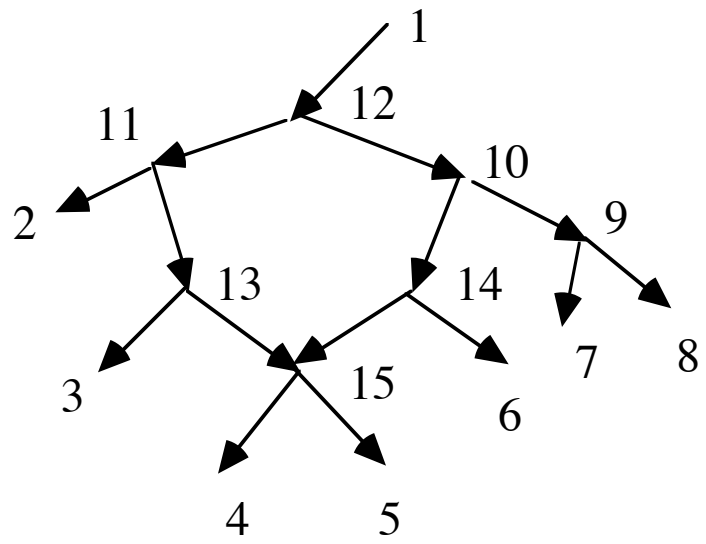
Restrictions on meaningful phylogenetic networks

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

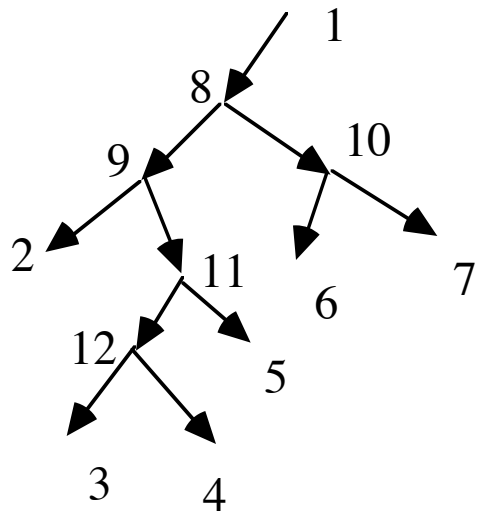
A **phylogenetic network** is an acyclic rooted directed graph *G* in which the leaves are identified with known taxa.



$X = \{1,2,3,4,5,6,7,8\}$ is the **base-set**.
X contains the vertices corresponding to taxa about which direct information is known. Other vertices are reconstructions.

The problem is to learn about *G* from information on *X*.

What makes phylogenetic trees meaningful?



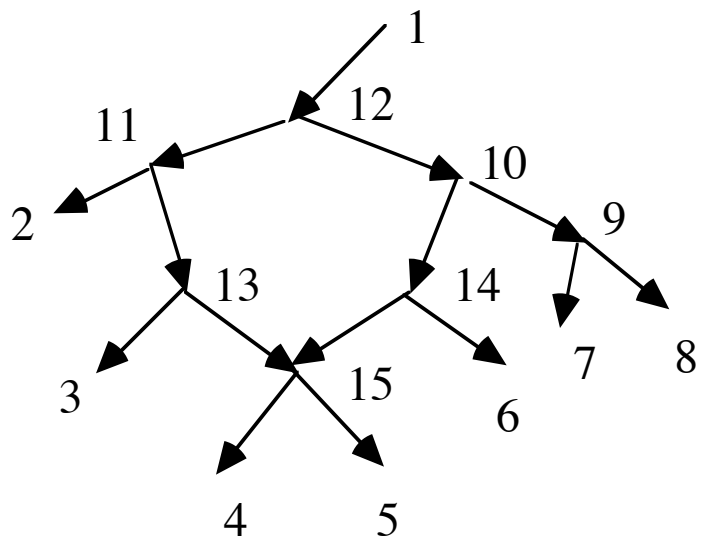
- Look for a most recent common ancestor:

  mrca(2,6) = 8.

- Predict where some character evolved:

  If a character is observed in taxa 2, 3, 4, 5 and nowhere else, where did it most likely arise?

Write $v \leq w$ if there is a directed path (possibly trivial) from v to w.

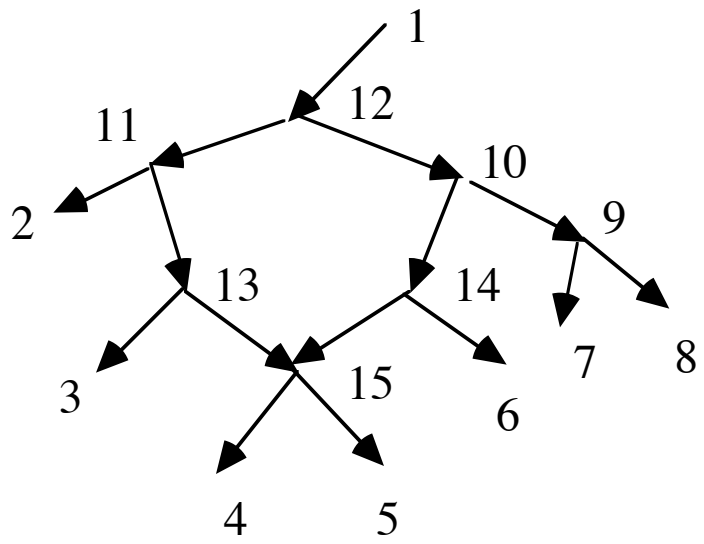


$11 \leq 4$
It is false that $11 \leq 6$.

Interpret $v \leq w$ to mean that v has the possibility of directly contributing to the genome of w.

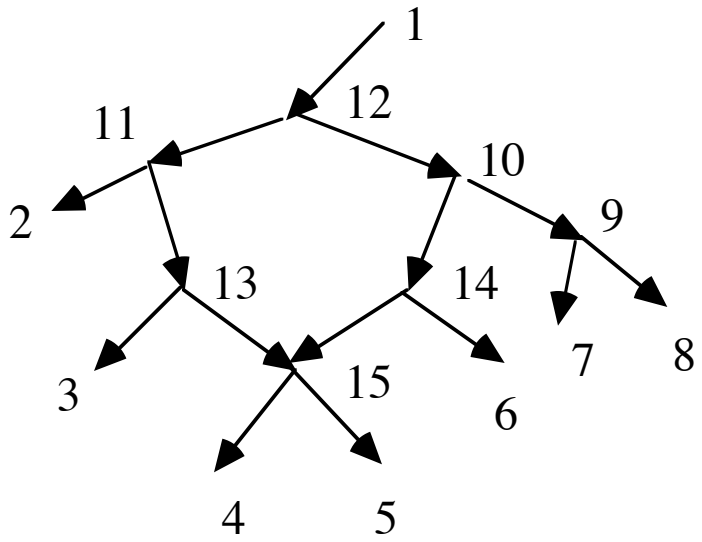
What we can learn depends on our model of evolution.

Let C be the set of **characters**. Assume
(1) Each character is binary. (An allele agrees with the root or disagrees.)
(2) Each vertex v has a **genome** $g(v) \subseteq C$
where $i \in g(v)$ iff v exhibits the allele of i different from the allele at the root.



$C = \{d,e,f,h,i,j,k\}$
$g(3) = \{d,e,h\}$ means:
(1) On characters d, e, h the genome of 3 disagrees with the genome at the root 1.
(2) On characters f, i, j, k the genome of 3 agrees with the genome at the root 1.

Evolution Model 1. (**Monotonic**, or **gene-aggregation**).
Every informative character i mutates exactly once (at u^i).
All descendents of u^i (and only those descendents) exhibit the modified character i.

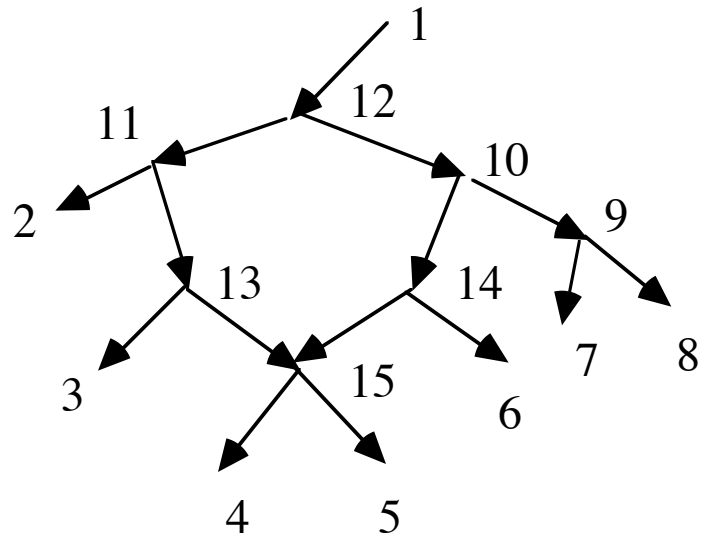


If $u^i = 10$, then modified character i is present in precisely {4,5,6,7,8,9,10,14,15}.

Following Baroni, Semple, and Steel 2004, form the **cluster map**

$c: V \to \mathcal{P}(X)$ by

$c(v) = \{x \in X: v \leq x\}$.

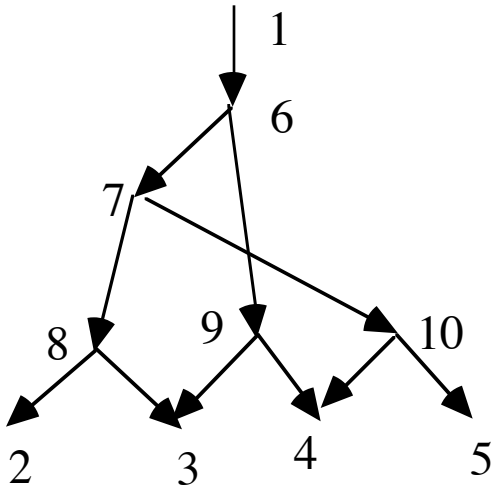


$c(11) = \{2,3,4,5\}$

$c(6) = \{6\}$

If a character originates at 11, we see it expressed at $\{2,3,4,5\}$.

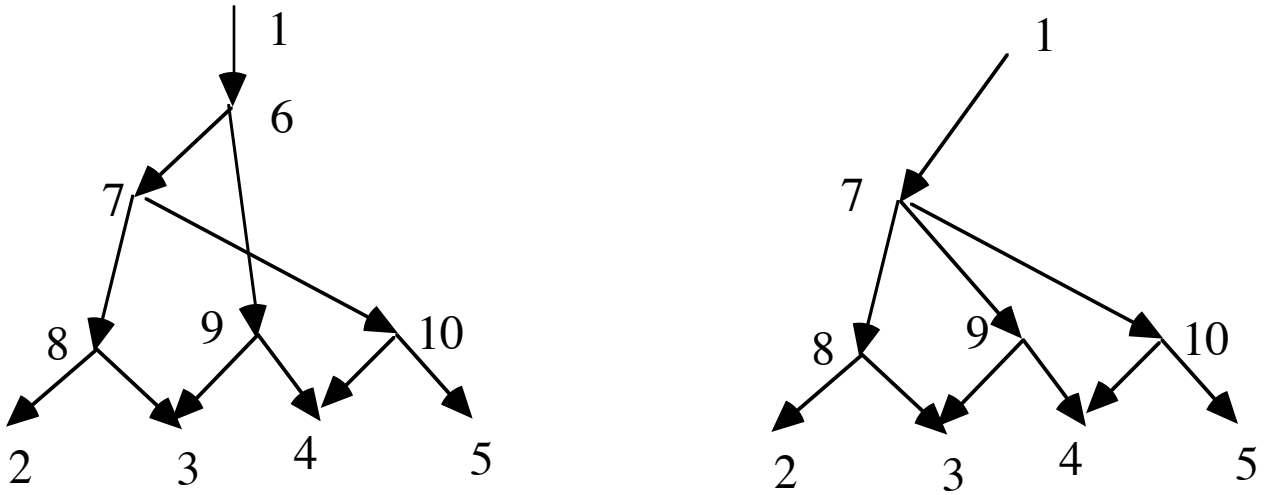
Example.



Here $X = \{1,2,3,4,5\}$
$c(6) = \{2,3,4,5\} = c(7)$

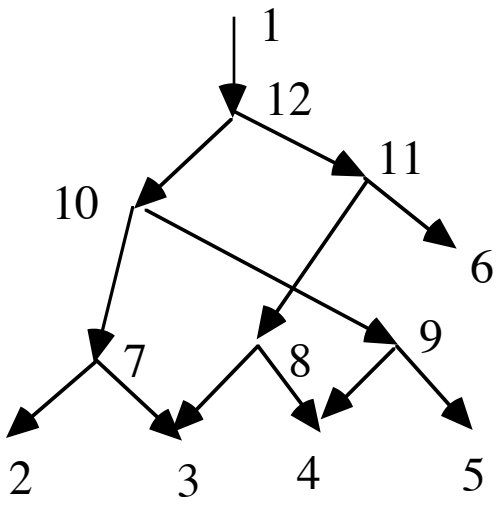
We could not tell whether a character in $\{2,3,4,5\}$ originates at 6 or at 7 by looking at genomes of members of X.

These two networks will have the same patterns on X:

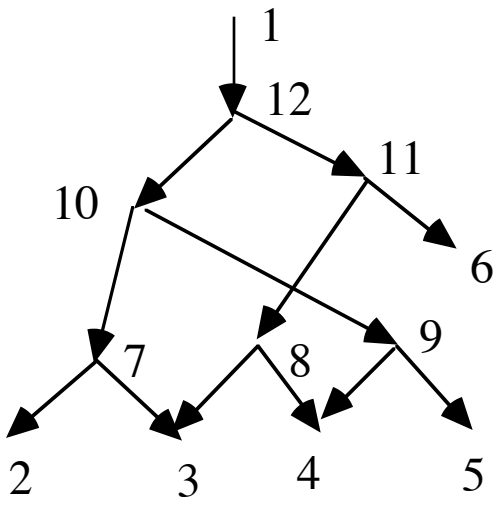


To be able to determine where a character i originates under this simple model we must have that c is one-to-one.

Example.



c(1) = {1,2,3,4,5}
c(2) = {2}
c(3) = {3}
c(4) = {4}
c(5) = {5}
c(6) = {6}
c(7) = {2,3}
c(8) = {3,4}
c(9) = {4,5}
c(10) = {2,3,4,5}
c(11) = {3,4,6}
c(12) = {2,3,4,5,6}.

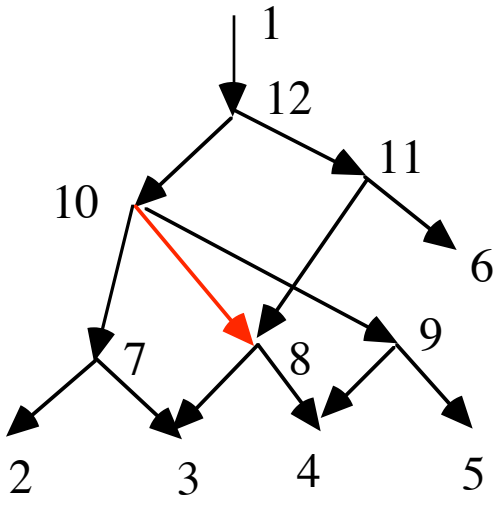


Hence *c* is one-to-one.

c(8) = {3,4}
c(10) = {2,3,4,5}

Note *c(8)* $\subseteq$ *c(10)* but it is false that *10* $\leq$ *8*.

But compare:



Note that *c(10) = {2, 3, 4, 5}*, exactly as before.

The genomes at all the members of *X* are exactly the same as before.

Assuming Model 1, we can't distinguish the two graphs on the basis of information on *X*.

Moral. Assume Model 1. If G is to exhibit all the genetic influences that could determine the genomes at members of X, then

$u \preceq v$ iff $c(v) \subseteq c(u)$.

Proof.
If $u \preceq v$ then $c(v) \subseteq c(u)$, because if $x \in X$ and $v \preceq x$, then $u \preceq v \preceq x$.

Suppose $c(v) \subseteq c(u)$ but it is false that $u \preceq v$.
Then there is some possible influence of u on v that is not described by the network.
Adding the arc (u,v) would not change any genomes of X.
QED

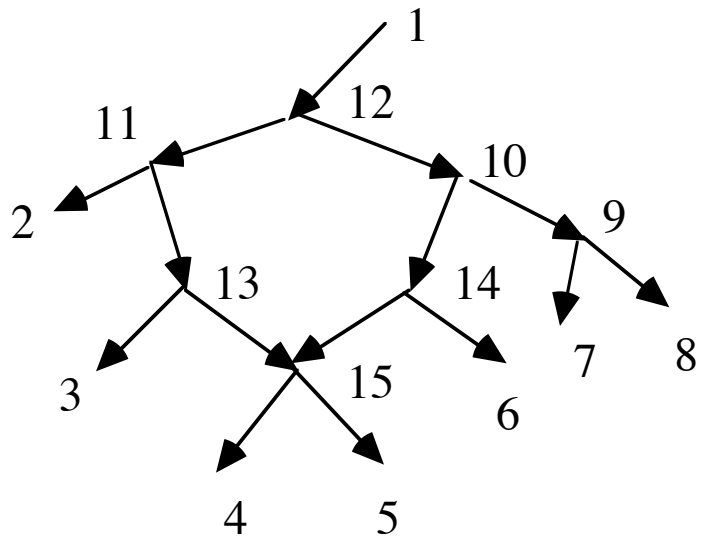
Definition. (Baroni, Semple, Steel 2004) The network G is **regular** iff
(1) c is one-to-one; and
(2) $u \preceq v$ iff $c(v) \subseteq c(u)$.

Theorem. Assume Model 1. Suppose
(1) G is to have the genomes of all vertices determined by the genomes at members of X; and
(2) G is to exhibit all the genetic influences that could determine the genomes at members of X.
Then G must be regular.

I argue that any meaningful network must be regular.
Otherwise, it is not justified by the data on X alone in this simplest of models for inheritance.

Model 1 is very strong, especially at hybrid vertices.

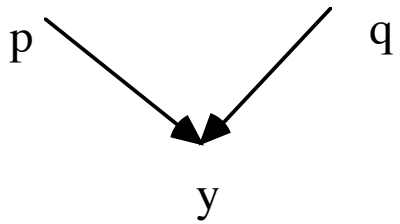
A vertex *v* is **hybrid** if it has at least two parents. A vertex *v* is **normal** if it has at most one parent.



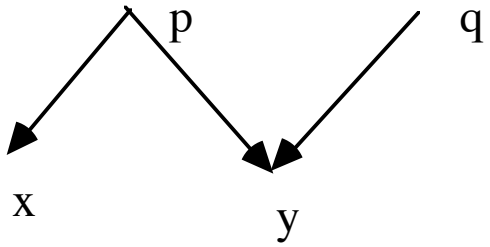
15 is hybrid.
All other vertices are normal.

Objection to Model 1:

If *y* has parents *p* and *q*, it is unlikely that it inherits all the modified genes in both *p* and *q*. Such inheritance might happen in polyploidy but not in general. We should allow homoplasies at hybrid vertices.



Another difficulty.



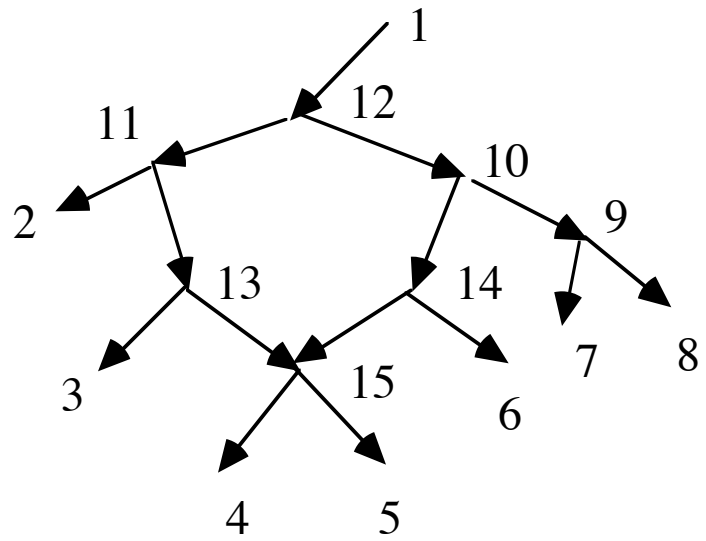
If $i \in g(x)$ but $i \notin g(y)$, maybe
(1) i originates at x; or
(2) i originates at p, but doesn't get inherited by y.

Such an appearance at p but immediate disappearance in p's child y is an **immediate homoplasy**.

Assume there are no immediate homoplasies.

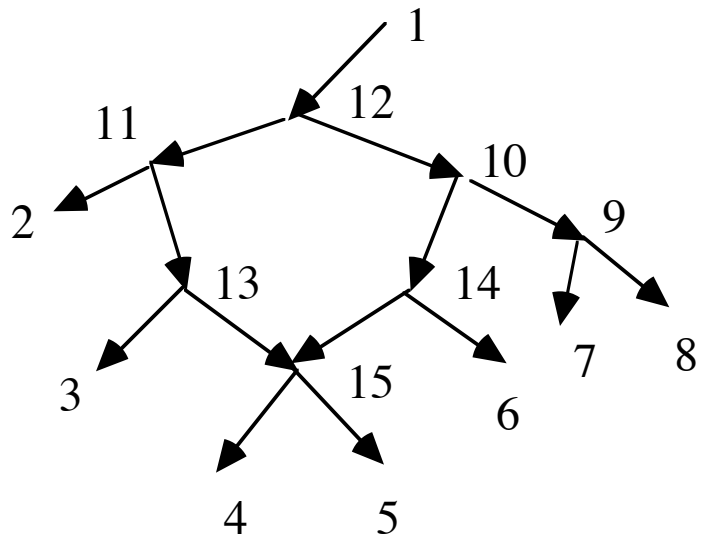
Model (2) Monotonic with hybrid homoplasies

Assume normal vertices inherit all the modified characters in their parent.
Assume hybrid vertices may or may not inherit a character from a parent.
Assume there are no immediate homoplasies.

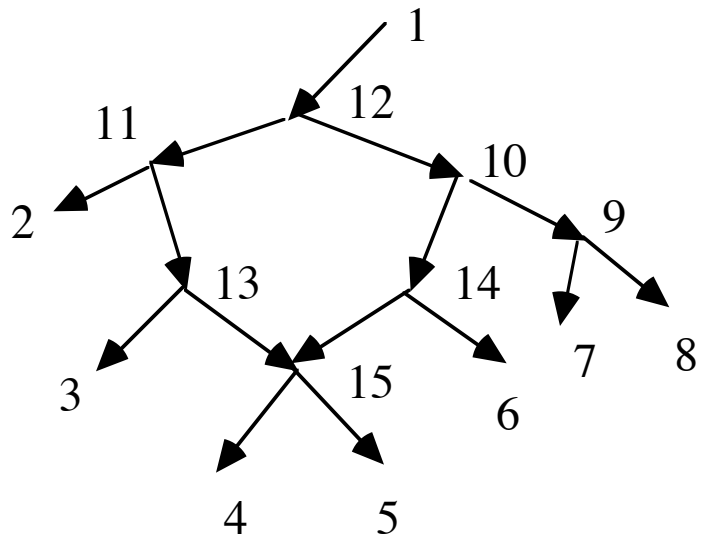


If $u^i = 9$ then possible patterns are {7, 8}
If $u^i = 10$ then possible patterns are {4, 5, 6, 7, 8}, {6, 7, 8}
If $u^i = 11$ then possible patterns are {2, 3, 4, 5}, {2, 3}
If $u^i = 12$ then possible patterns: {2, 3, 4, 5, 6, 7, 8}, {2, 3, 6, 7, 8}
If $u^i = 13$ then possible patterns are {3, 4, 5}
If $u^i = 14$ then possible patterns are {4, 5, 6}
If $u^i = 15$ then possible patterns are {4, 5}

Note {6} is not a possibility if $u^i = 14$ since it requires an immediate homoplasy.



If $u^i = 2$ then possible patterns are {2}
If $u^i = 3$ then possible patterns are {3}
If $u^i = 4$ then possible patterns are {4}
If $u^i = 5$ then possible patterns are {5}
If $u^i = 6$ then possible patterns are {6}
If $u^i = 7$ then possible patterns are {7}
If $u^i = 8$ then possible patterns are {8}



Each of the indicated patterns occurs exactly once in the example.

Define the **generalized cluster map**

$gc: V\text{-}\{r\} \rightarrow \mathcal{P}(\mathcal{P}(X))$

$gc(v) = \{ U: U \subseteq X$ and U arises under Model (2) as the pattern that tells where some character i originating at v is exhibited$\}$.

$gc(10) = \{\{4,5,6,7,8\}, \{6,7,8\}\}$

The generalized cluster map gc is **generalized injective** if
(1) no $U \subseteq X$ appears in both $gc(u)$ and $gc(v)$, and
(2) $\varnothing$ never appears in $gc(u)$.

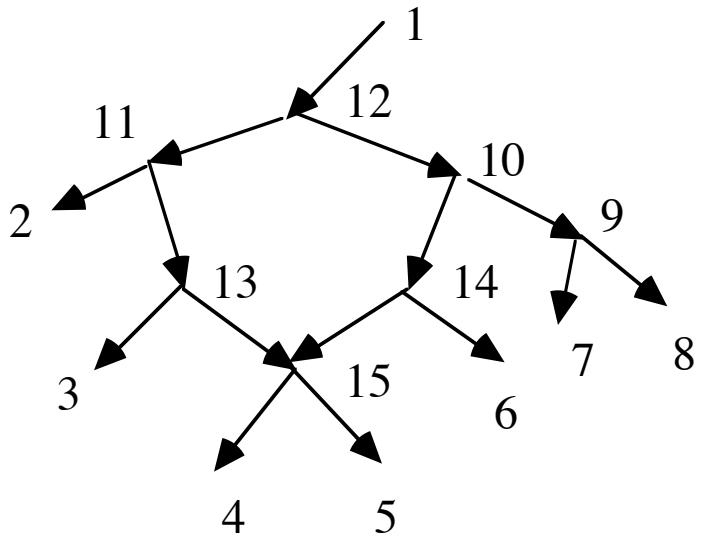
Moral. If the genomes of the vertices are to be uniquely determined by information on X under Model 2, then gc is generalized injective.

Proof. If $U \in gc(u)$ and $U \in gc(v)$, then when $\{x \in X: i \in g(x)\} = U$, we cannot tell whether $u = u^i$ or $v = u^i$. QED

I argue that if G is a meaningful network, then its generalized cluster map should be generalized injective.

How do we recognize such networks?

A **normal path from *u* to *v*** is a directed path from *u* to *v* such that every vertex after *u* is normal. Note *u* may or may not be hybrid.



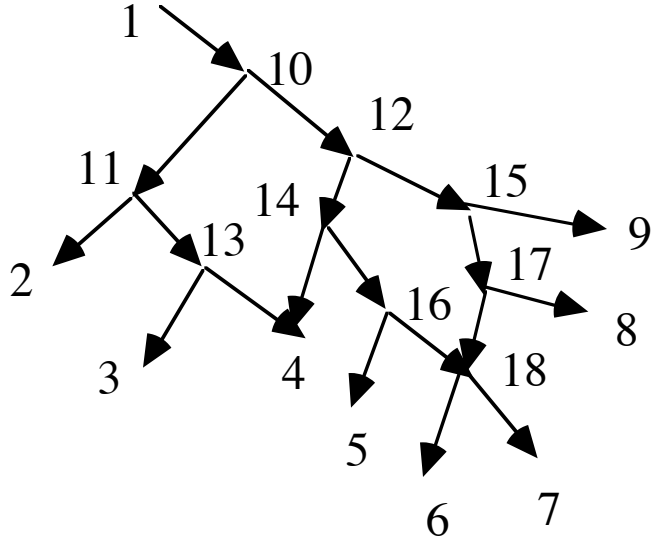
The path 12, 11, 2 is normal.
The path 14, 15, 5 is not normal.

There is a **normal path from *u* to *X*** if there is a normal path from *u* to some $x \in X$.
There is a normal path from 10 to *X* given by 10, 9, 7.
There is a normal path from 15 to *X* given by 15, 5.

Definition. A network is **normal** if from every vertex *u* there is a normal path to *X*.

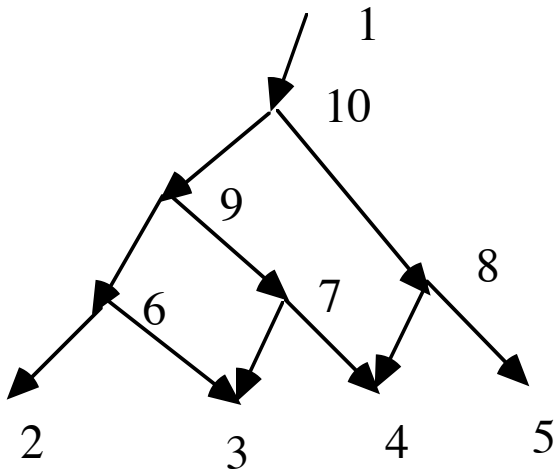
Main Theorem. Suppose *G* is a normal network. Then
(1) *G* is regular.
(2) *gc* is generalized injective.

Example.



G is normal. Hence *G* is regular, and *gc* is generalized injective.

Example.



Here $X = \{1,2,3,4,5\}$. There is no normal path from 7 to X.
gc(9) contains {2,3} and *gc(6)* contains {2,3}. Hence *gc* is not generalized injective.

Theorem.
Let X consist only of the leaves and the root. Suppose that every vertex other than the root has outdegree 0 or 2, while the root has outdegree 1. Assume that the generalized cluster map gc is generalized injective. Then the network is normal.

Moral. I propose that meaningful networks are normal.

How complicated are normal networks?

Theorem.
Suppose that $G = (V,A,r,X)$ is an acyclic rooted network and $|X| = n$ (including the root). Let $v = |V|$.

(1) If G is not regular, then v is unbounded.
(2) If G is regular, then $v \leq 2^{n-1}$.
(3) If G is a tree, then $v \leq 2n-2$.

How complicated are normal networks?

Theorem.
Suppose that $G = (V,A,r,X)$ is an acyclic rooted network and $|X| = n$ (including the root). Let $v = |V|$.

(1) If G is not regular, then v is unbounded.
(2) If G is regular, then $v \leq 2^{n-1}$.
(3) If G is a tree, then $v \leq 2n-2$.
(4) If G is normal, then $v \leq (n^2-n+2)/2$

Thus the assumption of normality is a large simplification.

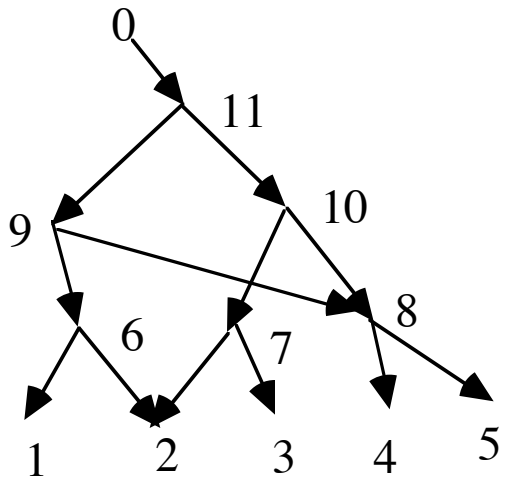
Let K be a nonempty set of vertices. Then the
most recent common ancestor of K
= mrca(K)
is the vertex v such that

(1) For all $k \in K$, $v \leq k$;

(2) Suppose that for all $k \in K$, $u \leq k$; then $u \leq v$

if it exists.

Example.



mrca(3,4) = 10.
mrca(2,4) does not exist since both 9 and 10 are candidates.

Some properties of normal networks:

(1) For each vertex v, $v = \text{mrca}(c(v))$.

(2) For each vertex v that is not in X, there exist x and y in X such that $v = \text{mrca}(x, y)$.

(3) There may exist x and y in X such that $\text{mrca}(x, y)$ does not exist.

Future work:

(1) Find more properties of normal networks.

(2) Models 1 and 2 are kinds of perfect phylogeny. Find analogous restrictions for other models of evolution and for characters that are not binary.

(3) What is the maximum number of vertices in a normal network with $|X| = n$? We know $v \leq (n^2-n+2)/2$.

(4) What are properties of "regularization" and "normalization"?

Summary

I suggest that we should seek normal networks, i.e., networks such that from every vertex u there is a normal path to X.