Viral Quasispecies Assembly and Haplotype Frequency Estimation

Irina Astrova\-skaya1, Kelly Westbrook2, David Campo Rendon3, Yuri Khudyakov3, Piotr Berman4, Ion Mandoiu5, Gilberto Vaughan3, Laura Maia3, Joseph Forbi3, Alexander Zelikovsky1+

1Department of Computer Science, Georgia State University, 2Life Technologies, 3Centers for Disease Control and Prevention, 4Penn State University, 5University of Connecticut; * supported by Molecular Basis of Disease Program, GSU, **- supported by NSF Grant IIS-0916401

Abstract

In epidemiology, it is critical to understand the evolution of RNA viruses within a host to design effective drugs and treatments against newly evolved strains. Complete sequencing of a RNA virus sample from a single host is a promising approach to obtain the viral genome structure and to analyze its diversity. We explore sequence assembly of Hepatitis C Virus (HCV) based on 454 Life Sciences system and propose a scalable assembling method that recovers the most frequent quasispecies in the sample.

Background

Many viruses (e.g., SARS, HCV) encode their genome in RNA rather than DNA. RNA viruses are unable to detect and repair mistakes during replication due to the lack of DNA polymerase.

Mutations are passed down to descendants, producing a family of related variants of the ancestral genome referred as quasispecies. Many quasispecies are generated in a host with a single infection. This is due to mistakes during transcription and translation resulting in a mixture of different genome copies.

Pyrosequencing

GS FLX Titanium is a 454 Life Sciences High Throughput Pyrosequencing System:

- Divides the source genetic material into reads (300-800 bp)
- Sequences the reads
- Assembles the reads into the original genome via software

The software was originally designed to sequence a single organism; all reads are assembled to a single genome.

We need a software that assembles reads to multiple genomes for quasispecies sequencing.

Data

44 real quasispecies sequences (1739 bp long) from the 1E12 region of Hepatitis C virus (von Hahn et al. 2006)

Simulated reads:

- 4 populations sizes: 10, 20, 30, and 40 sequences.
- 3 population distributions: geometric, skewed normal, uniform.

4444 reads from 341 quasispecies sequences (309bp long) of HCV virus.

454 alignment: 48.78% of deletions and 28.83% of all insertions are likely to be due to 454 sequencing error.

Experimental Results

Positive Predictive Value (PPV)= \( \frac{TP}{TP+FP} \) and Sensitivity=\( \frac{TP}{TP+FN} \)

Fig 2. PPV and sensitivity for shortest path and max bandwidth path per vertex candidate selection on geometric population of 10 quasispecies with 60K reads being generated.

Run time:

- 2000 superreads (1 min) versus 5000 original reads (30 min)
- 25000 superreads (100 min) versus 25000 original reads (>12hours)

Candidate Path Selection

Choose the most probable source-sink path through each vertex.

Min Cost Flow

The Shortest Path per vertex

- Union of the shortest paths tree from the source and the shortest paths tree from the sink

\[
\varphi_l = \frac{e_{kl}}{c_{l}}
\]

No correct candidate quasispecies are lost with increasing steepness!

- The Max Bandwidth Path per vertex \((n=0)\)

Expectation Maximization for Viral Frequency Estimation

Bipartite graph:

- vertices are quasispecies and superreads
- an edge exist if quasispecies contains a read

E step:

- \( f_{kr} = \frac{a_{kr}}{\sum_i a_{ir}} \) per edge

M step:

- \( f_{kr} = \frac{p_{kr}}{\sum_i p_{ir}} \) per quasispecies

References