

# Progress in Combinatorial Haplotyping

Dan Gusfield

The results in this talk are in collaboration with D. Brown, Z. Ding,  
V. Filkov, C. Langley, Y. Frid, Y. Song, Y. Wu,

USC January 28, 2007

# Three “Post-HGP” Topics

In the past five years my group has addressed three topics in Population Genomics

- SNP Haplotyping in populations
- Reconstructing histories of recombinations and mutations through phylogenetic networks
- The intersection of the two problems

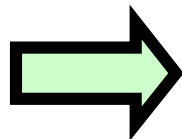
# Themes of Combinatorial Haplotyping

- The use of **objective functions and exact** optimization provide precise **semantics** of **what** is computed, even if the **how** is heuristic or ad hoc.
- **Progression of combinatorial haplotyping problems to incorporate more biological reality, or context:** PPH; MinIncompat, Parsimony, MinPPH; IPPH; Galled-tree haplotyping; recombination LB haplotyping, MinHK, MinCC, MinRecMin, MinRec, ...
- **Efficient computation:** Linear-time PPH, ILP formulations and experiments, exact Min recombination...

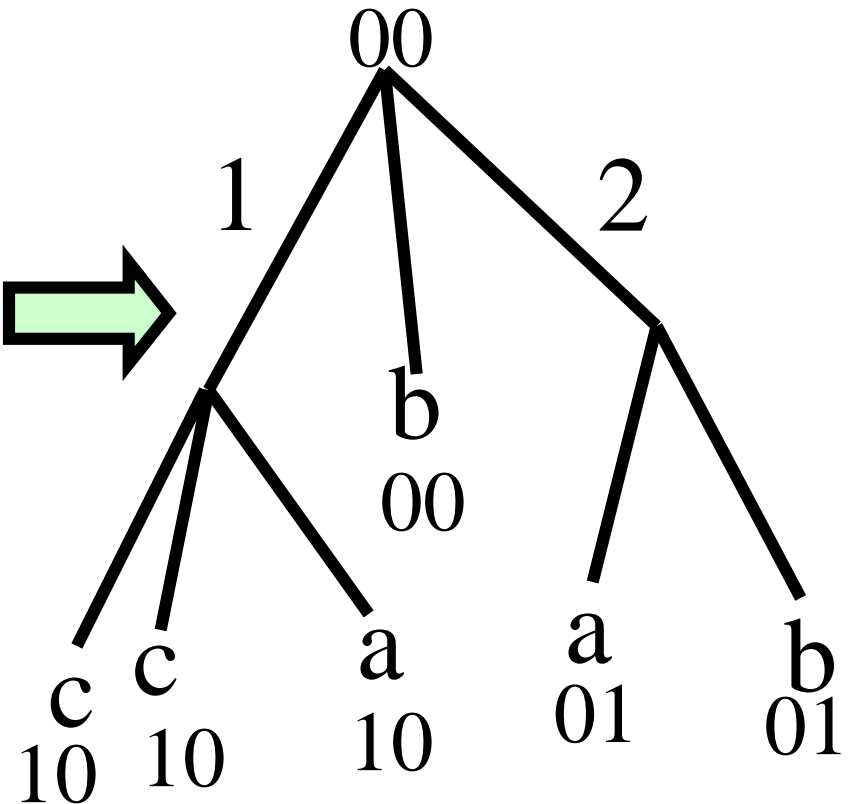
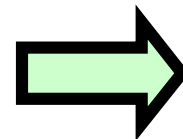
# The PPH Problem

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



# Equivalently

The PPH problem is: Given a set of genotypes, find an explaining set of haplotypes that passes the “four gamete test”: no pair of sites contains all four binary pairs (gametes):

0,0; 0,1; 1,0; 1,1

A pair of sites that has all four gametes is called **incompatible**, otherwise is called **compatible**.

# 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 in  $O(nm \alpha(nm))$  time. Connects to Matroid theory and matroid literature. 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, Lancia, Yooseph JCB 03)
- Berkeley (Eskin, Halperin, Karp) approach -  $O(nm^2)$  time.
- **Linear-time** solutions - (Ding, Gusfield, Filkov) Recomb 2005, and two other linear time solutions. These are graph-theory methods. The matrix-centric solutions require finding all potentially incompatible pairs, and this is equivalent to Boolean matrix multiplication, and so are unlikely to be implementable in linear time (BGHY).

When can a set of **haplotypes** 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 in Genetics

A pair of sites that has all four gametes is called **incompatible**, otherwise is called **compatible**.

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 1982-1991).

So determining if there are any incompatible pairs can be done in linear time. However, the problem of finding all incompatible pairs is equivalent to Boolean matrix multiplication (BGHY 05), and so is unlikely to be doable in  $O(nm + m^2)$  time.

# Related Imputation Problem

- **Perfect Phylogeny with Missing data:** Given a ternary matrix, as in the PPH problem, can the 2's be changed to 0's and 1's so that the resulting matrix has a perfect phylogeny? Sounds similar to PPH, but is NP-hard. (Steel?) However, it can be solved very efficiently in practice by ILP - detailed in this talk.
- If the root of the required perfect phylogeny is specified, then the problem is polynomial-time solvable (Pe'er, Sharan, Shamir).

# Relaxing and extending the perfect phylogeny model

**MinIncompat Problem:** Haplotype to minimize the resulting number of incompatible pairs of sites.

A natural generalization of PPH, which is the zero case.

NP-hard, but we have seen it can be very efficiently solved in practice by a simple ILP formulation.

# Generalized missing data problem

Given a ternary matrix (0s, 1s, 2s), change the twos to zeros and ones in order to **minimize** the resulting number of incompatible pairs of sites.

Perfect Phylogeny with Missing Data:  
determine if zero incompatibilities are possible.

# Naive ILP for the generalized missing data problem

Create a binary variable  $Y(i,p)$  for a 2 in cell  $(i,p)$ ,  
indicating whether the cell will be set to 0 or to 1.

For each pair of sites  $p, q$  that could be made  
incompatible, let  $D(p,q)$  be the set of missing or  
**deficient** gametes in site pair  $p,q$ .

For each  $a,b$  in  $D(p,q)$ , create binary variables:  
 $X(i,p,q,a,b)$ , which will be set to 1 if the  $(i,p)$   $(i,q)$   
pair of sites is set to gamete  $a,b$

## Example

$$D(p,q) = \{0,1; 1,1\}$$

QuickTime™ and a  
TIFF (LZW) decompressor  
are needed to see this picture.

## Example

p q	D(p,q) = {0,1; 1,1}
0 0	
2 1	
1 0	
2 2	
2 0	
0 2	

The ILP will have the following inequalities for row 2:

$$X(2, p, q, 1, 1) + X(2, p, q, 0, 1) = 1 \quad \text{one pair must be picked}$$

$$X(2, p, q, 1, 1) \leq Y(2, p) \quad \text{set the } Y(2, p) \text{ variable}$$

consistent with the setting

$$X(2, p, q, 0, 1) + Y(2, p) \leq 1 \quad \text{of the pair}$$

And similar, but more involved inequalities for row 4.

QuickTime™ and a  
TIFF (LZW) decompressor  
are needed to see this picture.

And ones for row 6, but not row 5, which can't create a gamete  
in  $D(p,q)$ .

Next, we have an inequality for each a,b gamete in  $D(p,q)$  that sets variable  $B(p,q,a,b)$  to 1 if the a,b gamete is created in **some** row of site-pair  $p,q$ .

$$X(2, p, q, 1, 1) + X(4, p, q, 1, 1) - 2B(p, q, 1, 1) \leq 0$$

$B(p,q,1,1)$  gets set to 1 if 1,1 is created in row 2 or 4.

$$X(2, p, q, 0, 1) + X(4, p, q, 0, 1) + X(6, p, q, 0, 1) - 3B(p, q, 0, 1) \leq 0$$

$B(p,q,0,1)$  gets set to 1 if 0,1 is created in row 2, 4 or 6

Then, variable  $C(p,q)$  is set to 1 if **every** gamete in  $D(p,q)$  is created for site-pair  $(p,q)$ .

$$B(p, q, 1, 1) + B(p, q, 0, 1) - C(p,q) \leq 1$$

So,  $C(p,q)$  is set to 1 if the missing entries are set so that sites  $p$  and  $q$  become incompatible.

Finally, we have the objective function:

$$\text{Minimize } \sum_{(p,q) \text{ in } P} C(p, q)$$

Where  $P$  is the set of site-pairs that could be made to be incompatible.

It is easy to extend this ILP to solve the MinIncompat problem. Empirically, these ILPs solve in fractions of seconds, or seconds even for  $m = n = 100$  and percentage of 2's up to 30%.

For the missing data problem, the ILP solution correctly imputes the missing data with 2% - 5% error rate. Lower error rates with less missing data - slightly less accurate than the results Andy Clark mentioned on Saturday

# The PPH ILP and other extensions

1. The MinIncompat ILP becomes an ILP for **PPH** with the addition of a constraint that requires the solution to have value 0. The resulting ILP is feasible if and only if there is a PPH solution.
2. Many other extensions. For example, haplotype in order to minimize the number of sites that have to be **removed** so that all remaining site pairs are compatible. This is the haplotyping version of the **site-consistency** problem in phylogenetics.

# Problems derived from MinIncompat

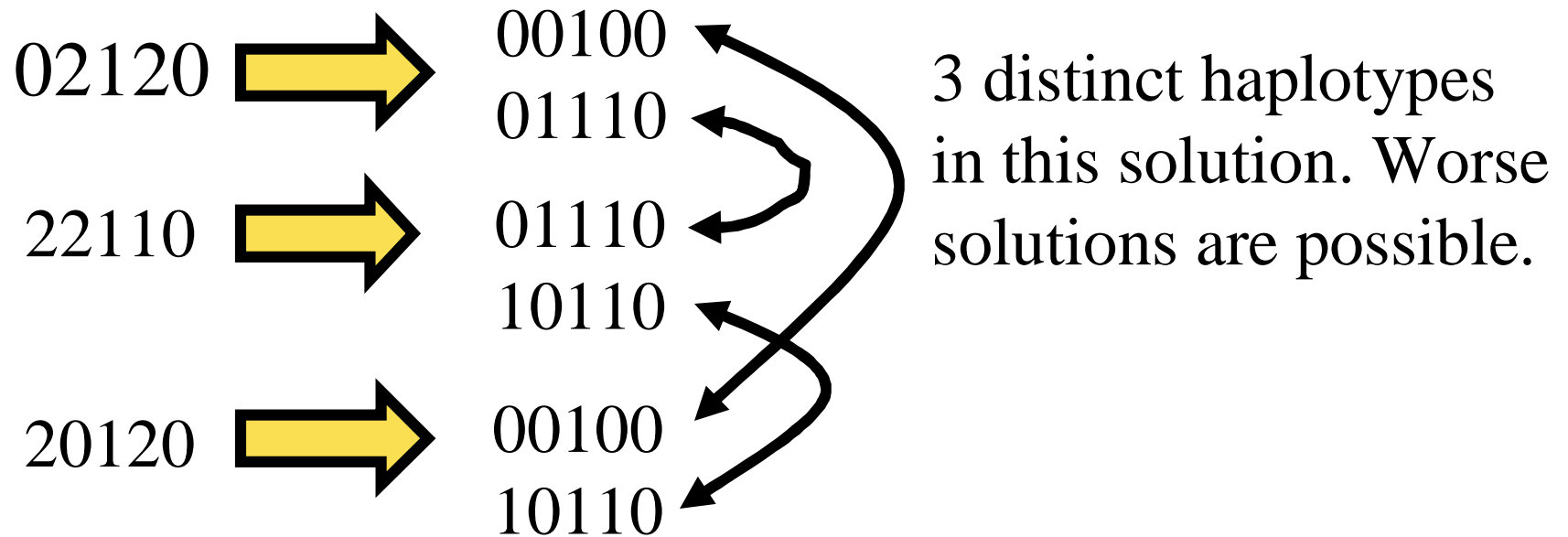
- MinPPH, discussed next
- MinHK: input missing data or haplotype to minimize the Hudson-Kaplan bound on the amount of recombination. The missing data problem is NP-hard, but efficiently solved in practice; the haplotyping problem can be solved in polynomial time using PPH (Wiuf).
- MinCC: Haplotype in order to minimize the connected-component lower bound on recombination. Polytime solution (Y. Wu)

# Another extension: Min PPH

Problem: If there is a PPH solution, find one **minimizing** the number of **distinct** haplotypes used.

Justified by empirical and theoretical grounds: very few distinct haplotypes observed in real populations; pure parsimony criteria works well; and is implicit in PHASE and other haplotyping methods.

# Example of Pure Parsimony



There is a naïve ILP for Parsimony that works (G 03) for moderate sized instances, and very clever (worst-case small) ILPs that, unfortunately only solve very small instances (BH and others)

# Practical Solution of MinPPH

- MinPPH problem is NP-hard (Bafna, Gusfield, Hannenhali, Yooseph)
- But, it can be solved very efficiently in practice by ILP. The ILP just combines the PPH ILP and the Brown, Harrower ILP for Parsimony Haplotyping.

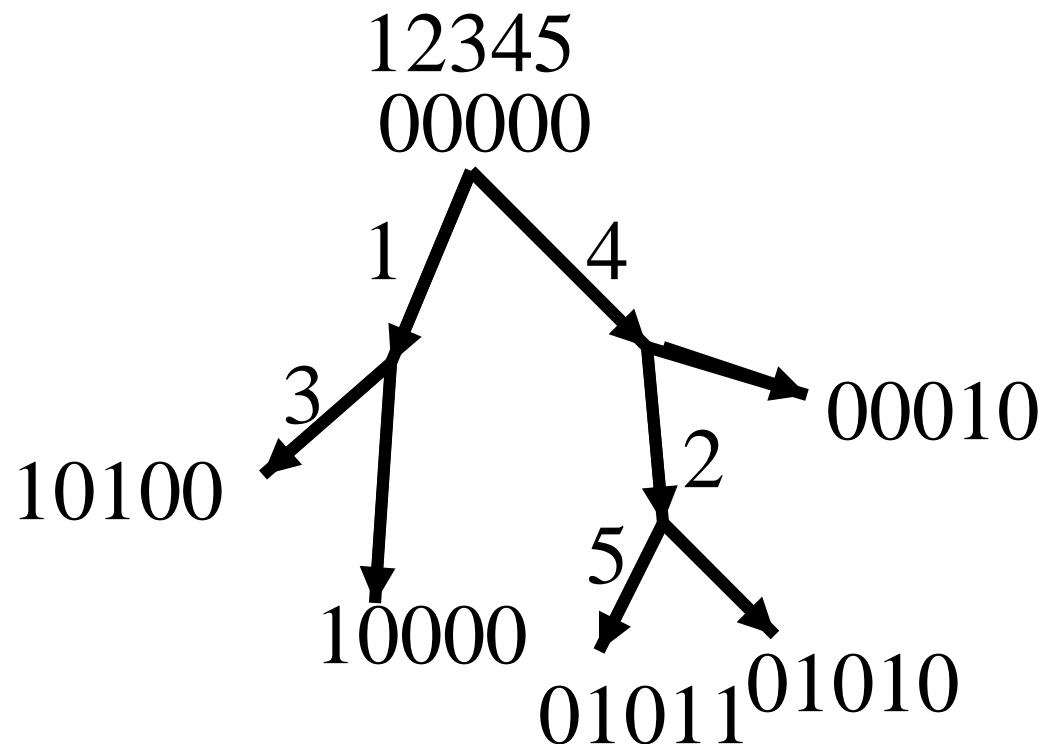
The interesting point is that the BH ILP for Parsimony can only solve tiny problem instances, but the addition of the PPH ILP inequalities makes it solve very efficiently on large problem instances.

# A richer model of haplotype evolution

M  
 10100  
 10000  
 01011  
 01010  
 00010

10101 added

Pair 4, 5 fails the four gamete-test. The sites 4, 5 ``conflict''.



Real sequence histories often involve **recombination**.

# 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.

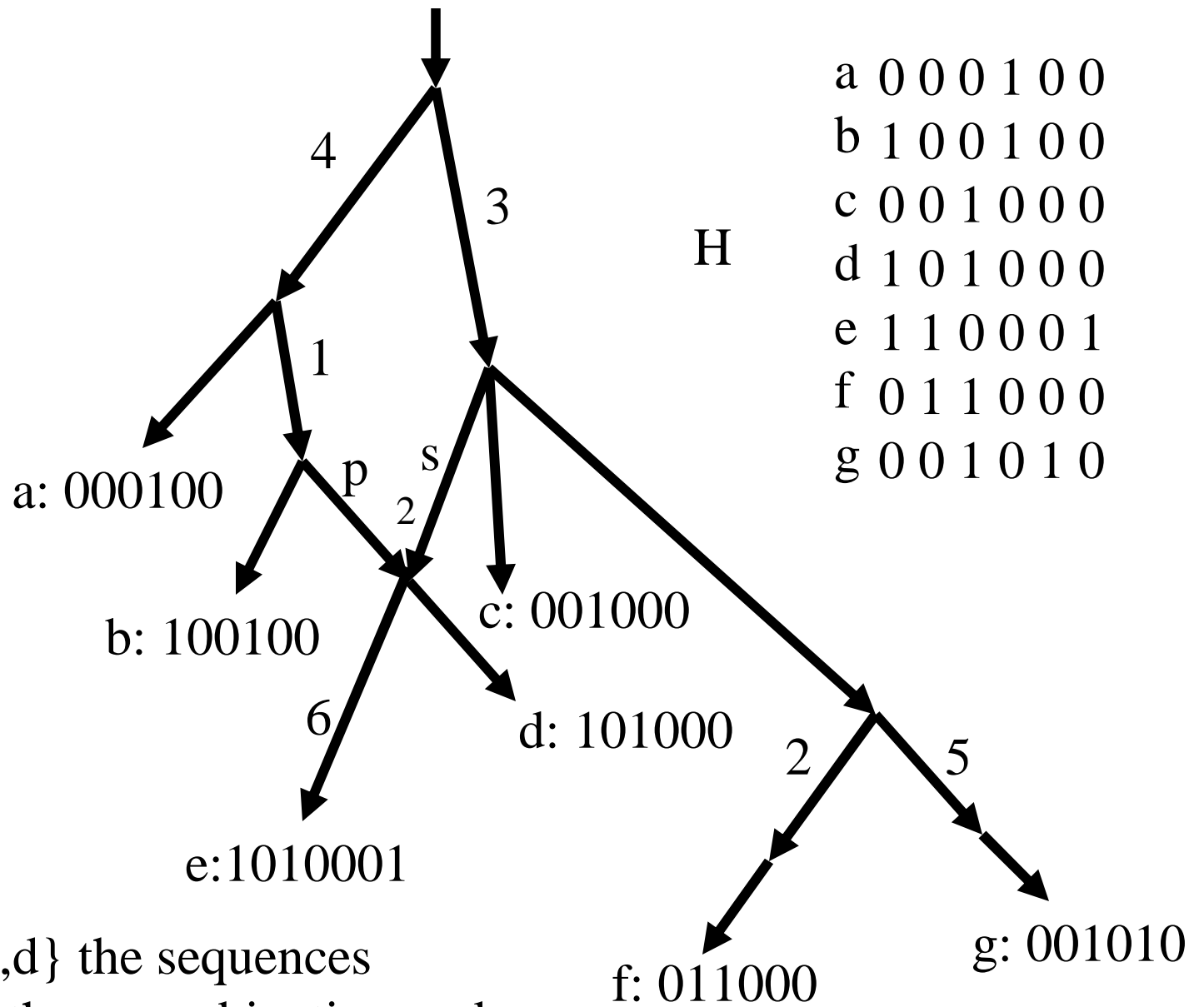


# Galled-Tree Haplotyping

Problem: Given genotype matrix  $G$ , if there is no PPH solution for  $G$ , is there a haplotyping  $H$  for  $G$  such that  $H$  can be derived on a Galled-Tree?

# A Necc. Condition for a one-gall tree

1. There exists a set of sequences  $S$  such that for every pair of incompatible sites  $p, q$ , a single  $p, q$  state-pair appears in all sequences in  $S$ , and does not appear in any sequence outside  $S$ .
2. There must be a number  $x$  such that  $p < x < q$ , for each incompatible pair  $p, q$ .



$S = \{e,d\}$  the sequences  
below the recombination node.

# Surprising Result - Yun Song

The necessary condition is also sufficient.

Yun S. Song in TCBB 2006

# Coming full circle - back to genotypes

When can a set of genotypes be explained by a set of haplotypes derived on a **galled-tree**, rather than on a perfect phylogeny?

The Song NASC can be translated into an ILP, using the part of the MinIncompat ILP that identifies which site pairs are incompatible.

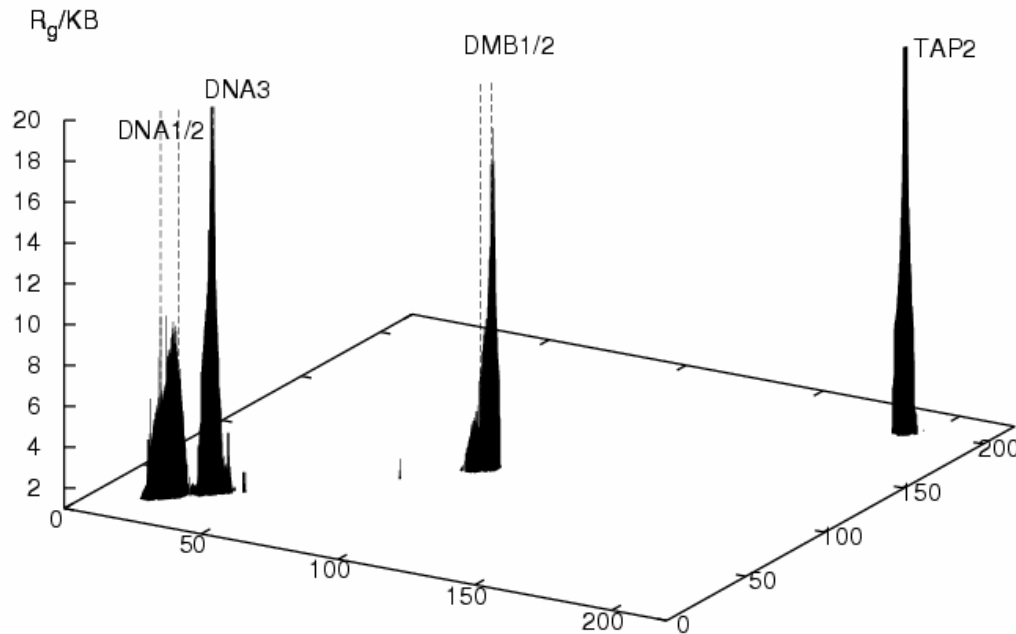
For the one gall problem, the formulation solves very efficiently (200 rows x 40 sites in seconds to minutes). So far, the 2-gall case does not solve well (ongoing work). (Dan Brown, Gusfield 2006).

# Haplotyping With Minimum Number of Recombinations

- Compute  $R_{min}(G)$  = the exact min number of recombinations used in any network generating any set of explaining haplotypes, for a set of genotypes  $G$
- NP-hard
- CSB 2006 (Wu, Gusfield) A branch and bound method computing exact  $R_{min}(G)$  for data with small number of sites
- APOE data: 47 non-trivial genotypes, 9 sites
  - Our method: 2 minutes,  $R_{min}(G) = 5$

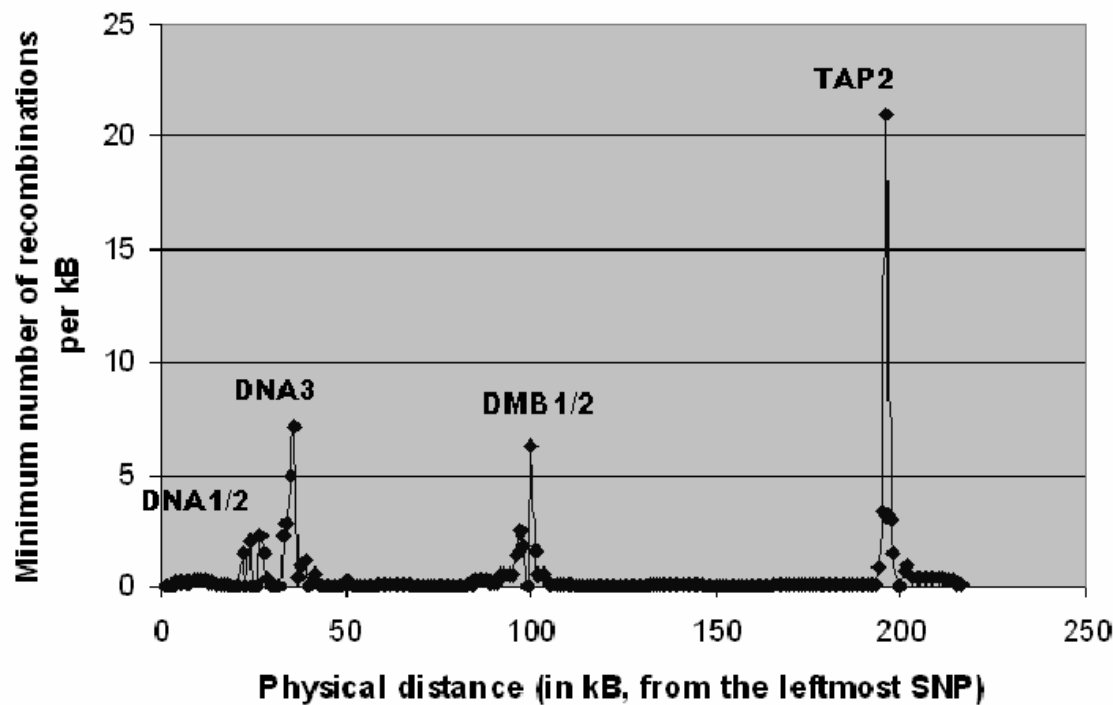
# Application: Recombination Hotspot

- Recombination **hotspot**: regions where recombination rate is much higher than neighboring regions
- Previous study (Bafna and Bansal, 2005): a recombination lower bound with inferred *haplotypes* were used to identify recombination hotspots
- Our work: compute the exact  $R_{\min}(G)$  with *genotypes* for a sliding window of a small number of SNPs to detect recombination hotspots



MS32 data (Jeffreys, et al. 2001)

Result from *haplotypes* (Bafna and Bansal, 2005)



Result from original *genotypes* (this paper)

# Other Applications

- Finding true *Rmin* from genotypes  $G$ 
  - Two stage approach: run PHAS to get an HI solution  $H$ , and compute  $Rmin(H)$
  - One stage approach: directly compute  $Rmin(G)$
- Accuracy of haplotype inference on a minimum network
- Simulation results: comparable, slightly weaker and non-conclusive

Papers and  
Software on [www.csif.cs.ucdavis.edu/~gusfield](http://www.csif.cs.ucdavis.edu/~gusfield)

Possible 2007-8 Postdoc at UCD on SNP/HAP  
problems (biological, math/stat/, comp. sci.)