IMPLEMENTING THE COPHYLOGENETIC LIKELIHOOD MODEL

Joshua Stretton (supervisor: Assoc. Prof. Michael Charleston)
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. Introduction and Motivation

Cophylogenetics

Cophylogenetics is the study of how the relationships between ecologically linked species unfold throughout history, and aims to model the interactions between these species in order to gain a better understanding of past evolutionary events.

Cophylogeny Reconstruction Problem

Given: two phylogenies H (hosts) and P (parasites), and a set of associations φ between their leaves; Figure 1.

Find: the mapping of P *into* H that shows how the ancestral parasites in P most likely evolved alongside the hosts in H; Figure 2.

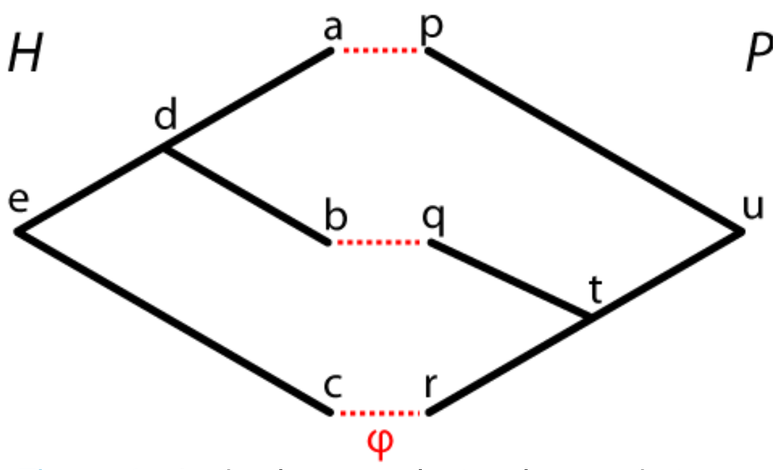

Figure 1: A simple example tanglegram between phylogenies H and P. The dashed red lines show the relationships observed between current species. Labels on vertices represent species names,

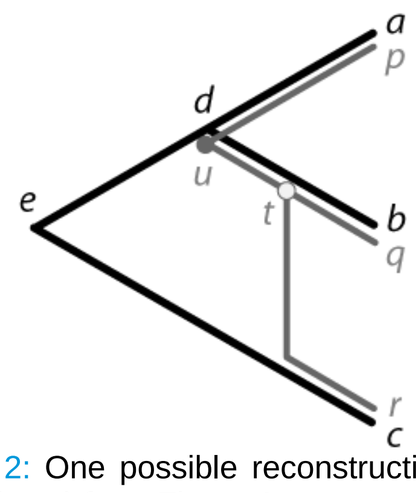

Figure 2: One possible reconstruction of H (black) and P (grey) from Figure 1.

NP-Completeness

The cophylogeny reconstruction problem is NP-complete [1], so to find an optimal solution efficiently, the problem space must be constrained.

A model to combine efficiency in calculations with biological accuracy is needed.

2. Cophylogeny Likelihood Model

Segments

Due to the continuous nature of time along a lineage in a phylogeny, some evolutionary events occur at certain *rates*, so finding the likelihood of any given reconstruction is mathematically complex.

Dividing the time scale into discrete segments allows the use of event *probabilities*, and overall likelihood can be calculated simply by multiplying the probabilities over the whole reconstruction.

Figure 3 shows how the reconstruction given in Figure 2 can be divided into segments, where there can be a maximum of one evolutionary event per lineage per segment.

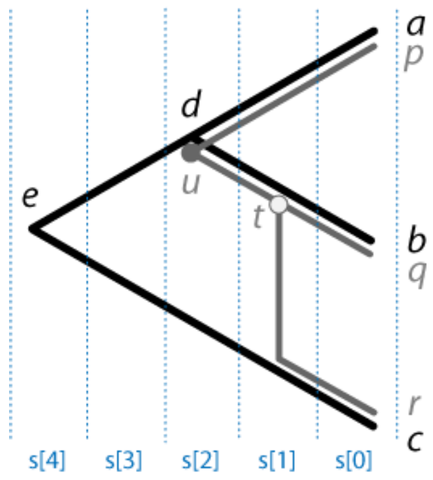

Figure 3: Dividing the reconstruction of Figure 2 into segments.

Likelihood Calculation

Modern methods of reconstruction calculation use events and event costs in order to find a good reconstruction.

Each kind of event is given a cost, and then these costs are combined to find the overall cost of a reconstruction.

We prove that the Cophylogeny Likelihood Model (CLM) [3] allows events to be determined simply, and that the costs (specifically the probabilities) of each event are optimally set as the ratio of the number of times an event occurred to the number of times it could have occurred in a given reconstruction.

This removes the need for a complex cost-optimising algorithm, such as presented in Merkle and Middendorf's paper [2].

Markov Chain Monte Carlo (MCMC)

Using this model, the likelihood of a given reconstruction can be determined quickly, and so many reconstructions can be examined in a short time in order to find the *most likely* one.

An MCMC algorithm allows the fast traversal of the reconstruction space, provided some method for moving between reconstructions is provided; Figure 4.

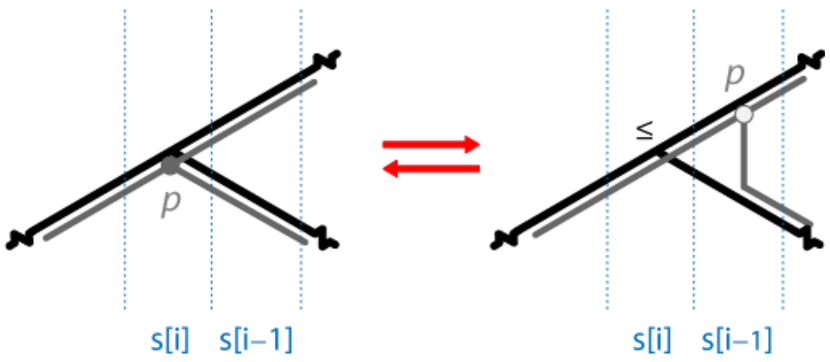

Figure 4: An example of a simple change to a reconstruction which results in a new possible history to examine.

If enough of these small perturbations are examined, *theoretically* every possible reconstruction can be examined, and the most likely one presented.

The MCMC algorithm proceeds to a new state if it presents a higher likelihood, as well as accepting a lower likelihood with some small probability, and so the reconstructions with the highest likelihoods will be revisited many times.

The reconstruction that the MCMC spends the most time on is considered the most likely reconstruction.

3. The Algorithm

The following is the broad-level algorithm used to calculate a solution to the cophylogeny reconstruction problem using the CLM:

```
function MaxLikelihoodReconstruction(H, P, φ, itLimit)
    h ← Reconcile(H, P, φ)
    C ← InitialCounts(h)
    ℓ_h ← Likelihood(h, C)
    states ← {}
    itCounter ← 0
    while itCount < itLimit do
        h_t ← Perturb(h)
        ℓ_t ← Likelihood(h_t, C)
        if ℓ_t ≥ ℓ_h OR Accept(h_t, h) then
            h ← h_t
            C ← UpdateCounts(C, h, ℓ_h, ℓ_t)
            ℓ_h ← ℓ_t
        end if
        itCount ← itCount + 1
        PromoteReconstruction(states, h)
    end while
    return MaxReconstruction(states)
end function
```

4. Improvements

How does the CLM improve results?

Many current approaches attempt to find an optimal reconstruction using methods such as dynamic programming or parsimony analysis.

These methods require simplifications of the problem space (such as disallowing host transfers, or ignoring divergence timing information).

The calculation of reconstruction likelihood in linear time allows more time to be spent on problem space complexities, such as multi-host parasites, and distance-based host transfer costs.

Is the time discretisation realistic?

The use of segments is justified by noting that (i) speciation does not take place instantaneously; (ii) estimates of divergence times are known to have wide confidence intervals; and (iii) if the time intervals were to become small enough, the problem would approach the continuous case.

5. Opportunities for Research

The simplifications provided by the CLM allow further work to focus on specific aspects of the biological world, as opposed to reconstruction calculating.

As an example, while this implementation does not allow multi-host parasites and failure to diverge events, these could be simply added to the CLM with little modification to the overall algorithm.

References

[1] Ovadia, Y., Fielder, D., Conow, C., and Libeskind-Hadas, R. "The cophylogeny reconstruction problem is np-complete". *Journal of Computational Biology 18* (2011), 59-65.

[2] Merkle, D., Middendorf, M., and Wieseke, N. "A parameter-adaptive dynamic programming approach for inferring cophylogenies". *BMC Bioinformatics 11* (2010).

[3] Charleston, M. A. "A new likelihood method for cophylogenetic analysis". Tech. rep.,
School of Information Technologies, The University of Sydney, 2009.