Алгоритмы веб

АЛГОРИТМ РЕКОНСТРУКЦИИ ГЕНОМОВ ПОПУЛЯЦИИ РНК-ВИРУСОВ

Артёменко А. В., Скумс П. В.

БГУ, Минск, Беларусь, e-mail: artyomenkoav@gmail.com

Многие РНК-вирусы, такие как ВИЧ и Гепатит С, существуют в инфицированном организме в виде совокупности генетически близких и постоянно мутирующих друг в друга вариантов (квазивидов) [4]. Это позволяет вирусам избегать имунного ответа и является причиной того, что эффективных методов лечения вызываемых этими вирусами заболеваний до сих пор не существует. Изучение генетической структуры популяций РНК-вирусов представляет собой исключительно важную для здравоохранения задачу. Длина геномов большинства РНК-вирусов составляет примерно 9000-10000 пар нуклеотидов, в то время как современные технологии секвенирования позволяют получить лишь фрагменты последовательностей РНК с длиной, не превышающей несколько сотен пар нуклеотидов. Реконструкция полноразмерных геномов из секвенированных фрагментов – одна из важнейших и наиболее известных задач биоинформатики. Подавляющее большинство существующих алгоритмов изначально разрабатывалось для восстановления фиксированных геномов (в частности, генома человека), и поэтому неприменимо для задачи реконструкции квазивидов, где требуется восстановить все генетически близкие варианты.

В настоящей работе предлагается алгоритм восстановления квазивидов, основанный на применении теории многопродуктовых потоков в сетях. Первым этапом алгоритма является простроение так называемого графа фрагментов, т.е. взвешенного ориентированного графа G, вершины которого (за исключением специально определенных источника и стока) соответствуют фрагментам геномов, ребра – перекрывающимся фрагментам, вес ребра пропорционален оценке вероятности того, что его концы соответствуют одному и тому же квазивиду. Можно показать, что каждый квазивид соответствует некоторому пути из источника в сток в графе G. На первом этапе алгоритма с помощью метода, предложенного в [1, 2], генерируются множество X всевозможных последовательностей-кандидатов, которые с достаточной вероятностью могут быть квазивидами. Затем для каждого кандидата C_k в граф G вводится пара вершин (s_k,t_k), а задача восстановления частот последовательностей-кандидатов сводится к задаче о нахождении многопродуктового потока максимальной стоимости в полученной сети с множеством пар источник-сток $\{(s_k,t_k) : C_k \in X\}$.

Литература

1. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe and A. Zelikovsky, “Inferring Viral Quasispecies Spectra from 454 Pyrosequencing Reads,” BMC Bioinformtaics 12(Suppl 6):S1 (2011).
2. K. Wesbrooks, I. Astrovskaya, D. C. Rendon, Y. Khudyakov, P. Berman, and A. Zelikovsky, “HCV Quasispecies Assembly using Network Flows,” Lecture Notes in Bioinformatics 4983, pp. 159-170.
3. Zagordi O, Klein R, Däumer M, Beerenwinkel N (2010). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Research, vol. 38 (21) pp. 7400-9
4. E. Domingo, "Biological significance of viral quasispecies", Viral Hepatitis Rev. 2, 1996, pp. 247-261.