Guest lecture for course:
Computational Genetics (236608)

Linkage Disequilibrium Mapping and HaploBlock

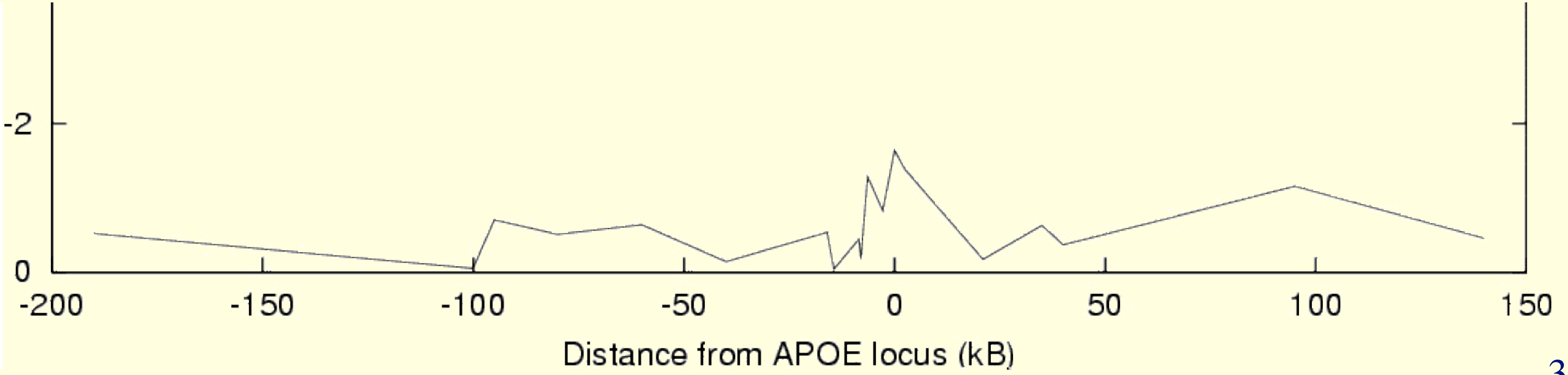
Gideon Greenspan
gdg@cs.technion.ac.il

Part 1: LD Mapping

- Basic LD Mapping
  - χ-squared test for individual SNPs
- Mapping with Haplotypes
  - Population phenomena
- Haplotyping
  - Clark algorithm
  - EM algorithm

Linkage Disequilibrium

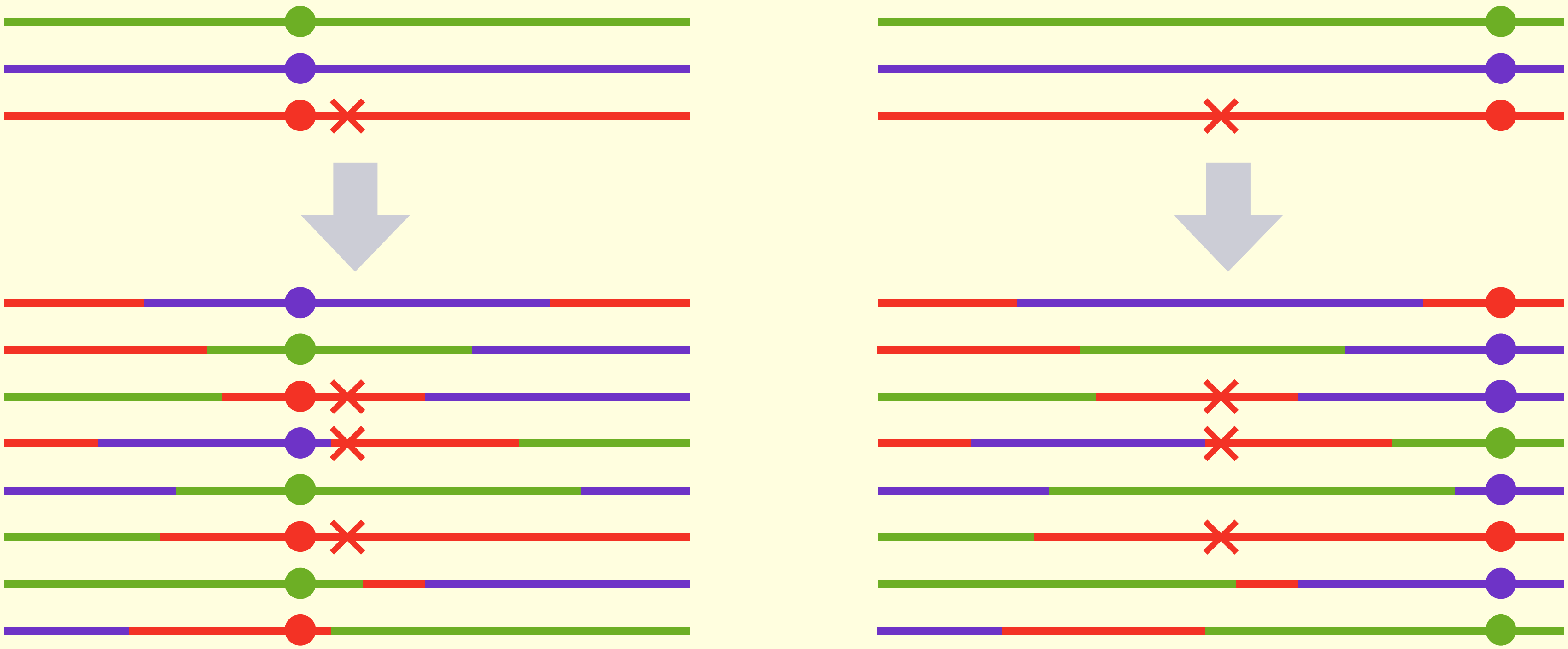
- LD = Another word for ‘correlation’
  - Correlation between markers in a population
- Random recombination destroys correlation
  - Close markers *may* have high LD
  - Above 1 Mb, LD disappears



LD Mapping: The Basics

- Take set of unrelated individuals
  - Ideally from a small, inbred population
- Measure markers at high resolution
  - Single Nucleotide Polymorphisms are ideal
- Test marker–disease correlations
  - Non-parametric disease model
  - Suitable (in theory) for low penetrance

LD Mapping in Action



	🟢	🟣	🔴
–	3	1	1
✕	0	1	2

	🟢	🟣	🔴
–	1	3	1
✕	1	1	1

Chi-Squared Test

Observed Counts

	Case	Control	Σ
A	69	236	305
a	31	264	295
Σ	100	500	600

Expected Counts

	Case	Control	Σ
A	50.83	254.17	305
a	49.17	245.83	295
Σ	100	500	600

$$\chi^2 = \sum \frac{(o-e)^2}{e} = 15.85$$

1 degree of freedom

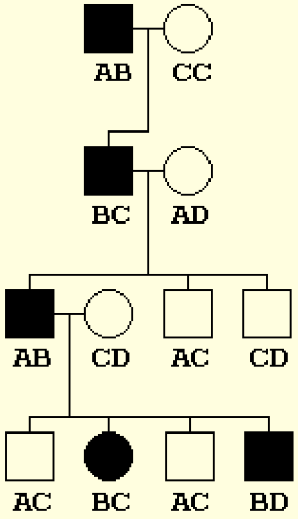
⇒ p-value = **0.0001**

SNPs

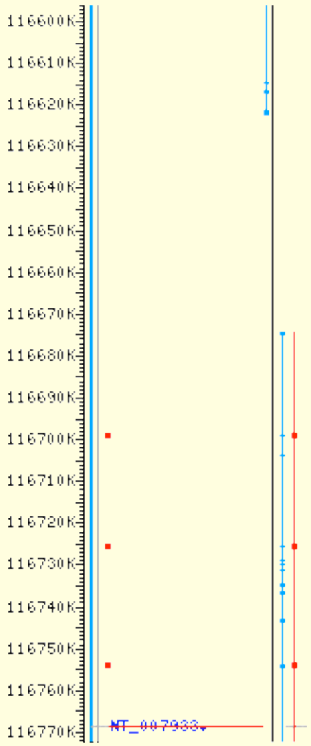
- Single base pair which exhibits variation
  - Caused by point mutations during meiosis
  - Variation almost always biallelic
- dbSNP contains ~ 4.3×10^6 SNPs
  - Over 1 SNP per 1,000 base pairs
  - About half with minor allele frequency > 20%
  - This number is still growing rapidly!

LD Mapping in Context

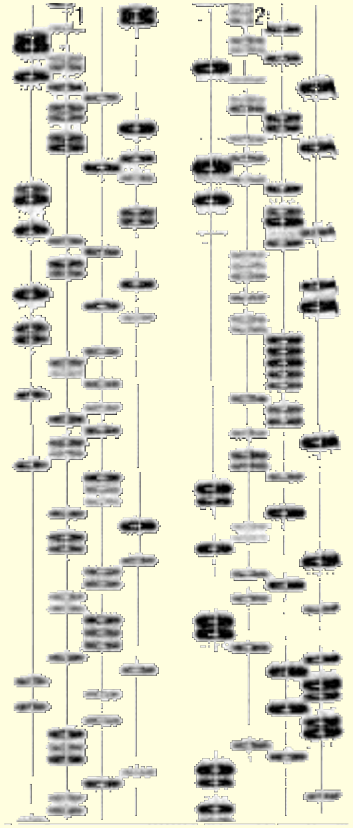
Identify chromosome
(10^8 bp)



Linkage analysis
(10^6~10^7 bp)



Identify genes
(10^5~10^6 bp)

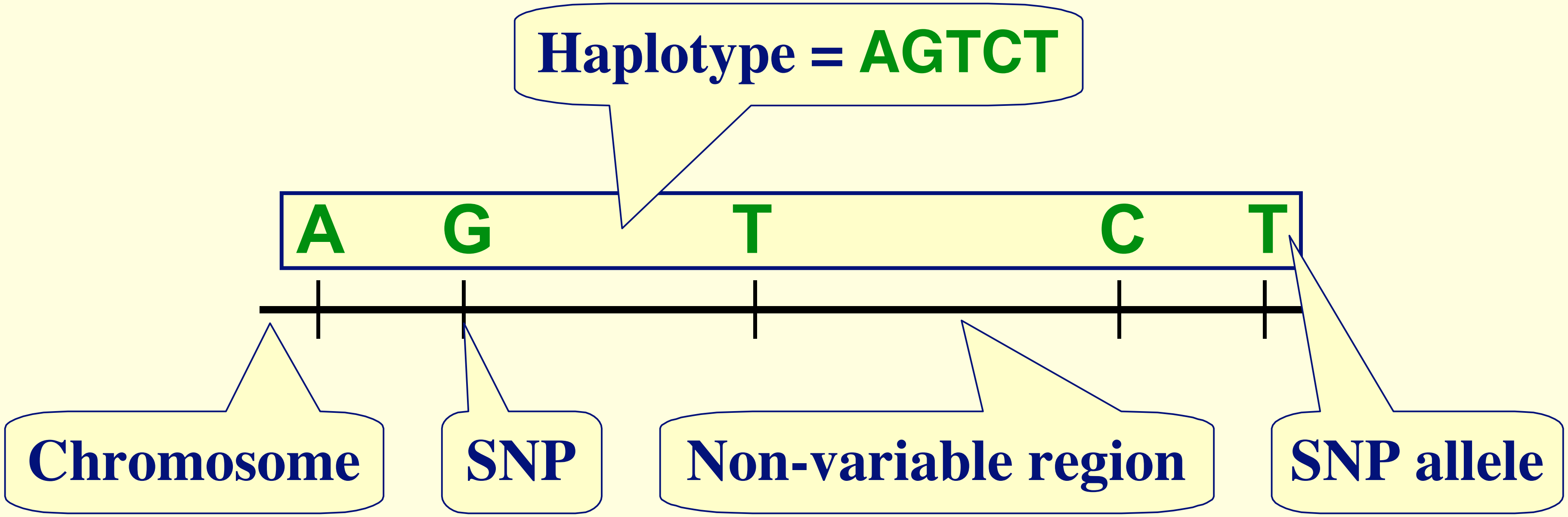


Resequencing
(10^0 bp)

False Positives

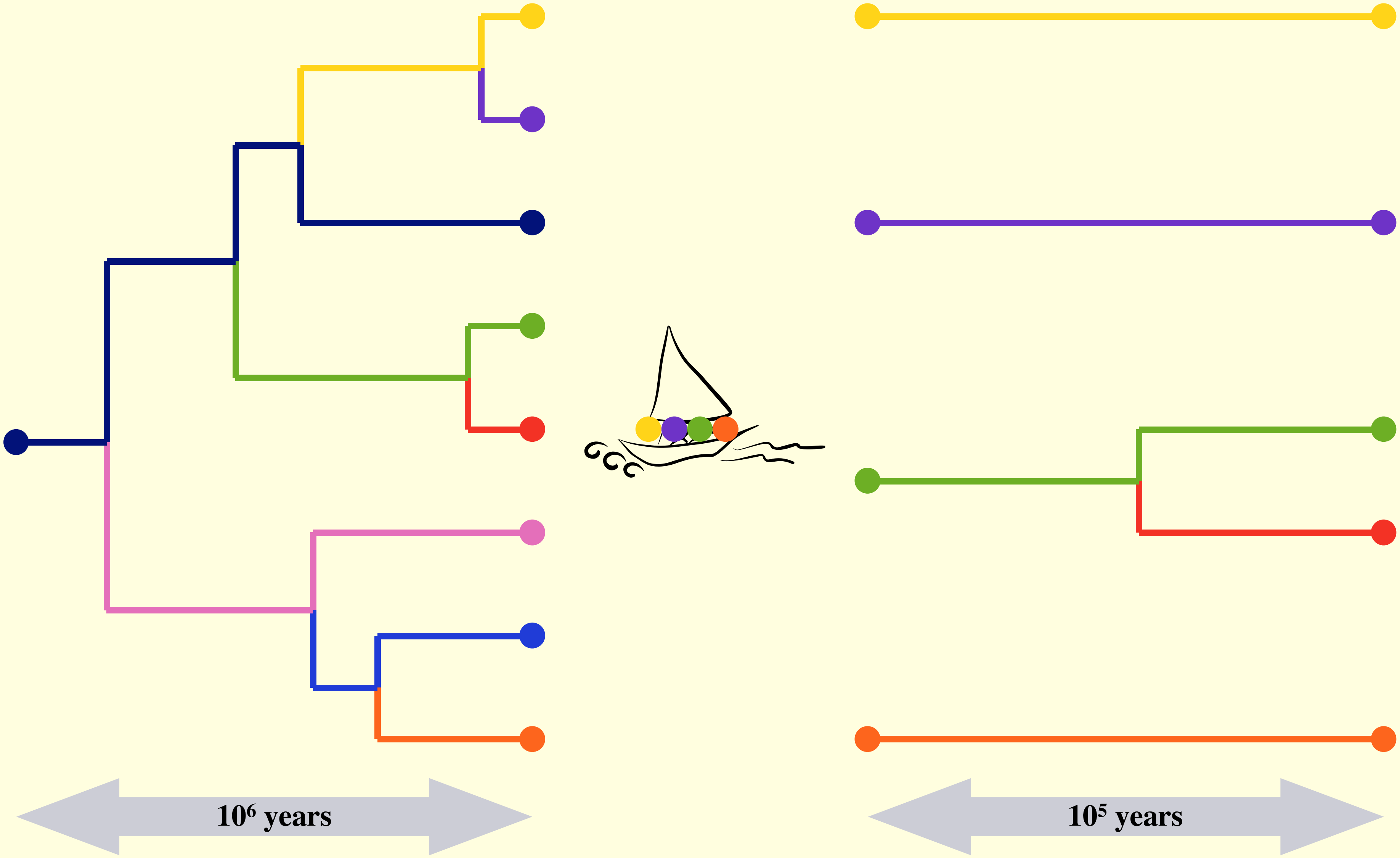
- Causes of spurious LD
  - Population structure
    - Migration and admixture
    - Preferential mating
  - Phenotypic site interaction
    - Disease epistasis
- Key problem: too many SNP tests
  - Bonferroni correction

Haplotypes

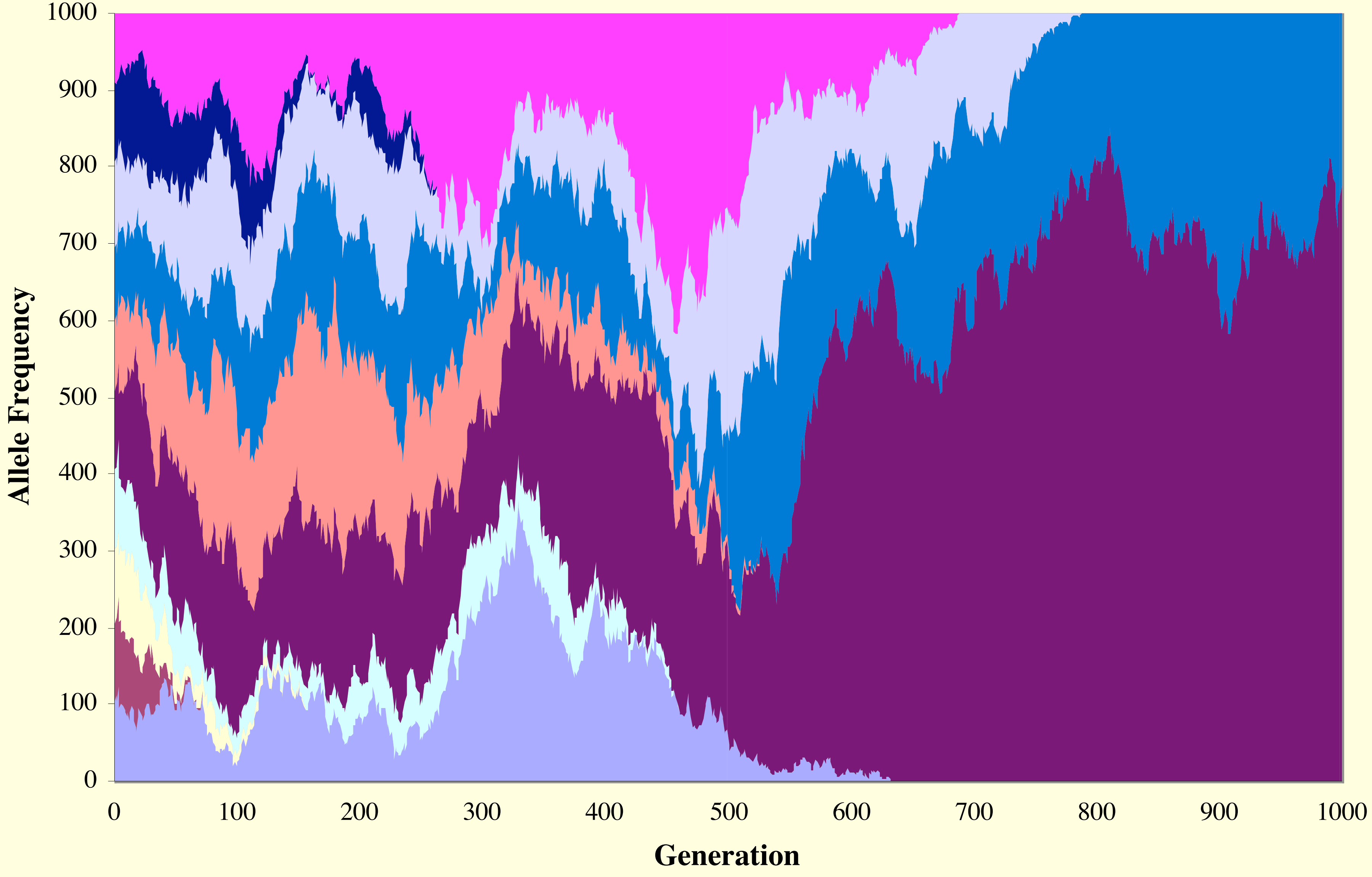


Generally, only a few of the 2^{loci} possible haplotypes cover >90% of a population, due to bottleneck effects and genetic drift.

Bottleneck Effects



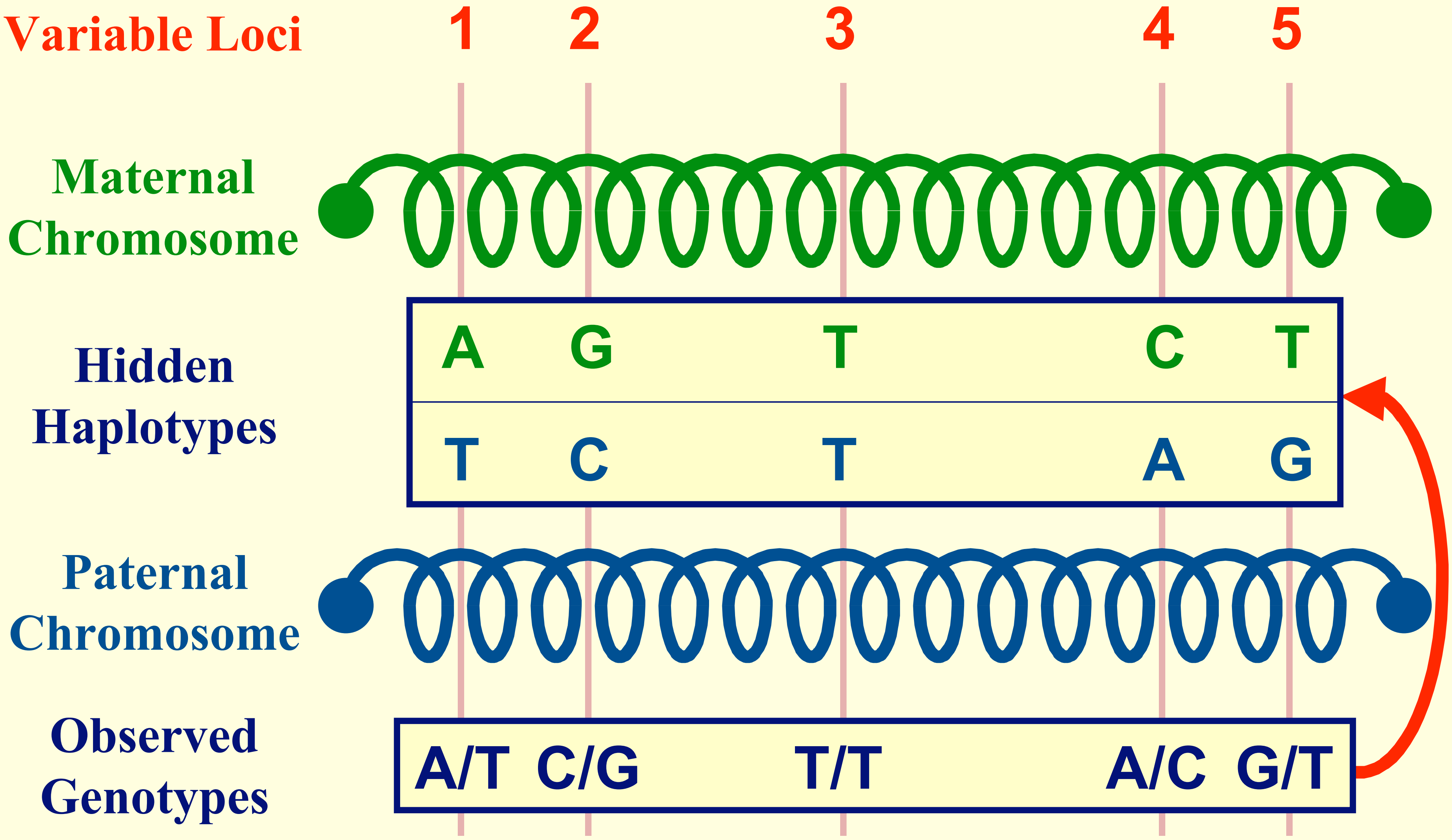
Genetic Drift



LD Mapping with Haplotypes

- Obtain haplotypes for a genomic region
  - Treat haplotype as correlated allele
- Advantage: fewer tests
  - Reduced false positive rate
- Disadvantage: ignores recombination
  - Different haplotypes could contain target
- Best: consider partial haplotypes…

The Haplotyping Problem



Variable Loci 1 2 3 4 5

Maternal Chromosome

Hidden Haplotypes

A G T C T

T C T A G

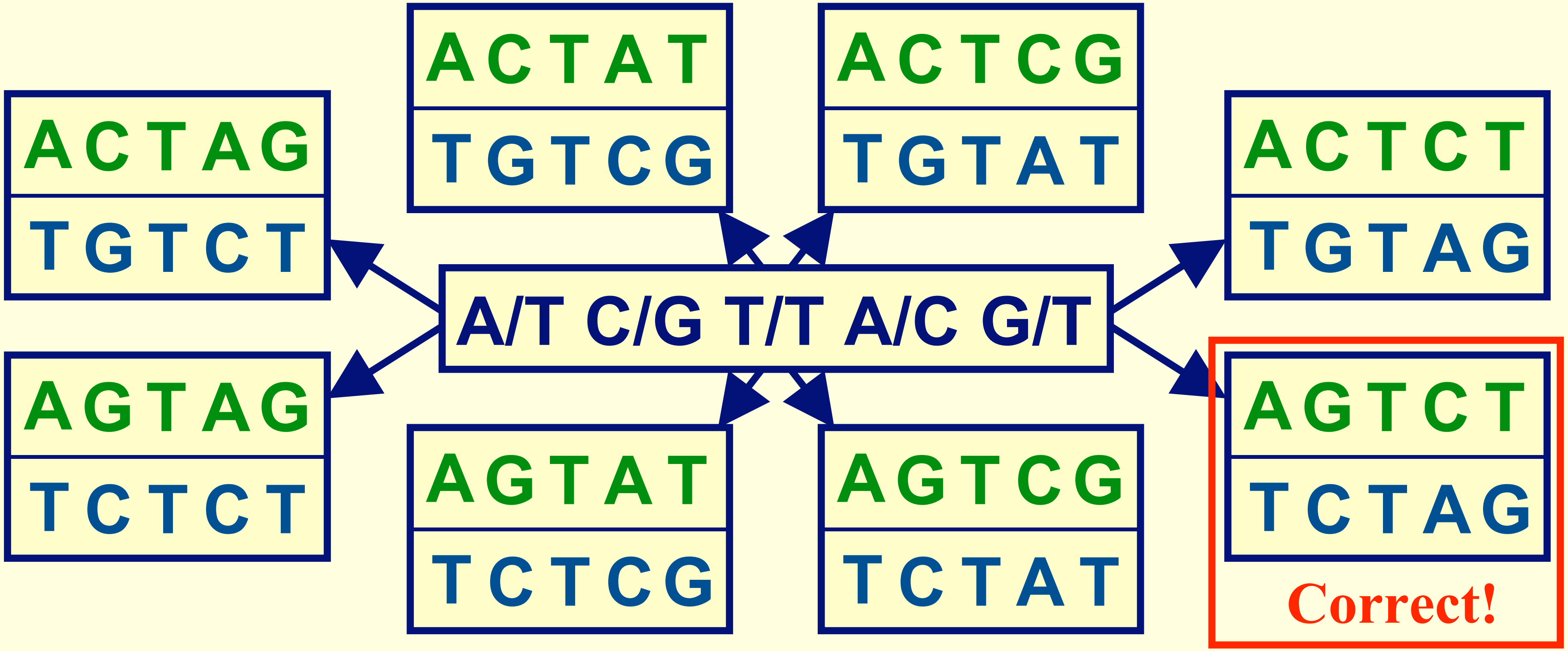
Paternal Chromosome

Observed Genotypes

A/T C/G T/T A/C G/T

Why is it hard?

- A series of joint measurements containing *h* heterozygous loci can be divided 2^{h-1} ways (we don’t care which is maternal or paternal).



A/T C/G T/T A/C G/T

ACTAG
TGTCT

ACTAT
TGTCG

ACTCG
TGTAT

ACTCT
TGTAG

AGTAG
TCTCT

AGTAT
TCTCG

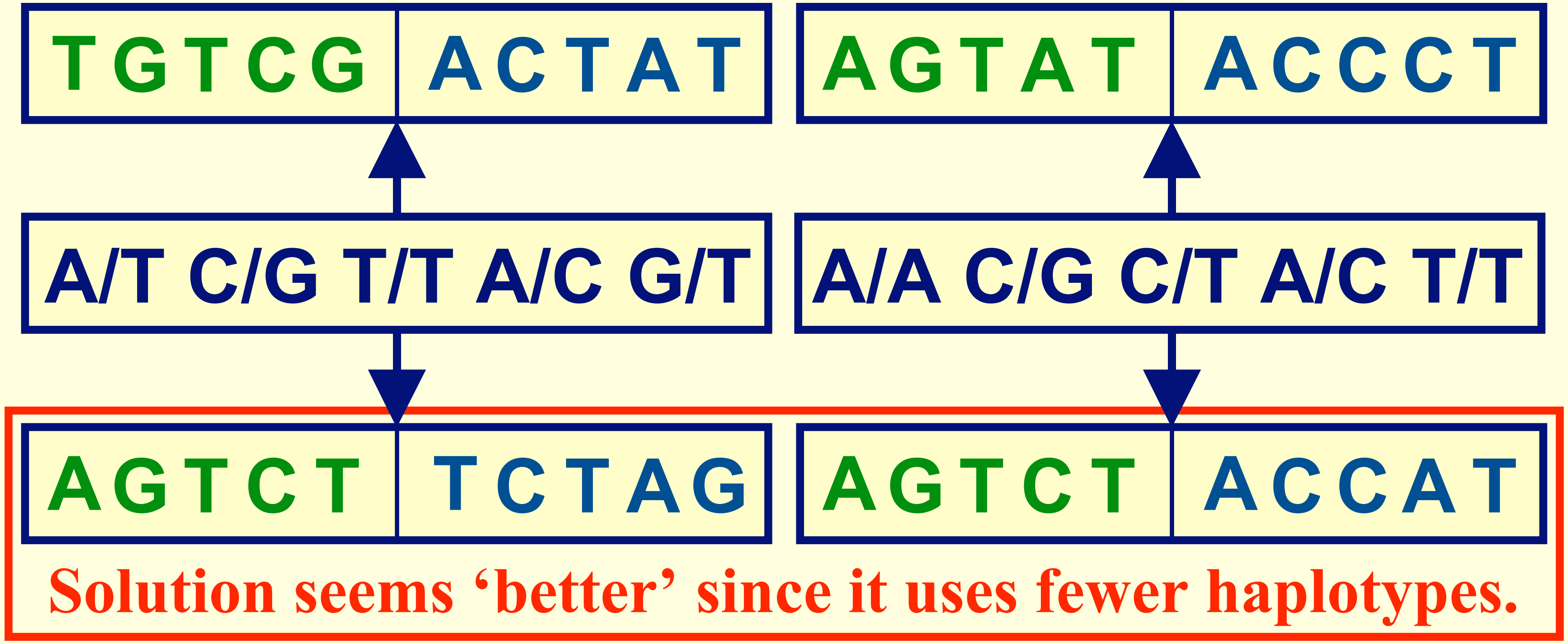
AGTCG
TCTAT

AGTCT
TCTAG

Correct!

Why is it approachable?

- Many of the haplotypes appear many times.
- Data for many individuals allows inference.



TGTCG ACTAT

AGTAT ACCCT

A/T C/G T/T A/C G/T

A/A C/G C/T A/C T/T

AGTCT TCTAG

AGTCT ACCAT

Solution seems 'better' since it uses fewer haplotypes.

Formalization 1

- Assume all loci biallelic (realistic).
- Individuals numbered $1\ldots n$
- Loci numbered $1\ldots l$
- Possible alleles $B=\{0,1\}$
- Possible haplotypes $H= B^l$
- Possible locus observations $L=\{[B,B]\}$
- Possible genotypes $G=L^l$
- Possible haplotype pairs $D=\{[H,H]\}$

Formalization 2

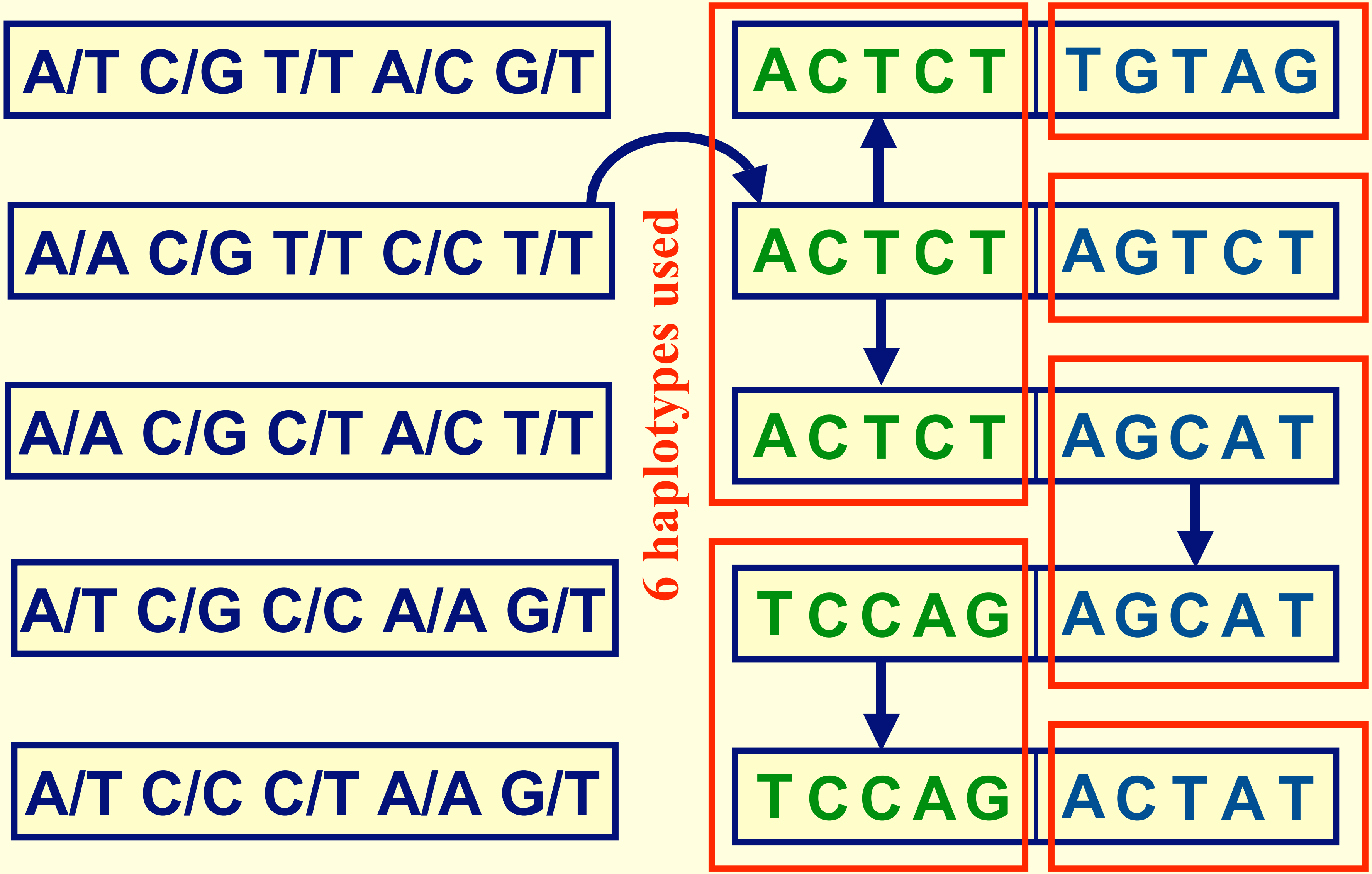
- Given a true haplotype pair $[h_1,h_2] \in D$, $\mathrm{G}(h_1,h_2) \in G$ is the genotype observed.
- Given an observed genotype $g \in G$, $\mathrm{D}(g) \subseteq D$ is set of possible haplotype pairs.

- Problem input: $(g_1,\ldots,g_n)$ where $g_i \in G$
- Problem output: $(d_1,\ldots,d_n)$ where $d_i \in \mathrm{D}(g_i)$

Clark’s Algorithm

1. Initialize set S to {}.
2. For genotypes g_i with a single possibility $[h_1,h_2]$ assign $d_i=[h_1,h_2]$ and add h_1,h_2 to S.
3. For genotypes g_i with a possibility containing a member $h_1 \in$ S and another haplotype h_2, assign $d_i=[h_1,h_2]$ and add h_2 to S.
4. Repeat step 3 until all haplotypes are assigned or we add nothing new to S.
5. Assign any remaining d_i arbitrarily.

Clark: Run



A/T C/G T/T A/C G/T ACTCT TGTAG

A/A C/G T/T C/C T/T ACTCT AGTCT

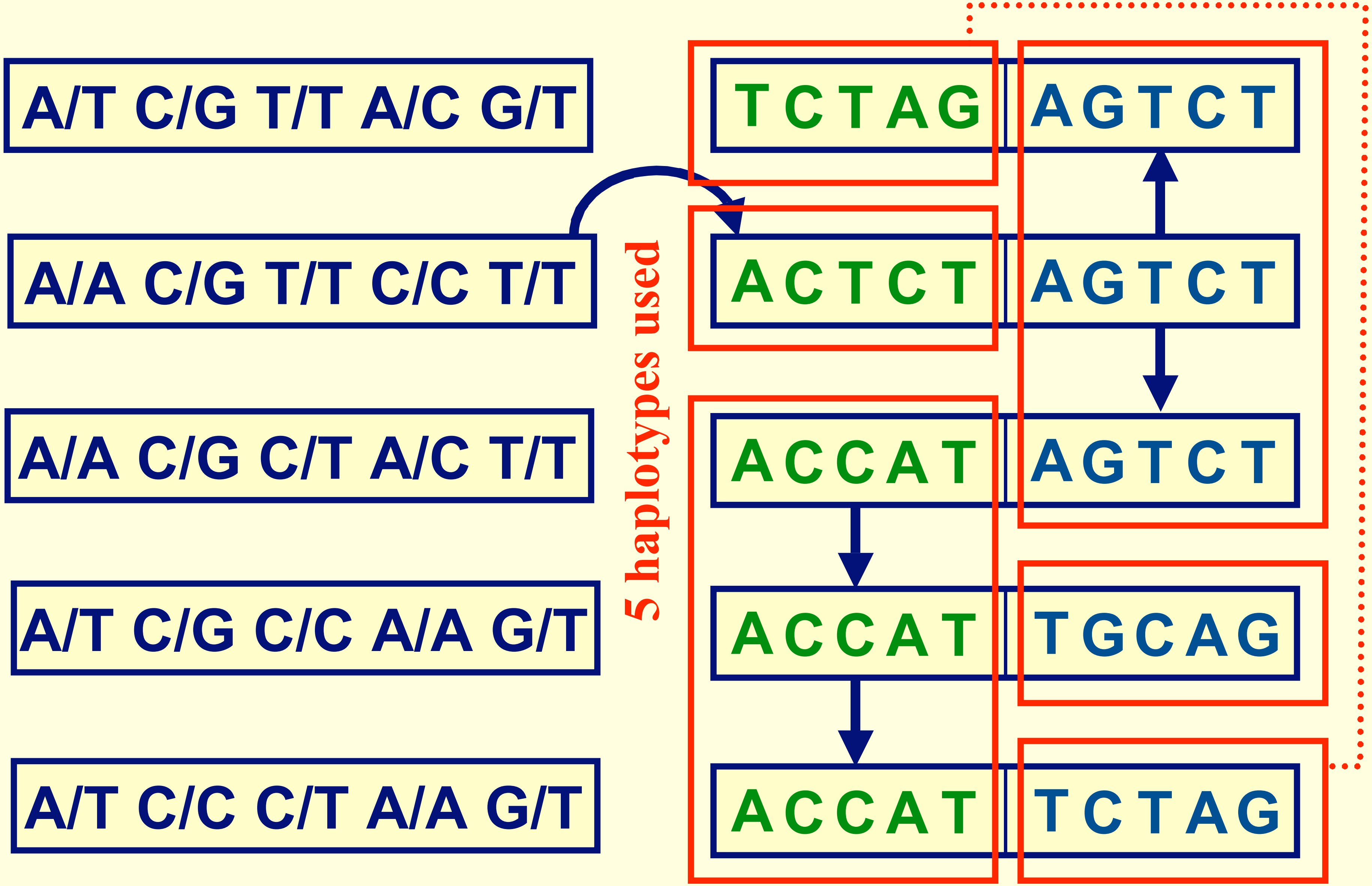
A/A C/G C/T A/C T/T ACTCT AGCAT

A/T C/G C/C A/A G/T TCCAG AGCAT

A/T C/C C/T A/A G/T TCCAG ACTAT

6 haplotypes used

Clark: Rerun (same input)



A/T C/G T/T A/C G/T

A/A C/G T/T C/C T/T

A/A C/G C/T A/C T/T

A/T C/G C/C A/A G/T

A/T C/C C/T A/A G/T

TCTAG AGTCT

ACTCT AGTCT

ACCAT AGTCT

ACCAT TGCAG

ACCAT TCTAG

5 haplotypes used

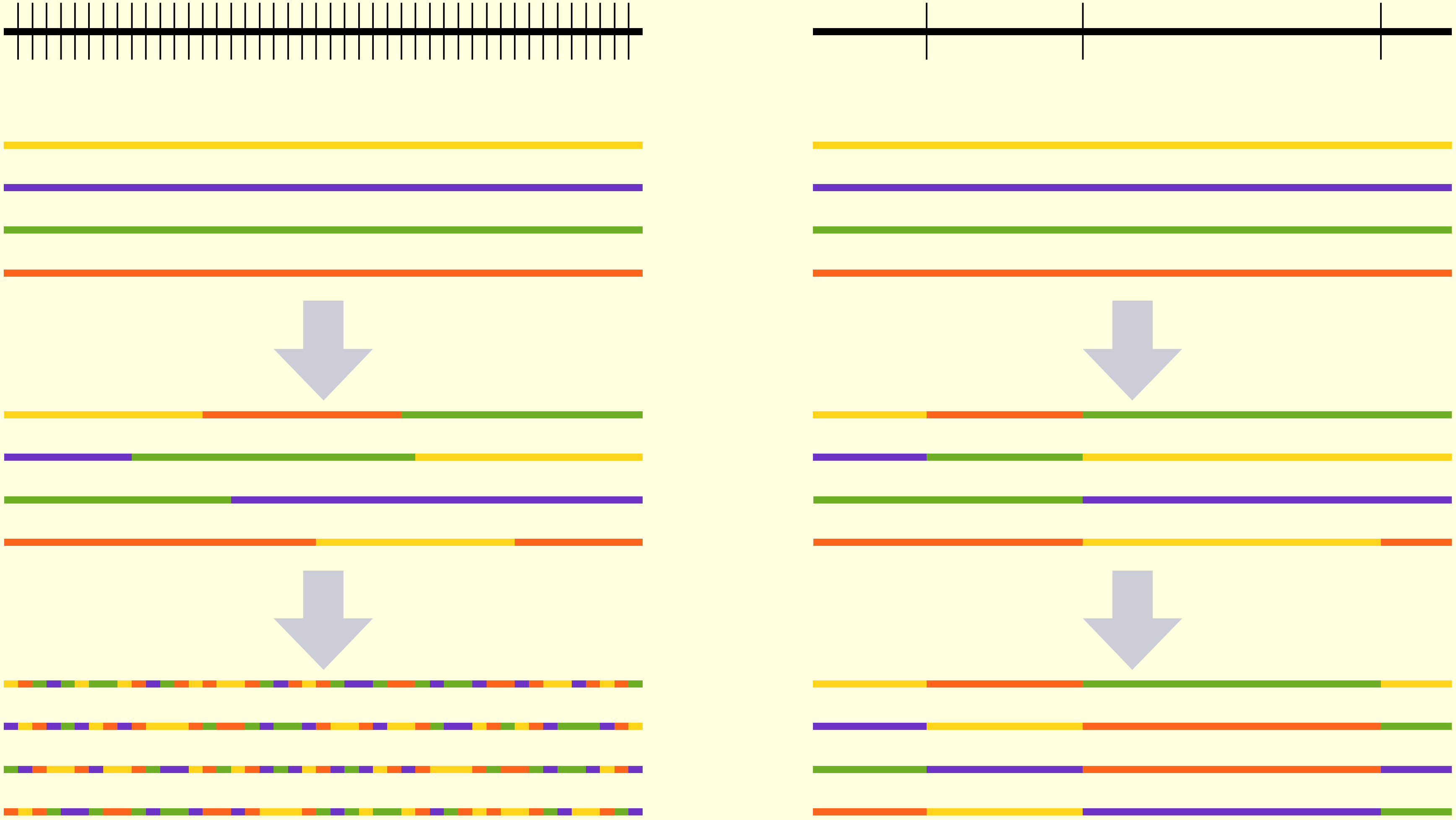
Clark: Comments

- Implementation is very fast, $O(ln^2)$
- Total failure if no starting point.
- Blind haplotyping of ‘orphans’ at end.
- Arbitrary selections based on input order.
  - Try multiple orderings, select best results.
- Or formulate choices as integer program
  - Solve approximately by linear relaxation.

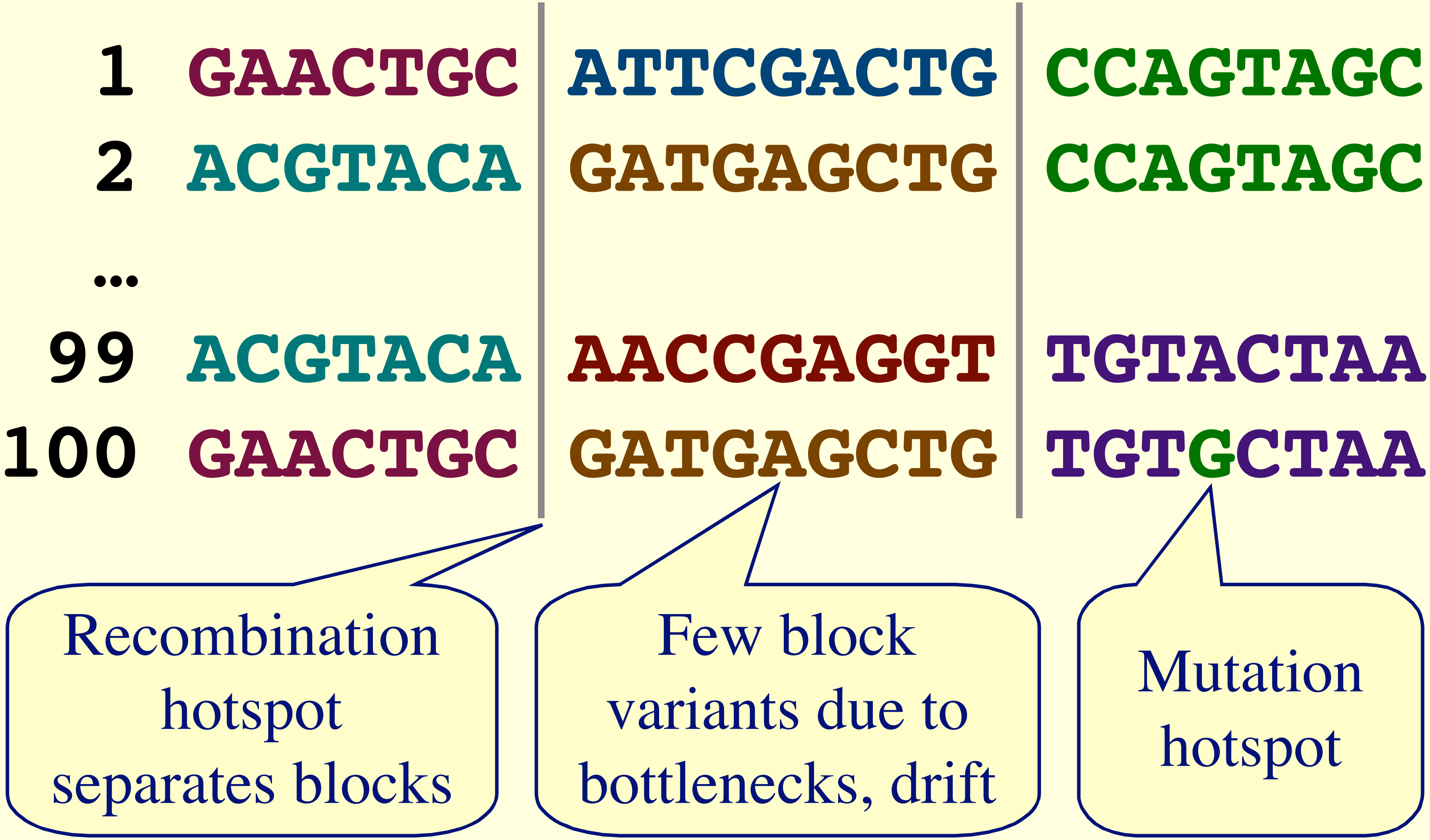
Part 2: HaploBlock

- Haplotype blocks
- Statistical model
- Model inference
- Model criterion
- Applications
  - Haplotyping
  - Block-based LD mapping

Recombination Hotspots



Haplotype Blocks



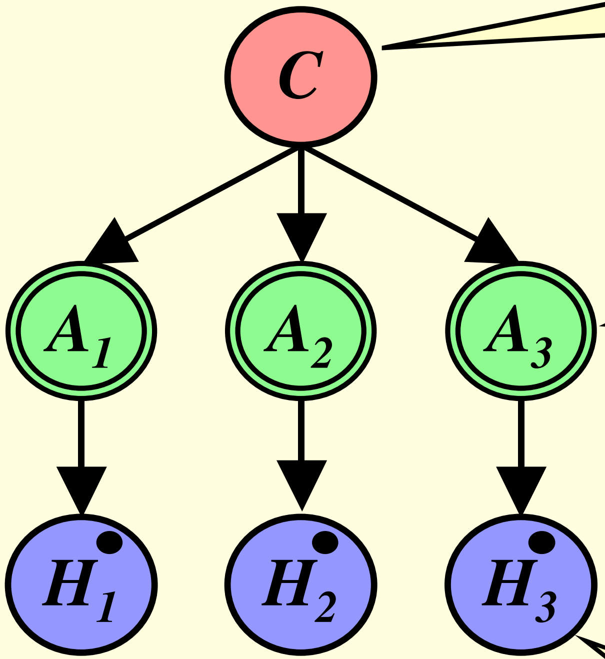
1	GAACTGC	ATTCGACTG	CCAGTAGC
2	ACGTACA	GATGAGCTG	CCAGTAGC
…			
99	ACGTACA	AACCGAGGT	TGTACTAA
100	GAACTGC	GATGAGCTG	TGTGCTAA

Recombination hotspot separates blocks

Few block variants due to bottlenecks, drift

Mutation hotspot

Bayesian Network Model



Pr($\boldsymbol{C} = \boldsymbol{c}$) is frequency of haplotype $\boldsymbol{c}$

Values of variable $\boldsymbol{C}$ are 1…$\boldsymbol{q}$ denoting index of block's haplotype

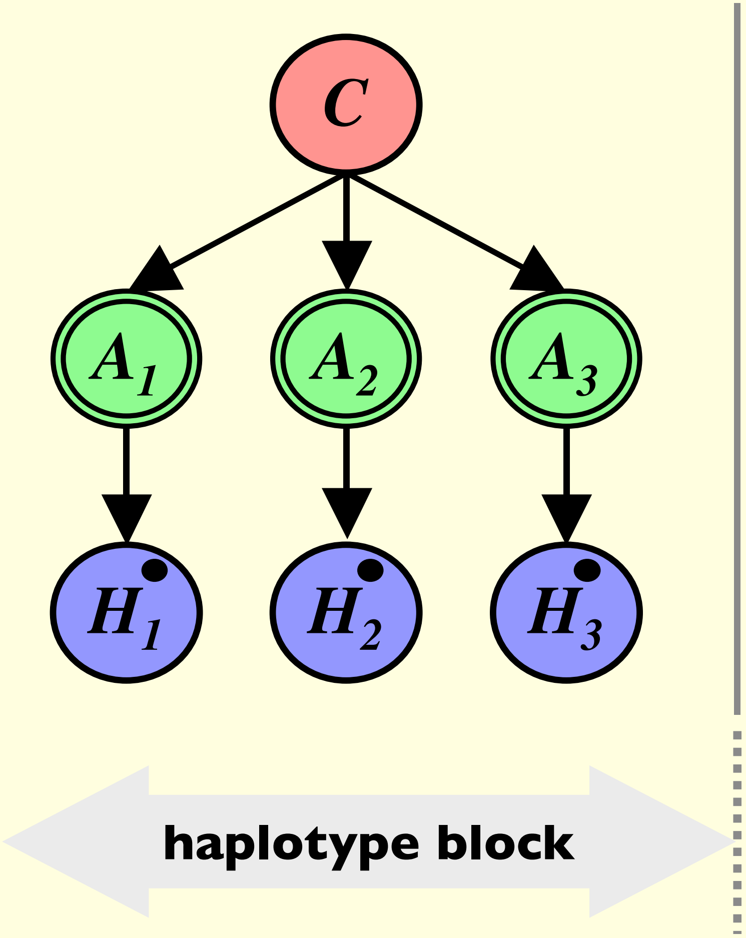
Pr($\boldsymbol{a}_j \mid \boldsymbol{c}$) is deterministic

Values of variable $\boldsymbol{A}_j$ are $A,C,G,T,$– denoting allele at site $\boldsymbol{j}$ of haplotype. Example: $\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3 = CTA$ for $\boldsymbol{C} = 2$

Pr($\boldsymbol{h}_j \mid \boldsymbol{a}_j$) is cumulative mutation rate

Values of variable $\boldsymbol{H}_j$ are $A,C,G,T,$– denoting allele at site $\boldsymbol{j}$ observed after possible haplotype mutations

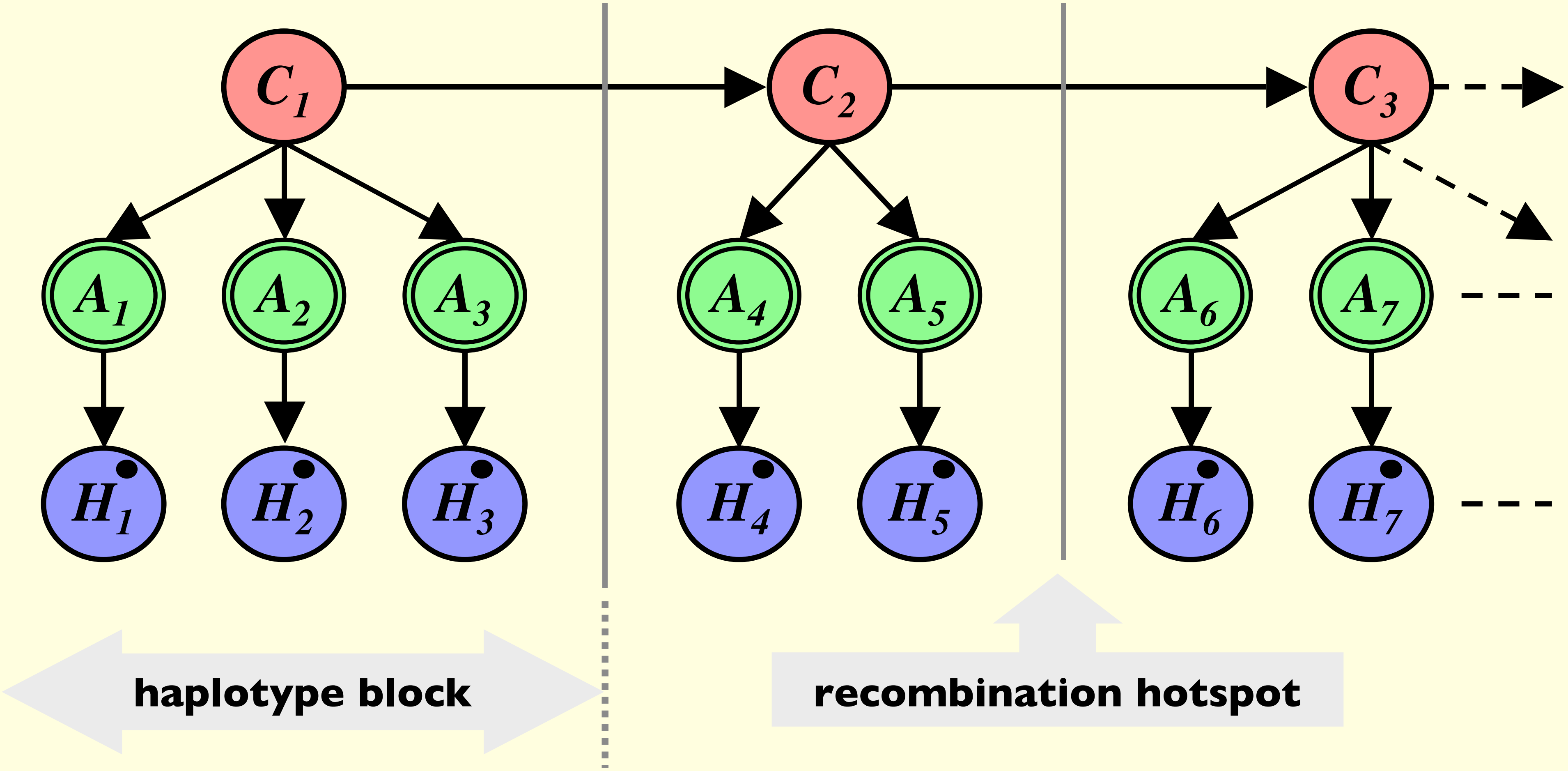
Bayesian Network Model



$$\Pr(c, a_1, a_2, a_3, h_1, h_2, h_3) =$$

$$\Pr(c) \times$$
$$\Pr(a_1 | c) \times$$
$$\Pr(a_2 | c) \times$$
$$\Pr(a_3 | c) \times$$
$$\Pr(h_1 | a_1) \times$$
$$\Pr(h_2 | a_2) \times$$
$$\Pr(h_3 | a_3)$$

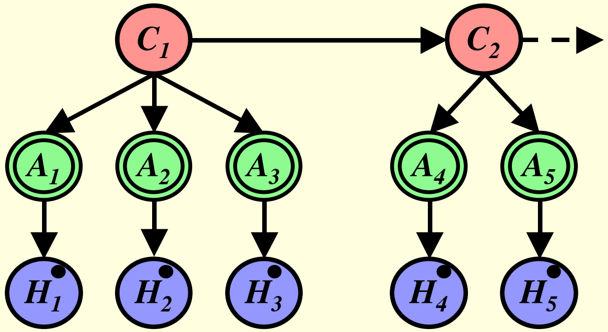
Bayesian Network Model



haplotype block

recombination hotspot

Data Likelihood

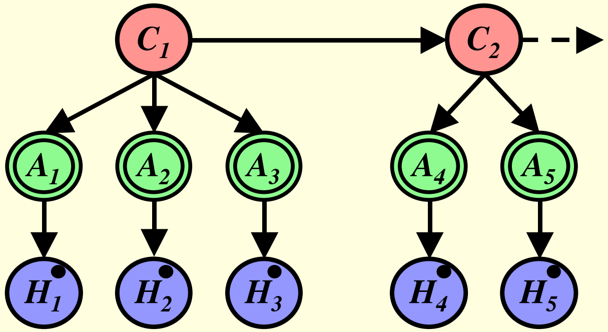


- For haplotypes *H*, likelihood is:

$$\Pr(H) = \prod_{h \in H} \left[\sum_{c_1} \cdots \sum_{c_b} \sum_{a_1} \cdots \sum_{a_l} \left[\begin{array}{l} \Pr(c_1) \prod_{k=2}^{b} \Pr(c_k \mid c_{k-1}) \\ \prod_{k=1}^{b} \prod_{j=s_k}^{e_k} \Pr(a_j \mid c_k) \Pr(h_j \mid a_j) \end{array} \right] \right]$$

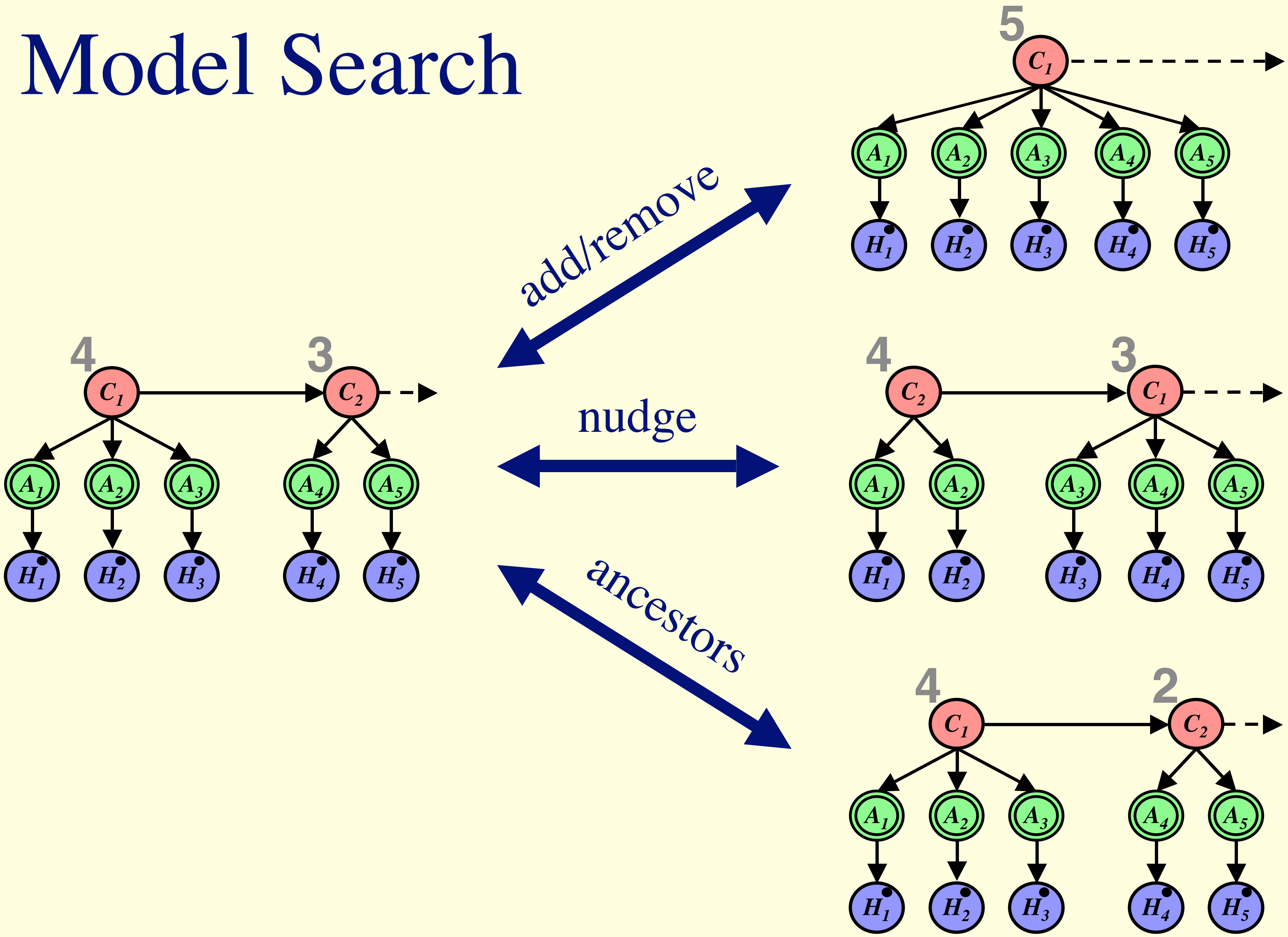
But we can calculate this efficiently using a suitable elimination ordering!

Data Criterion



- Maximum Likelihood leads to over-fitting
  - – No hotspots, no mutations, many ancestors
  - – Need to consider model complexity
  - – Min $\mathrm{DL}(H,M)=\mathrm{DL}(M)-\log_2\Pr(H|M)$
- DL(M) considers variable elements only
  - – Ancestor block sequences
  - – Markov chain parameters

Model Search



add/remove

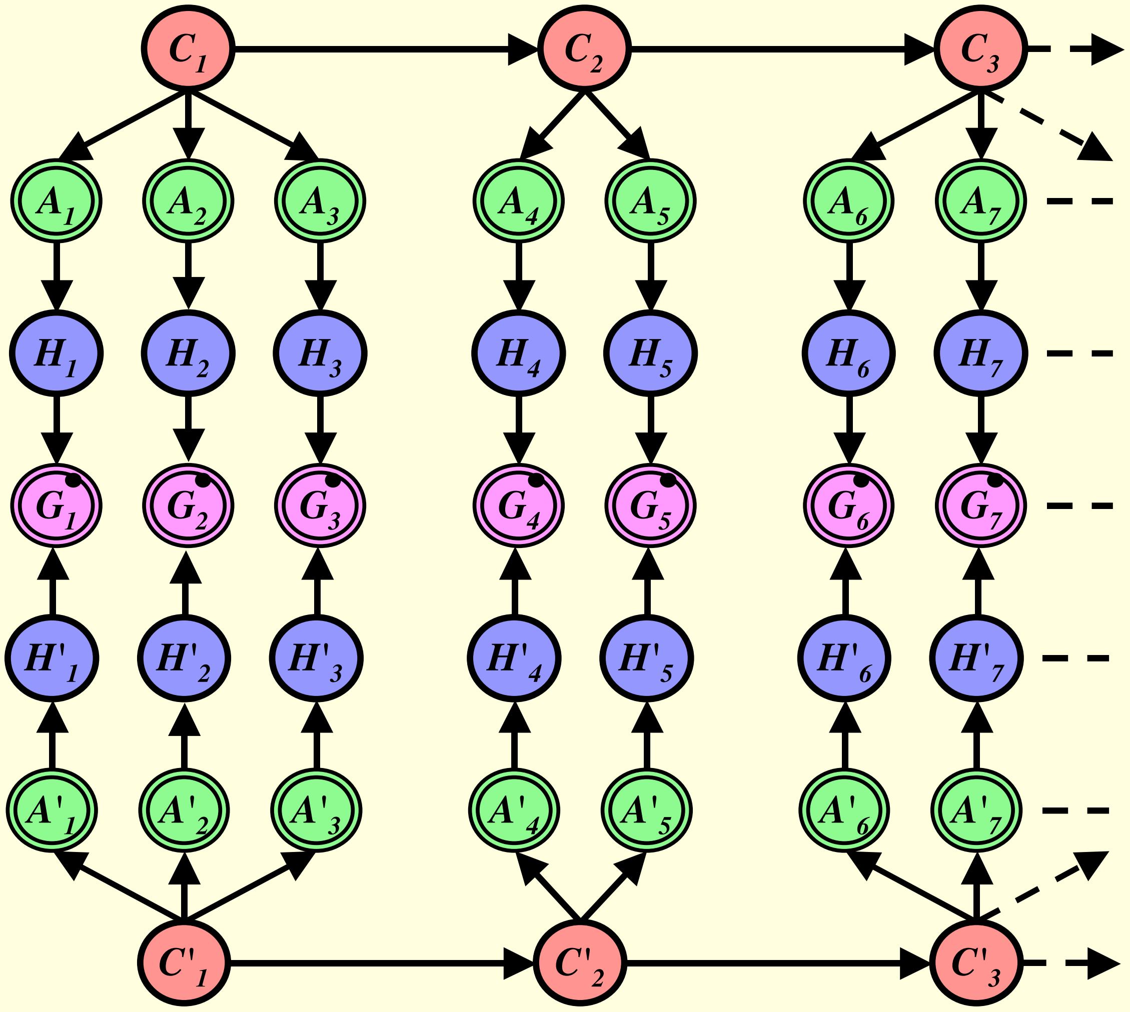
nudge

ancestors

5
C_1
A_1 A_2 A_3 A_4 A_5
H_1 H_2 H_3 H_4 H_5
4
3
C_1
C_2
A_1 A_2 A_3 A_4 A_5
H_1 H_2 H_3 H_4 H_5
4
3
C_2
C_1
A_1 A_2 A_3 A_4 A_5
H_1 H_2 H_3 H_4 H_5
4
2
C_1
C_2
A_1 A_2 A_3 A_4 A_5
H_1 H_2 H_3 H_4 H_5

Model for Haplotyping

- Learn model directly from genotypes
- Haplotype pair: choose most likely under model



Haplotyping Results

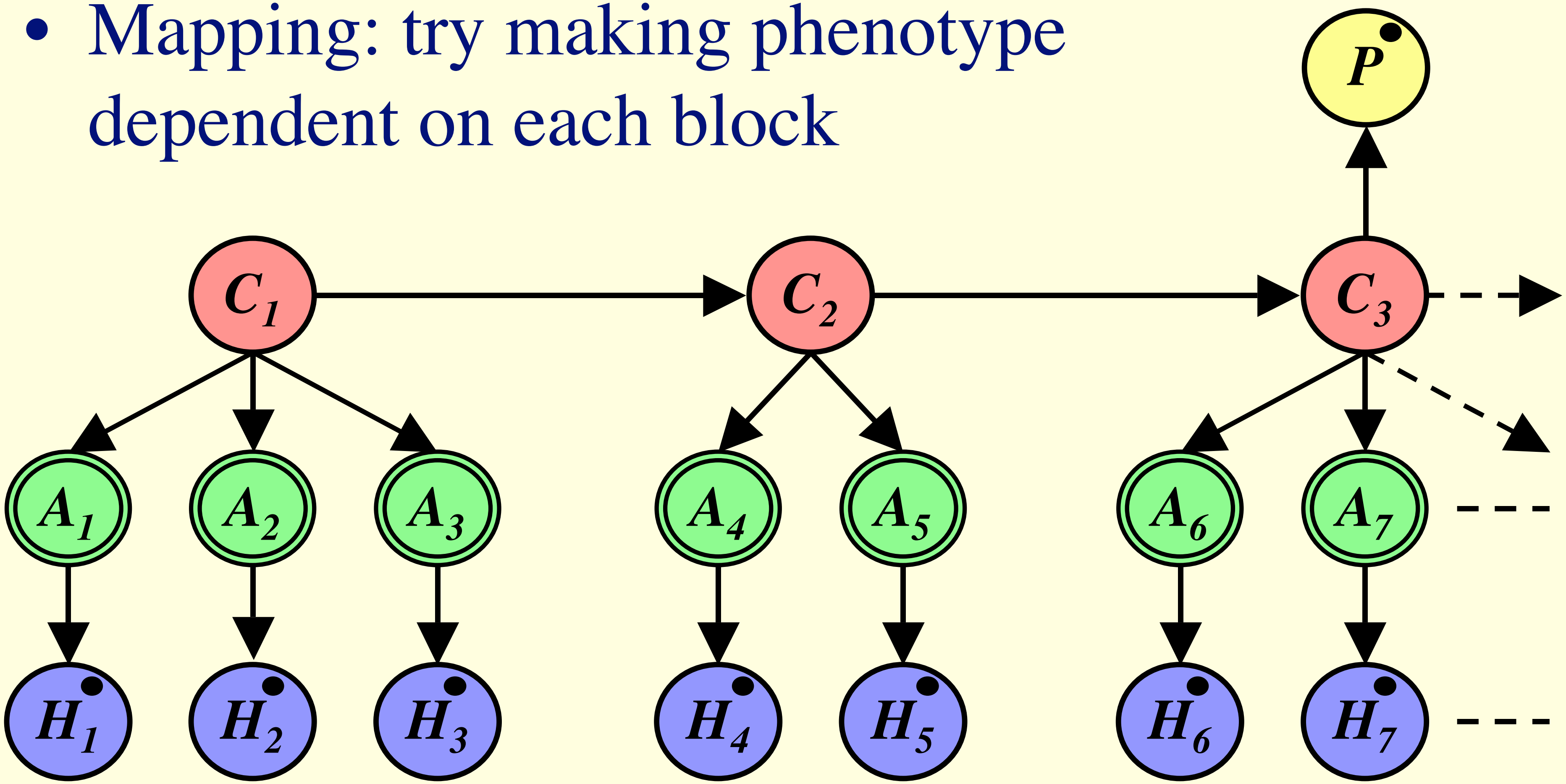
Site pairwise error rate	C21a	C21b	C21c	C21d	C21e	ACE
Clark	.0548	.0251	.0280	.0329	.0234	.0381
Hierarchical EM	.0095	.0042	.0009	.0047	.0083	.0152
HAPLOTYPER	.0224	*failed*	.0204	.0077	*failed*	.0102
PHASE	.0669	.0403	.0655	.0262	.0183	.0419
HaploBlock	**.0047**	**.0020**	**.0005**	**.0014**	**.0048**	**.0098**
Improvement factor	**2x**	**2x**	**2x**	**3x**	**2x**	**=**

C21x data: 20 haplotypes, 100 SNPs over ≤ 35kb, Patil *et al*. (2001)
ACE data: 22 haplotypes, 52 SNPs over 24kb, Rieder *et al*. (1999)

Average shown for 10 random pairings of true haplotypes

Model for LD Mapping

- Learn model from marker data
- Mapping: try making phenotype dependent on each block



LD Mapping Results

Resequencing required	5q31 haplos	5q31 genos	Chr 21
BLADE	144 kb	–	107 kb
No Blocks	131 kb	105 kb	33 kb
HaploBlock	**40 kb**	**37 kb**	**24 kb**
Improvement factor	**3x**	**3x**	**1.4x**

5q31 data: 258 haplotypes, 98 SNPs over 464kb, Daly *et al*. (2001)
Chr 21 data: 20 haplotypes, 5 sets of 200 SNPs, Patil *et al*. (2001)

Average shown for 5 random selections of target SNP

HaploBlock: Comments

- Our model boils down to an HMM
  - Calculations have linear complexity
  - Forward/backward probability caching
- Better to infer multiple models
  - Prevent getting stuck in local minima
  - Account for uncertainty of block identification
  - Use Gibbs-style iterations on hotspots
  - Take ‘average’ result over set of models