

Построение филогенетических сетей для штаммов вирусов

Лебедев Р. В., Новик Д. С., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Алексеев Н. В., к.ф.-м.н., ведущий научный сотрудник ФИТиП
Университета ИТМО

Введение

Геном представляет из себя совокупность наследственного материала, заключенного в клетке организма. Его можно представить в виде символической последовательности, в которой каждому элементу соответствует одно из 4 азотистых оснований: аденин, гуанин, тимин и цитозин. Имея такую геномную последовательность организма и зная каким образом ее характеризует та или иная ее область, можно получить огромное количество полезной информации о ее носителе.

Проблема заключается в том, что получить правильную последовательность целиком на сегодняшний день технически невозможно. Процесс обработки биологических данных называется секвенированием, результатом его работы является набор относительно небольших подпоследовательностей, называемых ридами, в которых к тому же могут присутствовать ошибки секвенирующей платформы. Гаплотипированием называется процесс получения набора геномов (гаплотипов) представленных в обрабатываемых биологических данных. На основе гаплотипов вирусов можно решать задачу построения их филогенетических сетей, имеющую ряд практических применений, например, восстановление сценариев и очагов заражения на основании клинических данных.

Поскольку качество построения филогенетических сетей напрямую зависит от качества сборки гаплотипов, появляется необходимость в качественном гаплотипере, способном хорошо находить и различать похожие гаплотипы, так как вирусам свойственна высокая скорость мутации, из-за чего геномы их штаммов могут быть очень близки.

Цель работы

Целью данной работы является разработка продвинутых алгоритмов гаплотипирования и построения филогенетических сетей штаммов вирусов.

Промежуточные результаты

Были изучены и проанализированы существующие методы гаплотипирования, среди которых SAVAGE[1], CliqueSNV[2], PredictHaplo[3], aBayesQR[4]. Определены их слабые и сильные стороны, поставлены задачи для дальнейшего исследования.

Описание предлагаемого подхода

Ведется адаптация подхода TCS[5] под NGS данные. Исследуются идеи в направлении применения теории гиперграфов к построению гаплотипов. Для построения филогенетических сетей планируется попробовать использовать не гаплотипы, а сами данные секвенирования.

Список литературы

1. Jasmijn A Baaijens, Amal Zine El Aabidine, Eric Rivals, Alexander Schönhuth. De novo assembly of viral quasispecies using overlap graphs //Genome research. – 2017.
2. Sergey Knyazev, Viachaslau Tsyvina, Andrew Melnyk, Alexander Artyomenko, Tatiana Malygina, Yuri B Porozov, Ellsworth Campbell, William M Switzer, Pavel Skums, Alex Zelikovsky. CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads //bioRxiv. – 2018. – С. 264242.

3. S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(1):182–191, Jan 2014.
4. Soyeon Ahn and Haris Vikalo. abayesqr: A bayesian method for reconstruction of viral populations characterized by low diversity. In *International Conference on Research in Computational Molecular Biology*, pages 353–369. Springer, 2017.
5. Templeton A. R., Crandall K. A., Sing C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation // *Genetics*. – 1992. – T. 132. – No. 2. – C. 619-633.