Algorithms for Analysis and Applications of High-Throughput Sequencing of Intra-Host Viral Populations

Alex Zelikovsky, Georgia State University

10th International Conference on Bioinformatics
Genomic and Evolution of Pathogens
11/21/2015 at Georgia Tech
RNA Virus: Intra-Host Population

High mutation rate (~10^{-4})

Lauring & Andino, PLoS Pathogens 2011
Intra-Host Viral Population: Curse or Blessing?

Curse – for sequencing:
• very low variability vs error rate
• analogue to very low signal-to-noise ratio
  • Mutation = signal rate 0.05%
  • Error rate = noise rate 0.1% / 2%

Blessing – for transmission inference
• Just a single sequence/no variation ➔ no information
  • limited information for relatedness
  • for inferring direction of transmission

This talk: deal first with Curse and then with Blessing
Intra-Host Viral Population Reconstruction from Single Amplicon NGS Reads
Introduction

- Viral spectrum reconstruction for RNA virus
- Technology: SMRT sequencing technologies (PacBio)
  - Long (up to 10 000bp)
  - High error rate (~15% → 3%)
  - Low coverage (30k → 100k reads)
Existing Algorithms

• **PredictHaplo** (Francesca Di Giallonardo et al.)
  – Probabilistic (Bayesian mixture) model with Dirichlet process to estimate number of haplotypes
  – Markov chain via Monte Carlo sampling for inference

• Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing (Dario A. Dilernia et al.)
ML Problem Formulation

• **Given:** set of reads $R$ from unknown haplotype set $H'$

• **Find:** set of haplotypes $H=\{H_1,...,H_k\}$ with corresponding frequencies $F=\{f_1,...,f_k\}$ maximizing $\Pr(R \mid H)$

**NOTE:** Given haplotypes, the frequencies can be reliably estimated via Expectation-Maximization

   – similarly to transcriptome quantification
Alignment

• Ideal: Multiple Sequence Alignment of all reads

• Challenge:
  – too many indels (10% of 2300bp sequences)
  – in too many reads (10K-30K)

• Solution:
  – Pairwise alignment to reference BWA (Li H. and Durbin R. (2009))
  – B2W (Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2009))

• Error rate:
  – After alignment the error rate reduces significantly
    • Majority of errors are random lengthy insertions
    • Alignment removes random insertions
Extract signal from noise

**Assumption:** Noise is random / signal is not!

For 2 positions I and J:
- Major/Major haplotype 11
- Major/Minor haplotype 12
- Minor/Major haplotype 21
- Minor/Minor haplotype 22

**Theorem:** Minor/Minor does not exist, then
for expected number of reads $E_{kl}$ ($k,l=1,2$)

$$E_{11} \cdot E_{22} \leq E_{12} \cdot E_{21}$$

**Definition:** let $X$ be binomial distribution with
$p = (A_{12} \cdot A_{21} / A_{11} \cdot n)$,
$A_{kl}$ ($k,l=1,2$) and $n =$ observed number of reads

If $\text{Prob}(X > A_{22}) < 0.01/(N \ choose \ 2)$, then
These two minor alleles are **linked**
Haplotyping

- **Initial cluster** C contains all reads
- **Label** C complex
- **Repeat**
  - If there is a complex cluster
    - **Find** pair of linked SNVs
    - **If** it exists split that cluster on 2 parts
    - **else** label current cluster as simple.
- **Until all clusters are simple**
- **Calculate frequencies** with kGEM
Experimental Setup

Clones Frequency Distribution

<table>
<thead>
<tr>
<th>Clone</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone1</td>
<td>50.00%</td>
</tr>
<tr>
<td>Clone2</td>
<td>25%</td>
</tr>
<tr>
<td>Clone3</td>
<td>12.5%</td>
</tr>
<tr>
<td>Clone4</td>
<td>6.25%</td>
</tr>
<tr>
<td>Clone5</td>
<td>3.125%</td>
</tr>
<tr>
<td>Clone6</td>
<td>1.56%</td>
</tr>
<tr>
<td>Clone7</td>
<td>0.78%</td>
</tr>
<tr>
<td>Clone8</td>
<td>0.39%</td>
</tr>
<tr>
<td>Clone9</td>
<td>0.19%</td>
</tr>
<tr>
<td>Clone10</td>
<td>0.097%</td>
</tr>
</tbody>
</table>
Edit Distance Heatmap
Clone6

Frequency vs. Number of reads

- 2SNV
- PH
- True
Clone7

Frequency vs. Number of reads

- 2SNV
- PH
- True
False Positive

Frequency vs. Number of reads

- 2SNV
- PH
- True
Inferring Viral Transmissions from Intra-Host Viral Populations
NGS of HOC outbreak

- 18 patients, 154233 reads and 33767 unique sequences.
- Each node is a unique sequence.
- Different patients are shown in different colors.
- Two sequences are linked if they differ in a single nucleotides.
NGS of HOC outbreak

- 18 patients, 154,233 reads and 33,767 unique sequences.
- Each node is a unique sequence.
- Different patients are shown in different colors.
- Two sequences are linked if they differ in a single nucleotide.

Sequences of the source patient are shown in green.
The main challenge:

- Finding consensus sequence is not enough
- It is crucial to get the **whole viral quasispecies spectrum** (all sequences and their relative frequencies), since minor variants can be responsible for viral transmission
Advanced Molecular Detection of viral transmissions and outbreaks

• Phylogenetic analysis
• Threshold-based methods
• Random processes
• Nonparametric methods
Threshold-based methods
Outbreak detection and display

• Step 1: calculate distances among patients
• We can measure distances among patients in different ways
  • Distance between representatives (consensus or most frequent)
  • Average distance
  • Minimal distance
• Step 2: Link populations with distances smaller than a cutoff
Distance between consensuses


Cutoff: 1.5% (approximate level of intrahost diversity early in infection known from a literature)

**Pros:**
- Easy to automate
- Simple and computationally efficient (linear)

**Cons:**
- Does not take into account structure of quasispecies population
- Does not allow detection of directions of transmissions
- May not detect transmissions of minor viral subpopulations
Nonparametric detection of transmissions
NHANES III Participants

Clinic-acquired HCV Infection and Potential Source

The July cluster

NVC46
NVC30

The September cluster

NVC31
NVC41
NVC44
NVC42
NVC29
NVC45

5.0% Nucleotide Variation
NHANES III Participants

Clinic-acquired HCV Infection and Potential Source

The July cluster

NVC46
NVC30

The September cluster

NVC41
NVC31
NVC01
NVC42
NVC45

5.0% Nucleotide Variation
NHANES III Participants
Clinic-acquired HCV Infection and Potential Source

The July cluster

NVC45
NVC46
NVC30
NVC29
NVC42
NVC41
NVC31
NVC01

Nucleotide Variation
5.0%

Strain intersection
Strain intersections

Given: two viral populations $P_1$ and $P_2$

1) partition the union $P_1 \cup P_2$ into clusters $C_1, \ldots, C_k$

2) $P_1 \cap P_2 = \bigcup_{i \in B} C_i$, where $B = \{i \in \{1, \ldots, k\}: C_i \cap P_1 \neq \emptyset, C_i \cap P_2 \neq \emptyset\}$

$P_1 \cap P_2$ is the union of clusters that contain sequences from both $P_1$ and $P_2$
Relatedness depth

\[ d(P_1, P_2) = \begin{cases} 
0, & \text{if } I = P_1 \cap P_2 = \emptyset \\
+\infty, & \text{if } P_1 \cap P_2 = P_1 \cup P_2 \\
1 + d(P_1|_I, P_2|_I), & \text{otherwise}
\end{cases} \]

**Input** Two sets of viral sequences \( P_1, P_2 \).

**Output** Separation coefficient \( d(P_1, P_2) \)

1: \( d \leftarrow 0 \)
2: \( k \leftarrow 2 \)
3: \( I \leftarrow P_1 \cap P_2 \)
4: while \( I \neq \emptyset \) and \( k \leq |P_1| + |P_2| \) do
5: \( d \leftarrow d + 1 \)
6: if \( I \neq P_1 \cup P_2 \) then
7: \( P_1 \leftarrow P_1|_I, P_2 \leftarrow P_2|_I \)
8: \( k \leftarrow 2 \)
9: else
10: \( k \leftarrow k + 2 \)
11: end if
12: \( I \leftarrow P_1 \cap P_2 \)
13: end while
Relatedness depth

Two populations $P_1$ and $P_2$ are genetically related, if $d(P_1, P_2) > 0$

Direction of transmission: if at some iteration of Algorithm 1
$P_1 \setminus I \neq \emptyset$, $P_2 \subseteq I$

Clustering: hierarchical clustering
Transmission clusters: weakly connected components
Sources: vertices with highest eigenvector centrality
Simulation/Random Walk Method
Motivation:
• To take full advantage of the knowledge of populations structures
• To estimate time of transmission

Method: simulate viral evolution
0 generations
1 to 25 generations
26 to 699 generations
700 to 2500 generations
Sequence Space Traversing (SST) Algorithm

**Input:** two viral populations $P_1$ and $P_2$, $P_1 = \{v_1,...,v_n\}$, $P_2 = \{v_{n+1},...,v_{n+m}\}$

**Output:** time distance $t(P_1, P_2)$

1. For a sequence set $P_1 \cup P_2$ construct median-joining network $G= (V,E)$

2. Simulate viral evolution using ODE model:
   \[
   \frac{dx_i}{dt} = \left(1 - \frac{\sum_{j=1}^{\|V\|} x_j}{M}\right) \left(rx_i + q \sum_{ji \in E} x_j\right), \quad i = 1, ..., \|V\|
   \]
   \[
   x_i(0) = \begin{cases} x_0, & i = 1, ..., n \\ 0, & i = n + 1, ..., \|V\| \end{cases}
   \]
   where $r = (1 - \varepsilon)^L$, $q = (\varepsilon/3)(1 - \varepsilon)^{L-1}$

3. $t(P_1, P_2) = \min\{t : x_i(t) \geq x_0 \text{ for } i = n + 1, ..., n + m\}$
Two populations $P_1$ and $P_2$ are genetically related, if $t(P_1, P_2) \leq T^*$

**Direction of transmission:** if $t(P_1, P_2) \leq t(P_2, P_1)$

**Transmission clusters:** weakly connected components

**Sources:** vertices with highest eigenvector centrality
Experimental Results: Data

- **Epidemiologically related samples.** 142 HCV HVR1 samples from 33 epidemiologically curated outbreaks reported to CDC in 2008-2013. Sources are known for 10 outbreaks as a result of epidemiological investigations.

- **Unrelated samples.** 193 HCV HVR1 samples from infected individuals without any known epidemiological relationship.
Algorithms for comparison

• Relatedness Depth (ReD)

• Sequence Space Traversing (SST)

• Consensus with 4.5% cutoff

• Consensus with 6.5% cutoff
Source estimation accuracy (related samples)

ReD

SST

Consensus (4.5%)
<table>
<thead>
<tr>
<th>Methods</th>
<th># predicted clusters</th>
<th>Related samples</th>
<th>Unrelated samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TPR</td>
<td>FPR</td>
</tr>
<tr>
<td>ReD</td>
<td>37</td>
<td>98.96%</td>
<td>0%</td>
</tr>
<tr>
<td>SST</td>
<td>37</td>
<td>96.03%</td>
<td>0%</td>
</tr>
<tr>
<td>CBC[4.5%]</td>
<td>43</td>
<td>81.84%</td>
<td>0%</td>
</tr>
<tr>
<td>CBC[6.5%]</td>
<td>38</td>
<td>96.66%</td>
<td>0%</td>
</tr>
</tbody>
</table>

True Positive Rate (TPR) = % of truly related pairs predicted as related
False Positive Rate (FPR) = % of truly unrelated pairs predicted as related
CONCLUSIONS

• Molecular analysis is one of the major tools used for investigations of viral outbreaks and inference of transmission networks. It generates novel bioinformatics problems and challenges.

• Replacement of simple consensus-based approaches and expert phylogenetic analyses with novel automatic algorithms is a major advancement in molecular surveillance of viral infections.

• Superior performance of the new algorithms over the traditional consensus-based methods indicates importance of full-edged quasispecies analyses for viral molecular surveillance and outbreaks investigation.