

АЛГОРИТМЫ ДЛЯ АНАЛИЗА ДАННЫХ СЕКВЕНИРОВАНИЯ ДНК

Савик О.В.

*Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь*

Задедюрин Е.В. – к.т.н., доцент

В статье описан общий принцип работы методов секвенирования ДНК, приведены основные алгоритмы секвенирования, а также отмечены основное направление текущих исследований разработки новых методов разбиения и сборки ДНК.

ДНК (дезоксирибонуклеиновая кислота) — это длинная закрученная молекула, находящаяся в хроматине каждой клетки. Она может быть названа планом развития тела: это химическая запись роста человека, строения его костей, цвета волос, химического состава тела и всех других унаследованных характеристик. Когда клетки делятся и размножаются, ДНК обеспечивает их точное копирование путем правильного считывания заключенной в ней наследственной информации. ДНК скомплектована из четырех химических соединений и похожа на закрученную веревочную лестницу, которую называют двойной спиралью. Входящие в состав четырех химических соединений ДНК аденин и тимин всегда комбинируются вместе, образуя один тип перекладин «лестницы». Два других соединения ДНК — цитозин и гуанин — также всегда комбинируются вместе и образуют второй тип перекладин. [2]

Секвенирование (sequencing) – это общее название методов, которые позволяют установить последовательность нуклеотидов в молекуле ДНК. В настоящее время нет ни одного метода секвенирования, который бы работал для молекулы ДНК целиком; все они устроены так: сначала готовится большое число небольших участков ДНК («разрезание» её в случайных местах), а потом читается каждый участок по отдельности. При этом эффективность таких методов становится тем выше, чем длиннее и точнее получаются фрагменты в результате секвенирования. Затем происходит непосредственно сборка последовательности — применение алгоритмов для восстановления генома из его фрагментов. [1]

Различные методы секвенирования отличаются друг от друга лишь тем, как именно они позволяют прочесть получившийся набор из многочисленных копий одной и той же ДНК.

Существует несколько популярных методов разбиения и чтения ДНК, такие как:

– секвенирование по Сэнгеру;

- секвенирование лигированием;
- секвенирование второго поколения Illumina;
- секвенирование PacBio

Как показывает практика, старые методы секвенирования, так называемые “первого поколения”, выдают более подходящие результаты для сборки исходной ДНК. Это в основном выражается в длине тех участков ДНК, которые удаётся последовательно прочесть и которые, собственно, и нужно собрать в одну большую строчку. Секвенаторы первого поколения выдавали проанализированные участки длиной более пятисот нуклеотидов, обычно около тысячи. Современные секвенаторы выдают участки длиной около ста нуклеотидов.

Зачем же вообще использовать секвенаторы второго поколения, чем они лучше? Причина здесь, как это нередко бывает, чисто экономическая: современные секвенаторы гораздо дешевле. Проект по сборке первого человеческого генома, завершённый в 2003 году, занял 13 лет и стоил 3.8 миллиардов долларов. С тех пор цена секвенирования уменьшалась экспоненциально. Новые технологии секвенирования обещают научиться обрабатывать геном человека за \$1000 и даже меньше, что открывает возможности для массового секвенирования в медицинских целях.

На таком уровне становится важной и цена алгоритмической стороны вопроса. Чтобы сборка геномов не занимала дольше и не стоила дороже, чем само их секвенирование, нужно разработать очень быстрые алгоритмы для решения задачи сборки.

Список использованных источников:

1. Разработка системы секвенирования ДНК с использованием paired-end данных, интернет ресурс: <https://docplayer.ru/30883723-Razrabotka-sistemy-sekvenirovaniya-dnk-s-ispolzovaniem-paired-end-dannyh.html> Подключаемые по Wi-Fi пылесосы, интернет ресурс: <https://www.trendhunter.com/trends/neato-botvac/>.

2. Геномика: постановка задачи и методы секвенирования, интернет ресурс: <https://postnauka.ru/longreads/468>