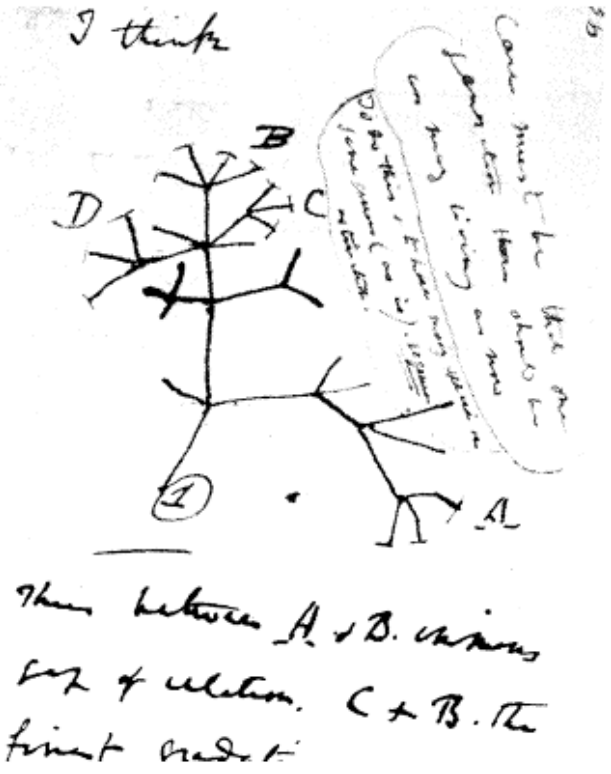
UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Fixed-Parameter Algorithms for the Subtree Distance Between Phylogenies

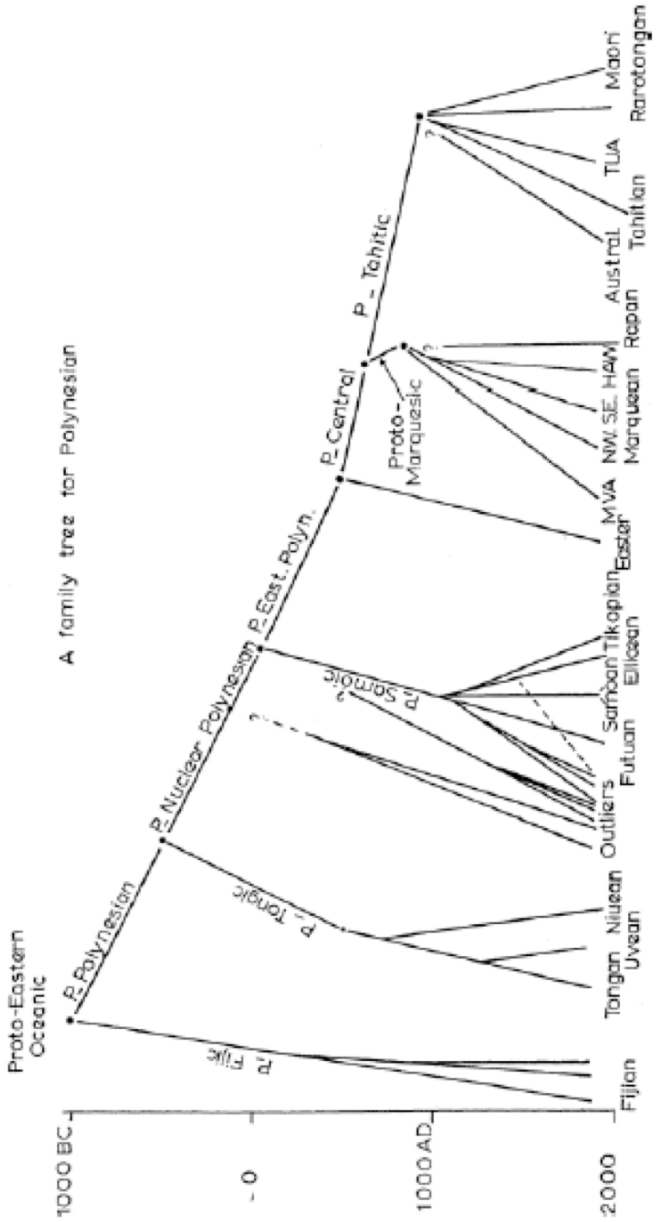
Charles Semple
Biomathematics Research Centre
Department of Mathematics and Statistics
University of Canterbury
New Zealand

Algorithmics Meeting, Napier 2008

ALLAN WILSON CENTRE



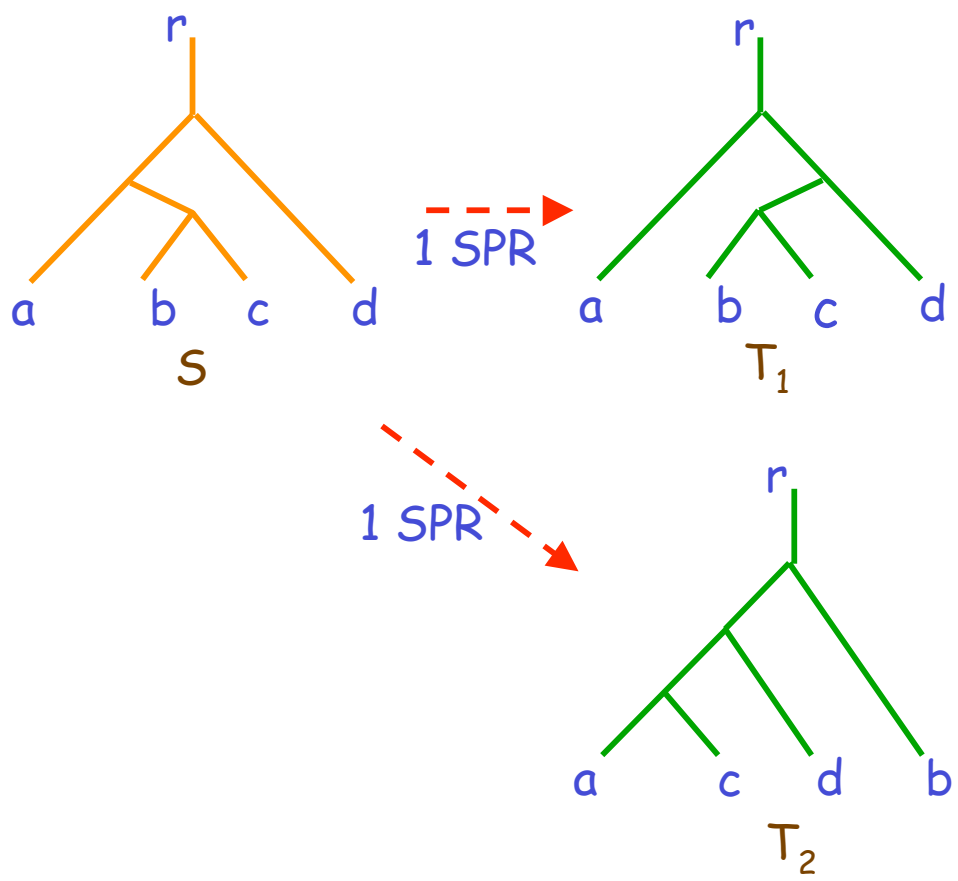
Charles Darwin, 1837



R C Green, 1966

Subtree Prune and Regraft (SPR)

Example.



Applications of SPR

Used

I. As a search tool for selecting the best tree in reconstruction algorithms.

II. To quantify the dissimilarity between two phylogenetic trees.

III. To provide a lower bound on the number of reticulation events in the case of non-tree-like evolution.

For II and III, one wants the minimum number of SPR operations to transform one phylogeny into another.

This number is the SPR distance between two phylogenies S and T.

The Mathematical Problem

MINIMUM SPR

Instance: Two rooted binary phylogenetic trees S and T.

Goal: Find a minimum length sequence of single SPR operations that transforms S into T.

Measure: The length of the sequence.

Notation: Use $d_{SPR}(S, T)$ to denote this minimum length.

Theorem (Bordewich, S 2004)

MINIMUM SPR is NP-hard.

Overriding goal is to find (with no restrictions) the exact solution or a heuristic solution with a guarantee of closeness.

Algorithms for NP-Hard Problems

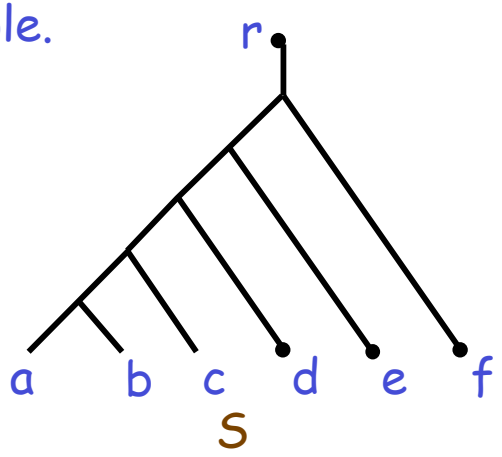
Fixed-parameter algorithms are a practical way to find optimal solutions if the parameter measuring the hardness of the problem is small.

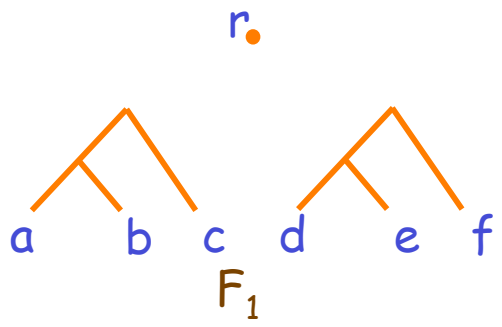
Polynomial-time approximation algorithms can efficiently find feasible solutions that are sometimes arbitrarily close to the optimal solution.

Agreement Forests

A forest of T is a disjoint collection of phylogenetic subtrees whose union of leaf sets is X U r.

Example.

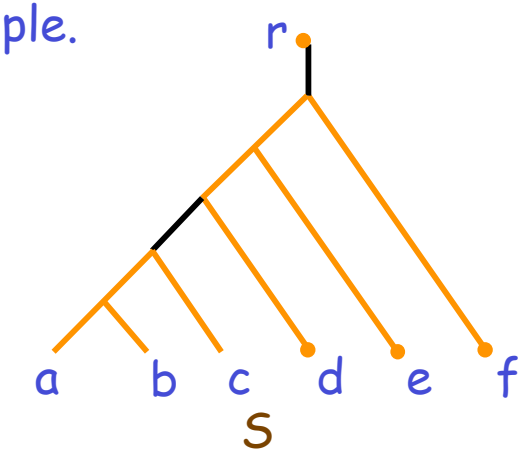


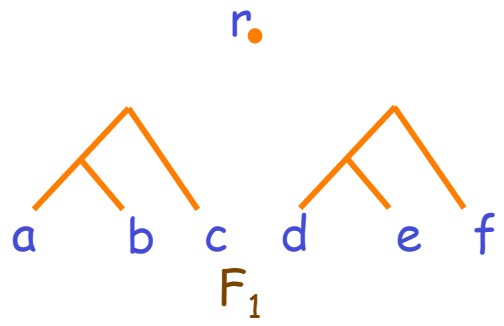


Agreement Forests

A forest of T is a disjoint collection of phylogenetic subtrees whose union of leaf sets is X U r.

Example.

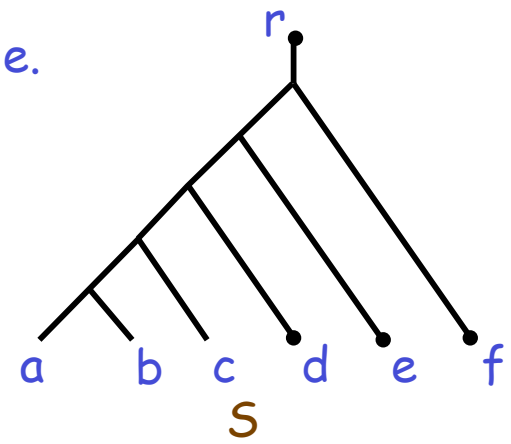


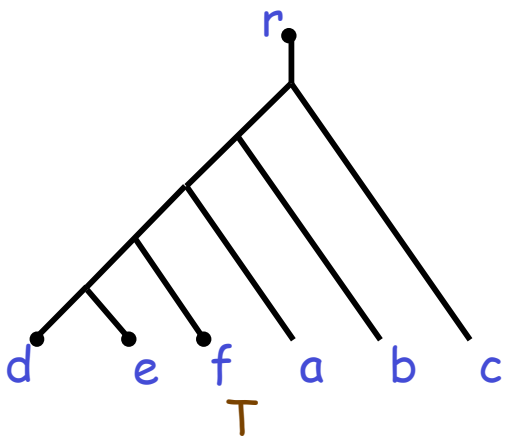


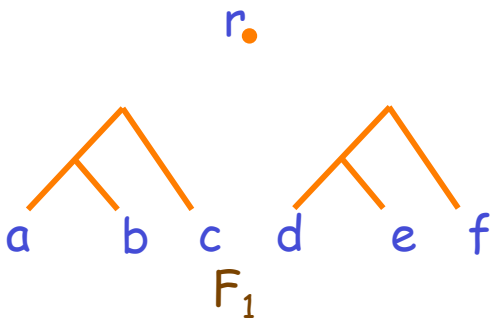
Agreement Forests

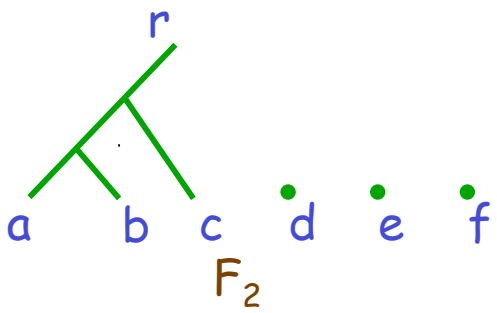
An agreement forest for S and T is a forest of both S and T.

Example.





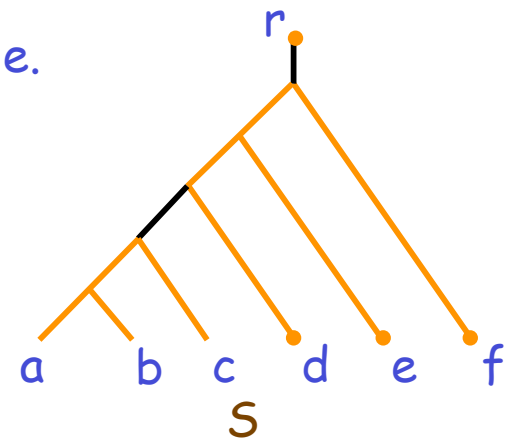


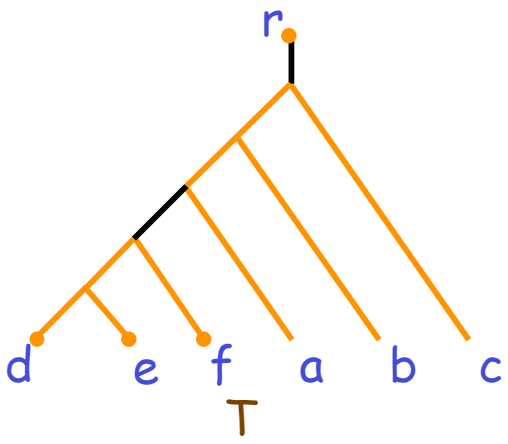


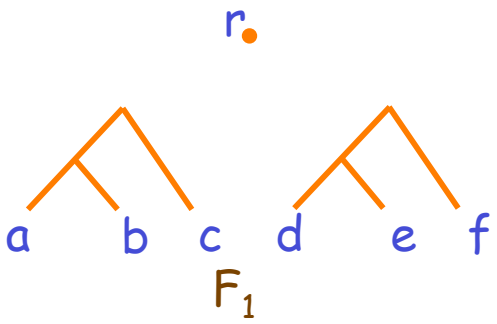
Agreement Forests

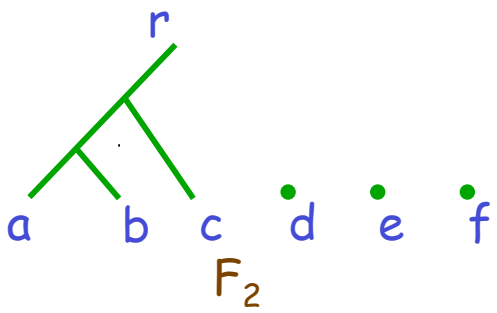
An agreement forest for S and T is a forest of both S and T.

Example.





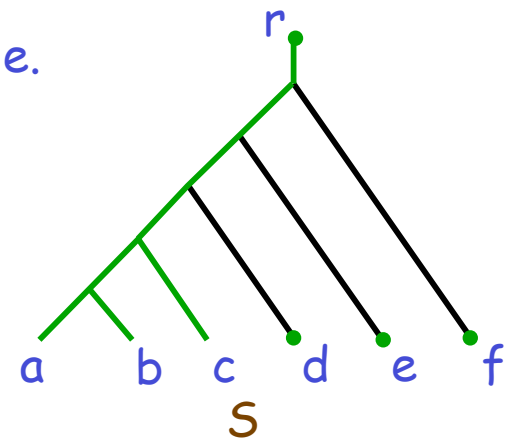


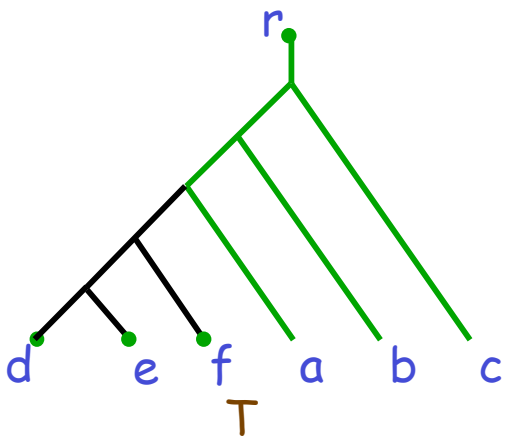


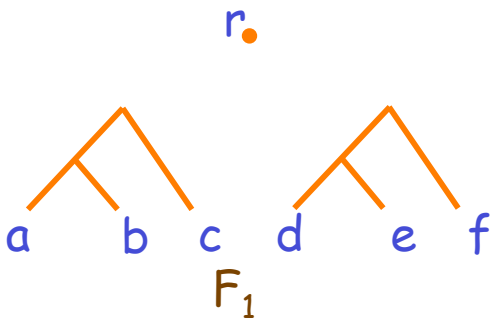
Agreement Forests

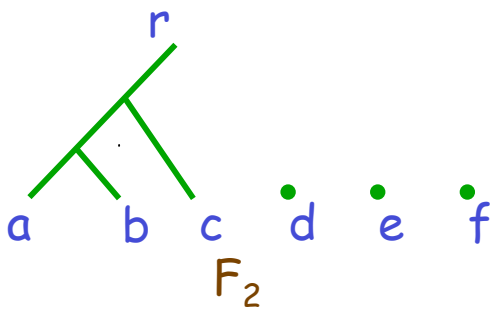
An agreement forest for S and T is a forest of both S and T.

Example.









Theorem. (Bordewich, S, 2004)

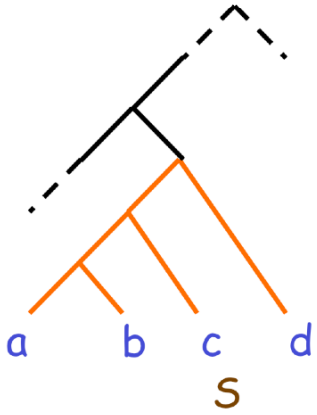
Let S and T be two binary phylogenetic trees. Then

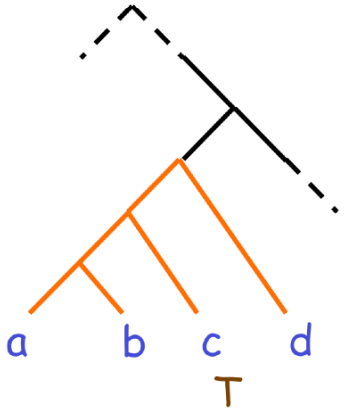
$d_{SPR}(S,T)$ = size of maximum-agreement forest - 1.

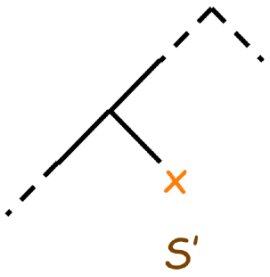
- It's fast to construct from a maximum-agreement forest for S and T a sequence of SPR operations that transforms S into T.

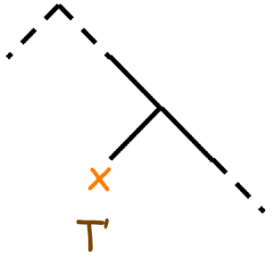
Reducing the Size of the Instance

Subtree Reduction

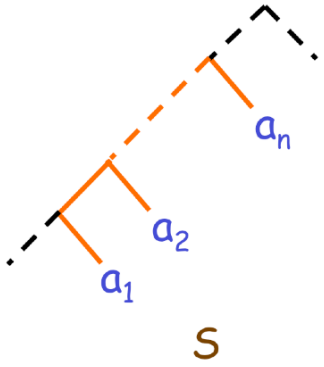


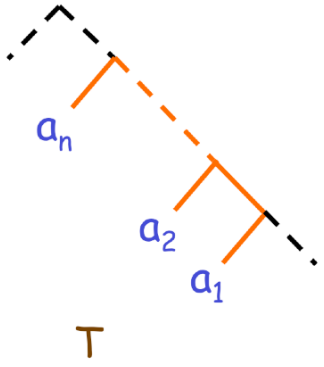


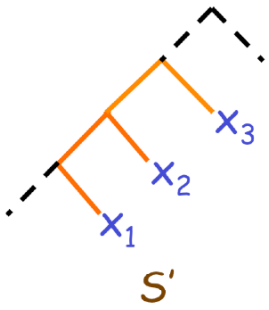


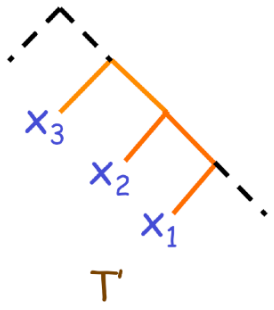


Chain Reduction









Fixed-Parameter Algorithms

The underlying idea is to find an algorithm whose running time separates the size of the problem instance from the parameter of interest.

One way to obtain such an algorithm is to reduce the size of the problem instance, while preserving the optimal value (kernalizing the problem).

Are the subtree and chain reductions enough to kernalize the problem?

Fixed-Parameter Algorithms

Lemma. If n′ denotes the size of the leaf sets of the fully reduced trees using the subtree and chain reductions, then

$$n' < 28d_{SPR}(S,T).$$

Corollary. (Bordewich, S 2004) MINIMUM SPR is fixed-parameter tractable.

1. Repeatedly apply the subtree and chain rules.
2. Exhaustively find a maximum-agreement forest by deleting edges in S and comparing with T.

Running time is $O((56k)^k + p(n))$ compared with $O((2n)^k)$, where $k=d_{SPR}(S,T)$ and $p(n)$ is the polynomial bound for reducing the trees using the subtree and chain reductions.

Modelling Hybridization Events with SPR Operations

Reticulation processes cause species to be a composite of DNA regions derived from different ancestors.

Processes include

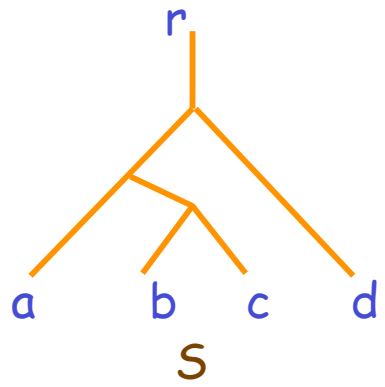
- horizontal gene transfer,
- hybridization, and
- recombination.

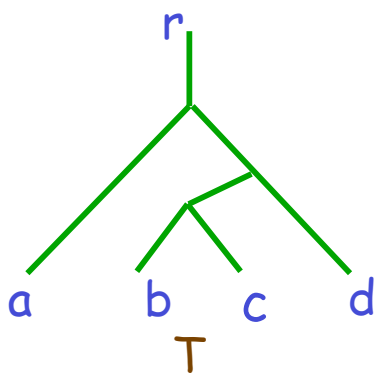
> *… molecular phylogeneticists will have failed to find the `true tree', not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot be properly represented as a tree.*
>
> Ford Doolittle, 1999
> (Dalhousie University)

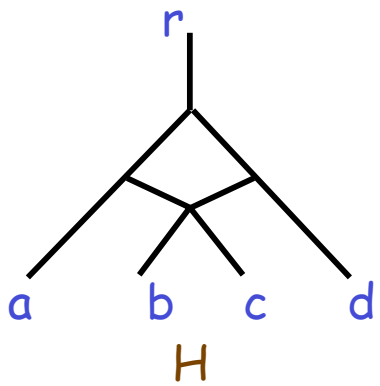
Modelling Hybridization Events with SPR Operations

A single SPR operation models a single hybridization event (Maddison 1997).

Example.



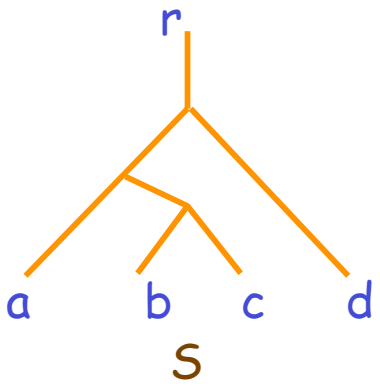


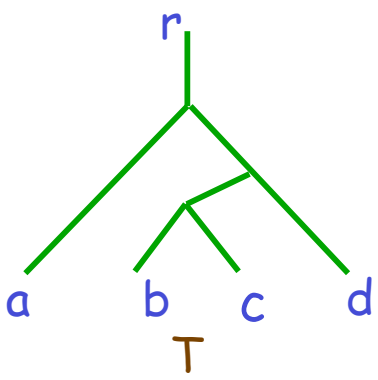


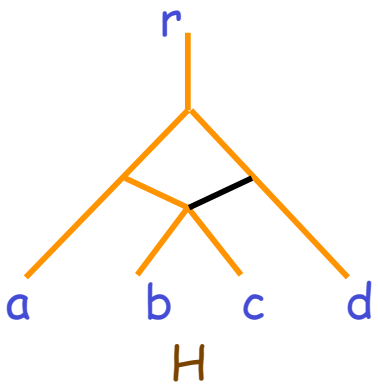
Modelling Hybridization Events with SPR Operations

A single SPR operation models a single hybridization event (Maddison 1997).

Example.



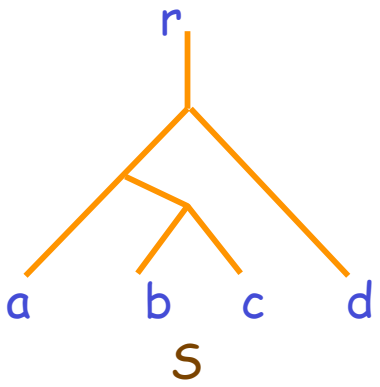


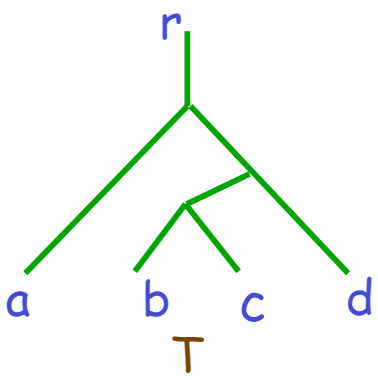


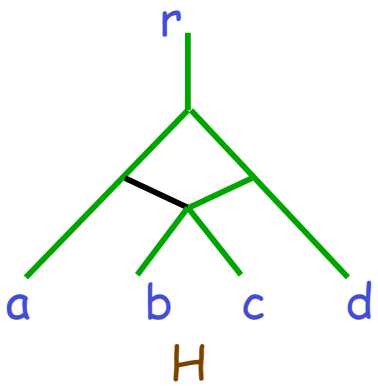
Modelling Hybridization Events with SPR Operations

A single SPR operation models a single hybridization event (Maddison 1997).

Example.



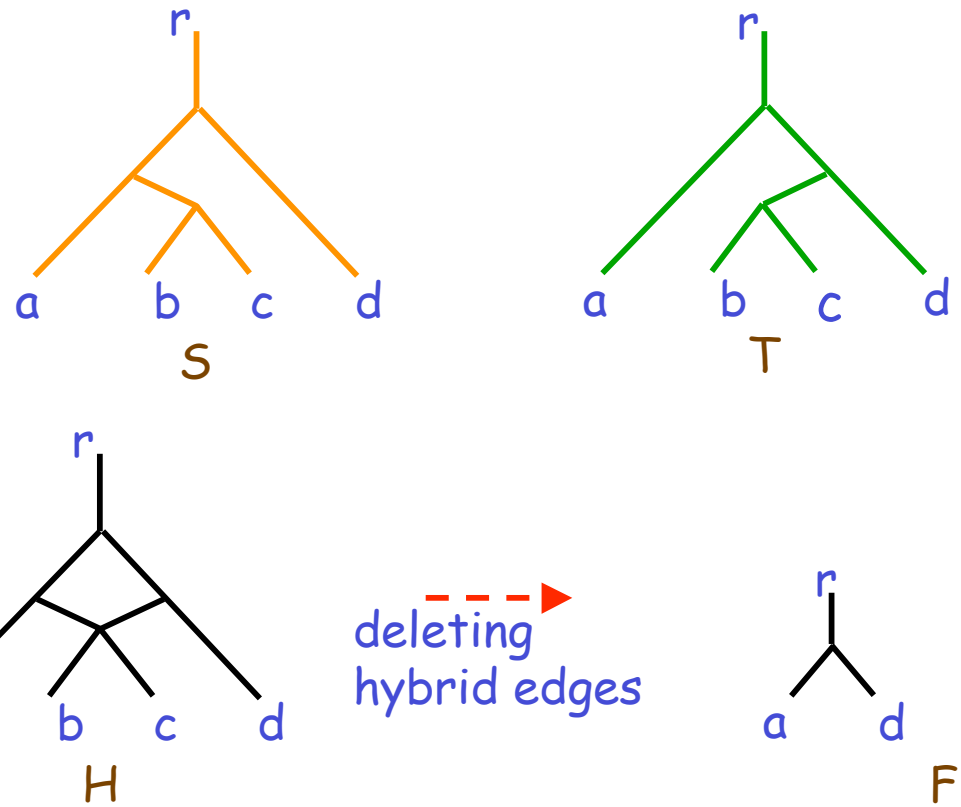




Modelling Hybridization Events with SPR Operations

A single SPR operation models a single hybridization event (Maddison 1997).

Example.



A Fundamental Problem for 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 evolution of the species?

This smallest number

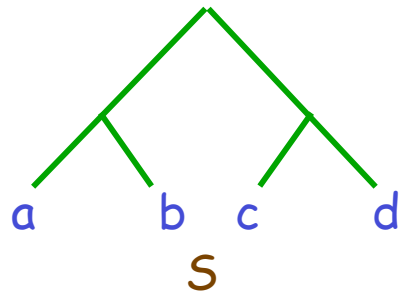
- Provides a lower bound on the number of such events.
- Indicates the extent that hybridization has had on the evolutionary history of the species under consideration.

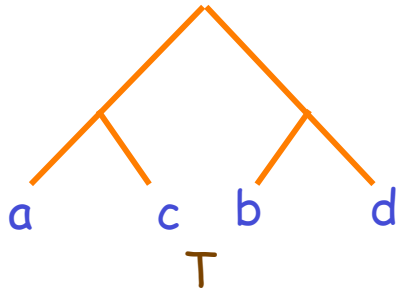
Since 1930's botantists have asked the question: How significant has the effect of hybridization been on the New Zealand flora?

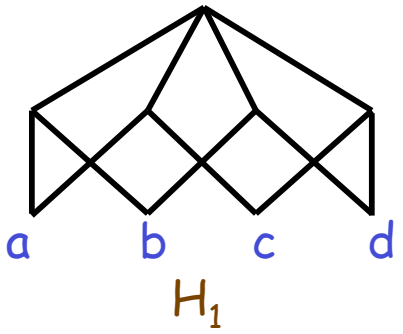
Trees and Hybridization Networks

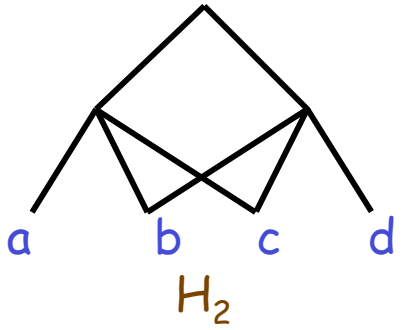
H explains T if T can be obtained from a rooted subtree of H by suppressing degree-2 vertices.

Example.





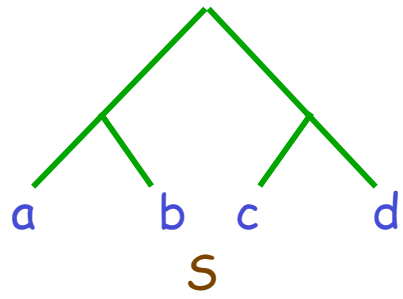


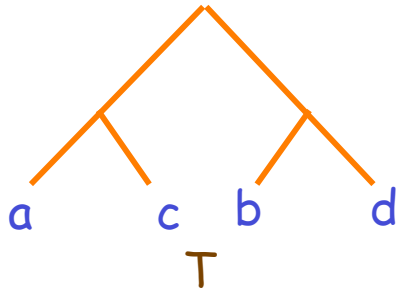


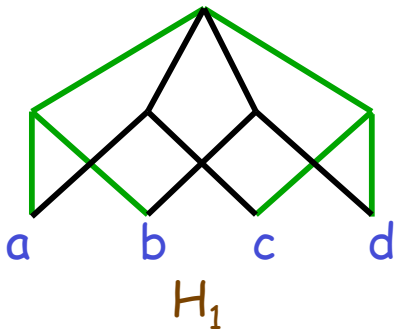
Trees and Hybridization Networks

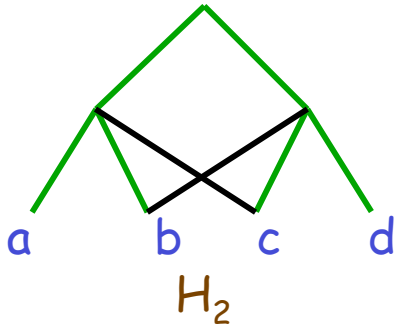
H explains T if T can be obtained from a rooted subtree of H by suppressing degree-2 vertices.

Example.





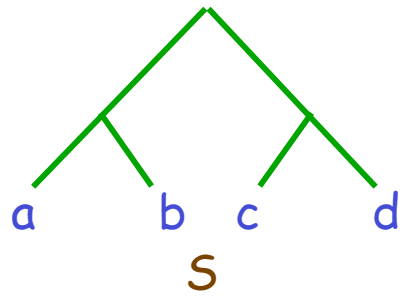


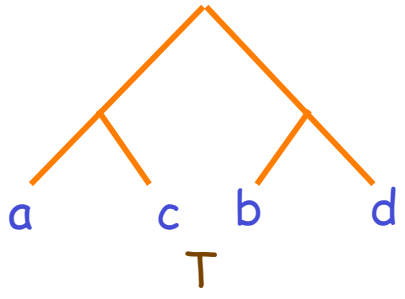


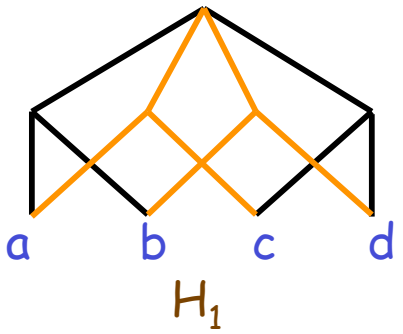
Trees and Hybridization Networks

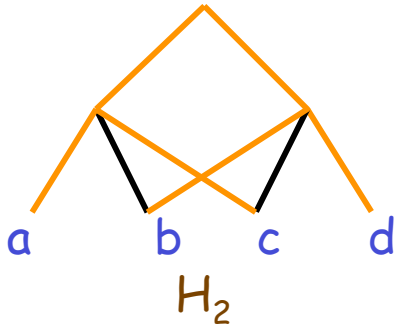
H explains T if T can be obtained from a rooted subtree of H by suppressing degree-2 vertices.

Example.









The Mathematical Problem

MINIMUM HYBRIDIZATION

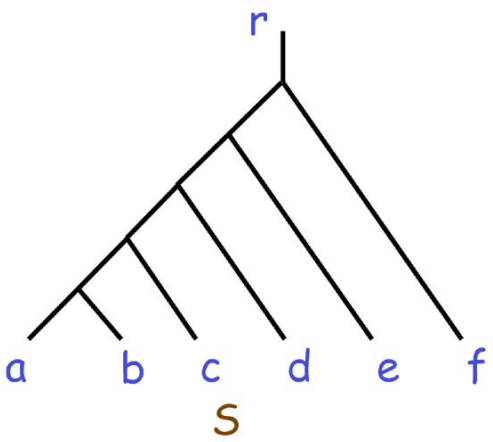
Instance: Two rooted binary phylogenetic trees S and T.

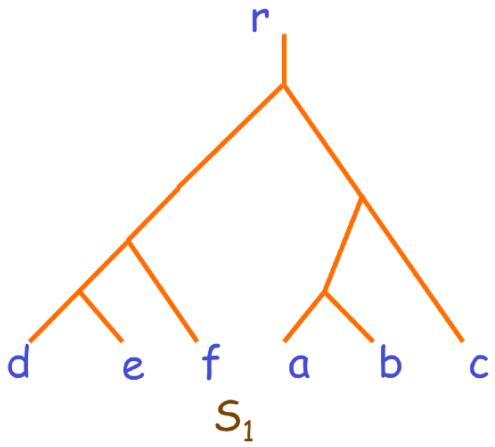
Goal: Find a hybridization network H that explains S and T, and minimizes the number of hybridization vertices.

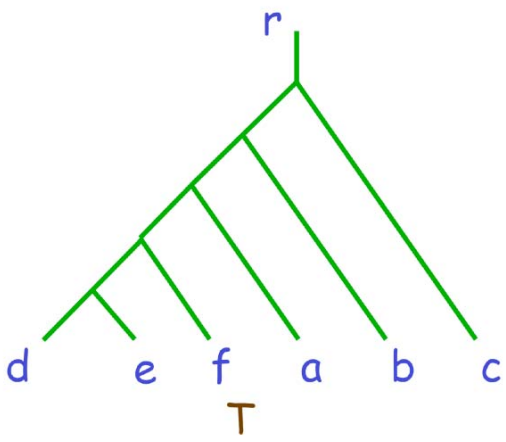
Measure: The number of hybridization vertices in H.

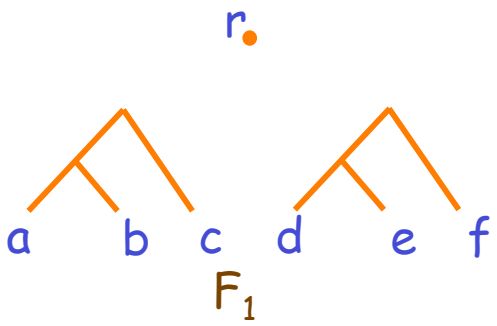
Notation: Use $h(S, T)$ to denote this minimum number.

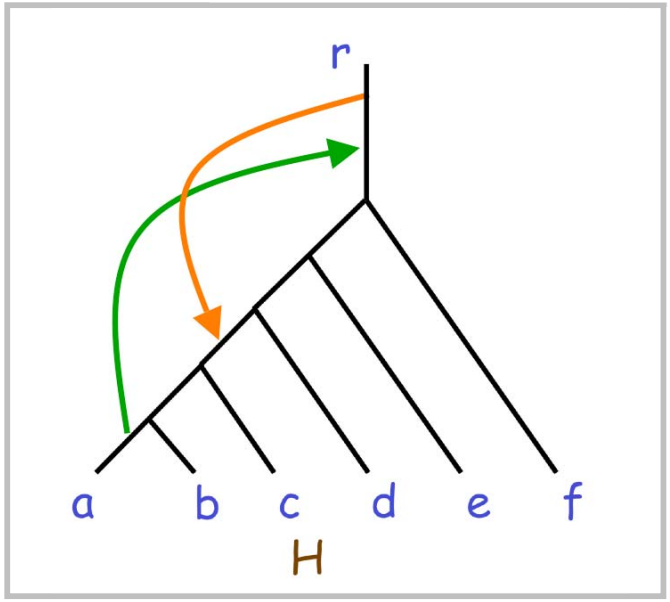
Example: Arbitrary SPR operations not sufficient.











- A sequence of SPR operations that avoids creating directed cycles to make a hybridization network that explains S and T.

- If one minimizes the length of an (acyclic) sequence, does the resulting network minimize the number of hybridization events to explain S and T ?

- YES, and such a sequence corresponds to an acyclic-agreement forest.

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

Let S and T be two binary phylogenetic trees. Then

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

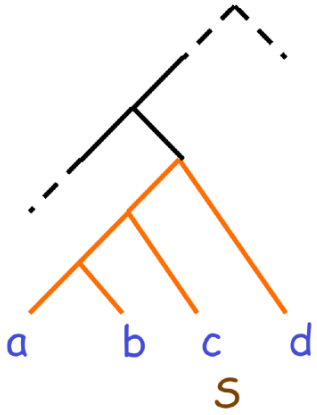
- It's fast to construct from a maximum-acyclic agreement forest for S and T a hybridization network that realizes h(S,T).

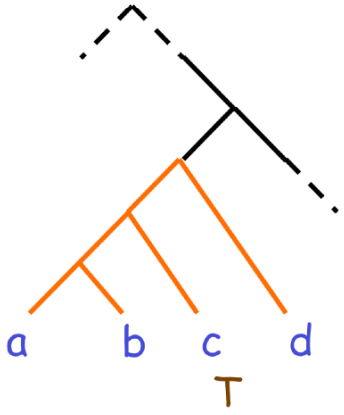
Theorem. (Bordewich, S, 2007)

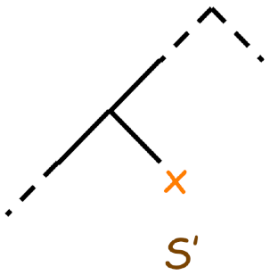
MINIMUM HYBRIDIZATION is NP-hard.

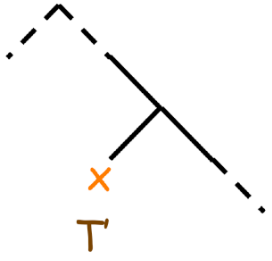
Reducing the Size of the Instance

Subtree Reduction

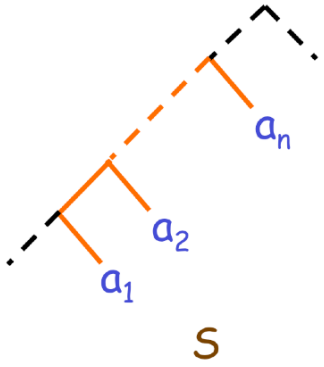


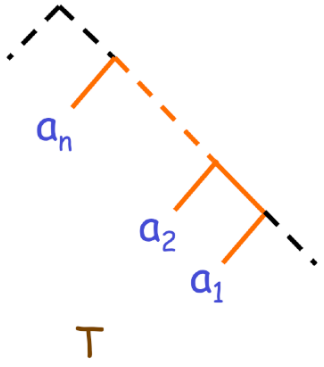


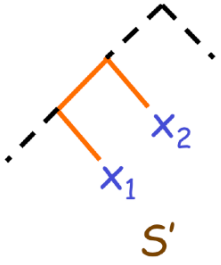


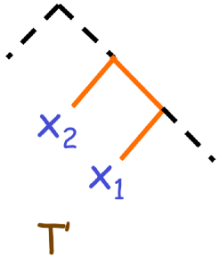


Chain Reduction









Fixed-Parameter Algorithms

Are the subtree and chain reductions enough to kernalize the problem?

Lemma. If n′ denotes the size of the leaf sets of the fully reduced trees using the subtree and chain reductions, then

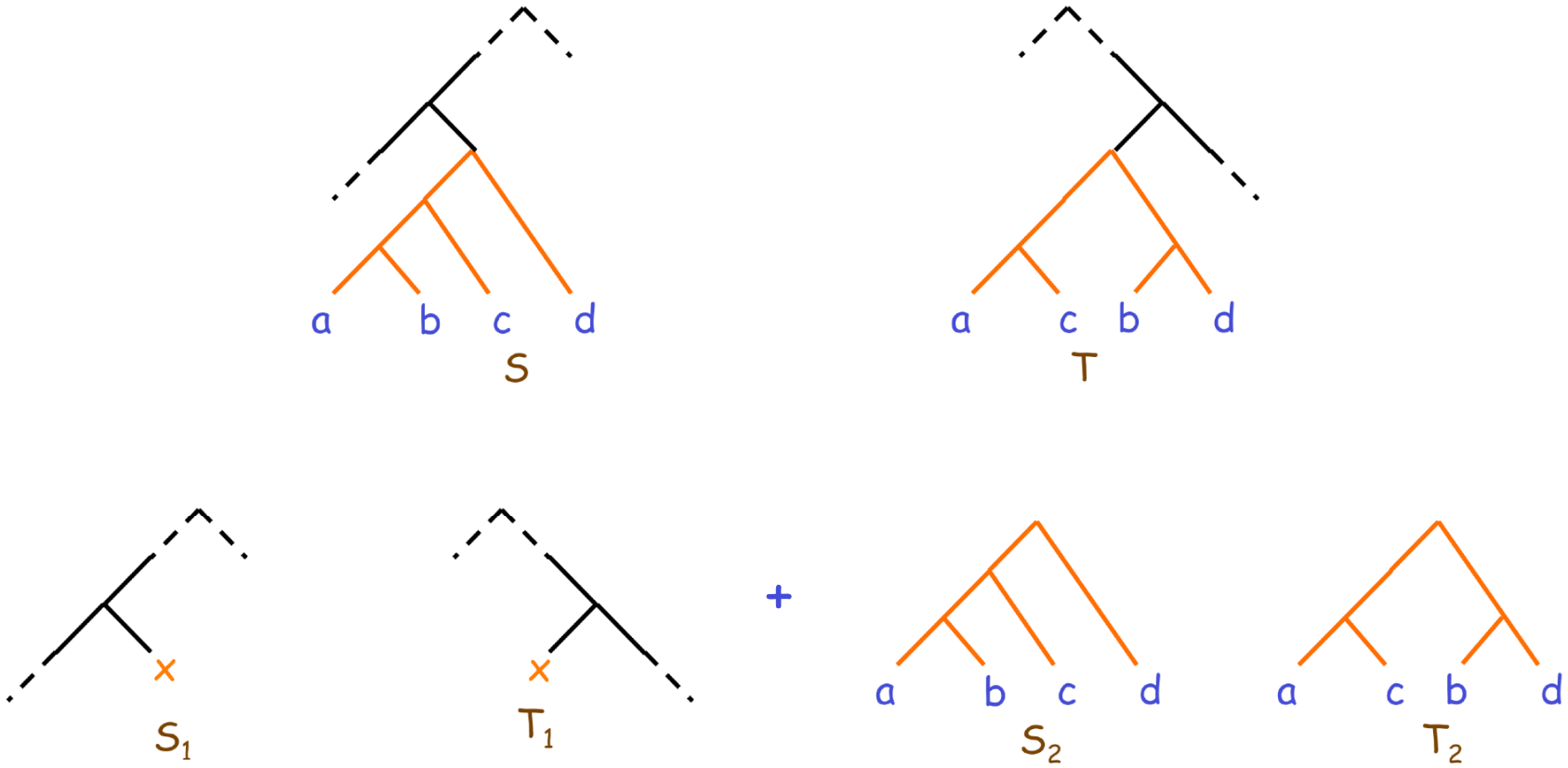
$$n' < 14h(S,T).$$

Corollary. (Bordewich, S 2007) MINIMUM HYBRIDIZATION is fixed-parameter tractable.

Running time is $O((28k)^k + p(n))$ compared with $O((2n)^k)$, where $k=h(S,T)$ and $p(n)$ is the polynomial bound for reducing the trees using the subtree and chain reductions.

Reducing the Size of the Instance

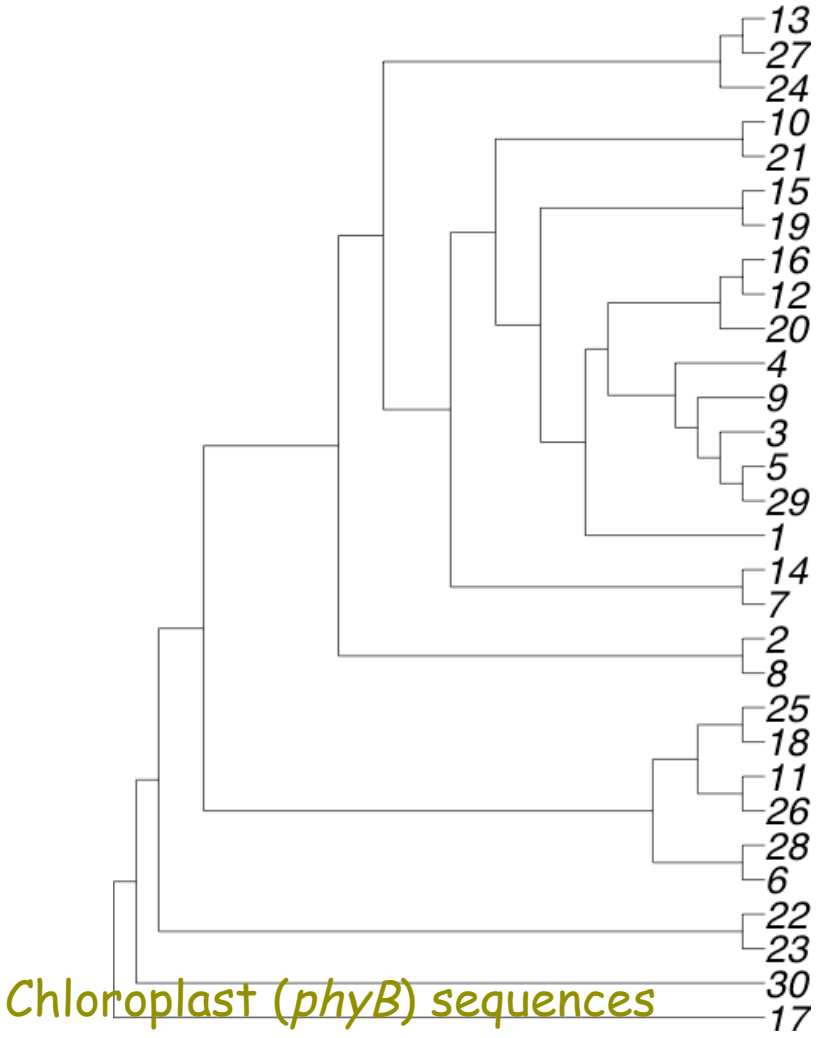
Cluster Reduction (Baroni 2004)



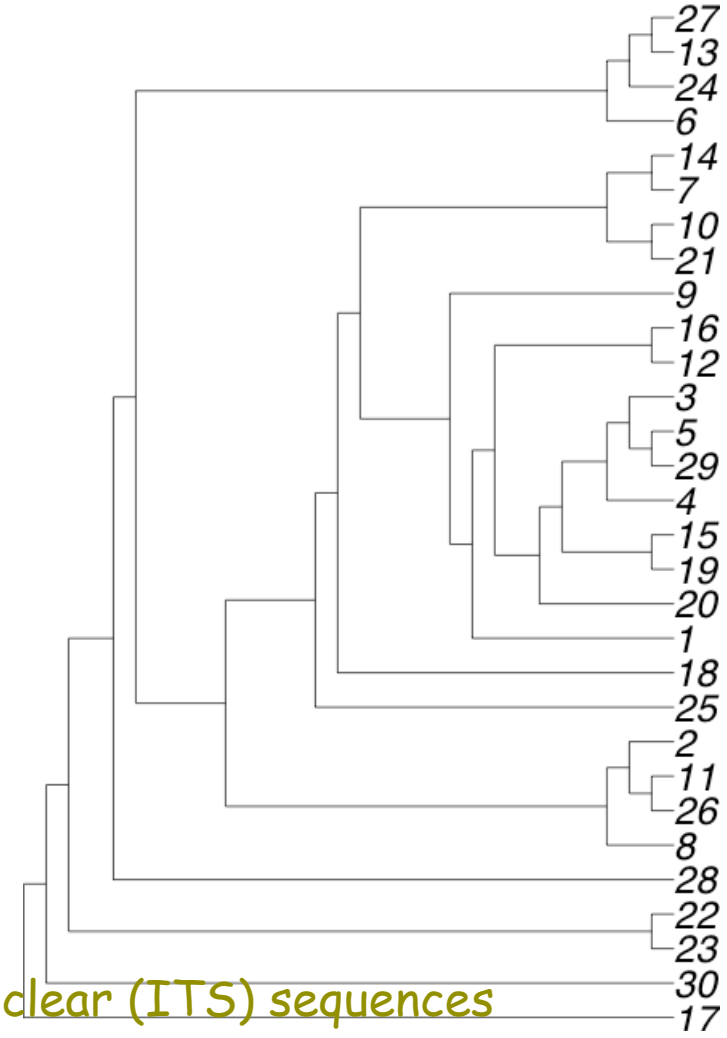
A Grass (Poaceae) Dataset (Grass Phylogeny Working Group, Düsseldorf)

loci	sequence origin	# sequences	alignment length
ITS	nucleus	47	322
ndhF	chloroplast	65	2210
phyB	nucleus	40	1182
rbcL	chloroplast	37	1344
rpoC2	chloroplast	34	777
waxy	nucleus	19	773

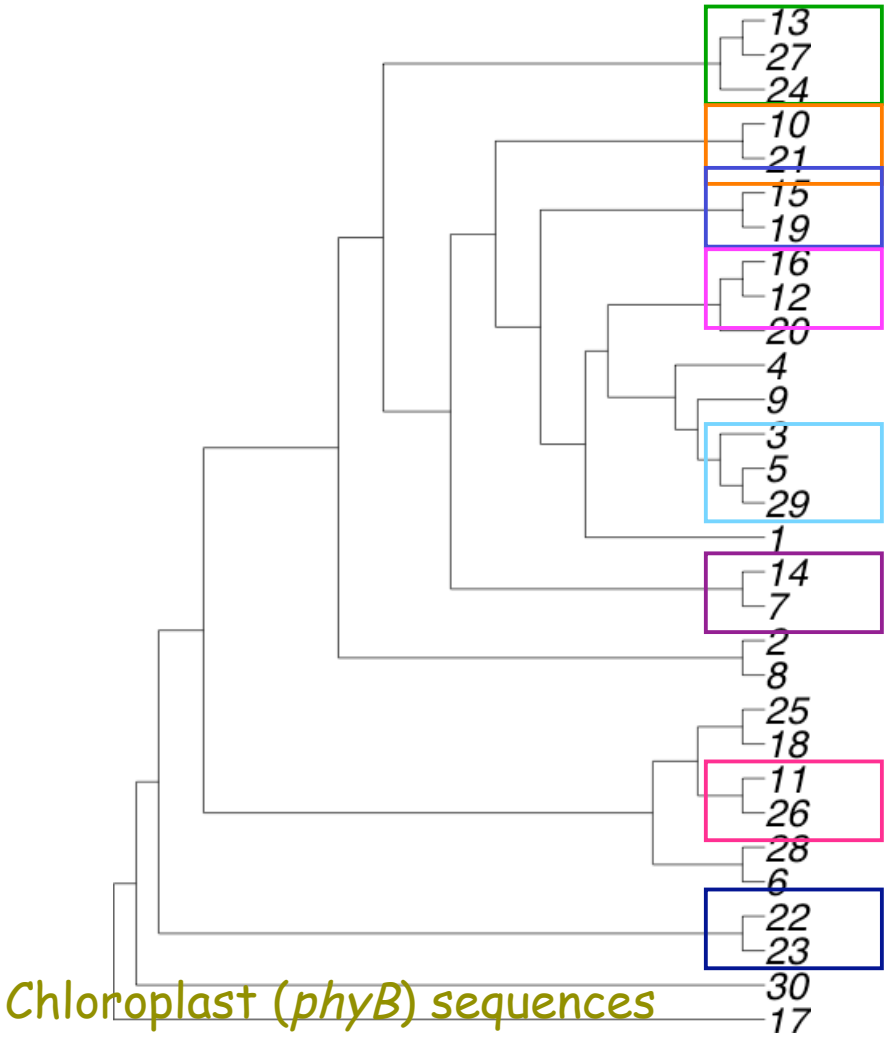
- Ellstrand, Whitkus, Rieseburg 1996 (Distribution of spontaneous plant hybrids)
- For each sequence, used fastDNAml to reconstruct a phylogenetic tree (H. Schmidt).



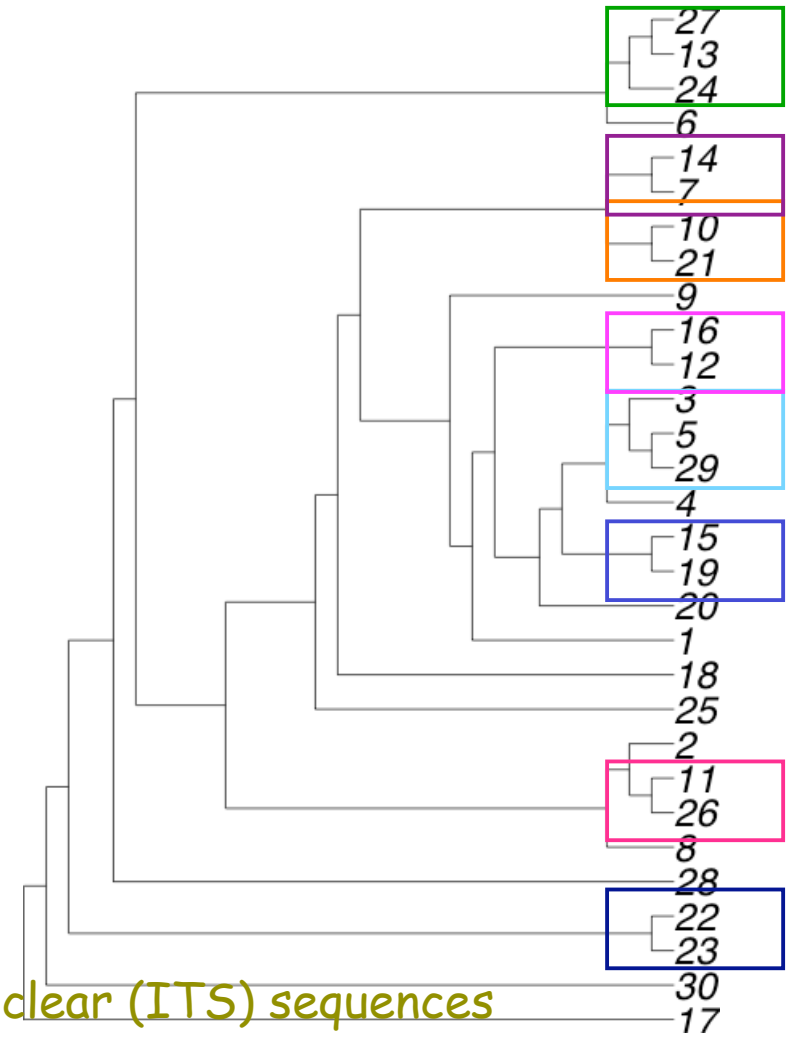
Chloroplast (*phyB*) sequences



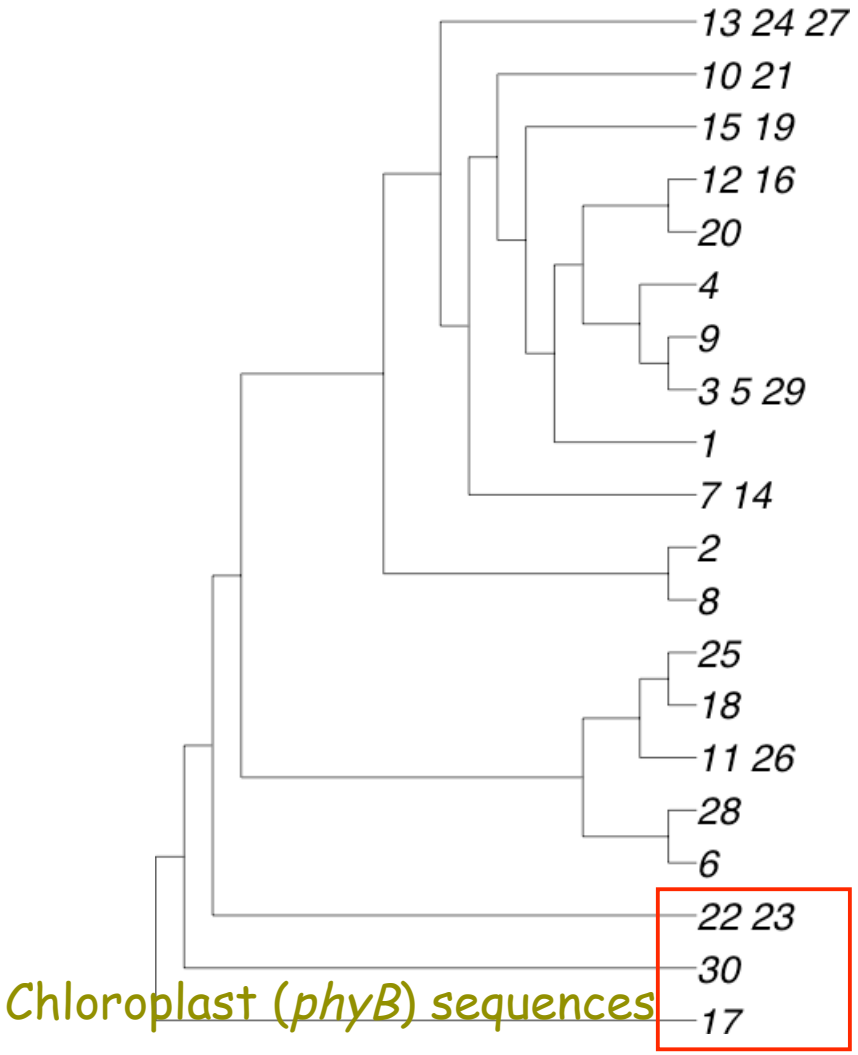
Nuclear (ITS) sequences



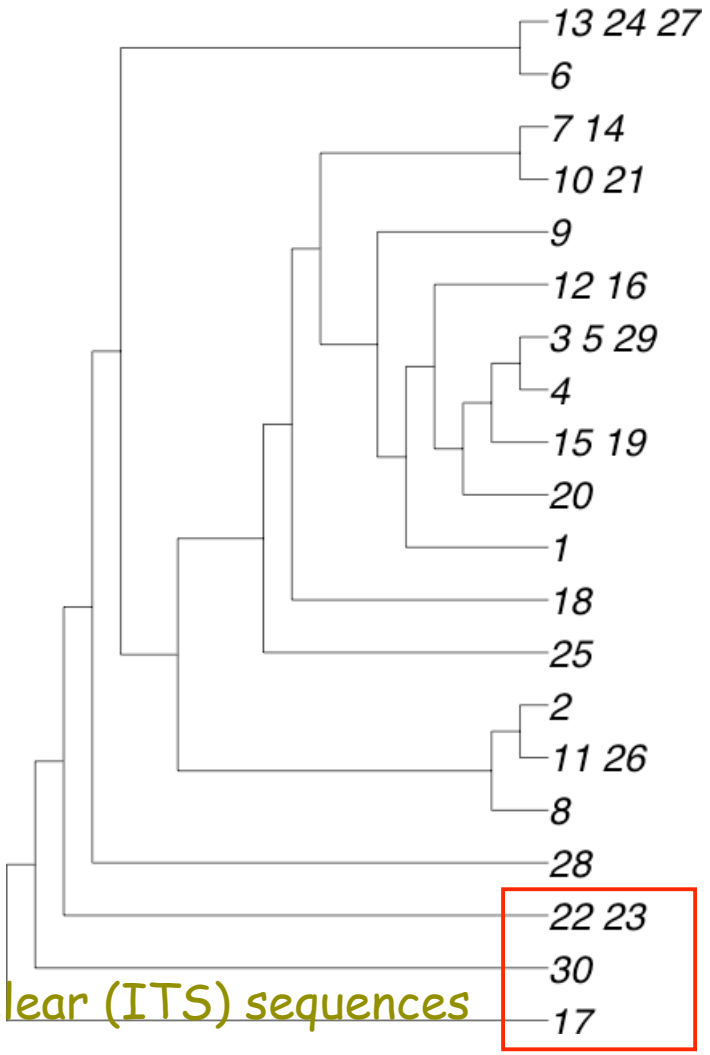
Chloroplast (*phyB*) sequences



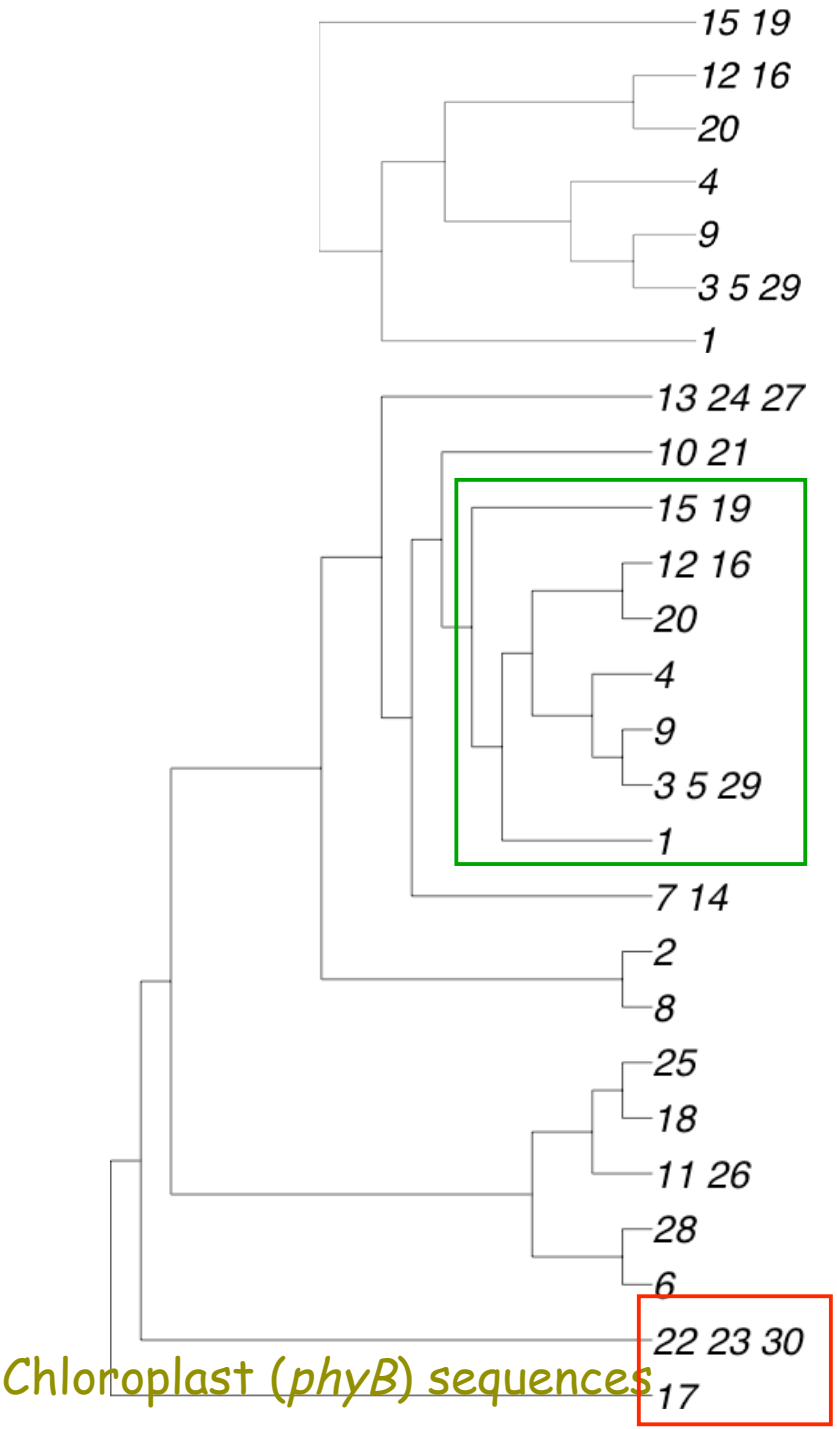
Nuclear (ITS) sequences



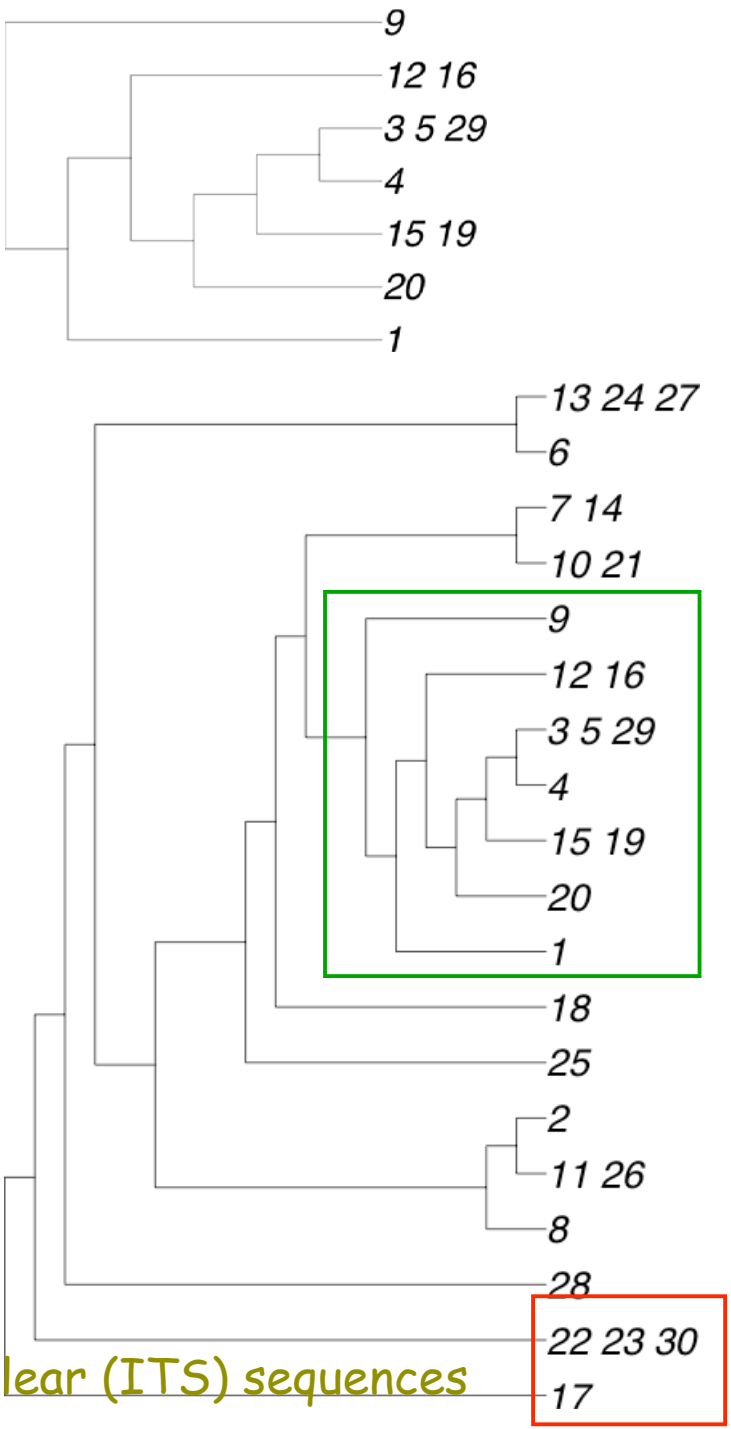
Chloroplast (*phyB*) sequences



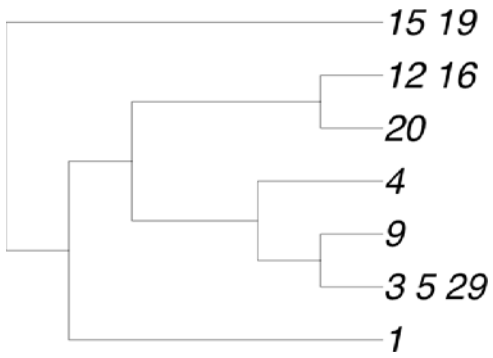
Nuclear (ITS) sequences

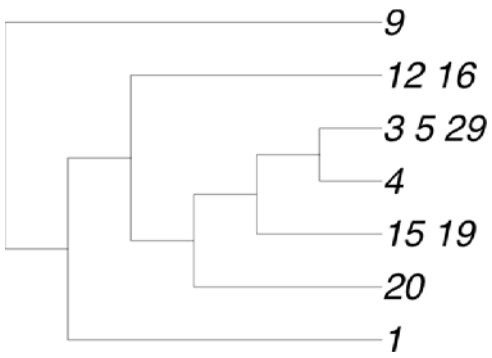


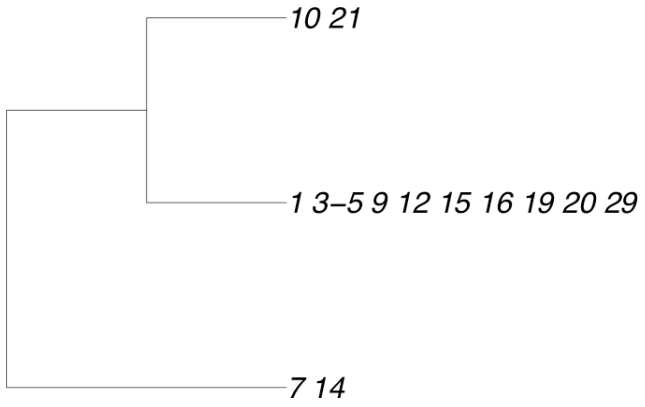
Chloroplast (*phyB*) sequences

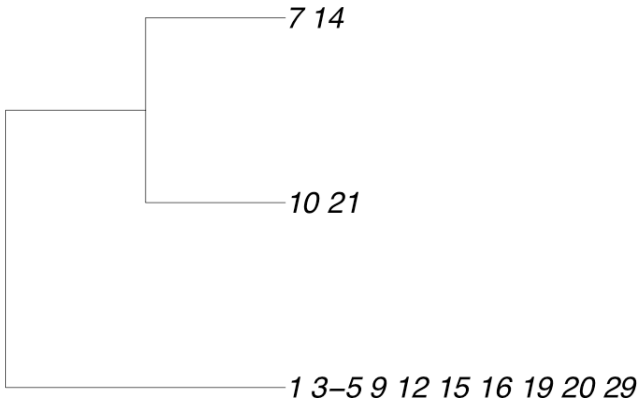


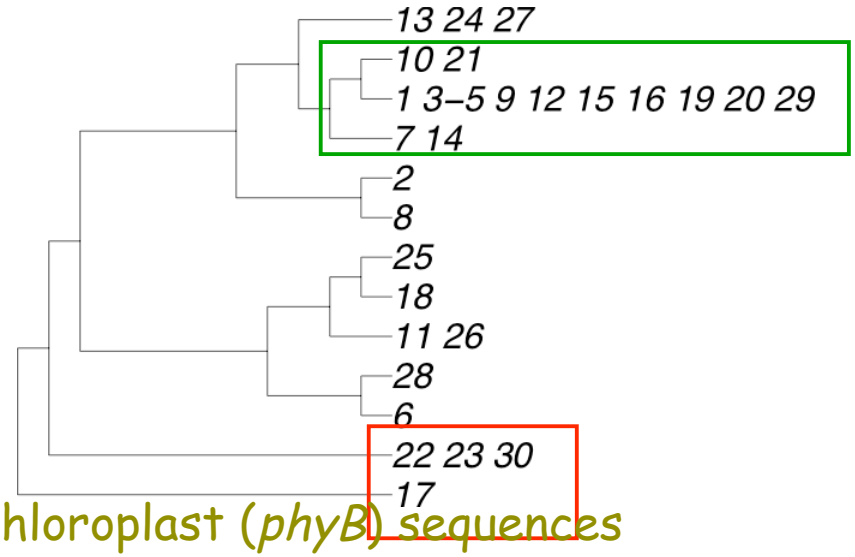
Nuclear (ITS) sequences



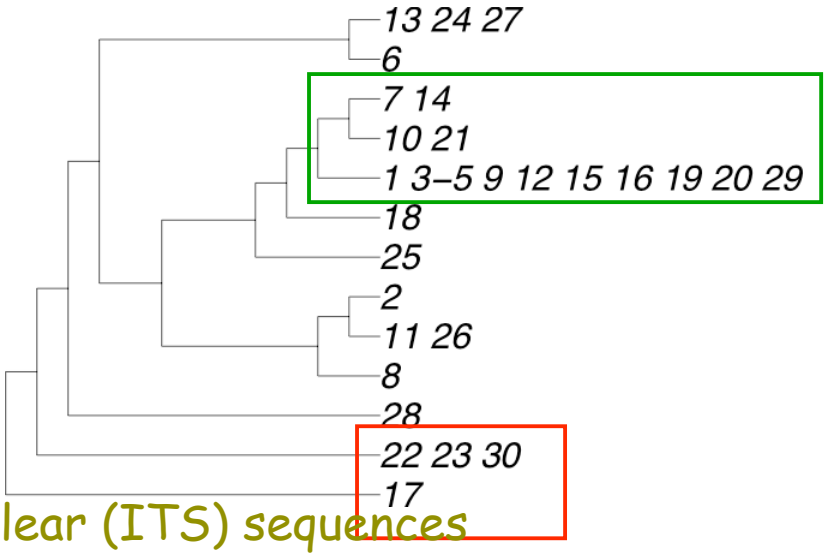




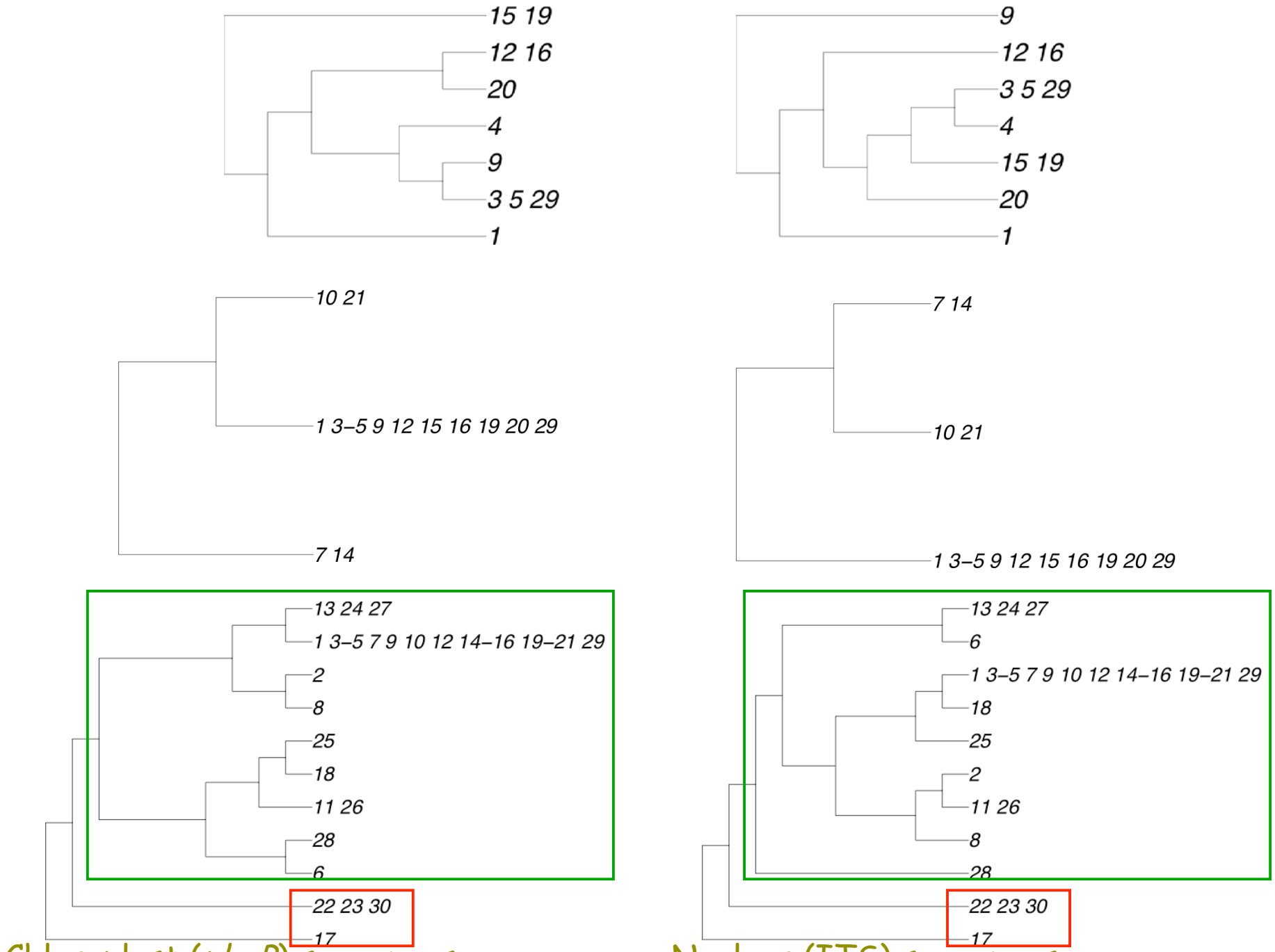


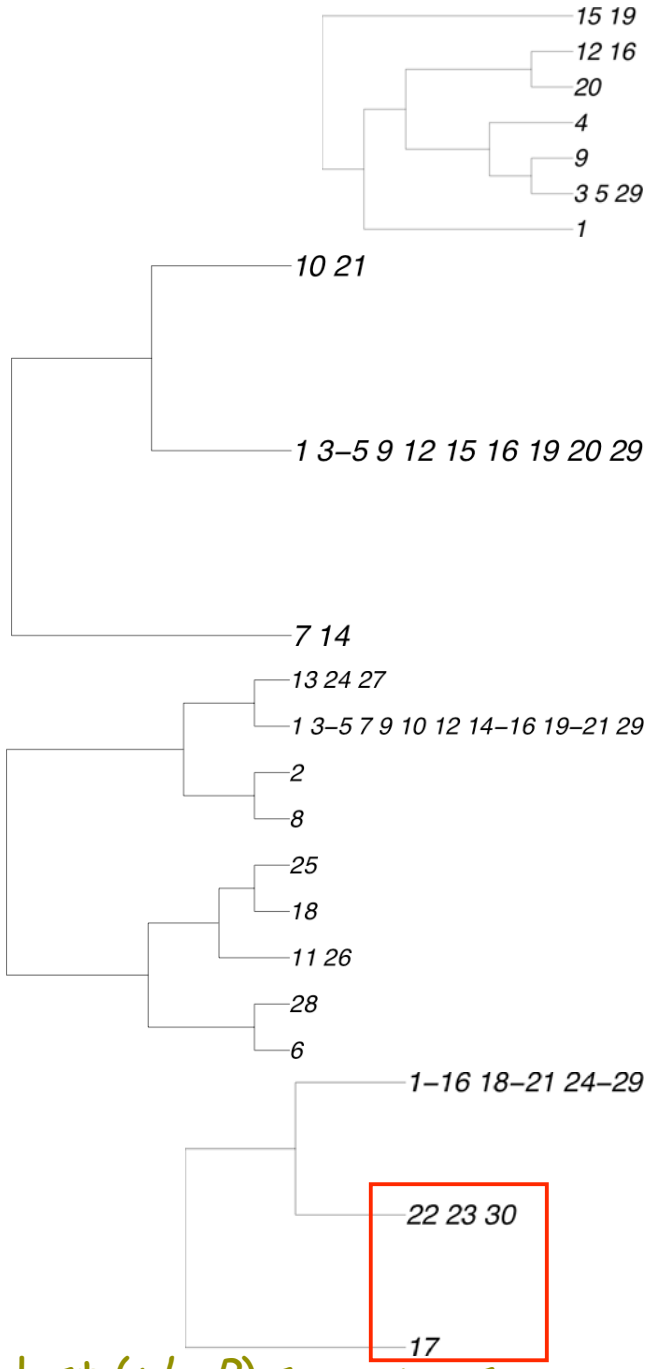


Chloroplast (*phyB*) sequences

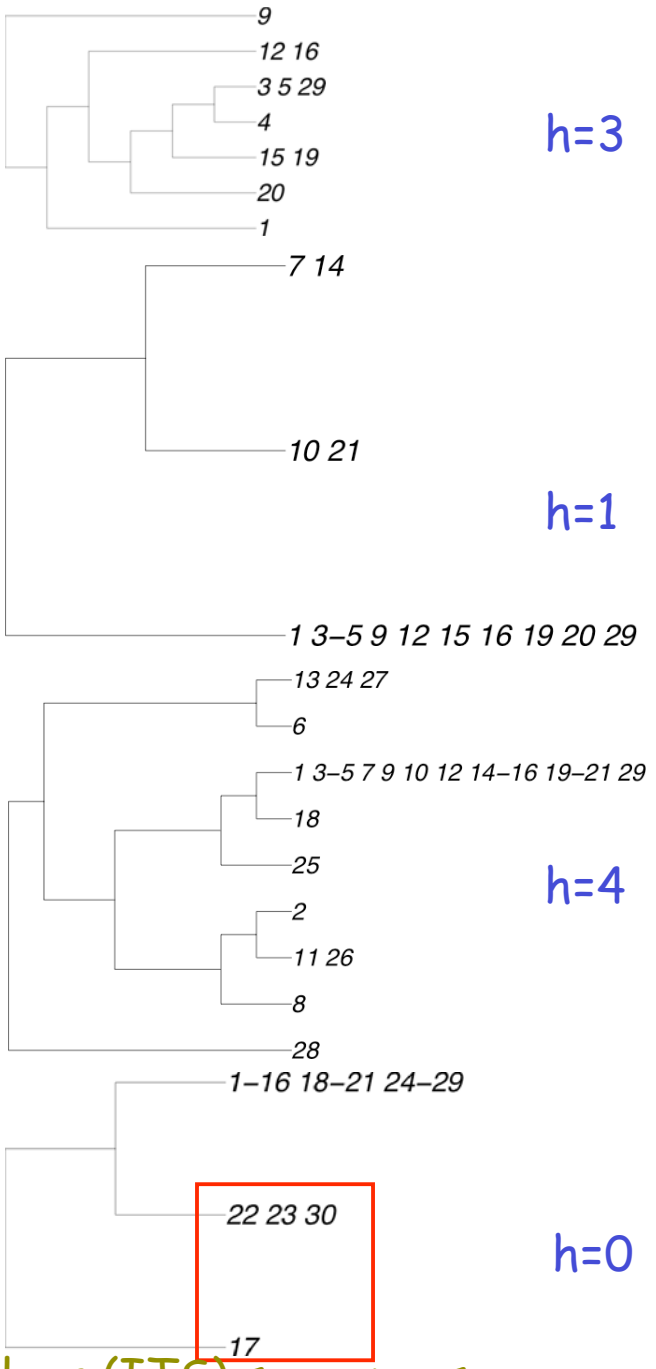


Nuclear (ITS) sequences


15 19
12 16
20
4
9
3 5 29
1
10 21
1 3–5 9 12 15 16 19 20 29
7 14
13 24 27
1 3–5 7 9 10 12 14–16 19–21 29
2
8
25
18
11 26
28
6
22 23 30
17
Chloroplast (phyB) sequences
9
12 16
3 5 29
4
15 19
20
1
7 14
10 21
1 3–5 9 12 15 16 19 20 29
13 24 27
6
1 3–5 7 9 10 12 14–16 19–21 29
18
25
2
11 26
8
28
22 23 30
17
Nuclear (ITS) sequences



Chloroplast (*phyB*) sequences



Nuclear (ITS) sequences

h=3

h=1

h=4

h=0

pairwise combination		# overlapping taxa	h(S,T)	running time 2000 MHz CPU, 2GB RAM
ndhF	*phyB*	40	14	11h
ndhF	*rbcL*	36	13	11.8h
ndhF	*rpoC2*	34	12	26.3h
ndhF	*waxy*	19	9	320s
ndhF	ITS	46	at least 15	
phyB	*rbcL*	21	4	1s
phyB	*rpoC2*	21	7	180s
phyB	*waxy*	14	3	1s
phyB	ITS	30	8	19s
rbcL	*rpoC2*	26	13	29.5h
rbcL	*waxy*	12	7	230s
rbcL	ITS	29	at least 9	
rpoC2	*waxy*	10	1	1s
rpoC2	ITS	31	at least 10	
waxy	ITS	15	8	620s

Bordewich, Linz, St John, S, 2007

Computing $d_{SPR}(S,T)$ and $h(S,T)$

$d_{SPR}(S,T)$

1. FPT using kernalization ($O((56k)^k + p(n))$).
2. FPT using a bounded search tree method ($O(4^k n^4)$) (Bordwich, McCartin, S 2008). Combining with 1. gives $O(4^k k^4 + p(n))$ FPT algorithm.
3. No cluster-based reduction.
4. 3-approximation algorithm (Bordwich, McCartin, S 2008).

$h(S,T)$

1. FPT using kernalization ($O((28k)^k + p(n))$).
2. Unknown if a bounded search tree method exists.
3. Cluster-based reduction.
4. Unknown if there is an approximation algorithm.

Acknowledgements

Magnus Bordewich, Durham University (UK)

Simone Linz, Heinrich-Heine Universität, (Germany)

Catherine McCartin, Massey University, (NZ)

Katherine St John, City University of New York, (USA)