УДК 577.21

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ, ПОСТРОЕННЫЕ НА ИНФОРМАЦИИ О ПОСЛЕДОВАТЕЛЬНОСТЯХ И СТРУКТУРЕ, ДЛЯ РАСПОЗНАВАНИЯ СТРУКТУР СТЕБЕЛЬ-ПЕТЛЯ НА 3'-КОНЦАХ ТРАНСПОЗОНОВ L1 И ALU В ГЕНОМЕ ЧЕЛОВЕКА

А. Заикин, А. Шеин, М. Попцова

Лаборатория биоинформатики, Департамент больших данных и информационного поиска, Национальный исследовательский университет Высшая школа экономики, Москва, Россия
125319, Москва, Кочновский проезд, 3
e-mail: mpoptsova@hse.ru; avzaikin@hse.ru; avshein@edu.hse.ru

Мы построили и исследовали два типа моделей, основанных на информации о последовательностях и структурных свойствах, для распознавания структур типа стебель-петля на 3`-концах L1 и Alu человека, и обнаружили параметры, дающие наибольший вклад в распознавание: Shift, Tilt, Rise и гидрофильность.

Ключевые слова: транспозоны, L1, Alu, стебель-петля, машинное обучение, случайный лес

Как белки LINE распознают свою собственную РНК и РНК SINE остается до сих пор неизвестным [1]. Для некоторых видов экспериментально показано, что белки LINE распознают вторичные структуры, такие как стебель-петля на 3`-конце РНК транспозона. Более того, показано наличие идентичных структур типа стебель-петля на 3`-концах последовательностей LINE и SINE для некоторых организмов, при этом структуры являются значимыми для распознавания белками транспозонов [2-4]. Мы обнаружили консервативную вторичную структуру на 3`-конце для L1 (LINE) и Alu (SINE) человека [5], а также и для других видов дерева жизни (неопубликованные результаты). При отсутствии схожести на уровне последовательностей, консервативное расположение этой структуры предполагает ее функциональность.

В настоящей работе мы исследовали структурные свойства 3`-конца L1 и Alu структур стебель-петля посредством моделей машинного обучения. Мы построили два типа моделей, используя два различных набора признаков: взятых из информации о последовательностях и из информации о структуре. Для модели, основанной на последовательностях, мы использовали частоты ди- и три-нуклеотидов, подсчитывая включения для каждого k-мера с шагом 1 вдоль последовательности. Для структурной модели мы рассматривали ножку, петлю и внутреннюю петлю ножки. Для ножки мы брали динуклеотидные свойства РНК из базы данных DiProDb [6], содержащей значения структурных параметров Shift, Slide, Rise, Tilt, Roll, Twist, а также физических и химических характеристик, таких как энтальпия, энтропия и гидрофильность. Мы построили модель Случайного Леса на 2000 деревьях с помощью библиотеки sklearn. Реализация модели и все данные для анализа доступны на github: https://github.com/AlexShein/transposons/.

Качество обеих моделей согласно метрике AUC варьирует в диапазоне 95-99%. Однако структурная модель позволяет извлекать важные структурные свойства. Анализ важности признаков структурной модели показал структурно-значимые характеристики для структуры стебель-петля из нескольких групп. Так оказалось, что параметры Shift, Rise и Tilt оказывают наибольший вклад в распознавание L1 и Alu 3`-структур. В дополнении к этому, ближайшие к петле нуклеотиды оказались более важными для распознавания Alu. Полученные результаты ясно демонстрируют наличие структурных ограничений для 3`-структур L1 и Alu, которые предположительно играют важную роль в распознавании пары L1-Alu машинерией L1. Сконструированные модели машинного обучения могут быть использованы для de novo обнаружения структур, относящихся к транспозонам.

Литература:
1. 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.
2. Hayashi, Y., et al., Mechanism by which a LINE protein recognizes its 3` tail RNA. Nucleic Acids Research, 2014. 42(16): p. 10605-10617.
3. Kajikawa, M. and N. Okada, LINEs Mobilize SINEs in the Eel through a Shared 3' Sequence. Cell, 2002. 111(3): p. 433-444.
4. 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.
5. 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.
6. Friedel, M., et al., DiProDB: a database for dinucleotide properties. Nucleic Acids Res, 2009. 37(Database issue): p. D37-40.

UDC 577.21

SEQUENCE-BASED AND STRUCTURE-BASED MACHINE-LEARNING MODELS FOR RECOGNITION OF 3'-END L1 AND ALU STEM-LOOPS IN HUMAN GENOME

A. Zaikin, A. Shein, M. Poptsova

Laboratory of Bioinformatics, Big Data and Information Retrieval School, Faculty of Computer Science, National Research University Higher School of Economics, Moscow, Russia
125319, Moscow, Kochnovsky proezd, 3
e-mail: mpoptsova@hse.ru

We built and evaluated two types of models: sequence-based and structure-based for recognition of 3'-end stem-loops of human L1s and Alus and found most important parameters contributing to recognition: Shift, Tilt and Rise, and aslo hydrophilicity.

Key words: transposons, L1, Alu, stem-loop, machine-learning, Random Forest

How LINE proteins recognize their own and SINE RNA remains unclear [1]. For several species it was experimentally shown that the LINE protein recognizes a secondary structure, such as a stem-loop, at the 3'-end of a transposon RNA. Moreover it was shown that in some organisms LINEs and SINEs have identical 3'-end sequences containing a stem-loop structure, which is essential for transposon RNA recognition by transposon proteins [2-4]. We discovered a conservative secondary structure at the 3'-end of human L1 (LINE) and Alu (SINE) transposons [5], as well as in different species across the tree of life (unpublished results). Despite the absence of similarity at the level of sequences, conservative position of this structure suggests its functionality.

In the present study we explored structural properties of 3'-end L1 and Alu stem-loops with machine-learning models. We built two types of models using two different types of features: sequence-based and structure-based. For sequence-based model we took frequencies of di- and trinucleotides counting occurrences of each k-mer moving with the 1 bp step along the sequence. For structure-based models we considered a stem, a loop, and a bulge. For stem we took RNA dinucleotide properties from DiPRoDb [6], which include structural parameters Shift, Slide, Rise, Tilt, Roll, Twist and physical and chemical properties such as enthalpy, entropy, free energy, and hydrophilicity.

We built Random Forest models with 2000 trees using scikit-learn library. The model implementation and all data analysis is available at github: https://github.com/AlexShein/transposons/.

The performance for two types of models are high with AUC in a range of 95-99%. However structure-based models allows extracting important structural properties.

The feature importance analysis of the structure-based model revealed structurally important characteristics in stem-loop structures of different groups of elements. Thus the parameter Shift followed by Rise and Tilt appeared to be more influential for recognizing the joint set of L1 and Alu stem-loops while Shift was shown to be more important for recognizing L1 3'-UTR stem-loops and Rise is more important for distinguishing between L1 and Alu 3'-end stem-loops. Additionally, the parameters of stem positions adjacent to the loop appeared to be more important for Alu recognition. The obtained results clearly demonstrate the existence of structural constrains for 3'-end stem-loops of L1 and Alu, which presumably play an important role in recognition of L1-Alu pairs by the L1 machinery. The constructed machine-learning models can be used for de novo discovery of transposon-related stem-loops.

References:
1. 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.
2. Hayashi, Y., et al., Mechanism by which a LINE protein recognizes its 3' tail RNA. Nucleic Acids Research, 2014. 42(16): p. 10605-10617.
3. Kajikawa, M. and N. Okada, LINEs Mobilize SINEs in the Eel through a Shared 3' Sequence. Cell, 2002. 111(3): p. 433-444.
4. 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.
5. 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.
6. Friedel, M., et al., DiProDB: a database for dinucleotide properties. Nucleic Acids Res, 2009. 37(Database issue): p. D37-40.