УДК 51-76, 519.688

Поиск коротких фрагментов в анализе нуклеотидных последовательностей*

Л.У. Ахметзянова[1,2], О.Ю. Кирьянова[2], И.М. Губайдуллин[1,2]

Институт нефтехимии и катализа – обособленное структурное подразделение Федерального государственного бюджетного научного учреждения Уфимского федерального исследовательского центра Российской академии наук (УФИЦ РАН)[1], Уфимский государственный нефтяной технический университет[2]

В работе рассматривается задача поиска коротких фрагментов (праймеров) в нуклеотидной последовательности цепи ДНК. Праймеры являются необходимым компонентом для проведения полимеразной цепной реакции (ПЦР). ПЦР – экспериментальный метод увеличения концентрации определенного фрагмента ДНК в биологическом материале. Праймеры – это короткие синтетические нуклеотидные последовательности, длина которых 10-30 нуклеотидов [1].

Необходимо найти позиции включения прямого и обратного праймера в нуклеотидной последовательности таким образом, чтобы расстояние между ними варьировалось в диапазоне от 50 до 500 нуклеотидов. Пример такого поиска представлен на рис.1.

Рис. 1. Схематичное изображение праймеров и ампликона при анализе генома для ПЦР.

Если представить последовательность ДНК в виде огромной строки (порядка нескольких миллиардов элементов), то задачу поиска праймеров можно сформулировать следующим образом.

Необходимо найти позицию полного совпадения некоторого образца A в более длинной строке T, также определить все включения образца A в строке T. Эта задача схожа с поиском определенного слова в фрагменте текста.

Самый просто способ поиска – это пошаговое сравнение элементов образца в строке.

В данном случае «наивный» метод имеет ряд недостатков, а именно:

- Высокая сложность $O(m \cdot n)$, где m – длина образца A, n – длина строки T;
- Сдвиг на одну позицию в случае несовпадения элемента образца и строки. Таким образом происходит многократное сравнение заведомо неравных элементов;
- Не учитывается информация о предыдущих сравнениях при дальнейшем поиске.

*Работа выполнена при финансовой поддержке гранта РФФИ 17-44-020120

Данные недостатки учитывает алгоритм Бойера-Мура, который позволяет увеличить сдвиг в случае несовпадения элементов образца и строки [2]. Для этого вводится эвристика стоп-символа. Также посимвольное сравнение проводится справа налево, что также ускоряет процесс поиска.

Данный алгоритм был реализован средствами языка программирования Python с использованием библиотек Numpy для работы с математическими функциями и Biopython для работы с файлами, содержащими информацию о геномах.

Исследования проводились на геномах растений, данные о которых представлены в таблице 1.

Таблица 1.Перечень исследуемых геномов,
с указанием величины нуклеотидных пар (н.п.).

Геном	Размер генома (н.п.)
Arabidopsis thaliana (L.) Heynh	130 000 000
Solanum tuberosum L	1 000 000 000
Triticum aestivum L	17 000 000 000

Для ускорения расчетов был применен JIT-компилятор Numba. Numba – это компилятор just-in-time для Python, применяемый для работы с кодом, в котором используются массивы NumPy, функции и циклы [3]. Также расчеты проводились с применением параллельной директивы Numba prange(). Для более качественной оценки времени расчета поиск проводился на 6 различных праймерах размером 10 нуклеотидов.

На рис. 2 представлено сравнение времени расчета с применением компилятор Numba и параллельной директивой компилятора Numba.

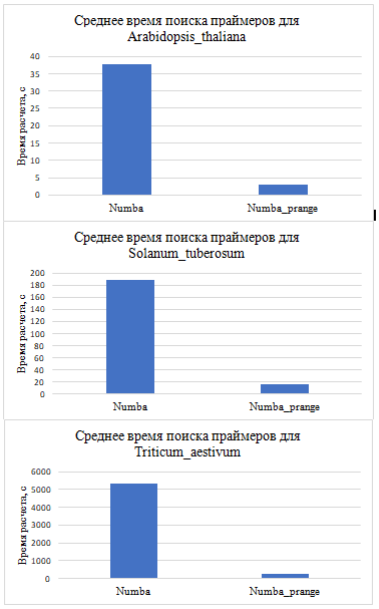

Рис. 2. Сравнение времени расчета поиска праймеров с использованием компилятора Numba и использованием функции prange().

Таким образом применение функции prange() позволило ускорить время расчета в среднем в 11 раз. В целом, реализация алгоритма Бойера-Мура с применением компилятора Numba позволяет значительно сократить время поиска праймеров в задачи планирования ПЦР.

Литература

1. Гарафутдинов Р. Р., Баймиев А. Х., Малеев Г. В., Алексеев Я. И., Зубов В. В., Чемерис Д. А., Кирьянова О. Ю., Губайдуллин И. М., Матниязов Р. Т., Сахабутдинова А. Р., Никоноров Ю.М ., Кулуев Б. Р., Баймиев А. Х., Чемерис А. В. Разнообразие праймеров для ПЦР и принципы их подбора // Биомика. 2019. Т. 11, № 1. С. 23-70.

2. Гасфилд Д. Строки, деревья, и последовательности в алгоритмах. Информатика и вычислительная биология. Пер. с англ. И. В. Романовского. СПб.:Невский Диалект; БХВ-Петербург. 2003. 654 с.

3. Numba: A High Performance Python Compiler. URL: http://numba.pydata.org/ (10.04.2020).

MSC2020 92B05, 68W32

Search for short fragments in nucleotide sequence analysis

L.U. Akhmetzyanova[1,2], O.Yu. Kiryanova[2], I.M. Gubaidullin[1,2]

Ufa Branch of the Russian Academy of Sciences[1], Ufa State Petroleum Technical University [2]