A NEW LIKELIHOOD METHOD FOR COPHYLOGENETIC ANALYSIS
TECHNICAL REPORT 636

MICHAEL A CHARLESTON
SCHOOL OF INFORMATION TECHNOLOGIES
THE UNIVERSITY OF SYDNEY

MARCH 2009
A new likelihood method for cophylogenetic analysis

Michael A. Charleston*
Technical Report #636
March 18, 2009

Abstract

I present a new method of calculating the likelihood in a cophylogenetic analysis. Given an instance of the cophylogeny problem, the cophylogeny likelihood model (CLM) relies on discretizing the time intervals in both phylogenies, such that each internal node of the host phylogeny $H$ exists in exactly one time interval and each host branch contains one or more time intervals. The approximation is justified by noting (i) that speciation takes place over a finite (non-instantaneous) time, (ii) that estimates of divergence times in phylogenetics generally have wide confidence intervals, and (iii) that given small enough time intervals the approach will converge upon the continuous case with arbitrary accuracy. The probability calculation is very fast as it involves a number of atomic calculations that grows approximately linearly with the size of the phylogenies involved. Through the use of Markov chain Monte Carlo it provides not only estimates of the most likely reconstruction, but also permits the direct estimation of model parameters by numerically integrating over all pertinent solutions, and is trivially extended to include uncertainty in the phylogenies $P$ and $H$ and other complexities such as reticulation and failure-to-diverge events.

1 Introduction

Cophylogeny is the study of the relationships among ecologically linked evolutionary trees (phylogenies) such as shown in Figure 1. These may be host and parasite phylogenies, or host / pathogen, species / gene, geographical area / endemic species (the biogeography problem) – even peoples and their languages can be considered in this paradigm. Simply, a phylogeny is a rooted, binary tree whose tips are labelled with extant species or taxa (in a more general setting, which can be accommodated by this approach, the phylogeny may be a directed acyclic graph: it just needs a root and a path to every leaf). Herein, $H$ will be the host phylogeny, and $P$ the parasite or pathogen phylogeny. All these systems involve an asymmetric comparison between an independent phylogeny of “host” taxa (e.g., primates) and one or more dependent phylogenies of pathogen or parasite taxa. I will use ‘pathogen’ and ‘host’ here to represent any such kinds of organisms; $H$ is the independent, and $P$ the dependent, phylogeny.

Given that there is in general no fossil record of parasitism or infection, cophylogeny rests on knowledge of what is current – observed associations and infections – and what can be inferred – the phylogenies of both groups of taxa.

Several methods have been tried to solve the cophylogeny problem, with varying success: parsimony methods that count coevolutionary events like Brooks’ Parsimony Analysis [2,3] and Ronquist’s TreeFitter program [18], and distance-based methods such as Parafit [14], but these methods, while fast, cannot reconstruct histories of events and have uncertain statistical validity.

The most successful method of estimating ancient associations among ecologically linked taxonomic groups is cophylogeny mapping, where one or more “dependent” evolutionary trees are mapped into an “independent” one (Figure 3) [5,8,17,19]. The possible events that can be recovered are codivergence, duplication, host switching and loss. In a cophylogeny map the ancient relationships are immediately clear, and a score can be assigned to each cophylogenetic event.

*School of Information Technologies, University of Sydney, NSW 2006, Australia; e-mail: michael.charleston@usyd.edu.au; telephone: +61 (0)2 9351 4459
that is implied; the optimal maps have minimal cost. This mapping is computationally very complex \([7, 10, 11, 15, 16]\), particularly as the number of host switches increases \([5]\). A further complication is that there are no known general values for these four events, so all Pareto-optimal maps must be kept, any of which ‘could be’ optimal for some feasible set of event costs \([5]\).

The obvious solution is to develop a \textit{cophylogeny likelihood model} (CLM) for a map, assigning probabilities to discrete events of codivergence and host switching, duplication, extinction and taxon sampling failure for the parasite / pathogen tree, and therefore a total probability for a map, conditional on some priors. Huelsenbeck \textit{et al.} attempted this \([12, 13]\), but that method was limited because it only permitted one pathogen lineage per host, and thus led to over-estimation of the number of host switches that took place. That method was not developed further. A more recent model permits more than one pathogen / gene lineage per host / organism, but does not allow host switching, and so is limited to use in gene tree / species systems with no lateral transfer \([1]\).

These models are also limited because they do not take into account the undoubted interactions that exist between pathogen species: for example, recombination in HIV, Dengue, and many other viruses is known to occur, but so far there is no probabilistic model of this process. Further, recombination leads to non-treelike phylogenies, and although non-treelike phylogenies can theoretically be included in a cophylogenetic analysis \([6]\) there is no implementation of this method.

The majority of researchers are forced to use fast but statistically unsound parsimony-based reconstruction methods, or limit the sizes of their studies to include only a small number of species and deal with Pareto-optimality.

## 2 Methods

\textbf{The new approach:} I present an efficient, statistically robust method of uncovering the likelihood of such observations, given biologically realistic models. The new approach can be applied to existing host/parasite and host/pathogen systems, both to understand better their dynamics and to refine the models.
The calculation of the likelihood of a given mapping is in general made difficult by the existence of events that cannot be observed in the present, such as extinction events and sampling failure or host switching outside the current set of host lineages. Such unseen events would normally have to be integrated out in a continuous framework; however, within a model that includes these events explicitly we can in principle calculate the likelihood for all histories that are consistent with our observations (see Figures 2 and 3).

We can think of a (cophylogenetic) history as a collection of evolutionary events of both host and pathogen, and associations between them, where not all of the host or pathogens need survive until the present (or are simply not sampled). Figure 3 has four such histories. Those on the right have events that we cannot view; those on the left, with no such events, are called reconstructions, as they could be constructed by reference to the observed $P$, $H$ and associations $\varphi$. Both the histories on the right present the reconstructions on the left: that is, they are consistent with what we could possibly reconstruct given only the extant taxa and their phylogenies.

2.1 Definitions

**Definition 1** A phylogeny is a directed acyclic graph (DAG) with a single root vertex with in-degree zero, all of whose internal vertices have either in- or out-degree greater than one, but not both, and whose leaves (vertices of out-degree 0) are labelled uniquely.

We denote the leaves of a graph $G$ with $L(G)$, and the arcs with $A(G)$. An association is a pair $(p, \ell)$ where $p$ is an element of $V(P)$ and $\ell \subset V(H) \cup A(H)$. The set of vertices and arcs, $V(H) \cup A(H)$, is called the set of locations of $H$ and is denoted $\Gamma(H)$. In the simplest cases $\ell$ is always a singleton, that is, $p$ is associated with a single location in $H$. For the purposes of this report I constrain $\ell$ to be such singletons, though the principle findings here can be applied easily to cases where $\ell$ is a composite location, that is, $|\{\ell\}| > 1$. The children of vertex $v$ are denoted $\kappa(v)$ and the parent(s) are $\pi(v)$. For the present purpose we will deal with phylogenies that are rooted trees; thus $|\pi(v)| \leq 1$ with the strict inequality holding only for the root of a given DAG.

**Definition 2**

- A codivergence event occurs when internal vertices $p \in V(P)$ and $h \in V(H)$ are coincident, and the children of $p$ diversify on the children of $h$, and we prohibit cases of more than one child of $p$ being mapped to the same child of $h$.

- A duplication occurs when $p$ is associated with an arc of $H$ rather than a vertex; this corresponds to a speciation or divergence of $p$ that is independent of a divergence event in the host.

- A host switch occurs for some arc $(p, q) \in A(P)$ where the parent $p$ is associated with a location in $H$ that precedes, but is not ancestral to, the location in $H$ with which $q$ is associated. In my approach in order for a host switch to occur $p$ must duplicate and a successful invasion into a new host lineage must occur.

- A loss occurs as the result of one of three things, which given a problem instance (below) are indistinguishable: extinction of some $p$, failure to track both hosts after a host divergence event ("missing the boat") and simple failure to sample the pathogen $p$. These definitions are illustrated in Figure 3.
Figure 3: Cophylogenetic histories consistent with tanglegram (Fig 2)

| Key: | • codivergence; ◦ duplication; ≤ miss the boat; † extinction; ◊ sampling failure |

(a) (b)

(c) (d)
Problem instance: Typically we are presented with phylogenies $H$ and $P$ and a set $\varphi$ of associations between the leaves $L(P)$ of $P$ and the leaves $L(H)$ of $H$. The goal is to construct the most plausible explanation(s) for the similarities and differences between $P$ and $H$, based on what we know about the current distribution of pathogens $p$ on leaves of $H$. We will estimate the maximum likelihood reconstruction, given some model that includes a set of probabilities for all the cophylogenetic events that are possible in such a reconstruction.

**Definition 3** Given a host phylogeny $H$ and associate phylogeny $P$, and associations given by the mapping $\varphi : V(P) \rightarrow V(H) \cup A(H)$, a reconstruction is a collection $\Phi$ of associations with certain properties:

1. $\Phi$ extends $\varphi$, that is, $\Phi|_{L(P)} \equiv \varphi$ (“lifting” condition);
2. If $P \cong H$ and $\varphi$ preserves the isomorphism, then $\Phi$ preserves the isomorphism also (“consistency” condition) subject to timing constraints on both phylogenies;
3. $\Phi$ is an isomorphism of $P$ (“isomorphism” condition);
4. If $p \in V(P) \setminus L(P)$ is mapped by $\Phi$ to $\ell$ in $H$, then at least one child of $p$ must also be mapped by $\Phi$ to a descendant of $\ell$ (“traceable” condition);
5. If $p \in V(P) \setminus L(P)$ is mapped to $\ell \in V(H)$, corresponding to a codivergence event, then all the children of $p$ must be mapped to descendants of $\ell$ (“interpretable” condition).

(Adapted from [6].)

**Definition 4** A history $\mathcal{A}$ is a collection of associations of pathogen vertices and host locations that contains as a subgraph some $\Phi$. We say that $\mathcal{A}$ presents $\Phi$ if there is such a $\Phi$ that satisfies the conditions defined earlier. Note that every mapping $\Phi$ is a history.

The idea of the history is that we can calculate the probability explicitly of a single case where the observed reconstruction is what we really see. We can calculate the probability of all histories $\mathcal{A}_i$ that present a given reconstruction $\Phi$, and thereby arrive at an estimate of the likelihood that that reconstruction is correct. This can be achieved using Markov chain Monte Carlo simulation (MCMC).

### 2.2 Discretization

The cophylogeny reconstruction problem is known to be NP-complete [15] even for the case where the number of possible times of divergence events is restricted to one of two for each internal node; hence the number of maps in the more general case, where time is permitted to range over many values, precludes exact computation. The goal then is not to perfectly model the continuous process, but to approximate it efficiently with a discrete process, in such a way that with a sufficiently fine-grained discretization we can obtain sufficient accuracy.

With this approach we may convert branch lengths reflecting time from continuous to discrete variables; in that way we can use the probability of events occurring in a given time interval, rather than the rate at which they occur over that interval. In the limit as the intervals become smaller and smaller this approach would converge on having continuous values, but at significant computational expense: therefore we use a relatively coarse-grained discretization initially.

In the first instance the time intervals used can be the same over branches and edges of $H$ (Figure 3). This is simply justified by the observation that speciation is not an instantaneous process and lineages which are separated from each other by segregating environments or behaviour need some time to accumulate sufficient divergence to be considered as separate species. For trees with relatively long branch lengths this approach will still work as multiple time intervals can be used for longer branches; the requirement then is for longer sampling time in the MCMC.

Ideally one would integrate over all possibilities and conclude with a probability of a given reconstruction, but given that there are a super-exponential number of possible reconstructions (Little, in prep.), this is simply not practicable. Instead we can approximate this integral using MCMC. This allows us to effectively meander through solution space and calculate the likelihood
of each history, and thus the relative likelihood of each reconstruction that is consistent with that history. The histories in Figure 3a,b are consistent with the same reconstruction (requiring one codivergence event and one host switch), and those in Figure 3c,d are consistent with another reconstruction involving no host switches. Evaluating the probability of each history using this model leads to the likelihood of each such reconstruction being correct.

A simple example (Figure 3) shows the basics of the calculation.

The leaves of the host tree \((H, \text{heavy lines})\) and of the pathogen tree \((P, \text{light lines})\) are considered to be identified by unique labels, but the labels do not need to be shown here. At a host node any pathogens currently on it can either codiverge with probability \(p_c\) or “miss the boat” – undergo lineage sorting – with probability \((1 - p_c)\). We may also include include failure to diverge events at this point by considering a pathogen vertex to be initially associated below a diverging host vertex with both (all) the host’s children, though for clarity of illustration we do not do that here. At an edge of the host tree, pathogens may duplicate with probability \(p_d\), or go extinct with probability \(p_e\). Duplicating pathogens either switch to a randomly chosen contemporaneous host, with probability \(p_s\), or remain on the current host.

At the leaves of \(P\) each pathogen is sampled with probability \(p_s\). Note that the three processes of missing the boat, extinction and sampling all contribute to the observable event of “loss”; using just phylogenetic information it is not possible to distinguish which of these has happened [19], but with this new probabilistic approach it is possible to infer the relative likelihoods of each of them, for a given problem.

Note that the histories on the left and right sides correspond to precisely the same reconstructions, but with different probabilities. The two top solutions begin with a codivergence event at the root of \(P\) and the bottom two have a duplication event in \(P\) prior to the root of \(H\). All of them represent possible ways in which the tanglegram (Fig 2) could arise.

In the first figure, putting \(p_n := (1 - p_d - p_e)\), the probability of the history shown is (each factor corresponding to a time segment as labelled in the figure)

\[
Pr = p_c \times p_n p_d p_w / 2 \times p_s^3,
\]

in the second it is

\[
Pr = p_c \times p_d^2 p_w (1 - p_w) / 2 \times 2 p_s^3 (1 - p_s).
\]

In Figure 1c there are no host switches but there are “missing the boat” events, and

\[
Pr = p_d \times p_c (1 - p_c) \times p_n^3 \times p_n (1 - p_c)^2 / 4 \times p_n^3 \times p_s^3
\]

(taking into account symmetries in the history).

The last history is the most complex, with

\[
Pr = p_d \times p_c (1 - p_c) \times p_n^3 p_d p_w \times p_n^2 p_c (1 - p_c) / 2 \times p_d (1 - p_w) p_s p_n^3 / 2 \times p_s^3 (1 - p_s)^2 / 2.
\]

Calculating the likelihood of the reconstruction requires that we integrate over all the histories that correspond to that reconstruction.

In the example above we could insert estimates of \(p_c, p_d\) etc., and estimate using MCMC the likelihood of the reconstruction for those specific parameter values. Further, we could allow the parameters to vary and find an estimate of the likelihood of the reconstruction, thus leading to the first statistically robust estimate of the most probable evolutionary route of the pathogens. For instance if we put \(p_c = 0.9\), \(p_d = 0.3\), \(p_e = 0.1\), \(p_w = 0.5\) (conditional on there being a duplication event), and \(p_s = 0.8\), we find the probabilities come to 0.020736, 0.0020736, 9.67588 × 10⁻⁷ and 3.26517 × 10⁻⁸ respectively. The latter two histories are considerably more complex, contributing to their low probabilities, but there are many more such histories than simple ones, so the overall likelihood of the reconstruction consistent with Fig 3c,d is not obvious.

We can expand on the above procedure to estimate distributions of reconstructions, with their likelihoods, as one does when obtaining a confidence set of trees.
3 Discussion

It should be noted that any history that presents a given reconstruction can be considered to contribute to the likelihood of that reconstruction, but the more complex that history is, with more time intervals and/or more hidden events, the lower its contribution will be. It is thus to be expected that the MCMC will spend relatively little time with these more complex histories, and at first glance it appears the landscape of histories presenting a given reconstruction will be “well-behaved”, in the sense of being unimodal [4], due to the way the probabilities combine.

Further the method enables particular parameters to be estimated from given problem instances. For example if \( H \) and \( P \) are perfectly matched with large numbers of taxa, then a model that has high values of \( p_c \) and \( p_s \), and low values of \( p_d \), \( p_x \), will fit the observations much better than will a model in which \( p_c \) is low, \( p_d \) high, etc. Through numerical integration afforded by MCMC we can obtain a confidence interval of parameter values by integrating out the reconstructions themselves.

The CLM allows estimates of risk of zoonosis: One output will be the estimation of model parameters, including the probability of host switching. This can be uniform or extended to more complex models such as to accommodate phylogenetic distance among hosts [9].

can be extended to include competition, cross-immunity, and geographic effects: These effects can be built in to specific problems with negligible computational cost as the likelihood calculations will be discretized.

enables solution of much larger cophylogeny problems than were previously possible: Though the problem of finding the optimal map is NP-complete [15], this will be obviated by performing estimation using MCMC.

accommodates phylogenetic uncertainty: The perennial problem of cophylogeny mapping relying on the phylogenies of host and parasite/pathogen being correct will be avoided by using Bayesian methods to estimate these trees and building in their posterior probabilities into the cophylogenetic likelihood equations.

deals with hybridization events (reticulation) The model will not require either phylogeny to be a tree, as evinced in [6]. The problem of interpretation in ambiguous cases is also avoided as each reconstruction will have its own likelihood.

allows heuristic searching for optimal maps: The CLM will be very fast to calculate so will allow heuristic searching to find the most likely solution(s).

I have presented a method that can be used to calculate quickly the probability of any history that presents a given reconstruction of a cophylogeny problem. The method is extensible and straightforward to implement in the existing programs (TreeMAP3, development).

It is expected that this approach will become the standard method for performing statistical analysis in cophylogenetics.

References


