The status of cophylogenetic analysis

Michael Charleston

University of Sydney

Phylomonia 2010.11.04-05
Some motivation

Figure 1: Numbers of papers cited in PubMed with co[-](speciat|diverg)* in the title or abstract

- About 75% of emergent human diseases are *zoonoses*, (SARS, HIV, Ebola, H1N1, ...).
- Understanding where an organism came from (e.g., invading pests) can tell us how better to *combat* them.
Different systems can coevolve at the macroscopic level

hosts and their parasites or pathogens; geographical areas and the species which inhabit them; whole organisms and their genes;
Introduction

The goal is to determine, for two groups of ecologically linked taxa, what were the evolutionary paths they took with respect to each other.

We aim to answer questions like:

- How long is the association between host and parasite?
- Did they cospeciate?
- Were there host switches or lateral gene transfers?
- What kind of risk of cross-infection does this pathogen present to its sister species?
Problem Instance

Given

- a *host* phylogeny $\mathcal{H}$
- an *associate* phylogeny $\mathcal{P}$
- known associations $\varphi$ of the tips of $\mathcal{P}$ with those of $\mathcal{H}$

We call a problem instance a *tanglegram*, such as $T = (H, P, \varphi)$.

The object is to find out the ancestral relationships between $\mathcal{P}$ and $\mathcal{H}$. This mostly comes down to an optimization problem.
Coevolutionary events

- Codivergence
- Miss the boat
- Extinction
- Host switch
- Unsuccessful host switch
- Duplication
- Ghost
- Failure to diverge
- Host-switch

Legend:
- Black: Host
- Gray: Pathogen
- Hatched: Untraceable
**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 $h$. 
**Definition**

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.
Definition

A host switch occurs for some arc \((p, q) \in A(P)\) where \(p\) is associated with a location in \(H\) that is contemporary with, but not ancestral to, the location in \(H\) with which \(q\) is associated.
Definition

A loss occurs as the result of one of three things, which 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 $p$. 
What we can recover

Ronquist confirmed in 2002\textsuperscript{[9]} that these are the 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.

We would like also to recover \textit{failure to diverge} events, where a parasite of a speciating host continues to parasitize it, without divergence.
Event costs

We can assign a cost to each event type, subject to simple constraints:[$^2$]

the biological rule, $c < d, \ell, w$ (for codivergence, duplication, loss and host switch respectively)

and the pragmatic rule, $0 \leq c, d, w, \ell$ (which allows a dynamic program to solve the optimization problem).
Jane 1 & 2 use dynamic programming to minimise total cost, with event costs prescribed.

<table>
<thead>
<tr>
<th>Event</th>
<th>Jane Cost</th>
<th>TreeMap Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cospeciation $c$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Duplication $d$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Host Switch $w$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Loss/Sorting $\ell$</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

TreeMap puts default costs on events but does not normally use them: it finds a Pareto front of solutions, and has worst-case exponential running time.

Jane 1 uses an $O(n^7)$ algorithm to find minimal cost reconstructions; Jane 2 has an algorithm down to $O(n^3)$.

The penalty in $O(n^7) \rightarrow O(n^3)$ is a loss of approximately 0.1% in performance, which is definitely acceptable.
Part II

Everything Is Against Us
Farenholz’ Rule,[5] that “parasite phylogeny mirrors host phylogeny,” has misled us. The classic cophylogeny example of gophers and lice (left) is wonderful, but in fact such cases are rare: more and more studies are showing lack of evidence for codivergence, despite similarity.

$p < 0.01$; much apparent congruence & codivergence
Many studies look for congruence with inappropriate tools

Program crashes: :: problem with program
The number of feasible maps increases rapidly for even modest numbers of taxa.

The number of maps in the Pareto front—those which could be optimal for some scheme of event costs—also increases quickly.

from Charleston 2003[3]
Empirical evidence of complexity

- The number of feasible maps is also highly correlated with the degree of incongruence.

from Charleston 2003[3]
Cophylogeny mapping is NPC

We begin with the Generalized Cophylogeny Reconstruction Problem (GCRP).

This is a 6-tuple

\[(H = (V_H, E_H), P = (V_P, E_P), t_H, t_P, \varphi, \kappa)\]

where \(H\) is the host network, \(P\) the parasite network, \(t_H\) and \(t_P\) are timing functions for \(H\) and \(P\) that map each vertex to a set of permitted times, \(\varphi\) is defined as before, and \(\kappa\) is a 4-tuple cost vector \(\kappa = (c, d, w, \ell)\) for codivergence, duplication, host switch and loss respectively.

The objective is to find a mapping \(\Phi : P \mapsto H\) that extends \(\varphi\), can be constructed using the usual events with respect to the timing functions, and is of minimum total cost.
Theorem

**GCRP** is solvable in polynomial time for the set of instances 

\((H = (V_H, E_H), P = (V_P, E_P), t_H, t_P, \varphi, \kappa)\) such that (i) \(P\) is a tree and 

(ii) for all \(v \in V_H\), \(|t_H(v)| = 1\)

(Proof is by construction of a polynomial time algorithm for this case using a dynamic program. See Libeskind-Hadas & Charleston\(^7\) for details.)
We first define the Generalized Cophylogeny Reconstruction Decision Problem (GCRDP) as follows:

**Instance:** Given \( (H = (V_H, E_H), P = (V_P, E_P), t_H, t_P, \varphi, \kappa) \) and a cost \( K \).

**Question:** Does there exist a reconstruction whose cost is \( K \) or less?

**Theorem**

The decision problem associated with GCRP is NP-complete for the set of instances \( (H = (V_H, E_H), P = (V_P, E_P), t_H, t_P, \varphi, \kappa) \) such that (i) \( P \) is a tree and (ii) for all \( v \in V(H) \), \( |t_H(v)| \leq 2 \).

(Proof is by reduction to 3-SAT: see Libeskind-Hadas & Charleston[7] for details.)
Cophylogeny mapping is NP-Complete

Ovadia et al. define the Cophylogeny Reconstruction Decision Problem as follows:

Definition

An instance of the Cophylogeny Reconstruction Decision Problem (CRDP) is a 4-tuple \((H, P, \varphi, B)\) where

- \(H\) and \(P\) are rooted host and parasite trees,
- \(\varphi : L(P) \leftrightarrow L(H)\) maps the tips of \(P\) to the tips of \(H\), and
- \(B\) is a 4-tuple \((B_C, B_D, B_S, B_L)\) of upper bounds on the number of cospeciation, duplication, loss and host switch events respectively.

The decision question is: Does there exist a mapping \(\Phi\) that extends \(\varphi\) and whose cost is strictly less than \(B\)?
Cophylogeny mapping is NP-Complete

Theorem

The CRDP is NP-complete.

(Proof is by showing that a related problem, where \( \psi : L(P) \mapsto 2^{L(H)} \), is NP-complete by a reduction from 3-SAT; then such an instance can be transformed into a corresponding instance of CRDP with the same answer, thus implying CRDP is NP-complete. See Ovadia et al.\cite{8} for details.)
Cophylogeny is probably in APX-Hard

A related problem is the *Lateral Gene Transfer Problem*, LGTP.

**Definition**

A *lateral transfer scheme* for a species tree $S$ is a pair $(S', A')$ where $S'$ is a subdivision of $S$ and $A \subseteq \{ \langle x, y \rangle : x, y \in V(S') \setminus V(S), x \neq y \}$ such that

1. the mixed graph $S' \cup \epsilon(A')$ does not contain a directed mixed cycle;
2. the tail of each arc $A'$ has in-degree 1 and out-degree 2 in $S' \cup A'$;
3. the head of each arc in $A'$ has in-degree 2 and out-degree 1 in $S' \cup A'$.

The lateral transfer scheme indicates where *gene* transfers could have occurred in the *species* tree.

A scheme is called “$\alpha$-active” if there are at most $\alpha$ gene copies on any one lineage of the species tree at any one time.
Cophylogeny is probably hard to approximate

_Sensu_ DasGupta _et al._ (2005), the LGT for a species tree $S$ and gene tree $G$ is to find a gene transfer scheme that has minimal cost.

**Theorem**

There does not exist a polynomial time approximation scheme for the 1-active LGT problem with performance guarantee of $1 + \epsilon$ where $\epsilon \geq 3/370024$

(See DasGupta _et al._ 2005[4] for details.)

This suggests that the cophylogeny reconstruction problem probably doesn’t have a polynomial time approximation scheme (PTAS) / is APX-Hard.
Part III

Method Comparison
Brooks came up with an early solution, coined Brooks’ Parsimony Analysis (BPA)\textsuperscript{[1]}.

BPA recodes the known associations $\varphi$ and the parasite/pathogen trees $H, P$ as binary characters and then puts them into a parsimony-based tree reconstruction method.
Brooks’ Parsimony Analysis works by

1. assigning IDs to internal nodes of both $P$ and $H$, and
2. using those to create an “alignment” of binary characters,
3. which are then used by a parsimony program to find the most parsimonious assignment of states to the internal branches of $H$.
4. The pattern of these states is then interpreted to determine what (co)evolutionary events must have taken place.
**Algorithm 1: BPAPathAssign** \((T)\)

1. /* \(T\) is a bifurcating tree */

1. let \(n = |L(T)|\)

2. let \(k = |V(T)|\)

3. without loss of generality let the nodes be labelled \(v_1 \ldots v_k\)

4. let \(S\) be a list of binary strings \((S_1, \ldots, S_k)\), initially all 0’s

5. for each \((v_i \in L(T))\) do {

6. 

6. for \((j = 1 \ \text{up to} \ k)\) do {

7. 

7. if \((v_j\) is on the path from \(v_i\) to the root) then {

8. 

8. \(S_{i,j} \leftarrow 1\)

9. 

9. }

10. 

10. }

11. }


There is no freely available implementation of Brooks’ Parsimony Analysis (BPA) method, or Secondary BPA (SBPA) which was proposed later as a fix for some of the issues with BPA.

An implementation was created to obey Brooks’ descriptions in the literature (there is no pseudocode description of BPA/SBPA in the literature either).

BPA works by assigning IDs to nodes in both $H$ and $P$, based on their position in the tree. This is from tip to root, and takes in worst case $O(n^2)$ runtime.

In order to increase efficiency we implemented a randomized version of SBPA, Randomized SBPA, with hoped-for runtime of $\Theta(n \log n)$.

---

\(^1\) i.e., Ben
Encoding order matters

<table>
<thead>
<tr>
<th>Code</th>
<th>SBPA</th>
<th>Randomized SBPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100001001</td>
<td>100000011</td>
</tr>
<tr>
<td>2</td>
<td>010000111</td>
<td>010001011</td>
</tr>
<tr>
<td>3</td>
<td>001000111</td>
<td>001001011</td>
</tr>
<tr>
<td>4</td>
<td>000100011</td>
<td>000100101</td>
</tr>
<tr>
<td>5</td>
<td>000011001</td>
<td>000010101</td>
</tr>
</tbody>
</table>

These sequences can be interpreted post-hoc to infer the coevolutionary events.

\( ^2 = \text{have to be} \)
SBPA suggests 1 codivergence, 3 duplications and 3 host switches.

SBPA suggests 2 codivergences, 2 duplications and 2 host switches.
ParaFit\textsuperscript{[6, 10]} works by converting both trees to distance matrices

**Algorithm 2: ParaFit**

1. let $\mathcal{T} = (H, P, \varphi)$ be a tanglegram
2. $A \leftarrow$ the associations $\varphi$ expressed as a matrix
3. $B \leftarrow$ the principal coordinates matrix of $P$
4. $C \leftarrow$ the principal coordinates matrix of $H$
5. $D \leftarrow CA^\top B$
6. $M \leftarrow \begin{bmatrix} A & B \\ C & D \end{bmatrix}$

The test statistic is $trace(D^\top D) = \sum (d_{ij}^2)$ and is gained by recalculating $D$ using the original $B$ and $C$ matrices, and randomizing the associations represented by $A$\textsuperscript{3}.

\textsuperscript{3}NB: in AxParaFit the random number seed is the same each time!

\textsuperscript{4}And its faster version, AxParaFit\textsuperscript{[10]}
Confidence is expressed as a percentage, as $100\% \times (1 - p)$

250 tanglegrams were created at random for each trial

red: random

blue: identical
Data: A huge sample with hundreds of arthropods and Wolbachia infecting them, data collected by Patricia Simões in Tahiti, Moorea, Raiatea.
Wolbachia on Moorea
Wolbachia on Moorea

With missing associations removed it’s not much better:

Moorea (cleaned)
Wolbachia on Raiatea

Host_Raiatea associations Wolbachia_Raiatea

Raiatea

The status of cophylogenetic analysis

Phylomania
The butterflies: *Heliconius*

*Heliconius* butterflies have a complex mimic/model system. Here we show two clades of the *Heliconius* genus, being different races of the target/model species *erato* and the mimic species *melpomene*.

We can identify the (race×region) tips with almost no difficulty (there result in two different resolutions), into *mimicry complexes*: just the race and the region matter.
Heliconius mimicry complexes
Heliconius mimicry complexes

![Diagram of Heliconius mimicry complexes]

- erato
  - eraPanhyd
  - eraCospet
  - eraCospet
  - eraFrehyd
  - eraFreera
  - eraPerfav
  - eraPeremm
  - eraBraphy
  - eraEasety

- associations
  - melPanmel
  - melCosros
  - melFremel
  - melFrethe
  - melPerama
  - melPeragl
  - p-26
  - melBranan

- melpomene
Heliconius mimicry complexes

The status of cophylogenetic analysis

MAC (USyd)  The status of cophylogenetic analysis  Phylomania  40 / 50
<table>
<thead>
<tr>
<th>erato</th>
<th>associations</th>
<th>melpomene</th>
</tr>
</thead>
<tbody>
<tr>
<td>eraPanhyd</td>
<td>melPanmel</td>
<td></td>
</tr>
<tr>
<td>eraCospet</td>
<td>melCosros</td>
<td></td>
</tr>
<tr>
<td>eraPanpet</td>
<td>melPanros</td>
<td></td>
</tr>
<tr>
<td>eraWescyr</td>
<td>melWescyt</td>
<td></td>
</tr>
<tr>
<td>eraPerfav</td>
<td>melPerama</td>
<td></td>
</tr>
<tr>
<td>eraPeremm</td>
<td>melPeragl</td>
<td></td>
</tr>
<tr>
<td>eraBraphy</td>
<td>p-26</td>
<td>melBranan</td>
</tr>
<tr>
<td>eraEasety</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MAC (USyd) | The status of cophylogenetic analysis | Phylomania | 40 / 50 |
Heliconius mimicry complexes

The status of cophylogenetic analysis
Heliconius mimicry complexes

![Diagram of Heliconius mimicry complexes]

- erato
  - eraPanhyd
  - eraCospet
  - eraPanpet
  - eraWescyr
  - eraTrihyd
  - eraColhyd
  - eraFrehyd
  - eraFreera
  - eraPerfav

- associations
  - melPanmel
  - melCosros
  - melPanros
  - melWescyt
  - melColmel
  - melTrimel
  - melFremel
  - melFrethe

- melpomene
  - melPerama

The status of cophylogenetic analysis

Phylomania 40 / 50
Heliconius mimicry complexes

\[ \text{erato} \quad \text{associations} \quad \text{melpomene} \]

- \text{eraPanhyd} \quad \text{melPanmel}
- \text{eraCospet} \quad \text{melCosros}
- \text{eraPanpet} \quad \text{melPanros}
- \text{eraWescyr} \quad \text{melWescyt}
- \text{eraTrihyd} \quad \text{melColmel}
- \text{eraColhyd} \quad \text{melTrimel}
- \text{eraFrehyd} \quad \text{melFremel}
- \text{eraFreera} \quad \text{melFrethe}
- \text{eraPerfav} \quad \text{melPerama}
- \text{eraPeremm} \quad \text{melPeragl}
Possible histories

map 1/4-31.204105 meliconia->erato

14 Codivergences 7 Duplications 1 Host switch 13 Losses
Possible histories

map 2/4-31.204105 meliconia->erato

14 Codivergences 7 Duplications 1 Host switch 13 Losses
Possible histories

map 3/4-29.781637 meliconia->erato

16 Codivergences 6 Duplications 13 Losses
Possible histories

map 4/4-31.096415 meliconia->erato

12 Codivergences 8 Duplications 2 Host switches 13 Losses
Randomizing both $P$ and $\varphi$ in Jane 2 we can estimate significance. All 1000 randomizations of $P$ have higher cost than the original: $p < \approx 0.0005$ in both cases.
Part V

Closing
There often isn’t the kind of quality of data that we initially hoped for. Codivergence is not the norm.

There are more algorithms emerging, and more programs becoming available; these will need thorough testing.

There has been some interesting progress in understanding the complexity of the cophylogeny mapping problem.

There remain many challenging open questions in biology, mathematics & statistics, and computer science in this area.

Thanks!
Thanks to...

Jennifer Hoyal Cuthill
(mimicry analysis)

Patricia Simões
(wolbachia)

INRIA (France)

Ben Drinkwater
(simulations)

Australian Research Council

MAC (USyd)
D. R. Brooks.
How to do BPA, really.

M. A. Charleston.
Jungles: a new solution to the host/parasite phylogeny reconciliation problem.

M. A. Charleston.

Inapproximability results for the lateral gene transfer problem.

H. Fahrenholz.
Ectoparasiten und abstammungslehre.

P. Legendre, Y. Desdevises, and E. Bazin.
A statistical test for host-parasite coevolution.

R. Libeskind-Hadas and M. Charleston.
On the computational complexity of the reticulate cophylogeny reconstruction problem.
Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas.
The cophylogeny reconstruction problem is np-complete.

F. Ronquist.
Parsimony analysis of coevolving species associations.

A. Stamatakis, A. F. Auch, J. Meier-Kolthoff, and M. Göker.
Axpcoords & parallel axparafit: statistical co-phylogenetic analyses on thousands of taxa.
Work underway

- Implementing the discrete likelihood model in TreeMap*
- Implementing the optimization problem as a Linear Program/SAT

and other methods
Desirable

- Faster heuristics for optimizing solutions
- Consensus methods
Open questions

- Is there a bigger $\epsilon$ such that the LGT problem can be solved to within $1 + \epsilon$ in polynomial time?
- What to do about multi-host parasites?