БИОЛОГИЧЕСКИЕ НАУКИ

BIOLOGICAL SCIENCES

УДК 577.323 DOI 10.23683/0321-3005-2017-4-1-63-69

РАСПОЗНАВАНИЕ СТРУКТУР СТЕБЕЛЬ – ПЕТЛЯ ТРАНСПОЗОНОВ ЧЕЛОВЕКА И ПРОГНОЗИРОВАНИЕ ИХ ФУНКЦИИ ПРИ ПОМОЩИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

© 2017 г. Д.А. Гречишникова[1], М.С. Попцова[1,2]
[1]Московский государственный университет им. М.В. Ломоносова, Москва, Россия,
[2]Высшая школа экономики, Москва, Россия

IDENTIFYING STEM-LOOP STRUCTURES IN HUMAN TRANSPOSONS AND PREDICTING THEIR FUNCTION BY MACHINE LEARNING MODEL

D.A. Grechishnikova[1], M.S. Poptsova[1,2]
[1]Lomonosov Moscow State University, Moscow, Russia,
[2]Higher School of Economics, Moscow, Russia

Гречишникова Дарья Александровна – аспирант, кафедра биофизики, физический факультет, Московский государственный университет им. М.В. Ломоносова, Ленинские Горы, 1, стр. 2, г. Москва, 119991, Россия, e-mail: daria.grechishnikova@gmail.com

Daria A. Grechishnikova - Postgraduate, Department of Biophysics, Faculty of Physics, Lomonosov Moscow State University, Leninskie Gory, 1, Bld. 2, Moscow, 119991, Russia, e-mail: daria.grechishnikova@gmail.com

Попцова Мария Сергеевна – кандидат физико-математических наук, доцент, факультет компьютерных наук, Высшая школа экономики, ул. Мясницкая, 20, г. Москва, 101000, Россия; старший научный сотрудник, кафедра биофизики, физический факультет, Московский государственный университет им. М.В. Ломоносова, Ленинские Горы, 1, стр. 2, г. Москва, 119991, Россия, e-mail: maria.poptsova@gmail.com

Maria S. Poptsova - Candidate of Physics and Mathematics, Associate Professor, Department of Computer Science, Higher School of Economics, Myasnitskaya St., 20, Moscow, 101000, Russia; Senior Researcher, Department of Biophysics, Faculty of Physics, Lomonosov Moscow State University, Leninskie Gory, 1, Bld. 2, Moscow, 119991, Russia, e-mail: maria.poptsova@gmail.com

Во многих процессах, происходящих в клетке, важную роль играют вторичные структуры РНК/ДНК. Часто такие структуры служат опознавательным знаком для белков. Ранее нами было установлено, что транспозоны человека имеют на 3'-хвосте консервативную вторичную структуру типа стебель – петля. Мы предположили, что она может быть необходима для связи с белком, осуществляющим транспозицию. Аргументом в пользу этой гипотезы могло бы стать значимое отличие физических характеристик структур из транспозонов и из других областей генома. В данной работе мы определяем физические и геометрические свойства структур стебель – петля на 3'-конце транспозонов человека и сравниваем их со свойствами структур из других областей генома. Каждой структуре стебель – петля сопоставлялся набор из 10 характеристик: свободной энергии Гиббса, энтальпии, энтропии, гидрофильности, Shift, Slide, Rise, Tilt, Roll и Twist. С помощью многомерного дисперсионного анализа мы отвергли гипотезу о неразличимости физических характеристик структур из транспозонов и из других областей генома. Построена модель логистической регрессии, способная распознавать структуры из транспозонов по физическим свойствам с точностью 89 %. Наибольшим весом в модели обладают гидрофильность, параметры Rise и Twist. Предполагается, что именно эти свойства в первую очередь влияют на распознавание белком структуры.

Ключевые слова: *транспозоны, структуры стебель – петля, динуклеотидные характеристики, энтропия, свободная энергия Гиббса, машинное обучение, логистическая регрессия.*

Many cellular processes are governed by the secondary RNA/DNA structures. These structures often play a role of a marker sign for the proteins. Earlier we found that human transposons have a conservative stem-loop structure at the 3'-end. We made a conjecture that this structure is important for the binding with the protein, which performs transposition. Statistically significant differences in physical properties of transposon stem-loop structures from other genomic regions may support our proposal. In the present work we identify physical and geometrical properties of the stem-loop structures at the 3'-end of human transposons and compare their properties with the properties of structures from other genomic regions. The set of 10 characteristics (Gibbs free energy, enthalpy, entropy, hydrophilic property, Shift, Slide, Rise, Tilt, Roll and Twist) is assigned to every stem-loop structure. With the help of multivariate analysis of variance (MANOVA) method we rejected an equality hypothesis for the physical characteristics of transposons and other genomic regions. We have built a logistic regression model, which is able to recognise transposon structures by their physical properties with 89 % accuracy. Hydrophilic property, Rise and Twist have the greatest weights in the model. We suppose that these properties are of the first importance for the structure recognition by a protein.

Keywords: *transposons, stem-loop structures, dinucleotide properties, entropy, Gibbs free energy, machine learning, logistic regression.*

Введение

Транспозоны – это участки ДНК, способные перемещаться по геному и увеличивать число своих копий. Транспозоны есть в геноме каждого эукариота. Например, они занимают 46 % генома человека [1, 2]. В настоящее время транспозоны активно исследуются, выявляются все новые и новые примеры их участия в различных биологических процессах. Зафиксировано 96 заболеваний, причиной которых являлся скачок транспозона. Прослеживается связь между активностью мобильных элементов и онкологическими заболеваниями. Активно разрабатываются диагностические методы с участием транспозонов. Так, например, транспозон LINE-1 может быть использован в качестве онкомаркера [1, 3]. Было обнаружено повышенное число копий LINE-1 в клетках мозга людей, страдающих шизофренией [4]. Скачки транспозонов способны вызывать перестройки генома, делеции, дупликации, инверсии [5]. Существует предположение, что транспозоны являются своего рода «скульпторами» генома, помогая организму приспособиться к окружающей среде в ходе эволюции [6]. Кроме того, известно, что транспозоны способны регулировать свою экспрессию и экспрессию близлежащих генов [7].

Выделяют два типа мобильных элементов. Первый перемещается методом «вырезать и вставить», а 2-й – методом «копировать и вставить». В последнем методе на матрице РНК синтезируется последовательность ДНК, которая затем при помощи обратной транскриптазы вставляется в новое место генома. Такие транспозоны называют ретротранспозонами. В геноме человека наиболее распространены LINE- и SINE-ретротранспозоны (далее транспозоны). LINE занимают 17 % генома (500 000 копий), SINE – 11 (> 1 000 000 копий).

LINE кодируют собственный молекулярный аппарат, который осуществляет копирование и вставку последовательности в геном. У SINE такого аппарата нет, они «заимствуют» его у LINE, подставляя свой РНК-транскрипт вместо РНК-транскрипта LINE. До сих пор остается неясным, как именно белки LINE узнают собственную РНК и РНК SINE [7]. Для нескольких организмов экспериментально было показано, что белок распознает вторичную структуру типа стебель – петля на 3'-конце РНК транспозона. Кроме того, выявлено, что некоторые организмы имеют одинаковые последовательности на 3'-конце LINE- и SINE-транспозонов, а значит, и одинаковые вторичные структуры, которые, по предположению, и распознаются белком [8–10]. Принято считать, что в случаях, когда последовательности на 3'-концах разные, распознается поли-А-хвост, имеющийся и у LINE-, и у SINE-транспозонов. Однако такие хвосты есть у всех мРНК, за небольшим исключением, а это ставит под сомнение возможность избирательного распознавания белком РНК-транскрипта. Механизм ретротранспозиции пока недостаточно изучен. Остается открытым вопрос о способе распознавания белком своего РНК-транскрипта и РНК-транскрипта SINE. Ранее нами было установлено, что LINE- и SINE-транспозоны человека имеют на 3'-хвосте консервативную вторичную структуру типа стебель – петля [11]. Мы предположили, что эта структура может быть необходима для распознавания РНК-транскрипта белком. В таком случае очень вероятно, что характер взаимодействия белка с вторичной структурой типа стебель – петля будет определяться в том числе и физическими свойствами структуры. Логично предположить, что физические свойства вторичной структуры, узнаваемой белком, должны отличаться от свойств других вторичных структур РНК. Такие структуры, скорее всего, находятся под действием эволюционного отбора.

Большинство белков взаимодействует с небольшим участком последовательности. Длина такого участка, как правило, составляет 15–20 нуклеотидов [12]. Его локальные физические свойства играют важнейшую роль во взаимодействии с белком. В последнее время появляется все больше работ, использующих физические характеристики

динуклеотидов для предсказания вовлеченности последовательности в тот или иной клеточный процесс. В частности, такой подход позволяет предугадать с некоторой степенью точности, возможно ли взаимодействие данной структуры с белком [13, 14]. В работе [14] авторы построили модель машинного обучения, способную автоматически определять сайты редактирования РНК. Редактирование осуществляет специальный белок ADAR (РНК-зависимая аденозиндезаминаза), который связывается с молекулой РНК и заменяет аденозин на инозин. Кроме того, было показано, что модели машинного обучения с использованием физических динуклеотидных характеристик способны эффективно определять горячие точки рекомбинации [15], сплайс-сайты [16], регуляторные малые РНК, произошедшие из транспозонных последовательностей [17].

В данной работе проведено сравнение по физическим характеристикам вторичных структур из транспозонов со вторичными структурами из других мест генома и со структурами, сгенерированными случайным образом. Построена модель машинного обучения, способная распознавать структуры транспозонов по физическим свойствам с 89%-й степенью точности.

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 подробна описана в [11].

Составление выборок структур стебель – петля. Сформировано 3 набора данных – консервативные вторичные структуры из 6622 L1-транспозонов человека, вторичные структуры, взятые из случайных мест генома, и сгенерированные случайным образом.

Физические характеристики структур стебель – петля. Динуклеотидные физические характеристики были взяты из базы данных DiProDB [18]. Рассмотрены 10 характеристик: свободная энергия Гиббса, энтальпия, энтропия, гидрофильность, Shift, Slide, Rise, Tilt, Roll и Twist. Последние 6 – это геометрические параметры, характеризующие относительное пространственное расположение нуклеотидов. Смысл этих параметров разъясняется на рис. 1.

Нуклеотидная последовательность структуры типа стебель – петля разбивалась на динуклеотиды. Затем каждому динуклеотиду сопоставлялось значение одной из характеристик, взятое из базы DiProDB.

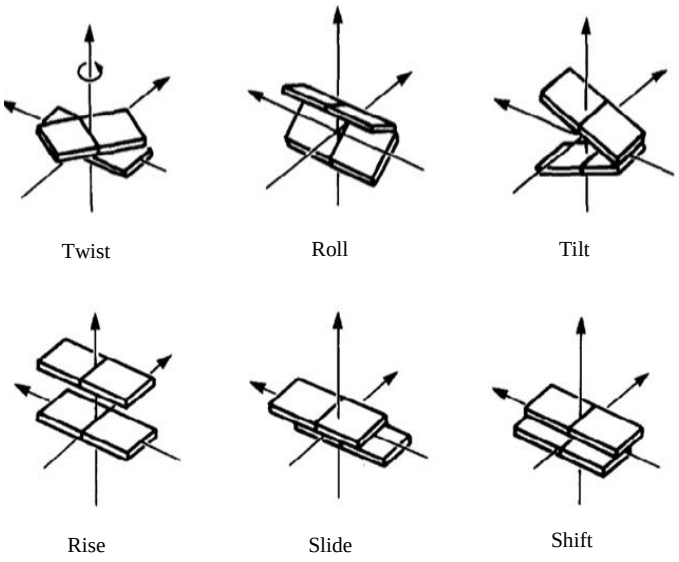


Рис. 1. Геометрические параметры, характеризующие пространственное расположение нуклеотидов относительно друг друга / Fig. 1. Geometric parameters characterizing special relationships between nucleotides

Посчитана медиана всех полученных для данной последовательности значений характеристики. Описанная процедура проведена для каждой последовательности и каждой характеристики.

Для анализа использовались 3 набора данных – последовательности структур типа стебель – петля из 6622 транспозонов человека, из случайных областей генома, а также случайно сгенерированные последовательности, образующие структуру стебель – петля.

Статистический анализ значимости различий. Проверка гипотез о неразличимости структур транспозонов и структур из других мест генома, а также структур, сгенерированных случайным образом, проводилась методом многомерного дисперсионного анализа. Расчеты произведены в среде R.

Построение модели машинного обучения. Применялась модель логистической регрессии для классификации последовательностей. В качестве зависимой переменной использовалась переменная, принимающая значение 0 в случае, если структура принадлежит транспозону, и 1, если не принадлежит. 10 физических характеристик представлены в качестве независимых переменных (предикторов), на основе значений которых требуется сделать вывод о принадлежности структуры транспозону. Расчеты проведены в среде R с использованием функции glm, а также функции train из пакета caret.

Для оценки модели применён метод кросс-валидации, а именно Leave Group Out Cross-validation (LGOCV). Из набора данных отбирается случайным образом группа образцов и используется для проверки эффективности модели, остальные данные – для обучения.

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

Результаты

Проведен многомерный дисперсионный анализ данных, в ходе которого 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 отличаются, проведены тесты Стьюдента. Полученные p-value скорректированы методом Бонферони, так как сразу несколько гипотез проверено на одних и тех же данных. Результаты представлены в табл 1а, б. По всем характеристикам наблюдаются значимые различия.

Для определения, является ли структура стебель – петля транспозонной или нет, мы предложили модель логистической регрессии с динуклеотидными физическими характеристиками в качестве предикторов. Построенная модель оказалась очень эффективной. Ее точность составила 89 % при сравнении структур из транспозонов и случайных мест генома и 93 % при сравнении структур из транспозонов и структур, сгенерированных случайным образом. В табл. 2 приведены другие параметры модели, на рис. 2 – ROC-кривые. Полученная модель с высокой точностью определяет вторичные структуры, принадлежащие 3’-концу транспозонов в геноме человека. Стоит отметить, что модель способна распознавать такие вторичные структуры в любой заданной последовательности РНК.

Таблица 1а

Физические характеристики (геометрические параметры и гидрофильность) структур стебель – петля / Stem-loop structure physical properties (geometric characteristics and hydrophilicity)

Свойство / датасет	Shift, 1 нм	Slide, 1 нм	Rise, 1 нм	Tilt, °	Roll, °	Twist, °	Hydrophilicity
Структуры из транспозонов	0,059	–1,456	3,212	0,553	8,133	31,752	0,232
Структуры из случайных мест генома	0,016	–1,458	3,237	0,380	8,330	31,681	0,243
Структуры, сгенерированные случайным образом	0,036	–1,480	3,243	0,395	8,310	31,487	0,251

Таблица 1б

Физические характеристики (термодинамические параметры) структур стебель – петля / Stem-loop structure thermodynamic properties

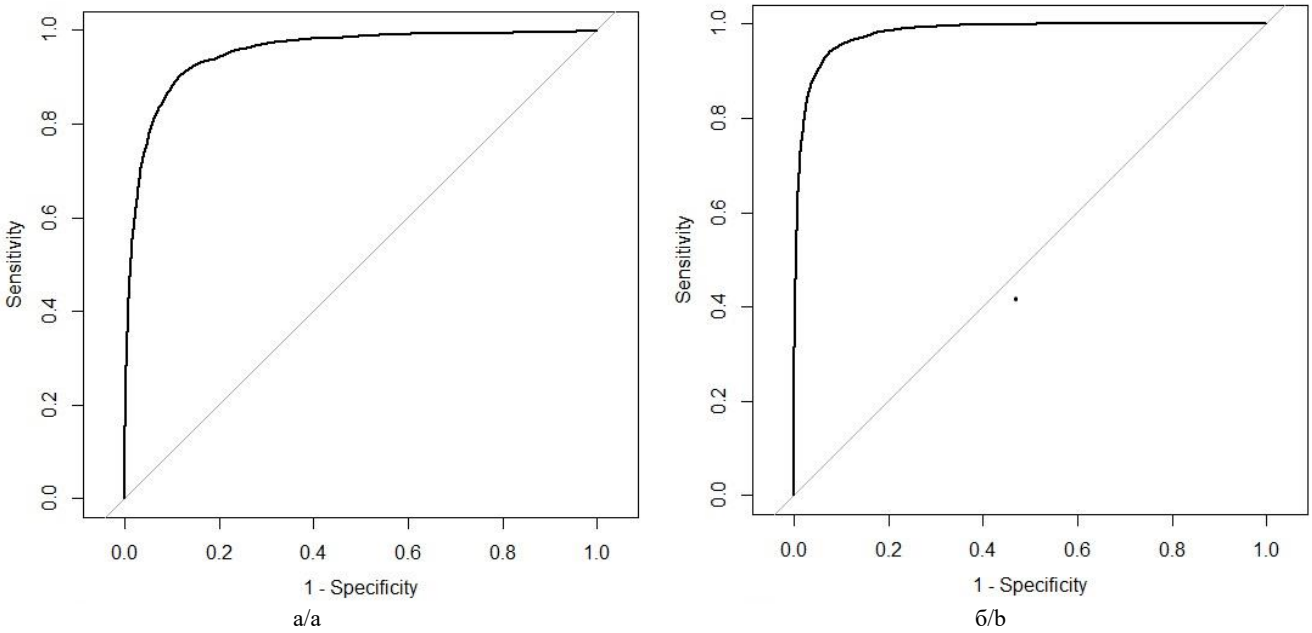
Свойство / датасет	Энтропия, Дж×К$^{-1}$ × ×моль$^{-1}$	Энтальпия, кДж/моль	Свободная энергия, кДж/моль
Структуры из транспозонов	–113,7	–43,5	–8,3
Структуры из случайных мест генома	–109,4	–41,5	–7,5
Структуры, сгенерированные случайным образом	–111,2	–42,4	–7,9

Таблица 2

Параметры модели / Predictor importance

	Точность	Специфичность	Чувствительность	AUC
Структуры из транспозонов vs структуры из случайных мест генома	0,89	0,89	0,89	0,95
Структуры из транспозонов vs структуры, сгенерированные случайным образом	0,93	0,93	0,93	0,98

Модель логистической регрессии позволяет выявить характеристики, по которым наблюдается наибольшее различие между структурами. Для каждого предиктора рассчитываются веса. Чем больше вес, тем больше вклад предиктора в разделяющую способность модели. Наибольший вклад при разделении транспозонных структур и структур из случайных мест генома внесли гидрофильность, параметры Rise и Twist, а при разделении транспозонных структур и структур, сгенерированных случайным образом, – Rise, Tilt и гидрофильность (табл. 3).



а/a б/b

Рис. 2. ROC-кривые: а – структуры из транспозонов vs структуры из случайных мест генома; б – структуры из транспозонов vs структуры, сгенерированные случайным образом / Fig. 2. ROC curve: a - transposon structure vs structures from other genome regions; b - transposon structure vs randomly generated structures

Таблица 3

Разделяющая способность характеристик / Predictor importance

Структуры из транспозонов vs структуры из случайных мест генома		Структуры из транспозонов vs структуры, сгенерированные случайным образом	
Характеристика	Разделяющая способность, %	Характеристика	Разделяющая способность, %
Rise	100,0	Rise	100,0
Hydrophilicity	45,3	Tilt	97,5
Twist	30,0	Hydrophilicity	77,7
Tilt	28,2	Shift	72,8
Slide	19,8	Enthalpy	66,1
Shift	8,9	Entropy	65,1
Roll	6,7	Twist	2,2

Обсуждение

Если некоторые характеристики последовательности отличаются от характеристик случайно сгенерированных последовательностей, можно предположить, что первые несут некую функциональную нагрузку. В таком случае это отличие поддерживается действием эволюционного отбора, сохраняющего последовательность от накопления случайных мутаций, влияющих на ключевые характеристики. Для многих белков экспериментально показана важная роль физических характеристик структуры во взаимодействии с белком. Поскольку это взаимодействие, как правило, специфично, структура должна обладать особым выделенным набором свойств. Полученный нами результат как раз показывает, что вторичные структуры из транспозонов обладают таким набором. Это является косвенным доказательством выдвинутой нами ранее гипотезы о роли вторичной структуры на конце транспозона в распознавании белком.

Построенная в данной работе модель может быть применима для решения важных задач биоинженерии. Она способна находить структуры, подобные транспозонным, в любой заданной последовательности РНК. Интересно было бы проверить экспериментально, может ли белок LINE, осуществляющий транспозицию, распознать любую последовательность, имеющую на 3’-конце вторичную структуру с выявленными в данной работе

свойствами. Если наша гипотеза верна, то разработанная модель может иметь огромное значение для решения задач встраивания любых заданных последовательностей в геном.

Показано, что характеристики гидрофильность, Rise и Twist обладают наибольшей разделяющей способностью. Эта информация важна, так как, очень вероятно, именно наиболее отличающиеся характеристики играют определяющую роль при взаимодействии с белком. Зная эти характеристики, можно управлять процессом связывания. Кроме того, подобная информация может оказаться полезной для задач молекулярного моделирования, например для оценки вероятности взаимодействия с белком.

Заключение

Показано, что транспозонные структуры статистически значимо отличаются от структур из других мест генома и структур, сгенерированных случайным образом. Построена модель машинного обучения, способная находить транспозонные структуры или им подобные в любой заданной последовательности РНК. Найдены характеристики вторичной структуры РНК, которые наиболее вероятно являются определяющими при связывании с белком. Полученные результаты, а также процедура их получения могут иметь важное практическое применение в биоинженерии (для встраивания заданных последовательностей в геном) и в задачах молекулярного моделирования.

Литература

1. *Huang C.R., Burns K.H., Boeke J.D.* Active transposition in genomes // Ann. Rev. Genet. 2012. Vol. 46. P. 651–675.

2. *Lander E.S., Linton L.M., Birren B.* [et al.]. International Human Genome Sequencing C. Initial sequencing and analysis of the human genome // Nature. 2001. Vol. 409, № 6822. P. 860–921.

3. *Hancks D.C., Kazazian H.H.* Active human retrotransposons: variation and disease // Current Opinion in Genetics & Development. 2012. Vol. 22, № 3. P. 191–203.

4. *Bundo M., Toyoshima M., Okada Y., Akamatsu W., Ueda J., Nemoto-Miyauchi T., Sunaga F., Toritsuka M., Ikawa D., Kakita A., Kato M., Kasai K., Kishimoto T., Nawa H., Okano H., Yoshikawa T., Kato T., Iwamoto K.* Increased l1 retrotransposition in the neuronal genome in schizophrenia // Neuron. 2014. Vol. 81, № 2. P. 306–313.

5. *Kazazian H.H.* Mobile elements: drivers of genome evolution // Science. 2004. Vol. 303, № 5664. P. 1626–1632.

6. *Beck C.R., Garcia-Perez J.L., Badge R.M., Moran J.V.* LINE-1 elements in structural variation and disease // Ann. Rev. Genomics Hum. Genet. 2011. Vol. 12. P. 187–215.

7. *Richardson S.R., Doucet A.J., Kopera H.C., Moldovan J.B., Garcia-Perez J.L., Moran J.V.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes // Microbiol Spectr. 2015. Vol. 3, № 2. P. MDNA3-0061-2014.

8. *Hayashi Y., Kajikawa M., Matsumoto T., Okada N.* Mechanism by which a LINE protein recognizes its 3' tail RNA // Nucleic Acids Research. 2014. Vol. 42, № 16. P. 10605–10617.

9. *Kajikawa M., Okada N.* LINEs Mobilize SINEs in the Eel through a Shared 3′ Sequence // Cell. 2002. Vol. 111, № 3. P. 433–444.

10. *Osanai M., Takahashi H., Kojima K.K., Hamada M., Fujiwara H.* 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. Vol. 24, № 18. P. 7902–7913.

11. *Grechishnikova D., Poptsova M.* Conserved 3' UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition // BMC Genomics. 2016. Vol. 17, № 1. P. 992.

12. *Luscombe N.M., Austin S.E., Berman H.M., Thornton J.M.* An overview of the structures of protein-DNA complexes // Genome Biol. 2000. Vol. 1, № 1. P. REVIEWS001.

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

14. *Chen W., Feng P., Ding H., Lin H.* PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions // Sci. Rep. 2016. Vol. 6. P. 35123.

15. *Chen W., Feng P.M., Lin H., Chou K.C.* iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition // Nucleic Acids Res. 2013. Vol. 41, № 6. P. e68.

16. *Chen W., Feng P.M., Lin H., Chou K.C.* iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition // Biomed. Res. Int. 2014. Vol. 2014. P. 623149.

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

18. *Friedel M., Nikolajewa S., Suhnel J., Wilhelm T.* DiProDB: a database for dinucleotide properties // Nucleic Acids Res. 2009. Vol. 37. C. D37–40.

References

1. Huang C.R., Burns K.H., Boeke J.D. Active transposition in genomes. *Ann. Rev. Genet*. 2012, vol. 46, pp. 651-675.

2. Lander E.S., Linton L.M., Birren B. et al. International Human Genome Sequencing C. Initial sequencing and analysis of the human genome. *Nature*. 2001, vol. 409, No. 6822, pp. 860-921.

3. Hancks D.C., Kazazian H.H. Active human retrotransposons: variation and disease. *Current Opinion in Genetics & Development*. 2012, vol. 22, No. 3, pp. 191-203.

4. Bundo M., Toyoshima M., Okada Y., Akamatsu W., Ueda J., Nemoto-Miyauchi T., Sunaga F., Toritsuka M., Ikawa D., Kakita A., Kato M., Kasai K., Kishimoto T., Nawa H., Okano H., Yoshikawa T., Kato T., Iwamoto K. Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron*. 2014, vol. 81, No. 2, pp. 306-313.

5. Kazazian H.H. Mobile elements: drivers of genome evolution. *Science, Science*. 2004, vol. 303, No. 5664,
pp. 1626-1632.

6. Beck C.R., Garcia-Perez J.L., Badge R.M., Moran J.V. LINE-1 elements in structural variation and disease. *Ann. Rev. Genomics Hum. Genet.* 2011, vol. 12, pp. 187-215.

7. Richardson S.R., Doucet A.J., Kopera H.C., Moldovan J.B., Garcia-Perez J.L., Moran J.V. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol. Spectr*. 2015, vol. 3, No. 2, pp. MDNA3-0061-2014.

8. Hayashi Y., Kajikawa M., Matsumoto T., Okada N. Mechanism by which a LINE protein recognizes its 3' tail RNA. *Nucleic Acids Research*. 2014, vol. 42, No. 16, pp. 10605-10617.

9. Kajikawa M., Okada N. LINEs Mobilize SINEs in the Eel through a Shared 3′ Sequence. *Cell*. 2002, vol. 111, No. 3, pp. 433-444.

10. Osanai M., Takahashi H., Kojima K.K., Hamada M., Fujiwara H. 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, vol. 24, No. 18, pp. 7902-7913.

11. Grechishnikova D., Poptsova M. Conserved 3' UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition. *BMC Genomics.* 2016, vol. 17, No. 1, p. 992.

12. Luscombe N.M., Austin S.E., Berman H.M., Thornton J.M. An overview of the structures of protein-DNA complexes. *Genome Biol.* 2000, vol. 1, No. 1, p. REVIEWS001.

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

14. Chen W., Feng P., Ding H., Lin H. PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Sci. Rep.* 2016, vol. 6, p. 35123.

15. Chen W., Feng P.M., Lin H., Chou K.C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res*. 2013, vol. 41, No. 6, p. e68.

16. Chen W., Feng P.M., Lin H., Chou K.C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res. Int*. 2014, vol. 2014, p. 623149.

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

18. Friedel M., Nikolajewa S., Suhnel J., Wilhelm T. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.* 2009, vol. 37, pp. D37-40.

Поступила в редакцию / Received ***9 августа 2017 г. / August 9, 2017***