Accurate reconstruction of transmission history using analysis of deep sequencing data for intra-host viral populations

Pavel Skums\textsuperscript{1}, Olga Glebova\textsuperscript{2}, June Zhang\textsuperscript{1}, Zoya Dimitrova\textsuperscript{1}, David Campo\textsuperscript{1}, Leonid Bunimovich\textsuperscript{3}, Alex Zelikovsky\textsuperscript{2}, and Yury Khudyakov\textsuperscript{1}

\textsuperscript{1} Centers for Disease Control and Prevention, Division of Viral Hepatitis, Atlanta GA 30333, USA, kki8@cdc.gov.
\textsuperscript{2} Georgia State University, Department of Computer Science, Atlanta GA 30303, USA
\textsuperscript{3} Georgia Institute of Technology, School of Mathematics, Atlanta GA 30332, USA

Abstract. Currently, molecular analysis has become one of the major tools used for viral outbreak investigation and transmission network inference. We present novel algorithms for accurate identification of transmission clusters, detection of sources of infection and inference of transmission history for highly heterogeneous viruses such as HIV and HCV. Our framework VITRAQ (Viral Transmission inference from the analysis of Quasispecies) for the first time incorporates into analysis the structure of intra-host viral populations, which allows not only for the identification of genetic relatedness among viral samples but also for the accurate inference of transmission history from molecular data alone. Evaluation conducted using experimental data obtained from HCV outbreaks shows that the proposed algorithms outperform the state-of-the-art consensus-based methods both in true and false positive rates for detection of transmission clusters as well as in accuracy of source identification, and allow for the accurate inference of transmission history ("who infected whom").

Keywords: RNA virus, transmission network, viral evolution

1 Introduction

Replication of RNA viruses is error prone – their genomes mutate at extremely high rates [7]. Since mutations are generally well tolerated, many RNA viruses such as Human Immunodeficiency Virus (HIV) and Hepatitis C virus (HCV) exist in infected hosts as populations of closely related variants, known to virologists as \emph{quasispecies} [5, 6].

The study of viral quasispecies has been revolutionized by the advent of sequencing technologies that allow for sampling viral quasispecies at great depth [8]. DNA sequencing has already been used for inference of transmission networks and for outbreak investigations for Influenza A [11], HIV [15], Hepatitis A virus...
【4】，Hepatitis B virus【13】 and HCV【9】. However, contribution of sequencing technologies to molecular surveillance of viral infections has mainly been hindered by the lack of reliable computational methods for inferring transmission networks directly from sequence data.

In this paper, we address the problem of designing an accurate and scalable algorithms for the prediction of viral transmissions and outbreaks, which allows for fast, reliable and automatic identification and analysis of transmission networks in public health laboratories during outbreak investigations. The state-of-the-art software tools for automated transmission network analysis are mainly based on using consensus sequences – only one sequence per patient represents the whole intra-host viral population【15】. Primarily, the consensus-based cutoff (CBC) algorithm is used to predict direct transmissions【15】; i.e., two individuals are considered linked by transmission if the genetic distance between the corresponding consensus viral sequences does not exceed a certain cutoff value.

Although such methods are extremely useful and produce important results, they have major limitations. In particular, it is known that minority viral variants are frequently responsible for transmission of HCV infections【1】，and such transmissions may not be effectively detected using consensus sequences. Moreover, analysis of consensus sequences and genetic-distance cutoff-based methods does not allow for detecting the direction of transmissions, which is crucial for the identification of outbreak sources and superspreaders.

We present novel methods for identification of transmission networks, transmission clusters and sources of outbreaks, which resolve the aforementioned limitations. The proposed algorithms allow for prediction of viral transmission and its direction, identification of transmission clusters and sources of outbreaks, and inference of transmission history.

Evaluation of the proposed algorithms using experimental data from HCV outbreaks showed their superior performance over the consensus-based approach in detecting transmission clusters and outbreak sources.

2 Methods

2.1 Inferring transmission and its direction

Due to continuous viral evolution, the intra-host viral populations at the moment of transmission may differ significantly from the sampled viral populations, with the intermediate parts of the sequence space not being sampled at all. To approximate unsampled parts of sequence space between two viral populations, we use Median Joining network【2】 implemented in SplitsTree【10】. Let \( P_1, P_2 \in \mathcal{P} \) be two viral populations, and \( G_{mnj} = (V_{mnj}, E_{mnj}, l_{mnj}) \) be a median-joining network with edge lengths \( p_{mnj} \). We consider only single-point mutations; therefore every edge \( e \) of length \( l_{mnj}(e) > 1 \) is subdivided by \( l_{mnj}(e) - 1 \) vertices. Without loss of generality, we assume that \( V_{mnj} = \{v_1, ..., v_n\}, P_1 = \{v_1, ..., v_n\}, P_2 = \{v_{n+1}, ..., v_{n+n_2}\} \). Let \( L \) be the length of the genomic region under consideration and \( \epsilon \) be the mutation rate. Viral evolution is modeled using the following quasispecies logistic growth system of equations:
\[
\frac{dx_i}{dt} = (1 - \sum_{j=1}^{n} x_j/M) (rx_i + q \sum_{ji \in E_{m,n}} x_j), \quad i = 1, \ldots, n
\] (1)

with initial conditions \(x_i(0) = x_0\) for \(i = 1, \ldots, n_1\) and \(x_i(0) = 0\) for \(i = n_1 + 1, \ldots, n\). Here \(r = (1 - \epsilon)L\) is the probability of mutation-free viral replication, \(q = (\epsilon/3)(1 - \epsilon)L^{-1}\) is the probability of a single mutation between two adjacent vertices in \(G_{m,n}\) and \(M\) is the maximal viral population size.

The time-distance between populations \(P_1\) and \(P_2\) is defined as follows:

\[
T(P_1, P_2) = \min \{ t : x_{n_1+i}(t) \geq x_0, i = 1, \ldots, n_2 \} \quad (2)
\]

Two viral populations are genetically related, if \(T(P_1, P_2) \leq T^*\). Further, if \(T(P_1, P_2) \leq T(P_2, P_1)\), then the most probable direction of transmission is from \(P_1\) to \(P_2\). Thus, pairs of related samples form a genetic relatedness digraph \(G_r\).

2.2 Bayesian reconstruction of transmission history using Markov Chain Monte Carlo sampling

We consider a tree of samples, where leafs represent sampled individuals and interior nodes represent transmission events. The objective is to find a transmission tree \(T\) that maximizes the probability \(p(T|G_r)\) of observing the tree \(T\), given the genetic relatedness digraph \(G_r\) estimated at the previous step. We estimated this probability in a Bayesian fashion as follows:

\[
p(T|G_r) = \frac{p(G_r|T)p(T)}{p(G_r)}, \quad (3)
\]

where \(p(G_r|T)\) is likelihood of the genetic relatedness digraph \(G_r\) given the tree, and \(p(T)\) is likelihood of the transmission tree \(T\), assuming that all transmission trees follow a prior distribution.

We estimate \(p(G_r|T)\) by fitting the edge lengths of \(G_r\) into a tree \(T\) under the assumption that all intra-host populations are sampled at the same time. This problem can be formulated as the constrained linear least-square problem.

To infer the probability of a given tree \(p(T)\), we first calculate labels of internal nodes of \(T\) using the following rule: if \(v\) is an internal node of \(T\) and \(i, j\) are its children with the known labels \(l_i\) and \(l_j\) such that \(ij \in A(G_r)\), then \(v\) receives the label \(l_v = l_i\). Using this rule, labels of all internal nodes of \(T\) are calculated recursively starting from leafs, which labels are known.

Further, it is known that RNA viruses transmission networks are social networks. Consequently there are many statistical similarities between transmission networks and social networks such as scale-free degree distribution, small diameter and presence of hubs (superspreaders) [15]. Therefore we assume, that transmission trees are distributed in such a way that trees, which give rise to scale-free transmission networks have higher probabilities to be observed. To measure scalefreeness of the transmission network \(G_r\) we use a metric proposed in [12]:
where $d_i$ is a (undirected) degree of a vertex $i$ in $G_t$ and $C_2$ is a normalization constant. The value $s(G_t)$ is used as a likelihood estimation for $T$.

Using (3), we implemented Markov Chain Monte Carlo sampler from the transmission tree distribution with tree neighborhoods obtained using nearest neighbor interchange operation and an acceptance ratio $\alpha = \min\{1, \frac{p(T'|G_{r})}{p(T|G_r)}\}$.

3 Results

For algorithms testing and comparison, we used a collection of HCV data containing 142 HCV samples from 33 epidemiologically curated outbreaks reported to Centers for Disease Control and Prevention in 2008-2013 and 193 HCV samples from infected individuals without any known epidemiological relationship [3]. We compared the quality of transmission clusters identification for the proposed algorithm and the consensus-based cutoff (CBC) algorithm with cutoffs 4.5% and 6.5%. Performance of algorithms was evaluated using the true positive clustering rate (TPR) and the false positive clustering rate (FPR) [14]. As Table 1 indicates, VITRAQ clearly outperforms standard consensus-based methods.

### Table 1. Combined results for related samples (33 clusters) and unrelated samples (193 samples)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Related samples</th>
<th>Unrelated samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># predicted</td>
<td>TPR</td>
</tr>
<tr>
<td>VITRAQ</td>
<td>37</td>
<td>96.03%</td>
</tr>
<tr>
<td>CBC[4.5%]</td>
<td>43</td>
<td>81.84%</td>
</tr>
<tr>
<td>CBC[6.5%]</td>
<td>38</td>
<td>96.66%</td>
</tr>
</tbody>
</table>

The quality of VITRAQ transmission history inference was evaluated using the known history for 9 outbreaks revealed by epidemiological and forensic investigations. VITRAQ correctly identified sources of all 9 outbreaks and was able to correctly infer 79.6% of transmission events.

References


