Методика оценивания качества геномных сборок на основе частот k-меров

Романенков Кирилл Владимирович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: kromanenkov2@yandex.ru

Современные высокотехнологичные автоматические методы расшифровки последовательностей ДНК (секвенирование) позволяют в течение относительно короткого времени (несколько дней) получить сотни миллиардов коротких последовательностей (длиной 100–500 символов) из четырех букв A, T, G, C, полученных прочтением фрагментов входного образца ДНК.

Для сборки генома используют специальные программы - сборщики генома, которые объединяют короткие фрагменты, полученные на этапе секвенирования. Большинство из них основаны на концепции графа де Брюйна, которая заключается в следующем. Из коротких фрагментов, полученных на этапе секвенирования, формируются всевозможные подстроки длины k (k-меры). Таким образом, из строки длины l получается $l-k+1$ k-меров. Затем сборщик строит граф де Брюйна, вершинами которого являются k-меры, между двумя вершинами проводится ребро, если они имеют общий (k-1)-мер. После этого выполняется упрощение этого графа и все найденные в нем пути без разветвлений (контиги) попадают в ответ.

Обычно для работы геномных сборщиков необходимо задать несколько параметров. Достаточно распространена ситуация, когда результаты применения сборщиков или одного сборщика с разными параметрами существенно отличаются для одних и тех же входных данных. В настоящее время не существует единой методики выбора наилучшей сборки, одной из самых распространенных практик остается запуск сборщиков с различными параметрами, а затем выбор наилучшего варианта согласно метрике N50. Метрика N50 определяет величину, при которой контиги длиннее значения этой величины составляют половину собранного генома, но данная метрика никак не учитывает степень близости полученного генома к набору коротких фрагментов. Практически любой геном содержит повторяющиеся участки, однако, начиная с определенного значения k, k-меры в некотором роде однозначно идентифицируют геном; если посчитать частоты распределения k-меров для, к примеру, $k = 17$, получится, что большая часть из них встречается в геноме в единственном

экземпляре.

Исходя из этих соображений, предложена следующая методика оценки качества геномной сборки. Сначала строится гистограмма частот встречаемости k-меров в коротких фрагментах, полученных в результате секвенирования. Пример такой гистограммы для организма Encephalitozoon cuniculi fungus [1] при $k = 21$ изображен на рис. 1. По оси X отложены частоты встречаемости k-меров в наборе чтений, а по оси Y отложено количество всевозможных k-меров с данной частотой. Видно, что гистограмма имеет два пика: первый связан с ошибками чтения генома, а второй соответствует уникальным k-мерам в исходном геноме. Частота второго пика (около 116) обусловлена покрытием при чтении генома: количеством прочтений каждого символа. На рис. 2 изображена гистограмма частот встречаемости k-меров собранного генома Encephalitozoon cuniculi fungus с помощью программы Velvet. Из-за того, что каждый участок генома прочитывается несколько раз, каждый уникальный k-мер из собранного генома встречается во множестве k-меров коротких чтений около 116 раз.

Предлагается вычислять долю уникальных k-меров для собранного генома среди различных k-меров, взятых из некоторой окрестности второго пика на гистограмме распределения частот встречаемости k-меров в чтениях, по следующей формуле:

$$Q = \frac{\sum\limits_{i=1}^{|K_1^g|} [k_1^g(i) \in K_{uniq\_reads}]}{|K_1^g|}, \qquad (1)$$

где $K_1^g = \{k^g : k^g \in K^g, abundance(k^g) = 1\}$, $K_i^r = \{k^r : k^r \in K^r, abundance(k^r) = i\}$, $K_{uniq\_reads} = \bigcup_{i=a_{p\_l}}^{a_{p\_r}} K_i^r$, $[a_{p\_l}; a_{p\_r}]$ - некоторая окрестность из второго пика на гистограмме распределения k-меров коротких чтений, $abundance(k)$ - частота встречаемости k-мера k, K^r - множество всех k-меров коротких чтений, K^g - множество всех k-меров собранного генома.

Чем больше значение Q, тем лучше полученная сборка соотносится с результатами секвенирования. Таким образом, предложенная методика устанавливает соответствие между набором коротких чтений, полученных в результате секвенирования, и собранным геномом, позволяя более точно оценивать результат геномной сборки.

Иллюстрации

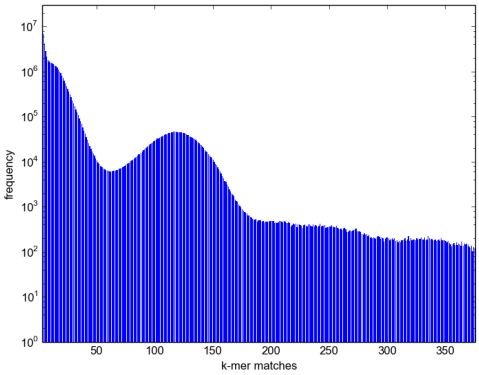

Рис. 1. Частоты встречаемости k-меров в наборе чтений для Encephalitozoon cuniculi fungus. Логарифмическая шкала

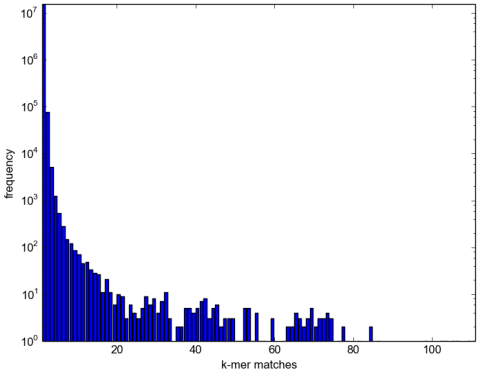

Рис. 2. Частоты встречаемости k-меров в собранном геноме Encephalitozoon cuniculi fungus. Логарифмическая шкала

Литература

1. European Nucleotide Archive. Encephalitozoon cuniculi: `http://www.ebi.ac.uk/ena/data/view/SRR122309`