

На правах рукописи

**Виноградова Светлана Владимировна**

**Предсказание структурных элементов РНК с  
использованием экспериментальных данных**

Специальность: 03.01.09 – математическая биология,  
биоинформатика

**АВТОРЕФЕРАТ**

диссертации на соискание ученой степени  
кандидата биологических наук

Москва – 2016

Работа выполнена в Учебно-научном центре «Биоинформатика» Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель:

кандидат физико-математических наук,  
доктор биологических наук, профессор  
**Миронов Андрей Александрович**  
(Федеральное государственное бюджетное  
образовательное учреждение высшего образования  
Московский государственный университет имени  
М.В.Ломоносова)

Официальные оппоненты:

доктор биологических наук, профессор  
**Карягина-Жулина Анна Станиславовна**  
(федеральное государственное бюджетное  
учреждение «Федеральный научно-  
исследовательский центр эпидемиологии и  
микробиологии имени почетного академика  
Н.Ф.Гамалеи» Министерства здравоохранения  
Российской Федерации)

кандидат физико-математических наук  
**Касьянов Артем Сергеевич**  
(Федеральное государственное бюджетное  
учреждение науки Институт общей генетики им. Н.И.  
Вавилова Российской академии наук)

Ведущая организация:

Федеральное государственное бюджетное учреждение  
науки Институт математических проблем биологии  
Российской академии наук

Защита состоится 16 мая 2016 года на заседании Диссертационного совета Д.002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук, расположенном по адресу: 127051, г. Москва, Большой Каретный переулок, 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН), а также на сайте ИППИ РАН по адресу [www.iitp.ru](http://www.iitp.ru).

Автореферат разослан «    » \_\_\_\_\_ 2016 г.

Ученый секретарь  
диссертационного совета,  
доктор биологических наук,  
профессор

Рожкова Г.И.

## Общая характеристика работы

**Актуальность работы.** Молекулы РНК играют важную роль в различных клеточных процессах, регулируя их ко- и пост-транскрипционно. Более того, последние двадцать лет мы наблюдаем прорыв в области биологии РНК, сопровождающийся открытием десятков новых классов некодирующих РНК.

Многие функциональные некодирующие РНК обладают вторичной структурой, а наличие консервативной и функциональной вторичной структуры чаще всего говорит об её участии в биологических процессах клетки (Mortimer и др., 2014). Анализ вторичных структур и поиск новых структур в масштабе целого генома является фундаментальной и актуальной задачей молекулярной биологии, 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. Методы сравнительной геномики позволяют находить функциональные структурные элементы РНК (Parker и др., 2011, Smith и др., 2013), детектируя давление эволюционного отбора на структуру (Pedersen и др., 2006) или опираясь на ковариационные модели (Nawrocki, Eddy, 2013). 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 (Soldatov и др., 2014). Однако во многих случаях свободная энергия вторичной структуры РНК является недостаточным сигналом для эффективного распознавания функциональных структурированных элементов РНК (Rivas, Eddy, 2000).

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. 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 (Deigan, Weeks, 2008, Zarringhalam и др., 2012). Несмотря на наличие большого количества 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.

**Цель диссертационной работы** состояла в разработке метода, позволяющего 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 практическая значимость.** 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 было проанализировано несколько типов экспериментальных данных по определению вторичной структуры РНК. Был разработан теоретический подход для преобразования экспериментальных данных 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.

**Апробация работы:** Материалы исследований по теме диссертации 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 Biology – 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).

**Структура и объем диссертации.** Диссертация состоит из введения, обзора литературы, 2 глав, выводов и библиографии. Общий объем диссертации 103 страницы, из них 91 страница текста, включая 36 рисунков и 3 таблицы. Библиография включает 107 наименований на 8 страницах.

## **Содержание работы**

### **Введение**

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

### **Обзор литературы**

В разделе содержится мотивировка поставленных задач, а также обзор современной литературы по проблемам, рассмотренным в диссертации.

### **Свойства экспериментальных данных**

Экспериментальные подходы к определению вторичной структуры РНК позволяют для каждого нуклеотида последовательности РНК оценить его склонность находиться в спаренном состоянии. В типичном эксперименте молекула РНК обрабатывается специальным химическим веществом, модифицирующим, например, только нуклеотиды, находящиеся в петлях вторичной структуры РНК (эксперименты SHAPE (Aviran и др., 2011), ДМС (Rouskin и др., 2013)). В случае другого типа эксперимента (эксперимент PARS (Kertesz и др., 2010, Wan и др., 2014)) производится обработка РНК двумя нуклеазами, предпочтительно разрезающими только одноцепочечные или только двухцепочечные РНК. Далее с помощью обратной транскрипции или других методов в ходе эксперимента определяют положения модификаций или разрывов и, следовательно, положения нуклеотидов, более склонных к тому, чтобы находиться в спаренном или неспаренном состоянии. Таким образом, после нормализации данных, полученных в ходе эксперимента, получают профиль структурированности для данной последовательности РНК в виде значения реактивности для каждого нуклеотида. Однако данная величина является вероятностной по своей природе и не дает однозначного ответа о статусе нуклеотида. На рисунке 1а показано распределение реактивностей, полученных в ходе эксперимента SHAPE, для спаренных и неспаренных нуклеотидов (статус нуклеотида был определен согласно референсной структуре). Мы наблюдаем, что, с одной стороны, реактивности спаренных и неспаренных нуклеотидов различаются, но с другой стороны, любой порог на

реактивность, выбранный с целью разделить нуклеотиды на «определенные как спаренные» и «определенные как неспаренные» будет неверно определять статус многих нуклеотидов. Для эксперимента PARS (рисунок 1б) мы наблюдаем похожую картину. Таким образом, для использования экспериментальных данных для предсказания вторичной структуры РНК, необходимо преобразование реактивности в некоторую универсальную для всех типов эксперимента меру, отражающую вероятностные характеристики реактивности.

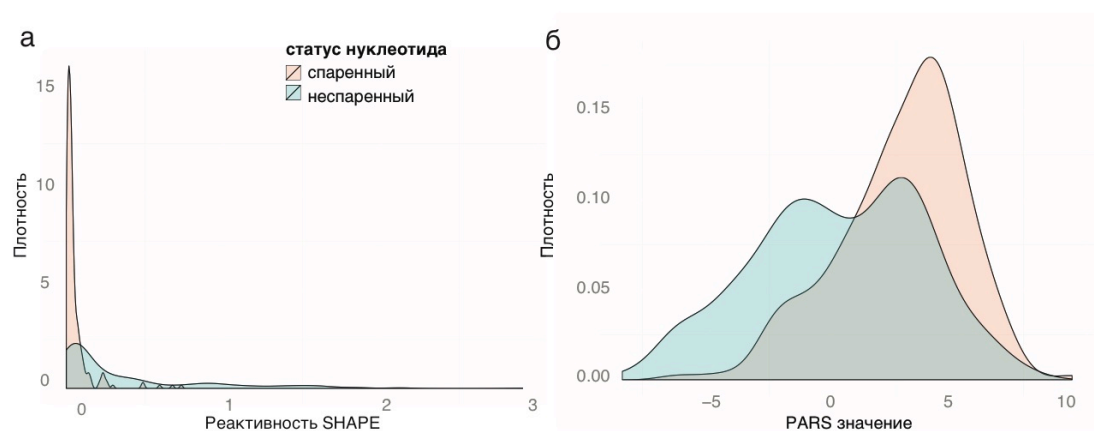


Рис. 1. Плотность вероятности распределений реактивностей спаренных и неспаренных нуклеотидов. а. Данные эксперимента SHAPE, функциональные структурированные РНК *E.coli*: глициновый, адениновый и цикло-ди-GMP рибопереключател, 5S рРНК и фенилаланин тРНК (Kladwang и др., 2014). б. Данные эксперимента PARS (Wan и др., 2014), РНК человека: малые ядрышковые РНК двух типов, малые ядерные РНК и РНК, локализующиеся в тельцах Кахаля.

Мы аппроксимировали распределения реактивностей  $r$  и, таким образом, получили плотности вероятностей  $P(r|paired)$  и  $P(r|unpaired)$ , специфичные для каждого эксперимента и отражающие способность данного эксперимента различать спаренные и неспаренные нуклеотиды. Количественная характеристика, отражающая структурные свойства нуклеотида, может быть выведена как логарифм отношения правдоподобия:

$$L(r) = \log \frac{P(r|paired)}{P(r|unpaired)} \quad (1)$$

На основании полученных распределений для спаренных и неспаренных нуклеотидов мы рассчитали  $L(r)$  и построили кривые зависимости  $L(r)$  от SHAPE реактивностей (рисунок 2а) и PARS значений (рисунок 2б). Зависимость  $L(r)$  от реактивности адекватно отражает статус нуклеотида в эксперименте.

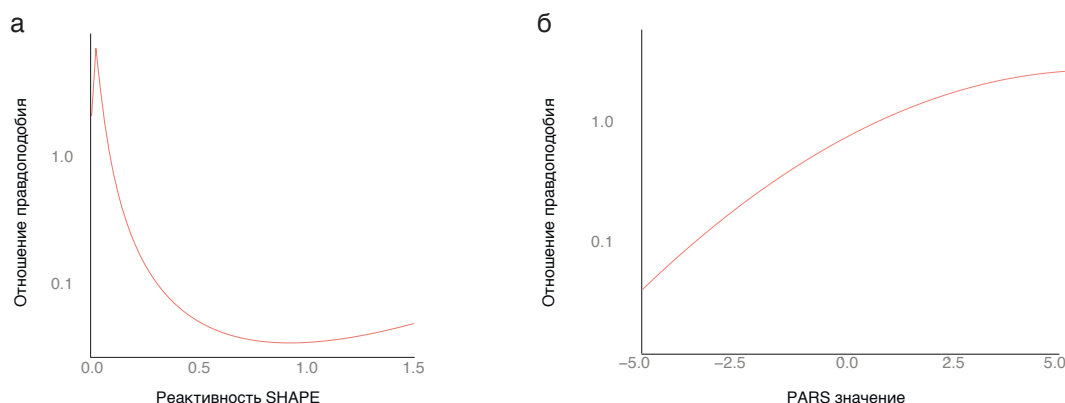


Рис. 2. Отношение правдоподобия в зависимости от SHAPE реактивности (а) и PARS значения (б).

Таким образом, мы ввели меру (логарифм отношения правдоподобия), отражающую склонность нуклеотида находиться в спаренном состоянии, которую можно использовать для включения экспериментальных данных в энергетическую модель при предсказании структур РНК.

Кроме того, мы проанализировали данные ДМС эксперимента по определению вторичной структуры РНК из статьи (Rouskin и др., 2013). В данной статье исследовали структуры мРНК в условиях *in vivo*, *in vitro* и денатурирующих условиях, на клетках дрожжей *S. cerevisiae* и клеточных линиях человека. Для оценки структурированности РНК мы использовали меру, отражающую склонность нуклеотида находиться в спаренном состоянии, – структурное значение:

$$\log \left( \frac{r^1_{/R1} + 2_{/(R1+R2)}}{r^2_{/R1} + 2_{/(R1+R2)}} \right) \quad (2)$$



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.

Мы рассчитали *in vivo/in vitro* структурные значения нуклеотидов относительно денатурированного контроля для фрагментов транскриптов с высоким покрытием. Было показано, что кодирующие фрагменты демонстрируют раскручивание *in vivo* относительно *in vitro*, в отличие от нетранслируемых областей (НТО) (рисунок 3).

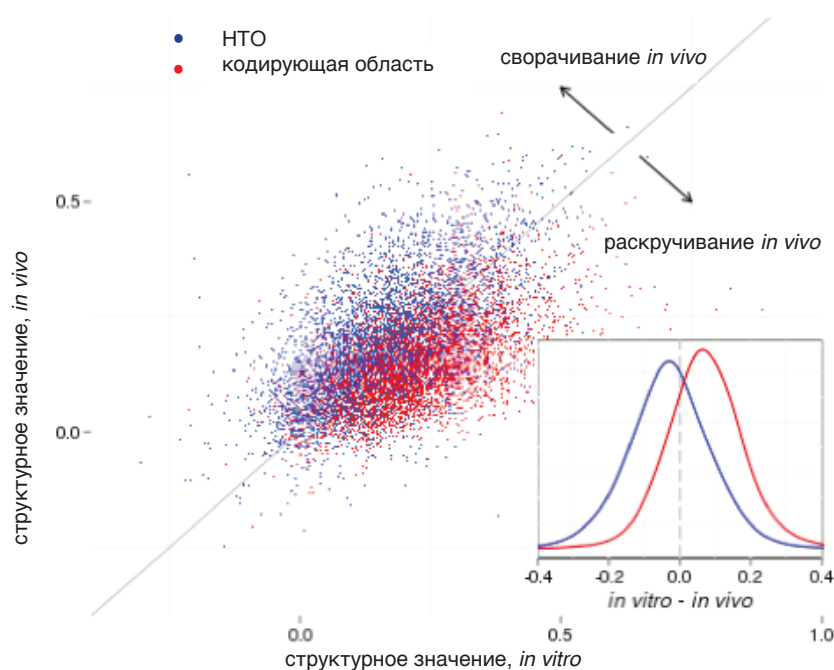


Рис. 3. Точечный график, демонстрирующий различия в структурных значениях для *in vivo* и *in vitro* состояний, для кодирующих и некодирующих областей. Данные ДМС эксперимента на фибробластах человека (Rouskin и др., 2013).

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 анализе

результатов включения экспериментальных данных в предсказание структур РНК и поиске структурированных элементов РНК.

### Поиск структурированных элементов РНК

На первом этапе нашей работы мы исследовали возможность распознавания структурированных элементов генома без использования экспериментальных данных, а также недостатки подхода, связанные с таким ограничением. Нами был детально исследован подход, основанный на методе сравнительной геномики, и предложено его уточнение, связанное с отказом от необходимости выравнивания. Отказ от необходимости выравнивания позволяет анализировать последовательности с невысоким уровнем сходства, но при этом 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. В результате сравнения с известными классами функциональных РНК было показано, что данный алгоритм позволяет анализировать полногеномные выравнивания для поиска структурированных РНК-элементов, однако в зависимости от класса РНК сильно варьируется чувствительность метода, и для большинства классов чувствительность остается низкой (таблица 1).

Таблица 1. Чувствительность предсказания различных типов некодирующих РНК (нкРНК) и длина окна для лучшего предсказания.

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

Целью диссертационной работы было разработать метод, позволяющий использовать экспериментальные данные при поиске структурированных элементов в длинных последовательностях РНК. При разработке метода мы опирались на алгоритм RNASurface (Soldatov и др., 2014), позволяющий искать локально структурированные участки в последовательностях РНК. Для возможности учёта экспериментальных данных мы модифицировали энергетическую модель программы, введя в формулу расчета свободной энергии стеблей дополнительный член  $\Delta G'_{ij}$ , отражающий склонность пары нуклеотидов  $i$  и  $j$  находиться в спаренном состоянии:

$$\Delta G'_{ij} = RT \cdot (L_i + L_j), \quad (3)$$

где  $L_i$  соответствует логарифму отношения правдоподобия,  $T$  – абсолютная температура, а  $R$  – универсальная газовая постоянная. Данная величина  $\Delta G'_{ij}$  имеет размерность свободной энергии, далее она обозначена как псевдо-свободная энергия, отражающая вклад экспериментальных данных.

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. Таким образом, для вычисления  $Z$ -значений на основании новой энергетической модели необходима новая фоновая модель.

Для построения фоновой модели мы использовали случайные фрагменты мРНК, покрытые данными PARS, предполагая, что в среднем такие фрагменты являются неструктурированными. На рисунке 4 представлен точечный график зависимости псевдо-МСЭ от МСЭ для случайных фрагментов мРНК. 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.

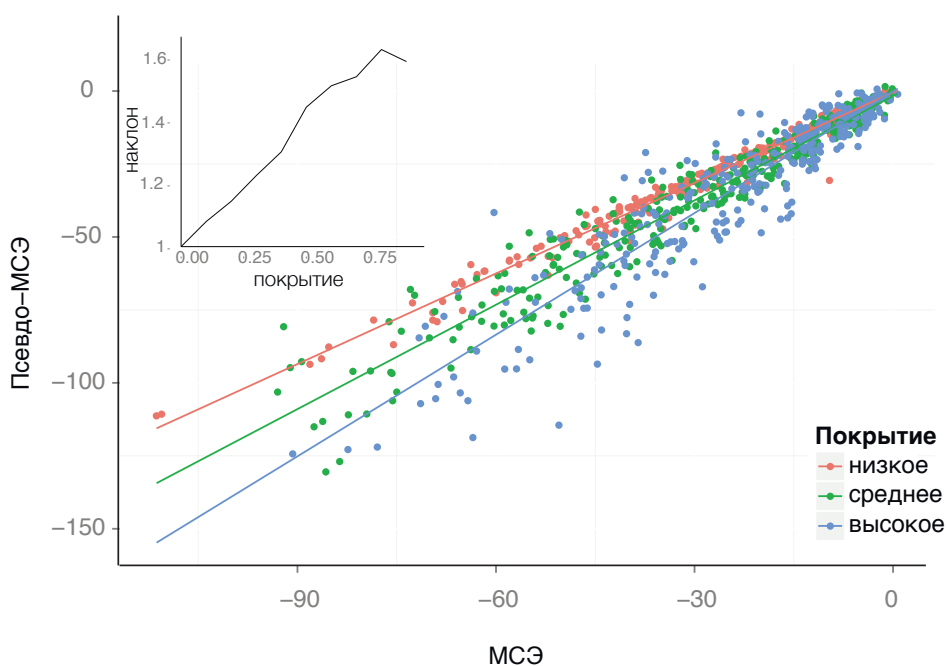


Рис. 4. Зависимость псевдо-МСЭ от МСЭ. Цвет соответствует уровню покрытия в эксперименте PARS. Линии – регрессионные прямые, для каждого уровня покрытия. Вставка на графике демонстрирует зависимость коэффициента пропорциональности  $a$  в линейной регрессии от уровня покрытия.

Рассмотрим, как именно меняется МСЭ при учете экспериментальных данных. На рисунке 5 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.

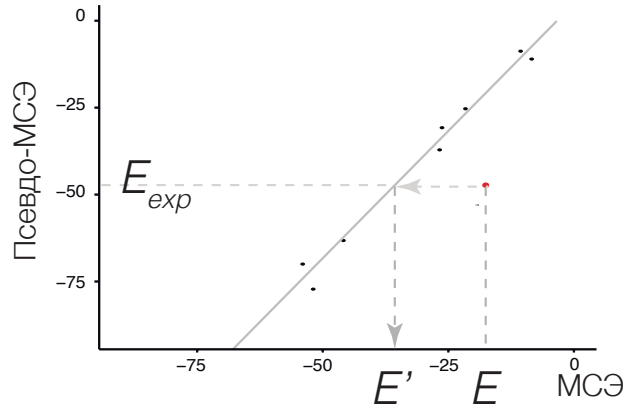


Рис. 5. Схема вычисления МСЭ' через псевдо-МСЭ на основании фоновой модели.

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:

$$\text{псевдоМСЭ} = a \cdot \text{МСЭ} + b \quad (4)$$

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 (вставка на рисунке 4).

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

$$a(cvr) = a'cvr + c' \quad (5)$$

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

$$\text{псевдоМСЭ} = a(cvr) \cdot \text{МСЭ} + b \quad (6)$$

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

$$\text{МСЭ}' = \frac{\text{псевдоМСЭ} - b}{a' \cdot cvr + c'} \quad (7)$$

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

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. 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.

В отличие от мРНК фрагментов, пробинг-Z-значения функциональных структурированных РНК статистически значимо ниже, чем Z-значения (тест Вилкоксона, р-значение равно  $2.4 \times 10^{-6}$ ), со средним сдвигом 0.64 стандартных отклонений. Большинство функциональных некодирующих РНК имеют пробинг-Z-значения более низкие, чем Z-значения (рисунок 6).

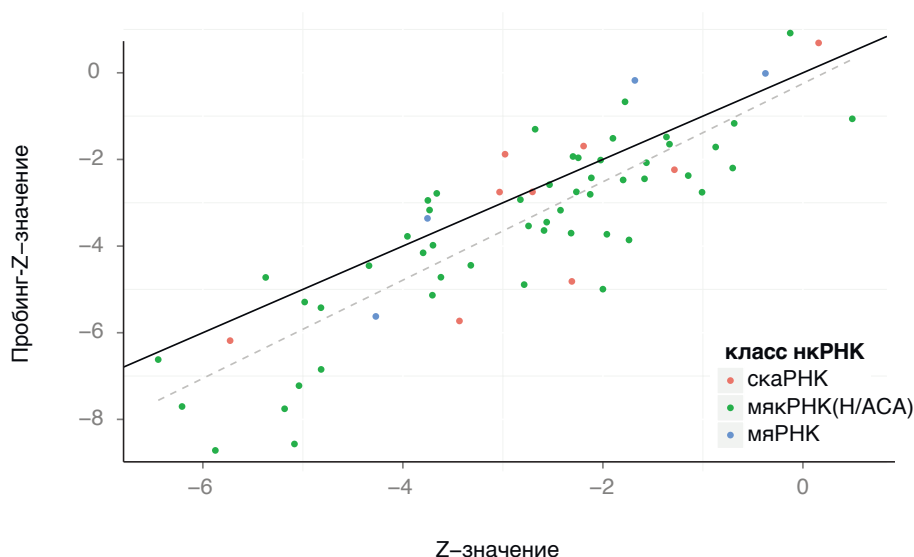


Рис. 6. Z-значения и пробинг-Z-значения функциональных некодирующих РНК. Черная прямая  $y=x$ , пунктирная серая прямая – регрессионная прямая по точкам.

Для оценки эффективности работы модифицированного алгоритма мы провели поиск локально структурированных элементов РНК в масштабе человеческого транскриптома с использованием экспериментальных данных PARS (Wan и др., 2014). 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).

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 для этих двух случаев, но только около 70% элементов совпадает (с точностью до границ) между двумя случаями. 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 7 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 (Агацjo и др., 2012). Интересно, что на границу 5' НТО и кодирующей части попадает меньшее количество структурированных элементов при использовании экспериментальных данных, чем без использования экспериментальных данных.

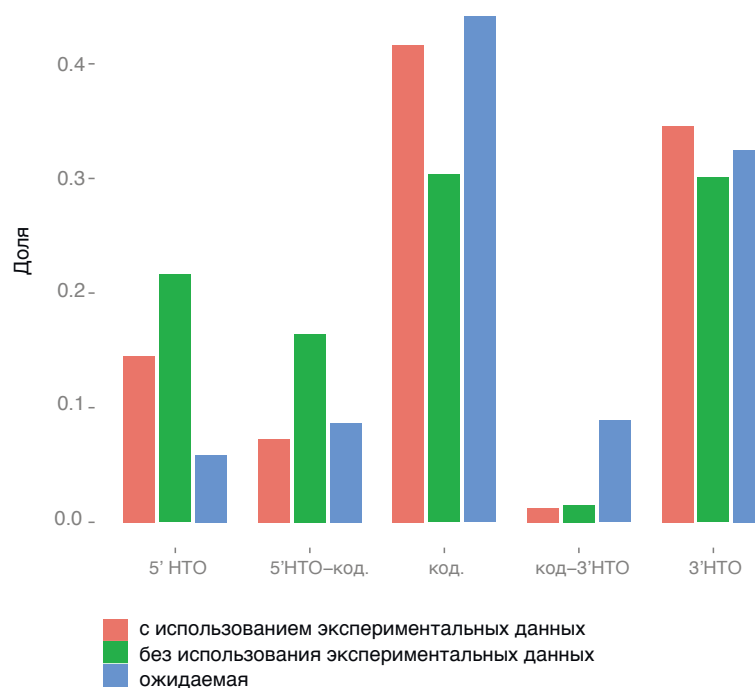


Рис. 7. Распределение структурированных элементов по разным областям мРНК, усреднение по всем транскриптам.

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* (Kertesz и др., 2010) и *Arabidopsis thaliana* (Li и др., 2012, The Plant cell) 5' НТО и 3' НТО менее структурированы, чем кодирующие области, в то время как обратное наблюдается для мРНК человека, *Drosophila melanogaster* и *Caenorhabditis elegans* (Li и др., 2012, Cell reports). 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.

Для дальнейшего анализа найденных структур и доказательства их функциональности мы сравнили результаты нашего анализа с результатами работы программы Evofold (Pedersen и др., 2006), использующей



множественные выравнивания и контекстно-свободные грамматики для поиска структурированных элементов РНК. Среди предсказаний Evofold, лежащих в рассматриваемых нами транскриптах, 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 (рисунок 8).

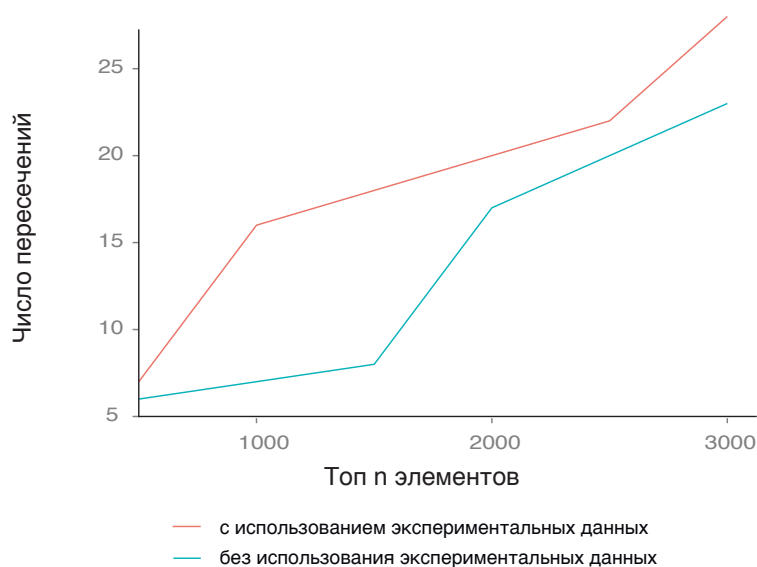


Рис. 8. Число уникальных пересечений между предсказаниями структур Evofold и структурированными элементами, найденными с учетом (красная кривая) и без учета (синяя кривая) экспериментальных данных.

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 случаев: 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 нуклеотидов мы определяли на основании структуры, полученной в программе Evofold). Было показано, что спаренные нуклеотиды в «новых» структурах имеют значительно большие PARS значения ( $p$ -значение=0.013), чем неспаренные, что говорит об общей согласованности экспериментальных данных и структур (рисунок 9). Напротив, PARS значения спаренных и неспаренных нуклеотидов в «потерянных» структурах не различаются, что говорит об отсутствии информации о структурированности в данных.

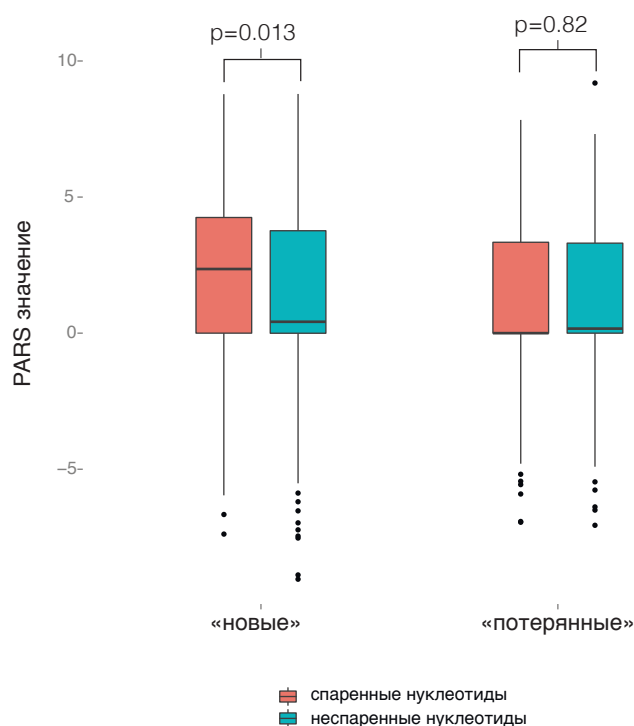


Рис. 9. Распределения PARS значений спаренных и неспаренных нуклеотидов для «новых» и «потерянных» структур.

Таким образом, в ходе исследования результатов включения экспериментальных данных в поиск структурированных элементов РНК в транскриптом человека было показано, что использование экспериментальных данных при поиске структурированных элементов РНК позволяет улучшить качество предсказания за счёт предсказания элементов, которые обладают

структурой с низкой свободной энергией, согласующейся с экспериментальными данными. Данный вывод подтверждается сравнением результатов с результатами программы, являющейся на данный момент «золотым стандартом» предсказания структур РНК на основании выравниваний.

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

## 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 реферируемых 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 Biology – ECCB'14). Сборник тезисов, Страсбург, Франция – 2014.
- 4) Vinogradova S. Probing data facilitates 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 (МССМВ'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