БИОФИЗИКА, 2017, том 62, вып. 6, с. 1045–1053

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

ФИЗИЧЕСКИЕ И ГЕОМЕТРИЧЕСКИЕ СВОЙСТВА СТРУКТУР «СТЕБЕЛЬ–ПЕТЛЯ» ТРАНСПОЗОНОВ ЧЕЛОВЕКА НАХОДЯТСЯ ПОД ДЕЙСТВИЕМ ЭВОЛЮЦИОННОГО ОТБОРА

© 2017 г. Д.А. Гречишникова*, М.С. Попцова* **

**Физический факультет Московского государственного университета имени М.В. Ломоносова, 119991, Москва, Ленинские горы, 1/2*

***Национальный исследовательский университет «Высшая школа экономики», 101000, Москва, ул. Мясницкая, 20*

E-mail: daria.grechishnikova@gmail.com, maria.poptsova@gmail.com

Поступила в редакцию 11.07.17 г.

Вторичные структуры РНК играют важную роль в транспозиции, в частности, в распознавании РНК белками транспозонов. Ранее мы обнаружили консервативную структуру на 3′-конце транспозонов человека и выдвинули гипотезу о ее роли в транспозиции. Несмотря на полное отсутствие сходства на уровне последовательностей, консервативное положение структуры говорит о наличии свойств, на которые действует положительный эволюционный отбор. В данной работе определены физические и геометрические свойства структур «стебель–петля» на 3′-конце транспозонов человека и произведено их сравнение со свойствами структур из других областей генома. Каждая структура «стебель–петля» была охарактеризована набором из десяти характеристик: свободной энергией Гиббса, энтальпией, энтропией, гидрофильностью, Shift, Slide, Rise, Tilt, Roll и Twist. С помощью методов машинного обучения построена модель, которая распознает структуры транспозонов по физическим и геометрическим свойствам с 94% степенью точности. Наибольший вклад в распознавание структур внесли гидрофильность, энтальпия, параметры Rise и Twist. Предполагается, что именно эти свойства структур транспозонов находятся под действием положительного эволюционного отбора.

Ключевые слова: транспозоны, структуры «стебель–петля», динуклеотидные характеристики, энтропия, свободная энергия Гиббса, машинное обучение.

Транспозоны – это фрагменты ДНК, способные каким-либо способом размножаться и перемещаться в геноме. Транспозоны содержатся в геномах всех эукариот и занимают значительную их часть (46% генома человека, 37% генома мыши, 10% генома плодовой мушки, 85% генома кукурузы) [1,2]. 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 транспозонов в качестве раковых маркеров [1,3]. Установлено, что транспозоны вносят существенный вклад в вариативность геномов различных органов, например, мозга и иммунной системы [4]. Также есть предположение, что транспозоны являются инструментом эволюции, так как они способны вызывать как масштабные перестройки (например, рекомбинацию между двумя неаллельными элементами в двух разных хромосомах), так и небольшие изменения генома (дупликации, инверсии, делеции) [5]. 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 [6].

По типу перемещения транспозоны можно разделить на те, которые перемещаются методом «вырезать и вставить», и те, которые используют метод «копировать и вставить». С ДНК копируется РНК, а потом с помощью обратной транскрипции с РНК создается копия ДНК, которая и вставляется в геном. Транс-

Сокращения: LINE – long interspersed elements, SINE – short interspersed elements.

позоны второго типа называются ретротранспозонами. Самые распространенные классы ретротранспозонов человека – LINE (long interspersed elements) и SINE (short interspersed elements). У человека ретротранспозоны LINE имеют порядка 500000 копий и занимают 17% генома, а SINE имеют больше миллиона копий и занимают 11% генома.

Для осуществления транспозиции LINE используют собственный молекулярный аппарат, закодированный ими для копирования своей последовательности и вставки ее в геном. SINE являются паразитами, они не кодируют белки, для транспозиции им требуются белки LINE. Остается неясным, каким образом белок распознает собственную РНК и РНК SINE [6]. Для нескольких организмов было показано, что белок распознает вторичную структуру типа «стебель–петля» на 3′-конце РНК. Более того, было показано, что некоторые организмы имеют одинаковые последовательности на 3′-конце LINE- и SINE-транспозонов [7–9]. 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 одинакового 3′-конца, принято считать, что распознается поли-А-хвост, имеющийся и у LINE-, и у SINE-транспозонов. Однако 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. Ранее мы обнаружили консервативную вторичную структуру на 3′-конце РНК LINE- и SINE-транспозонов человека [10], а также у разных видов, расположенных по всему дереву жизни (неопубликованные результаты). Нами была выдвинута гипотеза, что данная структура играет роль в процессе ретротранспозиции. Несмотря на полное отсутствие сходства на уровне последовательностей, консервативное положение структуры говорит о наличии свойств, на которые действует положительный эволюционный отбор. Наиболее вероятно, что структурные характеристики определяют характер связывания белка со структурой «стебель–петля», или шпилькой.

Большинство белков, взаимодействующих с молекулой ДНК, контактируют с участком длиной 15–20 пар оснований [11]. Важнейшую роль при взаимодействии с белком играют локальные физико-химические и геометрические свойства этого участка. Наименьшее приближение, на котором физико-химические и геометрические свойства последовательности имеет смысл рассматривать – это уровень динуклеотидов. Последние исследования показали, что динуклеотидные характеристики РНК или участка ДНК могут быть использованы для предсказания связи с белком [12,13]. Так, например, было показано, что с помощью динуклеотидных физико-химических и геометрических характеристик возможно построить модели машинного обучения, с высокой точностью распознающие горячие точки рекомбинации [14], сплайс-сайты [15], регуляторные малые РНК, происшедшие из транспозонных последовательностей [16], сайты ДНК-редактирования [13].

Целью данной работы является определение физических и геометрических свойств структур «стебель–петля» на 3′-конце транспозонов человека и сравнение их со свойствами структур из других областей генома. В качестве характеристик структур «стебель–петля» в геноме человека мы выбрали динуклеотидные свойства, доступные в базе данных DiProDB. Каждая структура «стебель–петля» была охарактеризована набором следующих характеристик: свободной энергией Гиббса, энтальпией, энтропией, гидрофильностью, Shift, Slide, Rise, Tilt, Roll и Twist. С помощью методов машинного обучения мы построили модель, которая распознает структуры транспозонов по физическим и геометрическим свойствам с 94%-й степенью точности. Наибольший вклад в распознавание структур внесли гидрофильность, энтальпия, Rise и Twist. Предполагается, что именно эти свойства структур транспозонов находятся под действием положительного эволюционного отбора.

МЕТОДЫ

Аннотация генома вторичными структурами. Аннотация генома вторичными структурами была проведена при помощи программного комплекса DNA Punctuation (www.dnapunctuation.org). Процедура поиска консервативных вторичных структур подробно описана в работе [10].

Составление выборок структур «стебель–петля». Было сформировано четыре набора данных – консервативные вторичные структуры из 6622 L1-транспозонов человека, из 39 самых активных L1-транспозонов человека, вторичные структуры, взятые из случайных мест генома, и сгенерированные случайным образом.

Физические и геометрические характеристики структур «стебель–петля». Для каждой струк-

туры «стебель–петля» были рассчитаны термодинамические характеристики, исходя из модели ближайших соседей. В этой модели свободная энергия структуры складывается из свободной энергии 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. Свободная энергия РНК дуплекса может быть представлена как

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

Первое слагаемое – это вклад i-й динуклеотидной пары оснований, n_i раз встречающейся в последовательности, i изменяется от 1 до 16 (число возможных динуклеотидных пар оснований). Второе слагаемое – это энергия инициации. В нее входят факторы, не зависящие от последовательности (конденсация контрионов, энтропийные потери при формировании дуплекса и т.д.). Третий член отвечает за энтропийные потери в случае если дуплекс образовался из одной нити ДНК (комплементарные участки находятся на одной нити).

Динуклеотидные физические характеристики для РНК были взяты из базы данных DiProDB [17]. Для построения модели были использованы 10 характеристик – нуклеотидная последовательность вторичной структуры разбивалась на динуклеотиды, каждому из которых сопоставлялось соответствующее число из DiProDB. Затем считалась медиана. Такая процедура проводилась для каждой из 10 рассмотренных характеристик. Таким образом каждой последовательности был сопоставлен вектор из 10 чисел, соответствующий 10 физическим характеристикам – предикторам. В качестве термодинамических предикторов были использованы два набора – средние удельные термодинамические характеристики (на каждую динуклеотидную пару), полученные описанным выше способом, и термодинамические характеристики образования структуры в целом.

Для анализа использовали четыре набора данных – последовательности структур типа «стебель–петля» из активных в настоящее время транспозонов L1 в геноме человека, из 6622 транспозонов с наиболее сохранившимися последовательностями, из случайных областей генома, а также случайно-сгенерированные последовательности, образующие структуру «стебель–петля».

Построение модели машинного обучения. Для распознавания вторичных структур по их физическим и геометрическим характеристикам была построена модель машинного обучения, использующая метод опорных векторов. Данный метод позволяет разделять точки в n-мерном пространстве (n – 1)-мерной гиперплоскостью. Из всех возможных разделяющих гиперплоскостей выбирается одна, расстояние от которой до элемента каждого класса максимально.

Для оценки эффективности модели были использованы следующие характеристики: ACC – точность, SN – чувствительность, SP – специфичность, AUC – ROC-кривая и площадь под ней.

Построение пространственной структуры «стебель–петля». Для построения пространственной структуры был использован программный комплекс 3DNA [18]. Он позволяет реконструировать структуру по динуклеотидным характеристикам ее последовательности.

РЕЗУЛЬТАТЫ

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

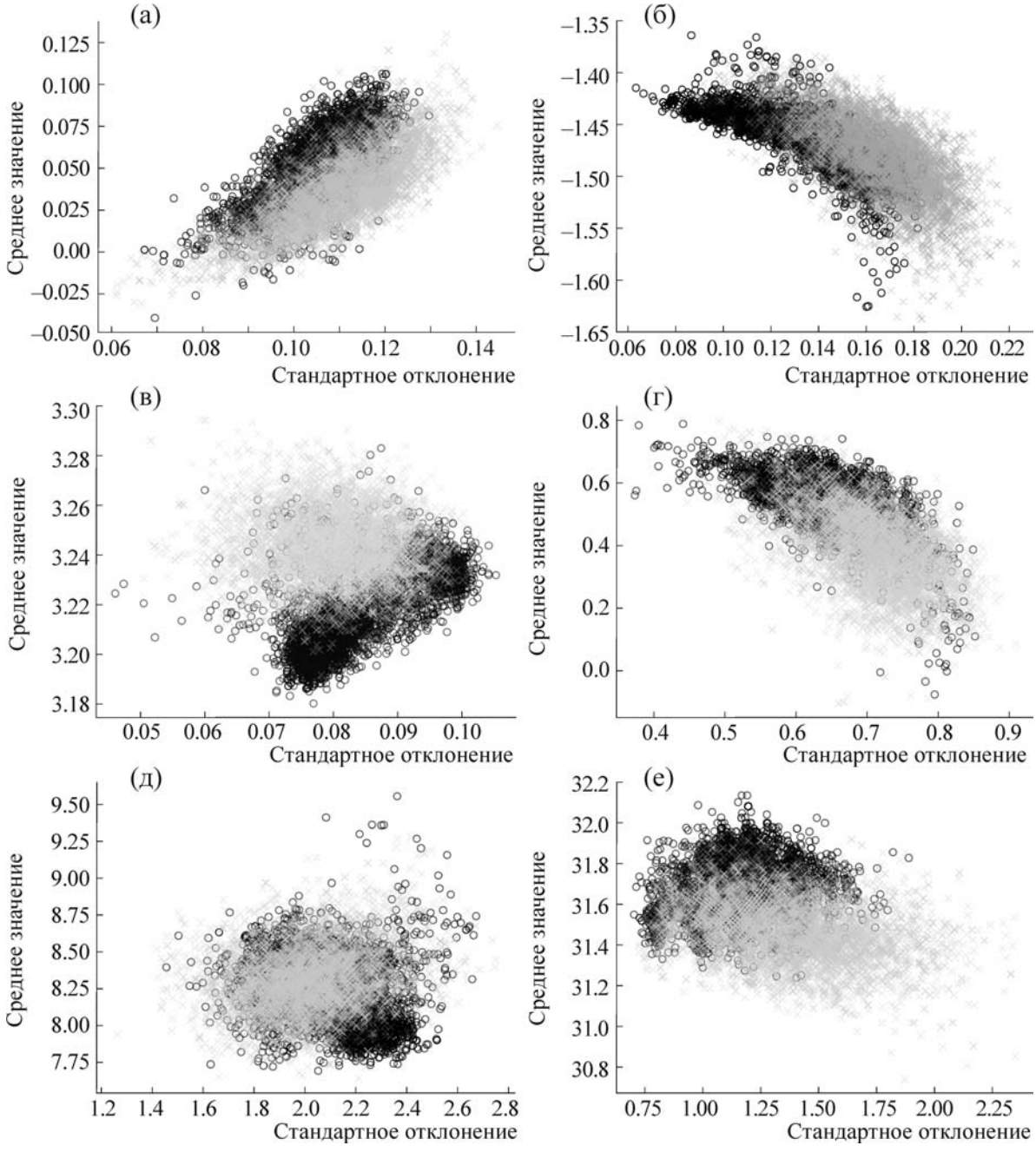
Для анализа использовали четыре набора данных – последовательности структур типа «стебель–петля» из активных в настоящее время транспозонов L1 в геноме человека, из 6622 транспозонов с наиболее сохранившимися последовательностями, из случайных областей генома, а также случайно-сгенерированные последовательности, образующие структуру «стебель–петля». Термодинамические характеристики были посчитаны для всех последовательностей структур. Результаты приведены в табл. 1. Прослеживается стремление к минимуму значений всех трех характеристик для структур активных транспозонов.

Зависимость средних значений термодинамических характеристик образования динуклеотидной пары от стандартных отклонений для разных классов представлена на рис. 1. Из графиков видно, что случайные транспозонные структуры могут быть разделены по их термодинамических характеристикам.

Таблица 1. Термодинамические характеристики последовательностей структур «стебель–петля»

Набор данных	ΔG, ккал/моль	ΔH, ккал/моль	ΔS, кал/моль/K
Структуры из активных L1	–13,5 ± 0,9	–130,0 ± 2,1	–375,7 ± 5,2
Структуры из хорошо сохранившихся L1	–11,1 ± 2,9	–119,0 ± 21,8	–347,8 ± 65,2
Структуры из случайных мест генома	–11,9 ± 6,4	–120,7 ± 38,5	–350,7 ± 79,3
Структуры, сгенерированные случайным образом	–9,5 ± 3,0	–108,8 ± 26,7	–320,3 ± 77,5

В терминах динуклеотидных геометрических характеристик возможно описать многие свойства, необходимые для связи с белком, например локальную упругость, гибкость, закрученность. Для каждой последовательности из трех наборов данных (структуры из хорошо сохра-



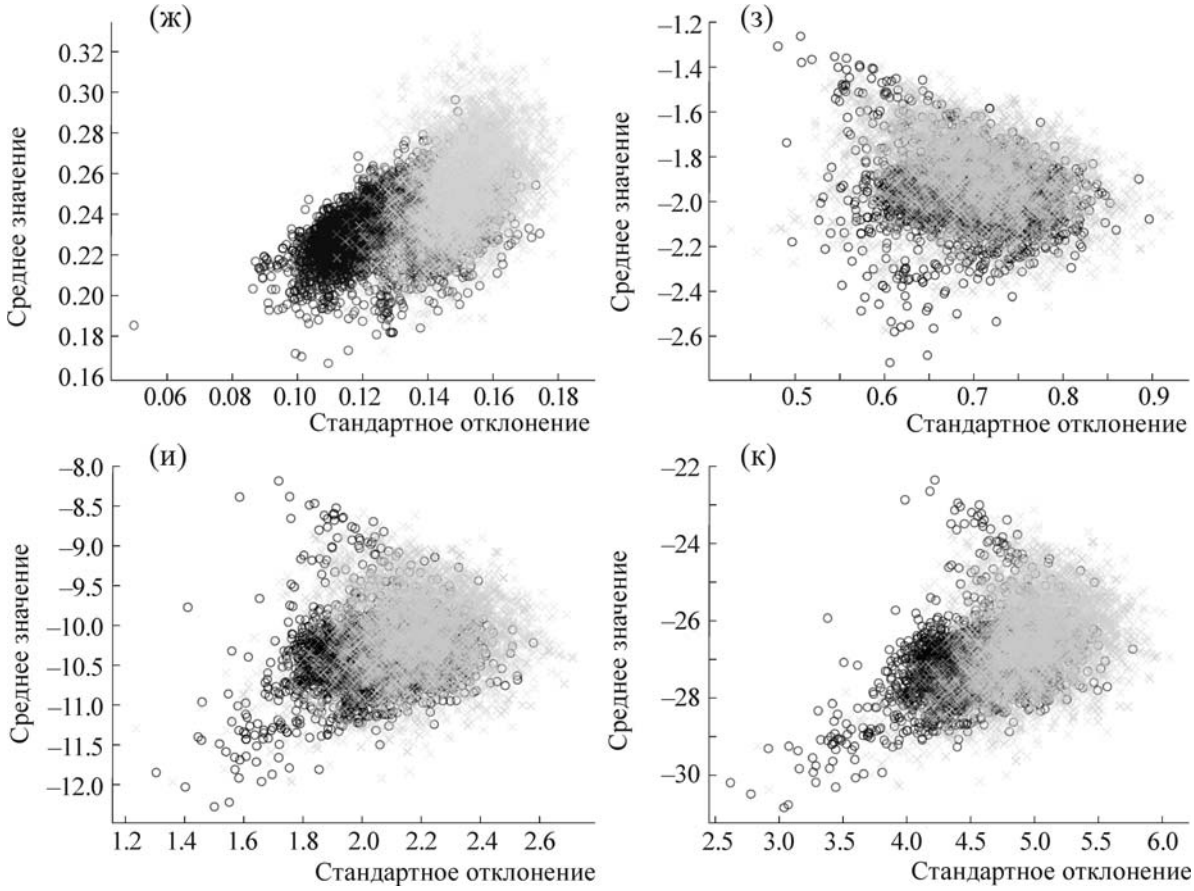


Рис. 1. Сравнение шпилек из 3′-конца L1-транспозона человека (черный цвет) со шпильками из случайным образом сгенерированной последовательности (серый цвет) по характеристикам: (а) – сдвиг пары оснований относительно соседней в направлении одного из желобков (Shift), (б) – сдвиг пары оснований относительно соседней в направлении сахарофосфатного остава (Slide), (в) – шаг спирали (Rise), (г) – угол раскрытия соседних пар оснований в сторону сахарофосфатного остава (Tilt), (д) – угол раскрытия соседних пар оснований в сторону одного из желобков (Roll), (е) – угол кручения (Twist), (ж) – гидрофильность, (з) – свободная энергия Гиббса, (и) – энтальпия, (к) – энтропия.

нившихся L1, из случайных областей генома, случайно сгенерированные структуры) были посчитаны среднее значение и стандартное отклонение каждой из шести геометрических характеристик (рис. 1). По всем шести геометрическим характеристикам структуры из активных и хорошо сохранившихся транспозонов L1 статистически значимо отделяются от структур из случайных областей генома и сгенерированных случайным образом. По аналогии с термодинамическими характеристиками, для каждого набора данных были вычислены средние значения шести геометрических характеристик. Они приведены в табл. 2. Видно, что структуры из L1 обладают определенными особенностями.

Каждой последовательности структуры «стебель–петля» была сопоставлена точка в десятимерном пространстве характеристик. Методом опорных векторов была построена девятимерная гиперплоскость, разделяющая структуры, принадлежащие концу L1 транспозонов человека, и структуры, взятые из случайных мест генома или сгенерированные случайным образом. Точность построенного классификатора составила 94%.

В табл. 3 приведены параметры построенной модели. На рис. 2 показана ROC-кривая.

Видно, что построенная модель оказалась очень 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.

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, наи-

Таблица 2. Геометрические характеристики последовательностей структур «стебель–петля» и величины их гидрофильности

Набор данных	Shift, Å	Slide, Å	Rise, Å	Tilt, град	Roll, град	Twist, град	Гидрофильность
Структуры из активных L1	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	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	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	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

Таблица 3. Параметры модели

Модель	Точность	Специфичность	Чувствительность	Площадь под ROC-кривой
SVM	0,94	0,94	0,93	0,98

Таблица 4. Разделяющая способность характеристик

Характеристика	Разделяющая способность, %
Гидрофильность	100,0
Энтальпия	45,3
Rise	30,0
Twist	28,2
Энтропия	19,8
Shift	8,9
Slide	6,7

более отличающихся между двумя вышеупомянутыми классами структур. 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. 4. Наибольший вклад в распознавание структур внесли гидрофильность, энтальпия, параметры Rise и Twist.

Исходя из динуклеотидных геометрических параметров, были построены модельные структуры для активного L1 транспозона и случайной структуры, взятой из генома человека (рис. 3). Ширина обеих бороздок L1 транспозона меньше, чем для случайной структуры. Многие белки связываются с участками, расположенными в большой бороздке. Возможно, меньшая ширина в случае L1 необходима белку для осуществления взаимодействия [19]. 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 [20,21]. Необходимо дальнейшее экспериментальное исследование пространственной структуры комплекса обратной транскриптазы с РНК транспозона для определения важнейших для связывания геометрических характеристик РНК.

Были сделаны некоторые выводы о закрученности структуры. Шпилька в геноме рыбы *Danio rerio* (рис. 4а), которая распознается обратной транскриптазой, является правозакрученной. Проведенный нами анализ аминокислотных последовательностей белка ORF2 35 организмов, стоящих на разных ступенях эволюционной лестницы, выявил консервативность доменов эндонуклеазы и обратной транскриптазы [10]. Логично предположить, что схожесть распознающего элемента влечет и схожесть распознаваемого. В базе PDB мы нашли структуру комплекса обратной транскриптазы ретровируса с молекулой РНК (рис. 4б). Общий характер укладки свидетельствует о том, что α-спирали, перекрываясь, имеют тенденцию к образованию левой суперспирали.

Таким образом, правозакрученная шпилька РНК взаимодействует с белком, распознающий домен которого имеет тенденцию к образованию левой суперспирали. Это еще раз подтверждает выявленную закономерность: взаимодействие разнотипных молекул осуществляется структурами, имеющими разный знак хиральности [22].

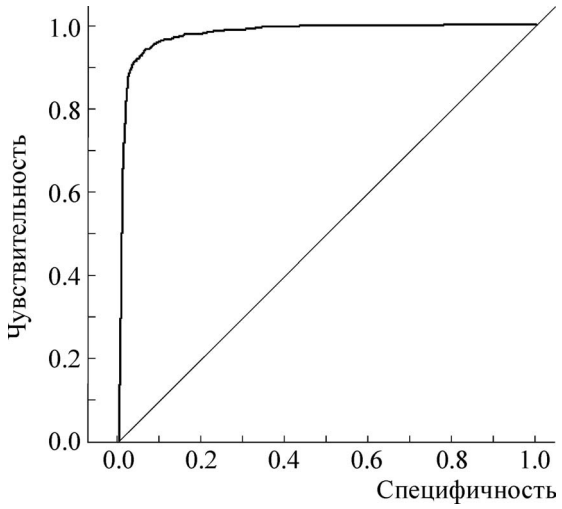


Рис. 2. Кривая ошибок модели логистической регрессии.

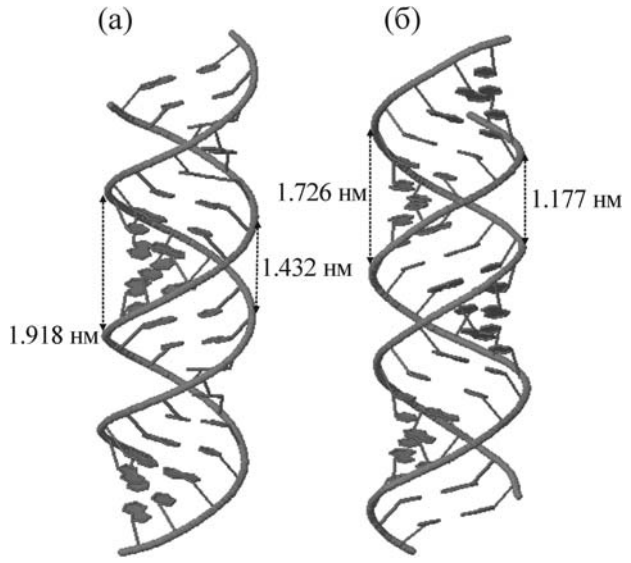


Рис. 3. (а) – Модель стебля структуры из случайной области генома человека; (б) – модель стебля структуры, расположенной на 3′-конце РНК-транскрипта транспозона L1 в геноме человека.

ОБСУЖДЕНИЕ

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 по средним значениям структурам 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.

Статистически значимое отличие геометрических характеристик структур из активных, хорошо сохранившихся транспозонов L1 от структур из случайных областей генома и сгенерированных случайным образом говорит о сохранении геометрических характеристик на динуклеотидном уровне в процессе эволюции. Такая консервативность может быть объяснена

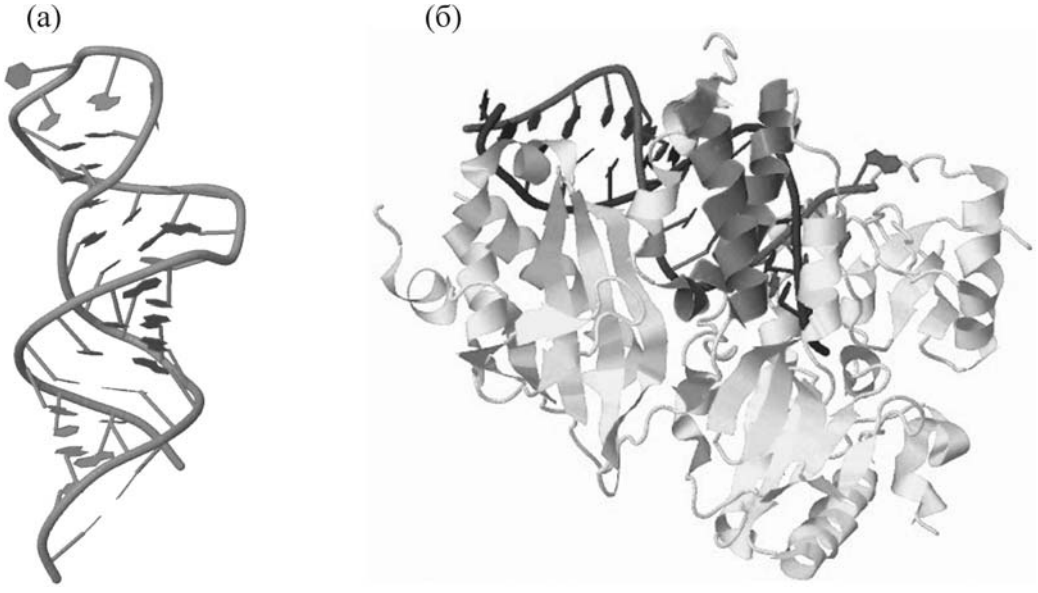


Рис. 4. (а) – Структура шпильки LINE транспозона в геноме рыбы *Danio rerio*, распознаваемой обратной транскриптазой; (б) – комплекс двойной спирали РНК и обратной транскриптазы ретровируса.

поддержанием структурных особенностей, необходимых для связи с белком.

Многие исследования показали, что комплексы связывания РНК–белок существенно зависят от формы и последовательности РНК (для обзора можно привести работы [23–25]). С параметром Rise связаны величины большой и малой бороздок. Известно, что величина и конфигурация бороздок играет 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 в процессе редактирования РНК [26]. Роль гидрофобности вместе с размерами бороздок была показана для белков, взаимодействующими с тетрапетлями структур «стебель–петля» [27]. 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, 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 [28]. Роль малых и больших бороздок при взаимодействии структур «стебель–петля» с белками была показана для рибосомального комплекса [29]. Роль степени закрученности спирали (параметр Twist) также была показана для РНК–белковых взаимодействий для бычьего вируса иммунодефицита [30]. 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 [31].

Построенная в работе модель машинного обучения показала, что транспозонные структуры «стебель–петля» с точностью около 95% могут быть распознаны по физическим и геометрическим характеристикам и отличаются от структур «стебель–петля» из других участков генома. Обученная модель также позволяет находить подобные структуры в любой заданной последовательности РНК. Представляет интерес проверить экспериментально, возможно ли распознавание обратной транскриптазой любой последовательности, имеющей на 3′-конце вторичную структуру с выявленными в данной работе свойствами. Выявленное свойство может иметь огромное значение в биоинженерии для встраивания любых заданных последовательностей в геном.

Определение наиболее отличающихся между группами структур характеристик программными методами является важной задачей при изучении процесса связывания с белками. Зная, какие именно характеристики определяют взаимодействие с белком, можно регулировать этот процесс. В данной работе мы демонстрируем возможность применения методов машинного обучения для решения этой задачи. Наибольший вклад в различие структур транспозонов от структур, взятых из случайных областей генома и сгенерированных случайным образом, вносят четыре характеристики – гидрофильность, энтальпия и геометрические параметры Rise и Twist. Гидрофильность и параметр Rise имеют более низкие 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.

ЗАКЛЮЧЕНИЕ

Выявлены физическо-химические и геометрические характеристики структур «стебель–петля» на 3′-конце L1-транспозонов человека, на которые действует эволюционный отбор. Эти характеристики 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 и Twist. 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. Определение ключевых характеристик вторичной структуры РНК, связывающейся с обратной транскриптазой, может иметь важное практическое применение в бионженерии для встраивания заданных последовательностей в геном.

СПИСОК ЛИТЕРАТУРЫ

1. C. R. Huang, K. H. Burns, and J. D. Boeke, Annu. Rev. Genet. **46**, 651 (2012).
2. E. S. Lander, et al., Nature **409** (6822), 860 (2001).

3. D. C. Hancks and H. H. Kazazian, Jr., Curr. Opin. Genet. Dev. **22** (3), 191 (2012).
4. C. R. Beck, et al., Annu. Rev. Genomics Hum. Genet. **12**, 187 (2011).
5. H. H. Kazazian, Jr., Science **303** (5664), 1626 (2004).
6. S. R. Richardson, et al., Microbiol. Spectr. **3** (2), MDNA3-0061-2014 (2015).
7. Y. Hayashi, et al., Nucl. Acids Res. **42** (16), 10605 (2014).
8. M. Kajikawa and N. Okada, Cell **111** (3), 433 (2002).
9. Osanai, M., et al., Mol. Cell Biol. **24** (18), 7902 (2004).
10. D. Grechishnikova and M. Poptsova, BMC Genomics **17** (1), 992 (2016).
11. N. M. Luscombe, et al., Genome Biol. **1** (1), REVIEWS001. (2000).
12. P. Barraud and F. H. Allain, Curr. Top. Microbiol. Immunol. **353**, 35 (2012).
13. W. Chen, et al., Sci. Rep. **6**, 35123 (2016).
14. W. Chen, et al., Nucl. Acids Res. **41** (6), e68 (2013).
15. W. Chen, et al., Biomed. Res. Int. **2014**, 623149 (2014).
16. B. Liu, F. Yang, and K. C. Chou, Mol. Ther. Nucl. Acids **7**, 267 (2017).
17. M. Friedel, et al., Nucl. Acids Res. **37** (Database issue), D37 (2009).
18. X. J. Lu and W. K. Olson, Nucl. Acids Res. **31** (17), 5108 (2003).
19. C. O. Pabo and R. T. Sauer, Annu. Rev. Biochem. **53**, 293 (1984).
20. P. C. van der Vliet and C. P. Verrijzer, Bioessays **15** (1), 25 (1993).
21. R. E. Dickerson, Nucl. Acids Res. **26** (8), 1906 (1998).
22. V. A. Tverdislov, Biofizika **58** (1), 159 (2013).
23. G. Masliah, P. Barraud, and F. H. Allain, Cell. Mol. Life Sci. **70** (11), 1875 (2013).
24. R. Stefl, L. Skrisovska, and F. H. Allain, EMBO Rep. **6** (1), 33 (2005).
25. J. R. Williamson, Nat. Struct. Biol. **7** (10), 834 (2000).
26. J. M. Thomas and P. A. Beal, BioEssays **39** (4), 1600187 (2017).
27. R. Thapar, A. P. Denmon, and E. P. Nikonowicz, Wiley Interdiscip. Rev. RNA **5** (1), 49 (2014).
28. K. Wild, I. Sinning, and S. Cusack, Science **294** (5542), 598 (2001).
29. G. L. Conn, Science **284** (5417), 1171 (1999).
30. D. Moras and A. Poterszman, Curr. Biol. **6** (5), 530 (1996).
31. A. Barik and R. P. Bahadur, Nucl. Acids Res. **42** (15), 10148 (2014).

Effects of Natural Selection on Physical and Geometrical Properties of Human Transposon Stem-Loop Structures

D.A. Grechishnikova and M.S. Poptsova

Faculty of Physics, Lomonosov Moscow State University, Leninskiye Gory, 1/2, Moscow, 119991 Russia

Secondary RNA structures play an important role in transposition, in particular, in RNA recognition by transposon proteins. Earlier we found the conservative structure at the 3'-end of human transposons and made a hypothesis about the role of this structure in transposition. Although there is no similarity at the sequence level, the conservative position of this structure points to the fact that there are structural properties, which are under positive natural selection. The aim of the present work is to identify physical and geometrical properties of stem-loop structures at the 3'-end of human transposons and to compare their properties with the properties of structures from other genomic regions. Every stem-loop structure was characterized by a set of 10 characteristics: Gibbs free energy, enthalpy, entropy, hydrophilic property, Shift, Slide, Rise, Tilt, Roll and Twist. Using machine learning methods we built a model, which recognize stem-loop structures by their physical and geometrical characteristics with 94% of accuracy. The most important parameters in recognition model are hydrophilic property, enthalpy, Rise and Twist. These properties of transposon structures are supposed to be under positive natural selection.

Keywords: transposons, stem-loop structures, dinucleotide characteristics, entropy, Gibbs free energy, machine learning