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

Pavel Skums[1], Olga Glebova[2], June Zhang[1], Zoya Dimitrova[1], David Campo[1], Leonid Bunimovich[3], Alex Zelikovsky[2], and Yury Khudyakov[1]

[1] Centers for Disease Control and Prevention, Division of Viral Hepatitis, Atlanta GA 30333, USA,
kki8@cdc.gov,
[2] Georgia State University, Department of Computer Science, Atlanta GA 30303, USA
[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 *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 consensuses 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 $G_{mjn} = (V_{mjn}, E_{mjn}, l_{mjn})$ be a median-joining network with edge lengths p_{mjn}. We consider only single-point mutations; therefore every edge e of length $l_{mjn}(e) > 1$ is subdivided by $l_{mjn}(e) - 1$ vertices. Without loss of generality, we assume that $V_{mjn} = \{v_1, ..., v_n\}, P_1 = \{v_1, ..., v_{n_1}\}, P_2 = \{v_{n_1+1}, ..., v_{n_1+n_2}\}$. Let L be the length of the genomic region under consideration and ϵ be the mutation rate. Viral evolution is modeled using the following quasispecies logistic growth system of equations:

$$\frac{dx_i}{dt} = (1 - \sum_{j=1}^{n} x_j/M)(rx_i + q \sum_{ji \in E_{mjn}} x_j), \;\; i = 1, ..., n \tag{1}$$

with initial conditions $x_i(0) = x_0$ for $i = 1, ..., n_1$ and $x_i(0) = 0$ for $i = n_1 + 1, ..., n$. 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_{mjn} and M is the maximal viral population size.

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

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

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 if from P_1 to P_2. Thus, pairs of related samples form a genetic relatedness digraph $\mathcal{G}_r$.

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 $\mathcal{T}$ that maximizes the probability $p(\mathcal{T}|\mathcal{G}_r)$ of observing the tree $\mathcal{T}$, given the genetic relatedness digraph $\mathcal{G}_r$ estimated at the previous step. We estimated this probability in a Bayesian fashion as follows:

$$p(\mathcal{T}|\mathcal{G}_r) = \frac{p(\mathcal{G}_r|\mathcal{T})p(\mathcal{T})}{p(\mathcal{G}_r)}, \tag{3}$$

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

We estimate $p(\mathcal{G}_r|\mathcal{T})$ by fitting the edge lengths of $\mathcal{G}_r$ into a tree $\mathcal{T}$ 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 $p(\mathcal{T})$, we first calculate labels of internal nodes of $\mathcal{T}$ using the following rule: if v is an internal node of $\mathcal{T}$ and i, j are its children with the known labels l_i and l_j such that $ij \in A(\mathcal{G}_r)$, then v receives the label $l_v = l_i$. Using this rule, labels of all internal nodes of $\mathcal{T}$ 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 $\mathcal{G}_t$ we use a metric proposed in [12]:

$$s(\mathcal{G}_t) = \frac{1}{C_2} \sum_{ij \in A(\mathcal{G}_t)} d_i d_j, \tag{4}$$

where d_i is a (undirected) degree of a vertex i in $\mathcal{G}_t$ and C_2 is a normalization constant. The value $s(\mathcal{G}_t)$ is used as a likelihood estimation for $\mathcal{T}$

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 $\alpha = \min\{1, \frac{p(\mathcal{T}'|\mathcal{G}_r)}{p(\mathcal{T}|\mathcal{G}_r)}\}$

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)

Methods	Related samples			Unrelated samples		
	# predicted clusters	TPR	FPR	# predicted clusters	TPR	FPR
VITRAQ	37	96.03%	0%	193	100%	0%
CBC[4.5%]	43	81.84%	0%	193	100%	0%
CBC[6.5%]	38	96.66%	0%	171	100%	1.37%

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

1. Andria Apostolou, Michael L Bartholomew, Rebecca Greeley, Sheila M Guilfoyle, Marcia Gordon, Carol Genese, Jeffrey P Davis, Barbara Montana, and Gwen Borlaug. Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011. *MMWR. Morbidity and mortality weekly report*, 64(7):165–170, 2015.

2. Hans-Jurgen Bandelt, Peter Forster, and Arne Röhl. Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution*, 16(1):37–48, 1999.
3. David S Campo, Guo-Liang Xia, Zoya Dimitrova, Yulin Lin, Joseph C Forbi, Lilia Ganova-Raeva, Lili Punkova, Sumathi Ramachandran, Hong Thai, Pavel Skums, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *Journal of Infectious Diseases*, page jiv542, 2015.
4. Melissa G Collier, Yury E Khudyakov, David Selvage, Meg Adams-Cameron, Erin Epson, Alicia Cronquist, Rachel H Jervis, Katherine Lamba, Akiko C Kimura, Rick Sowadsky, et al. Outbreak of hepatitis a in the usa associated with frozen pomegranate arils imported from turkey: an epidemiological case study. *The Lancet Infectious Diseases*, 14(10):976–981, 2014.
5. EJJH Domingo and JJ Holland. Rna virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51(1):151–178, 1997.
6. Esteban Domingo, Julie Sheldon, and Celia Perales. Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews*, 76(2):159–216, 2012.
7. John W Drake and John J Holland. Mutation rates among rna viruses. *Proceedings of the National Academy of Sciences*, 96(24):13910–13913, 1999.
8. Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W Shafer, and Niko Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS computational biology*, 4(5):e1000074, 2008.
9. Mark Holodniy, Gina Oda, Patricia L Schirmer, Cynthia A Lucero, Yury E Khudyakov, Guoliang Xia, Yulin Lin, Ronald Valdiserri, William E Duncan, Victoria J Davey, et al. Results from a large-scale epidemiologic look-back investigation of improperly reprocessed endoscopy equipment. *Infection Control*, 33(07):649–656, 2012.
10. Daniel H Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267, 2006.
11. Makoto Kuroda, Harutaka Katano, Noriko Nakajima, Minoru Tobiume, Akira Ainai, Tsuyoshi Sekizuka, Hideki Hasegawa, Masato Tashiro, Yuko Sasaki, Yoshichika Arakawa, et al. Characterization of quasispecies of pandemic 2009 influenza a virus (a/h1n1/2009) by de novo sequencing using a next-generation dna sequencer. *PLoS One*, 5(4):e10256, 2010.
12. Lun Li, David Alderson, John C Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.
13. Sumathi Ramachandran, Michael A Purdy, Guo-liang Xia, David S Campo, Zoya E Dimitrova, Eyasu H Teshale, Chong Gee Teo, and Yury E Khudyakov. Recent population expansions of hepatitis b virus in the united states. *Journal of virology*, 88(24):13971–13980, 2014.
14. David L Wallace. Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.
15. Joel O Wertheim, Andrew J Leigh Brown, N Lance Hepler, Sanjay R Mehta, Douglas D Richman, Davey M Smith, and Sergei L Kosakovsky Pond. The global transmission network of hiv-1. *Journal of Infectious Diseases*, 209(2):304–313, 2014.