Phylodynamic Inference of MERS-CoV Using Structured Markov Genealogy Processes

Peter Yang

Thesis Advisors

Professor Aaron King
Department of Ecology and Evolutionary Biology, University of Michigan
Professor Edward Ionides
Department of Statistics, University of Michigan

Graduate Student Mentor

Jesse Wheeler
Department of Statistics, University of Michigan

Abstract

Middle East respiratory syndrome coronavirus (MERS-CoV) is a novel coronavirus originating from the Arabian Peninsula in 2012. Despite the overall dynamics and epidemiology of MERS remaining poorly understood, it is widely believed that this virus is endemic in camels, and is the main cause of transmission of limited outbreaks of MERS in humans, with human-to-human transmission appearing insufficient to sustain widespread epidemic or pandemic spread. The goal of this thesis is to use a more exact representation of the tree likelihood on MERS viral sequence data via phylogenetic trees to conduct likelihood maximization. To do this, we use the Iterated Filtering 2 (IF2) algorithm within the framework of Phylodynamic Partially Observed Markov Process (PhyloPOMP) models.

Acknowledgements

I would like to express my gratitude to Professor Aaron King for his guidance throughout this project. His teaching introduced me to the field of phylodynamics, and his advice and acumen were instrumental when I encountered challenges during the research process. I am also grateful to Professor Edward Ionides for providing valuable input on the statistical methodology used in this work, and his thoughtful comments on my thesis drafts. I thank Jesse Wheeler for his early mentorship in learning time series analysis and the POMP framework, and his support with cluster computing and proofreading. Finally, I would like to thank my mother for her unwavering support in my studies and ambitions.

Contents

1 Introduction

Middle East Respiratory Syndrome (MERS) is a viral respiratory illness caused by the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) first identified in Saudi Arabia in September 2012. Since its emergence, MERS has posed a significant threat to public health and safety due to its relatively high mortality rate and potential for widespread human-to-human transmission. Most MERS-CoV cases are reported in the Arabian Peninsula, with 84.35% of the cases being reported from Saudi Arabia. From April 2012 to May 2024 there have been a reported 2,613 confirmed cases of MERS globally, spanning across 27 countries. As of May 2024, the World Health Organization (WHO) estimates that the case-fatality ratio (CFR) of MERS-CoV is 36% (Disease Control and Prevention 2025). Although in recent years, the number of cases has decreased, sporadic outbreaks, specifically in the Middle East, still occur with high mortality rates, highlighting the severity of the disease and the need for continued research, surveillance, and containment of the disease.

It is widely believed that MERS is primarily a zoonotic disease, with dromedary camels acting as a reservoir host and the main source of human infections (Gossner et al. 2014) (Reusken et al. 2016). Similarly, sustained human-to-human transmission is generally considered to be rare, and limited to specific settings (WHO 2025). However, despite this prevailing belief, the true extent of human-to-human transmission, and the underlying dynamics of MERS as a whole, remains poorly understood, due to the relatively recent emergence of the virus and limited case numbers. The transmission dynamics of MERS therefore remains a key area of focus for epidemiological research.

In recent years, advances in genome sequencing have enabled the widespread sampling of MERS-CoV genomic data in both camels and humans. As of August 2022, a total of 728 MERS-CoV genomes have been identified (Azhar et al. 2023). This genomic data provides an opportunity to study the evolutionary and transmission dynamics of the MERS Coronavirus through phylogenetic analysis. Coalescent or birth-death process models enable the reconstruction of a phylogenetic tree from sampled genomic data, facilitating the ability to perform statistical inference on key epidemiological parameters, such as the viral reproduction number (R_0). However, while these phylogenetic methods are useful to observe the evolution of genomic sequences and reconstruct genealogical trees by considering the relationships between viral sequences, these models often treat transmission dynamics and the evolution of the phylogenetic tree as separate from the underlying epidemiological process. However, recent methodological developments have introduced *phylodynamic* models, which

integrate transmission dynamics directly with the evolutionary history of the genome. This allows researchers to jointly infer the spread and evolution of a virus, leading to more robust and accurate understandings of transmission dynamics.

POMP (partially observed Markov process) models offer a flexible framework for the modeling and inference of stochastic compartmental models using time series data. PhyloPOMP builds on the POMP framework to incorporate time series phylogenetic data by taking in a phylogenetic tree as observed data. This incorporation allows for the use of the evolutionary history of a genome to be used to find parameters related to determinants of epidemic spread via an underlying compartmental model.

Recent advances in phylodynamics enable exact likelihood calculation for a phylogenetic tree via structured Markov genealogy processes for some underlying transmission model (King, Lin, and Ionides 2024). In this work, we leverage these advances to maximize the likelihood for a reconstructed human-camel phylogenetic tree of MERS (Dudas et al. 2018) using the Iterated Filtering (IF2) algorithm (Ionides et al. 2015). As this approach to fitting a PhyloPOMP model to real-world data has never been done, a primary goal of this work is to determine the practical utility of PhyloPOMP in real epidemic settings, specifically by assessing how effectively likelihood maximization can be performed via the IF2 algorithm on genealogies, and to potentially validate or challenge existing beliefs about MERS-CoV.

2 Methodology

2.1 Partially Observed Markov Processes

A Partially Observed Markov Processes (POMP) model, which is also known as a Hidden Markov Model (HMM) consists of a noisy observed process, and a latent (unobserved) state process. Suppose that we have data of the observed process $Y(t)$ denoted by $Y_{1:N} = y_1^*, \ldots, y_n^*$ at time points $t_1 < \cdots < t_n$, and let $X(t)$ be the unobserved Markov process of the latent state, and let the values $X_{1:N} = x_1, \ldots, x_n$ denote the values of $X(t)$ coinciding at the same time points $t_1 < \cdots < t_n$ of the observed process $Y(t)$. The model relies on defining a *process model* $f_{X_{n+1}|X_n}(x_{n+1}|x_n;\theta)$ that models the evolution of the latent process, where θ is a vector in our parameter space. A *measurement model* $f_{Y_n|X_n}(y_n|x_n;\theta)$ is then specified in order to connect the latent process model to the observable process. A figure of the described models and their interactions within the POMP framework can be seen in 1

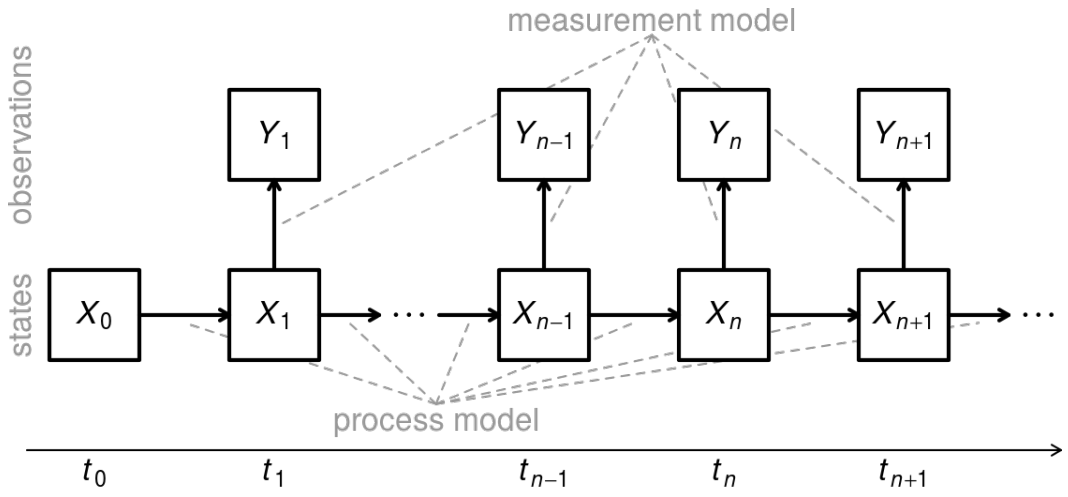

Figure 1: A diagram of a general POMP model. Figure Source: (King and Ionides 2024)

Using this setup, we can write the joint density as:

$$f_{X_{1:N},Y_{1:N}}(x_{1:N}, y_{1:N}; \theta) = f_{X_1}(x_1; \theta) \prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1}; \theta) f_{Y_n|X_n}(y_n|x_n; \theta) \tag{1}$$

We can integrate over $x_1, \ldots, x_n$, getting us the likelihood of our data given the model parameters

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y^*_{1:N}; \theta) = \int_{\mathbb{R}^{N+1}} f_{X_1}(x_1; \theta) \prod_{i=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1}; \theta) f_{Y_n|X_n}(y^*_n|x_n; \theta)\, dx_{1:N}. \tag{2}$$

The likelihood function $\mathcal{L}(\theta)$ plays a key role in statistical inference, as it provides a metric that explains how well the parameter set θ under some model explains the observed data. However, equation (2.2) is generally only tractable in simple cases. The model we consider in this thesis is, in contrast, nonlinear and non-Gaussian, making analytic computation of the likelihood infeasible, and motivating the use of Monte Carlo based inference methods.

The *particle filter/Sequential Monte Carlo* is one common way to compute an estimate of the likelihood via Monte Carlo simulation (Arulampalam et al. 2002).

2.2 IF2 Algorithm

Though the likelihoods of nonlinear POMP models can be effectively approximated by using a particle filter, using this approximation to perform likelihood maximization is imprac-

tical in most settings. In many such scenarios, Iterated filtering algorithms, which extend the particle filter, can be used for likelihood maximization. These algorithms work by iteratively perturbing the parameters of latent state X_n and recomputing likelihoods with particle filters to converge to the maximum likelihood estimate. (Ionides et al. 2015). Pseudocode for the algorithm is provided below

Model Input: Simulators for $f_{x_0}(x_0;\theta)$ and $f_{x_n|x_{n-1}}(x_n|x_{n-1};\theta)$; evaluator for $f_{y_n|x_n}(y_n|x_n;\theta)$; data, $y_{1:N}$
Algorithmic Parameters: Number of iterations, M; number of particles, J; initial parameter swarm, $\{\Theta_j^0, j = 1,\dots,J\}$; perturbation density, $h_n(\theta|\varphi;\sigma)$; perturbation scale, $\sigma_{1:M}$
Output: Final parameter swarm, $\{\Theta_j^M, j = 1,\dots,J\}$

1. For m in $1:M$
   (a) $\Theta_{0,j}^{F,m} \sim h_0(\theta|\Theta_j^{m-1},\sigma_m)$ for j in $1:J$
   (b) $X_{0,j}^{F,0} \sim f_{X_0}(x_0;\Theta_{0,j}^{F,m})$ for j in $1:J$
   (c) For n in $1:N$
      i. $\Theta_{n,j}^{P,m} \sim h_0(\theta|\Theta_{n-1,j}^{F,m},\sigma_m)$ for j in $1:J$
      ii. $X_{n,j}^{P,m} \sim f_{X_n|X_{n-1}}(x_n|X_{n-1,j}^{F,m};\Theta_{n,j}^{P,m})$ for j in $1:J$
      iii. $w_{n,j}^m = f_{Y_n|X_n}(y_n^*|X_{n,j}^{P,m};\Theta_{n,j}^{P,m})$ for j in $1:J$
      iv. Draw $k_{1:J}$ with $P[k_j = i] = w_{n,i}^m / \sum_{u=1}^J w_{n,u}^m$
      v. $\Theta_{n,j}^{F,m} = \Theta_{n,k_j}^{P,m}$ and $X_{n,j}^{F,m} = X_{n,k_j}^{P,m}$ for j in $1:J$
   (d) End For
   (e) Set $\Theta_j^m = \Theta_{N,j}^{F,m}$ for j in $1:J$
2. End For

A key feature of iterated filtering algorithms is that they retain the *plug-and-play* property of particle filters. That is, they only require the ability to simulate the latent process, and evaluate the measurement process, in order to reliably optimize model likelihoods.

2.3 Phylodynamics/PhyloPOMP

Phylodynamics is a field encompassing and integrating epidemiology and phylogenetics, which uses viral genomic data in order to infer factors influencing epidemic spread. When

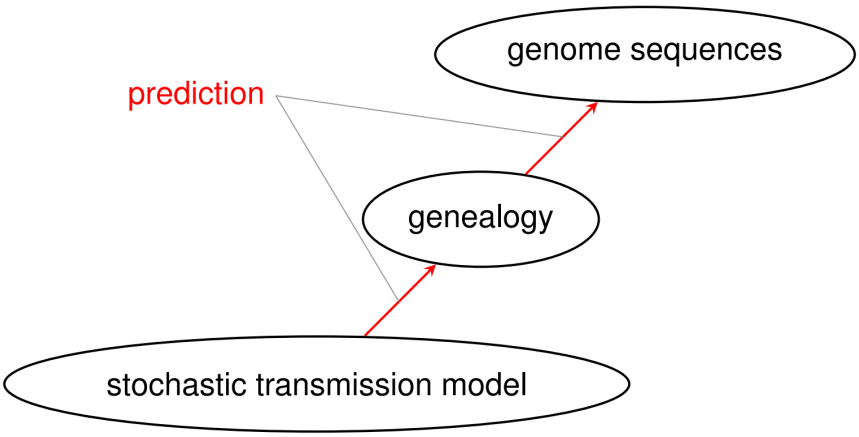

Figure 2: Schematic representation of the phylodynamic modeling process. Figure Source: (King 2024)

mutations occur at similar timescales to the transmission of a pathogen, the patterns in differences in mutated genomes provide information on the history of the pathogen's transmission through a population. This data can be examined to gain a better understanding of the dynamics of pathogen spread. Given a mathematical model of transmission, one can estimate parameters by comparing their ability to explain the underlying data, through conventional statistical methodology. This process is known as *phylodynamic inference*.

Most commonly, the data used in phylodynamic models is a reconstructed tree-like *phylogeny* or *genealogy* which captures temporal and ancestral relationships between different sampled genomes. This data is then related to an underlying mathematical transmission model to compute the likelihood of a genealogy tree given a model under a set of model parameters. Specifically, if S is a set of genome sequences, Φ a genealogical tree reconstructed from the sequences, E a model of the evolution of sequences, and D a dynamic transmission model, the likelihood is given as:

$$\mathcal{L}(D, E) = f(S|D, E) = \int f(S|\Phi, E)\ f(\Phi|D)\ d\Phi \tag{3}$$

where the integral is taken over all possible genealogies. The function $f(\Phi|D)$, is called the *phylodynamic likelihood* relates the phylogeny to the transmission model.

Existing approaches to computing the phylodynamic likelihood have been based on either the Kingman Coalescent (Kingman 1982) or the linear-birth-death process (Kendall 1948). The Kingman Coalescent looks at the tree backward in time, and looking at times where

lineages combine to calculate the phylodynamic likelihood. The model calculates the exact likelihood, assuming that the underlying model is the Moran model (Moran 1958), where the population size is constant. The Linear-birth-death process is another approach, which allows for a closed form expression of the likelihood. "Linear" here refers to the model's assumption that distinct lineages don't interact. This key assumption allows for the likelihood to be analytically tractable. Although the relative ease of computation of these two models makes them attractive, the under utilization of full information in the data and strong assumptions made by the models are areas of concern. Naturally, this has led to increased interest in improved methods in phylodynamic inference.

Work by King, Lin, and Ionides (2024) has led to the enabling of computing the phylodynamic likelihood for a broad range of dynamic models. Specifically, the authors introduce a framework using structured Markov Genealogy Processes that they use to derive an exact expression for the likelihood for a genealogy under a dynamic transmission model. This expression can then be evaluated with standard Monte Carlo methods.

The methods introduced in King, Lin, and Ionides (2024) are implemented in the PhyloPOMP R package, which builds upon the functionality of the existing POMP framework. PhyloPOMP enables likelihood evaluation and parameter inference for phylodynamic models by representing genealogical trees as POMP objects. This allows for the application of the particle filter and the iterated filtering (IF2) algorithms from the POMP package on phylogenetic data.

3 Data

The dataset consists of 274 MERS-CoV genomes, sourced from GenBank, with 174 coming from humans, and 100 from camels. Only sequences covering $\geq 50\%$ of the genome were included. The data was collected from 2012-2017, with the majority of camel sequences coming from Saudi Arabia, and human sequences originating primarily from Saudi Arabia, with additional cases from the United Arab Emirates, Qatar, Jordan, Egypt, South Korea, and the United States due to imported infections. Sequence data was labeled as human or camel using Multiple Alignment via Fast Fourier Transform (MAFFT). Each genome was annotated with host origin (human or camel), sampling date, and country of origin, with human cases including both domestic and travel-associated infections. In total, there were an estimated 56 independent camel-to-human spillover events. The phylogenetic tree was sourced from (Dudas et al. 2018); it was created by using BEAST v2.4.3 to construct a

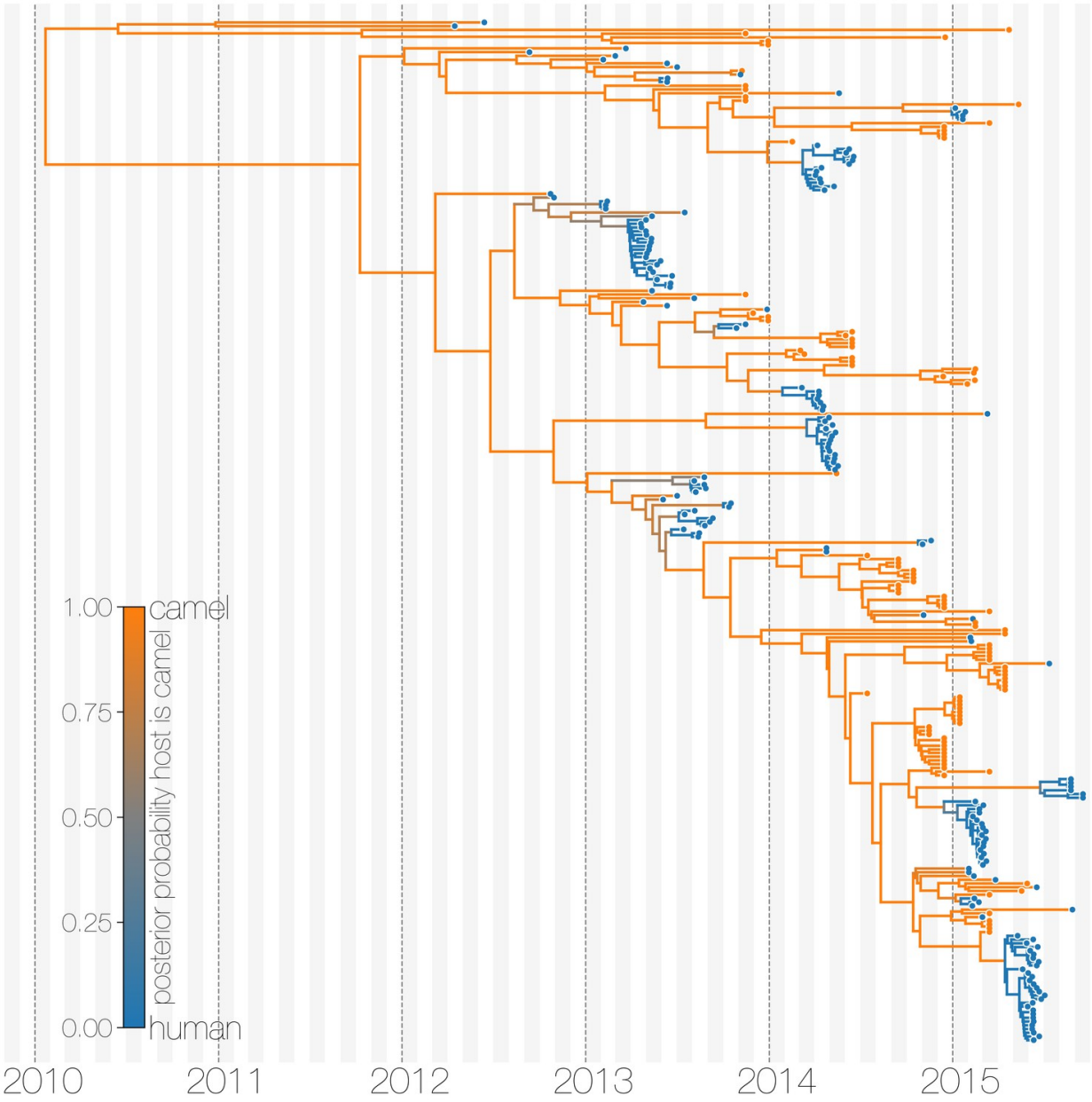

Figure 3: Typed maximum clade credibility tree of MERS-CoV genomes from (Dudas et al. 2018).

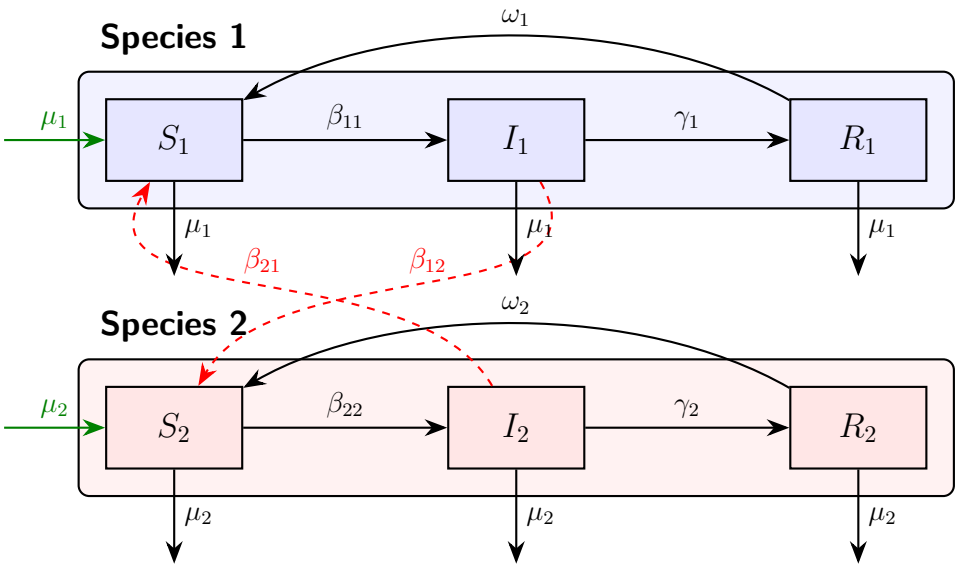

Figure 4: Compartmental 2-species SIRS model with spillover transmission. Solid arrows represent within-species transitions, dashed red arrows represent spillover infections between species. β_{ij} are transmission rates, γ_i are recovery rates, ω_i are immunity waning rates, and μ_i are birth/death rates.

maximum clade credibility (MCC) tree under a structured coalescent model. A figure of the tree can be seen in 3

4 Model

The underlying epidemiological model used was a two-species Susceptible-Infected-Recovered-Susceptible (SIRS) model. The two-species SIRS model is similar to two independent single-species SIRS models, with the additional feature that infection rates within these two populations are also influenced by infection spillover rates from one species infecting another, in addition to within population infection rates. We designate species 1 as camel, and species 2 as human. In the context of our study, we are interested in the value of β_{12}, the spillover infections from camels to humans. We label these rates as β_{12} and β_{21} to indicate transmission from species 2 to species 1 and vice-versa, respectively. A diagram of the two-species model can be seen in Figure 4, and the model rates in Table 1

Transition	**Compartment Transition**	**Rate**
Infection in species 1	$S_1 \to S_1 - 1,\ I_1 \to I_1 + 1$	$\beta_{11}\frac{I_1}{N_1}S_1$
Infection in species 2	$S_2 \to S_2 - 1,\ I_2 \to I_2 + 1$	$\beta_{22}\frac{I_2}{N_2}S_2$
Spillover from species 1 to species 2	$S_2 \to S_2 - 1,\ I_2 \to I_2 + 1$	$\beta_{21}\frac{I_1}{N_1}S_1$
Spillover from species 2 to species 1	$S_1 \to S_1 - 1,\ I_1 \to I_1 + 1$	$\beta_{12}\frac{I_2}{N_2}S_2$
Recovery in species 1	$I_1 \to I_1 - 1,\ R_1 \to R_1 + 1$	$\gamma_1 I_1$
Recovery in species 2	$I_2 \to I_2 - 1,\ R_2 \to R_2 + 1$	$\gamma_2 I_2$
Waning immunity in species 1	$R_1 \to R_1 - 1,\ S_1 \to S_1 + 1$	$\omega_1 R_1$
Waning immunity in species 2	$R_2 \to R_2 - 1,\ S_2 \to S_2 + 1$	$\omega_2 R_2$

Table 1: Stochastic Two-Species SIRS Model with Spillover

Parameter	**Description**	**Value**
c_1, c_2	Probability that a sampled host is culled	$c_1 = 1$, $c_2 = 1$
β_{11}	Transmission rate within species 1	365/10
β_{12}	Transmission rate from species 2 to species 1	0
β_{21}	Transmission rate from species 1 to species 2	10
β_{22}	Transmission rate within species 2	1–72
γ_1	Recovery rate for species 1	365/14
γ_2	Recovery rate for species 2	365/10
ψ_1, ψ_2	Per capita sampling rates for species 1 and 2	$\psi_1 = 0.1$, $\psi_2 = 0.2$
ω_1, ω_2	Rate of waning immunity for species 1 and 2	$\omega_1 = 1$, $\omega_2 = 0$
b_1, b_2	Per capita birth rates for species 1 and 2	$b_1 = 0$, $b_2 = 0$
d_1, d_2	Per capita death rates for species 1 and 2	$d_1 = 0$, $d_2 = 0$
S_1, S_2	Initial susceptible populations for species 1 and 2	$S_1 = 4950$, $S_2 = 5000$
I_1, I_2	Initial infected populations for species 1 and 2	$I_1 = 50$, $I_2 = 0$
R_1, R_2	Initial immune populations for species 1 and 2	$R_1 = 0$, $R_2 = 0$

Table 2: Initial Two-Species Model parameters. Species 1 represents camels, Species 2 represents humans.

5 Results

5.1 Likelihood Maximization with IF2

A two-species SIRS POMP model was made with the initial parameters shown in Table 2. $\beta_{11}, \beta_{21}, \beta_{22}, \psi_1, \psi_2$ were designated to be fitted, while the rest of the parameters were fixed. Our model assumes strict intervention measures upon detection of human cases, involving immediate isolation or quarantine, and similarly assumes that camels are immediately culled upon testing positive for MERS-CoV infection. As such, in our model, humans and camels are assumed to be immediately culled after a positive sampling of MERS ($c_1 = 1, c_2 = 1$).

The fixed MERS infection recovery and waning immunity parameters were obtained from previously published epidemiological estimates. The fixed value of infection recovery of 14 days for camels was taken from (Alharbi, Ibrahim, Alhafufi, et al. 2020). The fixed value of infection recovery of 10 days for humans was taken from (Zumla, Hui, and Perlman 2015). The waning immunity figure for camels was taken from (Meyer et al. 2016), and was increased to 1 year in order to account for variability across camels of different maturities. The population sizes for camels and humans were both set to 5000, with 1% of the camel population initialized as infected. The IF2 algorithm was run with the hyperparameters seen in 3. Two further refinement runs were conducted after the main run, first with a random-walk standard deviation (`rw.sd`) of 0.002, followed by a second run with a reduced `rw.sd` of 0.0005. Each refinement run was done by taking the top 6 log-likelihood trajectories from the previous IF2 run and duplicating each trajectory 6 times.

IF2 Hyperparameter	**Value**
Random Walk Standard Deviation (`rw.sd`)	0.01
Cooling Rate	0.1
Number of Iterations (`Nmif`)	50
Number of Trajectories	36
Number of Particles	20,000

Table 3: IF2 Hyperparameters used for inference.

5.2 Profile Likelihood

We re-estimate log-likelihoods while fixing β_{22} to create a profile likelihood plot for β_{22}, as seen in figure 6. The hyperparameters used to create the profile likelihood plot can be

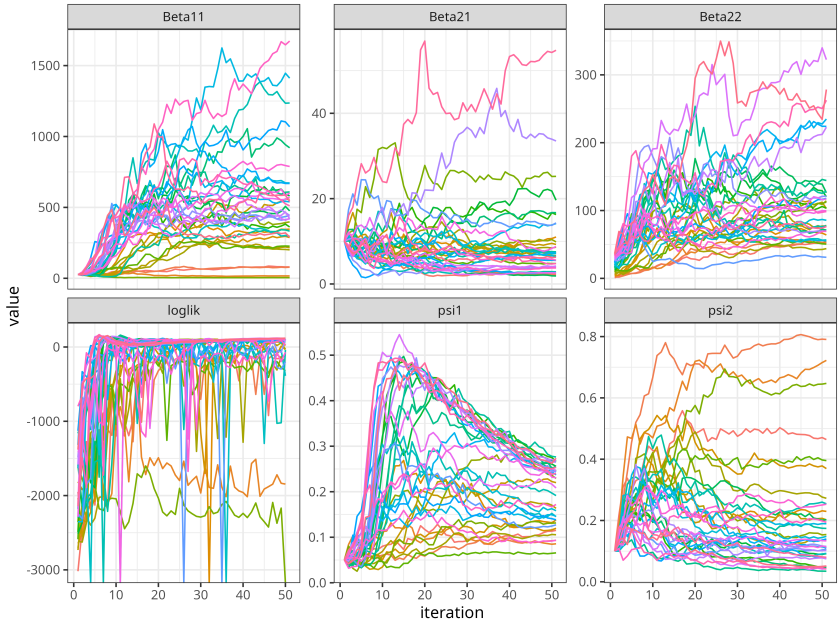

Figure 5: Trajectory plots from an IF2 (Iterated Filtering 2) run, showing parameter evolution over 50 iterations. Each line represents one of 72 particle trajectories used to estimate parameters β_{11}, β_{21}, β_{22}, ψ_1, ψ_2, and the log-likelihood.

Parameter	**Fitted Values**
β_{11}	338.20
β_{21}	16.19
β_{22}	96.45
ψ_1	0.218
ψ_2	0.031
$R_0^{(1)}$	12.97
$R_0^{(2)}$	2.64
Log-Likelihood	124.23

Table 4: Fitted values for parameters based on the IF2 trajectory with the largest Log-Likelihood

seen in 5. We can see from this plot that β_{22} values around 260-440 are the most consistent with the underlying data, attaining a maximum Log-Likelihood of 114.92.

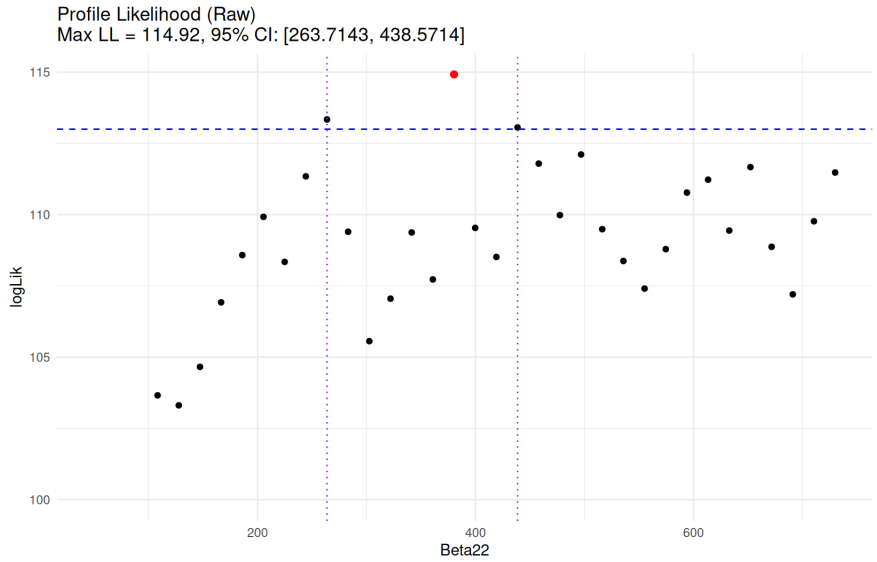

Figure 6: Profile Likelihood of β_{22}. For each value of β_{22}, the maximum logLik was taken. The blue and purple lines represent the 95% confidence interval.

Profile Likelihood Hyperparameters	**Value**
Parameter Profiled	β_{22}
Range of β_{22}	50 to 700
Number of Trajectories	180
Number of Particles	5,000
Random Walk Standard Deviation (`rw.sd`)	0.01
Cooling Rate	0.1
Number of Iterations (`Nmif`)	50

Table 5: Profile likelihood hyperparameters used to evaluate the likelihood over a range of β_{22} values.

5.3 Benchmarking

To assess the performance and plausibility of our methodology, we compared our results to the Moran model. This benchmarking process serves two purposes for this project. First, it provides a tractable baseline model with a closed form likelihood. Second, it allows for us to evaluate the impact of incorporating structured dynamics provides a better fit to the data by comparing likelihoods.

The continuous-time Moran model is a classical stochastic process model that assumes

a constant rate of births/deaths and population size. At large population sizes, the Moran model converges to the Kingman Coalescent, and can be used to compute likelihoods of phylogenetic trees. The Moran model has three main parameters: the population size n, the per capita event rate μ, and the per capita sampling rate ψ. For a given phylogenetic tree, the Moran model computes the exact likelihood under the inputted parameters. Obtaining the exact likelihood via the Moran model can be done in PhyloPOMP using the `moran_exact` function.

We optimize the Log-Likelihood of our MERS-CoV geneaology under the Moran model using the `optim` package in R. We fix $n = 10,000$ to reflect the population size of both camels and humans in our two species model, and optimize the model's log-likelihood by varying the per capita event and sampling rates. We optimize with the commonly used Limited-memory Broyden–Fletcher–Goldfarb–Shanno with Box constraints (L-BFGS-B) optimization algorithm, with the box constraints added to ensure positive parameter values. The optimized parameters and log-likelihood can be seen in 6.

Description	**Parameter**	**Estimate**
Per capita event rate	μ	1548.039
Per capita sampling rate	ψ	0.003130
Log-likelihood	$\log \mathcal{L}$	-95.200

Table 6: Optimized Parameters and Log-Likelihood under the Moran Model with $n = 10000$. μ denotes the per capita event rate and ψ the per capita sampling rate.

6 Discussion

This thesis aimed to evaluate the practicality of applying the novel PhyloPOMP framework developed in King, Lin, and Ionides (2024) for analyzing real-world tree-structured epidemiological data. We applied this framework in tandem with the Iterated Filtering 2 Algorithm (IF2) (Ionides et al. 2015) to perform likelihood maximization of a reconstructed phylogenetic tree of MERS genomes sourced from (Dudas et al. 2018). This thesis had two main goals: assessing the feasibility of using the PhyloPOMP framework on real-world epidemiological data using the IF2 algorithm, and gaining new insights on the transmission dynamics of MERS-CoV by observing the fitted parameters.

Our model achieved a higher Log-Likelihood value compared to the Kingman Coalescent

benchmark. Optimizing the Moran model with the `optim` R package, we achieved a maximum Log-Likelihood value of -95.2, whereas our PhyloPOMP two-species model fitted with the IF2 algorithm achieved a maximum Log-Likelihood value of 124.23.

Improved model fits compared to standard alternatives to the same data gives credibility to novel model based conclusions. Specifically, it supports interpreting parameter estimates as meaningful scientific findings. One particularly interesting finding from our parameter estimation is the relatively high reproduction number (R_0) of MERS-CoV in human-to-human transmission, which the model suggests is around 2.64. This estimate is notably high and contrasts with the general consensus that MERS-CoV typically exhibits an R_0 below 1 in human populations, implying limited potential for sustained community transmission. A possible explanation for this discrepancy is that the model dynamics of human transmission may be dominated by dynamics in higher transmission settings, such as hospital related outbreaks, related to healthcare facilities, such as the major hospital outbreaks in Saudi Arabia and South Korea in 2013, and 2015, respectively.

Despite this, our results demonstrate the effectiveness of PhyloPOMP models and highlight the utility of applying the IF2 algorithm for likelihood maximization on reconstructed phylogenetic trees. We extend the theoretical work done by King, Lin, and Ionides (2024) by applying the PhyloPOMP framework on a real dataset. Furthermore, we successfully utilized the IF2 algorithm on a phylogenetic tree to maximize the likelihood of a phylogenetic tree by perturbing the transmission and sampling rate parameters.

Due to the novelty of the approach used in this project and the complexity of the available data, our analysis has natural limitations. One potential improvement that could be made is refining the underlying model. The two species model treats the tree symmetrically in the context of species, i.e., the tree doesn't distinguish between species types at each node. The reconstructed tree from (Dudas et al. 2018) has labelings of the predicted species associated with the node. As such, one possible improvement to be made is to use this information in the context of the model in order to refine the two species model.

Another potential improvement is the introduction of multiple demes of camels at different ages. Tolah et al. (2020) states that compared to other ages of camel, camels that are aged from 1-2 years had a MERS-CoV Viral RNA detection of (28.4%, 95% CI 24.1–32.8), which is significantly higher compared to younger (9.3%, 95% CI 0.6–18.0) and older camels (16.8%, 95% CI 14.1–19.6). Introducing a deme for camels aged 1-2 years could allow for a greater understanding of the effect that young camels have on MERS-CoV transmission dynamics, and increase model accuracy overall.

An additional refinement involves increasing the number of particles used in the profile likelihood estimation to match the level used in the IF2 runs. The profile likelihood seen in 6 has a high amount of variance in the IF2 runs. Increasing the particle count can substantially reduce the variance of log-likelihood estimates, leading to smoother and more reliable profiles.

Finally, while the IF2 algorithm successfully identified a high likelihood region of the likelihood surface, there was substantial variance in the IF2 estimates compared to POMP models using more traditional time series data, even with high particle counts. A possible explanation is that simulating genealogies under a model is inherently complex and leads to higher Monte Carlo errors compared to traditional POMP models. Future work could involve tuning IF2 for PhyloPOMP or exploring more computationally efficient filtering strategies.

References

Alharbi, N.K., O.H. Ibrahim, A. Alhafufi, et al. (2020). "Challenge infection model for MERS-CoV based on naturally infected camels". In: *Virology Journal* 17, p. 77. DOI: 10.1186/s12985-020-01347-5. URL: https://doi.org/10.1186/s12985-020-01347-5.

Arulampalam, M.S. et al. (2002). "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". In: *IEEE Transactions on Signal Processing* 50.2, pp. 174–188. DOI: 10.1109/78.978374.

Azhar, Esam I. et al. (2023). "Middle East respiratory syndrome coronavirus—a 10-year (2012-2022) global analysis of human and camel infections, genomic sequences, lineages, and geographical origins". In: *International Journal of Infectious Diseases* 131, pp. 87–94. ISSN: 1201-9712. DOI: https://doi.org/10.1016/j.ijid.2023.03.046. URL: https://www.sciencedirect.com/science/article/pii/S120197122300125X.

Disease Control, Centers for and Prevention (2025). *Middle East Respiratory Syndrome (MERS): Clinical Overview*. Accessed: 2025-03-08. URL: https://www.cdc.gov/mers/hcp/clinical-overview/index.html.

Dudas, Gytis et al. (2018). "MERS-CoV spillover at the camel-human interface". In: *eLife* 7. © 2018, Dudas et al., e31257. ISSN: 2050-084X. DOI: 10.7554/eLife.31257. URL: https://doi.org/10.7554/eLife.31257.

Gossner, C et al. (Dec. 2014). "Human-Dromedary Camel Interactions and the Risk of Acquiring Zoonotic Middle East Respiratory Syndrome Coronavirus Infection". en. In: *Zoonoses Public Health* 63.1, pp. 1–9.

Ionides, Edward L. et al. (2015). "Inference for dynamic and latent variable models via iterated, perturbed Bayes maps". In: *Proceedings of the National Academy of Sciences* 112.3, pp. 719–724. DOI: 10.1073/pnas.1410597112. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1410597112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1410597112.

Kendall, David G. (1948). "On the Generalized "Birth-and-Death" Process". In: *The Annals of Mathematical Statistics* 19.1, pp. 1–15. ISSN: 00034851. URL: http://www.jstor.org/stable/2236051 (visited on 04/06/2025).

King, Aaron A. (Nov. 2024). *Exact phylodynamics via structured Markov genealogy processes*. Theoretical Ecology Seminar. Presented at the Theoretical Ecology Seminar, University of Michigan and Santa Fe Institute. URL: https://scholar.google.com/

scholar_lookup?title=Exact%20phylodynamics%20via%20structured%20Markov%20genealogy%20processes.

King, Aaron A. and Edward L. Ionides (2024). *Lesson 1: Introduction to Simulation-based Inference for Epidemiological Dynamics*. Available online.

King, Aaron A., Qianying Lin, and Edward L. Ionides (2024). *Exact phylodynamic likelihood via structured Markov genealogy processes*. arXiv: 2405.17032 [q-bio.QM]. URL: https://arxiv.org/abs/2405.17032.

Kingman, J.F.C. (1982). "The coalescent". In: *Stochastic Processes and their Applications* 13.3, pp. 235–248. ISSN: 0304-4149. DOI: https://doi.org/10.1016/0304-4149(82)90011-4. URL: https://www.sciencedirect.com/science/article/pii/0304414982900114.

Meyer, Benjamin et al. (Dec. 2016). "Time Course of MERS-CoV Infection and Immunity in Dromedary Camels". en. In: *Emerg Infect Dis* 22.12, pp. 2171–2173.

Moran, P. A. P. (1958). "Random processes in genetics". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54.1, pp. 60–71. DOI: 10.1017/S0305004100033193.

Reusken, Chantal Bem et al. (Jan. 2016). "Cross host transmission in the emergence of MERS coronavirus". en. In: *Curr Opin Virol* 16, pp. 55–62.

Tolah, Ahmed M et al. (May 2020). "Cross-sectional prevalence study of MERS-CoV in local and imported dromedary camels in Saudi Arabia, 2016-2018". en. In: *PLoS One* 15.5, e0232790.

WHO (2025). *Middle East respiratory syndrome coronavirus (MERS-CoV)*. Accessed: 2025-04-06. URL: https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers.

Zumla, Alimuddin, David S Hui, and Stanley Perlman (Sept. 2015). "Middle East respiratory syndrome". In: *The Lancet* 386.9997, pp. 995–1007. DOI: 10.1016/S0140-6736(15)60454-8. URL: https://doi.org/10.1016/S0140-6736(15)60454-8.