SHARK (SEQUENCE HANDLING AND RESAMPLING KIT) – НОВОЕ ПРОГРАММНОЕ СРЕДСТВО ДЛЯ ОБРАБОТКИ ВЫБОРОК НУКЛЕОТИДНЫХ И АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

П. Ю. Кветко, Н. В. Воронова

Белорусский государственный университет
Минск, Беларусь
e-mail: pavelkvetko@gmail.com

Проблема предварительной очистки данных в биоинформатических исследованиях при больших объемах обрабатываемой информации стоит особенно остро. Сегодня не существует универсального программного инструмента для подготовки выборок НК и АК-последовательностей для дальнейшего анализа. Нами создано и протестировано программное средство для решения данных задач – SHaRK (Sequence Handling and Resampling Kit).

Ключевые слова: программное средство; очистка данных; ресэмплинг; нуклеотидные последовательности; аминокислотные последовательности.

SHARK (SEQUENCE HANDLING AND RESAMPLING KIT) – NOVEL SOFTWARE TOOL FOR NUCLEOTIDE AND AMINO ACID SAMPLES HANDLING

P. Yu. Kviatko, N. V. Voronova

Belarusian State Unversity,
Minsk, Belarus

Data cleaning is crucial to bioinformatical analyses when dealing with large datasets. Yet, there is no universal data cleaning tool for nucleotide and amino acid sample files handling prior to analysis. We introduce a novel software for dealing with this task – SHaRK (Sequence Handling and Resampling Kit).

Keywords: software; data cleaning; resampling; nucleotide sequences; amino acid sequences.

ВВЕДЕНИЕ

Многие работы в области молекулярной филогении, эволюционной, популяционной и медицинской генетики, особенно связанные с оценкой полиморфизма любого рода или предусматривающие использование биоинформатического, статистического и математического анализа, требуют использования больших выборок нуклеотидных и аминокислотных (НК и АК) последовательностей, содержащих множество однотипных элементов, в том числе полученных из электронных баз данных. Однако «сырые»

данные, получаемые непосредственно «с прибора» или из электронных баз данных, часто содержат некорректные, нецелевые либо некачественные последовательности вследствие ошибок систем поиска последовательностей, секвенирования, автоматической аннотации и некоторых других [1]. В частности, в файлы, полученные из электронных баз данных, часто включаются последовательности нецелевых генов или фрагментов неподходящего диапазона, множество дублирующих идентичных последовательностей, записи, содержащие продолжительные вырожденные участки и т. д. В связи с этим практически любому виду анализа выборок аминокислотных и нуклеотидных последовательностей предшествует процедура очистки данных (англ. data cleaning), необходимая для исключения из анализа не подходящих по тем или иным параметрам или вовсе нецелевых последовательностей [2].

Поскольку структура данных и содержание в них посторонних элементов зачастую нерегулярны, их очистка должна производиться при непосредственном контроле со стороны человека, т. е. данную процедуру весьма сложно алгоритмизировать. В то же время «ручная» подготовка данных может становиться лимитирующим по времени фактором, значительно замедляющим процесс, поскольку обработка больших массивов данных сложна для человека. Для решения этой проблемы исследователи прибегают как к созданию скриптов, предназначенных для решения частных задач конкретного исследования, так и к использованию готовых программных продуктов с теми или иными функциями первичной обработки данных. К сожалению, поскольку центральная функциональность существующих программных продуктов не связана с первичной обработкой данных, исследователям приходится пользоваться разрозненными инструментами для проведения полной процедуры очистки данных перед целевым анализом. В связи с этим целью нашей работы стала разработка программного средства, эффективно решающего задачи, возникающие при подготовке выборок НК- и АК-последовательностей. Как уже было сказано, среди областей, где найдет применение такое средство, можно особенно выделить молекулярную филогенетику, популяционную, эволюционную и медицинскую генетику, т. е. различные направления исследований, связанные с анализом большого числа однотипных последовательностей.

МАТЕРИАЛЫ И МЕТОДЫ

Разработка программного обеспечения осуществлялась с использованием языка программирования Python версии 3.5.1. Графический интерфейс программы создавался с использованием фреймворка Qt версии 5.4.0. При создании программного инструмента были использованы модули из стандартной библиотеки языка Python 3.5.1, а также модули из библиотек в составе дистрибутива Anaconda3 версии 4.0.0, а именно библиотеки Biopython версии 1.66 и scipy версии 0.17.1; реализация графического интерфейса осуществлялась с помощью средств библиотеки PyQt версии 4.11.4.

Множественное выравнивание последовательностей было реализовано с использованием алгоритма MUSCLE версии 3.8.31.

Была проведена эмпирическая оценка времени работы программы в зависимости от объема обрабатываемых данных. Для этого в качестве входных данных использовали различного объема выборки нуклеотидных последовательностей гена COI и измеряли время работы созданной программы. Выбор объекта для тестирования продиктован большими объемами последовательностей данного гена, так как он используется в качестве маркера при проведении ДНК-баркодинга.

Последовательности гена COI получены из базы данных CBOL. Размер выборок варьировал от 10 последовательностей при размере файла 5 Кб до 403 274 последовательностей и 658 Мб соответственно.

Эмпирический анализ проводили с использованием двух аппаратных конфигураций (см. рис. 2).

В случаях обеих аппаратных платформ тестирование проводили на ОС Windows 7 Ultimate ×64.

РЕЗУЛЬТАТЫ

Программа SHaRK имеет версии для запуска в ОС Linux и Windows, при этом предусмотрены версии как для запуска с использованием системного интерпретатора Python, так и standalone-версия, независимая от наличия системного интерпретатора.

Выгодной чертой интерфейса программы является возможность выбора между интерфейсом командной строки и графическим. С помощью программы возможна обработка файлов, содержащих аминокислотные и нуклеотидные последовательности в формате FASTA и GenBank.

В общем смысле, используя программу SHaRK, пользователь приобретает возможность решить ряд задач, схематично представленных на рис. 1.

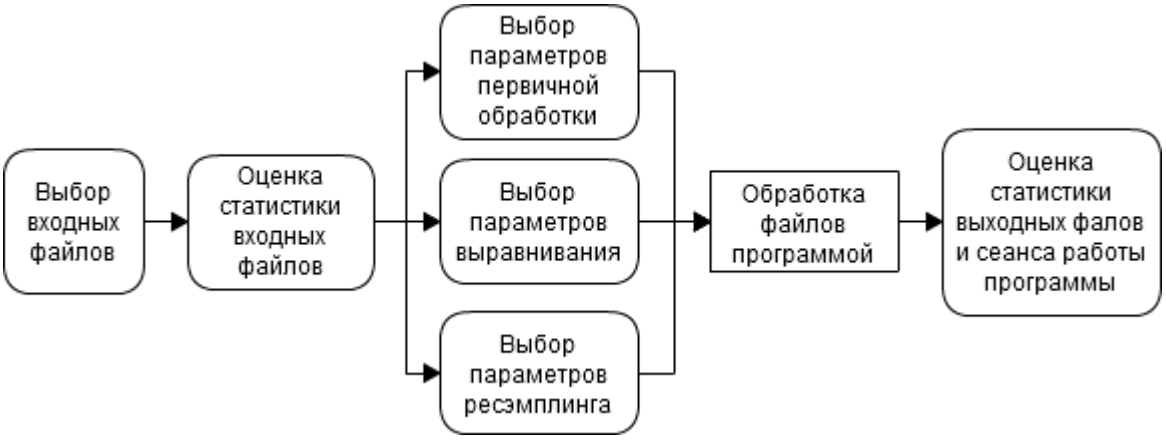

Рис. 1. Алгоритм действия пользователя при работе с выборками НК и АК последовательностей с использованием программы SHaRK

При помощи программы SHaRK пользователь может производить положительный отбор последовательностей по наличию в них целевых текстовых меток среди всех последовательностей одного файла. Одновременно можно задать функцию отрицательного отбора, т. е. удаление из файла последовательностей, содержащих метки, назначенные пользователем как «нежелательные», что повышает эффективность отбора и информативность результирующих выборок. В случаях, если целевыми метками являются названия генов либо белков, достаточно использовать любой из синонимов их названий, поскольку программа включает таблицу соответствия названий генов и их продуктов, которая может быть дополнена пользователем.

Для удобства пользователя программа позволяет ознакомиться со сводной статистикой по составу входных и выходных файлов. Программа производит многофакторную, основанную на множественном выравнивании, оценку качества последовательностей и удаляет не подходящие под пользовательские критерии. В случаях, если не-

обходимо гарантировать уникальность таксономического положения источников последовательностей, программа позволяет удалять повторяющиеся элементы, при удалении руководствуясь относительным качеством последовательностей: длиной, степенью перекрытия с референсом и другими последовательностями, количеством пробелов, содержанием вырожденных нуклеотидов и др. Референсная последовательность в формате FASTA или GenBank, используемая для выравнивания, может быть добавлена пользователем при необходимости, и может включаться в качестве первой последовательности в выходные файлы. При работе с файлами из нескольких источников пользователь может объединять данные в один выходной файл, параллельно производя конвертирование форматов последовательностей. Опционально выводится сводная информация по области перекрытия последовательностей во входных файлах, что необходимо при форматировании последовательностей по длине.

Поскольку результаты некоторых видов статистического анализа ДНК зависят от количества последовательностей в анализе (например, оценка уровня генетических различий, числа выявляемых гаплотипов и т. д.), в программе предусмотрена возможность нормализации выходных файлов по объему. Пользователь может создавать подвыборки рандомизированного состава и заданного объема (реплики), причем может задаваться как число, так и процент удаления или сохранения последовательностей в исходном файле.

По окончании работы программы опционально может выводиться отчет о прошедшем сеансе работы, включающий общее время работы, количество обработанных файлов, их объем, время, затраченное на каждый файл, количество созданных копий и др.

На рис. 2 графически представлены результаты эмпирической оценки времени, затрачиваемого на обработку файлов, в зависимости от их размера и количества содержащихся последовательностей.

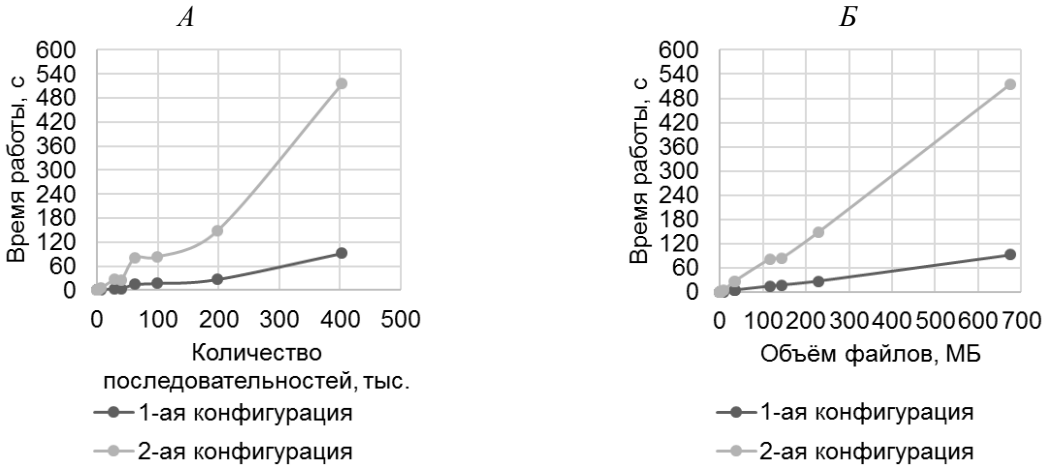

Рис. 2. График времени работы программы в зависимости от количества последовательностей (*А*) и объема (*Б*) обрабатываемых файлов

На рисунке отражены временные характеристики работы программы на двух аппаратных конфигурациях:

1) шестиядерный процессор AMD FX6300 с тактовой частотой 3,5 ГГц, 8 Гб оперативной памяти DDR3 с тактовой частотой 1866 МГц;

2) двухъядерный процессор AMD Fusion E-450 с тактовой частотой 1650 МГц, 4 Гб оперативной памяти DDR3 с тактовой частотой 671 МГц.

При работе с файлами, размер которых находился в диапазоне от 10 до 403 274 последовательностей (типичный размер рабочего файла от 100 до 10 000 последовательностей), зависимость времени работы от количества последовательностей приближалась к линейной логарифмической. Зависимость времени работы от объема файлов (от 5 Кб до 658 Мб) приближалась к линейной, что свидетельствует о высокой эффективности работы программы на использованных конфигурациях системы.

Созданное программное средство было использовано при подготовке данных в ряде работ, связанных с молекулярной филогенетикой животных [3–5].

ЗАКЛЮЧЕНИЕ

Создана программа SHaRK, представляющая собой эффективное средство предварительной обработки данных (очистки выборок и проведения ресэмплинга) для подготовки выборок НК- и АК-последовательностей к последующему анализу.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Weber G. M., Mandl K. D., Kohane I. S. Finding the Missing Link for Big Biomedical Data // JAMA. 2014. Vol. 311, № 24. P. 2479–2480.
2. Rahm E., Do H. H. Data cleaninig: Problems and current approaches // IEEE DATA Eng. Bull. 2000. Vol. 23, № 4. P. 3–13.
3. Воронова Н. В., Кветко П. Ю. Тестирование гипотезы о связи интенсивности метаболизма и скорости молекулярной эволюции в генах митохондриального генома животных // Труды БГУ. 2015. Т. 10, Ч. 1. С. 160–167.
4. Кветко П. Ю., Воронова Н. В. Новый подход к установлению относительного возраста таксонов с использованием молекулярно-генетических данных: на примере mammalia // Междунар. науч.-практ. конф. ученых «Актуальные проблемы биологии, биотехнологии, экологии и биобезопасности», посвященная 80-летнию заслуженного ученого, профессора В. Л. Зайцева, Казахстан, 13 июля 2015 г., РГП «Научно-исследовательский институт проблем биологической безопасности» КН МОН РК. Казахстан. 2015. С. 162−165.
5. Voronova N. V., Kviatko P., Krivaia A. Dramatic difference in the rate of molecular evolution in closely related taxa with similar life history // Book of abstract of International EMBO/EMBL Symposium: New Model Systems for Linking Evolution and Ecology. Heidelberg, Germany, 08–11 May, 2016. P. 56.