УДК 004.891.2

РАЗРАБОТКА МЕТОДОВ ПОСТРОЕНИЯ И ВИЗУАЛИЗАЦИИ ГЕНОМНОГО КОНТЕКСТА С УЧЕТОМ HI-C СВЯЗЕЙ В МЕТАГЕНОМНЫХ ДАННЫХ.

Шостина А.Д. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),
Иванов А.Б. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н., доцент, Ульянцев В.И.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Предложен метод построения геномного контекста с учетом Hi-C связей, который был реализован на основе приложения MetaCherchant. Также разработан способ визуализации информации о Hi-C прочтениях на графе де Брейна. Предложенный метод визуализации был реализован в приложении Bandage.

Введение.

Существуют различные методы получения метагеномных данных. В данной работе рассматривается метод полногеномного секвенирования (WGS) и метод определения конформации хромосом Hi-C. WGS используется для определения нуклеотидной последовательности генома. В ходе полногеномного секвенирования со случайных позиций считываются фрагменты генома, называемые ридами. Затем на основе полученных ридов происходит определение последовательности генома целиком. Это называется сборкой генома. Для сборки генома используются графы де Брейна. Граф де Брейна — это ориентированный граф, вершинами которого являются различные k-меры, а ребрами k+1-меры. К-мер — это нуклеотидная последовательность длины k. Ребро соединяет два k-мера, если один k-мер является его префиксом, а другой k-мер его суффиксом. Сжатый граф де Брейна — это граф де Брейна, вершинами которого являются однозначно определенные нуклеотидные последовательности максимальной длины. Hi-C секвенирование используется для определения пространственной организации генома. Hi-C риды соединяют участки ДНК, которые в пространстве расположены близко к друг другу. Это могут быть риды из одного генома или риды из двух различных геномов.

В области здравоохранения существуют проблемы, связанные с анализом геномного контекста, например, проблема устойчивости бактерий к некоторым видам антибиотиков. Антибиотико-резистентные гены (АРГ) могут располагаться в плазмидах. Плазмиды — это небольшие молекулы ДНК, физически обособленные от хромосом. В природе плазмиды обычно содержат гены, повышающие приспособленность бактерий к окружающей среде. Например, бактерия приобретает свойство устойчивости к антибиотику, если плазмида с АРГ находится в клетке бактерии. При этом геном бактерии никак не меняется, поэтому при помощи метода WGS нельзя выделить бактерии, клетки которых содержат плазмиды с исследуемым АРГ. Однако такие бактерии можно найти при помощи Hi-C секвенирования.
Геномный контекст, это подграф графа де Брейна, в котором содержится исследуемый ген. Таким образом, анализ геномного контекста, построенного с учетом с Hi-C связей, поможет решить данную проблему.

Основная часть.

Для возможности проведения анализа расширенного геномного контекста необходимо разработать метод построения и визуализации геномного контекста с учетом Hi-C связей.
Построение геномного контекста будем проводить в три этапа:

1. Построение геномного контекста вокруг анализируемого гена.

2. Поиск Hi-C ридов, которые лежат вне построенного геномного контекста, однако имеют Hi-C связи с ним. В результате, мы выделим Hi-C связи, которые расширяют исходный геномный контекст.
3. Построение и визуализация объединенного геномного контекст вокруг анализируемого гена и всех Hi-C ридов, выделенных на этапе два.

Приложение MetaCherchant позволяет построить геномный контекст вокруг одной или нескольких нуклеотидных последовательностей. Следовательно, используя функционал приложения MetaCherchant, можно выполнить этапы один и три.

На этапе два предлагается выполнить картирование пар Hi-C ридов на контиги геномного контекста, построенного на этапе один. Картирование выполним при помощи утилиты bwa. В результате картирования каждому Hi-C риду будет сопоставлен контиг, в котором он находится. При этом Hi-C риды, которые не содержатся ни в одном контиге, будут отмечены как неизвестные. Затем найдем такие неизвестные Hi-C риды, парный рид которых был сопоставлен с контигом из исходного контекста. Тем самым, мы найдем все интересующие нас пары Hi-C ридов, в которых один рид лежит в исходном контексте, а другой – нет .

Для визуализации геномного контекста будем использовать приложение Bandage. Данное приложение не поддерживает отображение Hi-C связей, поэтому необходимо разработать способ их визуализации. Предлагается изображать Hi-C связи пунктирной линией, соединяющей середины контигов. Насыщенность цвета будет зависеть от количества Hi-C связей между данными контигами, это значение будем называть весом Hi-C связи.

Большое количество ребер загромождает граф де Брейна и усложняет его анализ, поэтому далее предлагаются различные способы уменьшения числа отображаемых Hi-C связей. При анализе геномного контекста используются только Hi-C связи с весом больше некоторого порогового значения, так как при WGS и Hi-C секвенировании могут происходить ошибки чтения, из-за чего мы не можем доверять всем Hi-C связям. Также следует отметить, что для дальнейшего анализа, как правило, интересны только Hi-C связи, соединяющие разные компоненты связности, поэтому в приложение Bandage была добавлена возможность фильтрации Hi-C связей по минимальному весу и возможность отображения только Hi-C связей между различными компонентами связности. Часто бывает важен только факт наличия Hi-C связей между двумя компонентами, а не отдельные пары Hi-C ридов. Поэтому в Bandage была добавлена возможность отображать ровно одну Hi-C связь между двумя компонентами связности.

Помимо этого, были рассмотрены различные способы использования Hi-C ребер при укладке графа. Наиболее удачным способом оказалось фиксирование длины одного Hi-C ребра между каждой парой компонент связности. Благодаря этому способу укладки компоненты связности, имеющие Hi-C связи между собой, будут расположены близко друг к другу. При этом расположение контигов внутри одной компоненты связности не будет нарушено. Однако данный способ не уменьшает количество пересечений между ребрами и вершинами в случае отображения всех Hi-C связей.

Выводы.

Таким образом, был разработан метод построения геномного контекста с учетом Hi-C связей, а также способ визуализации графа де Брейна с использованием Hi-C связей. Разработанные методы были реализованы на основе приложений MetaCherchant и Bandage. Данные методы были проверены на сгенерированных метагеномных данных. В дальнейшем планируется тестирование разработанных методов на реальных метагеномных данных.

Шостина А.Д. (автор) Подпись

Ульянцев В.И. (научный руководитель) Подпись