Reticulate Evolution

Charles Semple

Biomathematics Research Centre

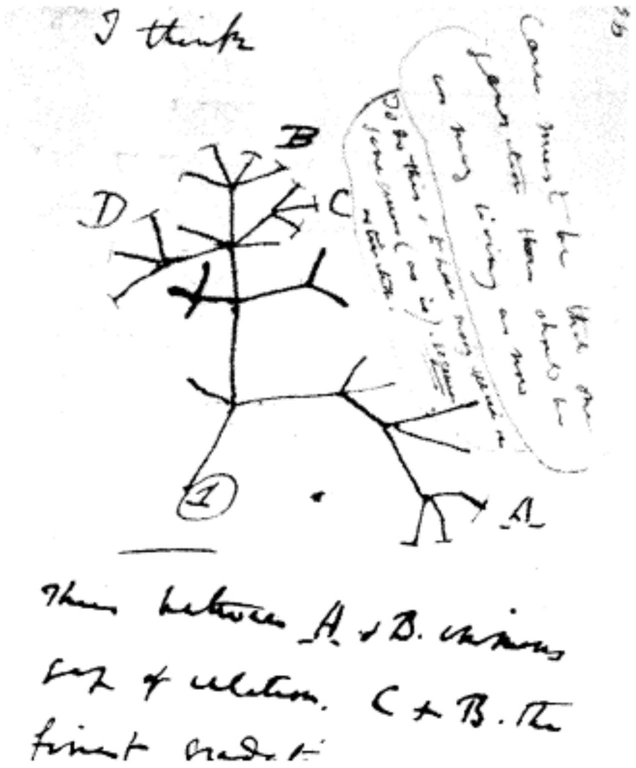
Department of Mathematics and Statistics

University of Canterbury, New Zealand

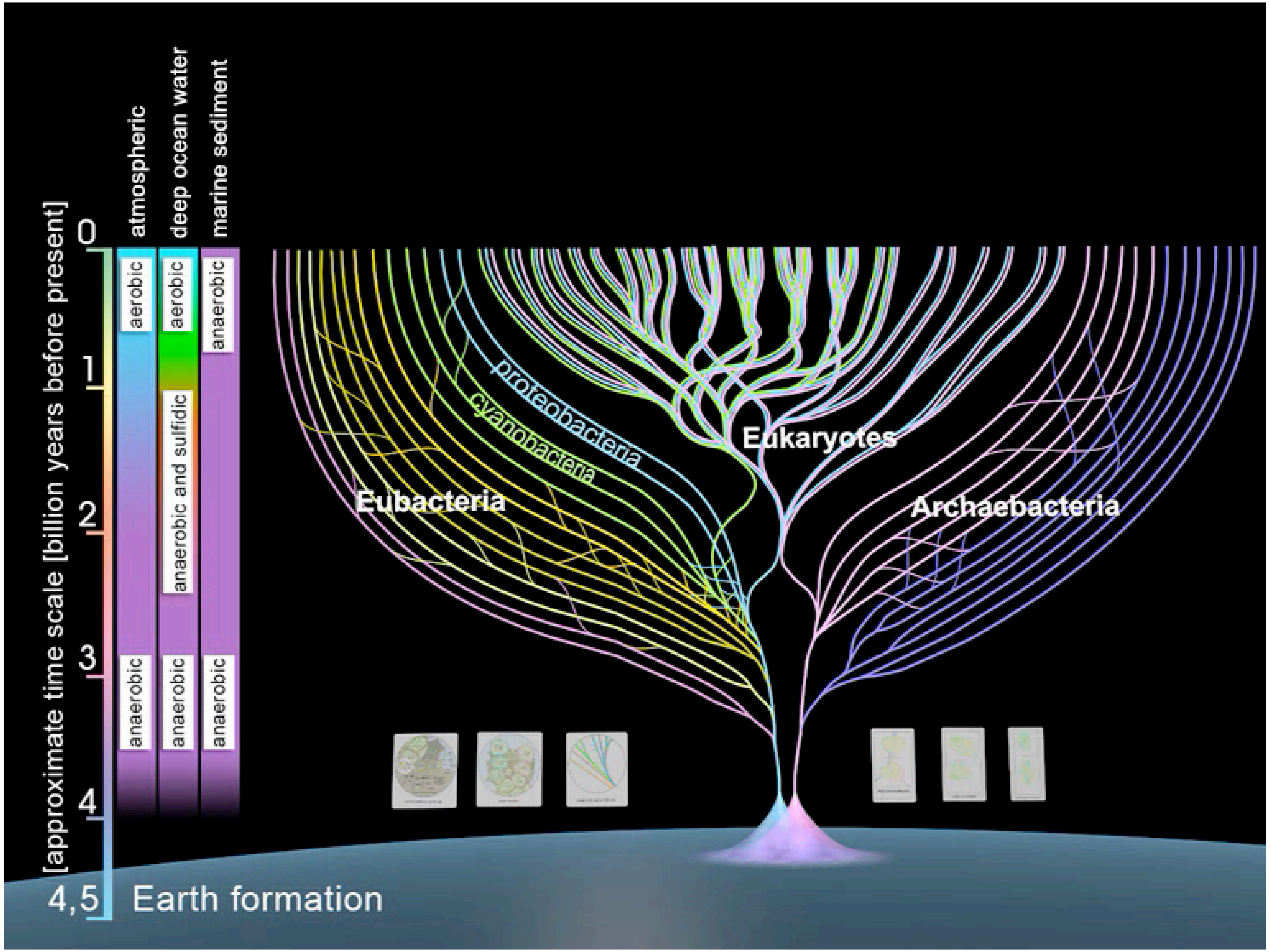
&

Allan Wilson Centre for Molecular Ecology and Evolution

Charles Darwin, 1837


I think
B
C
D
1
A
Then between A & B. immense gap of relation. C & B. the finest gradation

- Evolution is not always tree-like.
- Reticulation events cause species to be a mixture of genes from different ancestors.
- Evolutionary history is better represented using a rooted digraph.



Bill Martin, 2004

Basic Problem

A fundamental problem for evolutionary biologists:

Given an initial set of data that correctly repesents the tree-like evolution of different parts of various species genomes,

what is the smallest number of reticulation events required that simultaneously explains the variation in this collection?

How significant has the effect of hybridisation been on New Zealand flora?

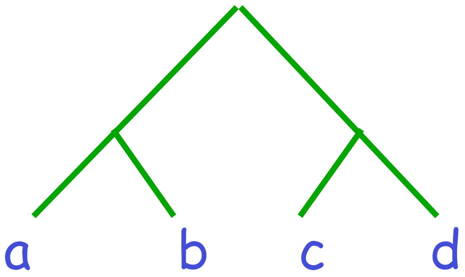
Some Terminology

A rooted binary phylogenetic X-tree is a rooted tree in which the root has degree 2, all other interior vertices have degree 3, and the set of leaves is X.

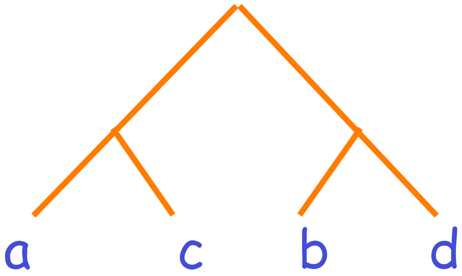
A hybrid phylogeny on X is a rooted acyclic digraph in which the root has out-degree at least two, the in-degree of any vertex is at most two, and the set of vertices of out-degree 0 is X.

H displays T if T can be obtained from a rooted subtree of H by suppressing degree-2 vertices.

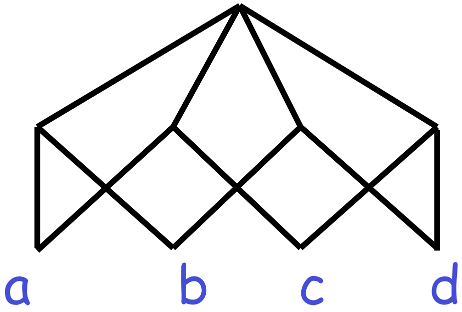
Examples: hybrid phylogenies



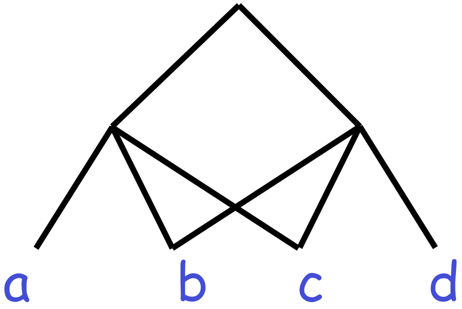
T_1



T_2

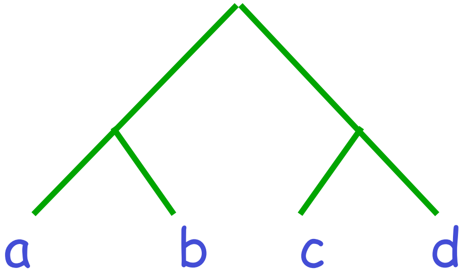


H_1

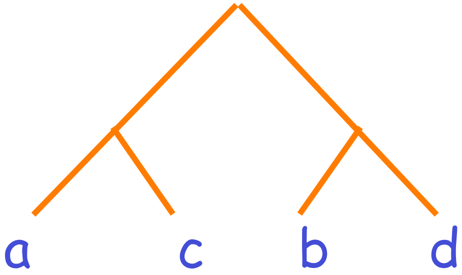


H_2

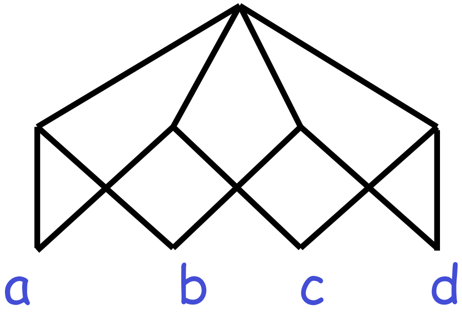
Examples: hybrid phylogenies



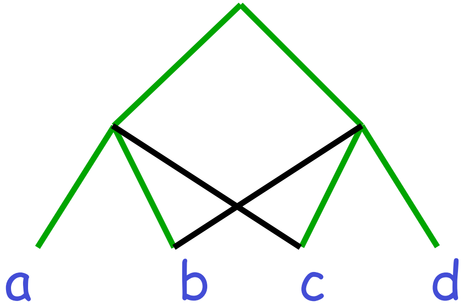
T_1



T_2

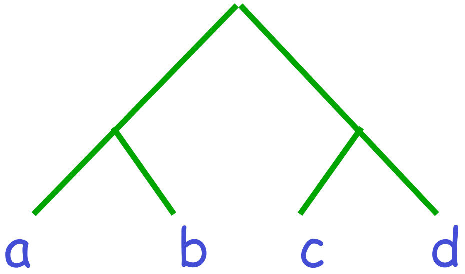


H_1

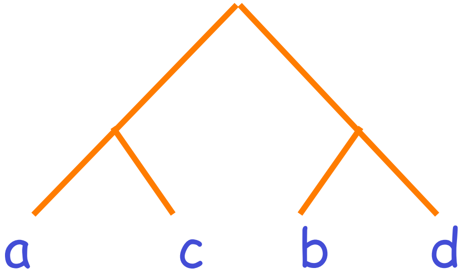


H_2

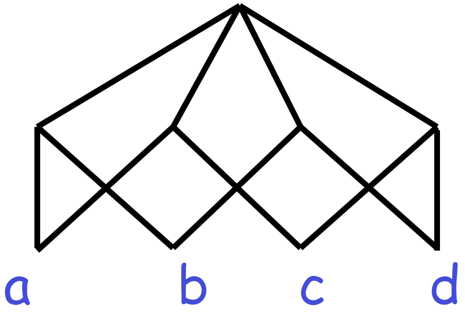
Examples: hybrid phylogenies



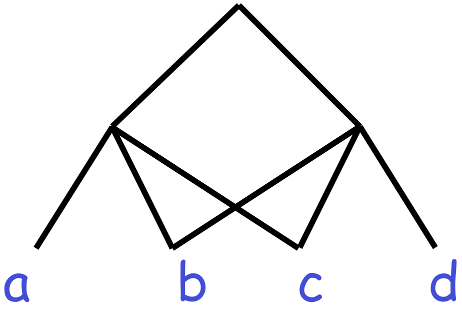
T_1



T_2

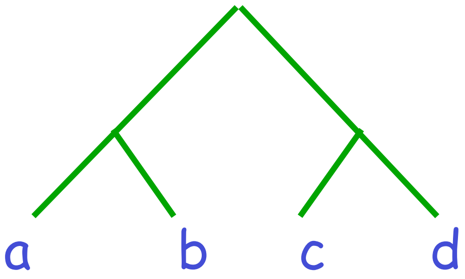


H_1

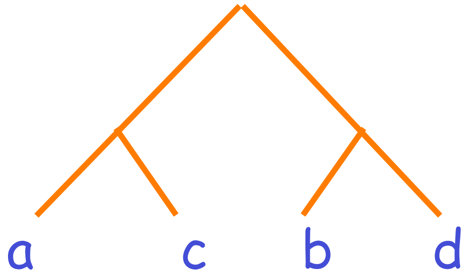


H_2

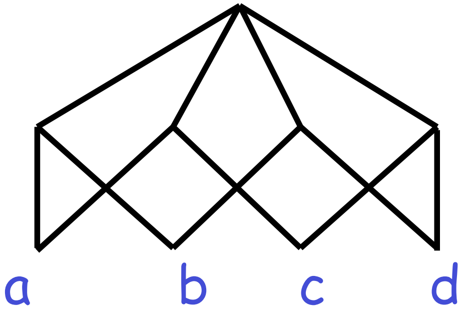
Examples: hybrid phylogenies



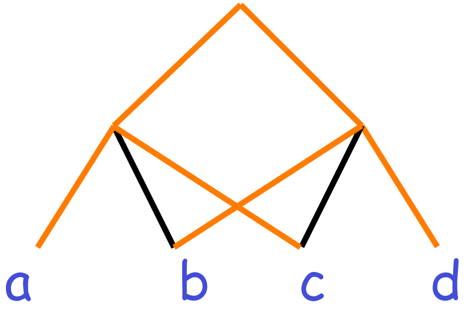
T_1



T_2



H_1



H_2

The Two Tree Problem

MINIMUM HYBRIDISATION

Instance: Two rooted binary phylogenetic trees S and T.

Goal: Find a hybrid phylogeny H that displays S and T, and minimises the number of hybridisation vertices.

Measure: The number of hybridisation vertices in H.

Notation: Use $h(S, T)$ to denote this minimum number.

Theorem. (Bordewich, Semple 2005)
MINIMUM HYBRIDISATION is NP-hard.

Because of NP-hardness:

- Identifying mathematical structures;
- Developing fast algorithms for special cases of the problem;
- Providing ways to bound the smallest number of hybridisation events.

Subtree Prune and Regaft

For a binary phylogenetic tree T, we plant T by adjoining an edge e to the root p of T and relocating p to the other end of e.

A binary phylogenetic T has been obtained from S by a subtree prune and regraft operation if it has been obtained from $P(S)$ by cutting a subtree of S and then reattaching this subtree to the resulting tree.

Notation: Use $d_{SPR}(S, T)$ to denote the minimum number of (single) SPR operations to obtain T from S.

For two trees S and T, there is a (seemingly) close connection between $d_{SPR}(S, T)$ and $h(S,T)$.

This closeness is recognised in a number of papers. For example, Hein (1990); Hein, Jiang, Wang, Zhang (1996); Maddison (1997); Nakhleh, Warnow, Linder (2004).

Theorem. (Baroni, Grünewald, Moulton, Semple 2005)

For all $n \geq 4$, there is a particular choice of S and T such that

$$d_{SPR}(S, T)=2 \text{ and } h(S, T)=n-n/2,$$

where n is the size of the leaf set of S and T.

Agreement Forests

A forest of T is a disjoint collection of phylogenetic subtrees of P(T) whose union of leaf sets is X U p.

An agreement forest for S and T is a forest of both S and T.

A maximum agreement forest for S and T is an agreement forest for S and T of smallest size.

Theorem. (Bordewich, Semple 2004)

Let S and T be two binary phylogenetic X-trees. Then

$$d_{SPR}(S, T) = \text{size of maximum agreement forest} - 1.$$

Acyclic Agreement Forests

Let F be an agreement forest for S and T.

The root-descendancy graph D_F of F is the digraph with vertex set F and arc set

$\{(T_i, T_j)$: for either S or T, the root of T_i is an ancestor of the root of $T_j\}$.

F is an acyclic agreement forest for S and T if D_F is acyclic.

A maximum-acyclic agreement forest for S and T is an acyclic agreement forest for S and T of smallest size.

Theorem. (Baroni, Grünewald, Moulton, Semple 2005)

Let S and T be two binary phylogenetic X-trees. Then

$h(S, T)$ = size of maximum-<u>acyclic</u> agreement forest - 1.

Corollary.

Let S and T be two binary phylogenetic X-trees. Then

$$d_{SPR}(S,T) \leq h(S,T).$$

Theorem. (Baroni, Grünewald, Moulton, Semple 2005)

Let S and T be two binary phylogenetic X-trees. Then

$h(S, T)$ = size of maximum-acyclic agreement forest - 1.

Corollary.

Let S and T be two binary phylogenetic X-trees. Then

$$d_{SPR}(S,T) \leq h(S,T) \leq n-2,$$

where $n=|X|$.

Hybrid Phylogenies from Acyclic Agreement Forests

- Let F be an acyclic agreement forest for S and T.
- Let T_p, T_1, T_2, …, T_k be an acyclic ordering of D_F.
- Let $H_0 = T_p$ and set $i = 1$.
- Attach T_i to H_{i-1} so that the resulting hybrid phylogeny H_i displays
  - T restricted to the union of the label sets of T_p, T_1, …, T_i and
  - S restricted to the union of the label sets of T_p, T_1, …, T_i.
- Increase i by 1 and repeat.

Some Remarks

Computing $d_{SPR}(S, T)$:

- NP-hard and APX-hard (Bordewich, Semple 2004);
- 3-approximation algorithm (Rodrigues, Sagot, Wakabayashi 2001);
- Fixed parameter tractable (Bordewich, Semple 2004).

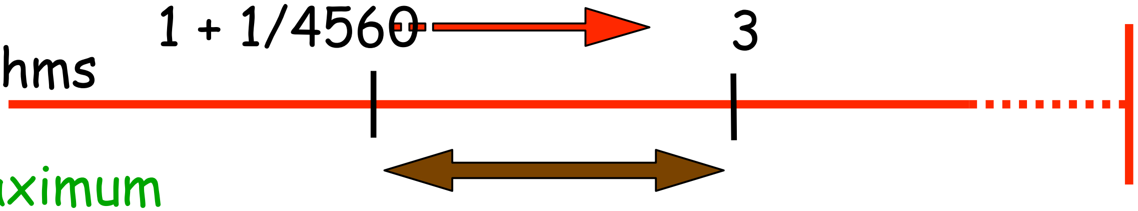
Computing $h(S, T)$:

- NP-hard and APX-hard (Bordewich, Semple 2005);
- Approximation algorithm?
- Fixed parameter tractable?

Approximation Algorithms

An r-approximation algorithm A for an optimisation problem means that the size of the feasible solution outputted by A, when applied to any instance I, is at most r times opt(I).

Example. If r=3, then any feasible solution returned by A when applied to I is at most 3 times the optimal solution.



Exact algorithms r=1

Maximum independent set in planar graphs (PTAS)

(B, S 2004)

1 + 1/4560

Increasing r

MINIMUM SPR

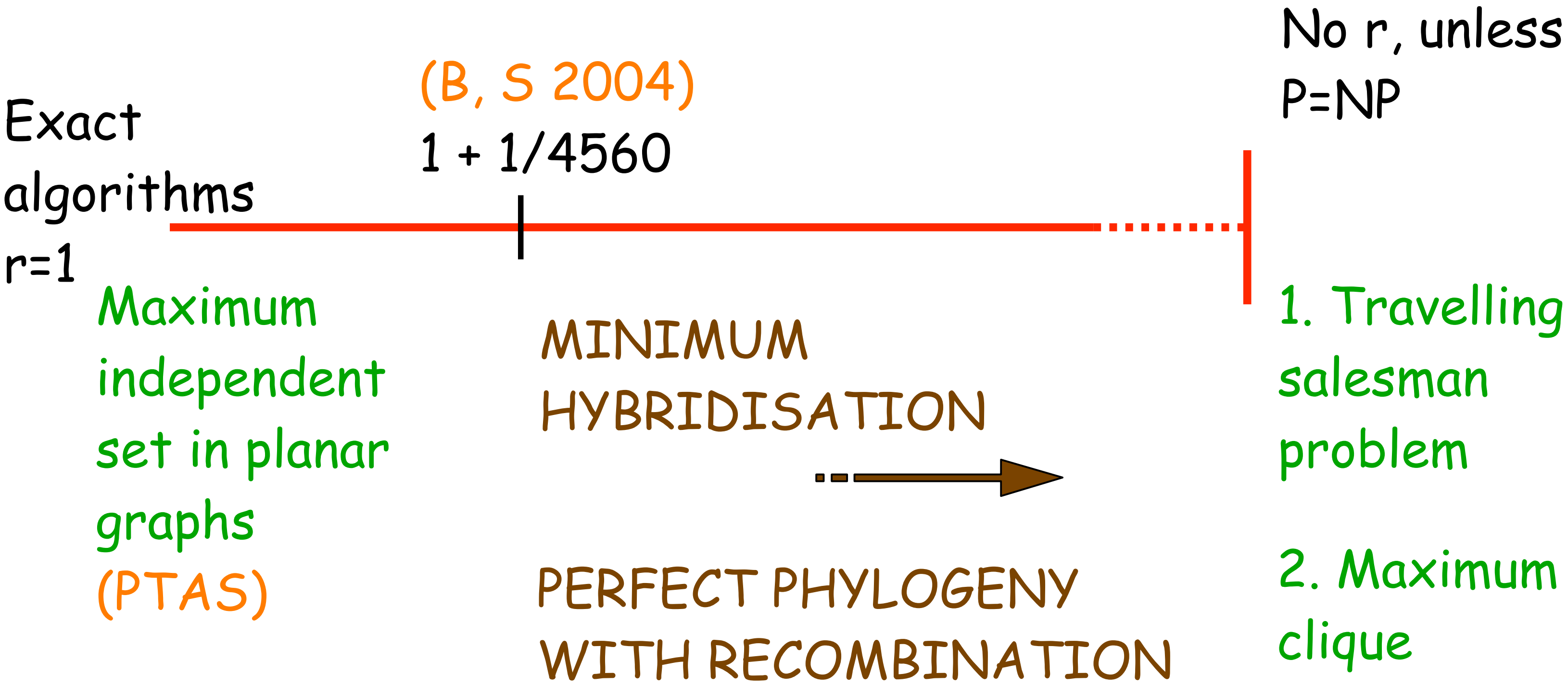
(R,S,W 2001)

3

No r, unless P=NP

1. Travelling salesman problem

2. Maximum clique



Exact algorithms r=1

Maximum independent set in planar graphs (PTAS)

(B, S 2004)

1 + 1/4560

MINIMUM HYBRIDISATION

PERFECT PHYLOGENY WITH RECOMBINATION

No r, unless P=NP

1. Travelling salesman problem

2. Maximum clique

The n Tree Problem

Two ways to count the hybridisation value of H:

1. Number of hybridisation vertices.
2. The sum of the indegree of v - 1 over all hybridisation vertices v.

   (Summing-up the number of additional parent vertices.)

An agreement forest for a collection P is a forest of each of the trees in P.

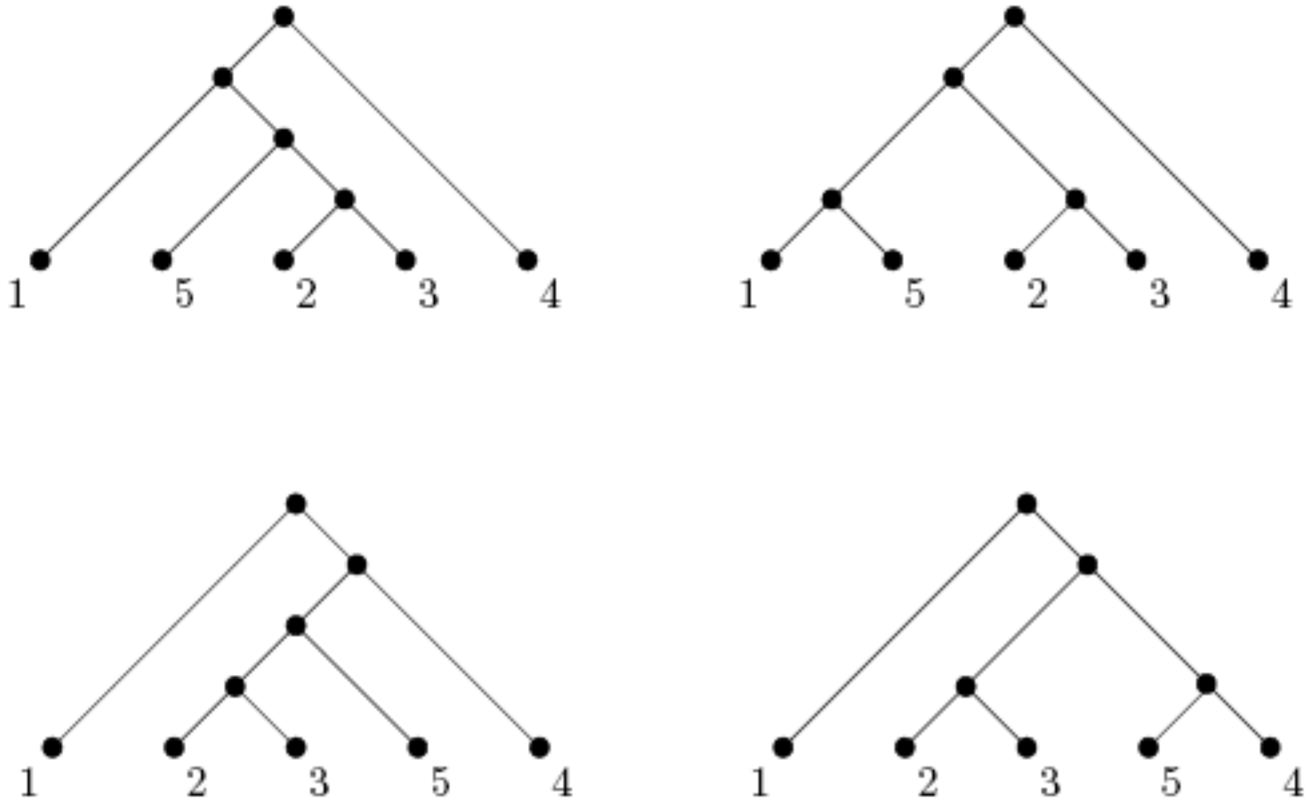
Analogously, we have the notion of a maximum-acyclic agreement forest for P.

Theorem.

Let P be a collection of binary phylogenetic X-trees. Then, using a type 1 hybrid count,

h(P) = size of maximum agreement forest - 1.

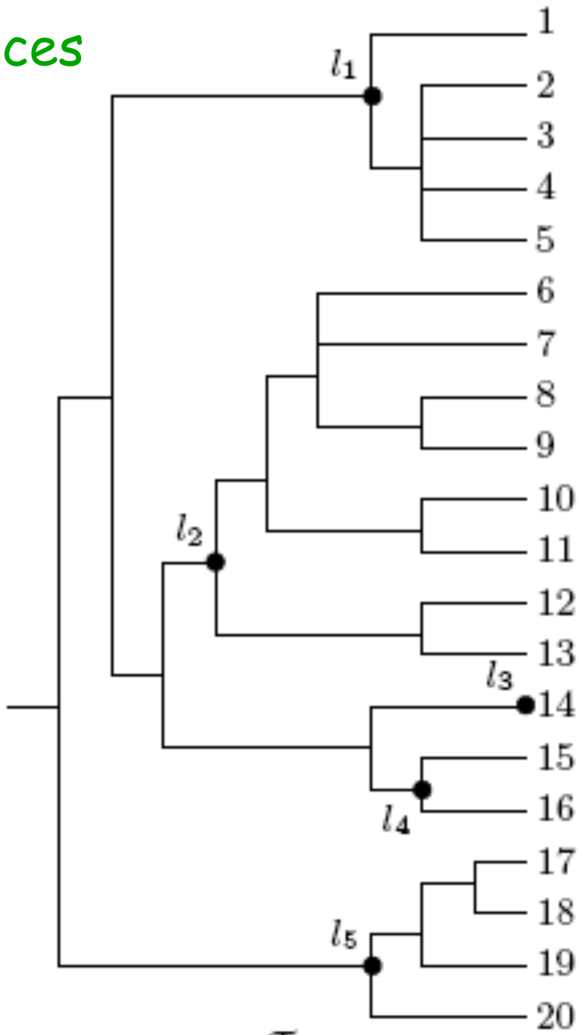
However, a type 2 hybrid count appears to be problematic.


1 5 2 3 4
1 5 2 3 4
1 2 3 5 4
1 2 3 5 4

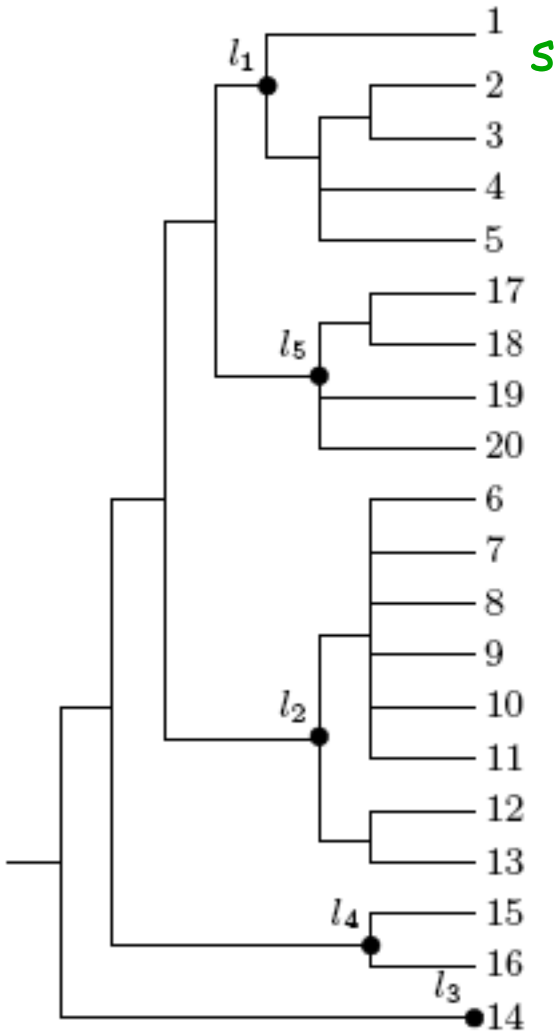
Real Data Example

nuclear ITS seqences

chloroplast sequences



$\mathcal{T}_1$



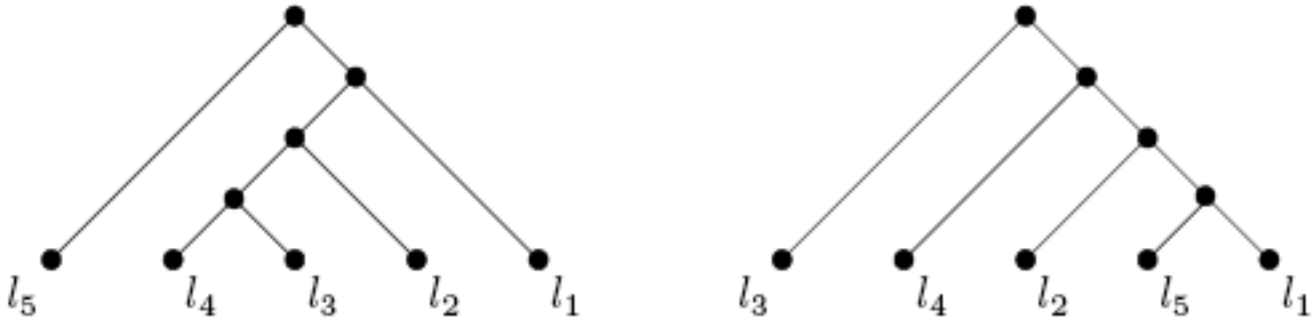
$\mathcal{T}_2$

Theorem. (Baroni 2004)

Let S and T be two binary phylogenetic X-trees. Suppose that A is a cluster of both S and T. Then

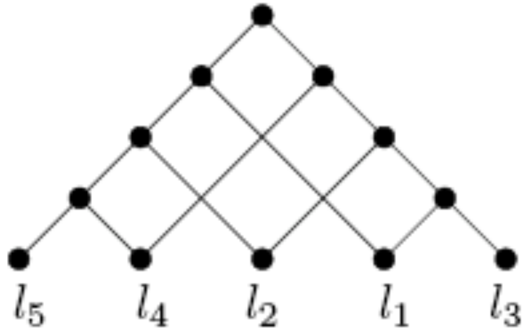
$$h(S, T)=h(S|A, T|A) + h(S_a, T_a),$$

where S_a and T_a are the rooted trees obtained by replacing the subtree with leaf set A with a new leaf a.

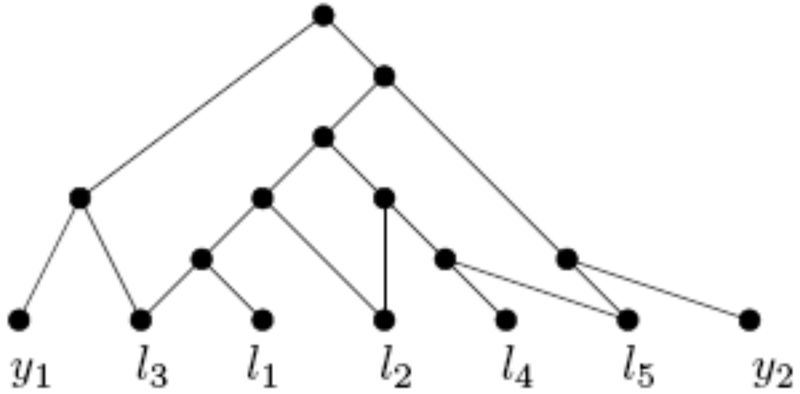


(a) Nuclear ITS sequences.

(b) Chloroplast sequences.



$\mathcal{H}$



$\mathcal{H}'$

No temporal labelling.

A temporal labelling, but includes additional taxa.

Real-Time Hybrids

A real-time hybrid phylogeny is one in which the vertices can be labelled with elements of the natural numbers N so that, for all v,

- if indegree of v is 1, then the element assigned to v is bigger than the element assigned to its parent, and
- if indegree of v is ≥2, then the element assigned to v is the same as each of its parents.

Such a labelling is called a temporal labelling.

Let H be a hybrid phylogeny on X with vertex set V.

Let π_H be the partition of V in which u and v are in the same part iff they are forced to have the same element of N assigned to them.

The time-descendancy digraph D_H of H is the digraph with vertex set π_H and arc set

$\{(A,B)$: if there is a tree edge (u,v) in H with u in A and v in $B\}$.

Which Hybrids are Real Time?

Theorem. (Baroni, Semple, Steel 2005)

Let H be a hybrid phylogeny. Then H is a real-time hybrid iff D_H is acyclic.

Simple algorithm:

- Let H be a hybid phylogeny, and suppose that D_H is acyclic.
- Let $A_1, A_2, \ldots, A_k$ be an acyclic ordering of D_H.
- For all i, assign the vertices in A_i the value i.
- The result assignment gives a temporal labelling of H.