
How Accurate is Pure Parsimony Haplotype Inferencing?

Sharlee Climer

Department of Computer Science and Engineering

Department of Biology

Washington University in Saint Louis

sharlee@climer.us

www.climer.us

Joint work with Weixiong Zhang and Gerold Jaeger

Pure Parsimony

- Pure Parsimony Haplotype Inferencing (PPHI)
 - Find smallest set of unique haplotypes that can resolve a set of genotypes
- Suggested by Earl Hubbell in 2000
- Cast as an Integer Linear Program (IP) by Dan Gusfield [CPM'03]
- Great research interest

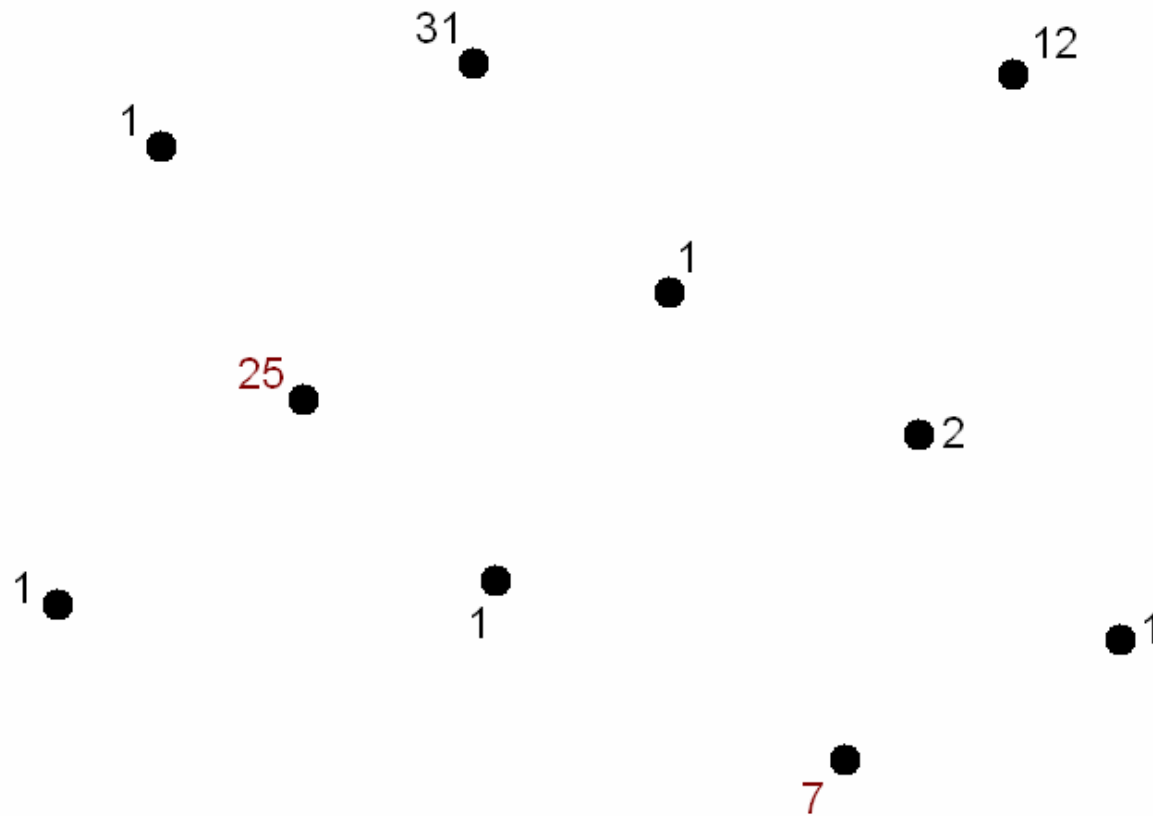
Overview

- Biological forces
- Haplotypes with low frequency
- Define haplotype classes
- Data sets
- Characteristics of real data

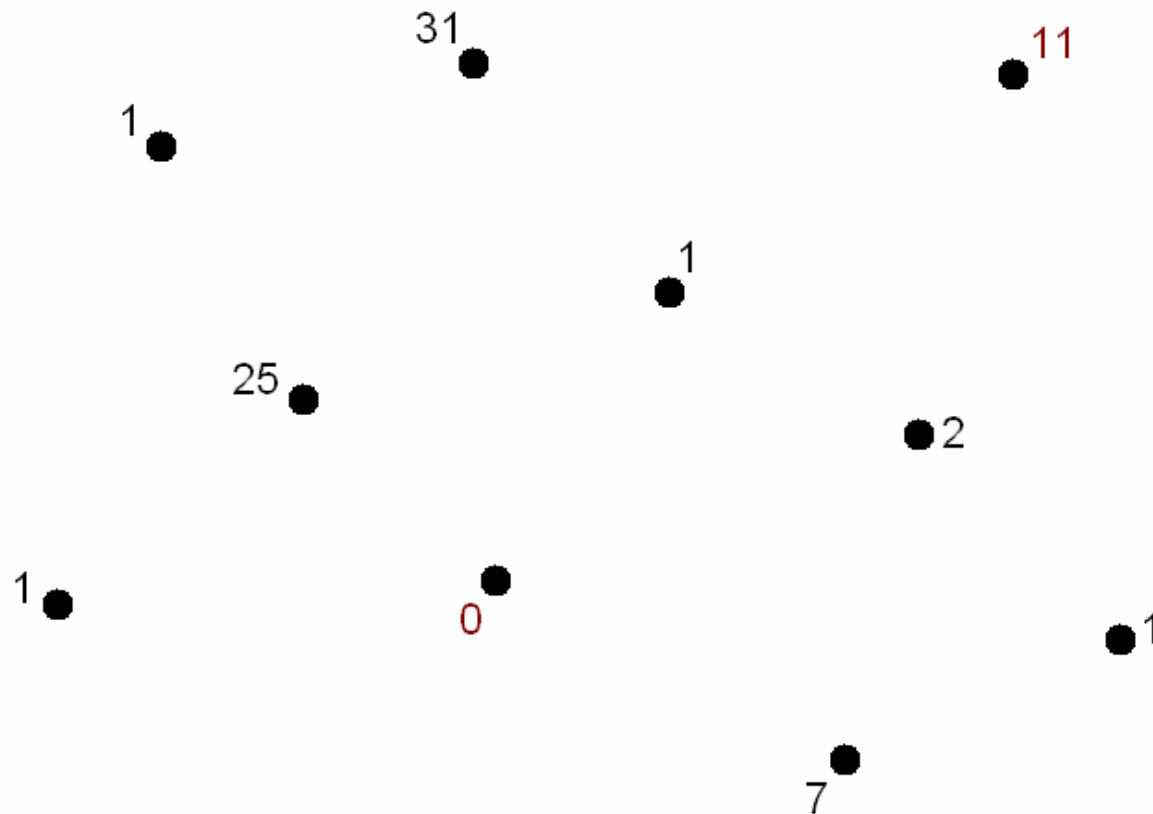
Biological forces



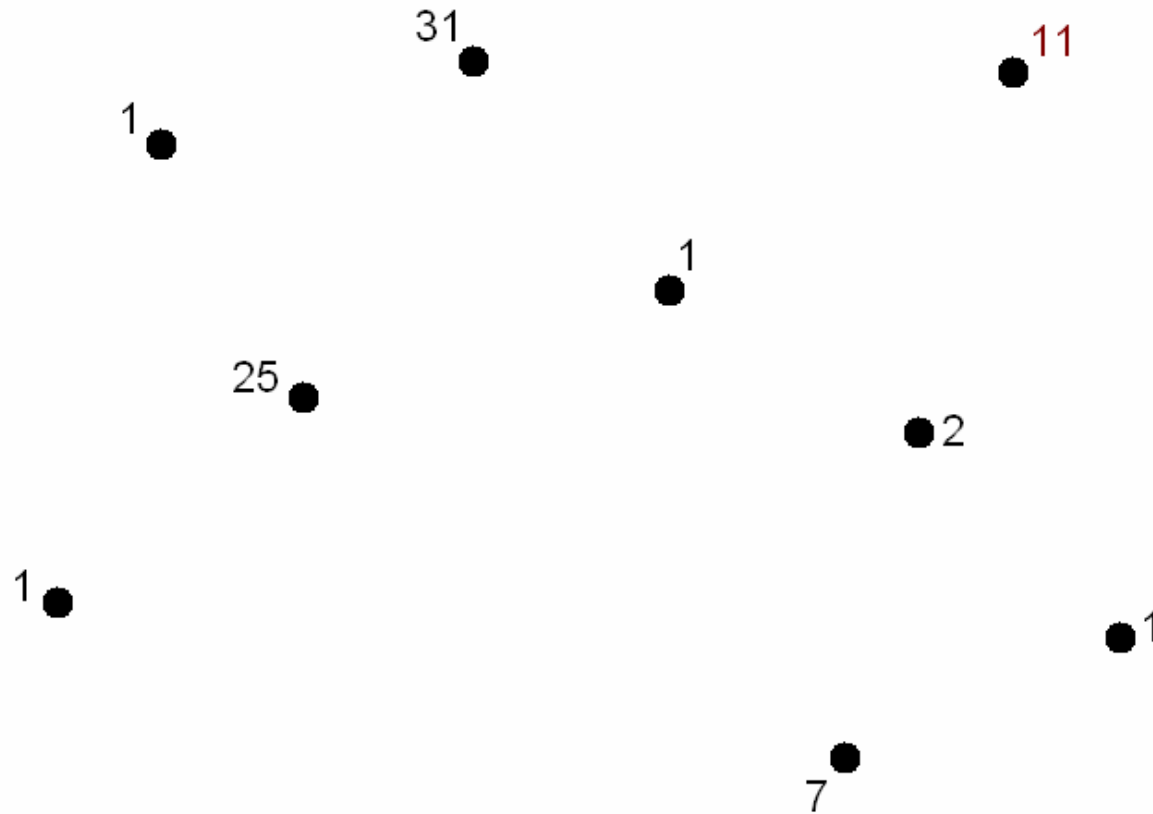
Biological forces



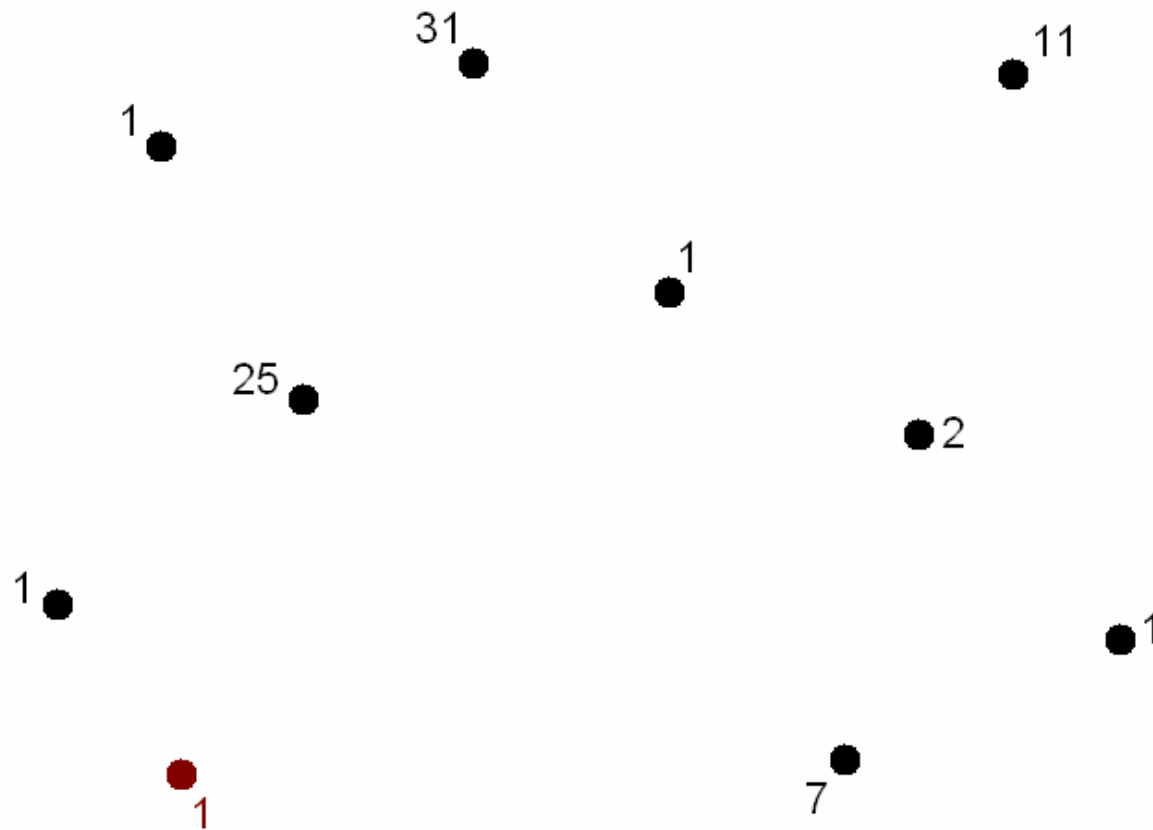
Biological forces



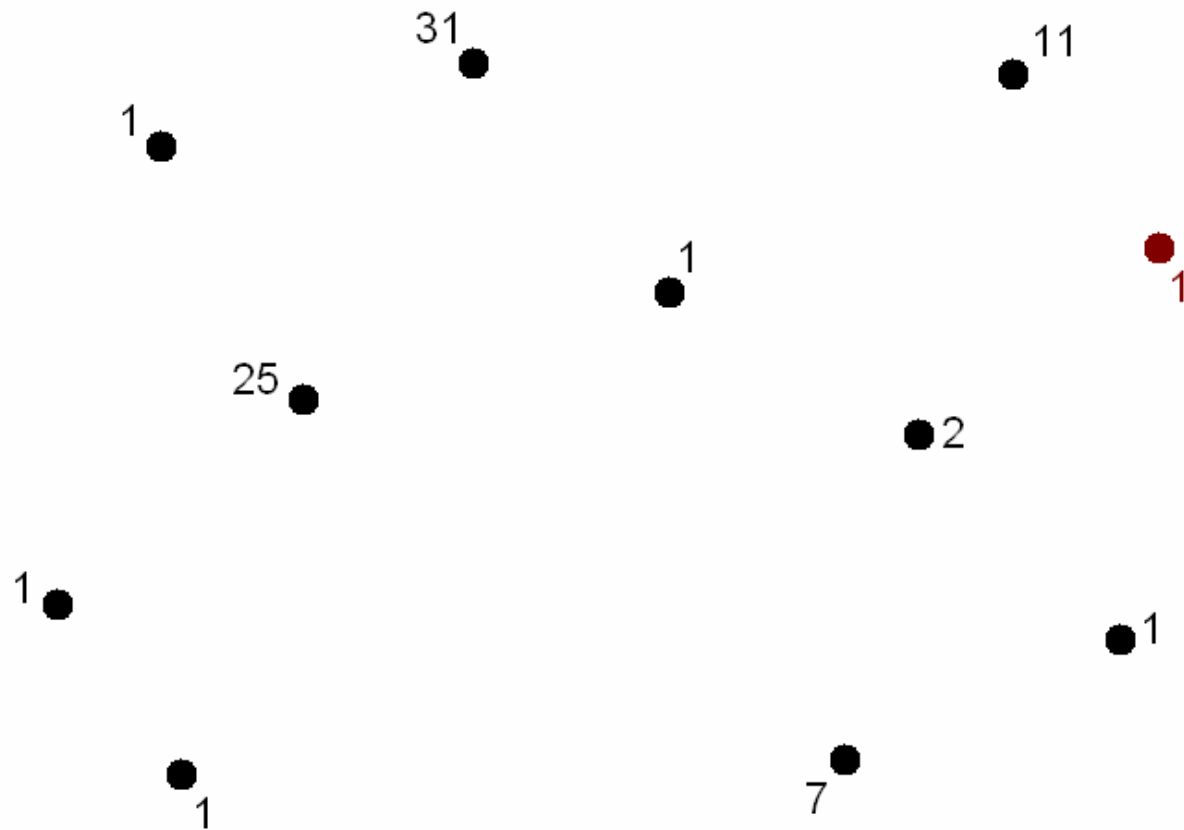
Biological forces



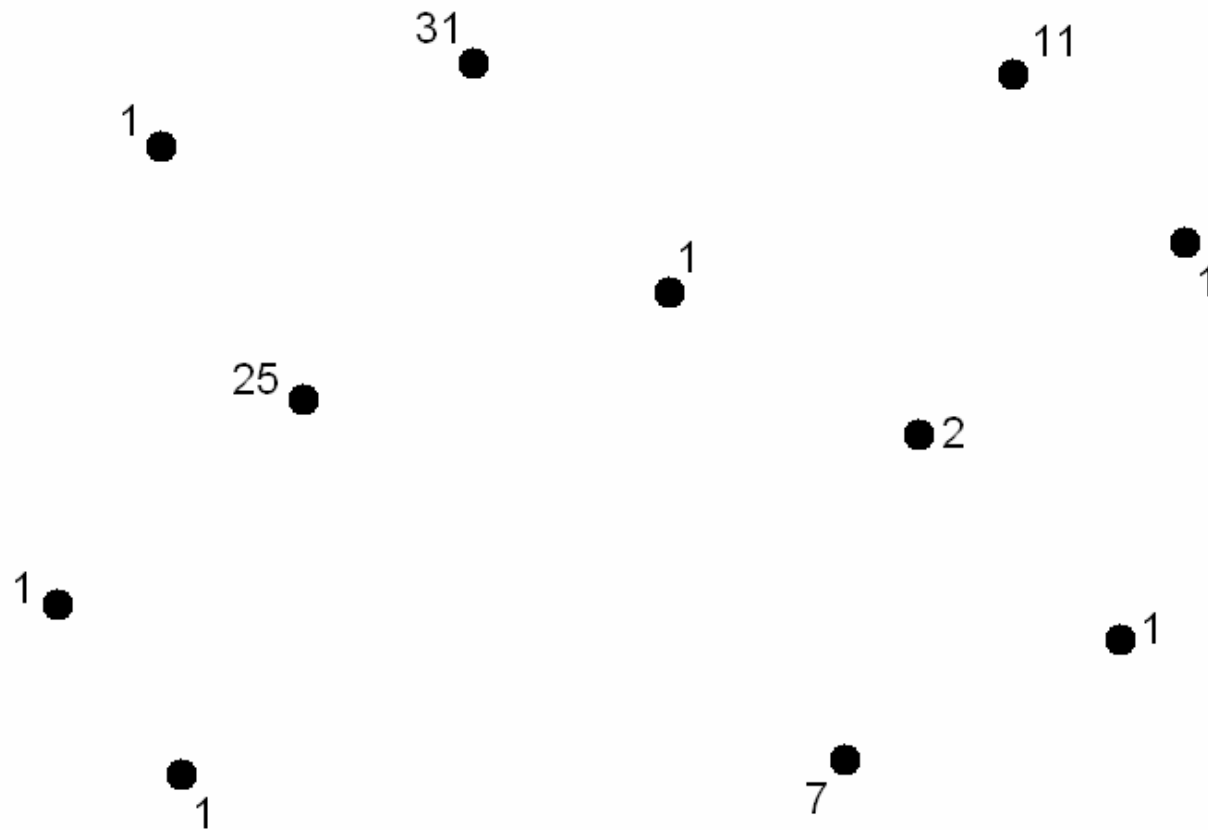
Biological forces



Biological forces



Biological forces



Biological forces

- Relatively few unique haplotypes
- Subset of haplotypes with low frequency
- Problems for PPHI
 - Large number of optimal solutions
 - True biological solution might not be parsimonious
- What are structural characteristics of optimal solutions?

Classes of haplotypes

- Set of possible haplotypes is exponentially large
- Partition similar to Traveling Salesman Problem
- Backbone haplotypes
 - Appear in every optimal solution
- Fat haplotypes
 - Do not appear in any optimal solution
- Fluid haplotypes
 - Appear in some, but not all, optimal solutions

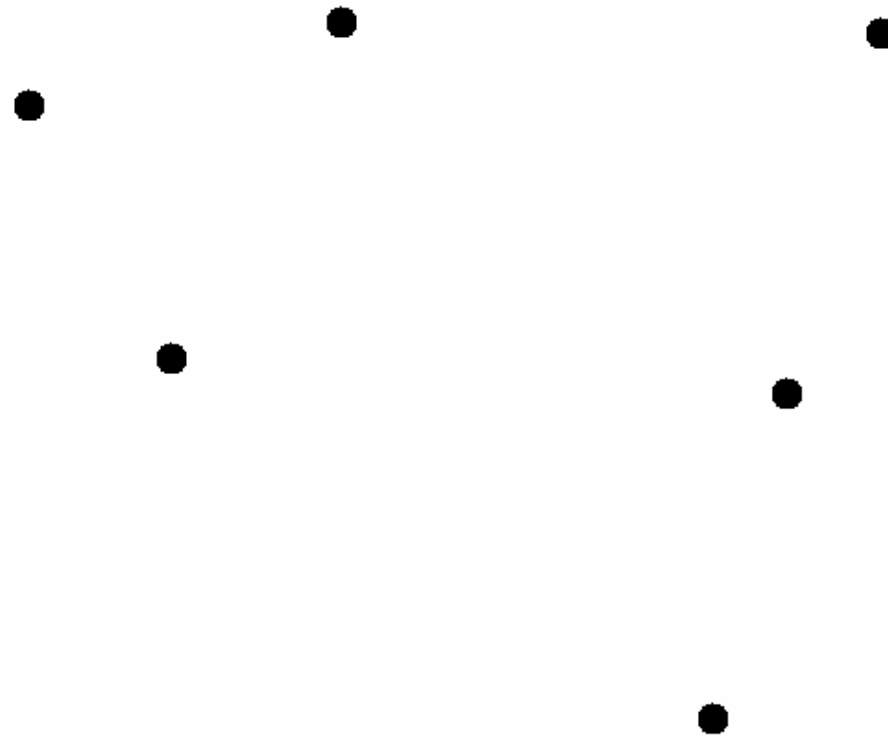
Backbone haplotypes

- Implicit backbones
 - All haplotypes that resolve unambiguous genotypes
- Explicit backbones
 - Can identify by solving at most one IP for each haplotype in solution that isn't implicit backbone

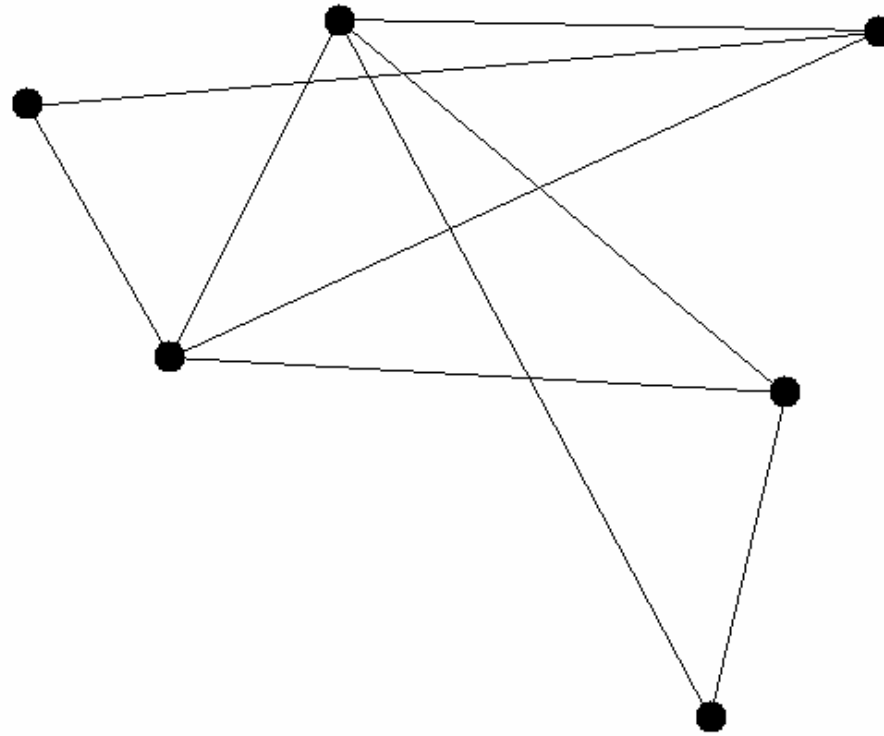
Backbone haplotypes

h3 h7 h15 h27 h39 h50 h55 h79 h91
bb bb bb bb

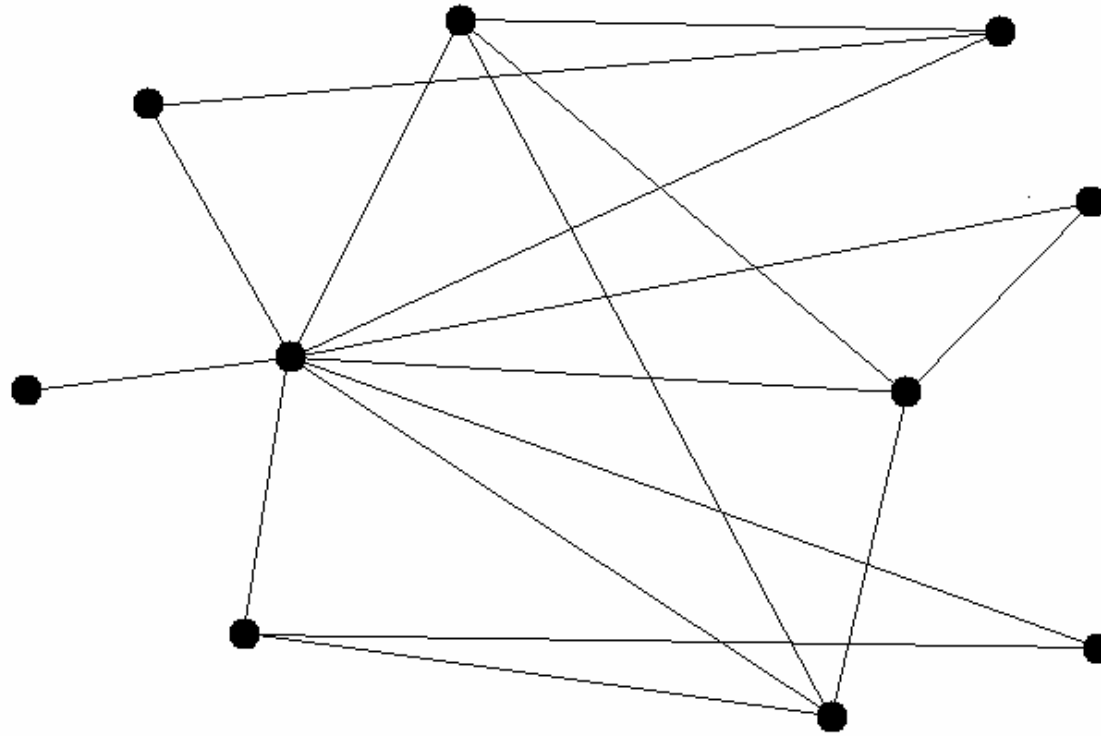
Backbone graph



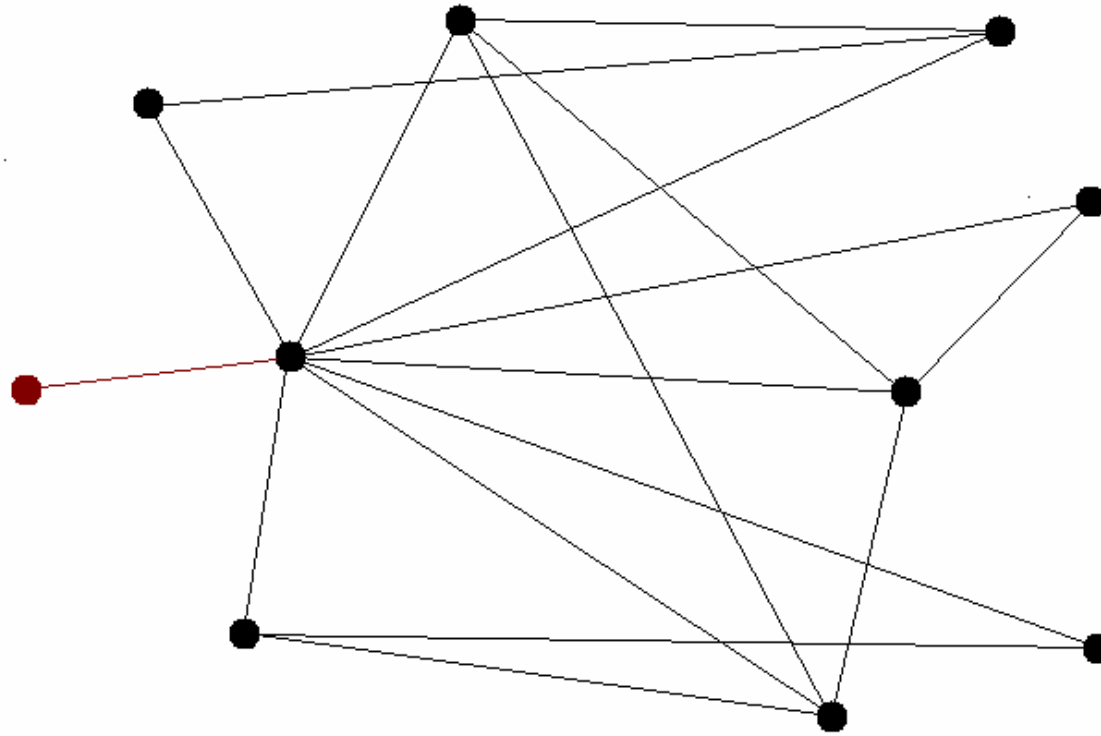
Backbone graph



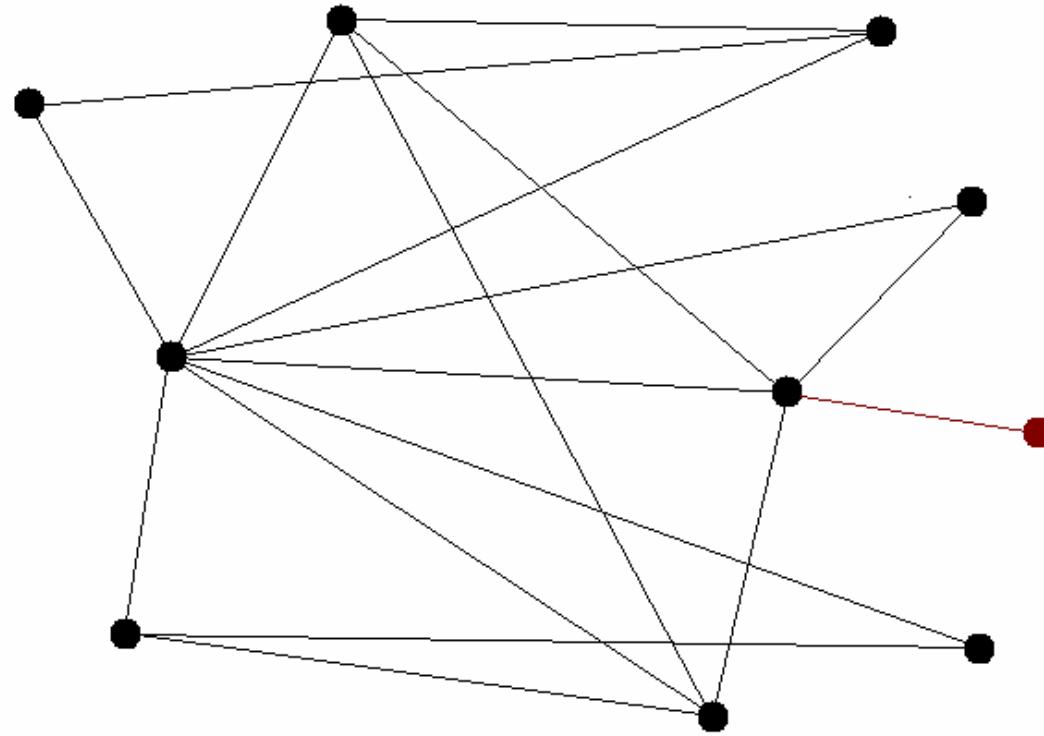
An optimal solution



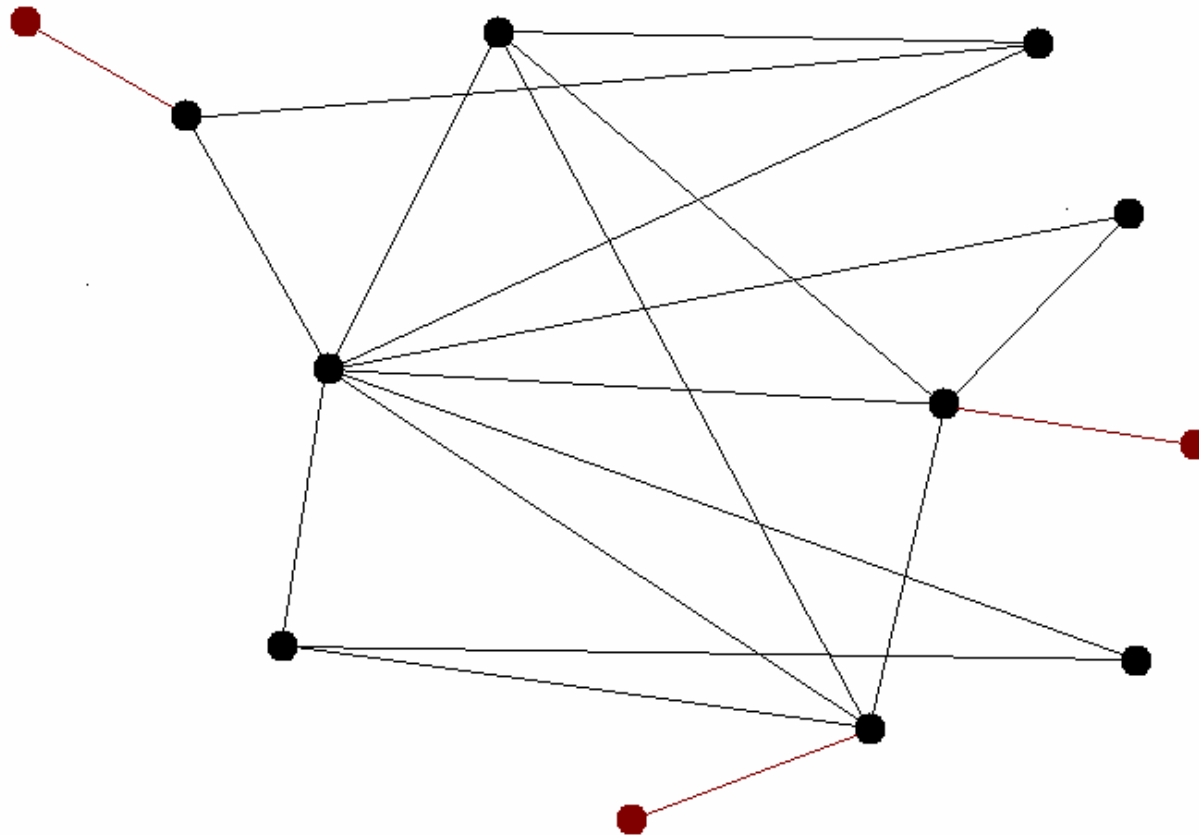
Low frequency haplotype



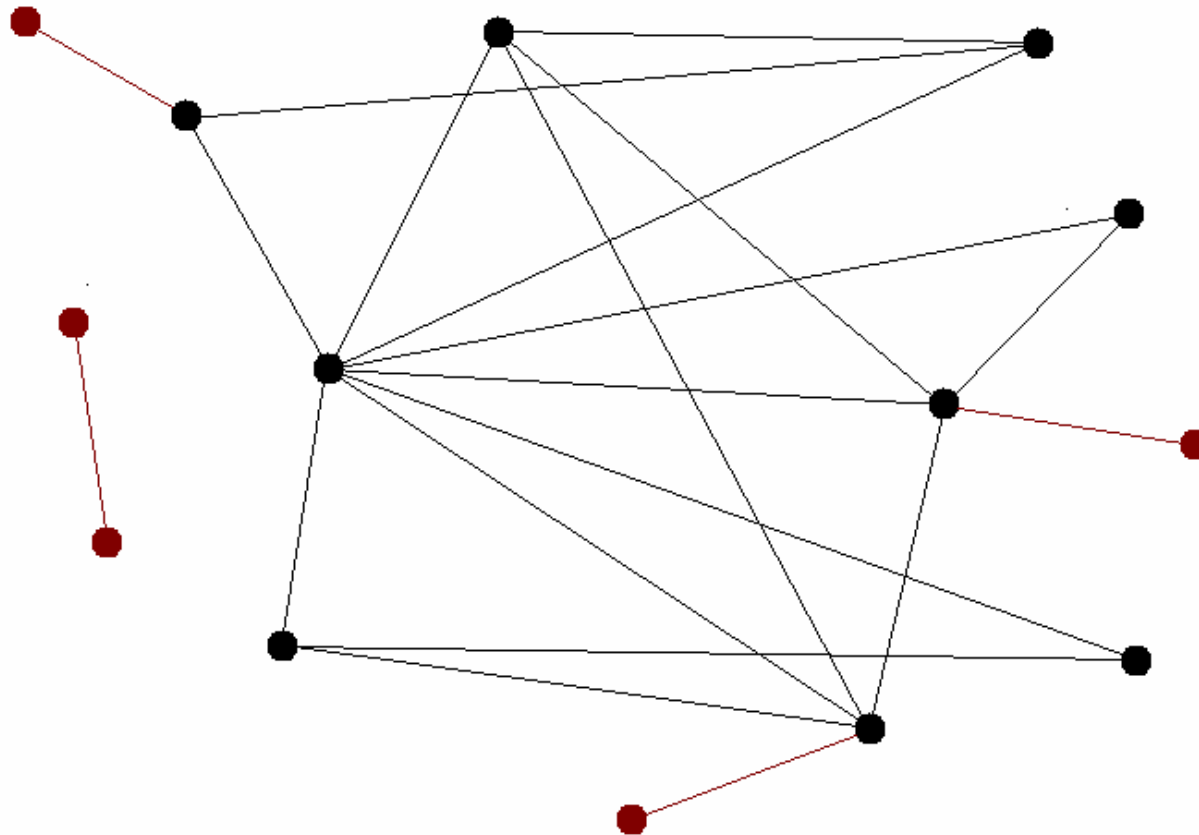
Low frequency haplotype



Low frequency haplotype



Low frequency haplotype



Data sets

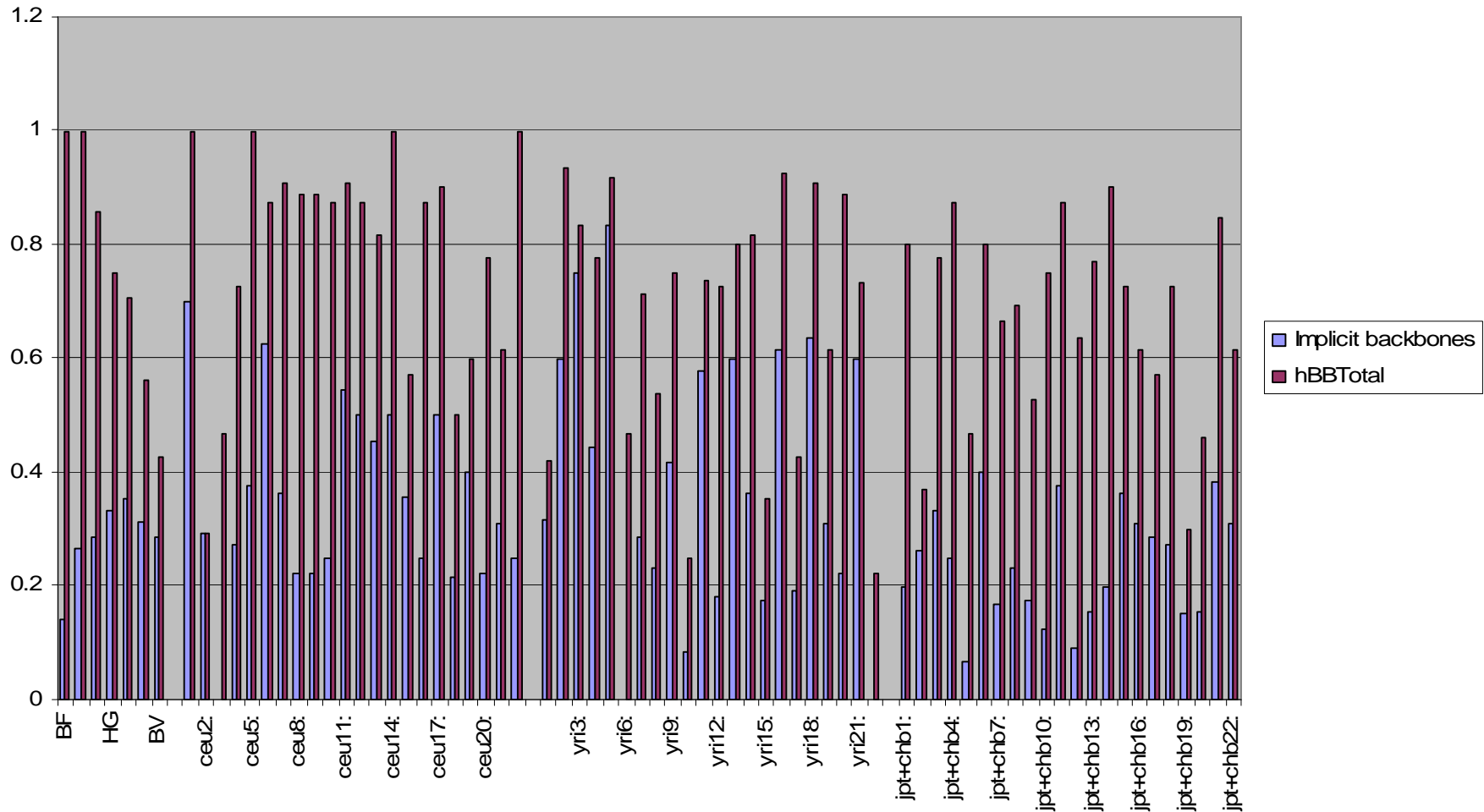
- 7 true haplotype data sets
 - Orzack et al. [*Genetics*, 2003]
 - 80 genotypes
 - 9 sites
 - ApoE
 - Andres et al. [*Genet. Epi.*, in press]
 - 6 sets of complete data
 - 39 genotypes
 - 5 to 47 sites
 - KLK13 and KLK14

Data sets

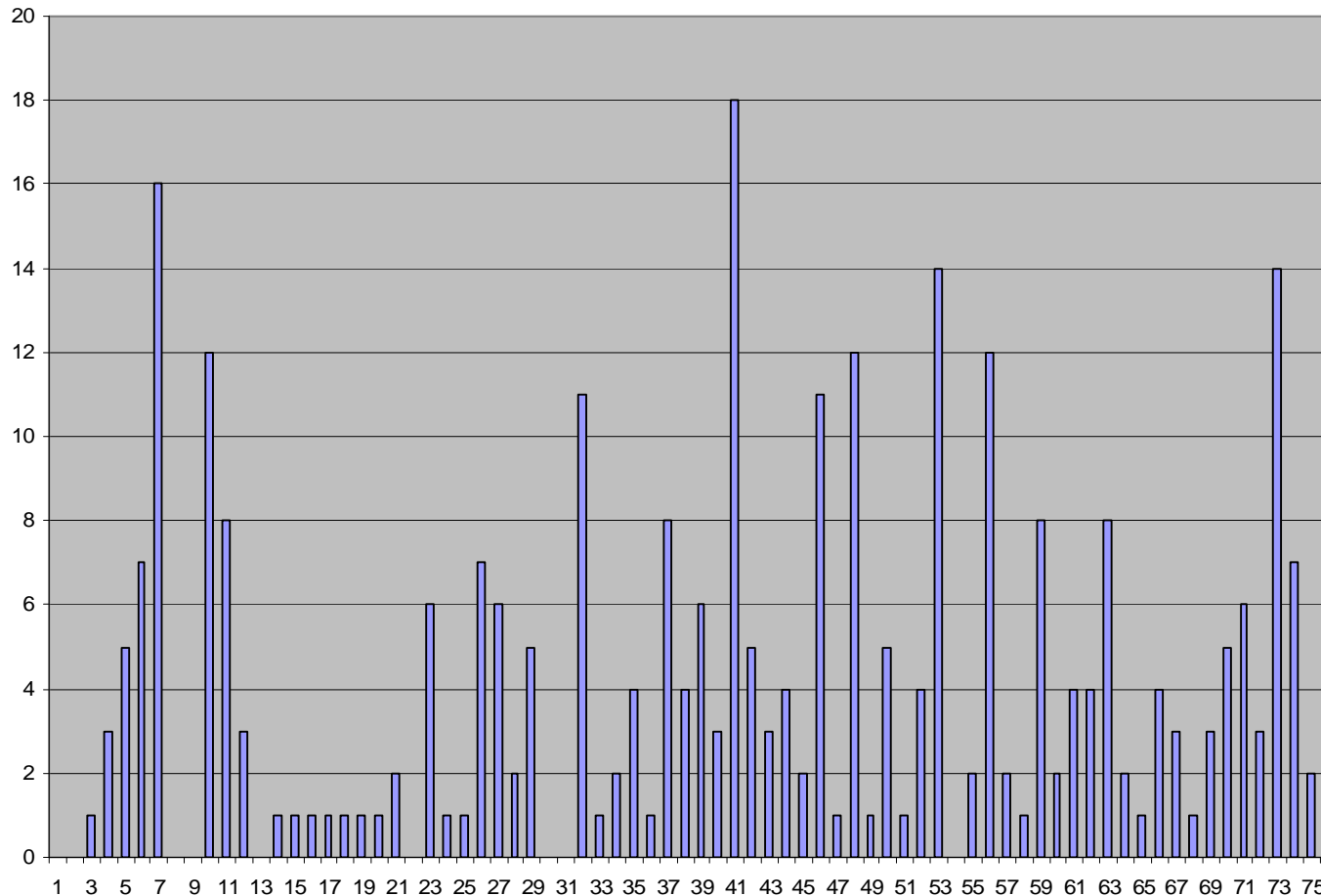
- HapMap data [*Nature* 2003, 2005]
 - Phase unknown
 - Random instance generator
 - 20 unique genotypes
 - 20 sites
 - Three populations
 - CEU
 - YRI
 - JPT+CHB
 - 22 chromosomes

Size of haplotype backbone

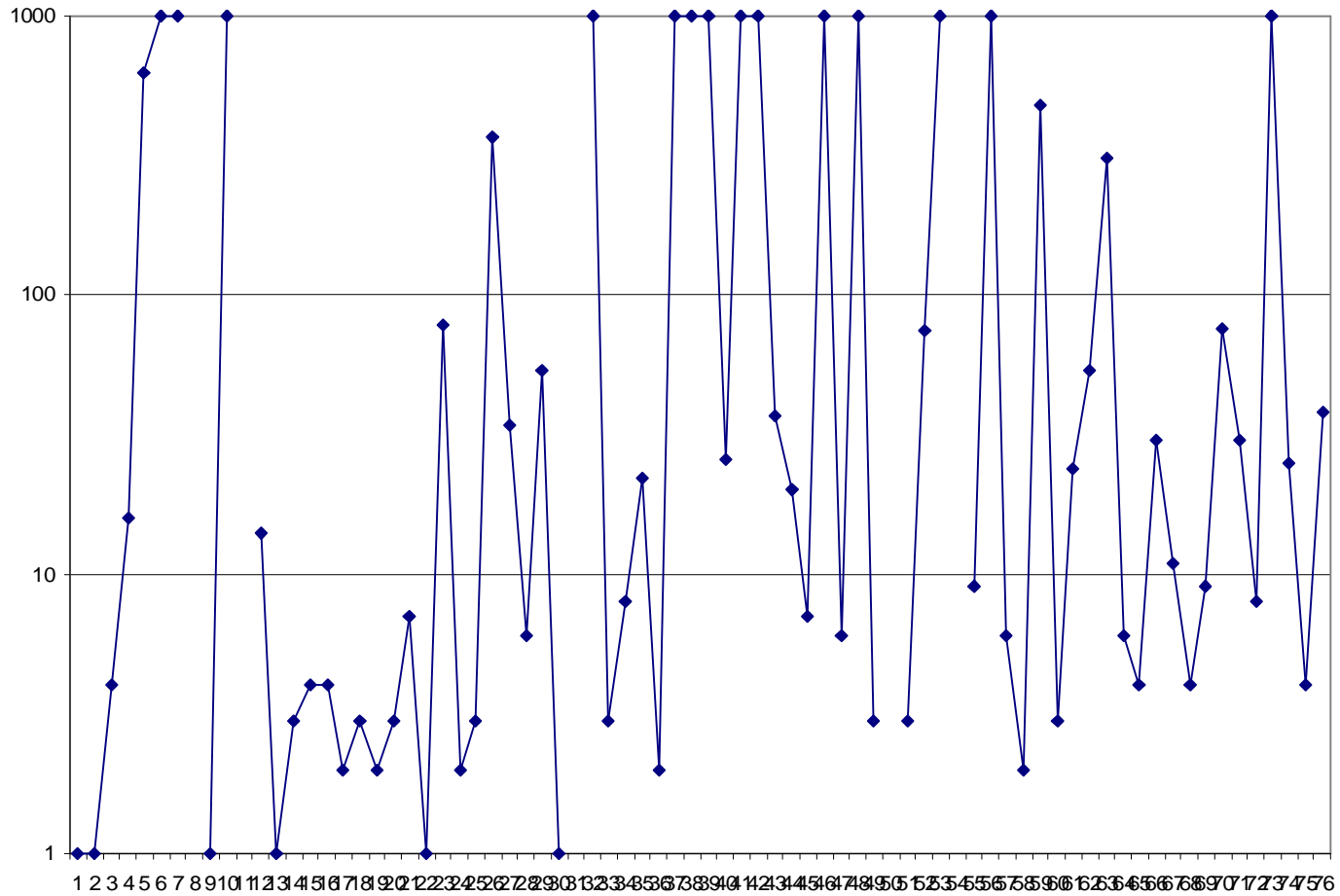
Percentage of haplotypes that are backbones



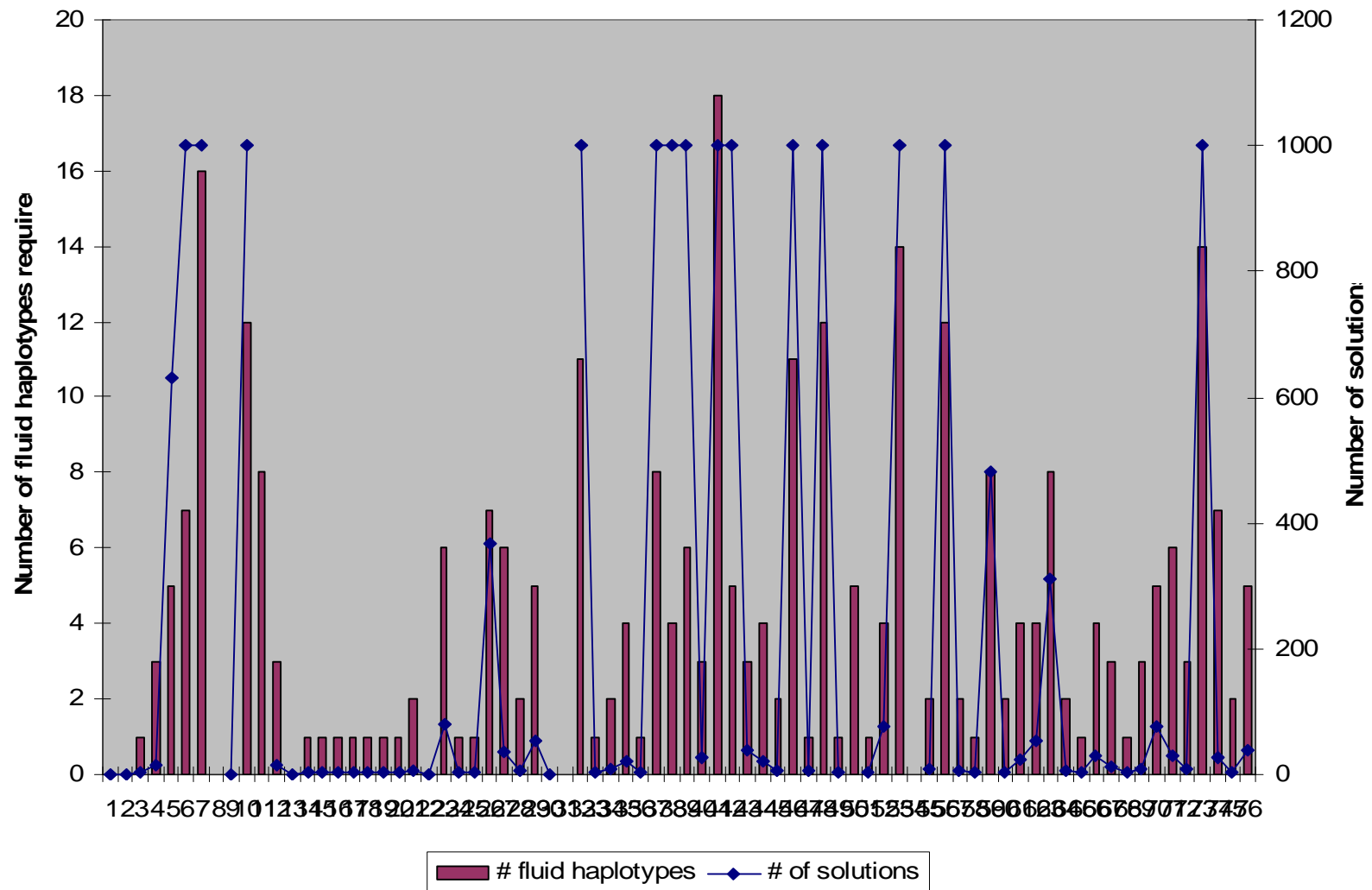
Number of fluid haplotypes in each solution



Number of optimal solutions



Number of fluid haplotypes and solutions



Biological correctness

Data set	# gen.	# sites	# BB hap.	#fluid hap.	# opt. sols.	Avg. distance to real
A	30	9	15	0	1	8
B	10	5	7	0	1	0
C	18	17	9	3	16	7.5
D	10	8	6	1	4	2.5
E	23	26	9	7	>1000	4.33
F	26	22	12	5	630	28.24
G	35	47	12	16	>1000	10.95

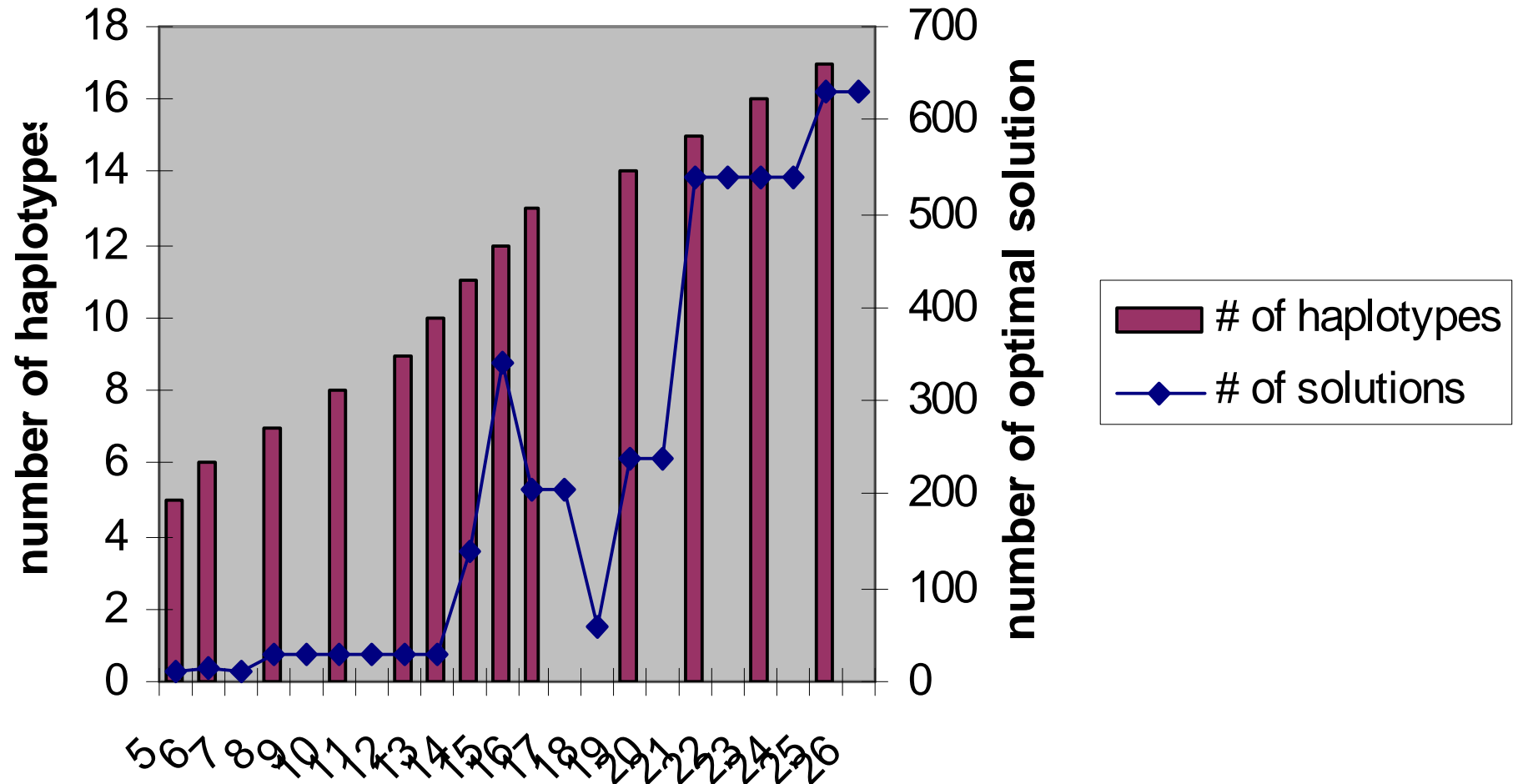
Biological correctness

Data set	Parsimony # of haplotypes	True # of haplotypes
A	15	17
B	7	7
C	12	12
D	7	7
E	16	16
F	17	18
G	28	32

Biological correctness

- Accuracy of backbone haplotypes
- Two data sets (F and G) had errors
 - One parsimony backbone haplotype not in real solution

Number of solutions vs. number of genotypes



Conclusions

- Biological forces tend to minimize cardinality, but also create low frequency haplotypes
- Low frequency in unique genotypes might not be low frequency in full set
- Low frequency haplotypes
 - Large number of optimal solutions
 - True solution not necessarily parsimonious
 - Combinatorial nature can lead to errors in backbones
- Parsimony combined with other biological clues