A likelihood method for cophylogenetics

Michael Charleston & Rob Little

School of IT
University of Sydney

Phylomania,
2009.10.29-30

Tracing history

I am, in point of fact, a particularly haughty and exclusive person, of pre-Adamite ancestral descent. You will understand this when I tell you that I can trace my ancestry back to a protoplasmic primordial atomic globule.

[The Mikado, Gilbert and Sullivan]

Cophylogenetics

Cophylogeny is the reconstruction of ancient relationships among ecologically linked groups of organisms, from their *phylogenetic information*.

Cophylogeny has broad applications throughout biological science.

Essentially the problem is that of how to relate phylogenies from ecologically related groups of organisms or habitats, given their (estimated) phylogenies and the known associations between the two.

Motivation

- An estimated 75% of emergent human diseases are *zoonoses*, that is, they switched hosts from other species into humans.
- If we can pinpoint episodes of codivergence (when two ecologically linked phylogenetic lineages split together) then we can *compare rates of evolution*.
- Understanding where an organism came from (e.g., invading pests) tells us how better to *combat* them.
- Discovery of linked divergent histories enables us to *find "missing" species* (e.g., New Zealand Beech trees *Nothofagus*).
- Possibly undiscovered *pathogens*?

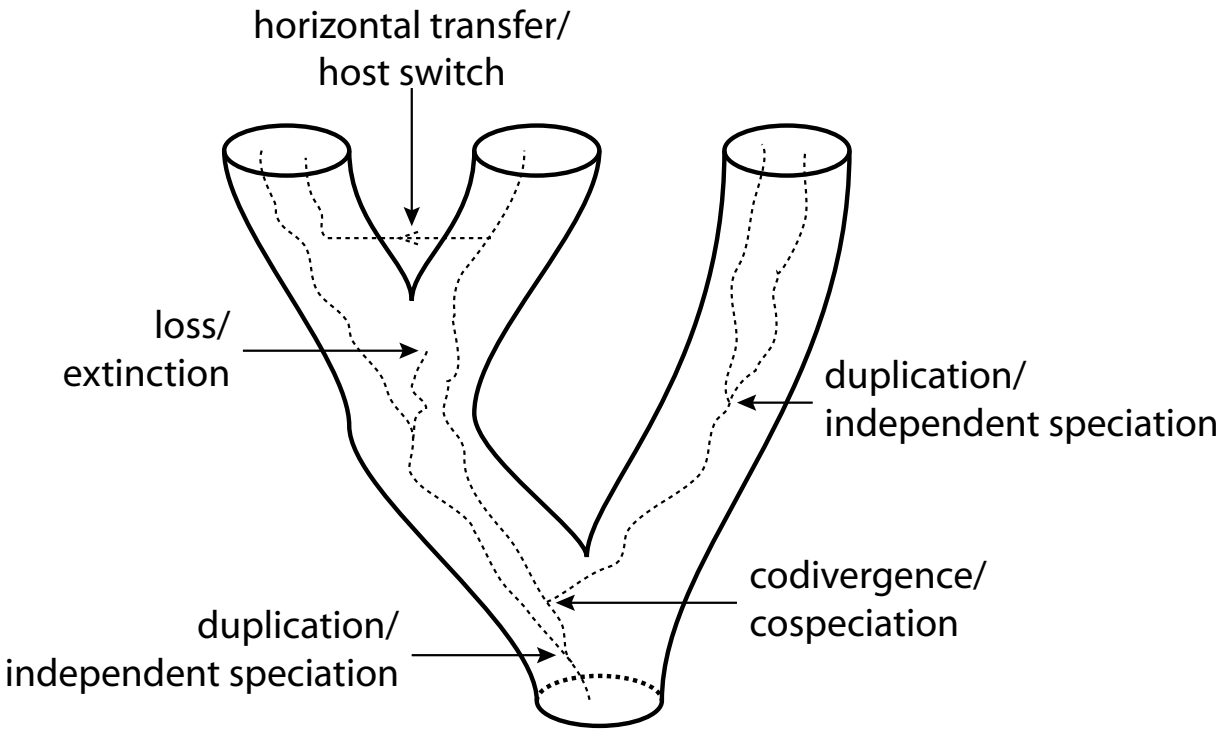
. . . and anyway,

- It's an interesting problem
- It's hard
- Hardly anyone else is doing it.
- The ARC now funds me to study it. (Woohoo!)

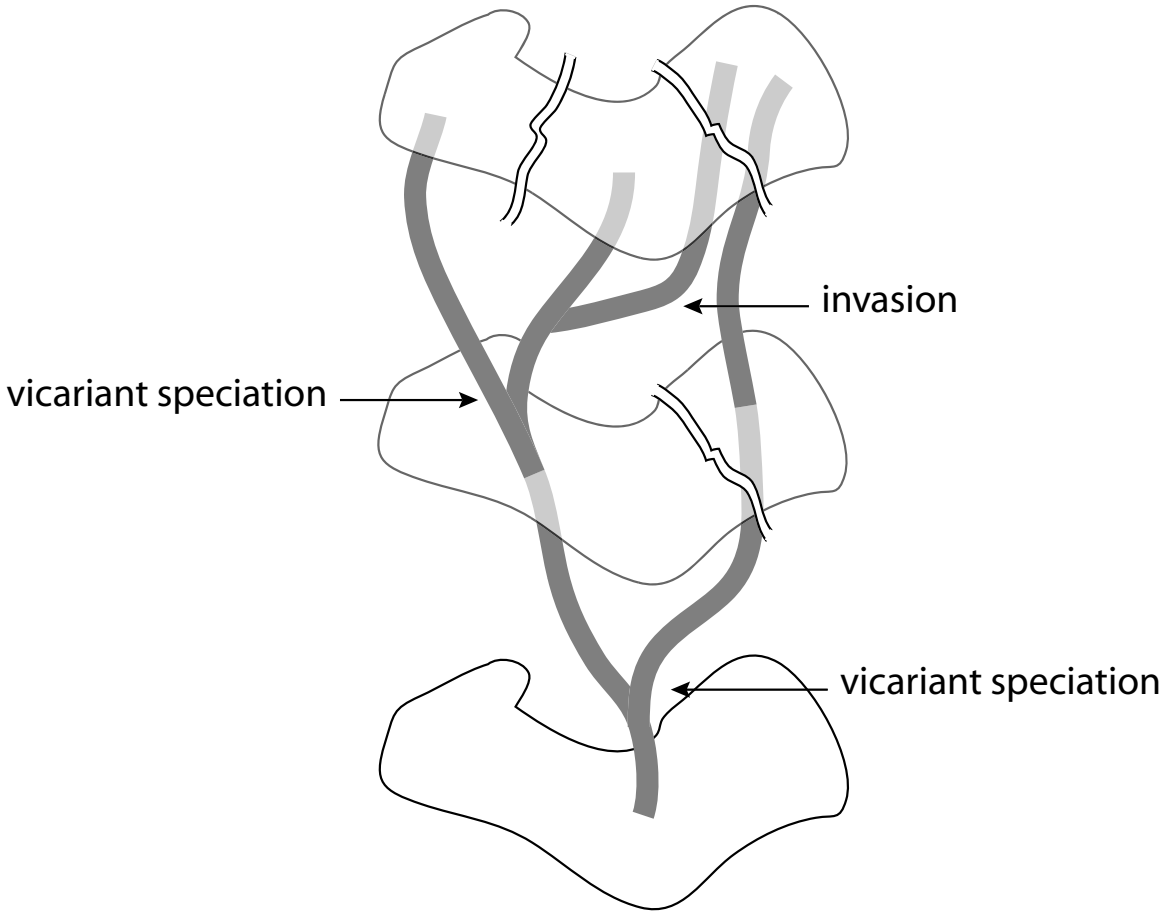
Different systems coevolve

For example,

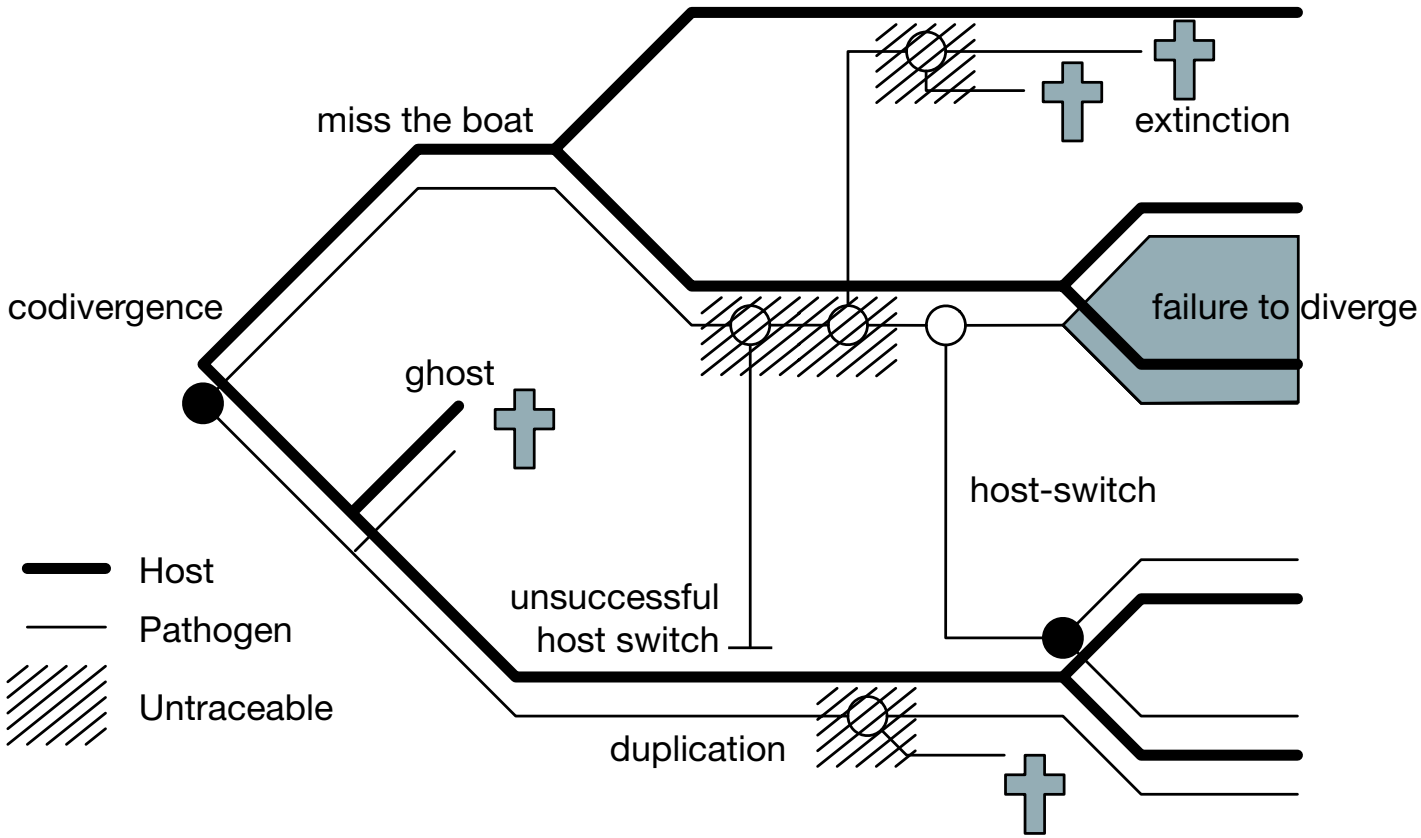
- hosts and their parasites or pathogens;
- whole organisms and their genes;



- geographical areas and the species which inhabit them.



General processes



Terminology of events

genes	**parasites**	**pathogens**	**biogeography**	**words**
codivergence	cospeciation / codivergence		vicariance	codivergence
duplication	independent speciation	emergence	sympatric speciation	duplication
horizontal transfer	*host switch*	host switch / zoonosis	migration	borrowing
loss (drift)	extinction	extinction / eradication	extinction	*loss*
	missing the boat / lineage sorting			loss

Loss arises from multiple processes, which cannot be distinguished from the phylogenies and associations.

codivergence & duplication

Definition

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.

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.

host switch & loss

Definition

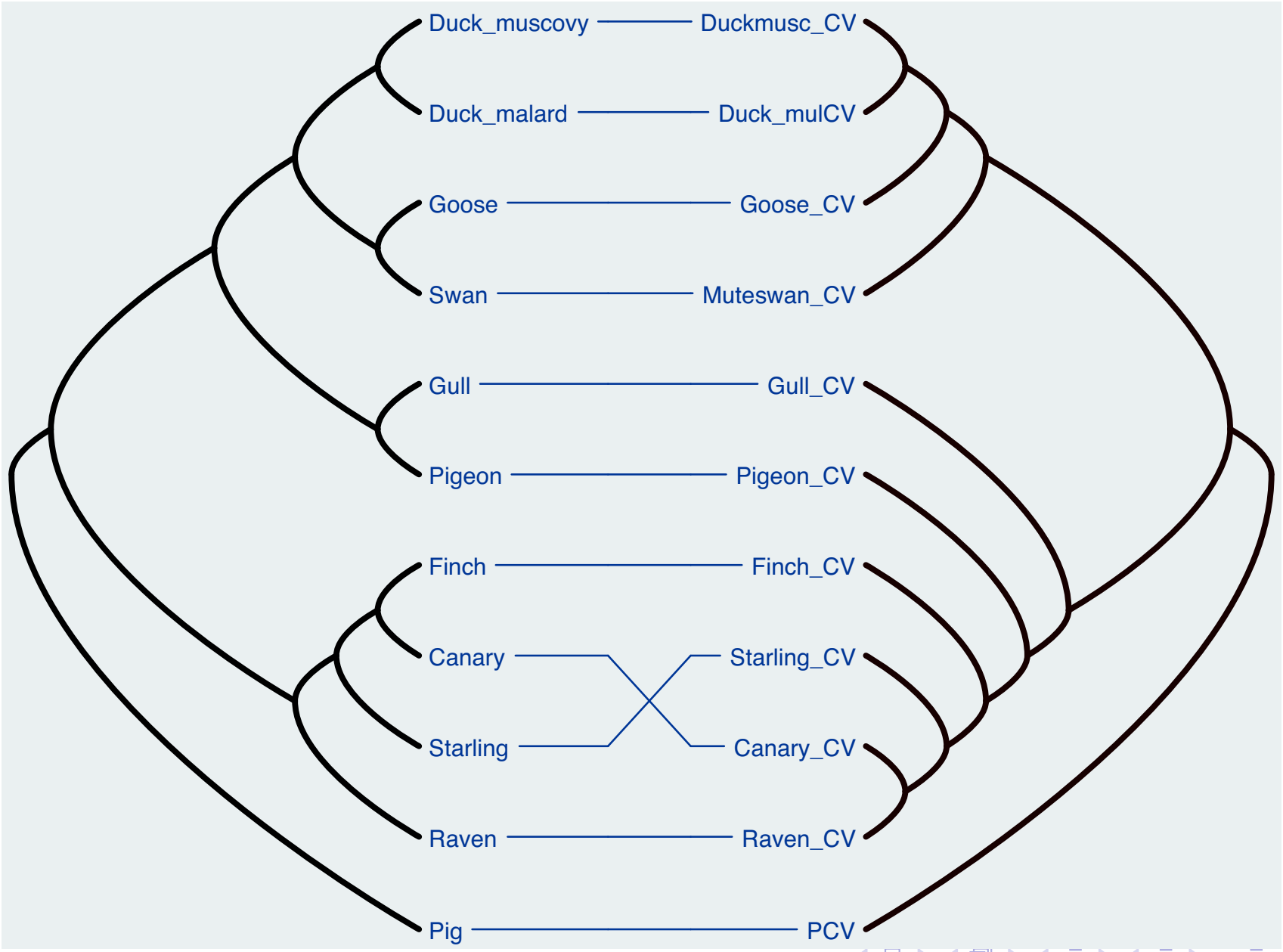
A host switch occurs for some arc $(p, q) \in A(P)$ where p is associated with a location in H that is contemprary with, but not ancestral to, the location in H with which q is associated. In my approach in order for a host switch to occur p must duplicate and a successful invasion into a new host lineage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

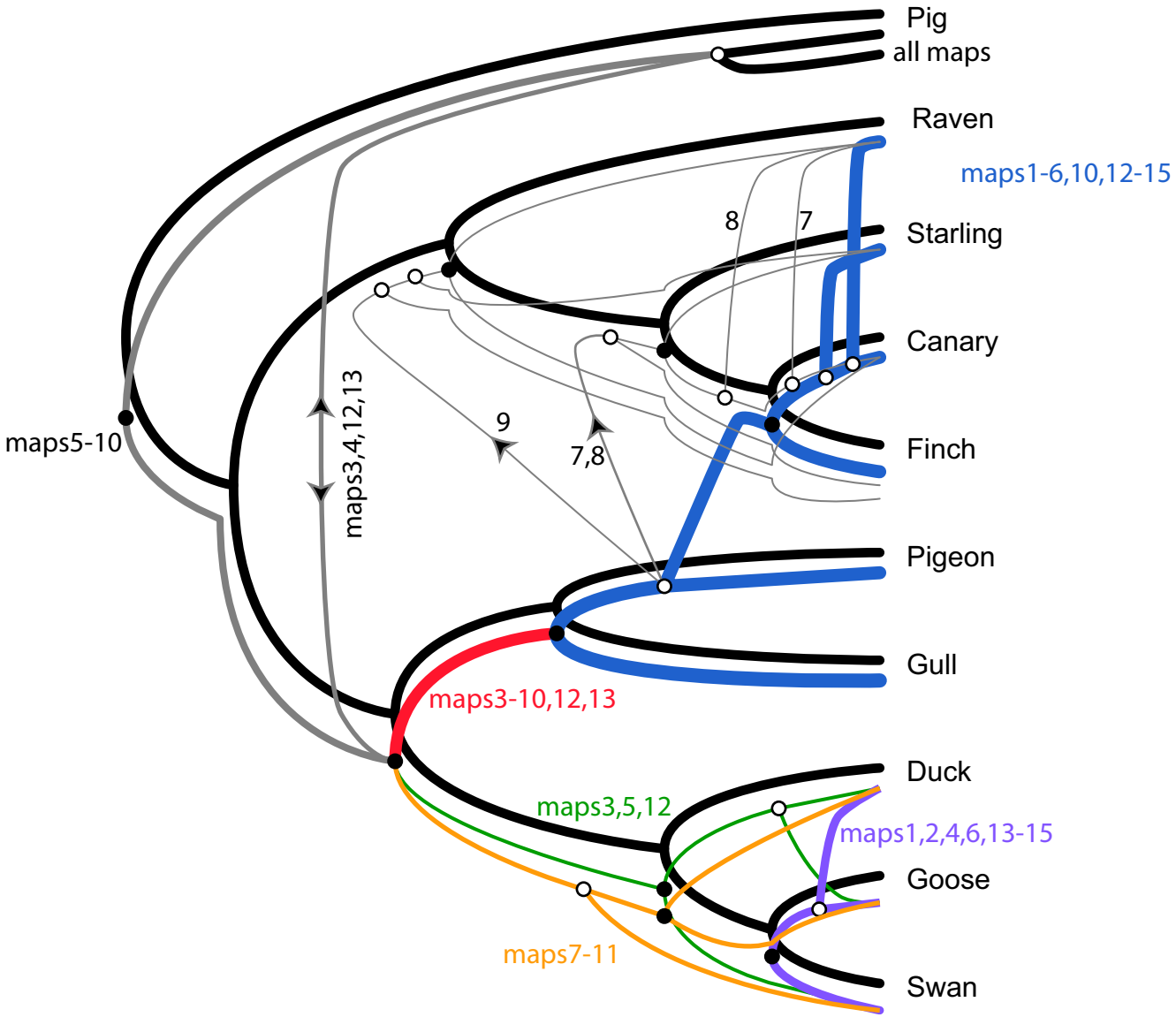
Circoviruses

- A virus of major economic importance is porcine circovirus 2 (PCV2).
- PCV2 is a single-stranded, non-enveloped, circular DNA virus about 1.76kb
- PCV2 is highly infectious and its prevalence is reported up to 40-60%, even up to 100% in some regions.
- PCV2 causes post-weaning multisystemic wasting syndrome (PMWS), which in Australia can lead to 50% mortality [8]

Tanglegram of circoviruses



Consensus-like map of relationships



Ramdsen et al., Mol. Biol. Evol. 2009

Basic terminology

Here all trees are *phylogenies*: each vertex has 0 or 2 children, and each tree has a defined internal *root*, from which all edges are directed.

The edges of T are $E(T)$; the vertices are $V(T)$.

Each vertex represents an operational taxonomic unit (OTU) or simply *taxon*.

A leaf is a vertex with out-degree 0. The leaves of T are $L(T)$.

The root represents the common ancestor of all the taxa in the tree.

If the edges are weighted then weights represent some measure of evolutionary distance (sequence divergence, time, morphological or DNA-DNA hybridisation distance etc.).

Problem Statement

Given

- a *host* phylogeny H
- an *associate* phylogeny P
- known associations φ of the tips of P with those of H

The object is to find out the ancestral relationships between P and H.

What we can recover

Ronquist confirmed in 2002 [9] that there are only four types of recoverable event for this problem:

1. codivergence,
2. duplication,
3. loss, and
4. host switching

All methods (attempt to) recover codivergence, but not all can recover host switching. Some only recover codivergence, duplication and loss.

Existing methods

There are three basic approaches to answering questions about cophylogeny:

- parsimony / event-based methods
- tree metrics
- mapping / reconcilation

Parsimony/event-based methods

Brooks came up with an early solution, coined Brooks' Parsimony Analysis (BPA) [1].

BPA recodes the known associations φ and the parasite tree P as binary characters and then puts them into a parsimony-based tree reconstruction method.

This suffers from many difficulties, including

- Requirement for *post hoc* reinterpretation (generally by Brooks himself).
- Inability to cope correctly with host switching without further *post-hoc* fiddling.
- Treatment of non-independent tree characters as independent binary characters.

Tree Metrics

Huelsenbeck and Rannala [4] came up with a method to test codivergence, but it was a *metric*: it treats P and H as equivalent, whereas we consider P to be "tracking" H.

Their method answered the question "did these trees evolve according to the same model of cladogenesis"?

Another (better?) question is "how strongly was the evolution of this parasite or pathogen governed by that of its hosts?"

A later method was maximum-likelihood based and asymmetric [5], but it was rather simplistic: one major constraint was that it permitted only one parasite lineage to exist per host lineage.

Mapping

The most intuitive (and computationally hardest) method is to reconcile trees using a mapping from P into H.

This requires no *post hoc* interpretation and can recover all events, depending on how the mapping method is implemented.

The Cophylogeny Mapping Problem

Given H, P and the map $\varphi : L(P) \mapsto L(H)$,

We want to recover a mapping

$$\Phi : V(P) \mapsto V(P) \times \xi,$$

where $\xi = V(H) \cup A(H)$.

An element x of ξ is a *location*, and an element (p, x) of $V(P) \times \xi$ is a *j-vertex*.

The relationships among associations of $V(P)$ and ξ define the historical events that took place.

Thus we need a set of j-vertices (associations), one for each vertex $p \in V(P)$, which could account for the coevolutionary history of P and H.

Definition

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

1. *Φ extends φ, that is, $\Phi|_{L(P)} \equiv \varphi$ ("lifting" condition);*
2. *If $P \cong H$ and φ preserves the isomorphism, then Φ preserves the isomorphism also ("consistency" condition);*
3. *Φ is an isomorphism of P ("isomorphism" condition);*
4. *If $p \in V(P) \setminus L(P)$ is mapped by Φ to ℓ in H, then at least one child of p must also be mapped by Φ to a descendant of ℓ ("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 ℓ ("interpretable" condition).*

(Adapted from [3].)

Complexity

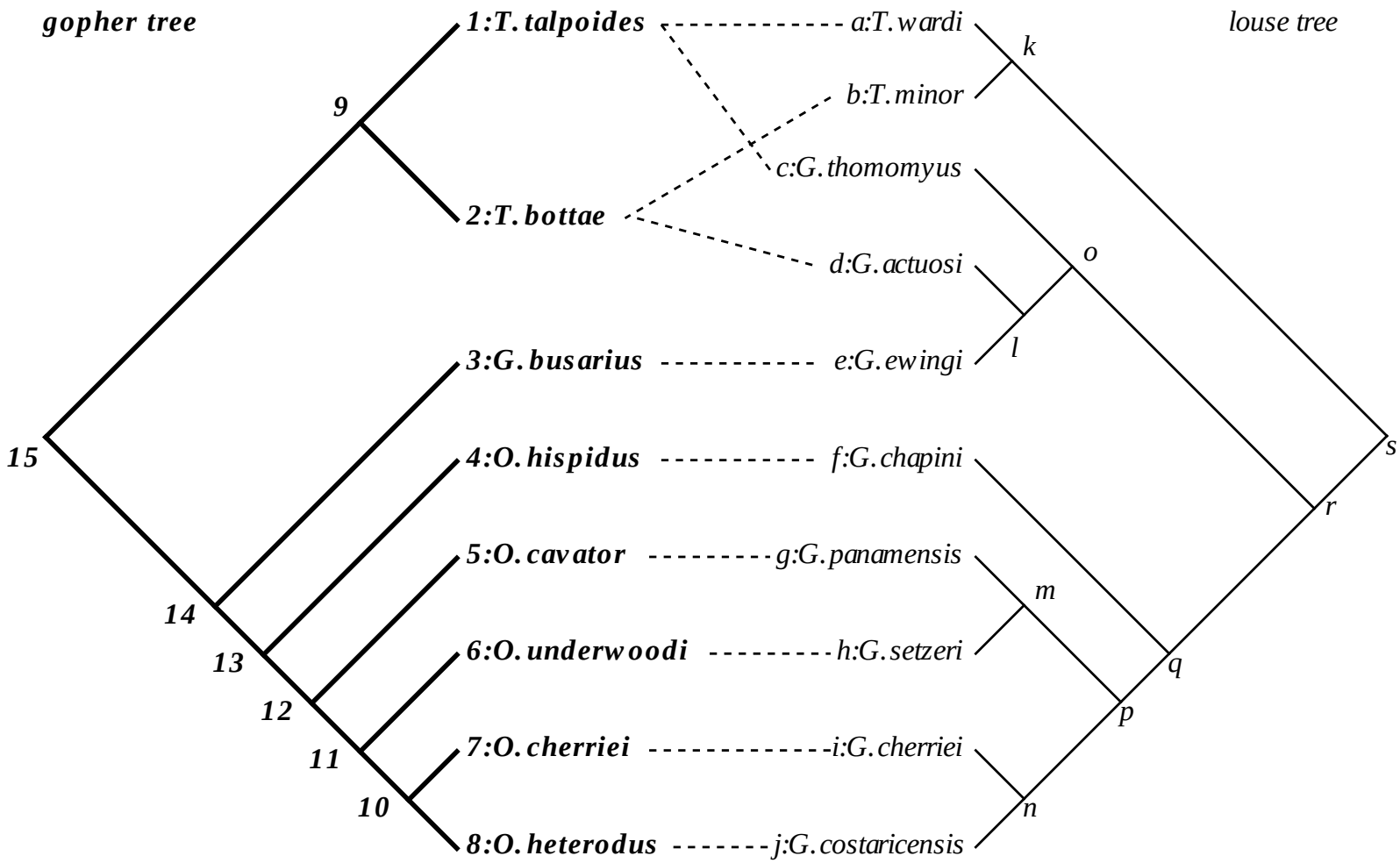
The cophylogenetic network reconcilation problem is NP-complete, even if there are only two possible locations in time for each host node [6];

In fact it has recently been shown that the tree-specific form (where $\mathcal{H}$ and $\mathcal{P}$ are trees) is also NP-complete (Libeskind-Hadas *et al.*, submitted to Jour. Comp. Biol.).

Jungles

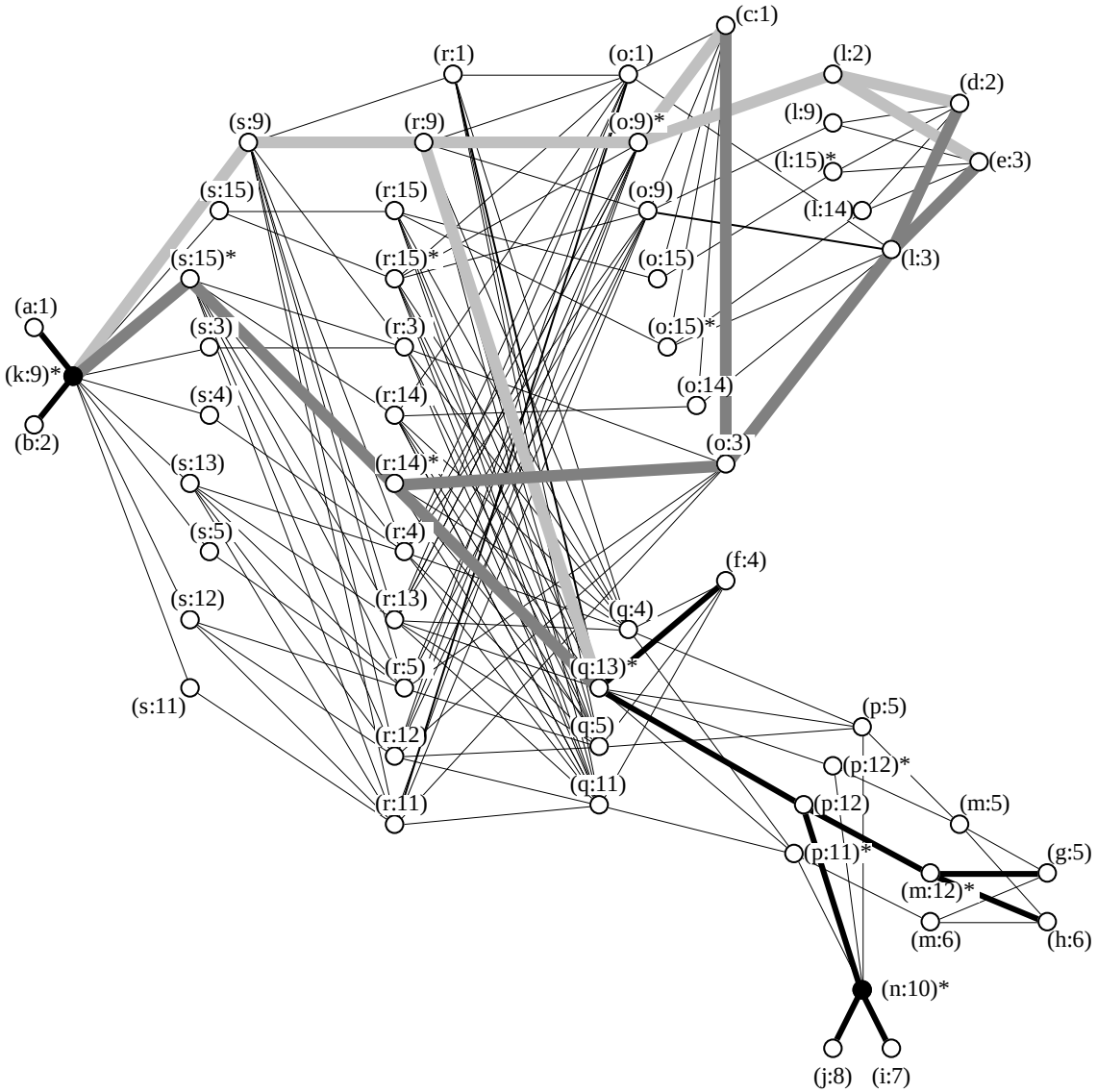
One solution to the mapping problem is to construct a *jungle*, which contains as subgraphs all the (Pareto) maps [2].

Treating parasite lineages as independent of each other and ignoring problems with time-travelling parasites we can traverse this jungle relatively quickly.



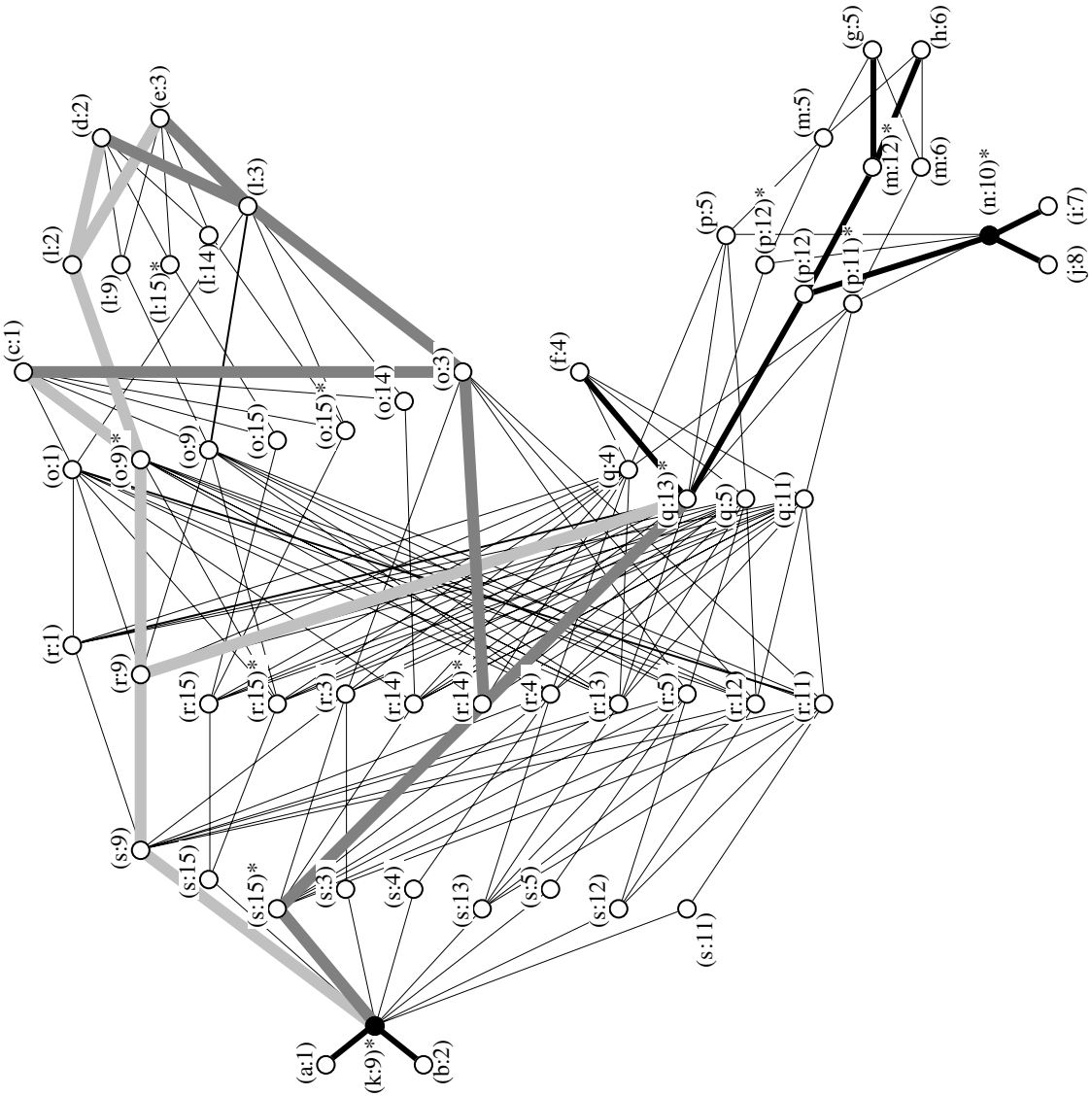
from Charleston, *Math. Biosc.* 1998

Jungle



from Charleston, *Math. Biosc.* 1998

Jungle Fowl



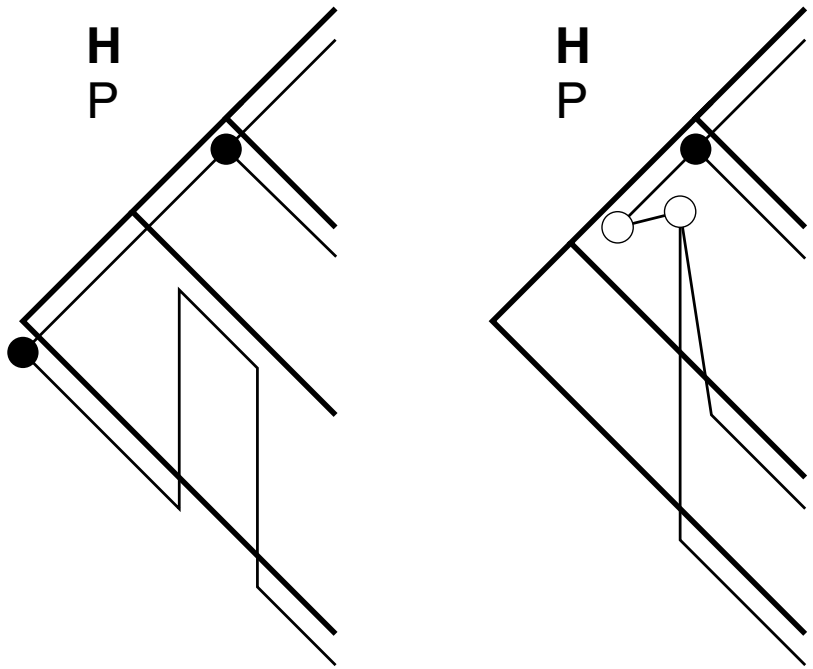
from Charleston, *Math. Biosc.* 1998

Untraceable Events

Not all coevolutionary events are *traceable*.

A parasite lineage that leaves no trace of itself on a host lineage, either by becoming extinct or by leaving that host, cannot be linked with that host unless in a probabilistic way.

These events are *untraceable*.



Event Likelihoods and Costs

The basic probabilistic model of codivergence has these ingredients:

- p_c: The conditional probability that if the host lineage of a parasite (or pathogen) diverges, so does that parasite.
- λ_P: The birth rate of the parasite tree, instantaneous rate of any extant lineage diverging independently of its host.
- μ_P: The death rate of the parasite tree, instantaneous rate of any extant lineage going extinct.
- p_w: The conditional probability that if a duplication event took place, one nascent parasite lineage switches hosts.

These can be roughly 'scored' with values c, d, e and w respectively, with $c < d, e, w$.

Definition

A history $\mathcal{A}$ is a collection of associations of parasite/pathogen vertices and host locations that contains as a subgraph some Φ.
We say that $\mathcal{A}$ presents Φ if there is such a Φ that satisfies the conditions defined earlier. Note that every mapping Φ is a history.
Note that a history may contain untraceable events

Combinatorics of segmented histories

So how many histories are there?

Consider a location ℓ in a segmented history, with first and second descendants ℓ', ℓ'' (if they exist). If a single parasite lineage p is associated with ℓ, we have the recurrence

$$F(\ell) = \begin{cases} 1 + F(\ell') + F(\ell') \sum_{a_i} F(a_i'), & \text{if } \ell \text{ is an arc of } H \\ 1 + F(\ell') + F(\ell'') + F(\ell') F(\ell''), & \text{if } \ell \text{ is a vertex of } H \end{cases}$$

for the total number of possible histories generated by the parasite lineage, where the a_i are all of the arcs in the segment containing ℓ.

This number increases alarmingly!

Combinatorics of segmented histories

It is also useful to the consider number of histories presenting a specific reconstruction

- Can count histories in much the same way, with special cases for reconstructed associations
  - Reconstructed null event may become duplication with or without switch
  - Reconstructed lineage sort may become a codivergence
  - Other reconstructed associations must be preserved

$$H(\ell, p) = \begin{cases} H(\ell', p')\, H(\ell'', p''), & \ell, p \text{ vertices} \\ H(\ell', p')\,(1 + F(\ell'')), & \ell, p \text{ reconstructed as sort} \\ H(\ell', p')\left(1 + \sum_{a_i} F(a_i')\right), & \ell, p \text{ reconstructed as null} \\ \ldots & \end{cases}$$

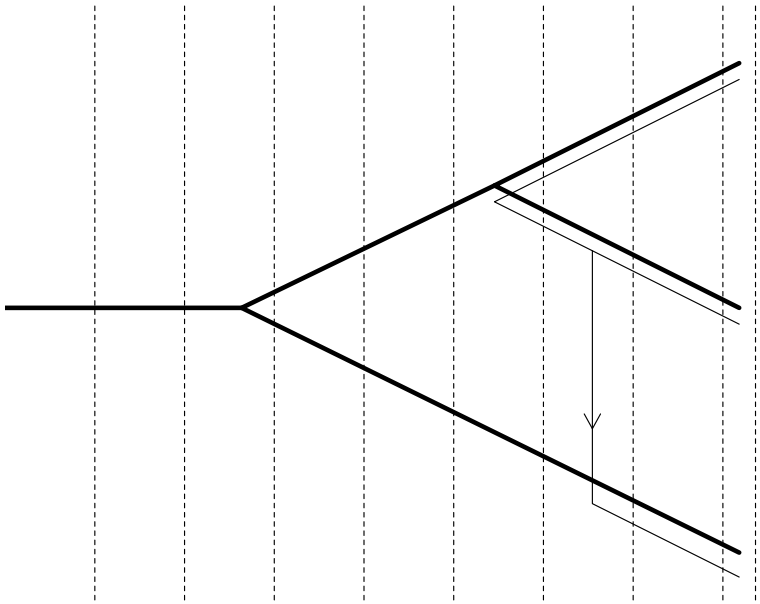
Philosophy

We can approximate time because

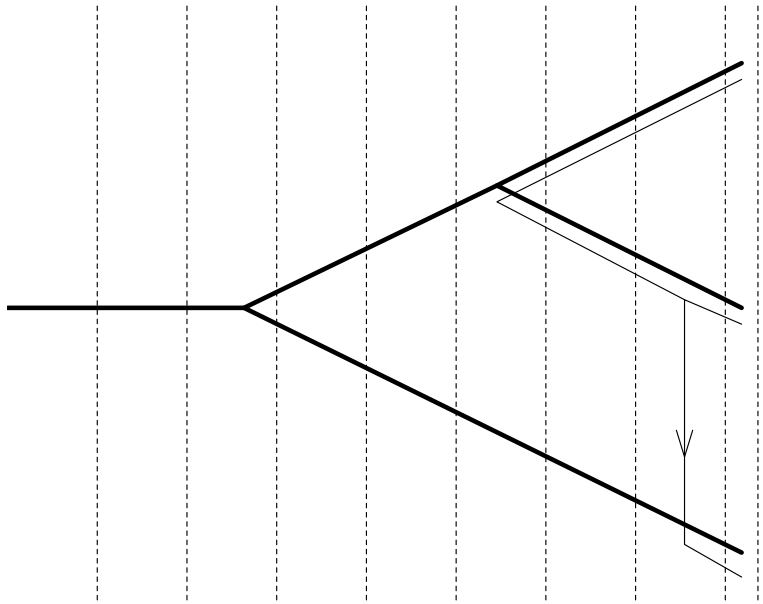
1. dating estimates tend to have very large confidence intervals
2. speciation does not happen in an instant
3. given small enough time increments we can approximate the continuous case

Discretization

1. Given an instance of the problem
2. If there are branch lengths, use them to segment the times into chunks (e.g., 10My) (as many as needed for the whole $\mathcal{P}$)
3. Begin with a map such as the *reconciled tree*, *sensu* Page [7]
4. Embed the map in a *history*
5. Search histories that present the original map *via* Markov chain Monte Carlo (MCMC) to estimate a likelihood of a given map
6. or Search all histories to find the maximum likelihood map



(a): map1a



(b): map1b

Map 1 and its two discretizations

Probabilities

$$
\begin{aligned}
Pr(map1a) &= (p_c^2) \cdot (p_n p_{hs}/2) \cdot (p_n^3) \cdot (p_s^3) \\
&= p_c^2 \cdot p_n^4 \cdot p_{hs} \cdot p_s^3/2 \qquad (1) \\
Pr(map1b) &= (p_c^2) \cdot (p_n^2) \cdot (p_n p_{hs}/2) \cdot (p_s^3) \\
&= p_c^2 \cdot p_n^3 \cdot p_{hs} \cdot p_s^3/2 \qquad (2)
\end{aligned}
$$

Where

- p_c probability of codivergence
- p_d probability of duplication
- p_w probability of host switch, given a duplication
- p_x probability of going extinct
- p_s probability of sampling extant taxon

Also define p_n as $(1 - p_d - p_x)$ and $p_{hs} \equiv p_d^2 p_w (1 - p_w)$ for brevity.

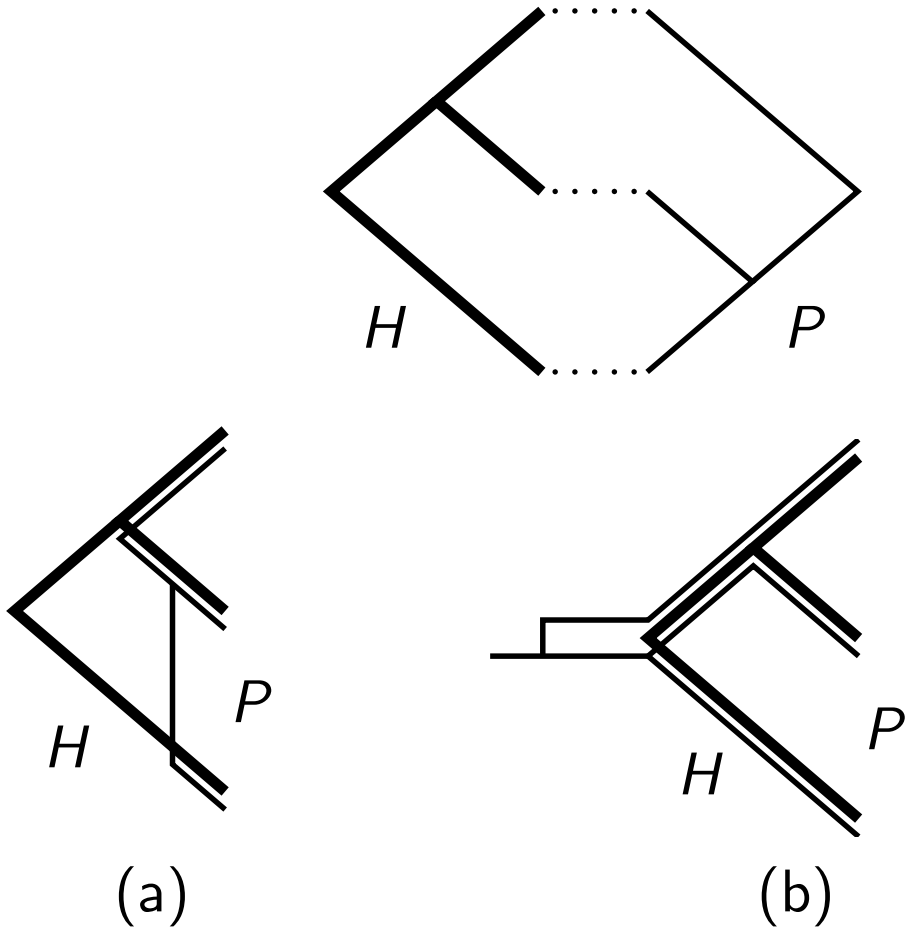
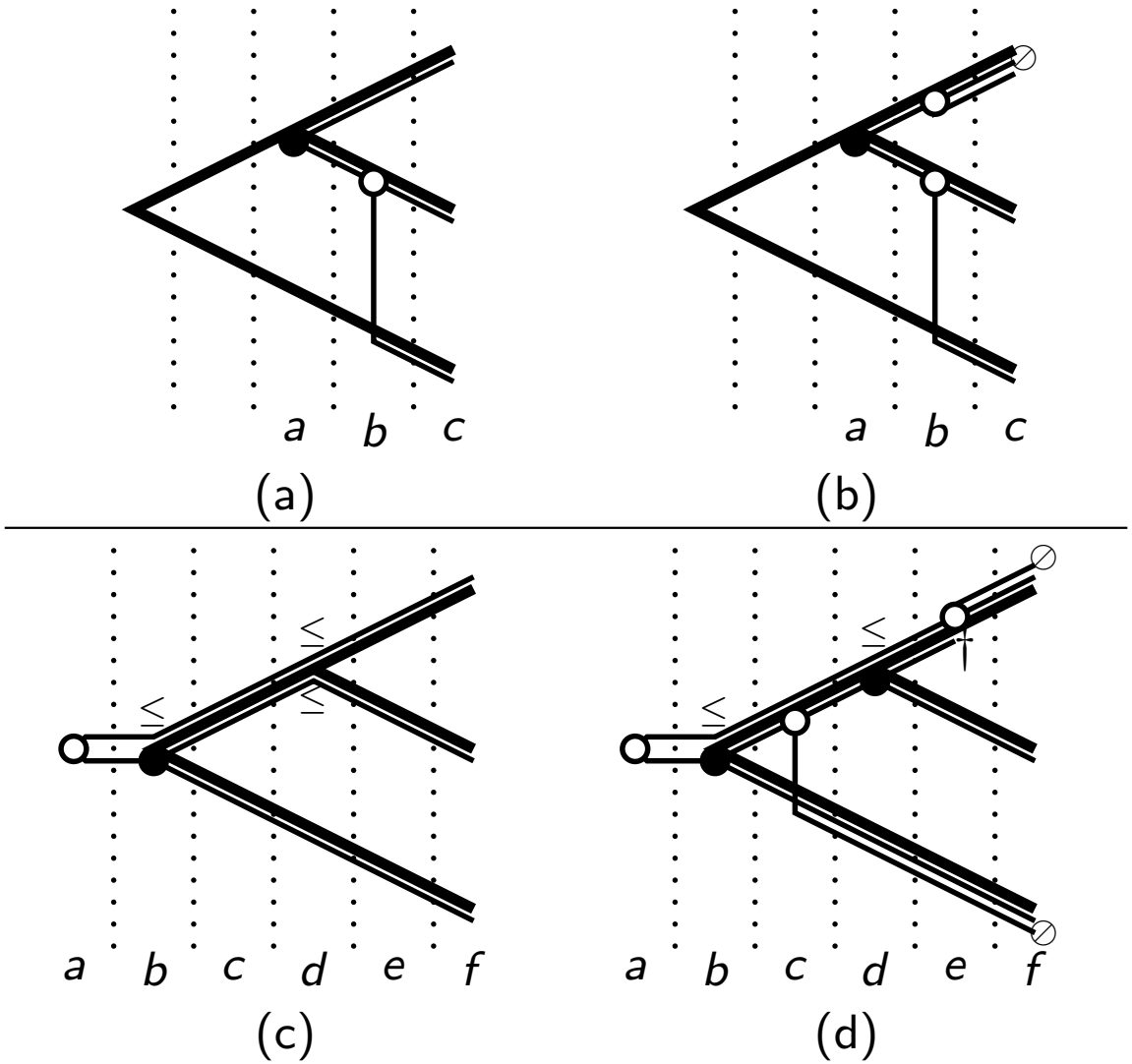


Figure: A simple tanglegram and two Pareto-optimal solutions

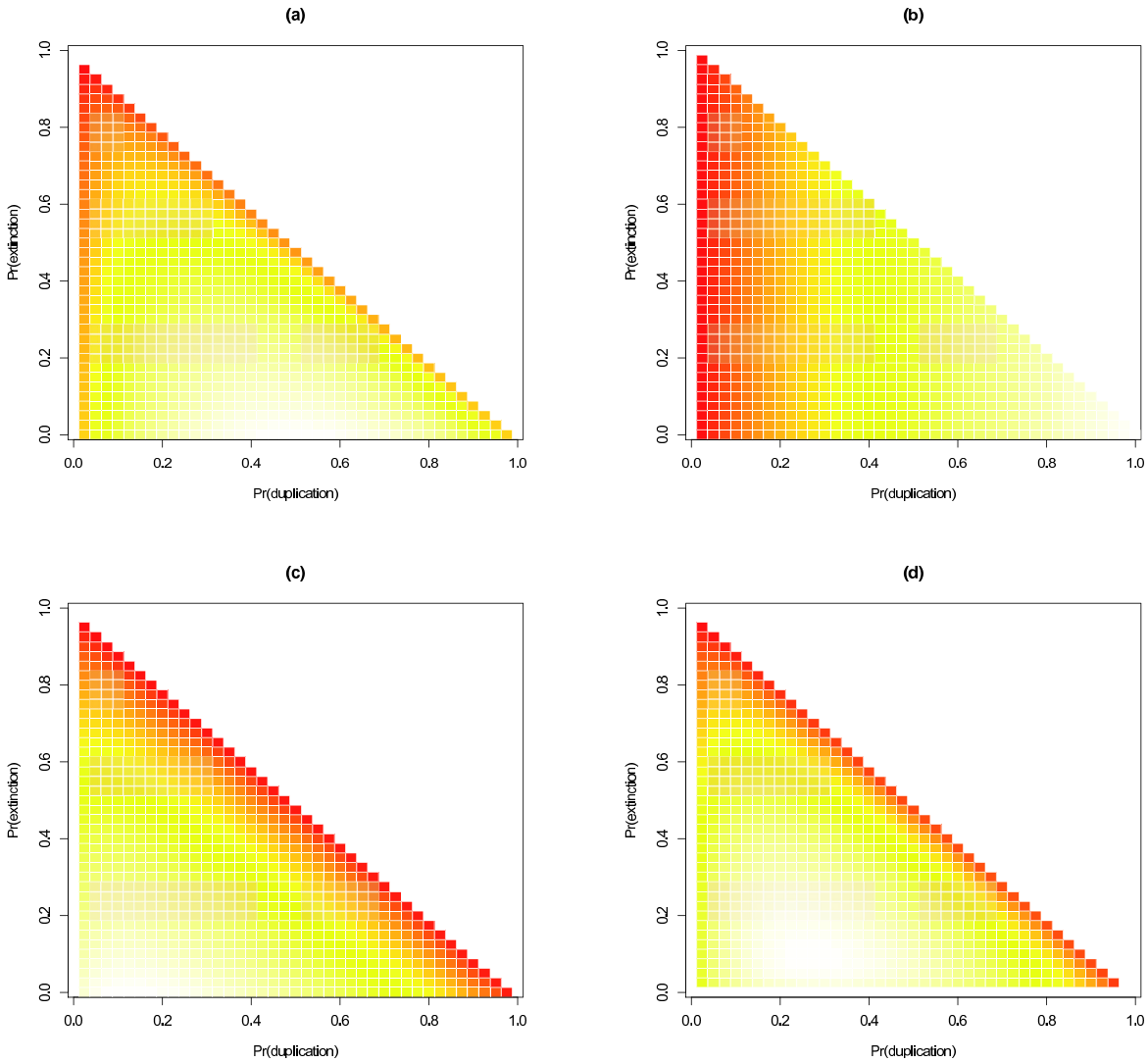
Histories



Key: •codivergence; °duplication; ≤miss the boat; †extinction; ⊘sampling failure

Probability distributions

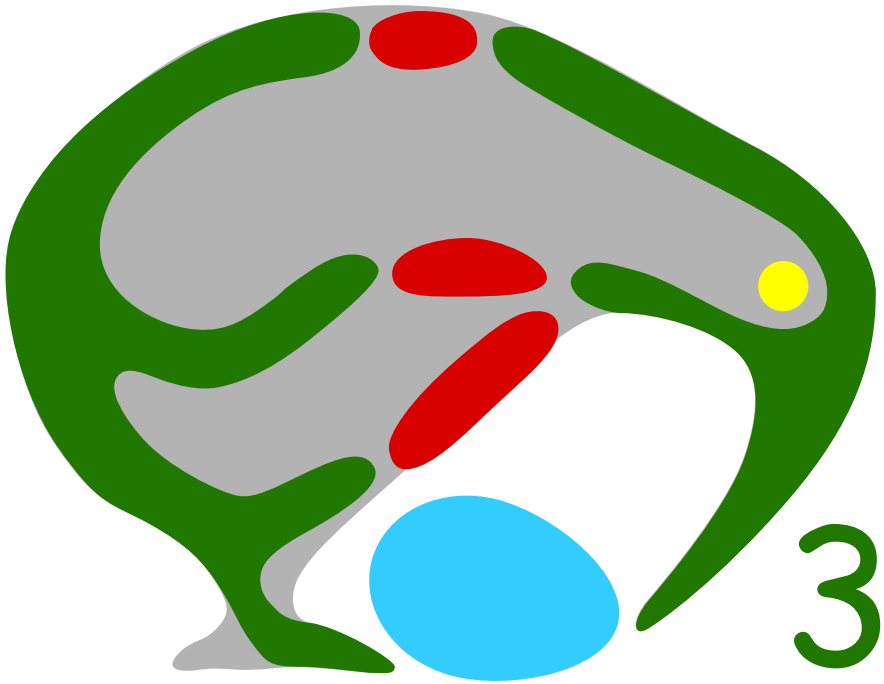
$p_c = 0.75,\ p_w = 0.2,\ p_s = 0.9$



Outcomes, we hope

- Estimates of model parameters, e.g., rates of zoonosis
- Easily extensible to include competition, cross-immunity etc.
- Fast heuristics for finding most likely maps, even on large problems
- Allowance for phylogenetic uncertainty
- Doesn't require either phylogeny to be a tree so can be implemented on DAGs

TreeMap 3



- In Java now (woo! GUI! Cross-platform!)
- Likelihoods and heuristics in progress
- Doesn't require phylogenies to be trees (just DAGs)
- Hasn't solved computational complexity by converting to Java (but some things *much* faster)

Job opportunities

- ARC-funded (DP1094891) post-doc: 3 years starting early 2010
- PhD top-up scholarships & projects

Acknowledgements

Support:
This work is supported by the Australian Research Council: parts of TreeMap3 by ARC DP0770991 and subsequent research by ARC DP1094891.

Collaborateurs:

- Eddie Holmes
- Cadhla Firth (Née Ramsden)
- Rod Page
- Iain Beeston
- Susan Perkins
- . . . and many more.

References

D. R. Brooks.
How to do BPA, really.
Jour. Biog., 28:345–358, 2001.

M. A. Charleston.
Jungles: a new solution to the host/parasite phylogeny reconciliation problem.
Math. Biosci., 149(2):191–223, May 1998.

Michael A. Charleston.
Principles of cophylogeny maps.
In Michael Lässig and Angelo Valleriani, editors, *Biological Evolution and Statistical Physics*. Springer-Verlag, 2002.

J. P. Huelsenbeck and B. Rannala.
Phylogenetic methods come of age: testing hypotheses in an evolutionary context.
Science, 276:227–232, 1997.

John P. Huelsenbeck, Bruce Rannala, and Ziheng Yang.
Statistical tests of host-parasite cospeciation.
Evolution, 51(2):410–419, 1997.

Ran Libeskind-Hadas and Michael Charleston.
On the computational complexity of the reticulate cophylogeny reconstruction problem.
Journal of Computational Biology, 16(1):05–117, 2009.
doi:10.1089/cmb.2008.0084.

R. D. Page and M. A. Charleston.
From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.
Mol. Phyl. Evol., 7(2):231–240, Apr 1997.

C. Ramsden, M. A. Charleston, S Duffy, B. Shapiro, and E. C. Holmes.
Insights into the evolutionary history of an emerging livestock pathogen: Porcine circovirus 2.
Journal of Virology.
(in press).