Demographic Inference with Coalescent Hidden Markov Model

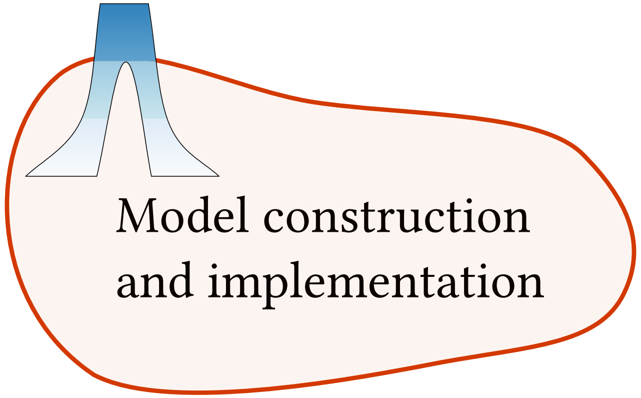
PhD Student:	Jade Y. Cheng
Supervisor:	Thomas Mailund
Institution:	Bioinformatics Research Centre Department of Computer Science Aarhus University Denmark

成玉

Presentation Outline

CoalHMM framework

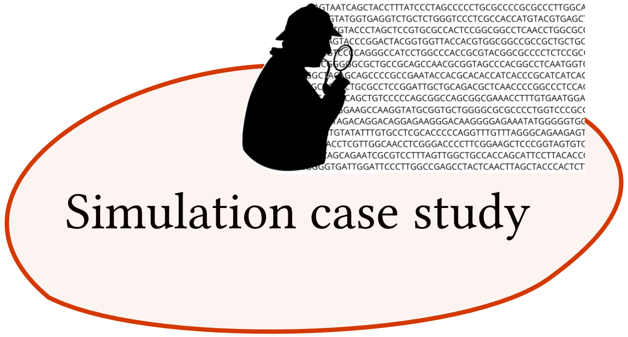
Continuous time Markov chain (CTMC)
Hidden Markov model (HMM)
Numerical optimizations



Model construction and implementation

CoalHMM with simulations

Simulation validation
Performance evaluation with simple to complex models



Simulation case study

CoalHMM with biological data

Data validation with various analyses
CoalHMM inference with Bears

Biological case study

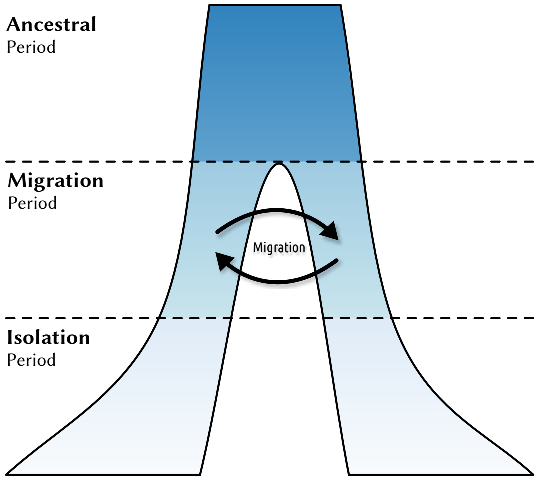
Admixture CoalHMM

General model construction
Three-population admixture model
Bear study case

Framework Overview

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

E.g.



Demorgraphic 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

Continuous time Markov chain (CTMC)

Hidden Markov model (HMM)

Numerical optimizations

CoalHMM with simulations

Simulation validation

Performance evaluation with simple to complex models

CoalHMM with biological data

Data validation with various analyses

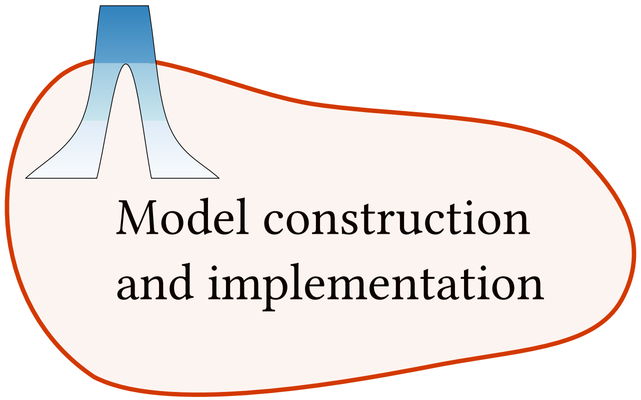
CoalHMM inference with Bears

Admixture CoalHMM

General model construction

Three-population admixture model

Bear study case



Model construction and implementation

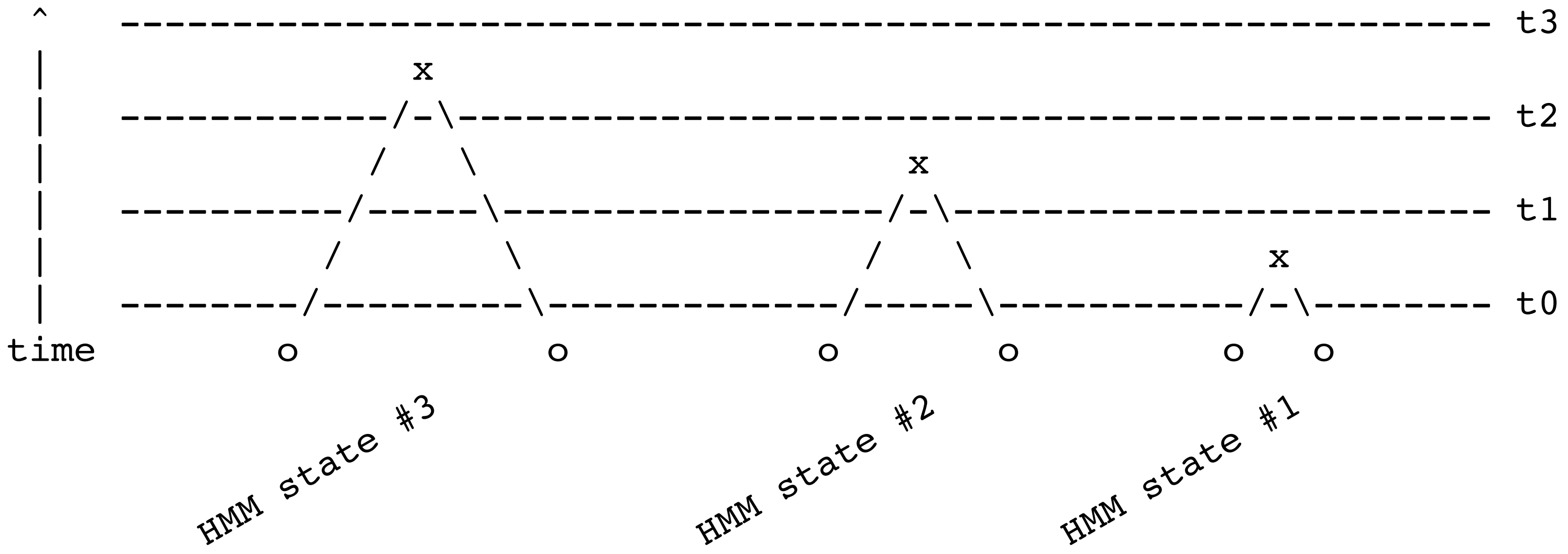
Simulation case study

Biological case study

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 $\mathcal{J}_{ij}$, 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.

```
 ^  --------------------- t4
 |     |  x x  |
 |  --------------------- t3
 |     |  x-x  |
 |  --------------------- t2
 |     |  o-o  |
 |     |  o o  |
 |  --------------------- t1
 |     |  o-o  |
 |     |  o-o  |
 |  --------------------- t0
time      o-o
          o-o
```

$\mathcal{J}_{33}$

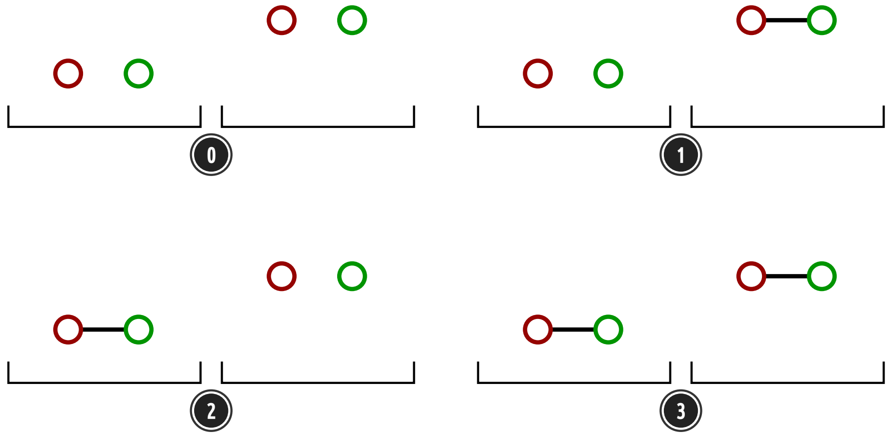
E.g.

```
 ^  ------------------------------ t4
 |     |        x-x        |
 |  ------------------------------ t3
 |     |    o  |   |       |
 |     |  x-o  |   |       |
 |  ------------------------------ t2
 |     |  o o  |   |       |
 |     |  o-o  |   |       |
 |  ------------------------------ t1
 |     |       |   |  o-o  |
 |     |  o-o  |   |       |
 |  ------------------------------ t0
time      o-o         o-o
```

$\mathcal{J}_{34}$

Continuous Time Markov Chain

CTMC state space for two samples in two isolated populations.

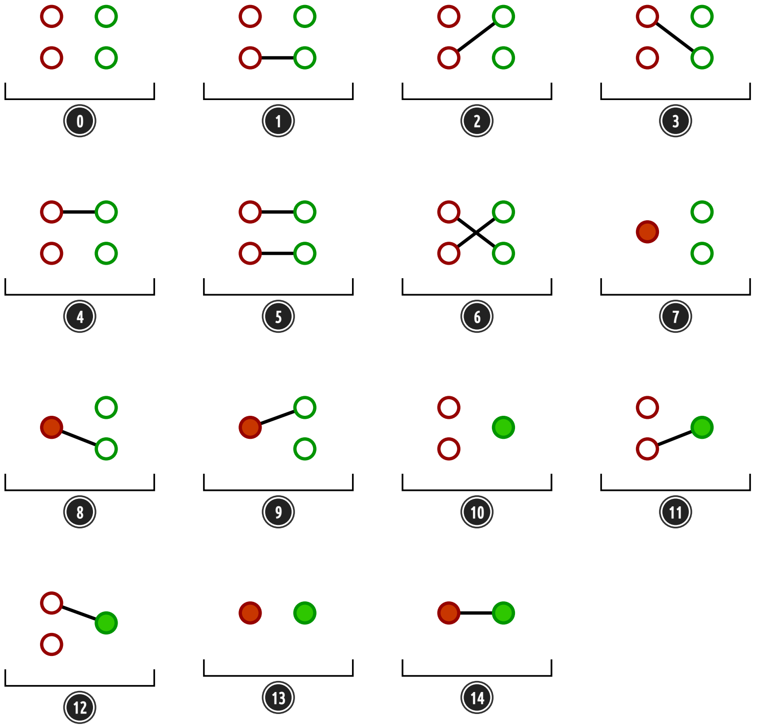


	0	1	2	3
0	-	C2	C1	0
1	R	-	0	C1
2	R	0	-	C2
3	0	R	R	-

```
0 [(1, ([1], [])), (1, ([], [1])), (2, ([2], [])), (2, ([], [2]))]
1 [(1, ([1], [])), (1, ([], [1])), (2, ([2], [2]))]
2 [(1, ([1], [1])), (2, ([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
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	C1	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	C1	0	0
5	0	R	0	0	R	–	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	–	C1	C1	C1	0
11	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	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	–

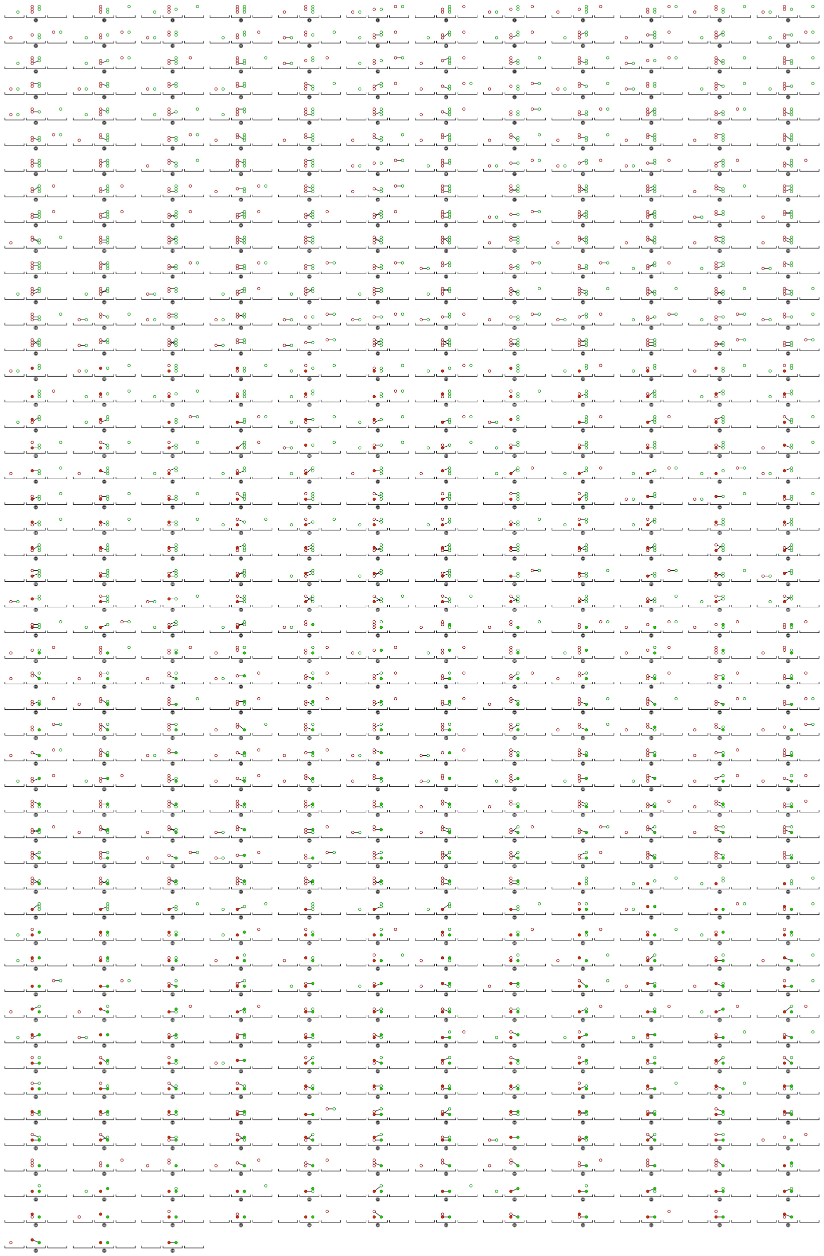
Continuous Time Markov Chain

The size of CTMC state space grows exponentially with the number of populations, samples, or loci.

E.g. CTMC for a

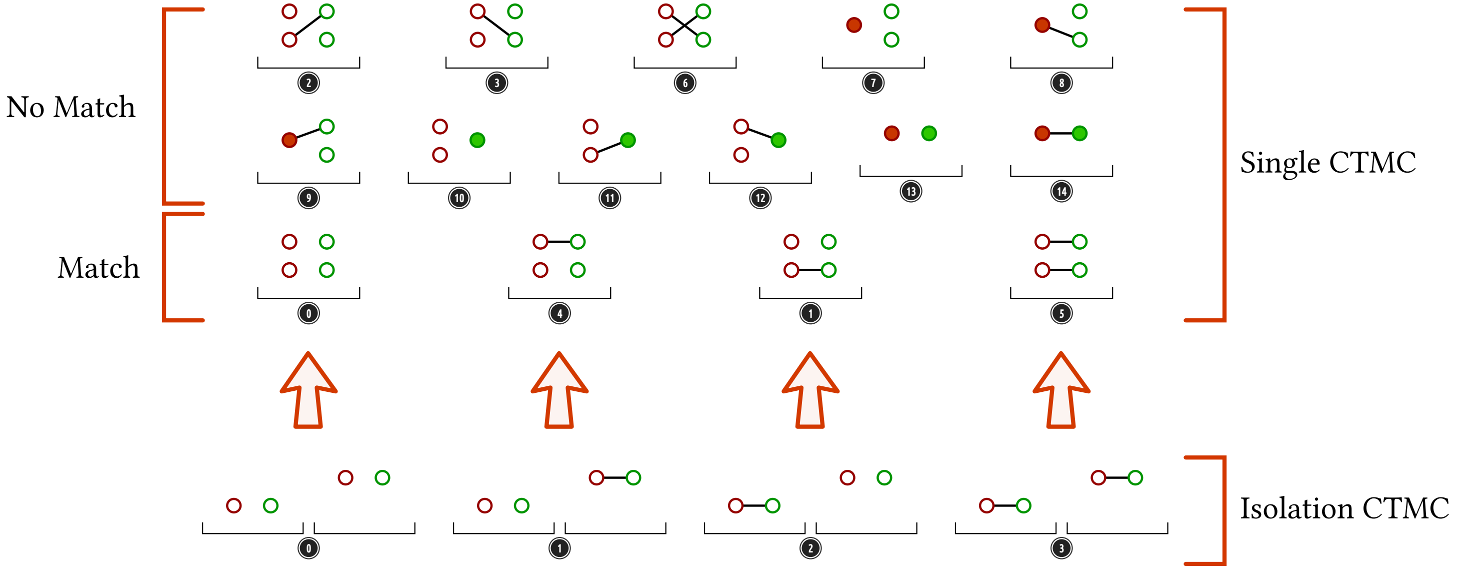
3 samples
3 populations
2 loci
2 donor populations
1 receiver population

demographic senario has 578 states.



CTMC Projections

We need projection matrices to move samples between time slices that have different CTMC state spaces.



No Match

Match

Single CTMC

Isolation CTMC

Isolation CTMC ⇨

```
  |  0  1  2  3  4  5  6  7  8  9  10 11 12 13 14
--+-----------------------------------------------
0 |  1  0  0  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  0  0
2 |  0  1  0  0  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  0  0
```

⇨ Single CTMC

HMM Transition Probabilities

Transition probability T_{ij} is the normalized joint probability $\mathcal{J}_{ij}$, 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.

```
 ^  ---------------------- t4
 |     |  x x  |
 |  ---------------------- t3
 |     |  x-x  |
 |  ---------------------- t2
 |     |  o-o  |
 |     |  o o  |
 |  ---------------------- t1
 |     |  o-o  |
 |     |  o-o  |
 |  ---------------------- t0
time      o-o
          o-o
```

$\mathcal{J}_{33}$

E.g.

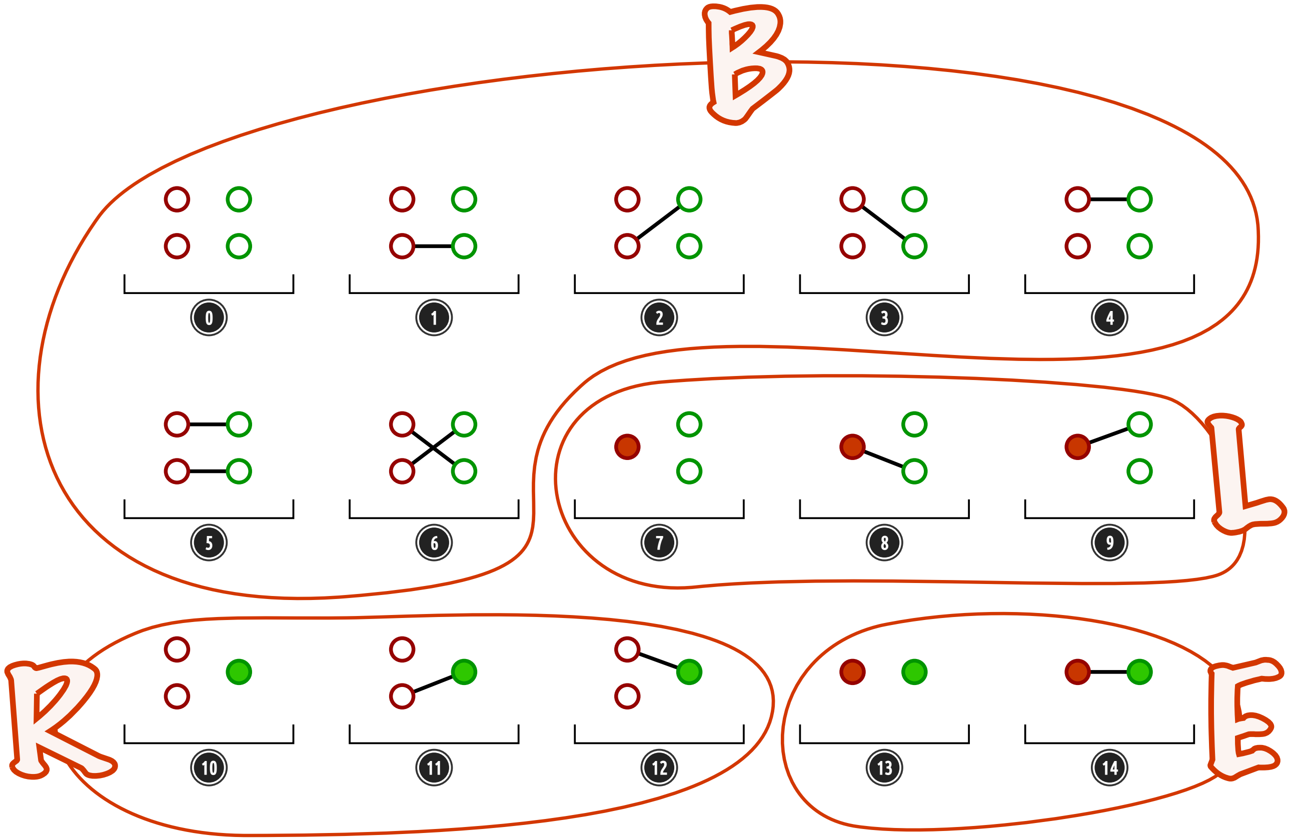
```
 ^  ------------------------------- t4
 |     |        x-x       |
 |  ------------------------------- t3
 |     |    o  |   |       |
 |     |  x-o  |   |       |
 |  ------------------------------- t2
 |     |  o o  |   |       |
 |     |  o-o  |   |       |
 |  ------------------------------- t1
 |     |       |   |  o-o  |
 |     |  o-o  |   |       |
 |  ------------------------------- t0
time      o-o         o-o
```

$\mathcal{J}_{34}$

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



HMM Transition Probabilities

There are three possible ways to reach an E state from a B state, which is the initial condition of two samples.

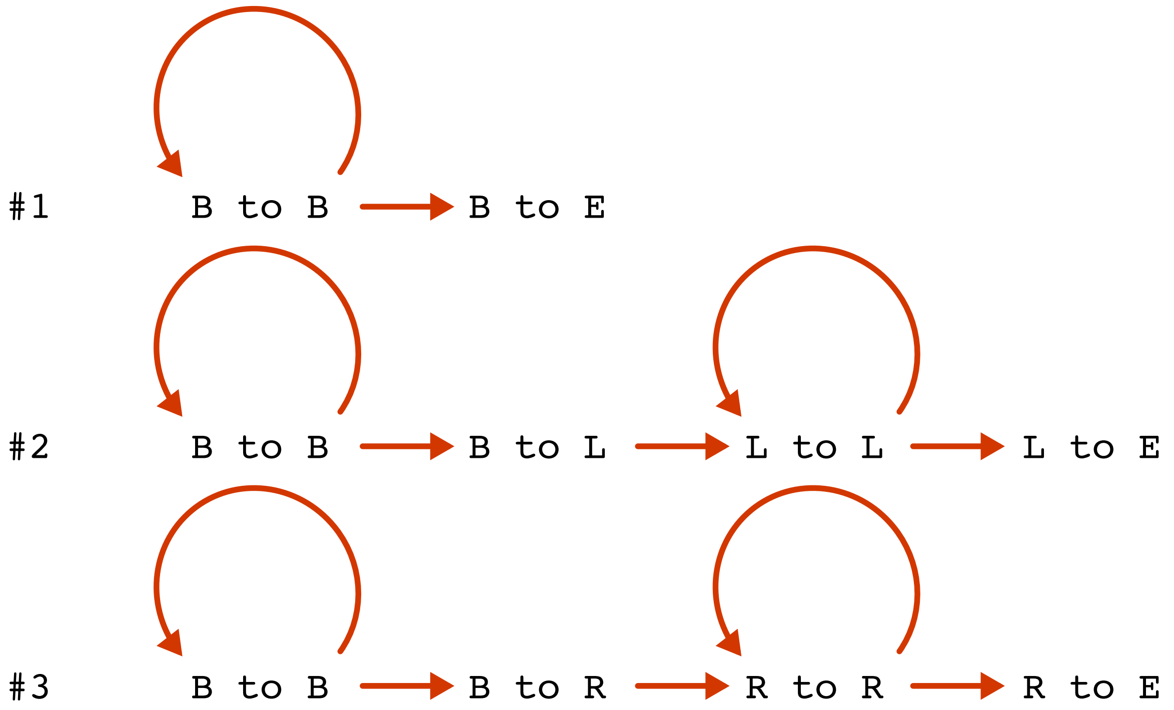
```
Goal:

    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

HMM Transition Probabilities

The probability of taking path 3 is the same as taking path 2. Joint probability matrix, $\mathcal{J}$, is symmetric.

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

$$M = \begin{cases} (P_0^t)_{\mathrm{BB}} \times (P_1^t)_{\mathrm{BB}} \times \cdots \times \left(P_{i-1}^t\right)_{\mathrm{BB}} \times (P_i^t)_{\mathrm{BE}} & \text{if } i = j \\ (P_0^t)_{\mathrm{BB}} \times \cdots \times \left(P_{i-1}^t\right)_{\mathrm{BB}} \times (P_i^t)_{\mathrm{BL}} \times \left(P_{i+1}^t\right)_{\mathrm{LL}} \times \cdots \times \left(P_j^t\right)_{\mathrm{LE}} & \text{if } i < j \end{cases}$$

```
           B     L     R     E
        |-----|-----|-----|-----|
      B | B-B | B-L | B-R | B-E |
        |-----|-----|-----|-----|
      L | L-B | L-L | L-R | L-E |
P =     |-----|-----|-----|-----|
      R | R-B | R-L | R-R | R-E |
        |-----|-----|-----|-----|
      E | E-B | E-L | E-R | E-E |
        |-----|-----|-----|-----|
```

P_{LE}

Presentation Outline

CoalHMM framework

- Continuous time Markov chain (CTMC)
- Hidden Markov model (HMM)
- Numerical optimizations

CoalHMM with simulations

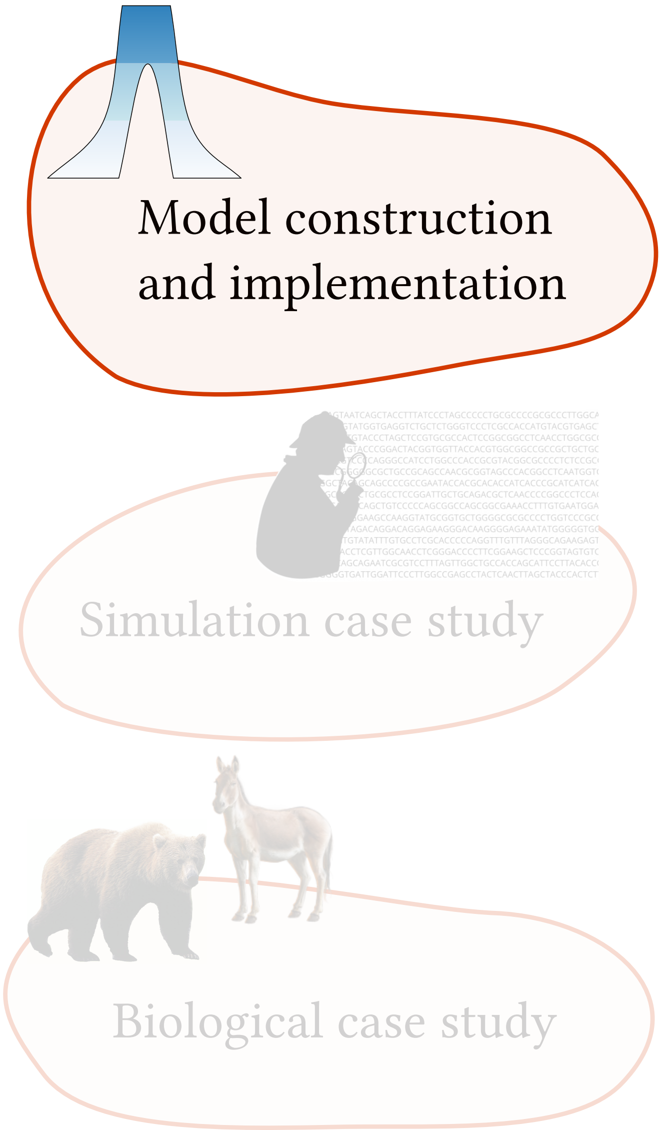
- Simulation validation
- Performance evaluation with simple to complex models

CoalHMM with biological data

- Data validation with various analyses
- CoalHMM inference with Bears

Admixture CoalHMM

- General model construction
- Three-population admixture model
- Bear study case



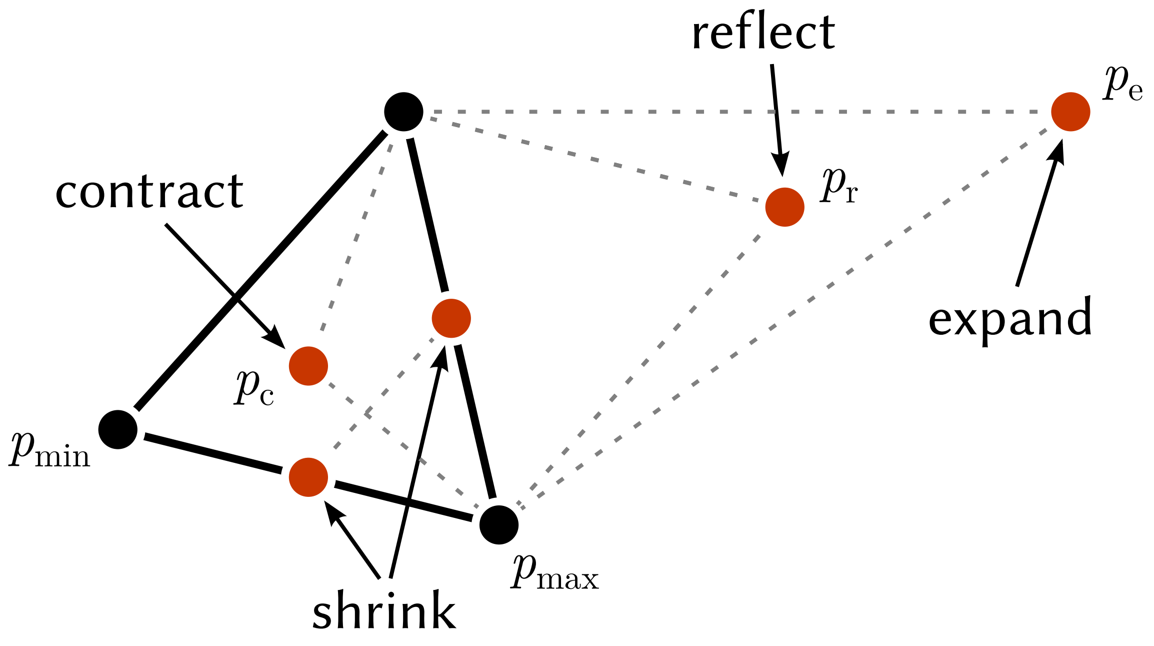
Model construction and implementation

Simulation case study

Biological case study

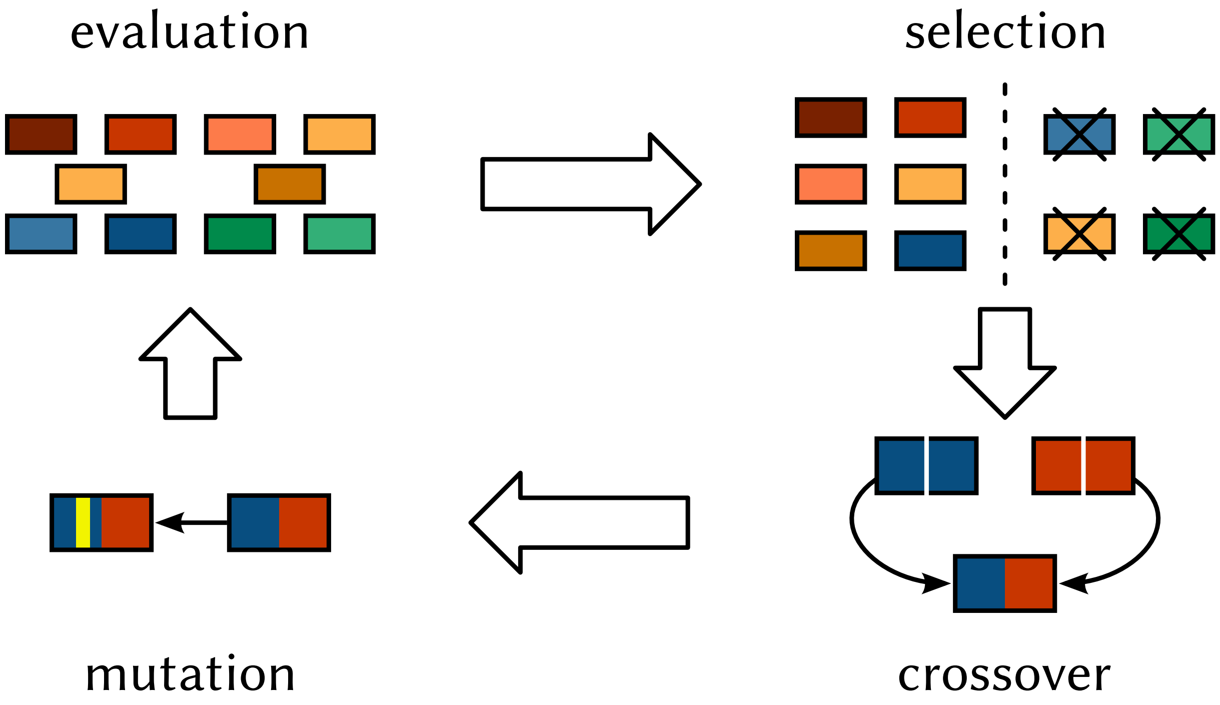
Nelder-Mead optimization

Nelder-Mead optimization minimises an objective function in a many-dimensional space by continuously refining a simplex.



Genetic Algorithm

A Genetic Algorithm is a type of evolutionary algorithm.

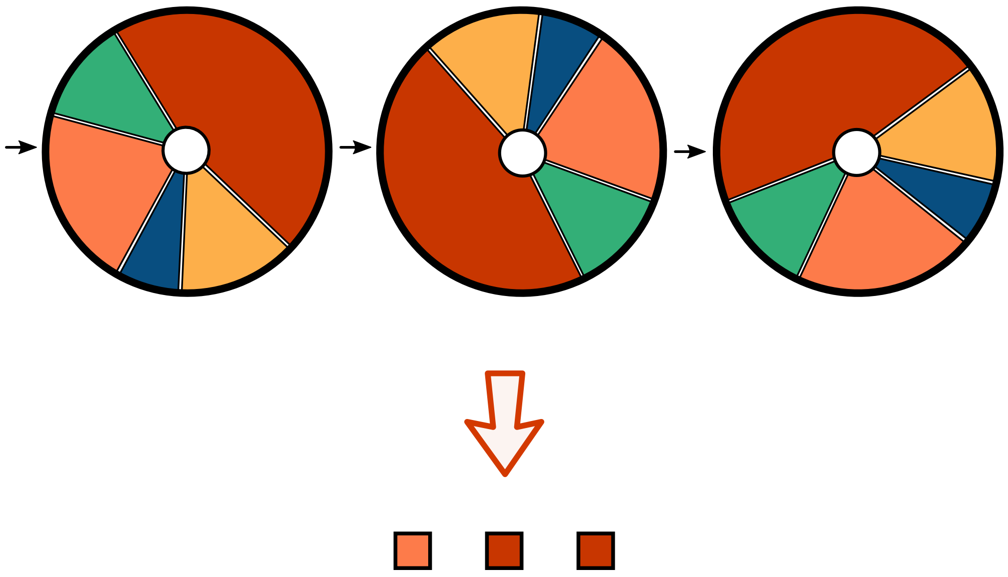


Fitness Proportion Selection

Selection: a GA chooses a relatively fit subset of individuals for breeding.

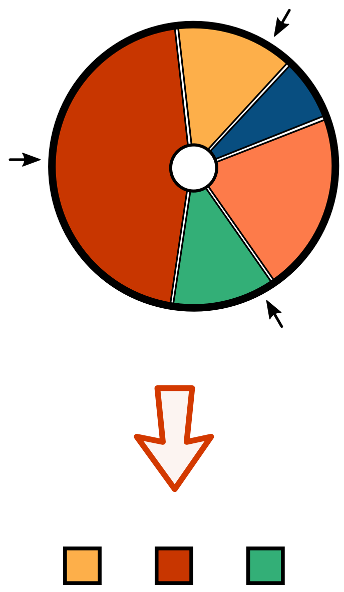
E.g. 🟥 fitness 40; 🟧 fitness 25; 🟨 fitness 18; 🟩 fitness 12; 🟦 fitness 5

Roulette Wheel Selection



Breeding Pool

Universal Sampling

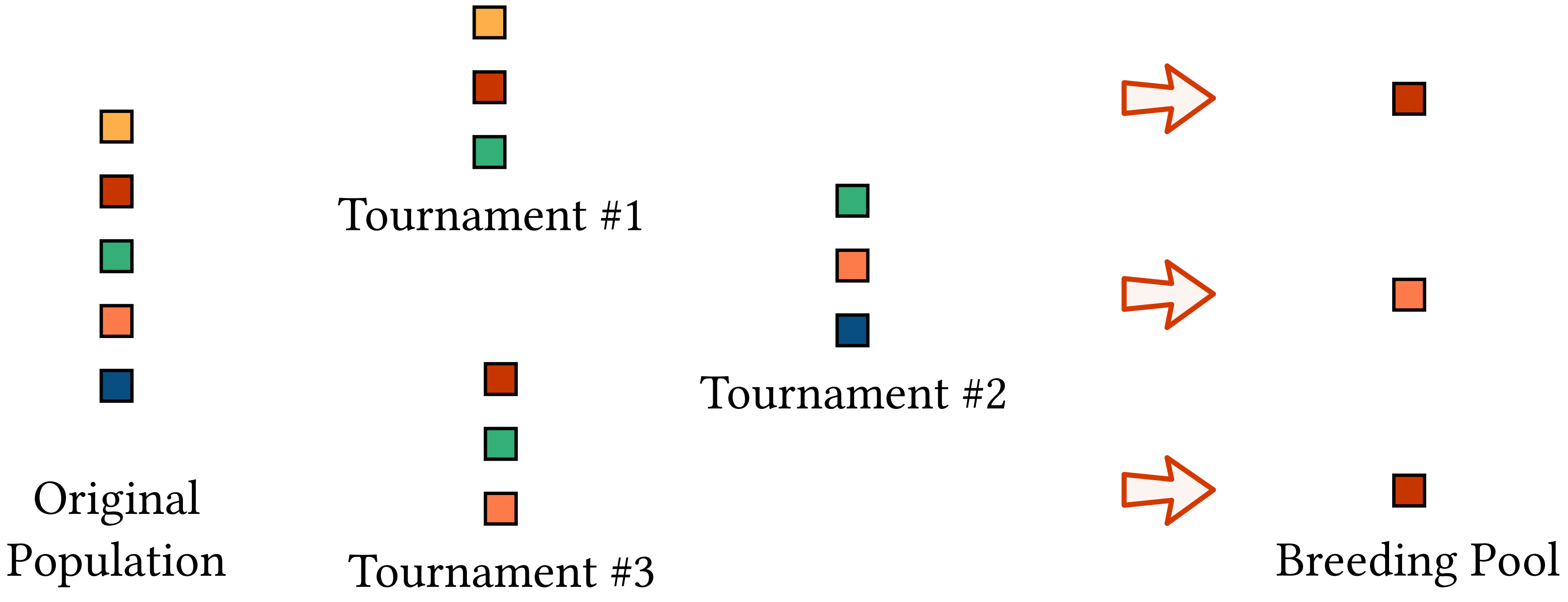


Breeding Pool

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



Tournament #1

Tournament #2

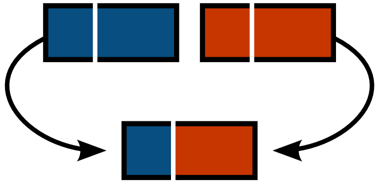
Tournament #3

Original Population

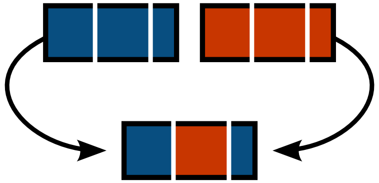
Breeding Pool

Crossover & Mutation

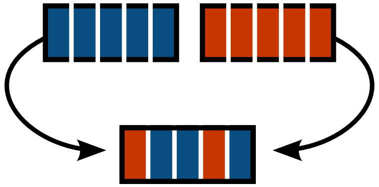
Crossover: it is a genetic operation used to combine pairs of individuals previously selected for breeding the following generation.



One-point crossover

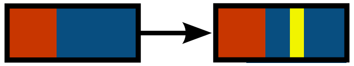


Two-point crossover



Uniform crossover

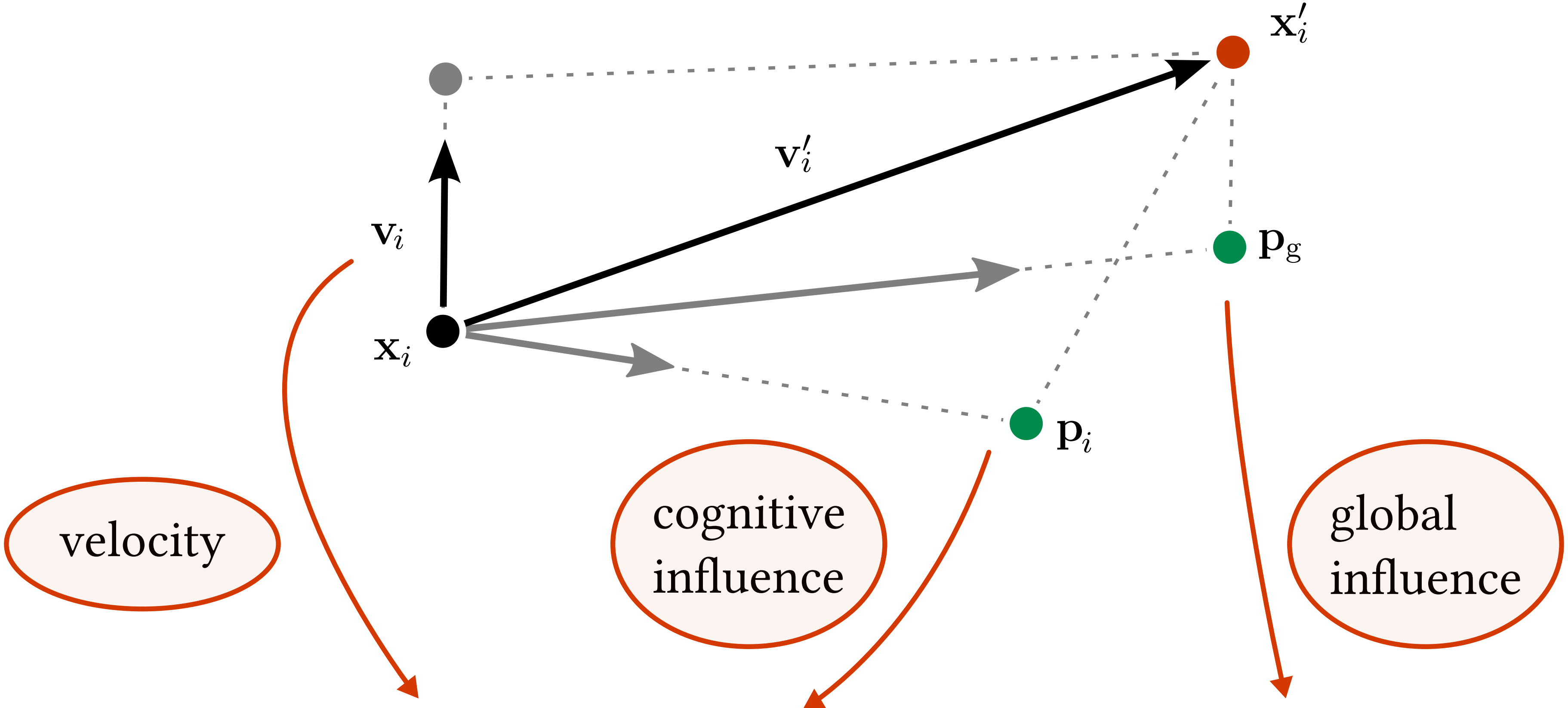
Mutation: each position has a certain probability to mutate,



mutation

Genetic Particle Swarm Optimization

PSO is another heuristic based search algorithm.



velocity

cognitive influence

global influence

$$\mathbf{v}_{i,d}' \leftarrow \omega \cdot \mathbf{v}_{i,d} + \phi_p \cdot r_p \cdot (\mathbf{p}_{i,d} - \mathbf{x}_{i,d}) + \phi_g \cdot r_g \cdot (\mathbf{p}_{g,d} - \mathbf{x}_{i,d})$$

$$\mathbf{x}_{i,d}' \leftarrow \mathbf{x}_{i,d} + \mathbf{v}_{i,d}.$$

Presentation Outline

CoalHMM framework

Continuous time Markov chain (CTMC)
Hidden Markov model (HMM)
Numerical optimizations

CoalHMM with simulations

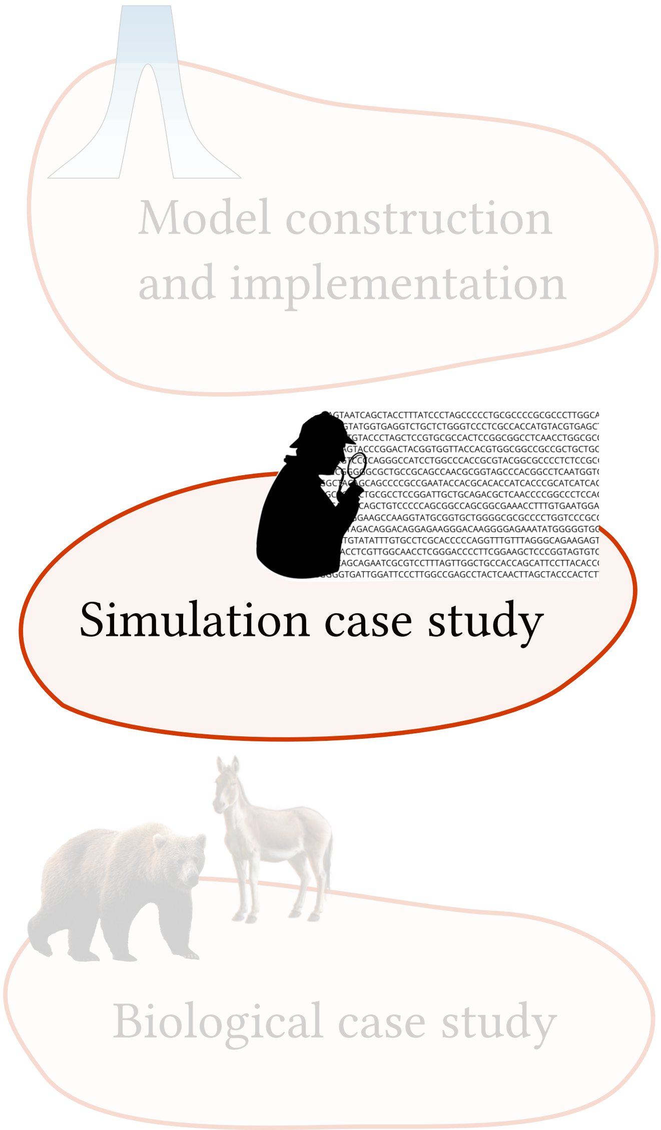
Simulation validation
Performance evaluation with simple to complex models

CoalHMM with biological data

Data validation with various analyses
CoalHMM inference with Bears

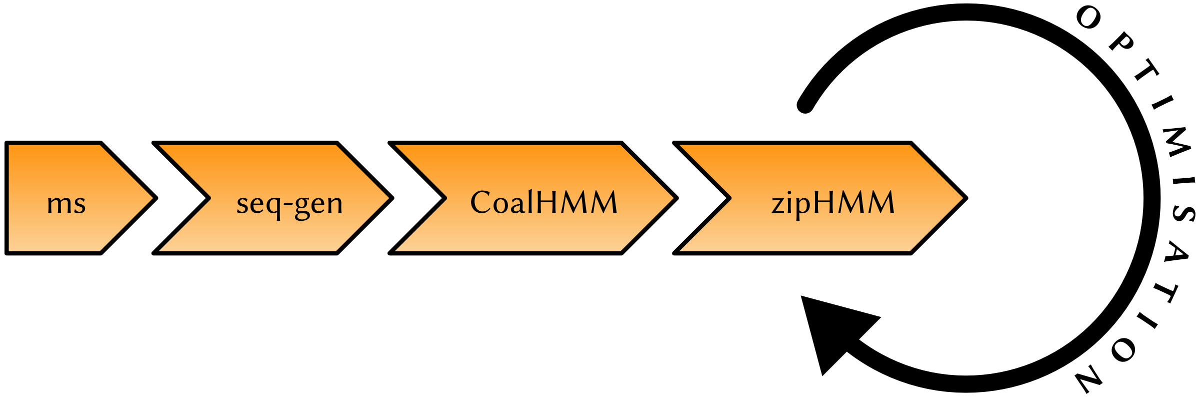
Admixture CoalHMM

General model construction
Three-population admixture model
Bear study case

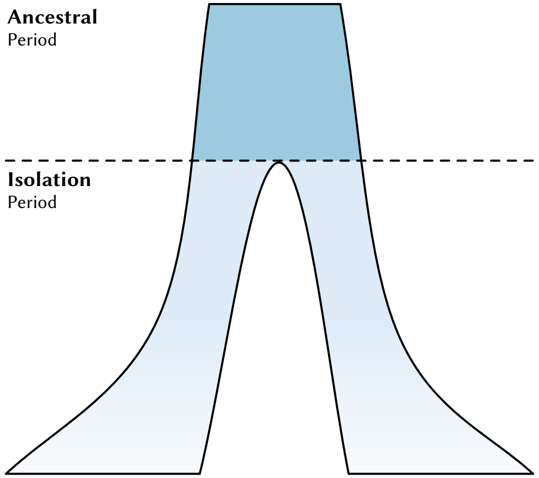


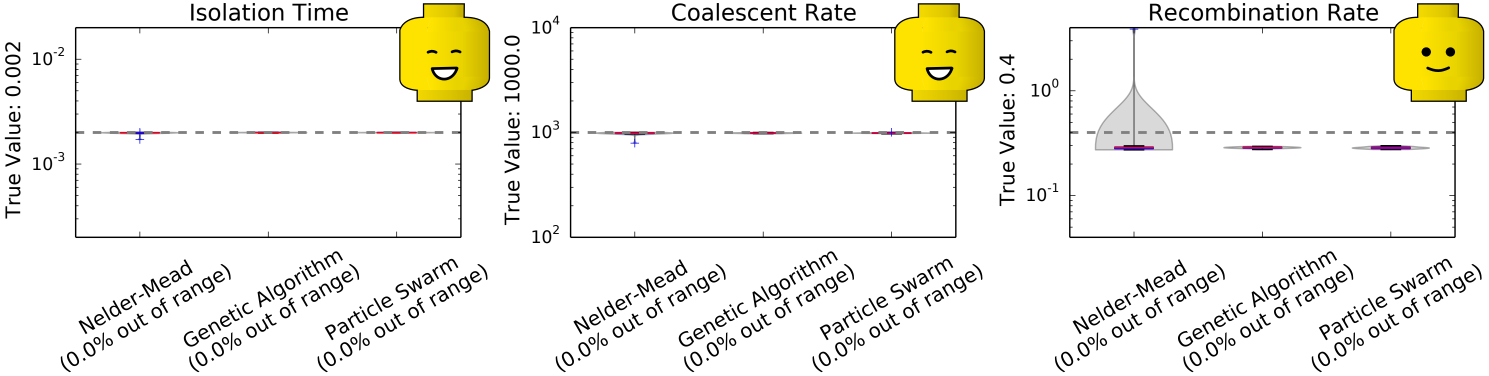
Pipeline & Isolation Model

Simulation study pipeline.

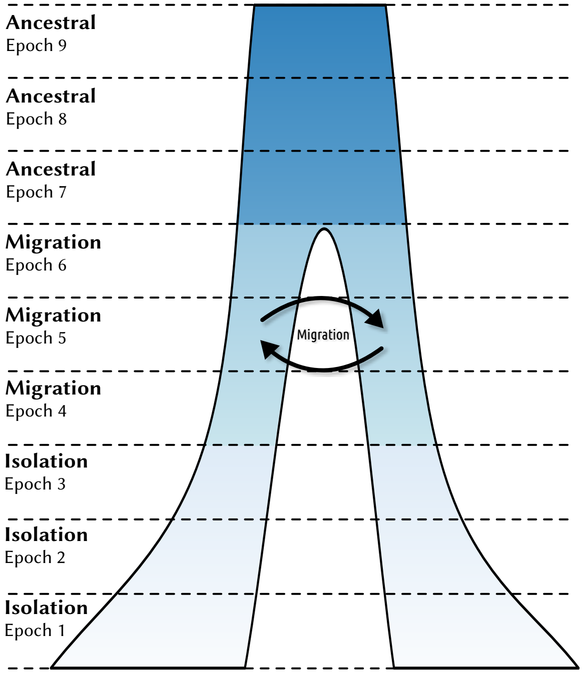


The simplest demographic model we consider is the clean isolation model. It has three parameters.





IIM-Nine Epoch Model



Ancestral Epoch 9

Ancestral Epoch 8

Ancestral Epoch 7

Migration Epoch 6

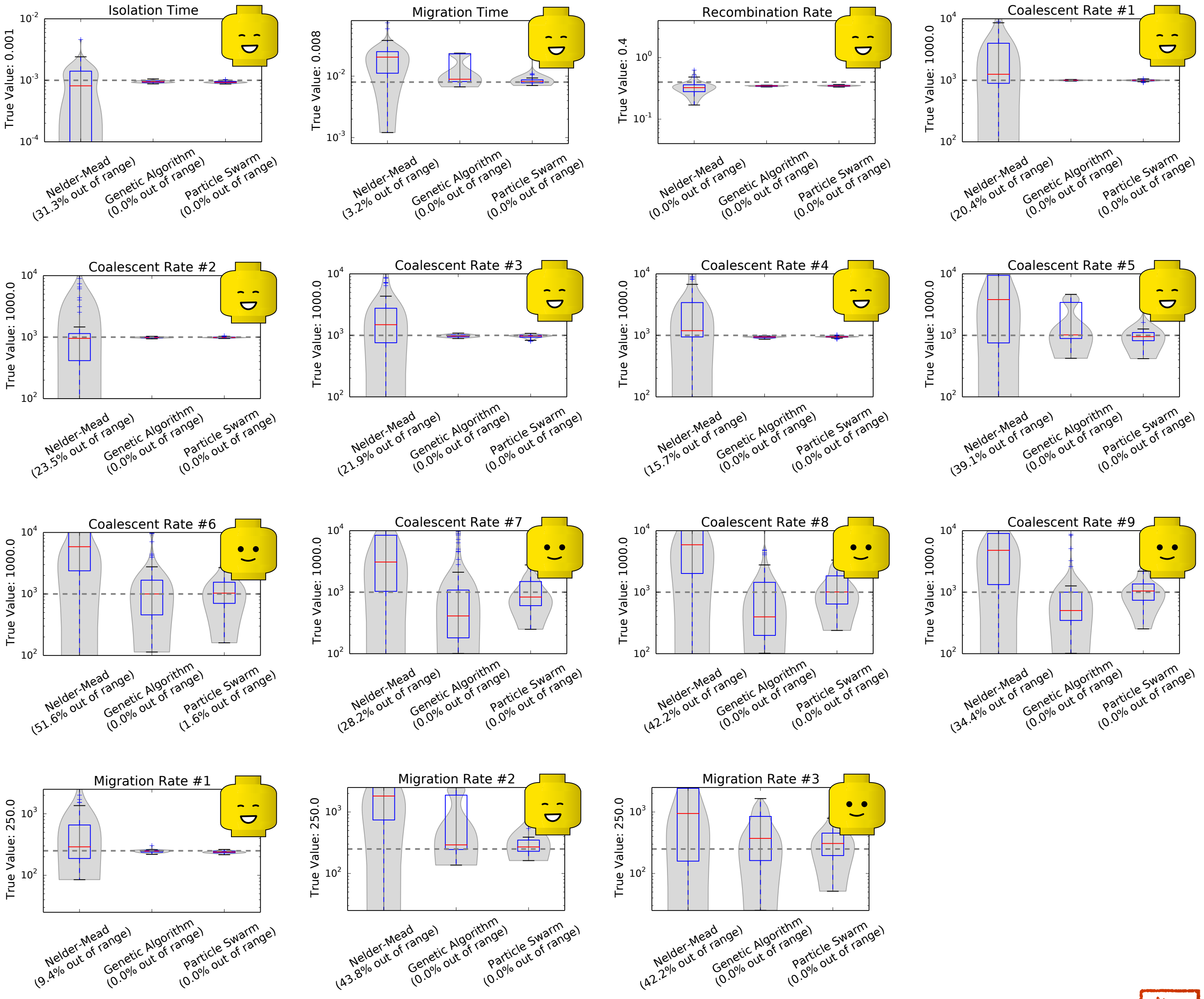
Migration Epoch 5

Migration Epoch 4

Isolation Epoch 3

Isolation Epoch 2

Isolation Epoch 1



Presentation Outline

CoalHMM framework

Continuous time Markov chain (CTMC)
Hidden Markov model (HMM)
Numerical optimizations

CoalHMM with simulations

Simulation validation
Performance evaluation with simple to complex models

CoalHMM with biological data

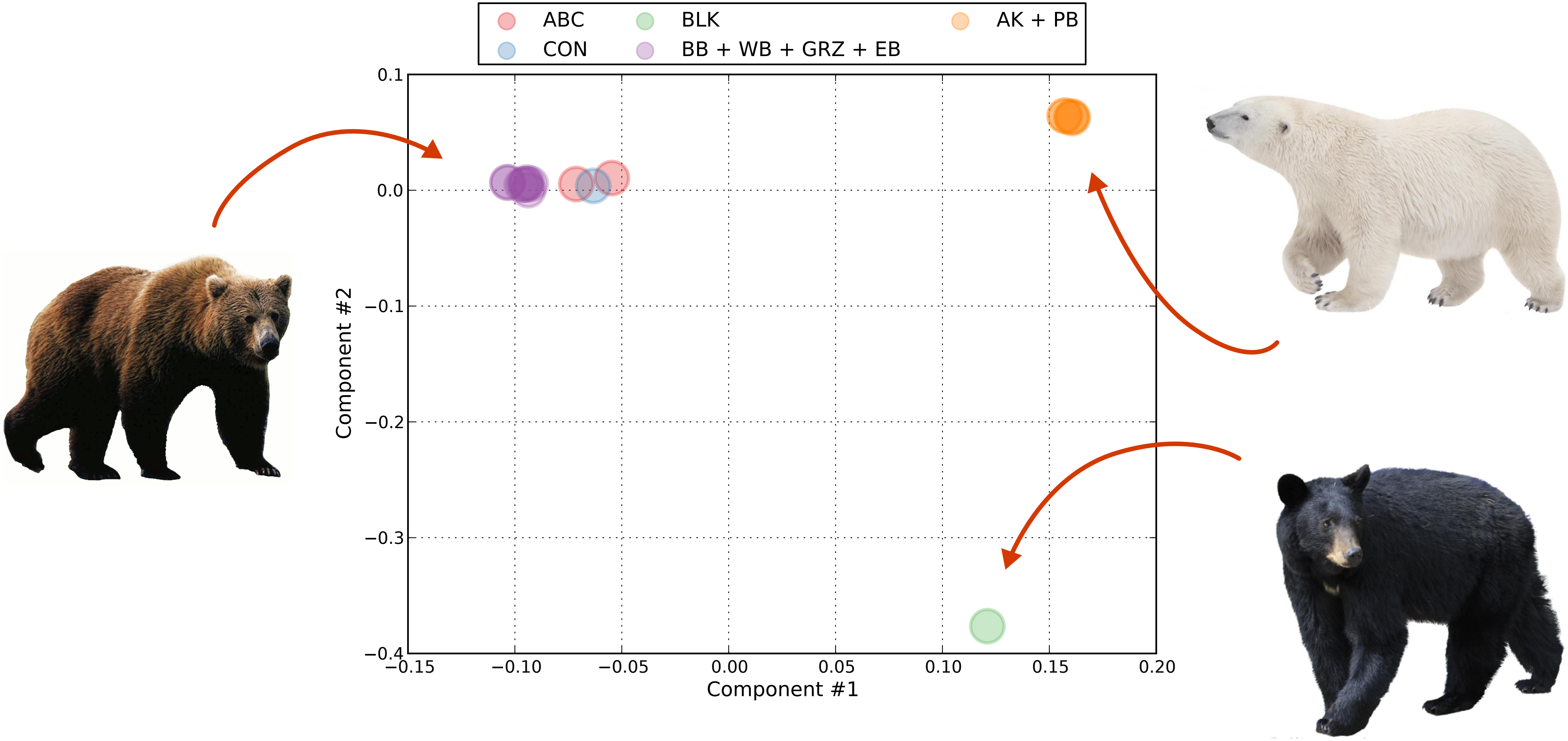
Data validation with various analyses
CoalHMM inference with Bears

Admixture CoalHMM

General model construction
Three-population admixture model
Bear study case

Principle Component Analysis

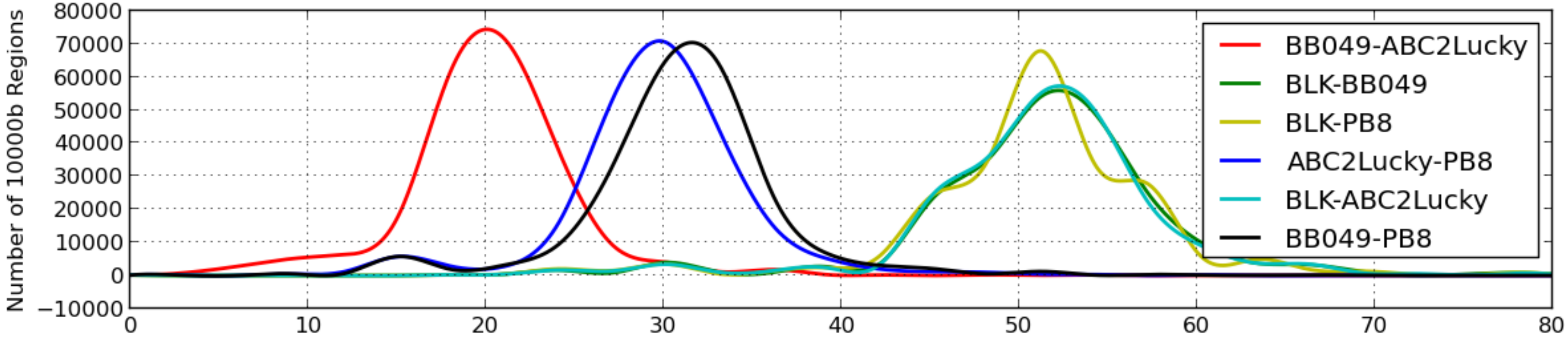
We have 17 bear samples including polar bears, brown bears, ABC island bears, and one black bear.



Pairwise Distance Comparison

For the CoalHMM analysis, we used four bears, a brown bear, an ABC bear, a black bear, and a polar bear.

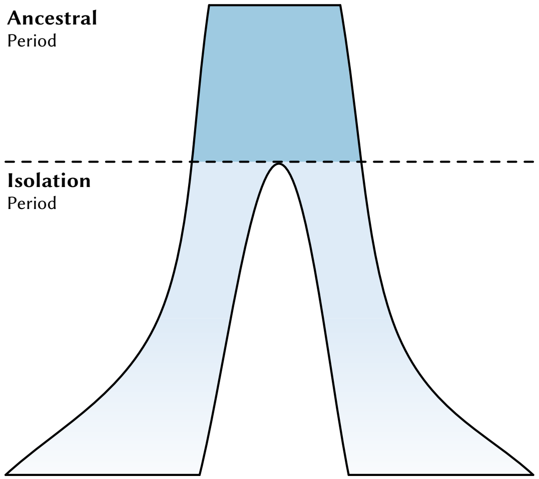
The black bear is the out group.

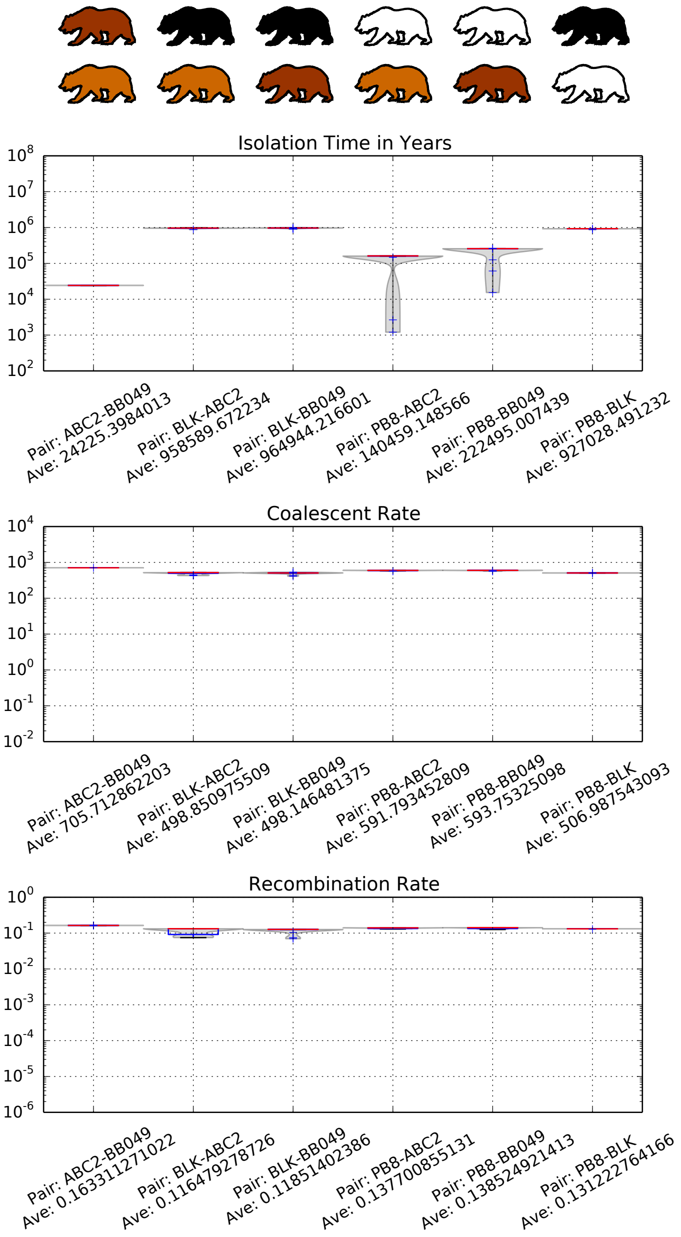


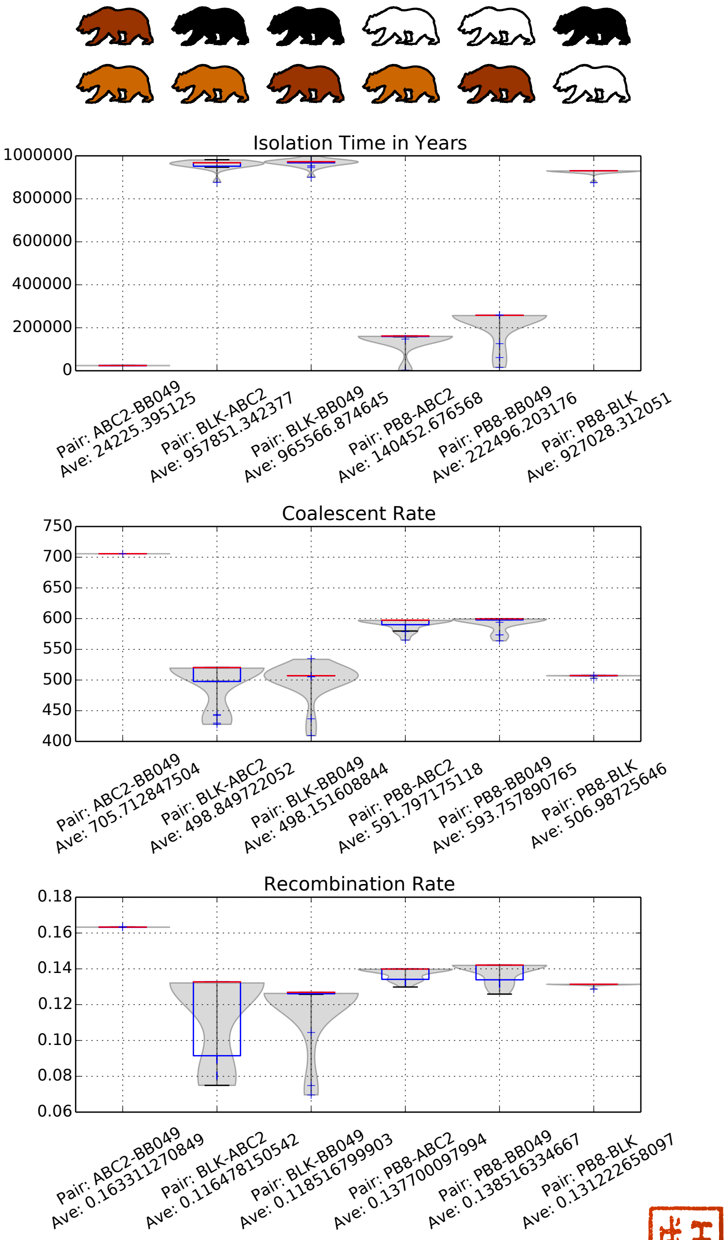
The ABC bear is the closest to the brown bear.

The ABC bear is closer to the polar bear than the brown bear is.

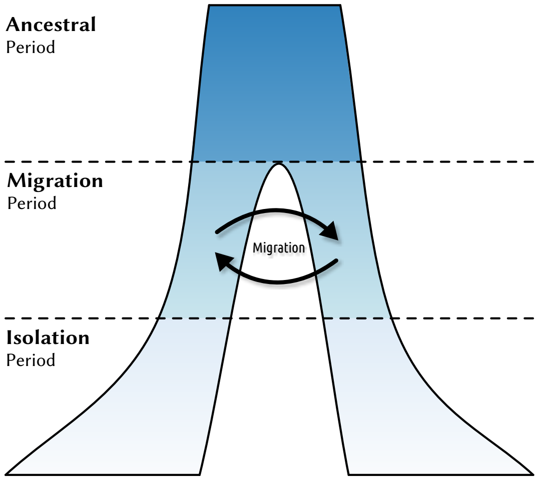
Bear CoalHMM — Isolation Model

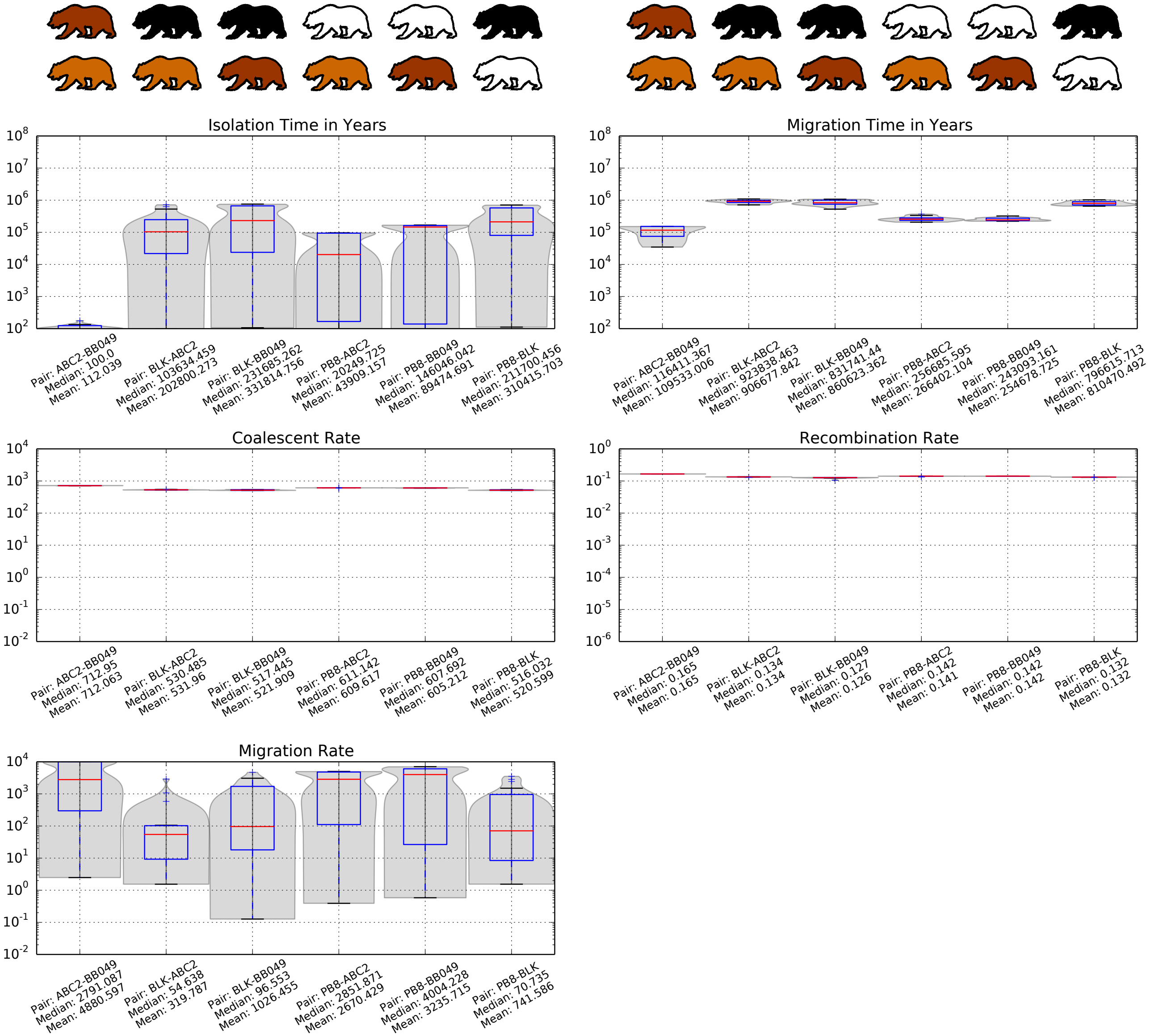






Bear CoalHMM — IIM Model





Bear CoalHMM Summary

Our estimates agree with the recent Cell paper on polar bears from Liu. They concluded that polar bears diverged from brown bears very recently — within the past 500,000 years.

The black bear is the out group.

```
Pair           Isolation-Time     Migration-Time     Split-Time
----------     --------------     --------------     ----------
BLK-PB8              ~210,000           ~800,000     ~1,010,000
BLK-ABC2             ~100,000           ~920,000     ~1,020,000
BLK-BB049            ~230,000           ~830,000     ~1,060,000
PB8-ABC2              ~20,000           ~260,000       ~280,000
PB8-BB049            ~150,000           ~240,000       ~390,000
ABC2-BB049               ~100           ~120,000       ~120,000
```

The ABC bear is the closest to the brown bear.

The ABC bear is closer to the polar bear than the brown bear is.

S. Liu, E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris, Z. Xiong, L. Zhou, T. S. Korneliussen, Somel M, Babbitt C, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785–794, 2014.

Presentation Outline

CoalHMM framework

- Continuous time Markov chain (CTMC)
- Hidden Markov model (HMM)
- Numerical optimizations

CoalHMM with simulations

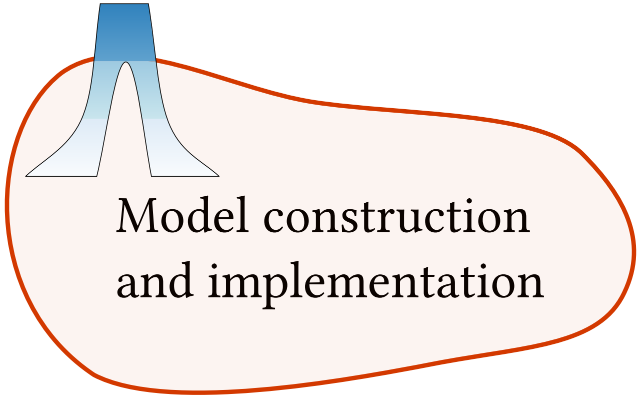
- Simulation validation
- Performance evaluation with simple to complex models

CoalHMM with biological data

- Data validation with various analyses
- CoalHMM inference with Bears

Admixture CoalHMM

- General model construction
- Three-population admixture model
- Bear study case



Model construction and implementation

Simulation case study

Biological case study

Admixture Models

We model gene flow caused by admixture events as single instantaneous events rather than epochs with migration rates.

E.g. A simple admixture model involving two populations. At some time point in the past the two populations exchanged genes.

```
   ---------------------------------------------------------
     b5            |        |                 |
   ----------------------------     Ancestral: 2 time slices
     b4            |        |                 |
ts ---------------------------------------------------------
     b3           /    /\    \                |
   ------------------------------             |
     b2         /     /  \     \     Middle: 3 time slices
   ------------------------------             |
     b1       /     /      \     \            |
ta ----------------------------------------------- [admix] -
     b0      |  *  |        |  *  |       Isolation
   ---------------------------------------------------------
  # of HMM state = 2 + 3 = 5.
```

Joint Probability

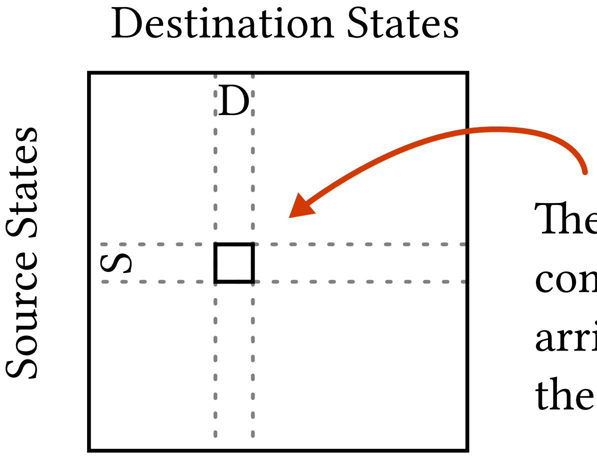
The calculation for joint probabilities differs. We need to inject admixture projection matrices when admixture events are modeled to occur.

Review:

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

$$M = \begin{cases} (P_0^t)_{\mathrm{BB}} \times (P_1^t)_{\mathrm{BB}} \times \cdots \times \left(P_{i-1}^t\right)_{\mathrm{BB}} \times (P_i^t)_{\mathrm{BE}} & \text{if } i = j \\ (P_0^t)_{\mathrm{BB}} \times \cdots \times \left(P_{i-1}^t\right)_{\mathrm{BB}} \times (P_i^t)_{\mathrm{BL}} \times \left(P_{i+1}^t\right)_{\mathrm{LL}} \times \cdots \times \left(P_j^t\right)_{\mathrm{LE}} & \text{if } i < j \end{cases}$$

Admixture Projection Matrix:



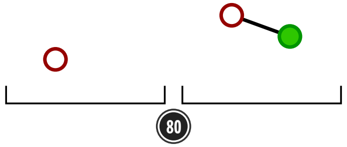
The probability of coming from S and arriving at D after the admixture event.

Admixture Projection Matrix

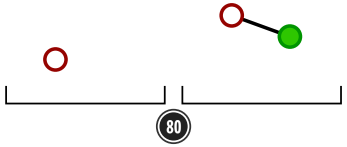
To do this, we first identify the number of pieces, *n*, in the source state. For each piece, we identify its location.

E.g.

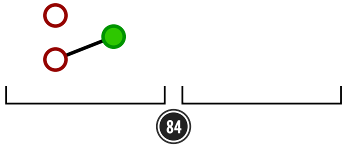
```
Source state:
     80. [1, ([1],[])] [2, ([2],[1,2])]
```



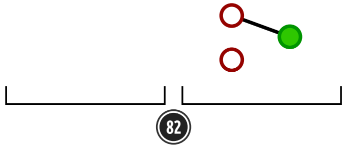
```
Destination states:
     80. [1, ([1],[])] [2, ([2],[1,2])]
```



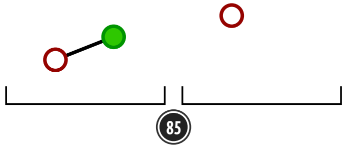
```
     84. [1, ([1],[])] [1, ([2],[1,2])]
```



```
     82. [2, ([1],[])] [2, ([2],]1,2])]
```



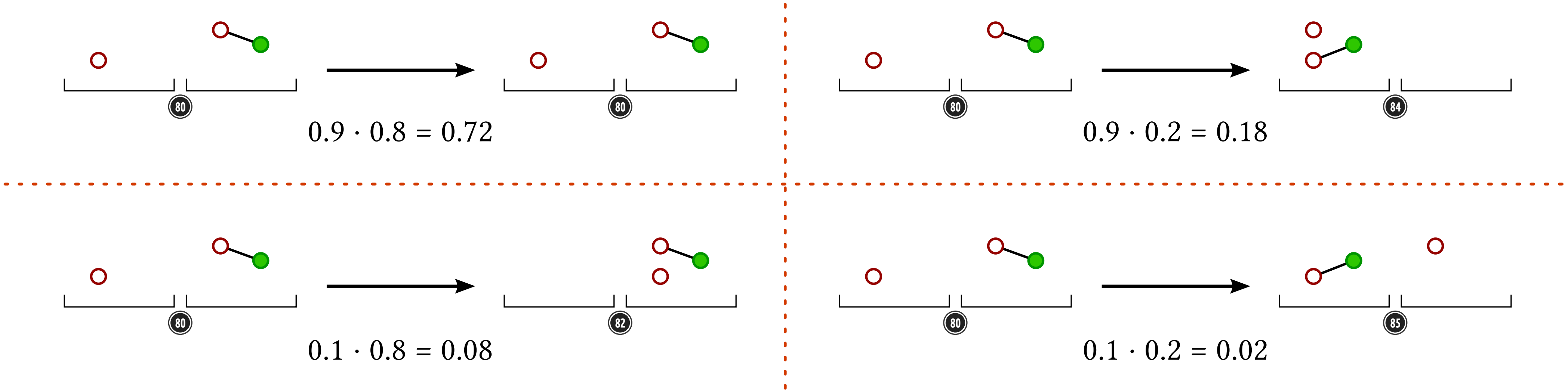
```
     85. [2, ([1],[])] [1, ([2],[1,2])]
```



Admixture Projection Matrix

During an admixture event, each piece may stay or move, so there are k^n destination states for a k-population demographic.

E.g. (continued) assume a probability of $p = 0.1$ for a sample to go from population 1 to 2 and a probability of $q = 0.2$ for a sample to go from population 2 to 1.



Admixture Projection Matrix

E.g. (continued) fill the corresponding row of the admixture project matrix.

```
...|. 80 ..... 82 ..... 84 ..... 85 .
---+---------------------------------
.  |  .        .        .        .
.  |  .        .        .        .
.  |  .        .        .        .
80 | 0.72     0.08     0.18     0.02
.  |  .        .        .        .
.  |  .        .        .        .
.  |  .        .        .        .
```

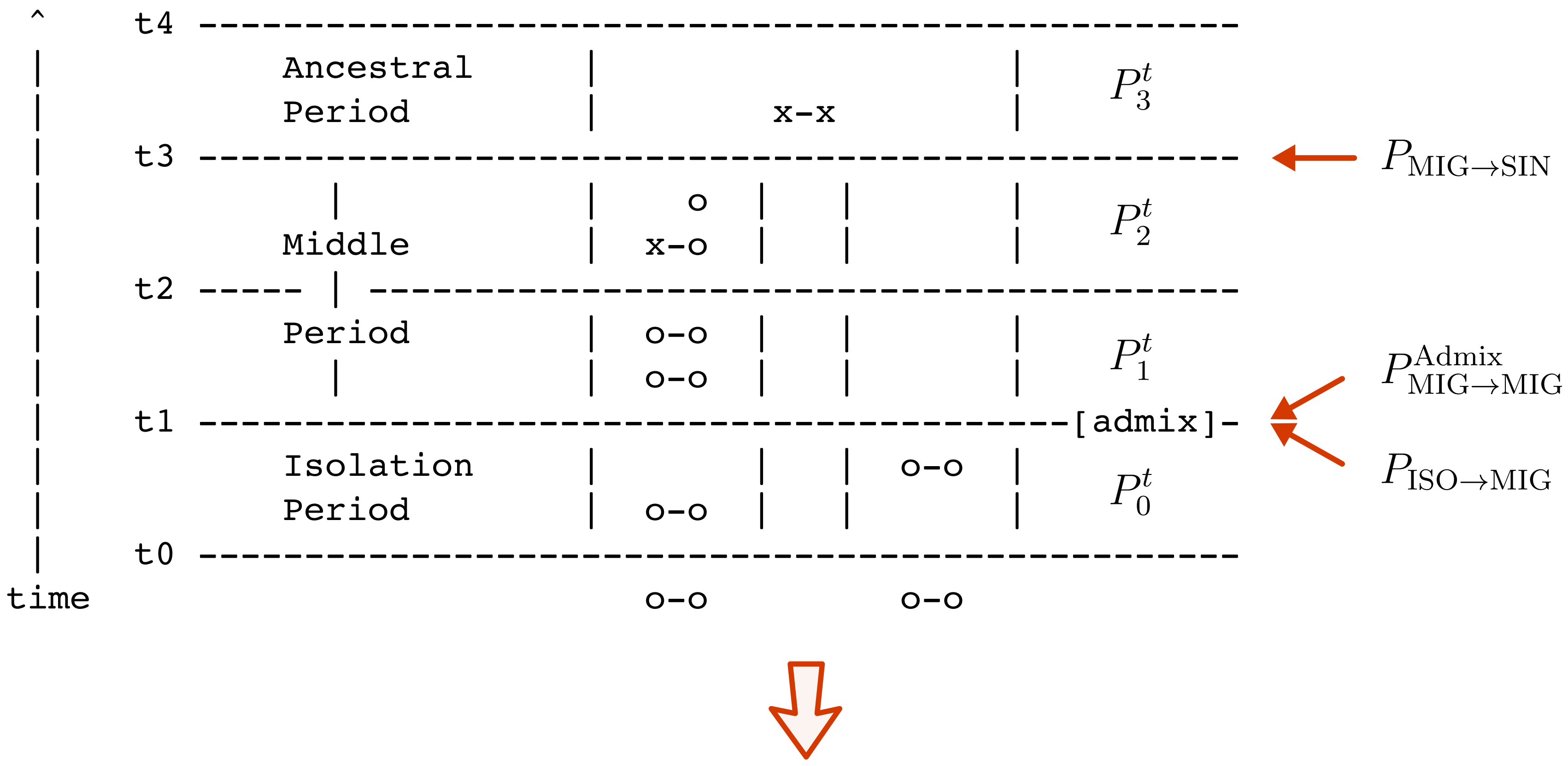
We programmatically determine whether or not a piece should move.

E.g. (continued)

decimal index	binary index	pieces
0d	00b	nobody moves
1d	01b	left piece stays; right piece moves
2d	10b	left piece moves; right piece stays
3d	11b	both pieces move

Admixture Projection Matrix

An example of how to compute the joint probability for admixture models.



```
 ^    t4 -------------------------------------------------
 |           Ancestral     |                   |
 |           Period        |        x-x        |
 |    t3 -------------------------------------------------
 |             |           |    o  |   |       |
 |           Middle        |  x-o  |   |       |
 |    t2 ----- | -----------------------------------------
 |           Period        |  o-o  |   |       |
 |             |           |  o-o  |   |       |
 |    t1 -----------------------------------------[admix]-
 |           Isolation     |       |   |  o-o  |
 |           Period        |  o-o  |   |       |
 |    t0 -------------------------------------------------
time                          o-o         o-o
```

P_3^t
$P_{\mathrm{MIG}\to\mathrm{SIN}}$
P_2^t
P_1^t
$P^{\mathrm{Admix}}_{\mathrm{MIG}\to\mathrm{MIG}}$
$P_{\mathrm{ISO}\to\mathrm{MIG}}$
P_0^t

$$J_{34} = \sum_{\alpha} \sum_{\beta} \left((P_0^t)_{\mathrm{BB}} \times (P_{\mathrm{ISO}\to\mathrm{MIG}})_{\mathrm{BB}} \times \left(P^{\mathrm{Admix}}_{\mathrm{MIG}\to\mathrm{MIG}}\right)_{\mathrm{BB}} \times (P_1^t)_{\mathrm{BB}} \times (P_2^t)_{\mathrm{BL}} \times (P_{\mathrm{MIG}\to\mathrm{SIN}})_{\mathrm{LL}} \times (P_3^t)_{\mathrm{LE}} \right)_{\alpha\beta}$$

Presentation Outline

CoalHMM framework

Continuous time Markov chain (CTMC)

Hidden Markov model (HMM)

Numerical optimizations

CoalHMM with simulations

Simulation validation

Performance evaluation with simple to complex models

CoalHMM with biological data

Data validation with various analyses

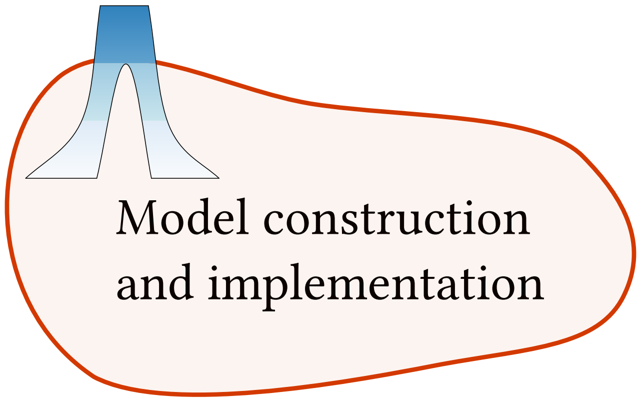
CoalHMM inference with Bears

Admixture CoalHMM

General model construction

Three-population admixture model

Bear study case



Model construction and implementation

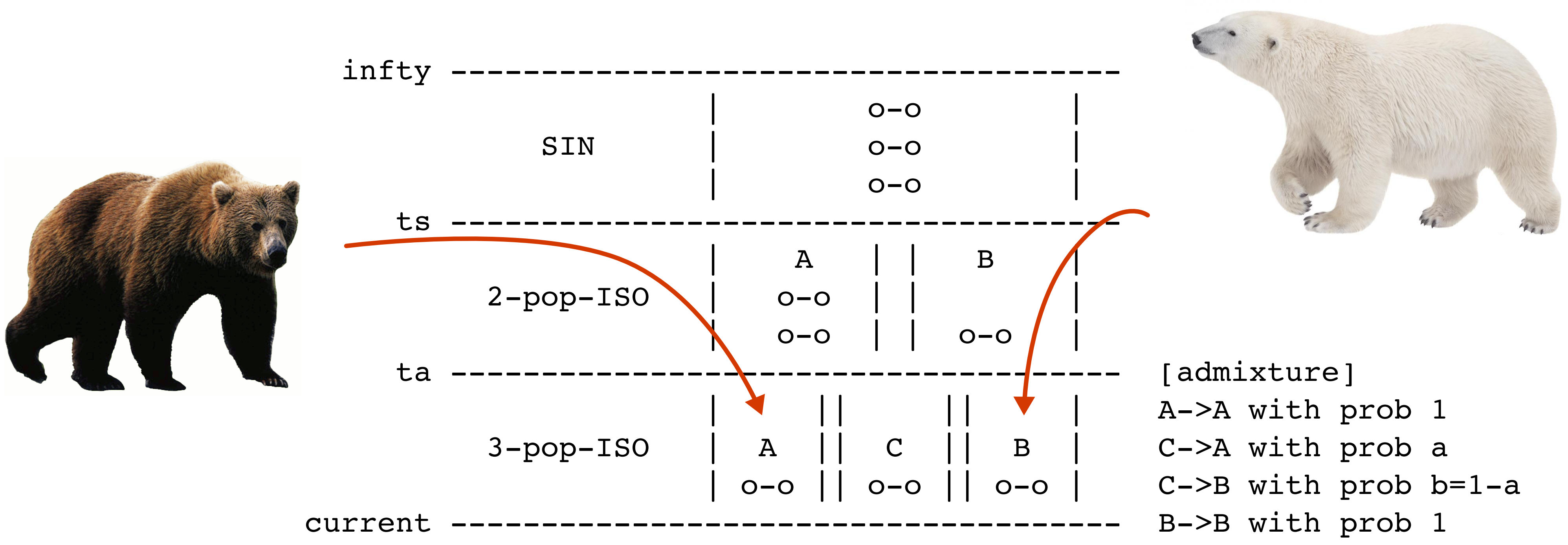
Simulation case study

Biological case study

Three-population Admixture Model

The ABC island bears are known to be an admixed population originated from brown bears and polar bears.

We construct a three-population admixture model. Two source populations, A and B, are not admixed. C is the admixed population.

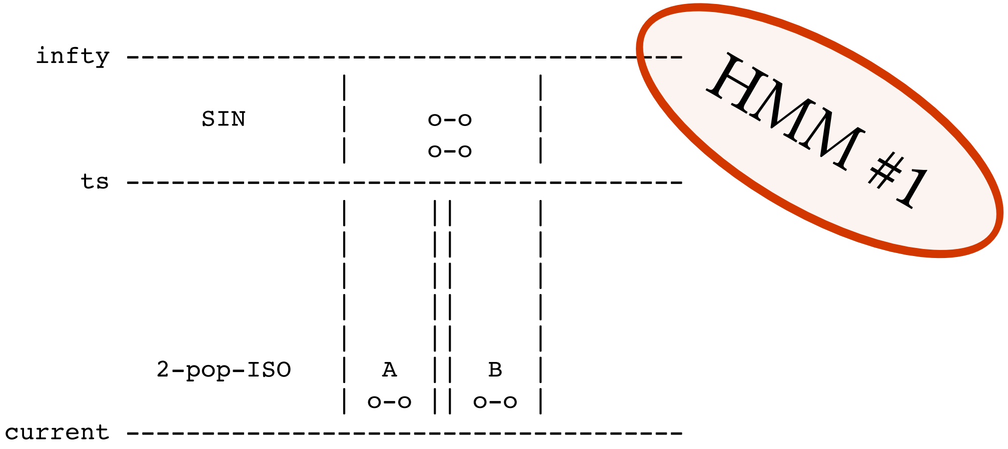


```
  infty ------------------------------------
                     |        o-o       |
            SIN      |        o-o       |
                     |        o-o       |
     ts ------------------------------------
                     |    A   | |   B   |
          2-pop-ISO  |   o-o  | |       |
                     |   o-o  | |  o-o  |
     ta ------------------------------------ [admixture]
                     |     ||     ||    |    A->A with prob 1
          3-pop-ISO  |  A  ||  C  || B  |    C->A with prob a
                     | o-o || o-o ||o-o |    C->B with prob b=1-a
current ------------------------------------ B->B with prob 1
```

Composite Likelihoods

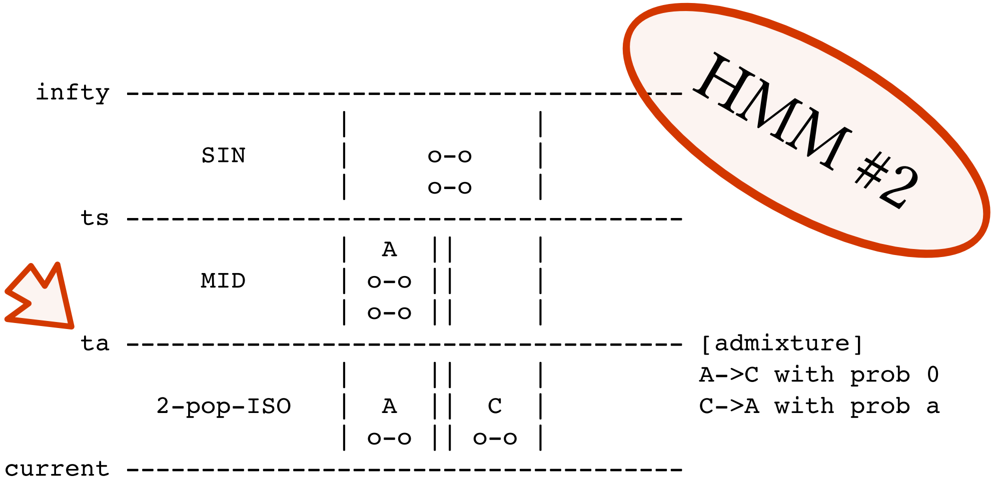
We apply the composite likelihood approach. We build one HMM for A and C, one HMM for B and C, and one HMM for A and B.

Two HMMs deal with admixing events. One models a straightforward Isolation demographic. We then optimize over the sum of the log likelihood values.



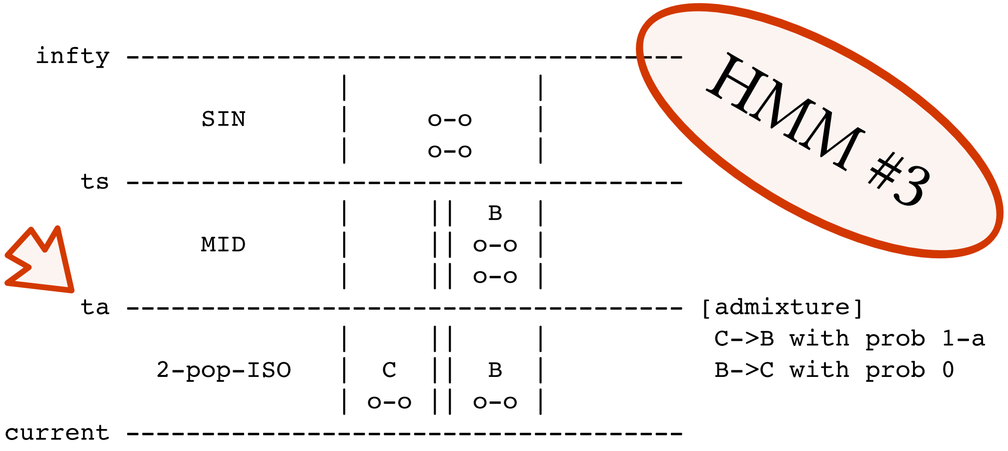
```
  infty -------------------------------------
                      |            |
             SIN      |     o-o    |
                      |     o-o    |
     ts -------------------------------------
                      |     ||     |
                      |     ||     |
                      |     ||     |
                      |     ||     |
                      |     ||     |
          2-pop-ISO   |  A  ||  B  |
                      | o-o || o-o |
current -------------------------------------
```

HMM #1



```
  infty -------------------------------------
                      |            |
             SIN      |     o-o    |
                      |     o-o    |
     ts -------------------------------------
                      |  A  ||     |
             MID      | o-o ||     |
                      | o-o ||     |
     ta ------------------------------------- [admixture]
                      |     ||     |          A->C with prob 0
          2-pop-ISO   |  A  ||  C  |          C->A with prob a
                      | o-o || o-o |
current -------------------------------------
```

HMM #2

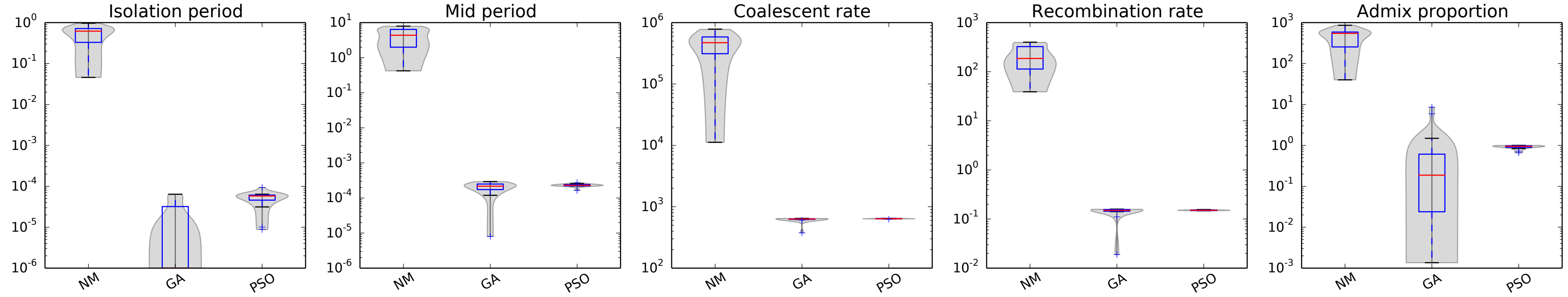


```
  infty -------------------------------------
                      |            |
             SIN      |     o-o    |
                      |     o-o    |
     ts -------------------------------------
                      |     ||  B  |
             MID      |     || o-o |
                      |     || o-o |
     ta ------------------------------------- [admixture]
                      |     ||     |           C->B with prob 1-a
          2-pop-ISO   |  C  ||  B  |           B->C with prob 0
                      | o-o || o-o |
current -------------------------------------
```

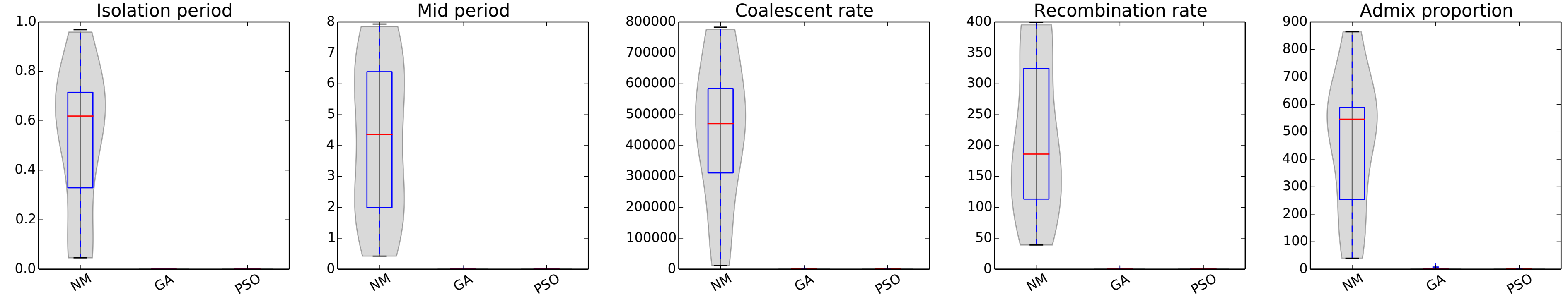
HMM #3

Bear Admixture CoalHMM

Nelder-Mead fails completely. GA and PSO both managed to improved the fitness values over generations. PSO reached very good convergence.



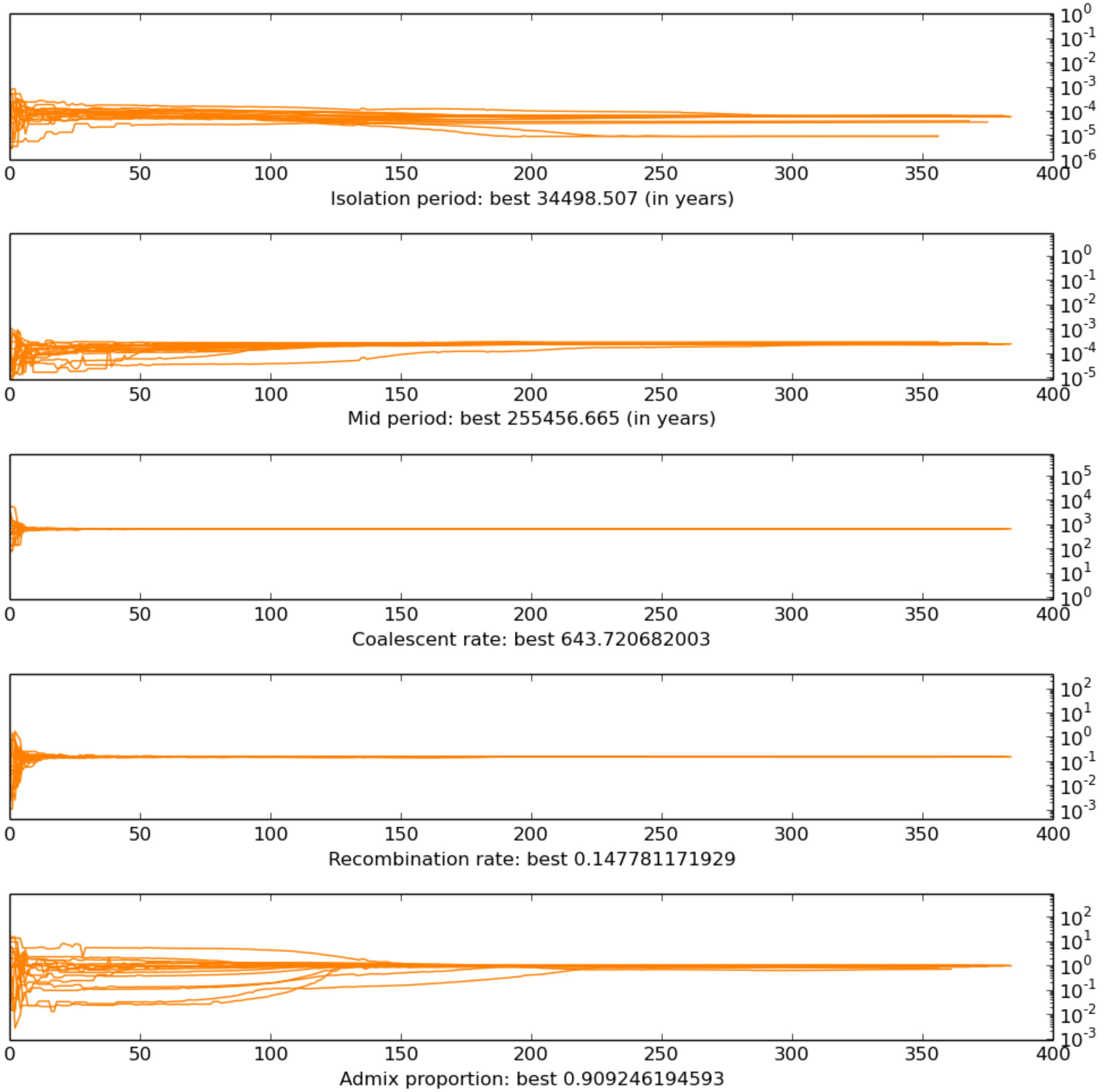
Log scale

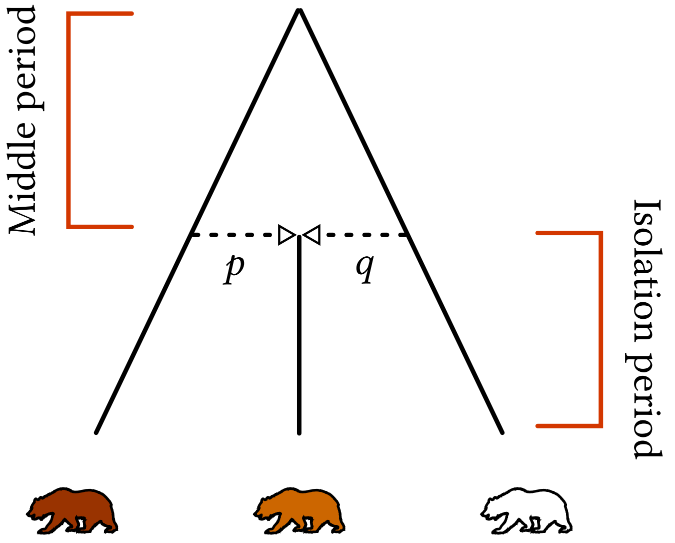


Linear scale

Admixture CoalHMM Inference

PSO executions over time.





Parameter	Estimate
Isolation period	(years) ~34,000
Middle period	(years) ~260,000
Coalescent rate	~640
Recombination rate	~0.15
Admix proportion BB to ABC	~0.91
Admix proportion PB to ABC	~0.09

Acknowledgement

Thanks!