ИДЕНТИФИКАЦИЯ СОВПАДАЮЩИХ ФРАГМЕНТОВ НЕКОДИРУЮЩИХ УЧАСТКОВ МИТОХОНДРИАЛЬНЫХ ГЕНОМОВ НАСЕКОМЫХ ОТРЯДА ПОЛУЖЕСТКОКРЫЛЫЕ (ЛАТ. HEMIPTERA)

В.И. Чесалин, Н.В. Воронова-Барте, С.С. Левыкина, В.Л. Крук, Н.С. Мешич, П.Е. Александрович

Белорусский государственный университет, г. Минск, Беларусь
E-mail: vitaliy.krukh@gmail.com, nikitameshich@gmail.com

Разработан алгоритм построения множества максимальных по включению общих фрагментов в некодирующих участках митохондриальных геномах насекомых.

Ключевые слова: митохондриальные геномы; молекулярная генетика; насекомые.

Введение. Митохондриальные геномы насекомых представляют собой кольцевую молекулу ДНК, размером от 15 до 18 kb. Они содержат 13 белок-кодирующих генов, 22 гена митохондриальных тРНК и 2 гена рРНК (12S-рРНК и 16S-рРНК). Как и у других животных, в митохондриальных геномах насекомых отсутствуют интроны и протяженные межгенные спейсеры. Однако в структуру мтДНК входят две некодирующие области: Repeat-регион, или область тандемных повторов, встречающийся только у части видов; а также обязательный участок – Control-region, или область формирования D-петли [1].

Repeat-регион – участок митохондриальных геномов, длиной в около 1000 п.н., который содержит длинные, тандемно расположенные повторы [2]. Наличие такого крупного участка в митохондриальных геномах представляет большой интерес, поскольку эволюция мтДНК идет по пути компактизации. В настоящее время функциональное значение региона-повторов, а также его происхождение не известно [3]. Control-регион, или область формирования D-петли – участок митохондриальных геномов, представляющий собой AT-богатую область, функциональное назначение которого – инициация репликации ДНК [4,5]. Известно, что положение в митохондриальном геноме и длина области формирования D-петли вариабельны, что обусловлено наличием в них тандемных повторов, функциональное значение которых остается дискуссионным [4].

В нашей работе мы сделали попытку установить наличие или отсутствие структурного сходства между repeat- и control-регионами, приняв в качестве гипотезы утверждение, что обнаруженное сходство в мотиве нуклеотидных последовательностей будет указывать на общность происхождения repeat-региона и D-петли. Для проверки нашей гипотезы мы по-

ставили перед собой задачу разработать алгоритм для обнаружения областей подобия с возможностью проведения множественного попарного сравнения между видами и между таксонами (в прямой, обратной и комплементарной ориентациях), а также с возможностью устанавливать нужный процент совпадений, поскольку на данный момент не существует оптимального пакета программ, позволяющего быстро и эффективно выполнить наш анализ с минимальными затратами ресурсов.

Материалы и методы. Для проверки гипотезы была создана выборка из 262 нуклеотидных последовательностей полных мтДНК насекомых, депонированных в открытом доступе в базе данных GenBank. Из полных геномов были извлечены последовательности, соответствующие областям control- и repeat-регионов. Сводные данные о составе выборки представлены в таблице.

Таксономическая выборка

Таксон насекомых	Количество геномов		Суммарное количество геномов
	С repeat-регионом	Без repeat-региона	
Coccoidea	3	4	7
Aleyrodidae	4	6	10
Psylloidea	4	32	36
Cicadellidae	1	57	57
Heteroptera	10	84	94
Aphididae	35	18	53

Для каждого таксона T с контрольной областью C и региона повторов R рассмотрим множество общих фрагментов последовательностей

$$X_T = X_C \cap X_R, \quad (1)$$

где X_C – множество всех фрагментов последовательностей из контрольной области C и X_R – множество всех фрагментов последовательностей из региона повторов R. На множестве X_T введем отношение порядка $\preccurlyeq$ следующим образом: пусть

$$\mathrm{a}, \mathrm{b} \in X_T, \mathrm{b} = (b_1, b_2, \ldots, b_n) \quad (2)$$

Тогда полагаем

$$\mathrm{a} \preccurlyeq \mathrm{b} \leftrightarrow \exists (i,j), 1 \le i \le j \le n, a = \left(b_i, b_{i+1}, \ldots, b_j\right) \quad (3)$$

Таким образом, пара $(X_T, \preccurlyeq)$ образует частично упорядоченное множество, максимальные элементы которого характеризуют степень сходства между последовательностями C и R. Для анализа сходства последовательностей мы используем алгоритм, изображенный на рис. 1. Данный алгоритм можно разделить на 2 этапа:

1. Создание множеств фрагментов для каждой последовательности. Поиск их пересечения.
2. Поиск множества максимальных элементов.

2.1. Создание сортированного по длине последовательностей списка S на основе пересечения множеств фрагментов.

2.2. Создание пустого списка M_s для хранения максимальных элементов.

2.3. Итеративный проход по отсортированному множеству строк. Для каждой строки из множества S проверяем, входит ли она в некоторую строку из M_s. Если нет, добавляем её в M_s.

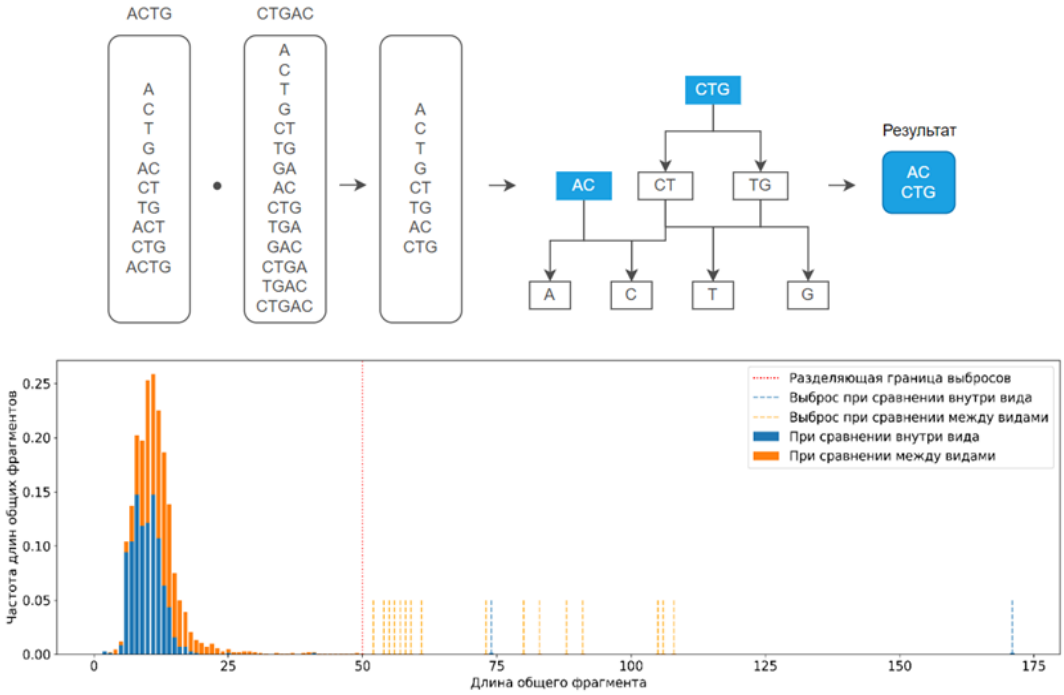

Рис. 1. Иллюстрация работы алгоритма

Результаты. Несмотря на установление жесткого порога для сходства в 100 % и минимальной длины последовательностей в 50 нуклеотидов (полагая, что при 100 % совпадении последовательностей длиной 50 нуклеотидов и выше мы можем исключить вероятность случайного совпадения) были обнаружили 2 случая, в которых в control- и repeat-регионах присутствовали идентичные участки длиной 279 нуклеотидов в геноме Aphididae (рис. 2) и три участка длиной по 185, 199 и 239 нуклеотидов в одном геноме Heteroptera (рис. 3). Наличие столь протяженных идентичных участков в двух отдельных областях генома, являющихся некодирующими, едва ли является случайным. Это не позволяет нам утверждать, что repeat-регион был образован в результате дупликации и транслокации участков control-региона, однако дальнейшие исследования с установлением меньшего порога сходства, вероятно, позволят нам ответить на этот вопрос.

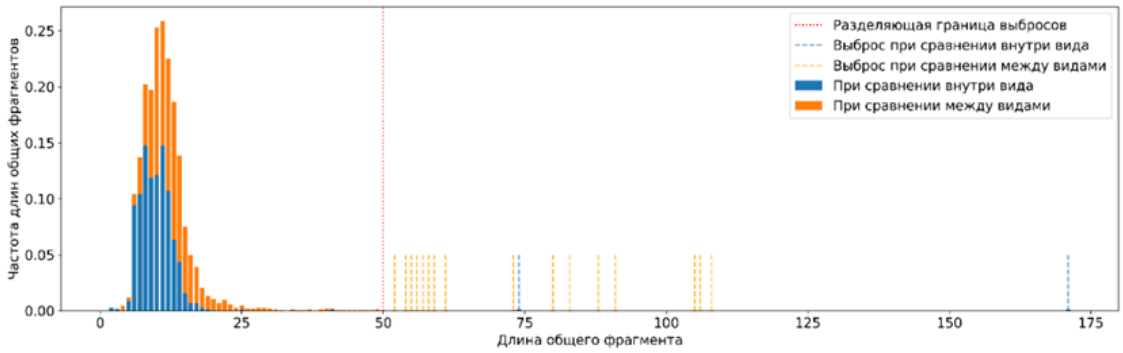

Рис. 2. Сгруппированная столбчатая гистограмма длин максимальных фрагментов таксона Aphididae.

Рис. 3. Сгруппированная столбчатая гистограмма длин максимальных фрагментов таксона Heteroptera.

Разделяющая граница выбросов – величина, задающая формальный критерий того, что мы считаем выбросом: выбросами считаются все величины, которые больше данной границы.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. The mitochondrial genome of Greenidea psidii van der Goot (Hemiptera: Aphididae: Greenideinae) and comparisons with other Aphididae aphids. / Chen J. [et al.] // International Journal of Biological Macromolecules. 2019. Vol. 122. P. 824-832.
2. *Воронова Н.В., Левыкина С.С., Бондаренко Ю.В., Шулинский Р.С.* Некодирующие области в митохондриальном геноме тли Aphis fabae mordvilkoi Börner & Janisch, 1922. // Молекулярная и прикладная генетика. 2019. Т. 26. С. 64-72.
3. Characteristic and variability of five complete aphid mitochondrial genomes. / Voronova N.V. [et al.] // International Journal of Biological Macromolecules. 2020. Vol. 149. P. 187-206.
4. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. / Zhang D.X. [et al.] // Biochemical Systematics and Ecology. 1997. Vol. 25. P. 99-120.
5. A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. / Kasamatsu H. [et al.] // PNAS. 1971. Vol. 68. P. 2252-2257.