Опубликовано в материалах 2-й межвузовской научной конференции по проблемам информатики СПИСОК-2011, с. 321-325.

А. А. Сергушичев, В. В. Исенбаев,
Ф. Н. Царев, А. А. Шалыто
Санкт-Петербургский государственный университет информационных технологий, механики и оптики
Е. Б. Прохорчук
Центр «Биоинженерия» РАН

Разработка метода восстановления фрагментов нуклеотидных последовательностей по парным чтениям

Многие современные задачи биологии и медицины требуют знания генома живых организмов, который состоит из нескольких нуклеотидных последовательностей ДНК – по одной в каждой хромосоме. В связи с этим возникает необходимость в дешевом и быстром методе секвенирования, то есть определения последовательности нуклеотидов в образце ДНК.

Существующие технические средства не позволяют считать разом всю молекулу ДНК организма, поэтому перед считыванием молекулы ДНК дробятся на маленькие фрагменты. Изначально, например, для проекта «Геном человека» [1], использовались фрагменты длиной около 800–1000 нуклеотидов, но такой подход является дорогим и обладает низкой пропускной способностью – число прочитанных нуклеотидов в единицу времени. По этим причинам в настоящее время используется следующая достаточно дешевая и эффективная технология: сначала вычленяется случайно расположенный в геноме фрагмент длиной около 500 нуклеотидов, а затем происходит

считывание двух последовательностей с его концов (длиной примерно по 100 нуклеотидов каждая). Эти последовательности называются *парными чтениями*. Процесс повторяется такое число раз, чтобы обеспечить достаточно большое покрытие генома чтениями. Такая технология реализуется, например, в секвенаторах компании Illumina [2]. Заметим, что при физическом чтении могут возникать ошибки, но для каждого прочитанного нуклеотида известно его качество – вероятность того, что он был прочитан правильно.

Кроме того, программные средства тоже не позволяют восстановить геном целиком. Поэтому результатом работы сборщиков является набор *контигов*, объединенных в *скэффолды*. Контигами называются максимальные подпоследовательности генома, которые удалось восстановить. Скэффолдами называются такие последовательности контигов с оценками на расстояния между ними, про которые предполагается, что они в той же последовательности и на таких же расстояниях находятся в геноме.

На данный момент существует множество сборщиков геном из парных чтений [3-6], большинство из которых полностью основаны на использовании графа Де Брюина. Граф Де Брюина размерности k над алфавитом Σ – это граф, вершинами которого являются все строки над алфавитом Σ длины k (они называются *k-мерами),* а ребра соединяют пары таких вершин, что суффикс длины $k-1$ первой вершины является префиксом второй вершины. Можно заметить, что есть однозначное соответствие между ребрами и $(k+1)$-мерами – каждому ребру соответствует $(k+1)$-мер, полученный конкатенацией k-мера начальной вершине ребра и последнего символа k-мера конечной вершины ребра. В графе Де Брюина над алфавитом $\{A, T, G, C\}$ геном представляет из себя набор

путей (возможно не простых), соответствующих хромосомам.

Как уже говорилось, эти сборщики не позволяют собрать весь геном полностью, поэтому имеет смысл искать методы, которые позволят улучшить результаты. Кроме того, желательно оптимизровать использование памяти и быстродействие, чтобы уменьшить общую стоимость секвенирования.

В работе [7] была предложена идея с помощью графа Де Брюина восстанавливать фрагменты, отвечающие парным чтениям, а затем использовать полученные последовательности по 500 нуклеотидов для сборки генома с помощью имеющихся сборщиков из «длинных» чтений, основанных, например, на алгоритме Overlap-Layout-Consensus [8], таких как Newbler [9] или Celera Assembler [10]. К сожалению, предложенный в этой работе алгоритм восстановления фрагментов оказался неприменим для больших геномов размером от миллиарда нуклеотидов.

В данной работе делается попытка развить метод восстановления фрагментов, предложенный в работе [7], чтобы он стал эффективным по используемой памяти и времени работы и его можно было применить к геномам размером несколько миллиардов нуклеотидов.

Важным предварительным шагом является исправление ошибок, например, с помощью частотного анализа. Этот шаг позволяет сильно снизить общий объем информации во входных данных, при этом увеличив полезный объем.

Таким образом, весь процесс восстановления генома по парным чтениям состоит из четырех этапов:

1. Исправление ошибок в исходных данных.
2. Восстановление фрагментов. Восстановленные фрагменты будем называть квазиконтигами.
3. Сборка контигов из квазиконтигов.

4. Сборка скэффолдов.

Данная работа сфокусирована на втором этапе.

В предлагаемом методе используется подграф графа Де Брюина, в котором множество ребер состоит только из тех *(k+1)*-меров, которые встречаются в чтениях достаточно большое число раз, чтобы их можно было с очень большой вероятностью считать входящими в геном, а множество вершин такое же, как и в исходном графе. Если участок нуклеотидной последовательности покрылся достаточно хорошо, то есть все входящие в него *(k+1)*-меры по много раз входят в исходные данные, то в этом подграфе существует путь между первым и последним *k*-мерами участка

Предлагаемый метод основан на поиске такого пути для фрагмента, соответствующего парным чтениям. Из всех путей нас интересуют только те, которые укладываются в априорные границы длин фрагментов, поэтому слишком короткие и слишком длинные пути можно отбросить. Из оставшихся путей следует выбрать те, которые достаточно «похожи» на парные чтения. Для этого можно сравнить все нуклеотиды из парных чтений с соответствующими им нуклеотидами из пути, посчитать вероятность таких ошибок и сравнить ее с математическим ожиданием этой вероятности. Те пути, для которых различие между этими величинами велико, отбрасываются. Оставшиеся пути – хорошие кандидаты на роль пути, соответствующего фрагменту в действительности. Если нашелся единственный такой путь, то можно с очень большой уверенностью сказать, что он отвечает реальному пути в геноме, поэтому этот фрагмент считается восстановленным, а найденный путь выводится как квазиконтиг. Если таких путей несколько, то не ясно какой из них на самом деле соответствует фрагменту, поэтому этот фрагмент не восстанавливается.

Если не нашлось ни одного такого пути, то данный фрагмент ДНК был плохо покрыт чтениями и его восстановление невозможно.

Для того, чтобы потребление памяти предлагаемого метода было не очень большим, необходимо иметь компактное представление используемого подграфа графа Де Брюина. Для его хранения достаточно хранить только множество его ребер, что можно эффективно делать, используя, например, хеш-таблицу с открытой адресацией. Преимуществами такого подхода хранения перед другими являются его простота, быстродействие и возможность балансировки между используемой памятью и скоростью. Более эффективными с точки зрения потребляемой памяти являются rank/select словари [11], которые позволяют сделать ее использование близким к энтропии, но из-за этого увеличивается время доступа.

Для поиска путей, соединяющих две заданные вершины и не превосходящих по длине некоторой максимальной длины L_{max}, применяется методика meet-in-the-middle, в которой происходит одновременный поиск путей из первой вершины по прямым ребрам и путей из второй вершины по обратным ребрам. Если конец какого-то пути из первой вершины совпадает с концом пути из второй вершины, то можно, объединив эти пути, получить путь из первой вершины во вторую. Для реализации такого подхода удобно запустить одновременно два обхода в ширину: из первой вершины по прямым ребрам и из второй – по обратным. Тогда на каждом шаге L можно поддерживать инвариант: для первой вершины известны все исходящие из нее пути длины L_1, а для второй – все входящие пути длины L_2, причем $L_1+L_2=L$. Таким образом, на каждом шаге можно пересечь множества конечных вершин путей и найти все пути длины L из первой вершины во вторую. Для перехода к следующей итерации

необходимо выбрать, в каком из обходов увеличивать на единицу длину путей. Самым простым является поочередное увеличение длин, но для более эффективного использования памяти и времени лучше производить увеличение в том обходе, в котором на данный момент число концевых вершин меньше, чем в другом. Кроме того, разумно параллельно отбрасывать те пути, которые на текущей итерации уже сильно отличаются от парных чтений.

Заметим, что это решение хорошо масштабируется на несколько ядер или даже на несколько компьютеров, так как все парные чтения можно разбить на группы, которые можно обрабатывать независимо, а построение хеш-таблицы занимает относительно немного времени.

Данный подход был разработан и применен в рамках проекта *dnGASP* [12], в котором предлагалось восстановить синтетический геном размером в 1,8 миллиардов нуклеотидов. Восстановление фрагментов работало на 24-ядерной машине с 64 ГБ оперативной памяти. Значение k было выбрано равным *30* – в этом случае один k-мер можно хранить в 64-битном целом числе. За сутки было обработано 350 миллионов парных чтений, для 67% которых удалось восстановить исходные фрагменты. Для 6% не удалось найти ни одного пути, для 27% фрагментов однозначно восстановить не получилось из-за большого числа путей

В дальнейшем планируется увеличить k, а также исследовать возможность внести дополнительные улучшения в этот метод для увеличения его эффективности и быстродействия.

Литература

1. *International Human Genome Sequencing Consortium*. Initial sequencing and analysis of the human genome.

Nature, Volume 409. 15 February 2001. P. 860-921. http://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf
2. Illumina, Inc. http://www.illumina.com/
3. *Zerbino D. R., Birney E.* Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18. 2008. P. 821–829.
4. *Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J., Birol I.* ABySS: A parallel assembler for short read sequence data. Genome Research 19. 2009. P. 1117–1123.
5. *Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K., et al.* De novo assembly of human genomes with massively parallel short read sequencing. Genome Research 20. 2010. P. 265–272.
6. *Pevzner P. A., Tang H., Waterman M. S.* An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences 98. 2001. P. 9748–9753.
7. *Исенбаев В. В., Шалыто А. А.* Разработка системы секвенирования ДНК с использованием paired-end данных. http://is.ifmo.ru/genom/_isenbaev_thesis.pdf
8. *Pevzner P.* Computational molecular biology: an algorithmic approach. MIT Press. 2000. P. 61–62.
9. Products & Solutions - Analysis Tools - GS De Novo Assembler : 454 Life Sciences, a Roche Company. http://454.com/products-solutions/analysis-tools/gs-de-novo-assembler.asp
10. SourceForge.net: wgs-assembler. http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
11. Okanohara D., Sadakane K. Practical Entropy-Compressed Rank/Select Dictionary. Computing

Research Repository, arXiv:cs/0610001v1. 2006. http://arxiv.org/pdf/cs/0610001v1
12. De novo Genome Assembly Project (dnGASP). http://cnag.bsc.es