A coalescent hidden Markov model for inferring admixture relationships

Jade Yu Cheng[1], Thomas Mailund[1*]

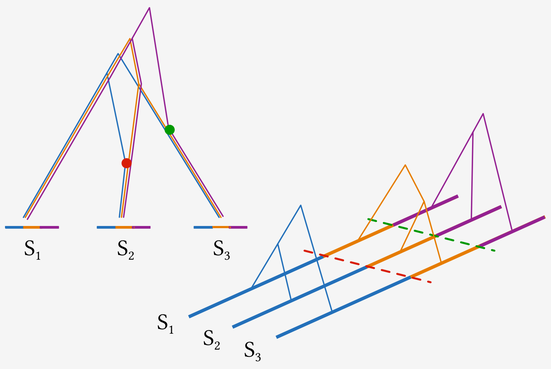
1. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark * Email: mailund@birc.au.dk

Introduction

With full genome data from several closely related species now readily available, we have the ultimate data for demographic inference. Exploiting these full genomes, however, requires models that can explicitly model recombination along alignments of full chromosomal length. Over the last decade a class of models, based on the sequential Markov coalescence combined with hidden Markov models, has been developed and used to make inference in simple demographic scenarios.

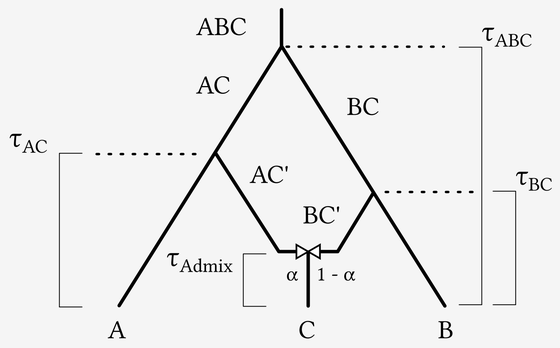
In this project, we develop a coalescence hidden Markov model for inferring parameters for admixture events. By tracing lineages in an admixed population and its source populations back in time and estimating the coalescence times of those lineages, we can infer the split time between the source populations, the time of admixture, and the admixture proportions.

Admixture CoalHMM

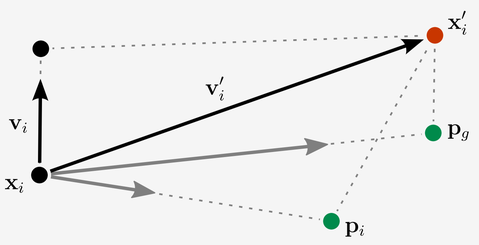


Coalescence theory describes the ancestry of a sample of present day genes and gives probabilities to all of the possible genealogies that could have created the variation seen in the samples. To the left, we demonstrate an ancestral recombination graph over three sequences, S_1, S_2, and S_3. The example shows the ancestry in the case where they have experienced two recombination events, shown in read and green. These recombinations segment the sequences into three regions, shown in blue, orange, and purple, each with different three genealogies.

A general admixture scenario involves extant populations A, B, C, and ancestral populations AC', BC', AC, BC, ABC. Population C is admixed from AC' and BC' which are related to A and B, respectively. Population AC is ancestral to A and AC'. Population BC is ancestral to B and BC'. Population ABC is ancestral to AC and BC. Our method infers the split times of A-AC', B-BC', and AC-BC, as well as the admixture time and the admixture proportions, which is α from C to AC' and 1-α from C to BC'.



Optimization



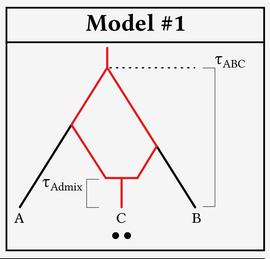
We use the Particle Swarm Optimization (Eberhart and Kennedy, 1995). It is a heuristic based search algorithm. In each iteration, three vectors are applied to a particle at position $\mathbf{x}_i$. A cognitive influence urges the particle towrad its previous best $\mathbf{p}_i$, a social influence urges the particle toward the swarm's previous best $\mathbf{p}_g$, and its own velocity $\mathbf{v}_i$ provides inertia, allowing it to overshoot local minima and explore unknow regions of the problem domain.

Composite Likelihood

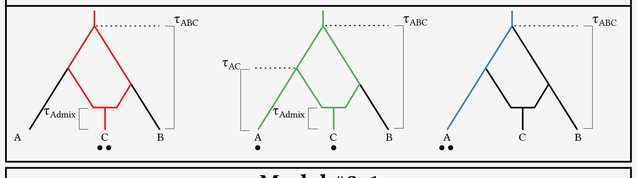
We apply the composite likelihood approach to deal with more than two samples. The more samples we have from each population the more HMMs we can construct and incorporate into our admixture CoalHMM models. We implemented and tested through simulations a range of models varying the availability of extant populations and samples per population.

Model \ Population	A	B	C	Samples per population	Note
#1	×	×	✓	●●	Missing both source populations
#2	✓	×	✓	●●	Missing a source population
#3-1	✓	✓	✓	●	One sample per population, 3 HMMs
#3-2	✓	✓	✓	●●	One pair per each configuration, 6 HMMs
#3-3	✓	✓	✓	★●	All pairwise, 15 HMMs

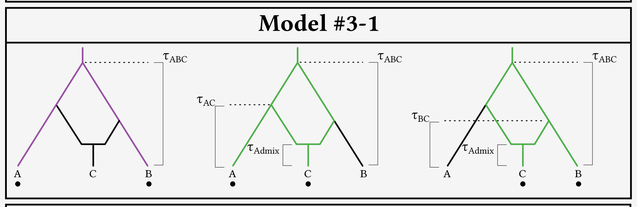
Model #1



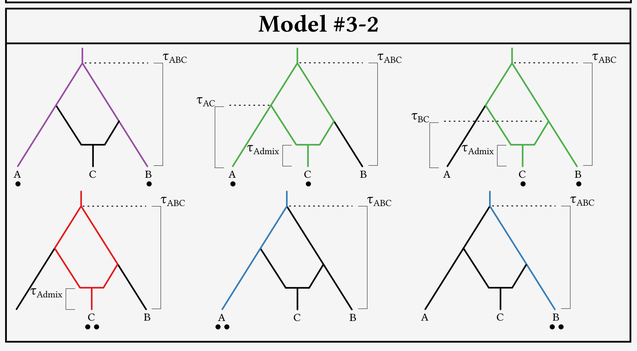
Model #2



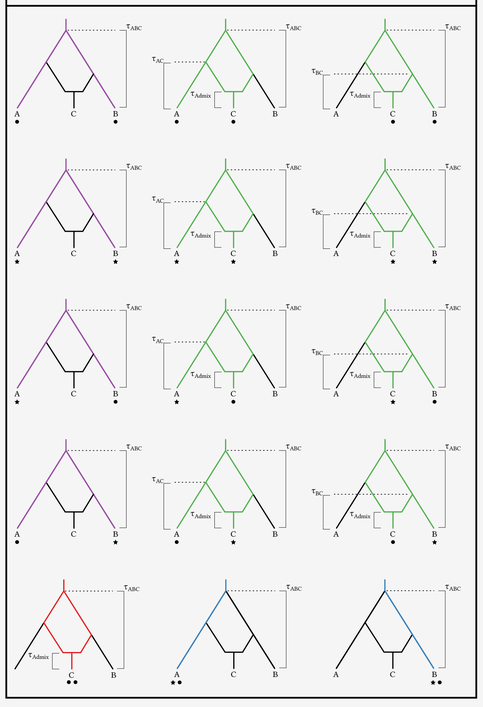
Model #3-1



Model #3-2



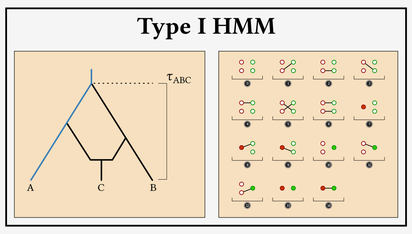
Model #3-3



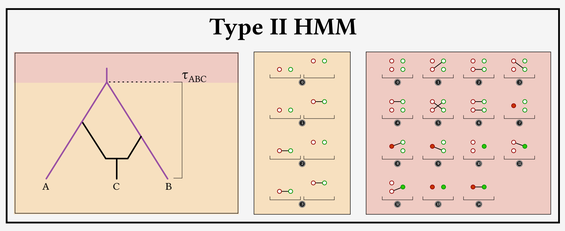
Continuous Time Markov Chain

For the three extant populations, we have four different pairwise configurations, two samples both from one of the two source populations, two samples both from the admixed population, two samples one from each of the two source populations, and two samples one from the admixed population the other from a source population. Each configuration requires a different sequence of CTMCs, and each CTMC has its own state spaces.

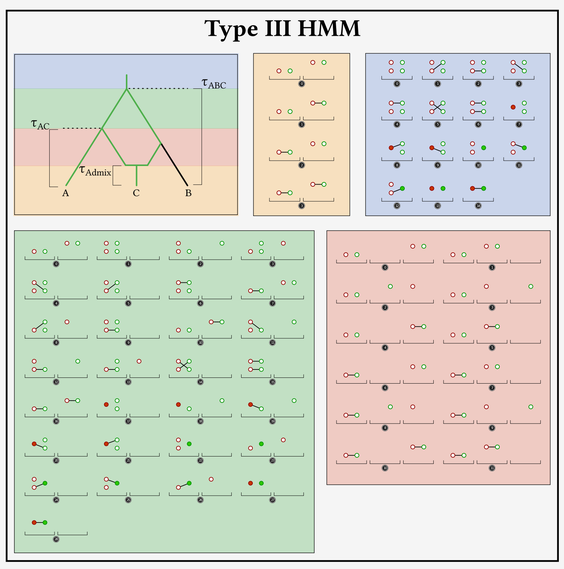
Type I HMM



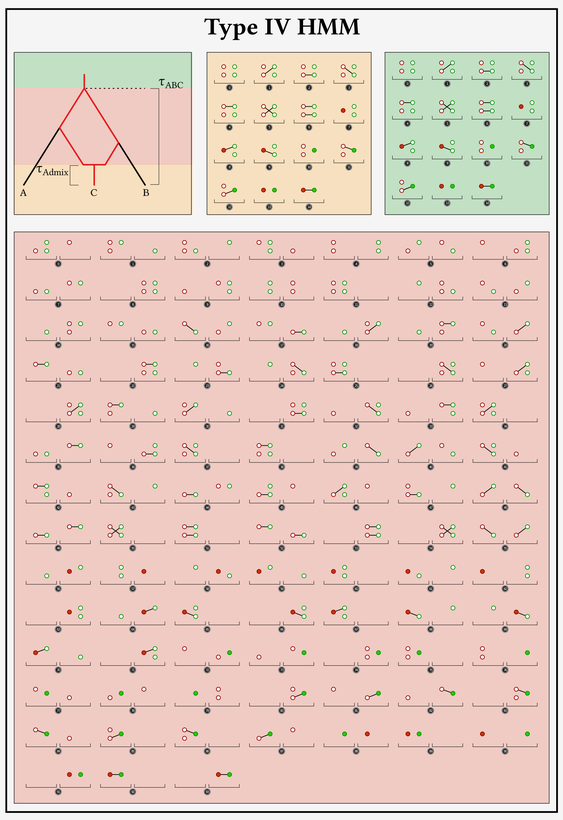
Type II HMM



Type III HMM

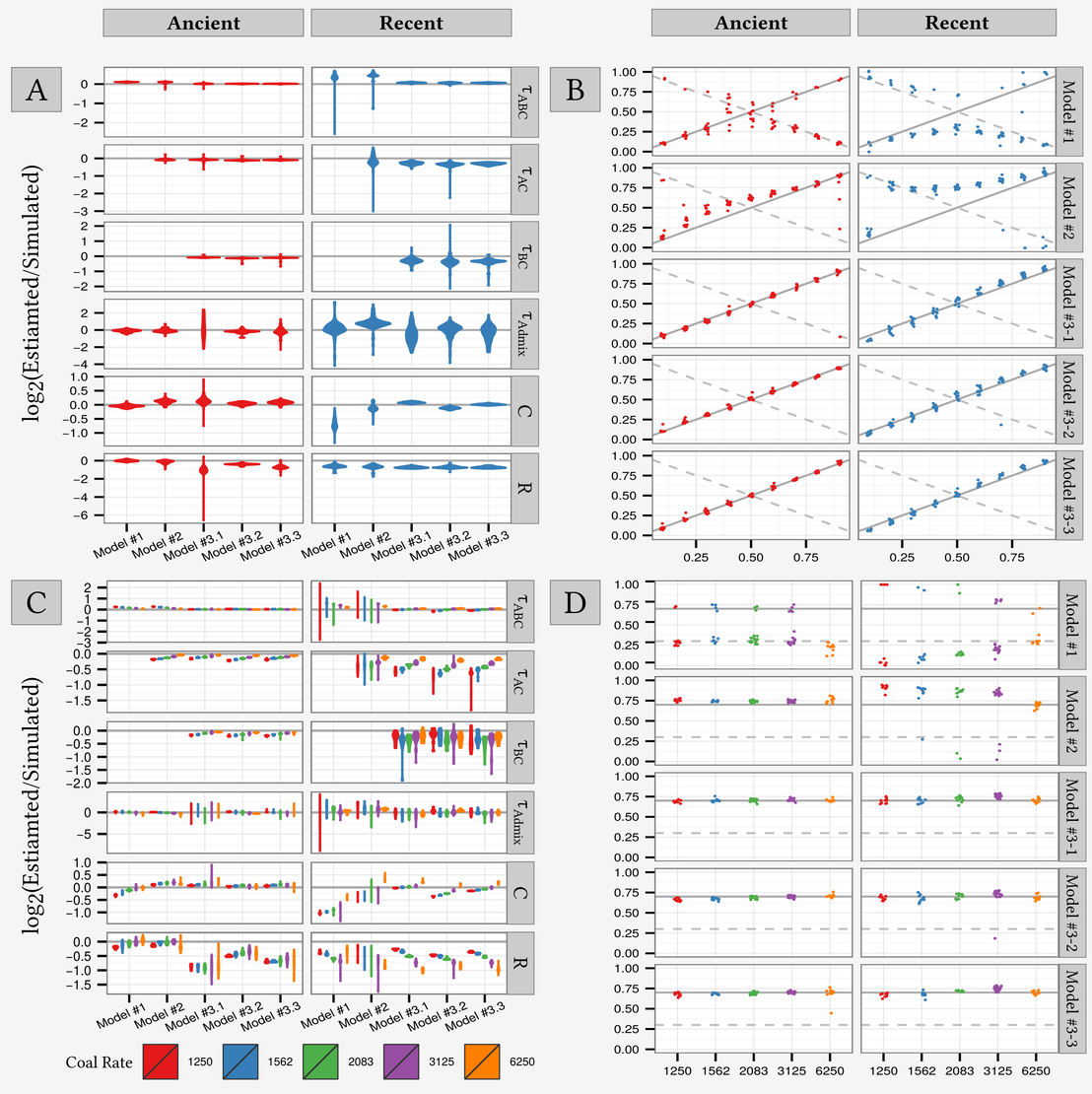


Type IV HMM



Simulation Study

We simulated data using the program *fastsimcoal2* (Excoffier, 2013). We tested two scenarios, differing in the time since the population divergence and admixture events: an "Ancient" scenario and a "Recent" scenario. Plots A and C show the accuracy of parameter estimation for time parameters, the coalescent rate and the recombination rate. Plots B and D show the accuracy of estimation of admixture proportions.



Conclusion

We have developed a coalescent hidden Markov model that enables us to estimate demographic parameters in the scenarios where one population is descendant from an admixture event between two ancestral populations, and we may or many not have samples from populations related to one or both of the source populations. Through simulations, we have shown that we recover most parameters well. The main focus of the admixture CoalHMM models is the admixture event and, hence, parameters related the admixture. Our full models successfully recovered all admixture proportions with good accuracy.

https://github.com/jade-cheng/Jocx

Acknowledgements

This research was funded by the Danish Council of Independent Research Sapere Aude grant.