mRNA *IGF2R* gene, 12 sites in the mRNA *SEMA6D* gene, six sites in the mRNA the *IRS1* gene and four sites in the mRNA *NCOA3* gene; for miR-574-5p, six sites in the mRNA *DMD* gene, three sites in the mRNA *KLF7* gene, 13 sites in the mRNA *VSNL1* gene, 11 sites in the mRNA *XRCC1* gene, seven sites in the mRNA *AFF3* gene, three sites in the mRNA *ARID3B* gene, nine sites in the mRNA *KIAA2018* gene.

References:
1. Siyanova E.Y., Mirkin S.M. Expansion of trinucleotide repeats // Molecular Biology. - 2001. - №2. - p. 208-223.
2. Ivashchenko A., Berillo O., Pyrkova A., Niyazova R., Atambayeva S. MiR-3960 binding sites with mRNA of human genes // Bioinformation. - 2014. - 10 (7). - P. 423-427

ПОИСК НЕКАНОНИЧЕСКИХ СТРУКТУР ДНК КАК НУКЛЕОСОМНЫХ БАРЬЕРОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Теванян Э.А., Попцова М.С.

Лаборатория биоинформатики, Департамент больших данных и информационного поиска, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Россия, 125319, г. Москва, Кочновский проезд, д.3
e-mail: etevanian@hse.ru; mpoptsova@hse.ru

Мы обучили алгоритм случайного леса для распознавания паттернов взаимного расположения нуклеосом и вторичных структур ДНК, которые могут служить барьерами для нуклеосом, в геноме мыши. Мы показали, что среди четырех типов рассмотренных структур (Z-ДНК, H-ДНК, G-квадруплексов и участков SIDD) лучшее качество модели достигается для G-квадруплексов и H-ДНК.

Ключевые слова: вторичные структуры ДНК, G-квадруплексы, H-ДНК, Z-ДНК, нуклеосомные барьеры, расположение нуклеосом, методы машинного обучения, случайный лес

Геномы обладают большим потенциалом образования вторичных структур ДНК, которые воздействуют на различные геномные процессы, включая транскрипцию. Одним из механизмов регуляции транскрипции является регуляция расположения нуклеосом. Хотя нуклеосомы наматываются только на каноническую, так называемую, B-форму, другие, отличные от B-формы, структуры ДНК могут конкурировать с нуклеосомами за расположение в геноме или служить барьерами, разделяющими нуклеосомные массивы.

В данной работе мы использовали данные перманганат/S1 нуклеаз-футпринтинга [1], который позволяет определить потенциальные сайты образования вторичных структур ДНК на участках расплетенной ДНК. Для определения расположения нуклеосом мы использовали данные MNAse-seq [1]. Компьютерная аннотация генома мыши вторичными структурами ДНК производилась с помощью следующих программ и алгоритмов: Z-ДНК – Zhunt [2], H-ДНК - Inverted Repeats Finder [3], G-квадруплексы – QuadParser [4], участки SIDD – алгоритм из [5]. В результате в геноме мыши существует потенциально 250-420 тысяч сайтов образования структур каждого типа, в то время как число вторичных структур ДНК, обогащенных участками расплетенной ДНК, которые были обнаружены при помощи метода перманганат/S1 нуклеаз-футпринтинга, составляет 4-8% от числа структур, предсказанных с помощью компьютерных программ.

Для анализа взаимного расположения нуклеосом и вторичных структур ДНК мы рассмотрели области в 500 п.о., центрированных на структуре. Анализ нуклеосомных профилей вокруг структур ДНК выявил три типа паттернов: 1) структура окружена нуклеосомой с двух сторон, 2) структура расположена только с одной стороны нуклеосомы и 3) на участке нет нуклеосом.

Используя статистику динуклеотидов и триплетов, мы построили модель машинного обучения (классификатор по алгоритму случайного леса), распознающую принадлежность региона к одному из паттернов. Качество модели по метрике ROC-AUC достигло 86% для G-квадруплексов, 79% для H-ДНК, 73% для SIDD и 63% для Z-ДНК.

Модель может быть улучшена добавлением новых характеристик, в числе которых физические и химические свойства ДНК последовательности, такие, как энтальпия, энтропия, энергия Гиббса, гидрофильность, а также структурные свойства спирали (Shift, Rall, Slide, Rise, Tild, Bend), доступные в базе данных DiProDB [6]. Возможно, добавление новых характеристик и не улучшит качество модели, но даст возможность определить, какие структурные свойства ДНК играют важную роль в классификации паттернов.

Литература:
1. Kouzine F., Wojtowicz D., Baranello L., et all. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome // Cell Syst. – 2017. – №4(3). – c.344-356.
2. Ho P.S., Ellison M.J., Quigley G.J., and Rich A.A. Computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences // EMBO J. – 1986 – №5(10). – c.2737-2744.
3. Warburton P.E., Giordano J., Cheung F., Gelfand Y., and Benson G. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes // Genome Res. – 2004. – №14 (10A) . – c.1861-1869.
4. Huppert J.L., and Balasubramanian S. Prevalence of quadruplexes in the human genome // Nucleic Acids Res. – 2005. – №33 (9) . – c.2908-2916.
5. Wang H., Noordewier M., and Benham C.J. Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters // Genome Res. – 2004. – №14(8) . – c.1575-1584.
6. Friedel M., Nikolajewa S., Suhnel J., and Wilhelm T. DiProDB: a database for dinucleotide properties // Nucleic Acids Res. – 2009. – №37 (Database issue). – c.37-40.

SEARCHING FOR NON-B-DNA STRUCTURES AS NUCLEOSOME BARRIERS WITH MACHINE LEARNING METHODS

E. A. Tevanyan, M. S. 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: etevanian@hse.ru; mpoptsova@hse.ru

We trained Random Forest model to recognize patterns of nucleosome and non-B DNA structures, considered as potential nucleosome barriers in the mouse genome. We showed that among four types of structures – Z-DNA, H-DNA, G-Quadruplexes and SIDD regions – recognition of G-Quadruplexes and H-DNA showed the best performance.

Key words: DNA structures, G-quadruplexes, H-DNA, Z-DNA, nucleosome barriers, nucleosome positioning, machine-learning methods, random forest

Non-B DNA structures have a great potential to form and influence various genomic processes including transcription. One of the mechanisms of transcription regulation is nucleosome positioning. Even though only B-DNA can be wrapped around a nucleosome, non-B DNA structures can compete with nucleosomes for a genomic location or serve as barriers separating arrays of nucleosomes.

Here we used the data for mouse genome from Permanganate/S1 Nuclease Footprinting [1] that could detect potential sites of unwound DNA with non-B DNA structures formed in the unwound region. For nucleosome maps we took MNase-Seq data from [1]. Computer annotations of mouse genome with DNA secondary structures were performed with the following programs: Z-DNA– Zhunt [2], H-DNA - Inverted Repeats Finder [3], G-quadruplexes – QuadParser [4], SIDD – the algorithm is taken from [5]. The number non-B DNA structures in mouse genome, inferred by computer methods ranges in 250-420 thousands structures for each type, and the number of non-B DNA structures enriched in the regions of single-stranded DNA detected with Permanganate/S1 Nuclease Footprinting [1] comprises 4-8% of the computer predicted structures.

To analyze an association of nucleosomes and DNA structures we considered a region of 500 bp centered on a DNA structure. Nucleosome profiles around DNA structures revealed three types of patterns: a structure is surrounded by nucleosomes from both sides, from one side, or the region around a structure is nucleosome free.

We built and trained machine learning models (Random forest classifier) to recognize regions containing a particular pattern based on di- and trinucleotide composition of 500 bp region containing the pattern. Model performance reached 86% AUC for G-quadruplexes, 79% for H-DNA, 73% for SIDD and 63% for Z-DNA.

The model can be improved by taking into account physical and chemical properties of DNA such as enthalpy, entropy, Gibbs energy, hydrophilicity, as well as helical structural properties of dinucleotides (Shift, Rall, Slide, Rise, Tild, Bend) available in DiProDB [6]. Even though it might not improve the model it could provide an understanding, which properties mainly contribute to the classification model.

References:
1. Kouzine F., Wojtowicz D., Baranello L., et all. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome // Cell Syst. – 2017. – №4(3). – c.344-356.

2. Ho P.S., Ellison M.J., Quigley G.J., and Rich A.A. Computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences // EMBO J. – 1986 – №5(10). – c.2737-2744.
3. Warburton P.E., Giordano J., Cheung F., Gelfand Y., and Benson G. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes // Genome Res. – 2004. – №14 (10A) . – c.1861-1869.
4. Huppert J.L., and Balasubramanian S. Prevalence of quadruplexes in the human genome // Nucleic Acids Res. – 2005. – №33 (9) . – c.2908-2916.
5. Wang H., Noordewier M., and Benham C.J. Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters // Genome Res. – 2004. – №14(8) . – c.1575-1584.
6. Friedel M., Nikolajewa S., Suhnel J., and Wilhelm T. DiProDB: a database for dinucleotide properties // Nucleic Acids Res. – 2009. – №37 (Database issue). – c.37-40.

УДК: 577.2, 577.1

ПОИСК НОВЫХ ГЕНОВ В «СКРЫТОЙ» ЧАСТИ ТРАНСКРИПТОМОВ СЕЛЬСКОХОЗЯЙСТВЕННЫХ РАСТЕНИЙ

М.А. Генаев[1], Н.А. Шмаков[1], З.С. Мустафин[1], А.М. Мухин[1,2], Д.К. Константинов[1,2], А.В. Дорошков[1,2], С.А. Лашин[1,2], Д.А. Афонников[1,2]

[1]Федеральный Исследовательский Центр Институт цитологии и генетики СО РАН,
[2]Новосибирский Государственный Исследовательский Университет

Проведен массовый анализ RNA-seq экспериментов, содержащихся в публичном архиве ENA для пяти видов сельскохозяйственных культур: рис, ячмень, картофель, кукуруза и томат. Всего было обработано ~1300 экспериментов. Для каждого эксперимента проведена реконструкция последовательностей транскриптов de novo, и на основе выравнивания определены транскрипты, которые не имеют сходства с референсным геномом. Доля последовательностей, которые выровнялись на геном, но не попали на аннотированные локусы, составляет 20-25% от всех транскриптов, а доля невыравненных последовательностей составляет до 5%. Анализ не выравненных и не аннотированных транскриптов показал, что некоторые из них имеют высокий уровень сходства с вирусами и другими патогенами растений, последовательностями некодирующих РНК растений, рибосомных РНК, повторов. Среди “новых” генов нами также были идентифицированы последовательности возможных генов устойчивости к патогенам.

Ключевые слова: сельскохозяйственные культуры, транскриптом, секвенирование, аннотация генов, омиксные базы данных

Анализ транксриптомов сельскохозяйственных культур на основе экспериментов RNA-seq является одним из эффективных направлений поиска генов, связанных с такими признаками как устойчивость к вредителям и факторам среды. Однако большинство результатов RNA-seq представляют собой анализ экспрессии генов на основе картирования прочтений на референсный геном. Мы предположили, что новые гены в геномах культурных растений могут быть обнаружены на основе идентификации транскриптов которые (а) не выравниваются на референсный геном либо (б) выравниваются на ранее неаннотированные геномные локусы. Чтобы оценить долю таких новых генов в геномах пяти сельскохозяйственных культур (кукуруза, рис, томат, картофель и ячмень) проведен массовый анализ транскриптомов, взятых из доступных SRA архивов ENA (всего ~1300 библиотек).

Для каждого транскриптома проведена реконструкция последовательностей de novo, и на основе выравнивания определены транскрипты, которые не имеют сходства с референсным геномом (новые транскрипты) или выравниваются в ранее неаннотированные его локусы (неаннотированные транскрипты). Оказалось, что для 5 организмов в среднем доля выравненных транскриптов, составила в среднем 96% от их общего числа. Доля последовательностей, которые выровнялись на геном, но не попали на аннотированные локусы, составляет 20-25% от всех транскриптов, а доля невыравненных последовательностей составляет до 5%.

Сравнение полученных транскриптов с последовательностями из специализированных баз данных аннотаций нуклеотидных последовательностей показало, что контаминация векторными последовательностями незначительна (в среднем около 20 транскриптов на каждый транскриптом). В то же время среди невыравненных последовательностей нами были обнаружены последовательности,