Секция «Биоинженерия и биоинформатика»

Эффективность применения парных ридов для сборки бактериальных геномов

Лежнина Ксения Владимировна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: oxia.com@gmail.com

В настоящее время существует несколько методов секвенирования геномов: классический, основанный на методе Сангера, с длиной ридов (коротких фрагментов генома) около 800 нуклеотидов и методы, основанные на новых технологиях, использующие риды длиной 35-500 нуклеотидов (Illumina, 454 Life Science) [4] [5], что на порядок быстрее и дешевле. Однако новые технологии имеют и существенный недостаток: маленькая длина ридов не позволяет получить достаточно длинные контиги (длинные фрагменты последовательности генома, построенные из ридов) из-за наличия повторов в геноме, длина которых больше длины рида, что затрудняет построение качественных полных геномов. Поэтому были разработаны новые методы сборки [2], результатом работы которых является не только набор контигов, но и построенный на них граф [1].

Для решения этой проблемы используют парные риды [3]. Парным ридом называется пара ридов с известным расстоянием между ними. Парные риды позволяют «склеить» соседние контиги, при этом существенную роль играет наличие графа.

Данная работа посвящена исследованию эффективности применения парных ридов для улучшения сборки генома, а также изучению зависимости качества сборки генома от параметров парных ридов. Были найдены и проанализированы условия, при которых парный рид может улучшить сборку генома, и на основе этого был разработан и реализован на языке C++ алгоритм. С помощью написанной программы для 761 бактериальных геномов базы данных RefSeq были построены модельные сборки, то есть сборки с максимально достижимым качеством для данной длины ридов. Сравнение результатов, полученных при использовании парных ридов, с данными, полученными без участия парных ридов, позволило исследовать зависимость основных характеристик сборки (количество контигов, количество повторов, кратность повторов) от параметров парных ридов.

Анализ результатов показал, что далеко не все парные риды улучшают сборку, но применение «полезных» парных ридов существенно увеличивает качество сборки геномов: количество контигов в среднем уменьшается в 3-5 раз, количество повторов - в 6-8 раз.

Литература

1. Pevzner, P.A., Tang, H., and Waterman, M.S. 2001. A Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. 98: 9748–9753

2. Chaisson, M.J. and Pevzner, P.A. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18: 324–330.

3. Mark J. Chaisson, Dumitru Brinza and Pavel A. Pevzner Does the read length matter? Genome Res. 2009 19: 336-346 originally published online December 3, 2008

4. 454 Life Science: http://www.454.com

5. Illumina: http://www.illumina.com

Слова благодарности

Автор выражает благодарность своему научному руководителю Николаеву В.К.