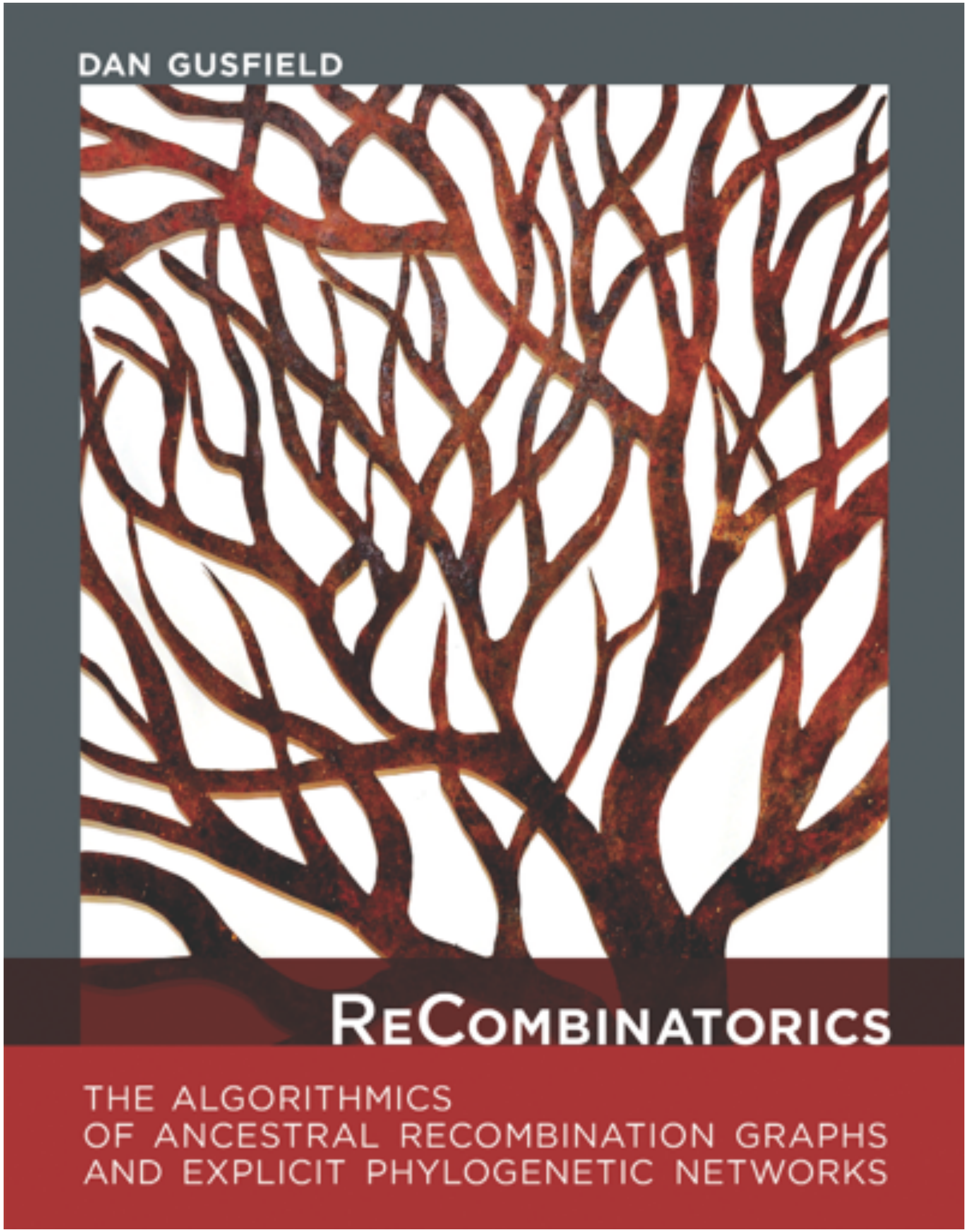
ReCombinatorics

The Algorithmics and Combinatorics of Phylogenetic Networks with Recombination

Dan Gusfield

NCBS CS and BIO Meeting

December 19, 2016



DAN GUSFIELD

RECOMBINATORICS

THE ALGORITHMICS
OF ANCESTRAL RECOMBINATION GRAPHS
AND EXPLICIT PHYLOGENETIC NETWORKS

SNP Data

- A SNP is a Single Nucleotide Polymorphism - a site in the genome where two different nucleotides appear with sufficient frequency in the population (say each with 5% frequency or more).
- SNP maps have been compiled with a density of about 1 site per 1000.
- SNP data is what is mostly collected in populations - it is much cheaper to collect than full sequence data, and focuses on variation in the population, which is what is of interest.

Geneological or Phylogenetic Networks

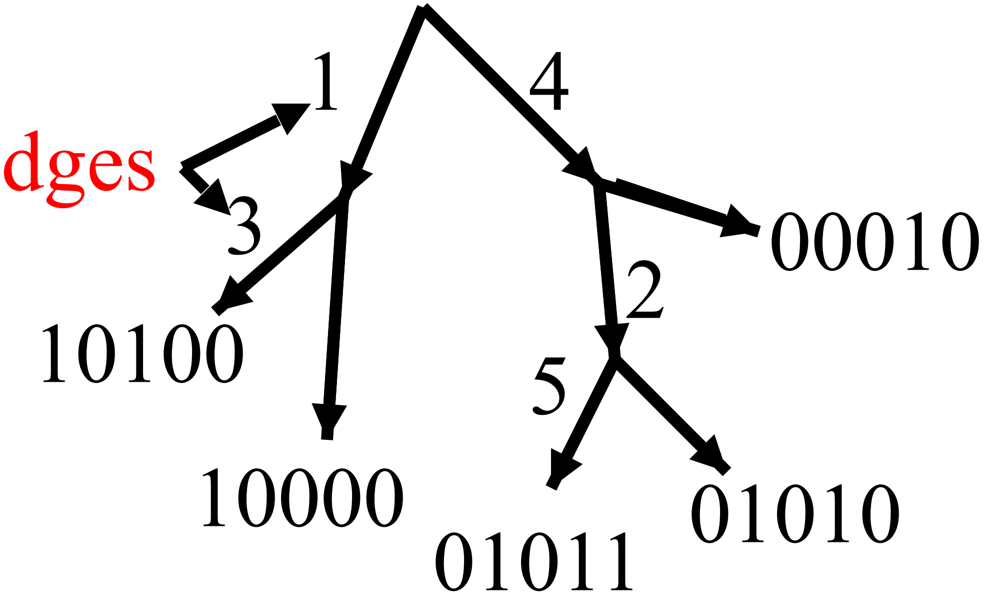
- The major biological motivation comes from genetics and attempts to reconstruct the history of recombination in populations.
- Also relates to phylogenetic-based haplotyping.
- Some of the algorithmic and mathematical results also have phylogenetic applications, for example in hybrid speciation, lateral gene transfer.

The Perfect Phylogeny Model
for binary sequences

Only one mutation per site allowed (infinite sites)

sites 12345

Ancestral sequence 00000



Site mutations on edges

1

4

3

2

5

10100

10000

01011

01010

00010

The tree derives the set M:

10100

10000

01011

01010

00010

Extant sequences at the leaves

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. That is the infinite sites model from population genetics.

$$M \quad \begin{array}{cc} & m \\ n & \boxed{\begin{array}{l} 01101001 \\ 11100101 \\ 10101011 \end{array}} \end{array}$$

When can a set of sequences be derived on a perfect phylogeny?

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 four pairs:

0,0 and 0,1 and 1,0 and 1,1

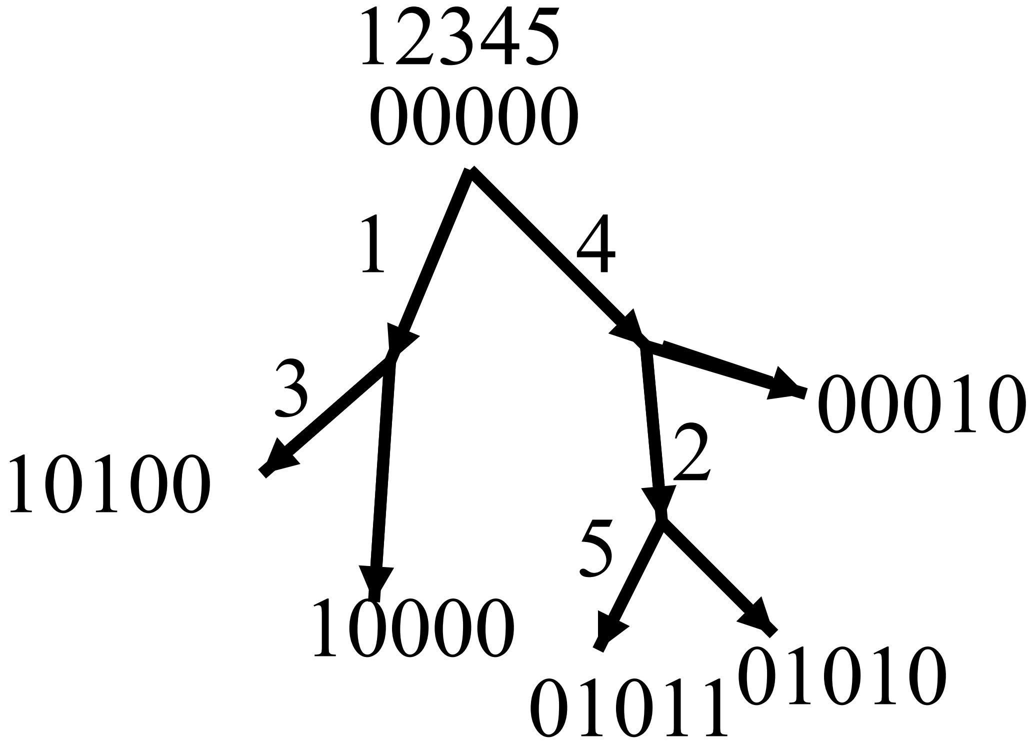
This is the 4-Gamete Test

So, in the case of binary characters, if each pair of columns allows a tree, then the entire set of columns allows a tree.

For M of dimension n by m, the existence of a perfect phylogeny for M can be tested in O(nm) time and a
tree built in that time, if there is one. Gusfield, Networks 91

A richer model

10100
10000
M 01011
01010
00010
10101 added

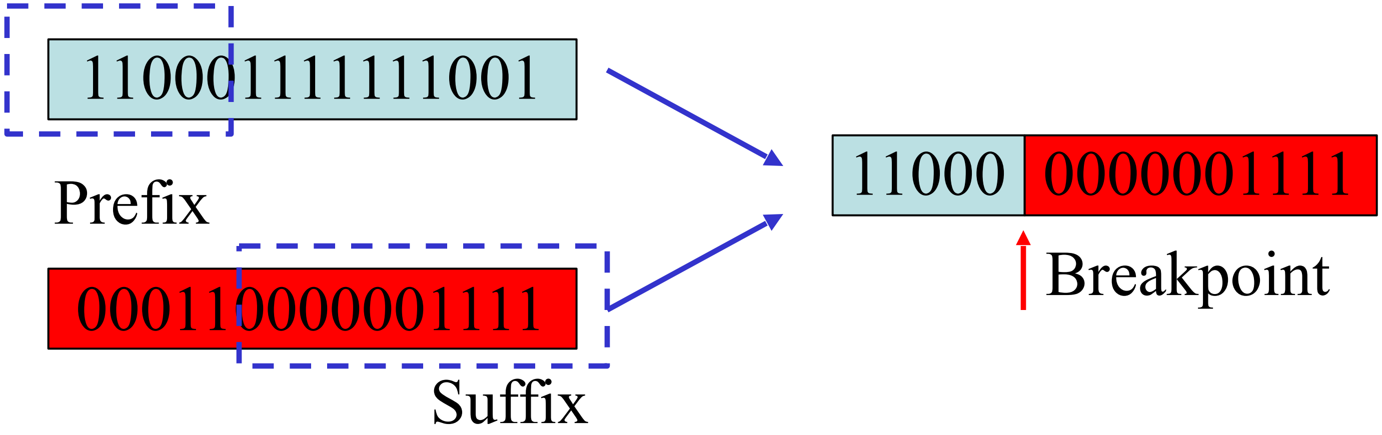


Pair 4, 5 fails the four gamete-test. The sites 4, 5 are incompatible.

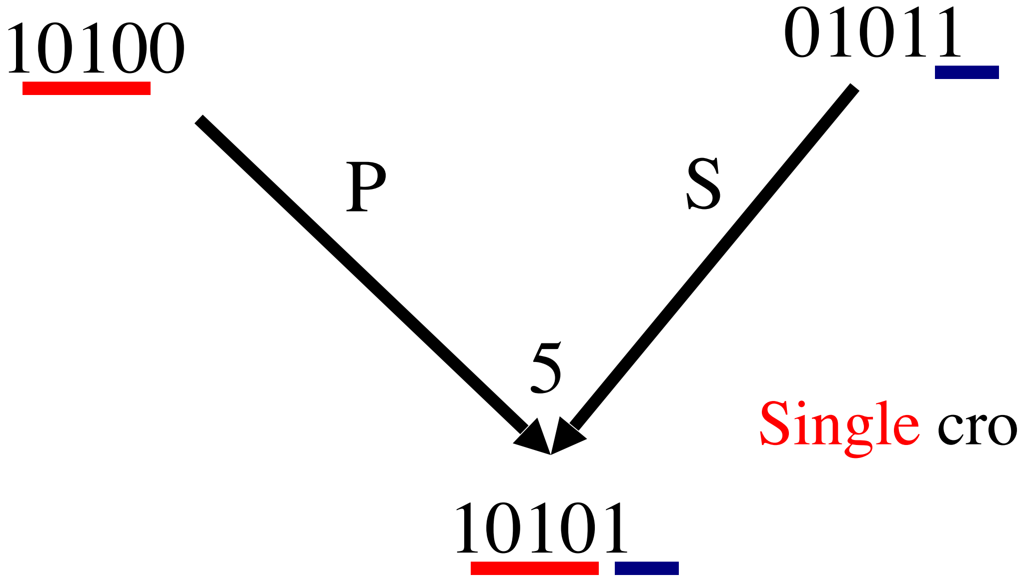
Real sequence histories often involve recombination.

Recombination

- Recombination: one of the principle genetic forces shaping sequence variations within species
- Two equal length sequences generate a third **new** equal length sequence during meiosis.



Sequence Recombination



Single crossover recombination

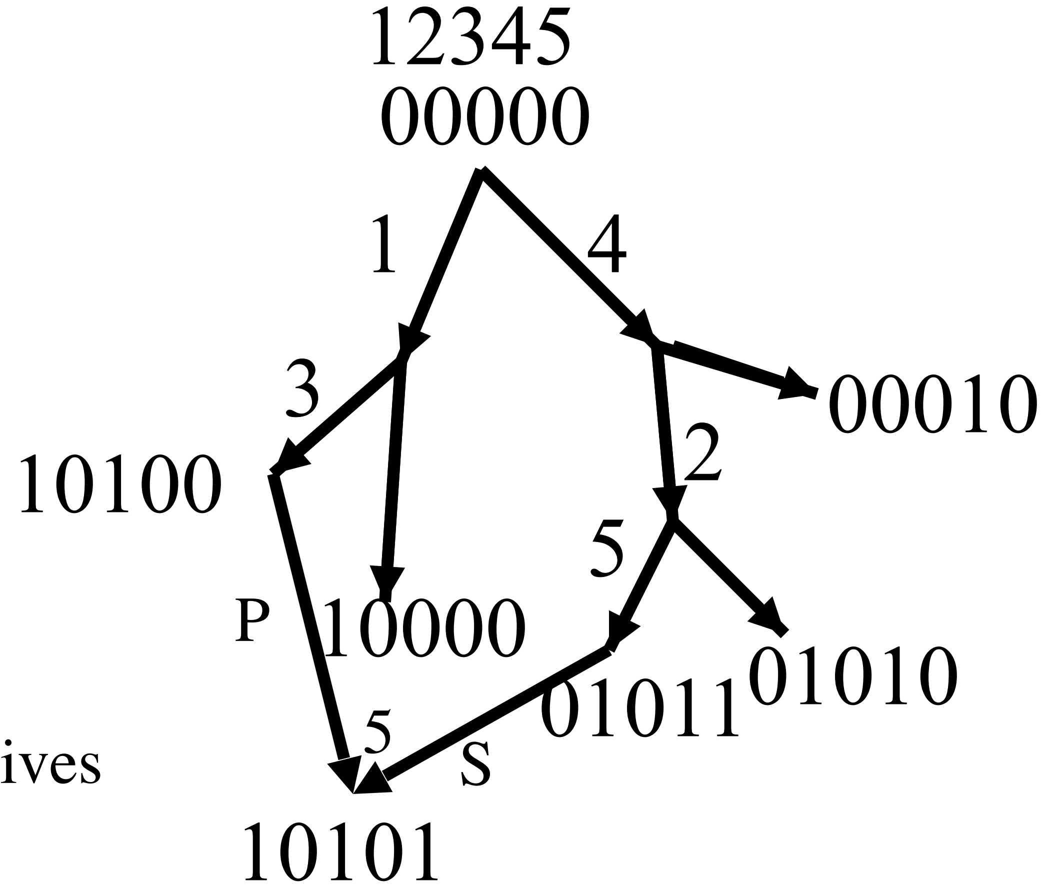
A recombination of P and S at recombination point 5.

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

Network with Recombination: ARG

M

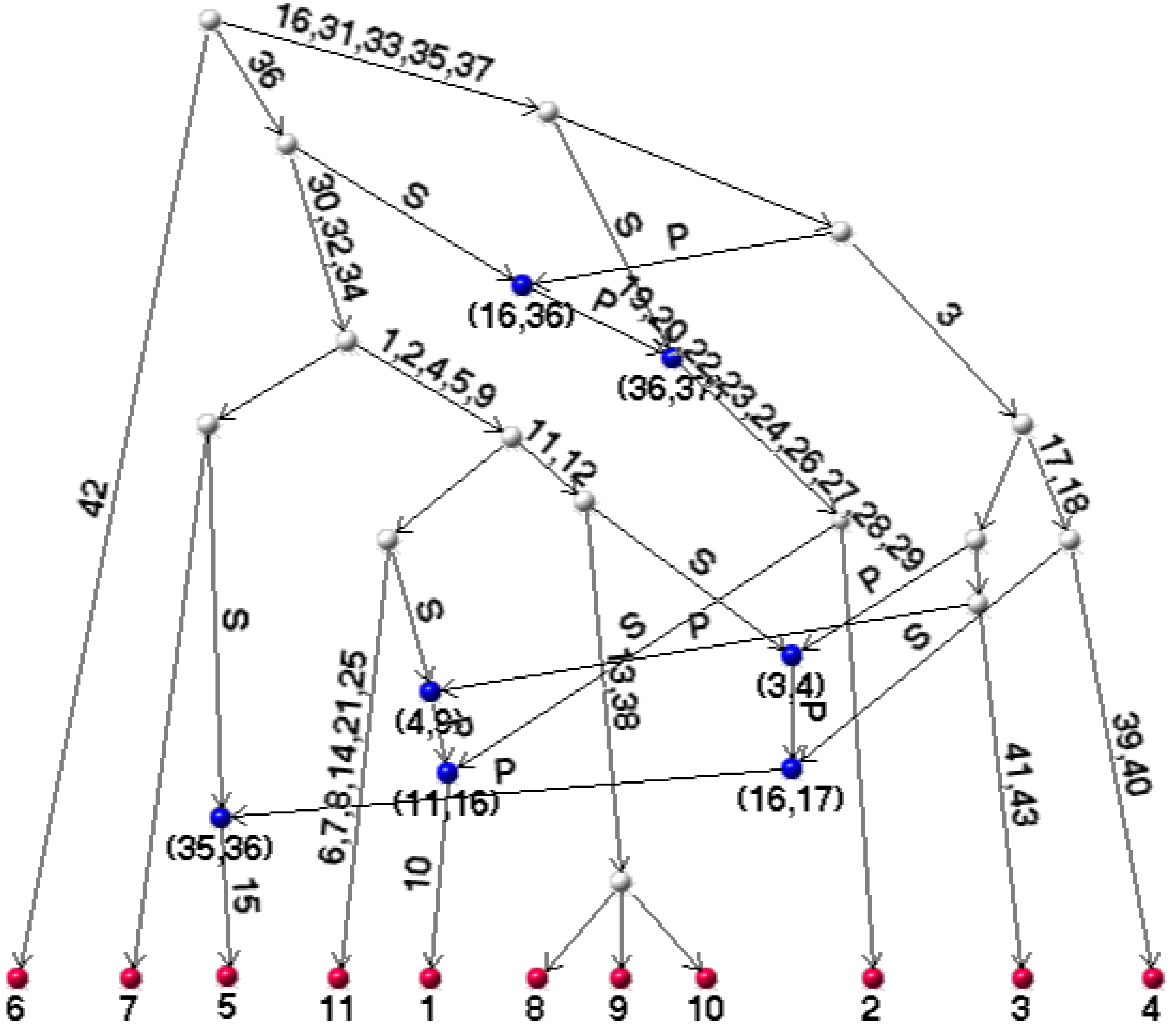
10100
10000
01011
01010
00010
10101 new



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

A Min ARG for Kreitman’s data

ARG created by
SHRUB



Results on Reconstructing the Evolution of SNP Sequences

- Part I: Clean mathematical and algorithmic results: Galled-Trees, near-uniqueness, graph-theory lower bound, the Decomposition theorem, sufficient conditions for fully-decomposed MinARGs

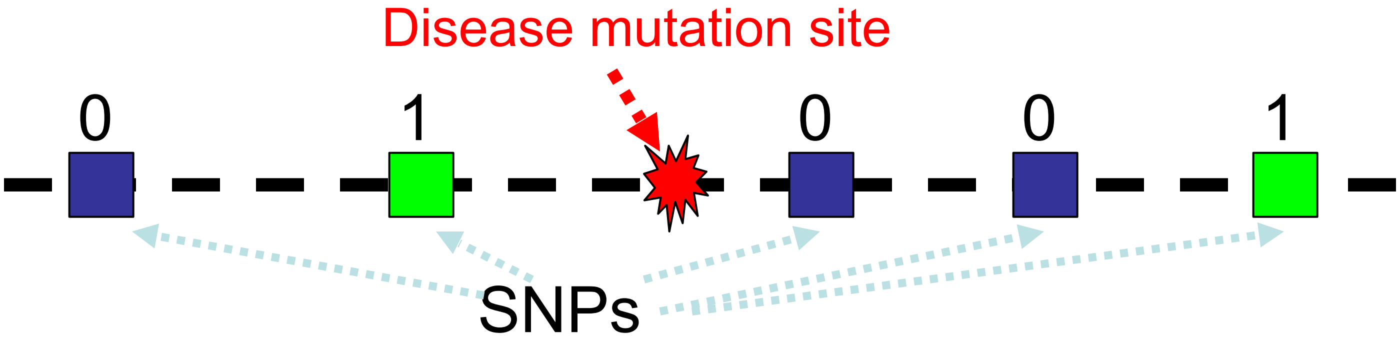
- Part II: Practical computation of Lower and Upper bounds on the number of recombinations needed. Construction of (optimal) phylogenetic networks; uniform sampling; haplotyping with ARGs; LD mapping …
- Part III: Varied Biological Applications
- Part IV: Extension to Gene Conversion
- Part V: The Minimum Mosaic Model of Recombination

An illustration of why we are interested in recombination and ARGs

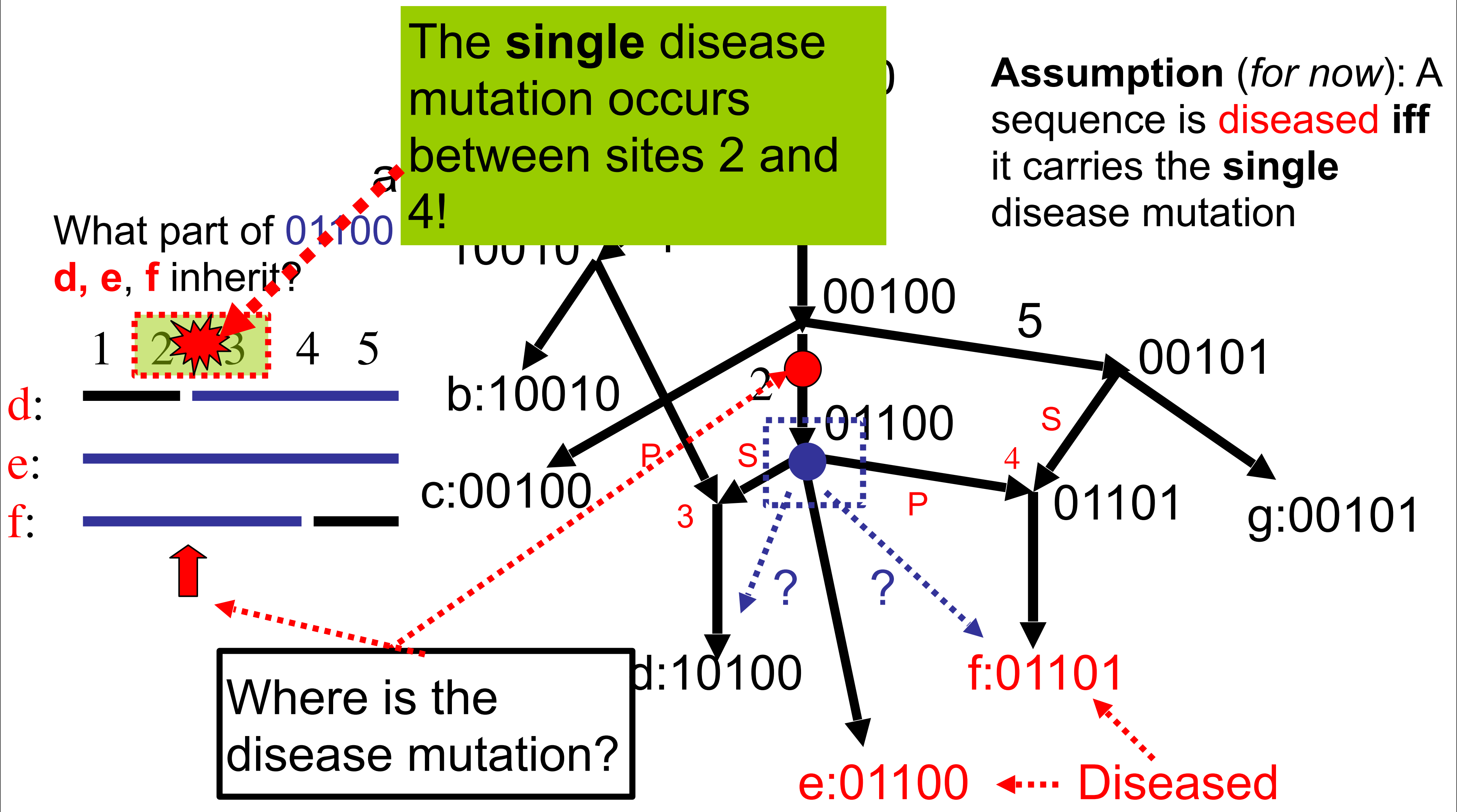
Association Mapping of Using **ARGs**

Association Mapping

- A major strategy being practiced to find genes influencing disease from haplotypes of a **subset** of SNPs.
  - Disease mutations: **unobserved.**
- A simple example to explain association mapping and why ARGs are useful, assuming the true ARG is known.



Very Simplistic Mapping the **Unobserved** Mutation of **Mendelian** Diseases with ARGs



The **single** disease mutation occurs between sites 2 and 4!

Assumption (*for now*): A sequence is diseased **iff** it carries the **single** disease mutation

What part of 01100 **d, e, f** inherit?

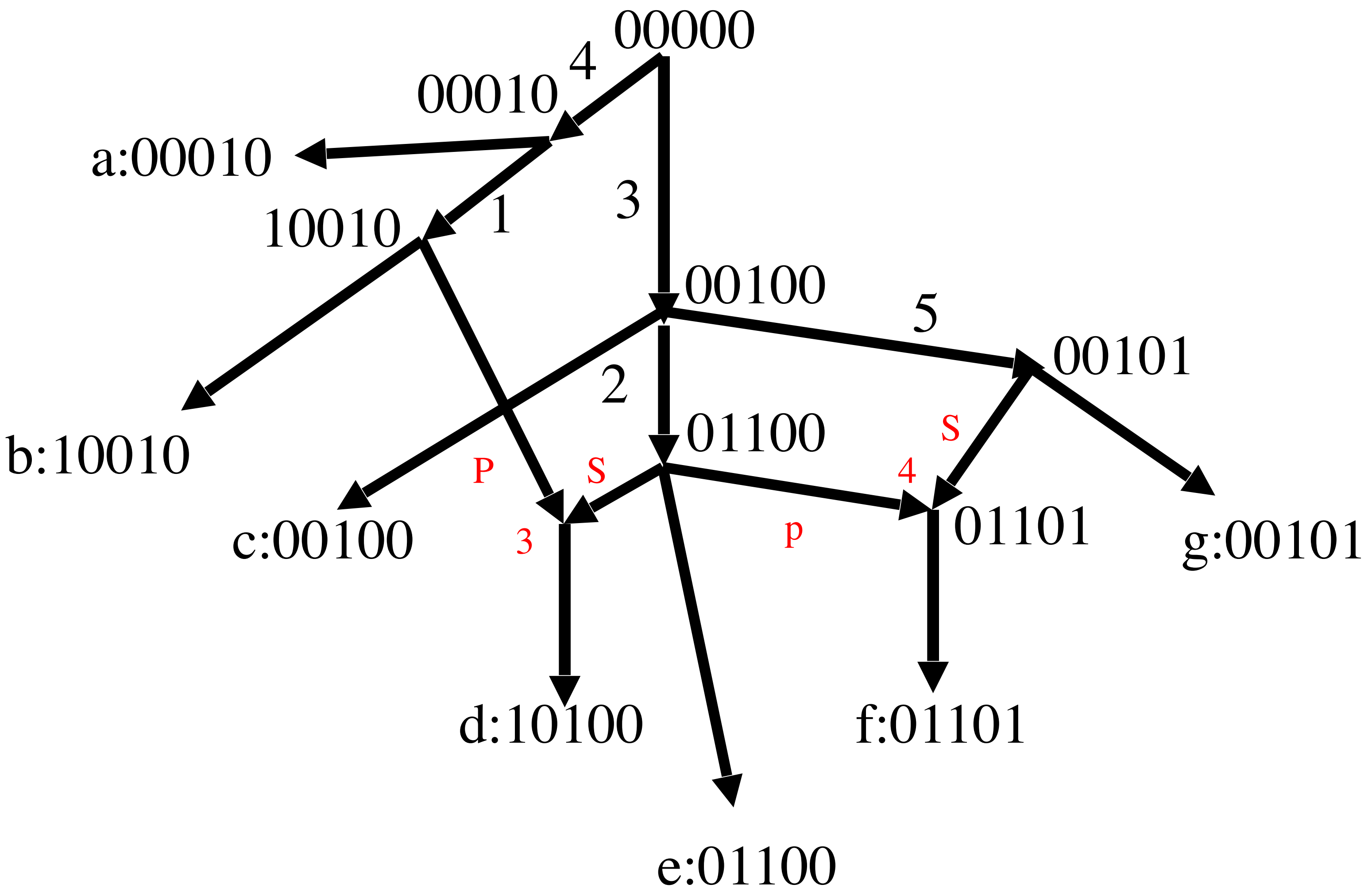
Where is the disease mutation?

Mapping Disease Gene with Inferred ARGs

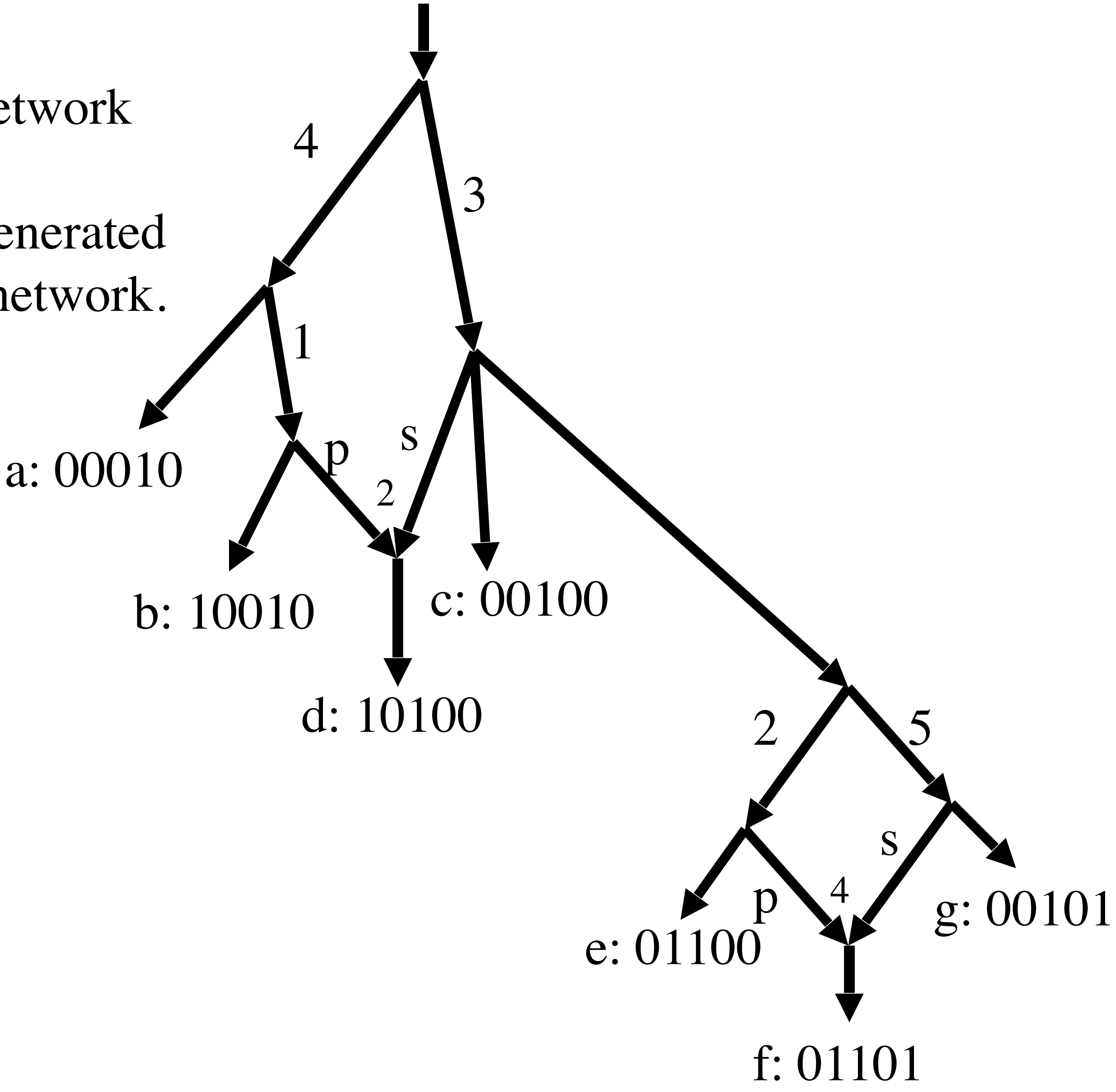
- "..*the best information that we could possibly get about association is to know the full coalescent* ***genealogy****…*" – Zollner and Pritchard, 2005
- But we do not know the true ARG!
- Goal: **infer** ARGs from SNP data for association mapping
  - Not easy and often approximation (e.g. Zollner and Pritchard)
  - Improved results to do the inference Y. Wu (RECOMB 2007)

Galled Trees

A Phylogenetic Network



A tree-like network for the same sequences generated by the prior network.

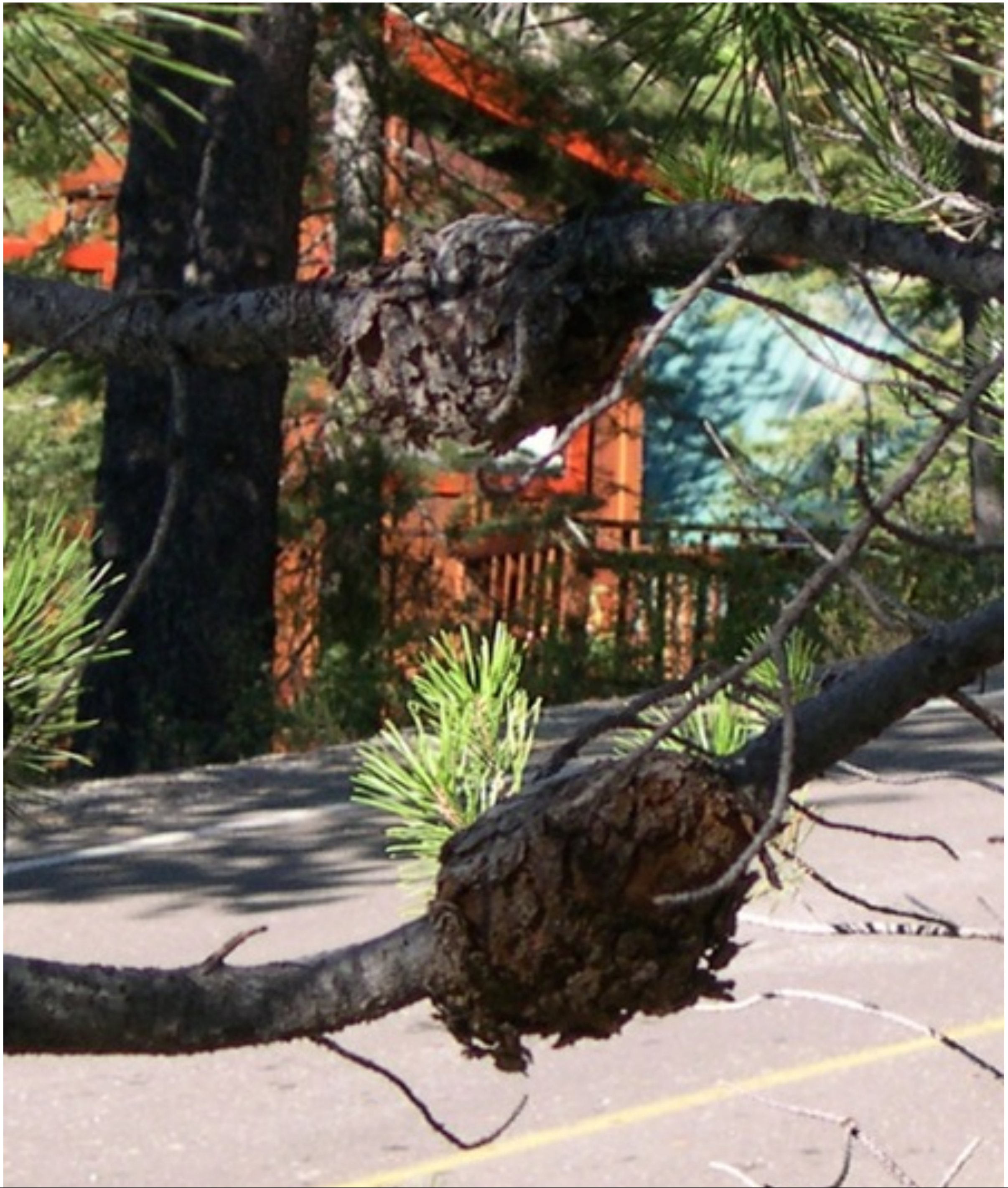


Recombination Cycles

- 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 phylogenetic network where no recombination cycles share an edge is called a galled tree.
- A cycle in a galled-tree is called a gall.
- Question: if M cannot be generated on a true tree, can it be generated on a galled-tree?



Sales pitch for Galled-Trees

Galled-trees represent a small deviation from true trees.

There are sufficient applications where it is plausible that a galled tree exists that generates the sequences.

Observable recombinations tend to be recent, and there are blocks in the human genome where recombinations are sparse.

The number of recombinations is never more than m/2. Moreover, when M can be derived on a galled-tree, the number of recombinations used is the minimum number over any phylogenetic network, even if multiple cross-overs at a recombination event are counted as a single recombination.

A galled-tree for M is ``almost unique” - implications for reconstructing the correct history.

Results about galled-trees

- Theorem: Efficient (provably polynomial-time) algorithm to determine whether or not any sequence set M can be derived on a galled-tree.
- Theorem: A galled-tree (if one exists) produced by the algorithm minimizes the number of recombinations used over all possible phylogenetic-networks.
- Theorem: If M can be derived on a galled tree, then the Galled-Tree is ``nearly unique”. This is important for biological conclusions derived from the galled-tree.

Papers from 2003-2007.

Elaboration on Near Uniqueness

Theorem: The number of arrangements (permutations) of the sites on any gall is
at most three, and this happens only if the gall has two sites.

If the gall has more than two sites, then the number of arrangements is at most two.

If the gall has four or more sites, with at least two sites on each side of the recombination point (not the side of the gall) then the arrangement is forced and unique.

Theorem: All other features of the galled-trees for M are invariant.

A whiff of the ideas behind the galled-tree results.

We take a more general view

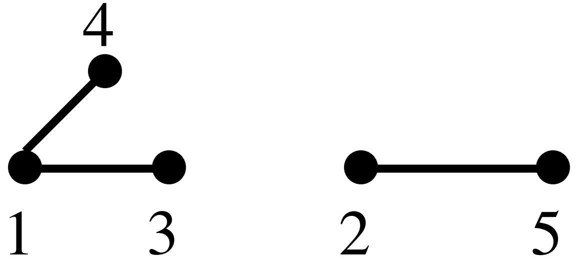
Incompatible Sites

A pair of sites (columns) of M that fail the 4-gametes test are said to be incompatible.

A site that is not in such a pair is compatible.

$$M \qquad \begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ a & 0 & 0 & 0 & 1 & 0 \\ b & 1 & 0 & 0 & 1 & 0 \\ c & 0 & 0 & 1 & 0 & 0 \\ d & 1 & 0 & 1 & 0 & 0 \\ e & 0 & 1 & 1 & 0 & 0 \\ f & 0 & 1 & 1 & 0 & 1 \\ g & 0 & 0 & 1 & 0 & 1 \end{array}$$

Incompatibility Graph G(M)

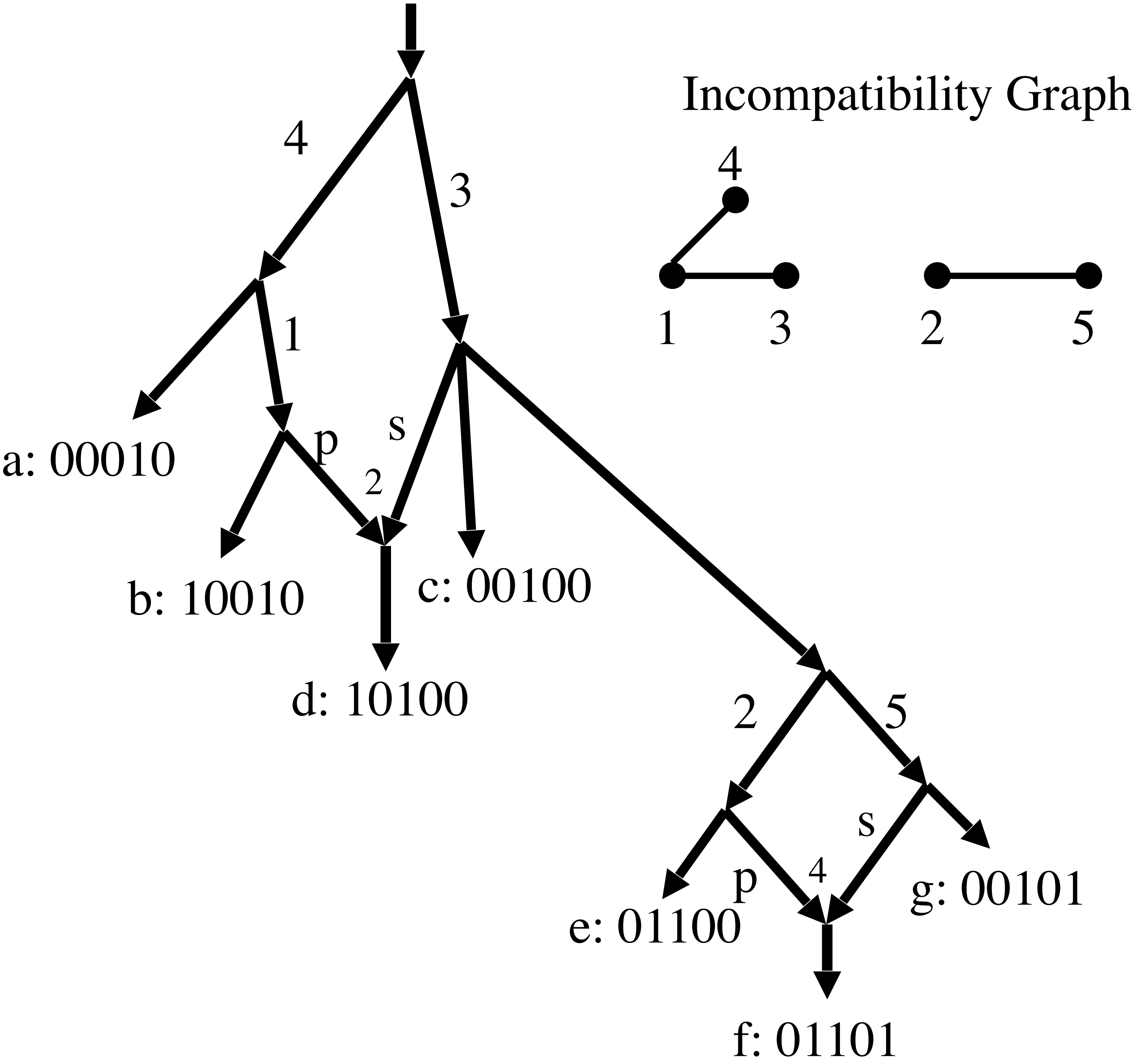


Two nodes are connected iff the pair of sites are incompatible, i.e, fail the 4-gamete test.

THE MAIN TOOL: We represent the pairwise incompatibilities in a incompatibility graph.

The connected components of G(M) are very informative

- Theorem: The number of non-trivial connected components is a lower-bound on the number of recombinations needed in any network.
- Theorem: When M can be derived on a galled-tree, all the incompatible sites in a gall must come from a single connected component C, and that gall must contain all the sites from C.
- So, in a galled-tree the number of recombinations is exactly the number of connected components in G(M), and hence is minimum over all possible phylogenetic networks for M.


Incompatibility Graph
4
1
3
2
5
4
3
1
a: 00010
p
s
2
b: 10010
c: 00100
d: 10100
2
5
s
p
4
e: 01100
g: 00101
f: 01101

Constructing Optimal Phylogenetic Networks in General

Optimal = minimum number of recombinations. Called Min ARG.

The method is based on the coalescent viewpoint of sequence evolution. We build the network backwards in time.

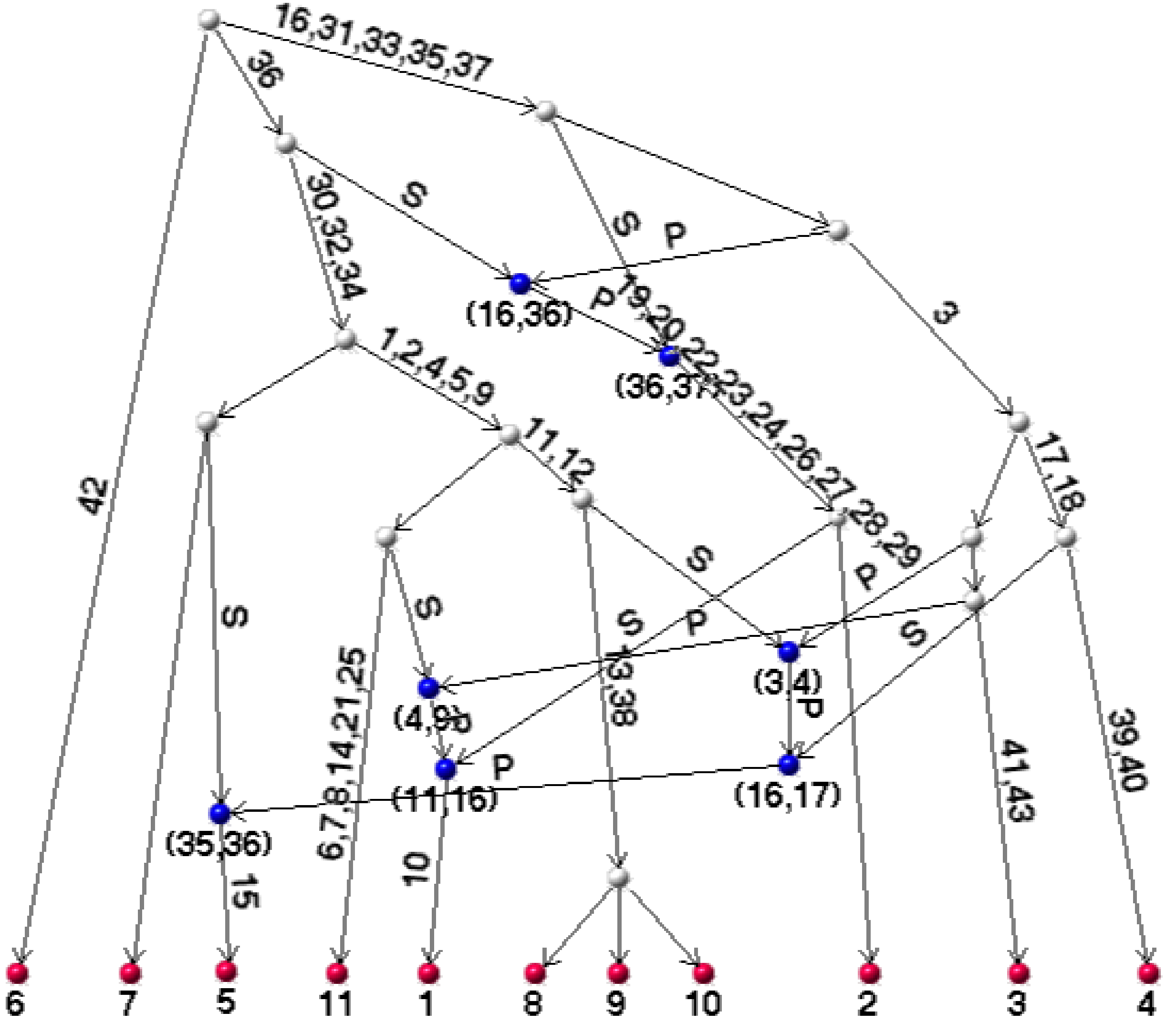
Kreitman's 1983 ADH Data

- 11 sequences, 43 segregating sites
- Both HapBound and SHRUB took only a fraction of a second to analyze this data.
- Both produced 7 for the number of detected recombination events

Therefore, independently of all other methods, our lower and upper bound methods together imply that 7 is the minimum number of recombination events.

A Min ARG for Kreitman’s data

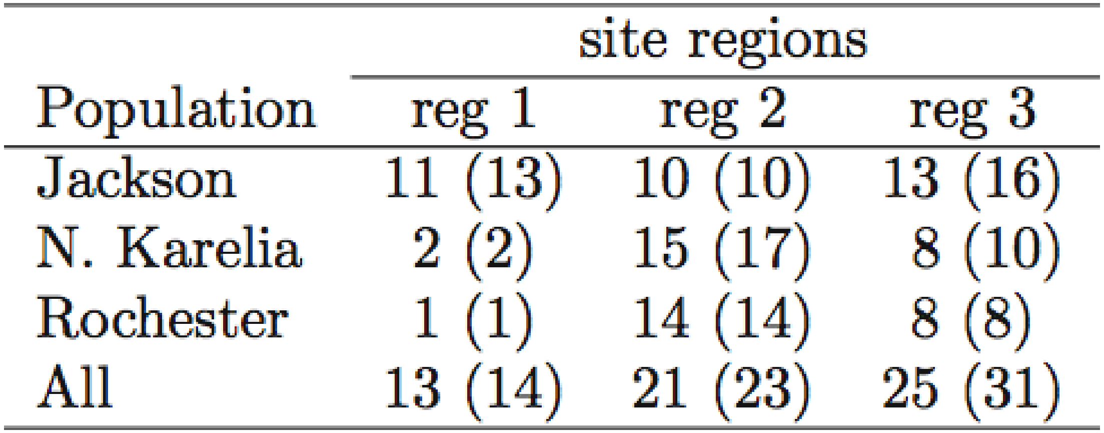
ARG created by
SHRUB



The Human LPL Data (Nickerson et al. 1998)

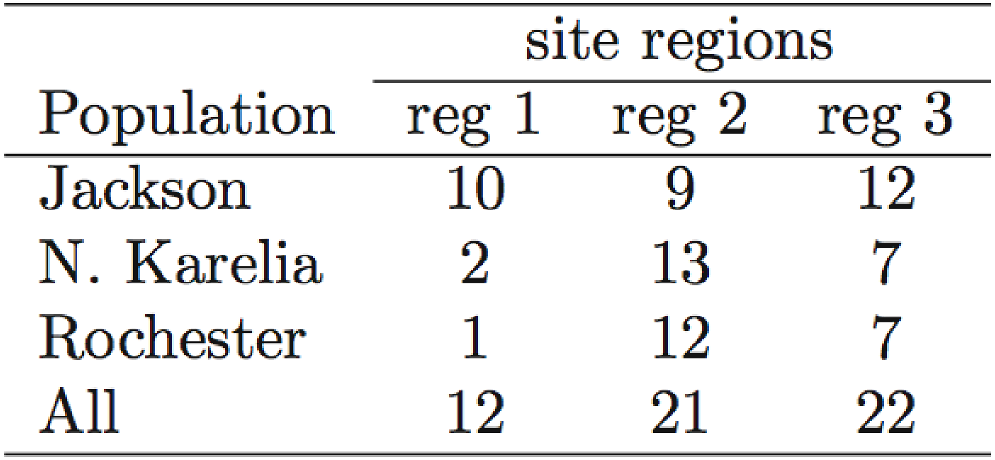
(88 Sequences, 88 sites)

Our new lower and upper bounds



	site regions		
Population	reg 1	reg 2	reg 3
Jackson	11 (13)	10 (10)	13 (16)
N. Karelia	2 (2)	15 (17)	8 (10)
Rochester	1 (1)	14 (14)	8 (8)
All	13 (14)	21 (23)	25 (31)

Optimal RecMin Bounds



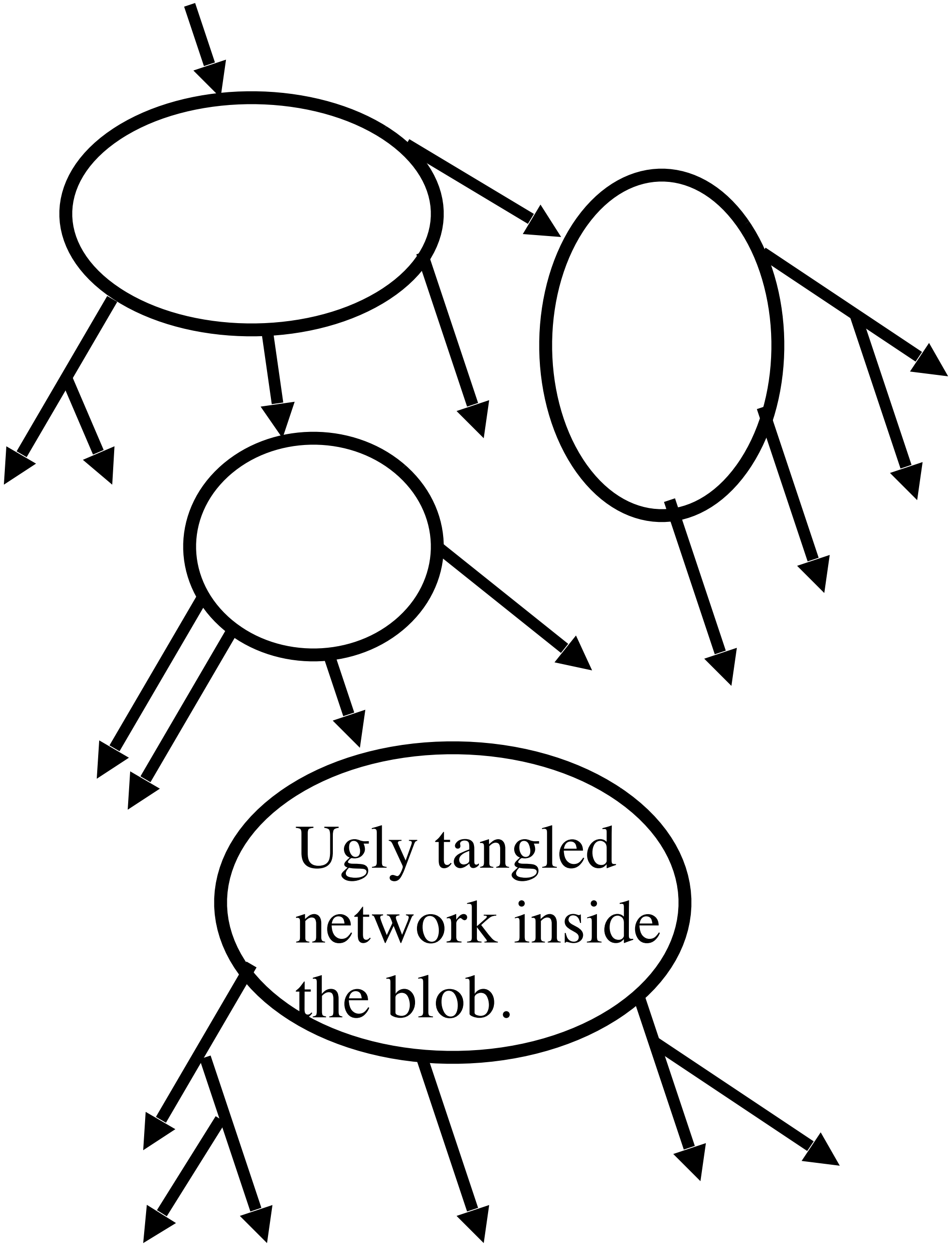
	site regions		
Population	reg 1	reg 2	reg 3
Jackson	10	9	12
N. Karelia	2	13	7
Rochester	1	12	7
All	12	21	22

(We ignored insertion/deletion, unphased sites, and sites with missing data.)

Blobbed-trees: generalizing galled-trees

- In a phylogenetic network a maximal set of intersecting cycles is called a blob.
- Contracting each blob results in a directed, rooted tree, otherwise one of the “blobs” was not maximal.
- So every phylogenetic network can be viewed as a directed tree of blobs - a blobbed-tree.

The blobs are the non-tree-like parts of the network.



Ugly tangled network inside the blob.

Every network is a tree of blobs.
How do the tree parts and the blobs relate?

How can we exploit this relationship?

Simple Fact

If sites two sites i and j are incompatible, then the sites must be together on some recombination cycle whose recombination point is between the two sites i and j.

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

Ex: In the prior example, sites 1, 3 are incompatible, as are 1, 4; as are 2, 5.

Simple Consequence of the simple fact

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

must be on the same blob in any blobbed-tree.

Follows by transitivity.

So we can’t subdivide a blob into a tree-like structure if it only contains sites from a single connected component of the incompatibility graph.

Key Result about Galls: For galls, the converse of the simple 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 a galled-tree T for M each gall contains all and only the sites of one (non-trivial) connected component of the incompatibility graph. All compatible sites can be put on edges outside of the galls.

This is the key to the galled-tree solution.

Optimality of Galled-Trees

Theorem: (G,H,B,B) The minimum number of recombination nodes in any phylogenetic network for M is at least the number of non-trivial connected components of the incompatibility graph.

Hence, when there is a galled-treee for the data, the galled-tree minimizes the number of recombinations, over all possible ARGs for the data. Hence the galled-tree is a MinARG for the data.

So in this case, the NP-hard problem of finding a MinARG has a polynomial-time solution.
the root-unknown galled-tree problem in polynomial time.

To complete the discussion of how to build
a galled-tree, we first take a more general
view and discuss the Full-Decomposition Theorem.

The Decomposition Theorem (Recomb 2005)

For any set of sequences M, there is a blobbed-tree T(M) that derives M, where each blob contains all and only the sites in one non-trivial connected component of G(M). The compatible sites can always be put on edges outside of any blob.

A blobbed-tree with this structure is called fully-decomposed.

General Structure

So, for any set of sequences M, there is a phylogenetic network T(M) that is fully decomposed.

Moreover, the tree part T of T(M) is unique. And it is easy to find the tree part.

Moreover

Since all sites from a single connected component must be together on some blob in any phylogenetic network, no network is more decomposed than the fully decomposed network.

What is a galled tree?

We can now see that a galled tree is just a fully-decomposed ARG, where in the tree

of blobs, each blob is just a single cycle.

We will later see how to efficiently determine

if each blob can be realized by a single cycle.

For now we concentrate on building the tree part of the tree of blobs.

Back to Galled-Trees

So a galled-tree is a Fully-Decomposed ARG where each blob has only one recombination node, and all and only the sites from a single non-trivial connected component of G(M) appear together on a single blob.

The backbone tree T of the Fully-Decomposed ARG is easy to find, as we will show later.

How to layout a gall

A gall B is associated with a single connected-component in the incompatibility graph, say component C.

Let M[C] be the sequences in M restricted
to the site in C.

The sequences in M[C] must be generated inside gall B, because mutations on sites in C
only occur in gall B.

More Formally:

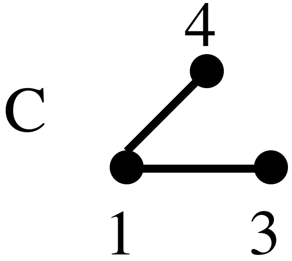
Let M[C] be the matrix M restricted to the sites in C. Let
S[C] be a sequence S restricted to the sites in C.

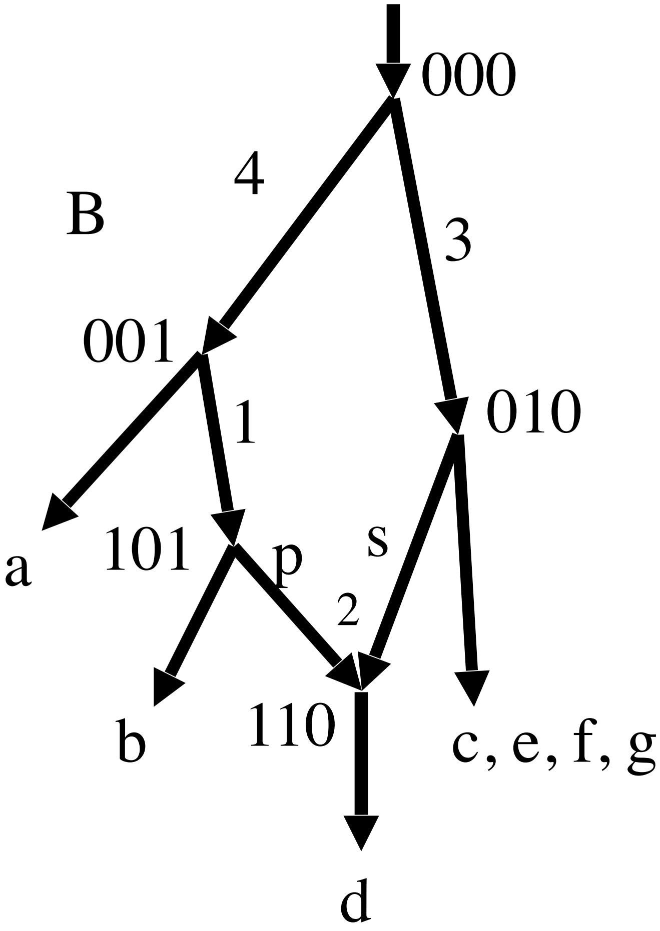
Key point:
Each distinct (non-zero) sequence in M[C] must be the sequence
S[C] for some sequence S labeling a branching node v on B.

So, although we don’t know much about the interior of B,
we know precisely the sequences
(restricted to sites of C) that
label the exterior nodes on B. And we know that
the states of the non-C sites are identical at each node in B.

M

$$\begin{array}{cccccc} & 1 & 2 & 3 & 4 & 5 \\ a & 0 & 0 & 0 & 1 & 0 \\ b & 1 & 0 & 0 & 1 & 0 \\ c & 0 & 0 & 1 & 0 & 0 \\ d & 1 & 0 & 1 & 0 & 0 \\ e & 0 & 1 & 1 & 0 & 0 \\ f & 0 & 1 & 1 & 0 & 1 \\ g & 0 & 0 & 1 & 0 & 1 \end{array}$$





$$\begin{array}{cccc} & 1 & 3 & 4 \\ a & 0 & 0 & 1 \\ b & 1 & 0 & 1 \\ c & 0 & 1 & 0 \\ d & 1 & 1 & 0 \\ e & 0 & 1 & 0 \\ f & 0 & 1 & 0 \\ g & 0 & 1 & 0 \end{array}$$

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

Punch Line

Each distinct sequence S[C] in M[C] corresponds to an edge e branching off of B.

Algorithmically

- Finding the tree part of the blobbed-tree is easy.
- Determining the sequences labeling the exterior nodes on any blob is easy.
- Determining a “good” structure inside a blob B is the problem of generating the sequences of the exterior nodes of B.
- It is easy to test whether the exterior sequences on B can be generated with only a single (possibly multiple-crossover) recombination. The original galled-tree problem is now just the problem of testing whether one single-crossover recombination is sufficient for each blob.

HapBound vs. RecMin on LPL from Clark et al.

Program	Lower Bound	Time
RecMin (default)	59	3s
RecMin –s 25 –w 25	75	7944s
RecMin –s 48 –w 48	No result	5 days
HapBound ORB	75	31s
HapBound -S	78	1643s

2 Ghz PC

Example where RecMin has difficulity in Finding the ORB on a 25 by 376 Data Matrix

Program	Bound	Time
RecMin default	36	1s
RecMin –s 30 –w 30	42	3m 25s
RecMin –s 35 –w 35	43	24m 2s
RecMin –s 40 –w 40	43	2h 9m 4s
RecMin –s 45 –w 45	43	10h 20m 59s
HapBound	44	2m 59s
HapBound -S	48	39m 30s