На правах рукописи

Солдатов Руслан Андреевич

Методы предсказания структурных элементов РНК

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва, 2015

Работа выполнена в Учебно-научном центре "Биоинформатика" Федерального государственного бюджетного учреждения науки Институт 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 физико-математических наук
Галзитская Оксана Валерьевна
(Федеральное государственное бюджетное учреждение науки Институт белка Российской академии наук)

кандидат физико-математических наук
Кулаковский Иван Владимирович
(Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук)

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт математических проблем биологии РАН

Защита диссертации состоится __ _____ 2015 года в __ часов на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институт проблем передачи информации им А.А. Харкевича Российской академии наук по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Институт проблем передачи информации им А.А. Харкевича Российской академии наук, а также на сайте ИППИ РАН по адресу: _____

Автореферат разослан __ _____ 2015 г.

Ученый секретарь диссертационного совета

доктор биологических наук, профессор Рожкова Г.И.

Общая характеристика работы

Актуальность работы. Развитие экспериментальных технологий привело к интенсивному росту количества секвенированных геномов. Появление большого количества данных выдвигает на передний план задачу эффективного и массового предсказания функциональных элементов, таких как белок-кодирующие области или регуляторные структуры РНК.

За последние двадцать лет произошел разительный прорыв в области биологии РНК, сопровождающийся открытием десятков новых классов некодирующих РНК. Как правило, РНК осуществляет регуляторную функцию с помощью своей вторичной структуры. В настоящий момент известно, что структурные РНК играют важную роль фактически во всех основных клеточных процессах, таких как сплайсинг, трансляция, вирусная репликация и модификация хроматина. Современные каталоги функциональных РНК существенно не полны и, как считается, геномы содержат значительно больше структурных РНК чем сейчас известно. К настоящему моменту не существует подходящих экспериментальных методов для предсказания и классификации новых структурных РНК. Как следствие, исследования фокусируются на вычислительных подходах по предсказанию функциональных структур РНК.

Если недоступен подходящий набор ортологичных последовательностей для сравнительного анализа, то главным сигналом функциональной структуры является её структурированность, то есть способность образовывать компактную пространственную структуру. Структурированность отражает наличие многих внутримолекулярных контактов и низкую свободную энергию молекулы. Новые структурные элементы РНК могут варьироваться от маленьких шпилек до больших многодоменных структур. Однако, несмотря на интенсивное развитие области, не существует общего алгоритма по предсказанию расположения и размера структурных элементов РНК только на основе их структурированности. Кроме того, выявления структурных элементов РНК различного диапазона размера и сложности на масштабе генома сталкивается с проблемой вычислительной сложности.

С другой стороны, неявным свидетельством функциональности структурного элемента РНК является факт сохранения структурированности в ходе эволюции. На практике, как правило, известно филогенетическое дерево с ортологичными последовательностями на листьях, и возникает задача детектирования отбора на структурные свойств РНК по наблюдаемым значениям структурированности на листьях.

Цели и задачи исследования. Целью исследования была разработка новых методов предсказания структурных элементов РНК в геноме на основе как анализа свойств

отдельных фрагментов генома, так и сравнительно-геномного подхода. Более конкретно, в работе решаются две задачи:

1. Создание алгоритма по предсказанию потенциальных структурных элементов РНК широкого диапазона размера и сложности на основе энергетических свойств, применимого для вычислительно эффективного полногеномного анализа.

2. Создание модели эволюции структурированности на основе диффузионных процессов. Разработка на основе модели сравнительно-геномного подхода к предсказанию функциональных структурных элементов РНК.

Научная новизна и практическая значимость. Традиционные подходы к выявлению структурных РНК сканируют геном с фиксированным окном, вычисляя статистическую достоверностью функциональной структуры РНК в каждом окне; далее окна с весомой статистической достоверностью отбираются как потенциальные структурные РНК. Сканирование фиксированным окном позволяет выявить структурные РНК с размером близким к длине окна, что расходится с представлением о широком диапазоне возможных длин регуляторных РНК. Мы разработали подход, который вычисляет статистическую "силу" вторичной структуры РНК (Z-значение) для каждого геномного сегмента до определенной длины и выбирает наиболее достоверные сегменты вне зависимости от их длины и расположения. Z-значение отражает структурированность сегмента, вычисленную на основе свободной энергии его вторичной структуры и учитывающую статистические свойства последовательности. На основе этого подхода определяются локально-оптимальные структурированные сегменты, что позволяет отказаться от априорного выбора размера окна. Подход реализован в виде программы RNASurface, которая демонстрирует значительное улучшение качества предсказания структурных РНК по сравнению с известным программами, имея при этом сравнимое время работы на масштабе генома.

RNASurface может применяться для массового анализа структурированных участков в геномах, или как основа для дальнейшего сравнительно-геномного анализа. RNASurface также вычисляет профиль структурированности РНК вдоль последовательности, который может быть использован для корреляции с другими профилями (например: границы белок-кодирующей области, профиль связывания рибосом или других РНК-связывающих белковых комплексов) для формирования новых биологических гипотез.

Неявным свидетельством функциональности структурного элемента РНК является факт сохранения структурированности в ходе эволюции. Наблюдаемые Z-значения набора ортологичных последовательностей являются следствием эволюционного процесса, в ходе которого происходят малые изменения последовательностей и их Z-значений, вдоль филогенетического дерева. Чтобы

аккуратно учесть этот процесс при предсказании функциональных структурных элементов РНК, была разработана диффузионная модель эволюции Z-значений и других количественных характеристик, неявно зависящих от последовательности. На основе модели введены статистики, описывающие статистическую значимость наблюдаемых Z-значений. В отличие от эвристических идей, на которые опираются стандартные сравнительно-геномные методы, данный подход опирается на строгую эволюционную модель. Наш метод реализован в виде библиотеки java-классов и, кроме Z-значений, применим для широкого класса сравнительно-геномных задач, таких как поиск сайтов связывания транскрипционных факторов, белок-кодирующих сегментов и т. д. Работа метода показала значительные преимущества использования диффузионной модели для повышения надежности предсказания структурных элементов РНК.

Степень достоверности и апробация результатов. По материалам диссертации опубликовано 2 статьи в рецензируемых научных журналах. Результаты работы были представлены на международных конференциях RECESS’12, RECESS’13, MCCMB’13, Benasque'15 и российских конференциях ИТИС’12, ИТИС’13, ИТИС'14 и 54-ая научная конференция МФТИ'11.

Структура и объем диссертации. Диссертация состоит из введения, обзора литературы, 2 глав, выводов и библиографии. Общий объем диссертации 109 страниц, из них 94 страниц текста, включая 36 рисунков и 4 таблицы. Библиография включает 113 наименование на 10 страницах.

Содержание работы

RNASurface: эффективный алгоритм предсказания локально-оптимальных структурированных сегментов РНК.

Вторичная структура РНК имеет регуляторную функцию в самых различных клеточных процессах, а предсказание сегментов РНК с функционально важными структурами является необходимым шагом для формирования гипотез о биологических механизмах. При отсутствии подходящего набора ортологичных последовательностей главным сигналом функциональной структуры является её компактная пространственная структура, выраженная низкой свободной энергией. Статистической мерой "необычности" свободной энергии является Z-значение относительно свободных энергий случайных последовательностей:

$$Z = \frac{E - \mu}{\sigma},$$

где E - минимальная свободная энергия, μ и σ - среднее и стандартное отклонение распределения минимальных свободных энергий для последовательностей такой же длины и с тем же динуклеотидным составом. Минимальная свободная энергия вычисляется алгоритмом Зукера, в основе которого лежит метод динамического программирования. Стекинг взаимодействия между соседними комплементарными парами вносят важный вклад в стабильность вторичной структуры, поэтому сохранение их энергетического вклада в случайных последовательностях требует поддержание динуклеотидного состава.

Сканирование длинных последовательностей в поисках локальных структурированных участков позволяет определить неизвестные регуляторные элементы, которые представляют особый интерес. Известные методы поиска структурированных РНК сканируют последовательность фиксированным окном: для каждой подпоследовательности фиксированного размера с некоторым шагом вычисляется стабильность вторичной структуры РНК (выраженная в виде Z-значения или свободной энергии), после чего выбираются наиболее значимые окна (Рисунок 1А). Как следствие, только структурные РНК с длинами близкими к длине окна будут предсказаны, а размер окна и шага имеют критическое значение на качество предсказания. Мы представляем подход, реализованный в виде программы RNASurface, который позволяет эффективно предсказывать локально наиболее структурированные участки в длинных последовательностях (Рисунок 1Б). По сравнению с известными программами, RNASurface демонстрирует значительно лучшее качество предсказания известных структурных РНК и представляет набор инструментов для формулирования новых биологических гипотез.

В основе подхода лежит следующая идея. При оценке оптимальной свободной энергии E последовательности S, алгоритм Зукера рекуррентно вычисляет оптимальные энергии E_{ij} всех подпоследовательностей S_{ij} и записывает их в матрицу энергий (Рисунок 1Б). Поэтому, вычислив оптимальную энергию последовательности, получаем и оптимальные энергии всех её подпоследовательностей. Если известны параметры μ_{ij} и σ_{ij} фонового распределения свободных энергий участка S_{ij} то для этого участка известен и $Z_{ij} = \frac{\mathrm{E_{ij}} - \mu_{ij}}{\sigma_{ij}}$ и можно восстановить матрицу Z-значений (Рисунок 1Б).

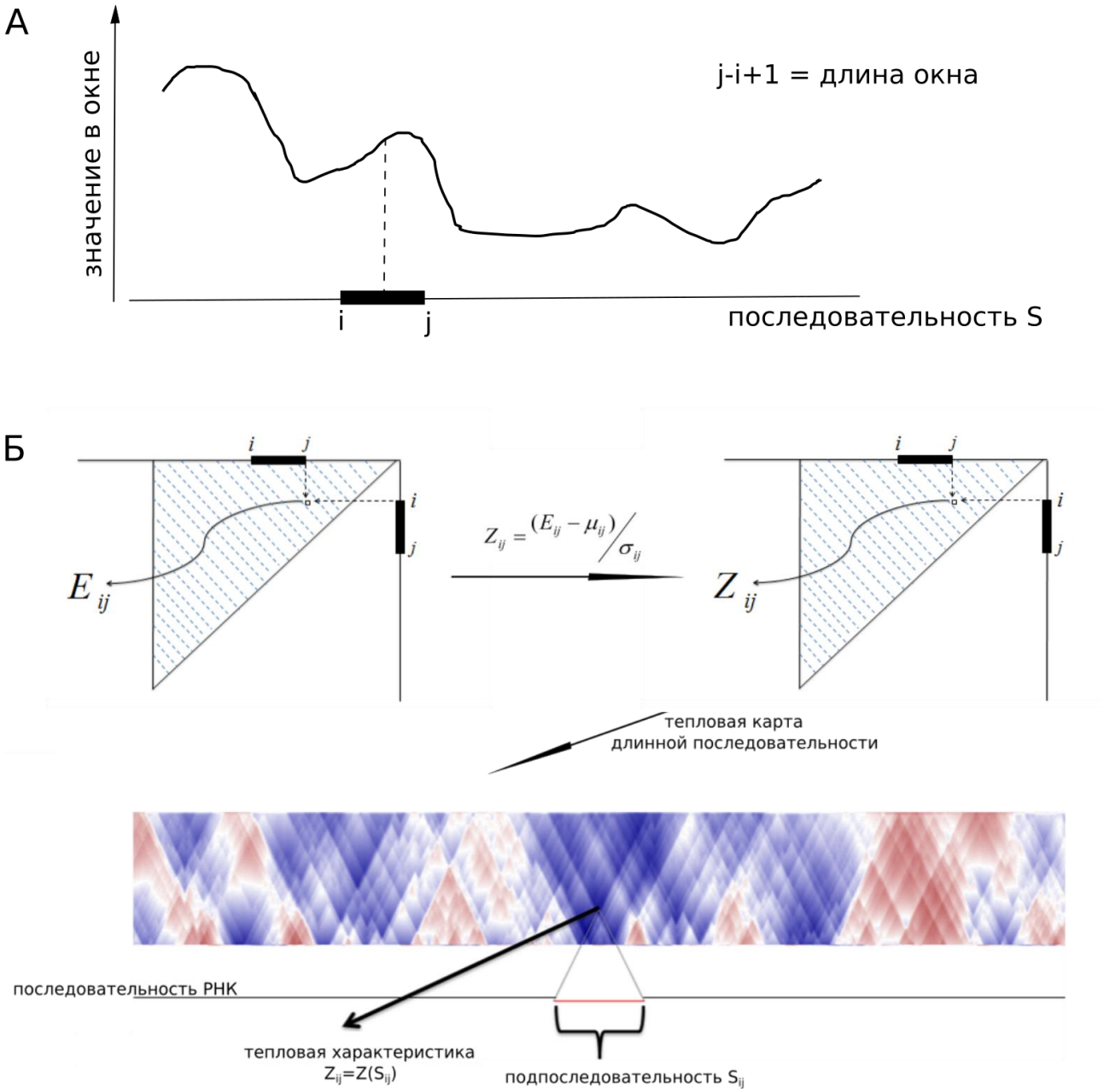


Рисунок 1. Сканирование последовательности фиксированным окном (А) и алгоритмом RNASurface (Б). А) Для каждого сегмента S_{ij}, имеющего размер заданного окна, вычисляется значение, после чего выбираются пики профиля. Б) По матрице энергий, полученной с помощью алгоритма Зукера, вычисляется

матрица Z-значений на основании набора регрессий. Для удобства визуализации матрица Z-значений представляется в виде тепловой карты, на которой структурированные участки соответствуют синему цвету, а неструктурированные - красному. Локальные пики тепловой карты соответствуют локально-оптимальным структурированным сегментам. Приведен пример тепловой карты последовательности в 1000 нк из генома *Bacillus subtilis* с окном L=200 нк, содержащей SAM рибо-переключатель посередине. Выделенная красным подпоследовательность имеет длину 100 нк, наклонные линии от неё ведут в точку соответствующую её структурированности Z_{ij}.

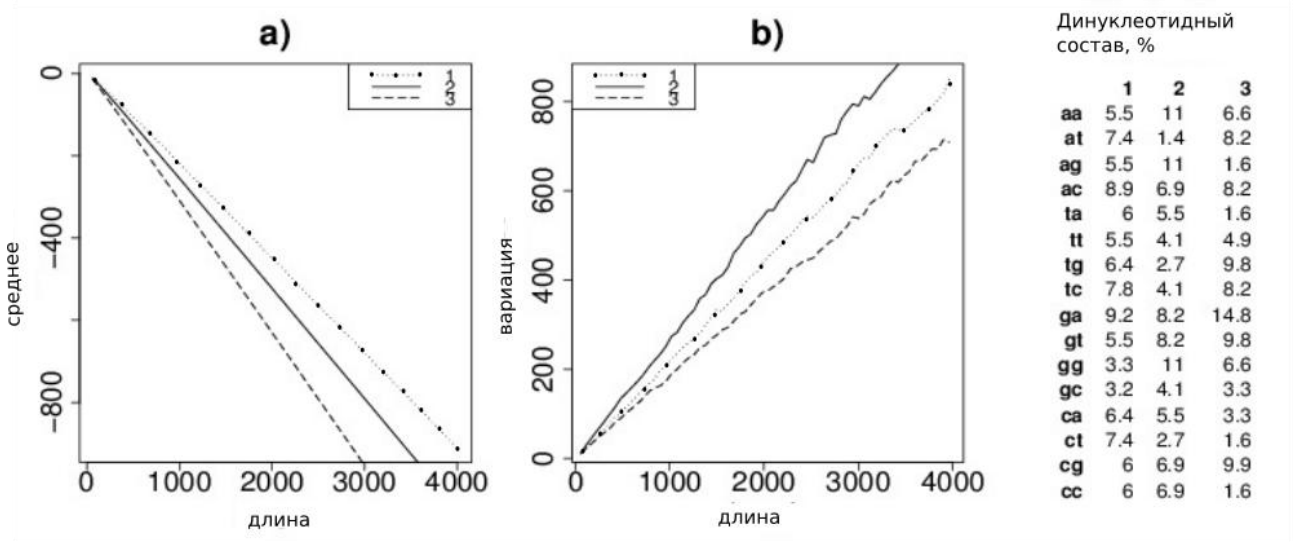
Свободная энергия отрицательна, и чем она ниже, тем более структурирована последовательность; поэтому низкие отрицательные Z-значения соответствуют хорошо структурированным последовательностям. Если точка на матрице Z-значений имеет Z-значение ниже чем её соседи, то она представляет локальный оптимум. Небольшие сдвиги границ подпоследовательности, соответствующей локально-оптимальной точке, только увеличивают Z-значение и ухудшают структурированность. Таким образом, локально-оптимальные сегменты являются логичными кандидатами на роль потенциальных структурных РНК, и их использование устраняет проблемы определения границ структурированных РНК, а также параметров окна и шага сканирования.

Аккуратное и быстрый метод вычисления Z-значений. Массовая оценка параметров μ_{ij} и σ_{ij} фонового распределения свободных энергий является вычислительно трудоемкой задачей, при этом передовой подход, использующий регрессию опорных векторов для оценки параметров, не позволяет адаптировать метод к полногеномному анализу. Параметры зависят от длины и динуклеотидного состава последовательности. Мы предлагаем следующий двухуровневый метод:

1) Мы показали, что для фиксированного динуклеотидного состава параметры фонового распределения энергий в первом приближении увеличиваются линейно с длиной последовательности. Как следствие, достаточно табулировать значения параметров для нескольких длин, а для остальных использовать линейные приближения.
2) При фиксированной длине, среднее и дисперсия достаточно сложно зависят от динуклеотидного состава, образуя некоторую поверхность в пространстве динуклеотидных частот $(d_1, \dots, d_{16})$. Тем не менее, эта поверхность хорошо приближается набором квадратичных регрессий. А квадратичную регрессию можно очень быстро рекуррентно пересчитывать при сканировании вдоль генома.

Использование этих соображений позволяет вычислять матрицу Z-значений точно и за малое время по сравнению с пересчетом энергий алгоритмом Зукера.

Как показывают симуляции, при фиксированном динуклеотидном составе $\mu(l)$ и $\sigma^2(l)$ приблизительно линейно зависят от длины последовательности l.



Динуклеотидный состав, %

	1	2	3
aa	5.5	11	6.6
at	7.4	1.4	8.2
ag	5.5	11	1.6
ac	8.9	6.9	8.2
ta	6	5.5	1.6
tt	5.5	4.1	4.9
tg	6.4	2.7	9.8
tc	7.8	4.1	8.2
ga	9.2	8.2	14.8
gt	5.5	8.2	9.8
gg	3.3	11	6.6
gc	3.2	4.1	3.3
ca	6.4	5.5	3.3
ct	7.4	2.7	1.6
cg	6	6.9	9.9
cc	6	6.9	1.6

Рисунок 2. Зависимость среднего (А) и дисперсии (Б) распределения минимальной свободной энергии в зависимости от длины последовательности. Последовательности генерировались с тремя различными динуклеотидными составами (марковскими моделями первого порядка). Вычисление проводились для длин от 20 нк до 4000 нк, для каждой длины были сгенерированы 500 последовательностей и вычислены их свободные энергии программой RNASlider, на основании которых оценены среднее и дисперсия распределения свободных энергий.

При фиксированной длине параметры зависят от вектора динуклеотидных частот $(d_1, \dots, d_{16})$. Пространство динуклеотидных частот было разбито на 27 областей, в каждой из которых зависимость среднего $\mu(d_1, \dots, d_{16}|l)$ и дисперсии $\sigma^2(d_1, \dots, d_{16}|l)$ от частот $(d_1, \dots, d_{16})$ было аппроксимировано квадратичной регрессией:

$$\sum_{1 \le i < j \le 16} c_{ij} d_i d_j.$$

Для оценки общего качества аппроксимации Z-значений, случайным образом были выбраны 24000 последовательностей в диапазоне длин от 60 до 2400 нуклеотидов из геномов человека и дрозофилы. Z-значения свободной энергии каждой последовательности были вычислен как с помощью интенсивных симуляций, так и с помощью регрессий. Аппроксимация демонстрирует очень высокое сходство с реальным Z-значением ($R^2 = 0.997$). При этом качество предсказания одинаково

высокое для всех длин, а средняя ошибка (модуль разницы между аппроксимацией и реальным Z-значением) составляет 0.1.

Качество предсказания и эффективность RNASurface.

Для оценки качества RNASurface был проведен полногеномный анализ грам-положительной бактерии *Bacillus subtilis,* обладающей разнообразными механизмами РНК-регуляции. В этой бактерии экспрессия генов регулируется широким набором рибо-переключателей и Т-бокс РНК, а более 1500 генов опираются на РНК-зависимую терминацию транскрипции на основе шпилек. Таким образом, *Bacillus subtilis* является удобным и интересным объектом для полногеномного анализа структурных РНК. Среди ряда подходов, которые сканируют последовательность в поисках структурированных РНК, только RNALfoldz алгоритмически адаптирован к массовому анализу и использует статистическую оценку значимости предсказаний (Z-значение). На основе известных структурных РНК из Bacillus subtilis мы сравнили качество предсказания RNASurface с RNALfoldz. Мы использовали 187 аннотированных структурных РНК из базы данных Rfam, из них 43 рибо-переключателя, 13 Т-бокс РНК, 6 лидерных РНК (лидерные регуляторные РНК рибосомальных белков), 20 малых РНК, 85 тРНК и 20 5S рРНК, а также более 2000 предсказанных ро-независимых терминаторов. Выдача обеих программ состоит из списка координат структурированных сегментов с низким Z-значений, однако RNALfoldz использует окно фиксированного размера. Если аннотированная структурная РНК имеет высокий уровень пересечения с каким-либо предсказанным сегментом, то считаем её правильно предсказанной. Более формально, мер Жаккара двух сегментов S_1 и S_2 определяется как:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

где $|S_1 \cap S_2|$ и $|S_1 \cup S_2|$ - количество нуклеотидов в пересечении и объединении двух сегментов соответственно. Считаем, что структурная РНК правильно предсказана, если для неё существует сегмент с $J \geq 0.75$. Количество правильных предсказаний обозначим через TP (true positive), а количество ложных предсказаний через FP (false positive). Все участки генома вне известных структурных РНК считаем заведомо не являющимися структурными РНК, их обозначим через TN (true negative). Отметим, что в геноме может быть немалое количество неаннотированных структурных РНК, что завышает TN и FP; однако общее количество структурных РНК в геноме мало по сравнению с его размером, поэтому погрешность в оценке TN мала. Качество предсказаний представлено в виде ROC кривой сравнения sensitivity (доля предсказанных структурных РНК) и FPR (доля неправильно предсказанных сегментов), и в виде соотношения sensitivity и PPV (доля структурных РНК среди всех

предсказаний) (Рисунок 3А,Б). Sensitivity, PPV и FPR определяются следующим образом:

$$PPV = TP/(TP + FP),$$

$$FPR = FP/(FP + TN),$$

$$sensitivity = TP/\text{все структурные РНК}.$$

При одинаковом количестве ложных предсказаний RNASurface детектирует значительно большую долю структурных РНК (Рисунок 3А), особенно при малом количестве ложных предсказаний; так на уровне FPR=1% RNASurface детектирует ~30% структурных РНК, в то время как RNALfoldz чуть более 15%. Доля структурных РНК среди всех предсказаний также значительно выше у RNASurface (Рисунок 3Б).

Мы изучили как изменяется доля предсказанных структурных РНК в зависимости от размера окна L как входного параметра RNASurface и RNALfoldz (Рисунок 3В). Чувствительность RNASurface монотонно растет с размером окна, в то время как чувствительность RNALfoldz монотонно растет до 150 нк, после чего падает. На уровне статистической достоверности FPR в 20%, RNASurface позволяет предсказать почти все структурные РНК при длине окна L от 200 нк, в то время как чувствительность RNALfoldz при любом размере окна принципиально не выходит за 75%.

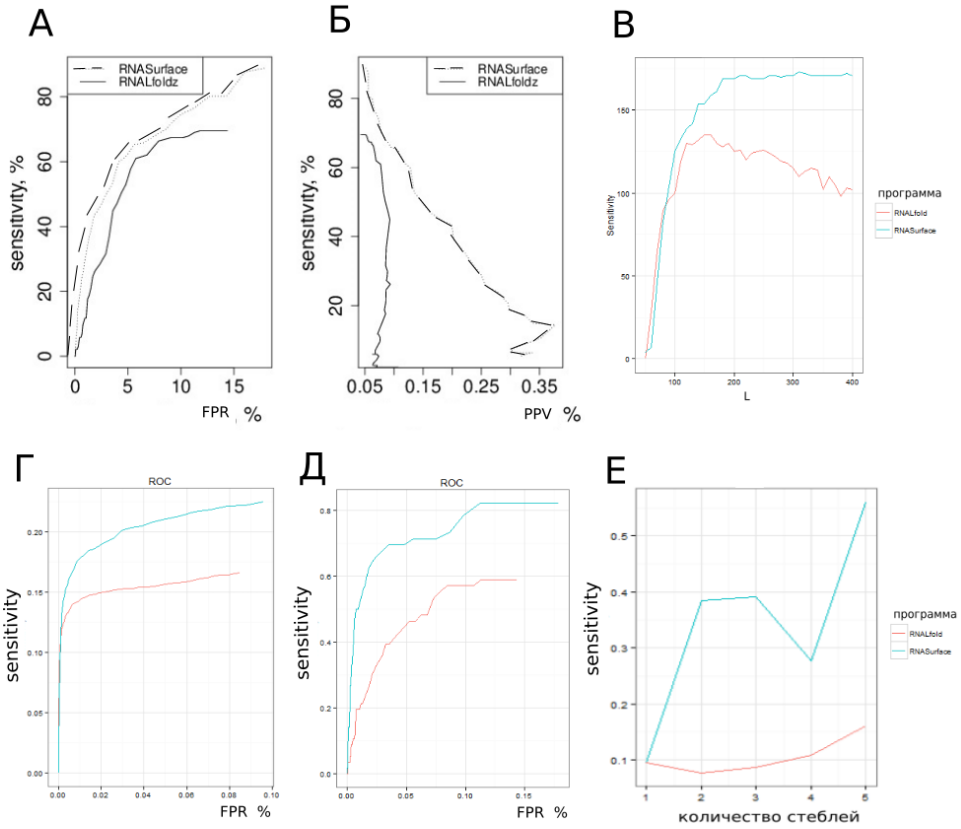


Рисунок 3. Сравнение качества предсказания структурных РНК программами RNASurface и RNALfoldz. А) ROC кривые сравнения sensitivity и FPR. Б) Кривая сравнения sensitivity и PPV. В) Зависимость чувствительности методов от размера окна. FPR зафиксирован на уровне 20%. Г) ROC кривая для ро-независимых терминаторов. Д) ROC кривая для рибо-переключателей и Т-бокс РНК. Е) Зависимость чувствительности методов от сложности структуры, определенной как количество составляющих её стеблей. FPR зафиксирован на уровне 1%

Регуляторные РНК имеют разнообразные формы, от маленьких шпилек до сложных многодоменных структур. Мы сравнили статистическую мощность программ (выраженную в ROC кривой) в зависимости от сложности вторичной структуры, при этом шпильки ро-независимых терминаторов были выбраны как пример простой структуры (Рисунок 3Г), а рибо-переключатели и Т-боксы как пример сложных структур (Рисунок 3Д). RNASurface превосходит RNALfoldz как на простых, так и на сложных вторичных структурах, однако обе программы значительно лучше определяют сложные структурные РНК, что имеет логичную вероятностную интерпретацию: один стебель имеет немалую вероятность образоваться по случайным причинам, в то время как формирование 4-5 стеблей на участке в несколько сотен нуклеотидов является очень маловероятным событием. Обобщая это наблюдение, сложность вторичной структуры, определенная как количество составляющих её стеблей, также коррелирует с качеством предсказания (Рисунок 3Е).

При пороге на Z-значение=-3, что охватывает 1% генома, предсказываются 30% структурных РНК. Количество предсказаний увеличивается до 62% при пороге на Z-значение=-2, что соответствует 5% генома. Таким образом, хотя RNASurface имеет значительное превосходство в своем классе задач, этот подход имеет слабую предсказательную мощность на масштабе генома. Отметим что классические структурные РНК рРНК и тРНК, участвующие в процессе синтеза белков, предсказываются плохо, в среднем 16%. В то время как регуляторные элементы в 5' нетранслируемой области - рибо-переключатели, Т-бокс РНК, регуляторные РНК в лидерных областях рибосомальных белков - имеют значительно лучшее качество предсказания, в среднем 48%. Это может означать, что для рРНК и тРНК важна не столько термодинамическая стабильность структуры, сколько определенный вид конформации для взаимодействия с другими белковыми комплексами.

Z-значение	Рибо-переключатели	Т-боксы	Лидерные	Малые РНК	тРНК	5S РНК	FPR, %	PPV, %
-1	79 (34)	92 (12)	67 (4)	75 (15)	95 (81)	100 (20)	18	0.05

-2	65 (28)	85 (11)	50 (3)	65 (13)	62 (53)	35 (7)	5	0.1
-2	44 (19)	69 (9)	33 (2)	35 (7)	16 (14)	15 (3)	1	0.25
Всего РНК	43	13	6	20	85	20		

Таблица 1. Количество и доля предсказанных структурных РНК, стратифицированных по классам, для разных порогов по Z-значению.

Теоретически, в среднем время работы RNASurface составляет $O(N \cdot L)$ на последовательности длины N с окном L. Сравнение практического времени работы RNASurface с программами RNASlider (на которую опирается RNASurface) и RNALfoldz на геноме *Bacillus subtilis* как функции от длины окна демонстрирует сравнимую скорость RNASurface с RNASlider, при этом ускорение в несколько раз относительно RNALfoldz.

Распределение структурированных сегментов по регионам

Регуляторные структуры РНК располагаются в специфических регионах геномы, например, вторичные структуры, регулирующие экспрессию (рибо-переключатели, Т-боксы и т. д.), находятся как правило в 5'НТО. Предсказанные локально-оптимальные сегменты были классифицированы по типам геномных регионов (Рисунок 4). Мы оценили перепредставленность структурированных сегментов в каждом из типов регионов (Таблица 2). В качестве нулевой модели предполагаем, что структурированные сегменты распределены равномерно вдоль генома.

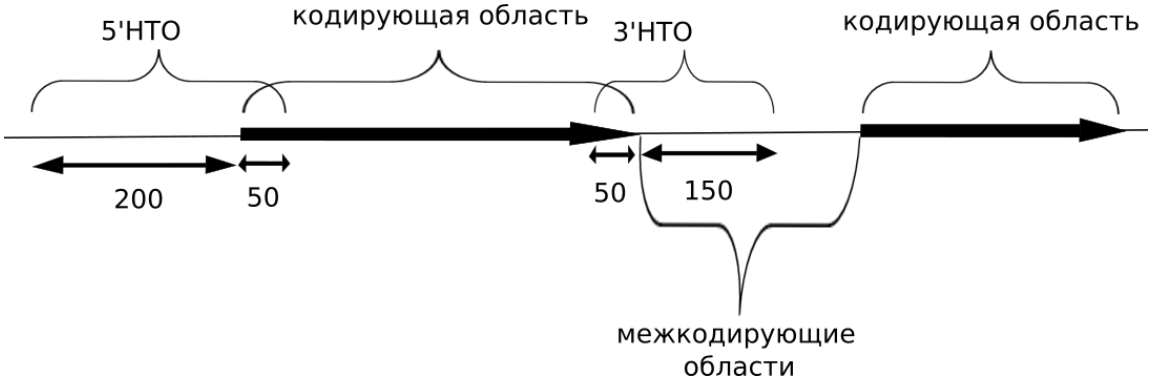


Рисунок 4. Разбиение на типы геномных регионов.

Распределение структурированных сегментов сильно неравномерно, с большой перепредставленностью в 5'НТО и 3'НТО генов (Таблица 2). Эта тенденция проявляется значительно сильнее при более строгих порогах на Z-значение. Так, при пороге -2 на Z-значение наблюдается качественная перепредставленность только в 3'НТО генах. Это связано с большим количеством (около 2000) структурированных ро-независимых терминаторов в этих областях, которые позволяют увидеть сигнал

структурированности даже при либеральном пороге. При низком пороге Z-значение=-5, наблюдается значительная перепредставленность в 3'НТО (более чем в 12 раз) и в 5'НТО (более чем в 3 раза). Перепредставленность в 5'НТО при строгом пороге на Z-значение связана с регуляторными РНК, их немного (поэтому нет сигнала при Z-значении=-2), но они обладают достаточно низким Z-значением. Строгий сигнал в межкодирующих областях отчасти объясняется их пересечением с 5'НТО и 3'НТО областями генов, но даже за вычетом их остается четырехкратная перепредставленность в межгенных областях. Это может быть вызвано сигналом от малых некодирующих РНК, которые по современным данным играют более весомую роль в бактериальном метаболизме чем считалось ранее.

Z-значение	-2	-3	-4	-5
Кодирующие области	0.91 (148441/162599)	0.68 (21050/30963)	0.37 (2010/5399)	0.15 (153/1017)
5'НТО	0.98 (6442/6601)	1.41 (1809/1280)	2.09 (480/230)	3.05 (131/43)
3'НТО	2.55 (6166/2420)	5.68 (2793/492)	10.11 (950/94)	12.5 (225/18)
Межкодирующие обасти	1.34 (16448/12249)	2.41 (5753/2387)	4.41 (1901/431)	12.52 (551/44)
Межкодирующие в опероне	1.71 (274/160)	3.18 (105/33)	5.86 (41/7)	13 (13/1)

Таблица 2. Плотность структурированных сегментов в различных регионах *Bacillus subtilis*. Для нескольких порогов на Z-значение вычислена перепредставленность структурированных сегментов в различных геномных регионах. Для каждой ячейки таблицы: первое и второе число в скобках соответствует наблюдаемому и ожидаемому количеству структурированных сегментов, а вне скобок представлено их отношение (перепредставленность структурированных РНК в данной области).

Сравнительно-геномный метод предсказания структурных элементов РНК на основе диффузионных процессов

Особый интерес представляют функциональные структурные элементы РНК. Косвенным признаком функциональности является факт сохранения структурированности РНК в ходе эволюции. Если дано филогенетическое дерево с ортологичными последовательностями на листьях, то сравнение Z-значений дает возможность выявить давление отбора на структурные свойства РНК. Нашей задачей является разработка эволюционного метода для предсказания и анализа функциональных структурных элементов РНК на основе Z-значений (Рисунок 5).

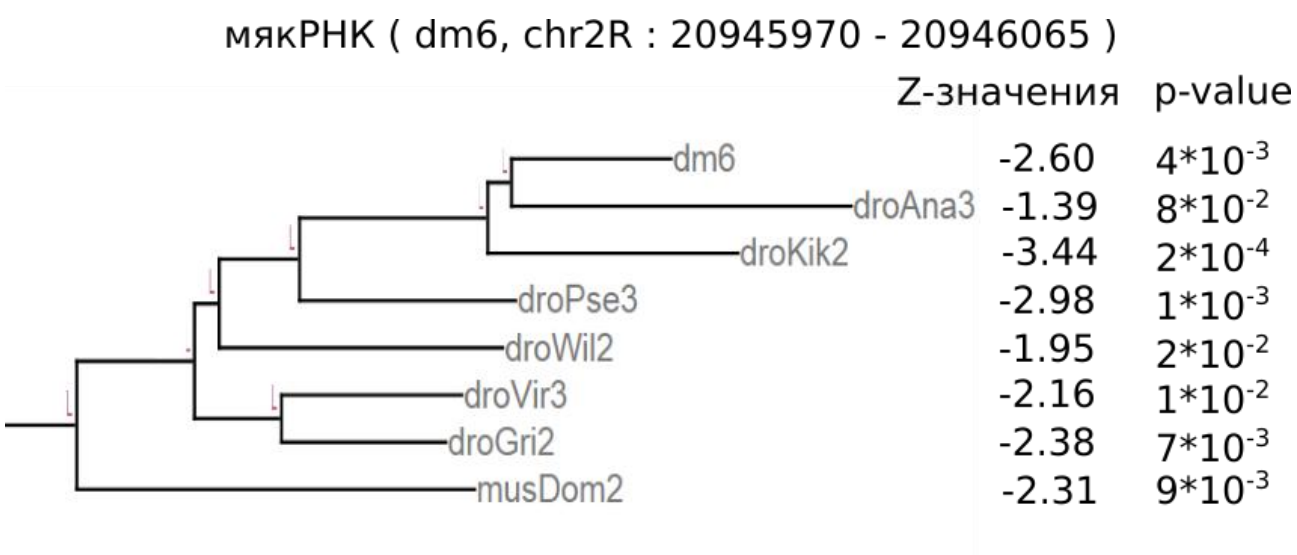


Рисунок 5. Демонстрация задачи на примере ортологичных последовательностей малой ядрышковой РНК (мякРНК) в геномах дрозофил. Для каждой последовательности вычислено Z-значение. Большое отрицательное Z-значение в каждом виде является косвенным свидетельством отбора на структурный элемент последовательности. Необходимо оценить статистическую значимость набора наблюдаемых Z-значений в последовательностях, связанных данной эволюционной историей.

Диффузионная модель Z-значений

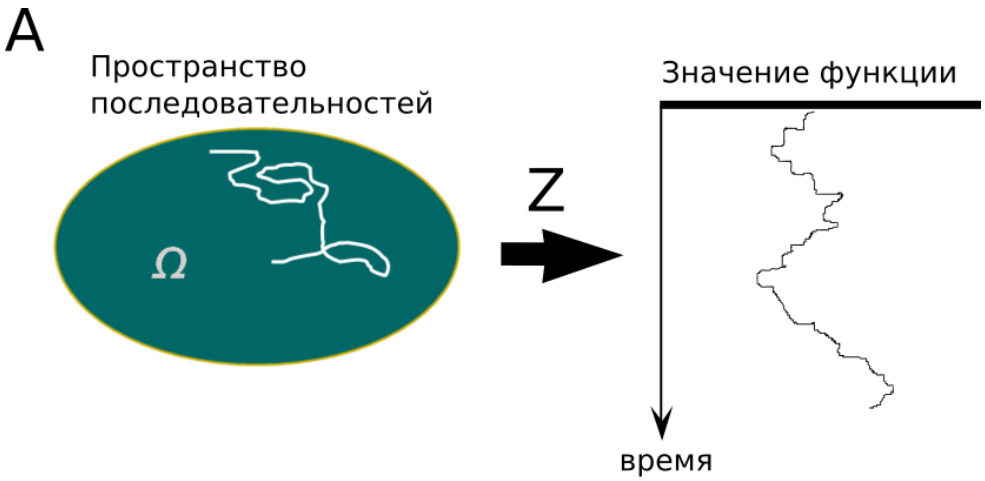
Диффузионный процесс является классической моделью для анализа эволюции частот аллелей в популяции и количественных характеристик видов (например, размер зубов). В данной работе мы адаптировали диффузионный процесс к анализу эволюции Z-значений. Z-значение вычисляется по последовательности

$$z: \Omega \to R,$$

где Ω - пространство последовательностей. а R - вещественные числа. Эволюцию последовательности можно представить как случайные блуждания в пространстве Ω. В таком случае, функция $z(s)$ будет описывать случайные блуждания в R (**Ошибка! Источник ссылки не найден.**А). При этом изменение $z(s)$ вероятней будет происходить в сторону значений с большей плотностью последовательностей (**Ошибка! Источник ссылки не найден.**Б). Качественно, случайное изменение $z(s)$ в "силовом поле" описывается диффузионным процессом:

$$dz = a(z)dt + b(\mathrm{z})dB_t$$

где $a(z)$ и $b(z)$ - функции сноса и диффузии соответственно, а $B_t \sim N(0, t)$.



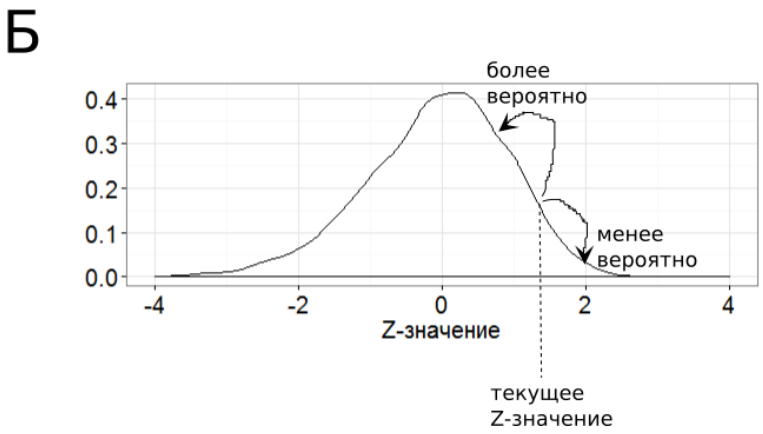


Рисунок 6. Модель эволюции Z-значений. А) Эволюция последовательности представляется в виде случайного блуждания, которое порождает случайные блуждания Z-значений. Б) Разные Z-значения имеют разную частоту встречаемости среди последовательностей. Изменение текущего значения в меньшую и большую сторону зависит от распределения Z-значений.

Пусть диффузионное уравнение имеет некоторое решение $\rho(z,t|y)$ с начальным условием $z(0)=y$: вероятность наблюдать значение z через время t, запустив процесс со значения y. Вид решения $\rho(z,t|y)$ зависит от $a(z)$ и $b(z)$. В то время как в предыдущих подходах эти параметры выбирались исходя из теоретических соображений, в данном методе зависимость Z-значений от последовательности позволяет вычислить $a(z)$ и $b(z)$ на основе модели эволюции последовательности. Мы разработали набор инструментов для оценки этих параметров по нарезке ортологичных участков генома, или исходя из модели эволюции последовательности. Анализ $a(z)$ и $b(z)$ в рамках разных подходов показывает, что параметры хорошо приближаются линейной и постоянной функциями соответственно (Рисунок **7**). Таким образом, в дальнейшем мы будем предполагать, что эволюция Z-значений описывается процессом

$$dz = -zdt + bdB_t,$$

который называется процессом Орнштейна-Уленбека (ОУ) и его решение $\rho(z,t|y)$ имеет нормальное распределение.

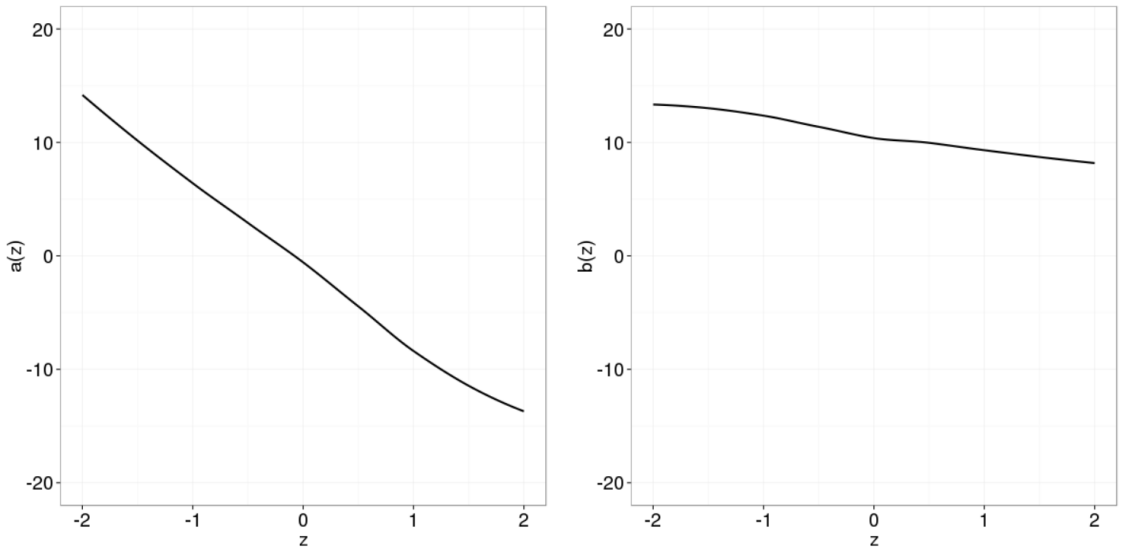


Рисунок 7. Вид параметров a(z) и b(z), вычисленный на основе нарезки ортологичных последовательностей геномов *Drosophila melanogaster, yakuba и ananassae*. Вид функций не меняется при симуляции в рамках заданных моделей эволюции последовательностей.

Для данного филогенетическое дерева с наблюдаемыми значениями $z_1, \dots, z_n$ на листьях (пример, Рисунок 5) плотность вероятности $\rho(z_1, \dots, z_n)$ вычисляется с использованием техники филогенетических марковских процессов и имеет многомерное нормальное распределение:

$$\rho(z_1, \dots, z_n) = \frac{\sqrt{detA}}{2\pi^{n/2}} e^{\frac{-1}{2}z'Az},$$

где $z = (z_1, \dots, z_n)$, A - матрица ковариаций, вычисляемая рекурсивно на основании филогенетического дерева.

На основании модели введем ряд статистик для анализа значимости наблюдений. Статистика

$$r^2 = z'Az.$$

монотонно изменяется с вероятностью $\rho(x_1, \dots, x_n)$ и подчиняется распределению Хи квадрат с n-1 степенью свободы.

Значение y предковой последовательности не известно, но если доступны наблюдения на листьях дерева, то апостериорное распределения предкового состояния вычисляется по формуле Байеса:

$$p(y|z_1, \dots, z_n) = \frac{\rho(z_1, \dots, z_n|y)p(y)}{\rho(z_1, \dots, z_n)},$$

и имеет нормальное распределение с некоторыми параметрами (μ, σ), где $p(.)$ - распределение Z-значений. Параметр μ можно интерпретировать как влияние "сноса" наблюдаемых значений признаков на предковое значение и представить следующим образом:

$$\mu = \sum a_i x_i,$$

где параметры a_i алгоритмически вычислимы на основе структуры дерева и длин веток. Эта статистика имеет нормальное распределение с нулевым среднем и алгоритмически вычислимой дисперсией. Так, примененный к мякРНК на Рисунок 5, диффузионный подход оценивает значимость наблюдений на уровне p-value < $5 \cdot 10^{-7}$. В анализе следующего раздела будет использоваться комбинация выше введенных статистик:

$$\log\big(1 - \mathrm{F(r)}\big) \cdot \big(1 - \mathrm{F(\mu)}\big),$$

где F(.) - кумулятивная функция распределения соответствующей статистики.

Таким образом, в данной работе разработан метод анализа Z-значений, который позволяет:

1) Оценить параметры диффузионного процесса в рамках данной модели эволюции или по нарезке ортологичных последовательностей их генома
2) Вычислить статистики для оценки значимости наблюдаемых значений.

В данной работе представлен подход для анализа Z-значений, при этом метод также применяется к другим количественным характеристикам, неявно зависящим от последовательности, например, силе сайта связывания транскрипционного фактора, белок-кодирующему потенциалу РНК.

Диффузионная модель улучшает надежность предсказания некодирующих РНК

Мы оценили надежность предсказания известных некодирующих РНК (нкРНК) генома *Drosophila melanogaster* при использовании диффузионной модели на филогенетическом дереве мух (*Drosophila melanogaster, ananassae, kikkawai, pseudoobscura, willistoni, virilis, mojavensis, grimshawi, Musca domestica*). Оценка параметров диффузионного процесса была получена на основе нарезки ортологичных последовательностей трех геномов дрозофил (Рисунок 7). Для того, чтобы оценить преимущества сравнительно-геномного анализа, мы также оценили качество предсказания нкРНК по их Z-значениям.

Для анализа были выбраны четыре широких класса нкРНК: микроРНК, малые ядерные РНК (мяРНК), малые ядрышковые РНК (мякРНК) и тРНК (Таблица 3). Из 849 нкРНК только 166 имеют ортологичные последовательности хорошего качества во всех геномах рассматриваемых видов. Для каждого класса нкРНК мы сформировали контрольную группу ортологичных участков с таким же распределением уровня консервативности вдоль рассматриваемого дерева. Качество предсказания представлено в виде ROC-кривой зависимости TPR от FPR, где TPR - доля предсказанных нкРНК среди всех нкРНК, а FPR - доля ложно предсказанных контрольных участков среди всех контрольных участков.

тип РНК	аннотированные	консервативные	расхождение	AUC, Z-значения	AUC, диффузия
микроРНК	238	47	0.06	0.97	0.98
мякРНК	288	62	0.12	0.67	0.98
мяРНК	31	6	0.05	0.79	0.92
тРНК	292	51	0.01	0.73	0.69
Общее	849	166	0.06	0.8	0.91

Таблица 3. Статистика разных классов некодирующих РНК. Для четырех классов нкРНК представлено их количество в геноме *Drosophila melanogaster* (аннотированные), количество консервативных нкРНК, прошедших фильтрацию (консервативные), средняя доля замен в нкРНК вдоль дерева (расхождение), качество предсказания с помощью Z-значений и диффузионной моделью выраженной в AUC.

Качество предсказания диффузионной модели, выраженное как площадь под ROC-кривой (AUC), значительно улучшается по сравнению с предсказанием на основе Z-значений (Рисунок 8A, Таблица 3Таблица 1). Отметим, что если предсказания практически не улучшаются для тРНК и микроРНК, то мы наблюдаем значительное улучшение качество предсказания мякРНК (Рисунок 8). Мощность сравнительно-геномного анализа зависит от степени расхождения последовательностей, при этом доля замен в мякРНК значительно превышает доли в микроРНК и тРНК (Таблица 3, столбец расхождение). В соответствии с этим, сохранение структурированности мякРНК при расхождении последовательности является сигналом отбора на структурные свойства, в то время как сохранение структурированности тРНК при сохранении последовательности не является дополнительным свидетельством эволюционного отбора на структурные свойства по сравнению с Z-значением. Отметим также, что при массовых анализах контрольная группа, как правило, значительно превосходит количество нкРНК, поэтому особый интерес представляет ROC-кривая в районе малых FPR. В этой области диффузионная модель демонстрирует систематическое улучшение надежности предсказания (Рисунок 8, вставки).

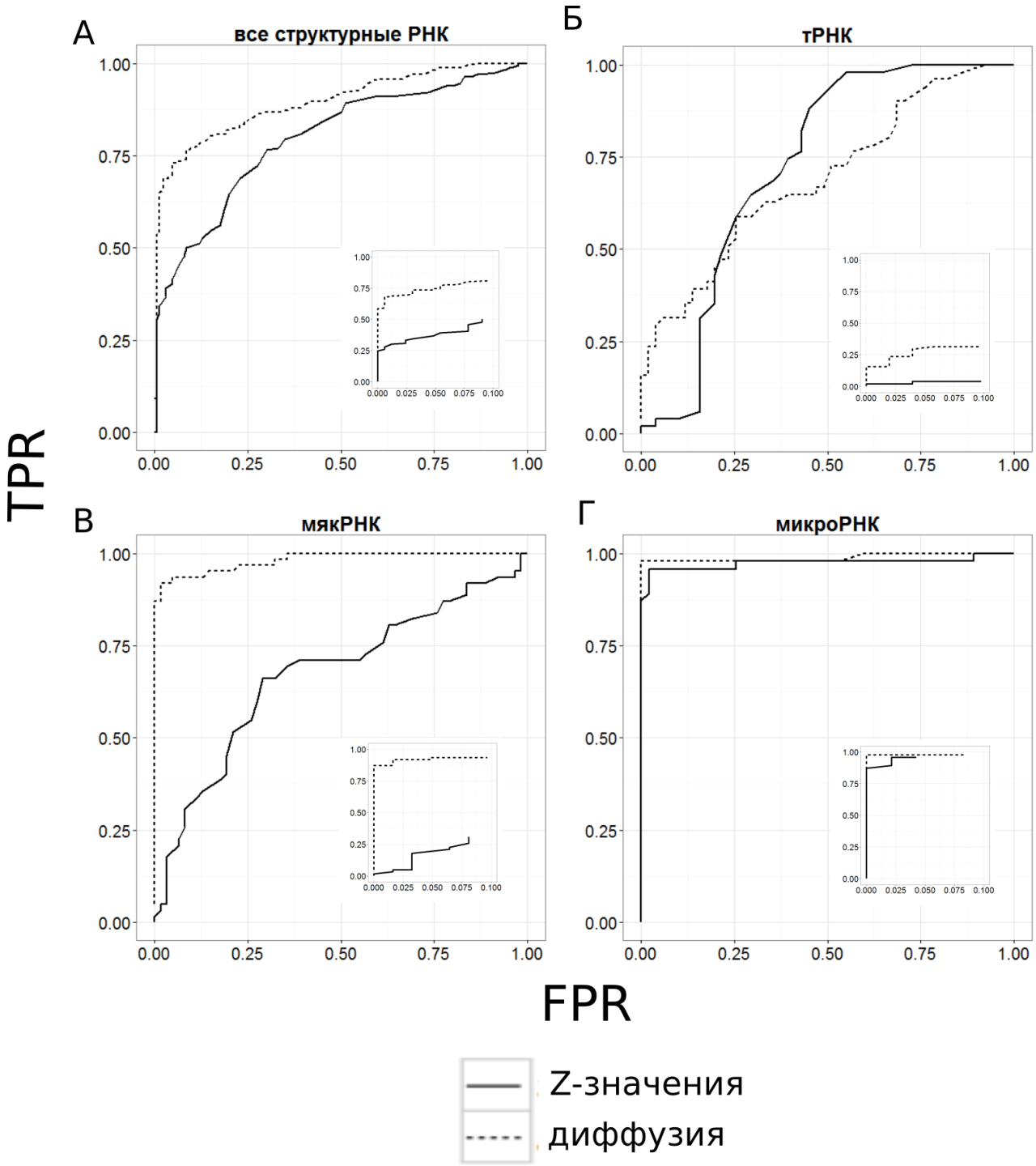


Рисунок 8. Качество предсказания некодирующих РНК для диффузионной модели (сплошная линия) и на основе Z-значений (пунктирная линия). Результаты показаны для всех структурных РНК (А) и отдельно для тРНК (Б), мякРНК (В) и микроРНК (Г).

Выводы

1. Формализовано понятие локально-оптимального структурированного сегмента РНК с использованием Z-значений. Предложен метод эффективного вычисления Z-значений с учетом статистических характеристик последовательностей.

2. Разработана программа RNASurface для поиска локально-оптимальных сегментов. Программа строит профили структурированности и тепловую карту структурированности. Теоретическое и практическое время работы и занимаемая память RNASurface не уступает самым эффективным программам данного класса.

3. Проведена апробация подхода и программы на полном геноме *Bacillus subtilis.* Апробация показала лучшее качество предсказания среди программ по предсказанию структурных элементов РНК на основе их энергетических свойств; показана устойчивость работы программы к выбору параметров. Анализ расположения предсказанных структурированных сегментов в геноме *Bacillus subtilis* выявил их сильную перепредставленность перед началом и после конца кодирующих областей.

4. Разработана и реализована диффузионная модель эволюции количественных характеристик последовательностей. Модель позволяет выявить давление отбора на исследуемую характеристику.

5. Применение диффузионной модели к анализу Z-значений структурированности РНК в *Drosophila melanogaster* показало значительное улучшение надежности предсказания некодирующих РНК.

Список публикаций по теме диссертации

Статьи в научных журналах

1. Солдатов РА, Миронов АА. Статистические методы сравнительно-геномного анализа, основанные на использовании диффузионных процессов // Биофизика. - 2013. - Т. 58. - С. 142–147.

2. Soldatov RA, Vinogradova SV and Mironov AA. RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments // Bioinformatics - 2014. - Vol. 30. - P. 457-463.

Тезисы конференций

1. Солдатов Р, Миронов А. Статистические методы анализа эволюции // 54-ая научная конференция МФТИ. - Москва. - 2011. - С. 92.

2. Soldatov R, Mironov A. Statistical methods of comparative genomic analysis based on diffusion approximation // Regulation and Evolution of Cellular Systems (RECESS). - Moscow. - 2012.

3. Солдатов Р, Миронов А. Статистические методы геномного анализа, основанные на использовании диффузионной модели // Информационные технологии и системы (ИТиС’12). Сборник тезисов. - Петрозаводск. - 2012. - С. 334-335

4. Soldatov R, Vinogradova S, Mironov A. RNASurface: fast and accurate identification of motifs with high structural potential // RECESS – Regulation and Evolution of Cellular Systems. - Venice. - 2013.

5. Soldatov R, Vinogradova S, Mironov A. RNASurface: fast and accurate identification of motifs with high structural potential // 6–th Moscow Conference on Computational Molecular Biology’13. Book of abstracts. - Moscow. - 2013.

6. Солдатов Р, Виноградова С., Миронов А. Поиск локально-оптимальных структурированных участков генома // Информационные технологии и системы (ИТиС’13). Сборник тезисов. - Калининград. - 2013 - С. 71-72.

7. Soldatov R, Vinogradova S, Mironov A. Detection of thermodynamically stable RNAs in long sequences with and without probing data // Computational Analysis of RNA Structure and Function. - Benasque. - 2015.