МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В. ЛОМОНОСОВА
ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

На правах рукописи

Гречишникова Дарья Александровна

РОЛЬ ВТОРИЧНЫХ СТРУКТУР РНК В РЕТРОТРАНСПОЗИЦИИ

Специальность` 03.01.02 – биофизика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук,
профессор, заведующий кафедрой биофизики
Физического факультета МГУ Твердислов В.А.

Москва, 2017

Оглавление

Введение 4

Цель диссертационной работы 6

Научная новизна 6

Научная и практическая значимость работы 7

Положения, выносимые на защиту 8

Глава 1. Обзор литературы 9

1.1. Роль вторичных структур ДНК и РНК в функционировании генома 9

1.2. LINE и SINE транспозоны: строение, механизм перемещения, эволюция 12

1.3. Применение динуклеотидных физических и геометрических характеристик для описания структуры РНК и ДНК 24

Заключение к Главе 1 34

Глава 2. Консервативная вторичная структура стебель-петля на конце РНК-транскриптов транспозонов и ее роль в ретротранспозиции 35

2.1. Метод поиска консервативных вторичных структур РНК/ДНК в геноме 35

2.2. Анализ транспозонов семейства L1 в геноме человека 41

2.3. Анализ транспозонов семейства Alu в геноме человека 44

2.4. Структура стебель-петля в нетранслируемой 3’-области LINE и SINE транспозонов 48

Заключение к Главе 2 59

Глава 3. Выявление физико-химических характеристик обнаруженных структур стебель-петля транспозонов L1 в геноме человека 60

3.1. Методы, используемые для анализа структуры стебель-петля 60

3.2. Физико-химические характеристики структуры стебель-петля L1 транспозонов человека. 66

3.3. Хиральная иерархичность в системе «РНК + белок». 78

3.4. Построение математической модели для автоматического распознавания вторичных структур транспозонов по их физико-химическим характеристикам 80

Заключение к Главе 3 91

Заключение 93
ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ ДИССЕРТАЦИИ 94
Приложения 95
Благодарности 106
Список литературы 107

Введение

Широкий ряд работ посвящен изучению роли структуры молекулы РНК в регуляции клеточных процессов [1-4]. В отличие от ДНК, РНК состоит из одной нити, однако ее участки способны взаимодействовать, придавая молекуле разнообразные структурные конфигурации: псевдоузлы, петли, шпильки (структуры стебель-петля), триплексы и квадруплексы. Эти образования относятся ко вторичным структурам, изучение которых выявляет их существенную роль во многих внутриклеточных процессах, таких как транскрипция ДНК, сплайсинг, экспрессия генов, транспорт транскриптов, взаимодействие с белками. Так, вторичная структура стебель-петля способна регулировать уровень стабильности РНК-транскрипта. Например, в геномах некоторых бактерий шпилька, расположенная на 3’-конце транскрипта, предотвращает расщепление мРНК экзонуклеазой [5]. На 5’-конце *ompA* транскрипта, кодирующего мембранный белок, располагается структура стебель-петля, благодаря которой данная мРНК является самым стабильным транскриптом в *Escherichia coli* [6]. Следует отметить важную роль вторичных структур в экспрессии генов. РНК способна «переключаться» между двумя конформациями, одна из которых блокирует экспресию, а другая, наоборот, допускает [1].

Множество взаимодействующих с РНК белков связываются именно с двунитиевыми участками РНК. Они принамают участие в широком спектре биологических процессов, включая редактирование РНК (ADAR2), РНК-интерференцию (TRBP, HYL1), функционирование рибосомальной РНК (Rnt1p и Aquifex aeolicus RNase III) и локализацию мРНК (Drosophila Staufen) [7]. Так, редактирование РНК осуществляет специальный белок ADAR (Adenosine deaminases acting on RNA). Он распознает структуру стебель-петля и связывается с ней. В наиболее изученной мРНК глутаматного рецептора сайт связывания с белком находится в шпильке длиной в 71 нуклеотид. Мутации, разрушающие структуру, приводят к отсутствию какого-либо взаимодействия [8].

Перенос генетического материала из одного участка молекулы ДНК в другой происходит в результате особого процесса – транспозиции, или премещения мобильных генетических элементов – транспозонов. В настоящее время активно исследуется роль транспозонов в возникновении тяжёлых заболеваний у человека, хотя долгое время считалось, что транспозоны как фрагменты ДНК, способные размножаться и перемещаться в геноме, являются так называемой «мусорной» ДНК и не несут никакой функциональной нагрузки. Но при этом транспозоны содержатся в геномах всех эукариот и занимают значительную их часть (46% генома человека, 37% генома мыши, 10% генома плодовой

мушки, 85% генома кукурузы) [9, 10]. Существует множество заболеваний человека, вызванных скачком транспозона. Так, известен случай заболевания гемофилией, вызванный скачком транспозона в ген коагуляционного фактора [11]. Исследуется роль транспозонов в возникновении рака, в поддержании вариативности геномов различных органов, в том числе мозга и иммунной системы [12]. Существует предположение, что транспозоны являются инструментом эволюции, так как они способны вызывать как масштабные перестройки (например, рекомбинации между двумя неаллельными элементами в двух разных хромосомах), так и небольшие изменения генома (дупликации, инверсии, делеции) [13]. Более того, транспозоны способны влиять на собственную экспрессию и на экспрессию близлежащих генов [14].

По типу перемещения транспозоны можно разделить на те, которые перемещаются методом “вырезать и вставить”, и те, которые используют метод “копировать и вставить”. Второй способ перемещения (посредством создания РНК-транскрипта и последующего синтеза последовательности ДНК по матрице РНК) называется ретротранспозицией, так как он аналогичен механизму размножения ретровирусов. Транспозоны, перемещающиеся подобным образом, называются ретротранспозонами. В данном исследовании рассматриваются только ретротранспозоны, которые для краткости в дальнейшем мы будем называть просто транспозонами.

Для осуществления транспозиции LINE используют собственный молекулярный аппарат, закодированный ими для копирования своей последовательности и вставки ее в геном [15]. SINE являются паразитами, они не кодируют белки, для транспозиции им требуются белки LINE [16]. Каким образом белок распознает собственную РНК и РНК SINE остается неясным. Для нескольких организмов было показано, что белок распознает вторичную структуру стебель-петля на 3’-конце РНК-транскрипта [17-20]. Более того, было показано, что некоторые организмы имеют одинаковые последовательности на 3’-конце LINE и SINE транспозонов [16]. В таких случаях предполагается, что белки LINE “узнают” вторичную структуру на 3’-конце, одинаковую для SINE и LINE РНК-транскрипта. В случаях, когда транспозоны не имеют общего 3’-конца, принято считать, что распознается поли-А хвост, имеющийся и у LINE, и у SINE транспозонов [21, 22]. Однако, поли-А хвосты имеют практически все мРНК, что ставит под сомнение возможность избирательного распознавания на основе связывания с ними. Механизм транспозиции пока недостаточно изучен. Одним из важнейших вопросов является вопрос о способе распознавания белком LINE своего РНК-транскрипта или РНК-транскрипта SINE.

Цель диссертационной работы - выяснение роли вторичных структур РНК в механизме распространения транспозонов и охарактеризование физико-химических свойств структур, участвующих в процессах эволюционного отбора.

Были поставлены следующие **задачи**:

- Разработать программный метод поиска консервативных вторичных структур РНК/ДНК в геноме.
- Разработать алгоритм и программу сравнения нуклеотидных последовательностей на основе вторичных структур РНК/ДНК.
- Выполнить поиск консервативных позиций вторичных структур РНК в транспозонах LINE и SINE в геноме человека, а также всех организмов, доступных в базе мобильных элементов RepBase.
- Для обнаруженных вторичных структур РНК, имеющих консервативное положение внутри транспозона, выявить физико-химические и геометрические, в том числе хиральные, характеристики и сравнить их с аналогичными структурами, сгенерированными случайным образом или взятыми из случайных областей генома.
- На основе отобранных физико-химических и геометрических характеристик вторичных структур РНК построить математическую модель, позволяющую выделять структуры, принадлежащие транспозонам.

Научная новизна

- Впервые разработан метод и реализована программа поиска консервативных вторичных структур в наборе последовательностей РНК/ДНК.
- Впервые найдена консервативная позиция вторичной структуры РНК на 3’-конце LINE и SINE транспозонов человека, а также ряда других организмов, распределенных по всему дереву жизни.
- Впервые выявлены физико-химические и геометрические характеристики вторичных структур РНК транспозонов, статистически значимо отличающиеся от контрольных структур, взятых из случайных мест генома или сгенерированных случайным образом. Впервые показано, что к наиболее существенным характеристикам относятся параметр Rise, энтропия и гидрофильность.
- Впервые определены физические свойства вторичных структур РНК транспозонов, находящихся под действием эволюционного отбора. Выдвинута

гипотеза, что данные физические свойства играют существенную роль в распознавании транспозона обратной транскриптазой.

- Впервые показано наличие знакопеременных иерархий хиральных структур в комплексах молекулы РНК и белка, осуществляющего транспозицию.
- Впервые построена математическая модель, позволяющая автоматически с высокой точностью определять вторичные структуры, принадлежащие 3’-концу транспозонов в геноме человека.

Научная и практическая значимость работы

Вторичные структуры РНК/ДНК играют принципиально важную роль во всех важнейших процессах функционирования генома. Разработанный в работе программный комплекс дает возможность выявить позиции консервативных вторичных структур в ДНК или РНК. Это особенно важно для определения эволюционного родства последовательностей. Не обнаруживаемое на уровне первичной последовательности сходство может проявляться при переходе на следующий уровень – уровень вторичных структур РНК/ДНК.

Позиции найденных консервативных структур транспозонов представляют важность для исследования комплекса белков LINE и РНК-транскрипта. Обнаружение консервативной структуры и определение ее физико-химических и геометрических свойств важно для понимания процесса селективного распознавания белком РНК транспозона.

Построенная модель машинного обучения может быть использована для выявления в произвольном участке генома вторичных структур РНК, распознаваемых обратной транскриптазой, а, следовательно, способных инициировать связывание с ней. Такими последовательностями могут быть не только известные транспозоны, но и произвольные участки генома, включая гены и межгенные области.

Впервые использованный подход с рассмотрением знакопеременных иерархий хиральных структур в комплексах РНК-белок может быть использован для изучения более широкого класса взаимодействий нуклеиновых кислот и белков.

Выявленные в данной работе свойства, характерные для вторичных структур стебель-петля на конце транспозонов, могут иметь большое значение в биоинженерии для встраивания любых заданных последовательностей в геном при помощи белка ORF2 транспозонов LINE.

Положения, выносимые на защиту

- Транспозоны, относящиеся к классам LINE и SINE, обладают консервативной вторичной структурой стебель-петля, расположенной на 3’-конце РНК-транскрипта.
- Связывание РНК-транскрипта транспозонов LINE и SINE с белком ORF2 транспозона LINE происходит через распознавание консервативной структуры стебель-петля на 3’-конце транскрипта.
- Функциональные характеристики консервативных вторичных структур стебель-петля, расположенных на 3’-конце транспозонов, находятся под действием физических факторов эволюционного отбора.

Глава 1. Обзор литературы

1.1. Роль вторичных структур ДНК и РНК в функционировании генома

Молекула РНК состоит из длинной цепи, разные части которой могут взаимодействовать, приводя к образованию сложных структур. Примером таких структур является так называемая РНК структура стебель-петля. Стебель составляют Уотсон-Криковские пары нуклеотидов, а петля – это цепочка одиночных нуклеотидов. Стебель может содержать так называемые пузырьки, участки с неспаренными нуклеотидами на одной нити молекулы РНК, и внутренние петли, участки с неспаренными нуклеотидами на обеих нитях (Рис. 1). Структура стебель-петля чаще всего образуется в таких участках генома, последовательность которых является палиндромом.

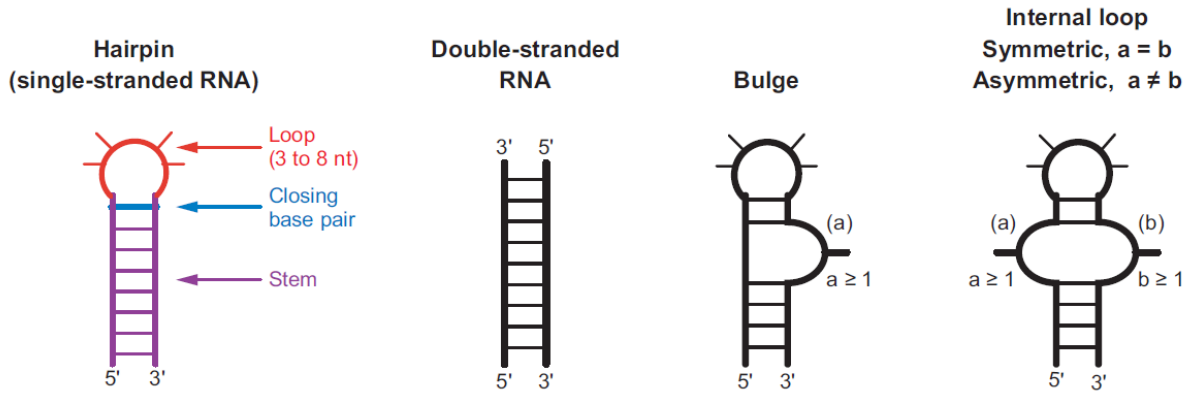


Рис. 1. Структура стебель-петля [1].

РНК структуры стебель-петля играют важную роль во многих процессах, проходящих в клетке. Например, они необходимы для РНК фолдинга, определяют взаимодействия с рибозимами, защищают мРНК от деградации, служат опознавательным знаком для белков, связывающихся с РНК, и субстратом в ферментативных реакциях, участвуют в регулировке трансляции и других ключевых процессах. [2]

Структуры стебель-петля могут образовываться и на молекуле ДНК. Известно, что они участвуют в регуляции репликации, генной экспрессии, являются необходимыми для связи с некоторыми белками [3].

Матричная РНК генов гистонов не полиаденилируется, 3'-некодирующая часть мРНК содержит структуру стебель-петля. Известно, что именно эта структура необходима для регуляции стабильности мРНК гистона [2]. Множество доказательств было приведено в пользу участия РНК шпильки в терминации транскрипции. Однако ее точную роль в этом процессе установить пока не удалось. Было выдвинуто несколько предположений:

структура стебель-петля может останавливать элонгационный комплекс, что повышает вероятность терминации транскрипции, или ухудшает взаимодействие между полимеразой и РНК [1].

Известно, что белки, при помощи которых происходит сплайсинг, связываются с определенными последовательностями в молекуле РНК. Структуры стебель-петля могут ингибировать сплайсинг, перекрывая доступ к этим участкам. Однако они не всегда стабильны. В таком случае возникает конкурентная борьба между ними и белком. Интересно, что структуры стебель-петля могут и помогать сплайсингу. Например, одним из распространенных механизмов является уменьшение расстояния между важными сигнальными последовательностями. Сигнальные последовательности могут оказаться рядом друг с другом из-за образования структуры стебель-петля между ними. Было показано, что мутации, разрушающие структуры стебель-петля, уменьшают эффективность сплайсинга [4].

Длинные структуры стебель-петля представляют угрозу для стабильности генома, так как они распознаются специальными белками, способными разрушать подобные структуры. Короткие палиндромные последовательности необходимы для основных геномных процессов, таких как репликация и транскрипция [1]. Структуры стебель-петля также могут играть не последнюю роль в эпигенетике. Например, известно, что последовательность, входящая в ДНК-структуру, защищена от метилирования. Так, DAM метилаза оказывается не способной модифицировать мотив-мишень в случае, если он находится в составе структуры стебель-петля [3, 23]. Было показано in vivo, что последовательность в центре палиндрома в геноме фага лямбды резистентна к метилированию. Таким образом, структуры стебель-петля могут напрямую управлять метилированием цепи ДНК. Это дает возможность выделить еще один уровень регуляции в геноме [3].

Для большинства организмов, от бактерий до млекопитающих, известно множество белков, способных распознавать структуры стебель-петля. Большая часть может быть разделена на несколько категорий: белки, участвующие в транскрипции и репарации (PARP, BRCA1, p53, 14-3-3), в репликации (MLL, WRN, 14-3-3, helicases), белки хроматина (DEK, BRCA1, HMG protein family, topoisomerases), белки, разрешающие нетипичные соединения (junction-resolving enzymes). Одним из наиболее важных механизмов поддержания стабильности генома является ДНК репарация. Белки, способные связываться с поврежденной ДНК и вторичнмыми структурами, играют важную роль в этом процессе. Часто в промоутерных участках генов содержатся обратные повторы, способные к образованию структур стебель-петля in vivo. Взаимодействующие с ДНК белки, такие как

HMGB-box family, Rad54, BRCA1 protein, (poly(ADP-ribose) polymerase-1), демонстрируют очень слабую способность распознавать нуклеотидную последовательность. При этом связываются они именно со структурами стебель-петля [3]. Было показано, что они служат опознавательными знаками начала репликации. Известно много белков, участвующих в репликации и связывающихся именно со структурами стебель-петля (14-3-3 семейство белков, MLL, WRN, AF10, S16, SMC семейство т.д.) [3].

Распознавание структур стебель-петля необходимо для поддержания стабильности генома и регуляции основных процессов, проходящих в клетке. Мутации и эпигенетические модификации, влияющие на способность того или иного участка ДНК к образованию таких структур, могут приводить к серьезным нарушениям клеточных процессов [1].

1.2. LINE и SINE транспозоны: строение, механизм перемещения, эволюция

Транспозоны – это участки ДНК, которые способны менять свое местоположение в геноме и увеличивать число своих копий. Транспозоны содержатся в геномах всех эукариот и занимают значительную часть генома (46% генома человека, 37% генома мыши, 10% генома плодовой мушки, 85% генома кукурузы (Рис. 2)) [9, 10].

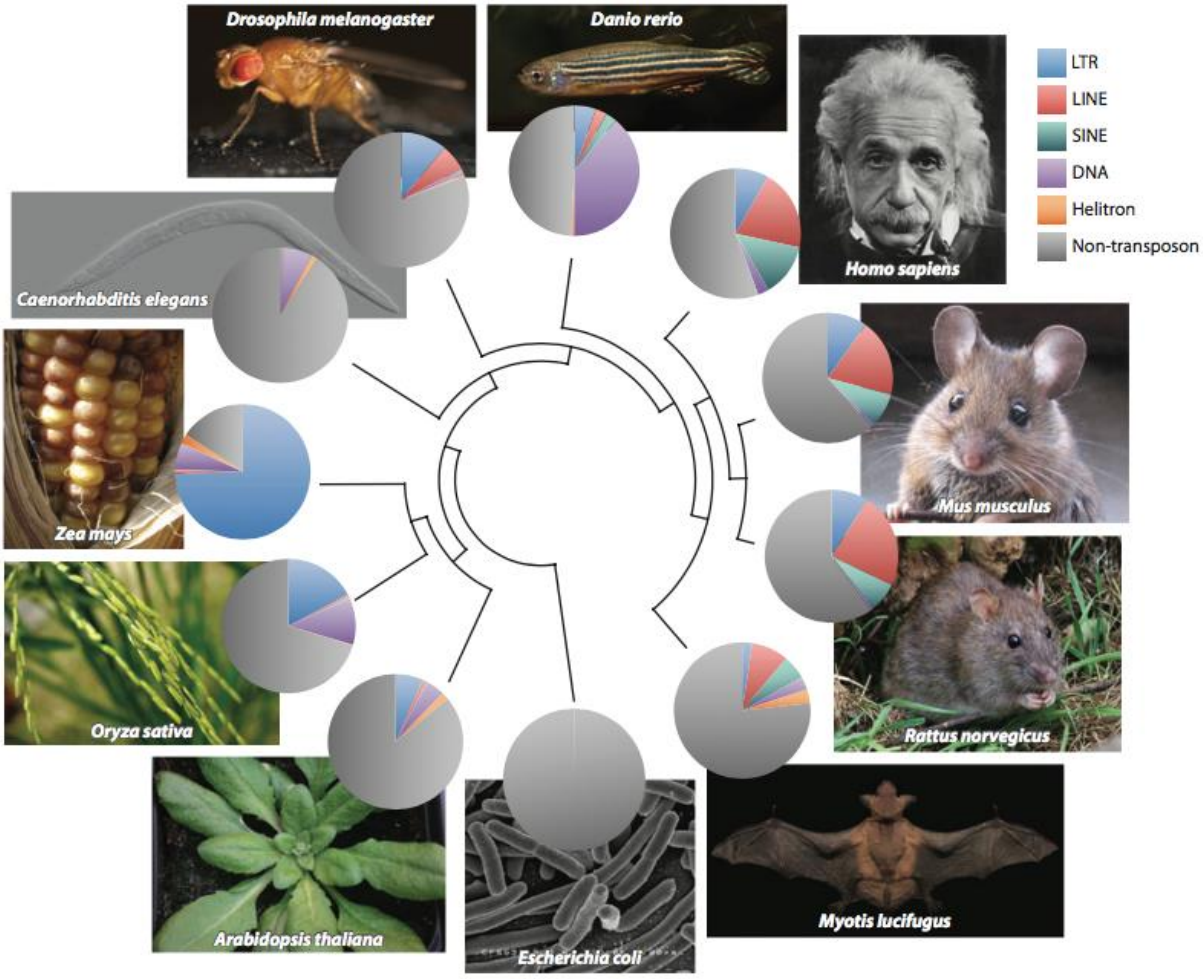


Рис. 2. Распространенность транспозонов в геномах различных организмов [9].

Долгое время считалось, что транспозоны являются так называемой “мусорной” ДНК и не несут никакой функциональной нагрузки. Однако после обнаружения у взрослого пациента гемофилии, вызванной скачком транспозона в ген коагуляционного фактора, стало появляться все больше и больше работ, посвященных изучению патогенной роли транспозонов. Известно 96 заболеваний человека, вызванных скачком транспозона. В настоящее время активно исследуется роль транспозонов в возникновении рака, разрабатываются методики использования мобильных элементов LINE-1 в качестве раковых маркеров [9, 11]. Установлено, что транспозоны вносят существенный вклад в

вариативность геномов различных органов, например, мозга и иммунной системы [12]. Было обнаружено повышенное число копий LINE-1 в клетках мозга людей, страдающих шизофренией [3]. Также есть предположение, что транспозоны являются инструментом эволюции, так как они способны вызывать как масштабные перестройки (например, рекомбинации между двумя неаллельными элементами в двух разных хромосомах), так и небольшие изменения генома (дупликации, инверсии, делеции) [13]. Более того, транспозоны способны влиять на собственную экспрессию и на экспрессию близлежащих генов. Уровень транскрипции зависит от типа ткани и подвержен влиянию внешних факторов, например, стресса в мозге млекопитающих [14].

По типу перемещения транспозоны можно разделить на те, которые перемещаются методом “вырезать и вставить”, и те, которые используют метод “копировать и вставить”. На основе ДНК синтезируется РНК, а затем, с помощью обратной транскрипции, по РНК шаблону создается копия ДНК, которая и вставляется в геном. Транспозоны второго типа называются ретротранспозонами. Самые распространенные классы ретротранспозонов человека - LINE и SINE. Данная диссертационная работа главным образом сфокусирована на изучении вторичных структур, принадлежащих двум самым распространенным и активным в настоящее время семействам транспозонов человека – L1 (принадлежит классу LINE) и Alu (принадлежит классу SINE). Поэтому в последующем изложении основное внимание будет уделено именно им. Однако следует отметить, что разнообразие семейств LINE и SINE очень велико, и, учитывая их эволюционную взаимосвязанность, рассмотрение конкретных семейств в отрыве от остальных невозможно, поэтому сравнения с другими семействами и общие закономерности также будут приведены.

L1 семейство является самым большим по массе семейством LINE транспозанов в геноме человека. В нашем геноме содержится более 500 000 копий, которые занимают примерно 17% всей его длины [24]. Большинство L1 неполноценны – они укорочены или повреждены с 5` конца. Такие транспозоны утратили способность менять свое местоположение [25]. По оценкам в нашем геноме осталось всего 80-100 активных L1, сохранивших способность к перемещению [26]. Типичный L1 имеет длину около 6000 пар оснований и состоит из 5` нетранслируемого региона, двух открытых рамок считывания (ORFs), нетранслируемой 3`-области и поли-А хвоста (Рис. 3).

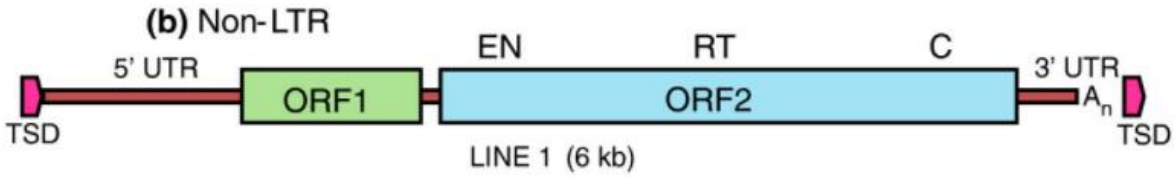


Рис. 3. Схема LINE транспозона [27].

Транскрипция L1 начинается с промоутера РНК полимеразы II, расположенного в нетранслируемой 5`-области [28, 29]. Первая открытая рамка считывания (ORF1) кодирует белок (ORF1p) с массой 40 кДа, состоящий из суперспирали (coiled-coil domain) [30], неканонического RRM домена (noncanonical RNA recognition motif domain) [31, 32] и карбоксильного конца [33]. Несмотря на то, что точная роль этого белка пока не установлена, было экспериментально показано, что он необходим для транспозиции [21]. Второй белок (ORF2p) имеет массу 150 кДа и совмещает в себе домены, обладающие активностью эндонуклеазы [34] и обратной транскриптазы [15]. Длина нетранслируемой 3`-области составляет приблизительно 200 пар оснований, последовательность нуклеотидов обладает низкой консервативностью. Она содержит полипуриновый тракт, который склонен к образованию G-квадруплексов [35], однако функция таких структур остается неизвестной. На конце L1 находится поли(A)-хвост, в эксперименте было установлено, что он очень важен для транспозиции [36]. Короткие прямые повторы, окаймляющие транспозон с обеих сторон, образуются при его интеграции в последовательность ДНК. Длина таких повторов варьируется от нескольких единиц до нескольких сотен нуклеотидов [25].

Второе по массе семейство транспозонов в геноме человека – это Alu. Alu принадлежат к классу SINE. По числу копий в геноме эти транспозоны являются наиболее распространенными – более 1 млн копий представлено в ДНК человека. Таким образом, Alu транспозоны занимают около 11% генома человека [24]. Длина типичного Alu составляет около 300 пар оснований. Alu произошли от 7SL РНК, являющейся одной из компонент сигналраспознающей частицы (Signal Recognition Particle, SRP), тогда как практически все остальные транспозоны, относящиеся к семейству SINE, произошли от транспортной РНК [37-39]. Alu состоит из двух мономеров, схожих с 7SL РНК, соединяющего их участка с высоким содержанием аденина и поли(A)-хвоста (Рис. 4).

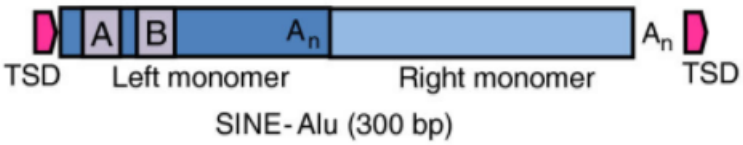


Рис. 4. Схема Alu транспозона [27].

В отличие от L1, Alu не способен к самостоятельному распространению по геному, так как не кодирует никаких белков. Вместо этого он заимствует молекулярный аппарат L1 и использует его для создания и вставки в геном собственных копий [40, 41]. Следует отметить, что все представители класса SINE не могут перемещаться самостоятельно, и для транспозиции они используют транспозоны LINE. Одним из важнейших и до сих пор открытых вопросов является вопрос о том, каким образом SINE удается “подсунуть” свой РНК транскрипт молекулярному аппарату LINE. Необычный жизненный цикл SINE транспозонов, являющихся по сути геномными паразитами, делает их эволюцию уникальной во многих отношениях. Интересно, что в ходе эволюции SINE много раз возникали de novo из доступных молекул (Рис. 5). Большинство SINE состоят из двух или более модулей: «голова» на 5’ конце, «тело» и «хвост» на 3’ конце (5’-terminal ‘head,’ ‘body’, 3'- terminal ‘tail’). «Голова» всех семейств SINE, известных на сегодняшний день, демонстрирует явное сходство с одним из трёх типов РНК: тРНК, 7SL РНК (РНК сигналраспознающей частицы) или 5S рРНК. Происхождение «тела» SINE непросто отследить, однако известно, что во многих случаях там присутствует последовательность, унаследованная от LINE. «Хвост» представляет собой последовательность переменной длины, состоящей из простых повторов [42].

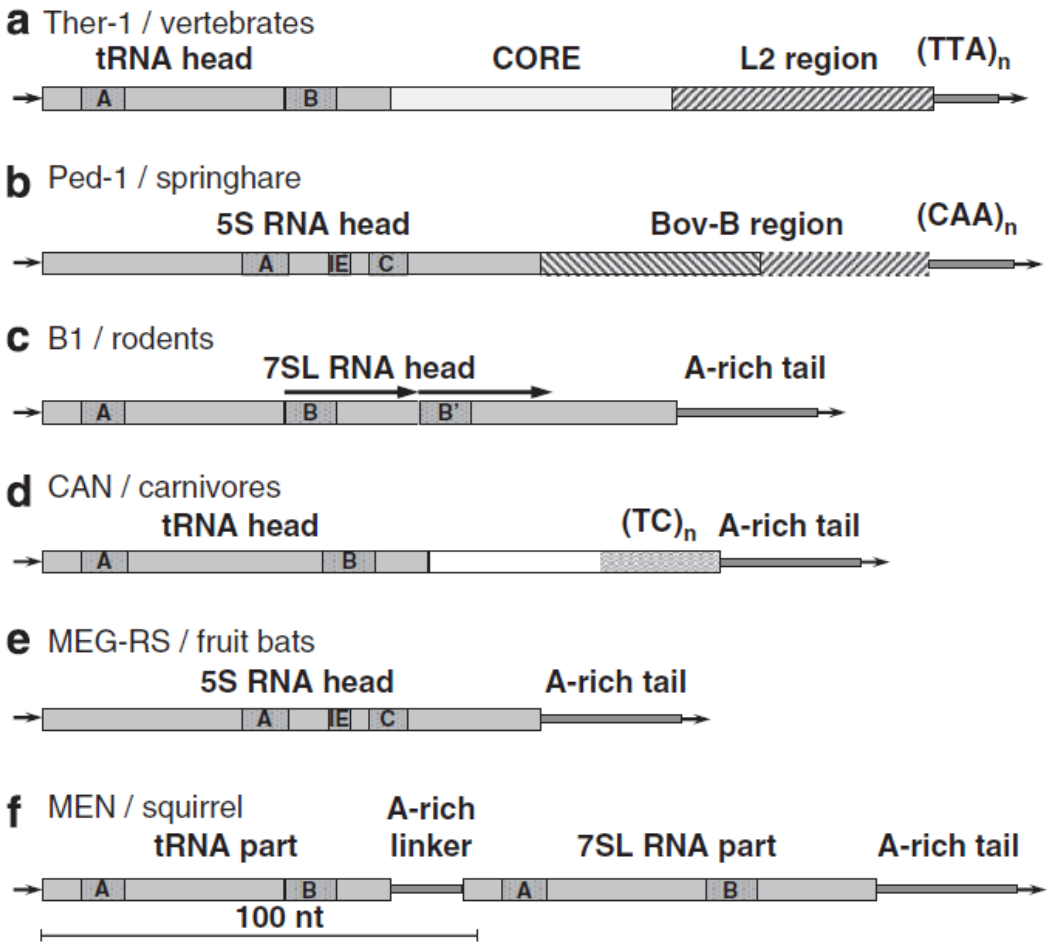


Рис. 5. Примеры структуры SINE элементов [42].

Схожесть нуклеотидной последовательности «головы» SINE с последовательностью одной из клеточных РНК указывает на ее происхождение из этой РНК. SINE транспозоны, происходящие от тРНК, особенно многочисленны (Рис. 5 a, d и f). Для многих семейств SINE транспозонов происхождение от тРНК может быть доказано с достаточной степенью точности, однако нуклеотидные замены и мутации, произошедшие в ходе эволюции, делают это невозможным для оставшихся семейств. На сегодняшний день SINE, произошедшие от 7SL RNA (Рис. 5 c и f), были идентифицированы только у грызунов [43], приматов [44] и тупай [45]. 7SL РНК встречается у всех эукариот как компонент сигналраспознающей частицы (SRP) – рибонуклеопротеина, который нацеливает секретируемые белки на эндоплазматический ретикулум. Число семейств SINE, происходящих из 5S рРНК, также невелико (Рис. 5 b и e); они были обнаружены у некоторых рыб [46] и у нескольких млекопитающих: летучих мышей[47] и некоторых грызунов [48].

Итак, ретротранспозиция – сложный многостадийный процесс, включающий транскрипцию, образование рибонуклеопротеиновой частицы, трансляцию, посттрансляционную модификацию, перемещение обратно в ядро и интеграцию в ДНК посредством механизма, называемого Target Primed Reverse Transcription (TPRT) [49]. После транскрипции мРНК LINE транспортируется из ядра в цитоплазму. В цитоплазме на рибосоме происходит синтез белков ORF1 и ORF2. Далее эти белки, предположительно прямо во время трансляции, связываются с мРНК. Важно отметить, что белки связываются именно с той РНК, с которой были транслированы. После этого комплекс из РНК и белков направляется обратно в ядро и связывается с сайтом-мишенью ДНК. Домен эндонуклеаза белка ORF2 расщепляет цепь ДНК, а домен обратная транскриптаза синтезирует ДНК последовательность по РНК матрице и вставляет последовательность в геном (Рис. 6). На каждом этапе РНК молекула транспозона взаимодействует с различными белками, в том числе и с теми, которые закодированы в последовательности самого транспозона. Часто структура стебель-петля РНК участвует в связывании белка с молекулой РНК [2].

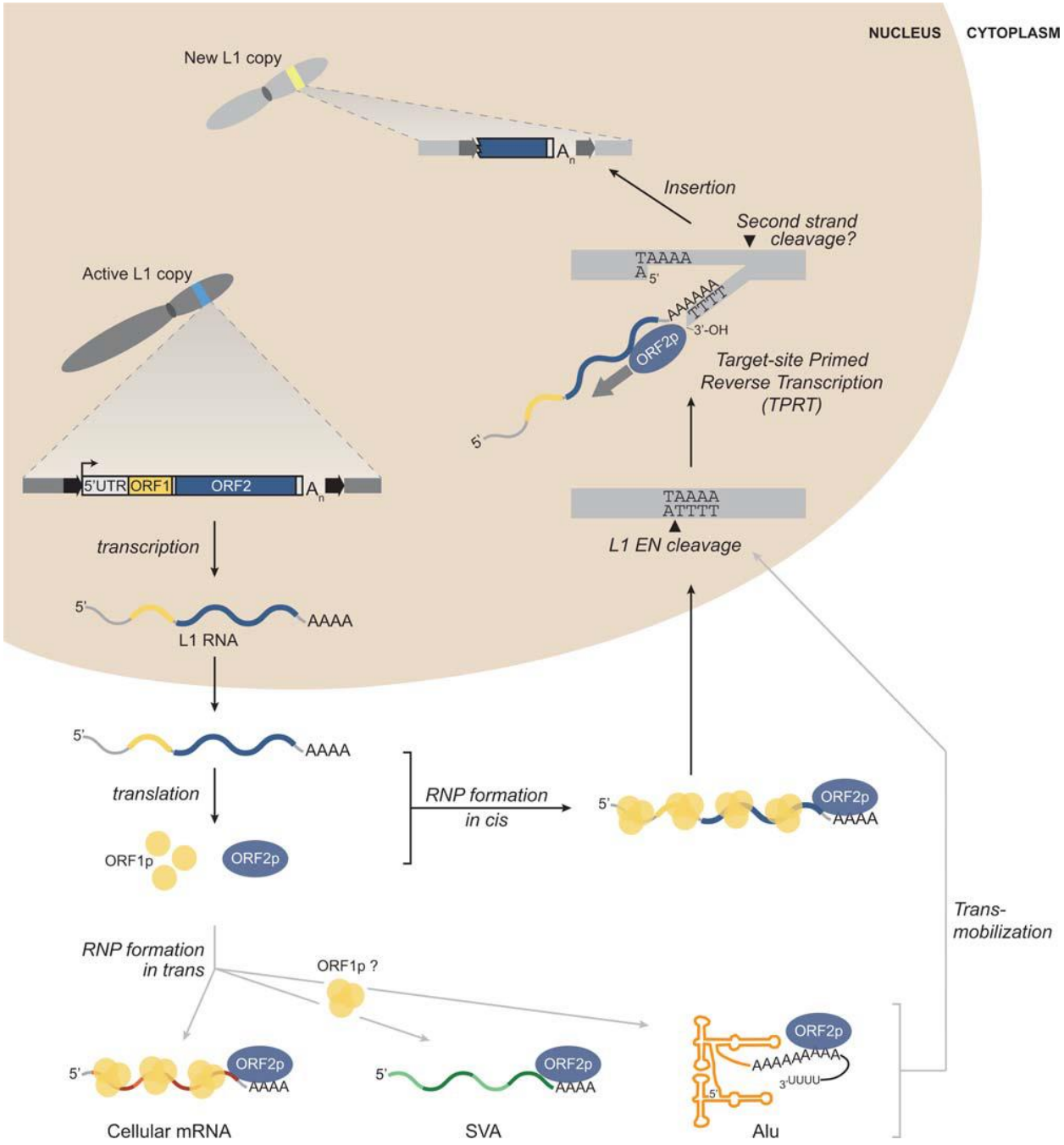


Рис. 6. Схема ретротранспозиции [14].

Вопрос о том, каким же образом обратная транскриптаза L1 распознает РНК-транскрипты самого себя и Alu, остается открытым. Так как транспозоны всего класса SINE для ретропозиции заимствуют молекулярный аппарат у транспозонов класса LINE, встает более общий вопрос – как обратная транскриптаза LINE распознает РНК и LINE, и SINE транспозонов? Этот вопрос становится еще более интригующим, если учесть тот факт, что белки ORF1 и ORF2 демонстрируют так называемое *cis*-предпочтение (феномен *cis*-preference в англоязычной литературе), то есть они связываются с тем РНК транскриптом, с которого были синтезированы [50-52]. Существует два типа пар активных LINE/SINE

транспозонов: пары, не имеющие в своем составе похожих участков, и пары, в последовательности которых такие участки прослеживаются. Схожесть LINE и SINE транспозонов была впервые обнаружена в геноме черепахи: было показано, что LINE транспозоны, принадлежащие CR1 семейству, имеют такую же последовательность нетранслируемой 3`-области, что и SINE транспозоны, произошедшие от транспортной РНК [53]. Позднее были найдены другие подобные примеры. Сейчас известно много LINE/SINE пар, имеющих одинаковые последовательности нетранслируемых 3`-областей [16, 54]. Такие пары известны в геномах растений [55, 56], рыб [18, 57, 58], насекомых [59] и млекопитающих [60]. На Рис. 7 показано филогенетическое дерево всех LINE транспозонов из всех изученных организмов. Красными овалами отмечены те LINE транспозоны, которые имеют общую нетранслируемую 3'-область с SINE.

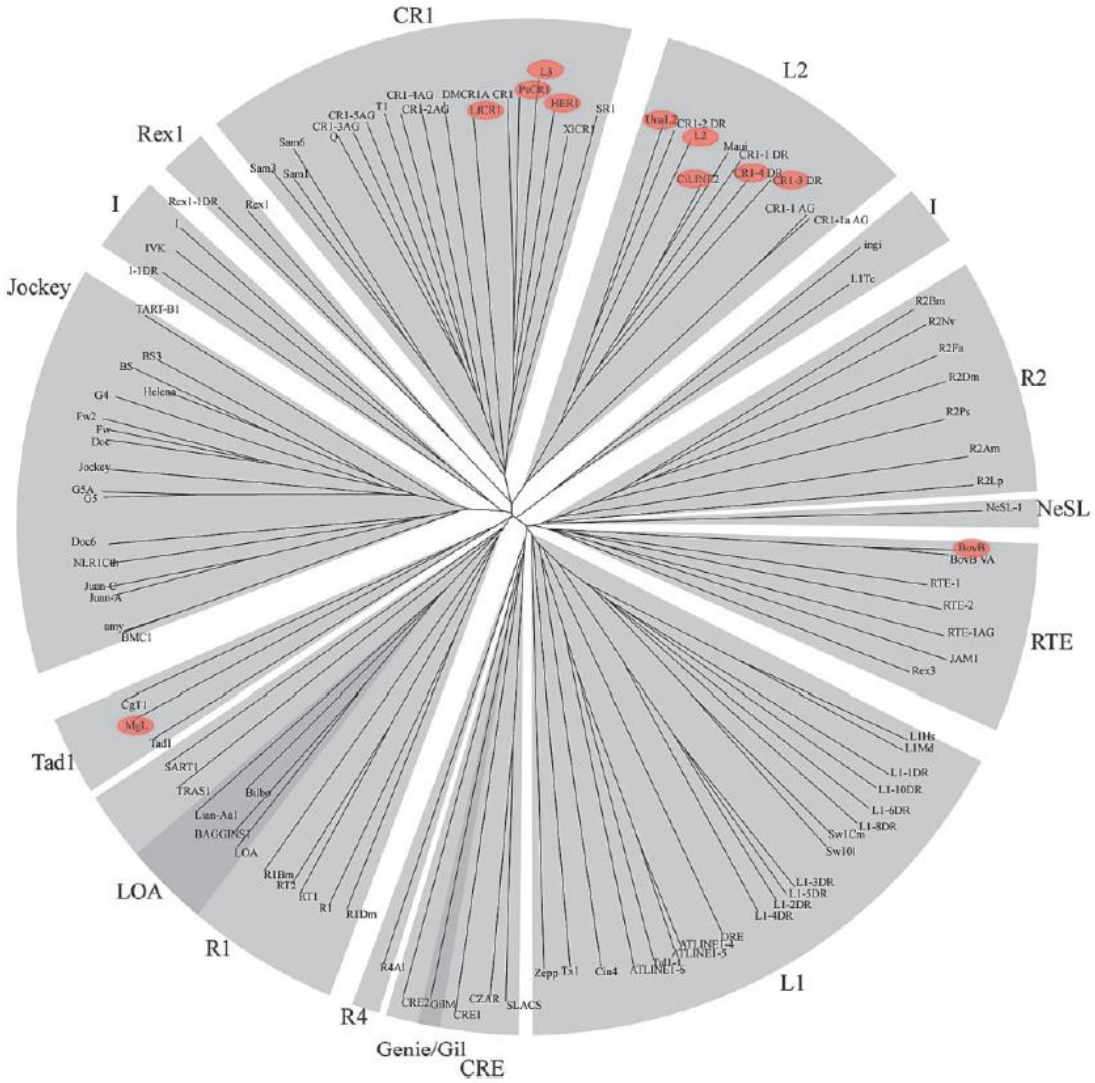


Рис. 7. Филогенетическое дерево LINE транспозонов. Красными овалами отмечены те LINE транспозоны, которые имеют общую нетранслируемую 3'-область с SINE [16].

Было предложено разделять все LINE транспозоны на две группы: строгую (stringent type) и нестрогую (relaxed type) [54]. Транспозоны, относящиеся к строгой группе, распознают только собственную нетранслируемую 3`-область, в то время, как транспозоны из нестрогой группы не требуют никакого специфичного распознавания 3`-конца, единственным необходимым условием для них является наличие поли(А)-хвоста. Так как L1 и Alu не содержат одинакового 3`-конца, L1 транспозоны были отнесены к нестрогой группе. Было предложено, что поли(А)-хвост нужен для распознавания молекулы РНК обратной транскриптазой и для эффективной ретропозиции при помощи молекулярного аппарата L1 [21, 22, 40, 55, 61, 62]. Прямой мониторинг Alu транспозона в клетке человека подтвердил необходимость поли(А)-хвоста для ретропозиции Alu [40].

Удлинение поли(А)-хвоста во время вставки Alu в новое место генома было продемонстрировано в работе [63]. По предположению, такое явление возможно при проскальзывании ORF2 белка. Существует гипотеза, что такое удлинение может играть важную роль в поддержании активности Alu транспозона [22]. Экспериментально было подтверждено, что поли(А)-хвост важен для успешной ретропозиции L1 [36]. Распознавание поли(А)-хвоста молекулярным аппаратом LINE может объяснять образование псевдогенов [50], однако многие псевдогены, не имеющие поли(А)-хвоста и произошедшие практически от всех типов РНК, были найдены в различных геномах [64, 65].

Во время трансляции ORF1p и ORF2p белков с РНК-транскрипта L1 белки и молекула РНК мгновенно образуют комплекс - рибонуклеопротеиновую частицу (RNP) [66]. Было показано, что обратная транкриптаза L1 связывается с РНК, и это связывание происходит перед поли(А)-хвостом [66]. Разветвленные молекулы, состоящие из связанных между собой 3`конца комплементарной ДНК и участка ДНК, в который будет вставлена новая копия транспозона, а также специфичное расположение L1 РНК внутри ORF2 белка были обнаружены во время наблюдения начальной стадии транспозиции in vitro [67]. Интересно, что удаление поли(А)-хвоста не оказало существенного влияния на ретропозицию, в то время как удаление более длинного участка привело к уменьшению числа транскриптов [67]. Кроме того, гипотеза о том, что в случаях, когда трансопзоны не имеют общего 3’-конца, распознается поли-А хвост, имеющийся и у LINE, и у SINE транспозонов, не согласуется с показанным экспериментально избирательным связыванием белка LINE именно со своим РНК транскриптом (транскриптом, с которого белок был синтезирован) [51, 68], так как поли-А хвосты имеются практически у всех мРНК [69].

Идея о том, что за распознавание и связывание с ORF2 белком, возможно, наряду с поли(А)-хвостом, может отвечать вторичная или третичная РНК структура, схожая для L1

и Alu, была предложена Boeke [62]. Тот факт, что ретропозиция может происходить даже в отсутствии поли(А)-хвоста [64, 65], дает право говорить о другом важном элементе, необходимом для распознавания РНК-транскрипта молекулярным аппаратом L1.

Для LINE транспозонов, входящих в строгую группу, одним из важных элементов является структура стебель-петля, расположенная в нетранслируемой 3`-области [18, 55-60]. К настоящему времени для различных видов собрано немало свидетельств того, что данная структура важна для ретропозиции. Так, экспериментальные доказательства необходимости данной структуры приведены для LINE транспозонов семейства SART1 тутового шелкопряда [59], LINE транспозонов семейства UnaL2 в геноме угря [20], LINE транспозонов ZfL2-1 и ZfL2-2 в геноме рыбы *Danio rerio* [17], SINE транспозонов семейства SmaI в геноме лосося [19], LINE R2 у насекомых [70]. Наличие структуры стебель-петля на 3`-конце LINE и SINE транспозонов, имеющих одинаковую последовательность нетранслируемой 3`-области, было продемонстрировано для значительно более широкого круга видов и семейств транспозонов, включая L1 в водорослях [55] и однодольных растениях [56], Tad1 в грибах [71], L2 у рыб [18, 57, 58], RTE у млекопитающих [60]. Эти находки заставляют задаться вопросом – была ли способность ORF2 распознавать структуру в нетранслируемой 3`-области пар LINE/SINE транспозонов, принадлежащих к строгой группе, сохранена в LINE транспозонах (в частности, в L1), относящихся к нестрогой группе.

Эволюционные исследования, основанные на сравнении доменов обратной транскриптазы в белке ORF2, выявили, что возраст LINE транспозонов совпадает с возрастом эукариот. Кроме того, такой анализ позволил выделить 11 монофилетических групп, или кладов, LINE транспозонов с вертикальным типом наследования, без признаков горизонтального переноса [72]. Однако стоит отметить, что случаи горизонтального переноса все же были зафиксированы для некоторых LINE кладов: перемещение элемента jockey в геноме Drosophila [73], перенос элемента Bov-B из генома чешуйчатых (Squamata) в геном предков жвачных (Ruminantia) [48, 74], вставка дополнительного карбоксильного конца (C-terminal domain) в ORF2 белок транспозона R1 насекомых из генома вируса растений [75] и L1 перенос из генома человека в геном бактерии Neisseria gonorrhoeae [76]. Филогенетический анализ, проведенный с использованием всей последовательности ORF2 LINE транспозона, для которого известен SINE-партнер [77], подтвердил разделение на 11 монофилетических групп, предложенное в [72], и несколько расширил его, доведя общее число кладов до 15 [77]. LINE транспозоны, для которых известны SINE партнеры, имеющие одинаковые с ними 3`-концы, в основном относятся к монофилетическим группам L2, CR1, RTE и Ted1. Однако набор данных, использовавшийся для анализа, не

включал в себя сведения о растениях и беспозвоночных, в которых также были найдены SINE/LINE пары [54, 56, 78]. В растениях L1 транспозон, имеющий одинаковую последовательность 3`-хвоста с SINE, был найден в геномах зеленых водорослей [55] и кукурузы [56]. Филогенетический анализ L1-клада, проведенный по полной последовательности ORF2, выявил монофилетические группы для зеленых водорослей, высших растений и позвоночных животных. Поскольку высшие растения произошли от зеленых водорослей, а L1 транспозоны, найденные в геноме зеленых водорослей, а также некоторых растений, относятся к строгой группе, Ошима предложил модель параллельного ослабления строго типа распознавания РНК молекулы у растений и млекопитающих [77].

В эволюционных исследованиях L1 и Alu транспозонов человека также были представлены доказательства вертикальной эволюции [79, 80]. L1 семейство транспозонов человека далее было подразделено на подсемейства. Подсемейства последовательно эволюционировали, в итоге придя к активной в настоящее время группе L1PA1 [79]. Анализ Alu последовательностей выявил более 200 различных семейств и 143 исходных транспозонов-источников [80].

Остановимся подробнее на примерах распознавания белком молекулы РНК с помощью шпильки. Так, для ретропозиции LINE транспозонов насекомых, R2, необходимы 250 нуклеотидов в нетранслируемой 3`-области. Без них распознавания не произойдет [70]. Интересно, что ORF2 белок транспозона R2 тутового шелкопряда Bombyx mori способен распознавать не только нетранслируемую 3`-область собственной РНК, но и 3`-конец R2 Drosophila melanogaster, хотя сходство нуклеотидных последовательностей на этих участках очень низко. Были построены модели вторичных структур, расположенных в нетранслируемых 3`-областях для обоих организмов. Оказалось, 3`-концы содержат сразу несколько шпилек [81]. Позднее в эксперименте было подтверждено, что в геноме тутового шелкопряда SART1 для ретропозиции важна структура стебель-петля, находящаяся в центре. Также было установлено, что транскрипция начинается, как правило, с нескольких GGUU участков, похожих на теломерные повторы и расположенных сразу после шпильки в направлении 3`-конца цепи ДНК [59].

В геноме угря транспозоны UnaL2 (LINE) и UnaSINE1 (SINE) имеют одинаковую нуклеотидную последовательность нетранслируемой 3`-области, что является залогом успешной транспозиции [18]. Обратная транскриптаза UnaL2 распознает 3`-конец UnaSINE1. Нетранслируемая 3`-область состоит из двух частей: шпильки с петлей GGAUA и пентануклеотидного повтора [TGTAA]n. Обе части необходимы для ретропозиции. Было высказано предположение, что 5`-часть петли GGAUA распознается обратной транскриптазой [82]. На стебле шпильки есть небольшая внутренняя петля, содержащая

одиночный неспаренный цитидин и U-U несовпадение. И то, и другое оказалось важно для ретропозиции. Внутренняя петля может служить для обеспечения необходимой для функционирования UnaL2 гибкости стебля [20].

В геноме полосатого данио ORF2 белок транспозона ZfL2 связывается со шпилькой на 3`-конце при помощи специального участка, расположенного между доменами эндонуклеазы и обратной транскриптазы. В данном случае распознавание может быть разделено на две составляющие, включающие общее распознавание стебля с помощью cTBR (консервативная область привязки к хвосту) ORF2 и специфическое распознавание цикла vTBR (переменная область привязки к хвосту) [17].

В геноме лосося SINE транспозон SmaI произошел от тРНК. Его 5`-участок принимает форму клеверного листа, подобно тРНК, в то время как на 3`-участке формируется шпилька, петля которой, по предположению, распознается обратной транскриптазой LINE [19]. Также было установлено, что три других SINE транспозона в геноме лосося, SlmI, HpaI и OS-SINE1, содержат одинаковый 3`-конец с LINE транспозоном RSg-1 [58].

В геноме черепахи последовательности 3`-конца SINE транспозонов семейства PolIII/SINE практически идентичны последовательностям на 3`-конце LINE транспозонов семейства CR1 [53]. Транспозоны семейства CR1 распространены в геномах птиц и рептилий. Примеры SINE транспозонов, произошедших от тРНК и имеющих одинаковые с LINE транспозонами последовательности на 3`-концах, встречаются в геномах черепах, рыб, млекопитающих и растений [54].

Гомологичность последовательностей 3`-концов SINE транспозона Vhc в геноме двустворчатого моллюска и LINE транспозона CR1 семейства в геноме моллюска Crassostrea gigas была установлена в работе [83]. Шпильки на 3’-конце Vhc SINE транспозонов из геномов различных видов моллюсков были предсказаны, однако пока не получены экспериментальные доказательства необходимости данных структур для транспозиции в геномах беспозвоночных. Также шпильки были найдены в центральной части V-домена. V-домен является консервативным участком в SINE транспозонах моллюсков и членистоногих. Однако его функции до сих пор не ясны.

Итак, экспериментально было показано, что шпилька в нетранслируемой 3’-области, одинаковая для LINE/SINE пары, играет важную роль в распознавании транспозона белком ORF2. Из этого факта возникает гипотеза о существовании единого механизма распознавания для LINE-SINE транспозонов, относящихся к нестрогой группе. Однако в настоящее время, принято считать, что распознается поли-А хвост, имеющийся и у LINE, и у SINE транспозонов. Но поли-А хвосты имеют практически все мРНК, что ставит под

сомнение возможность избирательного распознавания на основе связывания с ними. Механизм ретротранспозиции пока недостаточно изучен. Одним из важнейших вопросов является вопрос о способе распознавания белком LINE своего РНК-транскрипта или РНК-транскрипта SINE.

1.3. Применение динуклеотидных физических и геометрических характеристик для описания структуры РНК и ДНК

Последовательность нуклеотидов определяет пространственное строение РНК молекулы, а, следовательно, во многом и ее биологическую функцию. Трехмерная конформация РНК молекулы напрямую зависит от последовательности динуклеотидных пар, которая в свою очередь продиктована химическими свойствами нуклеотидов, образующих пару [84, 85]. Разнообразные формы и структуры молекулы ДНК могут быть описаны на динуклеотидном уровне, а именно в терминах локальных параметров цепи – сдвигов и углов между двумя нуклеотидами, расположенными друг за другом [86, 87]. Значительный прогресс достигнут в понимании геометрии динуклеотидных пар ДНК. Получены средние значения динуклеотидных параметров свободной ДНК и ДНК, взаимодействующей с белком [88]. Эти данные активно используются для моделирования структуры участков ДНК [89]. Распознавание участка ДНК белком, как правило, требует более подробного описания его упругих свойств, изогнутости, гибкости, чем описание в терминах модели упругого стержня. Молекула ДНК взаимодействует с белком, в том числе и посредством локальной деформации – изгиба, скручивания, интеркаляции, способность к которым напрямую определяется последовательностью нуклеотидов [90, 91]. Большинство белков, взаимодействующих с молекулой ДНК, контактируют с участком длиной 15-20 пар оснований [92]. Однако в их число не входят гистоны, область контакта для которых достигает 147 пар оснований [93]. Таким образом, гибкость молекулы ДНК на динуклеотидном уровне играет одну из ключевых ролей во взаимодействии с белком. Большинство теоретических и экспериментальных работ посвящено изучению упругих свойств В-формы ДНК [91]. Движения, которые молекуле ДНК необходимо совершить при взаимодействии с белком, совпадают с движениями, спонтанно происходящими в несвязанной молекуле ДНК. Существует предположение, что в процессе эволюции выработались такие структуры комплексов белок-ДНК, для которых ДНК молекула деформируется естественным для себя образом. Локальные константы жесткости, полученные для несвязанной В-ДНК и для комплексов ДНК-белок, очень близки. Это

позволяет сделать вывод о том, что глобальные изменения конформации ДНК, возникающие при связывании с белком, являются результатом суммирования небольших деформаций на уровне пар оснований [91].

Для описания пространственного расположения динуклеотидов друг относительно друга используют 3 трансляционных параметра – slide, shift, rise и 3 угла – tilt, twist, roll. Была введена следующая система координат. Ось *x* направлена вдоль короткой оси пары оснований, а ось *y* – вдоль длинной, ось *z* направлена перпендикулярно плоскости пары оснований таким образом, чтобы направляющие вектора осей *x, y, z* образовывали правую тройку векторов. Длинная ось пары оснований определяется как линия, соединяющая C-6 атом пиримидина с C-8 атомом пурина [94]. Ось *x* направлена в сторону большой бороздки (Рис. 8). Определение шести геометрических параметров приведено на Рис. 9.

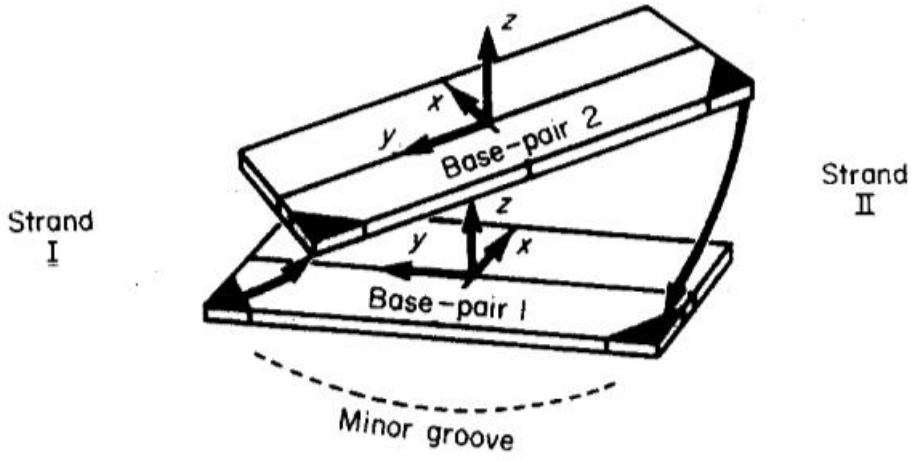


Рис. 8. Система координат, введенная для описания пространственного расположения оснований в динуклеотиде [94].

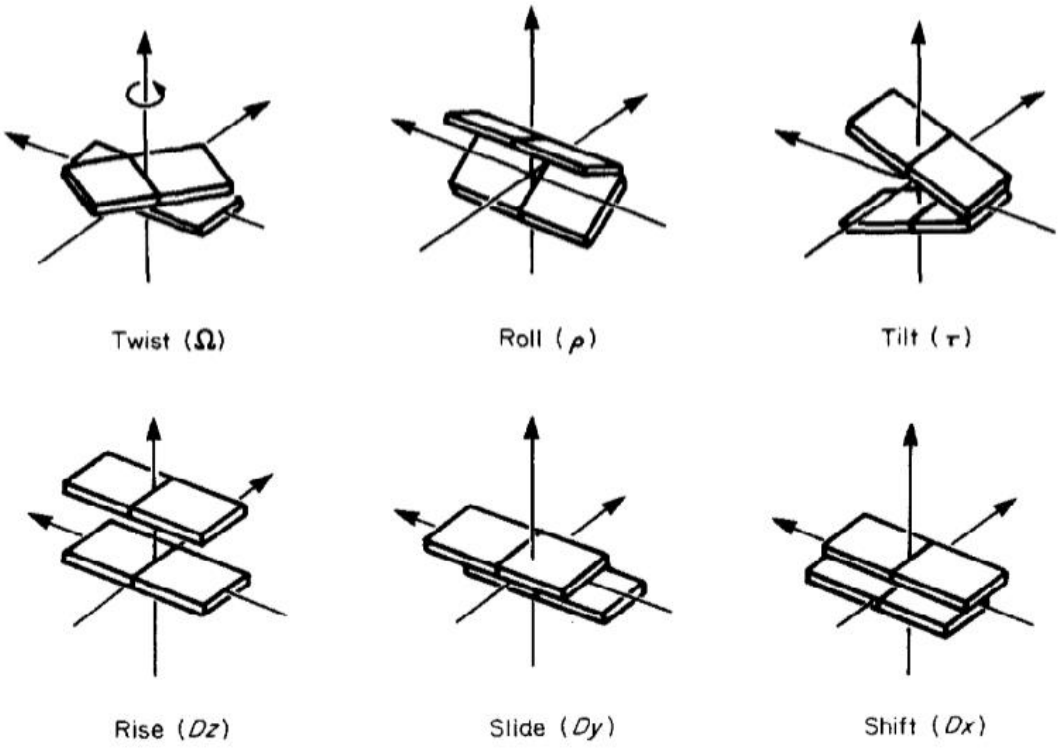


Рис. 9. Геометрические характеристики, описывающие взаимное расположение оснований в динуклеотиде.

Локальная (на уровне динуклеотидов) способность молекулы ДНК к изгибу является важнейшим фактором для правильного позиционирования белка. Так, например, белки CAP (catabolite activator protein), IHF (integration host factor), TBP (TATA-binding protein) не могут связаться с ДНК нужным образом, если подпоследовательность динуклеотидов, служившая шарниром, заменена на жесткий негнущийся А-тракт. Стоит отметить, что ДНК не находится постоянно в изогнутой форме, нужной белку, а принимает ее под его воздействием. Было показано, что, помимо уже упомянутых CAP, IHF, TBP, существенные изгибы происходят при связывании с Lac operator, PurR, gd-resolvase [86]. Меньшие деформации производят helix–turn–helix (HTH) белков прокариот – lambda, 434 repressors, и EcoRV репрессор. В работе [86] были проанализированы 86 комплексов ДНК-белок. Изгибы молекулы, возникающие при связывании, были описаны с помощью параметров roll, tilt, twist, slide, shifrt, rise и разделены на несколько типов. Таким образом, для успешного взаимодействия с белком молекула ДНК должна обладать определенными геометрическими параметрами. Локальный изгиб важен также для предсказания стабильности нуклеосомы [93, 95].

Деформируемость двойной спирали ДНК имеет решающее значение для ее упаковки в клетке, распознавания другими молекулами и временного разъединения цепи в ходе различных биохимических процессов. В работе [88] приводятся функции энергии, зависящие от нуклеотидной последовательности, способные описывать деформируемость молекулы ДНК. Функции были построены на основе знаний о флуктуациях и корреляциях структурных динуклеотидных параметров комплексов ДНК с белками. Они позволяют описывать локальные движения ДНК в процессе взаимодействия с белком – например, при распознавании или при деформации под действием белка. В частности, димеры пиримидин-пурина выделяются как наиболее изменчивые динуклеотиды в комплексах ДНК-белок, которые, по-видимому, действуют как гибкие «шарниры», приспосабливающие дуплекс к поверхности белка. Наблюдаемые корреляции изгибов и сдвигов нуклеотидов друг относительно друга важны для правильной укладки ДНК в нуклеосомах и других нуклеопротеиновых комплексах. Функции энергии также позволяют строить реалистичные трехмерные модели для изучения длинных участков ДНК на глобальном уровне, включающие структурные особенности, ускользающие из рассмотрения при использовании модели упругого стержня.

В дополнение к генетическому сообщению последовательность ДНК несет множество других сигналов, связанных с манипуляцией длинной нитевидной молекулой. Первичные последовательности оснований нуклеиновых кислот определяют трехмерные структуры, физические свойства которых отражают составляющие остатки. При рассмотрении кристаллических структур ДНК [96] обнаруживаются тонкие, зависящие от последовательности, неоднородности в ориентации и смещении смежных остатков [87]. Устойчивость дуплекса при заданных условиях окружающей среды также зависит от набора динуклеотидов, расположенных в непосредственной близости друг к другу [97]. Таким образом, линейная последовательность генетической информации представляет собой некий пространственный и энергетический код, который управляет глобальной организацией двойной спирали и ее восприимчивостью к взаимодействию с другими молекулами.

Интерес к пониманию физических свойств геномной ДНК вызвал развитие новых подходов к анализу и описанию зависящих от последовательности изгибов и скручиваний соседних пар оснований. Исследования миграции в геле [98], кинетики циклизации [98, 99] и фазирования нуклеосом [100] дали множество зависимых от последовательности моделей для описания наблюдаемых данных. Кроме того, собранные кристаллографические структуры олигонуклеотидов показывают сходные конформационные тенденции [87], хотя

для нескольких димеров существуют расхождения в направлениях изгибов, полученных из данных рентгеновских исследований и построенных моделей [101].

Тепловые флуктуации, данные о которых получены из исследований растворов длинной ДНК [102], представляют собой еще одну важную особенность двойной спирали. Эти макроскопические свойства, которые нелегко интерпретировать на уровне динуклеотидов, обычно описываются изменениями упругих констант идеализированных моделей, не зависящих от последовательности и соответствующих наблюдаемым данным.

Однако специфическое распознавание ДНК белками и другими лигандами требует более сложного описания гибкости цепей, чем может быть получено на основе модели гибкого стержня. В частности, ДНК реагирует на присутствие белка через зависящие от последовательности изгиб и интеркаляцию [103]. Эта деформируемость имеет важное значение как на глобальном, так и на местном уровнях, служит в качестве потенциального сигнала для молекулярного распознавания на расстоянии, а также для обеспечения локальных искажений двойной спирали, вызванных связыванием с белком, т. е. так называемого конформационного распознавания (indirect readout of DNA residues) [104]. Для анализа таких систем требуются модели, которые подробно описывают стереохимические особенности последовательностей ДНК. Одним из лучших источников этой информации является база данных ДНК-белковых комплексов [96]. В настоящее время она содержит достаточно данных для получения дуплексных свойств (средств и колебаний) на уровне ближайших соседей, т. е. 10 динуклеотидов. Хотя тримеры и тетрамеры могут оказаться еще более полезными, кристаллографических данных еще недостаточно для исследования деформируемости ДНК на этом уровне. Еще один недостаток имеющихся данных состоит в том, что «реальная» гибкость ДНК в любом заданном комплексе не может быть определена из-за недостатка данных. Тем не менее, можно использовать усредненные по ансамблю параметры, характеризующие вариабельность ДНК в этих комплексах, чтобы получить представление о движениях двойной спирали.

Этот подход можно обосновать следующим образом. Различные белки прилагают различные силы к молекуле ДНК, поэтому искажения ДНК, вызванные многими видами белков, эффективно отменяют друг друга. Таким образом, усреднение по большому числу комплексов позволит определить конформационный отклик ДНК. Однако эти общие принципы не обязательно будут предсказывать взаимодействие ДНК с аминокислотами в конкретных комплексах [88].

Средняя степень скручивания пар оснований в динуклеотиде увеличивается в одном и том же порядке в трёх стандартных химических классах: пиримидин-пурин, пурин-пурин, пурин-пиримидин. Например, среди пиримидин-пуринов порядок - Twist (AT) <Twist (AC)

<Twist (GC) характерен и для комплексов белок-ДНК, и для свободной B-ДНК (Табл. 1) Дисперсия параметров динуклеотидов обычно больше в образце, связанном с белком, чем в свободной B-ДНК. Отличия в значениях дисперсий для разных динуклеотидов как раз и являются показателем того, что способность молекулы к деформации зависит от нуклеотидной последовательности [88].

Step	Twist, deg		Tilt, deg		Roll, deg		Shift, Å		Slide, Å		Rise, Å	
CG	36.1	(5.5)	0	(4.2)	5.4	(5.2)	0	(0.87)	0.41	(0.56)	3.39	(0.27)
CA	37.3	(6.5)	0.5	(3.7)	4.7	(5.1)	0.09	(0.55)	0.53	(0.89)	3.33	(0.26)
TA	37.8	(5.5)	0	(2.7)	3.3	(6.6)	0	(0.52)	0.05	(0.71)	3.42	(0.24)
AG	31.9	(4.5)	-1.7	(3.3)	4.5	(3.4)	0.09	(0.69)	-0.25	(0.41)	3.34	(0.23)
GG	32.9	(5.2)	-0.1	(3.7)	3.6	(4.5)	0.05	(0.76)	-0.22	(0.64)	3.42	(0.24)
AA	35.1	(3.9)	-1.4	(3.3)	0.7	(5.4)	-0.03	(0.57)	-0.08	(0.45)	3.27	(0.22)
GA	36.3	(4.4)	-1.5	(3.8)	1.9	(5.3)	-0.28	(0.46)	0.09	(0.70)	3.37	(0.26)
AT	29.3	(4.5)	0	(2.5)	1.1	(4.9)	0	(0.57)	-0.59	(0.31)	3.31	(0.21)
AC	31.5	(4.2)	-0.1	(3.1)	0.7	(3.9)	0.13	(0.59)	-0.58	(0.41)	3.36	(0.23)
GC	33.6	(4.7)	0	(3.9)	0.3	(4.6)	0	(0.61)	-0.38	(0.56)	3.40	(0.24)

Табл. 1. Средние значения и дисперсии динуклеотидных геометрических параметров ДНК [88].

Пиримидин-пурины (YR) являются наиболее легко деформируемыми динуклеотидами [88]. Эта тенденция согласуется с эмпирическими предсказаниями энергии [105] и интерпретациями подвижности в геле [106]. По-видимому, повышенная «податливость» динуклеотидов YR воздействию белка связана с относительно слабыми стэкинг-взаимодействиями [107]. Примечательно, что резкие изгибы ДНК, наблюдаемые в комплексах белок-ДНК и ЯМР-комплексах, происходят почти исключительно на динуклеотидах YR [86]. Они, по-видимому, действуют как гибкие «шарниры», приспосабливающие дуплекс к поверхности белка. Неслучайное позиционирование динуклеотидов YR в сайтах связывания с белками также облегчает формирование петли ДНК [108], которая в свою очередь может служить в качестве сигнала для белка. Периодичность появления гибких участков (участков, обогащенных динуклеотидами YR) может быть использована для предсказания сайтов связывания белка в геномах.

Повышенная деформируемость динуклеотидов YR в комплексах белка-ДНК по сравнению со свободной ДНК потенциально может влиять на энтропию комплексообразования. Если поверхность соприкосновения белка и ДНК не препятствует движению динуклеотидов в области контакта, размещение гибких «шарниров» YR может быть выгодным с точки зрения энтропии [88]. Предполагаемая энтропийная роль гибких

«шарниров» YR, тем не менее, заслуживает дальнейшего исследования и может быть связана с опубликованными калориметрическими и ЯМР свойствами комплексов нуклеопротеинов [109] .

Корреляции между параметрами являются следствием сложной связи между двумя различными видами взаимодействий в ДНК: стэкинг-взаимодействия, зависящего от последовательности (которое в первую очередь ответственно за направление изгиба и сдвига динуклеотида), и, в значительной степени, независимого от последовательности взаимодействия, стабилизирующего сахарофосфатный остов (который имеет тенденцию сохранять свою оптимальную конформацию и, тем самым, служит эластичной струной, управляющей «муфтами» между тремя угловыми и тремя поступательными параметрами). Конкретные способы выявления этих взаимодействий на локальном уровне, очевидно, зависят от так называемой «морфологии» оснований и, следовательно, от базовой последовательности [87]. Среди упомянутых корреляций есть несколько, которые кажутся «почти универсальными» и, следовательно, применимы к различным последовательностям. К ним относятся корреляции Twist-Roll, Tilt-Shift, Roll-Slide, и Twist-Slide [88].

Взаимозависимость между изгибом ДНК и боковыми смещениями пар оснований, вероятно, имеет функциональные последствия. Когда ДНК изогнута в плотной петле, например, в нуклеосоме или в комплексе инициации транскрипции, такие перемещения динуклеотидов регулируют закрученность суперспирали. Подробные исследования геометрии ДНК [105, 110] показывают, что для образования левой суперспирали ось ДНК должна быть направлена таким образом, чтобы Slide был положительным, когда ДНК изгибается в сторону малой бороздки, и отрицательным, когда дуплекс сгибается в сторону большой бороздки.

В дополнение к связи между скручиванием ДНК и образованием суперспирали, как Slide, так и Shift регулируют пространственную траекторию ДНК. Поскольку изменчивость Roll и Slide больше, чем у Tilt и Shift, здесь рассматриваются только первые. Смещения оси дуплекса благодаря изменению параметра Shift также может играть важную роль в упаковке ДНК в хроматине [88].

В 80х годах прошлого столетия предполагалось, что по средним значениям динуклеотидных характеристик можно определить форму участка ДНК (A или B) [85, 111]. Первая кристаллическая структура B-формы ДНК была получена в 1981 [112]. Дисперсия динуклеотидных параметров для полученной структуры оказалась весьма велика. Среднее значение roll составило 0.5 ± 5.2°, twist - 35.6 ± 4.4° и slide - 0.2 ± 0.5 Å. Для A-формы ДНК наблюдались значения: roll - 6.8 ± 2.6°, twist - 30.8 ± 1.2° и slide - -1.5 ± 0.4 Å [85]. Позднее, с развитием методов кристаллографии и появлением все большего числа расшифрованных

структур, стало понятно, что однозначно отнести участок ДНК к той или иной форме, используя параметры roll и twist, нельзя, так как они изменяются в широком диапазоне значений для обеих форм ДНК [113]. Однако значения slide хорошо согласуются с формой [114]. При исследовании набора РНК олигомеров были получены следующие величины, хорошо согласующиеся с таковыми для А-формы ДНК: slide - -1.5 ± 0.4 Å, roll - 9.0 ± 4.0° и twist - 31.1 ± 3.4° [114, 115].

На протяжении нескольких десятилетий предпринимаются попытки охарактеризовать динуклеотиды по степени их предрасположенности к той или иной форме ДНК [116, 117]. Под предрасположенностью подразумевается их способность индуцировать переход от одной формы к другой в растворе. Так, например, динуклеотиды AA и GG сильно тяготеют к B- и A-формам соответственно. GA и AG преимущественно предрасположены к B-форме. Динуклеотиды AC, AT, GC, TA и CG не демонстрируют ярко выраженной предрасположенности [114].

Интересно, что в большинстве исследованных случаев участки ДНК, связанные с белком, принимают B-форму. Лишь в небольшой доле были обнаружены необычные (выходящие за пределы 3σ) значения динуклеотидных характеристик. В остальных же случаях динуклеотидные параметры были близки к величинам, характерным для свободной B-ДНК [114]. Для нуклеосомы, являющейся классическим примером сильно изогнутой структуры [118], было найдено всего 5 из 146 значений roll, выходящих за пределы 3σ. Отсутствие острых перегибов приводит к тому, что структура нуклеосом имеет плавную кривизну. Рассмотрев наборы структур свободной и связанной ДНК, авторы пришли к выводу, что на глобальном структурном уровне несколько связанных с белком ДНК-дуплексов, как предполагается, принимают изогнутую конформацию при отсутствии каких-либо больших искажений, что указывает на то, что ряд нормальных значений структурных параметров на уровне динуклеотида может вызывать кривизну на глобальном уровне [114].

Изучение геометрии динуклеотидных пар, а также ее изменений под действием различных факторов (например, связывании с белком) является одним из необходимых шагов на пути к пониманию общей структурной организации молекулы РНК. Локальная способность к деформации молекулы РНК изучена намного хуже, чем в случае с ДНК. Точно так же, как и для ДНК, гибкость молекулы РНК может определять ее взаимодействия с белками и другими молекулами. Несколько работ было посвящено изучению структурных изменений белка при взаимодействии с РНК, механизмам распознавания и описанию сайта связывания. В последнее время появляется все больше работ, посвященных конформационным особенностям РНК на уровне динуклеотидов и их изменений при

взаимодействий с белками. Так, в работе [119] был проведен анализ конформационных предпочтений динуклеотидов свободной РНК и РНК, связанной с белком. Были обработаны рентгеновские изображения структур 88 молекул РНК и 127 комплексов РНК-белок. Для каждого динуклеотида были посчитаны средние значения геометрических параметров. По нескольким из них наблюдались статистически значимые отличия между свободными и связанными РНК молекулами. Кроме того, было проведено сравнение РНК молекул с A- и B-формами двойной спирали ДНК. Корреляционные зависимости динуклеотидных геометрических параметров РНК молекул оказались схожими с A-формой двойной спирали ДНК и, в свою очередь, отличными от B-формы двойной спирали ДНК. Динуклеотиды из комплексов РНК-белок были разделены на две группы – те, которые непосредственно контактируют с белком, и те, которые не имеют прямого контакта. Оказалось, что геометрические параметры динуклеотидов обеих групп статистически значимо отличаются от аналогичных параметров свободной молекулы РНК. Авторы высказали предположение, что, помимо изменения локальной конформации в области непосредственного контакта с белком, имеет место и аллостерический эффект для динуклеотидов, не взаимодействующих с белком напрямую.

В последнее время активно изучается возможность применения моделей машинного обучения для предсказания роли данной нуклеотидной последовательности в том или ином клеточном процессе. Зачастую проведение эксперимента требует больших денежных и временных затрат. С другой стороны, благодаря стремительному развитию технологий секвенирования, к настоящему времени уже удалось накопить огромное количество геномных последовательностей. Они могут быть использованы для построения моделей, позволяющих определять роль заданной последовательности в функционировании клетки. Например, в работе [120] авторы определяют так называемые горячие и холодные точки рекомбинации – области генома, в которых рекомбинация происходит соответственно с повышенной и пониженной вероятностью. Построенная в исследовании модель использует для предсказания геометрические динуклеотидные параметры. В качестве классификатора был выбран метод опорных векторов, принимающий на вход вектор динуклеотидных характеристик последовательности. Точность модели составила более 82 %. В исследовании [121] удалось идентифицировать сайты начала репликации по динуклеотидным геометрическим характеристикам с точностью 98%. В качестве классификатора был использован случайный лес. Здесь необходимо подчеркнуть физические основания, послужившие базисом для такого рода методов. Большинство манипуляций с молекулами ДНК и РНК осуществляется белками, а значит запуск того или иного процесса непосредственно зависит от распознавания нужного сайта белком. Выше

было описано значение локальных физических характеристик в этом процессе. Можно предположить, что построенные модели в каком-то смысле имитируют процесс распознавания белком – они пытаются отделить искомые последовательности от всех остальных путем сравнения их физических характеристик. В качестве еще одного интересного примера можно привести определение сайтов сплайсинга по геометрическим динуклеотидным параметрам [122]. Сплайсинг осуществляется комплексом из более чем 300 белков и нескольких ядерных РНК молекул, называемым сплайсомой. Он распознает сайты сплайсинга, а затем удаляет их в ходе двух последовательных реакций переэтерификации. Для построения модели был использован метод опорных векторов, ее точность составила более 85 % [122]. В работе [123] авторам удалось разработать алгоритм распознавания сайтов редактирования РНК (модификация нуклеозида аденозина в инозин). Редактирование РНК осуществляет специальный белок ADAR (Adenosine deaminases acting on RNA). В наиболее изученной в контексте редактирования мРНК глутаматного рецептора сайт связывания с белком находится во вторичной структуре типа стебель-петля длиной в 71 нуклеотид. Данная структура важна для взаимодействия, так как при ее разрушении или изменении формы посредством ввода соответствующих мутаций связывания не происходит [124]. Для классификации был применен метод опорных векторов, где точность модели превысила 79% [123]. В исследовании была показана возможность определения piРНК среди других клеточных РНК по динуклеотидным физическим характеристикам. piРНК играют важную роль во многих клеточных процессах, например, они участвуют в подавлении активности транспозонов, регуляции генной экспрессии и трансляции некоторых белков. [125]

Заключение к Главе 1

Резюмируя все выше сказанное, следует отметить, что существует немало открытых вопросов, касающихся механизма транспозиции. Одним из важнейших из них является вопрос о способе распознавания белком LINE своего РНК-транскрипта или РНК-транскрипта SINE. В настоящее время выявлено несколько организмов, в геноме которых структура стебель-петля, расположенная на 3'-конце LINE и SINE транспозонов, необходима для успешной ретротранспозиции. Так, экспериментально было показано, что введение мутаций, разрушающих данную структуру, приводило к прекращению транспозиции или резкому снижению ее эффективности для LINE транспозонов семейства SART1 тутового шелкопряда, LINE транспозонов семейства UnaL2 в геноме угря, LINE транспозонов ZfL2-1 и ZfL2-2 в геноме рыбы *Danio rerio*, SINE транспозонов семейства SmaI в геноме лосося, LINE R2 у насекомых. Наличие структуры стебель-петля на 3`-конце LINE и SINE транспозонов, имеющих одинаковую последовательность нетранслируемой 3`-области, было продемонстрировано для значительно более широкого круга видов и семейств транспозонов, включая L1 в водорослях и однодольных растениях, Tad1 в грибах, L2 у рыб, RTE у млекопитающих. Однако до сих пор остается неясным, является ли принцип распознавания РНК-транскрипта по данной структуре глобальным, или он справедлив только для некотрой группы транспозонов.

Приведенный выше обзор литературы показывает, что физические характеристики последовательности, в том числе и свойства, играющие важную роль при взаимодействии с белком, могут быть рассмотрены на уровне динуклеотидов. Такой подход очень часто применяют для исследования участков ДНК. Для изучения РНК-молекул его стали использовать сравнительно недавно. Физической основой позиционирования белка является зависящая от последовательности склонность ДНК и РНК к принятию той или иной изогнутой конфигурации. Рассмотрение деформаций изгиба, скручивания, а также боковых смещений смежных пар оснований позволяет выявить структурные характеристики, играющие существенную роль во взаимодействии с белком, а, следовательно, определяющие способность ДНК или РНК принимать участие в том или ином клеточном процессе. Анализ литературы показал возможность использования математических моделей для задач классификации последовательностей и предсказания их функции по физичесиким характеристикам. Математические модели на основе динуклеотидных физических характеристик были успешно применены для определения точек рекомбинации, сайтов сплайсинга, и участков, подвергающихся РНК редактированию.

Глава 2. Консервативная вторичная структура стебель-петля на конце РНК-транскриптов транспозонов и ее роль в ретротранспозиции

Экспериментальные свидетельства необходимости структуры стебель-петля на 3`-конце LINE и SINE транспозонов для успешной транспозиции в геномах тутового шелкопряда, угря, рыбы полосатый данио, лосося подтолкнули к идее о наличии подобной структры в других организмах и о ее возможной роли в качестве «опознавательного знака» для белка, синтезирующего ДНК последовательность по РНК шаблону и вставляющего ее в геном. Не обнаруживаемое на уровне первичной последовательности, сходство может проявляться при переходе на следующий уровень – уровень вторичной структуры ДНК/РНК. Нуклеотидная последовательность 3’-конца транспозонов может различаться, однако структура, образующаяся на данном участке, сохраняется и служит для распознавания РНК-транскрипта белком. Кроме того, необходимо также сохранение позиции структуры относительно длины транспозона. Обоснование высказанных выше предположений потребовало анализа большого числа нуклеотидных последовательностей на наличие консервативной вторичной структуры на самом конце нетранслируемой 3’-области транспозонов. Для решения этой задачи был разработан программный комплекс, позволивший обнаружить такие структуры в L1 и ALu транспозонах человека. Далее был проведен расширенный анализ LINE и SINE транспозонов из 161 организма, показавший наличие консервативной структуры на 3’-конце.

2.1. Метод поиска консервативных вторичных структур РНК/ДНК в геноме

Для данного исследования были отобраны 6622 L1 транспозона при помощи USCS Genome Brouser, имеющих полную длину и не имеющих вставок, делеций и других масштабных изменений последовательности нуклеотидов. Был сформирован набор из 39 L1 транспозонов, активность которых была показана экспериментально. Шесть транспозонов из данного набора обладают активностью, в 10 раз превышающей активность остальных 33 транспозонов [26]. Также было отобрано 401 242 Alu транспозона человека.

Структура стебель-петля является палиндромной структурой. Стебель (stem) образует двойную спираль, сформированную полностью или частично комплементарными парами оснований. Петля (loop), напротив, состоит из неспаренных нуклеотидов (Рис. 10).

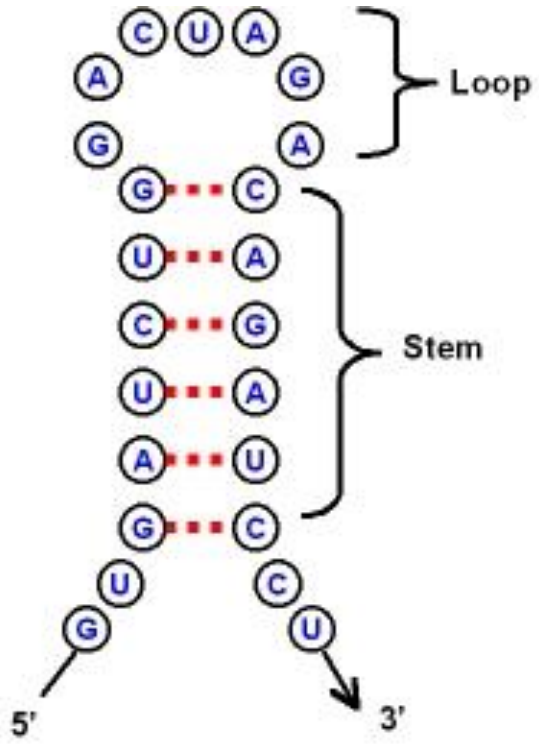


Рис. 10. Схема структуры стебель-петля. Рисунок взят с сайта http://www.mikeblaber.org/oldwine/BCH4053/Lecture21/Lecture21.htm.

Программный комплекс позволяет варьировать такие параметры искомых палиндромных структур, как минимальная длина стебля (*minStemLen*), максимальная длина стебля (*maxStemLen*), максимальная длина петли (*maxLoopLen*) и максимальное число некомплементарных нуклеотидов в стебле (*maxMismatch*). Задача поиска палиндромных структур может быть сформулирована таким образом: найти наибольшие комплементарные подстроки с длиной в промежутке от *minStemLen* до *maxStemLen* и содержащие не более *maxMismatch* некомплементарных нуклеотидов, расстояние между которыми меньше *maxLoopLen.*

Для решения задачи был применен метод динамического программирования. Основой для предложенного алгоритма является алгоритм Нидлмана-Вунша для выравнивания последовательностей. Однако вместо поиска выравнивания с наибольшей оценкой определяется выравнивание с минимальным штрафным баллом за несовпадения (некомплементарные нуклеотиды) и разрывы (на одной из цепей отсутствует нуклеотид). Для того, чтобы работать с положительными числами, балл не добавлялся, если нуклеотиды комплементарны, и балл добавлялся, если нуклеотиды некомплементарны или встретился разрыв.

Таким образом матрица M размеров [(*minStemLen* +1) × (*minStemLen* +1)] строится по следующему алгоритму:

Для 2≤i≤ *minStemLen* +1, 2≤j≤ *minStemLen* +1:

$$M(i,j) = \min \begin{cases} M(i-1,j-1)+S(i,j) \text{ (mismatch or match)} \\ M(i,j-1)+w \text{ (gap in sequence 1)} \\ M(i-1,j)+w \text{ (gap in sequence 2)} \end{cases}$$

, где S(i,j) = 0 (совпадение), S(i,j) = 1 (несовпадение), w=1 (разрыв).

Совпадение означает наличие комплементарных нуклеотидов на данной позиции, несовпадение соответствует некомплементарным нуклеотидам. Значения в первой строке и в первом столбце соответствуют координатам в подстроке. Первый столбец содержит позиции нуклеотидов в первой подстроке, а первая строка матрицы заполнена позициями нуклеотидов во второй подстроке.

Например, для двух подстрок TACG и AATGC исходный вид матрицы M такой (Рис. 11):

		A	**A**	**T**	**G**	**C**
	0	1	2	3	4	5
T	1					
A	2					
C	3					
G	4					

Рис. 11. Исходный вид матрицы M.

Затем матрица постепенно заполняется значениями в соответствии с представленной выше схемой. Например, M[2,2]=min(M[1,1]+0, M[1,2]+1,M[2,1]+1)=min(0,2,2)=0 (Рис. 12).

		A	C	T	G	C
	0	1	2	3	4	5
T	1	**0 (0+0,1+1,1+1)**	1 (1+1,0+1, 2+1)	2 (2+1,1+1,3+1)	3 (3+1,2+1,4+1)	4 (4+1,3+1,5+1)
A	2	1 (1+1,2+1,0+1)	1	1	2	3
C	3	2 (2+1,3+1,1+0)	2	1	2	3
G	4	3 (3+1,4+1,2+1)	3	2	2	2

Рис. 12. Матрица, заполненная в соответствии со схемой, представленной выше.

После заполнения матрицы программа начинает поиск максимальных координат, для которых соответствующее значение в ячейке матрицы M не превышает *maxMismatch.* Полученные координаты должны удовлетворять условию – длина соответствующей им подстроки превышает minStemLen. Например, пусть *maxMismatch* =2 и *minStemLen* =2, тогда программа найдет максимальные координаты, для которых значение в матрице M не превышает 2. В данном примере это нижняя правая ячейка. Таким образом, удалось получить максимальную оценку, удовлетворяющую критериям поиска среди всех возможных выравниваний. Восстановление самого выравнивания, давшего полученную оценку, производится аналогично тому, как это делается в алгоритме Нидлмана-Вунша. Разница состоит лишь в том, что мы будем использовать минимальные оценки (значания в

матрице M). В приведенном выше примере была найдена структура, представленная на Рис. 13.

```
ACTGC
|-|||
T_ACG
```

Рис. 13. Пример выравнивания.

Программа *DNA Punctuation,* в которой реализован описанный алгоритм, написана на языке C, доступна на сайте www.dnapunctuation.org.

Структуры стебель-петля были аннотированы при помощи программы *DNA Punctuation,* описанной выше. Программа осуществляла поиск структур с длиной стебля от 15 до 50 пар оснований, длиной петли до 10 пар оснований и числом несовпадений или разрывов, не превышающим 7 нуклеотидов. Был построен профиль покрытия транспозона структурами стебель-петля. По оси ординат откладывается 1, если нуклеотид входит в структуру, и 0, если нуклеотид не принадлежит структуре стебель-петля. Так как транспозоны имеют разную длину, по оси абсцисс длина была выражена в процентах, что дало возможность сравнивать относительные позиции структур стебель-петля внутри транспозонов.

Двумерные вторичные модели структуры стебель-петля были сконструированы и визуализированы при помощи пакета программ Vienna RNA Package 2.1.9.

Структура разделяется на составляющие (участки стебля, имеющие только спаренные нуклеотиды, пузырьки, петли, одиночные неспаренные нуклеотиды, и т.д.), для каждой из которых определяется энергия Гиббса (Рис. 14). Для ее подсчета используются энергии Гиббса образования динуклеотидов, а также эмпирически полученные расчетные формулы для определения энергий неспаренных участков. Энергия Гиббса всей структуры вычисляется путем суммирования энергий ее составляющих [126]. Полученные модели были использованы для проведения первичного, так называемого «разведочного», анализа, призванного выявить наличие или отсутствие общих закономерностей для большого числа структур. Трехмерная структура и ее физические характеристики будут рассмотрены в Главе 3. В данной части работы были проанализированы структуры из 39 активных в настоящее время в геноме человека L1 транспозонов, структуры из 6622 хорошо сохранившихся L1 транспозонов, а также структуры из 401 242 транспозонов Alu.

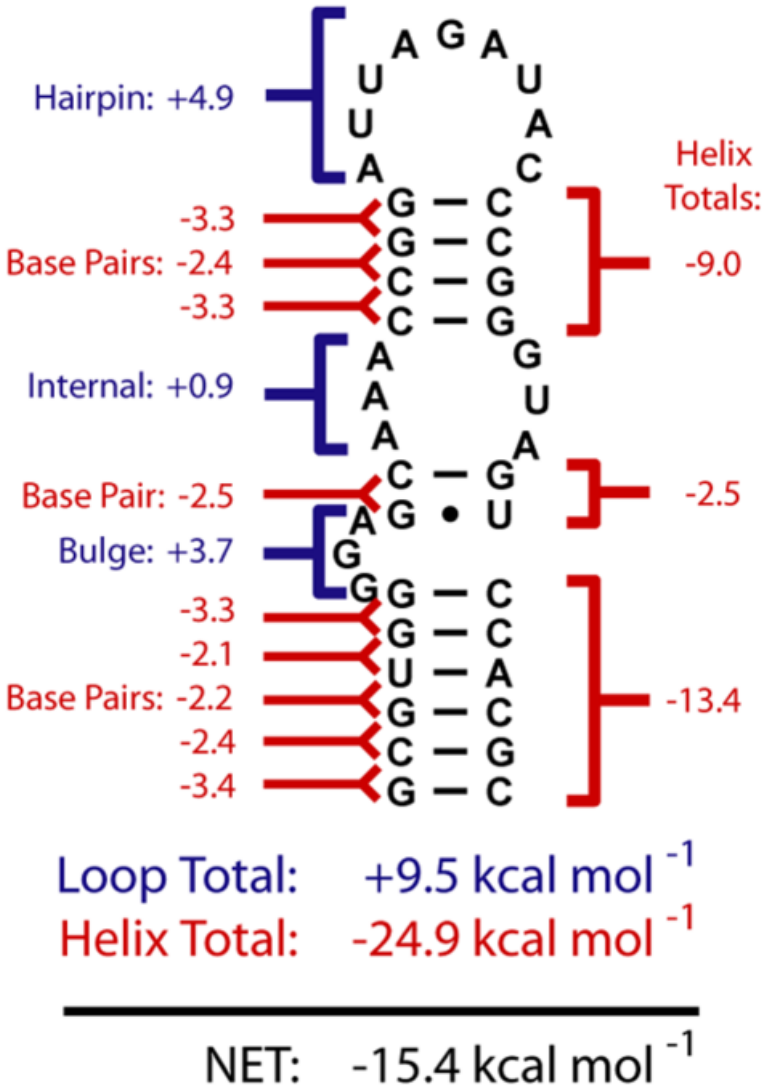


Рис. 14. Иллюстрация процесса разделения структуры на составляющие и подсчета суммарной энергии. Рисунок взят с сайта http://rna.urmc.rochester.edu/NNDB/help.html.

Для анализа такого большого набора структур был предложен способ, позволяющий выявлять консервативные позиции важных структурных особенностей, например, пузырьков и петель. По аналогии с профилями покрытия были построены профили спаренных/неспаренных нуклеотидов. В случае, если нуклеотид не спарен, добавляется 1, в случае, если спарен – 0. Полученный профиль отражает долю структур, имеющих неспаренную пару оснований на данной позиции. Были построены два типа профилей. В первом типе длина, отложенная по оси *x*, была выражена в процентах, а во втором – *x* координата соответствовала позиции, которая являлась числом нуклеотидов в направлении от петли ($x = 0$) к концу стебля.

2.2. Анализ транспозонов семейства L1 в геноме человека

С целью обнаружения консервативных вторичных структур стебель-петля был проведен анализ нуклеотидных последовательностей транспозонов семейств L1 и Alu в геноме человека. Под консервативными подразумеваются такие структуры, позиции которых относительно длины транспозона сохраняются. Для анализа были взяты последовательности активных в настоящее время транспозонов L1 (источником послужила публикация [26]), 6622 хорошо сохранившихся транспозонов L1, разделенных на 27 подсемейств в соответствии с классификацией, приведенной в статье [79], а также 401 242 Alu транспозона, разделенных на 213 семейств, как предложено в публикации [80]. Все последовательности транспозонов были аннотированы структурами стебель-петля. Полученная аннотация была использована для построения профилей покрытия вторичными структурами. Для различных наборов последоваетльностей пики на профилях покрытия соответствуют позициям консервативных структур стебель-петля (Рис. 15). На Рис. 15 A изображен профиль покрытия для 6 самых активных в настоящее время L1 транспозонов, взятых из исследования [26]. На Рис. 15 B-E представлены три самых репрезентативных профиля (соответствующих подсемействам L1PA, L1PB и L1MA) из 27 профилей подсемейств L1. Консервативная позиция структуры стебель-петля для наиболее молодого подсемейства L1PA1 совпадает с позицией структуры для 6 самых активных транспозонов. Видно, что степень консервативности уменьшается с возрастом подсемейства. Это происходит из-за вставок, делеций, усечений нетранслируемой 5’-области, что приводит к изменению длины транспозона, а также из-за случайных мутаций, разрушающих структуру стебель-петля. Профили покрытия для всех 27 подсемейств L1 транспозона приведены в конце диссертации (Приложение 1). Тенденция к «размыванию» позиции и уменьшению степени ее консервативности ясно прослеживается в направлении от молодых к более древним семействам и хорошо согласуется с представлением о вертикальном характере эволюции L1 транспозонов. Профили покрытия для 6 самых активных L1 транспозонов и для расширенного набора из 33 достаточно активных L1 транспозонов практически совпадают. Они также приведены в конце диссертации (Приложение 2).

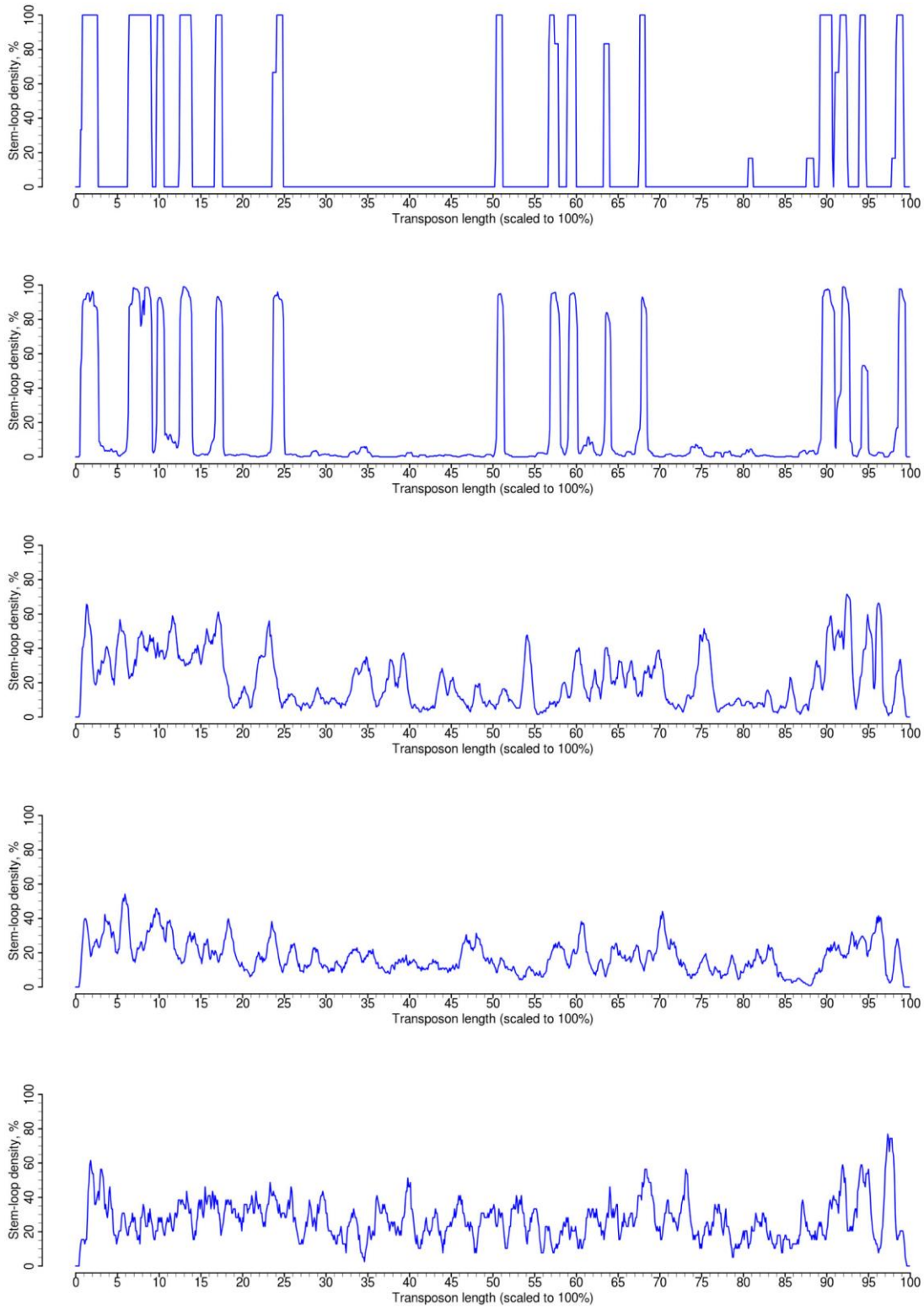


Рис. 15. Профили покрытия вторичными структурами стебель-петля. A) Профиль для 6 самых активных транспозонов; B) Профиль для транспозонов, принадлежащих самому молодому подсемейству L1 – L1PA1; C) Профиль для транспозонов, принадлежащих среднему по возрасту подсемейству L1 – L1PA8; D) Профиль для транспозонов, принадлежащих древнему подсемейству L1PB1; E) Профиль для транспозонов, принадлежащих древнему подсемейству L1MA1.

Построенные профили покрытия позволили выявить наличие консервативных структур стебель-петля по всей длине транспозонов. А именно, удалось обнаружить подобные структуры в нетранслируемой 5’-области, во второй рамке считывания (ORF2) и, что наиболее важно, в нетранслируемой 3’-области. В ходе анализа профилей покрытия активных в настоящее время L1 транспозонов, а также транспозонов, принадлежащих самому молодому подсемейству L1PA1, были выделены участки, обогащенные вторичными структурами – три характерных участка в нетранслируемой 5’-области (5′UTR-1, 5′UTR-2 и 5′UTR-3), два в конце второй рамки считывания ORF2 (ORF2-1 и ORF2-2), и участок в нетранслируемой 3’-области. Транскрипция L1 транспозона начинается с промоутера, расположенного в нетранслируемой 5’-области. Эта область занимает 1000 – 1200 пар оснований, ее последовательность не является консервативной. Участок длиною в 100 нуклеотидов, расположенный в самом начале 5’ области, обладает промотерной активностью [29]. Сайт связывания транскрипционного фактора YY1 занимает с 3го по 26й нуклеотид [127]. На позициях 472–477 и 572–577 пар оснований находится сайт связывания транскрипционного фактора семейства SRY [128], а с 81 по 101 пару оснований располагается сайт, способный связывать RUNX транскрипционные факторы [129].

В нетранслируемой 5’-области активных L1 транспозонов были обнаружены три консервативных структуры стебель-петля. Позиция первой структуры – 50-100 пар оснований. Она расположена сразу после сайта связывания YY1 транскрипционного фактора. Две другие структуры (423-461 пар оснований, 534-567 пар оснований) находятся непосредственно перед сайтами связывания соответственно SRY и RUNX транскрипционных факторов. Важная роль структур стебель-петля, расположенных в промоутерных областях, описывается в работах [130, 131].

Вторичные структуры стебель-петля, обнаруженные в ORF2 кодирующей области, могут играть роль при образовании нуклеопротеинового комплекса, в составе которого перемещается L1 транскрипт. Пока очень мало известно о структуре мРНК in vivo. Последние исследования выявили наличие большого числа двухцепочечных участков в молекулах мРНК [132].

Несмотря на то, что филогенетический анализ последовательностей открытых рамок считывания (ORF) L1 транспозонов указывает на вертикальный характер эволюции, это утверждение не является верным для некодирующих участков транспозонов. Между 3’ и 5’ нетранслируемыми областями не прослеживается сходства на уровне последовательностей различных подсемейств. Известно, что L1 транспозоны много раз меняли нетранслируемые

5’-области в ходе эволюции [79]. Для проведения анализа нетранслируемые 5’-области были объединены в группы, предложенные в [79]. Для каждой группы был построен профиль (Приложение 3). Наиболее молодая группа включала подсемейства L1PA1-8, профиль покрытия выявил четыре консервативные позиции в нетранслируемой 5’-области данной группы.

Был проанализирован эффект мутаций в кодирующих областях (рамки считывания ORF1 и ORF2) на структуры стебель-петля. Профили покрытия транспозонов L1 с неповрежденными и подвергшимися мутациям открытыми рамками считывания приведены в конце диссертации (Приложение 4). По меньшей мере половина мутаций, произошедших в открытых рамках считывания, затрагивает структуры стебель-петля, расположенные на этих участках. Однако структуры стебель-петля в 3’ нетранслируемой области демонстрируют высокую степень консервативности в отличие от таковых в других областях транспозонов.

2.3. Анализ транспозонов семейства Alu в геноме человека

Считается, что вторичная структура РНК мономеров Alu такая же, как у 7SL РНК (РНК сигналраспознающей частицы), поскольку Alu состоит из двух 7SL-Alu-доменов [39]. Каждый мономер в нетранслируемой 5’-области содержит две коротких структуры стебель-петля, соединенных U-образным поворотом, а в нетранслируемой 3'-области – длинную структуру стебель-петля с так называемым пузырьком (bulge) [39]. На Рис. 16A изображена структура активного в настоящее время AluYa5 транспозона. Двумерное представление структуры стебель-петля, обнаруженной на 3’-конце Alu транспозона, показано на Рис. 16B.

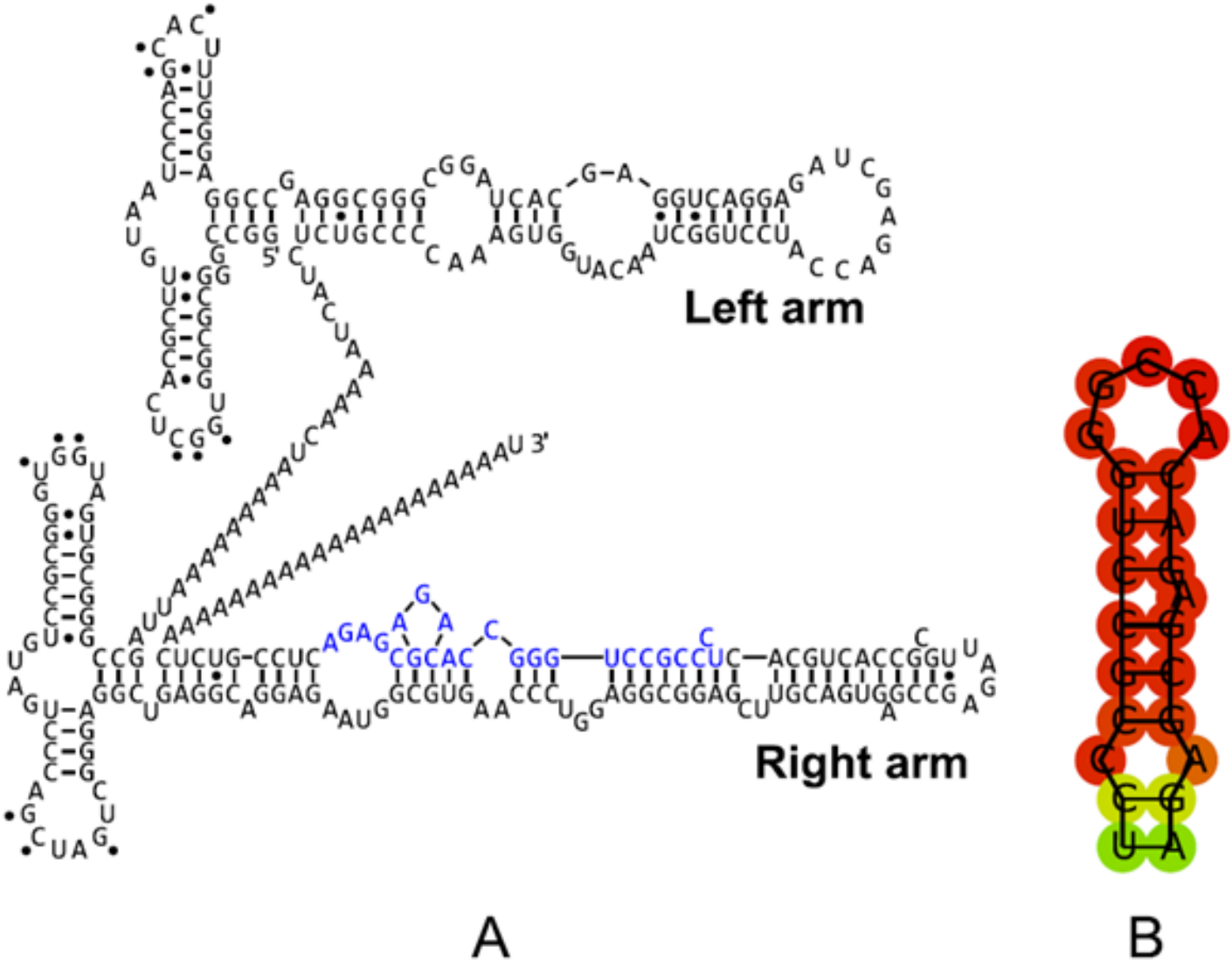


Рис. 16. A) Структура активного транспозона AluYa5, B) структура стебель-петля, расположенная на его 3’ конце (соответствующая ей последовательность нуклеотидов выделена синим).

Alu транспозон унаследовал от сигналраспознающей частицы трехмерную структуру, распознаваемую белковым комплексом SRP9/14. Белковый комплекс SRP9/14, связывающийся с РНК-транскриптом Alu, изображен на Рис. 17 [133].

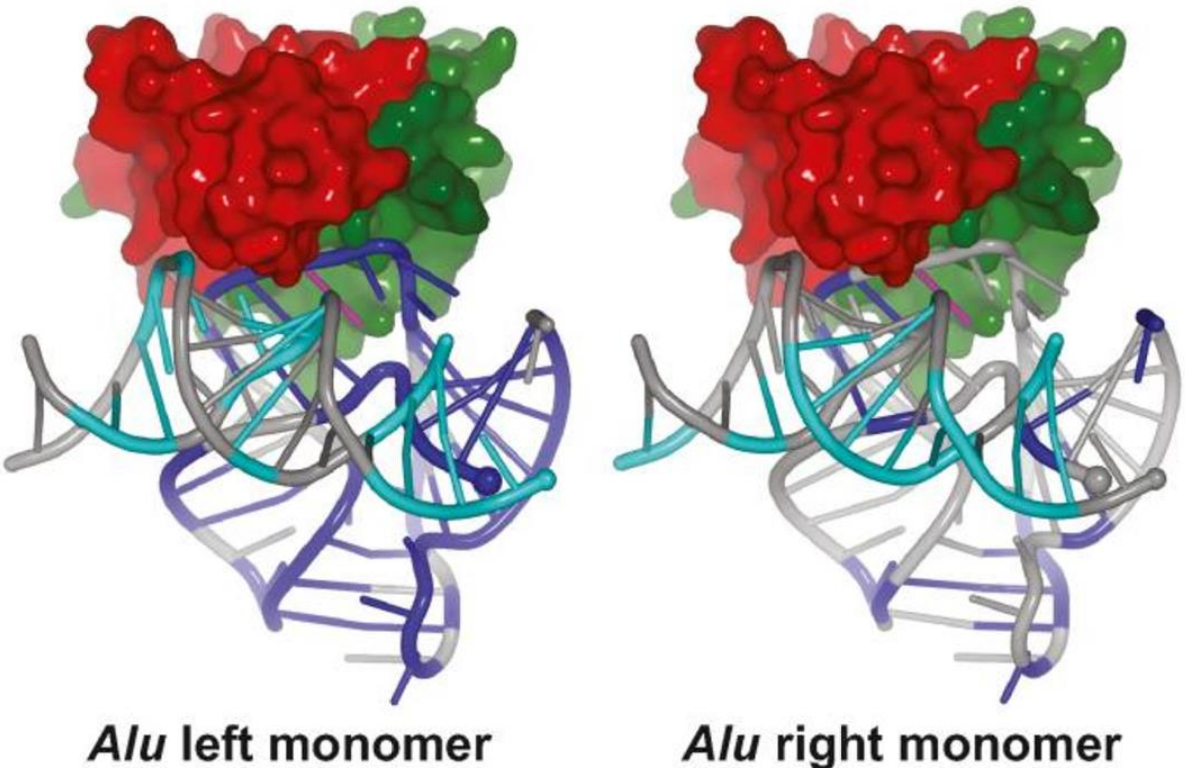


Рис. 17. Трехмерная структура комплекса РНК Alu и белков SRP9/14 [133].

Принятая на данный момент модель ретротранспозиции Alu предполагает, что именно этот белковый комплекс осуществляет так называемую «стыковку» транспозона Alu и рибосомы, на которой, в свою очередь, происходит синтез белков, кодируемых транспозоном L1. Оказываясь в непосредственной близости к белкам L1, Alu транспозон производит «захват» молекулярного аппарата L1 и использует его для вставки своей копии в геном вместо копии L1. Схема ретротранспозиции Alu приведена на Рис. 18 [133]

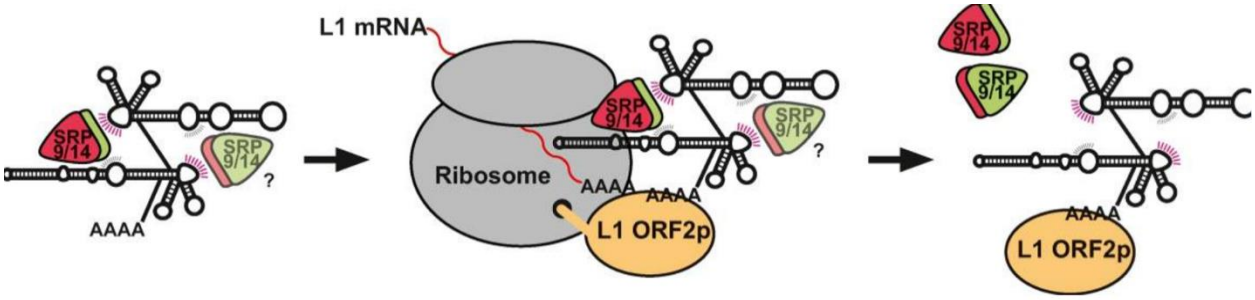


Рис. 18. Схема ретротранспозиции Alu [133].

Эксперименты по изучению влияния мутаций последовательности Alu на его способность связываться с белковым комплексом SRP9/14 показали, что мутации,

введенные в область связывания на левом мономере, намного сильнее уменьшают эффективность ретротранспозиции, чем мутации, введенные в область связывания, расположенную на правом мономере [133]. Выявленная закономерность указывает на различную роль левого и правого мономеров в ретротранспозиции. Скорее всего, левый мономер связан с белковым комплексом SRP9/14, в то время как правый мономер взаимодействует с только что синтезированным белком L1. Весьма вероятно, что распознавание специфической структуры стебель-петля на РНК транскрипте правого мономера играет важную роль в этом процессе. 7SL РНК содержит три структуры стебель-петля: две в нетранслируемой 5'-области и одну в нетранслируемой 3'-области. Каждая из этих структур потенциально может быть использована Alu для связывания с доменом белка ORF2 транспозона LINE – обратной транскриптазой. Известно, что обратная транскрипция начинается с самого конца нетранслируемой 3'-области. Наиболее вероятно, что структура стебель-петля, близкая к нему, и участвует в связывании с обратной транскриптазой.

На наличие вторичных структур стебель-петля были проверены 401 242 Alu транспозона, разбитые на 213 подсемейств. Для всех Alu транспозонов из всех семейств (AluJ, AluS и AluY) были обнаружены структуры стебель-петля на самом конце нетранслируемой 3'-области за несколько нуклеотидов до поли-А хвоста. Однако если правый мономер Alu принимает структуру 7SL РНК, то предсказанная структура стебель-петля скорее всего не будет образована, так как вся ее последовательность войдет в длинный стебель мономера (Рис. 16). Предсказанная структура стебель-петля приведена на Рис. 16 В). Она содержит петлю из 5 пар оснований и симметричный пузырек на расстоянии 6 пар оснований от петли.

Были проанализированы особенности двумерной структуры стебель-петля в зависимости от принадлежности транспозона к тому или иному подсемейству. Все предсказанные структуры стебель-петля можно разделить на две группы – транспозоны Alu, содержащие так называемый пузырек, и не содержащие его. Практически все структуры стебель-петля, принадлежащие транспозонам из древних семейств, не имеют пузырька, в то время как более молодые подсемейства, включая активные в настоящее время транспозоны AluY, содержат пузырек, схожий с представленным на Рис. 16. Интересно, что в семействе AluS, являющимся промежуточным между древним семейством AluJ и молодым AluY, примерно половина предсказанных структур стебель-петля имеет пузырек. Его возможная роль рассмотрена ниже.

2.4. Структура стебель-петля в нетранслируемой 3’-области LINE и SINE транспозонов

Нетранслируемая 3’-область занимает примерно 200-245 пар нуклеотидов и располагается сразу после второй рамки считывания (ORF2). Активные в настоящее время L1 транспозоны и транспозоны, принадлежащие самому молодому семейству L1PA1, имеют на профилях покрытия отчетливые пики в нетранслируемой 3’-области. Нуклеотидные последовательности структур стебель-петля из активных транспозонов были использованы для построения двумерных моделей вторичной структуры. Такой подход был использован для первичного «разведочного» анализа предсказанных структур. На Рис. 19A изображена структура стебель-петля, накрывающая 50 последних нуклеотидов активного L1 транспозона.

Для сравнения были построены модели структур стебель-петля, расположенных в конце нетранслируемых 3’-областей LINE транспозонов различных организмов, включая организмы, для которых важная роль структуры в процессе ретротранспозиции была показана экспериментально. Так, на Рис. 19 изображены структуры стебель-петля из транспозона семейства L2 в геноме угря (Рис. 19B) [18], две структуры из транспозона семейства L2 в геноме рыбы полосатый данио (Рис. 19 C-D) [134], структура из транспозона R1 в геноме тутового шелкопряда (Рис. 19E) [59], структуры из L1 транспозонов крысы (Рис. 19 F) [135] и кукурузы (Рис. 19G) [72], а также структура из R1 транспозона комара (Рис. 19H) [72]. Необходимость структур из транспозонов угря, рыбы и тутового шелкопряда для ретротранспозиции была показана экспериментально. Обратная транскриптаза транспозона L1 связывается именно со структурами стебель-петля на 3’-конце транскрипта [18, 59, 134]. Характерной чертой вышеперечисленных структур является ассиметричный пузырек, расположенный на расстоянии 4-6 нуклеотидов от петли (Рис. 19 A-H).

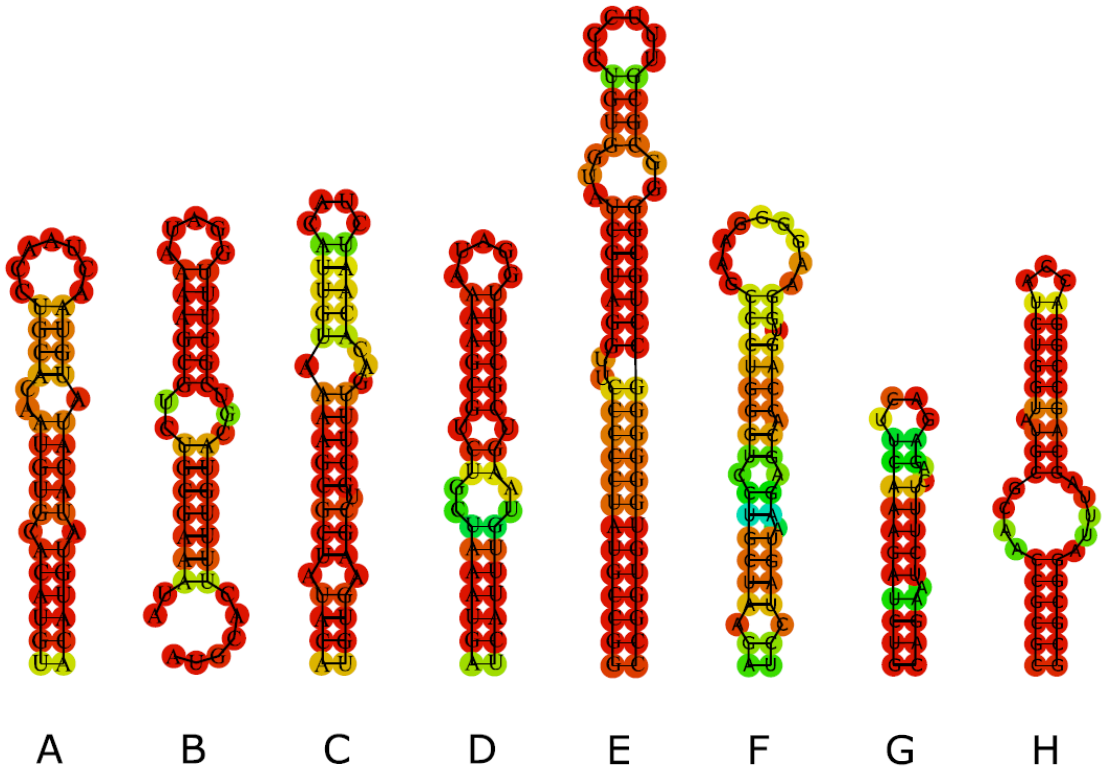


Рис. 19. Вторичная структура стебель-петля, расположенная на 3’-конце LINE транспозонов различных организмов. A) Структура стебель-петля, расположенная на 3’-конце 6 самых активных в настоящее время L1 транспозонов в геноме человека; B) Структура стебель-петля, расположенная на 3’-конце L2 транспозона в геноме угря. Было показано экспериментально, что именно с этой структурой связывается белок ORF2; C-D) Две структуры стебель-петля, расположенные на 3’-конце L2 транспозона в геноме рыбы полосатый данио. Существует экспериментальные подтверждения распознавания белком ORF2 обеих структур и последующего связывания с ними; E) Структура стебель-петля, расположенная на 3’-конце R1 траснпозона в геноме тутового шелкопряда. Экспериментально показано связывания ORF2 белка с данной структурой; F) Структура стебель-петля, расположенная на 3’-конце L1 транспозона в геноме крысы; G) Структура стебель-петля, расположенная на 3’-конце L1 транспозона в геноме кукурузы; H) Структура стебель-петля, расположенная на 3’-конце L1 транспозона в геноме комара.

Интересно, что сравнительный анализ двумерных моделей структур выявил некоторое сходство структур из активных транспозонов человека и рыбы *Danio rerio*. Под сходством мы понимаем позицию пузырька относительно петли и ее длину. Двумерные модели вторичных структур были построены и для полного набора из 6622 транспозонов L1 человека. Более чем в половине из них был обнаружен характерный пузырек.

Профили парных и непарных оснований были сконструированы для оценки распределения размеров пузырьков и их позиций в L1 и Alu транспозонах. Не нулевые значения соответствуют неспаренным основаниям (Рис. 20). Профили L1 транспозонов показали наличие двух пузырьков, первый из которых расположен на расстоянии 4-5 нуклеотидов от петли, второй – на расстоянии 11 нуклеотидов от петли (Рис. 20 A-B). Для Alu было проведено сравнение двух структур – образованной нетранслируемой 3’-областью и накрывающей 50 последних нуклеотидов 3’-конца. Профиль выявил наличие одного неспаренного нуклеотида на расстоянии 4 нуклеотидов от петли, а также пузырька на расстоянии 13-14 нуклеотидов от петли в первой структуре (Рис. 20 E-F) и наличие трёх пузырьков на расстоянии 3, 6-7 и 12 нуклеотидов от петли во второй структуре (Рис. 20 C-D). Профили показали, что структура стебель-петля из 3’ конца транспозона L1 в большей степени похожа на структуру, накрывающую 50 нуклеотидных пар на самом конце Alu транспозона, нежели на структуру, образованную всей 3’ нетранслируемой областью.

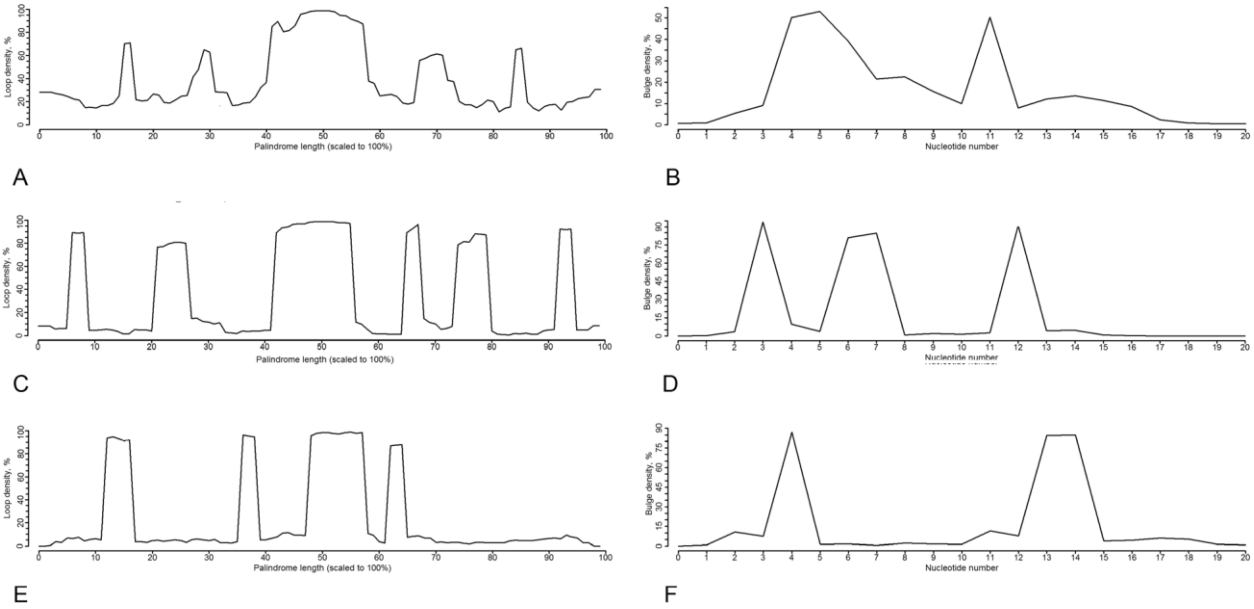


Рис. 20. Профили спаренных и неспаренных нуклеотидов на 3’-конце транспозонов L1 и Alu. По оси ординат отложена доля структур стебель-петля, имеющих неспаренный нуклеотид на заданной позиции. A) Структура, расположенная на 3’-конце L1 транспозона. Длина выражена в процентах (ось *x*); B) Стебель структуры, расположенной на 3’-конце L1 транспозона. По оси *x* отложены порядковые номера нуклеотидов, начиная от петли; C) Структура, расположенная на 3’-конце Alu транспозона. Длина выражена в процентах (ось *x*); D) Стебель структуры, расположенной на 3’-конце Alu транспозона. По оси *x* отложены порядковые номера нуклеотидов, начиная от петли; E) Структура всей нетранслируемой 3’-области Alu транспозона. Длина выражена в процентах (ось *x*); F) Стебель структуры, образованной нетранслируемой 3’-областью. По оси *x* отложены порядковые номера нуклеотидов, начиная от петли.

Эксперименты по введению мутаций в структуру стебель-петля, расположенную на 3’-конце LINE транспозона угря, показали, что удаление пузырька полностью блокирует процесс ретротранспозиции [18]. Ранее для интронов и вирусов было установлено, что пузырек на стебле структуры несет важную функциональную нагрузку. Например, интроны группы II в домене V имеют структуру с пузырьком на стебле. Именно внутри пузырька находятся сайты связывания ионов металла, необходимых для правильного размещения функциональных групп, участвующих в катализе [136]. Пузырек играет важную роль в процессе упаковки HIV-1, в связывании структуры стебель-петля с рибосомальным белком дрожжей L30 и иммуноглобулином М мыши [137]. Исследования реакций с участием ДНК структур стебель-петля, содержащих пузырек, свидетельствуют о значимости его размера, а также характеризуются уменьшением энергии Гиббса [138].

Наши исследования показали, что позиция пузырька относительно стебля найденной структуры на 3’-конце L1 транспозонов консервативна для всех активных L1. Более того, удалось показать ее консервативность и для более чем 50% проанализированных 6622 L1 транспозонов, разбитых на 27 разных подсемейств.

В принятой в настоящее время модели ретротранспозиции рассматривается два типа распознавания РНК транскрипта белком ORF2 транспозона LINE - строгое и нестрогое. При строгом распознавании домен обратной транскриптазы белка ORF2 «узнает» 3’-конец. Известно, что в геноме некоторых организмов LINE и SINE транспозоны имеют одинаковую последовательность 3’-конца. При нестрогом распознавании считается, что белок, осуществляющий ретротранспозицию, связывается с поли-А хвостом. Такое разделение по типам возникло после обнаружения LINE/SINE пар, имеющих один и тот же 3’-конец. В геномах нескольких организмов экспериментально была показана необходимость структуры стебель-петля для ретротранспозиции по строгому типу. Приведенные выше свидетельства важной роли пузырька, а также результаты экспериментов со структурой стебель-петля в геноме угря дают возможность выдвинуть предположение о его необходимости для распознавания РНК-транскрипта белком.

Эволюционные исследования показали, что транспозоны строгого и нестрогого типов «разбросаны» по всему дереву жизни, то есть нет каких-либо выделенных ветвей, где встречался бы только тот или иной тип [77]. В LINE транспозонах семейства L2 в геноме рыбы полосатый данио был обнаружен участок, связывающийся со вторичной структурой стебель-петля на 3’-конце РНК транскрипта. Он расположен между доменами эндонуклеазы и обратной транскриптазы ORF2. С целью изучения степени консервативности последовательности аминокислот участка, находящегося между двумя вышеупомянутыми доменами, нами было произведено выравнивание последовательностей

аминокислот белков ORF2 36 LINE транспозонов, принадлежащих различным семействам, (Приложение 6). Выравнивание было проведено для различных семейств LINE – L1, L2, R1, R2, CR1, I, Jockey и др. Оно показало наличие некоторой консервативности в последовательности рассматриваемого участка, в связи с чем можно предположить, что в ходе эволюции он сохранил функцию распознавания структуры стебель-петля. Следовательно, нельзя исключать того, что все транспозоны LINE, относящиеся и к строгому, и к не строгому типам, сумели сохранить способность узнавать вторичные структуры.

Как было отмечено выше, нетранслируемая 3’-область играет важнейшую роль в распознавании белком ORF2 РНК-транскрипта LINE транспозонов строгого типа. Экспериментально было показано, что структура стебель-петля на самом конце транскрипта необходима для ретротранспозиции. Так, LINE транспозону семейства R2 в геноме насекомых необходимо 250 последних нуклеотидов на 3’-конце для успешной ретротранспозиции [70]. Интересно, что R2 транспозон тутового шелкопряда может распознавать не только свой собственный 3’ хвост, но и хвост транскрипта R2 транспозона мушки дрозофилы, несмотря на отсутствие сходства на уровне последовательности. Показано, что этот участок содержит несколько структур стебель-петля [81].

В геноме угря была обнаружена LINE/SINE пара, имеющая одинаковую последовательность в нетранслируемой 3’-области. Показано, что белок ORF2 транспозона LINE способен распознавать вторичную структуру стебель-петля на 3’-конце транспозона SINE и связываться с ней [18, 20, 82].

В геноме рыбы *Danio rerio* белок ORF2 связывается с структурой стебель-петля на 3’ конце транскрипта LINE транспозона посредством специального участка, расположеного между доменами обратной транскриптазы и эндонуклеазы [17].

В геноме лосося SINE транспозон произошел от транспортной РНК. Его 5’-конец формирует структуру наподобие клеверного листа, а 3’-конец, в свою очередь, образует удлиненную структуру стебель-петля. Именно эта структура распознается LINE белком ORF2 [19].

Экспериментально подтвержденные свидетельства о распознавании белком ORF2 транспозона LINE строгого типа структуры стебель-петля и последующего связывания с ней наталкивают на вопрос о существовании подобного механизма для транспозонов нестрогого типа. В пользу правомерности предположения существования такового свидетельствует вертикальный характер эволюции траспозонов LINE. К тому же, установленная в ходе данной диссертационной работы гомологичность участка между доменами обратной транскриптазы и эндонуклеазы для 36 LINE транспозонов,

принадлежащих различным семействам, также говорит о возможном сохранении распознавания по вторичной структуре в ходе эволюции, так как известно, что именно этот участок связывается со структурой стебель-петля в геноме рыбы полосатый данио [134]. Alu и L1 не имеют одинаковой последовательности на 3’-конце за исключением поли-А хвоста. В обоих транскриптах он присутствует. Многие авторы полагают, что белок распознает именно поли-А хвост [62]. Было показано, что он играет важную роль в процессе ретротранспозиции Alu. Так, частота ретротранспозиции Alu уменьшается почти в 1000 раз при удалении поли-А хвоста [40]. Интересно, что при создании ДНК копии Alu и последующей ее вставки длина поли-А хвоста может увеличиться вследствие проскальзывания обратной транскриптазы. Этот эффект позволяет Alu поддерживать свою активность, действуя в противовес естественному процессу укорачивания поли-А хвоста [63]. После трансляции белки ORF1 и ORF2 связываются с транспозоном L1, образуя рибонуклеопротеиновую частицу, в составе которой они возвращаются обратно в ядро. Интересно, что белок PABP, связывающийся с поли-А хвостами мРНК в клетке, входит в состав рибонуклеопротеиновой частицы, и его наличие является существенным условием успешной ретротранспозиции [139].

Не все исследования поддерживают идею о необходимости поли-А хвоста для ретротранспозиции. Так, в работе [67] продемонстрировано in vitro, что удаление поли-А хвоста не привело к серьезному ухудшению эффективности ретротранспозиции, зато удаление большего участка в нетранслируемой 3’-области привело к значительному ее снижению. Кроме того, наличие в геномах большого числа ретропсевдогенов, не обладающих поли-А хвостом, также наталкивает на предположение о, по меньшей мере, второстепенной роли поли-А хвоста в ретротранспозиции. Сотни тысяч РНК-транскриптов, не обладающих поли-А хвостом, были успешно вставлены в геном молекулярным аппаратом LINE [64, 65]. В свете приведенных фактов все более правдоподобной представляется идея о существовании другого, более надежного, способа распознавания РНК-транскрипта белком, нежели простое связывание с поли-А хвостом.

Были проанализированы 2112 последовательностей транспозонов класса LINE (228 подсемейств) из базы данных RepBase и 217 последовательностей транспозонов SINE (80 подсемейств) из SineBase. Выявлено наличие консервативной структуры на 3’-конце LINE и SINE транспозонов. В Табл. 2 представлены позиции консервативных структур в парах транспозонов LINE и SINE.

SINE Name	Transposon Length	Stem-Loop Start	Stem-Loop End	References	LINE Name	Transposon Length	Stem-Loop Start	Stem-Loop End	References	Common Tail
ACar-1	279	234	250	SINE Base	VINGI-1_ACAR	3146	3098	3132	RepBase	1
AFC	325	317	329	SINE Base	CILINE2	1973	1861	1898	Oliveira 1999	1
AfroLA	168	158	173	SINE Base	BOV-B	3211	3153	3185	RepBase	1
AfroSINE	222	209	227	SINE Base	BOV-B	3221	3168	3201	RepBase	1
Alu	290	259	277	SINE Base	L1	6064	5956	5997	RepBase	
AnolisSINE2	481	421	439	SINE Base	L2_AC_4	6077	5996	6028	RepBase	1
B1	145	119	133	SINE Base	L1MdF	6330	6141	6175	RepBase	
B1-dID	221	165	183	SINE Base	L1-1_STR	6657	6602	6632	RepBase	
B2	209	148	166	SINE Base	L1_MUR2_3END	1154	1026	1058	RepBase	
B4	299	251	268	SINE Base	L1B_MM	6327	6140	6181	RepBase	
Bm1	447	384	401	SINE Base	ZENON-1_HMEL	3397	3284	3320	RepBase	1
BmSE	259	251	264	SINE Base	R1	5092	5043	5075	RepBase	1
Bovc-Ta2	329	319	333	SINE Base	BOVB	3847	3786	3823	RepBase	1
Bov-tA	210	178	191	SINE Base	BOVB	3847	3786	3823	RepBase	1
C	335	203	222	SINE Base	L1-1B_OPR	7223	7114	7147	RepBase	
CAN	175	97	111	SINE Base	L1_CARN5_3END	1021	973	1004	RepBase	
CELE45	264	257	272	SINE Base	RTE-1	3291	3230	3262	RepBase	
CHRS-1	119	85	103	SINE Base	L1-BT	8390	8302	8335	RepBase	
CHRS-2	321	197	208	SINE Base	L1-BT	8390	8302	8335	RepBase	
CM-1	320	312	324	SINE Base	CR1-2_CM	3620	3555	3589	RepBase	1
Cre-3	506	480	499	SINE Base	L1-1_CR	7343	7281	7308	RepBase	1
Cry	200	192	204	SINE Base	PSCR1	4480	4424	4458	RepBase	1

DANA	389	328	351	SINE Base	ZFL3	2039	1947	1977	Ogiwara 2002	1
DAS-I	87	42	76	SINE Base	L1	9630	9583	9617	RepBase	
DAS-IIa	214	35	50	SINE Base	L1	9630	9583	9617	RepBase	
DAS-IIb	372	228	243	SINE Base	L1	9630	9583	9617	RepBase	
DAS-III	412	394	407	SINE Base	L1	9630	9583	9617	RepBase	
DIP	208	124	141	SINE Base	L1	6600	6529	6560	RepBase	
EC-1	407	373	388	SINE Base	BOV-B	3111	2661	2701	RepBase	1
ERE-A	261	220	234	SINE Base	L1	5996	5925	5954	RepBase	
ERE-B1	190	153	175	SINE Base	L1	5996	5925	5954	RepBase	
ERE-B2	249	212	234	SINE Base	L1	7116	6952	6987	RepBase	
ERI-1	158	143	154	SINE Base	L1	7116	6952	6987	RepBase	
ERI-2	196	105	119	SINE Base	L1_CF	4251	4166	4206	RepBase	
Fc-1	115	88	106	SINE Base	L1_CF	4251	4166	4206	RepBase	
Feilai	286	247	260	SINE Base	JUAN-A	4709	4633	4669	RepBase	
Gecko	188	153	165	SINE Base	MOSQUI	7202	7145	7177	RepBase	
HaSE1	385	347	361	SINE Base	HARTE1	1646	1598	1633	Wang 2012	1
HaSE3	364	327	340	SINE Base	HARTE1	1646	1598	1633	Wang 2013	1
HE1	348	334	351	SINE Base	HER-1	1581	1496	1527	RepBase	1
HpaI	189	169	190	SINE Base	RSG-1	583	528	559	Winkfein 1988	1
ID type 1	85	48	66	SINE Base	L1	6600	6529	6560	RepBase	
IDL-Geo	194	165	183	SINE Base	L1-1_DOR	6599	6549	6584	RepBase	
ID-Spe	146	125	137	SINE Base	L1	6600	6529	6560	RepBase	
Lun1	315	284	301	SINE Base	LFR1	1708	1627	1662	Ogiwara 2002	1
Mad-1	169	152	164	SINE Base	L1	6081	6018	6053	RepBase	
Mad-2	203	153	188	SINE Base	L1	6081	6018	6053	RepBase	
MamSINE1	296	267	280	SINE Base	RTE1_LA	3233	3184	3217	RepBase	

Mar1	258	225	238	SINE Base	BOV-B	2576	2414	2455	RepBase	1
Mar3	139	72	94	SINE Base	L1	2579	2467	2500	RepBase	
MEG-RL	207	171	192	SINE Base	L1	6821	6770	6799	RepBase	
MEG-RS	135	91	109	SINE Base	L1	6821	6770	6799	RepBase	
MEG-T2	232	146	163	SINE Base	L1	6821	6770	6799	RepBase	
MEG-TR	196	152	170	SINE Base	L1	6821	6770	6799	RepBase	
MEN	269	245	259	SINE Base	L1-1_STR	6657	6602	6632	RepBase	
Mon-1	272	250	272	SINE Base	L2_PLAT1A	5056	4987	5021	RepBase	1
OS	477	445	466	SINE Base	RSG-1	583	528	559	Winkfein 1988	1
OsSN3	183	146	178	SINE Base	L1-22_OS	1981	1932	1975	RepBase	
Pca-1	342	303	319	SINE Base	BOVB_PC	3221	3168	3201	RepBase	1
Ped-1	243	204	220	SINE Base	BOV-B 3 END	94	49	86	RepBase	1
Ped-2	197	158	174	SINE Base	BOV-B 3 END	94	49	86	RepBase	1
PoaS-II	125	87	101	SINE Base	L1-5_BDI	5928	5874	5904	RepBase	1
PRE-1	243	169	185	SINE Base	L1-2_SSC	6016	5971	6000	RepBase	
RSINE1	149	86	121	SINE Base	L1	6600	6529	6560	RepBase	
SlmI	227	197	217	SINE Base	RSG-1	583	528	559	Winkfein 1988	1
SmaI	140	110	126	SINE Base	SALL2	2655	2599	2631	Oshima 2005	1
SOR	179	122	136	SINE Base	L1-2_SAR	5491	5415	5455	RepBase	
SP-D-Geo	237	179	195	SINE Base	L1-2_DOR	5792	5730	5759	RepBase	
SS-1	131	91	104	SINE Base	L1-2_SSC	6016	5971	6000	RepBase	
STRIDM	227	176	192	SINE Base	L1-1_STR	6657	6602	6632	RepBase	
TAL	263	246	259	SINE Base	L1-1_CCR	6503	6184	6217	RepBase	
Ther-1	263	240	262	SINE Base	L2	3419	3357	3391	RepBase	1
Ther-2	215	197	208	SINE Base	L3	4489	4430	4468	RepBase	1
Tu-I	189	148	178	SINE Base	L1-1_TBEL	6213	6161	6196	RepBase	

Tu-II	288	256	274	SINE Base	L1-1_TBEL	6213	6161	6196	RepBase	
Tu-III	258	214	227	SINE Base	L1-1_TBEL	6213	6161	6196	RepBase	
TUB	146	48	79	SINE Base	L1-1_TBEL	6213	6161	6196	RepBase	
UnaSINE1	225	178	194	SINE Base	UNAL2	3631	3570	3604	RepBase	1
VES	220	208	220	SINE Base	L1-1A_EF	7237	7182	7216	RepBase	
vic-1	150	138	150	SINE Base	L1-2_VPA	5792	5671	5700	RepBase	

Табл. 2. LINE/SINE пары транспозонов.

Ранее для каждой приведенной пары было показано экспериментально, что вставку в геном РНК-транскрипта SINE осуществляет белок, кодируемый соответствующим транспозоном LINE. В последней колонке Табл. 2 единицей обозначены случаи наличия общего 3’-конца у транспозонов LINE и SINE. Однако консервативная структура присутствует на конце всех указанных транспозонов, вне зависимости от степени сходства последовательностей их нетранслируемых 3’-областей. Транспозоны, имеющие консервативную структуру в 3’-конце, принадлежат разным организмам, в том числе и тем, которые далеко отстоят друг от друга на эволюционном дереве (Рис. 21). Был проанализирован 161 вид из домена эукариот. Обнаружено, что вторичная структура стебель-петля присутствует практически во всех ветвях дерева эволюции эукариот – млекопитающие, птицы, рыбы, земноводные, насекомые, черви, растения.

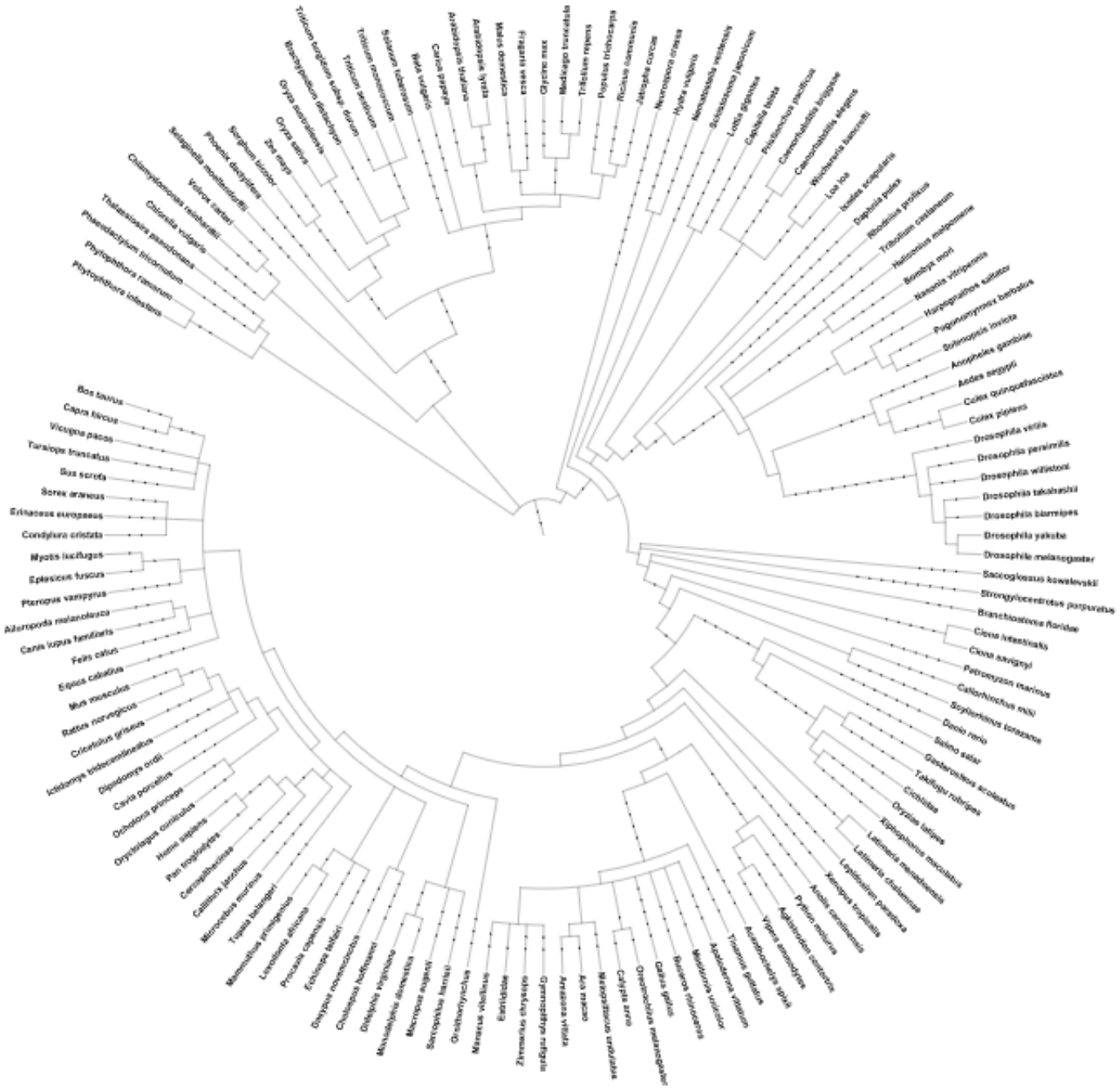


Рис. 21. Филогенетическое дерево проанализированных в данной работе организмов.

Заключение к Главе 2

В данной главе были приведены доказательства наличия консервативной структуры стебель-петля на самом конце 3’-нетранслируемой области траснпозонов L1 и Alu в геноме человека. Было продемонстрировано сходство структур стебель-петля из L1 транспозона и транспозонов LINE из различных организмов, в том числе и тех, для которых экспериментально была установлена необходимость данной структуры в процессе ретротранспозиции. Для 36 организмов, расположенных на разных ветвях эволюционного дерева, была обнаружена гомологичность последовательностей участка белка ORF2, связывающегося со структурой стебель-петля на 3’-конце LINE транспозона в геноме рыбы полосатый данио. Выявленный факт может свидетельствовать о том, что белок в ходе эволюции сохранил способность распознавать вторичную структуру стебель-петля на 3’-конце РНК-транскрипта LINE/SINE транспозонов, принадлежащих к нестрогой группе (то есть тех, для которых предполагается распознавание только поли-А хвоста). Была выдвинута гипотеза о том, что связывание домена обратной транскриптазы белка ORF2 L1 транспозона с L1 и Alu РНК-транскриптами происходит при помощи распознавания похожей структуры стебель-петля. Иными словами, найденная структура служит «опознавательным знаком», по которому белок отличает нужный РНК-транскрипт от множества других, присутствующих в клетке. Сходство структур на конце L1 и Alu РНК позволяет Alu «обмануть» белок L1 и заставить его переместить себя вместо транскрипта L1.

Были найдены другие консервативные структуры стебель-петля, расположенные в нетранслируемой 5’-области и в ORF2 участке транспозонов LINE. Очень вероятно, что и они играют роль во взаимодействии белка с молекулой РНК, однако до сих пор экспериментальных свидетельств этому получено не было.

Разработанные в результате проведенных исследований методы поиска консервативных структур, а также их сравнения могут быть применены к любым последовательностям ДНК и РНК. Их несомненным плюсом являются высокая скорость обработки больших наборов данных, а также наглядность полученного результата. Разработанный программный комплекс доступен на сайте www.dnapunctuation.org.

Глава 3. Выявление физико-химических характеристик обнаруженных структур стебель-петля транспозонов L1 в геноме человека

Роль обнаруженных структур в качестве «опознавательного знака» наталкивает на предположение о существовании физико-химических характеристик, на которые действует естественный отбор. Выявление таких характеристик необходимо в контексте определения свойств специфического связывания РНК-белок. Понимание топологии данной молекулярной системы требует анализа термодинамических, геометрических свойств входящих в нее структур.

3.1. Методы, используемые для анализа структуры стебель-петля

В качестве физико-химических характеристик структур стебель-петля в геноме человека были выбраны динуклеотидные свойства, доступные в базе данных DiProDB [140]. Каждая структура стебель-петля была охарактеризована набором из 11 характеристик: энергией Гиббса, энтальпией, энтропией, гидрофильностью и шестью геометрическими параметрами - Shift, Slide, Rise, Tilt, Roll and Twist. Они приведены в Табл. 3.

	AA	AC	AG	AT	CA	CC	CG	CT
Shift	-0.08	0.23	-0.04	-0.06	0.11	-0.01	0.3	-0.04
Hydrophilicity	0.023	0.083	0.035	0.09	0.118	0.349	0.193	0.378
Slide	-1.27	-1.43	-1.5	-1.36	-1.46	-1.78	-1.89	-1.5
Rise	3.18	3.24	3.3	3.24	3.09	3.32	3.3	3.3
Tilt	-0.8	0.8	0.5	1.1	1	0.3	-0.1	0.5
Roll	7	4.8	8.5	7.1	9.9	8.7	12.1	8.5
Twist	31	32	30	33	31	32	27	30
Free energy	-0.93	-2.24	-2.08	-1.1	-2.11	-3.26	-2.36	-2.08
Enthalpy	-6.82	-11.4	-10.48	-9.38	-10.44	-13.39	-10.64	-10.48
Entropy	-19	-29.5	-27.1	-26.7	-26.9	-32.7	-26.7	-27.1

	GA	GC	GG	GT	TA	TC	TG	TT
Shift	0.07	0.07	-0.01	0.23	-0.02	0.07	0.11	-0.08
Hydrophilicity	0.048	0.146	0.065	0.16	0.112	0.359	0.224	0.389
Slide	-1.7	-1.39	-1.78	-1.43	-1.45	-1.7	-1.46	-1.27
Rise	3.38	3.22	3.32	3.24	3.26	3.38	3.09	3.18
Tilt	1.3	0	0.3	0.8	-0.2	1.3	1	-0.8
Roll	9.4	6.1	12.1	4.8	10.7	9.4	9.9	7
Twist	32	35	32	32	32	32	31	31
Free energy	-2.35	-3.42	-3.26	-2.24	-1.33	-2.35	-2.11	-0.93
Enthalpy	-12.44	-14.88	-13.39	-11.4	-7.69	-12.44	-10.44	-6.82
Entropy	-32.5	-36.9	-32.7	-29.5	-20.5	-32.5	-26.9	-19

Табл. 3. Динуклеотидные характеристики из базы DiProDB.

Шесть геометрических параметров динуклеотидов, описывающих пространственное расположение пар оснований относительно друг друга, представлены на Рис. 9.

Для каждой структуры стебель-петля термодинамические характеристики были рассчитаны, исходя из модели ближайших соседей. В ней энергия Гиббса структуры складывается из энергии Гиббса стебля (спаренной части) и петли (неспаренные нуклеотиды). При подсчете энергии стебля для каждой пары нуклеотидов учитываются вклады от соседних пар с обеих сторон. Таким образом, модель предполагает, что вклад пары оснований в ту или иную термодинамическую характеристику зависит только от двух ближайших соседей. Термодинамическая характеристика линейно зависит от частоты появления в последовательности динуклеотидной пары. Энергия Гиббса РНК дуплекса может быть представлена:

$$\Delta G^{37^\circ}{}_{total} = \sum_i n_i \, \Delta G^{37^\circ}(i) \; + \Delta G^{37^\circ}{}_{initiation} + \; \Delta G^{37^\circ}{}_{sym}$$

Первое слагаемое – это вклад *i-й* динуклеотидной пары оснований, n_i раз встречающейся в последовательности. Величина i изменяется от 1 до 10 (число возможных динуклеотидных пар оснований); $\Delta G^{37^\circ}{}_{initiation}$ – это энергия инициации. В нее входят факторы, не зависящие от последовательности (конденсация противоионов, энтропийные потери при формировании дуплекса и т.д.), Третий член отвечает за энтропийные потери в случае, если дуплекс образовался из одной нити (комплементраные участки находятся на одной и той же нити молекулы РНК/ДНК) [124].

Например, в РНК дуплексе на Рис. 22 стрелки указывают на середину каждой динуклеотидной пары оснований.

↓ ↓ ↓ ↓ ↓

5' C-G-T-T-G-A 3'

3' G-C-A-A-C-T 5'

Рис. 22. РНК дуплекс.

Энергия Гиббса образования приведенного РНК дуплекса:

$$\Delta G^{37^\circ}{}_{predicted} = \Delta G^{37^\circ}{}_{initiation} + \; \Delta G^{37^\circ}(CG/GC) + \; \Delta G^{37^\circ}(GT/CA) + \; \Delta G^{37^\circ}(TT/AA) + \; \Delta G^{37^\circ}(TG/AC) + \; \Delta G^{37^\circ}(GA/CT)$$

Слагаемое $\Delta G^{37^\circ}{}_{sym}$ = 0, так как дуплекс образовался из двух нитей [141].

В работе [141] приведено подробное описание процедуры определения значений харакетристик динуклеотидов, представленных в базе DiProDB. Термодинамические

параметры динуклеоитидных пар были определены методом УФ плавления. Нагревание приводит к разрыву двойной цепи ДНК/РНК и переходу молекулы в клубкообразное состояние. При этом растет величина поглощения в УФ диапозоне. Основной характеристикой этого процесса является температура, при которой половина звеньев молекулы находится в клубкообразном состоянии (T_m). На кривой плавления она соответствует середине плавного перехода (Рис. 23). Температуру плавления измеряют при различных концентрациях, а затем находят энтальпию и энтропию из уравнения регрессии:

$$1/T_m = (R/\Delta_r H)\ln(C_T/4) + (\Delta_r S/\Delta_r H)$$

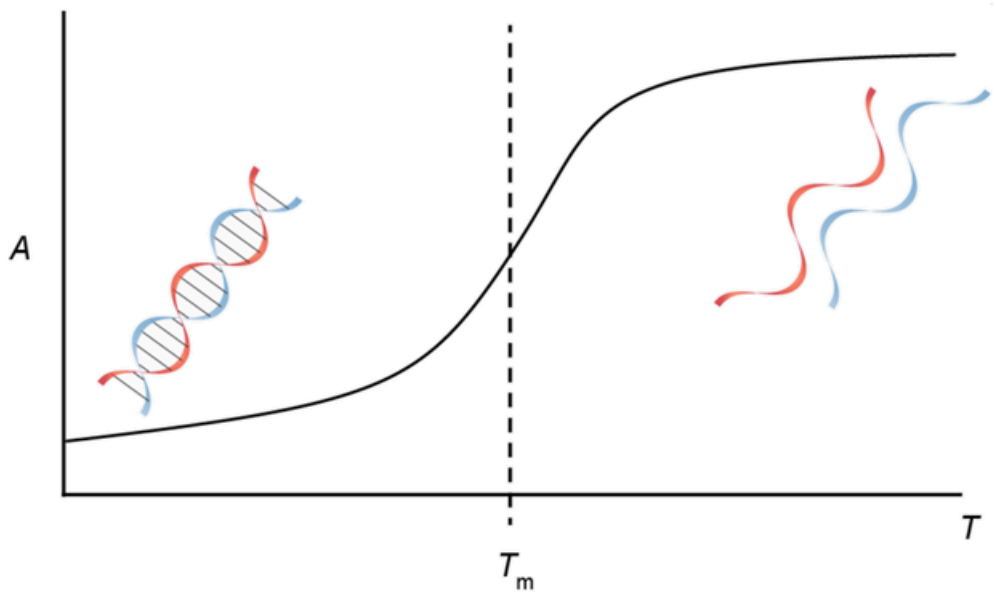


Рис. 23. Типичный вид кривой плавления. Рисунок взят с сайта https://www.atdbio.com/content/53/DNA-duplex-stability#Self-complementary-DNA.

Соответствующую им энергию Гиббса находят по формуле:

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

Чаще всего свободную энергию вычисляют при *T = 37°C*.

Для того, чтобы найти термодинамические характеристики динуклеотидных пар оснований в работе методом УФ плавления были измерены энтропия, энтальпия и энергия Гиббса для 90 олигонуклеотидных РНК дуплексов. Характеристики динуклеотидных пар оснований были получены из уравнения регрессии:

$$\Delta G^{37^\circ}{}_i(duplex) - \Delta G^{37^\circ}{}_{sym} = \Delta G^{37^\circ}(total) = \Delta G^{37^\circ}{}_{initiation} + \sum_j n_{ij}\,\Delta G^{37^\circ}{}_j(NN)\,,$$

где *i* обозначает номер дуплекса [141].

Приведенные выше выражения справедливы для дуплексов РНК, не имеющих петель, «пузырьков» и прочих особенностей структуры. Для их учета в выражения для термодинамических характеристик добавляются дополнительные члены. На Рис. 24 приведены возможные типы структурных особенностей РНК.

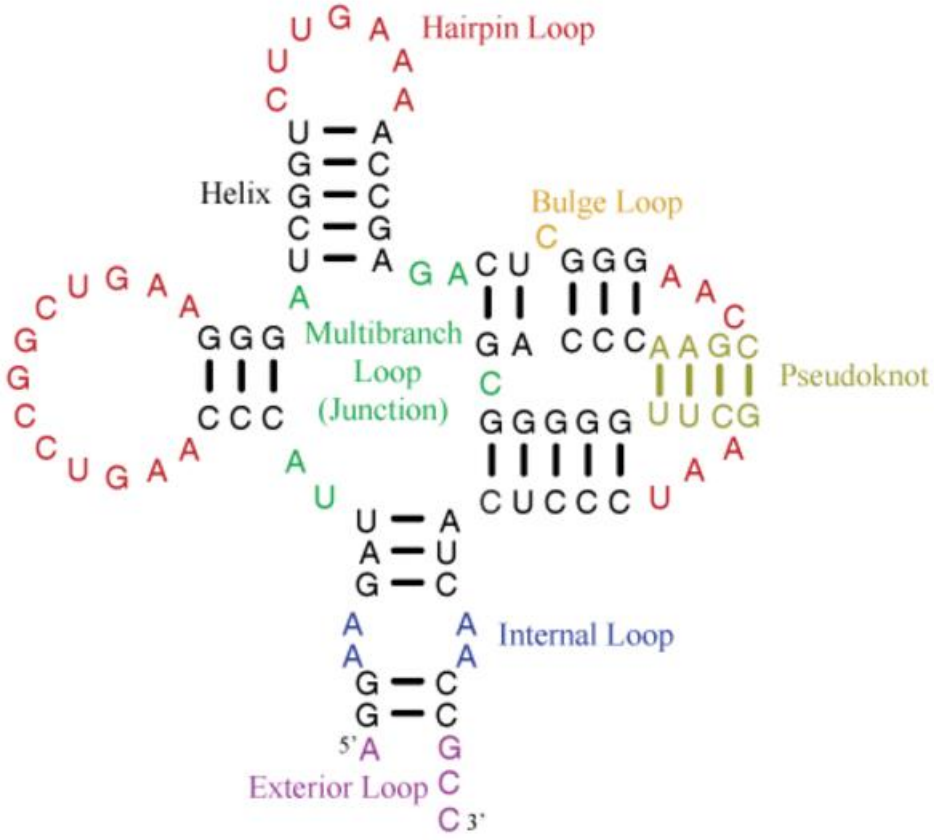


Рис. 24. Типы структурных особенностей РНК. Рисунок взят с сайта http://rna.urmc.rochester.edu/NNDB/help.html.

Для вычисления энергии Гиббса интересующих нас РНК структур типа стебель-петля необходимо было знать вклады, вносимые такими структурными особенностями, как петля шпильки (hairpin loop), внутренняя петля, или пузырек (internal loop), односторонняя петля, или выпуклость (bulge loop). Так, энергия Гиббса для hairpin loop описывается выражением:

$$\Delta G^{37^\circ}{}_{loop} = \Delta G^{37^\circ}{}_{initiation}(n) + \Delta G^{37^\circ}(first\ mismatch\ stacking) + \Delta G^{37^\circ}(UU\ or\ GA\ first\ mismatch\ but\ not\ AG) + \Delta G^{37^\circ}(GG\ first\ mismatch) + \Delta G^{37^\circ}(special\ G-U\ closure) + \Delta G^{37^\circ}(oligo-C\ loops)$$

, где n – число неспаренных нуклеотидов [142].

Слагаемое $\Delta G^{37^\circ}{}_{initiation}(n)$ – энергия Гиббса образования петли, состоящей из n нуклеотидов; $\Delta G^{37^\circ}(first\ mismatch\ stacking)$ – энергия Гиббса взаимодействия первой неспаренной пары нуклеотидов с последней нуклеотидной парой перед петлей (closing base pair); последующие три слагаемых – это бонусы, уменьшающие энергию Гиббса, если UU, GA или GG являются первыми неспаренными нуклеотидами (Рис. 25) [142, 143]; вклад $\Delta G^{37^\circ}(oligo-C\ loops)$ не равен нулю только в том случае, если петля состоит из нуклеотидов С.

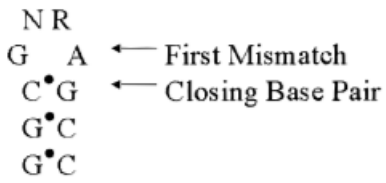


Рис. 25. Пример «закрывающей» нуклеотидной пары и первого несовпадения в петле.

Величины всех вкладов были получены авторами [144] экспериментально методом УФ плавления. Вклад $\Delta G^{37^\circ}{}_{initiation}(n)$ измерен только для $n\ \epsilon$ [3,9]. Если n > 9, то $\Delta G^{37^\circ}{}_{initiation}(n)$ аппроксимируется следующим образом:

$$\Delta G^{37^\circ}{}_{initiation}(n>9) = \Delta G^{37^\circ}{}_{initiation}(9) + 1.75RTln(n/9)$$

Для некоторых часто встречающихся пузырьков (internal loops) величины энергий Гиббса были измерены и занесены в базы данных. Для оценки вкладов остальных прибегают к аппроксимации. Пузырьки часто обозначают как $n \times m$, где n -число нуклеотидов на одной нити, m - на другой. Если n=m, то пузырек называют симметричным. В базах данных содержатся значения энергий Гиббса для многих пузырьков: 1x1, 1x2, 2x3. Они были получены методом УФ плавления. Данные из базы использовались для построения линейной регрессии. Так, например, для пузырька 2x3 была отобрана комбинация параметров, дающая очень близкое к экспериментальным данным значение энергии Гиббса:

$$\Delta G^{37^\circ}{}_{2x3} = \Delta G^{37^\circ}{}_{initiation}(n) + \Delta G^{37^\circ}{}_{asym} + \Delta G^{37^\circ}(AU/GU) + \Delta G^{37^\circ}(5'RG/3'YA) + \Delta G^{37^\circ}(5'YG/3'RA) + \Delta G^{37^\circ}(GG) + \Delta G^{37^\circ}(UU)$$

, где $\Delta G^{37^\circ}{}_{initiation}$ – энергия образования петли, содержащей 5 нуклеотидов; $\Delta G^{37^\circ}{}_{asym}$ – вклад в случае неравного количества нуклеотидов на каждом стренде; $\Delta G^{37^\circ}(AU/GU)$ – штраф за A-U, U-A, G-U, или U-G пару на границе с петлей; $\Delta G^{37^\circ}(5'RG/3'YA)$, $\Delta G^{37^\circ}(5'YG/3'RA)$ - энергии взаимодействия последней нуклеотидной

парой перед петлей с неспаренной парой нуклеотидов; $\Delta G^{37^\circ}(GG)$; $\Delta G^{37^\circ}(UU)$ – бонус в случае, если первой или последней парой в петле являются GG или UU [144].

Энергия Гиббса пузырьков, для которых нет экспериментальных данных, аппроксимируются по формуле:

$$\Delta G^{37^\circ}{}_{loop}(n) = \Delta G^{37^\circ}{}_{loop\ initiation}(n) + |n1 - n2|\ \Delta G^{37^\circ}{}_{asym} + \Delta G^{37^\circ}(AU/GU) + \Delta G^{37^\circ}{}_{first\ noncanonical\ pairs}(except\ for\ 1\ \text{x}\ (n-1)\ \ for\ n > 3)$$

, где $n1$ и $n2$ – число нуклеотидов на каждом стренде [142, 145].

Геометрические параметры динуклеотидов, представленные в базе DiProDB, были получены при помощи программы 3DNA [91]. Здесь приведено краткое описание процедуры их определения. На вход программы подавались координаты атомов, полученные рентгеноструктурным анализом. Основания идентифицировались при помощи сравнения координат атомов с эталонными. С каждой парой оснований связывался локальный базис, а для каждой динуклеотидной пары оснований вводился базис MST [146], при помощи которого в дальнейшем определялись относительные смещения и углы поворота соседних пар оснований. Достоинством этого базиса является инвариантность измеренных параметров относительно обращения последовательности. Было исследовано 209 динуклеотидных пар оснований РНК. Для каждой возможной динуклеотидной пары оснований находились средние значения смещений и углов поворота. Эти значения приведены в Табл. 3 [91].

Значения гидрофильности были получены авторами [147] методом бумажной хроматографии. Процедура измерения такова: на бумагу наносились капли раствора, содержащего нуклеотиды. В нашем случае раствор содержал молекулы 16 динуклеотидных пар. Далее бумага одним концом помещалась в растворитель, который под действием капиллярных сил продвигался вверх. Когда растворитель достигал пятна, образованного раствором с динуклеотидами, он увлекал часть динуклеотидов за собой. Насколько далеко динуклеотидам удается продвинуться, зависит от степени их растворимости в данном растворителе. В исследовании [147], результаты которого занесены в базу, использовался растворитель с высоким содержанием соли. По предположению более гидрофильная субстанция лучше конкурируют с солью за молекулы воды и, следовательно, дальше перемещается.

3.2. Физико-химические характеристики структуры стебель-петля L1 транспозонов человека.

Были рассмотрены 6 геометрических характеристик (3 трансляционных параметра и 3 угла, описывающих пространственное расположение одной пары оснований относительно другой), 3 термодинамических – энергия Гиббса, энтальпия, энтропия, а также гидрофильность.

Для анализа использовались четыре набора данных. В первый набор были включены последовательности структур стебель-петля из активных в настоящее время транспозонов L1 в геноме человека [26]. Во второй набор вошли структуры из 6622 транспозонов с наиболее сохранившимися последовательностями, отобранными из базы UCSC Genome Browser. Третий набор составили структуры из случайных областей генома. Геном человека был аннотирован вторичными структурами стебель-петля. Из полученной последовательности случайным образом были отобраны 6622 структуры, которые и вошли в третий набор. Четвертый набор был составлен следующим образом. Нуклеотиды в геноме были перемешаны с сохранением количества. Затем полученный геном был проаннотирован вторичными структурами. Далее случайным образом были выбраны 6622 позиции структур стебель-петля. Средние значения термодинамических характеристик полученных структур, рассчитанные исходя из модели ближайших соседей, приведены в Табл. 4. Прослеживается стремление к минимуму значений всех трёх характеристик для структур активных транспозонов.

Для каждой последовательности были рассчитаны медианные значения, средние значения, а также стандартные отклонения термодинамических характеристик образования динуклеотидной пары. Медианные значения совпали со средними с точностью до погрешности измерения. Зависимость медианных значений от стандартных отклонений для разных классов представлена на Рис. 26 и Рис. 27. Из графиков видно, что транспозонные структуры и структуры, сгенерированные случайным образом, могут быть разделены по их термодинамическим характеристикам. Данное утверждение верно и при сравнении транспозонных структур со структурами случайных областей генома.

Средние значения свободной энергии, энтальпии и энтропии для структур из активных в настоящее время транспозонов ниже, чем для структур из хорошо сохранившихся L1, из случайных областей генома и случайно сгенерированных структур. Таким образом, структуры из активных в настоящее время транспозонов L1 стабильнее остальных. Структуры из хорошо сохранившихся L1 близки к средним значениям структур из случайных мест генома. Возможно, что в выборку последних попали стабильные

функционально значимые структуры, что повысило среднее значение. Структуры, сгенерированные случайным образом, являются наименее стабильными.

Набор данных	Число последовательностей в наборе	ΔG, ккал моль$^{-1}$	ΔH, ккал моль$^{-1}$	ΔS, Дж моль$^{-1}$ К$^{-1}$
Структуры из активных L1	39	-14±1	-130±2	-376±5
Структуры из хорошо сохранившихся L1	6622	-11±3	-119±21	-348±65
Структуры из случайных мест генома	6622	-12±6	-121±38	-351±79
Структуры, сгенерированные случайным образом	6622	-9±3	-109±27	-320±78

Табл. 4. Термодинамические характеристики последовательностей структур стебель-петля.

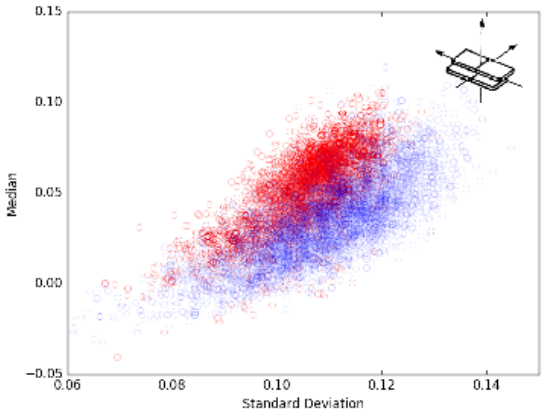
В терминах динуклеотидных геометрических характеристик возможно описать многие свойства, необходимые для связи с белком, например, локальную упругость, гибкость, закрученность. Для каждой последовательности были рассчитаны медианное значение, среднее значение и стандартное отклонение каждой из шести геометрических характеристик (Рис. 26 и Рис. 27). Медианные значения оказались практически равны средним. Далее, для сравнения значений физических характеристик между наборами данных были рассчитаны средние медианные значения каждой характеристики в том или ином наборе (Табл. 5). Гипотеза о равенстве средних значений физических характеристик структур из разных наборов данных была отвергнута при уровне значимости 0.05. Таким образом, по всем шести геометрическим характеристикам структуры из активных и хорошо сохранившихся транспозонов L1 статистически значимо ($p < 0.05$) отделяются от структур из случайных областей генома и сгенерированных случайным образом. Это говорит о сохранении геометрических характеристик на динуклеотидном уровне в процессе эволюции. Такая консервативность может быть объяснена поддержанием структурных особенностей шпильки, необходимых для связи с белком. Видно, что структуры из L1 обладают определенными особенностями, однако точнее охарактеризовать их, исходя из

средних значений, нельзя. Необходимо построение пространственной модели, о которой речь пойдет ниже.

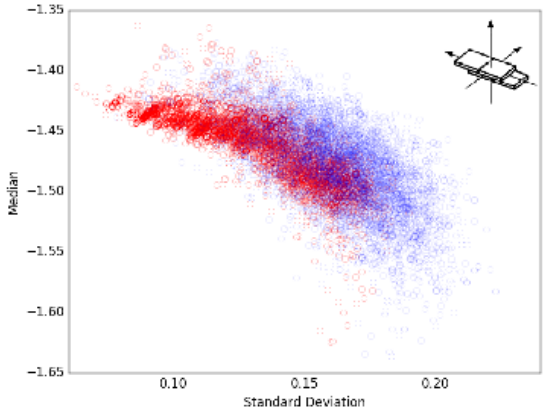
Проверка гипотез о неразличимости структур транспозонов и структур из других мест генома, а также структур, сгенерированных случайным образом, проводилась методом многомерного дисперсионного анализа. Расчеты были произведены в среде R. В ходе анализа были отвергнуты две гипотезы: 1) о равенстве средних значений физических характеристик структур из транспозонов и из случайных мест генома, 2) о равенстве средних значений физических характеристик структур из транспозонов и структур, сгенерированных случайным образом. Чтобы выяснить, какие именно характеристики статистически значимо различаются, были проведены тесты Стьюдента. Полученные p-значения были скорректированы методом Бонферрони, так как сразу несколько гипотез были проверены на одних и тех же данных. По всем характеристикам наблюдаются значимые различия (p-значение < 0,01).

Набор данных	N	Shift, Å	Slide, Å	Rise, Å	Tilt, °	Roll, °	Twist, °	Hydrophilicity
Структуры из активных L1	39	0.07± 0.01	-1.45± 0.01	3.20± 0.01	0.6± 0.1	8.0± 0.1	31.9± 0.1	0.24± 0.01
Структуры из хорошо сохранившихся L1	6622	0.06± 0.02	-1.46± 0.03	3.22± 0.02	0.5± 0.1	8.1± 0.2	31.7± 0.1	0.23± 0.01
Структуры из случайных мест генома	6622	0.02± 0.03	-1.46± 0.06	3.24± 0.02	0.4± 0.2	8.3± 0.3	31.6± 2.5	0.24± 0.03
Структуры, сгенерированные случайным образом	6622	0.04± 0.02	-1.48± 0.04	3.24± 0.02	0.4± 0.2	8.3± 0.2	31.4± 0.2	0.25± 0.22

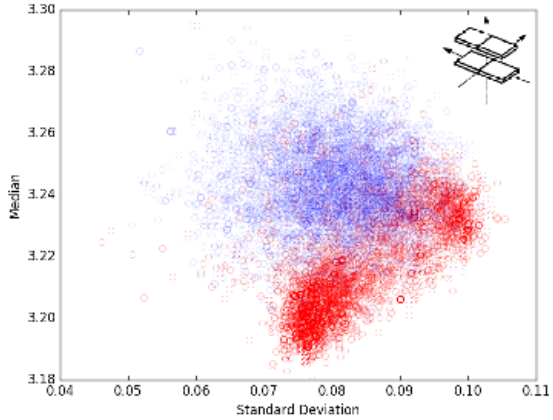
Табл. 5. Геометрические характеристики последовательностей структур стебель-петля и гидрофильность.


Median
Standard Deviation

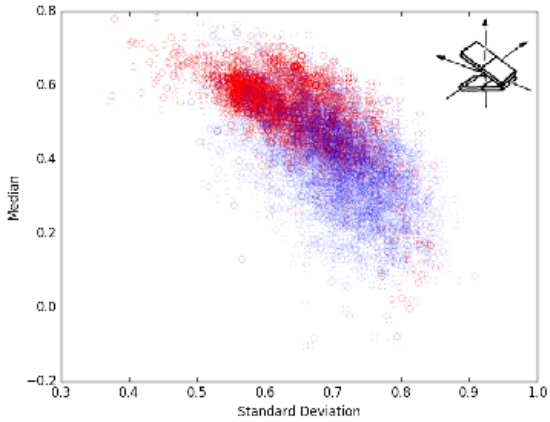
A.


Median
Standard Deviation

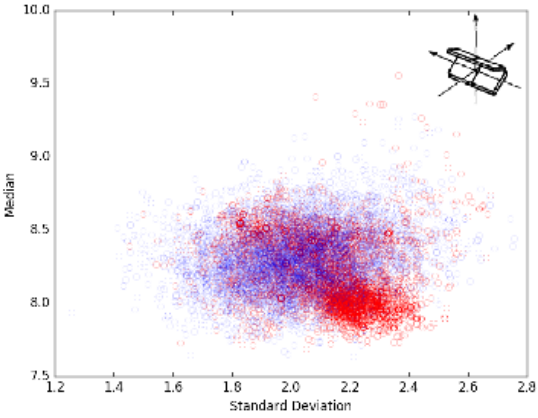
B.


Median
Standard Deviation

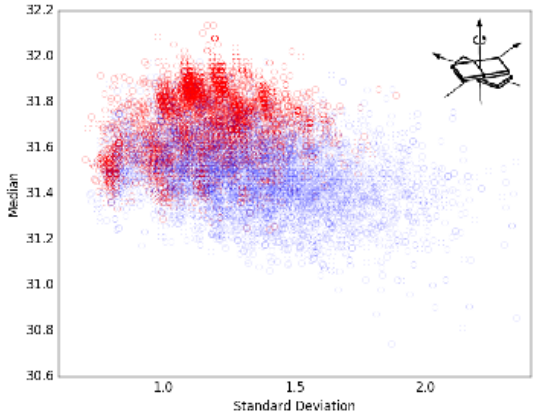
C.


Median
Standard Deviation

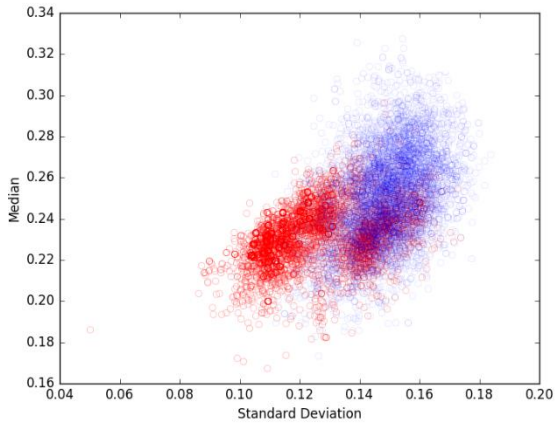
D.


Median
Standard Deviation

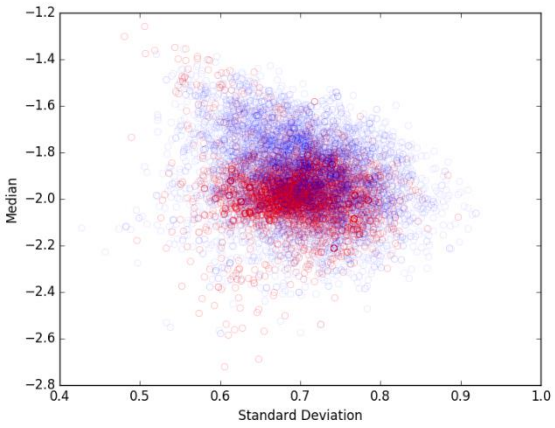
E.


Median
Standard Deviation

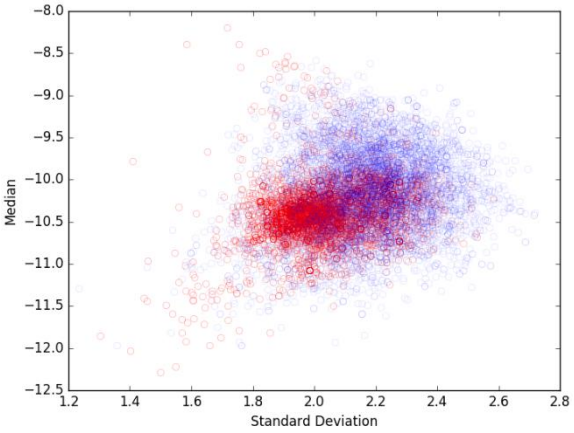
F.



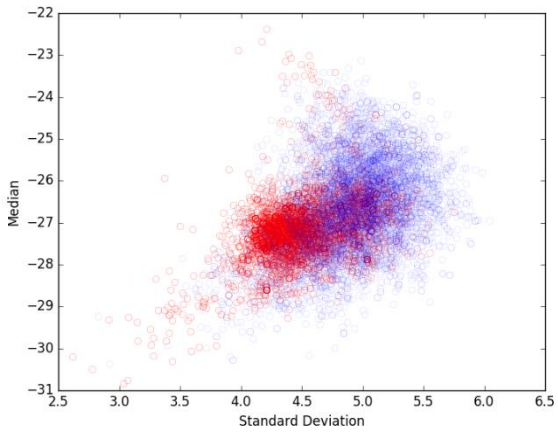
G.



H.



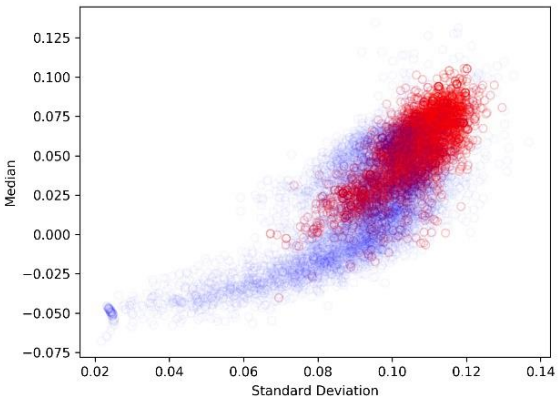
I.



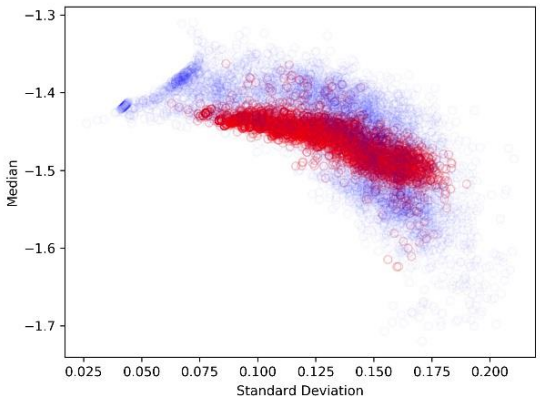
J.

Рис. 26. Сравнение шпилек из 3’-конца L1 транспозона человека (красный цвет) со шпильками из случайным образом сгенерированной последовательности (синий цвет). A. Shift, B. Slide, C. Rise, D. Tilt, E. Roll, F. Twist, G. Гидрофильность, H. Свободная Энергия Гиббса, I. Энтальпия, J. Энтропия.

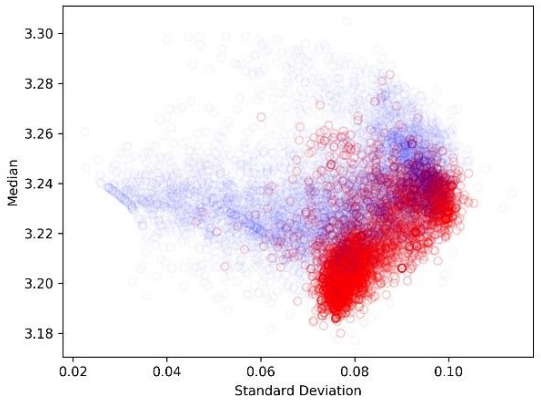
Статистически значимое отличие геометрических характеристик структур из активных, хорошо сохранившихся транспозонов L1 от структур из случайных областей генома, и сгенерированных случайным образом приводит к заключению о действии эволюционного отбора на физико-химические характеристики структур стебель-петля, расположенных на 3’-конце L1 транспозонов человека.


Median
Standard Deviation
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
0.02
0.04
0.06
0.08
0.10
0.12
0.14

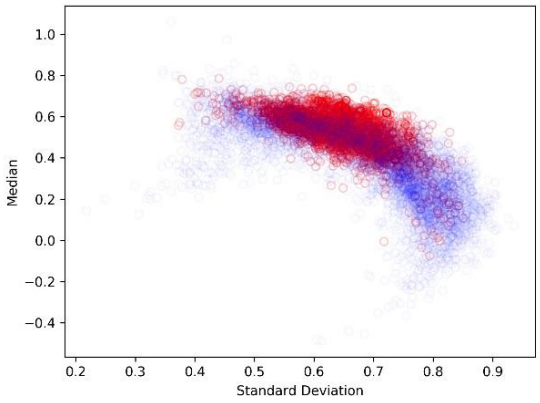
A.


Median
Standard Deviation
−1.3
−1.4
−1.5
−1.6
−1.7
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

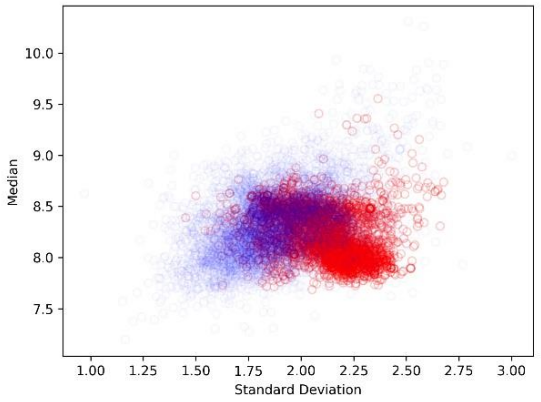
B.


Median
Standard Deviation
3.30
3.28
3.26
3.24
3.22
3.20
3.18
0.02
0.04
0.06
0.08
0.10

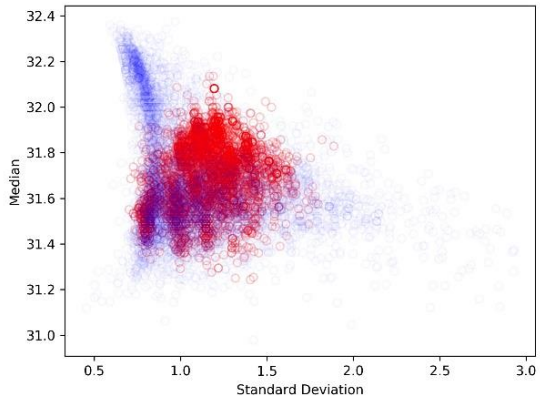
C.


Median
Standard Deviation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

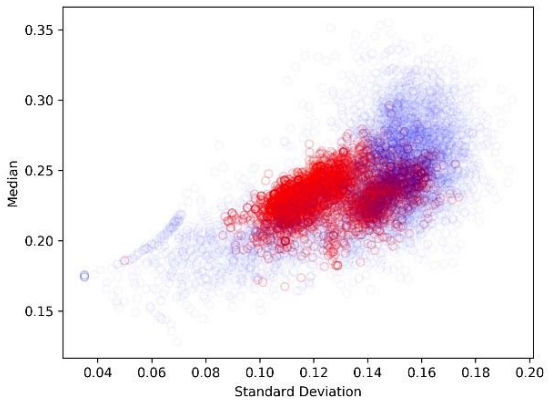
D.


Median
Standard Deviation
10.0
9.5
9.0
8.5
8.0
7.5
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

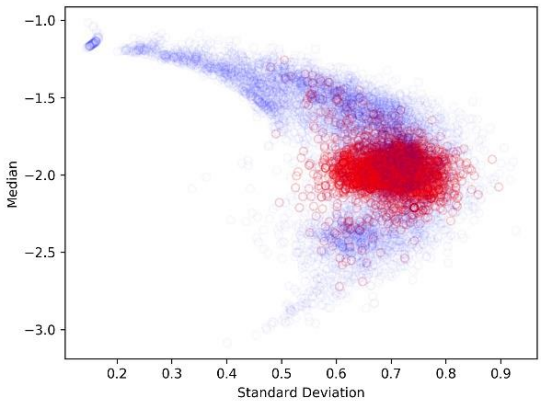
E.


Median
Standard Deviation
32.4
32.2
32.0
31.8
31.6
31.4
31.2
31.0
0.5
1.0
1.5
2.0
2.5
3.0

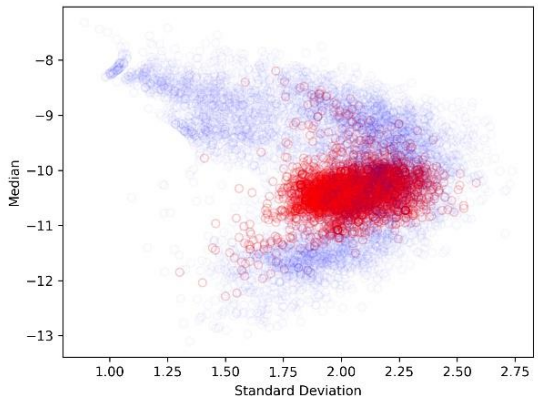
F.



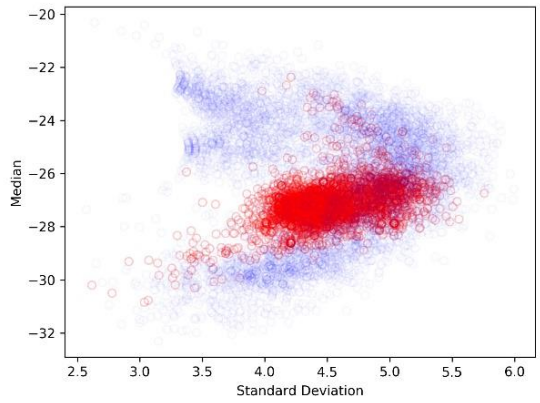
G.



H.



I.



J.

Рис. 27. Сравнение шпилек из 3’-конца L1 транспозона человека (красный цвет) со шпильками из случайных областей генома (синий цвет). A. Shift, B. Slide, C. Rise, D. Tilt, E. Roll, F. Twist, G. Гидрофильность, H. Свободная Энергия Гиббса, I. Энтальпия, J. Энтропия.

Значения динуклеотидных параметров позволяют сделать некоторые выводы о структуре РНК/ДНК молекулы [148]. Параметр Roll отражает величину изгиба в сторону той или иной бороздки, а параметр Tilt – в сторону главной цепи (англ. backbone) макромолекулы. Twist характеризует степень закрученности, кроме того, он тесно связан с числом пар оснований, приходящихся на виток ДНК спирали. Если бы молекула ДНК вела себя как идеальный резиновый шнур, то длина витка спирали полностью определялась бы значениями параметра Rise. Однако в действительности Rise дает менее чем 10% вклад в нее. Зато параметр Slide на 90% определяет длину витка. Интересно, что сильно изогнутые в сторону малой бороздки участки ДНК характеризуются отрицательными значениями Roll

и высокими положительными значениями Slide (например, около 2,5 Å для нуклеосомальной ДНК), изогнутые же в сторону большой бороздки участки обладают небольшими отрицательными значениями Slide (около -1 Å). В работе [148] были смоделированы изменения в структуре ДНК, вызванные занулением одного из параметров для всех динуклеотидов. Этот пример позволяет лучше понять степень влияния того или иного параметра на поведение молекулы. Рис. 28 иллюстрирует изменения, происходящие со спиралью ДНК при занулении либо параметра Shift, либо параметра Slide. Нулевые значения параметра Shift практически не влияют на траекторию цепи, зато зануление параметра Slide приводит к ее выравниванию, «уплощению» (Рис. 28).

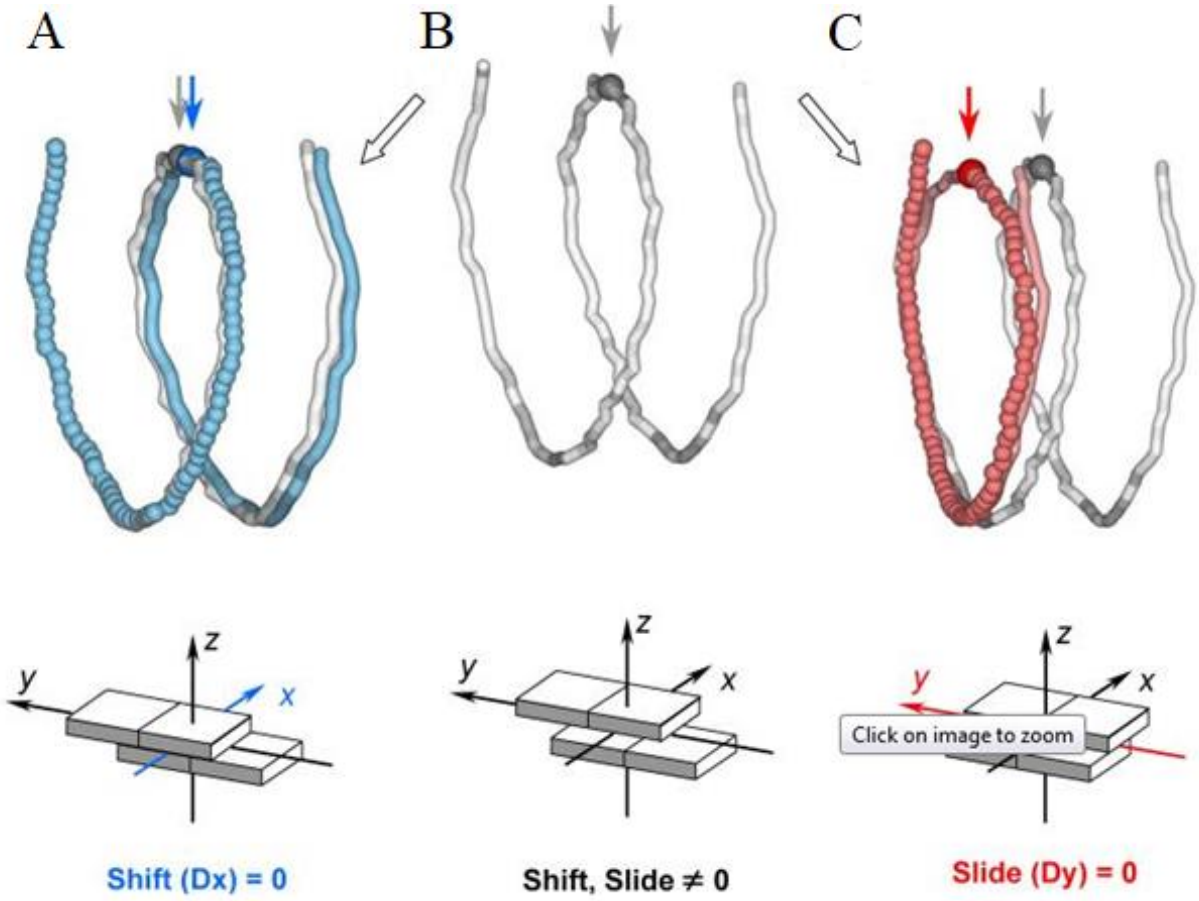


Рис. 28. Полученные после зануления того или иного параметра траектории цепи нуклеосомальной ДНК наложены на немодифицированные аналоги. Модифицированные траектории выделены цветом. A) Изменение траектории цепи при занулении параметра Shift B) Траектории цепи в нативном состоянии C) Изменение траектории цепи при занулении параметра Slide [148].

Зануление параметра Roll для всех динуклеотидов приводит к выпрямлению цепи ДНК, в то время как приравнивание к нулю параметра Tilt не изменяет существенным

образом траекторию цепи (Рис. 29). Иными словами, для нуклеосомальной ДНК справедливо утверждение об определяющей роли параметра Roll в изгибах ДНК.

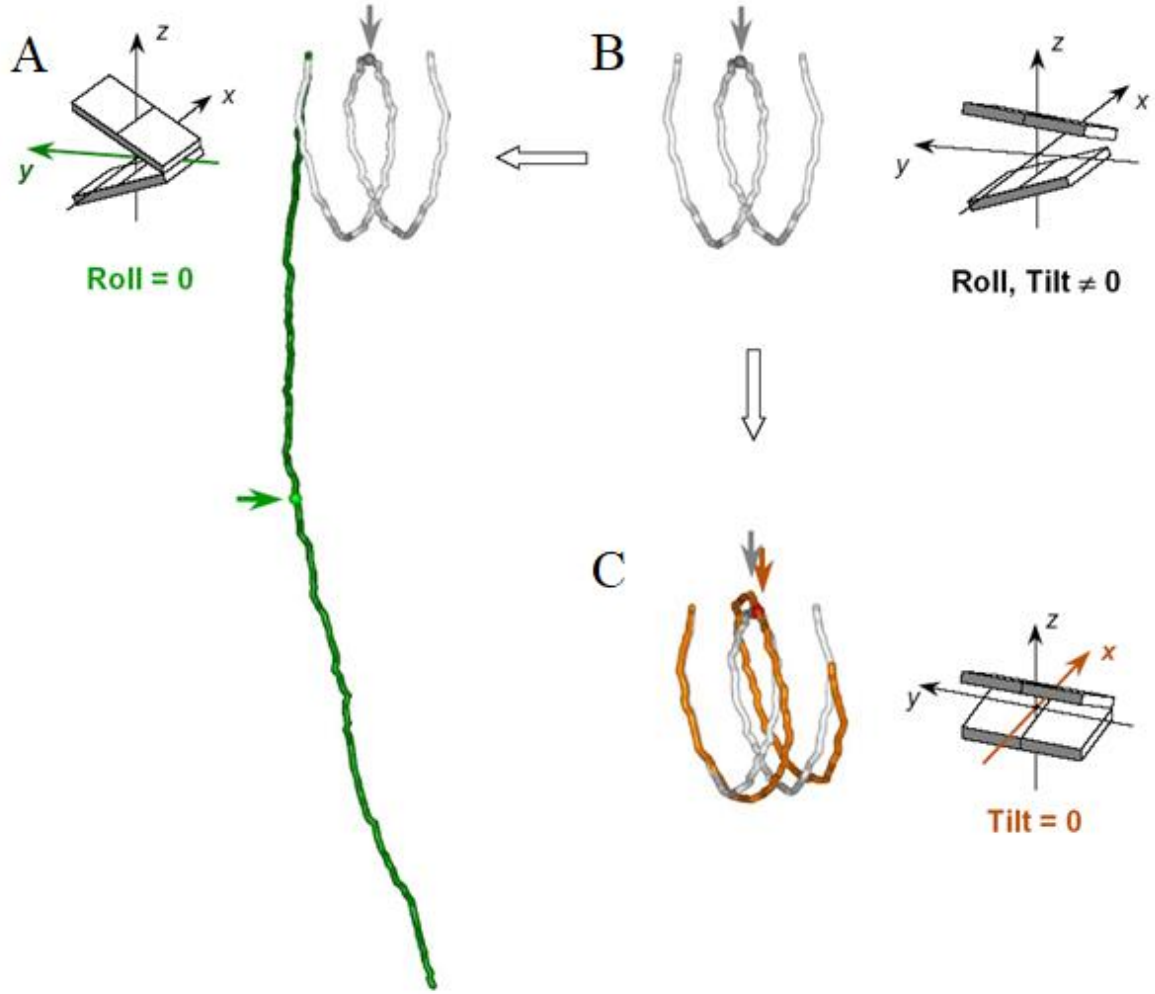


Рис. 29. Полученные после зануления того или иного параметра траектории цепи нуклеосомальной ДНК наложены на немодифицированные аналоги. Модифицированные траектории выделены цветом. A) Изменение траектории цепи при зануленииии параметра Roll B) Траектории цепи в нативном состоянии C) Изменение траектории цепи при зануленииии параметра Tilt [148].

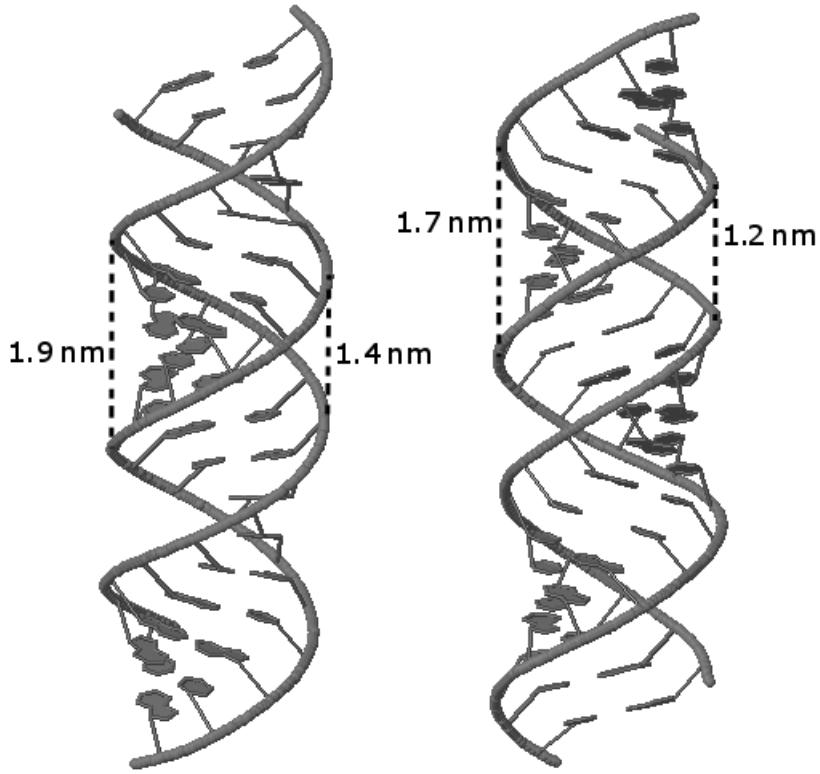
Обнаруженные в ходе данной диссертационной работы структуры стебель-петля, расположенные на конце транспозонов, характеризуются сравнительно небольшими значениями параметра Roll и отрицательными значениями параметра Slide, что может свидетельствовать о наличии плавного изгиба в сторону большой бороздки. Значения параметра Roll для транспозонных структур ниже, чем его значения для других структур из генома и структур, сгенерированных случайным образом. Это говорит о меньшей склонности транспозонных структур к изгибу. Это согласуется с концепцией представления

структуры стебель-петля в качестве «опознавательного знака», поскольку такая структура может служить прямым «стержнем», выступающим над поверхностью молекулы и хорошо заметным для белка. Транспозонные структуры характеризуются и более высоким значением параметра Twist по сравнению с остальными структурами, что может приводить к большей степени закрученности, а также к уменьшению длины витка, что и наблюдается для структур стебель-петля из транспозонов. Как было упомянуто выше, параметр Tilt отвечает за отклонения в сторону главной цепи макромолекулы. Его значения для транспозонных структур очень невелики, однако они все же выше его показателей для остальных структур. Скорее всего присутствует слабый наклон в сторону главной оси. Следует также отметить, что дисперсии всех динуклеотидных параметров транспозонных структур меньше аналогичных величин для структур, случайным образом взятых из других мест генома. Это, в свою очередь, может говорить о некоторой «консервации» структуры, стремлении к сохранению в ходе эволюции важных структурных особенностей, необходимых для выполнения определенной функции. Иными словами, этот факт является косвенным аргументом в пользу выдвинутой в Главе 2 гипотезы о необходимости обнаруженной структуры стебель-петля для распознавания белком ORF2.

Многие исследования показали, что комплексы связывания РНК-белок существенно зависят от формы и последовательности РНК (для обзора см., например, [7, 149, 150]). На величину бороздок влияют в разной степени все 6 геометрических параметров, однако в первую очередь стоит выделить параметры Twist, Roll, Rise, Slide. Известно, что величина и конфигурация бороздок играют существенную роль при связывании белка с РНК. Так, важность размеров большой и малой бороздок при связывании с РНК была показана для белка аденозиндезаминазы, действующей на РНК (ADAR) [8] в процессе редактирования РНК. Роль гидрофобности вместе с размерами бороздок была показана для белков, взаимодействующих с тетрапетлями структур стебель-петля [151] . Было показано, что механизмы взаимодействия могут быть разными и включают специфическое узнавание оснований в гидрофобном кармане, адаптивное связывание с мотивом GNRA в большой бороздке и специфическое связывание в малой бороздке в зависимости от геометрических размеров. Для частиц, распознающих сигнал (SRP), было показано, что РНК-белковые взаимодействия происходят благодаря специфическому связыванию с расширенной большой бороздкой и тетрапетлей без непосредственного контакта белка с нуклеиновыми основаниями, а через определенным образом упорядоченные молекулы воды [152]. Роль малых и больших бороздок при взаимодействии структур стебель-петля с белками была показана для рибосомального комплекса [153]. Роль степени закрученности спирали (Twist) также была показана для РНК-белковых взаимодействий для бычьего вируса

иммунодефицита [154]. Гидратация сайтов связывания при РНК-белковых взаимодействиях была исследована на 89 комплексах РНК-белок из базы данных PDB. Было показано, что во взаимодействиях белок-РНК большая бороздка оказывается более гидратирована, чем малая, в то время как обратная зависимость наблюдается для сайтов связывания ДНК с белком [155].

На Рис. 30 изображены построенные при помощи программы 3DNA модели РНК структуры стебель-петля активного L1 транспозона и аналогичной структуры из случайного места генома (данная структура была выбрана таким образом, чтобы средние по ее последовательности значения динуклеотидных характеристик были близки к средним значениям динуклеотидных характеристик, рассчитанных с использованием всего набора структур из случайных мест генома).



A. B.

Рис. 30. A. Модель стебля структуры из случайной области генома. B. Модель стебля структуры, расположенной на 3’-конце РНК-транскрипта транспозона L1.

Величины малой и большой бороздок, а таже длина витка, полученные при помощи программы 3DNA, представлены в Табл. 6.

	Размер малой бороздки, нм	Размер большой бороздки, нм	Длина витка, нм
Структура из L1 транспозона	1,2 ±0,1	1,7±0,1	2,9±0,2
Структура из случайной области генома	1,4±0,1	1,9±0,1	3,3±0,2

Табл. 6. Характеристики структур, полученные при помощи программы 3DNA.

В работе [156] авторам удалось рассмотреть двойную спираль РНК при помощи атомно- силового микроскопа. Длина витка исследуемой молекулы составила 3,1±0,3 нм. Оценки, полученные в данной диссертационной работе, хорошо согласуются с данным результатом. Структура из L1 транспозона характеризуется меньшими размерами обеих бороздок. Это согласуется с приведенными выше значениями динуклеотидных параметров. Так, параметр Twist принимает большие значения для транспозонных структур, что говорит о более высокой степени их закрученности по сравнению со структурами, отобранными из случайных мест генома. Транспозонные структуры также имеют несколько пониженные значения параметра Rise. Таким образом, структуры, расположенные на 3’ конце L1 транспозонов, являются несколько более сплюснутыми и закрученными, нежели их аналоги из случайных мест генома. Это еще раз подтверждает выделенность транспозонных структур среди всех остальных. Возможно, более узкие бороздки в сочетании с их особой конфигурацией способны принимать только определенный тип белка, что в свою очередь обеспечивает специфичность распознавания.

Видно, что для транспозонной структуры размеры и малой, и большой бороздок меньше, нежели для структуры из случайного места генома. В зависимости от типа белка связывание может проходить и в той, и в другой бороздке. Чтобы узнать, к какому типу белка относится ORF2, необходимо провести эксперимент. Только он даст однозначный ответ на данный вопрос. Однако важным результатом исследования, проведенного в данной диссертационной работы, является выявление различия в размере бороздок между транспозонной структурой и структурой из случайного места генома. Это говорит о неком выделенном положении транспозонной структуры и является аргументом в пользу гипотезы о ее необходимости для распознавания белком ORF2. Важность геометрических характеристик РНК для взаимодействия с белками была экспериментально показана для

большого класса белков, участвующих в разных процессах функционирования генома, таких как редактирование, процессинг, транспорт и интерференция РНК [86, 157]. Необходимо дальнейшее экспериментальное исследование пространственной структуры комплекса домена обратной транскриптазы белка ORF2 с РНК транспозона для определения важнейших для связывания геометрических характеристик РНК.

3.3. Хиральная иерархичность в системе «РНК + белок».

Хиральность является основополагающим фактором иерархического структурообразования в молекулярно-биологических системах [158-160]. В работах [158, 160] впервые было выявлено чередование знака хиральности при переходе на более высокий уровень организации молекулы ДНК, такая же тенденция наблюдается при рассмотрении структурной организации белковых структур. На первом уровне иерархии белков стоят L-аминокислот, а иерархия ДНК начинается с D-углевода дезоксирибозы [158]. Концепция хирального дуализма имеет общебиологическую значимость. Хиральная иерархичность может являться глубинным основополагающим принципом функционирования молекулярно-биологических систем. В представлении о белке как о машине, механизме она является «структурной основой выделенных механических степеней свободы» [159]. В работе [159] формулируется фундаментальное положение о том, что гомохиральность первичных структур «служит депо свободной энергии», которая используется молекулой для осуществления структурных превращений [159]. Предполагается, что выявленная последовательность структурных уровней однозначно определяет траекторию, по которой происходит фолдинг макромолекулы [159].

Спиральность является частным случаем хиральности. Авторы [159] выявили следующую закономерность: разнотипные молекулы как правило взаимодействуют посредством структур, обладающих разными знаками хиральности. По предположению белок ORF2, осуществляющий транспозицию, распознает структуру, расположенную на 3’ конце L1 транспозонов, и связывается с ней. Такое взаимодействие скорее всего осуществляется посредством спиральных структур.

В работе [161] был рассмотрен возможный механизм коммутации клеток через взаимодействие спиральных структур. Мы по аналогии рассмотрим связывании белка ORF2 L1 со спиральной структурой стебель-петля на конце РНК-транскрипта. Аргументом в пользу образования именно спиральной структуры на конце РНК-транскрипта, необходимой для связывания с белком, служит повышение прочности конструкции при перекручивании нитей. В данном случае речь идет о переплетении двух участков одной и

той же нити РНК. В работе [161] подход Эйлера был использован в рассмотрении задачи о взаимодействии спиральных структур в растворе хиральных биомиметиков. По формуле Эйлера сила трения F между нитью, намотанной на стержень, к концу которой приложена сила натяжения φ, и стержнем выражается через число витков нити n и коэффициент трения нити о стержень α следующим образом [161, 162]:

$$F = \varphi(\exp(2\pi n\alpha) - 1)$$

Данное соотношение можно также применить, с некоторыми допущениями, к описанию прочности спирально перекрученных нитей.

Формула Эйлера обобщается на случай, когда нить и стержень притягиваются друг к другу с силой f на единицу длины нити. В нашем случае в качестве сил притяжения выступают Вандерваальсовы силы, действующие в том числе между нуклеотидами. Рассмотрение равновесия бесконечно малого элемента нити, с учетом силы притяжения нити к стержню, направленную по нормали приводит к выражению [161]:

$$F = (\varphi + Rf)(\exp(2\pi n\alpha) - 1)$$

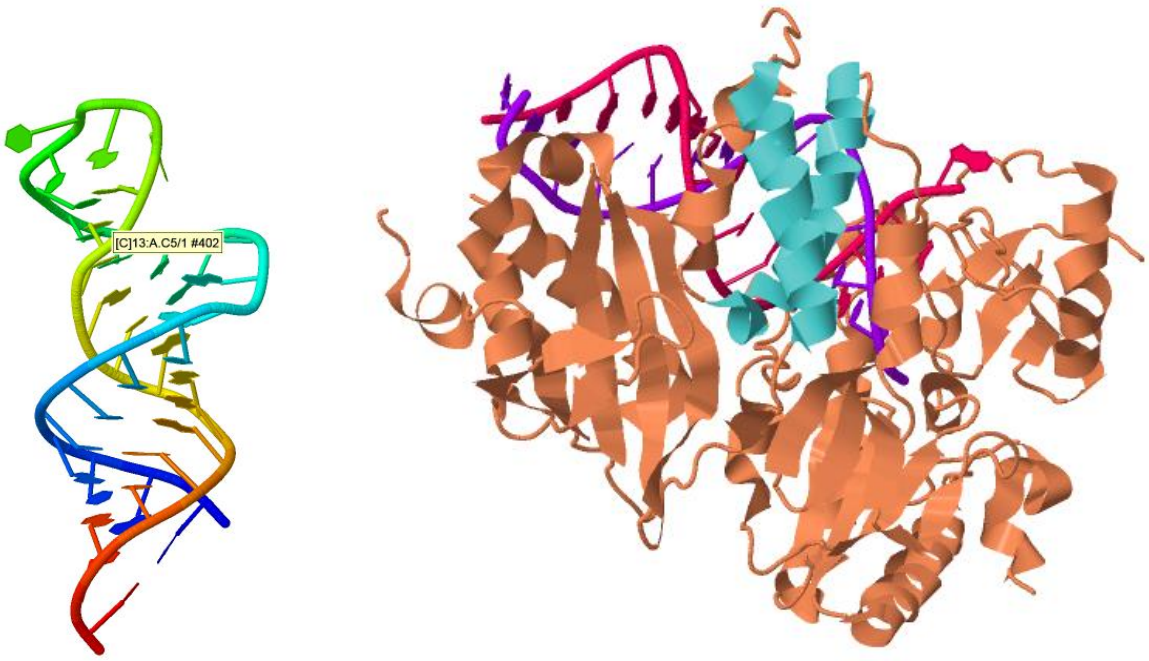
Таким образом, сила трения в системе перекрученных нитей экспоненциально зависит от числа витков спирали.

Вторичная структура стебель-петля на 3’-конце активных в настоящее время L1 транспозонов содержит, в среднем, 50 нуклеотидов, следовательно, примерно 25 нуклеотидных пар. Среднее расстояние между парами оснований для актвных транспозонов L1 составляет 0,32 нм. Из этих данных можно оценить, что длина спиральной структуры стебель-петля равна приблизительно 8 нм. Зная динуклеотидные характеристики структуры и используя программный комплекс 3DNA, мы сконструировали схематичную модель, которая позволила оценить длину витка спирали (Рис. 30). Она составила 3 нм. Таким образом, структура стебель-петля содержит два полных спиральных витка ($n = 2{,}6$). Даже при небольшом числе витков экспоненциальная зависимость силы трения от их числа обеспечивает большую прочность спиральной структуры в сравнении с параллельным расположением нитей.

Понимание топологии молекулярной системы «РНК – белок» требует анализа закрученности входящих в нее структур. Структура стебель-петля в геноме рыбы *Danio rerio*, которая распознается обратной транскриптазой, является правозакрученной (Рис. 31). Проведенный в данной работе анализ аминокислотных последовательностей белка ORF2 36 организмов, стоящих на разных ступенях эволюционной лестницы, выявил консервативность доменов эндонуклеазы и обратной транскриптазы [163]. Логично предположить, что схожесть распознающего элемента влечет и схожесть распознаваемого. Была проанализирована структура комплекса обратной транскриптазы ретровируса и

молекулы РНК, доступная в базе PDB. Общий характер укладки свидетельствует о том, что α-спирали, перекрываясь, имеют тенденцию к образованию левой суперспирали (Рис. 31).

Таким образом, правозакрученная шпилька РНК взаимодействует с белком, распознающий домен которого имеет тенденцию к образованию левой суперспирали. Это еще раз подтверждает выявленную В.А.Твердисловым и соавторами закономерность: взаимодействие разнотипных молекул осуществляется структурами, имеющими разный знак хиральности [158, 159].



A. B.

Рис. 31. А. Структура шпильки LINE транспозона в геноме рыбы *Danio rerio*, распознаваемой обратной транскриптазой. В. Комплекс двойной спирали РНК и обратной транскриптазы ретровируса. Изображения были взяты из базы данных PDB (https://www.rcsb.org/pdb/home/home.do).

3.4. Построение математической модели для автоматического распознавания вторичных структур транспозонов по их физико-химическим характеристикам

Для создания модели, способной автоматически распознавать транспозонные структуры по их физико-химическим свойствам, были использованы методы машинного обучения. Были построены два классификатора. Основой первого послужила логистическая регрессия, для второго был выбран метод опорных векторов.

Логистическая регрессия – это линейная модель классификации, позволяющая анализировать связь между несколькими независимыми переменными и одной зависимой [164] . В отличие от линейной регрессии, логистическая регрессия способна предсказывать значение бинарной переменной. В модели линейной регрессии предполагается следующий вид зависимости пременной-отклика от независимых переменных:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

где y – зависимая переменная (отклик), x_i – независимые переменные (предикторы), θ_i – неизвестные коэффициенты. Зачастую более простым и удобным является матричное представление уравнения:

$$Y = \theta^T X$$

Однако в такой модели $y \in [-\infty; +\infty]$.

В случае, если y является бинарной переменной, вместо линейной регрессии используется логистическая регрессия. Модель логистической регрессии позволяет оценить условную вероятность исхода y при условии наличия набора независимых переменных x:

$$P(y = 1|x) = p(x, \theta) = \frac{1}{1 + e^{-\theta^T x}} = g(\theta^T x)$$

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - p(x, \theta)$$

Функцию $g(\theta^T x)$ часто называют сигмоидальной или логистической функцией. Она переводит область определения функции от независимых переменных в отрезок [0;1].

Задача обучения модели ставится следующим образом: при известных y и x необходимо найти оптимальные значения коэффициентов θ. К асимптотически оптимальным значениям неизвестных параметров приводит метод максимального правдоподобия, который дает наилучшие оценки с точки зрения квадратичной функции ошибки модели. Функция правдоподобия для модели логистической регрессии имеет вид:

$$L(\theta) = \sum_{i=1}^{N} \{y_i \log p(x_i, \theta) + (1 - y_i) \log(1 - p(x_i, \theta))$$

Необходимо найти такой набор параметров θ, который максимизирует функцию правдоподобия. Для поиска экремума необходимо найти частную производную функции правдоподобия и приравнять ее к нулю:

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{N} x_i \left(y_i - p(x_i, \theta)\right) = 0$$

Решить уравнение можно методом Ньютона. Для его применения потребуется найти вторые производные, или гессиан:

$$\frac{\partial^2 L}{\partial \theta^2} = -\sum_{i=1}^{N} x_i x_i^T p(x_i, \theta)(1 - p(x_i, \theta))$$

Тогда,

$$\theta^{new} = \theta^{old} - \left(\frac{\partial^2 L}{\partial \theta \partial \theta^T}\right)^{-1} \frac{\partial L}{\partial \theta}$$

Значения частных производных рассчитывается в точке θ^{old}.

Для удобства следует переписать полученные уравнения в матричном виде. Пусть y – вектор, содержащий значения откликов y_i, X – матрица независимых переменных размерности $N \times k + 1$, где N - число наблюдений, k – число независимых переменных, p - вектор, содержащий вероятности отклика, его i-*ый* элемент равен $p(x_i, \theta^T)$, W – диагональная матрица весов размерности $N \times N$, i-*ый* элемент которой равен $p(x_i, \theta^T)(1 - p(x_i, \theta^T))$. Тогда,

$$\frac{\partial^2 L}{\partial \theta^2} = X^T(y - p)$$

$$\frac{\partial^2 L}{\partial \theta \partial \theta^T} = -X^T W X$$

Таким образом шаг алгоритма Ньютона выглядит так:

$$\theta^{new} = \theta^{old} + (X^T W X)^{-1} X^T (y - p) = (X^T W X)^{-1} X^T W\left(X\theta^{old} + W^{-1}(y - p)\right)$$
$$= (X^T W X)^{-1} X^T W z$$

Полученное выражение соответствует формуле для поиска коэффициентов методом наименьших квадратов с весами, если рассматривать $z = \left(X\theta^{old} + W^{-1}(y - p)\right)$ как отклик. Задача решается итерационно, причем на каждом шаге неизвестные параметры находятся при помощи метода наименьших квадратов с весами:

$$\theta^{new} = arg\min_{\theta}(z - X\beta)^T W (z - X\beta)$$

Коэффициенты θ характеризуют вклад соответствующей независимой переменной в разделяющую способность модели. Таким образом можно оценить, насколько важна та или иная переменная для предсказания исхода.

Еще одним методом, позволяющим определить важность переменных, является случайный лес. Случайный лес – это совокупность множества решающих деревьев. Окончательное решение о принадлежности наблюдения к тому или иному классу принимается путем голосования. Принимается исход, выбранный большинством деревьев.

Пусть имеется набор из N элементов, а также заданы M независимых переменных (предикторов, признаков). Каждое дерево строится независимо от других по следующему алгоритму:

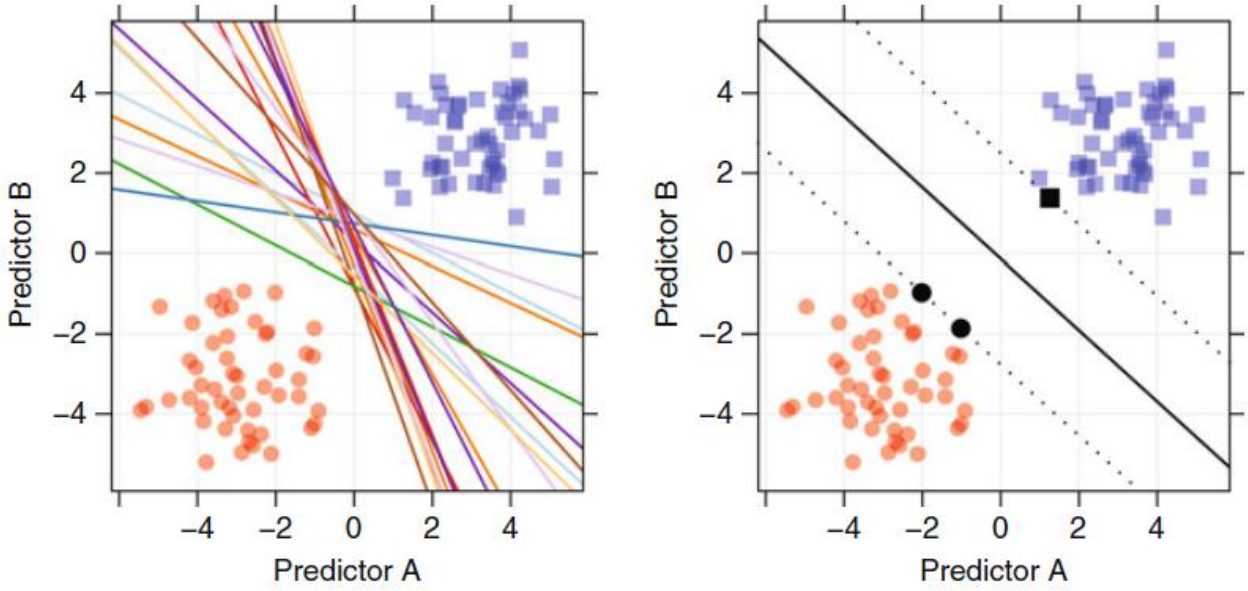
1. Из исходных данных генерируется случайное подмножество из N элементов с повторением.
2. Конструируется дерево, классифицирующее элементы данного подмножества. При создании каждой следующей точки ветвления (вершины) выбираются независимые переменные (предикторы), на основе которых может быть осуществлено разделение. Для выбора независимых переменных используется не полный набор M предикторов, а m случайно отобранных признаков из него.

Выбор наилучших предикторов может осуществляться разными способами. В исходной версии алгоритма для этой цели использовался критерий Джини. Он же используется для определения наиболее важных признаков, которые внесли основной вклад в разделяющую способность модели:

$$I_G(p) = 1 - \sum_{i=1}^{K} p_i^2,$$

где K – число классов, $i \in \{1,2, \ldots . K\}$ и p_i - доля элементов, принадлежащих к классу i, в листе дерева [164].

Метод опорных векторов позволяет разделить имеющиеся в наборе данных объекты на классы путем построения *(M-1)*-мерной гиперповерхности в пространстве признаков (характеристик, предикторов) размерности *M [164]*. Таких гиперповерхностей может быть построено много. Поэтому выбирается гиперповерхность, которая максимизирует размер области разделения между классами (Рис. 32).



A. B.

Рис. 32. А) Данные полностью разделимы. Существует бесконечное число плоскостей, разделяющих их безошибочно. В) Плоскость, максимизирующая размер зазора между объектами, принадлежащими разным классам. Черным цветом обозначены так называемые опорные векторы [165].

Постановка задачи и ее решение с точки зрения математики приведены ниже. Общая линейная модель классификации выглядит следующим образом:

$$a(x) = sign\left(\sum_{i=1}^{n} w_i x^i - w_0\right),$$

где $x = (x^1, \dots, x^n)$ – объект, который необходимо классифицировать, представленный в векторной форме

$w = (w_1, \dots, w_n), w_0$ – веса, параметры модели классификации.

$(x_1, y_1), \dots, (x_m, y_m)$ – набор данных, на котором осуществляется обучение модели. Объект x_i принадлежит тому или иному классу $y_i \in \{-1,1\}$.

Метод опорных векторов максимизирует размер разделяющей классы области. Его можно описать следующим набором уравнений:

$$\begin{cases} \|w\| \to min \\ y_i\big((w, x_i) - w_0\big) \geq 1, i = 1, \dots, m \end{cases}$$

Используя теорему Куна-Таккера, можно показать, что данный набор уравнений эквивалентен дуальной задачи Лагранжа:

$$\begin{cases} \mathcal{L}(w, w_0; \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \lambda_i \big(y_i\big((w, x_i) - w_0\big) - 1\big) \to \min\limits_{w, w_0} \min\limits_{\lambda} \mathcal{L} \\ \lambda_i \geq 0, i = 1, \dots, m \\ \lambda_i = 0 \ \vee (w, x_i) - w_0 = y_i, \ i = 1, \dots, m \end{cases},$$

где $\lambda = (\lambda_1, \dots, \lambda_m)$ – множители Лагранжа. Тогда искомые веса можно представить следующим образом:

$$\begin{cases} w = \sum\limits_{i=1}^{m} \lambda_i y_i x_i \\ w_0 = (w, x_i) - y_i, \forall i: \lambda_i \neq 0 \end{cases}$$

Поставленную задачу можно переформулировать таким образом, чтобы в конечном выражении присутствовали только множители Лагранжа:

$$\begin{cases} \mathcal{L}(\lambda) = \sum\limits_{i=1}^{m} \lambda_i - \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} \lambda_i \lambda_j y_i y_j \big(x_i, x_j\big) \to \max\limits_{\lambda} \mathcal{L} \\ \lambda_i \geq 0, i = 1, \dots, m \\ \sum\limits_{i=1}^{m} \lambda_i y_i = 0 \end{cases}$$

Построенный классификатор в результате выглядит так:

$$a(x) = sign\left(\sum_{\substack{i=1 \\ \lambda_i \neq 0}}^{m} \lambda_i y_i (x_i, x) - w_0 \right)$$

Для оценки эффективности моделей были использованы следующие характеристики:

Точность (ACC) – процент правильно предсказанных наблюдений. ACC рассчитывается по формуле:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Чувствительность (SN) – процент верно предсказанных событий (true positive). Рассчитывается по формуле:

$$SN = \frac{TP}{TP + FN}$$

Специфичность (SP) – процент верно предсказанных элементов, не являющихся событием (true negative). Рассчитывается по формуле:

$$SP = \frac{TN}{TN + FP},$$

где TN – true negative, TP – true positive, FN – false negative, FP – false positive.

Также важными характеристиками модели являются ROC кривая и площадь под ней (AUC). ROC кривая – это зависимость SN от 1 – SP. Если AUC близко к единице, то классификатор считается очень хорошим, если оно близко к 0,5, то классификатор делает предсказание случайным образом.

В построенной модели логистической регрессии в качестве предикторов были использованы динуклеотидные характеристики. Полученный классификатор показал высокую эффективность. Его точность составила 89% при сравнении структур из транспозонов и случайных мест генома. В Табл. 7 приведены другие параметры, на Рис. 33 приведены ROC кривые. Полученная модель с высокой точностью определяет вторичные структуры, принадлежащие 3’-концу транспозонов в геноме человека.

Модель	Точность	Специфичность	Чувствительность	AUC
SVM	0.94	0.94	0.93	0.98
Логистическая регрессия	0.89	0.89	0.89	0.95

Табл. 7. Параметры модели.

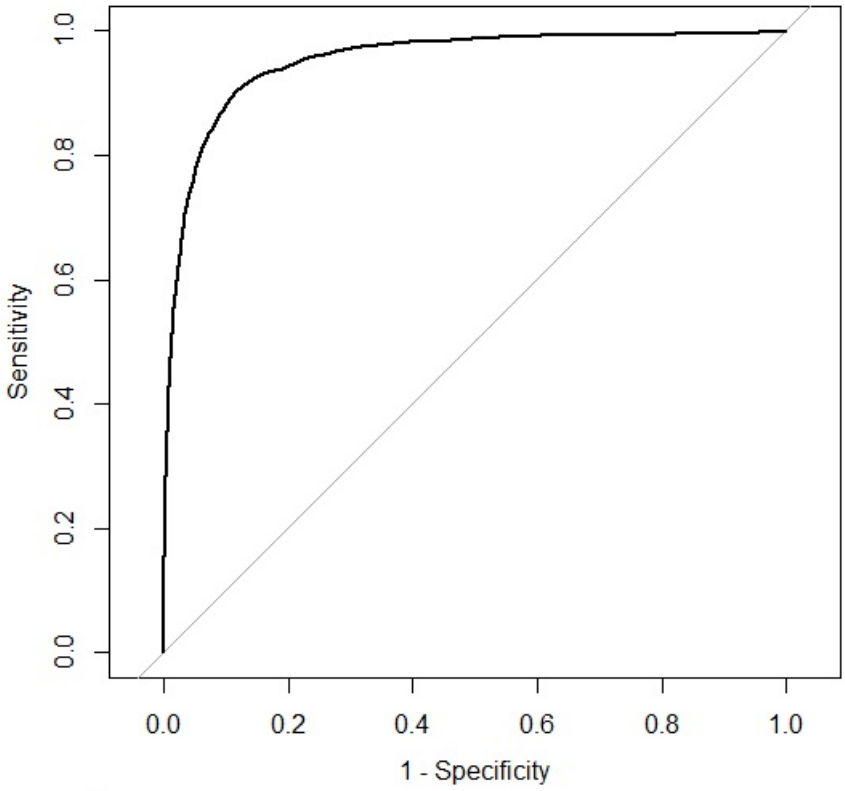


Рис. 33. ROC кривая модели логистической регрессии.

Для построения второго классификатора был использован метод опроных векторов. Каждой последовательности структуры стебель-петля была сопоставлена точка в 10-мерном пространстве характеристик. Построенный классификатор в результате выглядит так:

$$a(x) = sign\left(\sum_{\substack{i=1 \\ \lambda_i \neq 0}}^{m} \lambda_i K(x_i, x) - w_0\right),$$

где λ_i – множители Лагранжа, w_0 – константа, $K(x_i, x) = \exp(-\gamma|x_i - x|^2)$ – ядро, x_i – векторы обучающей выборки ($\lambda_i \neq 0$ только для опорных векторов), x - классифицируемый вектор.

Два настраиваемых параметра модели, γ, входящий в ядро, а также C - коэффициент регуляризации, приняли значения соответственно 0,026 и 16. Значение w_0 составило 9,5. В качестве признаков, по которым определяется принадлежность исследуемой структуры транспозону, были использованы следующие характеристики – средние значения энергии Гиббса, энтальпии, энтропии, гидрофильности динуклеотидных пар в последовательности, а также средние значения каждой из шести динуклеотидных геометрических характеристик последовательности. Была построена 9-ти мерная гиперповерхность, разделяющая

структуры, принадлежащие 3'-концу L1 транспозонов человека, и структуры, взятые из случайных мест генома или сгенерированные случайным образом. Данная гиперповерхность определяется 1513 опорными векторами и соответствующими им 1513 множителями Лагранжа (Рис. 34).

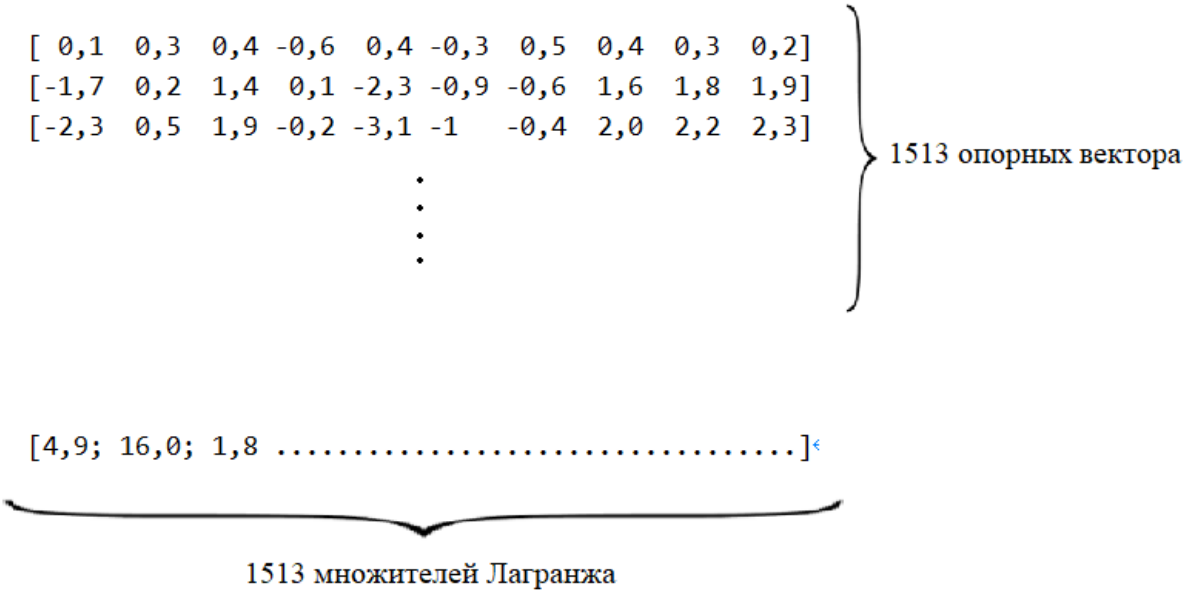


Рис. 34. Примеры найденных опорных векторов и множителей Лагранжа, определяющих 9-ти мерную разделяющую гиперповерхность.

Точность построенного классификатора составила 94%. В Табл. 7 приведены параметры модели. На Рис. 35 приведена ROC кривая.

Видно, что построенная модель оказалась эффективной. Она позволяет с высокой точностью определять вторичные структуры, принадлежащие 3’-концу транспозонов в геноме человека. Модель способна распознавать такие вторичные структуры в любой заданной последовательности РНК.

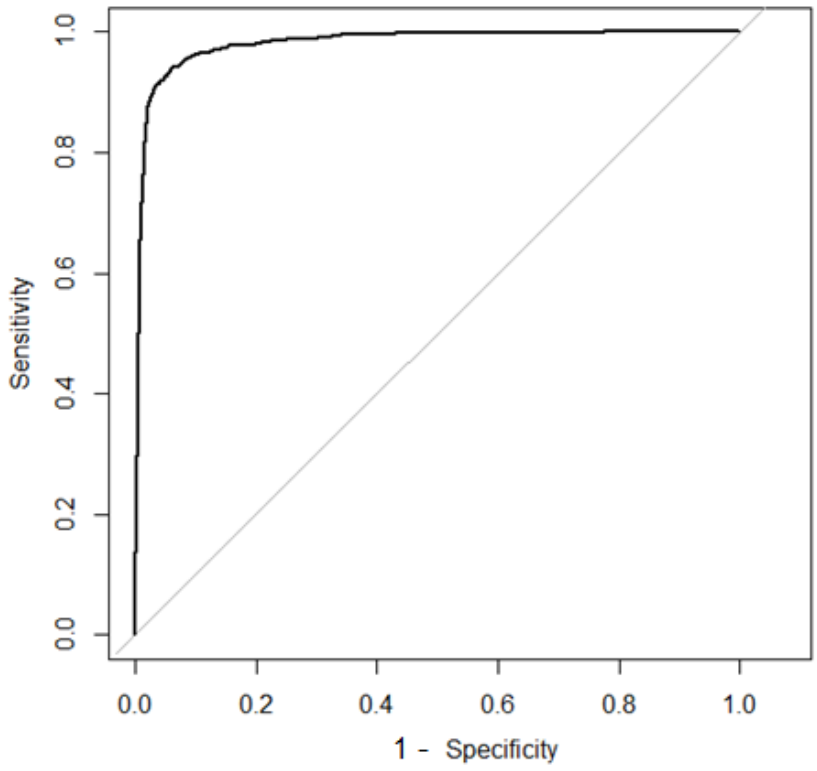


Рис. 35. ROC кривая модели опорных векторов.

Поскольку гипотеза о статистически значимой разнице между характеристиками структур из транспозонов и структур из случайных мест генома была принята, следующим шагом анализа является определение характеристик, наиболее отличающихся между двумя выше упомянутыми классами структур. Метод опорных векторов не обладает функционалом для определения величин вкладов каждой характеристики в способность математической модели к классификации. Для этого мы использовали алгоритм машинного обучения Случайный лес, позволяющий количественно оценить важность (разделяющую способность) каждой независимой переменной (в нашем случае – каждой характеристики). Результаты представлены в Табл. 8. Наибольший вклад в распознавание структур внесли гидрофильность, энтальпия, параметры Rise и Twist.

Таким образом, построенные в работе модели машинного обучения показали, что транспозонные структуры стебель-петля с высокой точностью могут быть распознаны по физическим характеристикам и отличаются от структур стебель-петля из других участков генома. Как показал анализ литературы, взаимодействие ORF2 белка с РНК транскриптом транспозона обладает высокой степенью специфичности. В таком случае структура, распознаваемая белком, должна обладать особым выделенным набором свойств. Полученный в данной работе результат как раз показывает, что вторичные структуры из транспозонов обладают таким набором. Это является косвенным доказательством

выдвинутой нами гипотезы о роли вторичной структуры на конце транспозона в распознавании белком.

Обученные модели также позволяют находить подобные структуры в любой заданной последовательности РНК. Для определения важнейших для связывания с белком геометрических характеристик РНК необходимо дальнейшее экспериментальное исследование пространственной структуры комплекса обратной транскриптазы с РНК транспозона. Выявленные в данной работе свойства могут иметь большое значение в биоинженерии для встраивания любых заданных последовательностей в геном.

Характеристика	Относительные вклады характеристик в способность математической модели к классификации, %
Гидрофильность	42
Энтальпия	19
Rise	13
Twist	12
Энтропия	8
Shift	4
Slide	3
Другие характеристики	<3

Табл. 8. Разделяющая способность характеристик.

Выявление с помощью программных методов наиболее значимых характеристик, необходимых для определения принадлежности той или иной структуры стебель-петля к классу вторичных структур на 3’ конце РНК L1 транспозона, является важной задачей при изучении процесса связывания РНК-транскрипта с белками. Зная, какие именно характеристики определяют взаимодействие с белком, можно регулировать этот процесс. Кроме того, подобная информация может оказаться полезной в задачах моделирования, например, для оценки вероятности взаимодействия с белком. В данной работе демонстрируется возможность применения методов машинного обучения для решения этой

задачи. Наибольший вклад в распознавание отличий структур транспозонов от структур, взятых из случайных областей генома, вносят четыре характеристики – гидрофильность, энтальпия и геометрические параметры Rise, Twist. Гидрофильность, и параметр Rise имеют меньшие значения для транспозонных структур. Параметр Twist, наоборот, максимален. Предполагается, что сохранение этих свойств структур транспозонов в группе активных транспозонов неслучайно, и именно они, а не последовательности, находятся под действием положительного эволюционного отбора, что в свою очередь может являтся свидетельством наличия у вышеупомянутых структур функциональной нагрузки.

Заключение к Главе 3

Подробно рассмотрены физические динуклеотидные характеристики обнаруженной на 3'-конце L1 транспозонов структуры стебель-петля. Они включают в себя термодинамические параметры, такие как свободная энергия Гиббса, энтальпия, энтропия, а также гидрофильность и шесть геометрических параметров структуры РНК - Shift, Slide, Rise, Tilt, Roll, and Twist. В ходе проведенного сравнения свойств транспозонных структур со свойствами структур из случайных областей генома, а также структур, сгенерированных случайным образом, выяснилось, что транспозонные структуры статистически значимо отличаются от своих не транспозонных аналогов. Более того, эти различия удалось количественно охарактеризовать и, используя полученные данные, сделать выводы о наличии структурных особенностей, которые могут быть необходимы для связи с белком. Например, транспозонные структуры демонстрируют более низкие значения параметра Roll по сравнению с другими подобными структурами из генома, а также структурами, сгенерированными случайным образом. Это говорит о меньшей склонности транспозонных структур к изгибу. Низкие значения параметра Roll в сочетании с отрицательными значениями параметры Slide указывают на наличие плавного небольшого изгиба в сторону большой бороздки. Более высокие значения параметра Twist для транспозонных структур свидетельствуют о большой степени закрученности. Используя геометрические динуклеотидные характеристики, удалось построить схематичные модели структур стебель-петля из транспозонов и случайных мест генома. Размеры малой и большой бороздок и длина витка при сравнении оказались ниже для транспозонных структур, чем для структур из случайных областей генома. Полученные длины витка, $2{,}9 \pm 0{,}2$ нм для транспозонной структуры и $3{,}3 \pm 0{,}2$ нм для структуры из генома, хорошо согласуются с экспериментальными данными для двойной спирали РНК - $3{,}1 \pm 0{,}3$. Значения свободной

энергии, энтальпии и энтропии для структур из транспозонов ниже, чем для структур из случайных областей генома и случайно сгенерированных структур. Таким образом, транспозонные структуры являются более стабильными, чем остальные исследованные структуры. Вышеперечисленные особенности указывают на выделенное положение транспозонных структур среди других геномных структур стебель-петля. Статистически значимые отличия свойств структур стебель-петля из L1 транспозонов от структур, сгенерированных случайным образом говорят о том, что наличие таких выделенных свойств скорее всего неслучайно. Описанные физическо-химические характеристики структур стебель-петля на 3'-конце L1 транспозонов человека подвержены действию эволюционного отбора. Это является аргументом в пользу выдвинутой в настоящей работе гипотезы о необходимости вторичной структуры на конце транспозона для специфичного распознавания доменом обратной транскриптазы белка ORF2 L1 транспозона. По указанным физико-химическим характеристикам структур стебель-петля на 3’-конце активного L1 транспозона с помощью методов машинного обучения возможно определять структуры стебель-петля со схожими свойствами в любой заданной последовательности РНК. Более того, удалось выявить свойства структур, вносящие наибольший вклад в разделяющую способность построенных моделей машинного обучения. Иными словами, найдены харакетристики, по которым в первую очередь можно отличать транспозонные структуры от всех иных. Определение ключевых характеристик вторичной структуры РНК, связывающейся с обратной транскриптазой, может иметь важное практическое применение в бионженерии для встраивания заданных последовательностей в геном, а также в задачах молекулярного моделирования, например, для оценки вероятности взаимодействия с белком. Также следует отметить, что примененный в данной работе подход к анализу найденной in silico структуры может быть полезен для решения аналогичных задач, касающихся любых типов структур и областей генома, а также для проведения так называемого разведочного анализа, целью которого является выявление наличия или отсутствия структурных особенностей, выделенных значений физико-химических характеристик. В последующей работе необходимо провести более детальное исследование структуры стебель-петля, расположенной на 3’-конце транспозонов, с привлечением экспериментальных методик и, возможно, методов молекулярного моделирования. Однако выявление основных закономерностей, структурных особенностей, возможное при помощи описанных и примененных в данной работе процедур, имеет важное значение для разработки стратегии исследования тех или иных структур в геноме.

Заключение

До сих пор остается открытым вопрос о том, каким образом ORF2 белок транспозонов класса LINE распознает с высокой специфичностью свой собственный РНК-транскрипт, а также РНК-транскрипты транспозонов-паразитов класса SINE, не кодирующих белки, но способных подставлять свой РНК-шаблон молекулярному аппарату LINE и, таким образом, увеличивать число своих копий в геноме. Ранее экспериментально была показана необходимость вторичной структуры стебель-петля на 3’-конце РНК-транскрипта для распознавания белком ORF2 в геномах трёх организмов. В данной работе разработан метод поиска консервативных вторичных структур стебель-петля, при помощи которого структура такого типа была обнаружена на 3’-конце РНК последовательностей L1 и Alu транспозонов человека, а также в других LINE и SINE транспозонах более чем в 161 организме. Всего было проанализировано более 460 000 последовательностей.

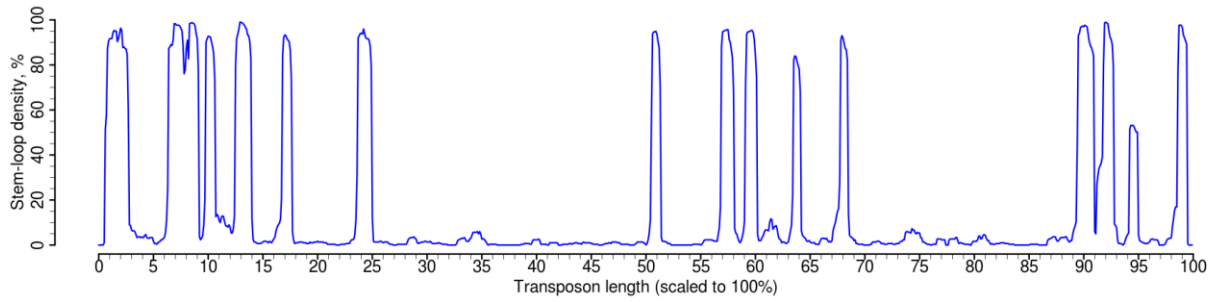
ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ ДИССЕРТАЦИИ

1. С целью выяснения роли вторичных структур в процессе транспозиции проведен анализ LINE и SINE транспозонов. Выявлено наличие консервативной вторичной структуры на 3’-конце LINE и SINE транспозонов человека, а также других организмов на всем дереве жизни.
2. Рассчитаны физико-химические и геометрические параметры структур стебель-петля на 3'-конце L1 транспозонов человека. Определены термодинамические характеристики: энергия Гиббса, энтальпия, энтропия, а также гидрофильность и шесть геометрических параметров структуры РНК: Shift, Slide, Rise, Tilt, Roll, and Twist. Показано, что по данным параметрам структуры на 3'-конце L1 транспозонов человека статистически значимо ($p < 0.05$) отличаются от структур стебель-петля из других областей генома, а также структур стебель-петля, сгенерированных случайным образом.
3. На основании физико-химических и геометрических характеристик структур стебель-петля на 3’-конце L1 транспозонов предложена математическая модель, определяющая структуры стебель-петля со схожими свойствами в любой заданной последовательности РНК. Построенная модель верно классифицировала 94% структур из тестовой выборки.
4. Выявлен вклад каждой характеристики в способность математической модели к классификации структур стебель-петля. Показано, что гидрофильность, энтальпия и параметр Rise являются наиболее значимыми характеристиками для определения принадлежности той или иной структуры стебель-петля к классу вторичных структур на 3’ конце РНК L1 транспозона. Вклады указанных характеристик составили соответственно 42%, 19%, 13%.

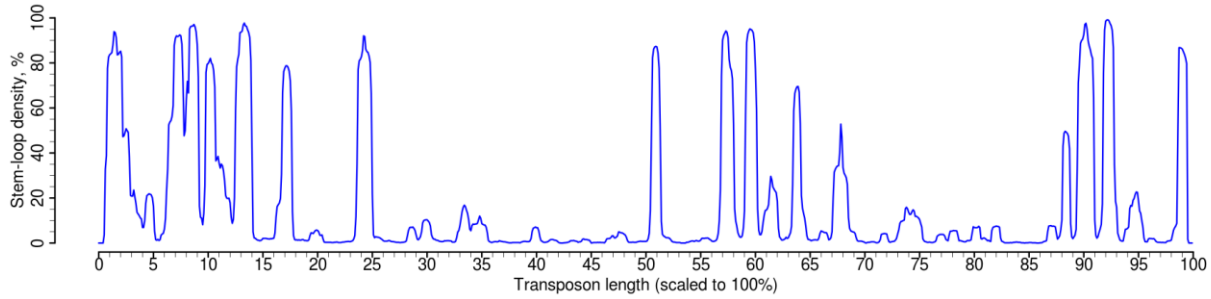
Приложения

Приложение 1. Профили покрытия для 27 подсемейств транспозонов L1 в геноме человека.

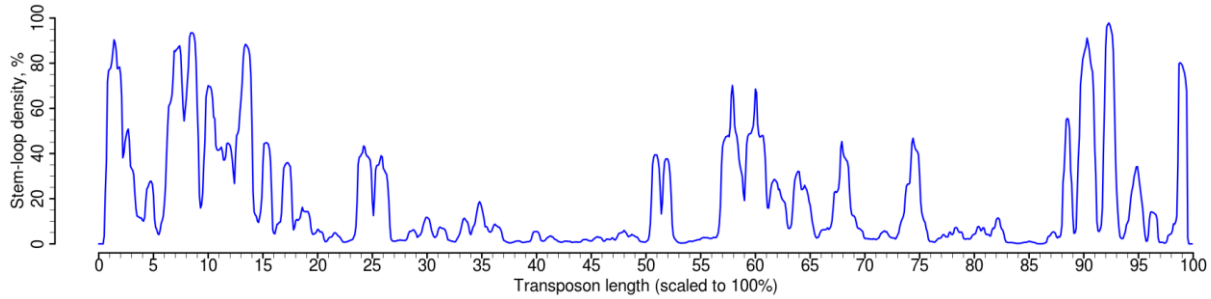
L1PA1



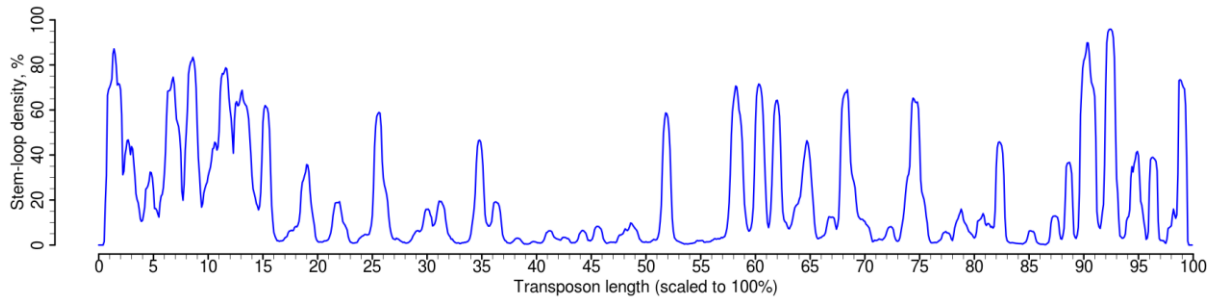
L1PA2



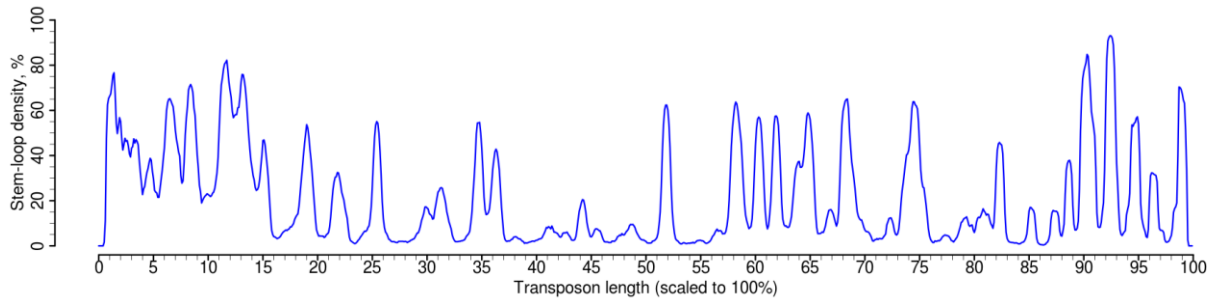
L1PA3



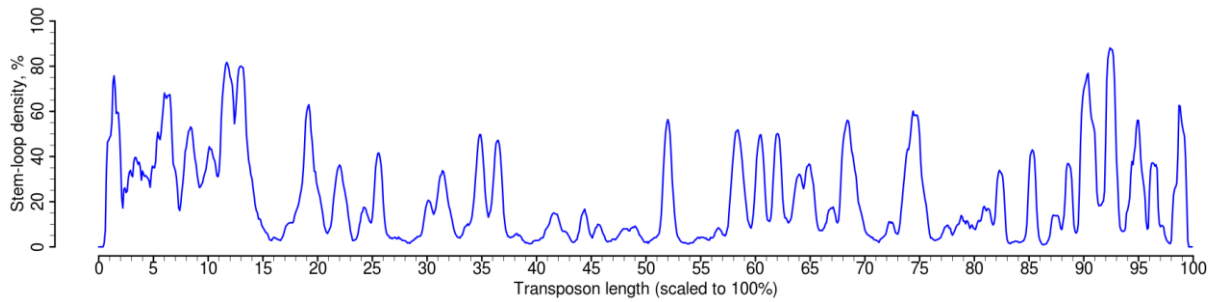
L1PA4



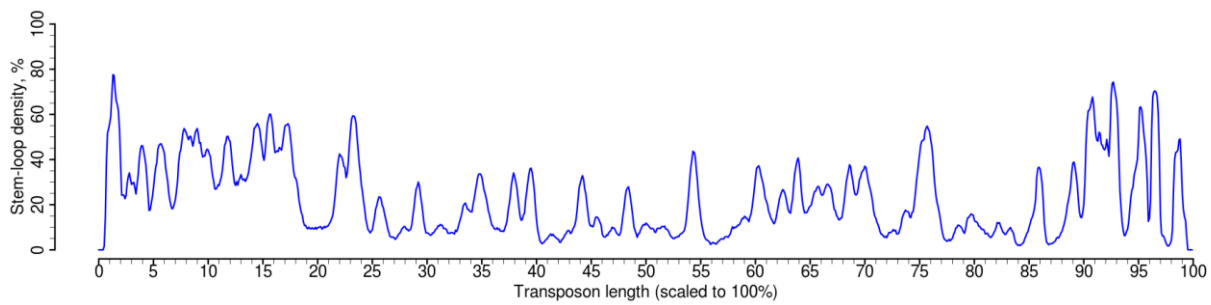
L1PA5



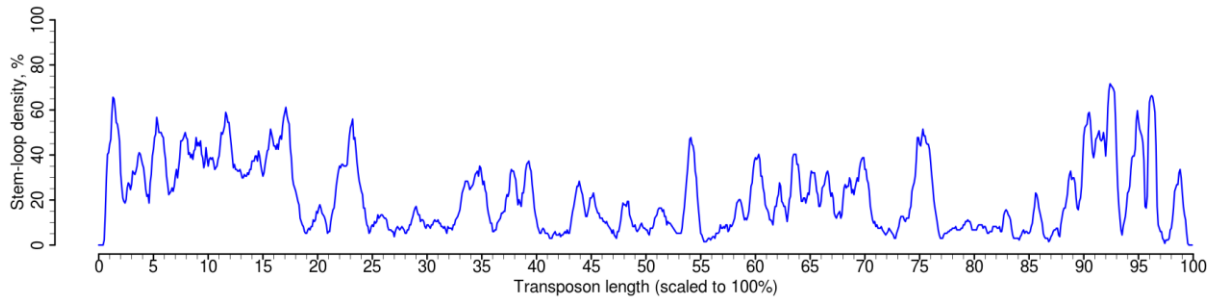
L1PA6



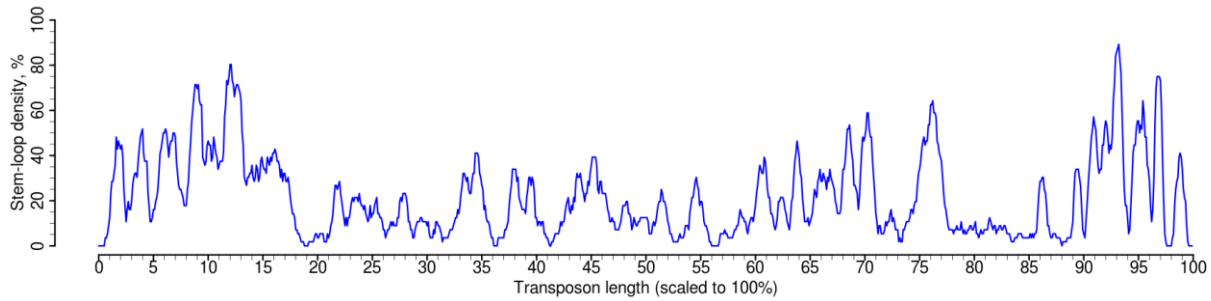
L1PA7



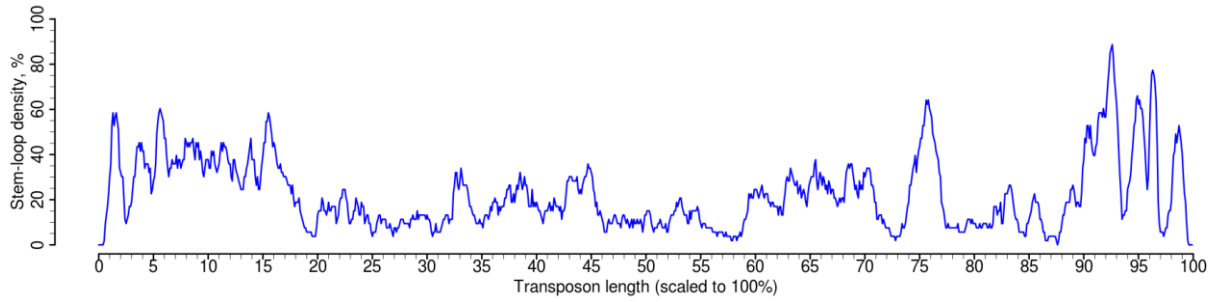
L1PA8



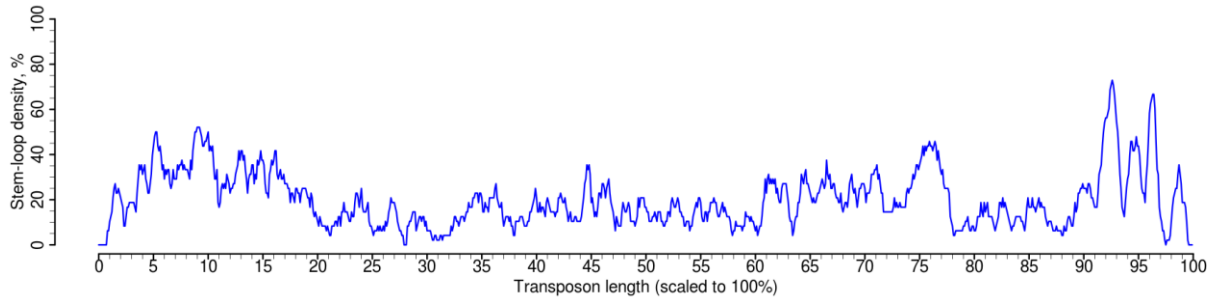
L1PA8A



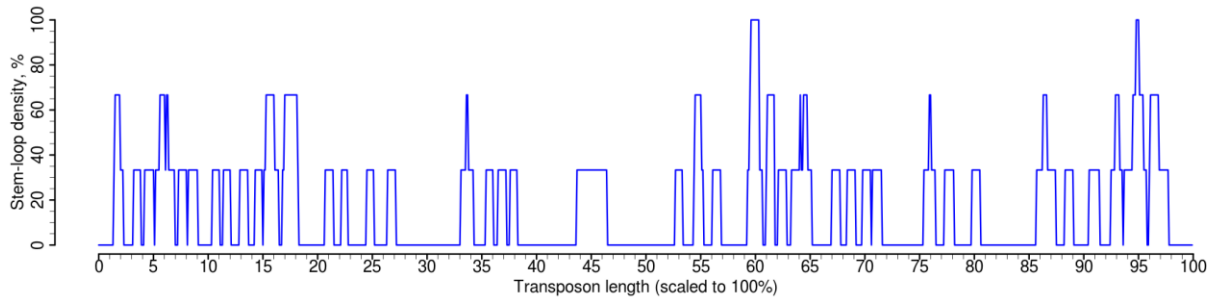
L1PA10



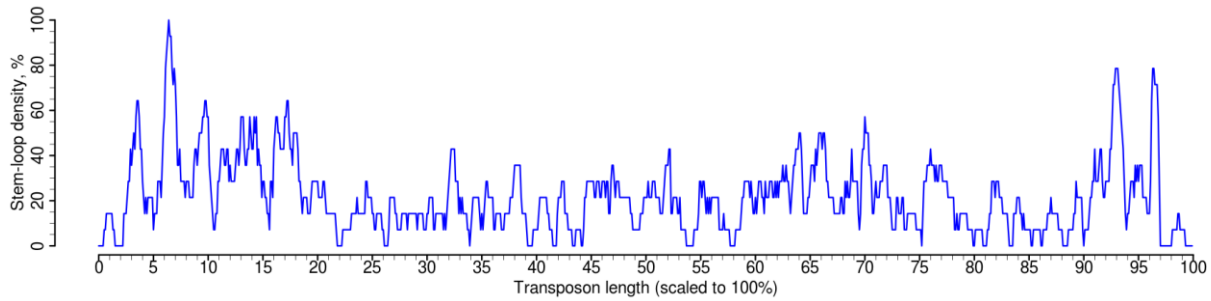
L1PA11



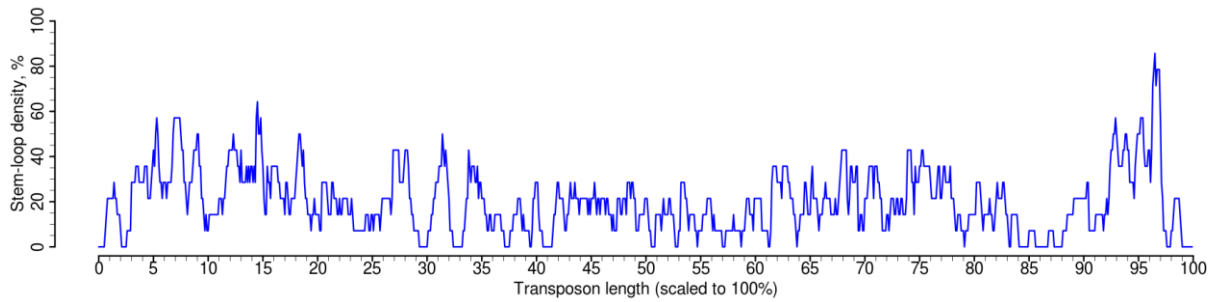
L1PA12



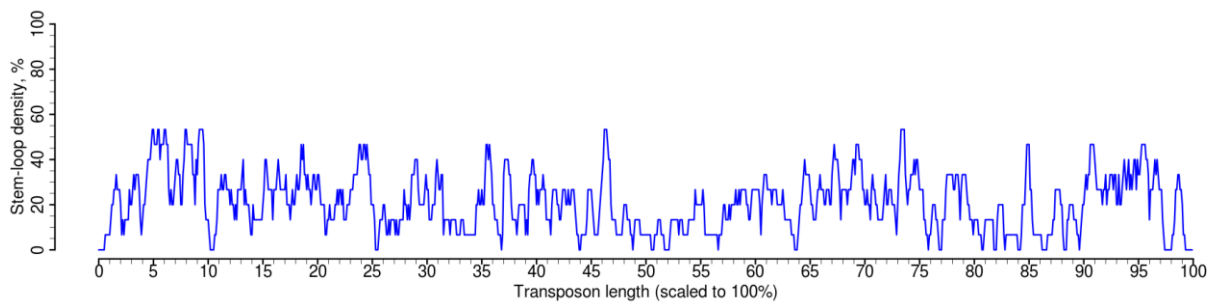
L1PA13A



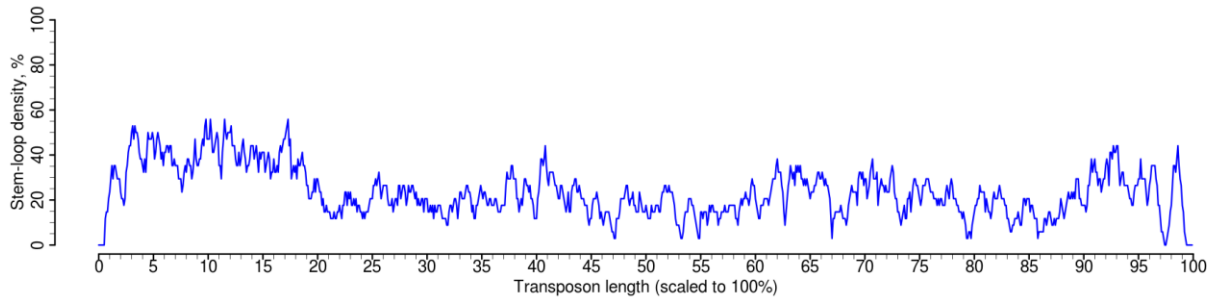
L1PA13B



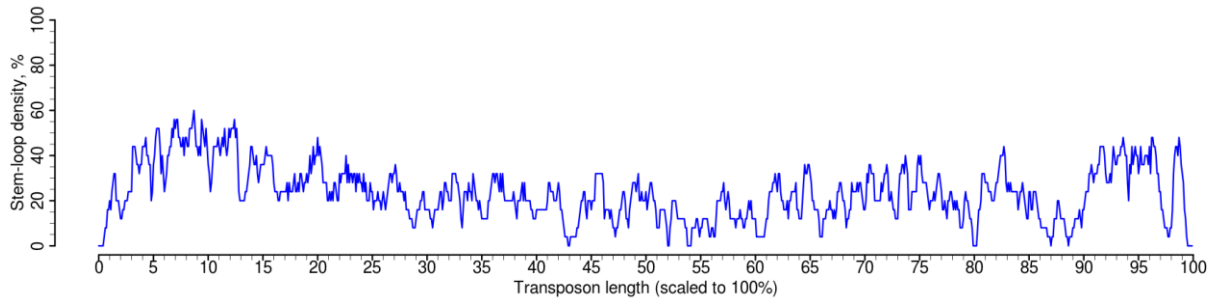
L1PA14



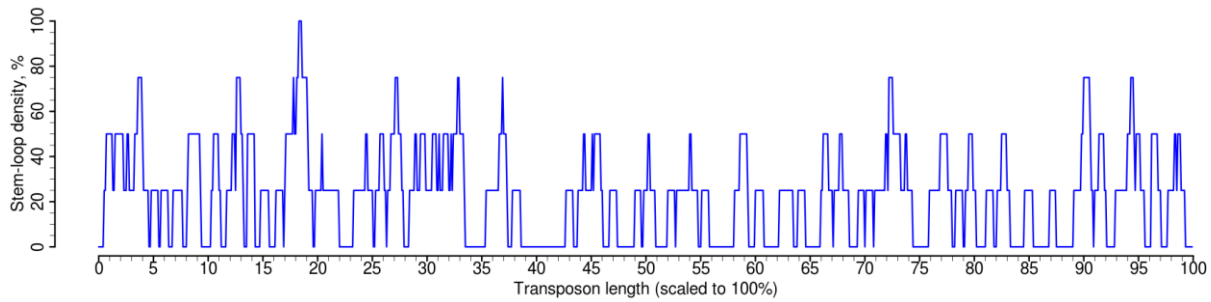
L1PA15



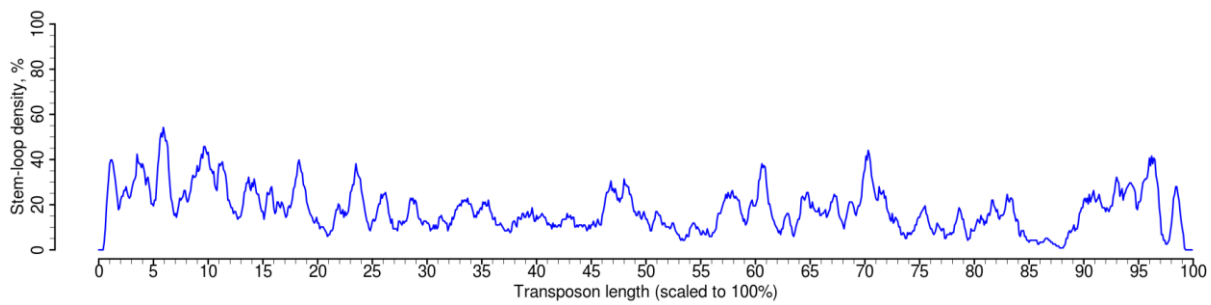
L1PA16



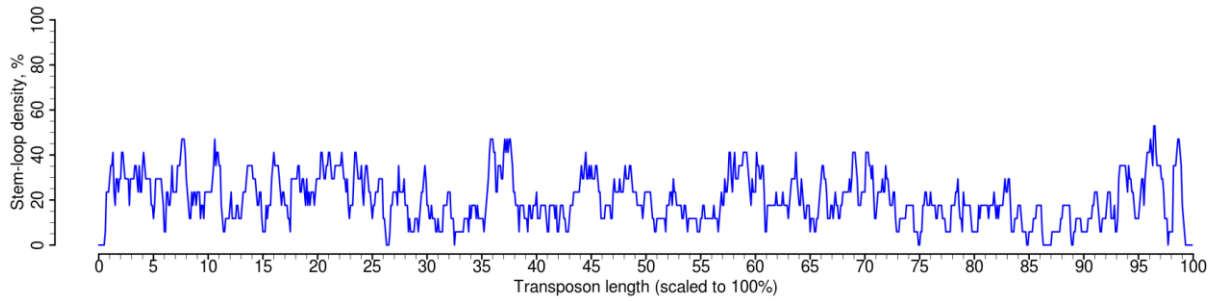
L1PA17



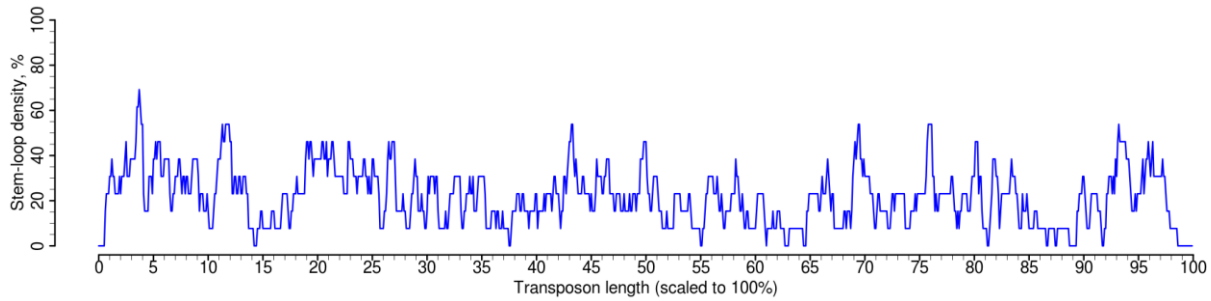
L1PB1



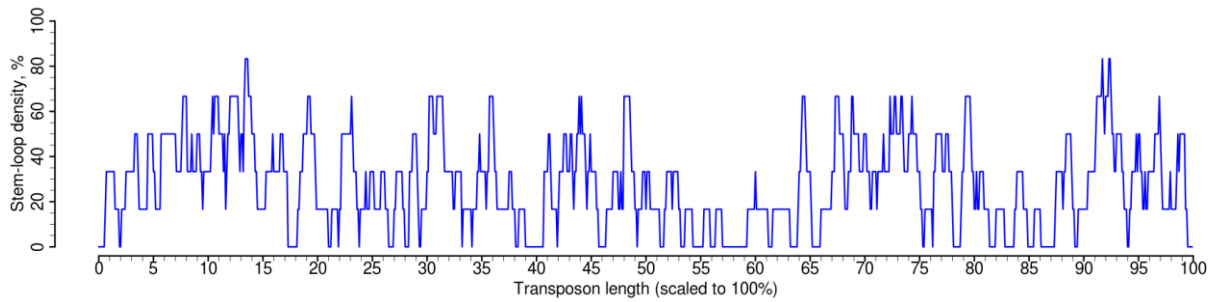
L1PB2



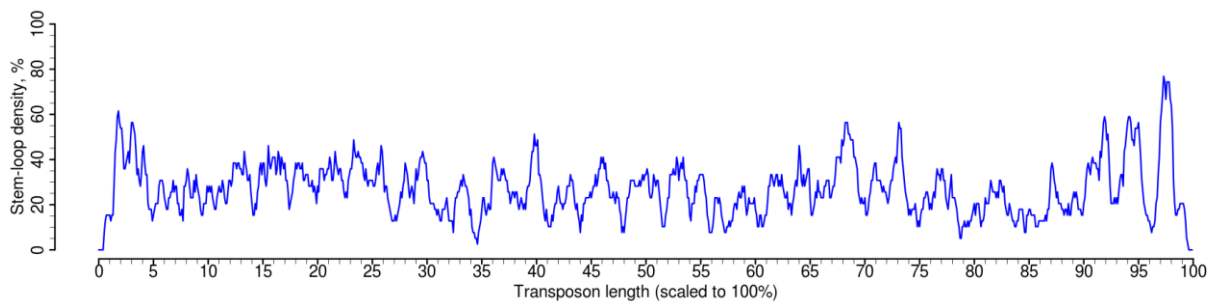
L1PB3



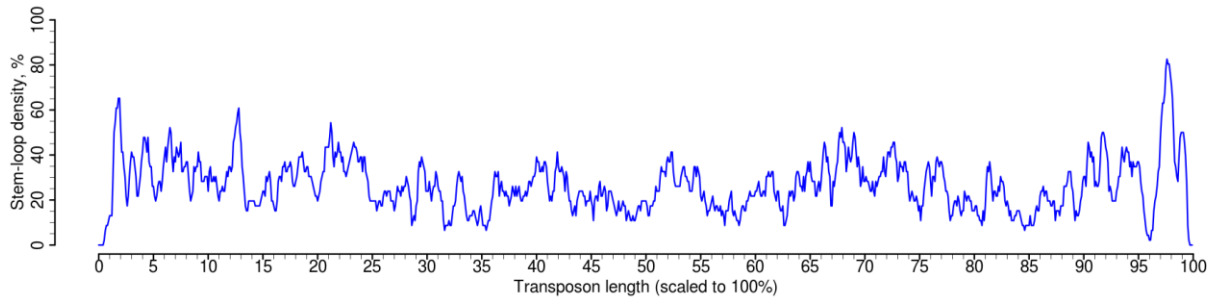
L1PB4



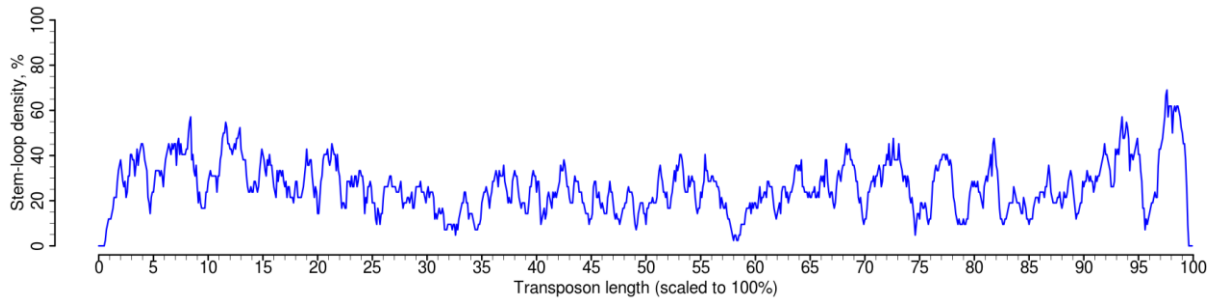
L1MA1



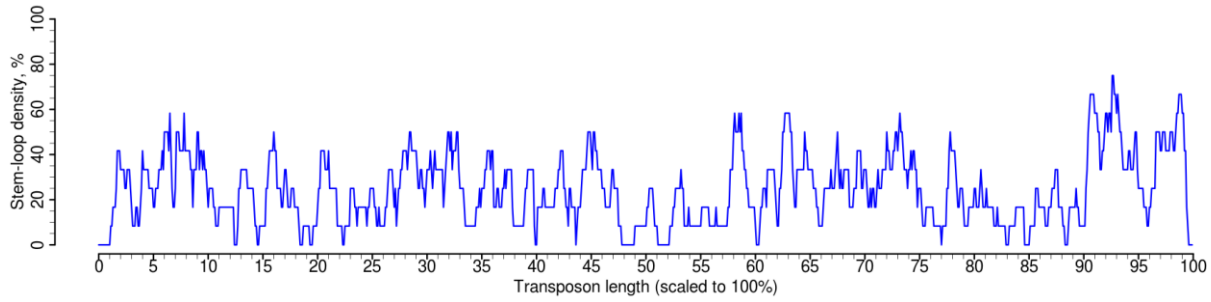
L1MA2



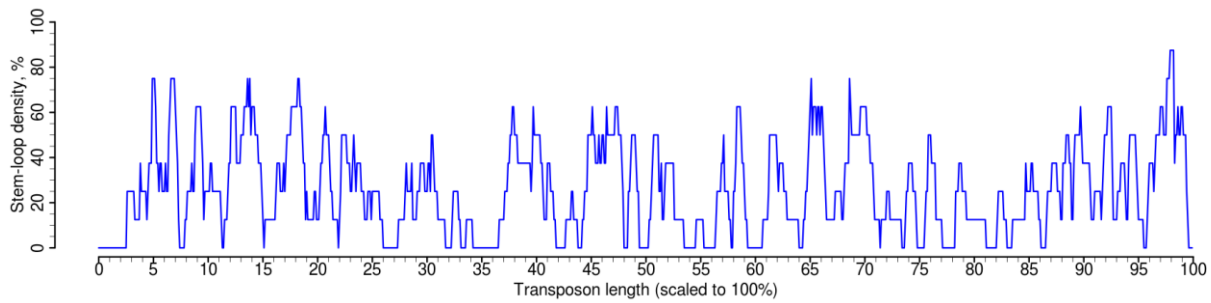
L1MA3



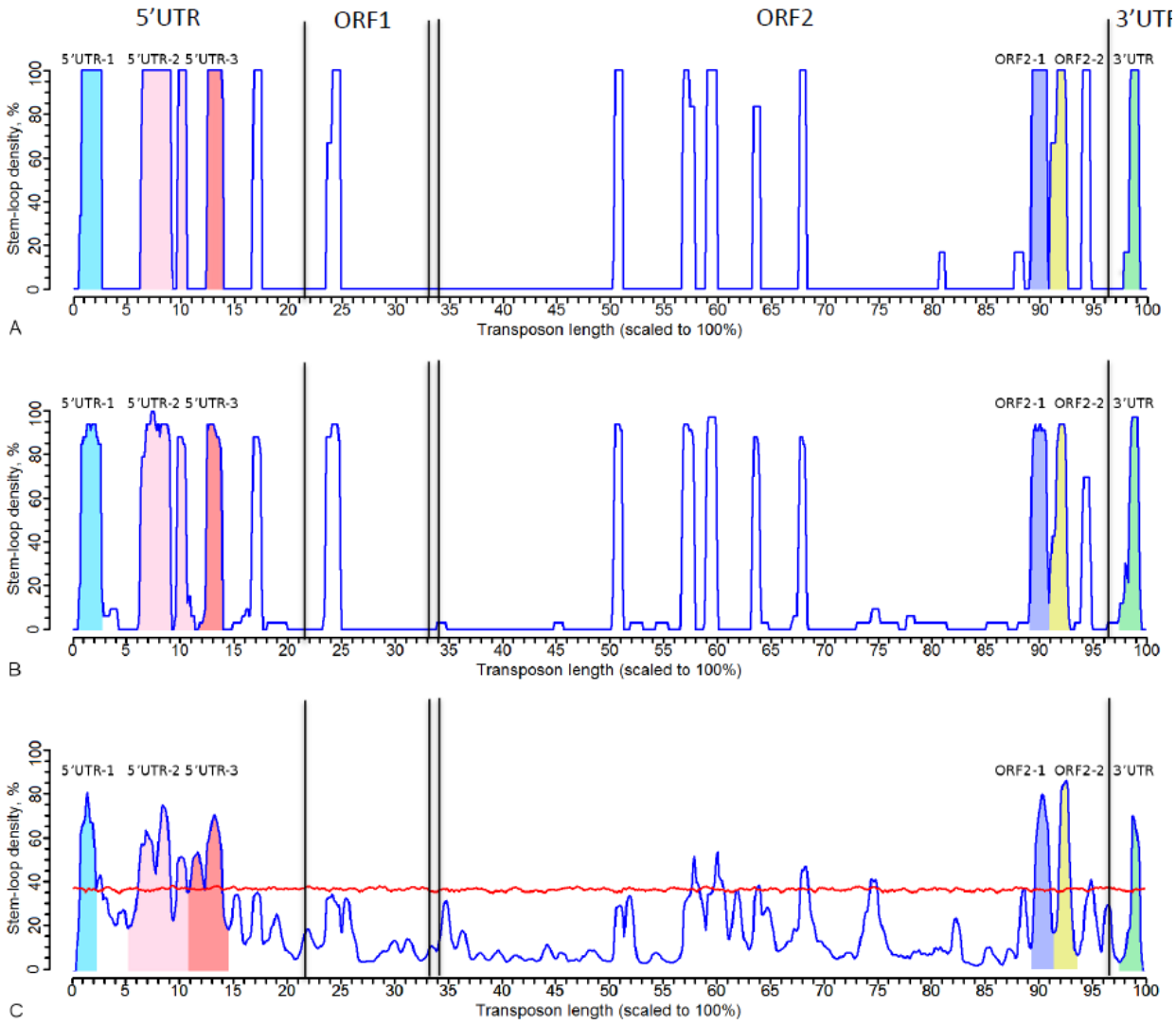
L1MA4



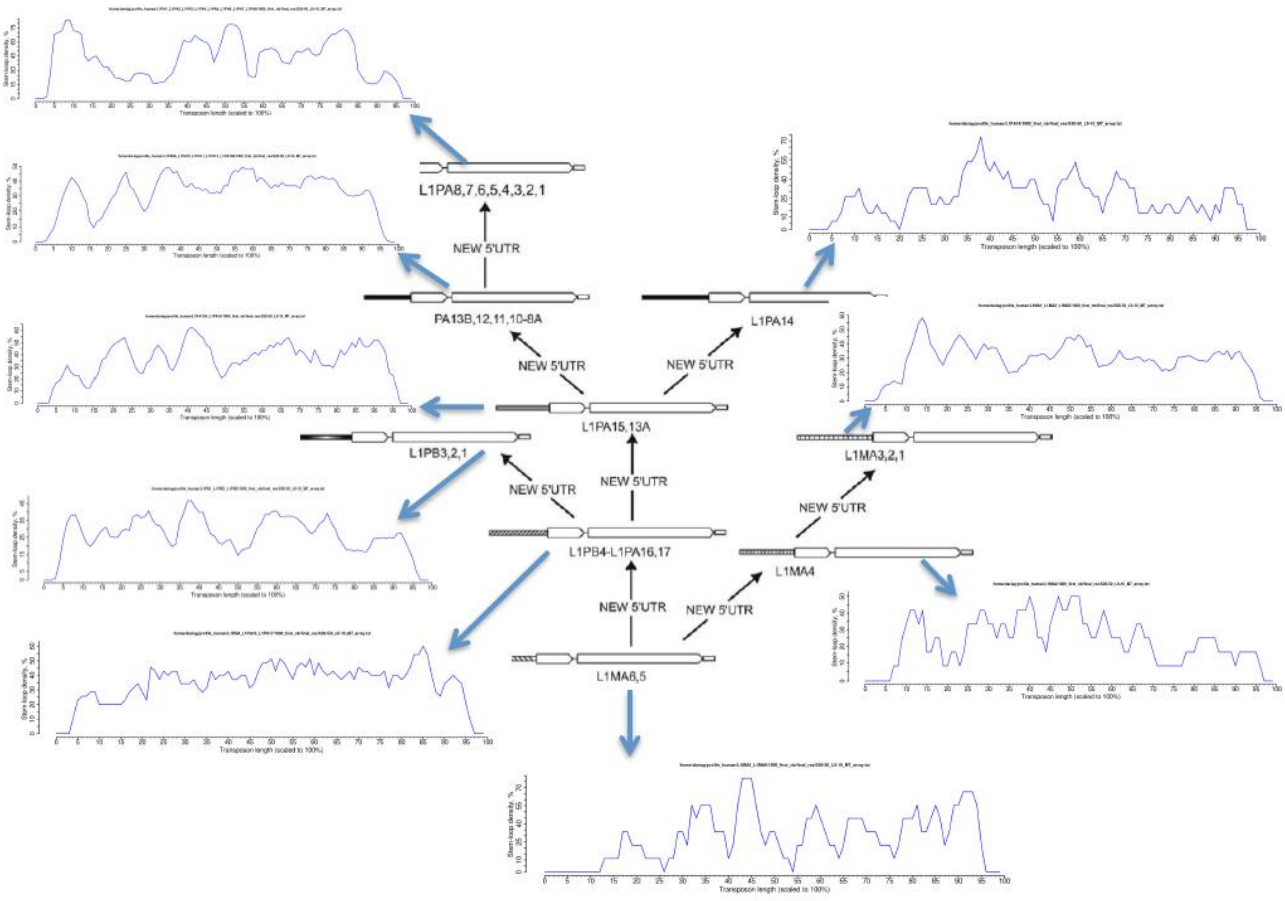
L1MA5



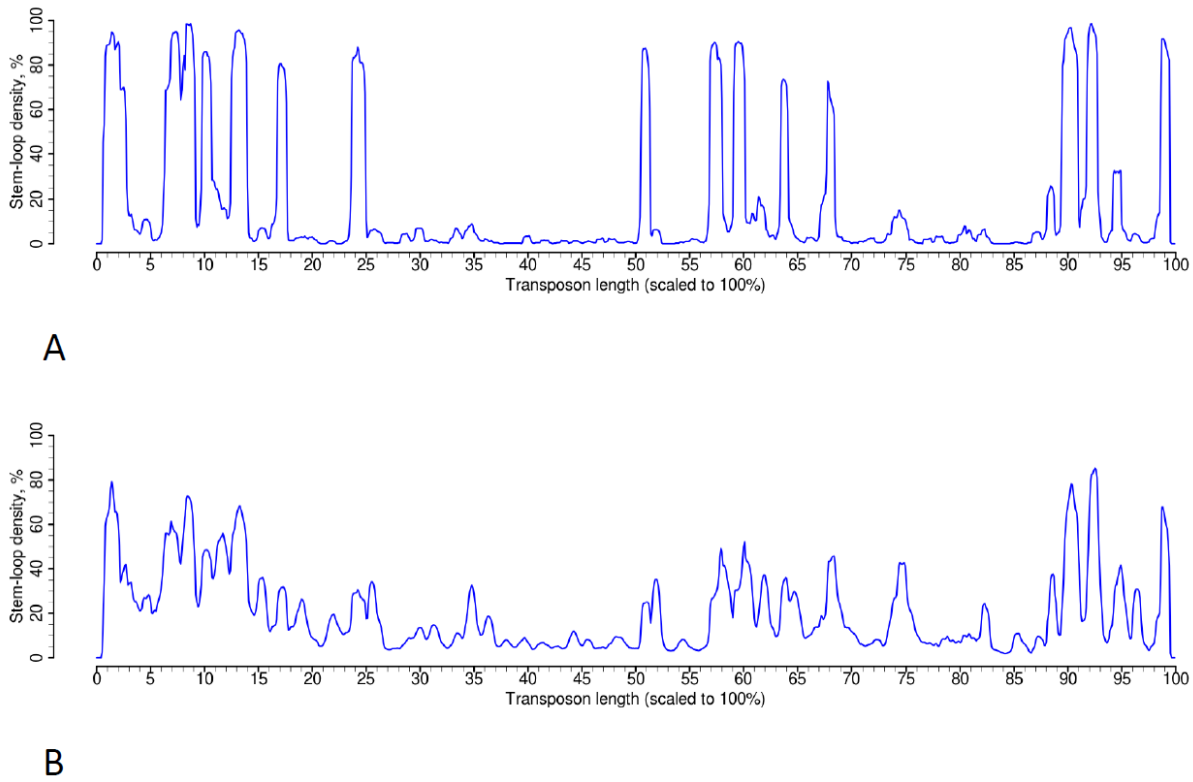
Приложение 2. Профили покрытия активных в настоящее время транспозонов L1 и хорошо сохранившихся 6622 транспозонов L1. A) Набор из 6 наиболее активных транспозонов L1; B) Набор из 33 менее активных транспозонов L1; C) Набор из 6622 хорошо сохранившихся транспозонов.



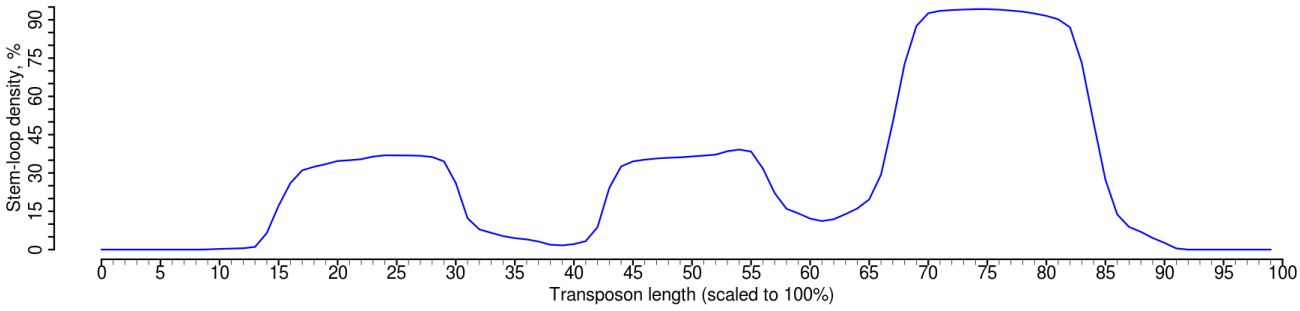
Приложение 3. Профили покрытия 5’ нетранслируемых областей групп подсемейств L1, имеющих один и тот же тип данной области (В ходе эволюции 5’ нетранслируемые области L1 транспозонов несколько раз менялись. Под типом понимается группа областей, обладающих сходством на уровне последовательности нуклеотидов.)



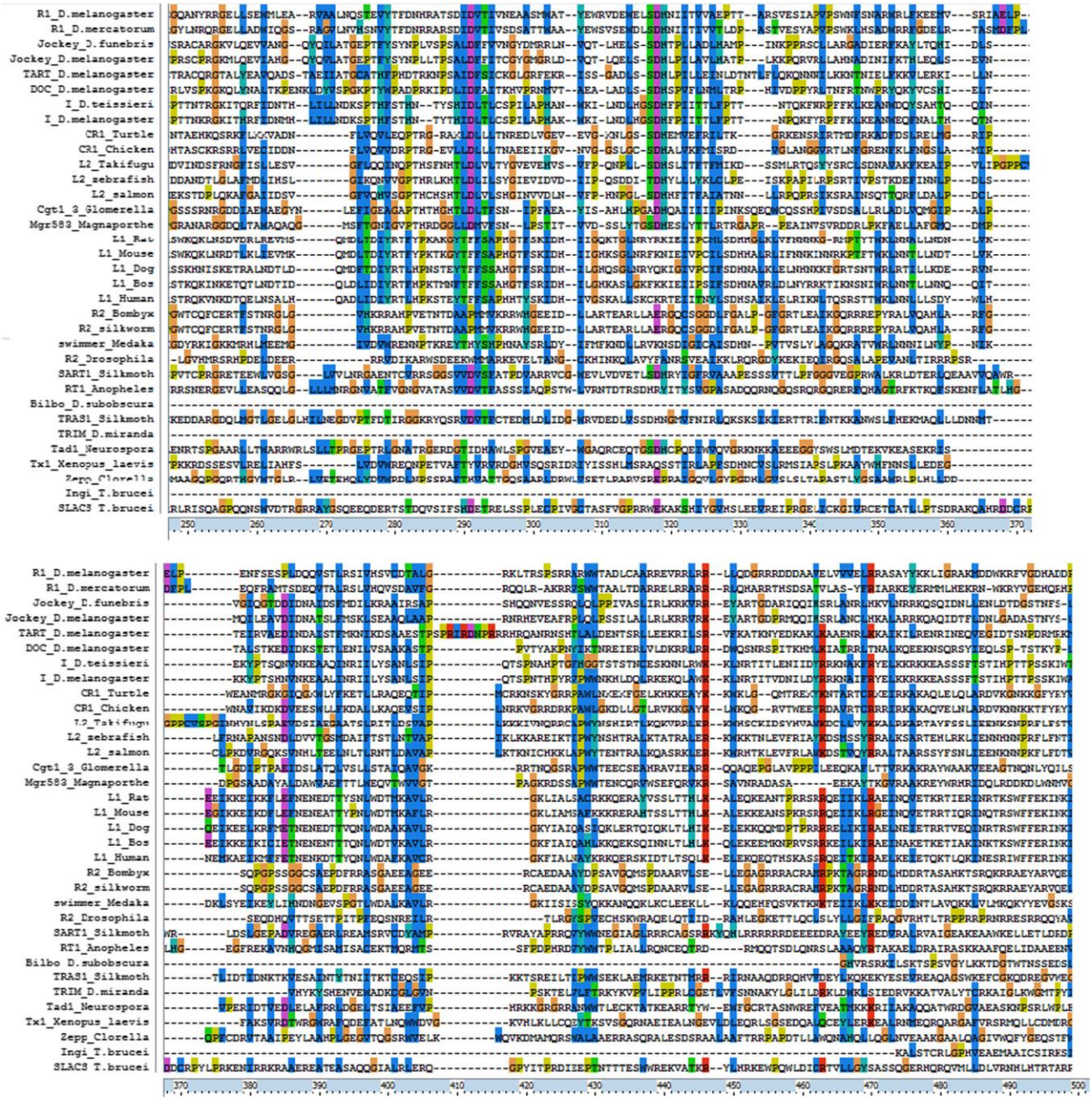
Приложение 4. Профили покрытия для L1 транспозонов, кодирующие области которых хорошо сохранились, а также для транспозонов, накопивших значительные мутации в кодирующих участках. A) Набор транспозонов, сохранивших интактные ORF участки; B) Набор транспозонов, накопивших мутации в ORF областях.



Приложение 5. Профиль покрытия 401 242 транспозона Alu.



Приложение 6. Выравнивание участка ORF2 белка LINE транспозонов различных организмов. было показано, что именно этот участок связывается со структурой стебель-петля в геноме рыбы полосатый данио. Проведенный анализ, помимо L1, включил также L2, R1, R2, CR1, I, Jockey и др. семейства.



Благодарности

Выражаю глубокую благодарность Всеволоду Александровичу Твердислову за развитие и расширение исследуемой задачи, существенно повысивших результативность и цельность работы, за всестороннюю поддержку и руководство проведенными исследованиями.

Выражаю глубокую благодарность Марии Сергеевне Попцовой за постановку интересной задачи, обсуждение полученных результатов, всестороннюю поддержку и руководство в ходе выполнения работы.

Приношу особую благодарность Леониду Владимировичу Яковенко за оказанную поддержку, детальное ознакомление с работой и исключительно ценные замечания и советы при ее обсуждении.

Выражаю глубокую признательность Алле Эдуардовне Сидоровой за всестороннюю помощь, ценные консультации, подробные разъяснения и поддержку.

Благодарю сотрудников кафедры биофизики физического факультета МГУ имени М.В.Ломоносова за внимание к работе, теплую атмосферу и поддержку.

Список литературы

1. Bevilacqua, P.C. and J.M. Blose, *Structures, kinetics, thermodynamics, and biological functions of RNA hairpins.* Annu Rev Phys Chem, 2008. **59**: p. 79-103.
2. Svoboda, P. and A. Di Cara, *Hairpin RNA: a secondary structure of primary importance.* Cell Mol Life Sci, 2006. **63**(7-8): p. 901-8.
3. Brazda, V., et al., *Cruciform structures are a common DNA feature important for regulating biological processes.* BMC Mol Biol, 2011. **12**: p. 33.
4. Warf, M.B. and J.A. Berglund, *Role of RNA structure in regulating pre-mRNA splicing.* Trends Biochem Sci, 2010. **35**(3): p. 169-78.
5. Belasco, J.G. and C.-Y.A. Chen, *Mechanism of puf mRNA degradation: the role of an intercistronic stem-loop structure.* Gene, 1988. **72**(1-2): p. 109-117.
6. Emory, S.A., P. Bouvet, and J.G. Belasco, *A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli.* Genes & Development, 1992. **6**(1): p. 135-148.
7. Masliah, G., P. Barraud, and F.H. Allain, *RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence.* Cell Mol Life Sci, 2013. **70**(11): p. 1875-95.
8. Thomas, J.M. and P.A. Beal, *How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs.* BioEssays, 2017. **39**(4): p. 1600187.
9. Huang, C.R., K.H. Burns, and J.D. Boeke, *Active transposition in genomes.* Annu Rev Genet, 2012. **46**: p. 651-75.
10. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
11. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.
12. Beck, C.R., et al., *LINE-1 elements in structural variation and disease.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 187-215.
13. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution.* Science, 2004. **303**(5664): p. 1626-32.
14. Richardson, S.R., et al., *The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes.* Microbiol Spectr, 2015. **3**(2): p. MDNA3-0061-2014.
15. Mathias, S., et al., *Reverse transcriptase encoded by a human transposable element.* Science, 1991. **254**(5039): p. 1808-1810.
16. Ohshima, K. and N. Okada, *SINEs and LINEs: symbionts of eukaryotic genomes with a common tail.* Cytogenet Genome Res, 2005. **110**(1-4): p. 475-90.

17. Hayashi, Y., et al., *Mechanism by which a LINE protein recognizes its 3' tail RNA.* Nucleic Acids Res, 2014. **42**(16): p. 10605-17.
18. Kajikawa, M. and N. Okada, *LINEs Mobilize SINEs in the Eel through a Shared 3' Sequence.* Cell, 2002. **111**(3): p. 433-444.
19. Kawagoe-Takaki, H., et al., *Probing the secondary structure of salmon SmaI SINE RNA.* Gene, 2006. **365**: p. 67-73.
20. Nomura, Y., et al., *Solution structure and functional importance of a conserved RNA hairpin of eel LINE UnaL2.* Nucleic Acids Res, 2006. **34**(18): p. 5184-93.
21. Moran, J.V., et al., *High Frequency Retrotransposition in Cultured Mammalian Cells.* Cell, 1996. **87**(5): p. 917-927.
22. Roy-Engel, A.M., *A tale of an A-tail: The lifeline of a SINE.* Mob Genet Elements, 2012. **2**(6): p. 282-286.
23. Harada, S., M. Uchida, and N. Shimizu, *Episomal high copy number maintenance of hairpin-capped DNA bearing a replication initiation region in human cells.* J Biol Chem, 2009. **284**(36): p. 24320-7.
24. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
25. Szak, S.T., et al., *Molecular archeology of L1 insertions in the human genome.pdf.* Genome Biology, 2002. **3**(10): p. research0052.1.
26. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
27. Kazazian, H.H., Jr., *Mobile DNA transposition in somatic cells.* BMC Biol, 2011. **9**: p. 62.
28. Becker, K.G., et al., *Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element.* Human Molecular Genetics, 1993. **2**(10): p. 1697-1702.
29. Swergold, G.D., *Identification, characterization, and cell specificity of a human LINE-1 promoter.* Molecular and Cellular Biology, 1990. **10**(12): p. 6718-6729.
30. Holmes, S.S., M; Swergold, G, *Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human line-1 transposable element.pdf.* Journal of Biological Chemistry, 1992. **267**(28).
31. Khazina, E., et al., *Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition.* Nat Struct Mol Biol, 2011. **18**(9): p. 1006-14.
32. Khazina, E. and O. Weichenrieder, *Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame.* Proc Natl Acad Sci U S A, 2009. **106**(3): p. 731-6.

33. Januszyk, K., et al., *Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1.* J Biol Chem, 2007. **282**(34): p. 24893-904.
34. Feng, Q., et al., *Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition.* Cell, 1996. **87**(5): p. 905-916.
35. Usdin, K.F., AV, *The structure of the guanine-rich polypurine:polypyrimidine sequence at the right end of the rat L1 (LINE) element.* J Biol Chem, 1989. **264**(26).
36. Doucet, Aurélien J., et al., *A 3′ Poly(A) Tract Is Required for LINE-1 Retrotransposition.* Molecular Cell, 2015. **60**(5): p. 728-741.
37. Vassetzky, N.S. and D.A. Kramerov, *SINEBase: a database and tool for SINE analysis.* Nucleic Acids Research, 2013. **41**(D1): p. D83-D89.
38. Ullu, E. and C. Tschudi, *Alu sequences are processed 7SL RNA genes.* Nature, 1984. **312**(5990): p. 171-172.
39. Sinnett, D., et al., *Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units.* J Biol Chem, 1991. **266**(14): p. 8675-8.
40. Dewannieux, M., C. Esnault, and T. Heidmann, *LINE-mediated retrotransposition of marked Alu sequences.* Nature Genetics, 2003. **35**(1): p. 41-48.
41. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
42. Kramerov, D.A. and N.S. Vassetzky, *Origin and evolution of SINEs in eukaryotic genomes.* Heredity (Edinb), 2011. **107**(6): p. 487-95.
43. Veniaminova, N.A., N.S. Vassetzky, and D.A. Kramerov, *B1 SINEs in different rodent families.* Genomics, 2007. **89**(6): p. 678-86.
44. Zietkiewicz, E., et al., *Monophyletic origin of Alu elements in primates.* J Mol Evol, 1998. **47**(2): p. 172-82.
45. Vassetzky, N.S., O.A. Ten, and D.A. Kramerov, *B1 and related SINEs in mammalian genomes.* Gene, 2003. **319**: p. 149-160.
46. Kapitonov, V.V. and J. Jurka, *A novel class of SINE elements derived from 5S rRNA.* Mol Biol Evol, 2003. **20**(5): p. 694-702.
47. Gogolevsky, K.P., N.S. Vassetzky, and D.A. Kramerov, *5S rRNA-derived and tRNA-derived SINEs in fruit bats.* Genomics, 2009. **93**(5): p. 494-500.
48. Gogolevsky, K.P., N.S. Vassetzky, and D.A. Kramerov, *Bov-B-mobilized SINEs in vertebrate genomes.* Gene, 2008. **407**(1-2): p. 75-85.
49. Ostertag, E.M. and H.H. Kazazian, Jr., *Biology of mammalian L1 retrotransposons.* Annu Rev Genet, 2001. **35**: p. 501-38.

50. Esnault, C., J. Maestre, and T. Heidmann, *Human LINE retrotransposons generate processed pseudogenes.* Nat Genet, 2000. **24**(4): p. 363-7.
51. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
52. Kulpa, D.A. and J.V. Moran, *Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition.* Hum Mol Genet, 2005. **14**(21): p. 3237-48.
53. Ohshima, K., et al., *The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements.* Mol Cell Biol, 1996. **16**(7): p. 3756-64.
54. Okada, N., et al., *SINEs and LINEs share common 3' sequences: a review.* Gene, 1997. **205**(1-2): p. 229-243.
55. Cognat, V., et al., *On the evolution and expression of Chlamydomonas reinhardtii nucleus-encoded transfer RNA genes.* Genetics, 2008. **179**(1): p. 113-23.
56. Baucom, R.S., et al., *Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome.* PLoS Genet, 2009. **5**(11): p. e1000732.
57. Sugano, T., M. Kajikawa, and N. Okada, *Isolation and characterization of retrotransposition-competent LINEs from zebrafish.* Gene, 2006. **365**: p. 74-82.
58. Matveev, V., H. Nishihara, and N. Okada, *Novel SINE families from salmons validate Parahucho (Salmonidae) as a distinct genus and give evidence that SINEs can incorporate LINE-related 3'-tails of other SINEs.* Mol Biol Evol, 2007. **24**(8): p. 1656-66.
59. Osanai, M., et al., *Essential motifs in the 3' untranslated region required for retrotransposition and the precise start of reverse transcription in non-long-terminal-repeat retrotransposon SART1.* Mol Cell Biol, 2004. **24**(18): p. 7902-13.
60. Okada, N. and M. Hamada, *The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: A new example from the bovine genome.* Journal of Molecular Evolution, 1997. **44**(S1): p. S52-S56.
61. Mandal, P.K., et al., *Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles.* Hum Mol Genet, 2013. **22**(18): p. 3730-48.
62. Boeke, J.D., *LINEs and Alus--the polyA connection.* Nat Genet, 1997. **16**(1): p. 6-7.
63. Wagstaff, B.J., et al., *Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion.* PLoS Genet, 2012. **8**(8): p. e1002842.
64. Noll, A., et al., *Ancient traces of tailless retropseudogenes in therian genomes.* Genome Biol Evol, 2015. **7**(3): p. 889-900.
65. Schmitz, J., et al., *A novel class of mammalian-specific tailless retropseudogenes.* Genome Res, 2004. **14**(10A): p. 1911-5.

66. Kulpa, D.A. and J.V. Moran, *Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles.* Nat Struct Mol Biol, 2006. **13**(7): p. 655-60.
67. Cost, G.J., *Human L1 element target-primed reverse transcription in vitro.* The EMBO Journal, 2002. **21**(21): p. 5899-5910.
68. Doucet, A.J., et al., *Characterization of LINE-1 ribonucleoprotein particles.* PLoS Genet, 2010. **6**(10).
69. Millevoi, S. and S. Vagner, *Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation.* Nucleic Acids Res, 2010. **38**(9): p. 2757-74.
70. Luan, D.D. and T.H. Eickbush, *RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element.* Mol Cell Biol, 1995. **15**(7): p. 3882-91.
71. Kachroo, P., S.A. Leong, and B.B. Chattoo, *Mg-SINE: a short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea.* Proc Natl Acad Sci U S A, 1995. **92**(24): p. 11125-9.
72. Malik, H.S., W.D. Burke, and T.H. Eickbush, *The age and evolution of non-LTR retrotransposable elements.* Mol Biol Evol, 1999. **16**(6): p. 793-805.
73. Loreto, E.L., C.M. Carareto, and P. Capy, *Revisiting horizontal transfer of transposable elements in Drosophila.* Heredity (Edinb), 2008. **100**(6): p. 545-54.
74. Kordis, D. and F. Gubensek, *Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10704–10709.
75. Lazareva, E., et al., *Acquisition of Full-Length Viral Helicase Domains by Insect Retrotransposon-Encoded Polypeptides.* Front Microbiol, 2015. **6**: p. 1447.
76. Anderson, M.T. and H.S. Seifert, *Neisseria gonorrhoeae and humans perform an evolutionary LINE dance.* Mob Genet Elements, 2011. **1**(1): p. 85-87.
77. Ohshima, K., *Parallel relaxation of stringent RNA recognition in plant and mammalian L1 retrotransposons.* Mol Biol Evol, 2012. **29**(11): p. 3255-9.
78. Van Dellen, K., et al., *LINEs and SINE-like elements of the protist Entamoeba histolytica.* Gene, 2002. **297**(1-2): p. 229-239.
79. Khan, H., A. Smit, and S. Boissinot, *Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates.* Genome Res, 2006. **16**(1): p. 78-87.
80. Price, A.L., E. Eskin, and P.A. Pevzner, *Whole-genome analysis of Alu repeat elements reveals complex evolutionary history.* Genome Res, 2004. **14**(11): p. 2245-52.
81. Mathews, D.H., et al., *Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element.* RNA, 1997. **3**(1): p. 1-16.

82. Baba, S., et al., *Solution structure of an RNA stem-loop derived from the 3' conserved region of eel LINE UnaL2.* RNA, 2004. **10**(9): p. 1380-7.
83. Matetovici, I., et al., *Mobile Element Evolution Playing Jigsaw - SINEs in Gastropod and Bivalve Mollusks.* Genome Biol Evol, 2016. **8**(1): p. 253-70.
84. Calladine, C., et al., *Understanding DNA, Third Edition: The Molecule and How it Works.* Academic Press, 2004.
85. Bhattacharyya, D. and M. Bansal, *Local variability and base sequence effects in DNA crystal structures.* J Biomol Struct Dyn, 1990. **8**(3): p. 539-72.
86. Dickerson, R.E., *DNA bending: the prevalence of kinkiness and the virtues of normality.* Nucleic Acids Res, 1998. **26**(8): p. 1906-26.
87. Gorin, A.A., V.B. Zhurkin, and K. Wilma, *B-DNA Twisting Correlates with Base-pair Morphology.* Journal of Molecular Biology, 1995. **247**(1): p. 34-48.
88. Olson, W.K., et al., *DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.* Proceedings of the National Academy of Sciences, 1998. **95**(19): p. 11163-11168.
89. Kameda, T., et al., *The 1-Particle-per-k-Nucleotides (1PkN) Elastic Network Model of DNA Dynamics with Sequence-Dependent Geometry.* Front Physiol, 2017. **8**: p. 103.
90. Marathe, A. and M. Bansal, *An ensemble of B-DNA dinucleotide geometries lead to characteristic nucleosomal DNA structure and provide plasticity required for gene expression.* BMC Struct Biol, 2011. **11**: p. 1.
91. Perez, A., et al., *The relative flexibility of B-DNA and A-RNA duplexes: database analysis.* Nucleic Acids Res, 2004. **32**(20): p. 6144-51.
92. Luscombe, N.M., et al., *An overview of the structures of protein-DNA complexes.* Genome Biol, 2000. **1**(1): p. REVIEWS001.
93. Segal, E., et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.
94. Dickerson RE, B.M., Calladine CR, et al., *Definitions and nomenclature of nucleic acid structure parameters.* EMBO J. , 1989. **8**: p. 1-4.
95. Anselmi, C., et al., *A Theoretical Model for the Prediction of Sequence-Dependent Nucleosome Thermodynamic Stability.* Biophysical Journal, 2000. **79**(2): p. 601-613.
96. Berman, H.M., et al., *The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.* Biophysical Journal, 1992. **63**(3): p. 751-759.
97. Breslauer, K.J., et al., *Predicting DNA duplex stability from the base sequence.* Proc Natl Acad Sci U S A, 1986. **83**(11): p. 3746-50.

98. Hagerman, P.J., *Straightening out the bends in curved DNA.* Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression, 1992. **1131**(2): p. 125-132.
99. Crothers, D.M., et al., *DNA bending, flexibility, and helical repeat by cyclization kinetics.* 1992. **212**: p. 3-29.
100. N. Trifonov, E., *DNA in profile.* Trends in Biochemical Sciences, 1991. **16**: p. 467-470.
101. Dickerson, R.E., D. Goodsell, and M.L. Kopka, *MPD and DNA bending in crystals and in solution.* J Mol Biol, 1996. **256**(1): p. 108-25.
102. Hagerman, P.J., *Flexibility of DNA.* Annu Rev Biophys Biophys Chem, 1988. **17**: p. 265-86.
103. Werner, M.H., A.M. Gronenborn, and G.M. Clore, *Intercalation, DNA Kinking, and the Control of Transcription.* Science, 1996. **271**(5250): p. 778-784.
104. Otwinowski, Z., et al., *Crystal structure of trp repressor/operator complex at atomic resolution.* Nature, 1988. **335**(6188): p. 321-9.
105. Ulyanov, N.B. and V.B. Zhurkin, *Sequence-dependent anisotropic flexibility of B-DNA. A conformational study.* J Biomol Struct Dyn, 1984. **2**(2): p. 361-85.
106. Bolshoy, A., et al., *Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles.* Proceedings of the National Academy of Sciences, 1991. **88**(6): p. 2312-2316.
107. Ornstein, R.L. and J.R. Fresco, *Correlation of Tm and sequence of DNA duplexes with delta H computed by an improved empirical potential method.* Biopolymers, 1983. **22**(8): p. 1979-2000.
108. Barber, A.M. and V.B. Zhurkin, *CAP binding sites reveal pyrimidine-purine pattern characteristic of DNA bending.* J Biomol Struct Dyn, 1990. **8**(2): p. 213-32.
109. Berger, C., I. Jelesarov, and H.R. Bosshard, *Coupled folding and site-specific binding of the GCN4-bZIP transcription factor to the AP-1 and ATF/CREB DNA sites studied by microcalorimetry.* Biochemistry, 1996. **35**(47): p. 14984-91.
110. Olson, W.K., *The flexible DNA double helix. II. Superhelix formation.* Biopolymers, 1979. **18**(5): p. 1235-60.
111. Fratini, A.V., et al., *Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG.* J Biol Chem, 1982. **257**(24): p. 14686-707.
112. Drew, H.R. and R.E. Dickerson, *Structure of a B-DNA dodecamer.* Journal of Molecular Biology, 1981. **151**(3): p. 535-556.
113. Ghosh, A. and M. Bansal, *A glossary of DNA structures from A to Z.* Acta Crystallographica Section D Biological Crystallography, 2003. **59**(4): p. 620-626.

114. Marathe, A., D. Karandur, and M. Bansal, *Small local variations in B-form DNA lead to a large variety of global geometries which can accommodate most DNA-binding protein motifs.* BMC Struct Biol, 2009. **9**: p. 24.

115. Arnott, S. and D.W.L. Hukins, *Optimised parameters for A-DNA and B-DNA.* Biochemical and Biophysical Research Communications, 1972. **47**(6): p. 1504-1509.

116. Ivanov, V.I., et al., *The B̄ to Ā transition of DNA in solution.* Journal of Molecular Biology, 1974. **87**(4): p. 817-833.

117. Peticolas, W.L., Y. Wang, and G.A. Thomas, *Some rules for predicting the base-sequence dependence of DNA conformation.* Proceedings of the National Academy of Sciences, 1988. **85**(8): p. 2579-2583.

118. Davey, C.A., et al., *Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution.* Journal of Molecular Biology, 2002. **319**(5): p. 1097-1113.

119. Kailasam, S., D. Bhattacharyya, and M. Bansal, *Sequence dependent variations in RNA duplex are related to non-canonical hydrogen bond interactions in dinucleotide steps.* BMC Res Notes, 2014. **7**: p. 83.

120. Chen, W., et al., *iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition.* Nucleic Acids Res, 2013. **41**(6): p. e68.

121. Xiao, X., et al., *iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition.* Oncotarget, 2016. **7**(23): p. 34180-9.

122. Chen, W., et al., *iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition.* Biomed Res Int, 2014. **2014**: p. 623149.

123. Chen, W., et al., *PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions.* Sci Rep, 2016. **6**: p. 35123.

124. Barraud, P. and F.H. Allain, *ADAR proteins: double-stranded RNA and Z-DNA binding domains.* Curr Top Microbiol Immunol, 2012. **353**: p. 35-60.

125. Liu, B., F. Yang, and K.C. Chou, *2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function.* Mol Ther Nucleic Acids, 2017. **7**: p. 267-277.

126. Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic Acids Research, 1981. **9**(1): p. 133-148.

127. Minakami, R., et al., *Identification of an internal cis-element essential for the human Li transcription and a nuclear factor(s) binding to the element.* Nucleic Acids Research, 1992. **20**(12): p. 3139-3145.

128. Tchenio, T., J.F. Casella, and T. Heidmann, *Members of the SRY family regulate the human LINE retrotransposons.* Nucleic Acids Res, 2000. **28**(2): p. 411-5.
129. Yang, N., *An important role for RUNX3 in human L1 transcription and retrotransposition.* Nucleic Acids Research, 2003. **31**(16): p. 4929-4940.
130. Carpenter, C.D. and A.E. Simon, *Analysis of sequences and predicted structures required for viral satellite RNA accumulation by in vivo genetic selection.* Nucleic Acids Res, 1998. **26**(10): p. 2426-32.
131. Haasnoot, P., et al., *A conserved hairpin structure in Alfamovirus and Bromovirus subgenomic promoters is required for efficient RNA synthesis in vitro.* RNA, 2000. **6**(5): p. 708-716.
132. Sugimoto, Y., et al., *hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1.* Nature, 2015. **519**(7544): p. 491-4.
133. Bennett, E.A., et al., *Active Alu retrotransposons in the human genome.* Genome Res, 2008. **18**(12): p. 1875-83.
134. Hayashi, Y., et al., *Mechanism by which a LINE protein recognizes its 3' tail RNA.* Nucleic Acids Research, 2014. **42**(16): p. 10605-10617.
135. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
136. Seetharaman, M., et al., *Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA.* RNA, 2006. **12**(2): p. 235-47.
137. Lawrence, D.C., et al., *Structure of the Intact Stem and Bulge of HIV-1 Ψ-RNA Stem-Loop SL1.* Journal of Molecular Biology, 2003. **326**(2): p. 529-542.
138. Prislan, I., et al., *The size of the internal loop in DNA hairpins influences their targeting with partially complementary strands.* J Phys Chem B, 2015. **119**(1): p. 96-104.
139. West, N.C., et al., *Shared Protein Components of SINE RNPs.* Journal of Molecular Biology, 2002. **321**(3): p. 423-432.
140. Friedel, M., et al., *DiProDB: a database for dinucleotide properties.* Nucleic Acids Res, 2009. **37**(Database issue): p. D37-40.
141. Xia, T., et al., *Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs.* Biochemistry, 1998. **37**(42): p. 14719-35.
142. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7287-92.
143. T Dale, R.S., and M J Serra, *<A test of the model to predict unusually stable.pdf>*. 2000.

144. Hausmann, N.Z. and B.M. Znosko, *Thermodynamic characterization of RNA 2 x 3 nucleotide internal loops.* Biochemistry, 2012. **51**(26): p. 5359-68.
145. Tanaka, F., et al., *Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop.* Biochemistry, 2004. **43**(22): p. 7143-50.
146. Lu, X.J., M.A. El Hassan, and C.A. Hunter, *Structure and conformation of helical nucleic acids: analysis program (SCHNAaP).* J Mol Biol, 1997. **273**(3): p. 668-80.
147. Weber, A.L. and J.C. Lacey, *Genetic code correlations: Amino acids and their anticodon nucleotides.* Journal of Molecular Evolution, 1978. **11**(3): p. 199-210.
148. Tolstorukov, M.Y., et al., *A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning.* J Mol Biol, 2007. **371**(3): p. 725-38.
149. Stefl, R., L. Skrisovska, and F.H. Allain, *RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle.* EMBO Rep, 2005. **6**(1): p. 33-8.
150. Williamson, J.R., *Induced fit in RNA-protein recognition.* Nat Struct Biol, 2000. **7**(10): p. 834-7.
151. Thapar, R., A.P. Denmon, and E.P. Nikonowicz, *Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins.* Wiley Interdiscip Rev RNA, 2014. **5**(1): p. 49-67.
152. Wild, K., I. Sinning, and S. Cusack, *Crystal structure of an early protein-RNA assembly complex of the signal recognition particle.* Science, 2001. **294**(5542): p. 598-601.
153. Conn, G.L., *Crystal Structure of a Conserved Ribosomal Protein-RNA Complex.* Science, 1999. **284**(5417): p. 1171-1174.
154. Moras, D. and A. Poterszman, *Protein–RNA interactions: Getting into the major groove.* Current Biology, 1996. **6**(5): p. 530-532.
155. Barik, A. and R.P. Bahadur, *Hydration of protein-RNA recognition sites.* Nucleic Acids Research, 2014. **42**(15): p. 10148-10160.
156. Ares, P., et al., *High resolution atomic force microscopy of double-stranded RNA.* Nanoscale, 2016. **8**(23): p. 11818-26.
157. van der Vliet, P.C. and C.P. Verrijzer, *Bending of DNA by transcription factors.* Bioessays, 1993. **15**(1): p. 25-32.
158. Твердислов, В.А., *Хиральность как первичный переключатель иерархических уровней в молекулярно-биологических системах.* Биофизика, 2013. **58**(1): p. 159-64.
159. Твердислов, В.А., et al., *Периодическая система хиральных структур в молекулярной биологии* Биофизика, 2017. **62**(3): p. 421–434.

160. Твердислов, В.А., А.Э. Сидорова, and Л.В. Яковенко, *От симметрий – к законам эволюции. I. Хиральность как инструмент стратификации активных сред.* Биофизика, 2012. **79**(12): p. 1728–1732.
161. Стовбун, С.В., et al., *Взаимодействие спиральных структур как молекулярная основа внутри- и межклеточных коммутаций.* ВМУ. Серия 3. ФИЗИКА. АСТРОНОМИЯ., 2015(1).
162. Артоболевский, И.И., *Теория механизмов и машин. Учебник для втузов. 4-е изд.* 1988.
163. Grechishnikova, D. and M. Poptsova, *Conserved 3' UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition.* BMC Genomics, 2016. **17**(1): p. 992.
164. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* 2009.
165. Kuhn, M. and K. Johnson, *Applied Predictive Modeling*. 2013: Springer New York.