Demographic Inference with CoalHMM and Heuristic Optimisations

Jade Cheng · Thomas Mailund

Introduction

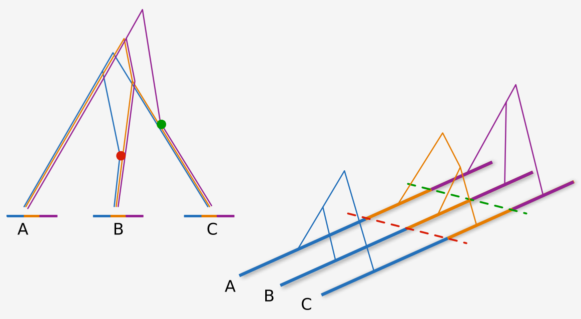
With full genome data from several closely related species now readily available, we have the ultimate data for demographic inference. Exploiting these full genomes, however, requires models that can explicitly model recombination along alignments of full chromosomal length. Over the last decade a class of models, based on the sequential Markov coalescence model combined with hidden Markov models, has been developed and used to make inference in simple demographic scenarios. To move forward to more complex demographic modelling we need better and more automated ways of specifying these models and efficient optimisation algorithms for inferring the parameters in complex and often high-dimensional models.

CoalHMM

In recent years a number of inference tools have been developed based on combining the sequential Markov coalescence with hidden Markov models, constructing so-called coalescence hidden Markov models or CoalHMMs, that have been constructed for the inference of speciation times, gene-flow patterns, changing population sizes or inference of recombination patterns and have been used in a number of whole genome analyses. Below we describe the coalescence process and an example of one of these models.

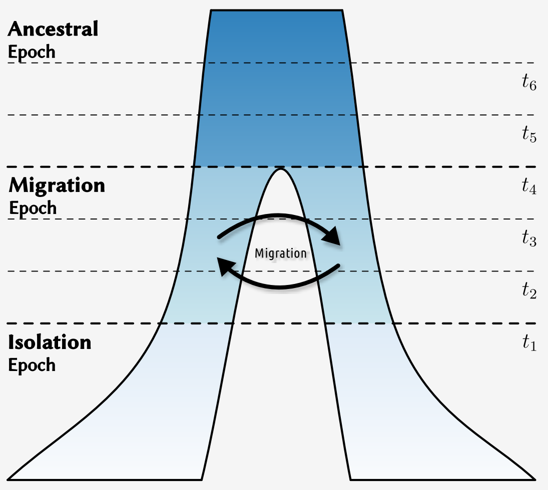
Coalescence Process

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



Demographic Inference

The demographic isolation-with-initial-migration model has five parameters and three epochs: an **ancestral** population epoch with one population and free coalescences, a **migration** epoch with two populations where lineages can only coalesce within the same population but can migrate between the populations, and an **isolation** epoch where the two populations are completely independent. The parameters are the time points where the system switches between the epochs, the coalscence and recombination rates (assumed to be the same in all populations) and a symmetric migration rate during the migration epoch. The time point t_1, t_2, ..., t_6 illustrates a possible discetisation of time into the intervals that becomes. the states of the hidden Markov model.

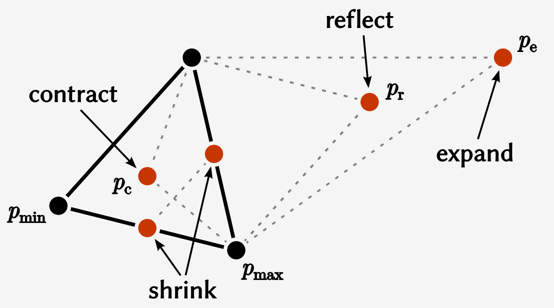


Numerical Optimisation

We can build more complex demographic models than previous frameworks and obtain more accurate parameter estimates using heuristic optimisation algorithms than when using gradient based approaches.

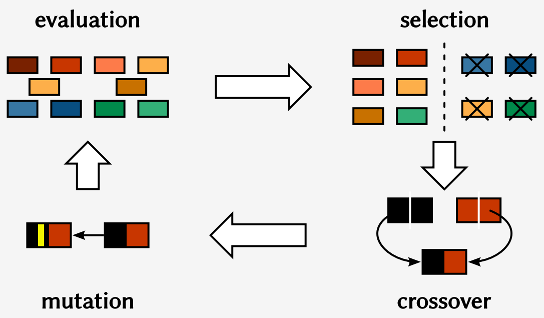
Nelder-Mead

Nelder-Mead optimization minimises an objective function in a many-dimensional space by continuously refining a simplex. In an iteration of the Nelder-Mead method over two-dimensional space, a point p_{min} is reflected to point p_r, expanded to point p_e, or contracted to point p_c. If these test points do not improve the overall score of the simplex, then it shrinks around the point p_{max} with the highest score.



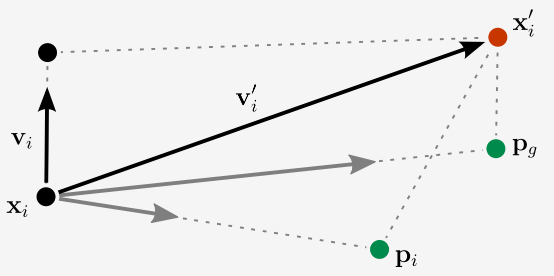
Genetic Algorithms

A Genetic Algorithm is a type of evolutionary algorithm that operates by encoding potential solutions as simple chromosome-like data structures and then applying genetic alterations. In one iteration of its evolution, a genetic algorithm operates in three stages: **Selection**, where it chooses a relatively fit subset of individuals for breeding; **Crossover**, where it recombines pairs of breeders to create a new population; and **Mutation**, where it potentially modifies portions of new chromosomes to help maintain the overall genetic diversity.



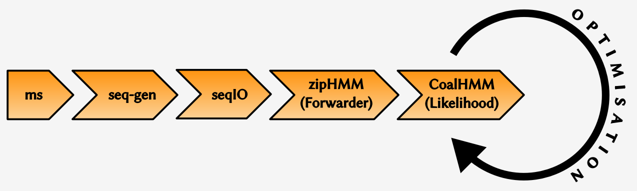
Particle Swarm Optimisation

Particle Swarm Optimization 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 toward 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 unknown regions of the problem domain.



Pipeline

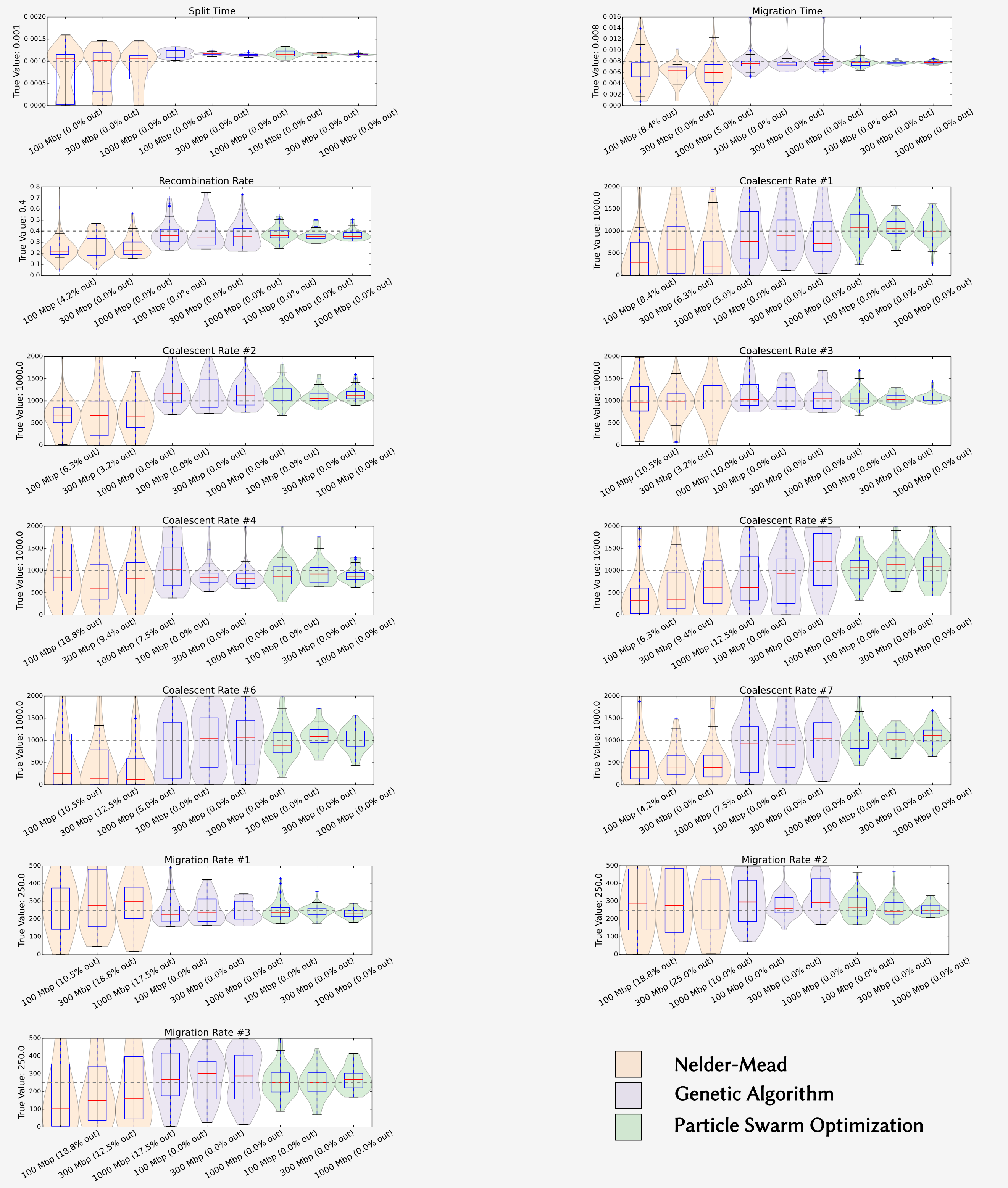
The program **ms** generates ancestral recombination graphs, and **seq-gen** produces sequence samples with 10 Mbp. Our framework imports alignments using **seqIO**, prepares **Forwarder** tables for HMM evaluation, and calculates likelihoods using models from **CoalHMM**. The optimisers use these likelihood values as fitnesses as many times as necessary.



Results

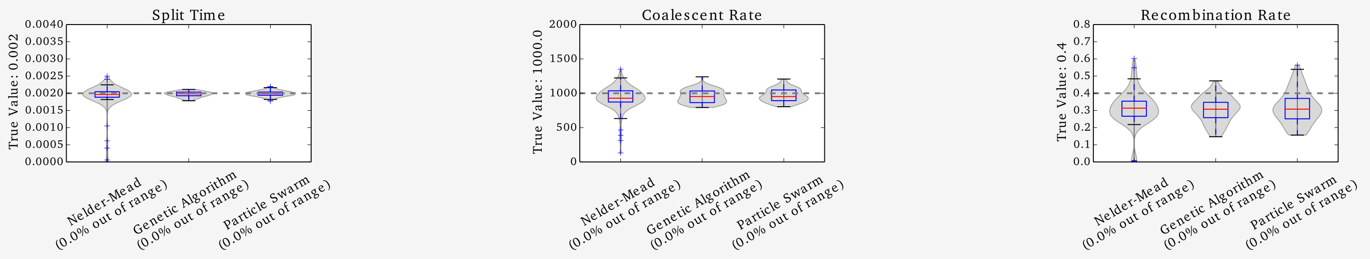
Isolation-with-Initial-Migration-Epochs Model

For complex demographic models, the accuracy and quality of the estimations diminish. With the given conditions, Particle Swarm Optimisation achieves the best inference. This experiment also demonstrates how inference results improve with data volume.



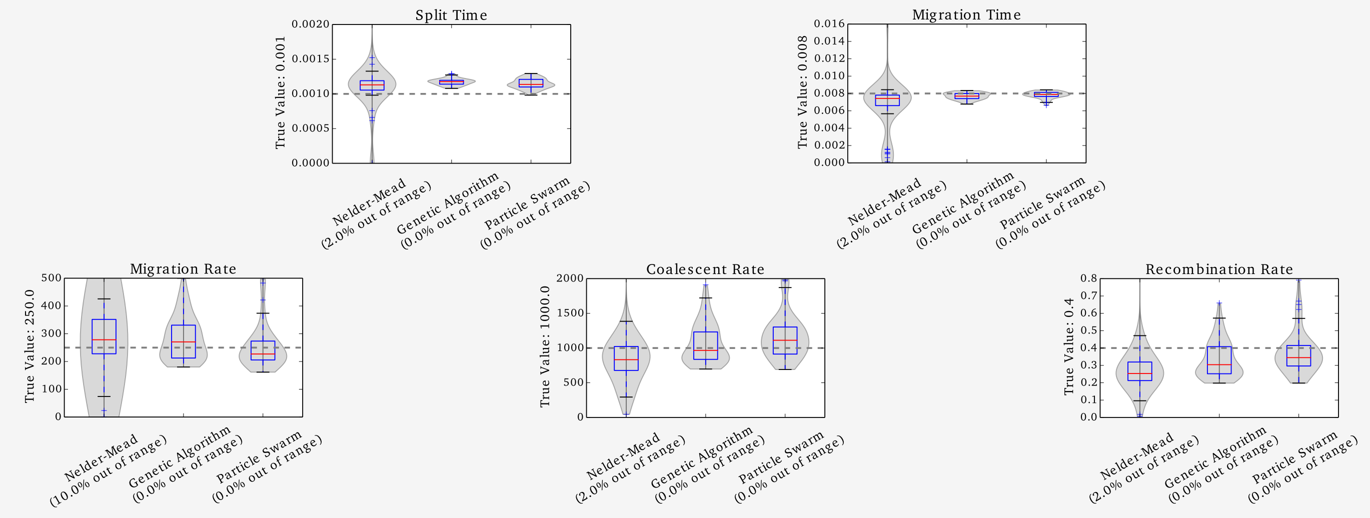
Isolation Model

For this simplest model, all three optimisers recover the simulated parameters, shown as dashed horizontal lines, reasonably well but with a higher variance for the Nelder-Mead optimiser.



Isolation-with-Initial-Migration Model

With more parameters to estimate, the variance in the estimates goes up as expected, but the parameters are still reasonably well estimated for the two heuristic optimisers but less so for the Nelder-Mead optimiser.



Conclusions

We have described a new framework for constructing coalescence hidden Markov models for demographic inference and showed that using heuristic optimisation algorithms we can accurately estimate parameters in a number of complex models. The framework is available under open source license GPLv2 at:

https://github.com/mailund/IMCoalHMM

Acknowledgements

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

BiRC