Using genetic sequences to infer population dynamics: Phylodynamic analysis of HIV transmission in SE Michigan

Edward Ionides

University of Michigan, Ann Arbor

ionides@umich.edu

ASC IMS 2014 session on "Biostatistics"
Thu 10th July, 2014

Overview

1. Phylodynamic modeling and inference.
2. An application to HIV transmission.
3. Relationships to spatio-temporal modeling and inference.

Phylogenetics, phylodynamics and epidemiology

- **Phylogenetics** is the use of genetic sequence data to infer evolutionary relationships represented by a phylogenetic tree.
- **Phylodynamics** uses phylogenetic methods to investigate population dynamics (rates of birth, death, migration, etc).
- **Infectious disease epidemiology** can be informed by phylodynamic study of a pathogen. Migration is disease transmission.

Relationship to spatio-temporal dynamics

- **Phylogeography** is the use of phylogenetic methods to investigate geographical dispersion of populations; closely related to phylodynamics.
- **Large dynamic systems**. In full generality, the space-time inference problem is equivalent to the complex system inference problem. For example, both can be framed as partially observed Markov process inference problems on a large space.
- **Weak coupling**. Tractability of space-time inference problems typically involves spatial interactions that become weak over large distances. Different branches of an evolving phylogeny also have weak interactions, typically only through competition for common resources.

Why phylodynamics?

- Fundamental infectious disease epidemiology concepts, like herd immunity and the dependence of prevalence on the reproductive ratio, R_0, exist only in the context of nonlinear dynamic models.
- More sophisticated questions, such as contact rates between age groups and the effectiveness of vaccines, cannot properly be answered outside the context of these dynamic models.
- The growing abundance of routinely collected genetic sequence data on pathogens should be able to inform many unresolved questions in epidemiology.
- Disclaimer: we find that the genetic data complements, but does not replace, the importance of traditional surveillance data.

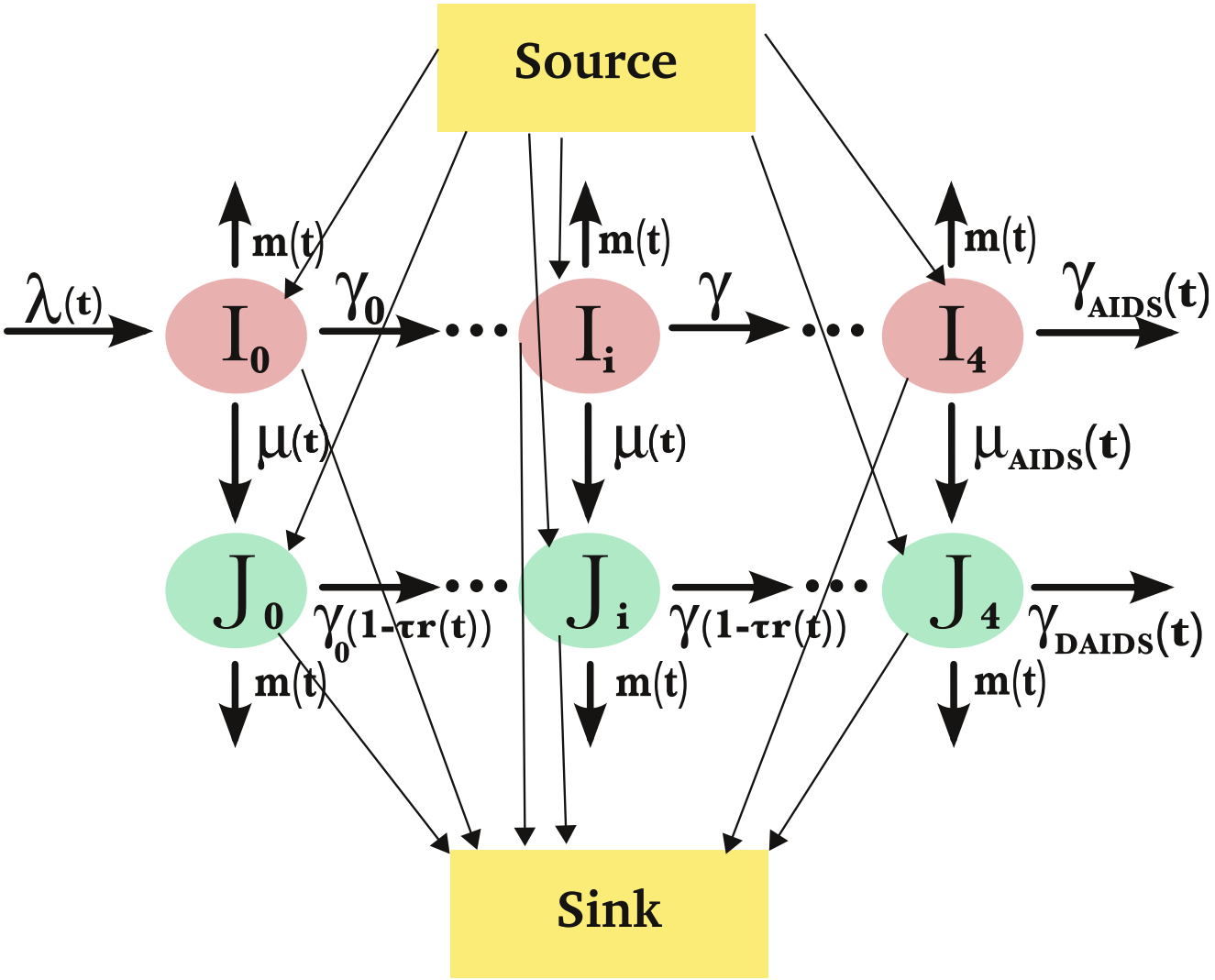
HIV transmission in early infection

- Early-stage HIV infection is characterized by high virus levels and possibly a continuation of high-risk behaviors associated with transmission.
- We studied the HIV epidemic in SE Michigan to quantify the fraction of transmissions from early HIV infection:

  > Erik M. Volz, Edward Ionides, Ethan O. Romero-Severson, Mary-Grace Brandt, Eve Mokotoff, James S. Koopman. "HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis." (*PLoS Medicine*, 2013).

- Data provided by the Michigan Dept of Community Health: 9,000 anonymized HIV sequences linked to clinical, demographic and behavioral covariates. Surveillance data for 30,000 diagnoses. Additional data on some individuals enrolled in observational studies.

An HIV compartment model flow diagram



I_k is infected & undiagnosed.

J_k is diagnosed.

k is stage of disease progression.

$k = 4$ is AIDS.

$k = 0$ is early HIV infection.

Interpretation of the flow diagram

- The rates were used to define a system of ordinary differential equations.
- The infection rate, $\lambda(t)$, and the diagnosis rate, $\mu(t)$, were modeled nonparametrically via cubic splines.
- A Poisson measurement model links the modeled number of diagnoses of HIV and AIDS to the surveillance data.
- How do we build a measurement model for the genetic sequence data?

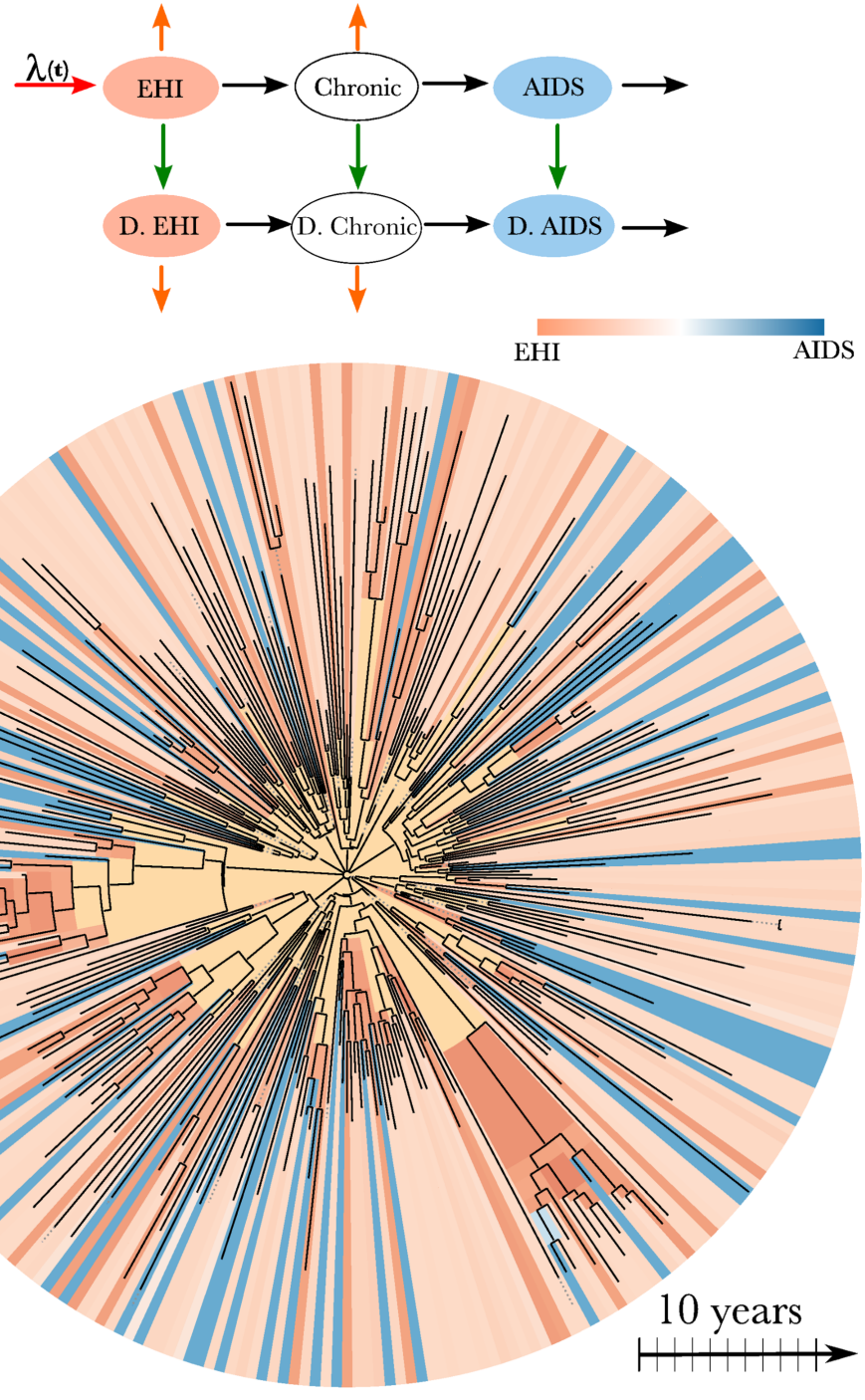
Likelihood for the coalescent in a dynamic, structured population

- Suppose the phylogeny for the sequence data is known (we estimated this via BEAST).
- "Coalescent times" are branch times in the phylogeny.
- Branches are assumed to correspond to transmission events between lineages ancestral to the observed sequences.
- A high transmission rate increases the coalescent rate.
- A small population increases the coalescent rate, since ancestral lineages are likely to coincide at population bottlenecks.
- Erik Volz (*Genetics*, 2009 & 2012) worked out the equations.

Differential system for the coalescent rate

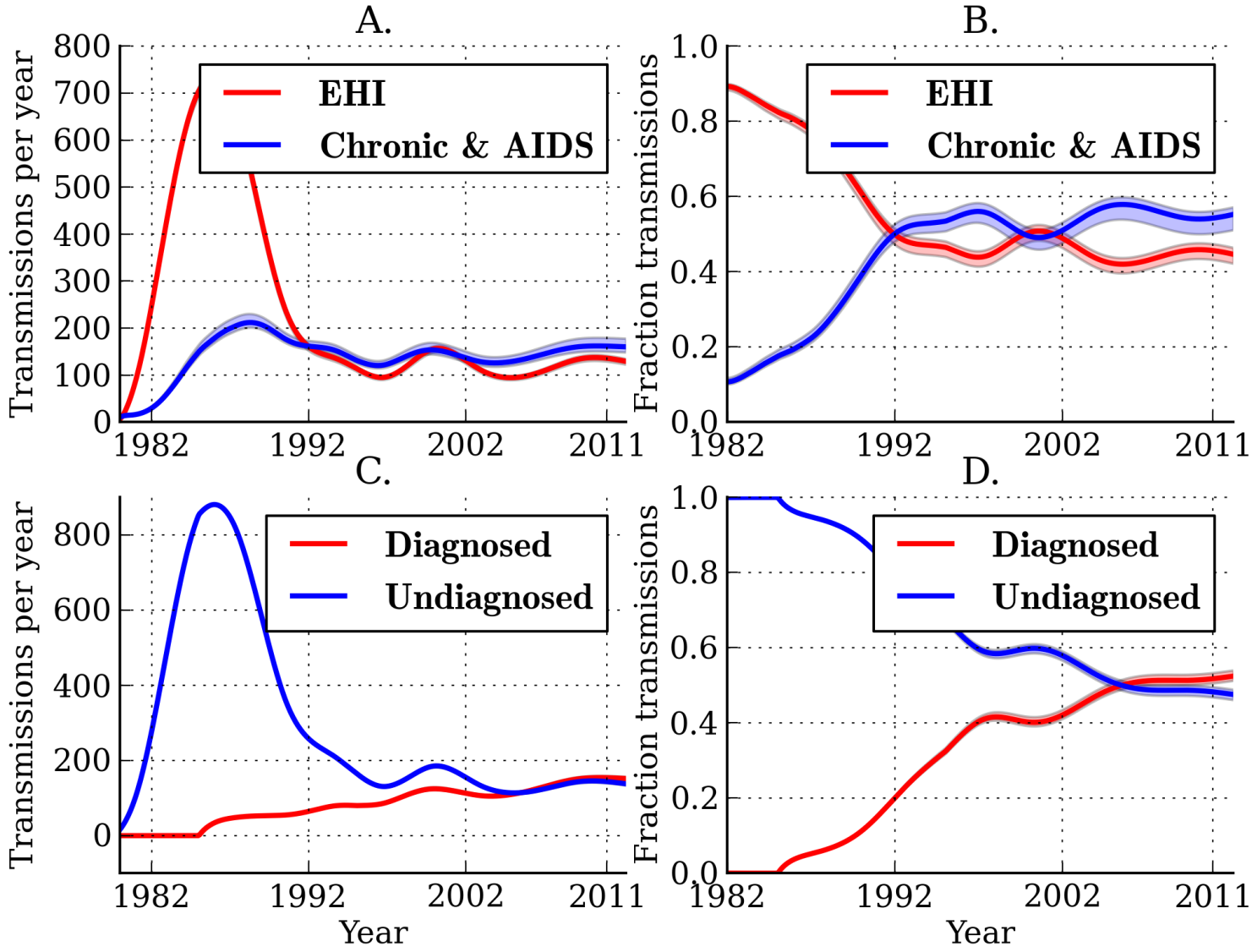
$$
\begin{aligned}
c_{ij} &= \sum_{k=1}^{m}\sum_{\ell=1}^{m} \frac{f_{k\ell}}{Y_k Y_\ell}(p_{ik}p_{j\ell} + p_{i\ell}p_{jk}) \\
\frac{d}{dt}p_{ik} &= \sum_{\ell=1}^{m}\left(\frac{p_{i\ell}}{Y_\ell}g_{k\ell} - \frac{p_{ik}}{Y_k}g_{\ell k} + \frac{p_{i\ell}}{Y_\ell}\frac{Y_k - A_k}{Y_k}f_{k\ell} - \frac{p_{ik}}{Y_k}\frac{Y_\ell - A_\ell}{Y_\ell}f_{\ell k}\right)
\end{aligned}
$$

- States (i.e., compartments) are numbered $1, \ldots, m$.
- $c_{ij}(s)$ is the coalescent rate between lineages i and j at time s, measuring time backwards from the leaf.
- $f_{k\ell}$ is the rate at which individuals in k have offspring in state ℓ.
- $g_{k\ell}$ is the rate at which individuals in k migrate to state ℓ.


λ(t)
EHI
Chronic
AIDS
D. EHI
D. Chronic
D. AIDS
EHI
AIDS
10 years

Inference

- The likelihood function was maximized using an iterative Nelder-Mead search, with 6000 simultaneous optimizations initially started in a large hyper-rectangle and re-started every 200 iterations in the vicinity of the 600 highest likelihood searches.
- Global maximization was validated by Monte Carlo replication.
- Profile likelihoods were computed to provide confidence intervals.
- Empirical Bayes methods were used, with a prior constructed from these confidence intervals, to investigate uncertainty in state estimates.
- The computation took around a week on a 200 core cluster.
- Around 45% $\pm$2% of HIV transmissions were estimated to originate from early infections in 2007.


A.
B.
C.
D.
Transmissions per year
Fraction transmissions
EHI
Chronic & AIDS
Diagnosed
Undiagnosed
Year
0
100
200
300
400
500
600
700
800
0.0
0.2
0.4
0.6
0.8
1.0
1982
1992
2002
2011

Conclusions

- Potential improvements on this methodology include:
  1. simultaneous estimation of the phylogeny and the dynamics.
  2. inclusion of stochasticity in the dynamics.
- Sequential Monte Carlo (SMC) approaches to this are being investigated (in collaboration with Alex Smith, Aaron King and James Koopman).
- Scaling up SMC methods for large numbers of sequences will require methods that take advantage of weak coupling. SMC methods for spatio-temporal systems and large complex systems are in their infancy (currently being investigated in collaboration with Joon Ha Park).

Thank You!

The End.