SOPHIE: viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework *

Pavel Skums[1,4], Fatemeh Mohebbi[1], Vyacheslav Tsyvina[1], Pelin Icer[2], Sumathi Ramachandran[3], and Yury Khudyakov[3]

[1] Georgia State University, Atlanta, GA, USA
[2] ETH Zurich, Basel, Switzerland
[3] Centers for Disease Control and Prevention, Atlanta, GA, USA
[4] Corresponding author. Email: pskums@gsu.edu

Abstract. Reconstruction of transmission networks from viral genomes sampled from infected individuals is a major computational problem of genomic epidemiology. For this problem, we propose a maximum likelihood framework SOPHIE (SOcial and PHilogenetic Investigation of Epidemics) based on the integration of phylogenetic and random graph models. SOPHIE is scalable, accounts for intra-host diversity and accurately infers transmissions without case-specific epidemiological data.

Keywords: Genomic Epidemiology · Transmission Network · Maximum Likelihood Inference.

1 Introduction

Advances of sequencing technologies have a profound effect on epidemiology and virology. In particular, genomic epidemiology is becoming a major methodology for investigation of outbreaks and surveillance of transmission dynamics [1].

The hallmark of viruses as species is an extremely high genomic diversity originating from their error-prone replication. First generation of genomic epidemiology methods largely ignored intra-host viral diversity, but later studies demonstrated that taking it into account greatly enhances the predictive power of transmission inference algorithms [3, 2]. Despite the significant progress achieved with the appearance of the next generation of transmission inference method, a number of computational, modelling and algorithmic challenges still need to be addressed. This includes development of scalable methodology based on maximum likelihood or Bayesian rather than maximum parsimony approach; problems with utilization of case-specific epidemiological information; accounting for non-independence of transmission events.

* PS was supported by the NIH grant 1R01EB025022 and by the NSF grant 2047828.

2 Methods

We propose to address aforementioned challenges by integrating two components: the evolutionary relationships between viral genomes represented by their phylogenies and the expected structural properties of inter-host social networks. Frequently cited properties of social contact networks include power law degree distribution, small diameter, modularity and presence of hubs. All of them are reflected by network vertex degrees. Thus, we model social networks as random graphs with given expected degree distributions (EDDs). The goal is to find transmission networks that are consistent with observed genomic data and have the highest probability to be subnetworks of random contact networks.

This methodology is implemented within a maximum likelihood framework SOPHIE (SOcial and PHilogenetic Investigation of Epidemics). SOPHIE samples from the joint distribution of phylogeny ancestral traits defining transmission networks, estimates the probabilities that sampled networks are subgraphs of a random contact network and summarize them accordingly into the consensus network. This approach is scalable, accounts for intra-host diversity and accurately infers transmissions without case-specific epidemiological data.

3 Results

We applied SOPHIE to synthetic data simulated under different epidemiological and evolutionary scenarios, as well as to experimental data from epidemiologically curated HCV outbreaks. The experiments confirm the effectiveness of the new methodology. We demonstrated that the proposed approach is capable of achieving a substantial accuracy improvement over state-of-the-art parsymony-based phylogenetic methods, while retaining their scalability and speed.

4 Disclaimer

The conclusions in this report do not necessarily reflect the official position of the Centers for Disease Control and Prevention. Experimental data were used as approved by the Institutional Review Board of the CDC (protocol 7270.0).

References

1. Armstrong, G.L., MacCannell, D.R., Taylor, J., Carleton, H.A., Neuhaus, E.B., Bradbury, R.S., Posey, J.E., Gwinn, M.: Pathogen genomics in public health. New England Journal of Medicine **381**(26), 2569–2580 (2019)
2. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., et al.: Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics **34**(1), 163–170 (2017)
3. Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M., Fraser, C., STOP-HCV Consortium, T.M.P.C., Collaboration, T.B.: Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. Molecular biology and evolution **35**(3), 719–733 (2017)