doi 10.18699/bgrs2024-01-08

Структурная организация эукариотического генома на примере *C. merolae*

Руденко В.*, Коротков Е.
ФИЦ Биотехнологии РАН, Москва, Россия
** v.m.rudenko@gmail.com*

Ключевые слова: дисперсные повторы; повторы в ДНК; семейства повторов; позиционно-весовая матрица; *Cyanidioschyzon merolae*

Мотивация и цель: Как известно, эукариотическая ДНК информационно избыточна. Она содержит большое число повторов, тандемных и дисперсных. В данной работе ставилась задача выявления и последующего изучения дисперсных повторов в геноме эукариотического организма *Cyanidioschyzon merolae*. Для обнаружения повторов использовался IP-метод [1], который ранее успешно применялся для поиска повторов в геномах бактерий. Обсуждаются достоинства и недостатки IP-метода для поиска дисперсных повторов, а также возможное функциональное значение повторов.
Методы и алгоритмы: Поиск дисперсных повторов проводился в геноме примитивного растительного организма – красной водоросли *Cyanidioschyzon merolae* (https://plants.ensembl.org/Cyanidioschyzon_merolae/Info/Annotation/#assembly).
Геном содержит 20 хромосом и имеет длину 16,546,747 н.п. Также исследовались последовательности ДНК хлоропласта и митохондрии. Геном *C. merolae* содержит относительно мало генов – 5,331, из которых аннотировано 4,984; практически все гены лишены интронов.
Метод IP, используемый в данной работе для поиска дисперсных повторов, имеет несколько преимуществ. Во-первых, он относится к категории *de novo* методов, т. е. не использует никакой информации о строении повтора. Во-вторых, он способен выявлять сильноразмытые дисперсные повторы. Это означает, что число замен на одно основание x между последовательностями отдельных представителей семейства повторов может достигать 1.5. В то же время, как было показано в [1], мощность традиционных методов аннотации повторов, таких как RepeatMasker [2], Repeat Detector (RED) [3], RECON [4] и т. д., ограничивается $x<=1.0$. Такая чувствительность IP-метода достигается за счет того, что для поиска подобий используется обобщенный профиль семейства повторов, который настраивается на конкретный геном в ходе итерационной процедуры. Также учитываются парные корреляции соседних оснований в профиле семейства повторов. В-третьих, метод строит консенсус семейства повтора по множественному выравниванию последовательностей, входящих в него. Это позволяет охарактеризовать семейство и сравнить его с существующей таксономией повторов.
Результаты: В геноме *C. merolae* было выделено 20 семейств повторов, общее число повторов составляет 33,938 [5]. Повторы равномерно распределены по хромосомам: около 2 повторов приходятся на каждые 1000 н.п., они покрывают 72.64 % генома. Длины повторов варьируют от 108 до 600 н.п., средняя длина повтора – 523 н.п.

Определялось пересечение повторов с генами. Оказалось, что 86 % генов пересекаются с повторами таким образом, что длина пересечения превышает половину длины гена. Поскольку гены в ядерной ДНК занимают около 44 % хромосомы, то можно сделать вывод, что найденные нами повторы предпочитают располагаться в кодирующих районах ДНК.

Полученные данные по дисперсным повторам мы сравнили с имеющейся по геному *C. merolae* аннотацией с сайта http://plants.ensembl.org. Для аннотации повторов там применялись программные средства Dust [6], TRF [7], RepeatMasker [2] и RED [3]. Повторы, идентифицируемые программами Dust и TRF, – это low complexity regions и тандемные повторы; они не являлись целью нашего исследования. Поэтому для корректного сравнения были выбраны только дисперсные повторы, определяемые RepeatMasker и RED. RepeatMasker относится к категории методов, которые используют существующие базы данных по дисперсным повторам. Для аннотации *C. merolae* применялись базы данных Redat и Repbase. RepeatMasker и RED детектировали 20,320 повторов, которые покрывают 28.09 % генома. Длина повторов варьируется в широких пределах: от 15 до 19,220 н.п. Однако в целом это более короткие повторы, чем найденные методом IP, средняя длина их составляет всего 260 н.п.

Также было изучено пересечение классов аннотированных повторов с найденными нами семействами. Пусть $v(i,j)$ – число пересечений i аннотированного класса и j семейства повторов найденного IP-методом. Определим величину $v`(i,j)$:

$$v'(i,j) = \frac{\left(v(i,j) - np(i,j)\right)}{\sqrt{np(i,j)\left(1 - p(i,j)\right)}}\ , \qquad (1)$$

где $x(i) = \sum_j v(i,j),\ y(j) = \sum_i v(i,j)$, $p(i,j) = x(i)y(j)/n,\ n = \sum_i \sum_j v(i,j)$

$v`(i,j)$ имеет приближенно стандартное нормальное распределение и показывает степень корреляции между существующими классами и нашими семействами. Большие позитивные значения матрицы $v`(i,j)$ были обнаружены для семейств № 11–13 и классов LTR-повторов Copia и Gypsy.

Для всех 20 семейств повторов на основании множественного выравнивания последовательностей, по которым генерировался профиль семейства, строились консенсусы семейств. Визуальные представления консенсусов были получены с помощью программы Weblogo [8]. Консенсус для самого крупного 1-го семейства повторов изображен на рис. 1. Позиции в консенсусе представлены нуклеотидами, высота которых пропорциональна частоте их встречаемости, преобразованной в биты. Можно видеть, что данное семейство достаточно размыто; некоторые участки внутри повторов жестко определены, другие же относительно случайны.

Выводы: Применение IP-метода к анализу генома *C. merolae* показало, что в этом геноме можно выделить 20 семейств повторов. Семейства включают в себя около 34 тысяч дисперсных повторов, которые покрывают более 72 % генома. Это гораздо больше, нежели было определено ранее. Однако этот факт не умаляет достоинств других программных средств аннотации повторов, таких как RepeatMasker и RED. Они ориентированы на поиск коротких сильногомологичных повторов, которые наш метод определяет не в полной мере. Тем не менее некоторые классы аннотированных повторов, как было показано, коррелируют с найденными IP-методом семействами. Мы рекомендуем для

полноценной геномной аннотации повторов использовать совокупность различных биоинформационных подходов.

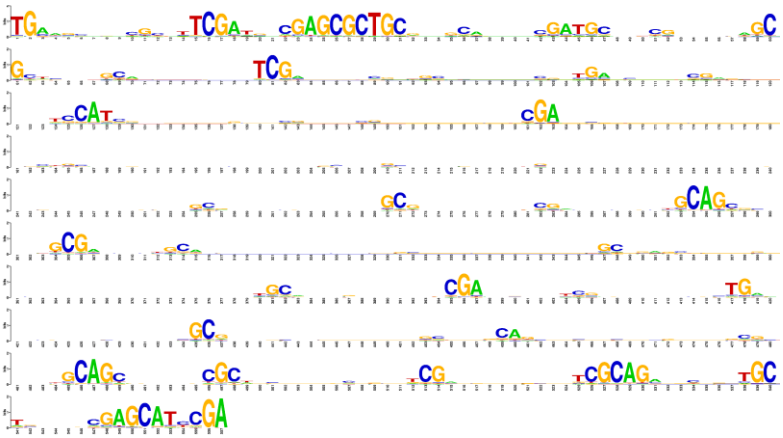

Рис. 1. Консенсус 1 семейства повторов

Выявленные в работе повторы покрывают значительную часть генома. Более правильно было бы назвать их структурными элементами или некоторой разметкой, которая накладывается на весь ядерный геном. В дальнейшем было бы интересно понять, является ли такая разметка уникальной в пределах биологического вида, или же это общее свойство генома.

Мы предполагаем, что подробная структурная организация может быть связана с компактизацией хроматина. Возможно, консервативные участки в повторах представляют собой сайты связывания гистонов в процессе образования нуклеосом.

Финансирование: Исследование поддержано грантом РНФ (№ 24-24-00031).

Structural organization of the eukaryotic genome by example of *C. merolae*

Rudenko V.*, Korotkov E.

Research Center of Biotechnology of RAS, Moscow, Russia

** v.m.rudenko@gmail.com*

Key words: dispersed repeats; repeated DNA; position weight matrix; repeat family; *Cyanidioschyzon merolae*

Motivation and Aim: As is known, eukaryotic DNA is informationally redundant. It contains a large number of tandem and dispersed repeats. The aim of this work is identifying and studying of dispersed repeats in the genome of the eukaryotic organism *Cyanidioschyzon merolae*. To detect repeats, we used the IP method [1], which was previously successfully used to search for repeats in bacterial genomes. The advantages and disadvantages of the IP method for searching for dispersed repeats are discussed, as well as the possible functional significance of the repeats.

Methods and Algorithms: The search for dispersed repeats was carried out in the genome of a primitive plant organism the red alga *C. merolae* (https://plants.ensembl.org/Cyanidioschyzon_merolae/Info/Annotation/#assembly). The genome contains 20 chromosomes; the length is 16,546,747 bp. Chloroplast and mitochondrial sequences were considered too. The *C. merolae* genome has only 5,331 genes 4,984 of which are annotated. Almost all genes lack introns.

The IP method used to search for dispersed repeats has several advantages. Firstly, it belongs to the category of *de novo* methods, so it does not use any information about the structure of the repeats. Secondly, it is capable of identifying highly dispersed repeats; the degree of similarity x between different repeats from the family can reach 1.5. It was shown in [1] the power of traditional repeats annotation methods such as RepeatMasker [2] and Repeat Detector (RED) [3], RECON [4], etc. is limited by $x<=1.0$. This sensitivity of the IP method is achieved due to the fact that a generalized repeat family profile is used to search for similarities, which is tuned to a specific genome during an iterative procedure. Pairwise correlations of neighboring bases in the repeat family profile are also taken into account. Thirdly, the method builds a consensus of the repeat family based on multiple alignment of the sequences included in it. This allows to characterize the family and compare it to existing repeat taxonomy.

Results: 20 repeat families have been identified in the *C. merolae* genome, with a total number of repeats of 33,938 [5]. The repeats are uniform distributed along the chromosomes, about 2 repeats for every 1000 bp and cover 72.64 % of the genome. Repeat lengths vary from 108 to 600 bp the average repeat length is 523 bp.

The intersection of repeats with genes was determined. It turned out that 86 % of genes intersect with repeats, such that the length of the intersection exceeds half the length of the gene. Since genes in nuclear DNA occupy about 44 % of the chromosome, we can conclude that the repeats prefer to be located in the coding regions of the DNA.

We compared the obtained data on dispersed repeats with the annotation available for the *C. merolae* genome from the website http://plants.ensembl.org. At this site repeats were annotated by the software tools Dust [6], TRF [7], RepeatMasker [2] and RED [3]. The repeats identified by the Dust and TRF programs are low complexity regions and tandem repeats; they were not the purpose of our study. Therefore, only dispersed repeats identified by RepeatMasker and RED were selected for comparison. RepeatMasker belongs to the category of methods that use existing databases of dispersed repeats. The Redat and Repbase databases were used to annotate *C. merolae*. RepeatMasker and RED detected 20,320 repeats, which covered 28.09 % of the genome. The length of the repeats varies widely: from 15 to 19,220 bp. However, in general, these are shorter repeats than those found by the IP method; their average length is only 260 bp.

We also examined the intersection of classes of annotated repeats with the families we found. Let $v(i,j)$ be the number of intersections of the i annotated class and the j family of repeats found by the IP method. Let's define the value $v`(i,j)$:

$$v'(i,j) = \frac{\left(v(i,j) - np(i,j)\right)}{\sqrt{np(i,j)\left(1 - p(i,j)\right)}}, \qquad (1)$$

where $x(i) = \sum_j v(i,j),\ y(j) = \sum_i v(i,j)$, $p(i,j) = x(i)y(j)/n,\ n = \sum_i \sum_j v(i,j)$

$v`(i,j)$ has an approximately standard normal distribution and shows the degree of correlation between existing classes and our families. Large positive values of the matrix

v`(i,j), indicating the presence of a correlation, were found between families No. 11–13 and the LTR repeat classes Copia and Gypsy.

For all 20 repeat families consensuses were constructed in text and numerical form Consensus based on multiple sequence alignment from which the family profile was generated. Visual representations of consensuses were obtained using the Weblogo program [8]. The consensus for the largest 1st repeat family is shown in Figure 1. Consensus positions are represented by nucleotides whose height is proportional to their frequency of occurrence converted to bits. It can be seen that this family is quite diverged, there is no strong homology between individual repeats over their full length. Some regions within repeats are strictly defined, while others are relatively random.

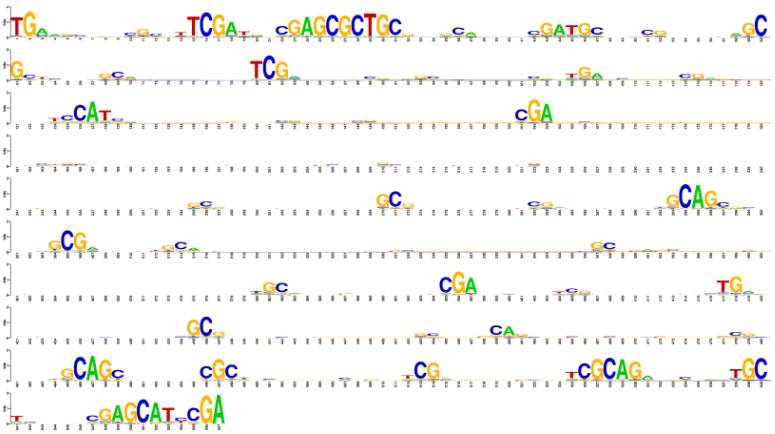

Fig. 1. Consensus for the first repeat family

Conclusion: The application of the IP method to the analysis of the *C. merolae* genome showed that 20 repeat families can be identified. The families include about 34 thousand dispersed repeats, which cover more than 72 % of the genome. This is much more than was previously determined. However, this fact does not detract from the merits of other repeat annotation software tools, such as RepeatMasker and RED. They are focused on searching for short, strongly homologous repeats, which our method does not fully detect. However, some classes of annotated repeats are correlating with families found by the IP method. We recommend using a combination of different bioinformatics approaches for complete genomic annotation of repeats.

The repeats identified in the work cover a significant part of the genome. It would be more correct to call them structural elements or some kind of marking that is superimposed on the entire nuclear genome, including both coding and non-coding regions. In the future, it would be interesting to understand whether such marking is unique within a biological species, or whether this is a general property of the genome.

We hypothesize that the structural organization of the DNA may be related to chromatin compaction. It is possible that conserved regions in the repeats represent histone binding sites in the formation of nucleosomes process. By changing the location of nucleosomes, the DNA molecule gets a chance to switch to the translation of alternative genes to adapt to changes in external conditions or ensure tissue specificity.

Funding: The study is supported by RSF (No. 24-24-00031).

Список литературы/References

1. Korotkov E., Suvorova Y., Kostenko D., Korotkova M. Search for dispersed repeats in bacterial genomes using an iterative procedure. *Int J Mol Sci*. 2023;24:10964. doi 10.3390/IJMS241310964/S1
2. RepeatMasker Home page. Available online: https://repeatmasker.org/ (accessed on Aug 11, 2022)
3. Girgis H.Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics.* 2015;16:227. doi 10.1186/S12859-015-0654-5/TABLES/4
4. Bao Z., Eddy S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*. 2002;12:1269-1276. doi 10.1101/GR.88502
5. Rudenko V., Korotkov E. Study of Dispersed Repeats in the Cyanidioschyzon merolae Genome. *Int J Mol Sci*. 2024;25:4441. doi 10.3390/IJMS25084441
6. Morgulis A., Gertz E.M., Schäffer A.A., Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol*. 2006;13:1028-1040. doi 10.1089/CMB.2006.13.1028
7. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573-580. doi 10.1093/nar/27.2.573 1999;27
8. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. WebLogo: a sequence logo generator. *Genome Res*. 2004;14:1188-1190. doi 10.1101/GR.849004