что эксперимент позволяет ответить на вопрос о том, насколько вероятно, что конкретный нуклеотид находится в спаренном состоянии, но не предоставляет информации о том, с каким нуклеотидом он спарен. Одной этой информации недостаточно для того, чтобы определить вторичную структуру РНК. Традиционным подходом является использование информации о спаренности нуклеотидов в качестве ограничения при процедуре минимизации энергии: нуклеотид, помеченный как неспаренный (т.е. доступный для химических модификаций) не должен оказаться в стебле вторичной структуры РНК. При этом различают жесткие (англ. hard) и мягкие (англ. soft)

ограничения. При использовании жестких ограничений нуклеотид, помеченный как неспаренный, должен гарантированно оказаться неспаренным в полученной в результате МСЭ структуре, и наоборот, нуклеотид, помеченный как спаренный, должен оказаться спаренным. Такой подход был применен Мэтью и коллегами, которые модифицировали алгоритм Зукера, добавив возможность введения жестких ограничений [56]. Однако эксперименты с использованием химических реагентов и ферментов вероятностны по своей природе: крайне редко можно с полной уверенностью говорить о спаренности или неспаренности отдельного нуклеотида. Таким образом, вместо жестких ограничений лучше использовать мягкие ограничения, которые позволяют использовать экспериментальные данные в качестве дополнительной информации при нахождении МСЭ.

Однако существует и другие подходы. Кваррье в 2010 году показал, что использование экспериментальных данных для выбора структур из ансамбля Больцманоских взвешенных структур позволяет лучше предсказывать структуры РНК, делая предсказание более устойчивым к шуму [69].

1.5.4 Программа RNAStructure

Программа RNAStructure [52] позволяет учитывать экспериментальные данные при поиске структуры РНК с минимальной энергией. Функция энергии в программе RNAStructure модифицируется с учетом экспериментальных данных: к свободной энергии прибавляется псевдо-свободная энергия, определяемая из реакциоспособности нуклеотидов в эксперименте SHAPE (англ. SHAPE reactivity). Данный подход основан на наблюдении, что реакциоспособность в эксперименте SHAPE сильно коррелирует с локальной гибкостью нуклеотидной цепи [58] и, таким образом, вероятностью того, что данный участок находится в одноцепочечном состоянии.

Псевдо-свободная энергия SHAPE описывается следующим уравнением:

$$\Delta G_{SHAPE}(i) = m \cdot ln[SHAPE\ reactivity(i) + 1] + b$$

Эмпирические параметры m и b служат для масштабирования реакциоспособности SHAPE в псевдо-свободную энергию. Свободный член b представляет вклад в энергию нуклеотида, чья реакциоспособность в эксперименте была оценена как ноль. Коэффициент наклона m является мерой штрафа за включение нуклеотидов с высокой реакциоспособностью в число спаренных.

Оптимальные значения для параметров b и m были получены в ходе анализа вторичной структуры 23S рРНК *E.coli*: b=-0.8 kcal/mol и m=2.6 kcal/mol.

Псевдо-свободная энергия ΔG_{SHAPE} вычисляется для каждого нуклеотида, а не для пары нуклеотидов: нуклеотиды с высокой реакциоспособностью получают положительные значения псевдо-свободной энергии, а нуклеотиды с низкой реакциоспособностью получают отрицательные значения псевдо-свободной энергии. ΔG_{SHAPE} учитывается при подсчете энергии только в случае спаренных нуклеотидов. Общая формула для энергии представляет собой сумму свободной энергии и псевдо-свободной энергии SHAPE.

Дальнейший анализ подобного метода учета экспериментальных данных подтвердил эффективность включения экспериментальных данных в процесс термодинамической оптимизации [88]. Однако уровень улучшения по сравнению с МСЭ зависит от последовательности и коррелирует с первоначальной точностью МСЭ. Было показано, что пары нуклеотидов полученные в МСЭ с высокой точностью обычно сохраняются и при добавлении экспериментальных данных, в то время как пары с низкой точностью – нет.

Другой метод использования экспериментальных данных при сворачивании РНК был предложен Вэшитлом в 2011 году [99]. Подход базируется на предположении, что и экспериментальные данные, и термодинамические энергетические параметры представляют собой неидеальную, шумную

аппроксимацию физической реальности. Авторы вводят вектор, минимизирующий взвешенную сумму пертурбаций энергий и отличий между измеренными и предсказанными вероятностями нуклеотидом находиться в спаренном состоянии:

$$F(\vec{\epsilon}) = \sum_{\mu} \frac{\epsilon_{\mu}^2}{\tau_{\mu}^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \cdot (p_i(\vec{\epsilon}) - q_i)^2 \ \longrightarrow min$$

Данный подход позволяет также улучшить точность предсказания вторичной структуры РНК.

Таким образом, экспериментальные данные являются крайне полезным дополнительным источником информации при определении вторичной структуры РНК, а также могут быть использованы при полногеномном поиске структурированных РНК.

Глава 2. Свойства экспериментальных данных

При экспериментальном определении вторичной структуры РНК мы хотим получить информацию о том, насколько вероятно, что конкретный нуклеотид находится в спаренном состоянии, и использовать эту информацию как дополнительный источник при поиске оптимальной вторичной структуры РНК. Для того, чтобы получить такую информацию, необходимо преобразовать данные, которые мы получаем в разных типах пробинг-экспериментов, то есть получить некоторую универсальную оценку. В этой главы мы рассмотрим, как устроены данные, получаемые в разных экспериментах, и насколько они согласуются со вторичной структурой РНК.

Кроме того, в зависимости от условий проведения эксперимента (например, *in vivo* или *in vitro*) экспериментально определенная реактивность нуклеотидов и, следовательно, уровень структурированности РНК в целом может значительно отличаться. В данной главе мы рассмотрим различия между результатами эксперимента ДМС *in vivo* или *in vitro* и сделаем вывод о возможных причинах этих различий.

2.1 Материалы и методы

2.1.1 PARS эксперимент

Для анализа мы использовали данные, полученные в работе [98]. Наиболее полный набор данных был доступен для условий «native deproteinized dataset», то есть эксперимента, в ходе которого РНК выделяли и использовали денатурирующие условия для очистки от белков. В этом случае структура РНК

наиболее близка к структуре РНК в клетке. Далее в эксперименте производили обработку S1 или V1 нуклеазами и секвенировали образцы с помощью Illumina's Hi-Seq. В данной работе была проведена стандартная процедура контроля качества образцов, картирования на транскриптом (сборка hg19, версия Gencode v12). Мы отобрали только транскрипты со средним покрытием, равным как минимум одно чтение на позицию. Далее мы применили второй фильтр, требуя, чтобы покрытие чтениями в обоих экспериментах было как минимум 10 чтений на позицию. В результате мы получили 558 транскриптов для дальнейшего рассмотрения (Рис. 2.1).

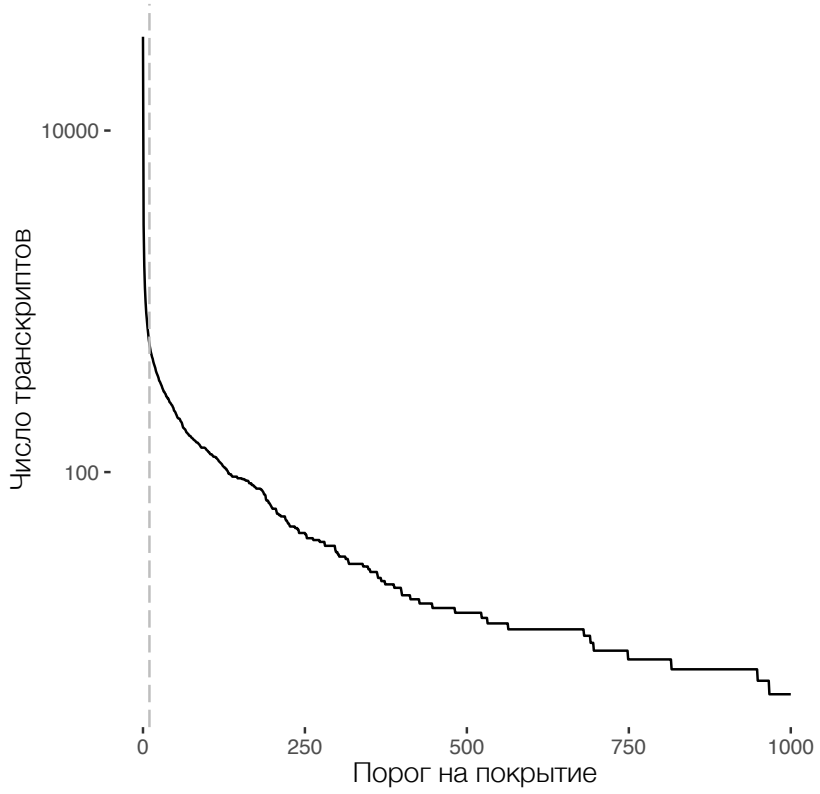


Рисунок 2.1. Количество транскриптов, в зависимости от порога на покрытие в эксперименте PARS.

Таким образом, для каждого нуклеотида в каждом из отобранных транскриптов мы знаем количество чтений, заканчивающихся на данной позиции

в эксперименте с V1 нуклеазой ($V1_i$) и в эксперименте с S1 нуклеазой ($S1_i$). Мы можем вычислить PARS значение для каждой позиции i по формуле:

$$PARS_i = \log_2 \frac{V1_i + 5}{S1_i + 5}$$

В данной формуле используются псевдокаунты для того, чтобы в случае слишком низких покрытий в одном или в обоих экспериментах не получать числитель и/или знаменатель, равный нулю.

Чтобы оценить, какие значения PARS имеют спаренные и неспаренные нуклеотиды, необходимы данные по вторичным структурам РНК, полученные из независимых экспериментов или вычислений. Первый вариант – использовать вторичные структуры для функциональных некодирующих РНК, предсказанные по консервативности вторичной структуры, например, с помощью программы Infernal [63].

Для оценки качества данных мы вычисляли чувствительность и специфичность определения в эксперименте спаренных нуклеотидов. В зависимости от того, реактивность была меньше или больше выбранного порога, нуклеотиды были отмечены как «предсказанные как спаренные» (обозначим их P, от англ. positives) или «предсказанные как неспаренные» (обозначим их N, от англ. negatives). Кроме того, все нуклеотиды в последовательности делили на те, для которых статус в эксперименте совпадает с их статусом во вторичной структуре (то есть верно предсказанные, обозначим из T, от англ. true) и те, для которых статусы не совпадают (неверно предсказанные, обозначим из F, от англ. false). Таким образом, все нуклеотиды оказались поделены на 4 группы: TP (и эксперимент, и структура свидетельствуют о том, что данный нуклеотид спарен), TN (и эксперимент, и структура свидетельствуют о том, что данный нуклеотид не спарен), FP (в структуре нуклеотид не спарен, но эксперимент свидетельствует о том, что нуклеотид спарен) и FN (наоборот, в структуре нуклеотид спарен, а эксперимент свидетельствует о том, что нуклеотид не спарен). Далее на основании этих величин мы вычислили специфичность и чувствительность:

$$\text{специфичность} = \frac{TN}{TN + FP}$$

$$\text{чувствительность} = \frac{TP}{TP + FN}$$

ROC-кривые были построены на основании вычисленных значений чувствительности (ось OX) и FPR (от англ. false positive rate; ось OY). FPR вычисляется следующим образом:

$$FPR = \frac{FP}{FP + TN}$$

Для подтверждения результатов на большем наборе данных, включающем РНК, для которых была недоступна достоверная вторичная структура, мы использовали программу RNAplfold [36]. Данная программа позволяет рассчитать вероятность каждого нуклеотида находиться в спаренном состоянии. Используя порог на минимальную вероятность, при которой нуклеотид считается спаренным, можно также разделить нуклеотиды на спаренные и неспаренные.

2.1.2 SHAPE эксперимент

В ходе SHAPE эксперимента химическое вещество модифицирует нуклеотиды, находящиеся в более гибких областях молекулы РНК. Далее используется один из методов определения позиций, подвергшихся модификации. Современные методы используют тот факт, что обратная транскриптаза или останавливается на модифицированных позициях, или совершает ошибки, а следовательно позиции модификации можно отследить при дальнейшем секвенировании. В ходе SHAPE-seq эксперимента подсчитывают количество чтений, заканчивающихся на данной позиции в эксперименте с реагентом и в контроле, и, используя модель и нормализацию, рассчитывают реактивность. Реактивность нуклеотидов скоррелирована с их склонностью к образованию

вторичной структуры, однако, как и в случае других экспериментов, не даёт однозначного ответа о статусе нуклеотида.

В настоящее время наиболее качественными данными являются данные по исследованию отдельных функциональных молекул РНК *E.coli*: глициновый, адениновый и цикло-ди-GMP рибопереключатели, 5S рРНК и фенилаланин тРНК [40].

2.1.3 Извлечение вероятностной информации из распределений реактивностей

Обозначим вероятность наблюдать реактивность нуклеотида r (PARS значение или другая мера, соответствующая статусу нуклеотида) при условии, что данный нуклеотид находится в спаренном состоянии (англ. *paired*) как $P(r|paired)$, а вероятность наблюдать реактивность нуклеотида r при условии, что данный нуклеотид находится в неспаренном состоянии (англ. *unpaired*) как $P(r|unpaired)$. Вероятности $P(r|paired)$ и $P(r|unpaired)$ зависят от типа эксперимента и отражают способность данного эксперимента различать спаренные и неспаренные нуклеотиды. Самый простой способ рассчитать $P(r|paired)$ и $P(r|unpaired)$ – использовать распределения реактивностей нуклеотидов для известных структур.

С помощью пакета R *fitdistrplus* мы аппроксимировали распределения SHAPE реактивностей и распределения PARS значений.

Количественная характеристика, отражающая структурные свойства нуклеотида, может быть выведена как логарифм отношения правдоподобия:

$$L(r) = \log \frac{P(r|paired)}{P(r|unpaired)}$$

Такой подход позволяет извлекать вероятностную информацию $L(r)$ на основании реактивности для любого типа эксперимента.

Для контроля адекватности данного преобразования мы также определяли статус нуклеотида не по известной вторичной структуре, а с помощью программ RNAplfold [36] или NUPACK [24].

2.1.4 Данные ДМС-пробинга

Для сравнения структурированности РНК в условиях *in vivo* или *in vitro* мы анализировали данные из статьи Рускина [76], полученные в результате эксперимента по пробингу структур РНК в условиях *in vivo*, *in vitro* и денатурирующих условиях, на клетках дрожжей S. cerevisiae и клеточных линиях человека (фибробласты). Данные для анализа скачаны из базы данных GEO (http://www.ncbi.nlm.nih.gov/geo/), идентификатор GSE45803.

Первый шаг анализа – вычисление уровня покрытия чтениями в зависимости от позиции нуклеотида, так как разница в покрытии может влиять на значение структурированности.

Далее для оценки структурированности РНК мы использовали меру структурное значение, отражающую склонность нуклеотида A/C находиться в спаренном состоянии:

$$-\log\left(\frac{r1/R1 + 2/(R1+R2)}{r2/R1 + 2/(R1+R2)}\right),$$

где r1 и r2 – число чтений, покрывающих нуклеотид в эксперименте (*in vivo* или *in vitro*) и в контроле соответственно, R1 и R2 – число чтений, покрывающих рассматриваемый фрагмент. Структурное значение фрагмента определено как среднее структурное значение нуклеотидов, входящих в данный фрагмент. Высокие структурные значения отвечают структурированным фрагментам мРНК.

При анализе человеческих клеточных линий мы рассматривали транскрипты с покрытием больше 5 чтений на позицию, которые мы разбили на

непересекающиеся фрагменты, содержащие 50 A/C нуклеотидов, и рассчитывали *in vivo*/*in vitro* структурные значения относительно денатурированного контроля для каждого фрагмента.

2.2 Результаты и обсуждение

2.2.1 Свойства экспериментальных данных

2.2.1.1 PARS эксперимент

PARS значение должно отражать склонность нуклеотида находиться в спаренном состоянии во вторичной структуре. Рассмотрим, какие PARS значения имеют нуклеотиды, отмеченные в структурах как спаренные и неспаренные. С помощью программы Infernal [63] мы получили структуры для пяти классов некодирующих РНК: малые ядрышковые РНК (мякРНК) двух типов, малые ядерные РНК, Vault РНК и РНК, локализующиеся в тельцах Кахаля (англ. Small Cajal body-specific RNAs - scaRNAs, далее скаРНК). Эти РНК обладают разной структурированностью, например, мякРНК типа C/D не имеет консервативной функциональной вторичной структуры, а то время как РНК, локализующиеся в тельцах Кахаля, включают в себя важный для функции стебель. Мы построили распределения PARS значений для высоко покрытых в эксперименте позиций (число чтений не менее 50 на нуклеотид) для пяти классов РНК, рисунок 2.2. Нуклеотиды, образующие пары в структурах, полученных с помощью программы Infernal, мы считали “спаренными”, остальные нуклеотиды – “неспаренными”.

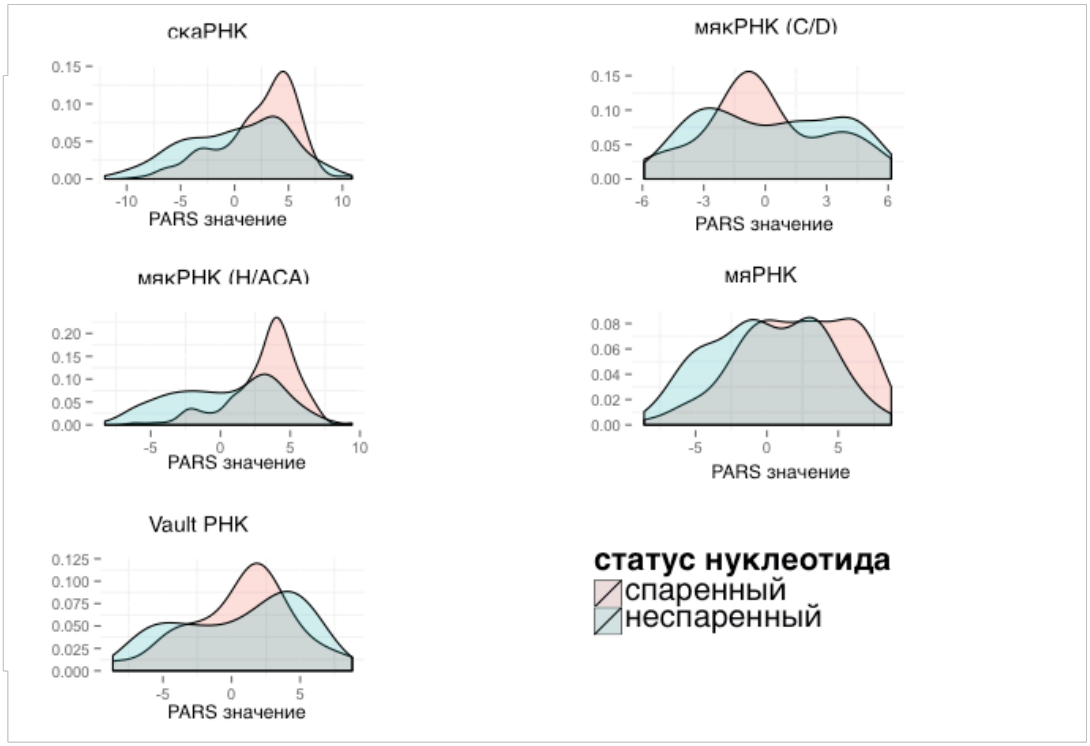


Рисунок 2.2. Распределения PARS значений для спаренных и неспаренных нуклеотидов разных классов некодирующих РНК. Порог на покрытие 50 чтений на позицию.

В идеальном случае можно было бы ожидать, что распределения PARS значений для спаренных и неспаренных нуклеотидов будут хорошо разделяться, однако для большинства рассмотренных РНК это не так. Это объясняется низким качеством эксперимента и высоким уровнем шума. Необходимо также заметить, что для РНК, обладающих функциональной вторичной структурой, распределения имеют разные медианы, более того, этот эффект зависит от порога на покрытие: чем выше порог на покрытие, тем лучше разделяются распределения (то есть мы смотрим на более достоверные данные), но мы, естественно, ограничены в повышении порога на покрытие количеством доступных данных.

Мы предположили, что можно провести порог на PARS значение такой, что нуклеотиды с PARS значением меньше данного порога можно рассматривать как

неспаренные, а выше – как спаренные. Для того, чтобы выбрать оптимальный порог и оценить, насколько хорошо согласуются данные по реактивности нуклеотидов в эксперименте PARS и вторичные структуры функциональных некодирующих РНК, мы вычислили чувствительность и специфичность определения в эксперименте спаренных нуклеотидов.

Далее на основании чувствительности и специфичности для разных порогов мы построили ROC-кривые, позволяющие оценить качество классификации нуклеотидов на спаренные и неспаренные для разных порогов на PARS значение (Рис. 2.3).

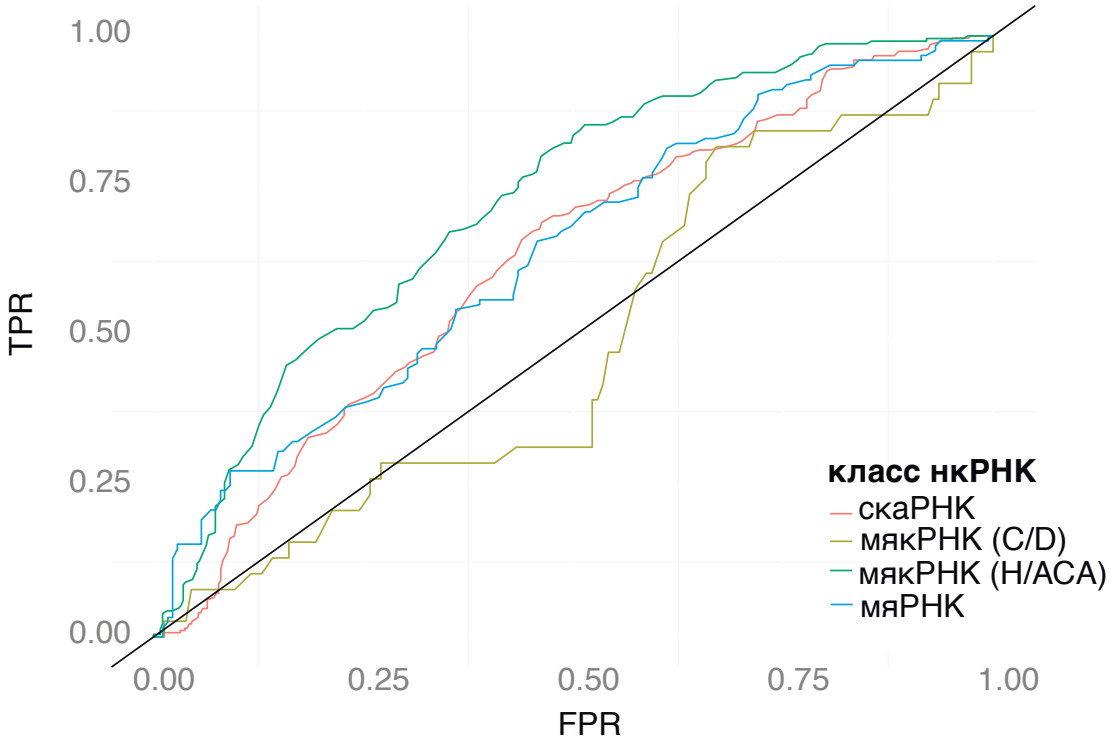


Рисунок 2.3. ROC-кривые классификации нуклеотидов на спаренные и неспаренные на основании их PARS значений, для нескольких классов нкРНК.

ROC-кривые демонстрируют, что разделение нуклеотидов на спаренные и неспаренные на основании некоторого порога приводит к весьма посредственным результатам для некоторых классов функциональных РНК. Для количественной оценки данного наблюдения мы вычислили площадь под кривой (AUC, от англ. area under curve), получив значения от 0.72 до 0.48. Таким образом, было показано, что разделение нуклеотидов на спаренные и неспаренные на основании фиксированного порога дает слабые результаты и необходима разработка более сложной модели.

Чтобы дополнительно подтвердить данные результаты на большем наборе данных, включающем РНК, для которых была недоступна достоверная вторичная структура, мы использовали программу RNAplfold [36]. Для различных порогов (от 0.7 до 0.95) мы получили результаты, сходные с результатами на основании известных структур.

2.2.1.2 SHAPE эксперимент

Как и в случае PARS эксперимента, реактивность нуклеотидов должна отражать склонность нуклеотида находиться в спаренном состоянии во вторичной структуре. Рассмотрим реактивности нуклеотидов, находящихся в спаренном и неспаренном состояниях.

Мы использовали данные для отдельных функциональных молекул РНК *E.coli,* для которых мы провели анализ, аналогичный анализу PARS данных. Распределения реактивностей спаренных и неспаренных нуклеотидов представлены на рисунке 2.4.

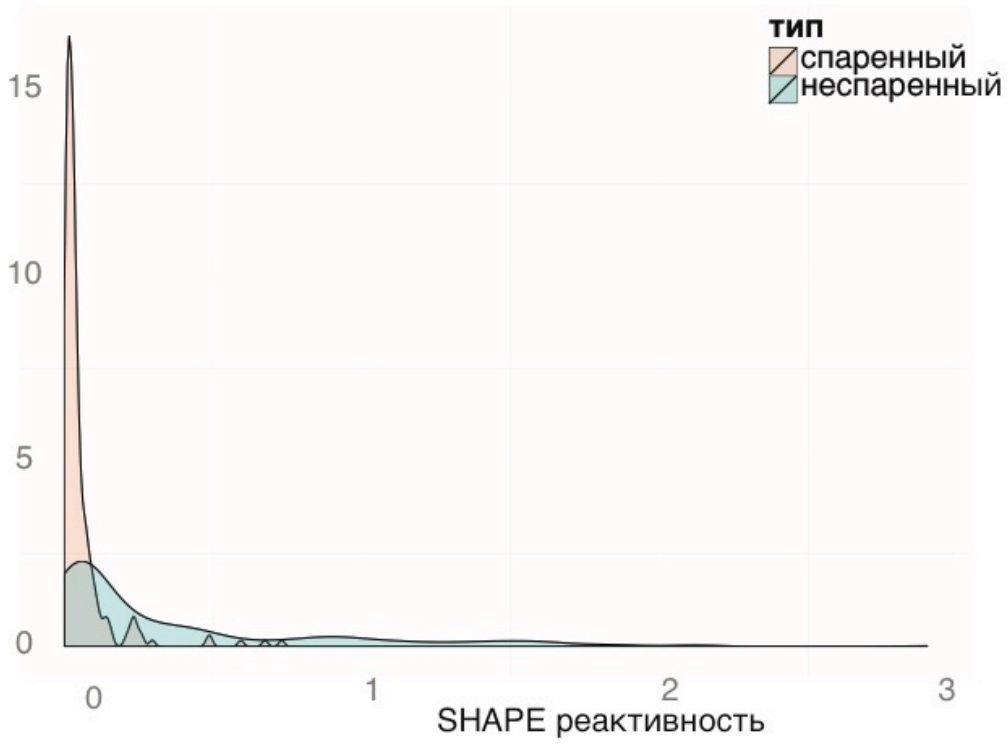


Рисунок 2.4. Распределения SHAPE реактивностей для спаренных и неспаренных нуклеотидов нескольких функциональных некодирующих РНК.

Несмотря на то, что данные SHAPE позволяют получить более достоверную информацию о статусе нуклеотидов, данные всё равно содержат шум, осложняющий разделение нуклеотидов на спаренные и неспаренные. Так же, как и в случае с PARS данными, использование порога не позволяет достигнуть высокого соотношения чувствительности и специфичности (рис. 2.5).

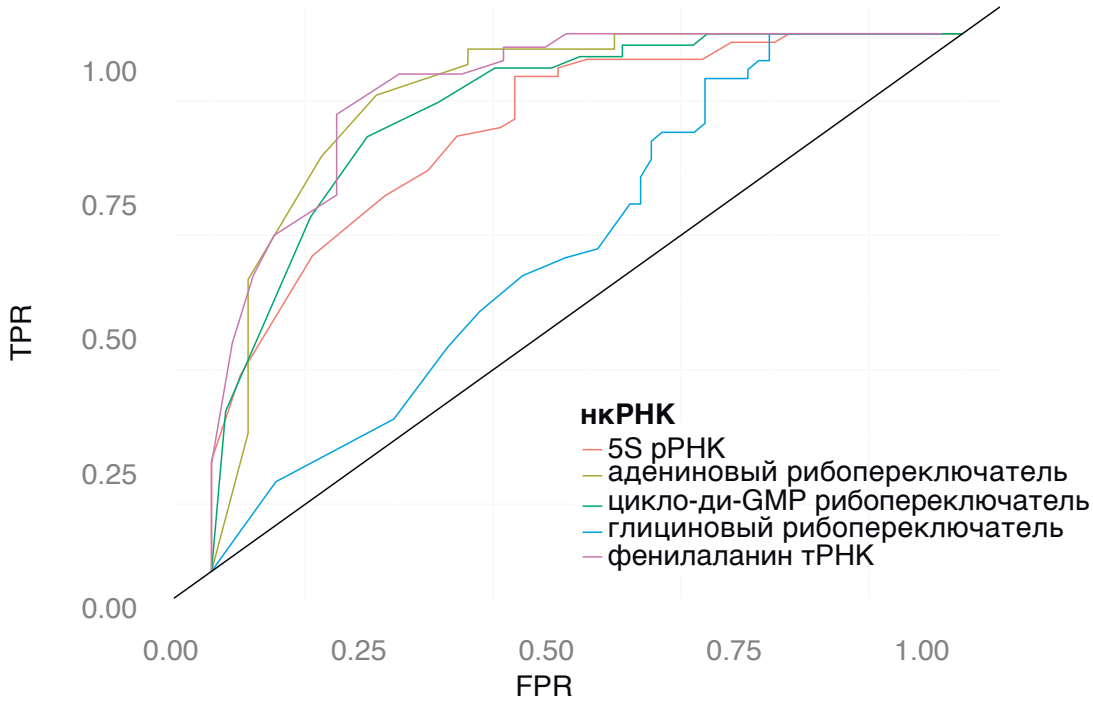


Рисунок 2.5. ROC-кривые классификации нуклеотидов на спаренные и неспаренные на основании их SHAPE реактивностей для нескольких функциональных некодирующих РНК.

Одним из источников шума при определении структурированности РНК с помощью химических реагентов является контекст: нуклеотиды в зависимости от крайнего или центрального положения в элементе вторичной структуры и длины этого элемента обладают разной реактивностью. Рис. 2.6 демонстрирует, что нуклеотиды, находящиеся на границах структурных элементов имеют сдвинутую реактивность. Например, крайние нуклеотиды в петлях имеют более низкую реактивность, чем нуклеотиды, удаленные от краёв, то есть крайние нуклеотиды менее реактивны, то есть менее склонны находиться в неспаренном состоянии. С другой стороны, крайние нуклеотиды в стеблях обладают большей реактивностью по сравнению с нуклеотидами в стеблях, удаленными от краёв, то есть такие нуклеотиды являются более реактивными и «гибкими».

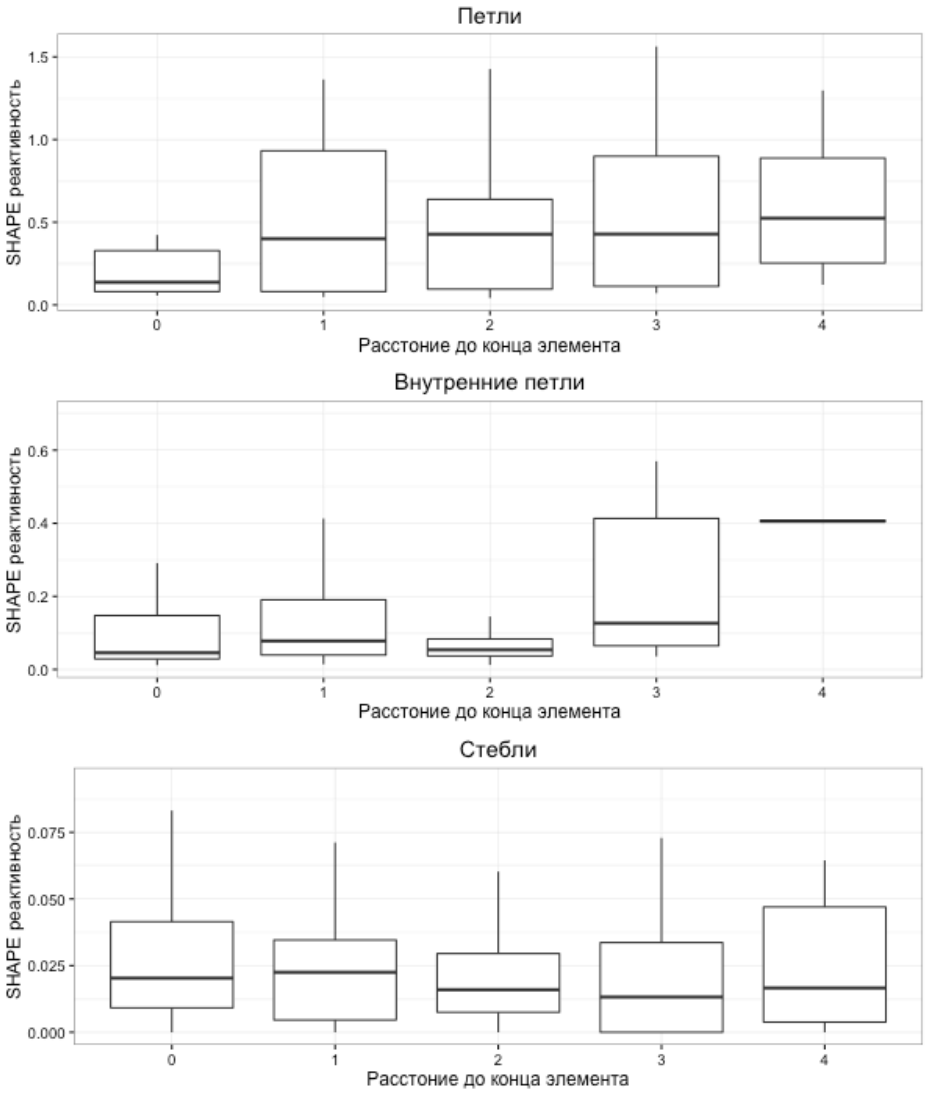


Рисунок 2.6. Неравномерность реактивности в зависимости от положения в элементе вторичной структуры.

2.2.2 Преобразование данных по реактивности

С помощью R мы аппроксимировали распределения SHAPE реактивностей распределением экстремальных значений для спаренных нуклеотидов и экспоненциальным распределением для неспаренных нуклеотидов. Распределения PARS значений были аппроксимированы с помощью двух нормальных распределений. Таблица 2.1 содержит значения параметров распределений.

Таблица 2.1. Параметры распределений, аппроксимирующих распределения реактивностей спаренных и неспаренных нуклеотидов, данные PARS и SHAPE.

Эксперимент	Тип	Распределение	Параметр	Значение параметра
SHAPE	спаренные нуклеотиды	экстремальных значений	μ σ χ	0.0191 0.0195 0.646
	неспаренные нуклеотиды	экспоненциальное	rate	2.729
PARS	спаренные нуклеотиды	нормальное	mean sd	2.317911 3.0326
	неспаренные нуклеотиды	нормальное	mean sd	0.3538203 3.0438

На основании полученных распределений для спаренных и неспаренных нуклеотидов мы рассчитали $L(r)$ и построили кривые зависимости $L(r)$ от PARS значений и SHAPE реактивностей (рис. 2.7). Зависимость $L(r)$ от реактивности адекватно отражает статус нуклеотида в эксперименте. Так, например, при PARS значении около нуля $L(r)$ примерно равен 1, что соответствует отсутствию информации о склонности нуклеотида к образованию вторичной структуры: нуклеотид с равной вероятностью может оказаться как спаренным, так и не спаренным. При PARS значении около 2 нуклеотид примерно в 5 раз более

вероятно находится в спаренном состоянии. Таким образом, мы переходим к оценке правдоподобия $L(r)$, которая является универсальной для всех видов экспериментов.

В случае определения статуса нуклеотида не по известной вторичной структуре, а с помощью программ RNAplfold [36] или NUPACK [24], для различных порогов и двух разных программ мы получили сходные кривые (рис. 2.7), что говорит об устойчивости разделения нуклеотидов на спаренные и неспаренные и адекватности подхода в целом.

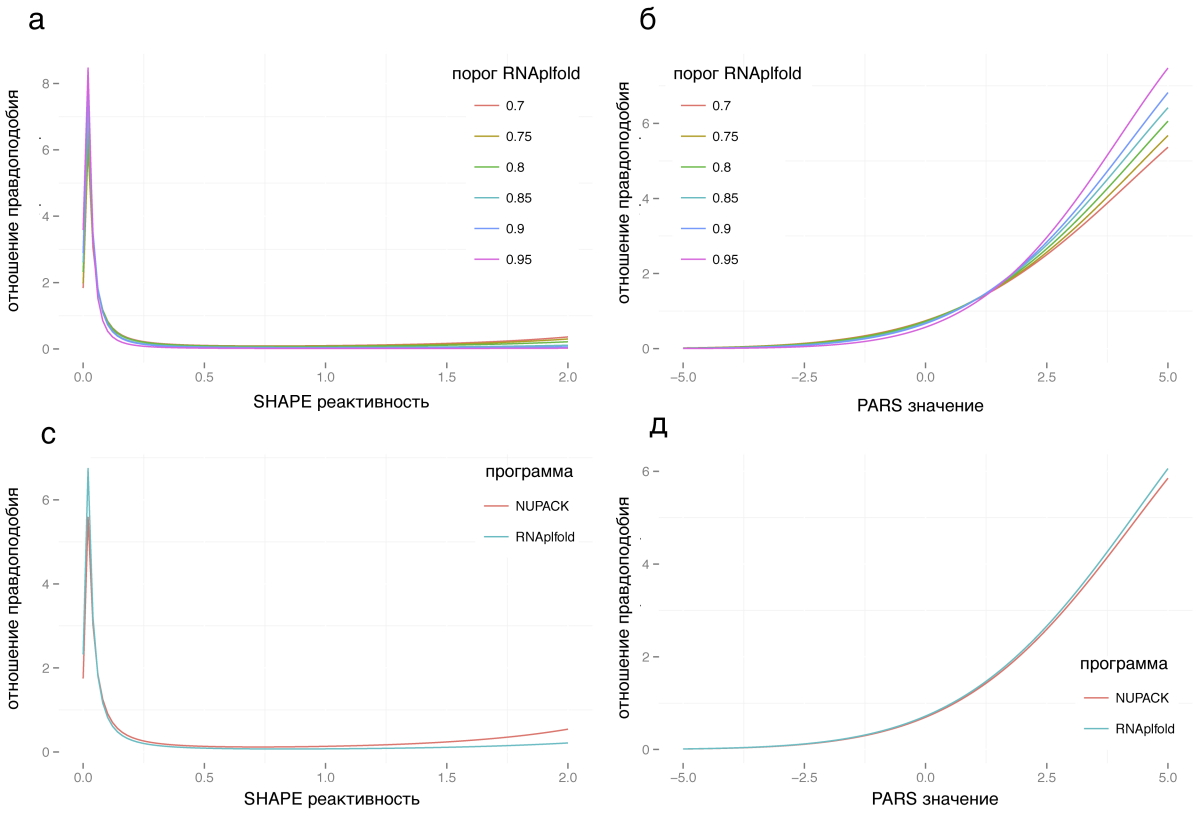


Рисунок 2.7. Отношение правдоподобия в зависимости от SHAPE реактивностей (а, с) и PARS значений (б, д), статус нуклеотидов определен с помощью программ RNAplfold (с разными порогами) и NUPACK.

2.2.3 Сравнение профилей *in vitro* и *in vivo*

На основании данных из статьи Рускина [76] мы провели сравнение структурированности транслируемых и нетранслируемых регионов мРНК, в разных условиях. Все эксперименты в данной статье были проведены с помощью метода ДМС-пробинга.

На первом шаге анализа мы сравнили уровень покрытия чтениями в зависимости от позиции нуклеотида относительно начала и конца транслируемой области. На рисунке 2.8 представлена зависимость покрытия чтениями от позиции на мРНК. 3’ нетранслируемые области (3' НТО) и 5’ нетранслируемые области (5' НТО) покрыты значительно меньше, чем кодирующая область мРНК, при этом покрытие примерно одинаковое для всех условий эксперимента. Так как структурированность связана с покрытием, далее при сравнении нетранслируемых областей с транслируемыми мы смотрим только на отношение экспериментов друг другу, не на абсолютные значения в одном эксперименте.

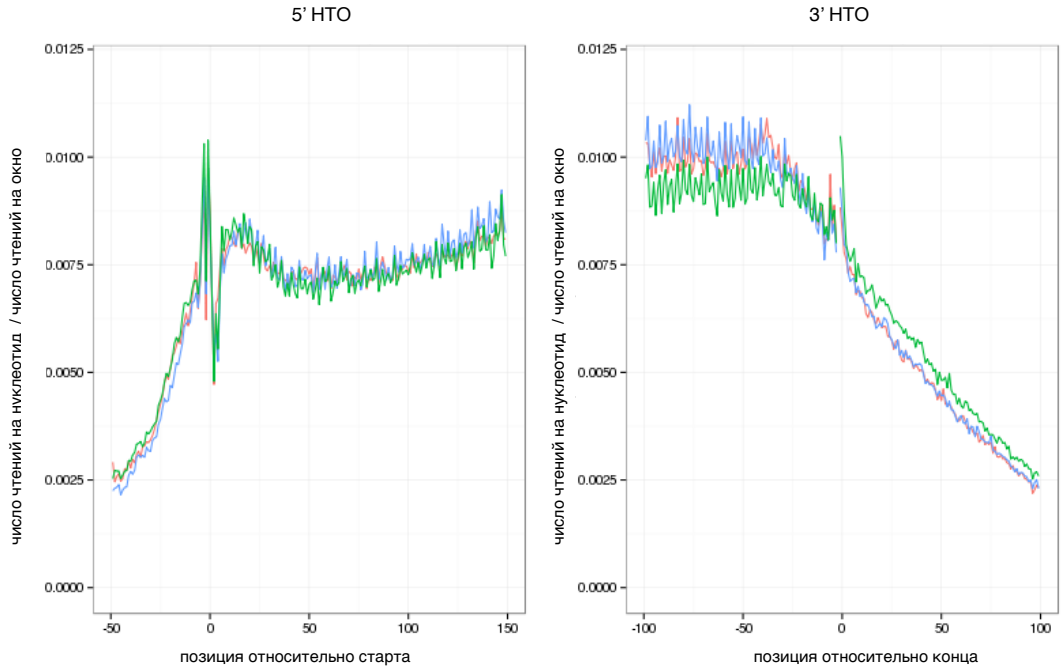


Рисунок 2.8. Зависимость покрытия чтениями от позиции нуклеотида на мРНК. По оси абсцисс – позиция мРНК, позиции выровнены относительно старта трансляции для левого рисунка и относительно стоп-кодона для правого рисунка. По оси ординат – число чтений, попавших на данную позицию, деленное на число чтений в рассматриваемом окне. Данные для ДМС-пробинга на дрожжах, [76]**.**

Далее мы рассчитали структурные значения для мРНК в условиях *in vitro* и *in vivo* относительно денатурированного контроля. Анализ распределения структурных значений по мРНК в *in vivo* эксперименте относительно денатурированного контроля показывает, что нетранслируемые области более структурированы относительно кодирующих областей (рис. 2.9 а и б). Для контроля того, что мРНК были верно выровнены относительно стартов трансляции и стоп-кодонов, мы также провели анализ периодичности и выявили три-периодичность в транслируемых областях, в отличие от нетранслируемых областей (рис. 2.9 с).

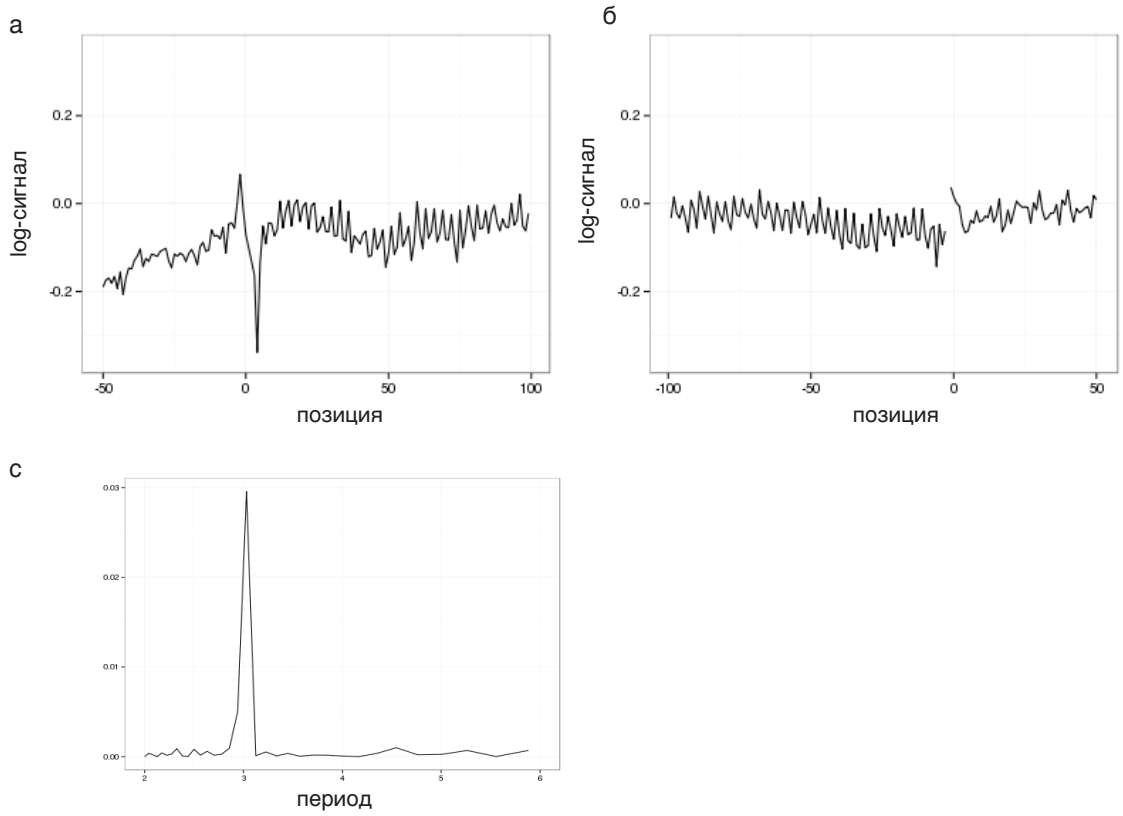


Рисунок 2.9. Структурное значение в зависимости от позиции на мРНК, *in vivo* эксперимент относительно денатурированного контроля. Данные для ДМС-пробинга на дрожжах, [76]**.**

В отличие от эксперимента *in vivo*, распределения структурных значений по мРНК в *in vitro* эксперименте относительно денатурированного контроля показывает, что нетранслируемые области менее структурированы относительно кодирующих областей (рис 2.10 а и б).

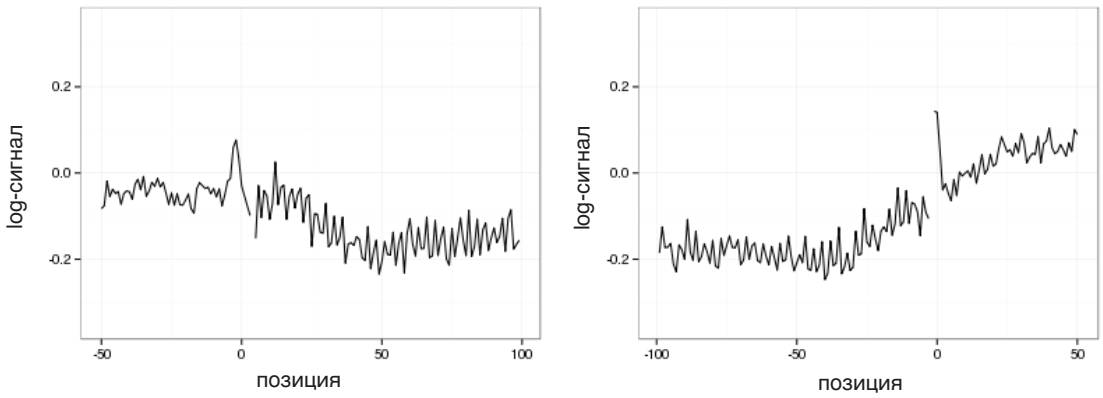


Рисунок 2.10. Структурное значение в зависимости от позиции на мРНК, *in vitro* эксперимент относительно денатурированного контроля. Данные для ДМС-пробинга на дрожжах, [76]**.**

Таким образом, сравнение транслируемых и не транслируемых областей в двух типах экспериментов показывает, что механизм, отвечающий за раскручивание мРНК *in vivo,* действует только на транслируемые области, и таким образом, вероятно, связан с трансляцией. При этом в случае эксперимента с неработающим АТФ эффект пропадает, что лишний раз подтверждает предположение о влиянии трансляции на раскручивание структуры *in vivo*: для трансляции, как и для любого другого активного процесса в клетке, необходимы молекулы АТФ.

Далее для дальнейшего подтверждения полученного результата мы провели следующий анализ на человеческих клеточных линиях. Мы рассчитали *in vivo*/*in vitro* структурные значения относительно денатурированного контроля для каждого фрагмента транскрипта, содержащего 50 A/C нуклеотидов. Как и в случае эксперимента и анализа на дрожжах, кодирующие фрагменты демонстрируют раскручивание *in vivo* относительно *in vitro*, в отличие от нетранслируемых областей (рис 2.11).

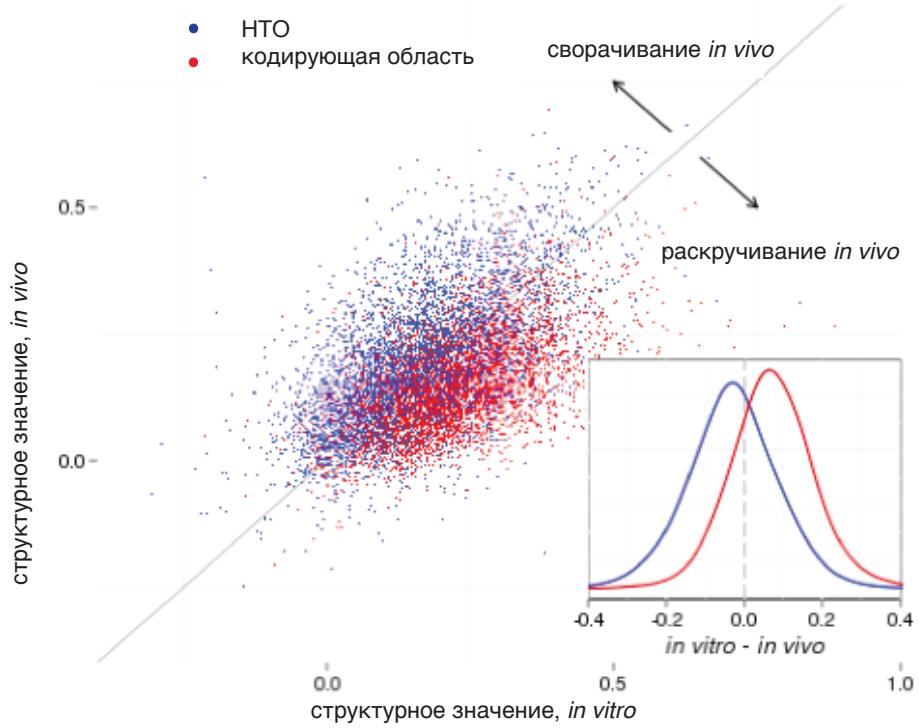


Рисунок 2.11. Точечный график, демонстрирующий различия в структурных значениях для in vivo и in vitro состояний, для кодирующих и некодирующих областей. Данные для ДМС-пробинга на фибробластах человека, [76]**.**

2.3 Выводы к главе 2

Мы провели анализ экспериментальных данных различных типов, показав, что экспериментальные данные содержат в себе вероятностную информацию, позволяющую судить о структурном статусе нуклеотидов. Мы ввели меру, отражающую вероятность нуклеотида находиться в спаренном состоянии, которую можно использовать для включения экспериментальных данных в энергетическую модель при предсказании структур РНК.

Кроме того, мы провели сравнение профилей данных ДМС *in vivo* и *in vitro* и показали, что кодирующие области мРНК являются менее структурированными в клетке вследствие процесса трансляции. Таким образом, в зависимости от условий эксперимента (*in vivo* или *in vitro*, присутствия или отсутствия белков и так далее) структурированность РНК может значительно отличаться, и этот факт необходимо учитывать при включении экспериментальных данных в предсказание структур РНК и поиске структурированных элементов РНК.

Глава 3. Поиск структурированных участков РНК

3.1 Материалы и методы

3.1.1 Поиск структурированных сегментов в ортологичных последовательностях

Чтобы оценить, насколько эффективно мы можем предсказывать структурированные РНК-элементы генома на основе сравнительно-геномного поиска, не использующего собственно выравнивание, а базирующегося только на факте ортологичности последовательностей, мы реализовали алгоритм, определяющий меру структурированности участка в “скользящем окне” путем подсчета вероятностей нуклеотидов находиться в спаренном состоянии и сравнения с фоновой моделью.

Веса спаривания нуклеотидов ξ вычисляли с помощью программы RNAplfold [36]. Максимизацию суммарной меры спаривания нуклеотидов $\xi_k[i, j]$ данного участка [i, j] последовательности k проводили с использованием стандартного алгоритма Нуссинов [64] с весами, полученными на первом шаге. Степень структурированности данного участка множественного выравнивания вычисляли как среднюю для всех последовательностей $x[i,j]=avr_k(\xi_k[i, j])$. Важно отметить, что здесь мы не использовали выравнивание как таковое, а только факт, что участки геномов сопоставлены. С учетом полученной степени структурированности применяли алгоритм поиска нкРНК: геномное выравнивание сканировали “скользящим окном” фиксированного размера, в каждом окне вычисляли структурированность множественного выравнивания. Для каждого участка вычисляли значение структурированности A:

$$A(i,j) = \frac{x[i,j] - E}{\sigma}$$

где x[i, j] – степень структурированности данного окна, E – математическое ожидание степени структурированности, σ – стандартное отклонение степени структурированности. В качестве оценки E и σ применяли значения, полученные в результате полногеномного анализа. Полученное значение структурированности использовали в дальнейшем статистическом анализе.

Чтобы вычислить долю ложных предсказаний (FDR, от англ. False Discovery Rate), использовали случайную модель, построенную следующим образом. Все рассмотренные выравнивания разбивали на короткие сегменты длиной 30 нуклеотидов. Далее колонки выравниваний в этих сегментах случайным образом перемешивали, и все вычисления проводили для этого набора выравниваний. Такой подход, с одной стороны, не сохраняет вторичные структуры, а, с другой, – сохраняет локальный уровень консервативности. FDR рассчитывался по стандартной формуле Бенджамини-Хочберга [6]:

$$FDR(x) = \frac{N \cdot F(x)}{n(x)}$$

где N – общее число наблюдений, n(x) – их число со значением веса (A), превышающем x, F(x) – ожидаемая доля числа наблюдений со значением A, превышающем x, оцененная из случайной модели.

Анализ проводили на следующих геномах рода *Drosophila* (идентификатор сборки по UCSC Genome Browser указан в скобках): *D. pseudoobscura* (dp4), *D. ananassae* (droAna3), *D. erecta* (droEre2), *D. grimshawi* (droGri2), *D. mojavensis* (droMoj3), *D. persimilis* (droPer1), *D. sechellia* (droSec1), *D. simulans* (droSim1), *D. virilis* (droVir3), *D. willistoni* (droWil1), *D. yakuba* (droYak2). Использование геномов разного уровня сходства позволяет предсказывать как консервативные по последовательности участки, так и разошедшиеся участки со стабильной вторичной структурой.

Поскольку важным критерием наличия функциональной вторичной структуры мы считали ее консервативность, то для анализа использовали только

те участки геномов, которые входили во множественные геномные выравнивания MULTIZ [7].

Мы рассматривали сегменты выравнивания, содержащие фрагменты геномов как минимум из 6 видов (включая *D. melanogaster*), а также обладающие степенью идентичности не менее 40% в “скользящем окне” 100 нуклеотидов.

Наши наблюдения мы сравнили с различными классами нкРНК. Использованы наборы: из 237 микроРНК *D. melanogaster* из miRBase [41]; из 279 тРНК *D. melanogaster* из Genomic tRNA Database [13]; из 286 мякРНК *D. melanogaster* из snoRNALBMEdb [46] и из 48 мяРНК *D. melanogaster* из аннотации UCSC Genome Browser.

3.1.2 Псевдо-свободная энергия

В разделе 2.2.2 мы рассмотрели преобразование реактивности нуклеотида в логарифм отношения правдоподобия L_i, отражающий склонность данного нуклеотида i находиться в спаренном состоянии. Далее мы вводим дополнительный член $\Delta G'_{ij}$, отражающий склонность пары нуклеотидов i и j находиться в спаренном состоянии:

$$\Delta G'_{ij} = RT \cdot (L_i + L_j),$$

где L_i соответствует логарифму отношения правдоподобия, T – абсолютная температура, а R – универсальная газовая постоянная.

Данная величина $\Delta G'_{ij}$ имеет размерность свободной энергии, назовем её псевдо-свободной энергией, отражающей пробинг-данные. В алгоритмах динамического программирования свободная энергия вторичной структуры оценивается как сумма свободных энергий отдельных элементов. Мы расширили энергетическую модель программы RNASurface, добавив член псевдо-энергии к свободной энергии спаренных оснований (рис. 3.1).

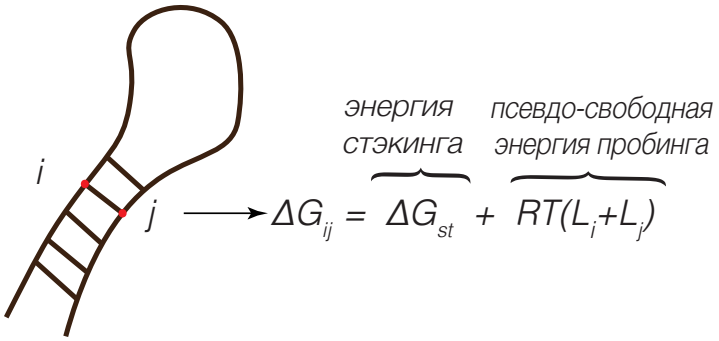


Рисунок 3.1. Схема энергетической модели.

Для оценки качества предсказаний вторичной структуры с учетом и без учета экспериментальных данных мы используем параметры чувствительность и PPV. Параметр «чувствительность» был определен в разделе 2.1.1, а параметр PPV (от англ. positive predictive value – уровень верных предсказаний) вычисляется следующим образом (значения TP и FP также определены в разделе 2.1.1):

$$PPV = \frac{TP}{TP + FP}$$

3.1.3 Построение фоновой модели

В нашем алгоритме мы изменяем энергетическую модель программы RNASurface и, таким образом, вычисленное значение свободной энергии для конкретной последовательности РНК. В этом случае необходимо построить новую фоновую модель для вычисления Z-значения на основе псевдо-свободной энергии и выделения структурированных элементов РНК.

Большинство подходов, оценивающих уровень структурированности РНК, в том числе программа RNASurface, на которую мы опираемся, используют перемешанные последовательности (с сохранением динуклеотидного состава) в качестве фоновой модели. Однако в нашем случае кроме последовательности РНК мы также используем экспериментальные данные, которые также необходимо перемешивать для «разрушения» паттернов вторичной структуры.

На рисунке 3.2 представлены результаты расчета Z-значений для сегментов РНК, полученные с помощью модифицированной версии программы RNASurface. На левом графике отображено распределение Z-значений с использованием экспериментальных данных PARS, в качестве фоновой модели мы использовали перемешивание последовательностей и данных. На правом графике в качестве фоновой модели мы использовали перемешанные последовательности, а экспериментальные данные перемешивали с сохранением марковских свойств данных, а также последующей квантильной нормализацией. Распределения Z-значений с использованием экспериментальных данных в обоих случаях смещены в отрицательную область, что говорит о том, что обе фоновые модели не являются нейтральными. Другие попытки перемешивания также не привели к созданию адекватной фоновой модели. Поэтому мы решили использовать другой подход.

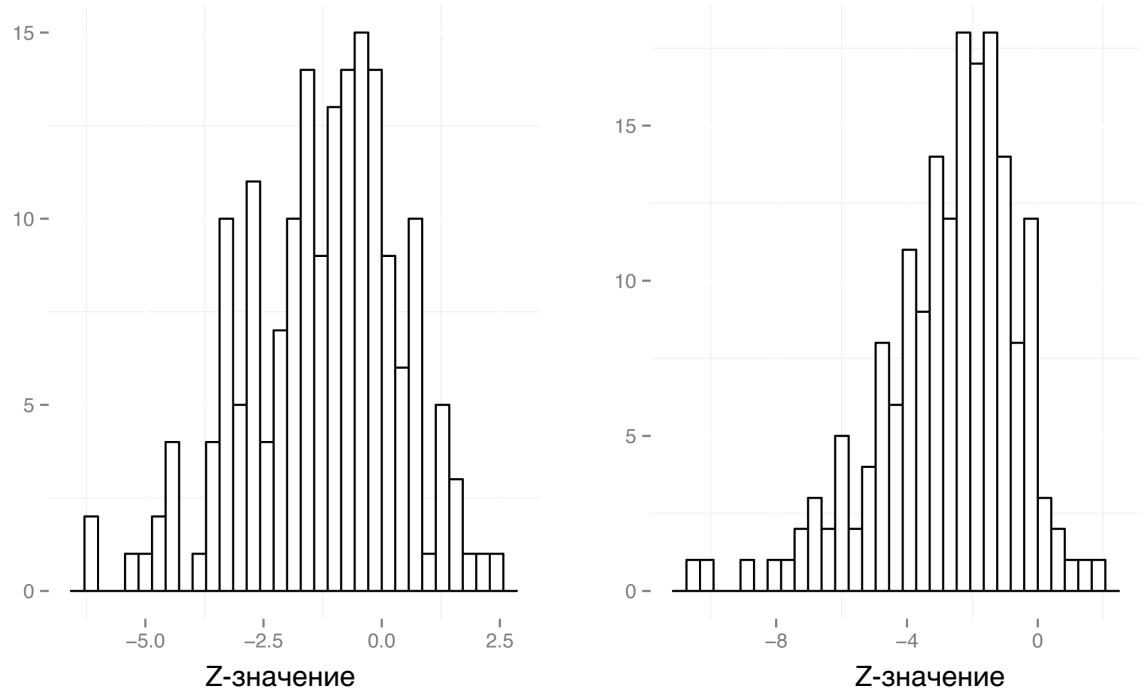


Рисунок 3.2. Результаты расчета Z-значений для сегментов РНК, с помощью модифицированной версии программы RNASurface с помощью двух разных фоновых моделей. На графике отображены распределения Z-значений с использованием экспериментальных данных PARS. Левый график: в качестве фоновой модели мы использовали перемешивание последовательностей и экспериментальных данных. Правый график: в качестве фоновой модели мы использовали перемешанные последовательности, а экспериментальные данные перемешивали с сохранением марковских свойств данных, а также последующей квантильной нормализацией

Мы построили фоновую модель с использованием случайных фрагментов мРНК, хорошо покрытых экспериментальными данными. Мы предполагаем, что в среднем мРНК менее структурирована, чем функциональные некодирующие РНК, и этот факт позволяет использовать фрагменты мРНК в качестве фоновой модели.

Чтобы оценить, как меняются значения энергии при расширении энергетической модели и добавлении экспериментальных данных для различных РНК, мы отобрали набор достаточно высоко покрытых транскриптов и разбили эти транскрипты на случайные отрезки длиной от 70 до 200 нуклеотидов. В итоге мы получили 23781 отрезок, и данный набор транскриптов мы считали базовым для построения фоновой модели.

Линейная зависимость между псевдо-МСЭ и МСЭ позволяет произвести следующее преобразование. На основе случайно выбранных фрагментов транскриптома мы построили линейную регрессию зависимости псевдо-МСЭ от МСЭ:

$$\text{псевдоМСЭ} = a \cdot \text{МСЭ} + b$$

Мы построили несколько регрессий для разных уровней покрытий (доля нуклеотидов в транскрипте, покрытых данными PARS) от 0.1 до 1 и увидели, что коэффициент пропорциональности *a* линейно зависит от покрытия.

На основании этого наблюдения мы ввели параметр, линейно зависящий от покрытия (обозначим как *cvr*, от англ. coverage):

$$a(crv) = a'cvr + c'$$

Таким образом, псевдо-МСЭ зависит и от МСЭ, и от покрытия:

$$\text{псевдоМСЭ} = a(cvr) \cdot \text{МСЭ} + b =$$

$$= (a' \cdot crv + c') \cdot \text{МСЭ} + b = a' \cdot \text{МСЭ} \cdot cvr + c' \cdot \text{МСЭ} + b$$

Рассчитав коэффициенты *a', b* и *c'* для фоновой модели, мы можем выразить МСЭ' из псевдо-МСЭ:

$$\text{МСЭ}' = \frac{\text{псевдоМСЭ} - \text{b}}{\text{a}' \cdot \text{crv} + \text{c}'}$$

На основании этого вычисленного МСЭ мы можем рассчитать Z-значения с учетом экспериментальных данных, для любых покрытий.

3.1.4 Полногеномный поиск на основании экспериментальных данных

При полногеномном поиске локально структурированных элементов РНК с использованием экспериментальных данных PARS мы рассматривали 558 транскриптов со средним покрытием 10 чтений на позицию. Для данных транскриптов мы проводили поиск с помощью программы RNASurface и её модифицированной версии, учитывающей экспериментальные данные. В качестве основного порога на Z-значение мы использовали порог, равный –3.

Для анализа относительного геномного расположения структурированных элементов на мРНК мы отобрали только транскрипты, содержащие нетранслируемые области длиной как минимум 100 нуклеотидов. Далее мы выровняли все транскрипты по старту трансляции или по концу кодирующей области и вычислили, сколько предсказанных структурированных элементов покрывает каждую позицию.

Для анализа относительного геномного расположения структурированных элементов на мРНК мы вычисляли, какая доля найденных элементов попадает в 5’ НТО, кодирующую часть мРНК и 3’ НТО. Элементы, относящиеся одновременно к 5’ НТО и кодирующей части или кодирующей части и 5’ НТО мы рассматривали отдельно. Для оценки ожидаемой доли элементов в каждом из возможных расположений мы провели 2 контроля. В первом случае ожидаемая доля была пропорциональна длине элемента (а для граничных расположений в качестве длины мы использовали длину 100 нуклеотидов). Во втором случае мы провели симуляцию: с помощью программы BedTools [71] мы случайно разбросали по мРНК элементы, предсказанные в алгоритме, и оценили их расположение. Процедуру повторили 100 раз, далее усреднили. В качестве основного контроля мы использовали второй контроль.

Мы использовали программу BedTools [71] для пересечения координат структурированных сегментов, предсказанных нашим алгоритмом, и сегментов, полученных с помощью Evofold [67]. Мы считали два сегмента

пересекающимися, если они пересекались на как минимум 50% от длины сегмента Evofold, и учитывали только уникальные сегменты Evofold (то есть если один сегмент Evofold пересекался с двумя нашими сегментами, то при подсчете количества пересечений мы учитывали это как одно пересечение).

Чтобы оценить соответствие между экспериментальными данными PARS и структурами, предсказанными с помощью Evofold, для каждой структуры мы разделили нуклеотиды на спаренные и неспаренные на основании Evofold структуры. Далее мы варьировали порог на реактивность PARS от -10 до 10, то есть в зависимости от порога мы считали нуклеотид спаренным, если он обладал реактивностью больше пороговой, и неспаренным, если он обладал реактивностью меньше пороговой. Для каждого порога все нуклеотиды были разделены на 4 группы: верно предсказанные спаренными на основании PARS данных (TP), верно предсказанные неспаренными (TN), неверно предсказанные спаренными (FP) и неверно предсказанные неспаренными (FN). Далее рассчитанные значения были использованы для вычисления чувствительности и специфичности и построения ROC-кривых. Лучший порог определялся как порог, при котором максимален коэффициент корреляции Мэтьюса *C*, вычисляемый как:

$$\mathrm{C} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FT) \cdot (TP + FP) \cdot (TN + FP) \cdot (TN + FN)}}$$

3.1.5 Веб-сервер RNASurface

Веб-сервер RNASurface представляет собой графический интерфейс к модифицированной версии алгоритма RNASurface. RNASurface написан на GWT (Google Webtool Kit). Сервер работает под операционной системой Linux на машине с процессором Modern Intel 1.6Ghz CPU (E5310) 6Gb RAM.

3.2 Результаты и обсуждение

3.2.1 Предсказание разных классов некодирующих РНК

Одним из самых простых способов поиска структурированных РНК-элементов генома является сравнительно-геномный поиск. Если при таком поиске не использовать собственно выравнивание (сопоставление символов), а опираться только на факт ортологичности последовательностей, то можно сильно снизить требования к качеству выравнивания, но при этом мы сможем находить эволюционно сохраняющиеся структуры. Для общей оценки эффективности подхода, опирающегося исключительно на последовательности РНК, мы реализовали алгоритм, определяющий меру структурированности участка в “скользящем окне” путем подсчета вероятностей нуклеотидов находиться в спаренном состоянии и сравнения с фоновой моделью. Данный алгоритм был применен к геному *D. melanogaster*: мы рассматривали окна длиной 40, 70, 100 и 150 нуклеотидов, что позволило найти структурированные элементы РНК разной длины.

Результаты расчета чувствительности для 4 известных классов РНК, таких как микроРНК, тРНК, мяРНК и мякРНК, представлены в таблице 3.1. Столбец «длина окна» содержит значение длины окна, при котором удалось достичь самой высокой чувствительности. В всех 4 случаях чувствительность не очень высокая, что говорит о средней предсказательной силе метода. Еще один важный вывод – это наблюдение о том, что чувствительность метода в рамках одного класса сильно зависит от длины используемого окна и, таким образом, окно является сильным ограничением метода. В дальнейшей работе мы использовали другой метод, снимающий данное ограничение.

Таблица 3.1

Тип нкРНК	Чувствительность, в %	Длина окна, в нуклеотидах
микроРНК	56	100
тРНК	44	40
мякРНК	21	150
мяРНК	53	40

Таким образом, описанный алгоритм позволяет анализировать полногеномные выравнивания для поиска структурированных РНК-элементов. Существенным минусом является наличие окна: мы можем искать только элементы примерно заданной длины, для других окон метод дает слабые предсказания. Кроме того, в зависимости от класса РНК сильно варьируется чувствительность метода, но для всех классов чувствительность остается низкой.

Использование экспериментальных данных пробинга вторичной структуры РНК для предсказания структурированных элементов РНК позволяет значительно улучшить эффективность предсказания. Далее мы рассмотрим, как можно использовать пробинг данные при полногеномных предсказаниях структурированных РНК. Первым шагом является расширение энергетической модели за счет включения данных пробинга.

3.2.2 Расширение энергетической модели

Если мы хотим использовать пробинг-данные в качестве дополнительного источника информации при определении структуры РНК или поиска структурированных участков РНК, данные должны быть учтены в энергетической

модели. Различные подходы к тому, как учесть пробинг данные, рассмотрены в разделе 1.5.3.

Анализ ROC-кривых позволяет сделать вывод о том, что использование порога на реактивность не позволяет достигнуть высокого соотношения чувствительности и специфичности разделения нуклеотидов на спаренные и не спаренные, вне зависимости от типа эксперимента. В этом случае использование строгих ограничений не представляется разумным: очень вероятно, что неправильное строгое определение статуса нуклеотида не только не улучшит качество предсказания всей структуры, но и значительно ухудшит его. Более правильным и обоснованным является использование мягких ограничений, то есть расширение энергетической модели так, чтобы она учитывала экспериментальные данные.

Вводя дополнительный член $\Delta G'_{ij}$, отражающий склонность пары нуклеотидов i и j находиться в спаренном состоянии, мы расширяем энергетическую модель, тем самым добавляя в алгоритм возможность учёта экспериментальной информации о структуре РНК.

Рассмотрим, как изменяется свободная энергия структуры при добавлении псевдо-энергии в энергетическую модель на примере структуры 5S рРНК. На рисунке 3.3 представлена вторичная структура данной молекулы, подтвержденная рентгеноструктурным анализом. Цвет нуклеотида отражает склонность данного нуклеотида находиться в спаренном состоянии согласно эксперименту SHAPE: нуклеотиды, имеющие низкую SHAPE реактивность (то есть склонные быть спаренными) отмечены синим цветом, а нуклеотиды, имеющие высокую SHAPE реактивность (то есть склонные быть неспаренными) отмечены красным цветом. Нуклеотиды, имеющие промежуточную реактивность, имеют менее насыщенный цвет. Данный рисунок демонстрируют среднюю точность данных SHAPE для рассматриваемой РНК. С одной стороны, большинство нуклеотидов имеет реактивность, согласующуюся со вторичной структурой и включение экспериментальной информации о структуре в энергетическую модель должно

улучшить предсказание. С другой стороны, есть несколько случаев неправильного определения реактивности, поэтому жесткие ограничения на статус нуклеотидов исходя из экспериментальных данных дадут неверную структуру.

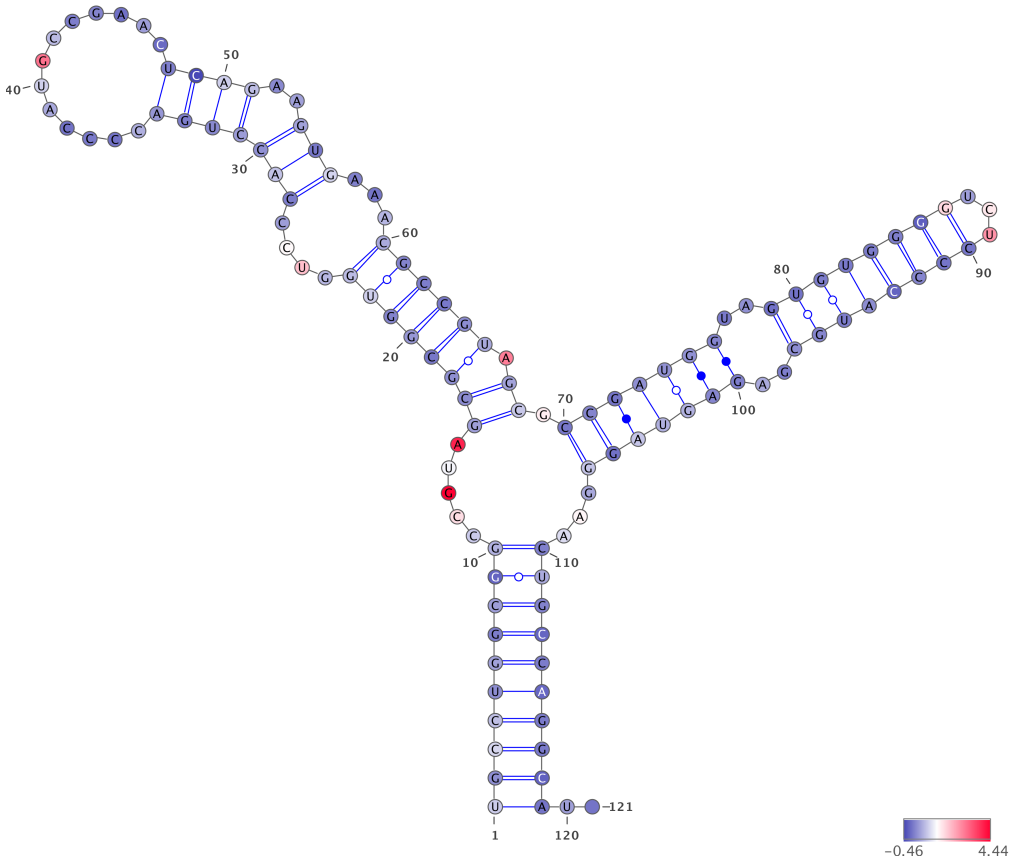


Рисунок 3.3. Вторичная структура 5S рРНК. Цвет нуклеотида отражает склонность данного нуклеотида находиться в спаренном состоянии согласно эксперименту SHAPE. Рисунок получен с помощью сервиса VARNA [17]**.**

В таблице 3.2 представлены результаты предсказания вторичной структуры РНК программой RNASurface, с учетом и без учета экспериментальных данных SHAPE.

Таблица 3.2. Результаты предсказания вторичной структуры РНК программой RNASurface, с учетом и без учета экспериментальных данных SHAPE.

нкРНК	Без SHAPE данных		С SHAPE данными	
	Чувствительность, %	PPV, %	Чувствительность, %	PPV, %
5S рРНК	50	53	64	61
тРНК	25	24	70	67
адениновый рибопереключатель	62	68	87	72
глициновый рибопереключатель	43	53	52	51

Во всех случаях мы наблюдаем улучшение чувствительности и PPV, в некоторых случаях незначительное, в некоторых – довольно сильное. Данные результаты свидетельствуют о том, что включение экспериментальных данных в энергетическую модель позволяет улучшить предсказание вторичной структуры отдельной молекулы, а значит может улучшить и предсказательную силу алгоритма при полногеномном поиске структурированных элементов.

3.2.3 Построение фоновой модели

Программа RNASurface рассчитывает Z-значение Z_{ij} для каждого сегмента S_{ij} последовательности до определенного размера. Для этого для каждого сегмента S_{ij} сначала вычисляется минимальная свободная энергия (МСЭ) вторичной

структуры, а далее эта свободная энергия используется для вычисления Z_{ij}. В нашем алгоритме мы изменяем энергетическую модель и, таким образом, меняется вычисленное значение свободной энергии для конкретной последовательности РНК.

Мы вычислили значения МСЭ и минимальной свободной энергии с учетом экспериментальных данных (псевдо-свободной энергии, псевдо-МСЭ) для каждого из рассматриваемых случайно выбранных фрагментов транскриптома. Далее мы построили зависимость МСЭ от покрытия (рисунок 3.4 а) и псевдо-МСЭ от покрытия (рисунок 3.4 б). Оба рисунка демонстрируют, что, как и ожидалось, нет зависимости между энергией и покрытием.

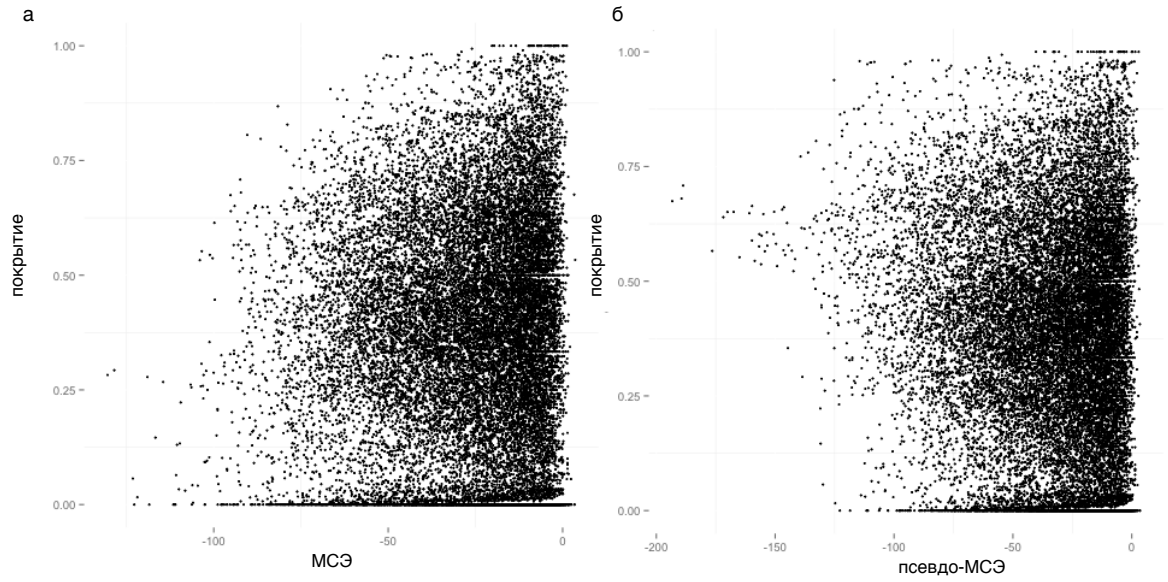


Рисунок 3.4. Зависимость МСЭ от покрытия (а) и псевдо-МСЭ от покрытия (б). Данные для случайных фрагментов транскриптома, экспериментальные данные по пробингу – PARS [98]**.**

Далее мы построили точечный график зависимости псевдо-МСЭ от МСЭ (рис. 3.5). Цветом отмечен уровень покрытия в эксперименте PARS [98]. Мы

наблюдаем, что псевдо-МСЭ примерно линейно зависит от МСЭ для случайных участков транскриптома. Вклад экспериментальных данных в энергетическую модель отражает соответствие данных и какой-либо вторичной структуры. Большинство случайно выбранных сегментов не обладает функциональной вторичной структурой, поэтому в среднем мы наблюдаем линейную зависимость псевдо-МСЭ от МСЭ.

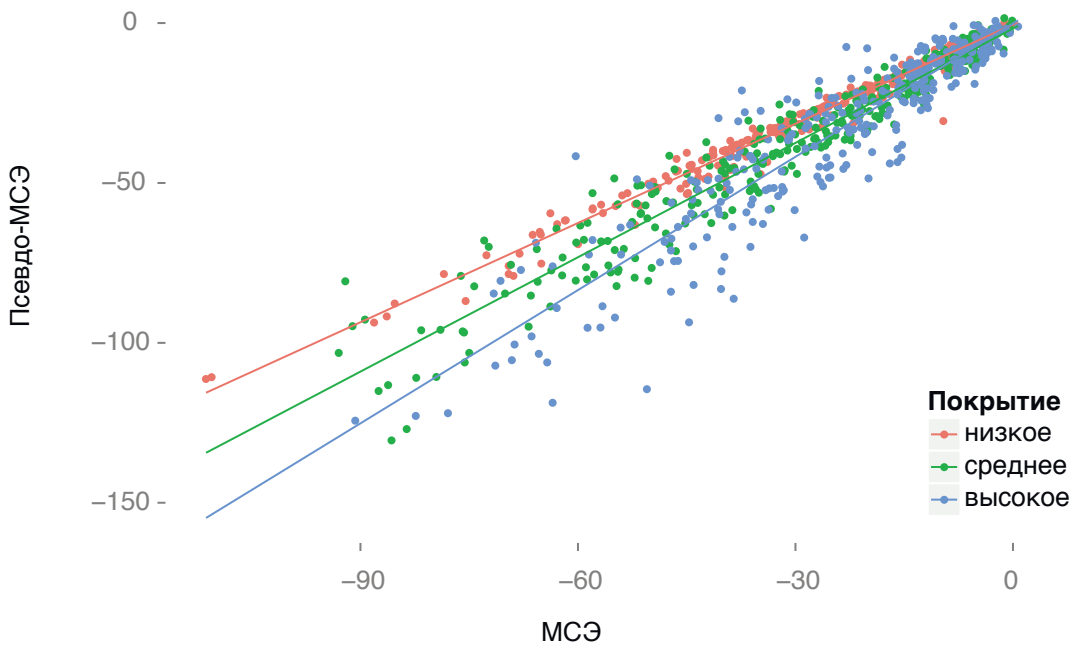


Рисунок 3.5. Зависимость псевдо-МСЭ от МСЭ. Цветом отмечен уровень покрытия в эксперименте PARS.

Рассмотрим, как именно может меняться МСЭ при учете экспериментальных данных. На рис. 3.6 представлена схема преобразования псевдо-МСЭ, полученной при включении экспериментальных данных в энергетическую модель, в МСЭ, которую далее мы используем при вычислении Z-значений. Каждая точка

соответствует фрагменту мРНК, для которого были вычислены МСЭ и псевдо-МСЭ. Синие точки на графики – случайные неструктурированные фрагменты мРНК, по ним была построена регрессионная прямая (обозначена серым) и они в среднем мало отклоняются от линейной зависимости.

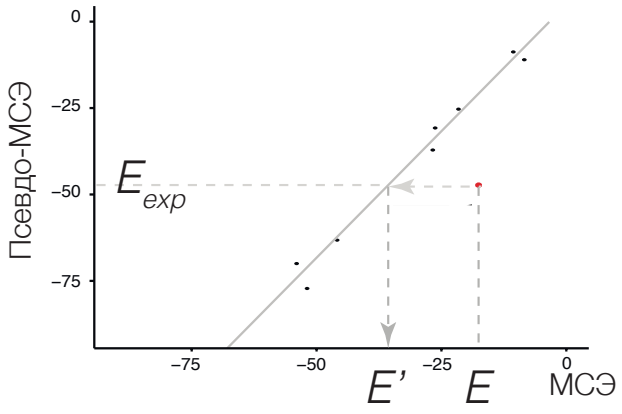


Рисунок 3.6. Схема вычисления МСЭ' через псевдо-МСЭ на основании фоновой модели.

Сильное отклонение псевдо-МСЭ от линейной зависимости фоновой модели отражает неслучайный вклад экспериментальных данных в энергетическую модель и, следовательно, соответствие некоторой структуры с низкой энергией и данных. В этом случае псевдо-МСЭ оказывается сильно сдвинута в сторону отрицательных значений. Например, красная точка на графике имеет псевдо-МСЭ примерно равную –50 ккал/моль, а МСЭ – примерно –20 ккал/моль. Последовательность из фоновой модели с МСЭ, равной –20 ккал/моль, имела бы псевдо-МСЭ в районе –25 ккал/моль. Таким образом, мы видим, что значение псевдо-МСЭ сильно сдвинуто в сторону отрицательных значений относительно

фоновой модели, что позволяет говорить о наличии функциональной структуры с низкой энергией, поддержанной и экспериментальными данными.

В случае противоречия энергетической модели и экспериментальных данных мы будем наблюдать обратную картину: значение псевдо-МСЭ будет сдвинуто в сторону положительных значений относительно фоновой модели.

Таким образом, мы можем вычислить для каждой точки ожидаемое значение МСЭ (обозначим как МСЭ’), с учетом фоновой модели. Разница МСЭ’ минус МСЭ показывает, насколько данные отвечают некоторой предположительно функциональной и стабильной структуре.

Кроме того, мы наблюдаем, что покрытие данными в эксперименте вносит вклад в зависимость псевдо-МСЭ от МСЭ, поэтому мы также учитываем и покрытие в фоновой модели. В разделе 3.1.3 представлена полная формула преобразования псевдо-МСЭ в МСЭ’.

Вычисление МСЭ’ позволяет вычислить Z-значения для фрагментов как хорошо покрытых данными, так и плохо покрытых, и сравнивать их между собой. Рассмотрим, какие Z-значения мы получаем для фрагментов из фоновой модели и функциональных структур РНК.

Мы вычислили Z-значения на основании МСЭ’ (обозначим как пробинг-Z-значения) для фрагментов из фоновой модели и сравнили с Z-значениями на основании МСЭ для тех же фрагментов (рис 3.7). Мы наблюдаем, что пробинг-Z-значения высоко коррелируют с Z-значениями (коэффициент корреляции Спирмана 0.81), при этом нет зависимости от покрытия. Это говорит о том, что фоновая модель построена верно, и случайные фрагменты генома не меняют сильно Z-значение при добавлении экспериментальных данных.

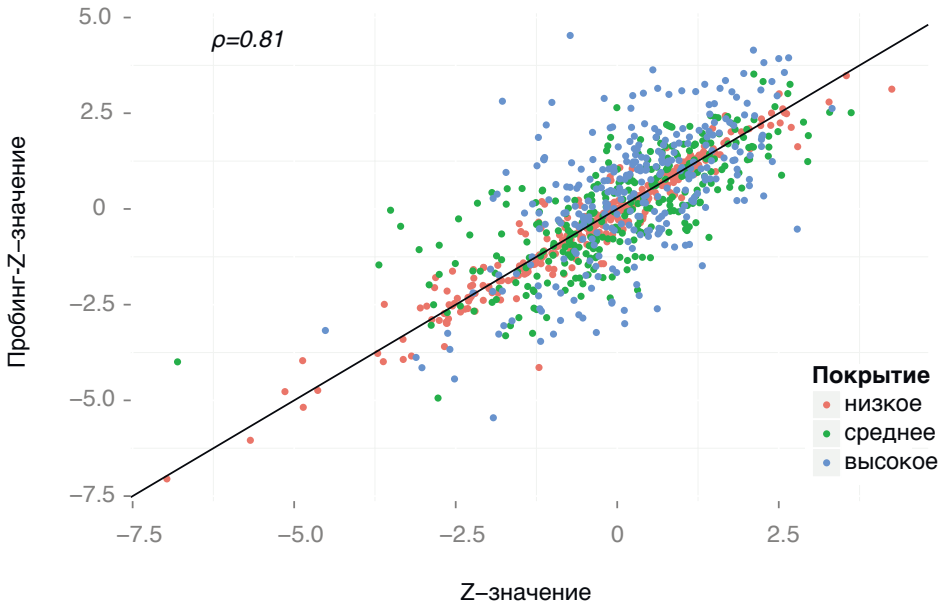


Рисунок 3.7. Z-значения, вычисленные на основании МСЭ' (обозначим как пробинг-Z-значения) для фрагментов из фоновой модели по сравнению с Z-значениями на основании МСЭ для тех же фрагментов.

Далее мы сравнили Z-значения и пробинг-Z-значения для нескольких классов функциональных некодирующих РНК (мякРНК, мяРНК, скаРНК) (рис. 3.8).

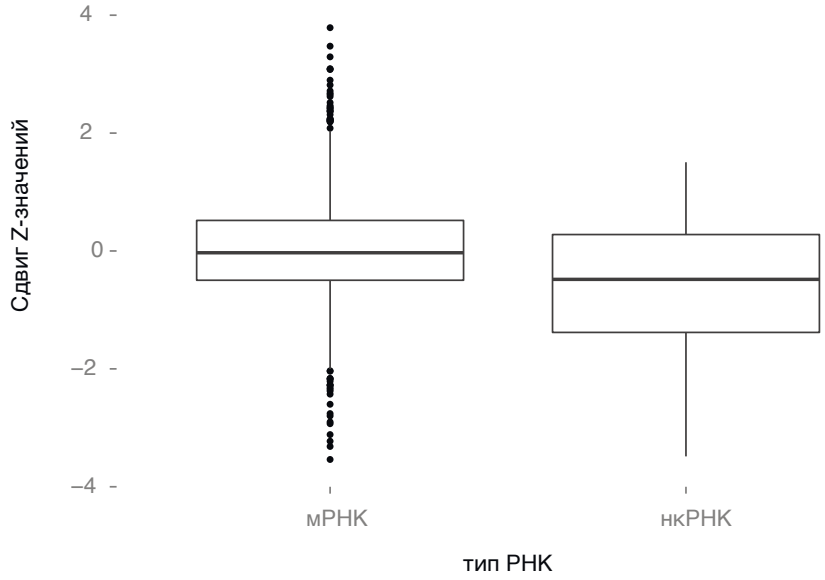


Рисунок 3.8. Распределений сдвигов в Z-значениях (пробинг-Z-значение минус Z-значение) для случайных сегментов мРНК и функциональных структурированных мРНК.

В отличие от мРНК фрагментов, пробинг-Z-значения статистически значимо ниже, чем Z-значения (тест Вилкоксона, p-значение равно 2.4×10^{-6}), со средним сдвигом 0.64 стандартных отклонений. Большинство функциональных некодирующих РНК имеют пробинг-Z-значения более низкие, чем обычные Z-значения (рис. 3.9).

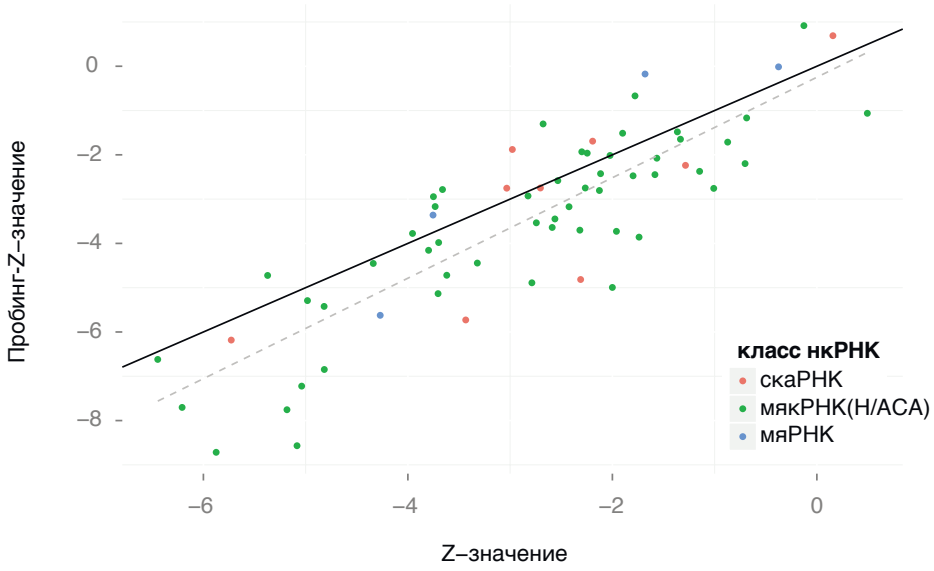


Рисунок 3.9. Z-значения и пробинг-Z-значения функциональных некодирующих РНК. Черная прямая y=x, пунктирная серая прямая – регрессионная прямая по точкам.

Учитывая, что рассматриваемые РНК уже имеют низкие Z-значения, даже указанный сдвиг значительно увеличивает предсказательную силу (рис. 3.10). Таким образом, функциональные некодирующие РНК являются структурированными, и экспериментальные данные соответствуют структуре, предсказываемой энергетической моделью. Данный факт позволяет лучше отличать функциональные структуры от фона, что и является целью включения экспериментальных данных в фоновую модель и модификации алгоритма RNASurface.

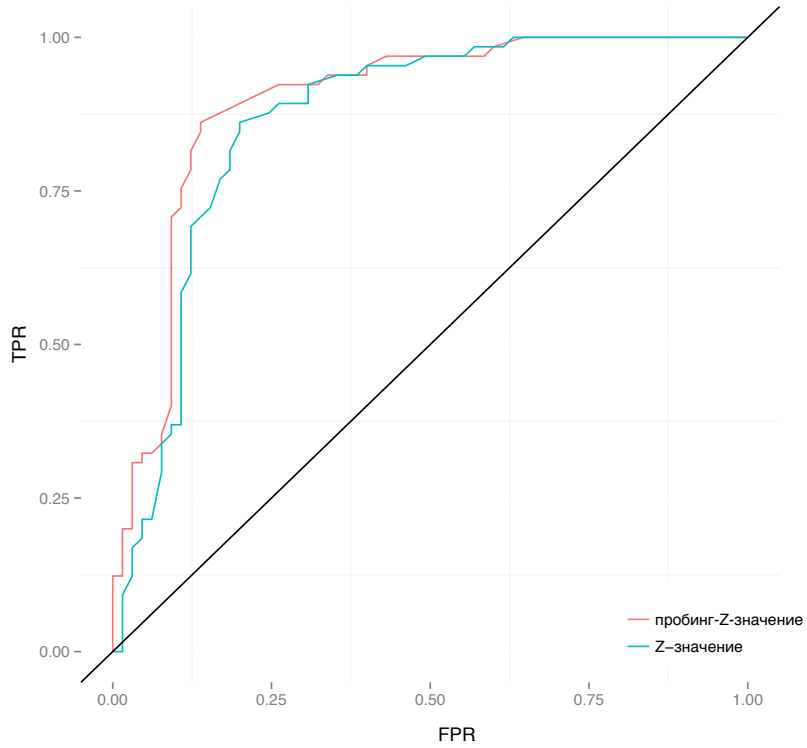


Рисунок 3.10. ROC-кривые, демонстрирующие увеличение предсказательной силы при использовании экспериментальных данных.

3.2.4 Полногеномный поиск с помощью PARS данных

Для оценки эффективности работы модифицированного алгоритма мы провели поиск локально структурированных элементов РНК в масштабе человеческого транскриптома с использованием экспериментальных данных PARS [98]. При поиске без учета экспериментальных данных программа RNASurface находит 3201 локально структурированный элемент РНК, а с учетом экспериментальных данных – 3587 элементов (порог на Z-значение в обоих случаях –3). Распределение длин и Z-значений не отличаются между двумя запусками, что еще раз подтверждает верное включение экспериментальных данных в работу алгоритма (рис. 3.11).

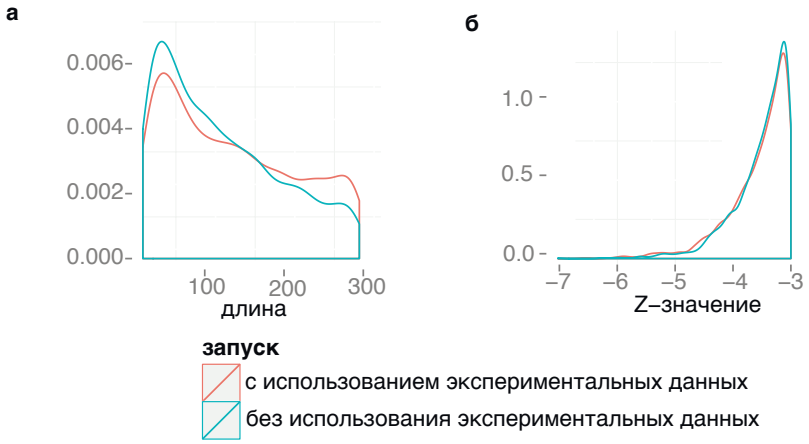


Рисунок 3.11. Распределения длин и Z-значений структурированных сегментов в запусках программы с использованием и без использования экспериментальных данных.

Нашей задачей было рассмотреть, чем отличаются предсказанные структурированные элементы с учетом экспериментальных данных от предсказанных структурированных элементов без учета экспериментальных данных. Мы нашли примерно одинаковое количество элементов в двух запусках, но только около 70 процентов элементов совпадает (с точностью до границ) между запусками. Важно было показать, что новые элементы (то есть элементы, которые находятся только при использовании экспериментальных данных) могут действительно являться функциональными.

Сначала мы проанализировали относительное геномное расположение структурированных элементов на мРНК. Мы вычислили, какая доля найденных элементов попадает в 5’ НТО, кодирующую часть мРНК и 3’ НТО. Также отдельно мы рассматривали элементы, попадающие на границы 5’ НТО и кодирующей части или кодирующей части и 3’ НТО. На рисунке 3.12 представлены результаты анализа: без использования экспериментальных данных основная часть структурированных элементов лежит в кодирующей области мРНК и в 3’ НТО. Также большое количество структурированных элементов

попадает в 5’ НТО, причем значительно больше, чем ожидается при случайном распределении элементов по мРНК. В случае использования экспериментальных данных большая часть структурированных элементов также попадает в кодирующую область и 3’ НТО. 5’ НТО также оказывается обогащена структурированными элементами, что согласуется с наблюдением о том, что 5’ НТО содержат большое количество регуляторных структурированных элементов. [3]. Интересно, что на границу 5’ НТО и кодирующей части попадает также меньшее количество структурированных элементов при использовании экспериментальных данных, чем без использования экспериментальных данных.

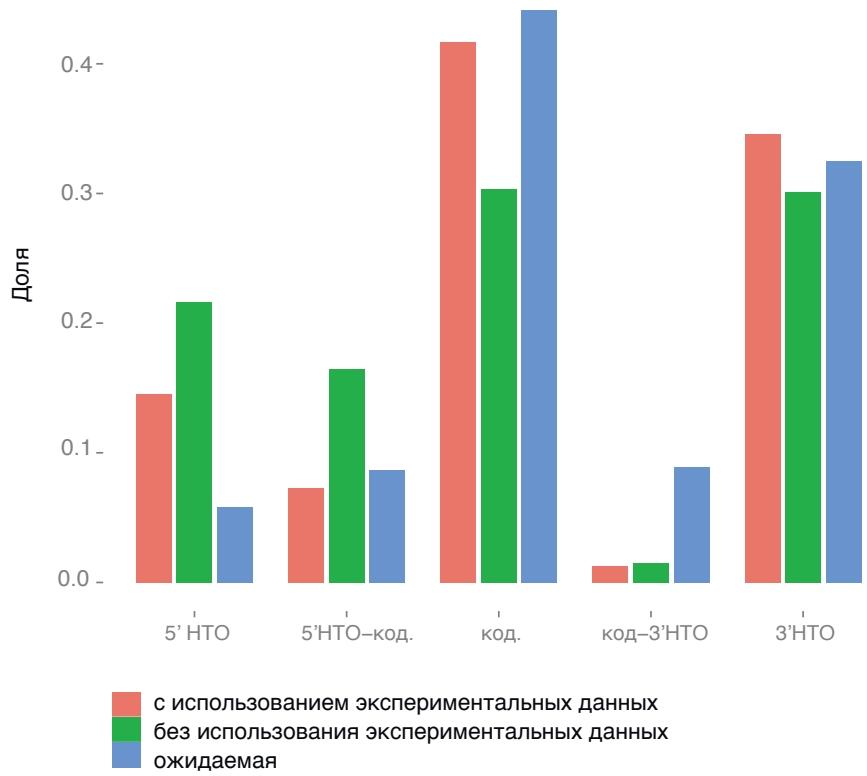


Рисунок 3.12. Распределение структурированных элементов по разным областям мРНК, усреднение по всем транскриптам.

Предыдущие исследования показали, что структурный состав 5’ НТО и 3’ НТО по отношению к кодирующим областям отличается от организма к организму. Было показано, что у организмов *Saccharomyces cerevisiae* [39] и *Arabidopsis thaliana* [48] 5’ НТО и 3’ НТО менее структурированы, чем кодирующие области, в том время как обратное наблюдается для мРНК человека, *Drosophila melanogaster* и *Caenorhabditis elegans* [47]. Отсутствие обогащения структурными элементами 3’ НТО может отчасти объясняться свойством данных PARS: экспериментальный протокол не позволяет определять уровень структурированности нуклеотидов, находящихся близко к 3’ концу мРНК.

Чтобы оценить относительное обогащение структурированными сегментами граничной области 5’ НТО и кодирующей области с разрешением в один нуклеотид, мы построили график отношения покрытия структурированными элементами с использованием экспериментальных данных к покрытию без использования экспериментальных данных (рис. 3.13).

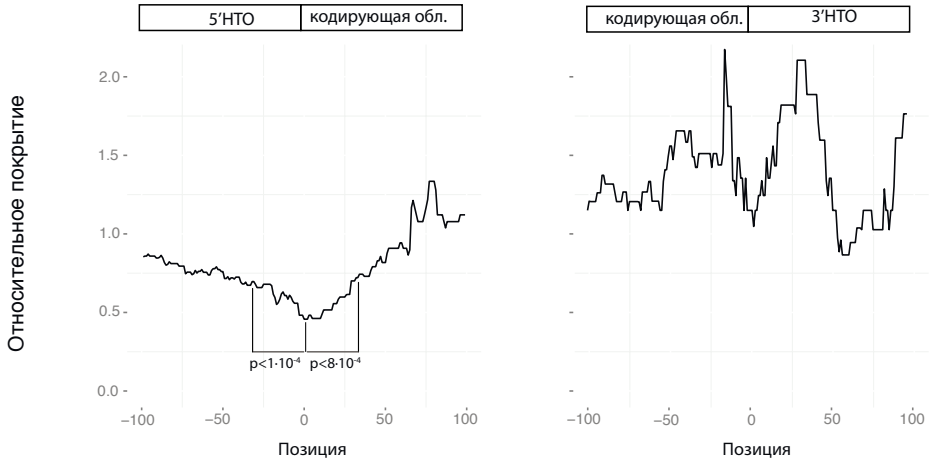


Рисунок 3.13. Относительное обогащение найденными структурированными элементами области старта и конца трансляции на мРНК, результаты при использовании экспериментальных данных

относительно результатов без использования экспериментальных данных. Усреднение по всем транскриптам.

Область вокруг стартов трансляции содержит мало структурированных элементов в запуске с использованием экспериментальных данных относительно запуска без использования экспериментальных данных (p-значение=0.00075). Это демонстрирует, что область вокруг стартов трансляции действительно менее структурирована *in vitro*, что согласуется с тем, что нуклеотиды вокруг стартов трансляции склонны находиться в неспаренном состоянии [61]. Кроме того, данные наблюдения согласуются с предыдущими вычислительными анализами относительного расположения структурированных элементов мРНК [81].

Для дальнейшего анализа найденных структур и доказательства их функциональности мы сравнили результаты нашего анализа с результатами работы программы Evofold [67], использующей множественные выравнивания и контекстно-свободные грамматики для поиска структурированных элементов РНК. Среди 16400 предсказаний Evofold в кодирующих и некодирующих областях мРНК 269 лежат в рассматриваемых нами транскриптах. Среди них 16% пересекаются с нашими предсказаниями с использованием экспериментальных данных и 12% – с предсказаниями без использования экспериментальных данных. Для различных порогов мы показали, что использование экспериментальных данных позволяет предсказывать больше структурированных элементов РНК, пересекающихся с предсказаниями Evofold (рис. 3.14). Кроме того, если наши предсказания объединить в кластеры (то есть пересекающиеся элементы рассматривать как один элемент, расширив границы), то тренд сохраняется (рис. 3.15).

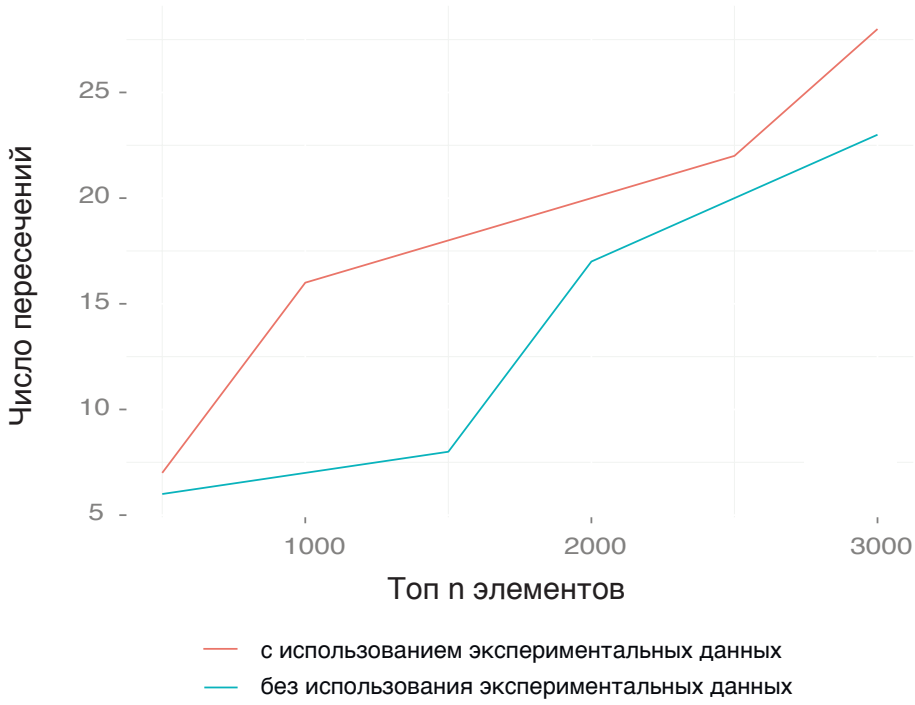


Рисунок 3.14. Число уникальных пересечений между предсказаниями структур Evofold и структурированными элементами, найденными с учетом (красная кривая) и без учета (синяя кривая) экспериментальных данных.

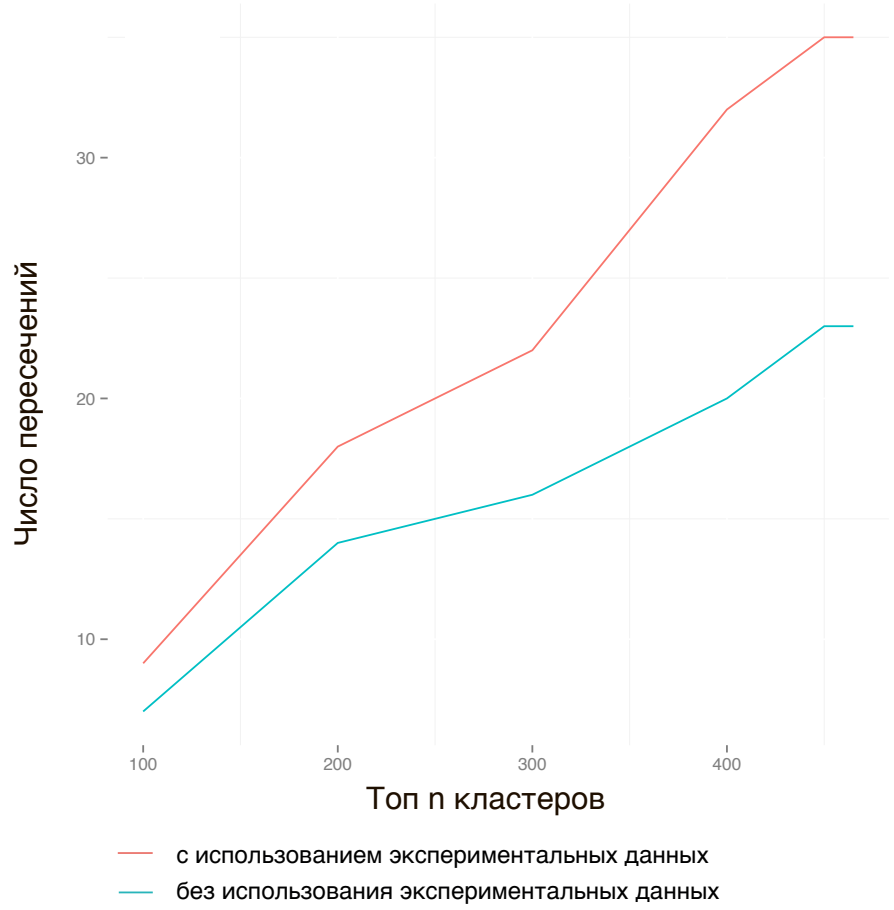


Рисунок 3.15. Число уникальных пересечений между предсказаниями структур Evofold и кластерами структурированных элементов, найденных с учетом (красная кривая) и без учета (синяя кривая) экспериментальных данных.

Далее мы исследовали, какие структуры мы находим только при использовании экспериментальных данных (так называемые «новые» структуры) и какие структуры мы находим только без использования экспериментальных данных (так называемые «потерянные» структуры). Для этого мы пересекли кластеры структурированных элементов из двух запусков: было найдено 9

«новых» структур и 5 «потерянных» структур. Мы хотели исследовать, какие структуры «теряются» при использовании экспериментальных данных и почему.

Для решения этой задачи мы проанализировали, какие PARS значения имеют спаренные и неспаренные нуклеотиды в «новых» и «потерянных» структурах (статус нуклеотиды мы определяли на основании структуры, полученной в программе Evofold). Было показано, что спаренные нуклеотиды в «новых» имеют значительно большие PARS значения (p-значение=0.013), чем неспаренные, что говорит об общей согласованности экспериментальных данных и структур (рис. 3.16). Напротив, PARS значения в «потерянных» структурах спаренных и неспаренных нуклеотидах не различаются, что говорит об отсутствии информации о структурированности в данных.

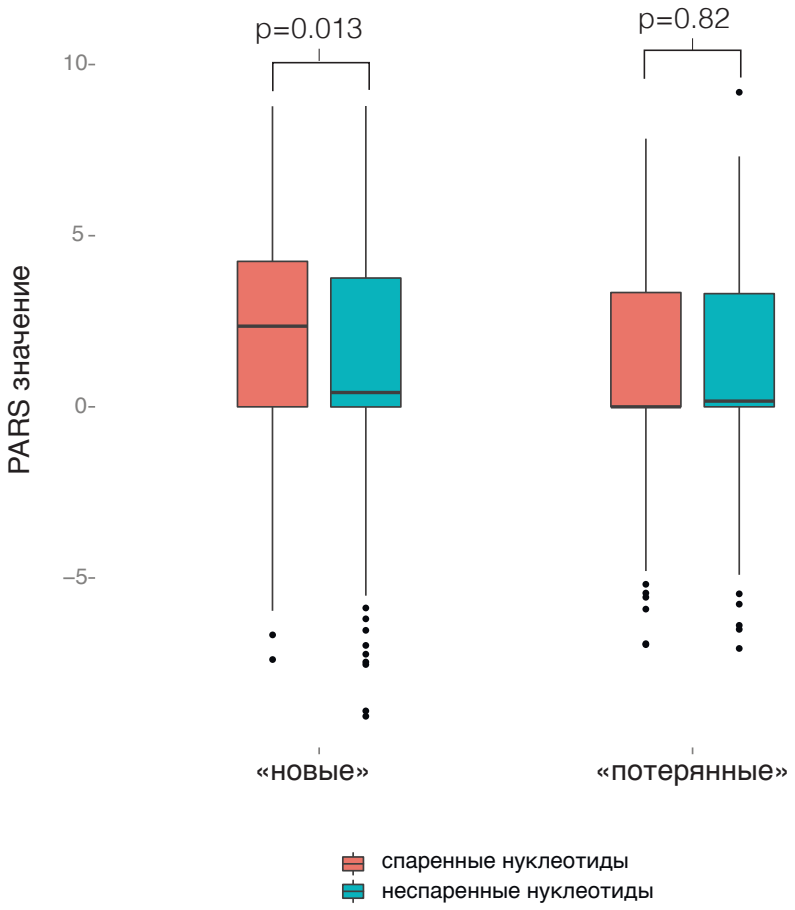


Рисунок 3.16. Распределения PARS значений спаренных и неспаренных нуклеотидов для «новых» и «потерянных» структур.

ROC-кривые демонстрируют, что разделение нуклеотидов на спаренные и неспаренные на основании порога, позволяет добиться лучших результатов в случае использования экспериментальных данных, по сравнению с результатами без использования экспериментальных данных (рис 3.17).

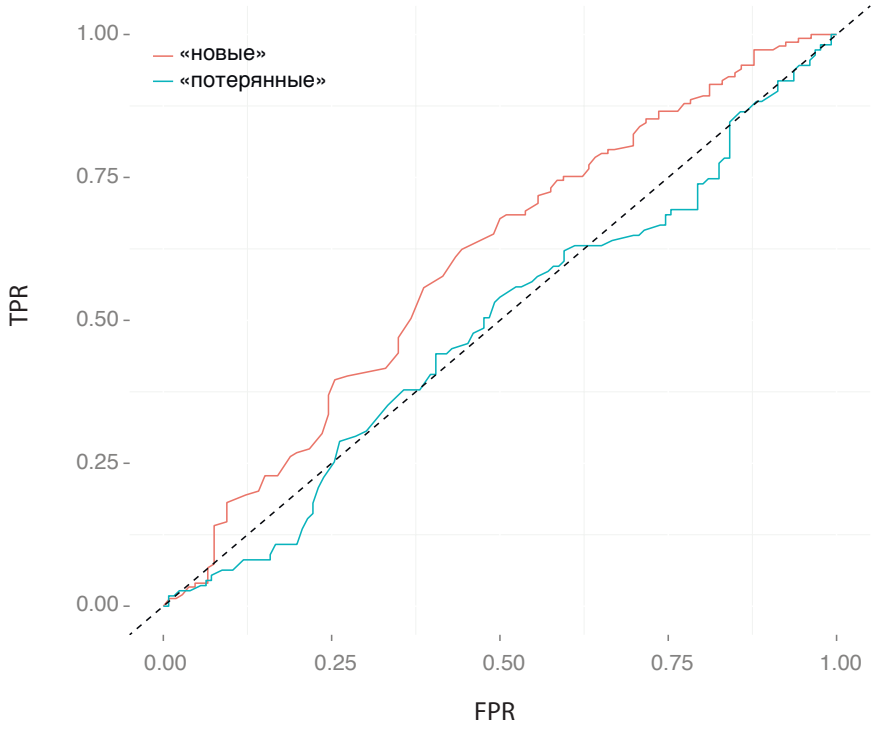


Рисунок 3.17. ROC-кривые для «новых» и «потерянных» структур, отражающие эффективность разделения нуклеотидов на спаренные и неспаренные относительно PARS значений.

На рисунке 3.18 приведены конкретные примеры, демонстрирующие разные уровни согласованности между данными PARS и структурами Evofold.

Для каждой структуры мы вычислили лучший порог на реактивность PARS, позволяющий разделять нуклеотиды на склонные находиться в спаренном состоянии и нуклеотиды, склонные находиться в неспаренном состоянии, на основании коэффициента корреляции Мэтьюса. В первой структуре все нуклеотиды, имеющие статус спаренных, демонстрируют PARS значение выше порогового (обозначены красным), а нуклеотиды, имеющие статус неспаренных, демонстрируют PARS значение ниже порогового (обозначены зеленым). В первом случае статус всех нуклеотидов, для которых есть информация из эксперимента, определен верно, то есть значение AUC равно 1 (идеальная согласованность эксперимента и структуры). Это объясняет, почему данная структура получила сильное «предпочтение» при поиске структурированных элементов с использованием экспериментальных данных. В случае второй и третьей структуры мы наблюдаем среднюю согласованность, статус части нуклеотидов определяется неверно. В случае последней структуры статус всех нуклеотидов определен неверно, то есть экспериментальные данные противоречат структуре: нуклеотиды в петлях имеют большие PARS значения, чем спаренные нуклеотиды, то есть невозможно подобрать порог, который адекватно разделил бы спаренные и неспаренные нуклеотиды. Эта структура оказывается «потерянной» при использовании экспериментальных данных, так как наблюдается несогласованность данных и структуры.

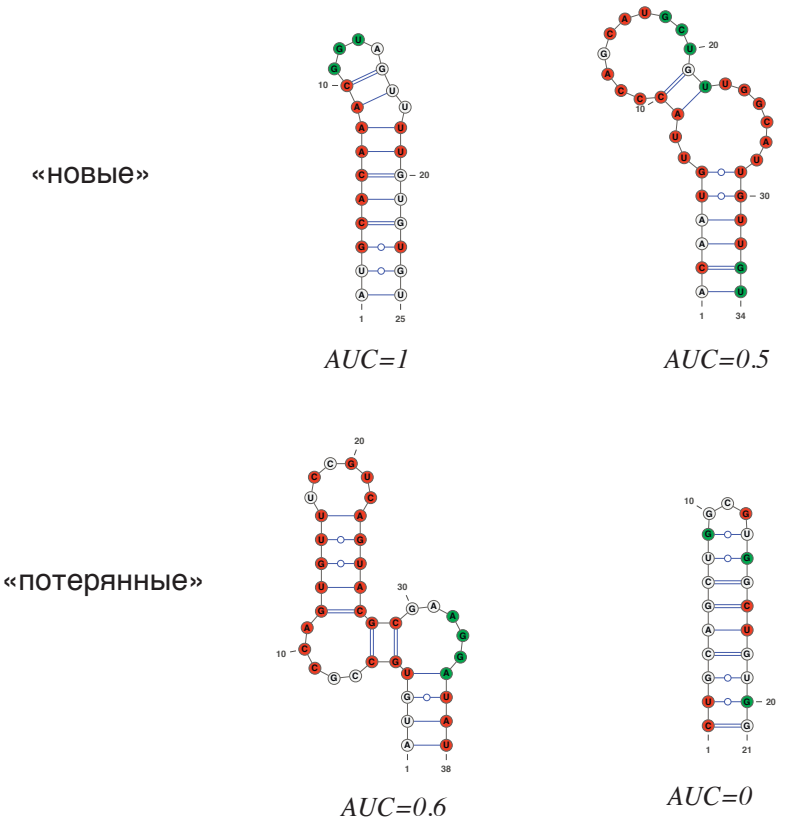


Рисунок 3.18. Структуры Evofold, цвет нуклеотида зависит от его PARS значения: красным обозначены нуклеотиды с PARS значением выше порогового, зеленым – нуклеотиды с PARS значением ниже порогового, без цвета – нуклеотиды, для которых отсутствуют данные.

Данный анализ является первым шагом в сторону глобального инкорпорирования экспериментальных пробинг данных в алгоритмы поиска структурированных РНК. Будущее улучшение протоколов получения экспериментальны данных *in vitro* и *in vivo* будет способствовать улучшению предсказательной силы метода и позволит находить функциональные структурированные элементы РНК с высокой точностью.

3.2.5 Веб-сервер RNASurface

Веб-сервер RNASurface расположен по адресу:

http://bioinf.fbb.msu.ru/RNASurface/index.html

Данный веб-сервер позволяет внешнему пользователю анализировать интересные ему длинные последовательности РНК (длиной до 10 тысяч нуклеотидов) с целью поиска структурированных сегментов. В качестве опции пользователь также может загружать экспериментальные данные по структуре рассматриваемой РНК.

В ответ на запрос пользователя веб-сервер выдает тепловую карту структурированности сегментов РНК, получаемую с помощью программы RNASurface. Кроме того, для скачивания доступны списки локально оптимальные сегментов РНК и все файлы, которые визуализированы на веб-сервере. Далее эти файлы можно использовать для загрузки в качестве отдельных треков на сторонние сервисы. Кроме того, пользователь может также загрузить свою разметку, например, координаты известных микроРНК для сравнения с результатами работы веб-сервиса.

3.3 Выводы к главе 3

Подход поиска структурированных РНК с использованием экспериментальных пробинг данных имеет определенные преимущества по сравнению с вычислительным подходом. Использование экспериментальных данных позволяет находить структуры, совместимые с данными эксперимента и одновременно обладающие низкой свободной энергией. Таким образом,

структуры с низкой энергией, но не совместимые с экспериментальными данными, не проходят отбор алгоритма на структурированность.

Мы разработали новый метод, позволяющий включить экспериментальные данные в энергетическую модель программы RNASurface. Данный подход позволяет работать с любыми источниками данных, что делает его универсальным. Расширение энергетической модели за счет включения псевдо-свободной энергии, рассчитанной на основе экспериментальных данных, возможно за счёт новой фоновой модели. Фоновая модель построена на основании случайной выборки неструктурированных сегментов мРНК и позволяет оценивать структурированность РНК элементов с учетом экспериментальных данных.

На основании экспериментальных данных PARS мы провели анализ структурированности РНК элементов в масштабах транскриптома человека и сравнили полученные результаты с результатами без использования экспериментальных данных. Вне зависимости от использования экспериментальных данных 5’ НТО мРНК обогащены структурированными элементами, а 3’ НТО мРНК имеют плотность структурных элементов на случайном уровне.

Кроме того, мы наблюдаем два региона с низким уровнем структурированности при поиске структурированных элементов РНК с использованием экспериментальных данных, но не в поиске без использования экспериментальных данных. Это интересное наблюдение, предполагающее, что использование экспериментальных данных позволяет получать более биологически осмысленные результаты: отсутствие структур в районе начала и конца кодирующей области может отвечать механизмам регуляции начала, элонгации или терминации трансляции.

Для удобства работы и доступа к модифицированной версии программы RNASurface мы также разработали веб-сервис, позволяющий визуализировать данные работы программы.

Выводы

1) Разработана методика анализа и преобразования экспериментальных данных, касающихся принадлежности отдельных нуклеотидов к вторичной структуре РНК. Методика основана на сопоставлении каждому нуклеотиду количественной характеристики, отражающей его склонность быть включенным во вторичную структуру.
2) Проведено сравнение данных эксперимента ДМС *in vivo* и *in vitro*. Показано, что кодирующие области мРНК являются менее структурированными в клетке по сравнению с состоянием *in vitro*.
3) Алгоритм RNASurface расширен на случай использования в оценке степени структурированности фрагмента экспериментальных данных. Построена фоновая модель для оценки структурированности РНК, учитывающая как энергетические параметры, так и экспериментальные данные.
4) Разработан и запущен веб-сервис RNASurface (http://bioinf.fbb.msu.ru/RNASurface/), позволяющий визуализировать результаты работы алгоритма по предсказанию структурированных элементов РНК с использованием экспериментальных данных.
5) На основании данных эксперимента PARS проведен анализ структурированности РНК элементов в масштабах транскриптома человека, показавший, что использование экспериментальных данных при поиске структурированных элементов РНК позволяет улучшить качество предсказания.

Список публикаций по теме диссертации

Статьи в научных журналах

1) Vinogradova SV, Sutormin RA, Mironov AA, Soldatov RA. Probing-directed identification of novel structured RNAs // RNA Biol. - 2016 – Vol.13(2) – P. 232-42.
2) Виноградова СВ, Солдатов РА, Миронов АА. Полногеномный поиск структурированных некодирующих РНК // Молекулярная биология – 2013 – Т. 47(4) – С. 689-96.

Тезисы конференций

1) Vinogradova S, Soldatov R, Sutormin R, Mironov A. RNASurface: a web-server for comprehensive reconstruction of RNA structural profile // Интеллектуальные системы молекулярной биологии (Intelligent Systems for Molecular Biology – ISMB'14). Сборник тезисов, Бостон, США – 2014
2) Солдатов РА, Виноградова СВ, Миронов АА. Процесс трансляции блокирует формирование вторичной структуры мРНК in vivo. // Информационные технологии и системы (ИТиС’14). Сборник тезисов, Нижний Новгород – 2014 – С. 26-27
3) Soldatov R, Vinogradova S, Mironov A. Translation causes global unfolding of mRNA structures in vivo // Европейская конференция по вычислительной биологии (European Conference on Computational Bilology – ECCB'14). Сборник тезисов, Страсбург, Франция – 2014.
4) Vinogradova S. Probing data facilitaties analysis of structure ensembles of RNA sequences // Симпозиум Европейской Организации Молекулярной Биологии и Европейского Института Биоинформатики «Некодирующий геном» (EMBO/EMBL Symposium'15 «The non-coding genome»). Сборник тезисов, Хайдельберг, Германия – 2015
5) Vinogradova SV, Soldatov RA, Mironov AA. Probing-directed structured elements detection in RNA sequences // Седьмая Московская международная

конференция по вычислительной молекулярной биологии (MCCMB'15). Сборник трудов, Москва – 2015.

6) Виноградова СВ, Солдатов РА, Миронов АА. Направленное РНК-зондирование генома для детектирования структурированных РНК. // Информационные технологии и системы (ИТиС’15). Сборник тезисов, Сочи – 2015 – С. 945-946

Благодарности

Хочу выразить искреннюю благодарность своему научному руководителю Андрею Александровичу Миронову за руководство и помощь при выполнении диссертации, а также коллегам из лаборатории биоинформатики ФББ МГУ и УНЦ «Биоинформатика» ИППИ РАН за ценные советы и помощь в выполнении работы.

Список литературы

1. Adilakshmi T., Lease R.A., Woodson S.A. Hydroxyl radical footprinting in vivo: mapping macromolecular structures with synchrotron radiation. // Nucleic Acids Research. 2006. № 8 (34). С. e64–e64.

2. Alkemar G., Nygård O. Probing the secondary structure of expansion segment ES6 in 18S ribosomal RNA. // Biochemistry. 2006. № 26 (45). С. 8067–8078.

3. Araujo P.R. [и др.]. Before It Gets Started: Regulating Translation at the 5' UTR. // Comparative and functional genomics. 2012. № 4 (2012). С. 475731–8.

4. Aviran S. [и др.]. Modeling and automation of sequencing-based characterization of RNA structure. // PNAS. 2011. № 27 (108). С. 11069–11074.

5. Bartel D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. // Cell. 2004. № 2 (116). С. 281–297.

6. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal Statistical Society Series B …. 1995.

7. Blanchette M. [и др.]. Aligning multiple genomic sequences with the threaded blockset aligner. // Genome Research. 2004. № 4 (14). С. 708–715.

8. Bonnet E. [и др.]. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. // Bioinformatics. 2004. № 17 (20). С. 2911–2917.

9. Brenowitz M. [и др.]. Probing the structural dynamics of nucleic acids by quantitative time-resolved and equilibrium hydroxyl radical "footprinting". // Current opinion in structural biology. 2002. № 5 (12). С. 648–653.

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

11. Caprara M.G., Myers C.A., Lambowitz A.M. Interaction of the Neurospora crassa mitochondrial tyrosyl-tRNA synthetase (CYT-18 protein) with the group I intron P4-P6 domain. Thermodynamic analysis and the role of metal ions. // Journal of molecular biology. 2001. № 2 (308). С. 165–190.

12. Chamberlin S.I., Weeks K.M. Mapping Local Nucleotide Flexibility by Selective Acylation of 2‘-Amine Substituted RNA // Journal of the American Chemical Society. 2000. № 2 (122). С. 216–224.

13. Chan P.P., Lowe T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. // Nucleic Acids Research. 2009. № Database issue (37). С. D93–7.

14. Clote P. [и др.]. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. // RNA. 2005. № 5 (11). С. 578–591.

15. Coppins R.L., Hall K.B., Groisman E.A. The intricate world of riboswitches // Current Opinion in Microbiology. 2007. № 2 (10). С. 176–181.

16. Costa M., Christian E.L., Michel F. Differential chemical probing of a group II self-splicing intron identifies bases involved in tertiary interactions and supports an alternative secondary structure model of domain V. // RNA. 1998. № 9 (4). С. 1055–1068.

17. Darty K., Denise A., Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. // Bioinformatics. 2009. № 15 (25). С. 1974–1975.

18. Das R. [и др.]. SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. // RNA. 2005. № 3 (11). С. 344–354.

19. Das R. [и др.]. Structural inference of native and partially folded RNA by high-throughput contact mapping. // PNAS. 2008. № 11 (105). С. 4144–4149.

20. Deigan K.E., Weeks K.M. Accurate SHAPE-directed RNA structure determination 2008. С. 1–6.

21. Ding Y. A statistical sampling algorithm for RNA secondary structure prediction // Nucleic Acids Research. 2003. № 24 (31). С. 7280–7301.

22. Ding Y. [и др.]. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features // Nature. 2013. С. 1–17.

23. Ding Y., Chan C.Y., Lawrence C.E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. // RNA. 2005. № 8 (11). С. 1157–1166.

24. Dirks R.M., Pierce N.A. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. // Journal of computational chemistry. 2004. № 10 (25). С. 1295–1304.

25. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach // Journal of molecular evolution. 1981. № 6 (17). С. 368–376.

26. Feng J. [и др.]. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6

ultraconserved region and functions as a Dlx-2 transcriptional coactivator. // Genes & Development. 2006. № 11 (20). С. 1470–1484.

27. Freier S.M. [и др.]. Improved free-energy parameters for predictions of RNA duplex stability. // Proceedings of the National Academy of Sciences of the United States of America. 1986. № 24 (83). С. 9373–9377.

28. Freyhult E., Gardner P.P., Moulton V. A comparison of RNA folding measures. // BMC bioinformatics. 2005. № 1 (6). С. 241.

29. Ge P., Zhong C., Zhang S. ProbeAlign: incorporating high-throughput sequencing-based structure probing information into ncRNA homology search. // BMC bioinformatics. 2014. № Suppl 9 (15 Suppl 9). С. S15.

30. Gherghe C.M. [и др.]. Strong Correlation between SHAPE Chemistry and the Generalized NMR Order Parameter (S2) in RNA // Journal of the American Chemical Society. 2008. № 37 (130). С. 12244–12245.

31. Gorodkin J., Hofacker I.L. From structure prediction to genomic screens for novel non-coding RNAs. // PLoS computational biology. 2011. № 8 (7). С. e1002100.

32. Guo F., Gooding A.R., Cech T.R. Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. // Molecular Cell. 2004. № 3 (16). С. 351–362.

33. Haller A. [и др.]. Conformational capture of the SAM-II riboswitch // Nature Chemical Biology. 2011. № 6 (7). С. 393–400.

34. Hartmuth K. [и др.]. An unusual chemical reactivity of Sm site adenosines strongly correlates with proper assembly of core U snRNP particles. // Journal of molecular biology. 1999. № 1 (285). С. 133–147.

35. Hofacker I.L., Fekete M., Stadler P.F. Secondary Structure Prediction for Aligned RNA Sequences // Journal of molecular biology. 2002. № 5 (319). С. 1059–1066.

36. Hofacker I.L., Priwitzer B., Stadler P.F. Prediction of locally stable RNA secondary structures for genome-wide surveys // Bioinformatics. 2004. № 2 (20). С. 186–190.

37. Horesh Y. [и др.]. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. // BMC bioinformatics. 2009. № 1 (10). С. 76.

38. Jády B.E., Kiss T. A small nucleolar guide RNA functions both in 2′−O−ribose methylation and pseudouridylation of the U5 spliceosomal RNA // The EMBO Journal. 2001. № 3 (20). С. 541–551.

39. Kertesz M. [и др.]. Genome-wide measurement of RNA secondary structure in yeast // Nature. 2010. № 7311 (467). С. 103–107.

40. Kladwang W. [и др.]. Standardization of RNA Chemical Mapping Experiments // Biochemistry. 2014. № 19 (53). С. 3063–3065.

41. Kozomara A., Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data // Nucleic Acids Research. 2010. № Database (39). С. gkq1027–D157.

42. Krol A. [и др.]. Solution structure of human U1 snRNA. Derivation of a possible three-dimensional model. // Nucleic Acids Research. 1990. № 13 (18). С. 3803–3811.

43. Latham M.P. [и др.]. NMR methods for studying the structure and dynamics of RNA. // Chembiochem : a European journal of chemical biology. 2005. № 9 (6). С. 1492–1505.

44. Lemieux S., Major F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. // Nucleic Acids Research. 2002. № 19 (30). С. 4250–4263.

45. Lescoute A., Westhof E. Topology of three-way junctions in folded RNAs. // RNA. 2006. № 1 (12). С. 83–93.

46. Lestrade L., Weber M.J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. // Nucleic Acids Research. 2006. № Database issue (34). С. D158–62.

47. Li F. [и др.]. Global analysis of RNA secondary structure in two metazoans. // Cell reports. 2012. № 1 (1). С. 69–82.

48. Li F. [и др.]. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. // The Plant cell. 2012. № 11 (24). С. 4346–4359.

49. Liebeg A., Waldsich C. Probing RNA structure within living cells. // Methods in enzymology. 2009. (468). С. 219–238.

50. Lorenz R. [и др.]. ViennaRNA Package 2.0. // Algorithms for molecular biology. 2011. № 1 (6). С. 26.

51. Loughrey D. [и др.]. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. // Nucleic Acids Research. 2014. С. gku909.

52. Low J.T., Weeks K.M. SHAPE-directed RNA secondary structure prediction // Methods. 2010. № 2 (52). С. 150–158.

53. Matera A.G., Terns R.M., Terns M.P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs // Nature reviews Molecular cell biology. 2007. № 3 (8). С. 209–220.

54. Mathews D.H. Revolutions in RNA secondary structure prediction. // Journal of molecular biology. 2006. № 3 (359). С. 526–532.

55. Mathews D.H. [и др.]. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure // Journal of molecular biology. 1999. № 5 (288). С. 911–940.

56. Mathews D.H. [и др.]. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure // PNAS. 2004. № 19 (101). С. 7287–7292.

57. McCaskill J.S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure // Biopolymers. 1990. № 6−7 (29). С. 1105–1119.

58. Merino E.J. [и др.]. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). // Journal of the American Chemical Society. 2005. № 12 (127). С. 4223–4231.

59. Mitra S. [и др.]. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. // Nucleic Acids Research. 2008. № 11 (36). С. e63–e63.

60. Moazed D., Stern S., Noller H.F. Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. // Journal of molecular biology. 1986. № 3 (187). С. 399–416.

61. Mortimer S.A., Kidwell M.A., Doudna J.A. Insights into RNA structure and function from genome-wide studies. // Nature Publishing Group. 2014. № 7 (15). С. 469–479.

62. Mueller F. [и др.]. The 3D arrangement of the 23 S and 5 S rRNA in the Escherichia coli 50 S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 A resolution. // Journal of molecular biology. 2000. № 1 (298). С. 35–59.

63. Nawrocki E.P., Eddy S.R. Infernal 1.1: 100-fold faster RNA homology searches. // Bioinformatics. 2013. № 22 (29). С. 2933–2935.

64. Nussinov R. [и др.]. Algorithms for Loop Matchings // SIAM Journal on Applied Mathematics. 1978. № 1 (35). С. 68–82.

65. Parker B.J. [и др.]. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. // Genome Research. 2011. № 11 (21). С. 1929–1943.

66. Peattie D.A., Gilbert W. Chemical probes for higher-order structure in RNA. // Proceedings of the National Academy of Sciences of the United States of America. 1980. № 8 (77). С. 4679–4682.

67. Pedersen J.S. [и др.]. Identification and classification of conserved RNA secondary structures in the human genome. // PLoS computational biology. 2006. № 4 (2). С. e33.

68. Poulsen L.D. [и др.]. SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data. // RNA. 2015. № 5 (21). С. 1042–1052.

69. Quarrier S. [и др.]. Evaluation of the information content of RNA structure mapping data for secondary structure prediction // RNA. 2010. № 6 (16). С. 1108–1117.

70. Quigley G.J., Rich A. Structural domains of transfer RNA molecules. // Science (New York, N.Y.). 1976. № 4267 (194). С. 796–806.

71. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. // Bioinformatics. 2010. № 6 (26). С. 841–842.

72. Rice G.M., Leonard C.W., Weeks K.M. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. // RNA. 2014. № 6 (20). С. 846–854.

73. Rivas E., Eddy S.R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. // Bioinformatics. 2000. № 7 (16). С. 583–605.

74. Rivas E., Eddy S.R. Noncoding RNA gene detection using comparative sequence analysis // BMC bioinformatics. 2001. № 1 (2). С. 1.

75. Romaniuk P.J. [и др.]. A comparison of the solution structures and conformational properties of the somatic and oocyte 5S rRNAs of Xenopus laevis. // Nucleic Acids Research. 1988. № 5 (16). С. 2295–2312.

76. Rouskin S. [и др.]. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo // Nature. 2013. С. 1–17.

77. Ruschak A.M. [и др.]. Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. // RNA. 2004. № 6 (10). С. 978–987.

78. Russell R. RNA misfolding and the action of chaperones. // Frontiers in bioscience : a journal and virtual library. 2008. (13). С. 1–20.

79. Sankoff D. Matching sequences under deletion-insertion constraints. // Proceedings of the National Academy of Sciences of the United States of America. 1972. № 1 (69). С. 4–6.

80. Seto A.G., Kingston R.E., Lau N.C. The Coming of Age for Piwi Proteins // Molecular Cell. 2007. № 5 (26). С. 603–609.

81. Shabalina S.A., Ogurtsov A.Y., Spiridonov N.A. A periodic pattern of mRNA secondary structure created by the genetic code. // Nucleic Acids Research. 2006. № 8

(34). С. 2428–2437.

82. Siegfried N.A. [и др.]. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). // Nature Publishing Group. 2014. № 9 (11). С. 959–965.

83. Smith M.A. [и др.]. Widespread purifying selection on RNA structure in mammals // Nucleic Acids Research. 2013. № 17 (41). С. 8220–8236.

84. Soldatov R.A., Vinogradova S.V., Mironov A.A. RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. // Bioinformatics. 2014. № 4 (30). С. 457–463.

85. Spitale R.C. [и др.]. RNA SHAPE analysis in living cells // Nature Chemical Biology. 2013. № 1 (9). С. 18–20.

86. Steen K.-A., Rice G.M., Weeks K.M. Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. // Journal of the American Chemical Society. 2012. № 32 (134). С. 13160–13163.

87. Stern S. [и др.]. RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA. // Science (New York, N.Y.). 1989. № 4906 (244). С. 783–790.

88. Sukosd Z. [и др.]. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions // Nucleic Acids Research. 2013. № 5 (41). С. 2807–2816.

89. Talkish J. [и др.]. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. // RNA. 2014. № 5 (20). С. 713–720.

90. Tijerina P., Mohr S., Russell R. DMS footprinting of structured RNAs and RNA-protein complexes. // nature protocols. 2007. № 10 (2). С. 2608–2623.

91. Tinoco I. [и др.]. Improved Estimation of Secondary Structure in Ribonucleic Acids // Nature New Biology. 1973. № 150 (246). С. 40–41.

92. Tinoco I., Uhlenbeck O.C., Levine M.D. Estimation of secondary structure in ribonucleic acids. // Nature. 1971. № 5293 (230). С. 362–367.

93. Toor N. [и др.]. Tertiary architecture of the Oceanobacillus iheyensis group II intron. // RNA. 2010. № 1 (16). С. 57–69.

94. Tranguch A.J. [и др.]. Structure-sensitive RNA footprinting of yeast nuclear ribonuclease P. // Biochemistry. 1994. № 7 (33). С. 1778–1787.

95. Tyagi R., Mathews D.H. Predicting helical coaxial stacking in RNA multibranch loops // RNA. 2007. № 7 (13). С. 939–951.

96. Underwood J.G. [и др.]. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing // Nature Publishing Group. 2010. № 12 (7). С. 995–1001.

97. Vitreschak A. Riboswitches: the oldest mechanism for the regulation of gene expression? // Trends in Genetics. 2004. № 1 (20). С. 44–50.

98. Wan Y. [и др.]. Landscape and variation of RNA secondary structure across the human transcriptome // Nature. 2014. № 7485 (505). С. 706–709.

99. Washietl S. [и др.]. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction // Nucleic Acids Research. 2012. № 10 (40). С. 4261–4272.

100. Washietl S., Hofacker I.L., Stadler P.F. Fast and reliable prediction of noncoding RNAs. // Proceedings of the National Academy of Sciences of the United States of America. 2005. № 7 (102). С. 2454–2459.

101. Watts J.M. [и др.]. Architecture and secondary structure of an entire HIV-1 RNA genome. // Nature. 2009. № 7256 (460). С. 711–716.

102. Wutz A., Gribnau J. X inactivation Xplained // Current opinion in genetics & development. 2007. № 5 (17). С. 387–393.

103. Yusupov M.M. [и др.]. Crystal structure of the ribosome at 5.5 A resolution. // Science (New York, N.Y.). 2001. № 5518 (292). С. 883–896.

104. Zarringhalam K. [и др.]. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction // PloS one. 2012. № 10 (7). С. e45160.

105. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. // Nucleic Acids Research. 2003. № 13 (31). С. 3406–3415.

106. Zuker M., Sankoff D. RNA secondary structures and their prediction // Bulletin of Mathematical Biology. 1984. № 4 (46). С. 591–621.

107. Zuker M., Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information // Nucleic Acids Research. 1981. № 1 (9). С. 133–148.