

Phylogenetic Networks of SNPs with Constrained Recombination

D. Gusfield, S. Eddhu, C. Langley

Nasty Typo Alert

Lemma 2.1 (page 4) in the proceedings paper omitted the key condition:

“Site i appears (mutates) on gall Q .”

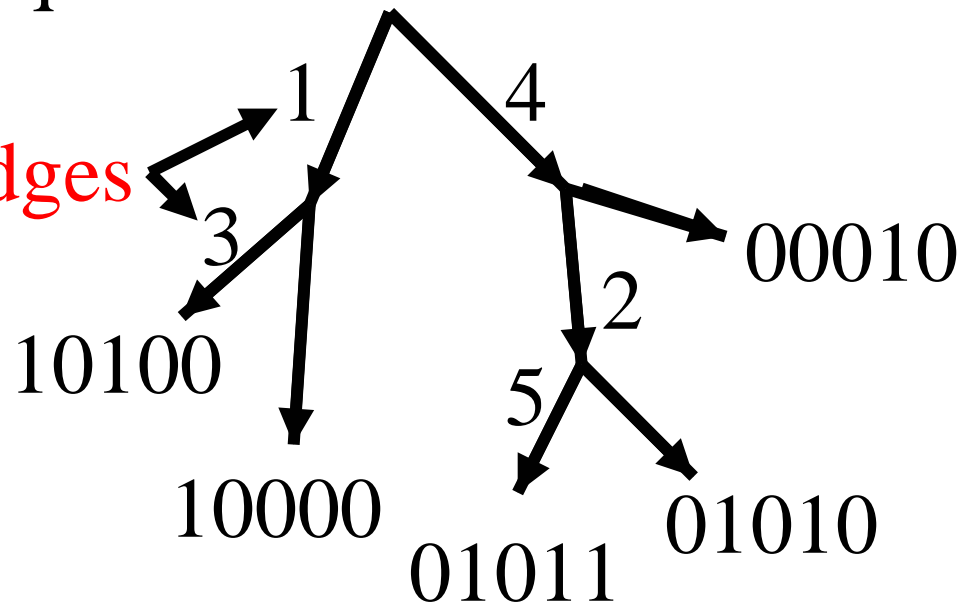
Reconstructing the Evolution of Binary Bio-Sequences (SNPs)

- Perfect Phylogeny (tree) model
- Phylogenetic Networks (DAG) with recombination
- Phylogenetic Networks with disjoint cycles: Galled-Trees
- Combinatorics of Galls and Galled-Trees
- Efficient Algorithms

The Perfect Phylogeny Model for SNPs - binary sequences

sites 12345
 Ancestral sequence 00000

Site mutations on edges



The tree derives the set M:

- 10100
- 10000
- 01011
- 01010
- 00010

Extant sequences at the leaves

Why SNPs?

SNPs imply that the sequences are binary, and that the order of the sites is fixed (on a chromosome). This is in contrast to a set of taxonomic characters, where the order is arbitrary.

The converse problem

Given a set of sequences M we want to find, if possible, a perfect phylogeny that derives M . Remember that each site can change state from 0 to 1 only once.

n will denote the number of sequences in M , and m will denote the length of each sequence in M .

		m			
M	n	<table border="1"><tr><td>01101001</td></tr><tr><td>11100101</td></tr><tr><td>10101011</td></tr></table>	01101001	11100101	10101011
01101001					
11100101					
10101011					

When can a set of sequences be derived on a perfect phylogeny with the all-0 root?

Classic NASC: Arrange the sequences in a matrix. Then (with **no** duplicate columns), the sequences can be generated on a **unique** perfect phylogeny if and only if no two columns (sites) contain all three pairs:

0,1 and 1,0 and 1,1

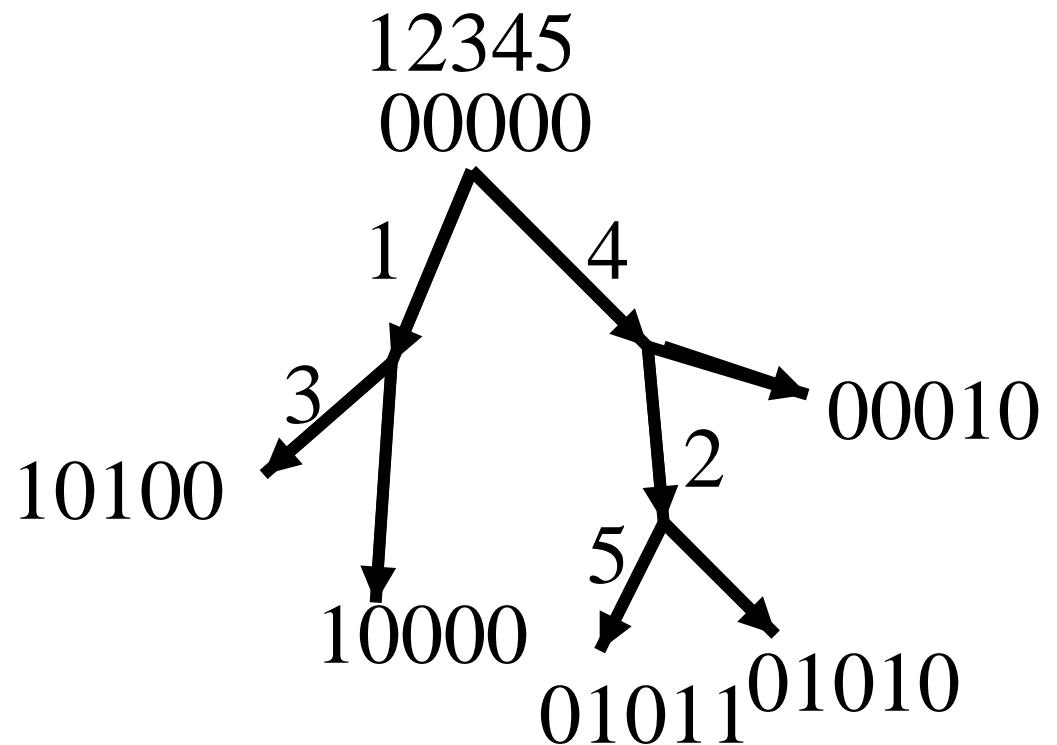
This is the 3-Gamete Test

A richer model

10100
10000
01011
01010
00010

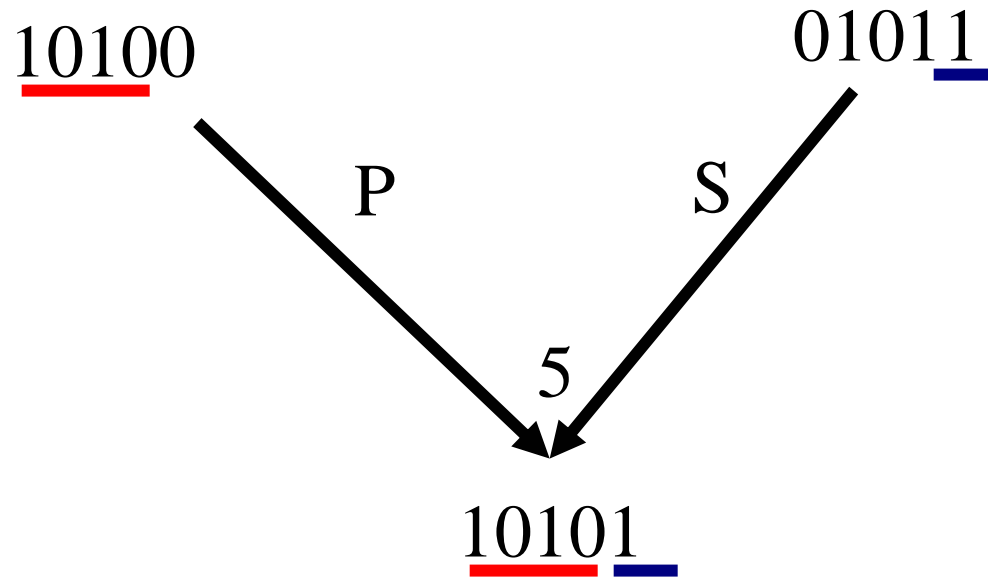
10101 new

pair 4, 5 fails the three gamete-test. The sites 4, 5 ``conflict''.



Real sequence histories often involve **recombination**.

Sequence Recombination



A recombination of P and S at recombination point 5.

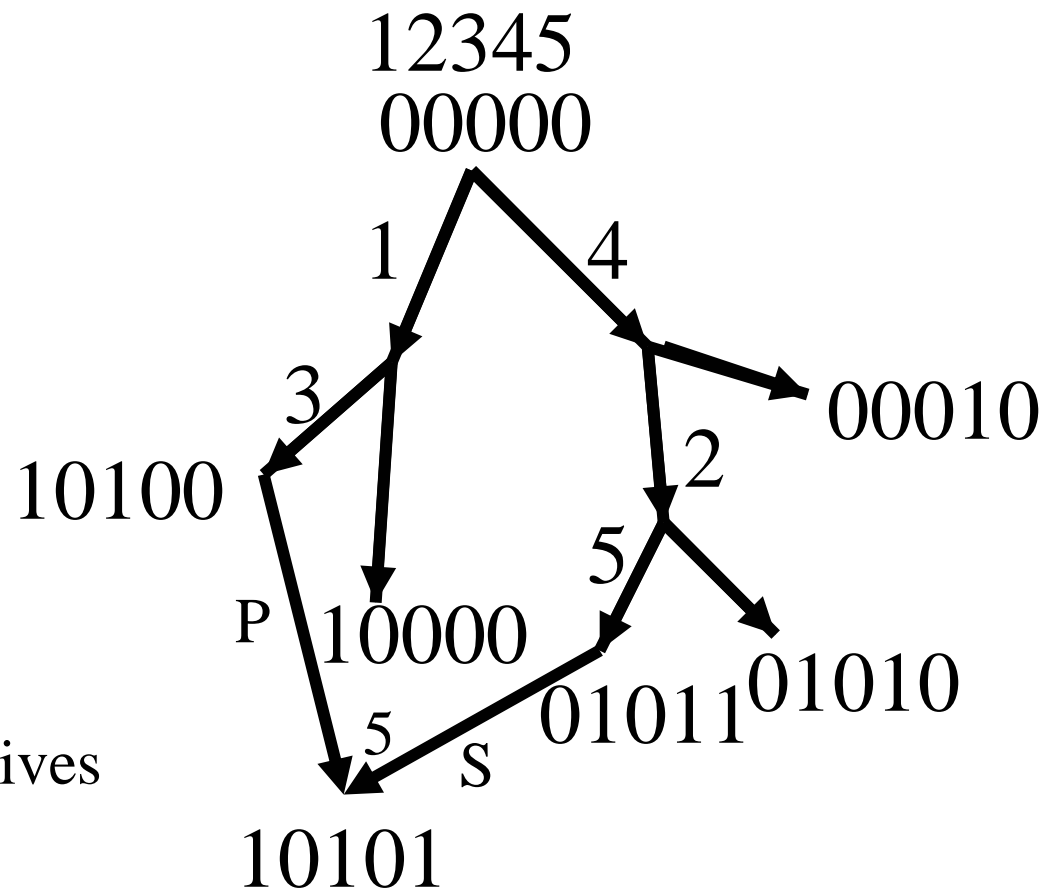
The first 4 sites come from P (Prefix) and the sites from 5 onward come from S (Suffix).

Perfect Phylogeny with Recombination

10100
10000
01011
01010
00010

10101 new

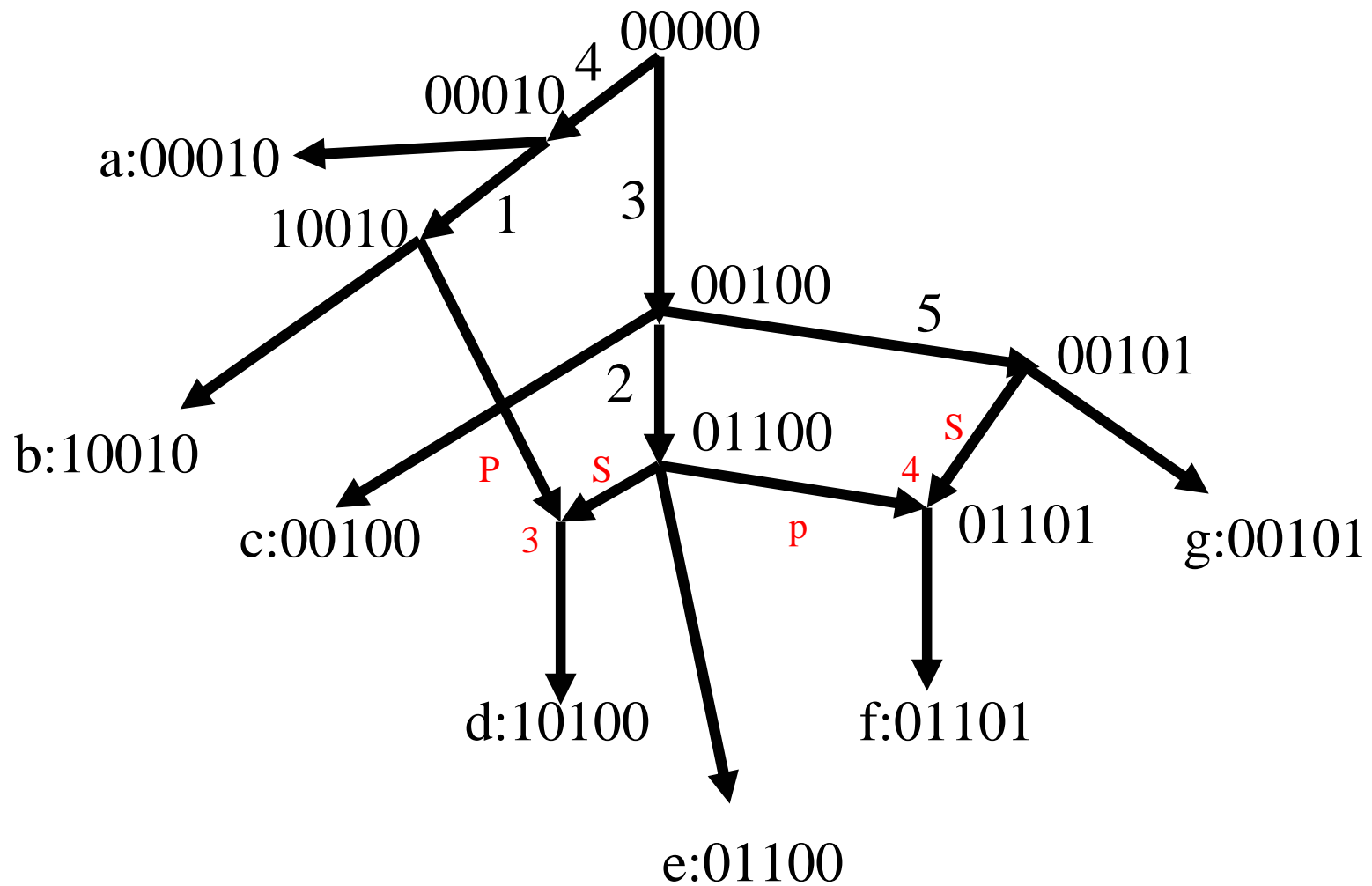
The previous tree with one recombination event now derives all the sequences.



Elements of a Phylogenetic Network

- Directed acyclic graph.
- Integers from 1 to m written on the edges. Each integer written only once. These represent mutations.
- Each node is labeled by a sequence obtained from its parent(s) and any edge label on the edge into it.
- A node with two edges into it is a “recombination node”, with a recombination point r . One parent is P and one is S .
- The network derives the sequences that label the leaves.

A Phylogenetic Network



Which Phylogenetic Networks are meaningful?

Given M we want a phylogenetic network that derives M , but which one?

A: A perfect phylogeny (tree) if possible. As little deviation from a tree, if a tree is not possible.

Minimizing recombinations

- Any set M of sequences can be generated by a phylogenetic network with enough recombinations, and one mutation per site. This is not interesting or useful.
- However, the number of (observable) recombinations is small in realistic sets of sequences. ``Observable'' depends on n and m relative to the number of recombinations.
- Two algorithmic problems: given a set of sequences M , find a phylogenetic network generating M , **minimizing** the number of recombinations. Find a network generating M that has some **biologically-motivated structural properties**.

Minimization is NP-hard

The problem of finding a phylogenetic network that creates a given set of sequences M , and minimizes the number of recombinations, is NP-hard. (Wang et al 2000)

They explored the problem of finding a phylogenetic network where the recombination cycles are **required to be node disjoint, if possible.**

They gave a sufficient but not a necessary condition to recognize cases when this is possible. $O(nm + n^4)$ time.

Recombination Cycles

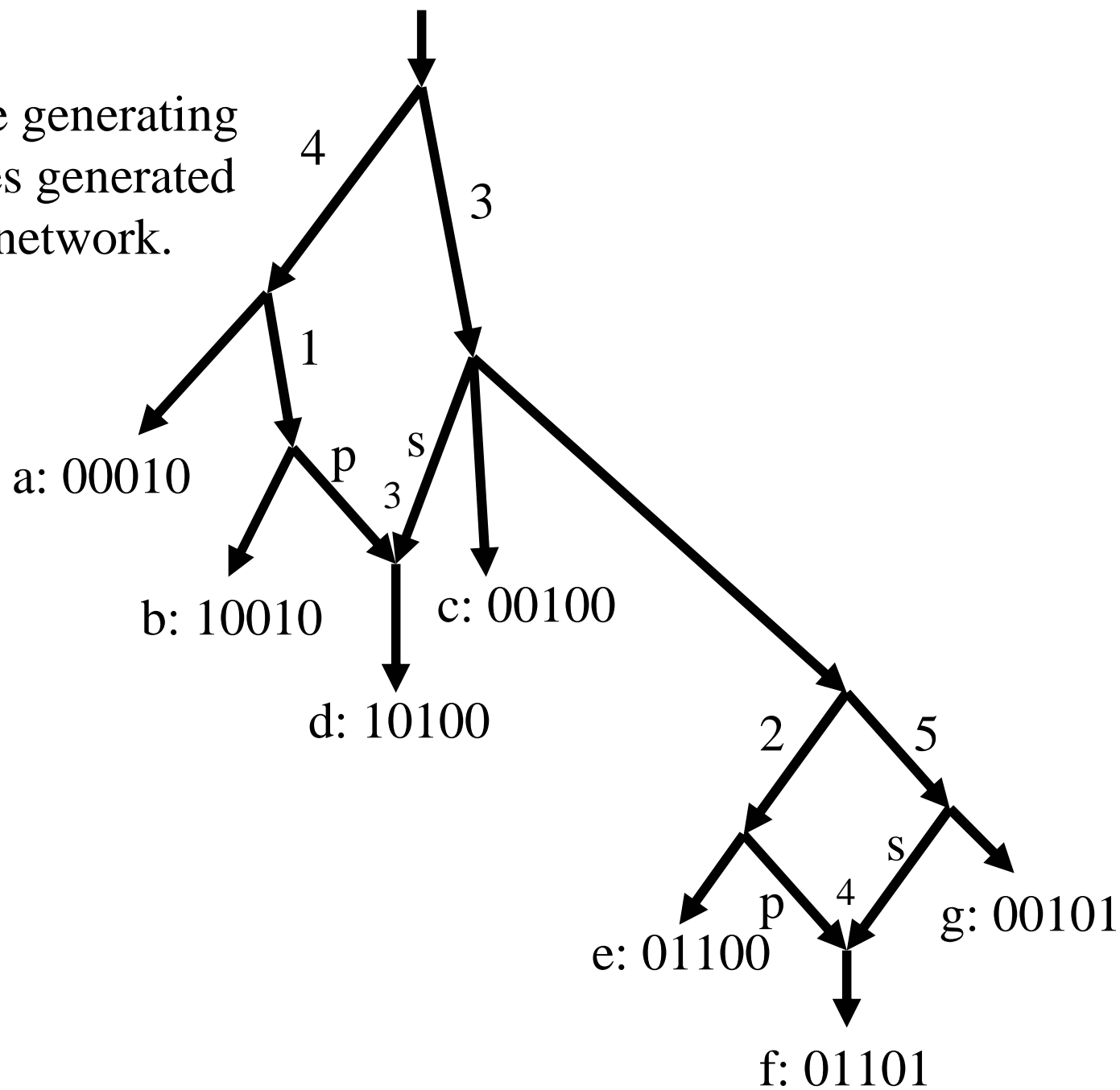
- In a Phylogenetic Network, with a recombination node x , if we trace two paths backwards from x , then the paths will eventually meet.
- The cycle specified by those two paths is called a “recombination cycle”.

Galled-Trees

A recombination cycle in a phylogenetic network is called a “gall” if it shares no node with any other recombination cycle.

A phylogenetic network is called a “galled-tree” if every recombination cycle is a gall.

A galled-tree generating the sequences generated by the prior network.



New Results

- $O(nm + n^3)$ -time algorithm to determine whether or not M can be derived on a galled-tree.
- Proof that the “canonical” galled-tree produced by the algorithm is a “nearly-unique” solution.
- Proof (not in the proceedings) that a canonical galled-tree (if one exists) minimizes the number of recombinations used, over all phylogenetic-networks that derive M .
- Understanding of some of the general structure of galls any phylogenetic network.

The start of technical stuff

Site Conflicts

A pair of sites (columns) of M that fail the 3-gametes test are said to **conflict**.

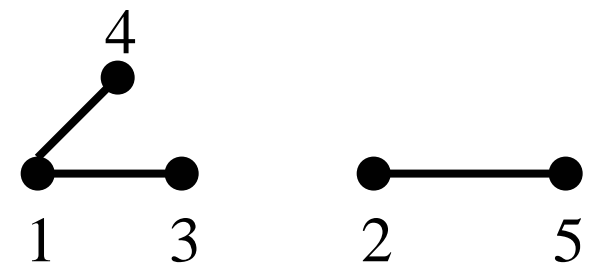
And each site in the pair is said to be **conflicted**.

A site that is not in such a pair is **unconflicted**.

	1	2	3	4	5
a	0	0	0	1	0
b	1	0	0	1	0
c	0	0	1	0	0
d	1	0	1	0	0
e	0	1	1	0	0
f	0	1	1	0	1
g	0	0	1	0	1

M

Conflict Graph



Two nodes are connected iff the pair of sites conflict, i.e, fail the 3-gamete test.

THE MAIN TOOL: We represent the pairwise conflicts in a conflict graph.

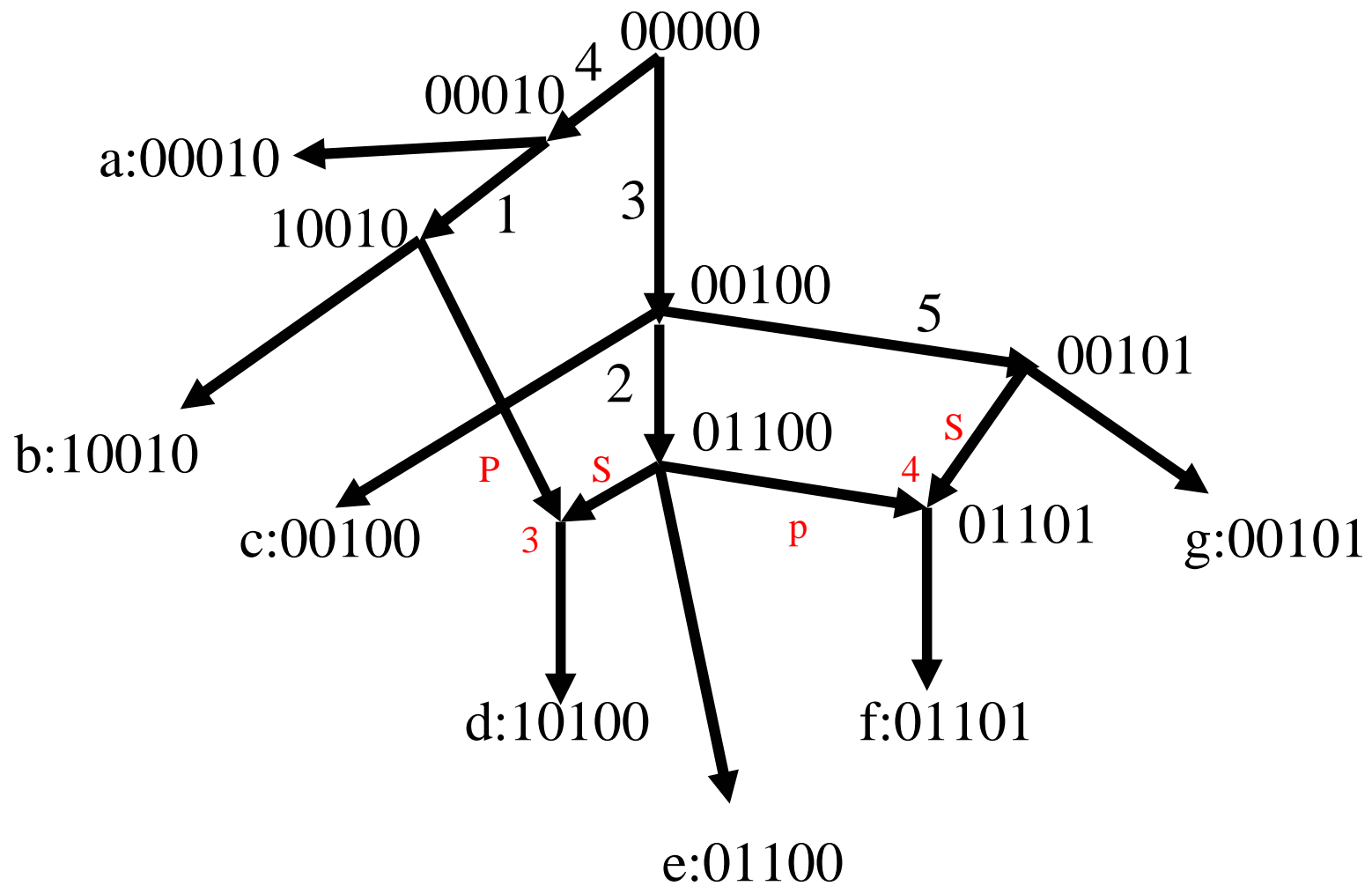
Simple Fact

If sites two sites i and $j > i$ conflict, then the sites must be **together** on some recombination cycle whose recombination point is between the two sites i and $j > i$.

(This is a general fact for all phylogenetic networks.)

Ex: In the prior example, site 1 conflicts with 3 and 4; and site 2 conflicts with 5.

A Phylogenetic Network



Simple Consequence of simple fact

All sites on the same (non-trivial) connected component of the conflict graph

must be on the **same gall** in any **galled-tree**.

Follows by transitivity and the fact that galls are node-disjoint recombination cycles.

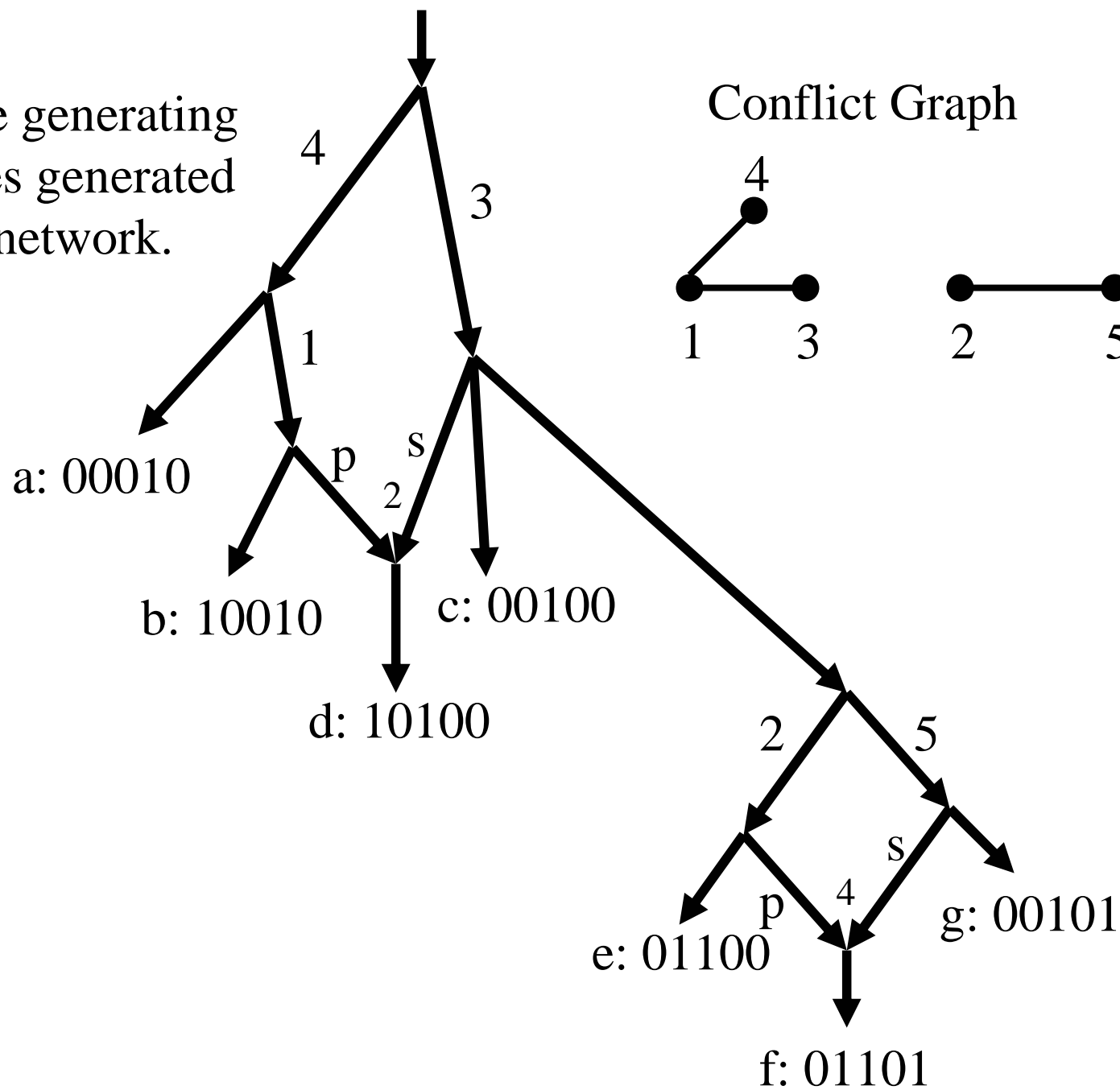
Key Result: For galls, the
converse consequence is also
true.

Two sites that are in **different** (non-trivial) connected components **cannot** be placed on the same **gall** in any phylogenetic network for M .

Hence, in any galled-tree T for M there is a **one-one** correspondence between the (non-trivial) connected components of the conflict graph for M and the galls of T .

These are the most important structural and algorithmic results about galls and galled-trees.

A galled-tree generating the sequences generated by the prior network.



Use of Key Result

- To build a galled-tree for M , if possible, focus on each connected component of the conflict graph separately.
- Determine how to arrange the sites on each gall, and then connect the galls.
- Add in any unconflicted sites, and any additional needed tree branches.

Canonical Galled-Trees

- A galled-tree is called **canonical** if every gall only contains conflicted sites.
- Theorem: If M can be derived on a galled-tree, it can be derived on a canonical galled-tree.
- The number of recombination nodes in a canonical galled-tree equals the number of connected components, which is the minimum number of recombinations possible in any galled-tree.

How to arrange the sites on a gall

Given a single connected component of the conflict graph with k sites, how do we arrange those k sites on a single gall, to generate the required sequences?

Arranging the sites

We will describe an $O(n^3)$ time method to arrange all of the galls. $O(n^2)$ time is possible with a more complex method.

A needed fact in words

Let Q be a gall for the sites on connected-component C of the conflict graph.

Let $M[C]$ be the matrix M **restricted** to the sites on C .

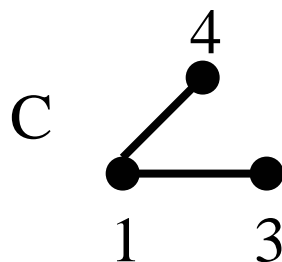
Let $LQ[C]$ be the sequences labeling the nodes of Q , **restricted** to the sites on C .

Claim: The two sets of sequences are identical, i.e.,

$$M[C] = LQ[C].$$

	1	2	3	4	5
a	0	0	0	1	0
b	1	0	0	1	0
c	0	0	1	0	0
d	1	0	1	0	0
e	0	1	1	0	0
f	0	1	1	0	1
g	0	0	1	0	1

M



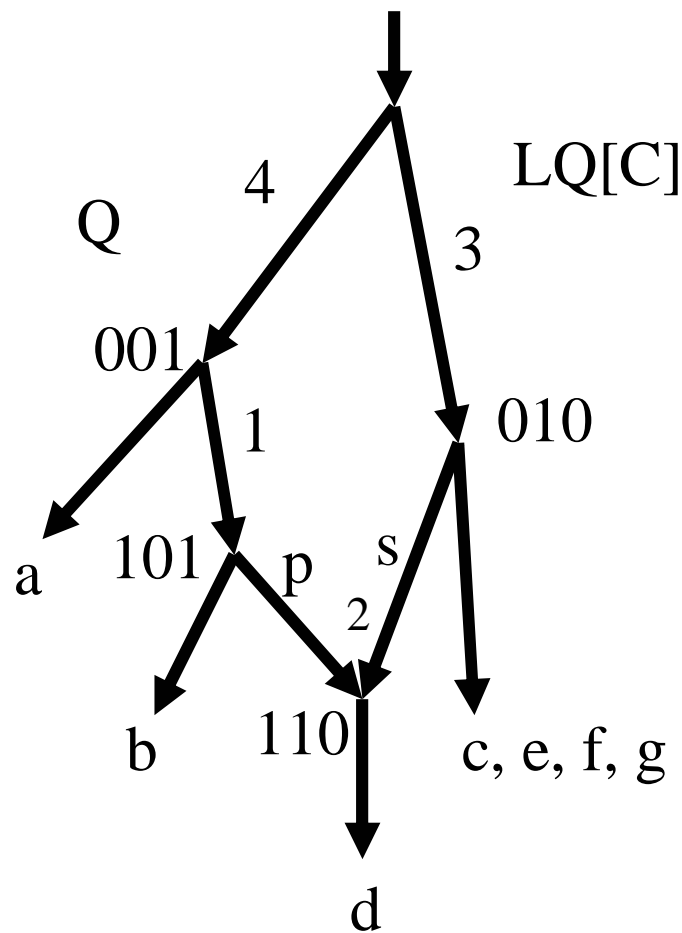
C

LQ[C] are the node labels on Q restricted to the sites in C

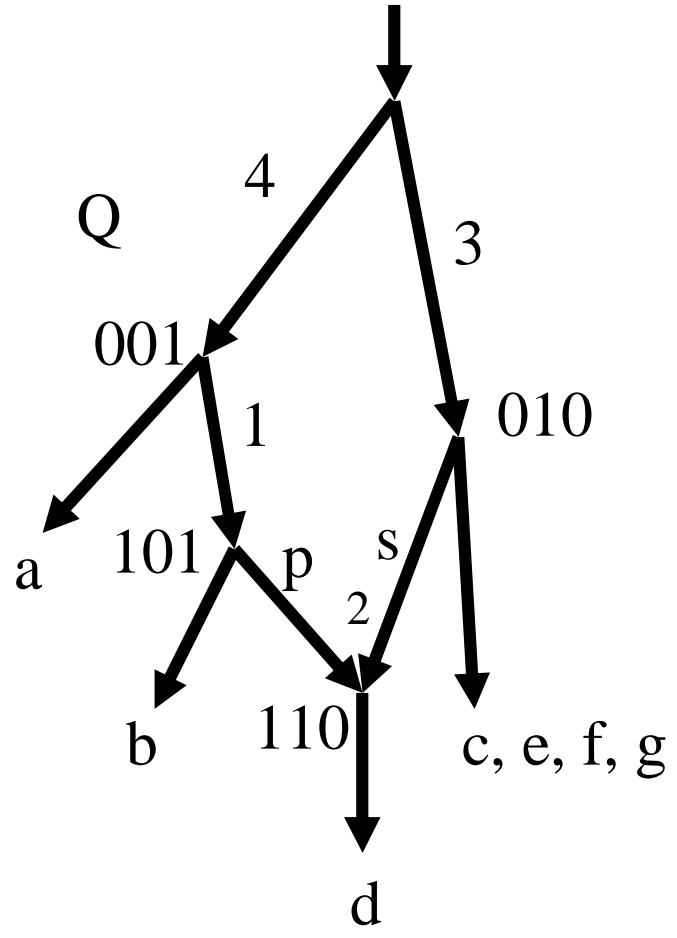
	1	3	4
a	0	0	1
b	1	0	1
c	0	1	0
d	1	1	0
e	0	1	0
f	0	1	0
g	0	1	0

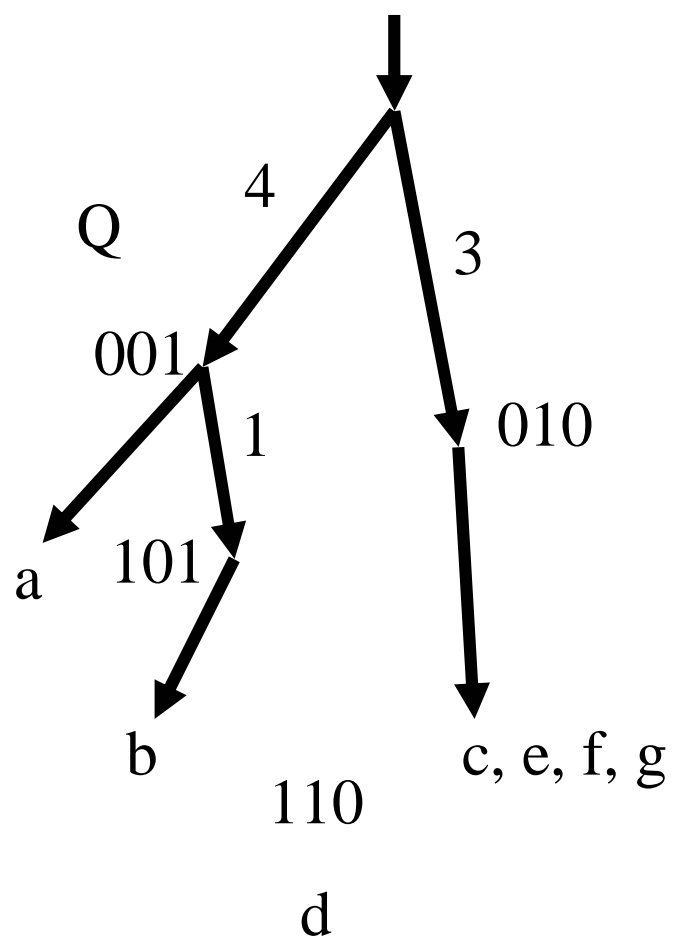
Matrix M[C] is
Matrix M restricted
to the columns in C.

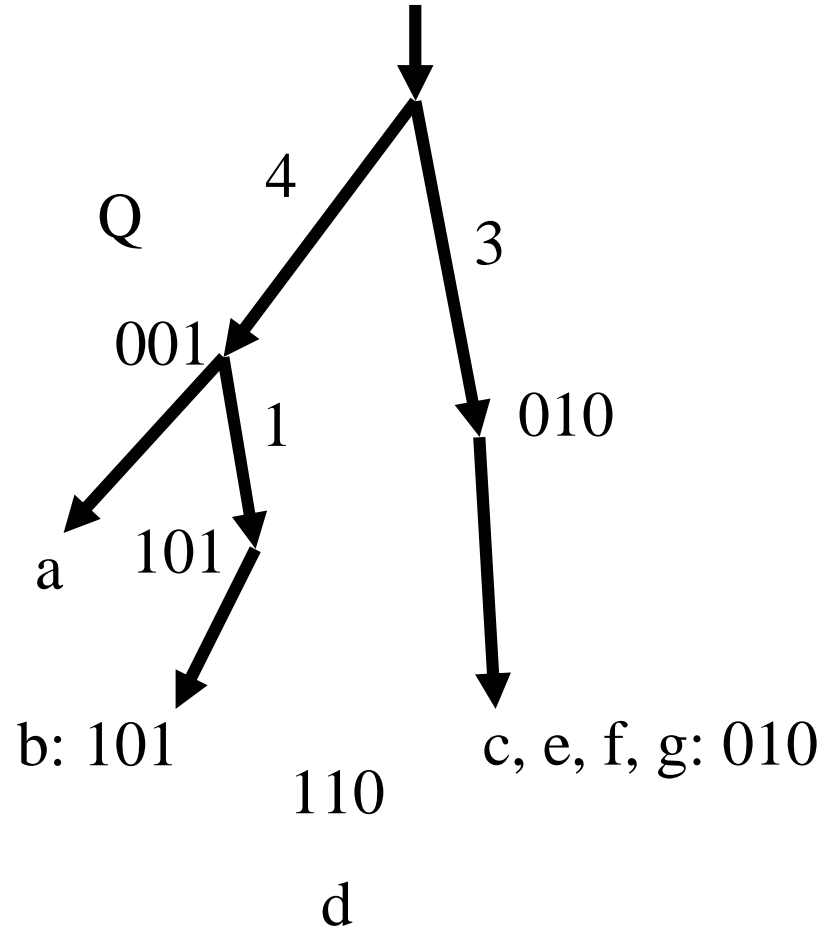
Fact: $M[C] = LQ[C]$



The idea for arranging the sites of
C on Q: via a short movie







Gall Q minus the recombination node is a perfect phylogeny for M[C] minus the recombinant sequence; all sites are on one or two paths from the root; and the two end sequences of those paths can recombine at point r to recreate the recombinant sequence.

The point

If we remove the recombinant node from Q , we have a phylogenetic tree (no cycles) for the remaining sequences in $LQ[C]$ and hence a perfect phylogenetic tree for the sequences in $M[C]$ minus the recombinant sequence of $LQ[C]$.

The sites in this tree are on one or two paths.

Moreover, the two end sequences on that perfect phylogeny can recombine to create the removed recombinant sequence.

The algorithm for arranging a gall Q for C , given r

1. Form the matrix $M[C]$.

2. For each row of $M[C]$, remove the row, see if there is a perfect phylogeny for the remaining rows.

If yes, see if the sites are in one or two paths, and the end sequences can generate the removed row by a recombination at r .

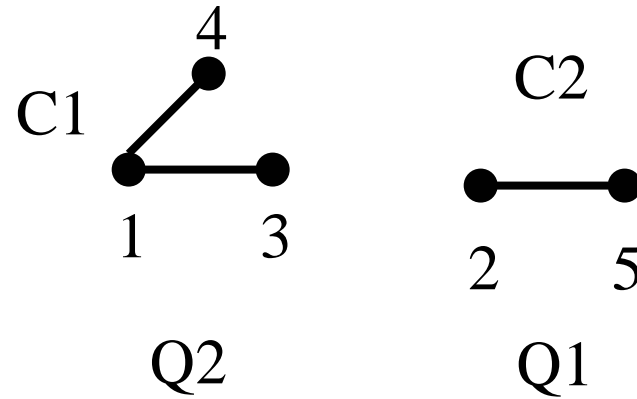
Fact: Every row that works gives a permitted arrangement of the sites on Ω .

How to connect the galls

Let C be a non-trivial connected component of the conflict graph. Let T be a galled-tree for the input M , and Q be the gall for C in T .

Idea: Any row j in $M[C]$ has a sequence that is **not** all-zero, if and only if the path to leaf j in T passes through gall Q .

		1	2	3	4	5
M	a	0	0	0	1	0
	b	1	0	0	1	0
	c	0	0	1	0	0
	d	1	0	1	0	0
	e	0	1	1	0	0
	f	0	1	1	0	1
	g	0	0	1	0	1



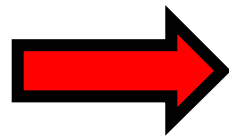
	1	3	4		2	5
a	0	0	1	a	0	0
b	1	0	1	b	0	0
c	0	1	0	c	0	0
d	1	1	0	d	0	0
e	0	1	0	e	1	0
f	0	1	0	f	1	1
g	0	1	0	g	0	1
M[C1]				M[C2]		

So the paths to every leaf pass through the gall Q1, but only the paths to e, f, g pass through gall Q2.

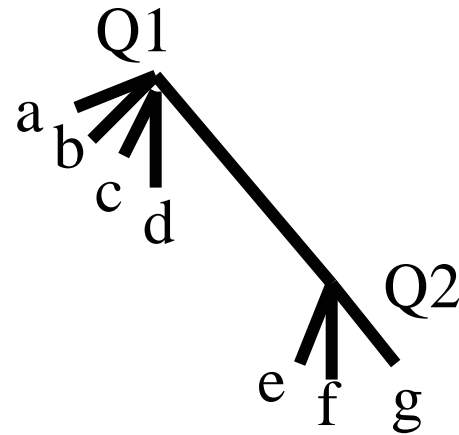
The “pass-through” information determines a perfect phylogeny of galls

	Q1	Q2
a	1	0
b	1	0
c	1	0
d	1	0
e	1	1
f	1	1
g	1	1

Pass-through matrix.



Apply a perfect phylogeny algorithm to the “pass-through” matrix.



Consequence

Every galled-tree for M has the same perfect phylogeny derived from the pass-through information. So the “pass-through” perfect phylogeny is **invariant** over all the galled-trees for M .

How to connect the galls - fine structure

If the path to j goes through Q , it enters at the top and exits Q at the node whose $LQ[C]$ label equals the row j sequence in $M[C]$.

Hence the only variation in the galled-trees for M is how the sites on each gall are arranged. That can be done in at most three ways per gall, and typically only one way.

Optimality

Theorem: A canonical galled-tree for M minimizes the number of recombinations over all phylogenetic networks that derive M .

The proof is not in the proceedings, where this issue was given as an open problem. The proof will appear in the journal version of the paper.

More Optimality

If M can be derived on a galled-tree, then a canonical galled-tree minimizes the number of “recombination events” over all possible phylogenetic networks for M , where a recombination event allows **any number of crossovers** between the strings, rather than just one.

More results

- There is a galled tree for the data M only if each connected component of the conflict graph is bi-convex, bipartite and all the nodes on one side have smaller index than the nodes on the other side.
- If there is a galled-tree for M , then the problem of finding the largest subset of columns that has a perfect phylogeny can be solved in $O(nm)$ time. (NP-hard in general)
- If there is a galled-tree for M then there is a tree generating M with at most one back mutation per site.

Finally

The approach of studying constrained or structured recombination in phylogenetic networks by looking for structure in the conflict graph opens a large area of exploration for graph enthusiasts. We are presently using this approach to study networks more complex than galled-trees.

For example, we can prove that the number of non-trivial connected components in the conflict graph is a lower bound on the number of needed recombination-events in any phylogenetic network for M .

Nasty Typo Alert

Lemma 2.1 (page 4) in the proceedings paper omitted the key condition:

“Site i appears (mutates) on gall Q .”