Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича

Российской академии наук

На правах рукописи

Солдатов Руслан Андреевич

Методы предсказания структурных элементов РНК

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата физико-математических наук

Научный руководитель:

кандидат физико-математических наук,

доктор биологических наук, профессор

Андрей Александрович Миронов

Москва, 2015

Оглавление

Введение

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 [1], трансляция [2], вирусная репликация [3] и модификация хроматина [4]. Современные каталоги функциональных РНК существенно не полны и, как считается, геномы содержат значительно больше структурных РНК чем сейчас известно [5]. 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 [5].

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.

Основные результаты и положения, выносимые на защиту

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.

Публикации. 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.

Глава 1. Обзор Литературы

1.1. Структурные РНК: основные классы и механизмы регуляции

Долгое время центральную роль в изучении клетки занимали белки, а главным механизмом регуляции их экспрессии считались транскрипционные факторы. Это было связано с парадигмой один ген - один белок, при которой матричная РНК (мРНК) воспринималась как промежуточная молекула в процессе синтеза белка [7]. Помимо мРНК, четыре главных известных типа РНК составляли транспортная РНК (тРНК), рибосомальная РНК (рРНК), малая ядрышковая РНК (мякРНК) и малая ядерная РНК (мяРНК). Эти классы РНК имеют жесткие структурные формы, необходимые для выполнения функции через взаимодействия с различными РНК и белковыми комплексами. Несмотря на различные роли этих структурных РНК, их функция сосредоточена на разных стадиях процесса синтеза белка. Это определило взгляд о роли РНК как "помощнике" при синтезе белков.

Бурное развитие области РНК биологии привело к открытию десятков новых классов РНК, осуществляющих структурную, каталитическую и регуляторную функцию. Так, вторичная структура рибозимов важна для выполнения их каталитической функции. МикроРНК осуществляют пост-транскрипционное подавление экспрессии мРНК через комплементарное связывание в комплексе RISC [8]. В настоящий момент база miRbase содержит несколько тысяч аннотированных микроРНК в геноме *homo sapiens* [9], которые регулируют более 60% генов [10]. С микроРНК тесно связан феномен РНК-интерференции, при котором эндо- и экзогенные короткие дуплексы РНК подавляют экспрессию белок-кодирующих генов с комплементарными сайтами, а также участвуют в антивирусной защите [11]. Малые piРНК пост-транскрипционно подавляют

мобильные элементы в зародышевой линии, а также являются ярким примером эпигенетической регуляции [12].

Разные классы малых РНК имеют общие пути биогенеза или принципы регуляции. В последние несколько лет была открыта разнообразная по механизмам функционирования популяция длинных некодирующих РНК (днРНК), составляющая более 10 тысяч локусов [13], [14]. Несмотря на схожий с белок-кодирующими РНК биогенез (кэпирование, полиаденилирование и сплайснг), днРНК не имеют кодирующий потенциал и выполняют широкий спектр функций, например: XIST инактивирует X хромосому [15], HOTAIR регулирует транскрипцию других генов посредством модификации состояния хроматина [16], а MALAT1 участвует в ко-транскрипционной регуляции сплайсинга [17].

Регуляторное разнообразие возрастает со "сложностью" организма, однако бактериальные геномы также содержат большое количество регуляторных РНК. Например, рибо-переключатели и Т-боксы, некодирующие структурные РНК размером 100-300 нуклеотидов, регулируют транскрипцию и трансляцию мРНК в бактериях, принимая альтернативные конформации вторичной структуры РНК [18]. Регуляторы располагаются преимущественно в 5'НТО генов и действуют по принципу обратной связи: подавляют экспрессию гена, который участвует в биосинтезе молекулы (аминокислота, витамин, фермент), связывающейся с РНК [19].

Таким образом, молекула РНК осуществляет огромное разнообразие регуляторных функций в клетке, а вторичная структура РНК является основным механизмом выполнения этих функций.

1.2. Вторичная структура РНК

1.2.1. Термодинамический подход

Молекула РНК существует в одноцепочечном состоянии, поэтому комплементарные участки сворачиваются на себя образуя вторичную структуру. Комплементарные пары оснований в основном образуются между каноническими парами G-C, A-U и неоднозначной парой G-U [20]. Детальный анализ кристаллических структур известных РНК показал, что из комплементарных пар 68% являются каноническими парами и 7% являются G-U парой [21]. В дальнейшем анализе другие неканонические взаимодействия не будут учитываются. Спаренные участки образуют стебли, а неспаренные участки - петли, Рисунок 1.2.1 содержит более подробное описание элементов вторичной структуры. Ряд аргументов позволяет считать вторичную структуру РНК хорошим приближением третичной. Во-первых, основной вклад в свободную энергию трехмерной структуры вносят водородные и стекинг взаимодействия нуклеотидов, представленные во вторичной структуре [22]. Во-вторых, считается что сворачивание РНК происходит иерархически: сначала образуются локальные дуплексы, а затем формируются дальние взаимодействия и третичные контакты [23].

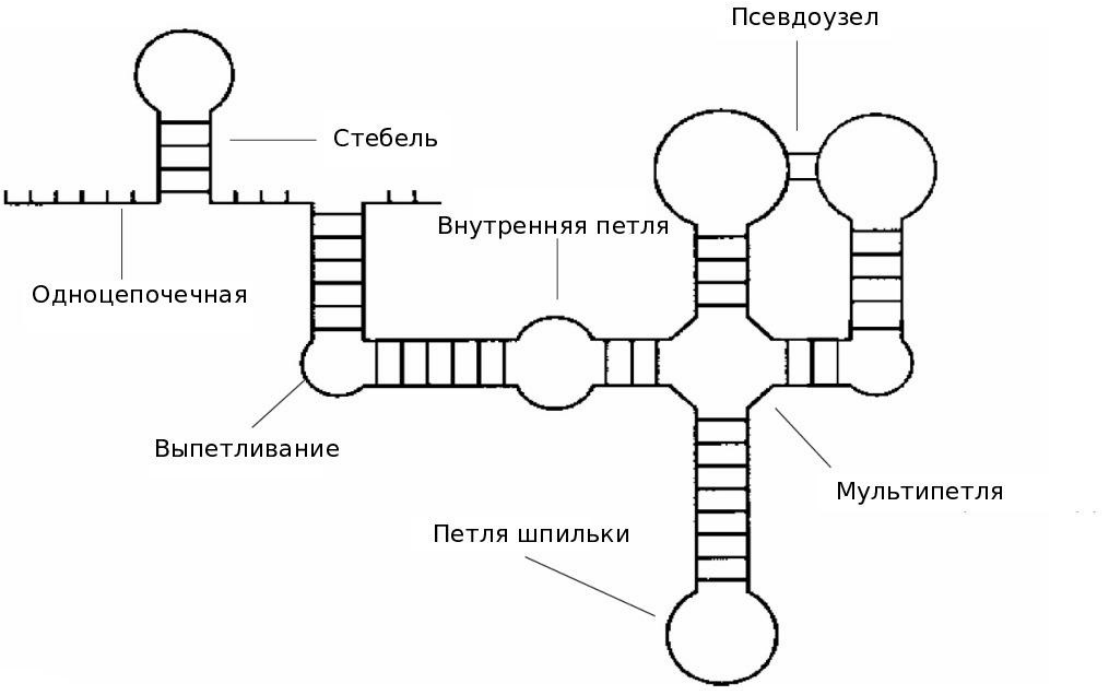


Рисунок 1.2.1. Элементы вторичной структуры РНК. Участки спирали делятся на стебли или псевдоузлы (формальное определение псевдоузла будет дано ниже). Неспаренные участки делятся на петли (петли шпилек), внутренние петли, выпетливания, мультипетли и одноцепочечные сегменты. (из [24])

Формально, вторичная структура последовательности РНК $s_1 s_2 \dots s_n$ представляет из себя набор пар нуклеотидов

$$P = \{(i,j) | i < j \text{ и } s_i - s_j \text{ образуют пару } G-C, A-U \text{ или } G-U\},$$

при этом каждый нуклеотид встречается не более чем в одной паре. Последнее условие исключает третичные взаимодействия.

Существует несколько способов удобной визуализации вторичной структуры (Рисунок 1.2.2). Кроме стандартного (Рисунок 1.2.2Б) и точечно-скобочного представления (Рисунок 1.2.2А), существует графовое представление, при котором последовательность изображена в виде

окружности и дуги соединяют спаренные основания. Такое представление особенно полезно при анализе и разработке алгоритмов предсказания структуры РНК.

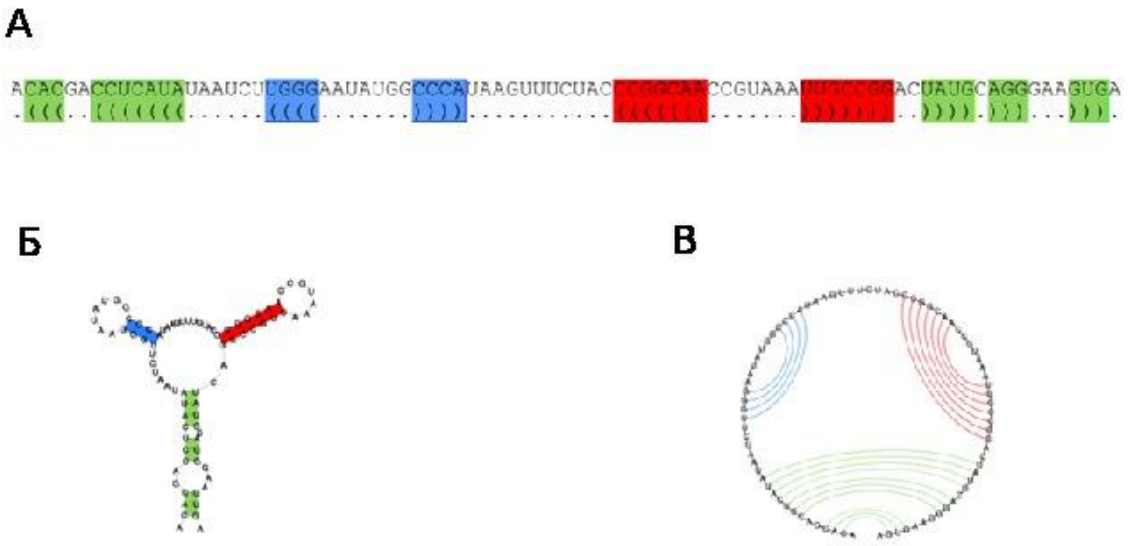


Рисунок 1.2.2. Визуализация вторичной структуры пуринового рибо-переключателя (Rfam RF00167). Цвета соответсвуют трем стеблям. А) Представлена последовательность РНК, а под ней точечно-скобочное обозначение вторичной структуры: точка соответствует неспаренному нуклеотиду, а открытая/закрытая скобка - левому/правому спаренному нуклеотиду. Б) Стандартный представление вторичной структуры РНК. В) Графовое представление: спаренные нуклеотиды соеденины дугой. (адаптировано из [25])

Алгоритмы предсказания структуры РНК почти без исключений не учитывают псевдоузлы. Две пары (i, j) и (k, l) образуют псевдоузел (будем считать $i < k$), если

$$i < k < j < l.$$

При графовом изображении структуры пары, образующие псевдоузел, обязательно пересекаются (Рисунок 1.2.3А). С алгоритмической точки зрения, наличие псевдоузлов делает задачу предсказания вторичной

структуры методом динамического программирования в общем случае NP-полной [26], то есть неразрешимой за обозримое время. С физической точки зрения, нет точной модели для описания термодинамики псевдоузлов [27]. Вторичная структура РНК без пседвоузлов называется вложенной, пример расположения пар такой структуры представлен на Рисунок 1.2.3Б-В. В дальнейшем будут рассматриваться только вложенные вторичные структуры, хотя эта модель имеет ограничения и известны примеры структурных РНК, в которых псевдоузлы консервативны и выполняют каталитическую или структурную роль [28].

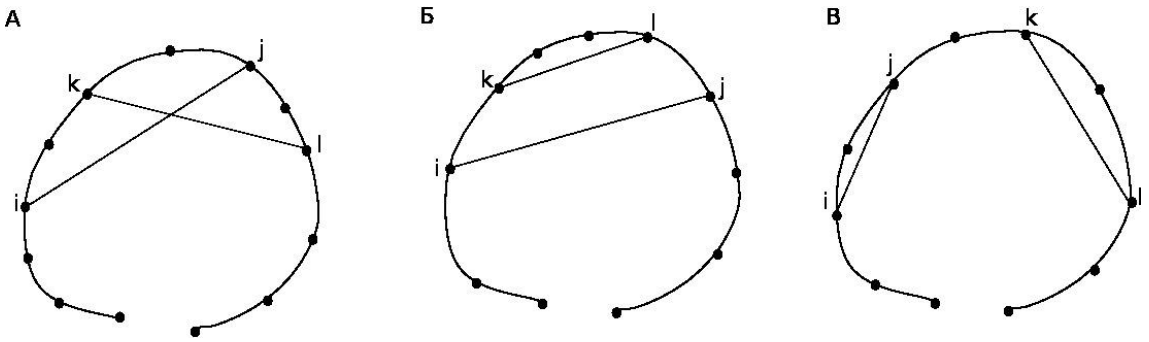


Рисунок 1.2.3. Взаимное расположение пар (i,j) и (k,l) во вторичной структуре РНК. А) Псевдоузел. Б) Вложенная пара. В) Непересекающаяся пара.

При предсказании структуры РНК необходимо установить критерий оптимальной структуры. Логично предположить, что молекула РНК сворачивается в термодинамически наиболее стабильную структуру. Стекинг и водородные взаимодействие между нуклеотидами стабилизируют структуру молекулы. Для простоты предположим, что оптимальная структура РНК это структура с наибольшим количеством спаренных нуклеотидов. Количество возможных вторичных структур растет экспоненциально с длиной последовательности [29], поэтому для поиска нужной структуры требуются эффективные алгоритмы. Нуссинов

разработала элегантный алгоритм [30], [31], основанный на методе динамического программирования, который представлен ниже.

Дана последовательность РНК $s_1 \dots s_n$. Через D_{ij} обозначим количество спаренных оснований оптимальной структуры подпоследовательности $s_i \dots s_j$. В оптимальной структуре $s_i \dots s_j$, нуклеотид s_i либо неспарен, либо спарен с некоторым другим нуклеотидом s_k. Тогда D_{ij} можно вычислить для этих двух случаев используя информацию о подструктурах (Рисунок 1.2.4):

$$D_{ij} = max \begin{cases} D_{i+1,j}, & \text{если } s_i - \text{неспарен} \\ D_{i+1,k-1} + 1 + D_{k+1,j}, & \text{если } s_i - \text{спарен с } s_k, k = i+1, .., j \end{cases}$$

Рисунок 1.2.4. Горизонтальной линией показана подпоследовательность, а дугой - спаренные нуклеотидов. Если в оптимальной структуре $\boldsymbol{i}$-ый нуклеотид неспарен, то вычисляем количество пар нуклеотидов в подструктуре сегмента $[\boldsymbol{i}+\boldsymbol{1}, \boldsymbol{j}]$. Если $\boldsymbol{i}$-ый нуклеотид спарен с $\boldsymbol{k}$-ым, то вычисляем независимо количество пар в подструктурах сегментов $[\boldsymbol{i}+\boldsymbol{1}, \boldsymbol{k}-\boldsymbol{1}]$ и $[\boldsymbol{k}+\boldsymbol{1}, \boldsymbol{j}]$ (из [25]).

Таким образом, определение оптимальной структуры сводится к определению оптимальных подструктур. В качестве начальных условий считаем, что $D_{ii} = 0, D_{i,i+1} = 0$; то есть нуклеотид не может быть спарен с собой и соседним. Эта задача решается эффективно методом динамического программирования, а именно заполнением матрицы $n \times n$ с элементами D_{ij}. Величина D_{1n} соответствует количеству спаренных

нуклеотидов в последовательности, а структура восстанавливается обратным ходом по матрице. Алгоритм требует $O(n^3)$ операций и $O(n^2)$ памяти.

Алгоритм Нуссинов отражает основную идею применения динамического программирования при предсказании структуры РНК, тем не менее он работает крайне плохо. Причина в том, что алгоритм никак не учитывает энергетическую модель и кинетику вторичной структуры РНК.

Стабильность вторичной структуры РНК определяется энергией Гиббса (свободной энергией). Изменение энергии Гиббса, ΔG, характеризует изменение полной энергии системы при протекании химического процесса в условиях постоянного давления и температуры; ΔG определяется следующим образом:

$$\Delta G = \Delta H - T\Delta S,$$

где энтальпия ΔH характеризует количество выпущенного тепла при изменении состояния; $T\Delta S$ характеризует изменение энтропии системы.

Структура РНК состоит из набора простых элементов (Рисунок 1.2.5А), свободная энергия каждого из которых экспериментально определена группой Тёрнера [22], [32]. Стекинг взаимодействия вносят основной вклад в стабилизацию структуры [33], [34], а петли дестабилизируют структуру [22]. Энергия структуры вычисляется как сумма энергий набора её элементов (пример, Рисунок 1.2.5Б):

$$E = \sum_{\text{элементы } l \text{ структуры } S} E(l).$$

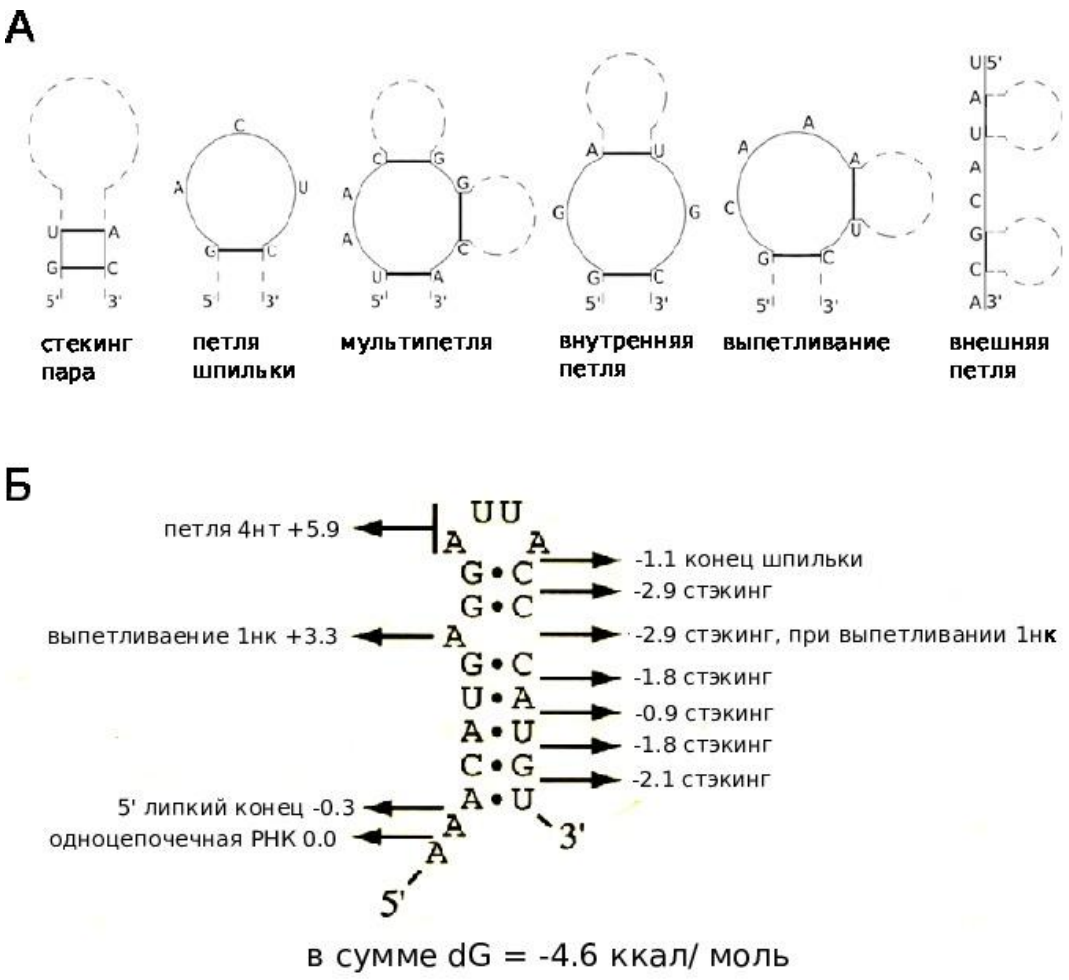


Рисунок 1.2.5. А) Элементы вторичной структуры. Стекинг взаимодействия стабилизируют структуру, петли - дестабилизируют. Б) Пример вклада разных элементов вторичной структуры в изменение энергиии Гиббса. (из [24])

Согласно законам статистической термодинамики, молекула РНК пребывает в данной конформации S с энергией Гиббса ΔG с частотой, описываемой распределением Больцмана:

$$P(S) = \frac{E^{-\Delta G(S)/kT}}{\Omega},$$

где $\Omega = \sum_S E^{-\Delta G(S)/kT}$ и называется статистической суммой.

Как следует из распределения Больцмана, наиболее вероятна структура РНК с наименьшей свободной энергией. В 1981 году Зукер предложил алгоритм предсказания структуры с наименьшей свободной энергией [35]. Подробная энергетической модель вторичной структуры РНК делает реализацию алгоритм Зукера достаточно громоздкой. Схематично, через $W(i,j)$ обозначим оптимальную энергию на участке $[i,j]$, а через $V(i,j)$ оптимальную энергию при условии, что нуклеотиды i и j образуют пару. Тогда, согласно алгоритму Зукера, $W(i,j)$ вычисляется как:

$$W(i,j) = \min\left\{V(i,j), \min_{i \le k < j}\{W(i,k) + W(k+1,j)\}\right\} \quad (1)$$

Для примера, представим действия алгоритма на участке последовательности $[i,j]$, в случае если i-ый и j-ый нуклеотиды спарены. В этом случае пара (i,j) либо является терминальной парой шпильки (Рисунок 1.2.6А), либо образует стекинг взаимодействие (Рисунок 1.2.6Б), либо является терминальной парой внутренней петли/выпетливания (Рисунок 1.2.6В), либо терминальной парой мультипетли (Рисунок 1.2.6Г). В каждом из этих случаев, вычисление структуры сводится к подструктурам (подробности в легенде, Рисунок 1.2.6).

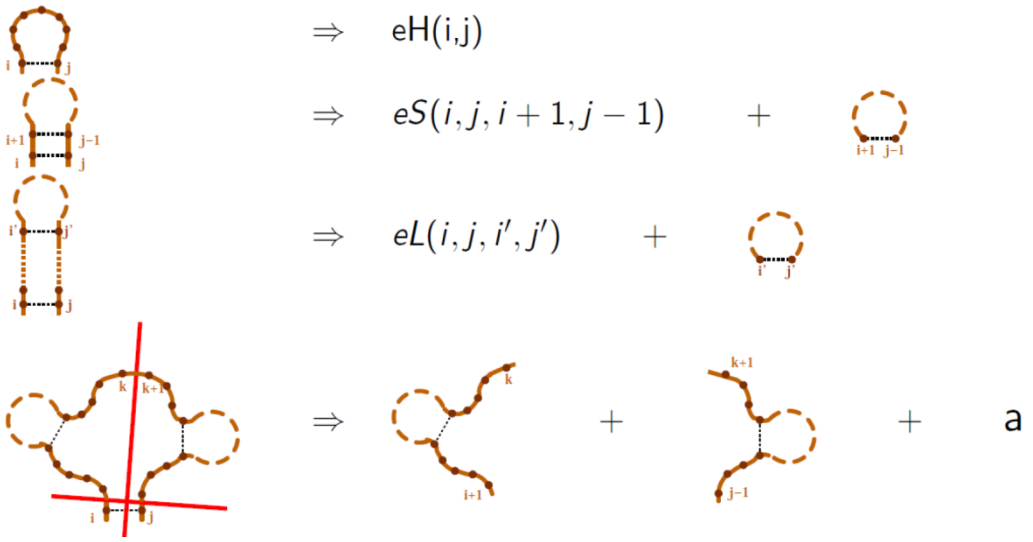


Рисунок 1.2.6. Схема алгоритма для подпоследовательности $[i,j]$, если пара нуклеотидов (i,j) образует водородную связь. Оптимальная структура подпоследовательности вычисляется как лучшая из четырех возможных вариантов А)-Г). А) Пара (i,j) образует терминальную связь шпильки: в этом случае энергия вычисляется, используя энергетическую модель, как энергия петли шпильки $eH(i,j)$. Б) Пара (i,j) образует стекинг взаимодействие: энергия складывается из стабилизирующего вклада стекинга $eS(i,j,i+1,j-1)$ и оптимальной структуры на участке $[i+1,j-1]$. В) Пара (i,j) образует терминальную связь во внутренней петле/выпетливании: энергия складывается из известной энергии петли $eL(i,j,i',j')$ и оптимальной энергии участка $[i',j^i]$. Г) Пара (i,j) образует терминальную связь мультипетли: в этом случае перебираются возможные разбиения мультипетли, и добавляется вклад a от закрывающей мультипетлю пары (i,j). (из [24])

Если i-ый и j-ый нуклеотиды неспарены, то проводится аналогичный комбинаторный перебор вариантов и сведение каждого к подструктурам. Таким образом, по аналогии с алгоритмом Нуссинов, алгоритм Зукера сводит задачу к подзадачам и эффективно решается динамическим программированием. Существуют несколько реализаций алгоритма Зукера, самыми известными являются RNAfold [36], Mfold [37] и RNAStructure [38].

Узким местом алгоритма является разбор мультипетель и внутренних петель [39], [40]; в случае использования простых эвристик, весь алгоритм требует $O(n^3)$ операций и $O(n^2)$ памяти. Качество предсказания, а именно чувствительность и специфичность, алгоритма Зукера на известных структурных РНК варьируется в пределах 60-80% [36]. Примечательно, ни использование различных наборов термодинамических параметров, ни модификации алгоритма не позволяют заметно улучшить качество предсказания [36]. Для этого существует ряд фундаментальных причин.

Во-первых, пространство вторичных структур молекулы РНК имеет, как правило, плотный энергетический спектр: существует много конформаций с энергией Гиббса близкой к минимальной [41]. Таким образом, в зависимости от начальных условий, молекула РНК способна сворачиваться в разные структуры обладающие схожей стабильностью. Например, в клетке РНК сворачивается в структуру ко-транскрипционно [42], что радикально изменяет её конформационное пространство [43]. Во-вторых, молекула РНК метастабильна: энергетическая поверхность структур РНК имеет низкие энергетический барьеры локальных оптимумов, что позволяет молекуле часто изменять конформацию [44]. В-третьих, погрешность эксперимента при определении термодинамических параметров значительно влияет на предсказание структуры. Так, при вариации параметров Тернера в пределах уровня погрешностей эксперимента, до 30% структур предсказываются неправильно [45].

Более аккуратные выводы о молекуле РНК можно получить, если учитывать весь ансамбль вторичных структур РНК [46]. Если рассматривать наиболее достоверные взаимодействия, входящие в наиболее вероятные конформации, то качество их предсказания повышается с 66% до 91% [47], [48].

Альтернативный подход к предсказанию вторичной структуры основан на сравнительно-геномном анализе. Структурные РНК имеют определенную вторичную структуру, необходимую для выполнения функции. Мутации, разрушающие спаривание нуклеотидов, нарушают вторичную структуру РНК и выметаются из популяции в силу естественного отбора [49], [50]. Таким образом, в ходе эволюции замены в структурных РНК сохраняют взаимодействие нуклеотидов. Сохранение связи может происходить через допустимую замену одного нуклеотида (например A-U ↔ G-U) или через компенсаторную замену двух нуклеотидов (например, A-U ↔ G-C или C-G ↔ G-C). В таком случае, ортологичные позиции спаренных оснований имеют высокую корреляцию паттернов замен [51]. Этот эволюционный сигнал можно использовать для предсказания общей вторичной структурой набора ортологичных последовательностей по их множественному выравниванию [52]: выбирается структура, в ходе эволюции которой происходили преимущественно допустимые и компенсаторные замены. Также существуют подходы к предсказанию, которые комбинируют эволюционный и термодинамический сигналы [53]. Эволюционный подход является мощным инструментом, однако применим только к предсказанию структур РНК, имеющих функцию и давление отбора на вторичную структуру. Кроме того, количество компенсаторных замен нередко не хватает для статистически достоверных выводов.

1.3. Предсказание структурных РНК

Большая часть генома млекопитающих транскрибируется [54], при этом значительная доля участков генома под эволюционным отбором находится за пределами белок-кодирующих областей [55]. С учетом этих соображений, обнаружение за последние двадцать лет разнообразных

классов регуляторных РНК [56] привело к осознанию, что геномы высших эукариот могут содержать большое количество неизвестных структурных РНК. Экспериментальные методы масштабного поиска структурных РНК не доступны, поэтому исследования сконцентрированы на биоинформатических предсказаниях.

Для предсказания структурных РНК необходимо выяснить свойства, которые выделяют их на фоне остального транскриптома и могут быть использованы как сигнал. Как и в случае с задачей предсказания вторичной структуры РНК, термодинамическая стабильность и эволюционный отбор на структуру являются двумя основными сигналами.

Геномы высших эукариот очень большие, например геном человека составляет примерно 3 миллиарда нуклеотидов [57]. Поэтому возникают серьезные препятствия по скорости работы алгоритмов на масштабе генома, и статистическому анализу их результатов.

1.3.1. Термодинамические свойства структурных РНК

Термодинамическая стабильность молекулы РНК может быть выражена в виде минимальной свободной энергии E, которая эффективно вычисляется с помощью алгоритма Зукера. Как упоминалось в предыдущей главе, настоящая структура РНК часто имеет не минимальную энергию, однако ввиду плотности энергетического спектра, имеет близкую к ней свободную энергию. Таким образом, минимальная свободная энергия является хорошим приближением к энергии структуры РНК. Чем ниже энергия вторичной структуры, тем более она стабильна.

В пионерских работах [58], [59] для оценки значимости наблюдаемой энергии E использовались энергии случайных последовательностей такой же длины и нуклеотидного состава. Для набора случайных

последовательностей можно построить распределение свободных энергий, и оценить насколько сильно энергии E смещена относительно этого фонового распределения (Рисунок 1.3.1). Формально, оценкой этого смещения служит величина Z-значение, определяемая как:

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

где μ и σ соответствуют среднему и стандартному отклонению фонового распределения энергий. Свободная энергия отрицательна, и чем ниже тем более стабильна молекула РНК; поэтому отрицательное Z-значение соответствует структурированным РНК.

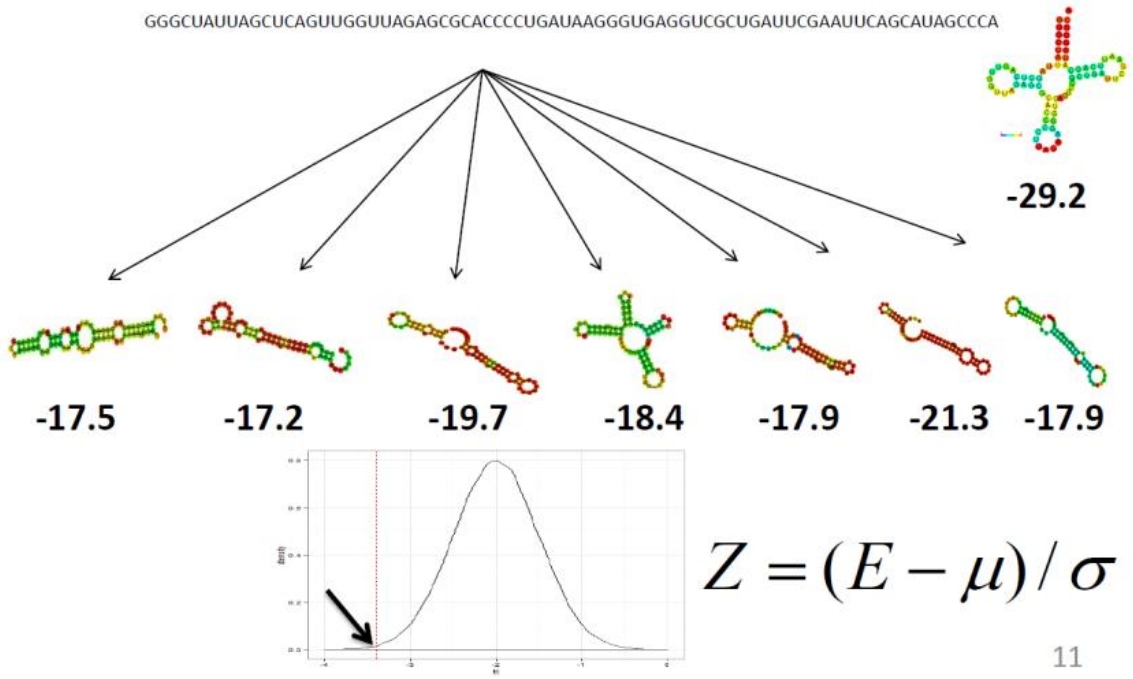


Рисунок 1.3.1. Метод оценки значимости свободной энергии. Если перемешать последовательность РНК (на верхней панели: последовательность и структура сбоку), то перемешанные последовательности также сворачиваются в структуры РНК с некоторыми энергиями (стрелки ведут к структурам перемешанных последовательностей). Цифры под структурой - свободная энергия. На основе энергии структур перемешанных

последовательностей строится распределение фоновых энергий, и оценивается "необычность" энергии исходной последовательности.

Три водородные связи G-C пары "прочнее" двух связей A-U пары, поэтому GC богатые последовательности РНК будут в среднем иметь более низкую свободную энергию, даже при похожей форме структур. В этом случае низкая свободная энергия может означать всего лишь высокий GC состав. Фоновая модель генерации последовательностей с таким же нуклеотидным составом позволяет сохранить ожидаемую энергию водородных связей. Однако ключевой энергетический вклад в стабильность структуры вносят стекинг взаимодействия между соседними парами нуклеотидов [33]. Для сохранения ожидаемой энергии стекинг взаимодействий необходимо генерировать последовательности с тем же динуклеотидным составом [60]. Так, при оценки стабильности вторичных структур матричных РНК было показано, что их энергия ниже чем у случайных последовательностей с тем же нуклеотидным составом [61], однако не отличаются от энергий случайных последовательностей с тем же динуклеотидным составом [60].

Классы структурных РНК имеют достаточно низкое Z-значение относительно последовательностей с тем же динуклеотидным составом [62], [63] (Рисунок 1.3.2А). Примечательно, транспортная РНК является мало структурированной относительно других классов РНК (Рисунок 1.3.2Б), несмотря на строго определенную структуру "клеверного листа". МикроРНК, наоборот, очень структурированы и имеют как правило Z-значение меньше -4 [64].

А

тип нкРНК	к-во посл-ей	среднее Z-значение
tRNA	579	−1.84
5S rRNA	606	−1.62
Hammerhead ribozyme III	251	−3.08
Group II catalytic intron	116	−3.88
SRP RNA	73	−3.37
U5 spliceosomal RNA	199	−2.73

Б

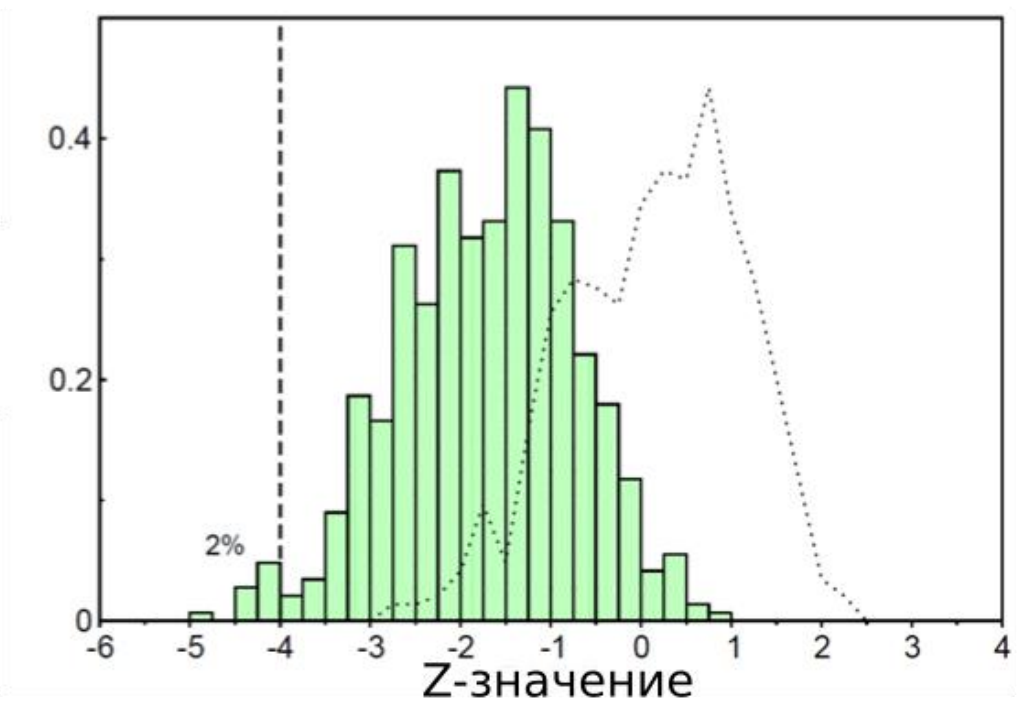


Рисунок 1.3.2. Структурные РНК обладают низким Z-значением. А) Среднее Z- значение для некоторых классов нкРНК. Первый столбец: тип РНК; второй столбец: количество последовательностей данного класса; третий столбец: среднее Z-значение по последовательностям данного класса РНК. Б) Гистограмма Z-значений для тРНК (зеленый), и для набора последовательностей с тем же динуклеотидным составом (пунктир). Только 2% случайных последовательностей имеют энергию ниже, чем у тРНК. (из [63])

Свободная энергия содержит лишь часть информации о вторичной структуре РНК, также имеет значение энергетическая поверхность структур: разнообразие локальных оптимумов и энергетические барьеры между ними. Например, чтобы молекула РНК продолжительное время находилась в одной структур, необходимо чтобы её локальный минимум имел высокий энергетический барьер. Существует несколько мер, которые отражают разнообразие структур последовательности РНК и доступны для эффективного вычисления. Однако системное исследование этих и ряд других мер показало, что на практике Z-значение лучше всего и достаточно полно отражает термодинамические свойства известных структурных РНК [65].

1.3.2. Поиск структурированных участков в последовательностях

Если требуется выявить структурированные участки длины L в большой последовательности, то логичных подходом будет сканировать последовательность окном L с некоторым шагом между соседними окнами (Рисунок 1.3.3). В этом случае пики значении Z-значения будут соответствовать наиболее структурированным сегментам. На этом небольшом наборе потенциальных структурных РНК уже можно проводить более тонкие анализы для подтверждения структурной роли: сравнение с доступными ортологами [66], анализ похожих мотивов в геноме [67] или базах данных [68], сравнение паттерна структуры с известными регуляторными структурами [69].

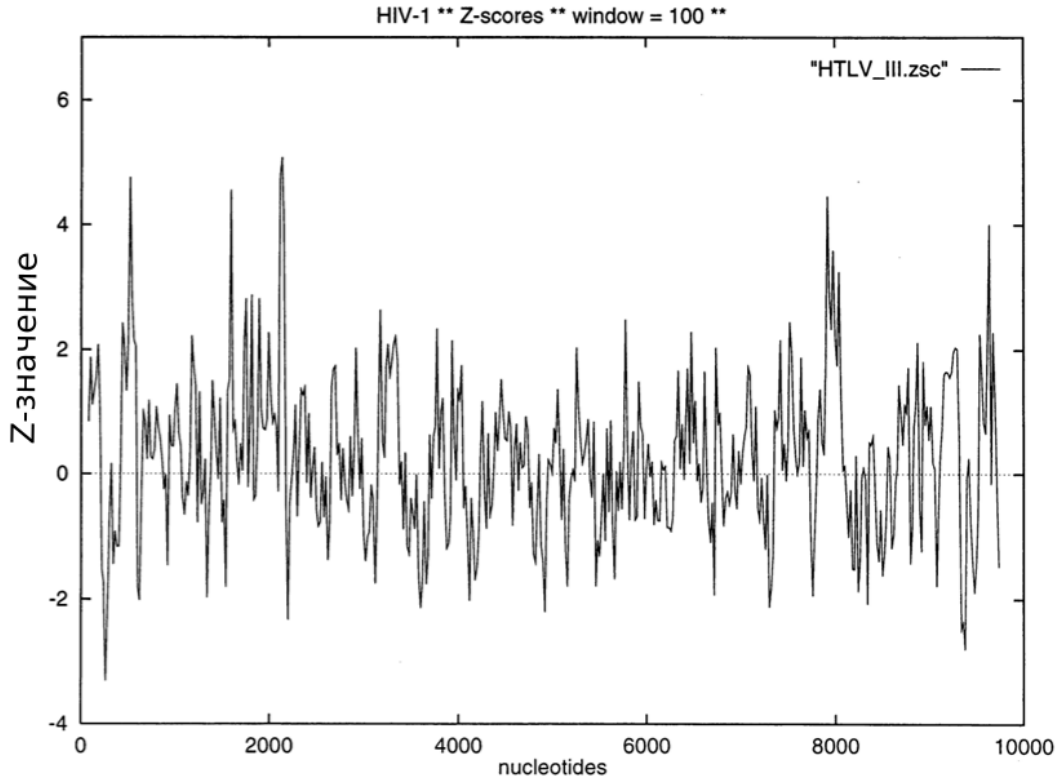


Рисунок 1.3.3. Пример сканирования генома ретровируса HIV-1 с фикисрованным окном. Пики на 500 нуклеотидах и 8000 нуклеотидах соответствуют известным цис-регуляторным вторичным структурам. (из [70])

Структурные РНК обладают большим разнообразием форм и размеров вторичных структур: от небольших, 20-30 нуклеотидов, шпилек ро-независимых терминаторов [71] до разветвленных Т-боксов и метастабильных рибо-переключателей [18] размером в сотни нуклеотидов. При сканировании последовательности с фиксированным окном и шагом есть риск пропустить структурную РНК значительно отличного размера, или "перешагнуть" регуляторную структуру малого размера. Более того, параметры окна и шага могут оказать решающие значение на качество предсказание. Так, в работе по поиску структурных РНК в геноме дрожжей с помощью Z-значения авторы выяснили что качество предсказания H/ACA

мякРНК варьируется от 6.9% до 62% в зависимости от выбора окна и шага [72] (Рисунок 1.3.4). Причем доля предсказанных H/ACA мякРНК всеми окнами составляла больше 72%, что говорит о принципиальном ограничении сканирования окном.

	Размер окна												
Размер шага	**80**	**90**	**100**	**110**	**120**	**130**	**140**	**150**	**160**	**170**	**180**	**190**	**200**
5	62.1	44.8	51.7	51.7	51.7	51.7	41.4	48.3	48.3	41.4	44.8	44.8	41.4
25	34.5	20.7	31.0	27.6	37.9	34.5	34.5	20.7	31.0	17.2	31.0	27.6	27.6
50	13.8	6.9	17.2	17.2	31.0	17.2	17.2	17.2	17.2	13.8	24.1	20.7	20.7

Рисунок 1.3.4. Процент предсказанных H/ACA мякРНК в зависимости от окна и шага при сканирования. Общий процент предсказанных H/ACA мякРНК всеми окнами с шагом 5 составляет 72.4%. (из [72])

Если длина последовательности S, а шага s, то количество подсчитанных Z-значений для окон будет приблизительно S/s. Даже для генома дрожжей и шага в 5 нуклеотидов необходимо вычислить Z-значение более 1 млн. раз. Таким образом, получение Z-значений с помощью Монте-Карло симуляций последовательностей с вычислением их энергий становятся узким местом подхода, ограничивая его применения к очень маленьким последовательностям. Для снятие этого ограничения Вашитль в программе RNAz [73] предложил следующую процедуру. Фоновая модель генерации энергий зависит от длины и динуклеотидного состава последовательности, поэтому параметры фонового распределения энергий зависят от 17 параметров:

$$\mu = \mu(l, d_1, \dots, d_{16}),$$

$$\sigma = \sigma(l, d_1, \dots, d_{16}),$$

где $(d_1, \dots, d_{16})$ - набор динуклеотидов, а l - длина (хотя независимыми можно считать только 13 параметров, так как из 16 динуклеотидных частот только 12 независимы).

Вашитль предварительно вычислил параметры для 1'155'737 векторов $(l, d_1, \dots, d_{16})$ и на их основе построил регрессию опорных векторов отдельно для μ и σ, и показал хорошее качество предсказания регрессии на тестовой выборке [74]. Параметры этой регрессии заданы в RNAz и позволяют эффективно оценивать μ и σ. В таком случае для оценки Z-значений каждого окна необходимо вычислить всего три величины: энергию последовательности E с помощью алгоритма Зукера, параметры μ и σ фонового распределения энергий с помощью встроенной регрессии опорных векторов.

Если при анализе маленьких геномов, например дрожжей или бактерий, вычисление энергии для всех окон осуществляется за обозримое время, то для больших геномов это нереализуемо практически. Классический алгоритм Зукера осуществляет $O(L^3)$ операций для одного окна длины L, соответственно $O(L^3 \cdot S/s)$ для всех окон вдоль последовательности длины S с шагом s. Однако, как показали исследования [25] и видно из Рисунок 1.3.4, при увеличении шага окна качество предсказания структурных РНК падает катастрофически. Поэтому необходимо брать шаг s близким к единице, что делает скорость алгоритма эффективно $O(L^3 \cdot S)$.

1.3.3. Эффективные техники ускорения алгоритма Зукера

Будем считать, что шаг окна вдоль последовательности равен 1. Через W_i обозначим окно, которое начинается с i-ого нуклеотида. При

оценки энергии в окне, алгоритм Зукера вычисляет энергии всех подпоследовательностей этого окна, сохраняя их в матрице Зукера $L \times L$ (Рисунок 1.3.5Б). Два соседних окна пересекаются по участку длины $L - 1$. Поэтому, если уже вычислена матрица Зукера для окна W_i, то автоматически известна почти вся матрица для окна W_{i+1} (Рисунок 1.3.5А-Б). А именно, необходимо вычислить только энергии подпоследовательностей оканчивающиеся $(i + L)$-ым нуклеотидом $E_{i+1,i+L}, \dots, E_{i+L-1,i+L}$.

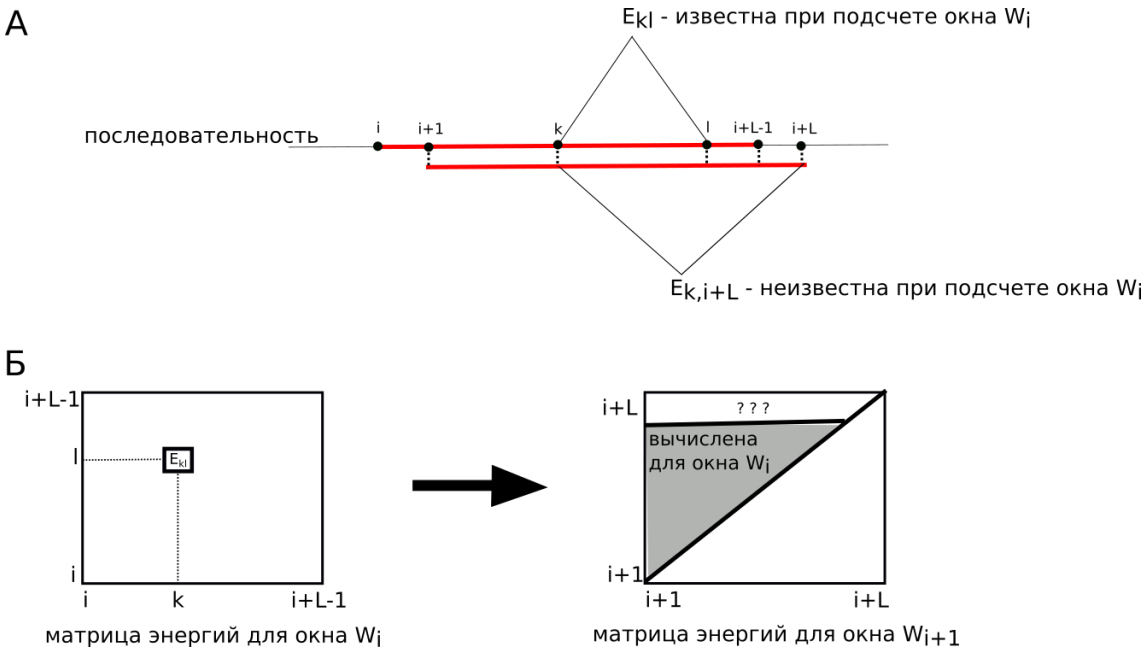


Рисунок 1.3.5. Быстрый пересчет энергий при сканировании окном. А) При сдвиге окна энергии почти всех подпоследовательностей известны. Красным обозначены окна W_i и W_{i+1}. Б) Для W_{i+1} достаточно вычислить одну строку энергий в матрице энергий Зукера.

В классическом алгоритме Зукера вычисление одной ячейки матрицы, что соответствует энергии некоторой подпоследовательности, занимает $O(L)$. Поэтому вычисление энергии нового окна, а для этого необходимо вычислить L ячеек (Рисунок 1.3.5Б), требует всего $O(L^2)$

операций. Таким образом, использование того факта, что соседние окна значительно пересекаются, позволяет ускорить алгоритм в L раз. Вариации этой техники сканирования последовательности окном реализованы в программах RNALL [75] и RNALfold [76]. А модификация RNALfold, программа RNALfoldz, сканирует последовательность пересчитывая энергии и вычисляя Z-значение с помощью встроенной регрессии опорных векторов [77]. До нашей программы RNASurface, RNALfoldz являлась единственным доступным инструментом для эффективного поиска структурированных участков в длинных последовательностях на основе термодинамического подхода.

Другая техника, метод разреженных матриц, позволяет еще *в среднем* сократить количество операций в L раз. Оказывается, для подсчета каждой ячейки матрицы энергий в среднем достаточно совершить несколько операций, а не $O(L)$. Структуру со спаренными крайними нуклеотидами будем называть закрытыми. Можно показать, что либо оптимальная структура является закрытая, либо представима в виде двух подструктур, первая из которых - закрыта и оптимальна [78] (Рисунок 1.3.6). Тогда при переборе в алгоритме Зукера (1) достаточно проходить только те промежуточные значения k, для которых левая подструктура является закрытой и оптимальной.

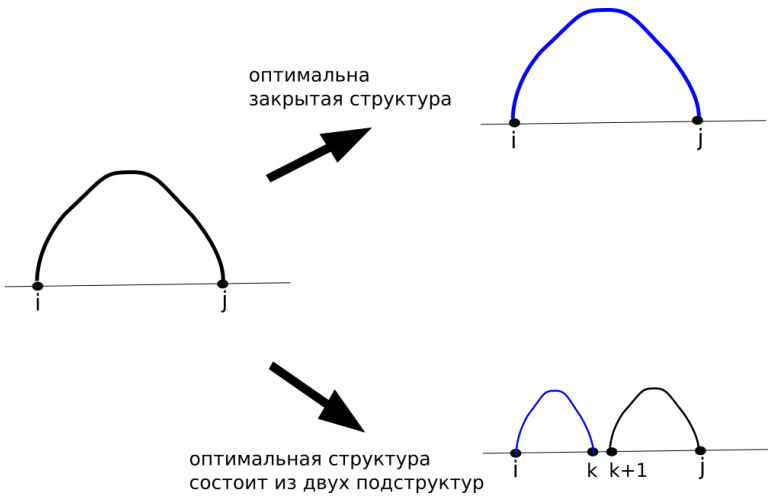


Рисунок 1.3.6. Идея метода разреженных матриц в случае предсказания структуры РНК. Оптимальная структура на сегменте $[i,j]$ либо является закрытой (верхний вариант), либо складывается из двух непересекающихся подструктур сегментов $[i,k]$ и $[k+1,j]$, первая из которых является оптимальной и закрытой на своем сегменте $[i,k]$ (нижний вариант). Закрытая структура обозначена синим цветом.

Количество кандидатов $V(i,k)$ с закрытыми оптимальными структурами определяет скорость алгоритма. Вероятность, что в оптимальной структуре крайние нуклеотиды спарены резко падает с длиной последовательности L, в первом приближении по степенному закону C/L^{α} [79]. Как показывают интенсивные симуляции оптимальных структур алгоритмом Зукера, $\alpha \approx 1.47$ [78], поэтому для сколь угодно длинных сегментов $[i,j]$ лишь конечное числа кандидатов будут иметь закрытые оптимальные подструктуры.

Если сочетать две вышеописанные техники, то ожидаемое время сканирования генома окном можно ускорить с $O(S \cdot L^3)$ до $O(S \cdot L)$. Эта комбинация реализована в программе RNASlider [80], которая легла в основу нашего алгоритма.

1.3.4. Другие методы поиск структурных РНК и ограничения подходов

Полногеномный поиск структурированных РНК на основе термодинамического подхода ограничен малыми геномами, такими как дрожжи или бактерии, или отдельными транскриптами. Чтобы избежать большого количества ложно-положительных предсказаний в геномах млекопитающих, необходимо устанавливать очень строгий порог на Z-значение < -6, который проходят малая доля реальных структурных РНК [77].

В этом случае наличие близких геномов усиливает сигнал и улучшает предсказание [63]. Главная идея в том, что если группа ортологичных локусов имеет общую функциональную вторичную структуру РНК, то количество компенсаторных и допустимых замен между ними будет значительно превышать ожидаемое количество для некоторой структуры [81]. Сравнительно-геномный подход статистически значительно более мощный чем термодинамический подход на одной последовательности [63], однако его область применимости очень ограничена.

Выравнивание накапливает ошибки при расхождении видов, так выравнивание геномов человека и мышки содержит $\sim 15\%$ ошибок [82]. С одной стороны, выравнивание должно содержать достаточно замен для статистически достоверных выводов, с другой не должно содержать много ошибок. Эмпирически исследования показали, что качество работы сравнительно-геномных подходов резко падает при среднем расхождении ортологов более 40% [83]. Как следствие, полногеномный анализ ограничен хорошо выравниваемой и консервативной частью генома, что составляет 3-10% в млекопитающих [81], [73]. При этом, как показывают последние работы, существует большая доля структурных РНК с высокой

скоростью эволюции и плохим качеством геномного выравнивания [84].

1.4. Эволюция количественных характеристик

Сравнительный анализ в биологии как правило осуществляется для изучения фенотипов. Фенотипом могут быть как характеристики индивидуумов (например: рост, размер зубов или длина хвоста), так и молекулярные свойства (например: сила сайта связывания транскрипционного фактора, белок-кодирующий потенциал РНК или энергия вторичной структуры РНК). При сравнении фенотипа внутри или между видами необходимо иметь представление об эволюционном процессе лежащем в основе его изменений. Масштаб на которых происходят эволюционные процессы варьируется от небольших изменений между поколениями, что изучает микроэволюция, и до крупных изменений между видами, что изучает макроэволюция.

1.4.1. Микроэволюция количественных характеристик

Первая полная модель динамики изменений фенотипа в популяции была разработана Ланде в 1976 [85]. К тому времени существовала достаточно подробно разработанная теория динамики частот аллелей в популяции [86], однако на практике можно было измерить только количественные характеристики, такие как размер зубов, а не частоту аллелей. Ниже в общих чертах представлена модель Ланде.

Рассмотрим фенотип z в популяции с непересекающимися поколениями $\{t\}$, имеющий функцию приспособленности $W(z)$. Функция приспособленности означает, что индивидуум с фенотипом z будет иметь $W(z)$ потомков. Чем выше $W(z)$, тем более адаптивна особь к условиям

окружающей среды и способна оставить больше потомства. Индивидуумы в популяции имеют различные фенотипы, а значит имеют разную приспособленность. Как следствие, представленность фенотипов в популяции в следующем поколении будет зависеть от их приспособленности. Таким образом, средний фенотип популяции $\bar{z}(t)$ будет изменяться в ходе эволюции. Фенотипом популяции или вида называется средний фенотип популяции. Как правило, исследования фокусируются на сравнении фенотипа между видами, поэтому важно изучить законы эволюции $\bar{z}(t)$ в рамках некоторых логичных предположений.

На реальное количество потомства, кроме приспособленности индивидуума, также влияют случайные факторы. Сначала предположим, что не существует случайных факторов и количество потомства определяется однозначно. В этом случае изменение фенотипа популяции между поколениями детерминировано и определяется, как показал Ланде, следующим образом:

$$\Delta\bar{z}(t) = \bar{z}(t+1) - \bar{z}(t) = \sigma^2 \frac{\partial \ln \overline{W}}{\partial \bar{z}(t)},$$

где $\overline{W}(t)$ - средняя приспособленность популяции. При этом делается предположение, что распределение фенотипов внутри популяции имеет нормальное распределение со средним $\bar{z}(t)$ и некоторой постоянной дисперсией σ^2.

Таким образом, средний фенотип изменяется увеличивая приспособленность популяции, пока не приходит в стационарное состояние с локально наилучшей приспособленностью. Для описание эволюции популяции широко используется понятие адаптивной

поверхности фенотипов [87]: по осям находятся частоты $p(z,t)$ каждого фенотипа в популяции, высота характеризует среднюю приспособленность $\bar{W}(t)$ такой популяции, а точка на этой поверхности характеризует популяцию в момент t. В описанной модели движение популяции происходит в сторону локального максимума адаптивной поверхности.

Теперь предположим, что количество потомства индивидуума также определяют множество случайных факторов. Изменение частот аллелей из-за случайных факторов называется генетическим дрейфом. Как следствие, вклад в эволюцию фенотипа популяции вносят как естественный отбор, так и генетический дрейф. В таком случае, изменение фенотипа популяции является стохастическим процессом, а $\bar{z}(t)$ описывается некоторым распределением вероятностей. Из общей теории известно [88], что в популяции эффективного размера N сила генетического дрейфа $\sim {}^{1}/_{N}$:

если большое количество индивидуумов имеют данный фенотип, то суммарно случайная вариация в количестве их потомства нивелирует друг друга; и наоборот, при малом количестве индивидуумов стохастический вклад каждого становится заметным. При логичных предположениях о нормальности функции приспособленности, фенотип $\bar{z}(t)$ в каждом поколении будет иметь нормальное распределение с параметрами:

$$\bar{z}(t) = \bar{z}(0)e^{-\frac{\sigma^2}{\sigma^2+w^2}\cdot t}, \tag{2}$$

$$\sigma^2(t) = \frac{\sigma^2+w^2}{2N}\left(1-e^{-2\frac{\sigma^2}{\sigma^2+w^2}\cdot t}\right), \tag{3}$$

где w^2 - дисперсия функции приспособленности.

Генетический дрейф порождает распределение возможных средних фенотипов $\Phi(\bar{z}(t))$, и это распределение экспоненциально сходится к

нормальному с нулевым средним фенотипом и дисперсией $\frac{\sigma^2+w^2}{2N}$ (Рисунок 1.4.1).

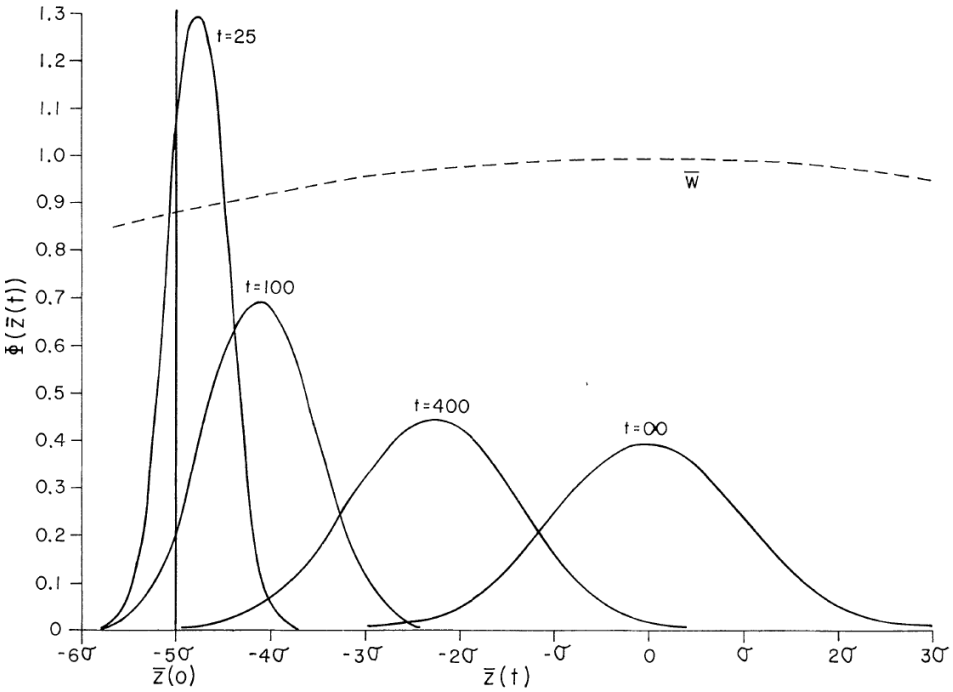


Рисунок 1.4.1. Динамика стохастической эволюции средних фенотипов $\Phi(\bar{z}(t))$: на рисунке изображено распределение вероятностей возможных фенотипов $\bar{z}(t)$ для промежутков времени $t = 0, 25, 100, 400, +\infty$ (число характеризует количество поколений). Адаптивная поверхность $\overline{W}$ обозначена пунктирной линией. (из [85])

Данный результат можно интерпретировать следующим образом: если "запустить" эволюцию популяции много раз, то каждый раз средний фенотип будет описывать некоторую стохастическую траекторию $\bar{z}(t)$, в то время как ансамбль "запусков" будет иметь распределение средних фенотипов $\bar{z}(t)$ описывающихся уравнениями (2), (3).

1.4.2. Процесс Орнштейна-Уленбека

Стохастические траектории среднего фенотипа в модели Ланде подчиняются случайному процессу Орнштейна-Уленбека:

$$dX(t) = \alpha\big(\theta - X(t)\big)dt + \sigma dB(t), \qquad (4)$$

где $dX(t)$ - изменение случайной величины $X(t)$ за малый промежуток dt, а $dB(t)$ - "белый шум" описываемый стандартным броуновским движением, то есть независимыми нормальными распределениями с параметрами $N(0, dt)$. При $\alpha = 0$, процесс Орнштейна-Уленбека превращается в процесс броуновского движения и характеризуется только "силой" флуктуаций σ (Рисунок 1.4.2А). На качественном уровне, броуновское движение объекта (или процесс Винера) возникает под действием большого количества случайных независимых возмущений, которые, по центральной предельной теореме, приводят к нормальному распределению ожидаемых траекторий. Согласно эволюционной интерпретации, броуновское движение описывает нейтральную эволюцию фенотипа под действием генетического дрейфа [89].

В применении к модели Ланде, α описывает силу отбора, θ - оптимальное значение фенотипа, а σ - силу генетического дрейфа. Детерминистическая часть уравнения Орнштейна-Уленбека

$$\alpha\big(\theta - X(t)\big)dt$$

имеет линейную форму: чем дальше процесс находится от оптимального значения, тем сильней его "тянет" в сторону оптимума (Рисунок 1.4.2Б). Рядом с оптимум, наоборот, главную роль играют стохастические флуктуации $\sigma dB(t)$. Эти свойства согласуются с качественной интерпретацией стабилизирующего отбора, поэтому процесс Орнштейна-Уленбека стал популярным средством для моделирования фенотипической эволюции: отбор "тянет" фенотип к оптимальному значению, но генетический дрейф препятствует этому движению.

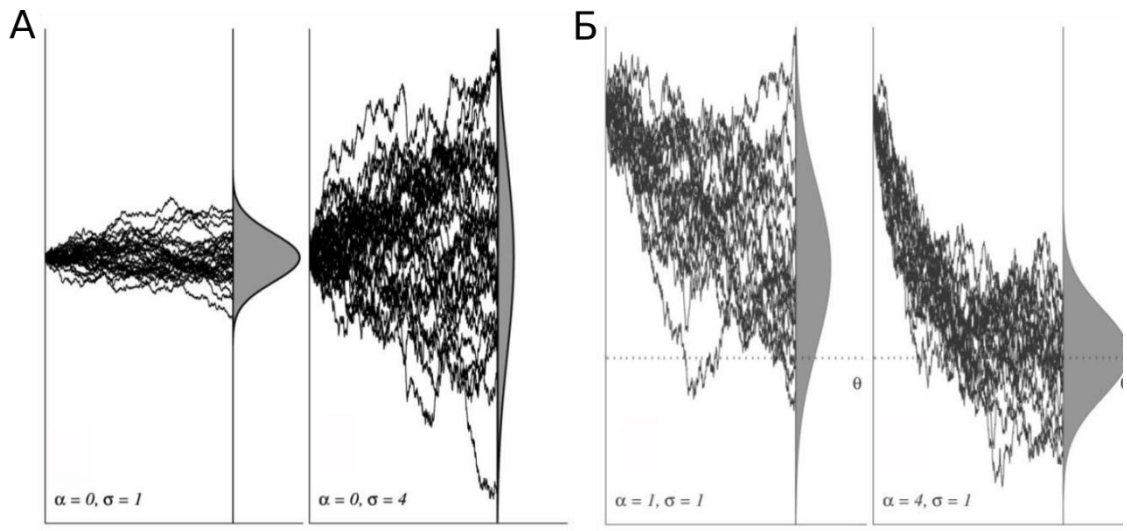


Рисунок 1.4.2. Динамика броуновского движения и процесса Орнштейна-Уленбека со временем. А) 30 реализаций броуновское движение. Параметр σ характеризует "силу" случайных флуктуаций. Б) 20 реализаций процесса Орнштейна-Уленбека в промежутке времени от t=0 до t=1. Параметр α характеризует скорость сходимости к стационарному состоянию и дисперсию стационарного распределения. (из [90])

Процесс Орнштейна-Уленбека является частным случаем стохастического дифференциального уравнения

$$dX(t) = a(x)dt + b(x)dB(t), \qquad (5)$$

однако общее уравнение не имеет аналитического решения. Поэтому процесс Орнштейна-Уленбека является компромиссным вариантом между математической простотой описания и качественным представлением эволюции.

1.4.3. Макроэволюция количественных характеристик

Богатый набор фенотипических данных между видами позволяет задавать вопросы о механизмах видообразовании и эволюции отдельных фенотипов. Микроэволюционная теория Ланде изучает изменения фенотипа внутри вида на протяжении сотен или тысяч поколений.

Нуклеотидная замена, то есть закрепление новой мутации в популяции, происходит значительно быстрее чем ожидаемое время новой мутации. В популяции с эффективной численностью N время закрепления нейтральной мутации в среднем составляет $4N$ [91], а ожидаемое время возникновения новой мутации - $1/\mu$, где μ - скорость мутаций на поколение. В человеческой популяции $\mu \sim 10^{-8}$ и $N \sim 10^{5}$ [92], поэтому время между двумя мутациями на несколько порядков превосходит время закрепления мутации. Это означает, что процесс закрепления мутации происходит на совсем иных временных промежутках, чем время расхождение между современными видами.

Опираясь на эти соображения, Хансен отметил, что модель Ланде не подходит для изучения межвидовых изменений фенотипов [93]. Популяция экспоненциально быстро сходится к стационарному распределению среднего фенотипа на адаптивной поверхности за счет сил отбора и дрейфа. А межвидовые изменения фенотипа наблюдаются из-за изменения адаптивной поверхности, или, проще говоря, сдвига оптимального (или наиболее приспособленного) значения фенотипа θ. Изменение оптимального фенотипа может диктоваться изменениями окружающей среды или соотношением фенотипа с другими фенотипами в организме. Хансен предложил использовать процесс Орнштейна-Уленбека для интерпретации изменения оптимального значения фенотипа на больших масштабах времени: в ходе эволюции вида различные факторы изменяют оптимальный фенотип случайным образом, однако существуют возможные оптимальные фенотипы при которых вид не выживет [94], что ограничивает дрейф оптимального фенотипа (Рисунок 1.4.3).

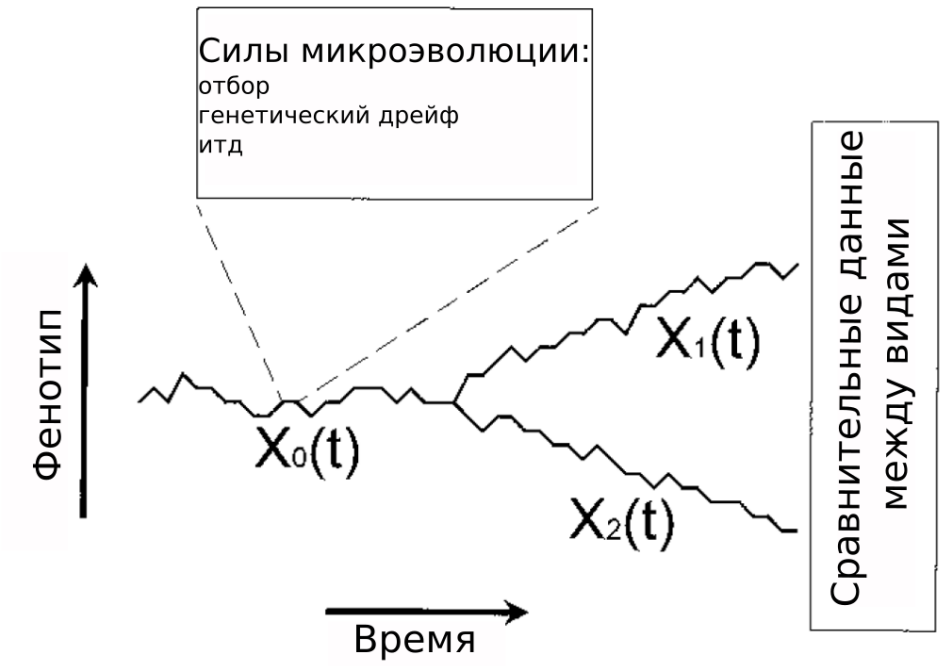


Рисунок 1.4.3. Эволюционные силы на разных масштабах времени. Эволюция фенотипа между видами диктуется сдвигом адаптивного ландшафта. В то время как силы отбора и генетического дрейфа оперируют лишь на малых промежутках времени, очень быстро адаптируя популяцию к новой адаптивной поверхности. $X_0(t)$, $X_1(t)$, $X_2(t)$ описывают эволюцию фенотипа в предковом и современных видах соответственно. (из [93])

Тем не менее, эволюционная интерпретация процесса Орнштейна-Уленбека при анализе фенотипов неоднозначна [90]: происходит ли случайное блуждание вокруг оптимального фенотипа, или самого оптимального фенотипа. Мы будем использовать интерпретацию, при которой случайное блуждание происходит вокруг фиксированного оптимального фенотипа. При сравнении видов ключевым является определение фенотипов, которые подверглись сильным адаптивным изменениям, то есть сдвигу оптимального фенотипа.

Представим использование процесса Орнштейна-Уленбека в случае двух видов (Рисунок 1.4.4). Пусть θ_0 - неизвестное предковое состояние

фенотипа, а θ_1 - оптимальное значение. Эволюция фенотипов представляется вектором $\boldsymbol{X}(t) = (x_1(T), x_2(T))$, где $x_1(T)$ и $x_2(T)$ описывают эволюционные траектории первого и второго вида соответственно.

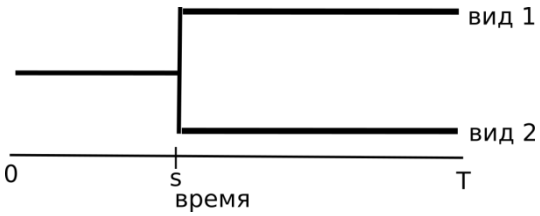


Рисунок 1.4.4. Эволюционное дерево двух видов.

Вектор $\boldsymbol{X}(T)$ имеет двумерное нормальное распределение со средним

$$E[x_i(T)] = \theta_0 e^{-\alpha T} + \theta_1(1 - e^{-\alpha T}), i = 1,2.$$

Ковариация между видами вычисляется как

$$cov\big(x_1(T), x_2(T)\big) = e^{-2\alpha(T-s)} \cdot (1 - e^{-2\alpha s}),$$

поэтому матрица ковариаций $\boldsymbol{X}(T)$ выглядит следующим образом:

$$\boldsymbol{V} = \frac{\sigma^2}{2\alpha} \cdot \begin{pmatrix} 1 - e^{-2\alpha T} & e^{-2\alpha(T-s)} \cdot (1 - e^{-2\alpha s}) \\ e^{-2\alpha(T-s)} \cdot (1 - e^{-2\alpha s}) & 1 - e^{-2\alpha T} \end{pmatrix}.$$

Аналогичным образом $E[\boldsymbol{X}(T)]$ и $\boldsymbol{V}$ определяются для произвольного филогенетического дерева видов. Так как $\boldsymbol{X}(t)$ многомерное нормальное распределение, то на основе $E[\boldsymbol{X}(T)]$ и $\boldsymbol{V}$ вычисляется правдоподобие данных

$$L(\alpha, \sigma, \theta_0, \theta_1) = (\boldsymbol{X}(T) - E[\boldsymbol{X}(T)]) \cdot \boldsymbol{V}^{-1} \cdot (\boldsymbol{X}(T) - E[\boldsymbol{X}(T)]) + N \log(2\pi det \boldsymbol{V}).$$

Оптимизируя правдоподобие $L(\alpha, \sigma, \theta_0, \theta_1)$, можно оценить оптимальные параметры и качество модели на основе нелинейной оптимизации. Отдельный вид или клада могут иметь другой оптимальный фенотип θ_2, что интегрируется в новую модель с правдоподобием $L(\alpha, \sigma, \theta_0, \theta_1, \theta_2)$. Сравнение правдоподобий с помощью теории вложенных моделей позволяет оценить статистическую достоверность события смены оптимального фенотипа [90].

Таким образом, применение макроэволюционной модели Орнштейна-Уленбека используется для

- определения силы стабилизирующего отбора и дрейфа фенотипа
- поиска или тестирования гипотезы событий смены оптимального фенотипа

В современных приложениях, процесс Орнштейна-Уленбека широко применяется при анализе эволюционных сил действующих на экспрессию генов [95], [96].

Глава 2. RNASurface: эффективный алгоритм предсказания локально-оптимальных структурированных РНК

2.1. Алгоритм и методы

2.1.1. Матрица Z-значений и локально-оптимальные сегменты

Известные методы поиска структурированных РНК сканируют последовательность фиксированным окном. Как следствие, только структурные РНК с длинами близкими к длине окна будут предсказаны. Грубый и неэффективный вариант решения этой проблемы - осуществить много запусков с разными размерами окон, и затем каким-либо образом объединить и профильтровать предсказания. Мы предлагаем другой подход: вычислить Z-значение Z_{ij} для каждого сегмента S_{ij} последовательности S до определенного размера (т.е. $j - i + 1 \leq L$). Лучшие "представители" этого набора $\{Z_{ij}\}$ соответствуют всем наиболее структурированным сегментам до размера L.

При оценке оптимальной свободной энергии E, алгоритм Зукера рекуррентно вычисляет оптимальные энергии E_{ij} всех подпоследовательностей S_{ij} и записывает их в матрицу энергий. Поэтому, вычислив оптимальную энергию последовательности, получаем и оптимальные энергии всех её подпоследовательностей. Если известны параметры μ_{ij} и σ_{ij} фонового распределения энергий участка S_{ij}, то для этого участка известно и Z-значение (Рисунок 2.1.1А):

$$Z_{ij} = \frac{E_{ij} - \mu_{ij}}{\sigma_{ij}} \qquad (6)$$

Определение μ_{ij} и σ_{ij} является отдельной проблемой и отдельно разбирается в следующем параграфе. Свободная энергия отрицательна, и чем она ниже, тем более структурирована последовательность. Поэтому

низкие отрицательные Z-значения соответствуют хорошо структурированным последовательностям.

Для удобства визуализации, полученную матрицу Z-значений можно представить в виде тепловой карты, сопоставив численным значениям Z-значениям градиент цветов (синий - низкое Z-значение, структурированные участки; красный - высокое Z-значение, неструктурированные участки) и повернуть на $45°$ для более удобного сопоставления Z-значения и участка (Рисунок 2.1.1Б). Таким образом, при сканировании длинной последовательности окном строится тепловая карта структурированности всех её подпоследовательностей (Рисунок 2.1.1В). Тепловая карта Z-значений представляет значительно больше информации о потенциальных структурных РНК, позволяя выделять пики Z-значений, анализировать взаимное расположение и размер структурных РНК.

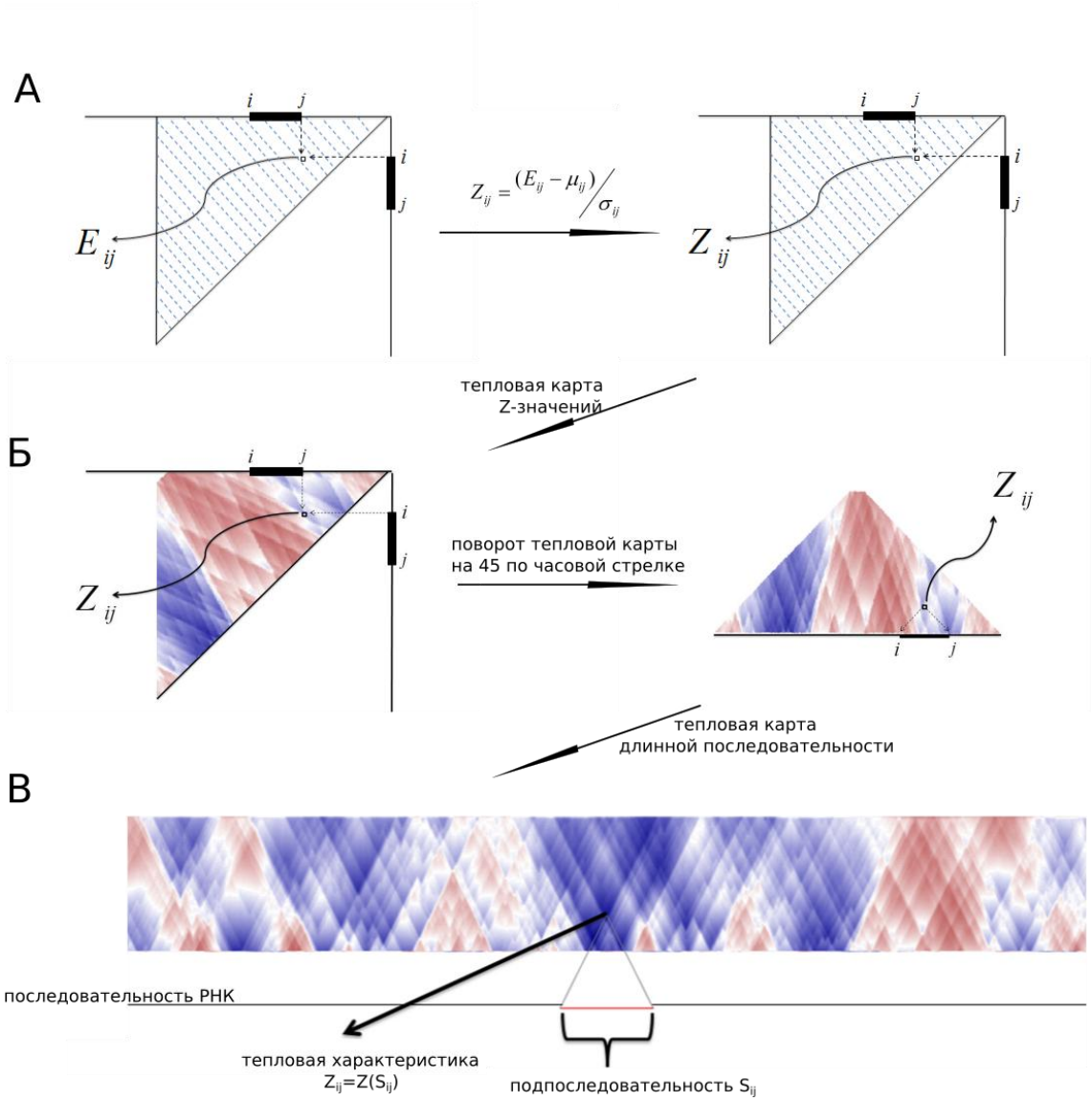


Рисунок 2.1.1. Создание тепловой карты Z-значений. А) По матрице энергий вычисляется матрица Z-значений. Б) Численные Z-значения заменяются градиентом цвета: от наиболее структурированных участков синим цветом до наиболее неструктурированных участков красным цветом. Диагональ можно сопоставить последовательности, а значит каждую точку тепловой карты можно сопоставить сегменту диагонали. Для удобства визуализации, поворачиваем на 45° и "кладем" тепловую карту на диагональ. В) Пример тепловой карты последовательности в 1000 нк из генома *Bacillus subtilis* с окном $L = 200$ нк, содержащей SAM рибо-переключатель посередине. Выделенная красным подпоследовательность S_{ij} имеет длину 100 нк, наклонные линии от неё ведут в точку соответствующую её структурированности Z_{ij}.

Если точка на тепловой карте имеет Z-значения ниже чем её соседи, то она представляет локальный оптимум. Небольшие сдвиги границ подпоследовательности, соответствующей локально-оптимальной точке, только увеличивают Z-значение и ухудшают структурированность. Получается, что подпоследовательность локально-оптимальной точки отвечает точному расположению структурированного сегмента, и сканирование любым окном с любым шагом не может улучшить её Z-значение. Более формально, сегмент S_{lo} называется k-локально-оптимальным, если

$$\forall S' : |S' \cup S_{lo}| - |S' \cap S_{lo}| \leq k \Rightarrow Z(S') > Z(S_{lo}),$$

где параметр k соответствует амплитуде изменения границ сегмента, или радиусу соседних точек на тепловой карте. Большие значения k захватывают более глобальные оптимумы, и значительно сокращают список предсказаний; таким образом, параметр k контролирует связь между аккуратностью и размером предсказаний.

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. Для построение тепловой карты Z-значений необходимо эффективно вычислять параметры фонового распределения энергий для всех подпоследовательностей. Подробное решение этой задачи представлена в следующем параграфе.

Отметим, что необычно высокие Z-значения могут свидетельствовать об одноцепочечном участке РНК, доступном для взаимодействия с другими молекулами, например РНК-связывающими белками [97]. Последующее

изложение фокусируется на анализе структурированных РНК, однако вся методология с минимальными изменениями переносится на анализ доступных участков РНК. Опция анализа локально-оптимальных доступных для взаимодействия сайтов также реализована в нашей программе RNASurface.

2.1.2. Эффективное вычисление Z-значений

Параметры фонового распределения энергий сегмента зависят от длины и динуклеотидного состава. Наиболее эффективный подход к вычислению параметров был предложен Вашитлем и реализован в программах RNAz [74] и RNALfoldz [77]. Параметры вычисляются с помощью регрессии опорных векторов по входящему набору свойств $(l, d_1, \dots, d_{16})$, где l - длина сегмента, а $d_1, \dots, d_{16}$ - набор частот динуклеотидов. Для получения тепловой карты Z-значений необходимо совершить $\sim N \cdot L^2$ вычислений Z-значений. Вычисление регрессия опорных векторов происходит за $O(1)$, однако требует большого количества операций C_r, поэтому при умеренной длине окна ($L < 1000$) на практике количество операций составляет:

$$C_r \cdot L^2 \cdot N \gg L^3 \cdot N.$$

В тоже время сканирование и пересчет матрицы энергий может быть произведен за $O(L \cdot N)$, поэтому оценка Z-значений становится "бутылочным горлышком" подхода. Практическую оценку времени вычисления тепловой карты Z-значений и вклад в весь алгоритм можно получить следующим образом. Программа RNALfoldz сканирует геном окном L и вычисляет Z-значение методом регрессии опорных векторов в некоторых, но не всех окнах. Если такая реализация RNALfoldz требует t единиц времени на вычисление Z-значения, то вычисление тепловой карты

для всех сегментов длины $\leq L$ занимает $T \geq t \cdot L/2$ единиц времени. Работа RNALfoldz на геноме *E.coli* с с окном 300 нк (которое захватывает почти все аннотированные структурные РНК в *E.coli*) занимает больше 1 часа на современном процессоре Intel Xeon Processor E5506, при этом 60-70% времени занимает вычисление Z-значений [77]. Получается, что вычисление всей тепловой карты на *E.coli* методом RNALfoldz занимает $T > 90$ часов и составляет больше 98% времени от общей работы алгоритма. Запуск с аналогичными параметрами на геноме человека составит 5-8 лет!

Таким образом, для практической реализации тепловой карты необходим принципиально новый подход к оценке Z-значения. 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 $(d_1, \dots, d_{16})$, образуя некоторую поверхность в пространстве динуклеотидных частот. Тем не менее, эта поверхность хорошо приближается квадратичной регрессией. А квадратичную регрессию можно очень быстро рекуррентно пересчитывать при сканировании вдоль генома.

Использование этих соображений позволяет вычислять тепловую карту точно и за малое время по сравнению с пересчетом энергий алгоритмом Зукера.

2.1.2.1. Зависимость энергетических параметров от длины

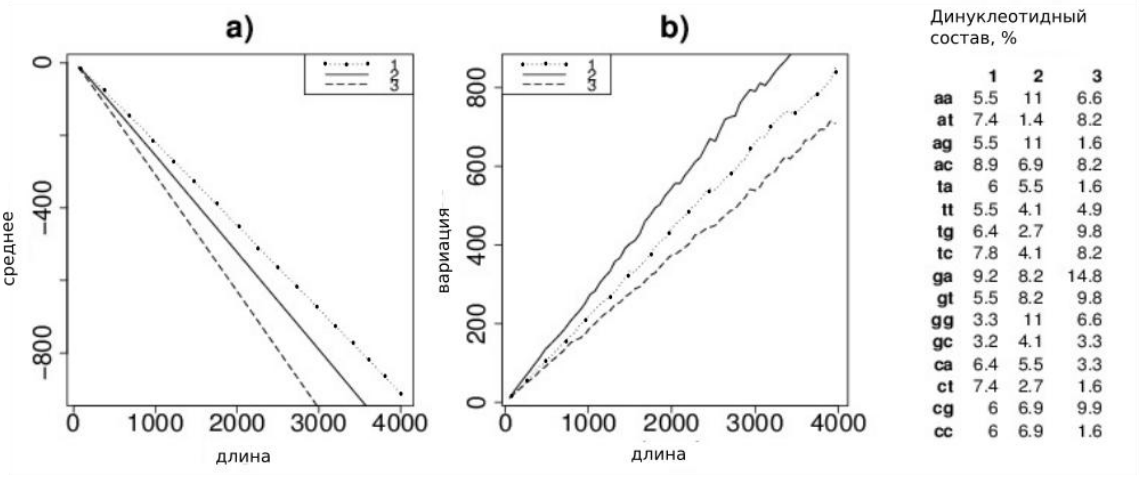
Зафиксируем динуклеотидные частоты, и будем генерировать последовательности марковской моделью первого порядка, то есть вероятность генерации следующего нуклеотида пропорциональна частоте динуклеотида, который он порождает. Ранее была доказана теорема [62], что при такой модели генерации, ожидаемая энергия ведет себя асимптотически линейно от длины последовательности:

$$\lim_{l \to +\infty} \frac{\mu(l)}{l} = M,$$

где $\mu(l)$ - ожидаемая энергия последовательностей длины l генерируемых общей марковской моделью 1-ого порядка. Этот факт имеет наглядное объяснение: с одной стороны, минимальная свободная энергия складывается из свободной энергии своих подструктур, поэтому не может понижаться медленней линейной функции; с другой стороны, минимальная энергия любой последовательности ограничена снизу GC периодичной шпильки, энергия которой понижается линейно. Поэтому ожидаемая энергия, будучи зажата линейными функциями, имеет линейную аппроксимацию.

Опираясь на это наблюдение, мы провели симуляции последовательностей от 20 до 4000 нк для разных наборов динуклеотидных частот с целью изучить зависимость параметров $\mu(l)$ и $\sigma(l)$ от длины l. Для каждой длины и динуклеотидного состава мы сгенерировали 500 последовательностей и оценили минимальную свободную энергию алгоритмом RNASlider, после чего вычислили её

среднее и дисперсию. Как и ожидалось, зависимость $\mu(l)$ имеет линейную аппроксимацию (Рисунок 2.1.2А). Более неожиданно, $\sigma^2(l)$ также имеет линейное приближение (Рисунок 2.1.2Б).



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 2.1.2. Зависимость среднего (А) и дисперсии (Б) минимальной свободной энергии от длины последовательности. Три прямые соответствуют трем различным динуклеотидным составам, с которыми генерировались последовательности.

Несмотря на визуально линейный вид, эти зависимости не являются точной линейной функцией. Это связано с тем, что свободная энергия складывается из стекинг взаимодействий, ожидаемая энергия которых увеличивается линейно с длиной, и петель, ожидаемая энергия которых увеличивается сильно нелинейно с длиной [98]. Поэтому мы разбили разумную протяженность длин на интервалы, внутри каждого из которых использовали линейное приближение. А именно, мы использовали 11 контрольных длин $\{l_i\}$:

$$l_i = \{,80,160,300,600,1000,1040,1080,1120,1160,1200\}$$

Аппроксимация $\mu(l)$ по контрольным длинам $\{l_i\}$ выглядит следующим образом:

$$\mu(l) = \begin{cases} \mu(l_i) \cdot (l_{i+1} - l) + \mu(l_{i+1}) \cdot (l - l_i) & , \quad l_i \le l \le l_{i+1} \\ \text{линейная регрессия по точкам } 1000, \dots, 1200, & l > 1200 \end{cases} \quad (7)$$

Плотное множество точек 1000-1200 выбрано для построение достоверной регрессии для очень длинных последовательностей.

Аппроксимация для $\sigma(l)$ по контрольным длинам $\{l_i\}$ выглядит следующим образом:

$$\sigma(l) = \begin{cases} \sigma(60) \cdot (80 - l) + \sigma(80) \cdot (l - 60) & , l \le 80 \\ c \cdot \sqrt{l + \alpha} & , l > 80 \end{cases} \quad (8)$$

Параметры c и α вычисляются так чтобы кривая совпадала со стандартным отклонением в точках $l = 80, 1000$:

$$\begin{cases} c \cdot \sqrt{80 + \alpha} = \sigma(80) \\ c \cdot \sqrt{1000 + \alpha} = \sigma(1000) \end{cases}$$

Таким образом, достаточно хранить значения параметров для небольшого набора контрольных длин, а для остальных длин вычислять по формулам (7)-(8).

2.1.2.2. Зависимость энергетических параметров от динуклеотидного состава

При фиксированной контрольной длине l, аппроксимируем зависимость параметров μ и σ от динуклеотидных частот $(d_1, \dots, d_{16})$ набором квадратичных регрессий:

$$\mu(d_1, \dots, d_{16}|l) = \sum_{1 \le i < j \le 16} c_{ij} d_i d_j.$$

Формулы и рассуждения приводятся далее для $\mu(d_1, \dots, d_{16}|l)$, для стандартного отклонения $\sigma(d_1, \dots, d_{16}|l)$ все результаты аналогичны.

Поверхность $\mu(d_1, \dots, d_{16}|l)$ в пространстве $(d_1, \dots, d_{16})$ имеет сложную форму. Квадратичная регрессия способна принимать широкий набор поверхностей из-за большого количества степеней свободы в виде параметров $\{c_{ij}\}$. Хотя одна квадратичная регрессия плохо описывает форму этой поверхности на всем пространстве $(d_1, \dots, d_{16})$, его разбиение на подмножества позволяет построить для каждого из них хорошее приближение квадратичной регрессией. Более формально, ограничим анализ последовательностями S с отношениями $\frac{G+C}{A+T+G+C}$, $\frac{G}{G+C}$, $\frac{A}{A+T}$ от 0.2 до 0.8:

$$\Gamma = \{ S \mid 0.2 \leq \frac{G+C}{A+T+G+C}, \quad \frac{G}{G+C}, \quad \frac{A}{A+T} \leq 0.8\},$$

как показано ранее [74], более 99% генома млекопитающих, включая почти все известные функциональные РНК, попадают в множество Γ. Разобьем множество Γ на 27 подмножеств $\Gamma_1, \dots, \Gamma_{27}$ поделив каждое из направлений $(\frac{G+C}{A+T+G+C}, \frac{G}{G+C}, \frac{A}{A+T})$ на три части: 0.2-0.4, 0.4-0.6, 0.6-0.8. Для каждого подмножества Γ_i построим свою квадратичную регрессию: вычислим значения параметра μ для 20'000 последовательностей, равномерно распределенных в области Γ_i; по этим 20'000 значениям μ оценим коэффициенты c_{ij} методом наименьших квадратов.

В итоге, для каждой из 11 контрольных длин построено 27 квадратичных регрессий для непересекающихся подмножеств пространства динуклеотидных частот.

Квадратичная регрессия позволяет эффективно пересчитывать параметры при сканировании вдоль генома. Предположим, что для подпоследовательности S_{ij}, $L = j - i + 1$, с частотами $(d_1, \dots, d_{16})$ уже вычислена квадратичная аппроксимация:

$$\mu(S_{ij}) = \sum_{k<l} c_{kl} d_k d_l.$$

У подпоследовательности $S_{i+1,j+1}$ только две динуклеотидные частоты отличаются от S_{ij}, для определенности d_{k_0} увеличилась на $\frac{1}{L-1}$, а d_{l_0} уменьшилась на $\frac{1}{L-1}$. Тогда выполнено:

$$\mu(S_{i+1,j+1}) - \mu(S_{ij}) = \frac{1}{L-1}\sum_{l} c_{k_0 l} d_l - \frac{1}{L-1}\sum_{k} c_{k l_0} d_k \; - \; c_{k_0 l_0} \cdot (d_{k_0} - d_{l_0} - \frac{1}{L-1}).$$

Две суммы справа аналогичным способом вычисляются рекуррентно, что в общей сложности позволяет осуществить пересчет параметра μ всего за несколько операций. Если длинная последовательность сканируется коротким окном (например, последовательность длины 10000 нуклеотидов сканируется окном 200 нуклеотидов), то после вычисления параметров в начальном окне их пересчет при дальнейшем сканировании происходит моментально.

Известно что структурные РНК более GC богаты чем соседние геномные участки [99], это объясняется небольшим давлением отбора на прочные GC связи и является дополнительном сигналом структуры [100]. Чтобы выделить этот сигнал, для каждого сегмента S_{ij} рассмотрим его геномное окружение: участок $S_{i-d,j+d}$, в который вложен сегмент. При оценке Z_{ij} будем рассматривать фоновые последовательности с динуклеотидным составом окружения $S_{i-d,j+d}$, а не самого сегмента S_{ij} (Рисунок 2.1.3). Эта процедура не влияет на эффективность алгоритма, и при $d = 0$ возвращает стандартную модель генерации случайных последовательностей. Также окружение длиной $\sim 500 - 2000$ нк позволяет с хорошей степенью достоверности оценить частоты, что невозможно для коротких сегментов.

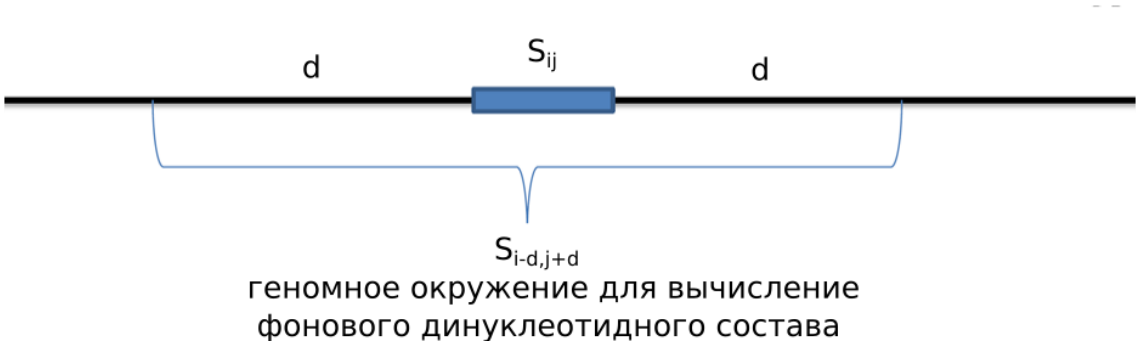


Рисунок 2.1.3. Геномное окружение сегмента используется для оценки динуклеотидного состава сегмента и генерации фонового распределения энергий.

2.1.2.3. Качество аппроксимации

Вычисление Z-значения последовательности представляет следующую процедуру:

- Определяем динуклеотидный состав, после чего вычисляем параметры распределения для контрольных длин, используя квадратичную регрессию.
- Вычисляем параметры для длины данной последовательности, используя линейную аппроксимацию контрольных длин.

Чтобы оценить качество всей аппроксимации, были выбраны случайным образом 24'000 последовательности из геномов человека и дрозофилы: по 25 последовательностей каждой длины от 60 до 99 нк, и по 10 последовательностей каждой длины от 100 до 2400 нк. Для каждой последовательности мы определили Z-значение согласно аппроксимации, и реальное Z-значение. Реальное Z-значение вычисляется с помощью интенсивных симуляций, а именно для каждой из 24'000 последовательностей были сгенерированы 2000 случайных последовательностей с тем же ожидаемым динуклеотидным составом и

длиной для вычисления параметров распределения энергий. Аппроксимация демонстрирует очень высокое сходство с реальными Z-значениями, с коэффициентом линейной регрессии $R^2 = 0.997$ (Рисунок 2.1.4А). При этом качество предсказания одинаково высокое для всех длин (Рисунок 2.1.4Б), и остается крайне высоким для длин больше 2000 нк, хотя крайняя контрольная длина равна 1200 нк. Средняя ошибка (модуль разницы между аппроксимацией и реальным Z-значением) при аппроксимации Z-значением составляет 0.1. Для небольших последовательностей, до 250 нк, средняя ошибка падает до 0.041, что сопоставимо с ошибкой 0.076 при аппроксимации регрессией опорных векторов.

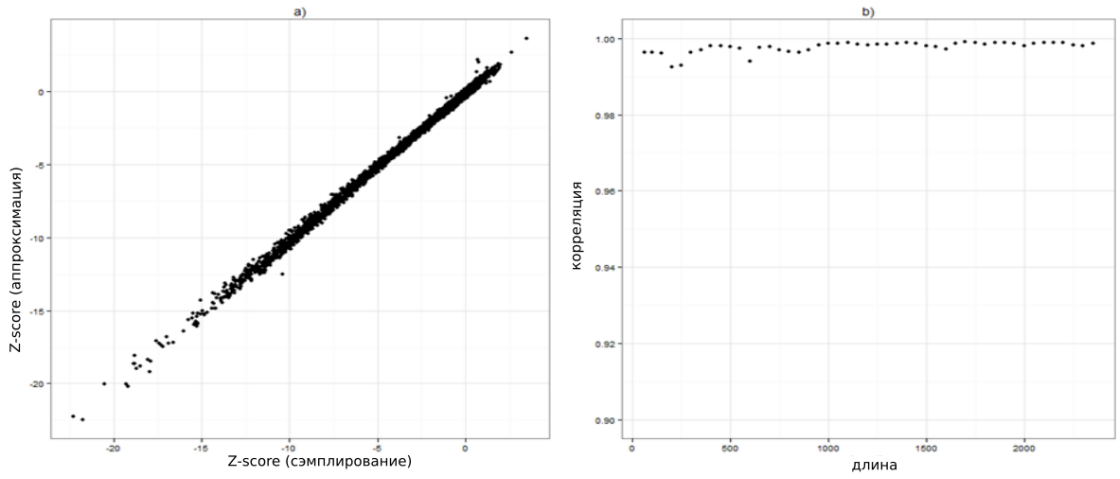


Рисунок 2.1.4. Качество аппроксимации. А) Высокий уровень соответствия Z-значений полученных аппроксимацией и из симуляций. Б) Корреляция Пирсона двух методов вычисления Z-значений одинаково высока для широкого диапазона длин. Для каждой длины L вычислялась корреляция Z-значений для последовательностей в диапазоне длин $[L - 50, L + 50]$.

Таким образом, наш подход к вычислению Z-значений значительно ускоряет эту процедуру, не уступает в качестве приближений, а также адаптирован для широкого диапазона длин.

2.1.2.4. Сглаживание матрицы Z-значений

Как показывают наблюдения, тепловая карта Z-значений имеет плотное множество локальных оптимумов. Это связано как со сложной организацией пространства вторичных структур РНК [101], так и с погрешностями при аппроксимации Z-значений. Как следствие, значимая вторичная структура РНК часто соответствует нескольким близким пикам. Чтобы устранить эту ситуацию, и сделать выдачу локально-оптимальных сегментов разреженной и отвечающей ситуации *различных* локально-оптимальных структур, мы сглаживаем матрицу Z-значений с экспоненциальным ядром. Определим расстояние между сегментами $[i, j]$ и $[k, l]$ следующим образом:

$$\delta(i, j; k, l) = |k - i| + |l - j|$$

Будем считать, что сегмент $[k, l]$ воздействует на сегмент $[i, j]$ с экспоненциальным весом $e^{-\lambda\delta}$, где параметр λ будет описан ниже. Чем больше два сегмента пересекаются, тем больше воздействуют друг на друга. По матрице Z_{ij} построим сглаженную на s нуклеотидов матрицу Z_{ij}^{s} следующим образом:

$$Z_{ij}^{s} = \frac{1}{C_s} \sum_{\delta=|k-i|+|l-j|\leq s} Z_{kl} e^{-\lambda\delta},$$

где C_s - суммарное воздействие соседних сегментов на расстояние $\delta \leq s$ до данного сегмента $[i, j]$:

$$C_s = \sum_{\delta=|k-i|+|l-j|\le s} e^{-\lambda\delta}$$

Параметр λ определим исходя из следующего соображения: при сглаживании учитываются только ближние соседи на расстояние $\le s$, поэтому логично считать, что их воздействие на сегмент должно определять значительную долю p общего воздействие всех сегментов. Более формально условие выглядит следующим образом:

$$C_s/C_\infty \ge p, \tag{9}$$

где p достаточно большая доля, например 0.9, а C_∞ определяет общее воздействие всех сегментов на сегмент $[i, j]$:

$$C_\infty = \sum_{\delta=0}^{\infty} e^{-\lambda\delta}. \tag{10}$$

Параметр λ определятся из условий (9)-(10). Примитивные подход вычисления Z_{ij}^s занимает $O(s^2)$, что при $s > 7$ на практике занимает большую часть времени работы программы. Поэтому мы используем эффективный алгоритм пересчета за $O(1)$. Основную идею продемонстрируем на примере экспоненциального сглаживания вектора чисел (а не матрицы). Дан вектор x_i и его сглаживание x_i^s:

$$x_i^s = \sum_{i-s\le k\le i} x_k e^{-k}$$

Если уже получено сглаживание i-ого элемент x_i^s, то сглаживание x_{i+1}^s определяется за несколько операций:

$$x_{i+1}^s = \left(x_i^s - x_{i-s}e^{-\lambda(i-s)}\right)e^{-\lambda} + x_{i+1}$$

В случае двумерной матрицы, вычисление рекуррентных соотношений между элементами матрицы занимают также ограниченное количество операций, однако требует десять рекуррентных уравнений.

2.1.3. Профили структурированности РНК

Корреляция структурированности РНК с профилем различных биологических свойств вдоль генома (например, границы гена и кодирующей области, плотность рибосом, уровень экспрессии, участки взаимодействия РНК-связывающих белков) является важным методом проверки или формирования биологических гипотез [102]. Для такого анализа необходимо иметь простое и удобное представление информации о структурированности РНК, что сложно осуществить с помощью двумерной тепловой карты Z-значений. Опираясь на тепловую карту, мы представляем две одномерные меры структурированности РНК, которые просто сравнивать с геномными профилями.

Мера MZ позиции i является максимумом по всем квадратам Z-значений последовательностей, накрывающих эту точку посередине:

$$MZ(i) = \max_{i-l=r-i} Z_{kl}^2 \cdot I\{Z_{kl} \leq 0\},$$

где I - функция-индикатор. Данная мера отражает самый структурированный сегмент, покрывающий нуклеотидную позицию, таким образом она является обобщением профилей построенных на фиксированном окне.

Другая мера, $\rho_w(i)$, отражает взвешенную плотность локально-оптимальных сегментов рядом с i-ым нуклеотидом:

$$\rho_w(i) = \frac{1}{w} \sum_{k,l} Z_{kl}^2 \cdot I\{S_{kl} - \text{ЛОС}\} \cdot I\left\{i - w \leq \frac{k+l}{2} \leq i + w\right\},$$

где параметр w - максимально допустимое расстояние между i-ой позицией и центром локально-оптимального сегмента. Эта мера учитывает взаимное расположение локально-оптимальных структурированных сегментов (например, несколько шпилек на небольшом расстоянии). Вычисление этих мер реализовано алгоритмически эффективно в RNASurface. На практике обе меры являются полезными и до некоторой степени дополняют друг друга (Рисунок 2.1.5).

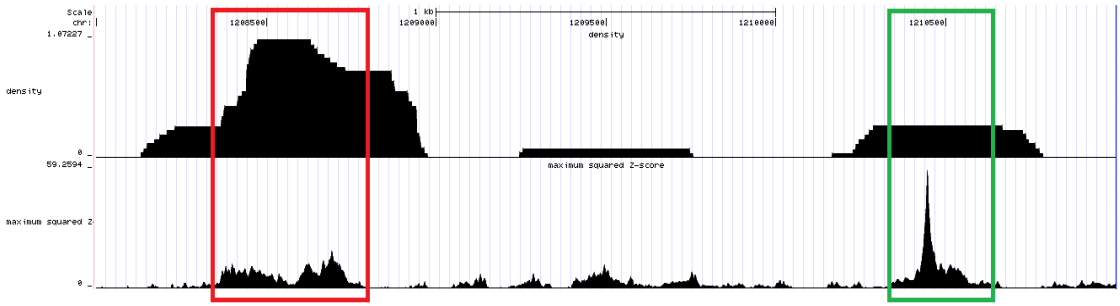


Рисунок 2.1.5. Качественная разница двух профилей. Верхняя панель: профиль ρ_w^i. Нижняя панель: профиль $MZ(i)$. Регион в красном прямоугольнике содержит несколько структурированных сегментов с Z-значением ~ -3, а правый регион содержит один очень структурированный сегмент с Z-значением ~ -7. По отдельности каждая из мер выделяет только один случай. На рисунке представлена последовательность протяженностью 3000 нк из генома *Bacillus subtilis* (координаты: 1208000-1211000).

2.1.4. Общая схема алгоритма и практическая реализация

Рисунок 2.1.1 представляет общую схему алгоритма. Алгоритм сканирует последовательность и вычисляет матрицу минимальных свободных энергий (Рисунок 2.1.1А). После этого вычисляется и сглаживается матрица Z-значений (Рисунок 2.1.1Б-В). На ее основе программа а) вычисляет два вышеописанных профиля структурированности и б) детектирует локально-оптимальные сегменты ниже заданного порога по Z-значениям. Поиск локально-оптимальных

сегментов является быстрой процедурой и осуществляется грубым перебором: для каждой ячейки матрицы Z-значений идет проверка её на оптимум. Также программа имеет возможность визуализировать матрицу Z-значений в виде тепловой карты.

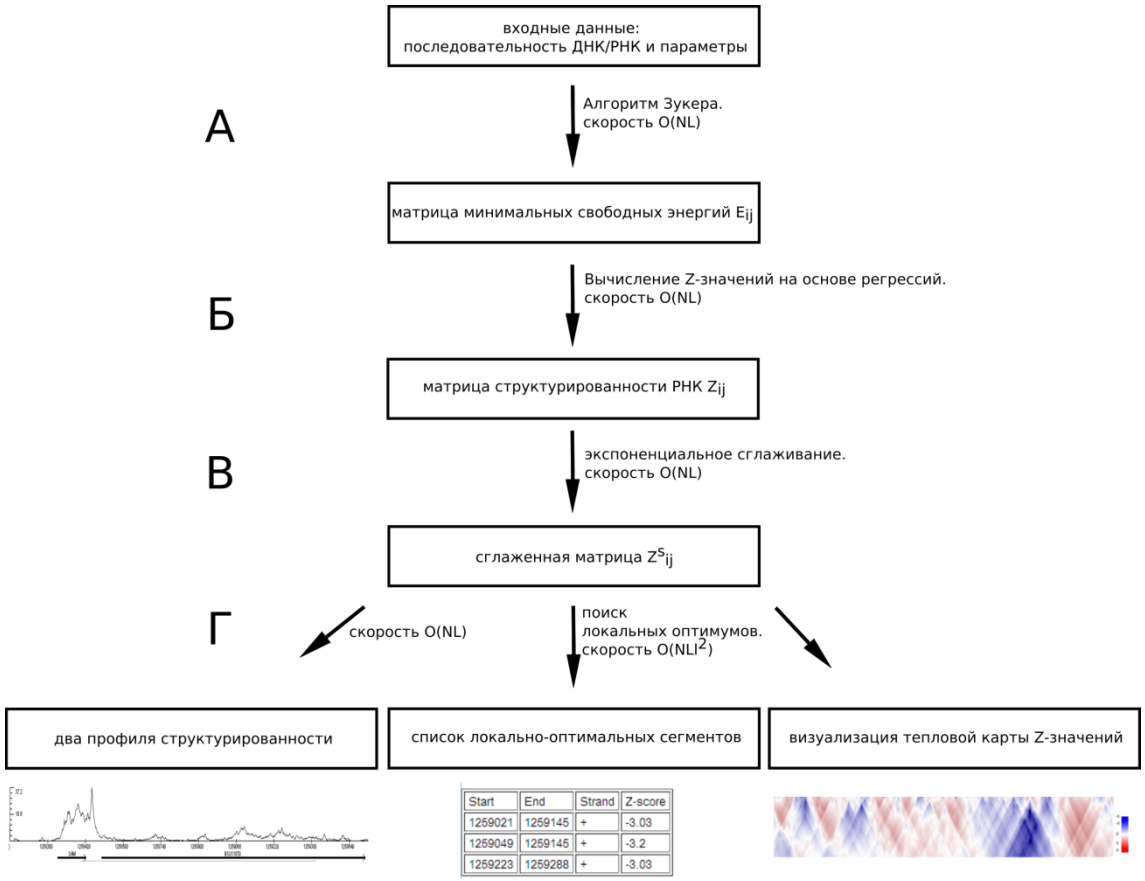


Рисунок 2.1.1. Схема алгоритма. N - длина последовательности, L - размер окна. А) Вычисление энергий алгоритмом Зукера, с использование эффективных ускорений. Б) Вычисление матрицы Z-значений на основе регрессий динуклеотидных частот и длин. В) Экспоненциальное сглаживание матрицы Z-значений. Г) Выдача программа: два одномерных профиля, список локально-оптимальных структурированных сегментов и визуализация матрицы Z-значений в виде тепловой карты.

Программа RNASurface написана на языке C, используя программу RNASlider [80] для реализации алгоритма Зукера со скользящим окном.

Программа RNASlider имеет очень быструю реализацию, так как использует эффективный пересчет энергий и технику разреженных матриц. Ниже перечислены параметры программы RNASurface:

параметр	определение	подробности и значения
-i	Файл с последовательностью ДНК или РНК	Файл в FASTA формате
-d, --diframe	Размер геномного окружения (для оценки динуклеотидного состава)	По умолчанию 500
-w, --winmax	Максимальная длина окна	По умолчанию 200
-m, --winmin	Минимальная длина окна	По умолчанию 30
-s, --smooth	Радиус экспоненциального сглаживания	По умолчанию 7
-l, --locopt	Радиус локального оптимума	По умолчанию 7
--structured	Режим анализа структурированных участков	Установлен по умолчанию
-a, --accessible	Режим анализа участков доступных для взаимодействия	По умолчанию выключен
-o, --output	Имя файл для записи локально-оптимальных сегментов	Файл в BED формате
--mapfile	Имя файл для записи матрицы Z-значений	По умолчанию не создается. По умолчанию матрица Z-значений не записывается в файл
--signfile	Имя файла для записи профиля MZ	По умолчанию не создается. Файл в WIG

		формате
--densfile	Им файла для записи профиля ρ_w	По умолчанию не создается. Файл в WIG формате
--densframe	Окно w для построения профиля ρ_w	По умолчанию 500

2.1.5. Геномные данные

Для анализа качества аппроксимации использовались геномы Homo Sapience (сборка hg19) и Drosophila melanogaster (сборка dm3) взятых с UCSC genome browser database [103].

Полногеномный анализ программой RNASurface был осуществлен на геноме *Bacillus subtilis* subsp. *subtilis* str. 168 взятом с ресурса http://www.microbesonline.org/ [104]. Оттуда также была взята аннотация белок-кодирующих генов.

Структурные РНК, соответствующие *Bacillus subtilis,* были взяты из базы данных Rfam [105]. Из анализа были исключены длинные рРНК и всего было использовано 187 аннотированных структурных РНК, 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. Предсказанные ро-независимые терминаторы были взяты из работы [106].

Для каждого класса РНК Rfam содержит ковариационную модель вторичной структуры [67], построенную по множественному выравниванию. Для получения вторичной структуры для каждой

последовательности РНК, мы вычислили структуру наиболее точно соответствующую ковариационной модели данного класса РНК.

2.2. Результаты и обсуждение

Bacillus subtilis - грам-положительная бактерия с 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. Анализ экспрессии межгенных областей показывает существование локусов малых некодирующих РНК в *Bacillus subtilis* [107], а аккуратный статистический анализ выявил сильную перепредставленность шпилек в геномной последовательности бактерии [108]. Таким образом, *Bacillus subtilis* является удобным и интересным объектом для полногеномного анализа структурных РНК

2.2.1. Качество предсказания RNASurface

Основными программами сканирования длинных последовательностей на основе какого-либо термодинамического сигнала являются RNALfoldz [77], RNASlider [80] и RNALL [75]. RNALL и RNASlider сканируют геном окном и выдают профиль свободных энергий, выбирая затем окна с низкими значениями энергии. Кроме RNASurface, только RNALfoldz сочетает быстрое сканирование и статистическую оценку наблюдений (Таблица 1).

Таблица 1. Свойства программ.

		RNASurface	RNALfoldz	RNAslider	RNALL
Скорость	Быстрый пересчет в окне	+	+	+	+

	Разреженное вычисление	+	-	+	-
Значимость энергий	Оценка для окна	+	+	-	-
	Оценка всех сегментов	+	-	-	-

На основе известных структурных РНК из *Bacillus subtilis* мы сравнили качество предсказания RNASurface с RNALfoldz. RNALfoldz, как и RNASurface, выдает структурированные сегменты с низким Z-значением, однако использует окно фиксированного размера. Нам не известны другие программы, которые определяют статистическую значимость вторичной структуры РНК и работают на масштабе большой последовательности. Практическое сравнение демонстрирует очевидное преимущество RNALfoldz по сравнению с RNALL и RNASlider [77], что согласуется с хорошей способностью отделять структурные РНК от остального транскриптома с помощью Z-значений по сравнению со свободной энергией [65]. Поэтому мы сфокусировались на сравнении с RNALfoldz, что позволит оценить преимущества от использования тепловой карты и локально-оптимальных сегментов. Почти все известные структурные РНК в *Bacillus subtilis* не превосходят 250 нк, поэтому при запуске RNASurface использовалось максимальное окно L=250 нк, и минимальное окно m=50 нк.

Выдача RNASurface и RNALfoldz состоит из списка координат структурированных сегментов. 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 [108], 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 2.2.1). Sensitivity, PPV и FPR определяются следующим образом:

$$PPV = TP/\text{все предсказания},$$

$$FPR = FP/\text{не предсказанные участки},$$

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

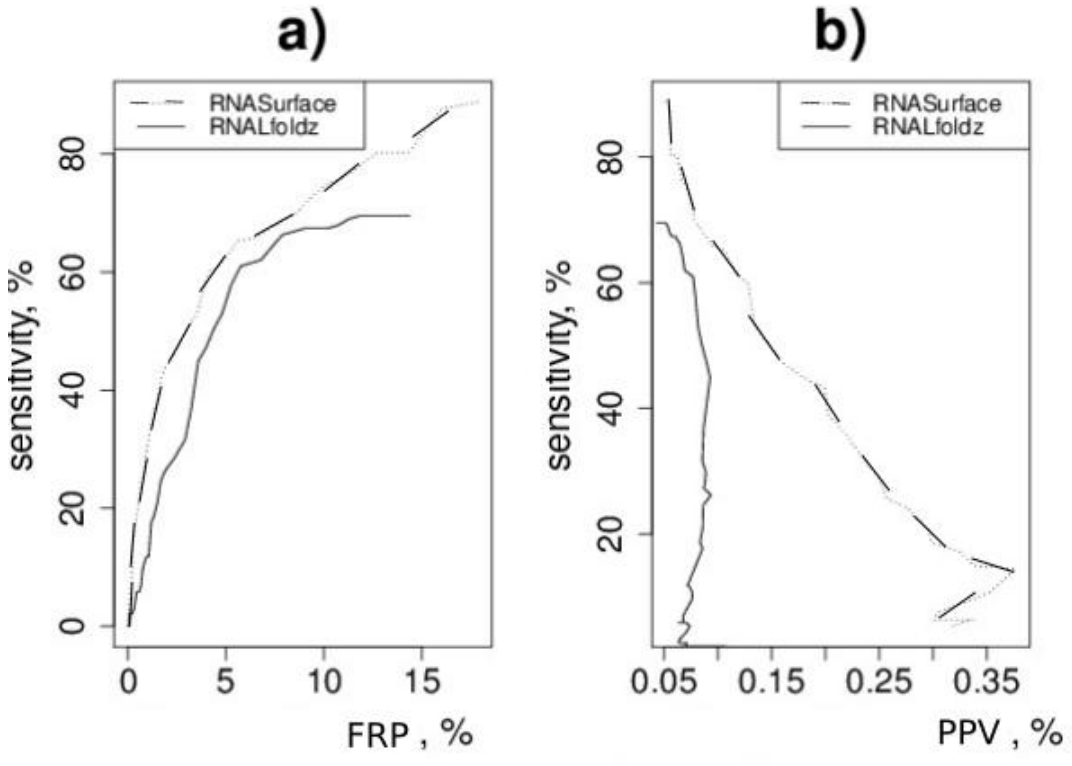


Рисунок 2.2.1. Сравнение качества предсказания структурных РНК программами RNASurface и RNALfoldz. А) ROC кривые сравнения sensitivity и FPR. Б) Кривая сравнения sensitivity и PPV.

RNASurface значительно лучше предсказывает структурные РНК. При одинаковом количестве ложных предсказаний RNASurface детектирует значительно большую долю структурных РНК (Рисунок 2.2.1А), особенно при малом количестве ложных предсказаний; так на уровне FPR=1% RNASurface детектирует $\sim 30\%$ структурных РНК, в то время как RNALfoldz чуть более 15%. Доля структурных РНК среди всех предсказаний также значительно выше у RNASurface (Рисунок 2.2.1Б).

RNASurface и RNALfoldz опирается на размер окна L, мы изучили как изменяется доля предсказанных структурных РНК в зависимости от L (Рисунок 2.2.2). Чувствительность RNASurface монотонно растет с размером

окна, в то время как чувствительность RNALfoldz монотонно растет до 150 нк, после чего падает. Количество предсказаний зависит от порога на Z-значение, что определяет уровень FPR. В данном примере был установлен очень либеральный FPR в 20%. На этом уровне статистической достоверности, 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%. При более строгих порогах на FPR количество предсказанных структурных РНК понижается, но общий вывод сохраняется. При окне больше 150 нк, доля структурных РНК с таким рамером всего 13%, качество предсказания RNALfoldz падает, но умеренно. Это связано со следующей особенностью этой программы: чтобы уменьшить количество вычислений Z-значений, авторы рассматривают только оптимальную закрытую подструктуру наибольшего размера в данном окне (её можно эффективно найти), а не само окно. С одной стороны, это позволяет RNALfoldz, также как RNASurface, детектировать сегменты различного размера и является плюсом подхода. С другой стороны, вероятность оптимальной структуры и спаривания крайних нуклеотидов убывает полиноминально с расстоянием между крайними нуклеотидами [78], [109]; при увеличении окна оптимальные закрытые структуры имеют хоть и максимальную, но крайне низкую вероятность, а поэтому предсказываются крайне неустойчиво.

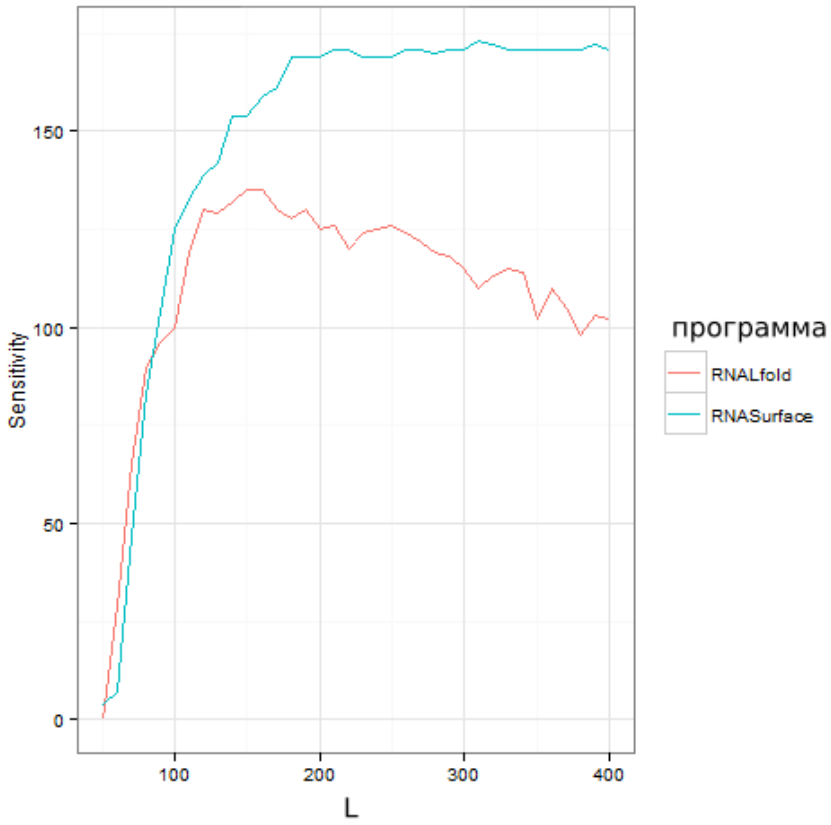


Рисунок 2.2.2. Зависимость чувствительности методов от размера окна. FPR зафиксирован на уровне 20%.

RNASurface базируется на RNASlider, поэтому для сравнения мы также приводим зависимость чувствительности предсказания RNASlider от окна на том же уровне FPR (Рисунок 2.2.3). Качество детектирования структурированных РНК значительно ниже чем у программ, использующих Z-значение, и чувствительность метода является приемлемой лишь для узкого диапазона окон $\sim 60 - 100$ нуклеотидов.

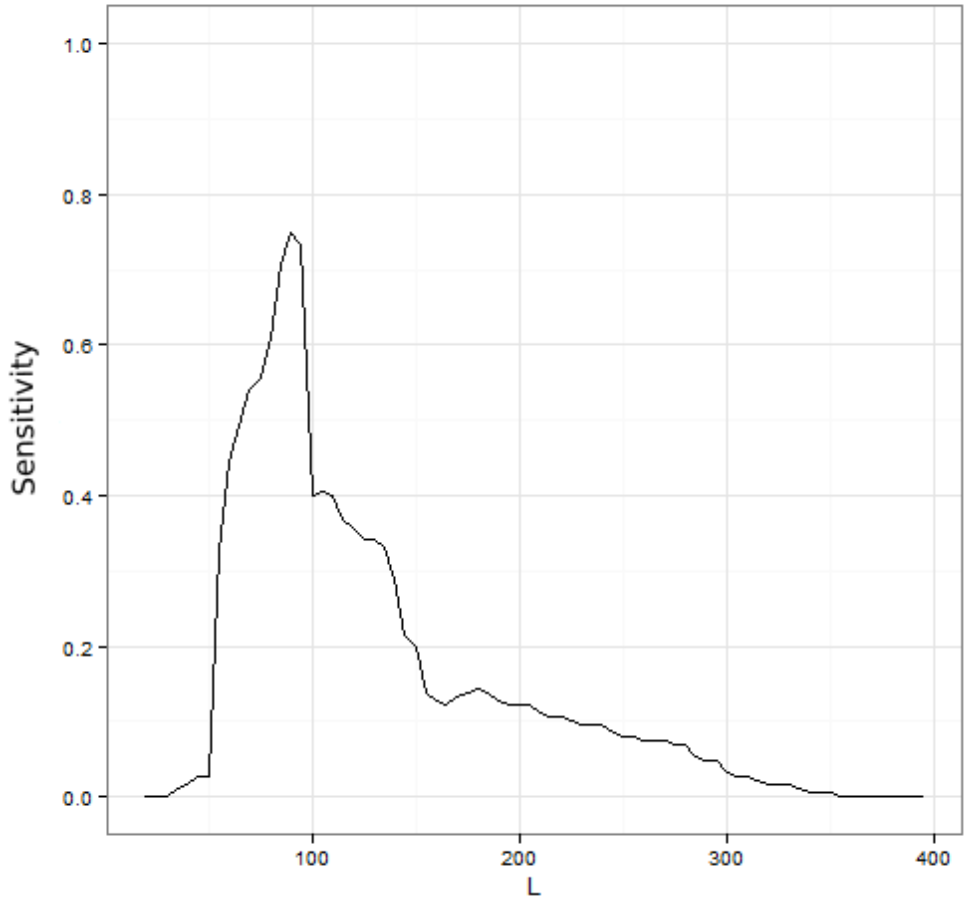


Рисунок 2.2.3. Чувствительность предсказания структурных РНК в зависимости от окна для программы RNASlider на уровне FPR=20%.

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. Ро-независимые терминаторы, образующие небольшие шпильки РНК, были выбраны как пример простой структуры (Рисунок 2.2.4А). В геноме *Bacillus subtilis* предсказано более 2000 ро-независимых терминаторов с аккуратностью 94% [106]. Рибо-переключатели и Т-бокс структуры были выбраны как

пример сложных структур (Рисунок 2.2.4Б). RNASurface превосходит RNALfoldz как на простых, так и на сложных вторичных структурах (Рисунок 2.2.4). 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.

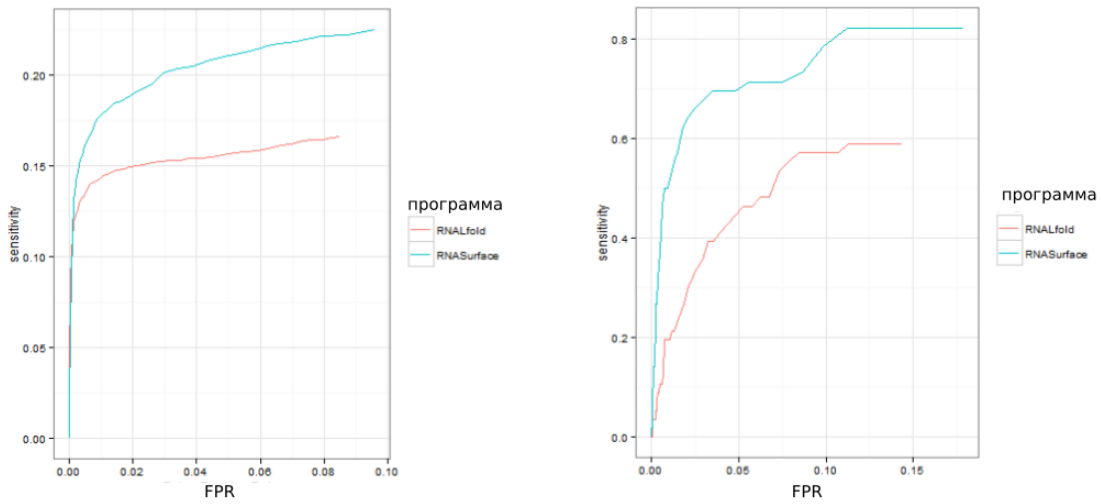


Рисунок 2.2.4. ROC кривые отдельно для А) ро-независимых терминаторов и Б) рибо-переключателей и T-box РНК.

Обобщая это наблюдение, определим сложность вторичной структуры через количество составляющих её стеблей. Для каждой структурной РНК вторичная структура определяется по ковариационной модели с использованием программы Infernal, а количество стеблей простым разбором элементов вторичной структуры. Качество предсказания сильно растет с ростом сложности структуры (Рисунок 2.2.5). Причем

лучшая чувствительность RNASurface особенно проявляется на сложных структурах.

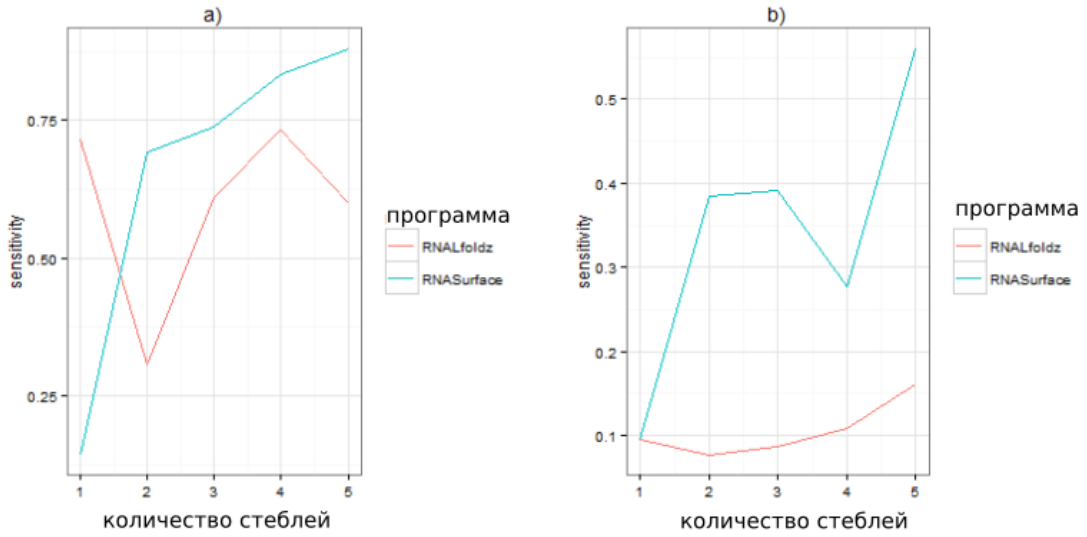


Рисунок 2.2.5. Сложность структуры значительно влияет на качество предсказания RNASurface и RNALfoldz. Ось X: количество шпилек во вторичной структуры. Ось Y: чувствительность предсказания для каждого класса сложности структуры. Результаты приведены для двух порог на FPR: А) FPR=10% Б) FPR=1%

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.

Таблица 2. Доля предсказанных РНК различных типов для трех порогов по Z-значению

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

					(81)			
-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		

Количество и доля предказанных структурных РНК, стратифицированных по классам, для разных порогов по Z-значению

Разные классы структурных РНК имеют как разную сложность, так и разные требования на термодинамическую стабильность вторичной структуры. Если в среднем предсказывается 30% структурных РНК на уровне FPR=1%, то для отдельных классов эта цифра варьируется от 15% до 69%. 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, T-box, 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.

2.2.2. Распределение по геномным регионам

Регуляторные структуры РНК располагаются в специфических регионах геномы. Так, ро-независимые терминаторы находятся в 3'НТО, а вторичные структуры, регулирующие экспрессию (рибо-переключатели, Т-боксы и т. д.), как правило в 5'НТО. Предсказанные локально-оптимальные сегменты были классифицированы по типам геномных регионов (Рисунок 2.2.6). 5'НТО определен как регион от -50 до +200 нк относительно старт

кодона, 3'НТО как регион от -50 до +150 нк относительно стоп кодона. Межкодирующие регионы - регионы между кодирующими областями.

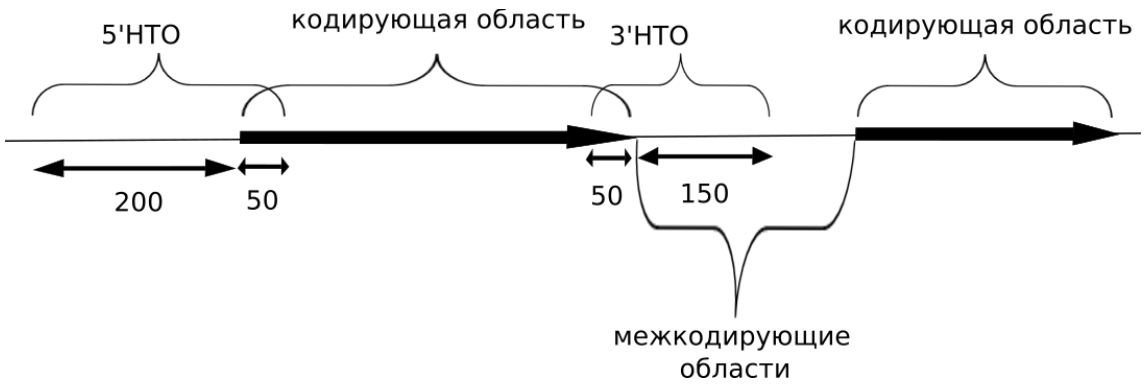


Рисунок 2.2.6. Разбиение на типы геномных регионов.

Мы оценили перепредставленность структурированных сегментов в каждом из типов регионов. В качестве нулевой модели предполагаем, что структурированные сегменты распределены равномерно вдоль генома. Перепредставленность региона определялась как отношение реального и ожидаемого количества предсказанных сегментов этого региона. Распределение структурированных сегментов сильно неравномерно, с большой перепредставленностью в 5'НТО и 3'НТО генов. Эта тенденция проявляется значительно сильнее при более низких порогах на Z-значение. Так, при пороге -2 на Z-значение наблюдается качественная перепредставленность только в 3'НТО генов. 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 3'НТО и 5'НТО областями генов, но даже за вычетом их остается четырехкратная перепредставленность в межгенных областях. Это может быть вызвано сигналом от малых некодирующих РНК, которые играют более весомую роль в бактериальном метаболизме чем считалось ранее [110], [107].

Таблица 3. Обогащение структурированных сегментов в различных областях генома *Bacillus subtilis*.

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)

Для нескольких порогов на Z-значением 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).

Кодирующие области имеют значительно меньшую плотность структурированных РНК по сравнению с 5'НТО и 3'НТО. Тем не менее, ранее показано что давление отбора на структурные свойства РНК существует и в кодирующих областях [111]. Однако кодирующие области имеют давление отбора на аминокислотную последовательность и три-периодичность нуклеотидного состава, что значительно осложняет построение нулевой модели и выделение структурированных РНК [112], [113], и является предметом отдельного исследования.

2.2.3. Время и память требуемые на выполнение RNASurface

Теоретически, среднее время работы RNASurface составляет $O(N \cdot L)$ на последовательности длины N с окном L. Мы сравнили практическое время работы RNASurface, RNASlider и RNALfoldz на геноме *Bacillus subtilis* как функция от длины окна. Время выполнения RNASurface в несколько раз меньше чем RNALfoldz. При этом время работы сравнимо с RNASlider, а значит процедуры вычисления тепловой карты и поиск локально-оптимальных сегментов на практике осуществляются за небольшое время по сравнению с вычислением свободной энергии. Дополнительные параметры фиксировались на следующем уровне: геномное окружение $d = 600$, ядро сглаживания $s = 1$, радиус поиска локально оптимального сегмента $l = 7$. Уменьшение $d = 0$ и увеличение $s = 10$ могут изменять практическое время в несколько раз, параметр l практически не влияет на время выполнения. RNASurface требудет $O(L^2)$ памяти.

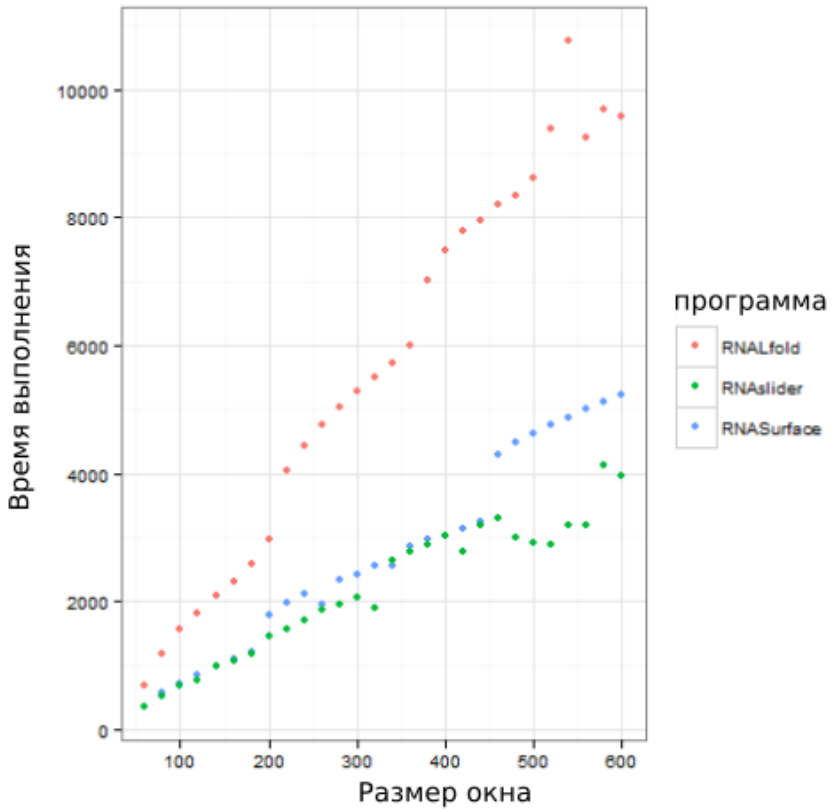


Рисунок 2.2.7. Сравнение времени выполнения программ на геноме *Bacillus subtilis* как функция от длины окна. Вычисления проводились на процессоре Intel Xeon Processor E5506.

Глава 3. Сравнительно-геномный метод предсказания структурных РНК на основе диффузионной модели

Особый интерес представляют функциональные структурные элементы РНК. Косвенным признаком функциональности является факт сохранения структурированности РНК в ходе эволюции. На практике, как правило, известно филогенетическое дерево с ортологичными последовательностями на листьях, и сравнение Z-значений последовательностей дает возможность выявить давление отбора на структурные свойства РНК (Рисунок 2.2.**1**). 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, в данной работе эволюция Z-значений представлена в виде диффузионного процесса. Диффузионный процесс является классической моделью для анализа эволюции частот аллелей в популяции и количественных характеристик видов (например, размер зубов). Мы адаптировали диффузионный процесс к анализу эволюции количественных характеристик, неявно зависящих от последовательности (например, белок-кодирующий потенциал РНК или GC состав последовательности), и применили его для предсказания структурных РНК на основе Z-значений. В дальнейшем изложение метода проводится для произвольной количественной характеристики последовательности.

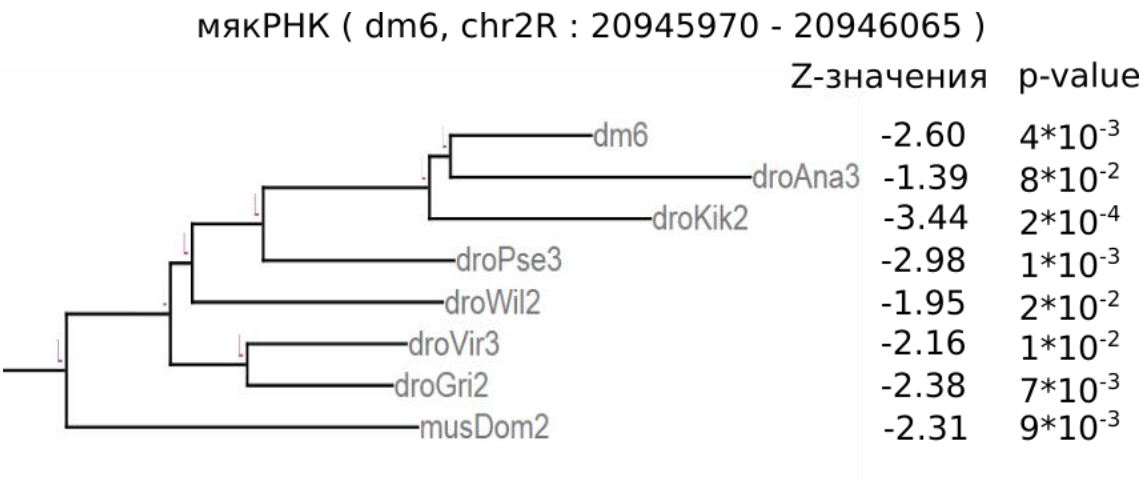


Рисунок 2.2.1. Демонстрация задачи на примере ортологичных последовательностей малой ядрышковой РНК (мякРНК) в геномах дрозофил. Для каждой последовательности вычислено Z-значение. Большое отрицательное Z-значение в каждом виде является косвенным свидетельством отбора на структурный элемент последовательности. Необходимо оценить статистическую значимость набора наблюдаемых Z-значений в последовательностях, связанных данной эволюционной историей.

3.1. Метод

3.1.1. Модель эволюции

Рассмотрим количественную характеристику, которая вычисляется по последовательности

$$x: \Omega \to \mathbb{R},$$

где Ω - пространство последовательностей. а $\mathbb{R}$ - вещественные числа. Например, это может быть GC состав, кодирующий потенциал РНК, или Z-значение структурированности РНК. Предположим, что функция является "непрерывной": малые изменения в последовательности приводят к малым изменениям в значениях функции. Эволюцию последовательности можно представить как случайные блуждания в пространстве Ω. В таком случае, функция $x(s)$ будет описывать случайные блуждания в $\mathbb{R}$ (Рисунок

3.1.1А). В случае броуновского движения, изменения $x(s)$ в большую и меньшую сторону равновероятны. В общем случае, признак имеет некоторое распределение значений $p(x)$, полученное, например, случайной нарезкой из генома или симуляцией в рамках заданной модели эволюции последовательности. Тогда изменение $x(s)$ вероятней будет происходить в сторону значений с большей плотностью последовательностей (Рисунок 3.1.1Б). Так, если последовательность имеет высокий GC состав, то последующие изменения вероятней его уменьшат. Качественно, случайное изменение $x(s)$ в "силовом поле" описывается диффузионным процессом:

$$dx = a(x)dt + b(x)dB_t, \quad (11)$$

где $a(x)$ и $b(x)$ - функции сноса и диффузии соответственно, а $B_t \sim N(0,t)$. Считаем, что эволюционные силы не меняются со временем, поэтому коэффициенты не зависят от времени. В дальнейшем считаем, что $p(x) \sim N(0,1)$. В противном случае, осуществляем преобразование количественной характеристики

$$\Phi^{-1}\big(F(x)\big),$$

которое переводит её стационарное распределение к стандартному нормальному, где $\Phi(.)$ - кумулятивная функция стандартного нормального распределения, $F(.)$ - кумулятивная функция количественной характеристики x.

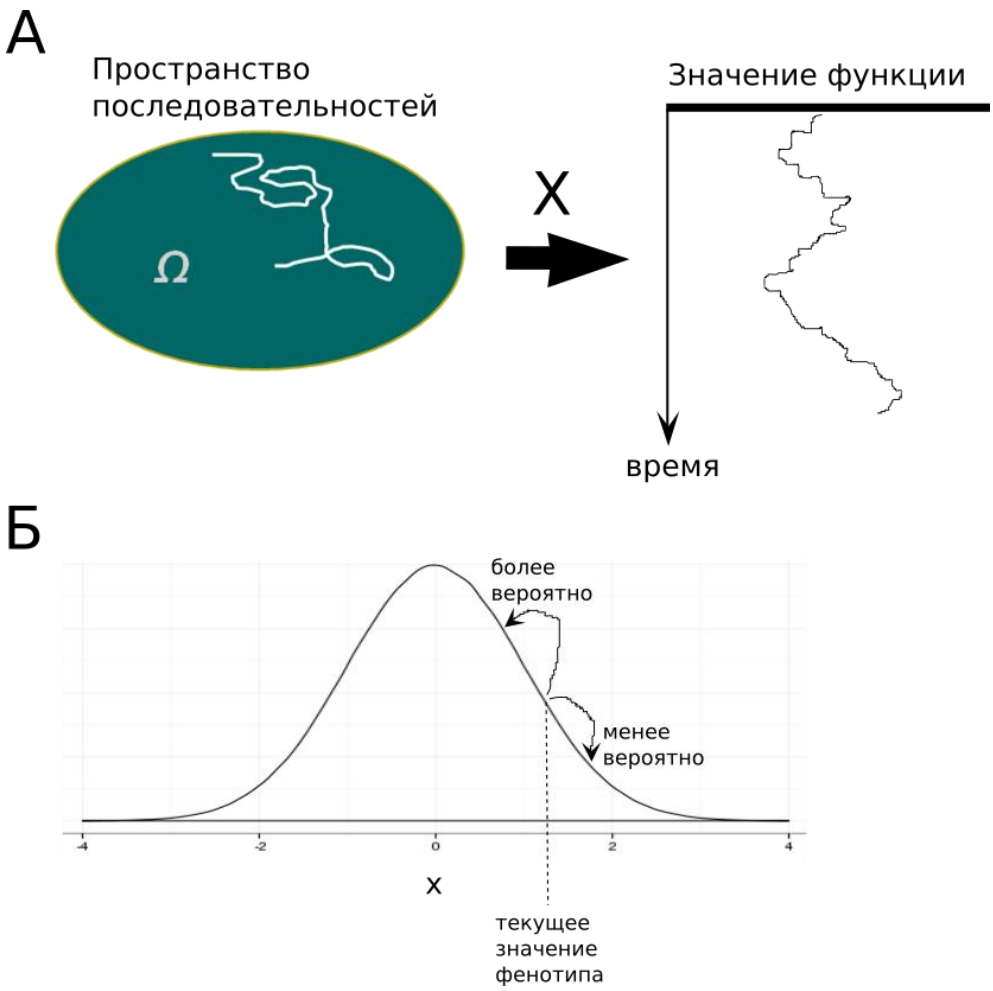


Рисунок 3.1.1. Модель эволюции количественной характеристики. А) Эволюция последовательности представляется в виде случайного блуждания, которое порождает случайные блуждание количественной характеристики. Б) Количественная характеристика имеет некоторое стационарное распределение: разные значения имеют разную частоту встречаемости среди последовательностей. Изменение текущего значения в меньшую и большую сторону зависит от стационарного распределения.

Пусть диффузионное уравнение имеет некоторое решение $\rho(x,t|y)$ с начальным условием $x(0)=y$: вероятность наблюдать значение x через время t, запустив процесс со значения y. Вид решения $\rho(x,t|y)$ зависит от $a(x)$ и $b(x)$. В то время как в предыдущих подходах эти параметры выбирались исходя из теоретических соображений, в данном методе

зависимость $x(s)$ от последовательности позволяет вычислить $a(x)$ и $b(x)$ на основе модели эволюции последовательности. Мы разработали набор инструментов для оценки этих параметров по нарезке ортологичных участков генома, или исходя из модели эволюции последовательности.

3.1.2. Оценка параметров диффузионного процесса

Если значение признака в начальный момент равно y, то через малый промежуток времени Δt возможные значения признака описываются распределением $\rho_\Delta(x|y)$. При этом среднее и дисперсия этого распределения приближенно определяются коэффициентами сноса и диффузии в точке y:

$$M\rho_\Delta(x|y) \approx y + a(y) \cdot \Delta t, \quad (12)$$

$$D\rho_\Delta(x|y) \approx b^2(y) \cdot \Delta t. \quad (13)$$

Приближенные равенства заменяется точным при предельном переходе $\Delta t \to 0$. В случае эволюции последовательности, время описывает процесс нуклеотидных замен и характеризуется долей замен в последовательности

$$t = \frac{m}{L},$$

где L - длина последовательности, а m - количество замен. В таком случае под малым временем будет подразумевать малую долю замен в последовательности. Из соотношений **Ошибка! Источник ссылки не айден.**-**Ошибка! Источник ссылки не найден.** следует, что отношение

$$g(y) = \frac{a(y)}{b^2(y)} = \frac{M\rho_\Delta(x|y) - y}{D\rho_\Delta(x|y)} \quad (14)$$

не зависит от единиц измерения Δt. Для вычисления $g(x)$ необходимо иметь набор векторов $(x, y, \Delta t)$, где x и y соответствуют значениям количественной характеристики последовательностей, а Δt - доле различий между ними. Данный набор может быть получен либо нарезкой ортологичных участков геномных выравниваний, либо симулирован в рамках заданной модели эволюции последовательности. По набору векторов, для каждого значения y оценивается среднее и дисперсия распределения $\rho_{\Delta}(x|y)$, после чего вычисляется функция $g(y)$ по формуле (14).

Для получения параметров $a(x)$ и $b(x)$ воспользуемся еще одним соотношением: распределение $p(x)$ является стационарным (то есть не зависящим от времени) решением диффузионного уравнения, поэтому

$$-a(x)p(x) + \frac{1}{2}\frac{d(b^2(x)p(x))}{dx} = 0, \quad (15)$$

Таким образом, учитывая что $p(x) \sim N(0,1)$, соотношение **Ошибка! сточник ссылки не найден.** представимо в виде:

$$(b^2(x))' - xb^2(x) = 2a(x),$$

отсюда $b^2(x)$ вычисляется на основе $g(x)$:

$$b^2(x) = c \cdot \exp\left(\int_0^x (2g(\xi) + \xi)d\xi\right),$$

а параметр $a(x)$ получается автоматически из $a(x) = g(x) \cdot b^2(x)$.

3.1.3. Расширение модели на филогенетическое дерево

На основе известных параметров $a(x)$ и $b(x)$ вычисляется решение диффузионного уравнения $\rho(x, t|y)$. Если дано филогенетическое дерево с наблюдаемыми значениями $x_1, \dots, x_n$ на листьях, то можно оценить их

вероятность $\rho(x_1, \dots, x_n)$ на основе техники марковских процессов на филогенетическом дереве. Рассмотрим простой сценарий, при котором ортологичные последовательности эволюционируют от общего предка независимо (Рисунок 3.1.2). В таком случае вероятность наблюдать значения их признаков $x_1, \dots, x_n$ на листьях дерева с предковым значением y определяется как:

$$\rho(x_1, \dots, x_n|y) = \prod_{i=1}^{n} \rho(x_i|y),$$

где вероятности $\rho(x_i|y)$ вычисляются на основании диффузионного уравнения. Подробный подход к приближенному вычислению этих вероятностей будет представлен ниже. Предковое значение признака неизвестно, однако оно имеет стационарное распределение $p(y)$, поэтому вероятности наблюденных признаков $x_1, \dots, x_n$ вычисляется как:

$$\rho(x_1, \dots, x_n) = \int_{-\infty}^{+\infty} \rho(x_1, \dots, x_n|y)p(y)dy.$$

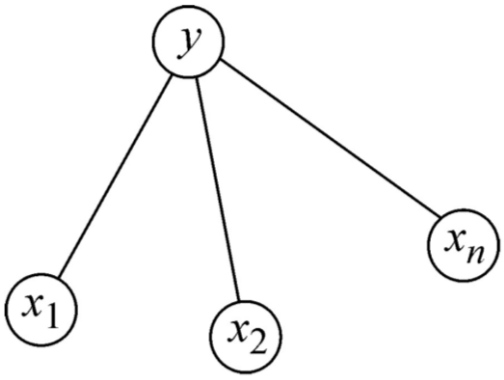


Рисунок 3.1.2. Независимая эволюция ортологичных последовательностей от общего предка.

В общем случае последовательности связаны более сложной эволюционной истории. В случае простейшего ветвящегося дерева (Рисунок 3.1.3) вероятности признаков при предковом состоянии y вычисляются как:

$$\rho(x_1, x_2, x_3|y) = \rho(x_1, x_2|y)\rho(x_3|y) =$$

$$= \rho(x_3|y) \int_{-\infty}^{+\infty} \rho(x_1, x_2|v)\, \rho(v|y)dv,$$

а общая вероятность признаков x_1, x_2, x_3:

$$\rho(x_1, x_2, x_3) = \int_{-\infty}^{+\infty} \rho(x_1|v)\rho(x_2|v)\, \rho(v|y)\rho(x_3|y)p(y)dvdy.$$

Аналогично, для каждого дерева вычисляется вероятности значений признаков. Вычисление формулы может осуществляться и реализовано в виде рекурсивного алгоритма на дереве.

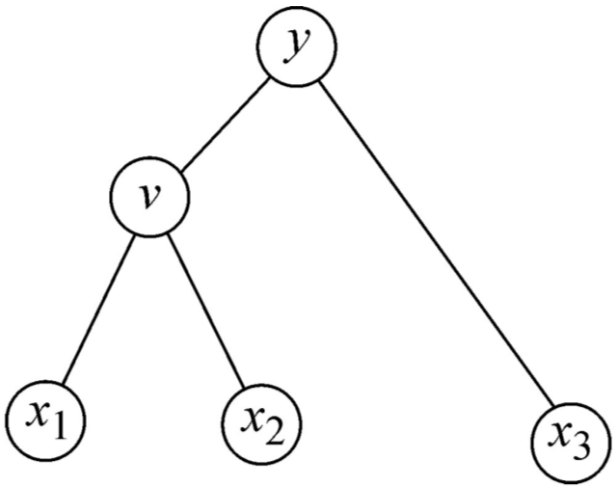


Рисунок 3.1.3. Сценарий зависимой эволюции ортологичных последовательностей.

Как показывают примеры практических количественных характеристик (раздел Результаты), их эволюция хорошо описывается, аналитически вычислимым, процессом Орнштейна-Уленбека

$$dx = -xdt + bdB_t.$$

В этом случае $\rho(x, t|y)$ имеет нормальное распределение, а вероятности значений признака на филогенетическом дереве имеют многомерное нормальное распределение:

$$\rho(x_1, \dots, x_n) = \frac{\sqrt{detA}}{2\pi^{n/2}} e^{-\frac{1}{2}x'Ax},$$

где матрица ковариаций признаков A вычисляется рекурсивно на основе структурны филогенетического дерева.

3.1.4. Статистический анализ на основе модели

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

$$r^2 = x'Ax.$$

монотонно изменяется с вероятностью $\rho(x_1, \dots, x_n)$. В случае, если признак эволюционирует согласно процессу Орнштейна-Уленбека, и соответственно $\rho(x_1, \dots, x_n)$ - многомерное нормальное распределение, то статистика r^2 имеет распределение хи-квадрат с $n-1$ степенью свободы (χ^2_{n-1}). Тогда возможно аналитическое вычисление p-value наблюдений, иначе для оценки значимости r^2 необходимо проводить симуляции для построения распределения статистики с последующей оценкой p-value.

Априорное распределение предкового состояния y соответствует стационарному распределению $p(y)$. Если известны наблюдения на

листьях дерева, то апостериорное распределения значения признака предкового состояния вычисляется по формуле Байеса:

$$p(y|x_1, \dots, x_n) = \frac{\rho(x_1, \dots, x_n|y)p(y)}{\rho(x_1, \dots, x_n)}.$$

Если признак эволюционирует согласно процессу Орнштейна-Уленбека, то $p(y|x_1, \dots, x_n)$ имеет нормальное распределение с параметрами (μ, σ), которые зависят от структуры дерева, длин веток и значений наблюдений. Например, μ можно интерпретировать как влияние "сноса" наблюдаемых признаков на предковую последовательность и вычислить 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.

Статистика r^2 оценивает "необычность" наблюдений. Однако она может показать значимое значение p-value из-за ошибок в построении дерева, проблем в установлении ортологичности последовательностей или неточной модели эволюции последовательности. Статистика μ, наоборот, показывает значимое p-value только в случае согласованного смещения наблюдаемых значений, что отражает функциональную роль последовательностей. Таким образом, предпочтительная схема использования статистика имеет следующий вид:

1) Проверяем, что $r^2(x_1, \dots, x_n) \geq r_0^2$, где r_0^2 - заданный квантиль значимость, например, 5%-квантиль распределения χ^2_{n-1}. Если статистика проходит барьер, то

2) проверяем, что $\mu^2(x_1, \dots, x_n) \geq \mu_0^2$, где μ_0^2 - заданный квантиль распределения статистики μ^2. Если статистика проходит барьер, то нулевая гипотеза о нейтральной эволюции признака отвергается и дается предсказание о функциональности признака. Иначе требуется более глубокий анализ расхождения в результатах статистик $r^2(x_1, \dots, x_n)$ и $\mu^2(x_1, \dots, x_n)$.

Чтобы усилить статистическую мощность, используется комбинация выше введенных статистик:

$$log\ (1 - F(r)\ ) \cdot (1 - F(\mu)\ ),$$

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

3.1.5. Реализация метода

Таким образом, в данной работе разработан метод анализа количественных характеристик, 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.

Описанные методы реализованы в виде библиотеки Java-классов. Реализованные модели эволюции последовательности позволяют симулировать марковский процесс нуклеотидных замен и поддерживать распределение других свойства последовательностей (например, неоднородной консервативности вдоль последовательности) с помощью алгоритма Метрополиса-Гастингса.

3.2. Результаты

3.2.1. Анализ модели на примере функции частот встречаемости нуклеотидов

Диффузионный метод был протестирован на функции $n(s)$ отклонения нуклеотидного состава от равномерного:

$$n(s) = \sqrt{(\pi_a - 0.25)^2 + (\pi_t - 0.25)^2 + (\pi_g - 0.25)^2 + (\pi_c - 0.25)^2},$$

где $\pi_a, \pi_t, \pi_g, \pi_c$ - частоты нуклеотидов a, t, g, c в последовательности s.

Оценка параметров диффузионного процесса этой характеристики проводилась в рамках модели эволюции последовательности с равными вероятностями замен нуклеотидов друг в друга. В таком случае, в стационарном состоянии нуклеотиды в последовательности имеют равную частоту $(0.25, 0.25, 0.25, 0.25)$. Длины последовательностей рассматривались в промежутке от 100 до 300 нуклеотидов. Анализ параметров $a(n)$ и $b(n)$ показывает, что эволюция этой количественной характеристики имеет качественно близкий вид к процессу Орнштейна-Уленбека (Рисунок 3.2.1А-Б): на промежутке [-2,2], за пределы которого случайное блуждания выходит редко, $a(n)$ ведет себя линейно, а вариация $b(n)$ мала и не превосходит 40%.

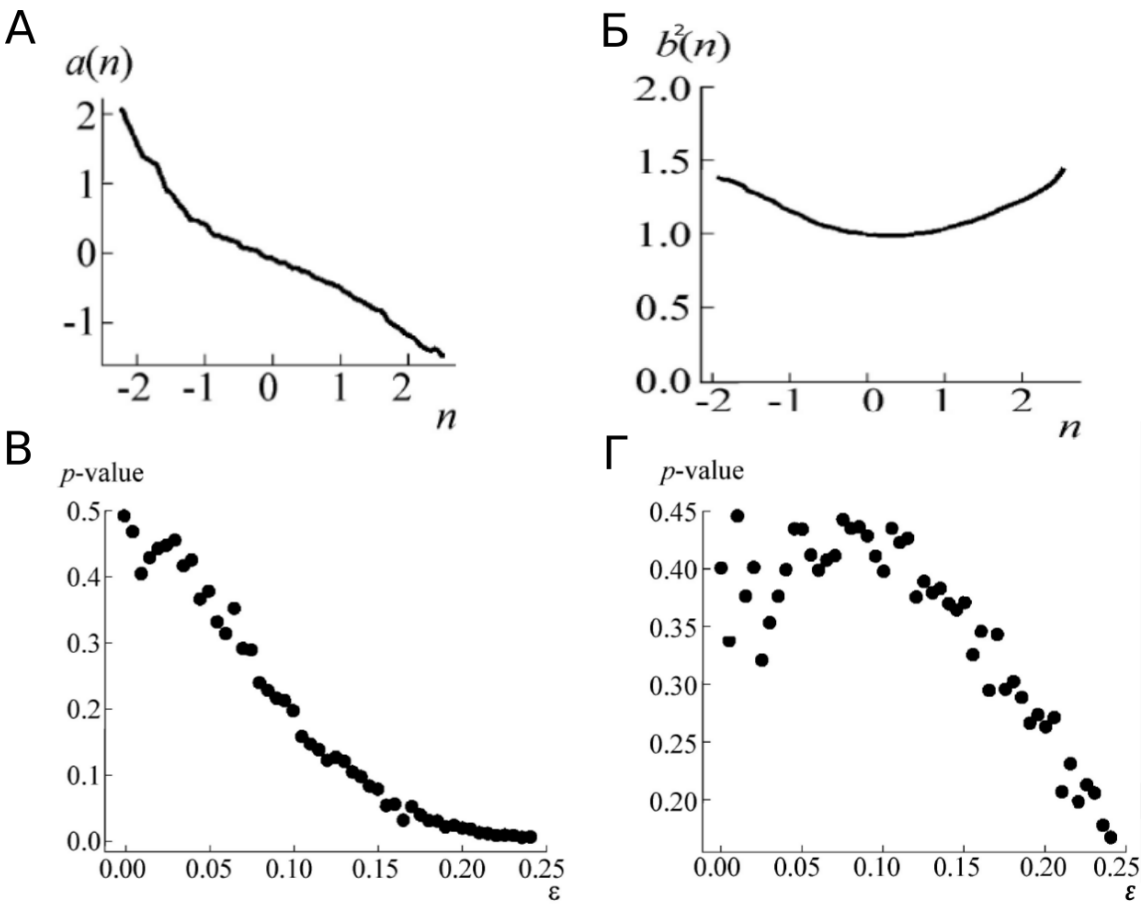


Рисунок 3.2.1. Диффузионная модель эволюции отклонения частот нуклеотидов от равномерных. Параметры А) сноса $a(n)$ и Б) диффузии $b(n)$. P-value значений статистик при симуляции смещенных на ε нуклеотидных частот для В) μ и Г) r^2.

Разработанный метод фокусируется на оценке статистической значимости наблюдаемых значений. Смещенные значения количественной характеристики на листьях филогенетического дерева свидетельствуют о давлении отбора на неё и выражаются в виде статистически значимых значениях статистик. Так, если изучаемые ортологичные последовательности имеют давление отбор на нуклеотидный состав (например, отбор на высокий GC состав), то их стационарный состав будет систематически отличаться от равномерного. Для оценки работы статистик

был проведен следующий компьютерный эксперимент: диффузионная модель строилась в предположении, что нейтральный нуклеотидный состав имеет равномерные частоты встречаемости нуклеотидов, а филогенетическое дерево с последовательностями на листьях симулировалось в рамках смещенных частот; после чего оценивалась значимость наблюдаемых статистик в зависимости от уровня смещения.

Более детально, в ходе анализа предполагалось, что матрица вероятностей замен нуклеотидов, $Q(\pi_a, \pi_t, \pi_g, \pi_c)$ со стационарными частотами нуклеотидов $(\pi_a, \pi_t, \pi_g, \pi_c)$, имеет равные стационарные частоты $\pi_a = \pi_t = \pi_g = \pi_c = 0.25$. Смещение частот симулировалось в рамках модели эволюции $Q(0.25 + \varepsilon, 0.25 + \varepsilon, 0.25 - \varepsilon, 0.25 - \varepsilon)$, где ε выражает уровень смещения стационарные частоты. Чем выше смещение ε, тем больше ожидаемое значение $n(s)$, а поэтому увеличивается статистическая достоверность эффекта в виде p-value. Эволюция последовательностей симулировалась на коротком промежутке времени на видовом дереве дрозофил в рамках моделей эволюции последовательности со смещением ε от 0 до 0.24 с шагом 0.01. Результаты показывают, что чем больше смещение ε, тем ниже p-value статистик (Рисунок 3.2.1), согласуясь с практическими ожиданиями от диффузионной модели.

3.2.2. 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) на основе их Z-значений на листьях. Оценка параметров $a(z)$ и $b(z)$ диффузионного процесса была получена на основе нарезки ортологичных последовательностей трех геномов дрозофил (Рисунок 3.2.**2Рисунок 2.2.1**), и показывает, что параметры хорошо приближаются линейной и постоянной функциями соответственно. Для того, чтобы оценить преимущества сравнительно-геномного анализа, мы также оценили качество предсказания нкРНК по их Z-значениям.

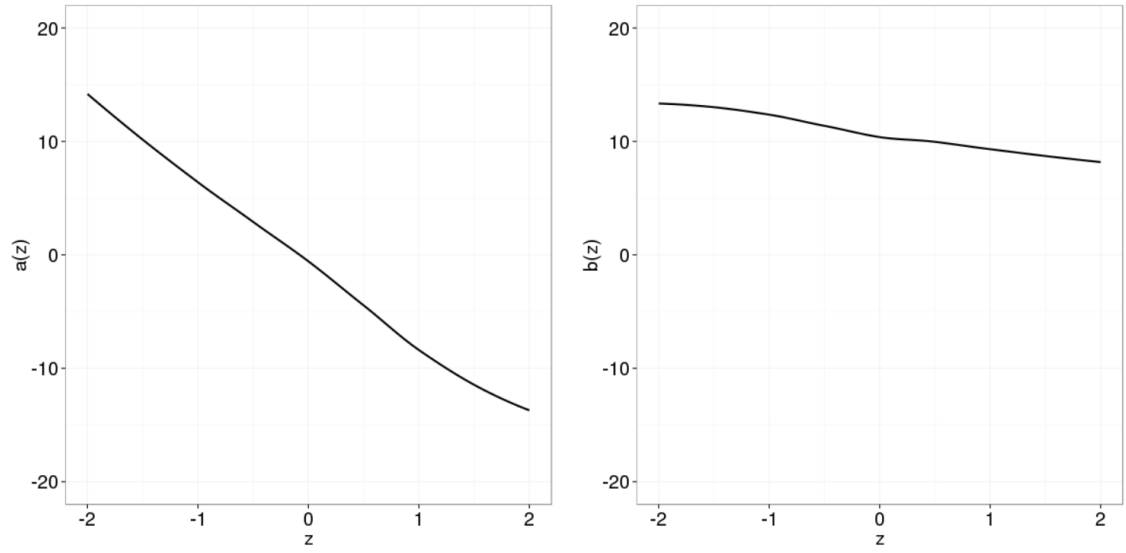


Рисунок 3.2.2. Вид параметров a(z) и b(z), вычисленный на основе нарезки ортологичных последовательностей геномов Drosophila melanogaster, yakuba и ananassae. Вид функций не меняется при симуляции в рамках заданных моделей эволюции последовательностей.

Для анализа были выбраны четыре широких класса нкРНК: микроРНК, малые ядерные РНК (мяРНК), малые ядрышковые РНК (мякРНК) и тРНК (Таблица 4). 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 4. 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 3.2.**3**А, Таблица 4). Отметим, что если предсказания практически не улучшаются для тРНК и микроРНК, то мы наблюдаем значительное улучшение качество предсказания мякРНК (Рисунок 3.2.**3**). Мощность сравнительно-геномного анализа зависит от степени расхождения последовательностей, при этом доля замен в мякРНК значительно превышает доли в микроРНК и тРНК (Таблица 4, 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 3.2.**з**, вставки).

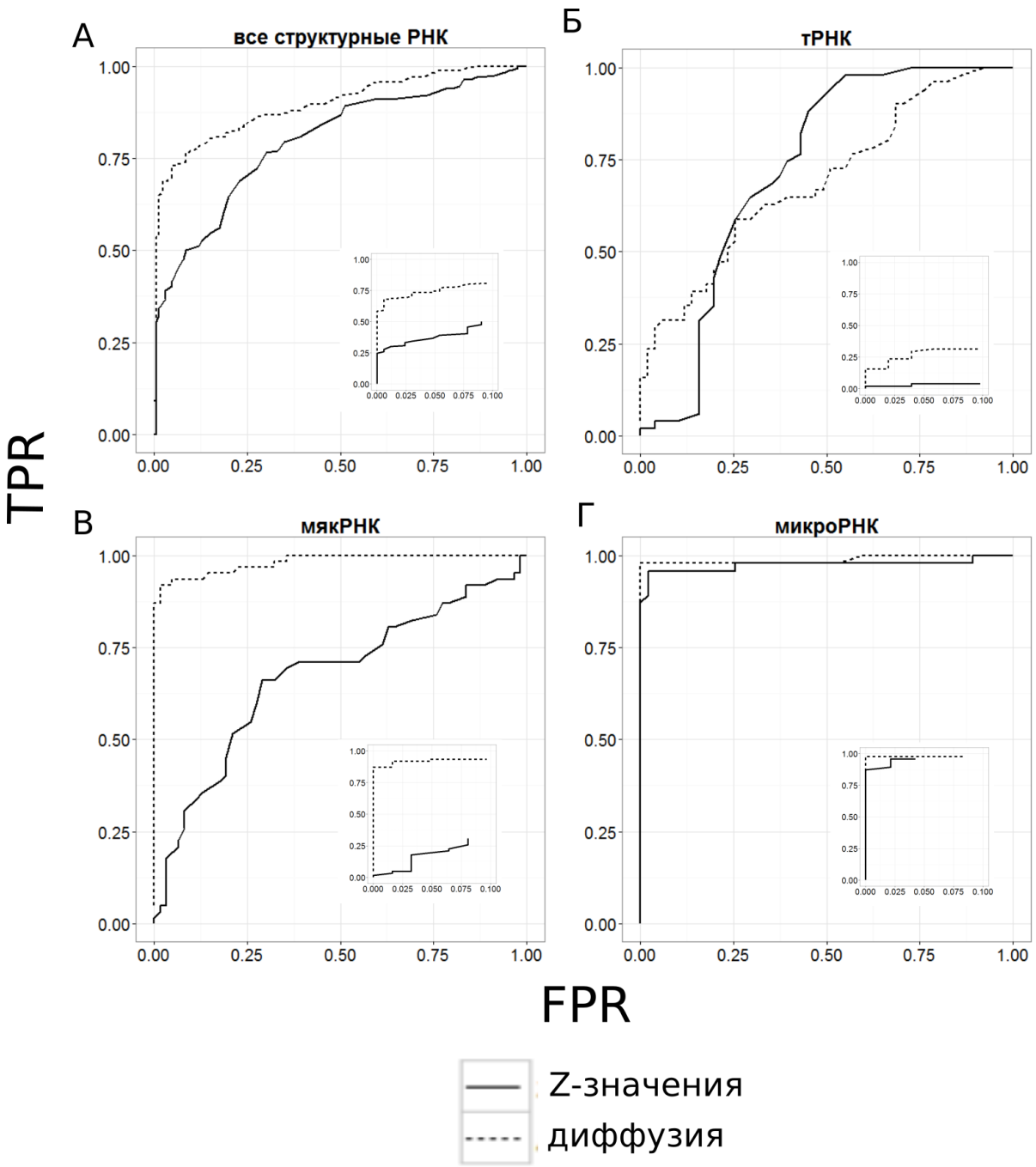


Рисунок 3.2.3. Качество предсказания некодирующих РНК для диффузионной модели (сплошная линия) и на основе Z-значений (пунктирная линия). Результаты показаны для всех структурных РНК (А) и отдельно для тРНК (Б), мякРНК (В) и микроРНК (Г).

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.

Список литературы

1 Pervouchine DD, Khrameeva EE, Pichugina MY, Nikolaienko OV, Gelfand MS, Rubtsov PM, *et al.* Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* 2012; **18**:1–15.

2 Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 2005; **361**:13–37.

3 Gultyaev AP, Fouchier RAM, Olsthoorn RCL. Influenza virus RNA structure: unique and common features. *Int Rev Immunol* 2010; **29**:533–556.

4 Lee JT. Epigenetic Regulation by Long Noncoding RNAs. *Science* 2012; **338**:1435–1439.

5 Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL. De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol* 2010; **28**:9–19.

6 Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 2011; **43**:513–518.

7 Beadle GW, Tatum EL. Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* 1941; **27**:499–506.

8 Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009; **136**:215–233.

9 Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014; **42**:D68–73.

10 Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009; **19**:92–105.

11 Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* 2009; **10**:126–139.

12 Ishizu H, Siomi H, Siomi MC. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev* 2012; **26**:2361–2373.

13 Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009; **458**:223–227.

14 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012; **22**:1775–1789.

15 Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. Requirement for Xist in X chromosome inactivation. *Nature* 1996; **379**:131–137.

16 He S, Liu S, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol* 2011; **11**:102.

17 Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, *et al.* RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. *Cell* 2014; **159**:188–199.

18 Mandal M, Breaker RR. Gene regulation by riboswitches. *Nat Rev Mol Cell Biol* 2004; **5**:451–463.

19 Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet TIG* 2004; **20**:44–50.

20 Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, *et al.* NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res* 2002; **30**:395–397.

21 Das J, Mukherjee S, Mitra A, Bhattacharyya D. Non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis. *J Biomol Struct Dyn* 2006; **24**:149–161.

22 Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999; **288**:911–940.

23 Tinoco I, Bustamante C. How RNA folds. *J Mol Biol* 1999; **293**:271–281.

24 Rose D. RNA secondary structures. 2011.http://www.bioinf.uni-freiburg.de/Lehre/Courses/2011_WS/V_BioinfoII/slides_Rose-RNA-nussinov-zuker.pdf

25 Gorodkin J, Hofacker IL. From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput Biol* 2011; **7**:e1002100.

26 Lyngsø RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. *J Comput Biol J Comput Mol Cell Biol* 2000; **7**:409–427.

27 Liu B, Mathews DH, Turner DH. RNA pseudoknots: folding and finding. *F1000 Biol Rep* 2010; **2**:8.

28 Staple DW, Butcher SE. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* 2005; **3**:e213.

29 Clote P, Kranakis E, Krizanc D, Salvy B. Asymptotics of canonical and saturated RNA secondary structures. *J Bioinform Comput Biol* 2009; **7**:869–893.

30 Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for Loop Matchings. *SIAM J Appl Math* 1978; **35**:68–82.

31 Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A* 1980; **77**:6309–6313.

32 Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* 2004; **101**:7287–7292.

33 Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry (Mosc)* 1998; **37**:14719–14735.

34 Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res* 2006; **34**:564–574.

35 Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 1981; **9**:133–148.

36 Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol AMB* 2011; **6**:26.

37 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003; **31**:3406–3415.

38 Bellaousov S, Reuter JS, Seetin MG, Mathews DH. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res* 2013; **41**:W471–474.

39 Waterman MS, Smith TF. Rapid dynamic programming algorithms for RNA secondary structure. *Adv Appl Math* 1986; **7**:455–464.

40 Ogurtsov AY, Shabalina SA, Kondrashov AS, Roytberg MA. Analysis of internal loops within the RNA secondary structure in almost quadratic time. *Bioinforma Oxf Engl* 2006; **22**:1317–1324.

41 Clote P. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J Comput Biol J Comput Mol Cell Biol* 2005; **12**:83–101.

42 Lai D, Proctor JR, Meyer IM. On the importance of cotranscriptional RNA structure formation. *RNA* 2013; **19**:1461–1473.

43 Danilova LV, Pervouchine DD, Favorov AV, Mironov AA. RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J Bioinform Comput Biol* 2006; **4**:589–596.

44 Voss B, Meyer C, Giegerich R. Evaluating the predictability of conformational switching in RNA. *Bioinforma Oxf Engl* 2004; **20**:1573–1582.

45 Layton DM, Bundschuh R. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res* 2005; **33**:519–524.

46 McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990; **29**:1105–1119.

47 Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA N Y N* 2004; **10**:1178–1190.

48 Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA N Y N* 2009; **15**:1805–1813.

49 Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* 2010; **464**:279–282.

50 Innan H, Stephan W. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics* 2001; **159**:389–399.

51 Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, *et al.* Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res* 1981; **9**:6167–6189.

52 Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 2003; **31**:3423–3428.

53 Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 2008; **9**:474.

54 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57–74.

55 Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet* 2014; **10**:e1004525.

56 Morris KV, Mattick JS. The rise of regulatory RNA. *Nat Rev Genet* 2014; **15**:423–437.

57 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**:860–921.

58 Le SV, Chen JH, Currey KM, Maizel JV. A program for predicting significant RNA secondary structures. *Comput Appl Biosci CABIOS* 1988; **4**:153–159.

59 Chen JH, Le SY, Shapiro B, Currey KM, Maizel JV. A computational procedure for assessing the significance of RNA secondary structure. *Comput Appl Biosci CABIOS* 1990; **6**:7–18.

60 Workman C, Krogh A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res* 1999; **27**:4816–4822.

61 Seffens W, Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* 1999; **27**:1578–1584.

62 Clote P, Ferré F, Kranakis E, Krizanc D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA N Y N* 2005; **11**:578–591.

63 Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* 2004; **342**:19–30.

64 Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinforma Oxf Engl* 2004; **20**:2911–2917.

65 Freyhult E, Gardner PP, Moulton V. A comparison of RNA folding measures. *BMC Bioinformatics* 2005; **6**:241.

66 Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001; **2**:8.

67 Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013; **29**:2933–2935.

68 Klein RJ, Eddy SR. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* 2003; **4**:44.

69 Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res* 1997; **25**:0955–964.

70 Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinforma Oxf Engl* 2000; **16**:583–605.

71 Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007; **8**:R22.

72 Kavanaugh LA, Dietrich FS. Non-Coding RNA Prediction and Verification in Saccharomyces cerevisiae. *PLoS Genet* 2009; **5**:e1000321.

73 Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 2005; **102**:2454–2459.

74 Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput Pac Symp Biocomput* 2010; :69–79.

75 Wan X-F, Lin G, Xu D. Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. *J Bioinform Comput Biol* 2006; **4**:1015–1031.

76 Hofacker IL, Priwitzer B, Stadler PF. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinforma Oxf Engl* 2004; **20**:186–190.

77 Gruber AR, Bernhart SH, Zhou Y, Hofacker IL. RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures. In: *German Conference on Bioinformatics 2010*. ; 2010. pp. 12–21.

78 Wexler Y, Zilberstein C, Ziv-Ukelson M. A study of accessible motifs and RNA folding complexity. *J Comput Biol J Comput Mol Cell Biol* 2007; **14**:856–872.

79 Kabakcioglu A, Stella AL. A scale-free network hidden in the collapsing polymer. *Phys Rev E* 2005; **72**. doi:10.1103/PhysRevE.72.055102

80 Horesh Y, Wexler Y, Lebenthal I, Ziv-Ukelson M, Unger R. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics* 2009; **10**:76.

81 Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, *et al.* Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol* 2006; **2**:e33.

82 Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Res* 2007; **18**:000–000.

83 Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* 2005; **33**:2433–2439.

84 Will S, Yu M, Berger B. Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res* 2013; **23**:1018–1027.

85 Lande R. Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution* 1976; **30**:314–334.

86 Kimura M. The neutral theory of molecular evolution: a review of recent evidence. *Idengaku Zasshi* 1991; **66**:367–386.

87 Wright S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc Sixth Int Congr Genet* 1932; **1**:356–366.

88 Crow JF, Kimura M. *An introduction to population genetics theory*. Burgess Pub. Co.; 1970.

89 Felsenstein J. Phylogenies and Quantitative Characters. *Annu Rev Ecol Syst* 1988; **19**:445–471.

90 Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am Nat* 2004; **164**:683–695.

91 Kimura M, Ohta T. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics* 1969; **61**:763–771.

92 Lawrie DS, Petrov DA. Comparative population genomics: power and principles for the inference of functionality. *Trends Genet TIG* 2014; **30**:133–139.

93 Hansen TF, Martins EP. Translating Between Microevolutionary Process and Macroevolutionary Patterns: The Correlation Structure of Interspecific Data. *Evolution* 1996; **50**:1404–1417.

94 Hansen TF. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution* 1997; **51**:1341–1351.

95 Bedford T, Hartl DL. Optimization of gene expression by natural selection. *Proc Natl Acad Sci* 2009; **106**:1133–1138.

96 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, *et al.* The evolution of gene expression levels in mammalian organs. *Nature* 2011; **478**:343–348.

97 Keller TE, Mis SD, Jia KE, Wilke CO. Reduced mRNA Secondary-Structure Stability Near the Start Codon Indicates Functional Genes in Prokaryotes. *Genome Biol Evol* 2012; **4**:80–88.

98 Trotta E. On the Normalization of the Minimum Free Energy of RNAs by Sequence Length. *PLoS ONE* 2014; **9**:e113380.

99 Schattner P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res* 2002; **30**:2076–2082.

100 Smit S, Knight R, Heringa J. RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic Acids Res* 2009; **37**:1378–1386.

101 Fontana null, Stadler null, Bornberg-Bauer null, Griesmacher null, Hofacker null, Tacker null, *et al.* RNA folding and combinatory landscapes. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top* 1993; **47**:2083–2099.

102 Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, *et al.* Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. *PLoS Comput Biol* 2012; **8**:e1002529.

103 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al.* The Human Genome Browser at UCSC. *Genome Res* 2002; **12**:996–1006.

104 Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, *et al.* MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* 2009; :gkp919.

105 Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* 2015; **43**:D130–D137.

106 De Hoon MJL, Makita Y, Nakai K, Miyano S. Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput Biol* 2005; **1**:e25.

107 Saito S, Kakeshita H, Nakamura K. Novel small RNA-encoding genes in the intergenic regions of Bacillus subtilis. *Gene* 2009; **428**:2–8.

108 Petrillo M, Silvestro G, Nocera PPD, Boccia A, Paolella G. Stem-loop structures in prokaryotic genomes. *BMC Genomics* 2006; **7**:170.

109 Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, Backofen R. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res* 2012; :gks181.

110 Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in Bacillus subtilis. *Nucleic Acids Res* 2010; **38**:6637–6651.

111 Katz L, Burge CB. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* 2003; **13**:2042–2051.

112 Park C, Chen X, Yang J-R, Zhang J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* 2013; **110**:E678–686.

113 Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* 2004; **32**:4925–4936.