Doing Cophylogenetics Fast

Michael Charleston *et al.*
University of Sydney

Phylomania, November 2013

Abstract

Cophylogeny mapping, the process of finding a set of plausible associations between ancestors of ecologically linked extant taxa, is both intuitive and valuable.

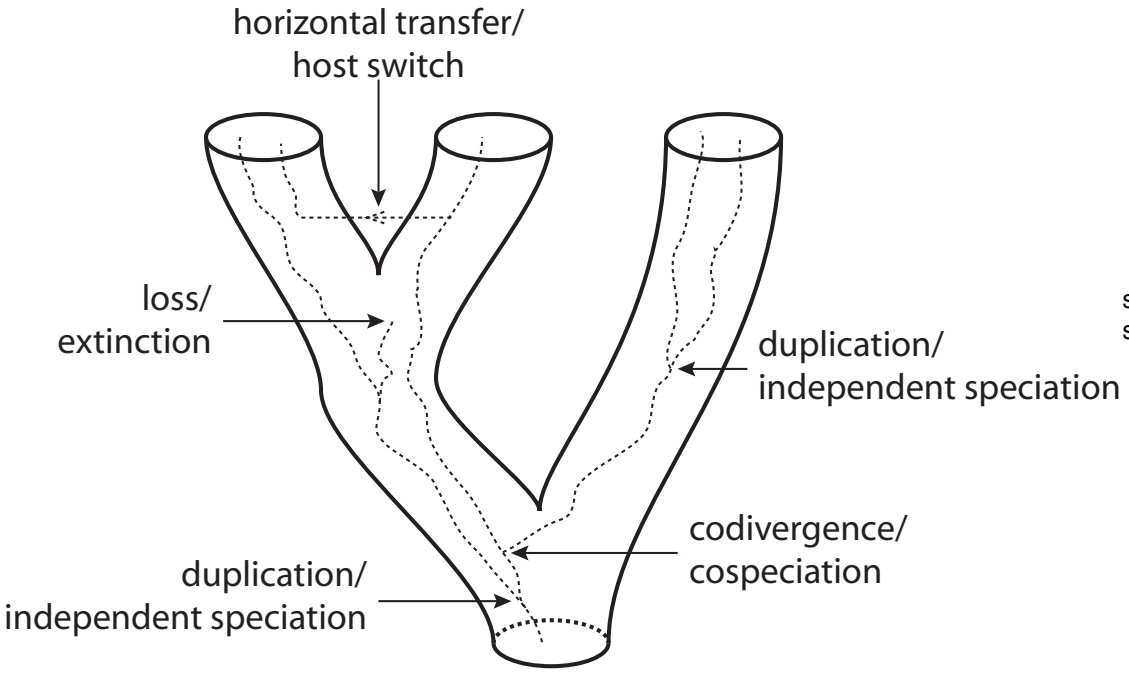
The fact that it's also NP-Hard (the decision problem is NP-Complete) is frustrating and unsurprising.

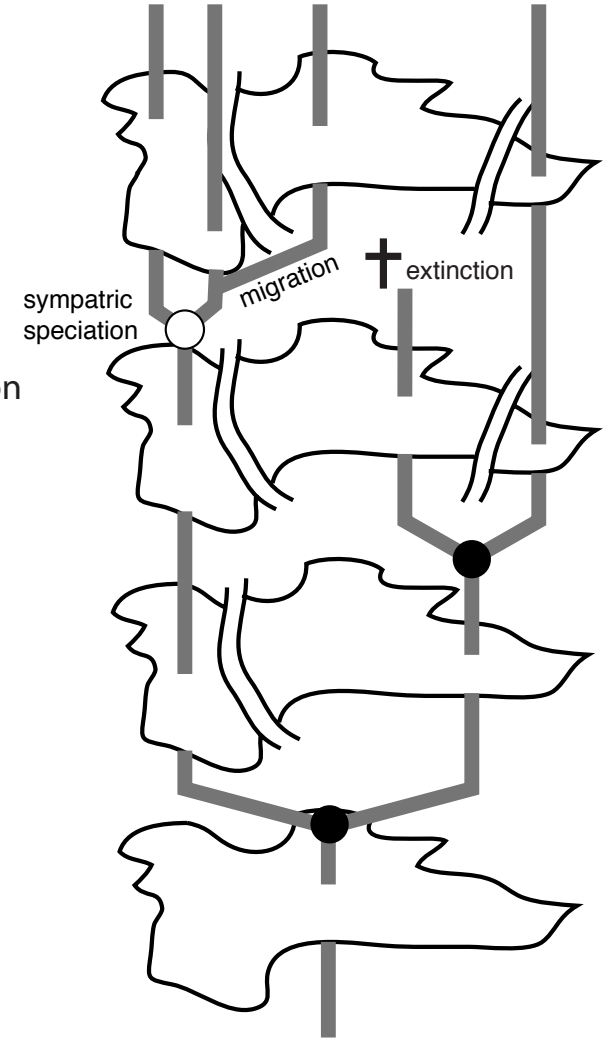
There is hope however in the form of graduate students (who can be parallelised!)

I present a set of approaches that can be used to (come close to) solving the cophylogeny mapping problem in good time. These approaches will enable researchers to investigate larger studies of coevolution.

Introduction to Cophylogenetics

Different Systems Coevolve

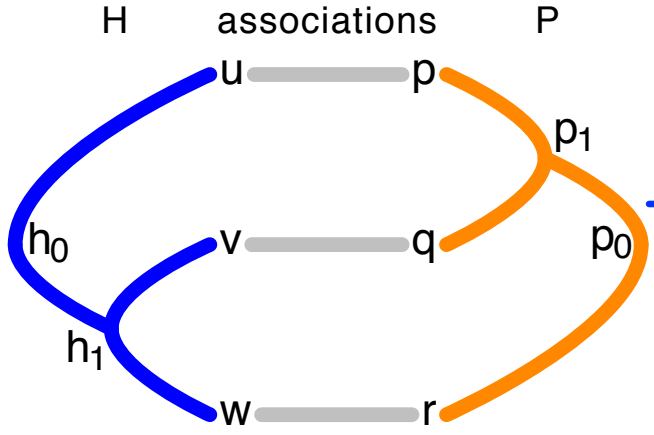




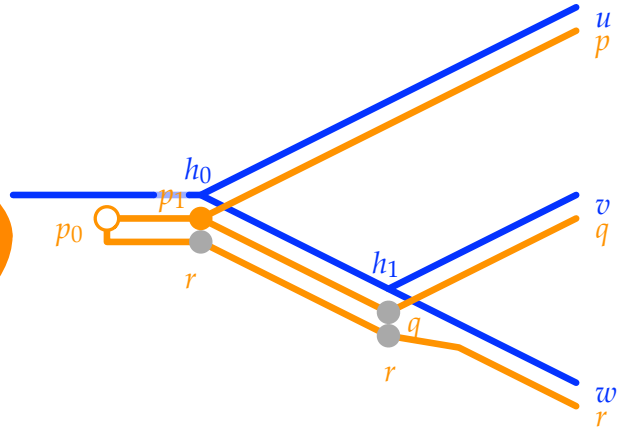
It's all pretty much the same problem in broad terms.

Mapping

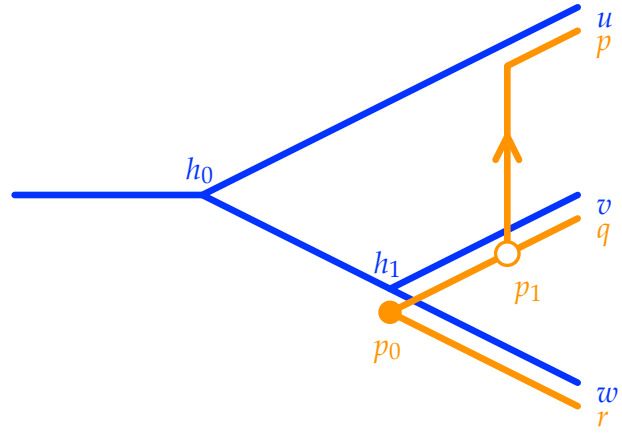
Given a *host* phylogeny (usually a rooted binary tree) H,
and a *parasite* or pathogen phylogeny (usually another such tree) P,
and a set of *associations* φ between their tips, we aim to answer questions about the coevolution of the parasites / pathogens with their hosts.



Given a tanglegram,



choose between this



. . . or this.

Above we can see codivergence, duplication, host switch and loss events.

Formalization

Cophylogeny Mapping Problem:
Find a minimal cost mapping from the dependent tree P into the independent tree H, subject to costs $\mathcal{C}$ and existing associations φ

Input: $H, P, \varphi, \mathcal{C}$

Output: A mapping Φ such that $\Phi|_{L(P)} \cong \varphi$ and is of minimal total event cost

Here H and P are rooted, leaf-labelled binary trees with leaf sets $L(H)$ and $L(P)$.
φ is a mapping from $L(P)$ into $L(H)$. $\varphi(p) = h$ means parasite or pathogen lineage p is found on / infecting host lineage h.

Assumptions

In general φ can be one to-many but most approaches assume

1. each parasite has only one host species
2. $\varphi(L(P)) = L(H)$

P and H are assumed to be "correct" working hypotheses.

P and H are assumed to be complete: there are no "ghost" lineages in either tree where invisible events can occur.

The decision problem of whether a map exists with a given cost is proved to be NP-complete [9, 12] but it would still be nice to solve the thing.

Recoverable Events

At present four coevolutionary events are recognised as recoverable:

Codivergence / cospeciation a parasite infecting a host lineage speciates with the host and infects both nascent host lineages;

Duplication a parasite speciates independently of the host and both new parasite lineages remain on its current host;

Loss a parasite is not present when it "should" be (caused by extinction, missing the boat or sampling failure);

Host switching a parasite successfully invades a new host species.

Host Timings

Although the general Cophylogeny Reconstruction Decision Problem is NP-Complete[12], if we *fix* the node times in H by giving them some unique integer values, the problem of mapping P into H is *polynomial*.

Libeskind-Hadas and MAC came up with such an algorithm for $P \mapsto H$ that is $O(n^7)$ by mapping parasite nodes to host edges[7, 9].

This algorithm was later modified to $O(n^3)$[8, 12] (see later).

An Integer Linear Programming solution

Zhou & Charleston

ILP

An Integer Linear Program (ILP) solution to this problem tries to assign true / false (Boolean) values to variables that are part of the problem statement, subject to a set of *constraints*, in order to optimise some *cost function*.

It doesn't make the underlying *complexity* any better, but there are good ILP solvers that can solve instances of the problem quickly.

Exact but still not too shabby

Our first attempt at an Integer Linear Program for the solution of the Cophylogeny Mapping problem was very slow[1].

Bin Zhou showed it was also incomplete — so generated his own and proved it to be complete and correct.

His ILP assumes two binary trees H and P and no widespread parasites.

[1] Libeskind-Hadas & Charleston, Tech. Report

ILP known variables

This the problem input:

- The set of host and parasite nodes and leaves $V(H), V(P), L(H), L(P)$;
- The partial orders of both trees $\preceq_H$, $\preceq_P$
- Non-relatedness of host nodes $\not\sim_H$
- The known leaf-leaf associations φ

ILP decision variables

These all have to be assigned Boolean values:

- A strict total ordering of the host nodes: $\ll_{h_1,h_2}$ is true $\iff$ h_1 speciated strictly before h_2
- The mapping itself: $\Phi_{h,p}$ is true $\iff$ p is associated with h at some point (recall p, h are nodes)
- Host switches: χ_{p,h_1,h_2} is true $\iff$ parasite p switched from h_1 to h_2
- Cospeciations: $C_{p,h}$ is true $\iff$ p and h cospeciated / codiverged.

ILP constraints

ILPs also require constraints (else it would all be too easy):

- Host nodes must be in total strict order and compatible with the ancestry relationships in the trees;
- Parasite ancestry relationships can't be broken by where they're mapped to ($p \prec q$ means $\Phi(q) \not\prec \Phi(p)$)
- Host switch take-off and landing must be contemporaneous (duh) and not imply time travel
- Parasites cannot map to unrelated hosts (must be associated with hosts on the same lineage)
- • • • (and some more — you get the idea).

ILP objective function

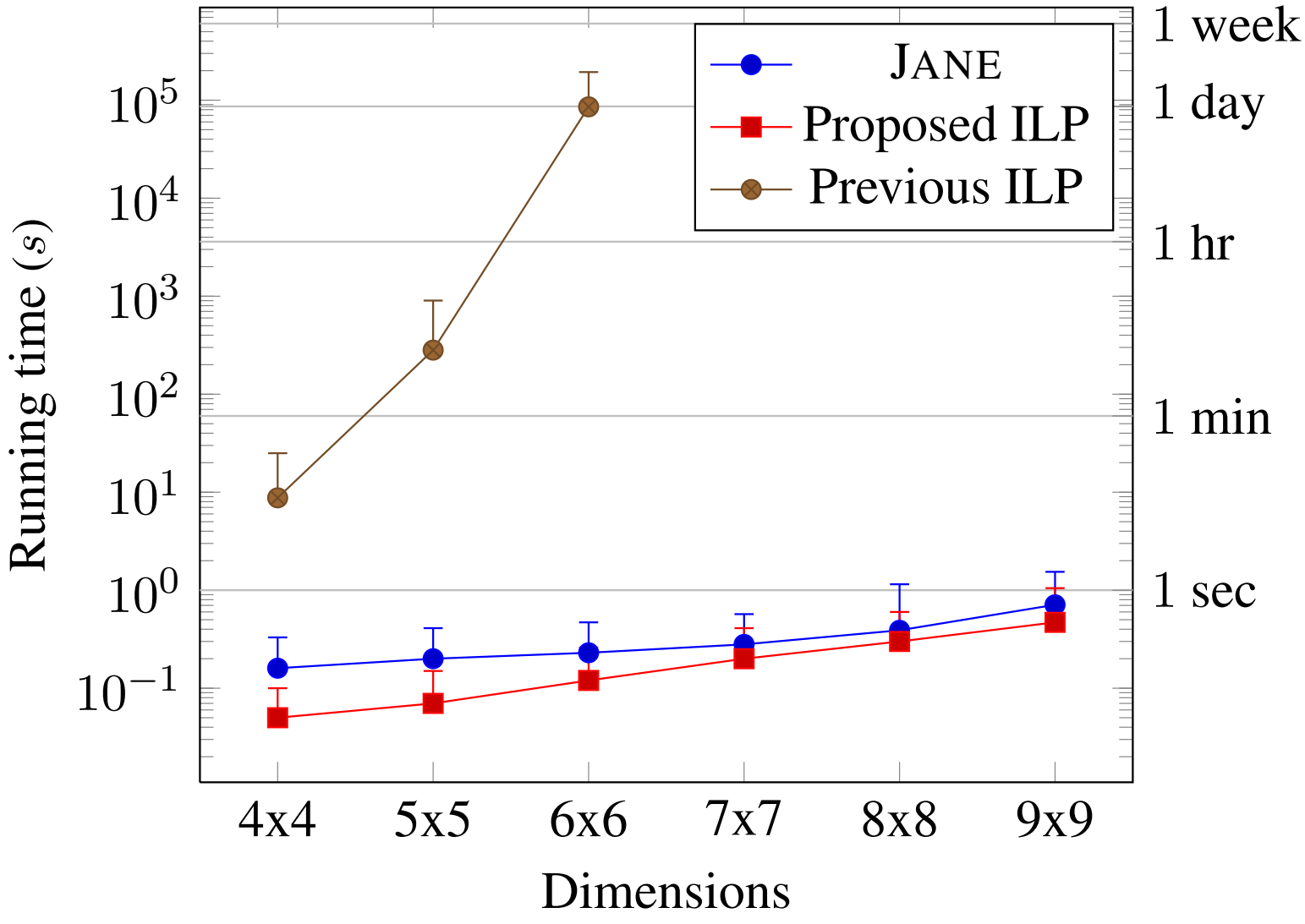
$\#C$ = number of codivergences; $\#D$ duplications; $\#W$ the number of host switches and $\#L$ losses.
The cost of X is $¥_X$.

Minimise total cost:

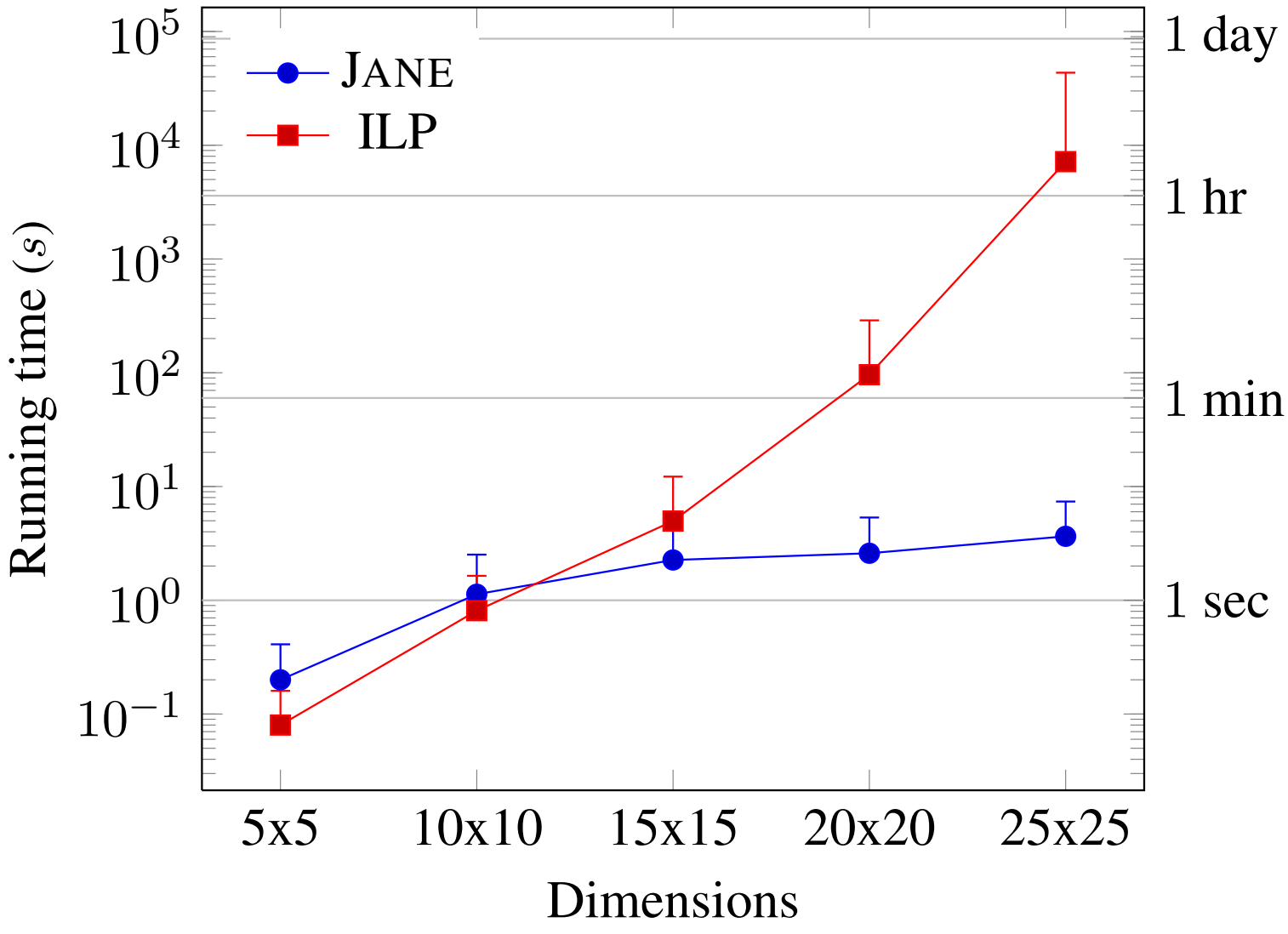
$$cost = ¥_C \#C + ¥_D \#D + ¥_W \#W + ¥_L \#L$$

Small data sets

(Jane is a Java program for cophylogeny mapping.)



Slightly larger data sets



Summary

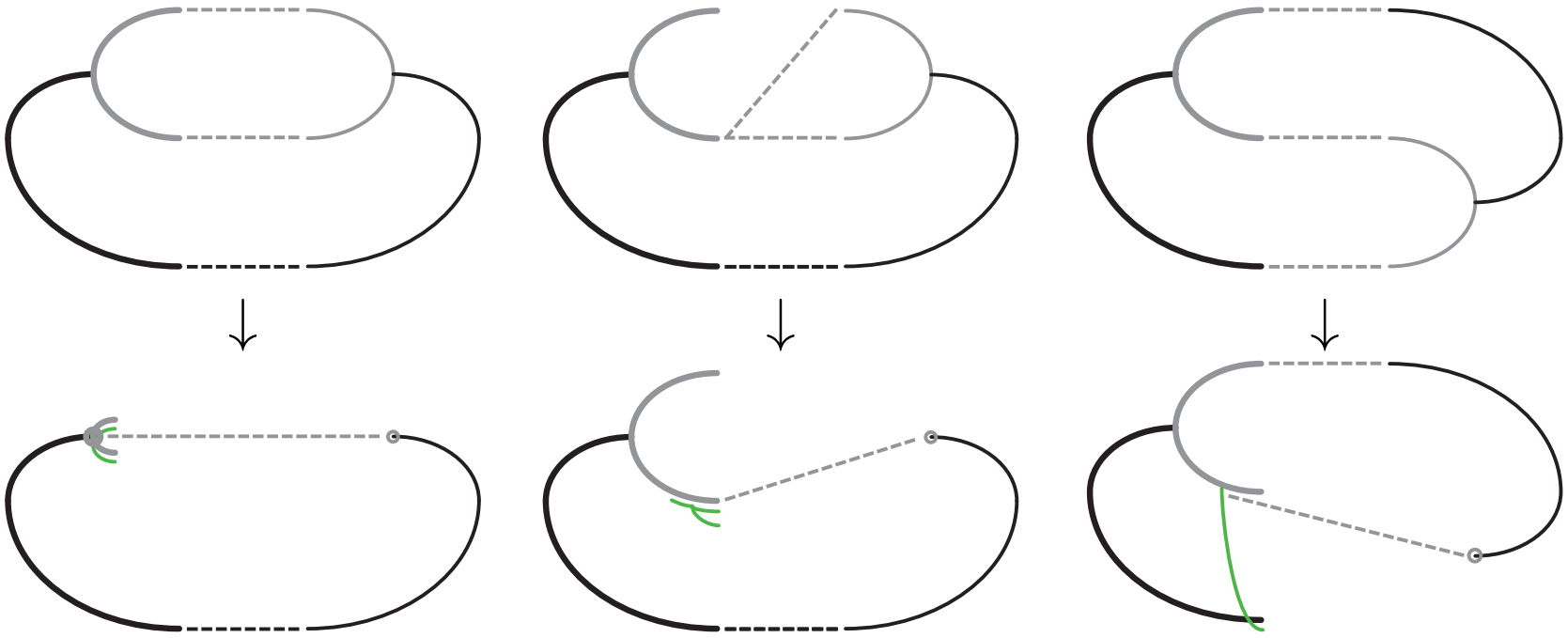
- Runs in reasonable time and guarantees minimal cost
- Practical for instances up to 40x40
- Shows Jane's very good accuracy
- Reveals some cases where even Jane fails
- Available as CPLEX solver

Tree Collapse

Drinkwater & Charleston

Tree Collapse

A different approach to solving this problem can be made if we exploit some common patterns in cophylogenetic analysis: for example, these ones:

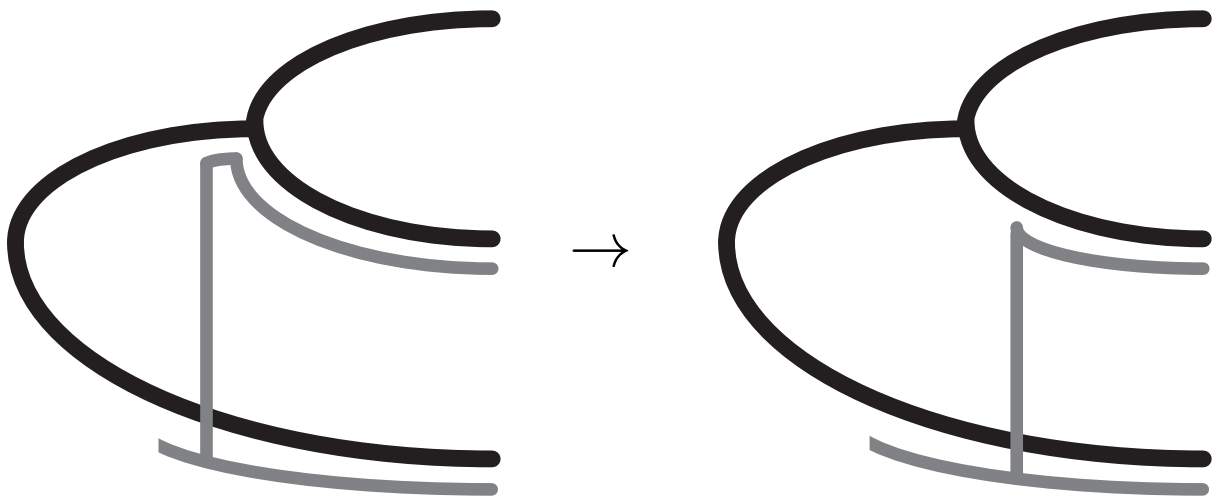


(There are four, more complex, patterns related to host switches. In the interests of time I'll skip these here.)

RightPush

The TreeCollapse pattern detection process can leave some nodes that are "too far back" in the host tree.

After the first phase, these are moved to the "right" (in the sense of the usual orientation) by RightPush:



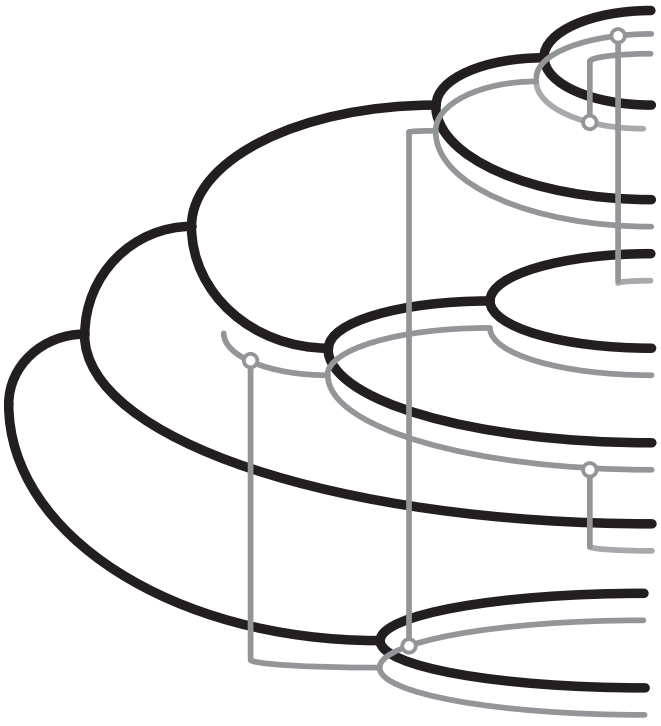
TreeCollapse accuracy

Table 1: Performance of the TreeCollapse Pattern Detection Framework over 150 real data sets.

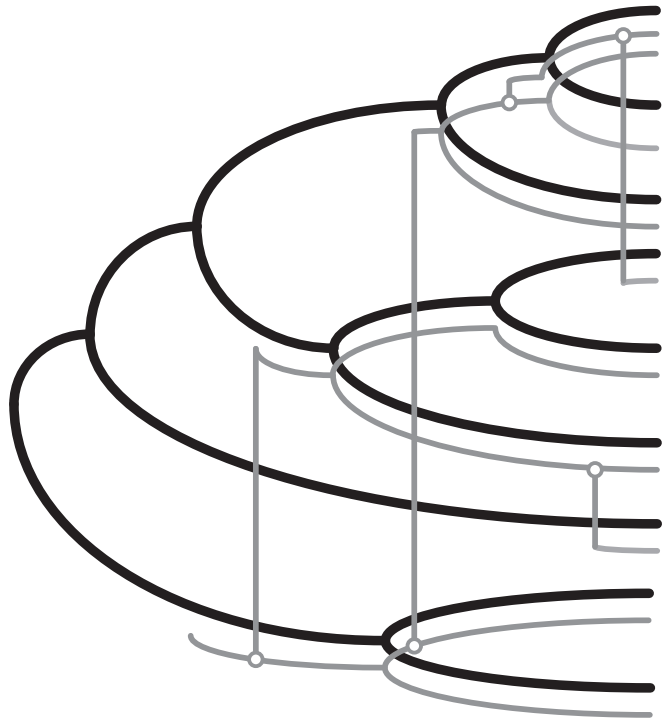
Distance from optimal	Number of Test Cases
0	113
1	17
2	5
3	5
4	4
5	3
≥ 6	3

(References and data available from the authors' web page)

TreeCollapse example



Jane



TreeCollapse

(One of the rare cases where TC actually beat Jane.)

TreeCollapse speed

TC is *linear* ($O(n)$), in both time and space, in the total number of nodes in both trees.

This complexity uses an application of the Level Ancestor Problem, for which, with (linear time and space) pre-processing, queries can be answered in O(1).

HOWEVER

TreeCollapse requires fixed host node ordering.

We use a Genetic Algorithm meta-heuristic such as is used in the Jane program, to search over host node orderings.

Widespread Parasites

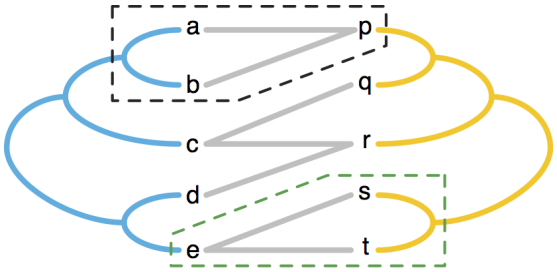
. . . because parasites & pathogens aren't that particular

Widespread Parasites

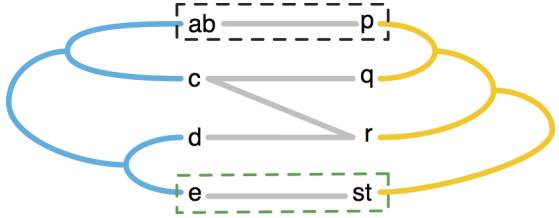
Not all parasites have a single host:

- Parasites often have complex life cycles
- Parasites/pathogens are frequently NOT highly host-specific, and can be found across several (un)related species [5]
- It's *difficult* to measure host specificity for these analyses.

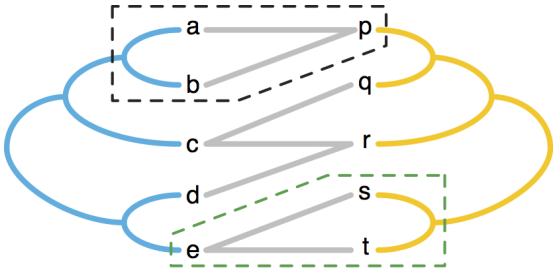
Introducing Failure To (Co)Diverge



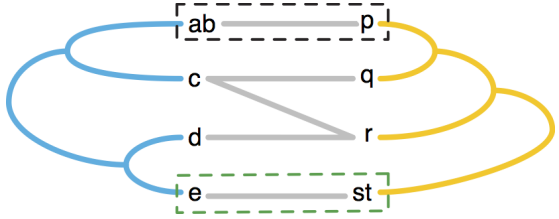
Monophyletic clades can be collapsed with no problem, to produce failure-to-(co)diverge or duplication $\rightarrow$



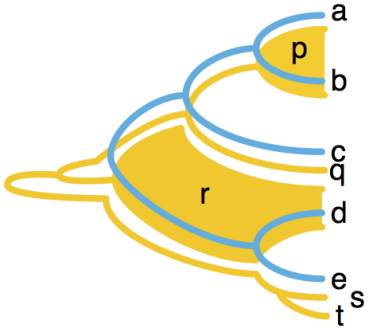
Introducing Failure To (Co)Diverge

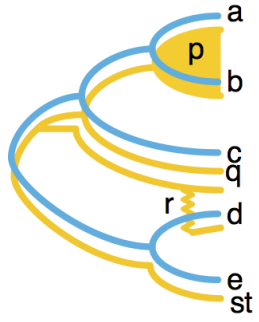


Monophyletic clades can be collapsed with no problem, to produce failure-to-(co)diverge or duplication →



But when the hosts of a parasite are less well related this causes problems:

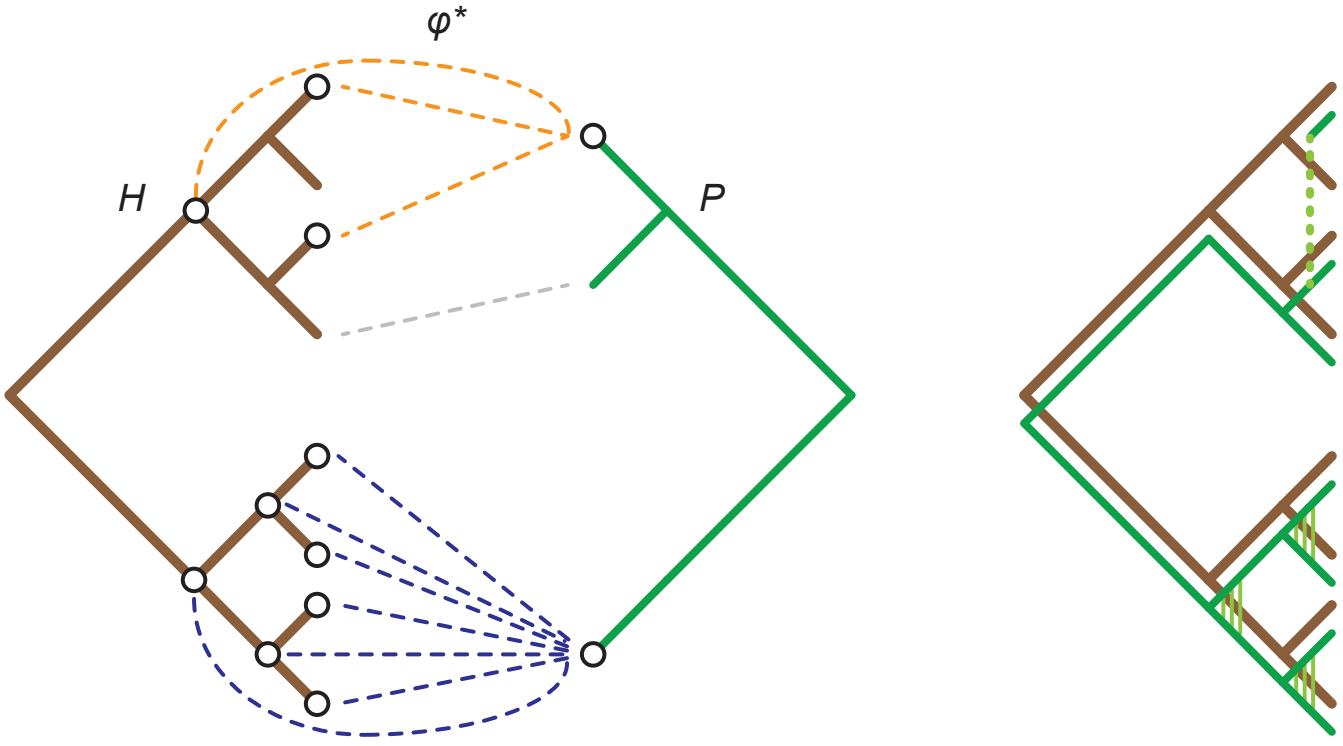




The solution currently (in Jane) is to push back FTDs to the common ancestor, inducing many losses$^{\{2\}}$.

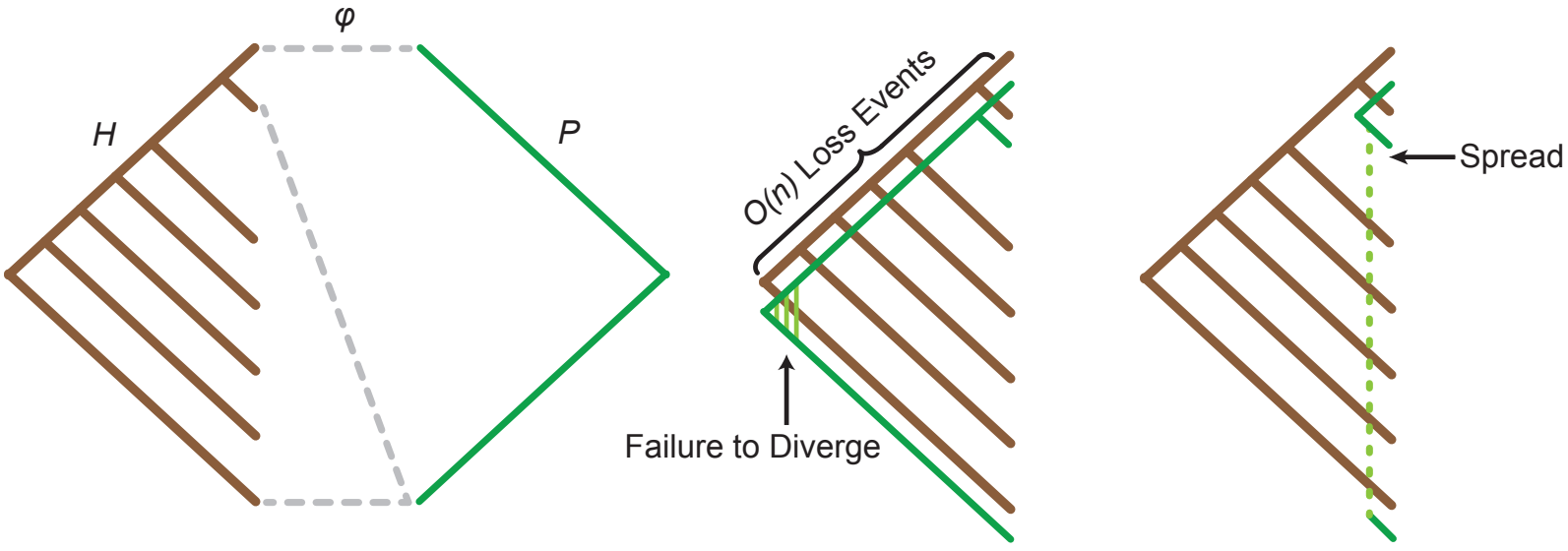
We propose a new event, "spread" (*sensu* Qiao) or "infestation" (*sensu* Libeskind-Hadas)

Spread events permit lower total cost



We can resolve the parasite tree in line with the host tree for widespread parasites, but this *doesn't* mean we are favouring a "codivergent" history.

Pushing FTD events back can be badly non-optimal



Cheeta

Earlier I mentioned an $O(n^3)$ solution to mapping P into H.

This method maps parasite nodes to host edges [2].

We have created another approach that is $O(n^3)$, mapping parasite nodes to host *nodes* — permitting dealing with widespread parasites with a further modification that's $O(n^4)$.

We still use the genetic algorithm to hunt for host node orderings.

We can solve the problem with widespread parasites by introducing the two new events "failure to (co)diverge" and "spread".

In keeping with the Jungle theme (leading to Tarzan and Jane software) we name our algorithm *Cheeta*.

Cheeta assumptions

It has been said that there is no problem that cannot be resolved by the judicious application of high explosives.

Adding widespread parasites to the cophylogeny reconstruction problem can be considered as a very *un*judicious application.

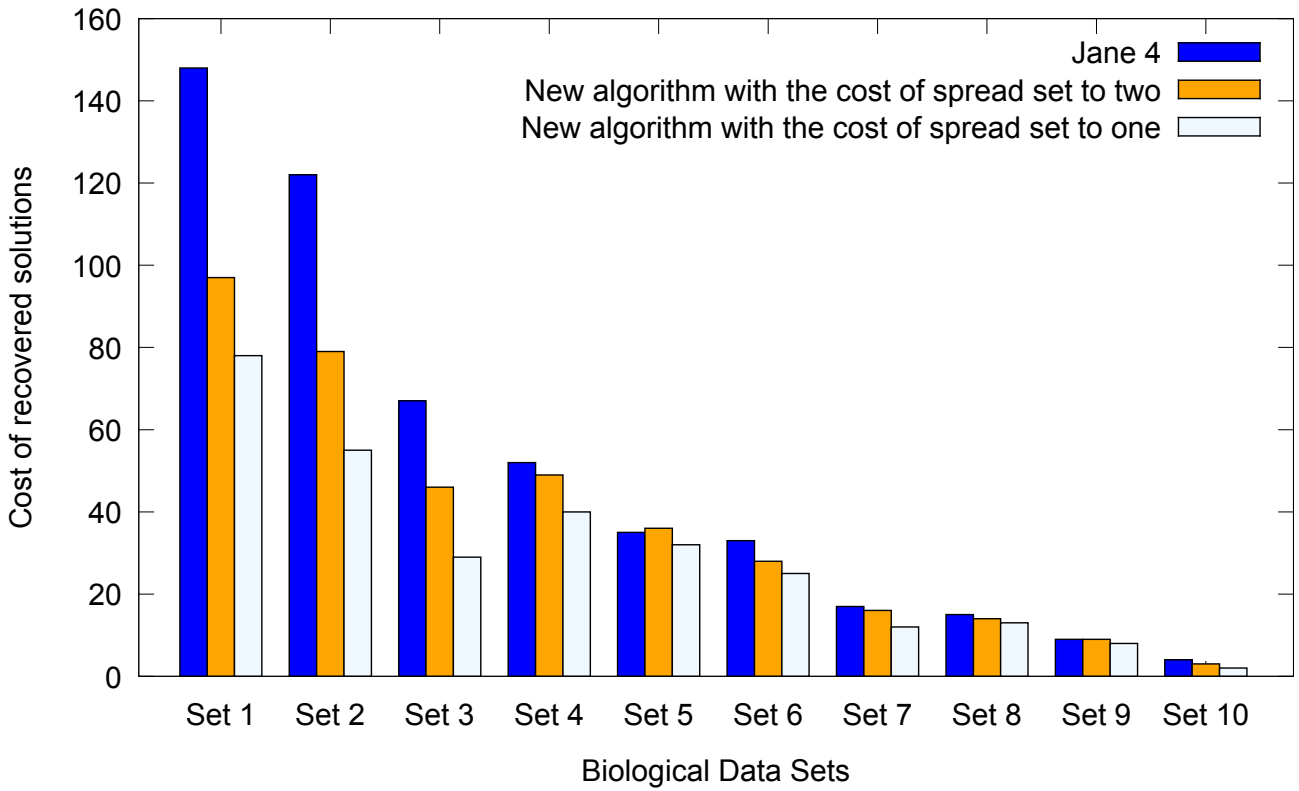
To avoid everything becoming unmanageable we make a *major simplifying assumption*: that

> *once a parasite has gained the ability to be widespread, that ability is retained.*

This assumption enables us to keep the solution to the reconstruction problem *polynomial*.

Cheeta results

Running our algorithm on the instances with widespread parasites we find a very good improvement on total cost.



References to data sets available on request

Cheeta results (now in a table!)

Table 2: Best solutions found with Cheeta and Jane

Data set	Jane's Best scheme 1 / scheme 2	Cheeta's Best scheme 1 / scheme 2
DS4 $\{11\}$	148 / 148	78 / 98
DS9 $\{10\}$	122 / 122	55 / 79
DS1 $\{3\}$	67 / 67	29 / 46
DS3 $\{15\}$	52 / 52	40 / 49
DS5 $\{1\}$	35 / 35	32 / 36
DS2 $\{14\}$	33 / 33	25 / 28
DS10 $\{16\}$	17 / 17	12 / 16
DS8 $\{6\}$	15 / 15	13 / 14
DS7 $\{4\}$	9 / 9	8 / 9
DS6 $\{13\}$	4 / 4	2 / 3

On average we see approximately 41% decrease in overall cost with our method.

Summary

- It is possible to get manageable speed even with exact methods, and reasonable accuracy with heuristics.
- We have built on previous work to create a better ILP and fast optimal mapping using dynamic programming;
- We have also created a novel pattern-finding method (TreeCollapse), which was inspired by Ronquist's work in 2002[{?}]
- It is possible even to handle the widespread parasites problem.

Thanks to

This work would not be possible without the talent of some wonderful people from my group & beyond:

Jennifer Hoyal Cuthill Sh**loads of work

Angela Qiao Widespread parasites and the "split and solve" approach

Bin Zhou Integer Linear Programming solution to cophylogeny mapping

Ben Drinkwater TreeCollapse and Cheeta

Lynden Shields Host specificity

Ran Libeskind-Hadas Computational Complexity and many excellent conversations

It was also supported by a grant from the Australian Research Council (DP1094981).

The End

J. Banks, R. Palma, and A. Paterson.

Cophylogenetic relationships between Penguins and their Chewing Lice.
Journal of evolutionary biology, 19(1):156–166, 2005.

C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas.

Jane: a new tool for the cophylogeny reconstruction problem.
Algorithms Mol Biol, 5(16):1–10, 2010.

S. Gómez-Acevedo, L. Rico-Arce, A. Delgado-Salinas, S. Magallón, and L. E. Eguiarte.

Neotropical mutualism between *Acacia* and *Pseudomyrmex*: Phylogeny and Divergence Times.
Molecular Phylogenetics and Evolution, 56(1):393–408, 2010.

M. S. Hafner, J. W. Demastes, T. A. Spradling, and D. L. Reed.

Cophylogeny between Pocket Gophers and Chewing Lice.
Tangled Trees: Phylogeny, Cospeciation, and Coevolution. University of Chicago Press, Chicago, pages 195–220, 2003.

J. Hoyal Cuthill and M. Charleston.

A simple model explains the dynamics of preferential host switching among mammal rna viruses.
Evolution, 2012.

A. P. Jackson and M. A. Charleston.

A cophylogenetic perspective of RNA-virus evolution.
Mol. Biol. Evol., 21(1):45–57, 2004.

R. Libeskind-Hadas.

Who is jane?
`http://www.cs.hmc.edu/~hadas/jane/Jane1/index.html`, September 2010.

R. Libeskind-Hadas.

Jane 4.
`http://www.cs.hmc.edu/~hadas/jane`, May 2013.

R. Libeskind-Hadas and M. 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.


A. Lockyer, P. Olson, P. Ostergaard, D. Rollinson, D. Johnston, S. Attwood, V. Southgate, P. Horak, S. Snyder, T. Le, et al.
The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of Schistosoma Weinland, 1858.
Parasitology, 126(3):203–224, 2003.


M. J. McLeish and S. Van Noort.
Codivergence and multiple Host Species use by Fig Wasp Populations of the Ficus Pollination Mutualism.
BMC Evolutionary Biology, 12(1):1, 2012.


Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas.
The cophylogeny reconstruction problem is np-complete.
Journal of Computational Biology, 2010.
doi:10.1089/cmb.2009.0240.


A. M. Paterson and R. Poulin.
Have Chondracanthid Copepods co-speciated with their Teleost hosts?
Systematic Parasitology, 44(2):79–85, 1999.


G. Refrégier, M. Le Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. Hood, and T. Giraud.
Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation.
BMC Evolutionary Biology, 8(1):100, 2008.


M. D. Sorenson, C. N. Balakrishnan, and R. B. Payne.
Clade-limited colonization in brood parasitic finches (vidua spp.).
Systematic Biology, 53(1):140–153, 2004.


J. D. Weckstein.
Biogeography explains Cophylogenetic patterns in Toucan Chewing lice.
Systematic Biology, 53(1):154–164, 2004.