Demographic Inference with Coalescent Hidden Markov Model

Jade Y. Cheng
Thomas Mailund

Bioinformatics Research Centre
Aarhus University
Denmark

The Thirteenth Asia Pacific Bioinformatics Conference
HsinChu, Taiwan, January 2015
Presentation Outline

CoalHMM framework
- Overview
- Model construction
  - Continuous time Markov chain (CTMC)
  - Hidden Markov model (HMM)
- Numerical optimization
  - Nelder-Mead
  - Genetic algorithm
  - Particle swarm optimization

Simulation case studies
- Isolation model
- IIM with nine epochs
Framework Overview

CoalHMM is a demographic inference framework based on combining the sequential Markov coalescence with hidden Markov models.

E.g.

Demographic parameters:

1. Isolation duration
2. Migration duration
3. Coalescent rate
4. Recombination rate
5. Migration rate

Our framework is available under open source licence GPLv2 at

https://github.com/mailund/IMCoalHMM
Presentation Outline

CoalHMM framework

Overview

Model construction
- Continuous time Markov chain (CTMC)
- Hidden Markov model (HMM)

Numerical optimization
- Nelder-Mead
- Genetic algorithm
- Particle swarm optimization

Simulation case studies

Isolation model
- IIM with nine epochs
Hidden Markov Model

In the context of CoalHMM, hidden states are different coalescence trees.

E.g. for two samples, the hidden states are the coalescence times.
HMM Transition Probabilities

Transition probability $T_{ij}$ is the normalized joint probability $\tilde{f}_{ij}$, which is the probability of observing coalescence of the left nucleotide in time period $i$ and coalescence of the right nucleotide in time period $j$. 

E.g.  
\begin{align*}
\text{time} & \quad t_0 \quad t_1 \quad t_2 \quad t_3 \quad t_4 \\
\text{J33} & \quad | \quad o-o \quad o-o \quad o-o \quad o-o \quad o-o \\
\end{align*}
Continuous Time Markov Chain

CTMC state space for two samples in two isolated populations.

0 [(1, ([1], [])), (1, ([], [1])), (2, ([2], [])), (2, ([], [2]))]
1 [(1, ([1], [])), (2, ([2], [2]))]
2 [(1, ([1], [1])), (2, ([2], [2]))]
3 [(1, ([1], [1])), (2, ([2], [2]))]
Continuous Time Markov Chain

CTMC state space for two samples in a single population.

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 - C1 C1 C1 C1 0 0 C1 0 0 C1 0 0 0 0 0
1 R - 0 0 0 C1 0 0 C1 0 0 C1 0 0 0
2 R 0 - 0 0 0 C1 0 0 C1 0 0 0 0 0
3 R 0 0 - 0 0 C1 0 C1 0 0 0 C1 0 0
4 R 0 0 0 - C1 0 0 0 C1 0 0 0 0 C1 0 0
5 0 R 0 0 R - 0 0 0 0 0 0 0 0 0 C1
6 0 0 R R 0 0 - 0 0 0 0 0 0 0 C1
7 0 0 0 0 0 0 0 - C1 C1 0 0 0 C1 0
8 0 0 0 0 0 0 0 R - 0 0 0 0 0 C1
9 0 0 0 0 0 0 0 R 0 - 0 0 0 0 C1
10 0 0 0 0 0 0 0 0 0 0 0 - C1 C1 C1 0
11 0 0 0 0 0 0 0 0 0 0 0 R - 0 0 C1
12 0 0 0 0 0 0 0 0 0 0 0 R 0 - 0 C1
13 0 0 0 0 0 0 0 0 0 0 0 0 0 - C1
14 0 0 0 0 0 0 0 0 0 0 0 0 0 R -
The size of CTMC state space grows exponentially with the number of populations, samples, or loci.

E.g. CTMC for a

<table>
<thead>
<tr>
<th>3 samples</th>
<th>3 populations</th>
<th>2 loci</th>
<th>2 donor populations</th>
<th>1 receiver population</th>
</tr>
</thead>
</table>

demographic senario has 578 states.
We need projection matrices to move samples between time slices that have different CTMC state spaces.

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\
\hline
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]
HMM Transition Probabilities

Transition probability $T_{ij}$ is the normalized joint probability $J_{ij}$, which is the probability of observing coalescence of the left nucleotide in time period $i$ and coalescence of the right nucleotide in time period $j$.

E.g.  
\[ \text{t0} \quad \text{t1} \quad \text{t2} \quad \text{t3} \quad \text{t4} \]

\[ \begin{array}{c}
| \text{time} | \\
| o-o | \\
| x-o | \\
| o-o | \\
\end{array} \]

\[ \begin{array}{c}
| \text{time} | \\
| o-o | \\
| o-o | \\
| x-x | \\
\end{array} \]

\[ \begin{array}{c}
| \text{time} | \\
| o-o | \\
| o-o | \\
| x-x | \\
\end{array} \]

\[ \begin{array}{c}
| \text{time} | \\
| o-o | \\
| o-o | \\
| o-o | \\
\end{array} \]

\[ J_{33} \quad J_{34} \]
HMM Transition Probabilities

For a two-sample CTMC, we can split its rate and probability matrices into 16 sections using the 4 state types: begin (B), left (L), right (R), and end (E)

E.g. for the two-sample single-population CTMC
There are three possible ways to reach an E state from a B state, which is the initial condition of two samples.

Goal: B to E

Legal Moves:
- B to B
- B to L
- B to R
- B to E
- L to L
- L to E
- R to R
- R to E

Possible Paths:

#1 B to B → B to E

#2 B to B → B to L → L to L → L to E

#3 B to B → B to R → R to R → R to E
The probability of taking path 3 is the same as taking path 2. Joint probability matrix, $J$, is symmetric.

$$J_{ij} = \begin{cases} J_{ij} & \text{if } i > j \\ \sum_{\alpha} \sum_{\beta} (M_{\alpha \beta}) & \text{if } i \leq j \end{cases}$$

$$M = \begin{cases} (P_{0t})_{BB} \times (P_{1t})_{BB} \times \cdots \times (P_{i-1t})_{BB} \times (P_{it})_{BE} & \text{if } i = j \\ (P_{0t})_{BB} \times \cdots \times (P_{i-1t})_{BB} \times (P_{it})_{BL} \times (P_{i+1t})_{LL} \times \cdots \times (P_{jt})_{LE} & \text{if } i < j \end{cases}$$
HMM Transition Probabilities

E.g. a simple joint probability calculation

\[ J_{24} = \sum_{\alpha} \sum_{\beta} (P_{0t}^{t})_{BB} \times (P_{1t}^{t})_{BL} \times (P_{2t}^{t})_{LL} \times (P_{3t}^{t})_{LE})_{\alpha\beta} \]
HMM Transition Probabilities

E.g. another joint probability calculation involving CTMC projection.

\[ J_{34} = \sum_{\alpha} \sum_{\beta} \left( (P_0^t)_{BB} \times (P_1^t)_{BB} \times (P_2^t)_{BL} \times (P_{MIG\rightarrow SIN})_{LL} \times (P_3^t)_{LE} \right)_{\alpha\beta} \]
Presentation Outline

CoalHMM framework
  Overview
  Model construction
    Continuous time Markov chain (CTMC)
    Hidden Markov model (HMM)
  Numerical optimization
    Nelder-Mead
    Genetic algorithm
    Particle swarm optimization

Simulation case studies
  Isolation model
  IIM with nine epochs
Nelder-Mead optimization minimises an objective function in a many-dimensional space by continuously refining a simplex.
A Genetic Algorithm is a type of evolutionary algorithm.
Selection: a GA chooses a relatively fit subset of individuals for breeding.

E.g.  
- fitness 40;  
- fitness 25;  
- fitness 18;  
- fitness 12;  
- fitness 5
Rank Based Selection

Tournament selection selects individuals with the highest fitness values from random subsets of the population.

E.g.  ■ fitness 40;  ■ fitness 25;  ■ fitness 18;  ■ fitness 12;  ■ fitness 5

Original Population

Tournament #1

Tournament #2

Tournament #3

Breeding Pool
Crossover & Mutation

Crossover: it is a genetic operation used to combine pairs of individuals previously selected for breeding the following generation.

One-point crossover
Two-point crossover
Uniform crossover

Mutation: each position has a certain probability to mutate,
Genetic Particle Swarm Optimization

PSO is another heuristic based search algorithm.

\[
v'_{i,d} \leftarrow \omega \cdot v_{i,d} + \phi_p \cdot r_p \cdot (p_{i,d} - x_{i,d}) + \phi_g \cdot r_g \cdot (p_{g,d} - x_{i,d})
\]

\[
x'_{i,d} \leftarrow x_{i,d} + v_{i,d}.
\]
Presentation Outline

CoalHMM framework
- Overview
- Model construction
  - Continuous time Markov chain (CTMC)
  - Hidden Markov model (HMM)
- Numerical optimization
  - Nelder-Mead
  - Genetic algorithm
  - Particle swarm optimization

Simulation case studies
- Isolation model
- IIM with nine epochs
The simplest demographic model we consider is the clean isolation model. It has three parameters.
IIM-Nine Epoch Model
Discussion

We discussed the construction of our statistic inference framework, CoalHMM and showed how to infer demographics with it.

CTMC + HMM + Numerical Optimization

We compared the estimation accuracy between the previously-used with newly developed optimization methods

Nelder-Mead 😞 GA 😊 PSO 😊

We presented the inference results on two demographic models, a simple one and a more complex one.
Acknowledgement

Thanks!