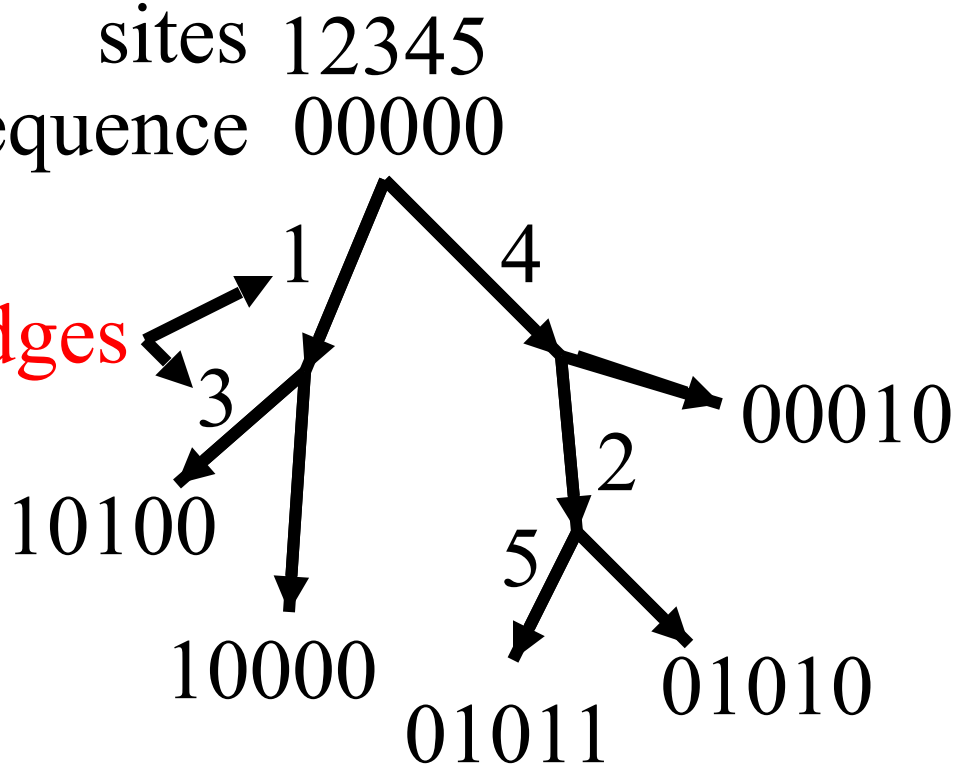
Combinatorial Optimization and Combinatorial Structure in Computational Biology

Dan Gusfield, Computer Science, UC Davis

Extended Perfect Phylogeny Problems and Applications using

- Trees
- Cycles
- Chordal Graphs
- Matroids
- Networks
- Bi-Convex graphs
- Connected components

The Perfect Phylogeny Model for binary sequences



Site mutations on edges

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.

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

M

n

m

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

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

We will use the classic theorem in two more modern and more genetic applications.

Haplotyping via Perfect Phylogeny - Model, Algorithms

Genotypes and Haplotypes

Each individual has two "copies" of each chromosome.

At each site, each chromosome has one of two alleles (states) denoted by 0 and 1 (motivated by SNPs)

0 1 1 1 0 0 1 1 0

1 1 0 1 0 0 1 0 0

Two haplotypes per individual

Merge the haplotypes

2 1 2 1 0 0 1 2 0 Genotype for the individual

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 bases of DNA.
- 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.

Haplotype Map Project: HAPMAP

- NIH lead project ($100M) to find common haplotypes in the Human population.
- Used to try to associate genetic-influenced diseases with specific haplotypes, to either find causal haplotypes, or to find the region near causal mutations.
- Haplotyping individuals is expensive.

Haplotyping Problem

- Biological Problem: For disease association studies, haplotype data is more valuable than genotype data, but haplotype data is hard to collect. Genotype data is easy to collect.

- Computational Problem: Given a set of n genotypes, determine the original set of n haplotype pairs that generated the n genotypes. This is hopeless without a genetic model.

The Perfect Phylogeny Model

We assume that the evolution of extant haplotypes can be displayed on a rooted, directed tree, with the all-0 haplotype at the root, where each site

changes from 0 to 1 on exactly one edge, and each extant haplotype is created by accumulating the changes on a path from the root to a leaf, where that haplotype is displayed.

In other words, the extant haplotypes evolved along a perfect phylogeny with all-0 root.

Perfect Phylogeny Haplotype (PPH)

Given a set of genotypes S, find an explaining set of haplotypes that fits a perfect phylogeny.

sites

S

	1	2
a	2	2
b	0	2
c	1	0

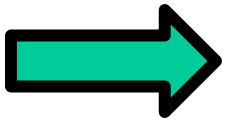
Genotype matrix

A haplotype pair explains a genotype if the merge of the haplotypes creates the genotype. Example: The merge of 0 1 and 1 0 explains 2 2.

The PPH Problem

Given a set of genotypes, find an explaining set of haplotypes that fits a perfect phylogeny

	1	2
a	2	2
b	0	2
c	1	0

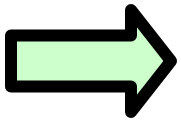


	1	2
a	1	0
a	0	1
b	0	0
b	0	1
c	1	0
c	1	0

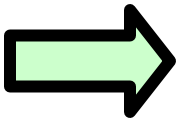
The Haplotype Phylogeny Problem

Given a set of genotypes, find an explaining set of haplotypes that fits a perfect phylogeny

	1	2
a	2	2
b	0	2
c	1	0

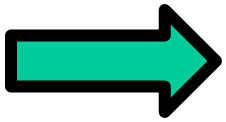


	1	2
a	1	0
a	0	1
b	0	0
b	0	1
c	1	0
c	1	0

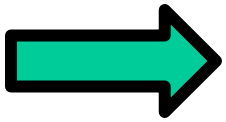


The Alternative Explanation

	1	2
a	2	2
b	0	2
c	1	0



	1	2
a	1	1
a	0	0
b	0	0
b	0	1
c	1	0
c	1	0



No tree possible for this explanation

Efficient Solutions to the PPH problem - n genotypes, m sites

- Reduction to a graph realization problem (GPPH) - build on Bixby-Wagner or Fushishige solution to graph realization O(nm alpha(nm)) time. Gusfield, Recomb 02
- Reduction to graph realization - build on Tutte's graph realization method O(nm^2) time. Chung, Gusfield 03
- Direct, from scratch combinatorial approach -O(nm^2) Bafna, Gusfield et al JCB 03
- Berkeley (EHK) approach - specialize the Tutte solution to the PPH problem - O(nm^2) time.

The Reduction Approach

The case of the 1's

1) For any row i in S, the set of 1 entries in row i specify the exact set of mutations on the path from the root to the least common ancestor of the two leaves labeled i, in every perfect phylogeny for S.
2) The order of those 1 entries on the path is also the same in every perfect phylogeny for S, and is easy to determine by "leaf counting".

Leaf Counting

In any column c, count two for each 1, and count one for each 2. The total is the number of leaves below mutation c, in every perfect phylogeny for S. So if we know the set of mutations on a path from the root, we know their order as well.

S

	1	2	3	4	5	6	7
a	1	0	1	0	0	0	0
b	0	1	0	1	0	0	0
c	1	2	0	0	2	0	2
d	2	2	0	0	0	2	0
Count	5	4	2	2	1	1	1

So Assume

The columns are sorted by leaf-count, largest to the left.

Similarly

In any perfect phylogeny, the edge corresponding to the leftmost 2 in a row must be on a path just after the 1's for that row.

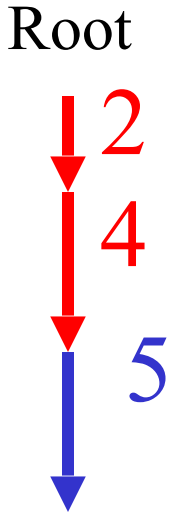
Simple Conclusions

sites

1 2 3 4 5 6 7

i:0 1 0 1 2 2 2

Subtree for row i data



The order is known for the red mutations together with the leftmost blue mutation.

But what to do with the remaining blue entries (2’s) in a row?

More Simple Tools

3) For any row i in S, and any column c, if S(i,c) is 2, then in every perfect phylogeny for S, the path between the two leaves labeled i, must contain the edge with mutation c.

Further, every mutation c on the path between the two i leaves must be from such a column c.

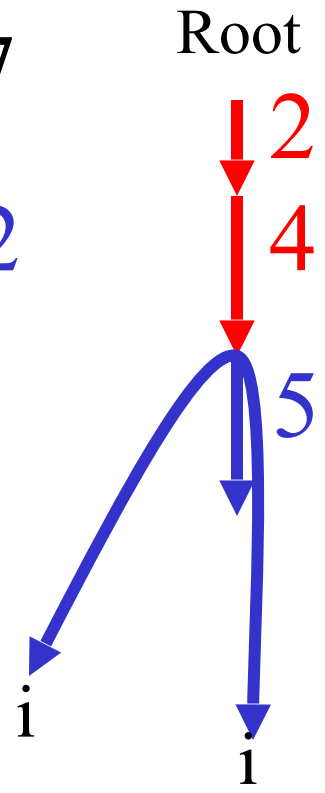
From Row Data to Tree Constraints

sites

1 2 3 4 5 6 7

i:0 1 0 1 2 2 2

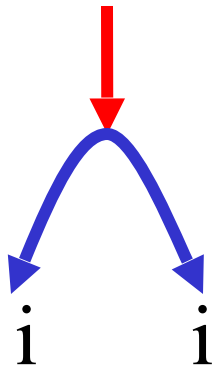
Subtree for row i data



Edges 5, 6 and 7 must be on the blue path, and 5 is already known to follow 4, but we don't where to put 6 and 7.

The Graph Theoretic Problem

Given a genotype matrix S with n sites, and a red-blue subgraph for each row i,



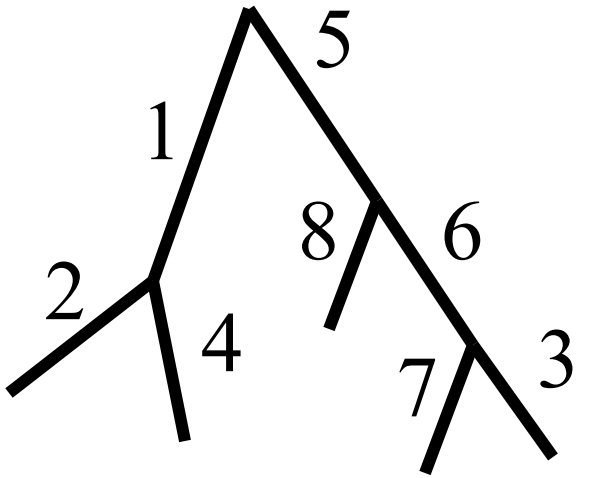
create a directed tree T where each integer from 1 to n labels exactly one edge, so that each subgraph is contained in T.

Powerfull Tool: Graph Realization

- Let Rn be the integers 1 to n, and let P be an unordered subset of Rn. P is called a path set.
- A tree T with n edges, where each is labeled with a unique integer of Rn, realizes P if there is a contiguous path in T labeled with the integers of P and no others.
- Given a family P1, P2, P3…Pk of path sets, tree T realizes the family if it realizes each Pi.
- The graph realization problem generalizes the consecutive ones problem, where T is a path.

Graph Realization Example

P1: 1, 5, 8
P2: 2, 4
P3: 1, 2, 5, 6
P4: 3, 6, 8
P5: 1, 5, 6, 7



Realizing Tree T

Graph Realization

Polynomial time (almost linear-time) algorithms exist for the graph realization problem – Whitney, Tutte, Cunningham, Edmonds, Bixby, Wagner, Gavril, Tamari, Fushishige, Lofgren 1930's - 1980's

Most of the literture on this problem is in the context of determining if a binary matroid is graphic.

The algorithms are not simple; none implemented before 2002.

Reducing PPH to graph realization

We solve any instance of the PPH problem by creating appropriate path sets, so that a solution to the resulting graph realization problem leads to a solution to the PPH problem instance.

The key issue: How to encode the needed subgraph

for each row, and glue them together at the root.

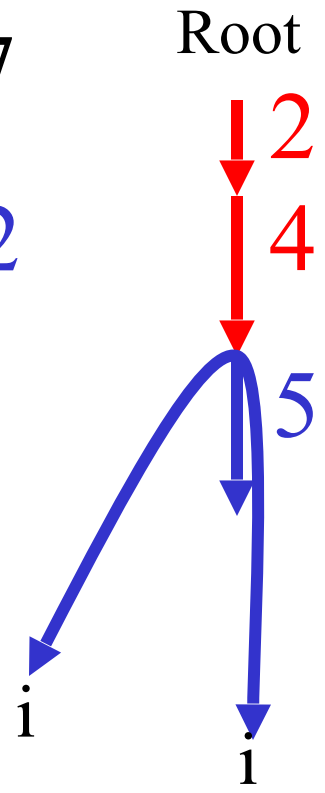
From Row Data to Tree Constraints

sites

1 2 3 4 5 6 7

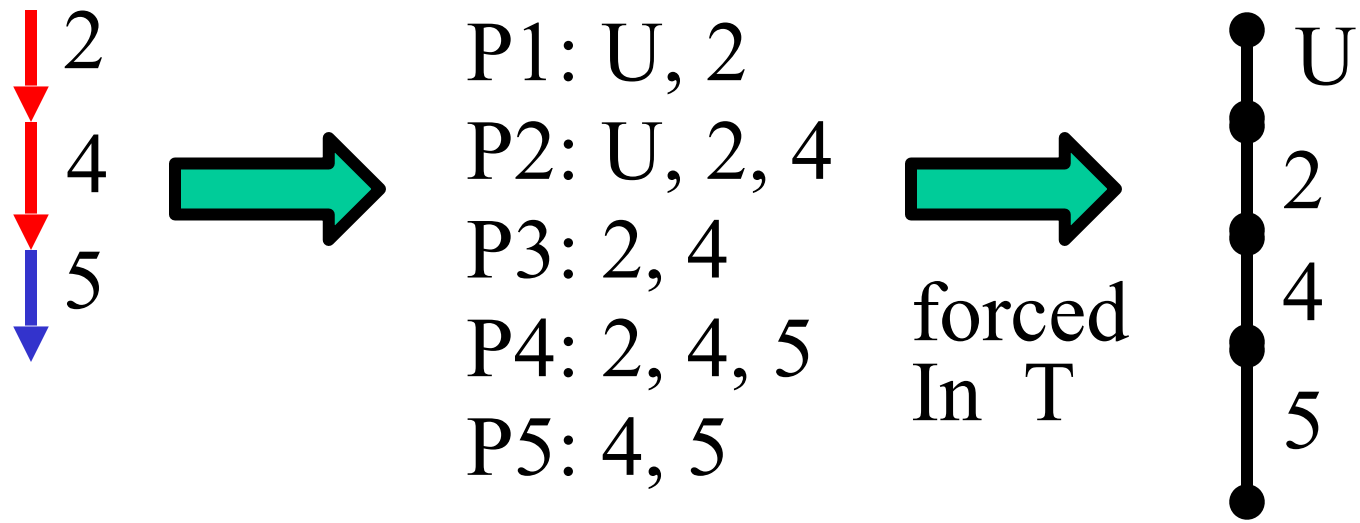
i:0 1 0 1 2 2 2

Subtree for row i data



Edges 5, 6 and 7 must be on the blue path, and 5 is already known to follow 4.

Encoding a Red-Blue directed path



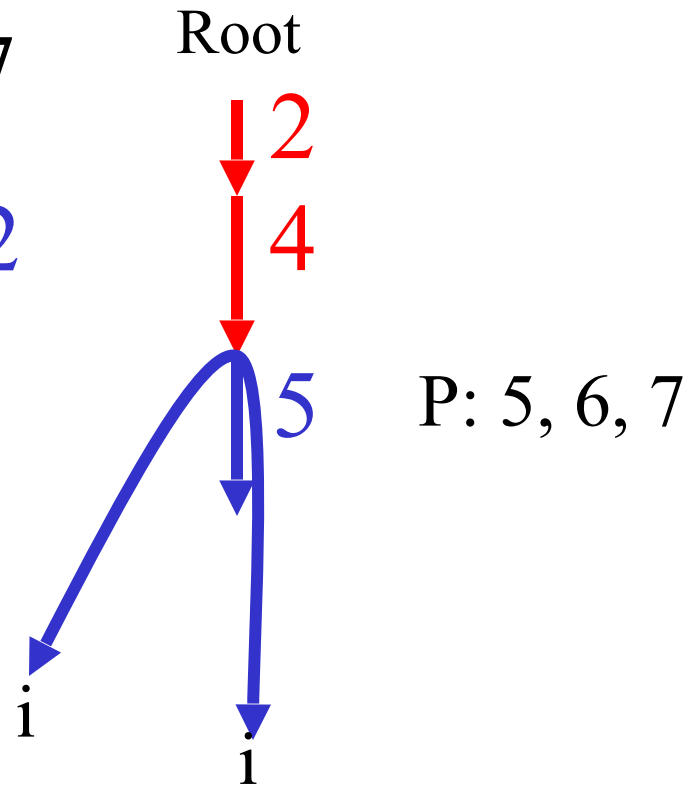
U is a glue edge used to glue together the directed paths from the different rows.

Now add a path set for the blues in row i.

sites

1 2 3 4 5 6 7

i:0 1 0 1 2 2 2



P: 5, 6, 7

That's the Reduction

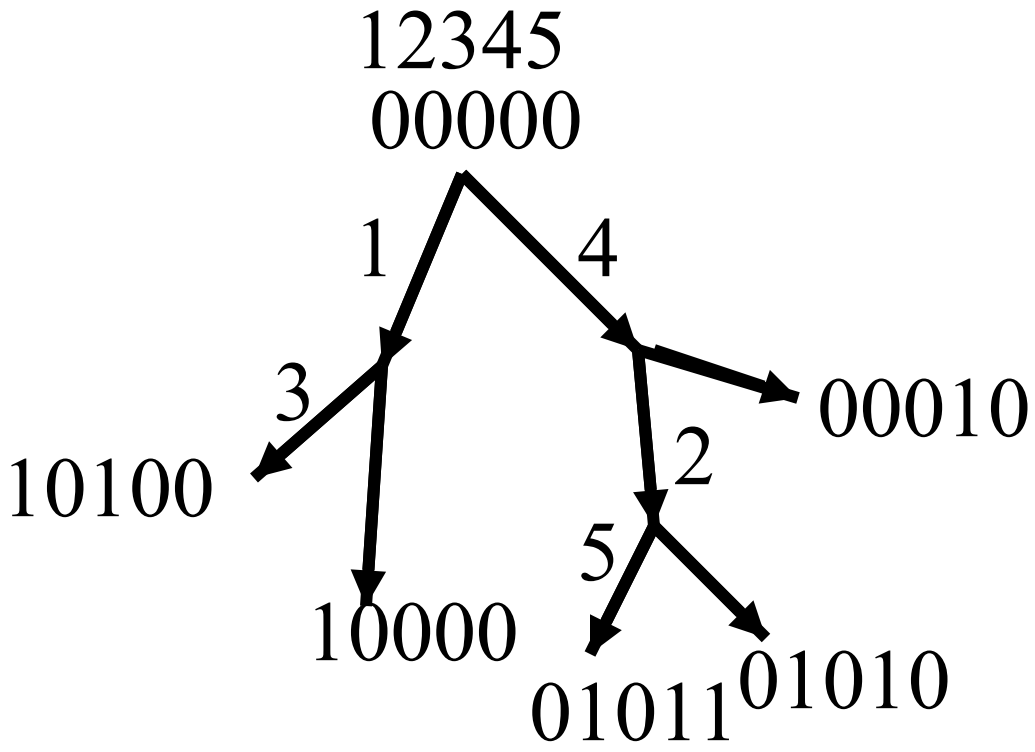
The resulting path-sets encode everything that is known about row i in the input.

The family of path-sets are input to the graph-realization problem, and every solution to the that graph-realization problem specifies a solution to the PPH problem, and conversely.

Whitney (1933?) characterized the set of all solutions to graph realization (based on the three-connected components of a graph) and Tarjan et al showed how to find these in linear time.

Phylogenetic Networks: A richer model of haplotype evolution

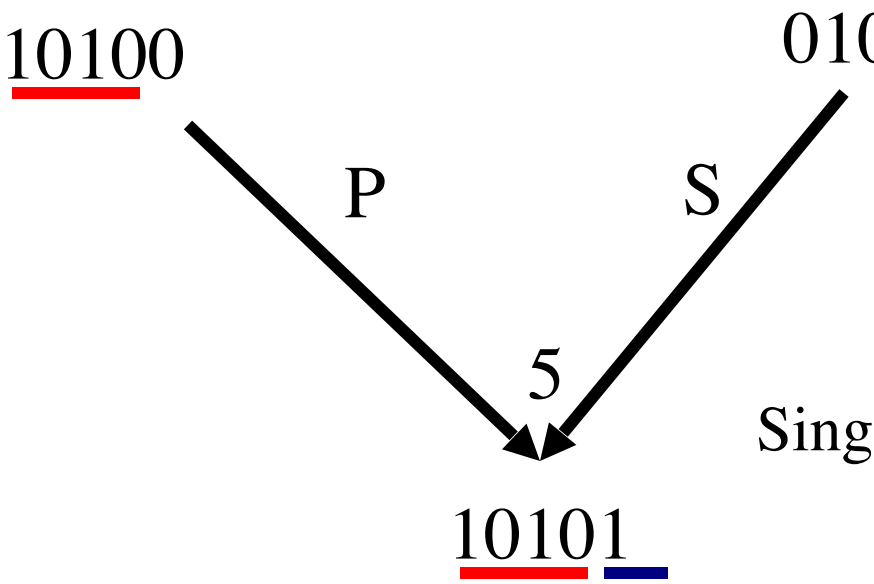
10100
10000
01011
01010
00010
10101 new



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

Real sequence histories often involve recombination.

Sequence Recombination



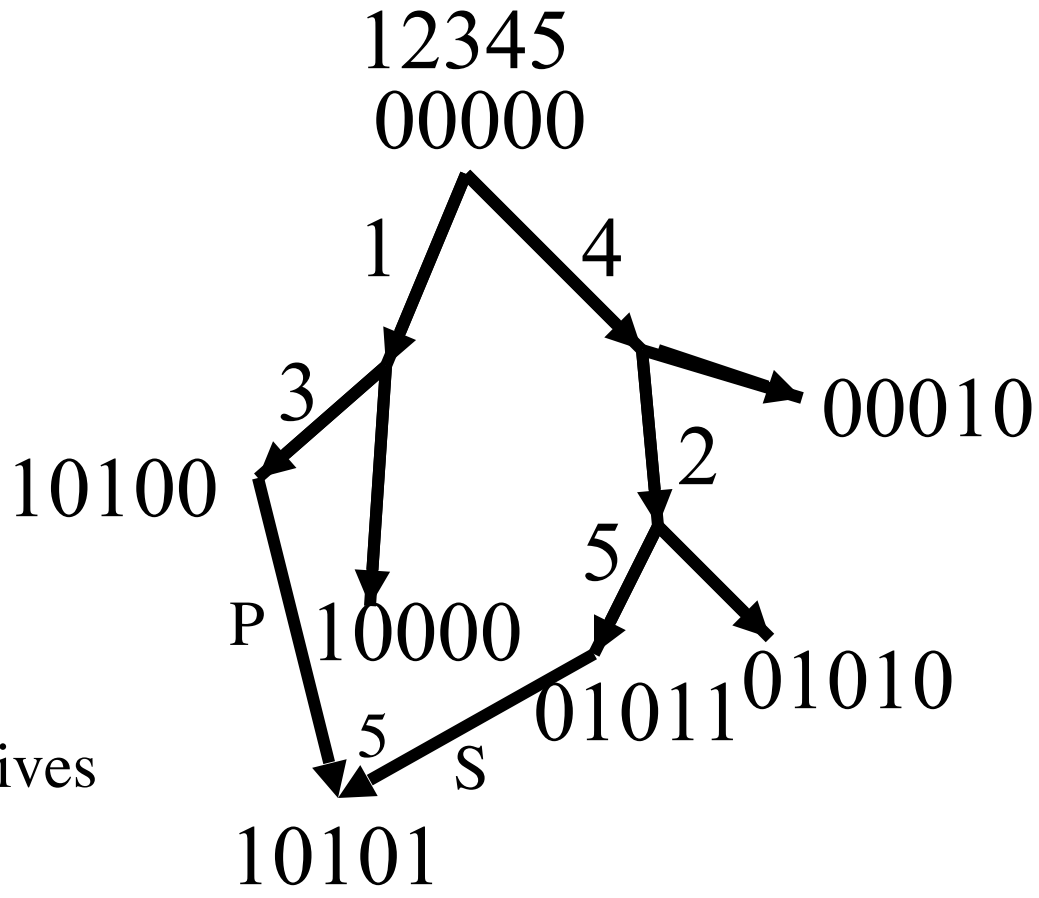
Single crossover recombination

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

10100
10000
01011
01010
00010
10101 new

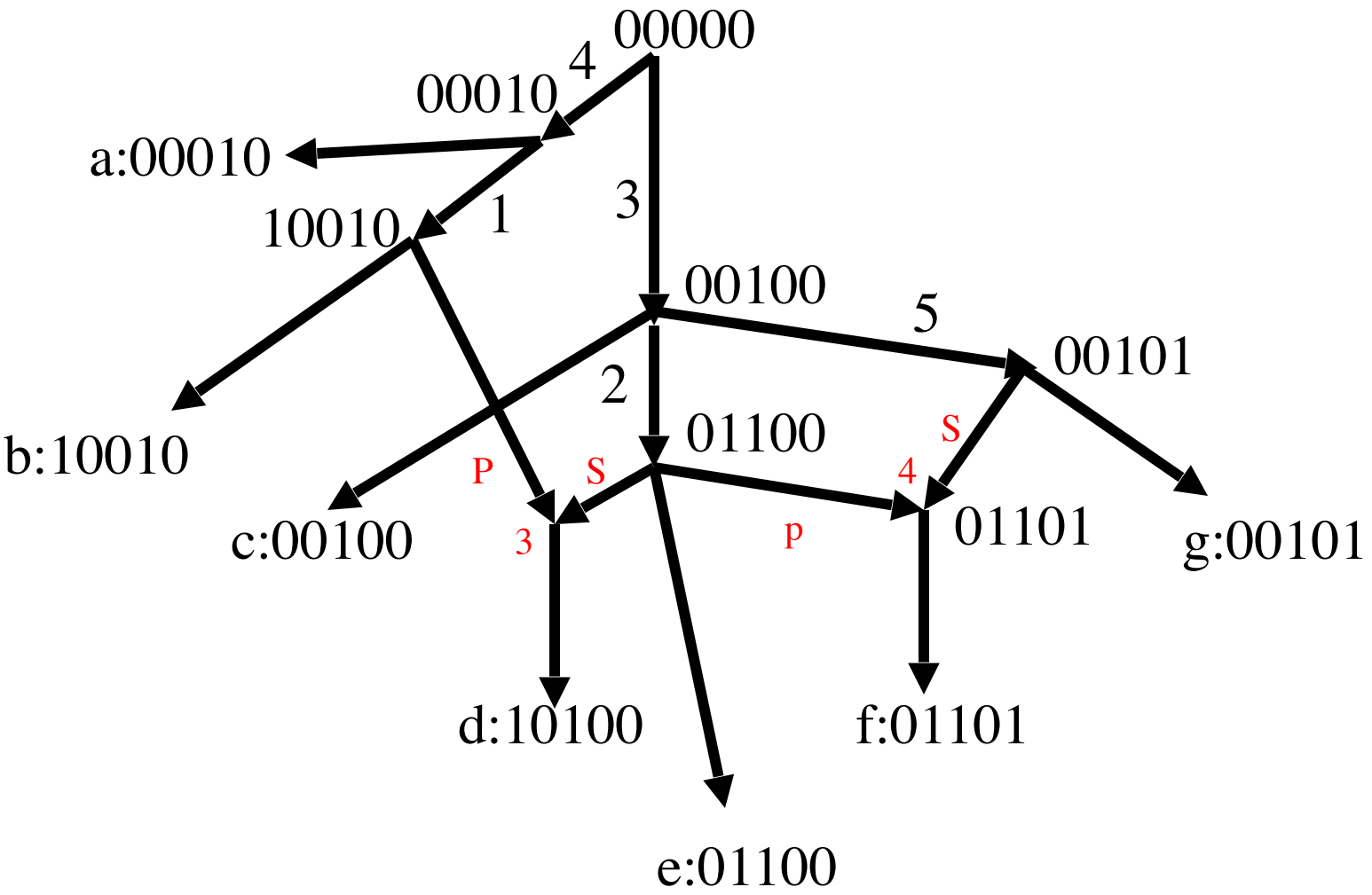


The previous tree with one recombination event now derives all the sequences.

Elements of a Phylogenetic Network (single crossover recombination)

- 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. Use as little recombination or gene-conversion as possible.

Recombination in a population is the key to gene-finding methods based on associating genetic markers with observed traits (disease or favorable trait). Nature, through recombination, has done many ``binary search'' experiments.

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

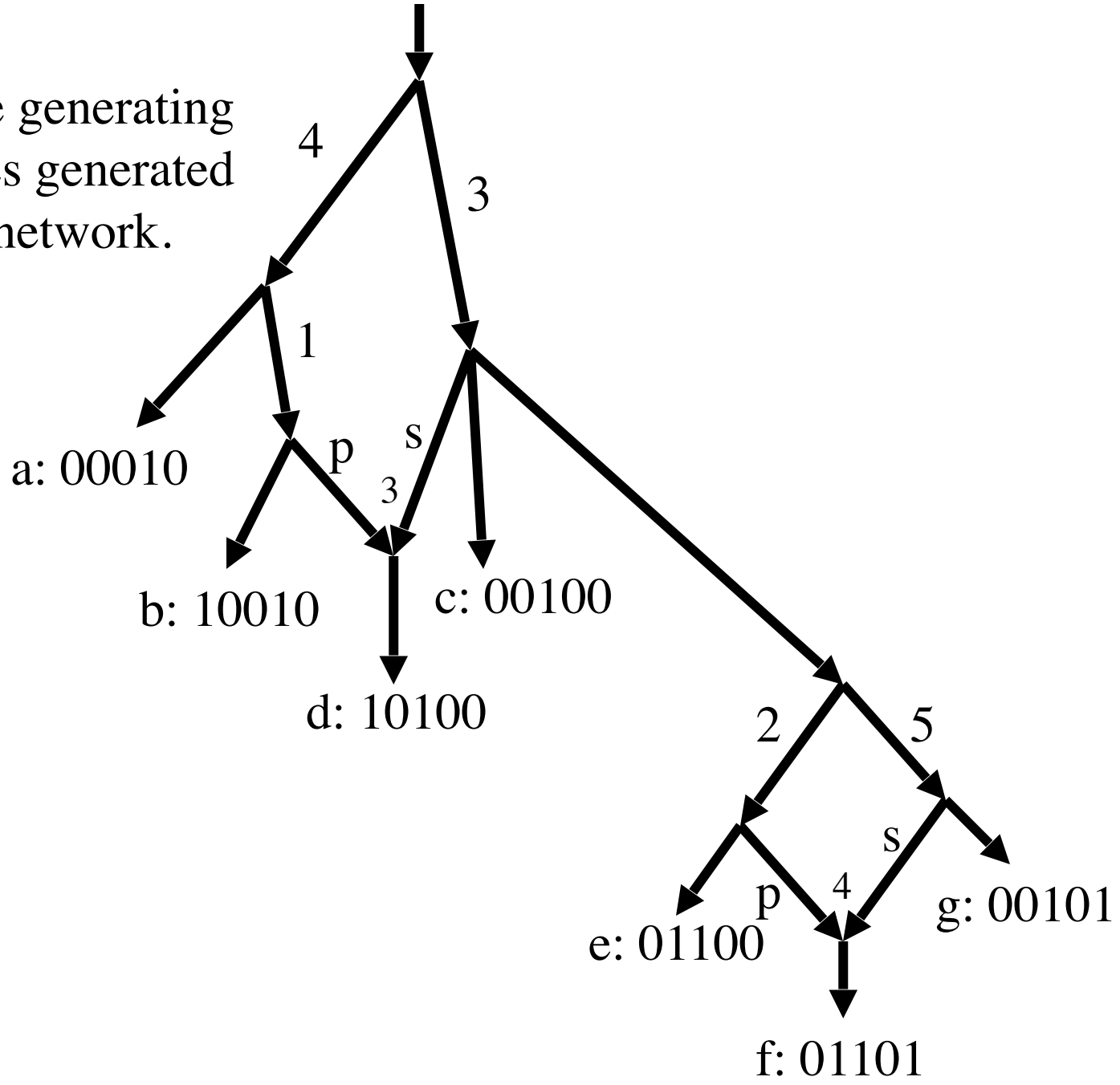
- In a Phylogenetic Network, with a recombination node x, if we trace two paths backwards from x, then the paths will eventually meet.
- The cycle specified by those two paths is called a ``recombination cycle”.

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.



Old (Aug. 2003) Results

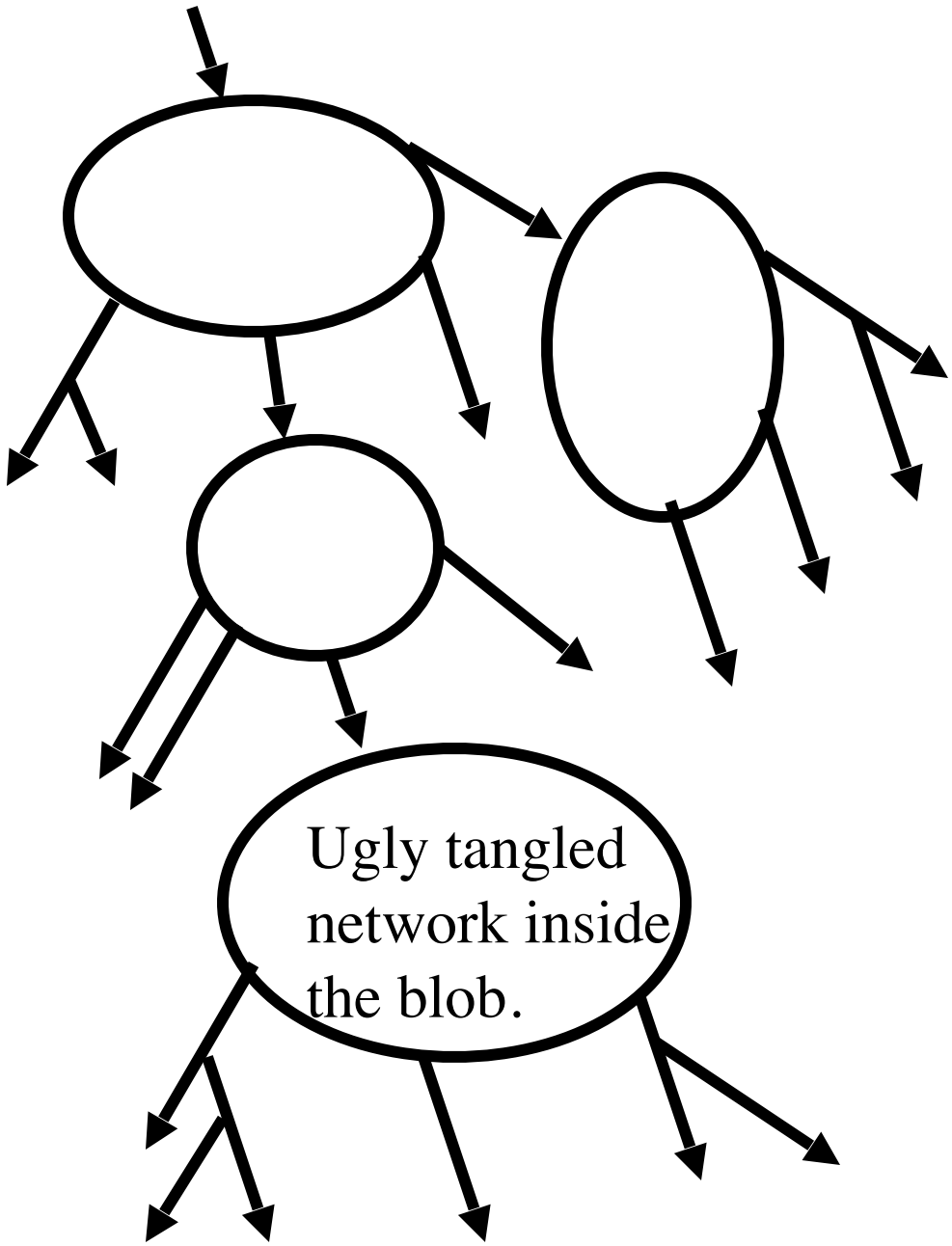
- O(nm + n^3)-time algorithm to determine whether or not M can be derived on a galled-tree.
- Proof that the galled-tree produced by the algorithm is a "nearly-unique" solution.
- Proof that the galled-tree (if one exists) produced by the algorithm minimizes the number of recombinations used, over all phylogenetic-networks with all-0 ancestral sequence.

To appear in J. Bioinformatics and Computational Biology, Gusfield, Edhu, Langley

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.



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

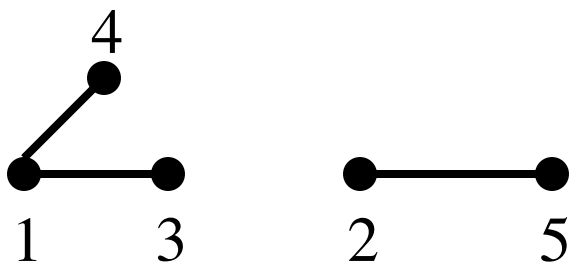
How can we exploit this relationship?

The Structure of the Tree Part

M

	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

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.

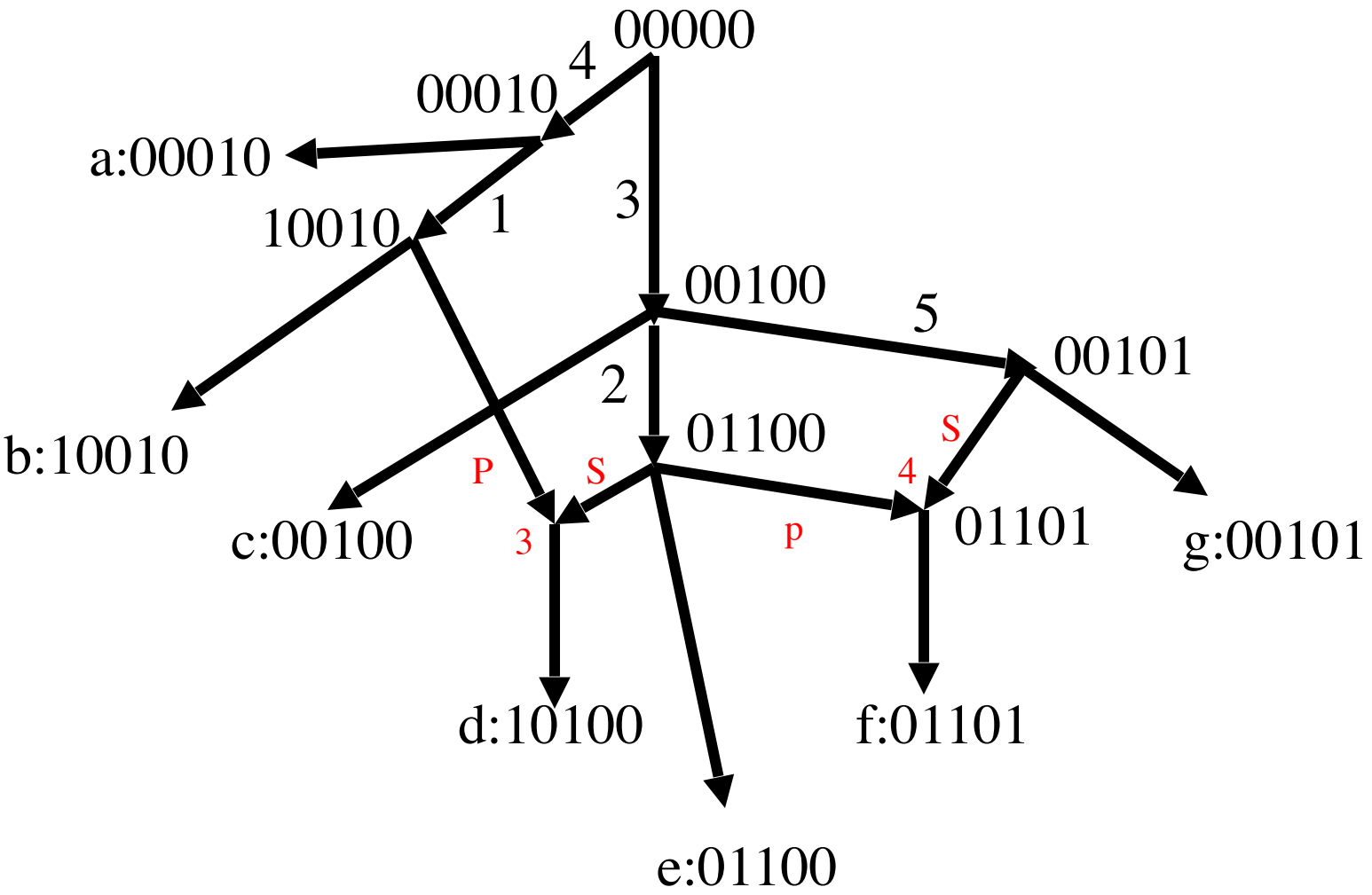
Simple Fact

If sites two sites i and j conflict, 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, site 1 conflicts with 3 and 4; and site 2 conflicts with 5.

A Phylogenetic Network



Simple Consequence of the simple fact

All sites on the same (non-trivial) connected component of the conflict 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 conflict 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 conflict graph. All unconflicted sites can be put on edges outside of the galls.

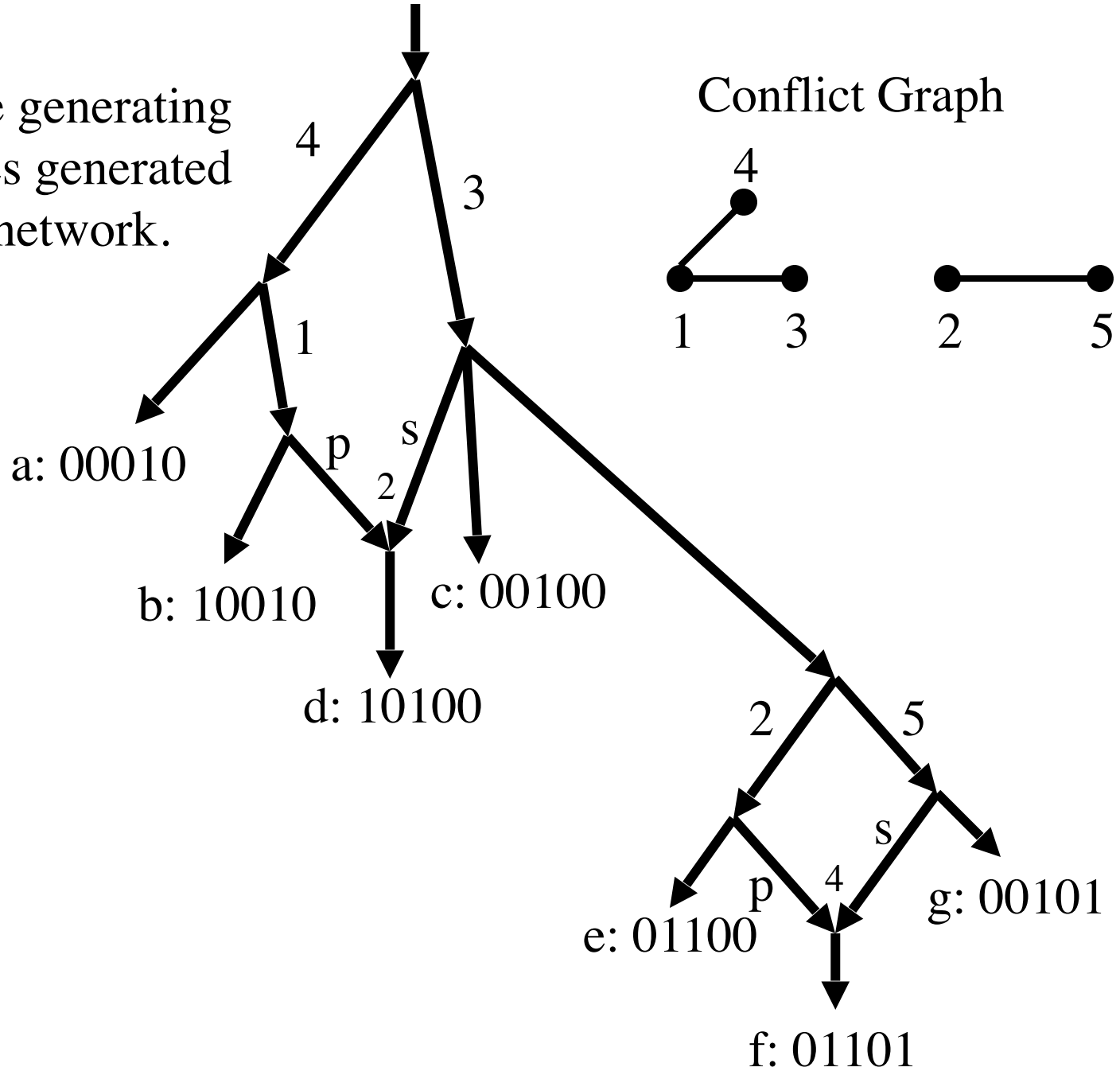
This is the key to the efficient solution to the galled-tree problem.

Optimality

The number of recombinations is

the number of non-trivial connected components, and this is the minimum possible.

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



Conflict Graph

The main new result

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 the conflict graph. The unconflicted sites can always be put on edges outside of any blob. Moreover, the tree part of T(M) is unique.

This is bit weaker than the result for galled-trees: it replaces "must" with "can".

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 recombination. The original galled-tree problem is now just the problem of testing whether one single-crossover recombination is sufficient for each blob.
- That can be solved by successively removing each exterior sequence and testing if the remaining sequences can be generated on a perfect phylogeny of the correct form.

Necessary Condition for a Galled-Tree

If M can be generated on a galled-tree, then the conflict graph must be a bipartite, bi-convex graph. Other structural properties of the conflict graph can be deduced and exploited.

Another extension of the basic Perfect Phylogeny model: non-binary characters

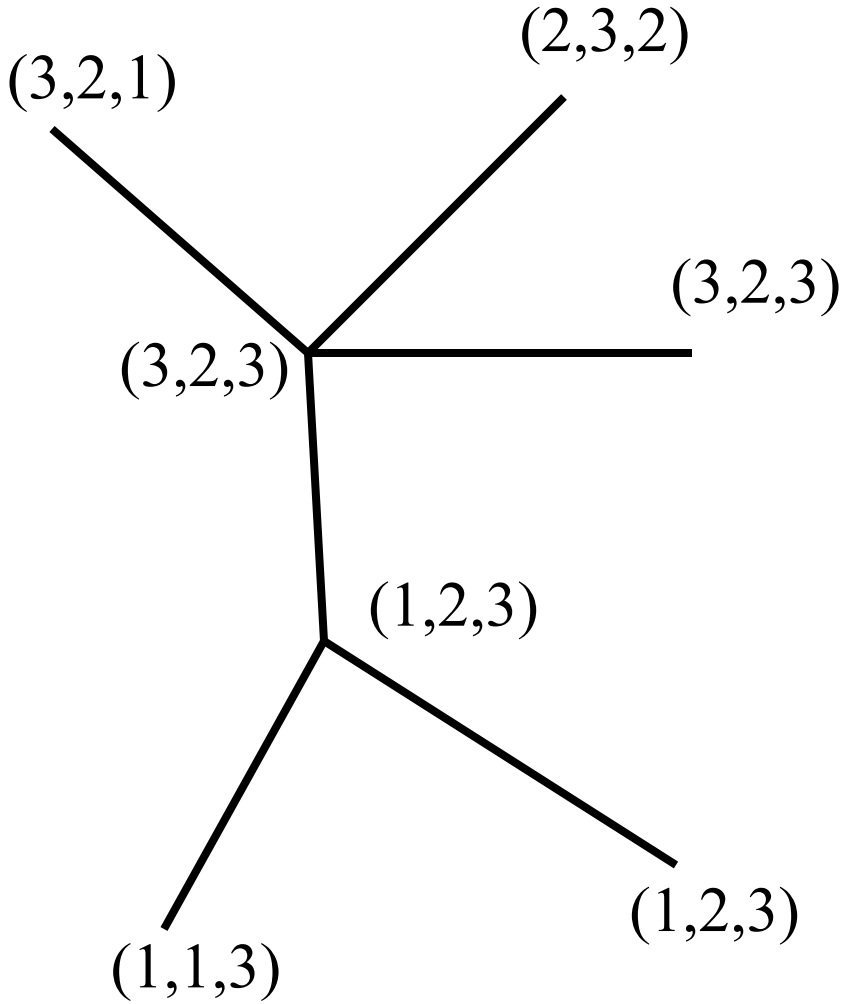
The pairwise theorem for binary characters does not hold for when there are more than two states per character.

That is, each pair of characters may allow a tree, but the entire set of characters does not.

What is a Perfect Phylogeny for non-binary characters?

- Given K characters (columns), with up to q states per character, and n rows (taxa) in a table E.
- In a Perfect Phylogeny T for E, each node of T is labeled with K-states, one from each of the K characters.
- T has n leaves, and each leaf is labeled with the states of a distinct row of E.
- For each character-state pair (C,i), the nodes of T that are labeled with state i for character C, form a connected subtree of T.

Example

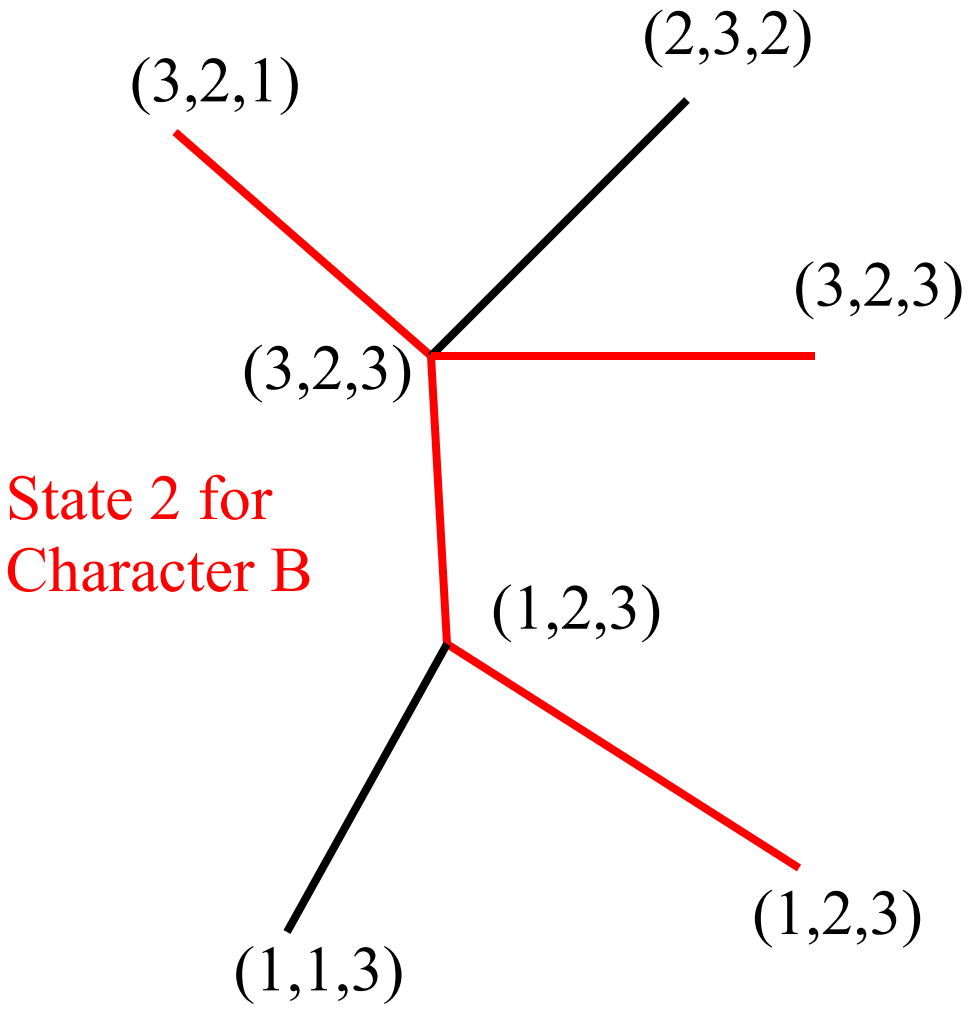


	A	B	C
1	3	2	1
2	2	3	2
3	3	2	3
4	1	1	3
5	1	2	3

Table E

n = 5
K = 3

Example



	A	B	C
1	3	2	1
2	2	3	2
3	3	2	3
4	1	1	3
5	1	2	3

Table E

n = 5
K = 3

Perfect Phylogeny Problem

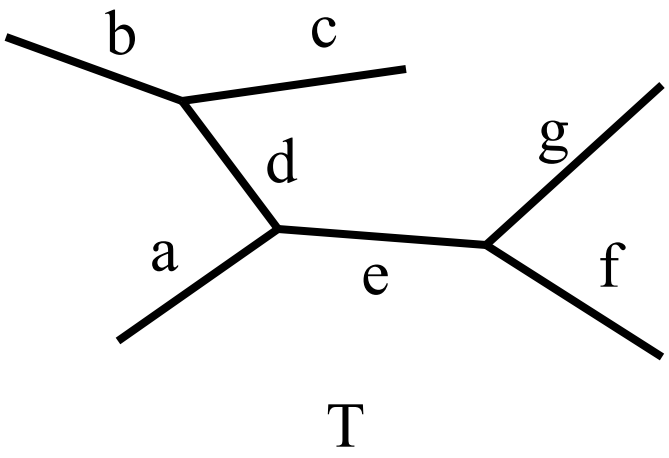
Given a table E, is there a Perfect Phylogeny for E?

Chordal Graphs

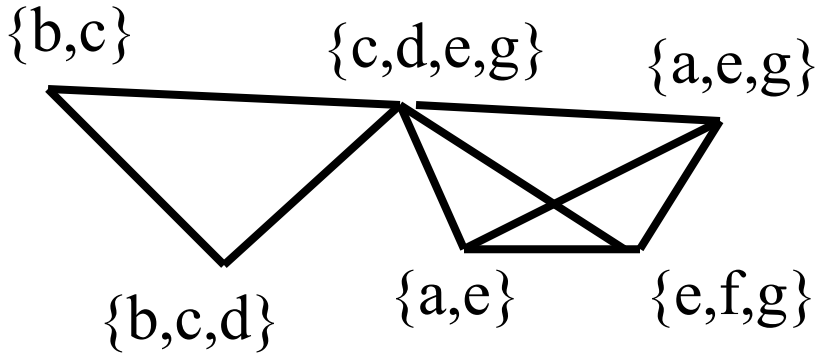
A graph G is called Chordal if every cycle of length four or more contains a chord.

Classic Chordal Graph Theorem

A graph G is chordal if and only if it is the intersection graph of a set S of subtrees of a tree T. Each node of G is a member of S.



T



G

Relation to Perfect Phylogeny

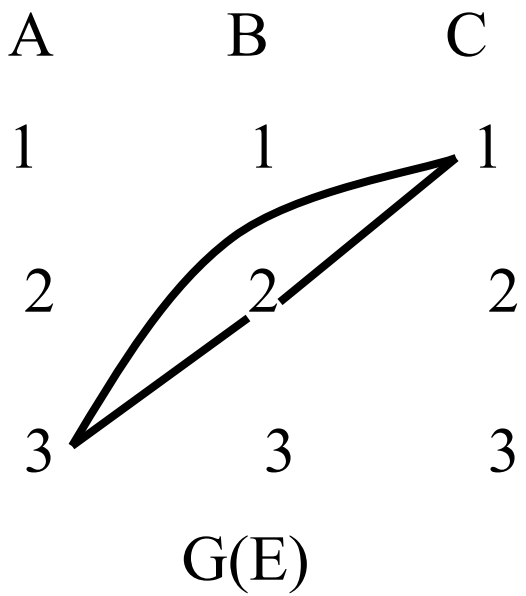
In a perfect phylogeny T for a table E, for any character C and any state X of character C, the subgraph of T induced by the nodes labeled (C,X) form a single, connected subtree of T.

So, there is a natural set of subtrees of T induced by E.

Chordal Completion Approach to Perfect Phylogeny

	A	B	C
1	3	2	1
2	2	3	2
3	3	2	3
4	1	1	3
5	1	2	3

Table E



G(E)

Graph G(E) has one node for each character-state pair in E, and an edge between two nodes if and only if there is a row in E with both those character-state pairs.

Each row of table E induces a clique in G(E).

Classic Theorem

Note that if table E has K columns, then G(E) is a K-partite graph.

Theorem (Buneman 196?)

There is a perfect phylogeny for table E if and only if edges can be added to graph G(E) to make it a chordal, K-partite graph.

The perfect phylogeny problem was open for about 20 years, but solved by Warnow, Kannan, Agarwalla and Fernandez-Baca.

Perfect Phylogeny Results

For any fixed bound on the number of states per character,
the Perfect Phylogeny Problem can be solved in polynomial time.

However, if the number of states per character is not bounded,
then the problem is NP-Complete.

Papers, Powerpoint and Programs

wwwcsif.cs.ucdavis.edu/~gusfield/

New Journal: IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)

see computer.org/tcbb/