Abstract. A network N is a rooted acyclic directed graph. A base-set X for N is a subset of vertices including the root (or outgroup), all leaves, and all vertices of outdegree 1. A simple model of evolution is considered in which all characters are binary and in which back-mutations occur only at hybrid vertices. It is assumed that the genome is known for each member of the base-set X. If the network is known and is assumed to be "normal," then the genome of every vertex is uniquely determined and can be explicitly reconstructed. Under additional hypotheses involving time-consistency and separation of the hybrid vertices, the network itself can also be reconstructed from the genomes of all members of X. A polynomial-time procedure is outlined for performing the reconstruction.

Reconstruction of certain phylogenetic networks from the genomes at leaves

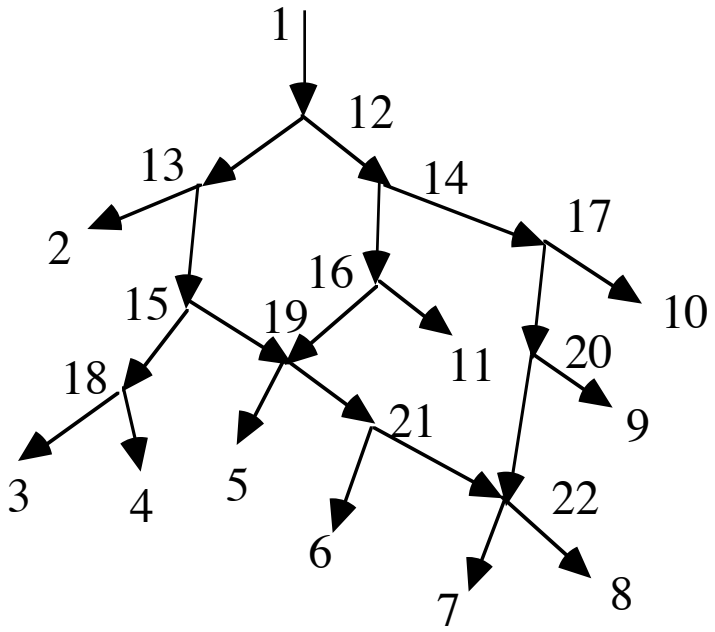
by
Stephen J. Willson
Department of Mathematics
Iowa State University
Ames, Iowa 50011
USA
swillson@iastate.edu

Workshop: Future Directions in Phylogenetic Methods and Models

Isaac Newton Institute, 20 December 2007

Evolution on networks that are not necessarily trees

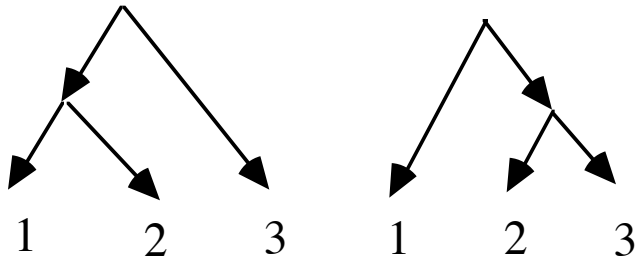
General Problem: Given characters on a set X of taxa, construct a species network that explains the data.



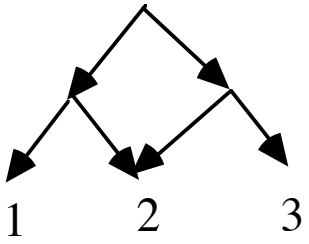
One approach:

Different genes give rise to different phylogenetic trees. Find some network that contains all these trees.

Given



find



Difficulties

1. Often there are many networks that can explain the trees. How do we choose among them?

2. The problem of finding one that has the minimum number of reticulation events is NP-hard. (Bordewich and Semple 2007)

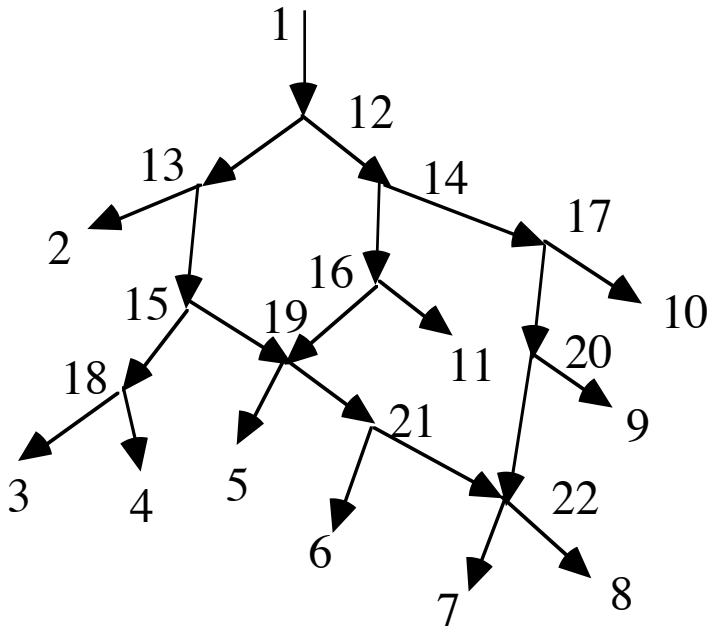
3. It is not clear that having the minimum number of reticulation events is biologically the right criterion.

This approach

Try to model evolution on a network N that is not necessarily a tree. Assume that the network itself is the underlying description of evolutionary history.

Then try to reconstruct N directly from the data.

This talk contains one such method. I will state theorems about the reconstruction.



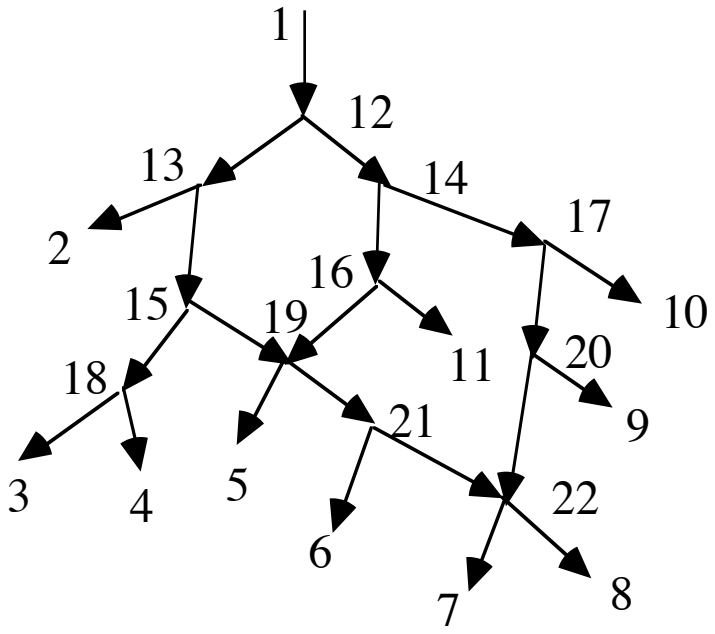
Summary of the talk

In order to prove the theorems, I will make strong assumptions:

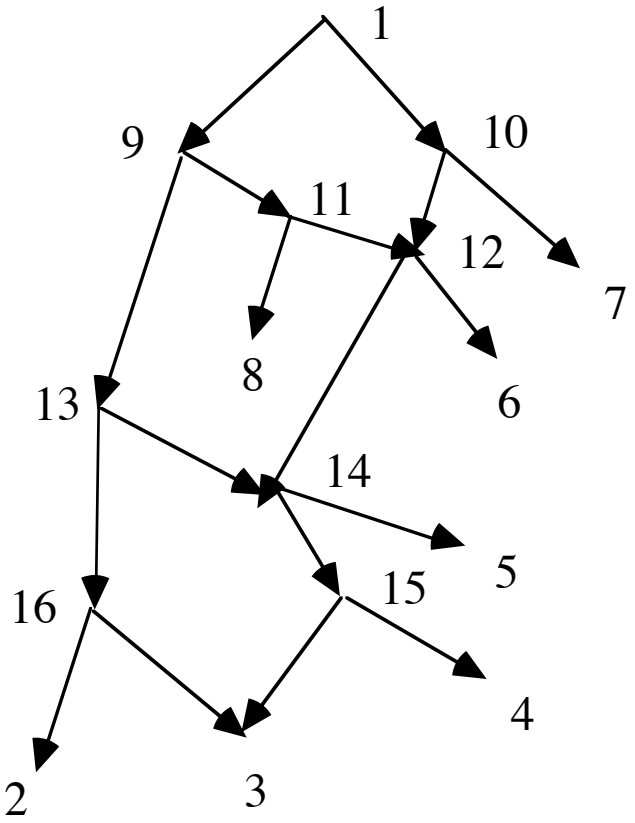
1. Restrict the network. It will be assumed to satisfy some strong conditions.

2. Assume a very simple unrealistic model of evolution.

Suppose that we are given the genome at the root and each leaf. I will then show how to reconstruct the network as well as the genome at each internal vertex.

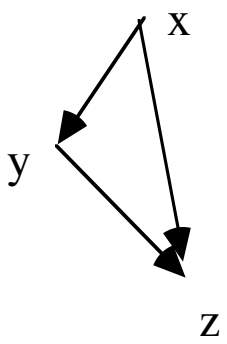


Assume that relationships among species are given by an **acyclic rooted directed graph**.
Vertices with two or more incoming arcs are **hybrid**; others are **normal** (or **tree-vertices** or **tree-children**).



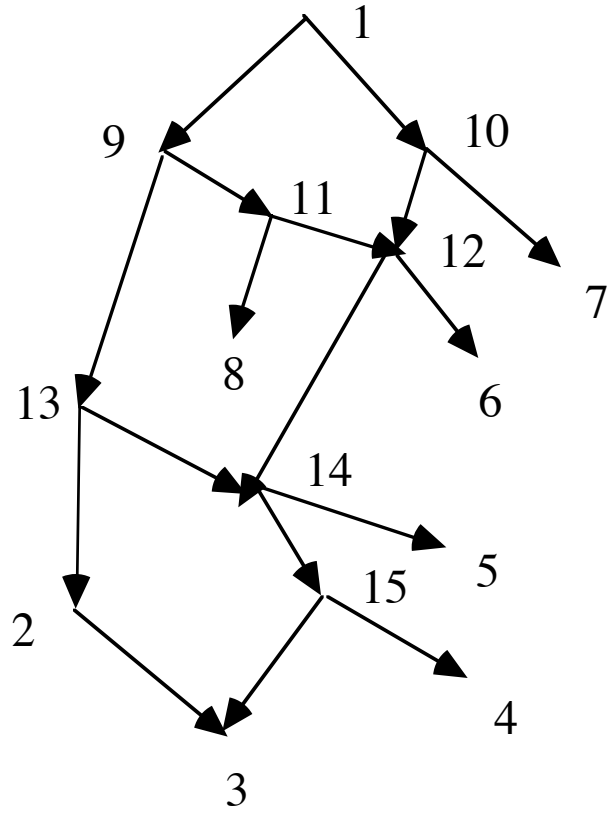
12, 14, and 3 are hybrid. The rest are normal.

If there exists both a nontrivial path from x to z and an arc from x to z, then the arc is **redundant**.



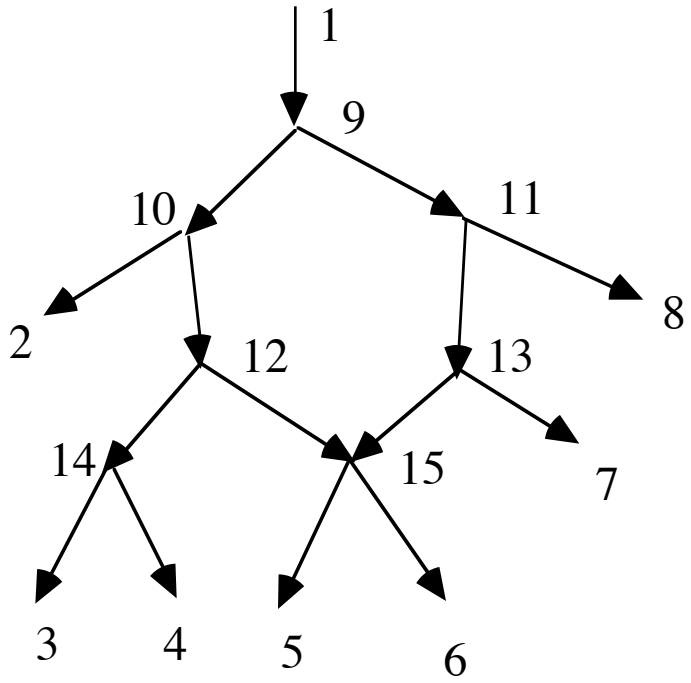
Assume that there are no redundant arcs.

A **base-set** X is a set of vertices that includes the root, all leaves, and all vertices of outdegree 1.



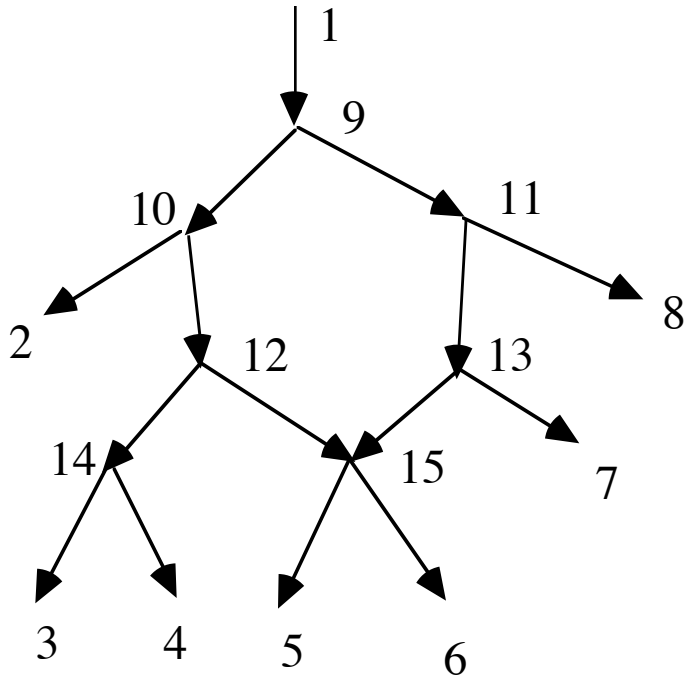
X = {1, 2, 3, 4, 5, 6, 7, 8} is a base set.
We will assume that we know the genome of x if $x \in X$.

A directed path $u = u_0, u_1, u_2, ..., u_k = v$ is a **normal path from u to v** provided for $i > 0$, u_i is normal. Note u may or may not be hybrid.



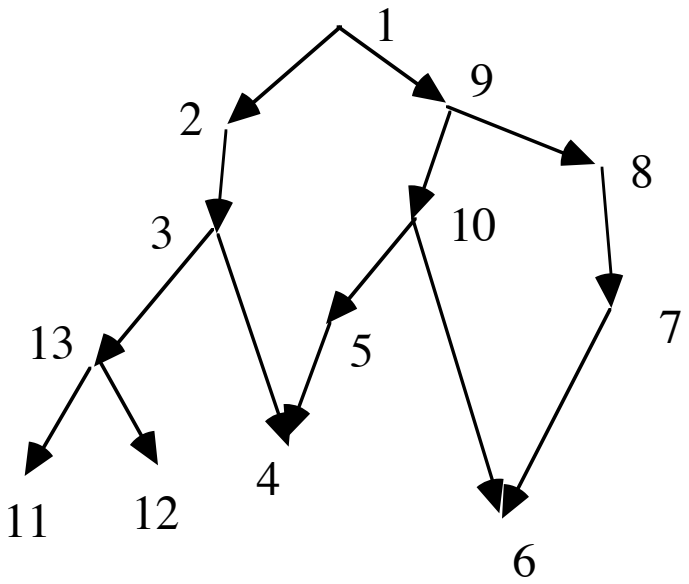
10, 12, 14, 4 is normal.
15, 5 is normal.
10, 12, 15, 5 is not normal.

The network N is **normal** if from every vertex u ∈ V there is a normal path to some member of X.



This network is normal, where X = {1,2,3,4,5,6,7,8}.

Model of evolution: genomes



Assume that each character is binary with two states (alleles), 0 or 1.
Maybe there are 10 characters (genes).

Assume that the root has only state 0, so Genome(1) = 0000000000.
Maybe vertex 4 has genome Genome(4) = 0010100011.
The 3rd, 5th, 9th, 10th characters have state 1; others have state 0.
M(4) = {3,5,9,10}

M(v) = the set of characters differing at v from the root
= mutation set of v.

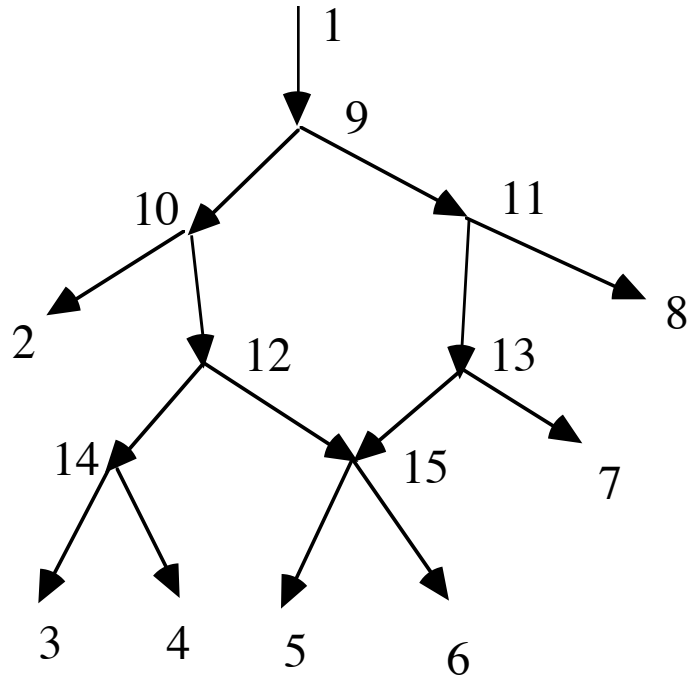
"Simple Homoplasy Model" of evolution

When a mutation of character i occurs for the first time at vertex v, then
v **originates** i.
The set of all characters originating at v is O(v).

(1) Assume that no character originates more than once (no parallel evolution).

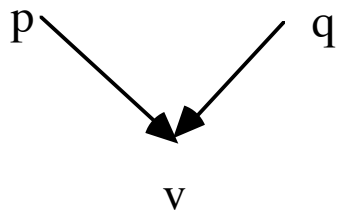
(2) **If v is normal with (unique) parent p then v inherits all the mutations of p.**

$M(v) = M(p) \cup O(v)$
(Assume any failures will be in the noise.)

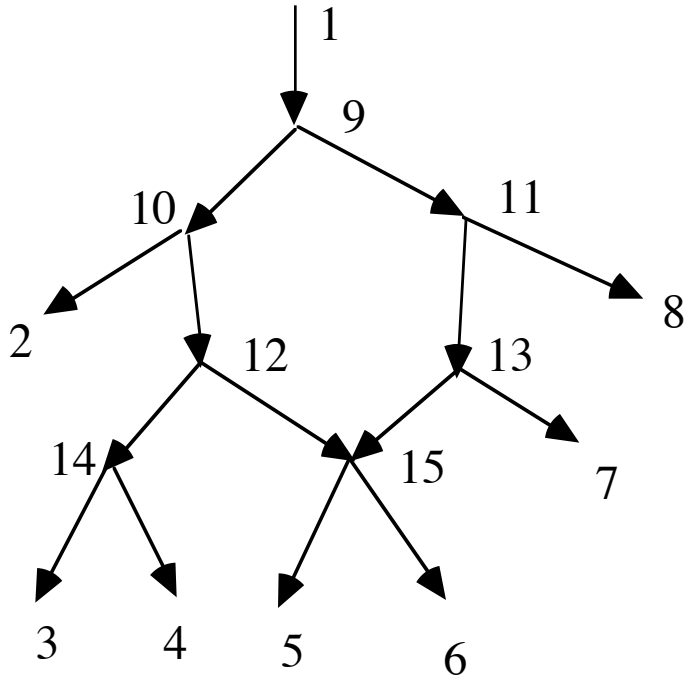


$M(12) = M(10) \cup O(12)$
$M(10) = M(9) \cup O(10)$
Hence $M(12) = M(9) \cup O(10) \cup O(12)$

(3) **If v is hybrid with parents p and q then M(v) consists of O(v) together with a contribution from parent p and a contribution from parent q.**



Immediate homoplasies



You can't distinguish i ∈ O(13) but not inherited by 15 from i ∈ O(7). If i ∈ O(13) is immediately deleted at 15, there was an **immediate homoplasy**.

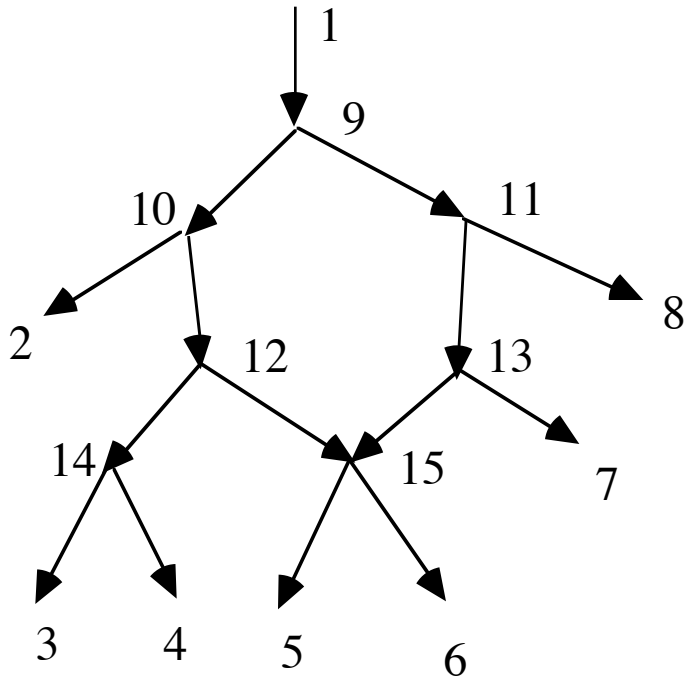
(4) Assume there are no immediate homoplasies.

If a child c inherits all the mutated characters of all its parents, then the model is "accumulation phylogeny" (Baroni and Steel 2006).

Main Theorem 1.

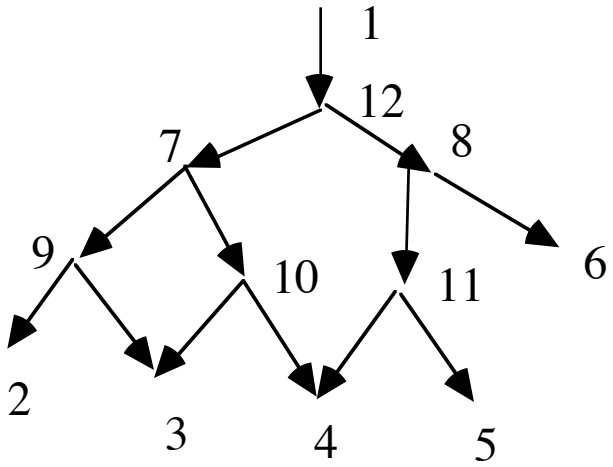
Assume the network N is normal (and given). Assume M(x) is known for each $x \in X$. Assume the Simple Homoplasy Model.
Then for all $v \in V$, M(v) is determined and O(v) is determined.

The genomes at all internal vertices can be reconstructed in polynomial time.



$X = \{1,2,3,4,5,6,7,8\}$. Assume M(x) is known for $x \in X$.
O(v) and M(v) are determined for all v.

The hypothesis on normal paths is needed:



One cannot distinguish between i ∈ O(9) and i ∈ O(7). Either way the only members of X containing i might be 2 and 3.

With additional assumptions, we can also reconstruct the network N.

Assumptions for reconstructing the network as well

(1) N is a normal network with base-set X.

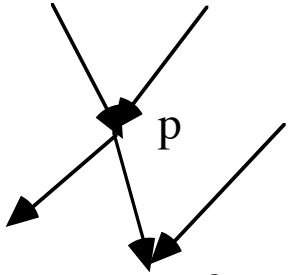
(2) The evolution satisfies the Simple Homoplasy Model.

(3) Every hybrid vertex has exactly two parents.

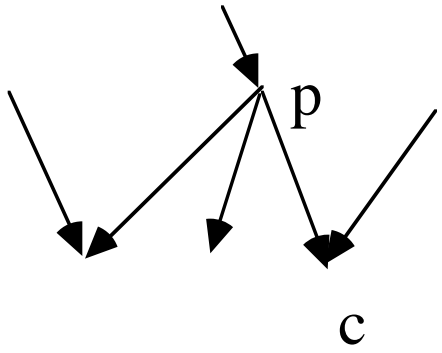
(4) For all $v \neq r$, if v is normal then $O(v) \neq \varnothing$.

(5) If p has a hybrid child c, then p is normal and every child of p other than c is normal.

We don't have

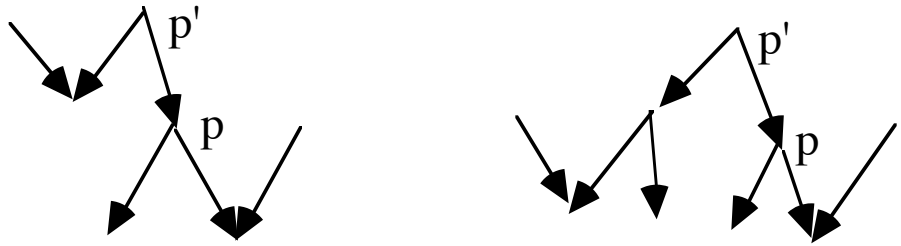


or



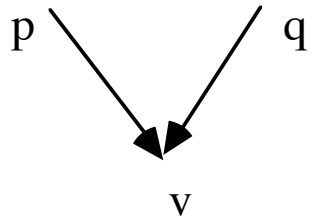
(6) If h is hybrid, then each grandparent of h doesn't have a hybrid child or other hybrid grandchild.

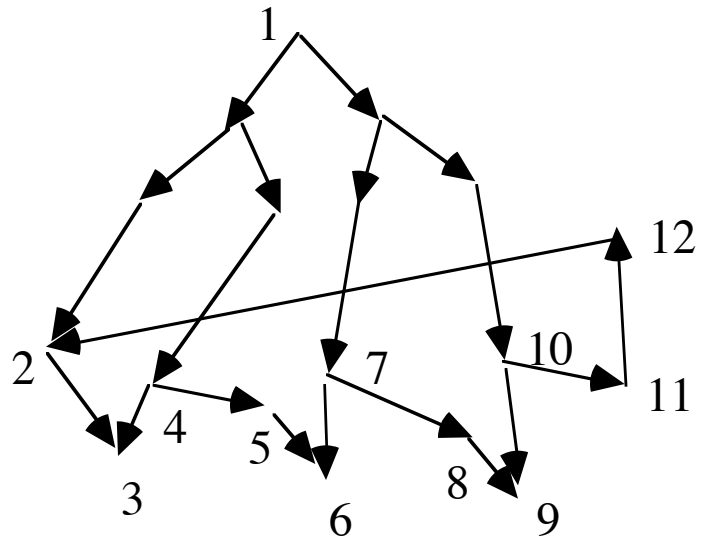
We don't have



Hybrid vertices are "separated".

(7) Suppose v is hybrid with parents p and q. Then some part of the genome of p is not inherited by v, and some part of the genome of q is not inherited by v.





A **pseudocycle** 2,3,4,5,6,7,8,9,10,11,12,2.

A pseudocycle is sufficiently similar to a cycle that it appears biologically unlikely. If there is a "temporal representation" (Baroni, Semple, Steel 2006) or "hybrid parents must co-exist" (Moret et al 2004) there can be no pseudocycle.

(8) Assume there are no pseudocycles.

Main Theorem 2.

Let N = (V,A,r,X) be a phylogenetic network that satisfies (1) through (8) above. Assume for all $x \in X$, M(x) is given.

Then N can be reconstructed uniquely.

Major tool:

Assume $|X| \geq 3$. Define the **stem function** δ whenever a, b, and x are distinct members of X by

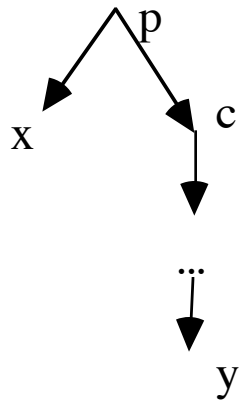
$$\delta(x,a,b) = M(x) - (M(a) \cap M(x)) \cup (M(b) \cap M(x))$$
$$= \text{the genome of x that is not in a or b}$$

Define

$$\delta(x) = \cap \{ \delta(x,a,b): x, a, b \text{ are distinct members of } X \}$$

Note $\delta(x,a,b) \subseteq M(x)$.

Lemma. Suppose x is a normal leaf with parent p and p has a normal child c distinct from x. Then $\delta(x) = O(x)$.



$\delta(x) = \delta(x,r,y) = O(x)$

In fact we can show that for every leaf x, $\delta(x) = O(x)$.

If $x \in X$ is not a leaf, then $\delta(x) = \varnothing$.

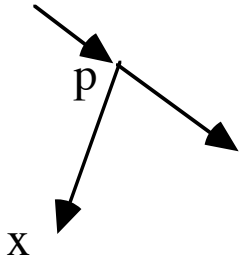
If x is a normal leaf with parent p,
then $\delta(x) = O(x) \neq \emptyset$.
Since

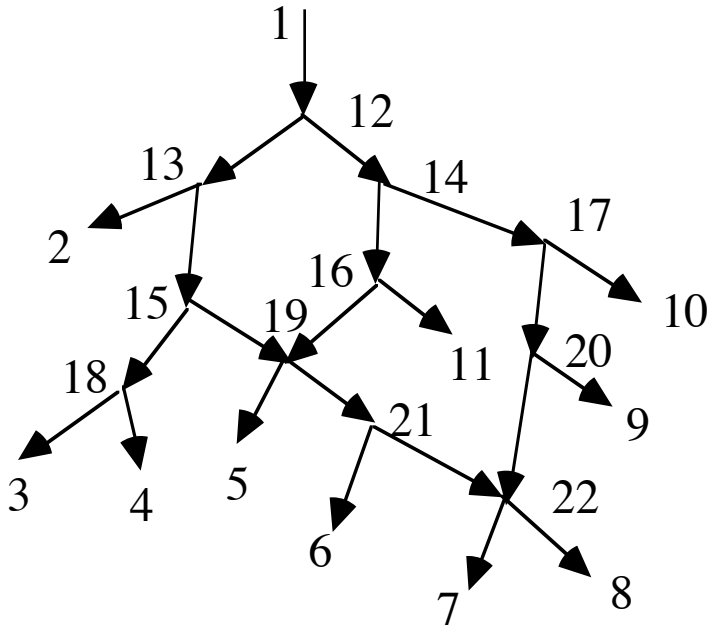
$$M(x) = M(p) \cup O(x)$$

it follows

$$M(p) = M(x) - O(x) = M(x) - \delta(x).$$

We know the genome of p.

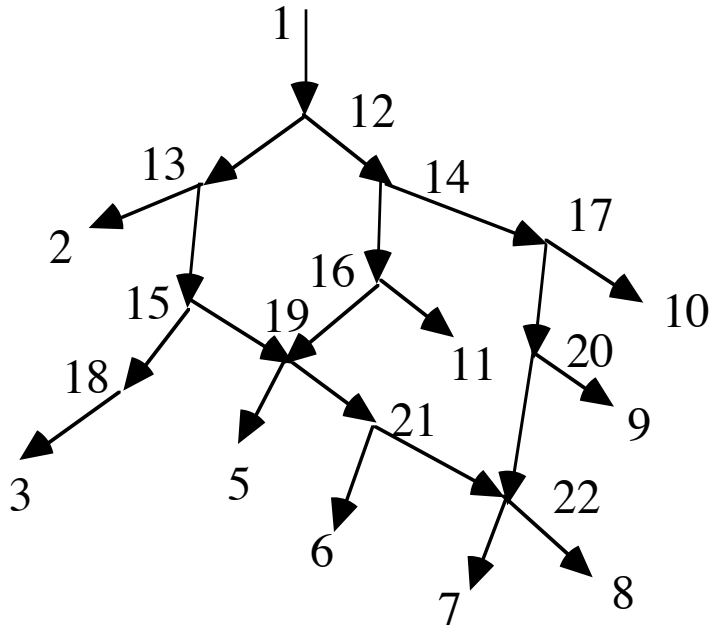




X = {1,2,3,4,5,6,7,8,9,10,11}

$\delta(4) = O(4) \neq \varnothing$
$M(18) = M(4) - \delta(4)$

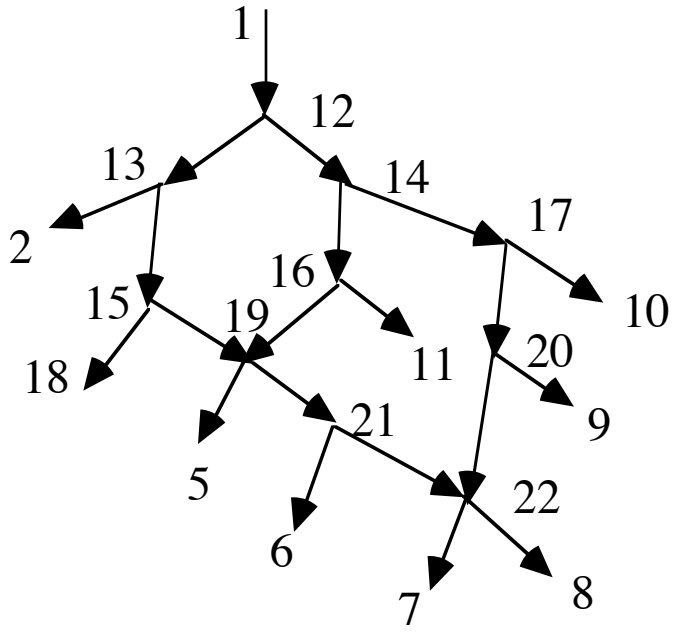
Remove 4 and insert 18. Remember the arc (18,4).



X = {1,2,3,5,6,7,8,9,10,11,18}

$\delta(3) = O(3) \neq \varnothing$
Its parent 18 has M(18) = M(3) - δ(3), which is already a genome in X.

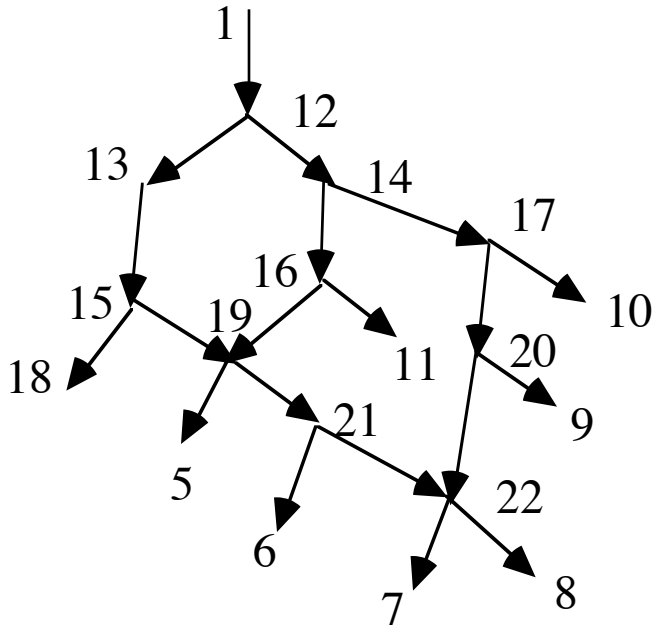
Remove 3. Remember the arc (18,3).



X = {1,2,5,6,7,8,9,10,11,18}

$\delta(2) = O(2) \neq \varnothing$
$M(13) = M(2) - \delta(2)$

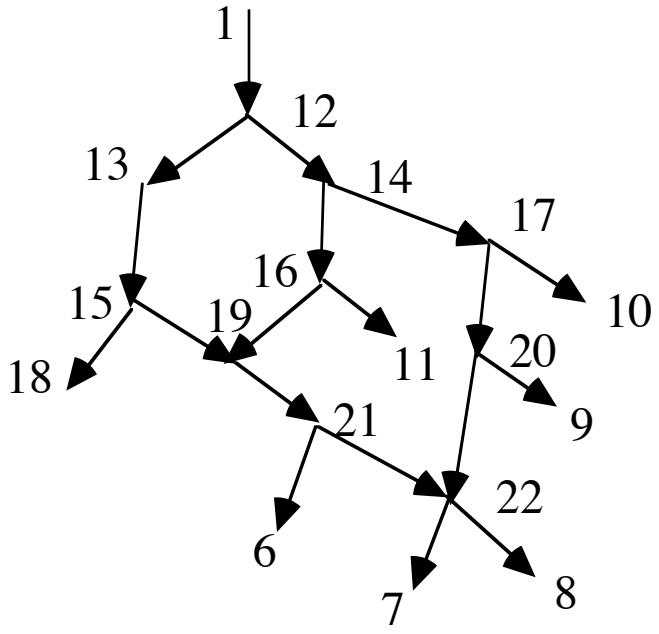
Remove 2, insert 13, and remember (13,2).



X = {1,5,6,7,8,9,10,11,13,18}

$\delta(5) = O(5) \neq \varnothing$
$M(19) = M(5) - \delta(5)$

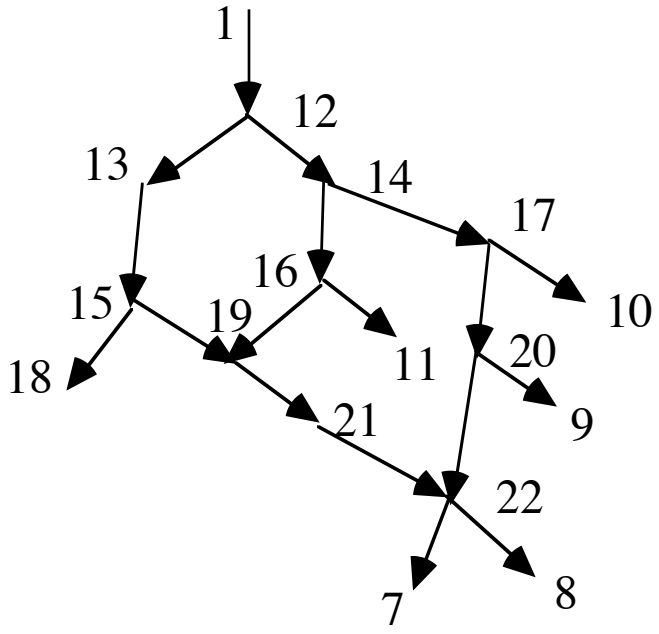
Remove 5, insert 19, and remember (19,5).



X = {1,6,7,8,9,10,11,13,18,19}

$\delta(6) = O(6) \neq \varnothing$
$M(21) = M(6) - \delta(6)$

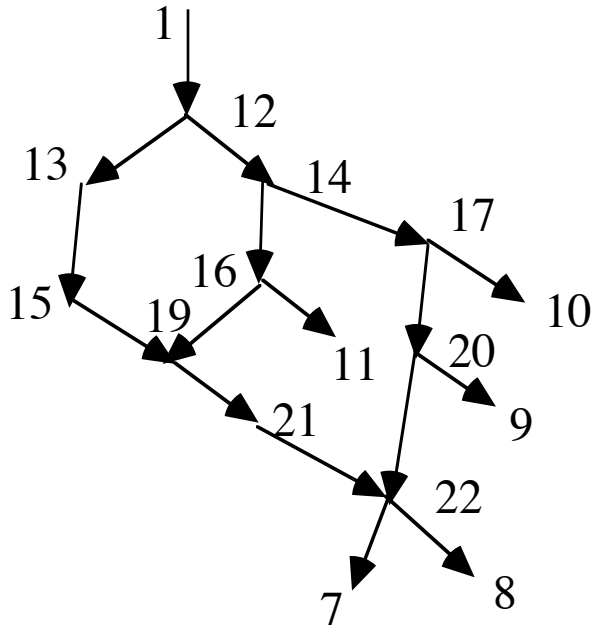
Remove 6, insert 21, and remember (21,6).



X = {1,7,8,9,10,11,13,18,19,21}

$\delta(18) = O(18) \neq \varnothing$
$M(15) = M(18) - \delta(18)$

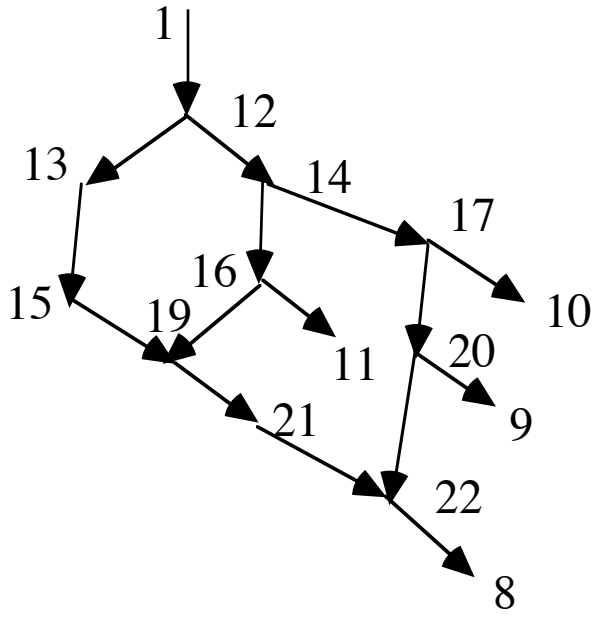
Remove 18, insert 15, and remember (15,18).



X = {1,7,8,9,10,11,13,15,19,21}

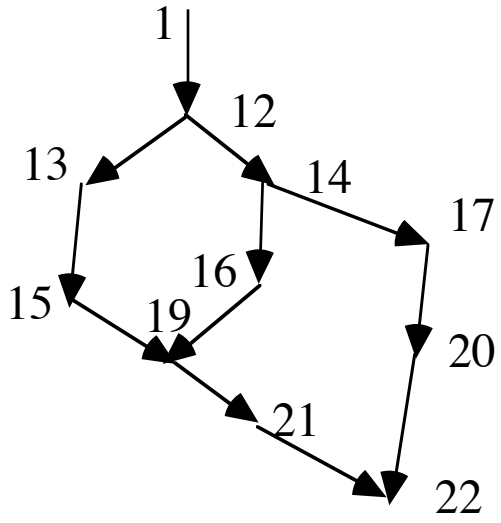
$\delta(7) = O(7) \neq \varnothing$
$M(22) = M(7) - \delta(7)$

Remove 7, insert 22, and remember (22,7).



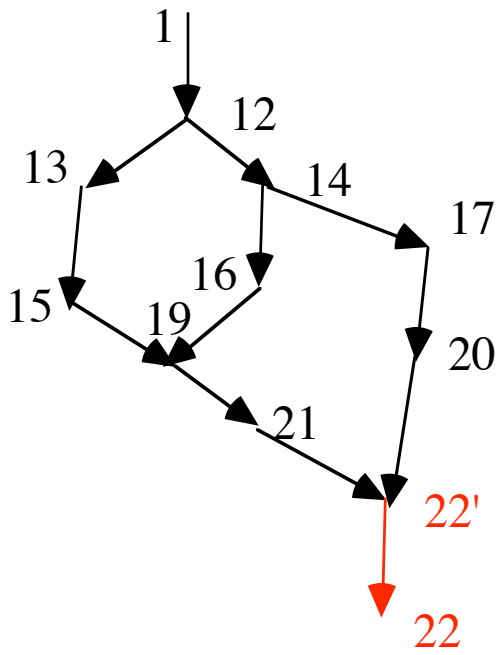
X = {1,8,9,10,11,13,15,19,21,22}

In this manner, remove recursively all the normal leaves.



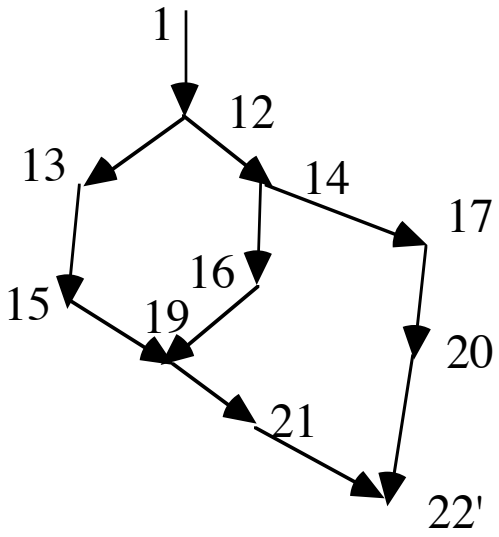
$X = \{1,13,15,16,17,19,20,21,22\}$

$\delta(22) = O(22) \neq \varnothing$
Give 22 a "separated parent" 22' with $M(22') = M(22) - \delta(22)$ and $O(22') = \varnothing$.



$O(22') = \varnothing$

Remove 22, insert 22', and remember the arc (22',22).

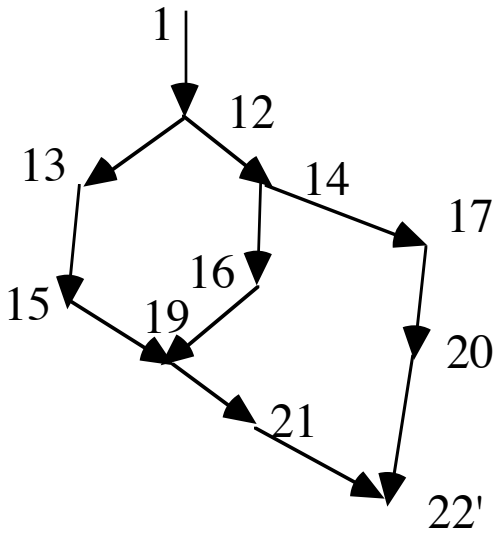


X = {1,13,15,16,17,19,20,21,22'}

Now for all x in X, $\delta(x) = \varnothing$. There are no normal leaves. Every hybrid leaf is separated.

Lemma. Assume for all $y \in X$, $\delta(y) = \emptyset$. Then there exists a hybrid leaf x with parents p and q such that x is the only child of p, and x is the only child of q.

Moreover, x, p, and q are in X.
Neither p nor q is equal to r.
If $\delta(x,a,b) = \emptyset$ then either (p=a and q=b) or (p=b and q = a).

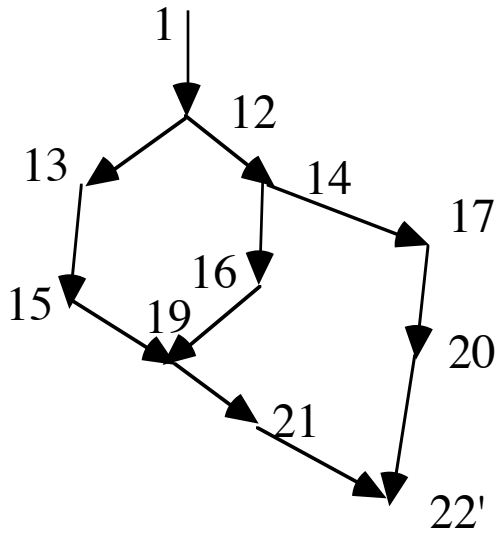


X = {1,13,15,16,17,19,20,21,22'}

Criterion to recognize a hybrid leaf x with parents p and q.

Lemma. Assume for all $y \in X$, $\delta(y) = \varnothing$.
Assume that x, p, and q are distinct members of X, all distinct from r. Assume
(1) $\delta(x,p,q) = \varnothing$.
(2) If $\delta(x,a,b) = \varnothing$ then either (p=a and q=b) or (p=b and q = a).

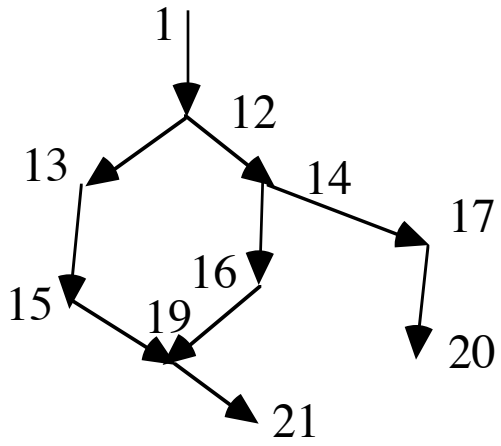
Then x is a hybrid leaf with parents p and q;
x is the only child of p;
and x is the only child of q.



X = {1,13,15,16,17,19,20,21,22'}

22' is a hybrid leaf with parents 20 and 21.

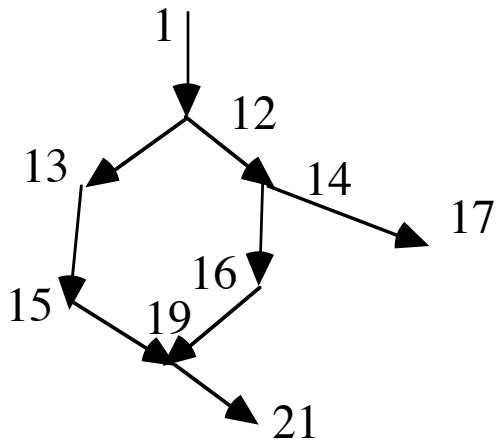
Remove 22' and remember (21,22'), (20,22').



X = {1,13,15,16,17,19,20,21}

$\delta(20) = O(20) \neq \varnothing$
M(17) = M(20) - δ(20) so the parent of 20 is already present.

Remove 20, and remember (17, 20).

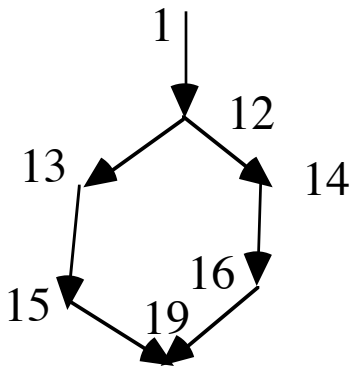


X = {1,13,15,16,17,19,21}

$\delta(21) = O(21) \neq \varnothing$
M(19) = M(21) - δ(21) so the parent of 21 is already present.

$\delta(17) = O(17) \neq \varnothing$
M(14) = M(17) - δ(17)

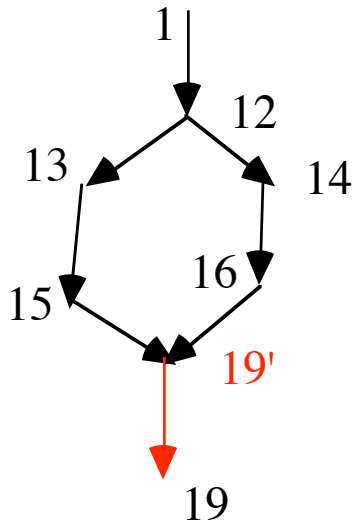
Remove 17 and 21; insert 14. Remember (14, 17) and (19, 21).



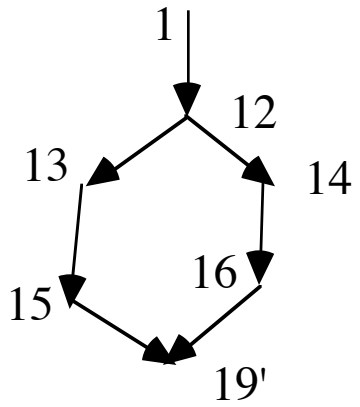
$X = \{1,13,14,15,16,19\}$

$\delta(19) = O(19) \neq \varnothing$

The procedure leads to a separated vertex 19'.



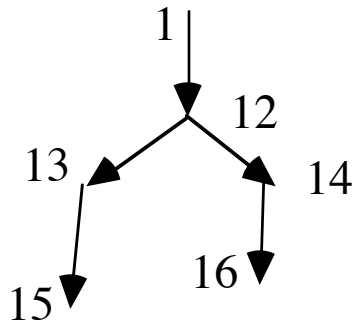
M(19') = M(19) - δ(19) so the parent of 19 is the separated hybrid vertex 19'.
Remove 19, insert 19', and remember (19', 19).



X = {1,13,14,15,16,19'}

Now $\delta(x) = \varnothing$ for all x.
Find x, p, q, so $\delta(x,p,q) = \varnothing$ uniquely, etc. This identifies 19', 15, 16.

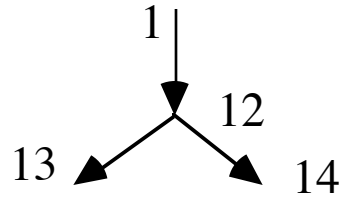
Remove 19' and remember (15,19'), (16, 19').



$X = \{1,13,14,15,16\}$

$\delta(15) = O(15) \neq \varnothing$
$\delta(16) = O(16) \neq \varnothing$

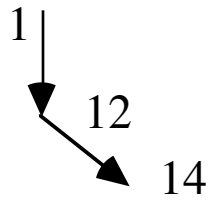
Remove 15 and 16. Their parents 13 and 14 are already present.



$X = \{1,13,14\}$

$\delta(13) = O(13) \neq \varnothing$
$M(12) = M(13) - \delta(13)$

Remove 13, insert 12, remember (12,13).

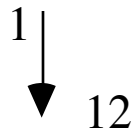


$X = \{1, 12, 14\}$.

$\delta(14) = \delta(14,1,12) = O(14) \neq \varnothing$

$M(12) = M(14) - O(14)$

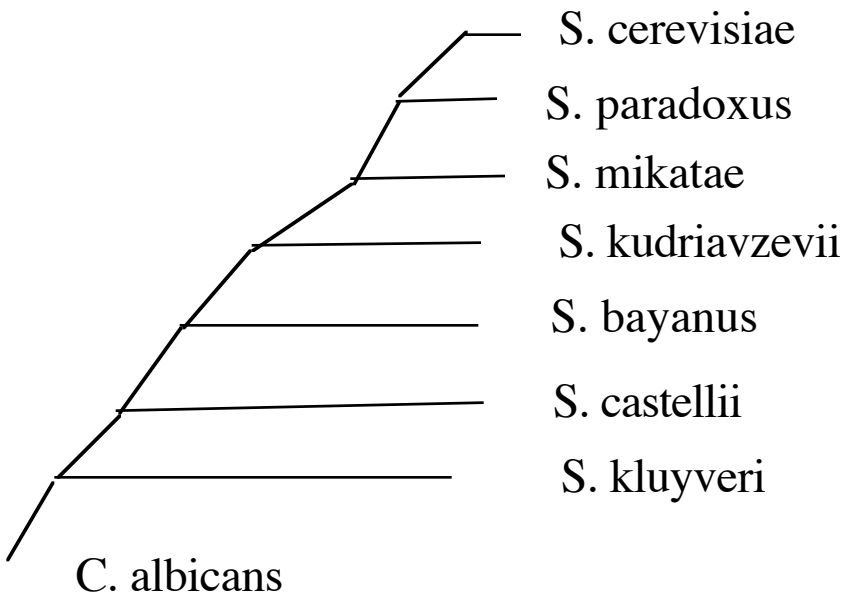
Remove 14 and remember (12,14).



There are only 2 vertices, one the root. Remember (1,12), and O(12) = M(12).

Implementation in a computer program

Rokas et al 2003 analyzed a set of 106 yeast genes from total database with 127026 nucleotide sites. There were 7 yeast genomes, genus Saccharomyces, and one outgroup from genus Candida. They concatenated the aligned genes and obtained a tree with 100% bootstrap support at each internal edge.



The tree found by Rokas et al 2003.

Holland, Huber, Moulton, Lockhart 2004 analyzed the data set. They found consensus networks for the 106 maximum-likelihood trees and also for the 106 maximum-parsimony trees.

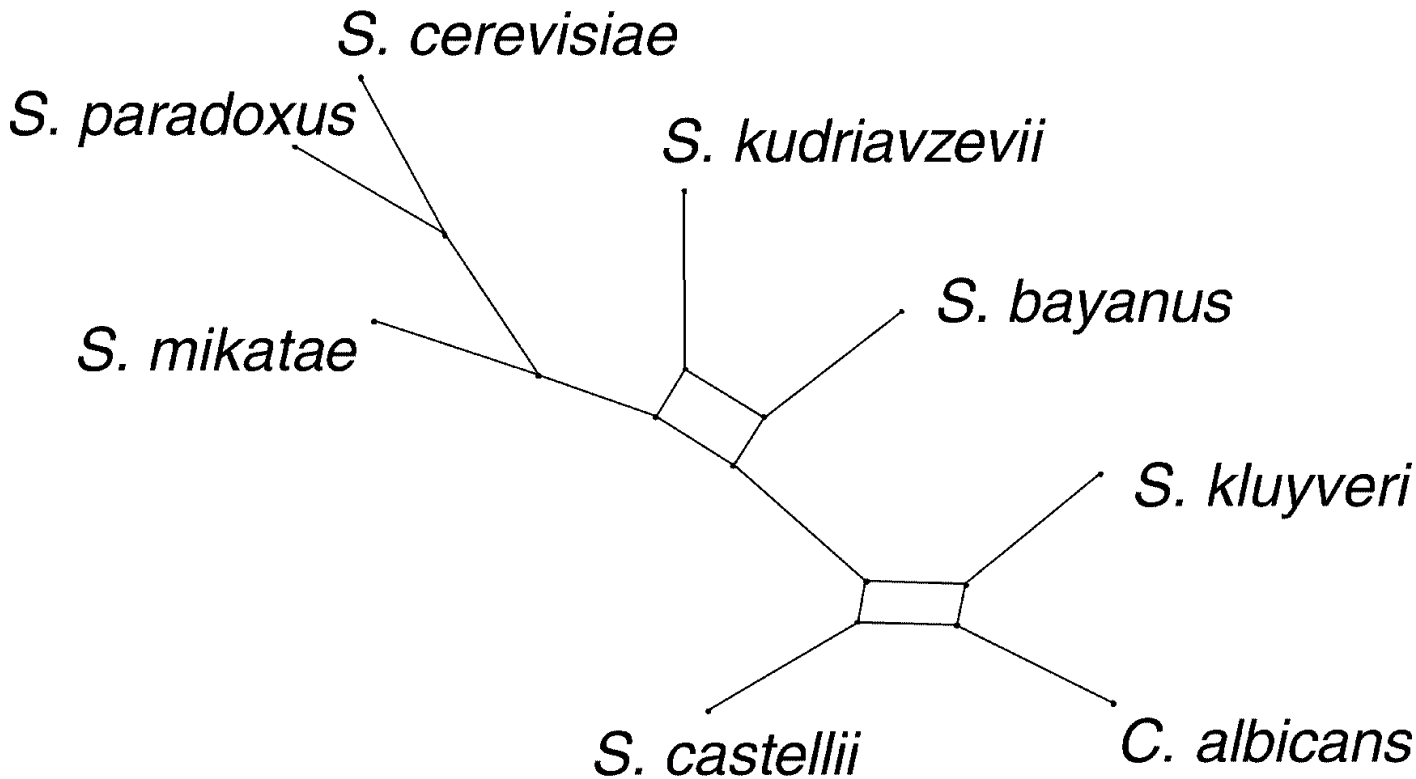
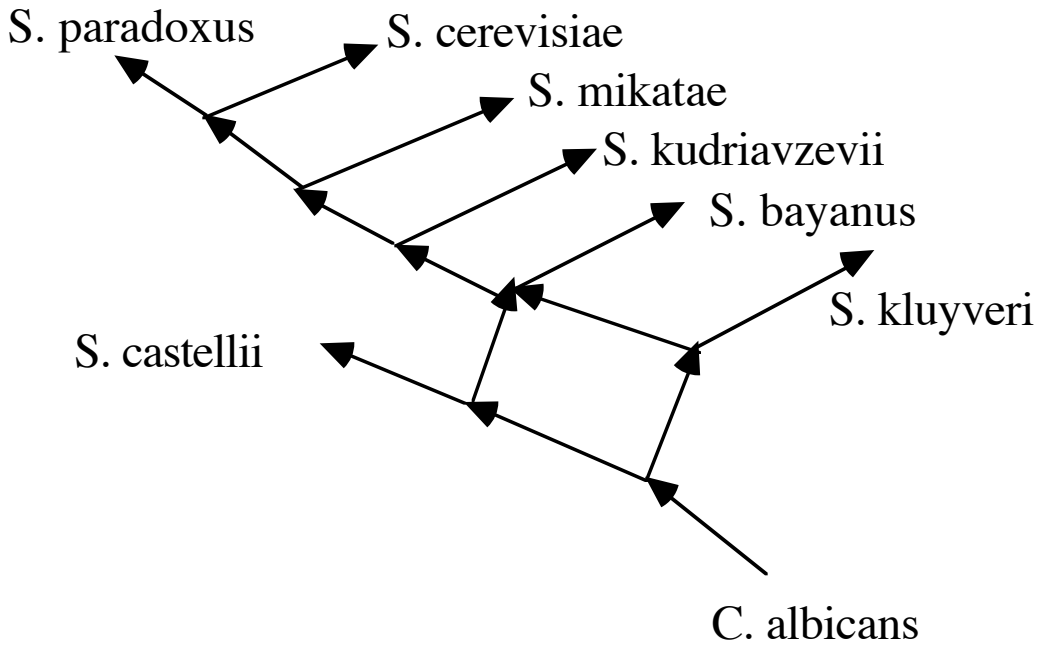


FIG. 1.—Consensus networks have been reconstructed for (*a*) 106 random bifurcating trees on eight taxa [A, B, C, . . . , H], (*b*) the 106 maximum-likelihood trees obtained in Rokas et al. (2003), and (*c*) the 106 maximum-parsimony trees obtained in Rokas et al. (2003). The presence of boxes in these networks indicates contradictory evidence for grouping certain species together. The lengths of the edges are proportional to the number of gene trees in which a particular edge occurs. Each network displays all those edges that are represented in at least 10 of the 106 trees.

Apply the interactive program to the concatenated sequences. It detects signal for



An Extension

There is a very similar result for reconstructing the network from distances between the members of X. Slightly stronger assumptions are needed both on the network and on the model of evolution.

Questions

1. Is there a better way to identify that a given member of X is hybrid?

2. What other families of networks have similar theorems?

3. Is there a way to include some (limited) homoplasies at normal vertices?

4. Can some kind of Markov process govern inheritance instead of the Simple Homoplasy Model?

Acknowledgments

Thanks to the **Isaac Newton Institute for Mathematical Sciences** for its hospitality at an excellent facility.

Thanks to the organizers **Vincent Moulton, Mike Steel, and Daniel Huson** of the Phylogenetics Programme.

Thanks to the organizers **Mike Steel, Vincent Moulton, and Katharina Huber** of this workshop.

Thank you for your attention.

References

M. Baroni, C. Semple, and M. Steel. Hybrids in real time. Syst Biol 55 (2006) 46-56.

M. Bordewich, C. Semple. Computing the minimum number of hybridization events for a consistent evolutionary history. Discrete Applied Mathematics 155 (2007), 914-928.

B. Holland, K.T. Huber, V. Moulton, P.J. Lockhart. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. 21 (2004) 1459-1461.

A. Rokas, B. Williams, N. King, S. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425 (2003) 798-803.

S.J. Willson. Unique solvability of certain hybrid networks from their distances. Annals of Combinatorics 10 (2006) 165-178.

S.J. Willson. Unique determination of some homoplasies at hybridization events. Bulletin of Mathematical Biology 69 (2007) 1709-1725.

S.J. Willson. Reconstruction of some hybrid phylogenetic networks with homoplasies from distances. To appear in Bulletin of Mathematical Biology.