УДК 631.523.13: 004.021

РАЗРАБОТКА АЛГОРИТМА ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА ИЗВЛЕЧЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ЦЕЛЕВЫХ ГЕНОВ ИЗ АННОТИРОВАННОГО ГЕНОМА

В.Р. Вертелко, Р.С. Шулинский

Белорусский государственный университет

Введение. Зачастую, в ходе своей работы исследователи сталкиваются с задачей анализа последовательностей генов или групп генов в уже аннотированном геноме. Данная задача является весьма трудоемкой, так как данные полногеномной сборки представляют собой огромный массив данных, при работе с которым существует вероятность допустить ряд ошибок, что негативно скажется на результатах работы. С подобной задачей мы сами сталкивались неоднократно, поэтому было принято решение о необходимости разработки программного модуля для извлечения целевых генов из полногеномных данных по ключевым идентификаторам, находящимся в стандартной аннотации: «CDS», «gene», «gene name» и т.д. Подобный подход позволит исключит потенциальные ошибки, возникшие в процессе поиска генов в длинных последовательностях, а также в значительной степени укорит сам процесс.

Материалы и методы. Для решения поставленной задачи использовали пакеты репозитория Bioconductor, написанные на языке программирования R (RBiostrings, Shortread и т.д.). Также, в ходе разработки, нами был использован API (Application Programming Interface), который позволяет автоматизировать запросы к базе данных NCBI (National Centre for Biotechnology Information).

Результаты и выводы. Для корректной работы модуля на используемом компьютере должен быть установлен R, а также все необходимые пакеты. В случае если таковых не обнаружено, будет произведена автоматическая установка.

После установки модуль предложит оператору заполнить поле «accession number» (уникальный номер, который присвоен каждому объекту базы данных NCBI), проводится проверка на наличие такового объекта, после чего модуль автоматически скачивает нужные файлы: файл с геномом или его частичными последовательностям, а также файл с его аннотацией в fasta и gff формате соответственно. После загрузки файлов программа останавливает работу и ждёт команды от

оператора. В случае, если скаченные файлы соответствуют искомым, исследователь подтверждает корректность выполнения предыдущего шага и программа продолжает выполнение. Пользователю предоставляется возможность указать ряд дополнительных параметров, на основе которых будет производиться формирование результирующей выборки последовательностей. Далее алгоритм преобразовывает файл аннотации в табличный вид, где каждый столбец таблицы соответствует полю записи из аннотации.

Исходя из заданных параметров, в сформированной ранее таблице осуществляется поиск интересующих исследователя генов (формирование выборки генов). Далее, на основе данных об их расположении в полногеномной нуклеотидной последовательности, производится копирование соответствующей области из генома.

Таким образом, был разработан модуль для автоматизации процесса извлечения последовательностей целевых генов из ранее аннотированного генома.

Литература

1. The R Project for Statistical Computing [Electronic resource]. – Mode of access: https://www.r-project.org/. – Date of access: 07.12.2018.

2. RStudio (IDE). – Mode of access: https://www.rstudio.com/. – Date of access: 08.12.2018.

3. Bioconductor - package repository for bioinformatics. – Mode of access: https://www.bioconductor.org/. – Date of access: 08.12.2018.