生物信息学

单体型组装与推断

吴凌云

中国科学院数学与系统科学研究院

SNP和单体型

多态性

- 人类在DNA水平上有99.9%的相似程度
- 不同个体或种群之间DNA序列的差异称为基因多态性
- 基因多态性决定着一些外部特征如绿眼睛和褐眼睛，或直发与弯法等
- 基因变异(Mutation)是主要的一种多态性
- 单体型是研究基因变异与疾病联系的重要工具

单核苷酸多态性(SNP)

- 在单个核苷酸上的差异称为单核苷酸多态性(Single Nucleotide Polymorphism)
- SNP是人类基因组DNA序列变异的主要形式
- SNP是决定人类疾病（尤其是多基因疾病）易感性和药物反应性差异的核心信息
- 90%的多态性都是由单个核苷酸的变异引起的
- 人的SNP密度：平均1000个碱基中含有1个SNP位点
- 在整个人类基因组中大约有300万个单核苷酸多态性存在

SNP的类型

- 大部分SNP都是以双等位基因的形式出现： major (wild type) / minor (mutant type)
- 非同义(Non-Synonymous) SNP是指那些导致编码的氨基酸发生变化的SNP
- 编码序列中的SNP有一半是非同义SNP
- 普通(Common) SNP是指那些minor等位基因出现频率大于5%的SNP

单体型和基因型

- 双倍体生物的染色体总是成对出现的
- 在双倍体生物的每一条染色体的每一个SNP位点上，都会出现两个可能的等位基因中的一个
- 每一条染色体上的所有等位基因按一定顺序排列构成一条**单体型**
- 两条染色体上的对应等位基因组合在一起构成一条**基因型**

例子

来自父方的染色体: ATAGCCTATTTCCAGGAGTCGAAGAC

来自母方的染色体: ATAGCGTATTTCCAGGAGTCGTAGAC

单体型 1→ C C A

单体型 2→ G C T

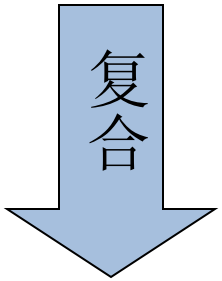
基因型→ {C,G} {C,C} {A,T}

单体型和基因型的表达

0 1 1 1 0 0 1 1 0

1 1 0 1 0 0 1 0 0

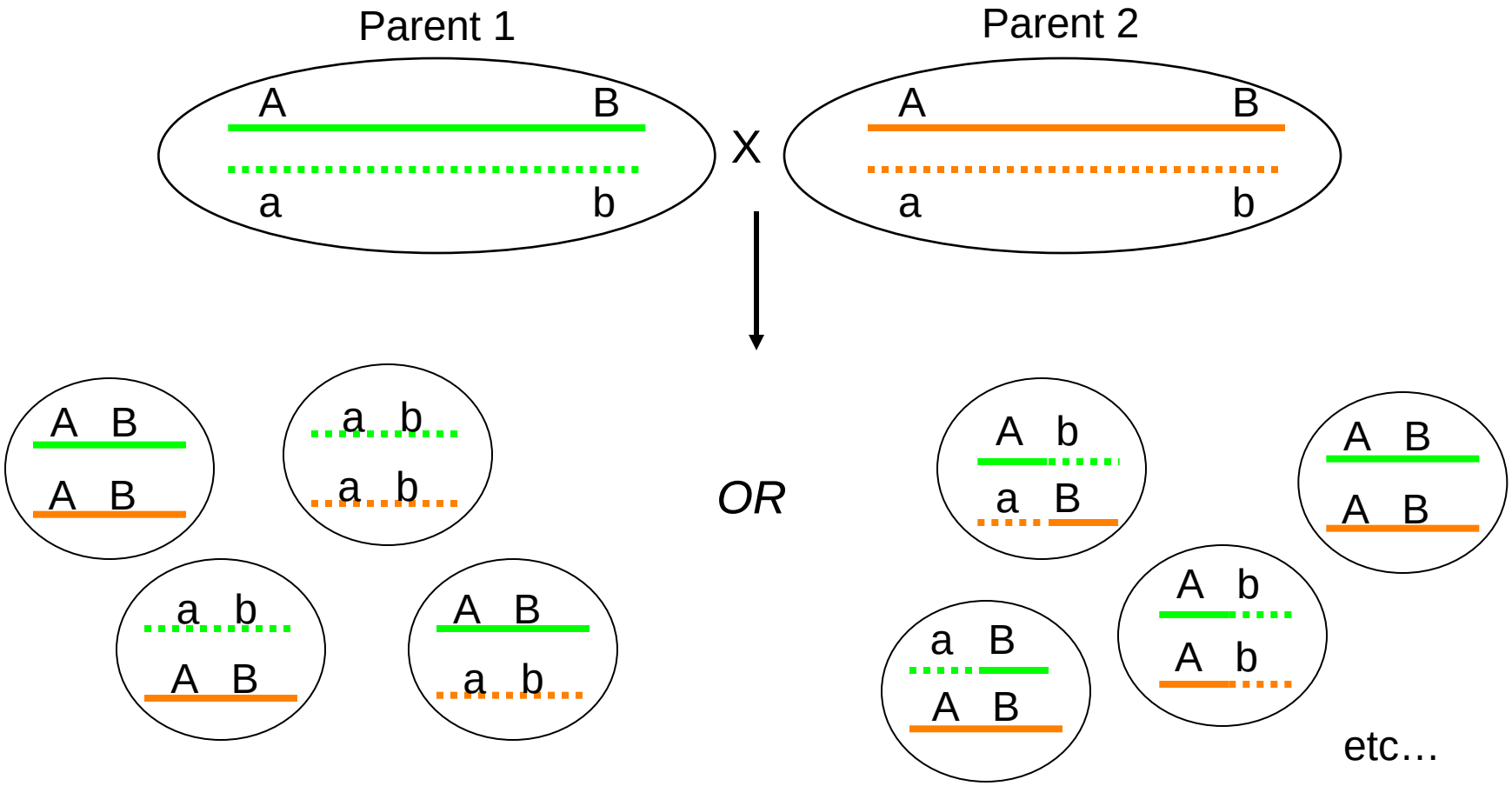
两条单体型



2 1 2 1 0 0 1 2 0

一条基因型

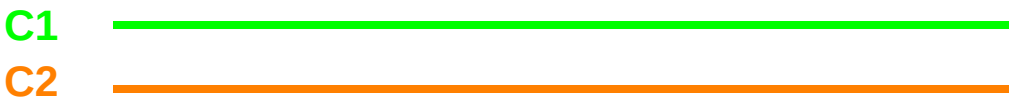
Linkage Disequilibrium



High LD -> No Recombination
($r^2 = 1$) SNP1 “tags” SNP2

Low LD -> Recombination
Many possibilities

alleles:	SNP1	SNP2
	A/a	B/b



POP allele freqs:		
	A (80%)	B (60%)
	a (20%)	b (40%)

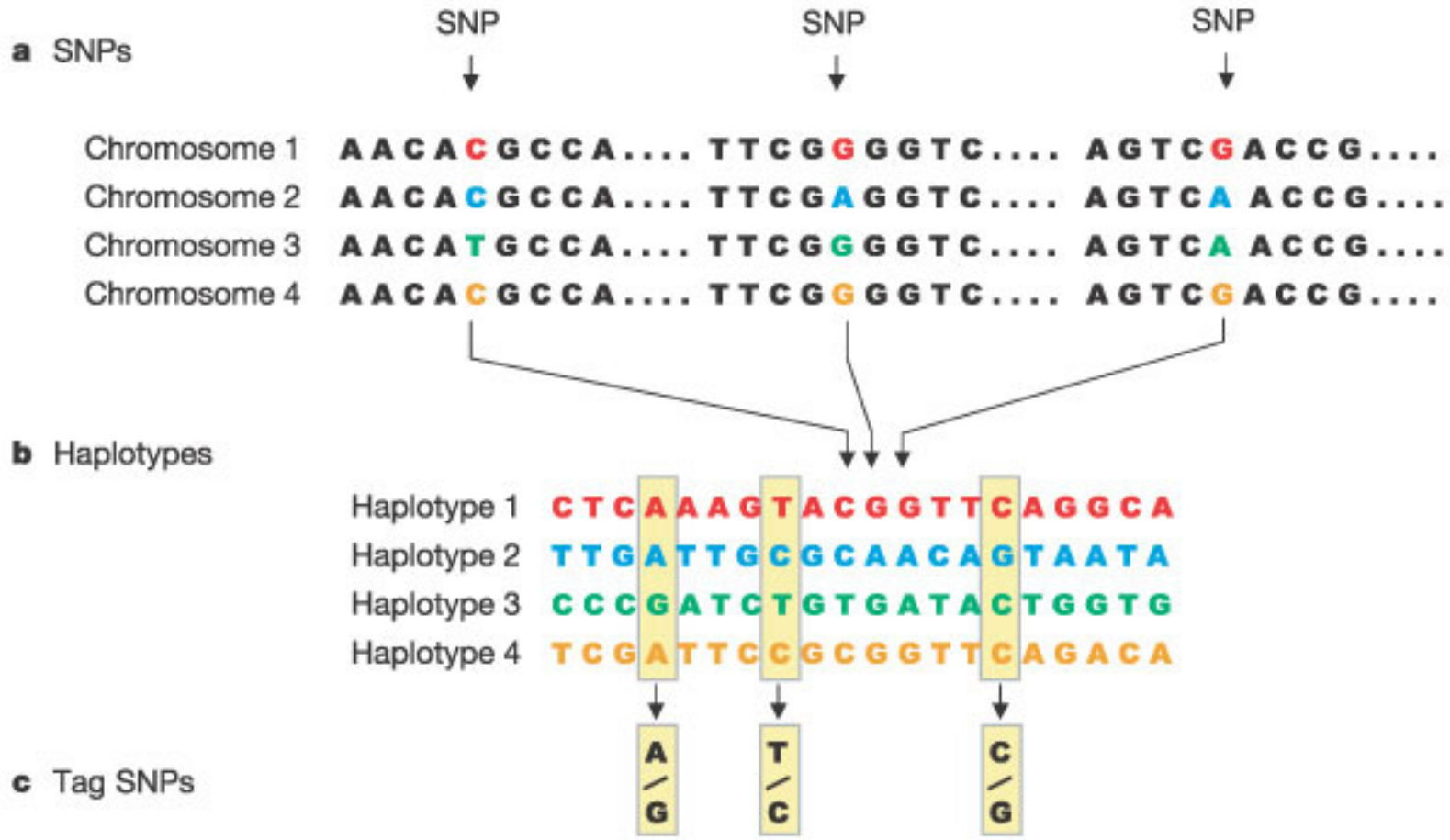
genotypes:

Person 1	Person 2	Person 3
AA	AA	Aa
BB	Bb	Bb

phased haplotypes (C1/C2):

Person 1	Person 2	Person 3
A — B	A — B	A — B
A — B	A — b	a — b
		OR
		A — b
		a — B

标签SNP


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

单体型图(HapMap)计划

- 检测染色体上所有的SNPs的费用极其昂贵
- 一些相互邻近的多态位点趋向于在一起共同遗传，这些变异连锁的区域就是单体型
- 在一个特定人群中，55％的人可能拥有同一种单体型，30％的人可能拥有另一种单体型，8％的人可能拥有第三种单体型，而其余的人可能拥有若干种稀有的单体型
- 通过标签SNPs来鉴定一个人的单体型集合
- 定位与重要医学特征相关的基因

单体型检测

- 目前检测单体型的方法主要有两种
- **单体型组装(Haplotype Assembly)**
  - 从较小的SNP片断来组装单体型
  - 优点：比较精确
  - 缺点：技术实现较难、费用昂贵、速度慢
- **单体型推断(Haplotype Inference)**
  - 从群体的基因型来推断单体型
  - 优点：技术实现容易、基因型容易获取
  - 缺点：不精确

单体型组装问题

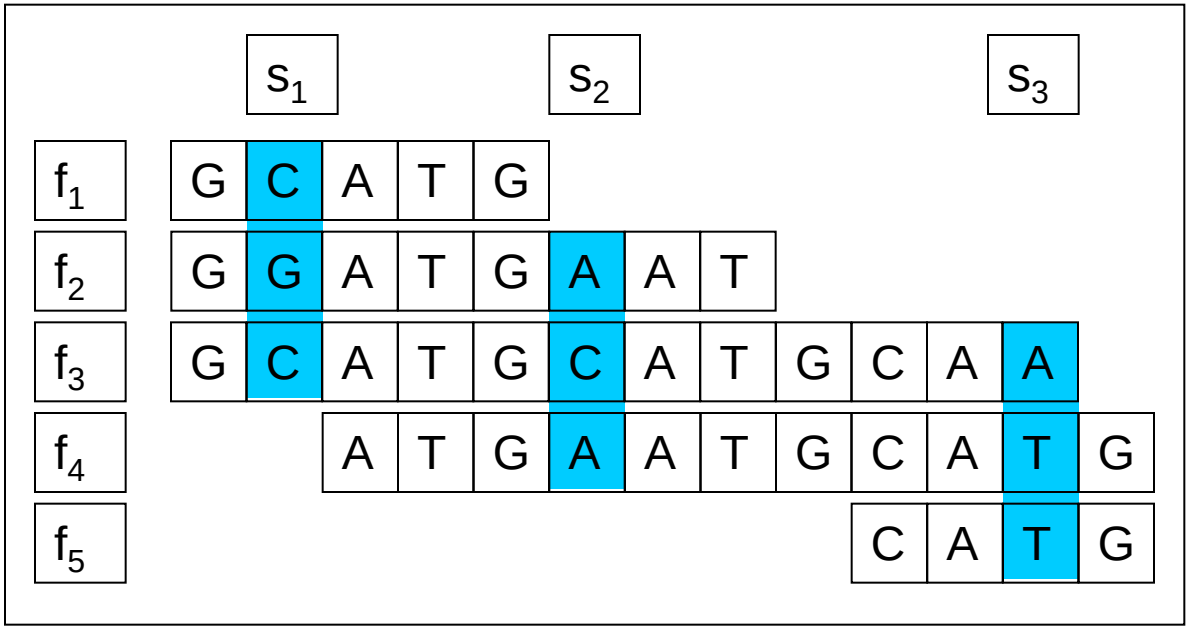
单体型组装问题

- Haplotype Assembly
- Haplotyping for individual
- 从Shotgun测序实验得到的序列片断来组装出一对单体型

困难

- 序列片断可能来自任何一条染色体
- 片断较短：每个片断可能包含2~3个SNP
- 测序错误
- 杂质污染

SNP矩阵



	s_1	s_2	s_3
f_1	0	–	–
f_2	1	0	–
f_3	0	1	0
f_4	–	0	1
f_5	–	–	1

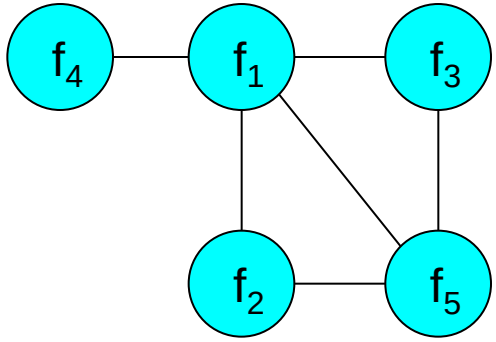
DNA fragments

SNP matrix

冲突图

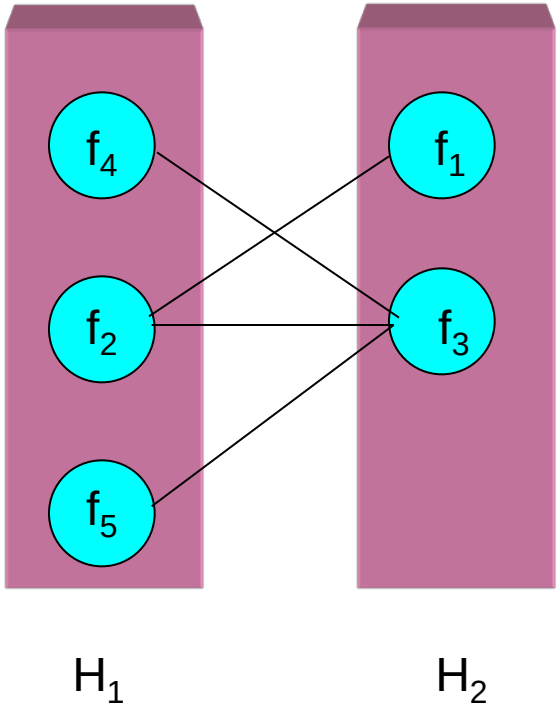
	s_1	s_2	s_3	s_4	s_5	s_6
f_1	0	1	–	0	0	1
f_2	1	0	–	–	1	–
f_3	–	0	1	0	1	0
f_4	–	0	–	–	–	0
f_5	1	–	0	1	0	–

SNP matrix



Fragment conflict graph G_F

二分图



图的二分化

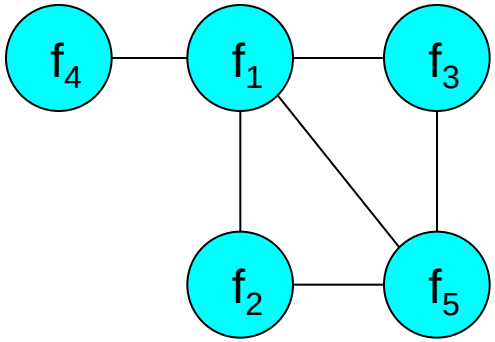
- 可以用Mathematica 5.1中的子程序 *BipartiteQ* 来计算判断一个图是否是一个二分图
- 一个图是否是二分图当且仅当它没有奇圈 (*a cycle with odd number of edges*) (S.Skiena,1990)
- 如何从带有误差的数据中恢复出一对单体型

  ⇔

  如何将一个图合理地变成一个二分图

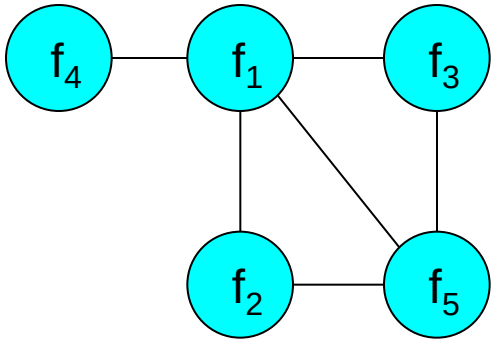
去除顶点

- 去掉一些顶点来得到二分图 (相当于去除一些受污染的片段)

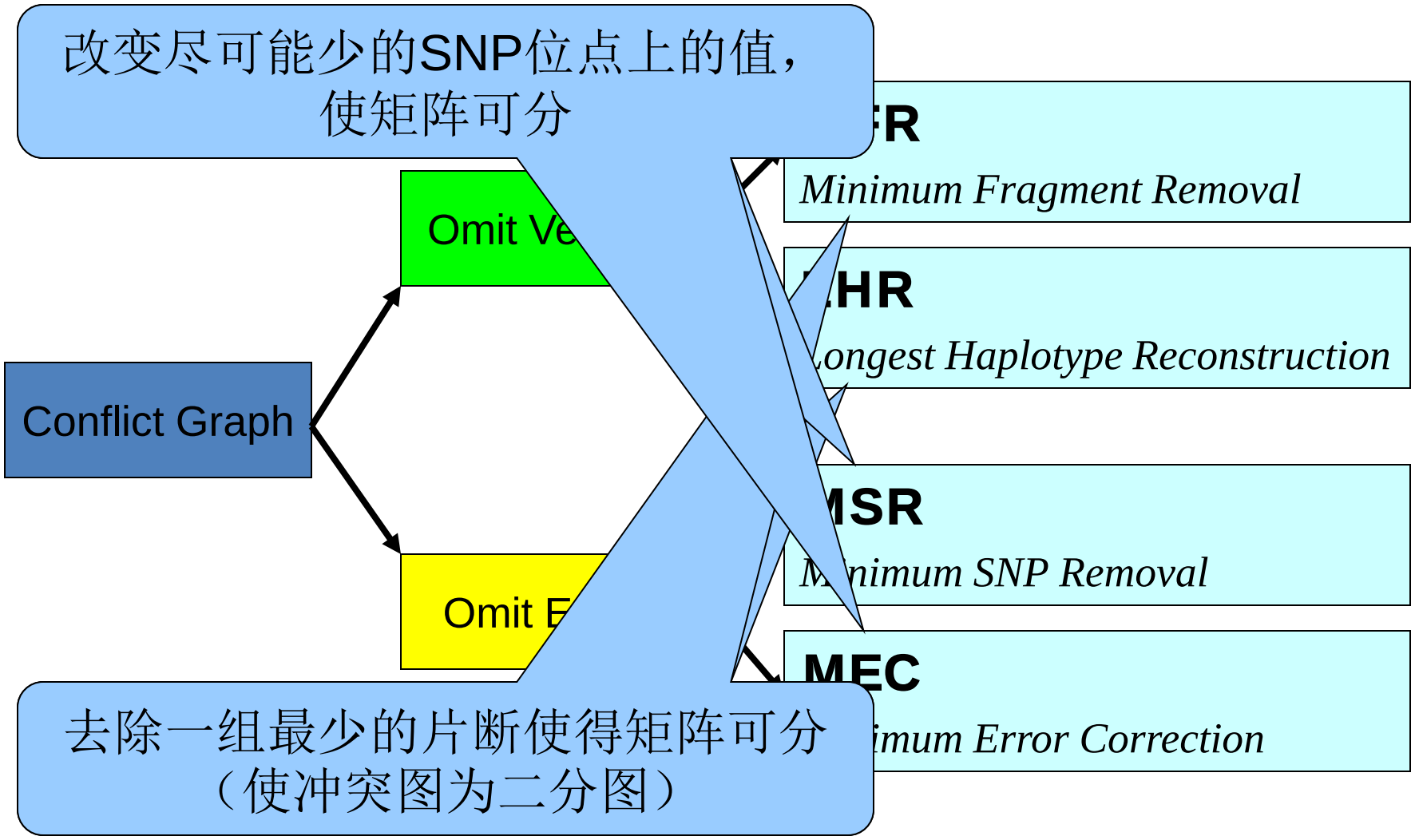


去除边

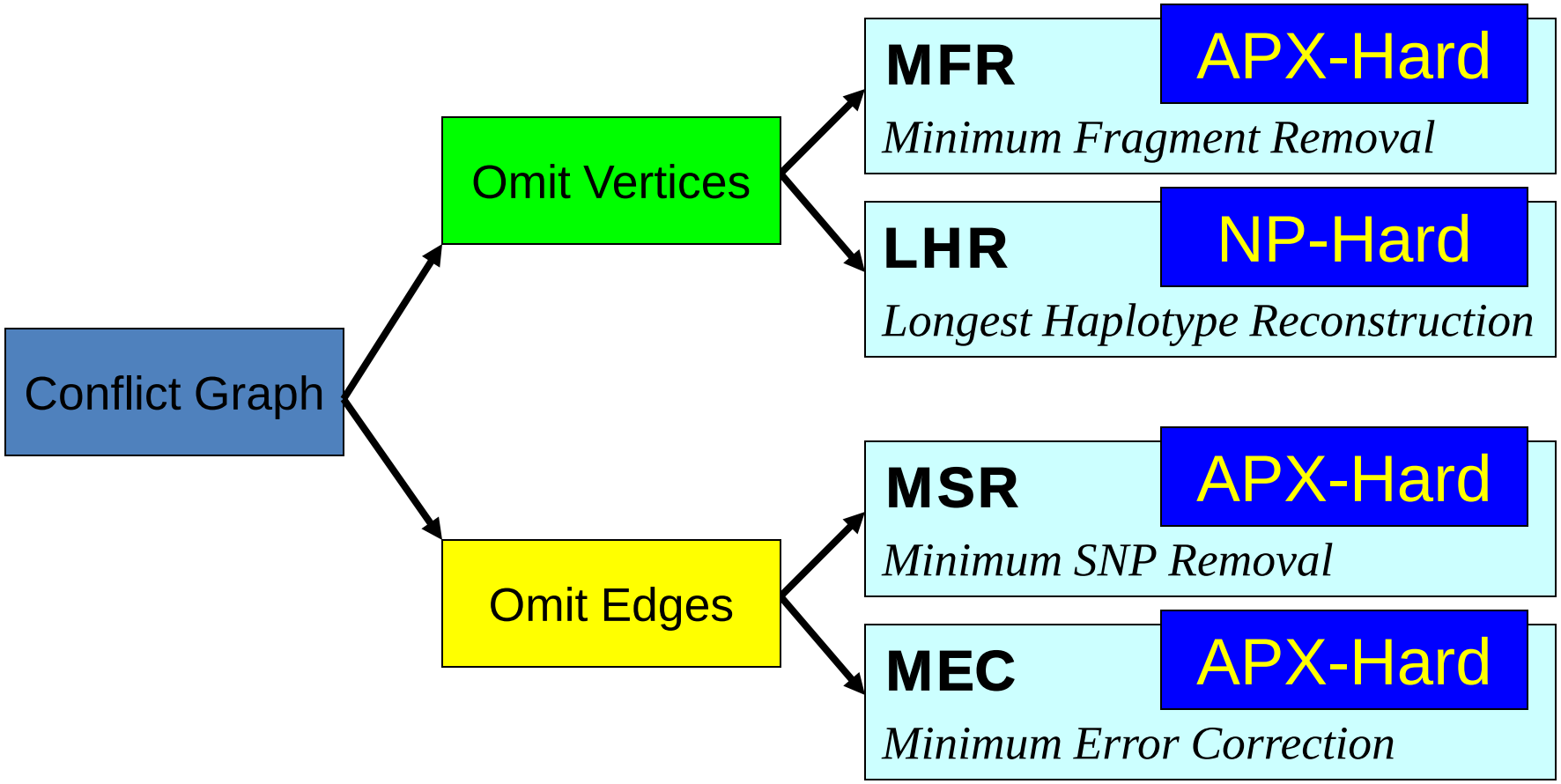
- 去掉一些边来得到二分图 (相当于去除一些SNP位点或改变某些片段上某些SNP位点的值)



单体型组装模型


改变尽可能少的SNP位点上的值，
使矩阵可分
Omit Ve
Conflict Graph
Omit F
去除一组最少的片断使得矩阵可分
（使冲突图为二分图）
FR
Minimum Fragment Removal
HR
ongest Haplotype Reconstruction
MSR
nimum SNP Removal
MEC
imum Error Correction

计算复杂性


Conflict Graph
Omit Vertices
Omit Edges
MFR
APX-Hard
Minimum Fragment Removal
LHR
NP-Hard
Longest Haplotype Reconstruction
MSR
APX-Hard
Minimum SNP Removal
MEC
APX-Hard
Minimum Error Correction

MEC的整数规划模型

$$\max \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij}$$

$$x_{il}(r_{ij}y_{ij} - h_{lj}) = 0 \qquad \begin{array}{l} i = 1, 2, \cdots, m; \\ j = 1, 2, \cdots, n; \\ l = 1, 2; \\ r_{ij} \neq 0 \end{array}$$

$$x_{i1} + x_{i2} = 1 \quad i = 1, 2, \cdots, m$$

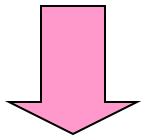
$$x_{il} \in \{0, 1\} \quad i = 1, 2, \cdots, m; l = 1, 2$$

$$y_{ij} \in \{-1, 1\} \quad i = 1, 2, \cdots, m; j = 1, 2, \cdots n$$

$$h_{lj} \in \{-1, 1\} \quad l = 1, 2; j = 1, 2, \cdots, n$$

线性约束

$$x_{il}(r_{ij}y_{ij} - h_{lj}) = 0 \qquad \begin{array}{l} i = 1, 2, \cdots, m; \\ j = 1, 2, \cdots, n; \\ l = 1, 2; \\ r_{ij} \neq 0 \end{array}$$



$$-2 + x_{il} \leq r_{ij}y_{ij} - h_{lj} \leq 2 - x_{il} \qquad \begin{array}{l} i = 1, 2, \cdots, m; \\ j = 1, 2, \cdots, n; \\ l = 1, 2; \\ r_{ij} \neq 0 \end{array}$$

算法思想

- 将MEC问题转化为分类问题
  - 将SNP片断分为两类：C_1和C_2
  - 对每一个分类C_i，用最大似然原则迅速求出对应的最优单体型H_i（即C_i中的SNP片断只需最少的修正次数即可与单体型H_i一致）
  - 搜索比较所有可能的分类
- MEC的最优解为满足下式的分类P^*

$$E(P^*) \leq E(P), \forall P$$

MEC模型算法

- 精确算法
  - 分支定界算法
- 近似算法
  - 动态聚类算法
  - 神经网络算法
  - 遗传算法

MEC模型的扩展

- MEC/GI
  - MEC with Genotype Information
  - 要求组装出的单体型和基因型数据一致

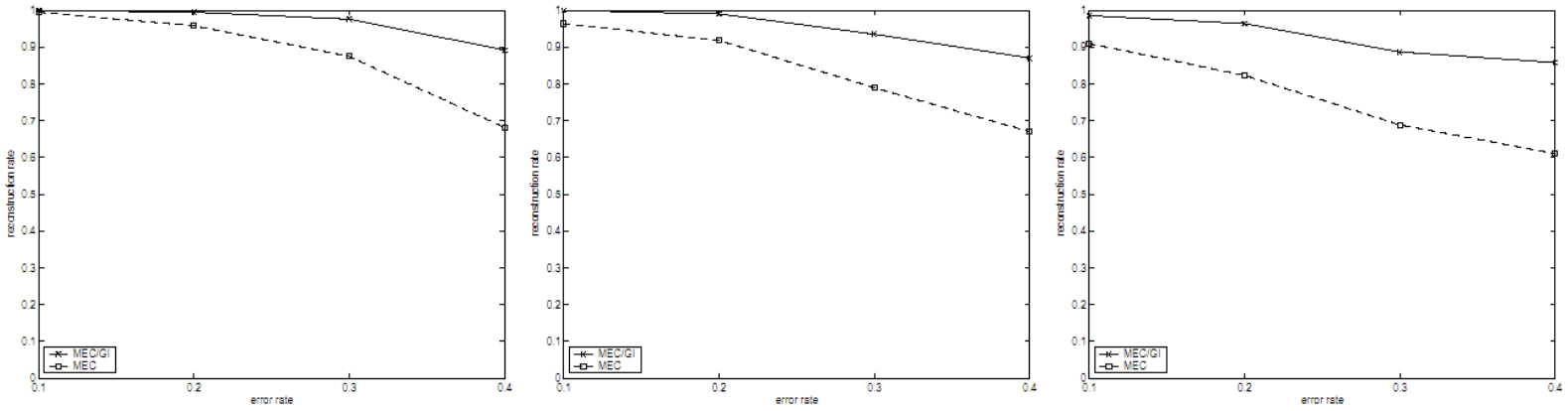


Figure 5: The comparison results of the MEC model and the MEC/GI model on ACE. From left to right, $g = 0.25$, $g = 0.5$, $g = 0.75$.

MEC模型的扩展

- WMEC
  - Weighted MEC
  - DNA测序时每个碱基都有一个置信度
  - 被纠正的SNP的权重之和最小
  - 当所有SNP的权重设为1时，即原始的MEC模型

MEC模型的扩展

- CWMEC
  - Complete Weighted MEC
  - 给定删除比例R_F和R_S，允许去掉不超过R_F比例的片断和R_S比例的SNP位点
  - WMEC和MFR、MSR的结合
    - $R_F=1, R_S=0$,即MFR模型
    - $R_F=0, R_S=1$,即MSR模型
    - $R_F=0, R_S=0$,即WMEC模型

图的二分化算法

- Reed B, Smith K, Vetta A. **Finding odd cycle transversals**. *Oper Res Letters* **2004**; 32(4): 299–301.
- Hüffner F. **Algorithm engineering for optimal graph bipartization**. In Proceedings of the 4th International Workshop of Efficient and Experimental Algorithms (WEA). Springer-Verlag **2005**; 240-252.
- Guo J, Gramm J, Hüffner F, Niedermeier R, Wernicke S. **Improved fixed parameter algorithms for two feedback set problems**. In Proceedings of the 9th Workshop on Algorithms and Data Structures (WADS). Springer-Verlag **2005**; 158-168.

评论

- 我们注意到，在 Reed *et al*. 的工作之前，在图论研究中极少有对图的最优二分化的研究，原因是大家知道这是一个NP-难问题。正是由于生物信息学研究的需要，推动图论学家回来研究这一问题并得到好的研究结果。

单体型推断问题

单体型推断问题

- **Haplotype Inference**
- **Haplotype Phasing**
- **Haplotyping in population**
  - 从一个人群的基因型来推断单体型
  - 给定一个人群的基因型数据
  - 为该人群中的每个人找到最可能的单体型配对
  - 估计在该人群中的单体型出现频率

基因型

- Genotype (基因型) 是指每个标签位置（Marker）的无序等位基因对（unordered pair of alleles for each marker）组成的一个序列
- Homozygous (纯合子): 在一个位置上的一对等位基因是相同的
- Heterozygous (杂合子):在一个位置上的一对等位基因是相异的

例子

来自父方的染色体: ATAGCCTATTTCCAGGAGTCGAAGAC

来自母方的染色体: ATAGCGTATTTCCAGGAGTCGTAGAC

单体型 1→ C C A

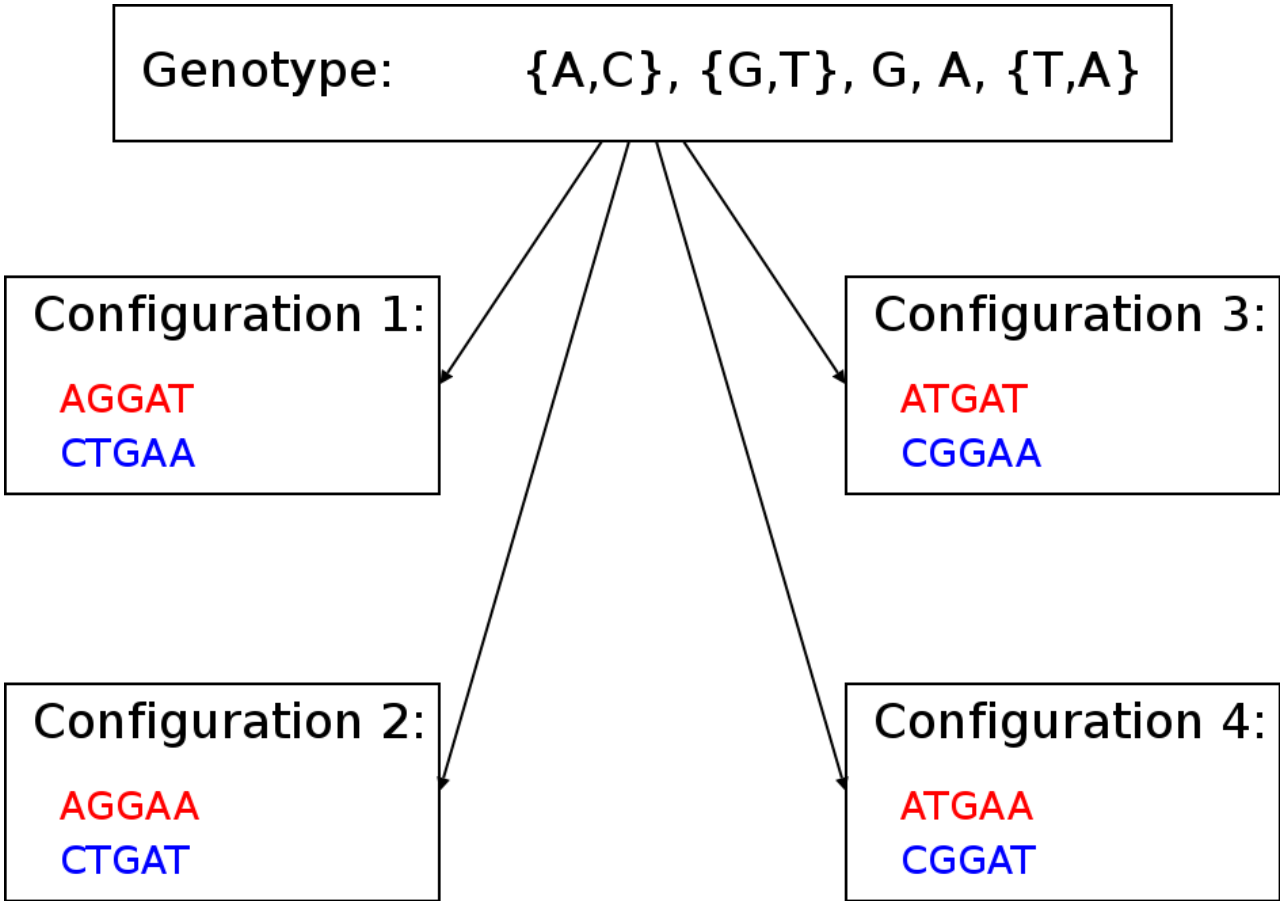
单体型 2→ G C T

基因型→ {C,G} {C,C} {A,T}

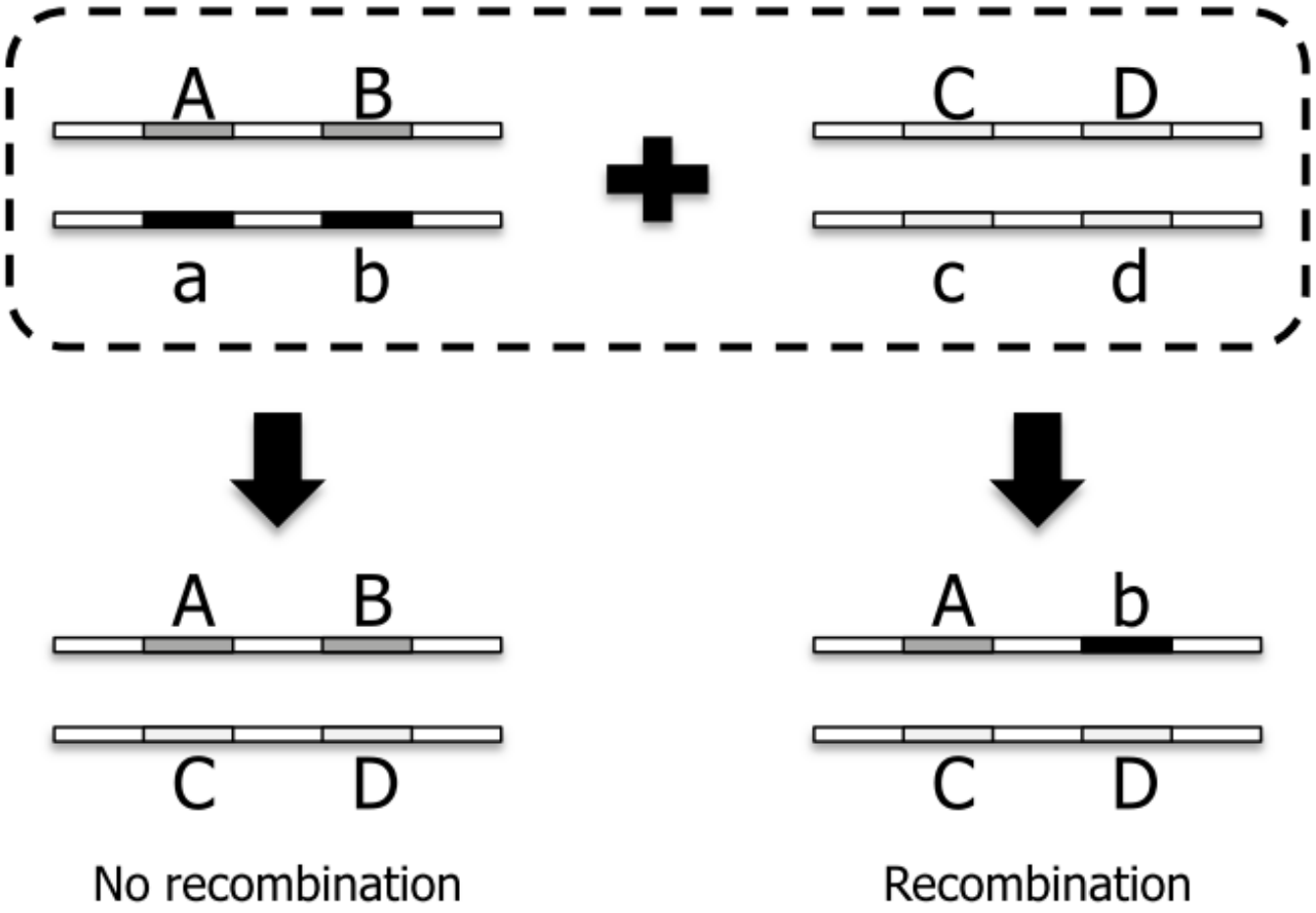
基因型的解析

- 与基因型一致的一对单体型称为这个基因型的解析（resolution)
- 一个基因型的杂合子个数为k，则对应的不同的解析有2^{k-1}个
- 2^{k-1}个解析中哪种才是正确的、合理的？

例子


Genotype: {A,C}, {G,T}, G, A, {T,A}
Configuration 1:
AGGAT
CTGAA
Configuration 3:
ATGAT
CGGAA
Configuration 2:
AGGAA
CTGAT
Configuration 4:
ATGAA
CGGAT

基因重组(Recombination)



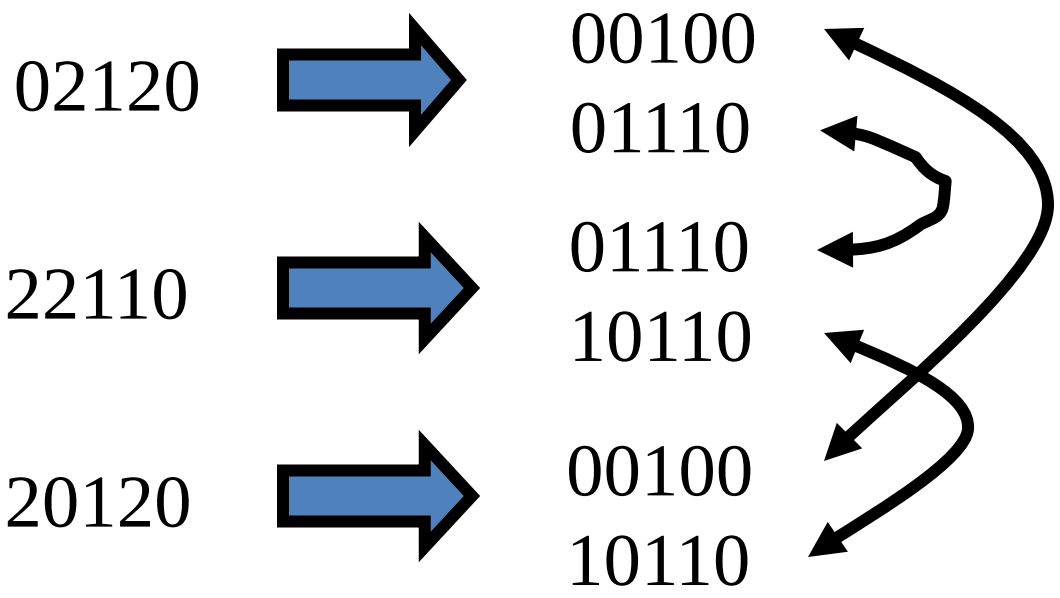
Pedigree Data

- Assumption
  - Pedigree graph (genetic relationship)
  - Mendelian law (no mutation)
- Difficulty
  - Additional genotyping costs
  - Potential recruiting problems
  - Not all SNP sites can be resolved

Pedigree Models

- **Minimum Recombination Haplotype Configuration (MRHC)**
  - Pedigree graph: NP-hard (Li and Jiang 2003)
  - Pedigree tree: NP-hard (Doi *et al.* 2003)
- **Zero Recombination Haplotype Configuration (ZRHC)**
  - Polynomial solvable (Li and Jiang 2003)
- ***k*-Minimum Recombination Haplotype Configration (*k*-MRHC)**
  - Pedigree graph: NP-hard (k>=1) (Chin *et al.* 2005)
  - Pedigree tree: Open

节俭原则



HIPP模型

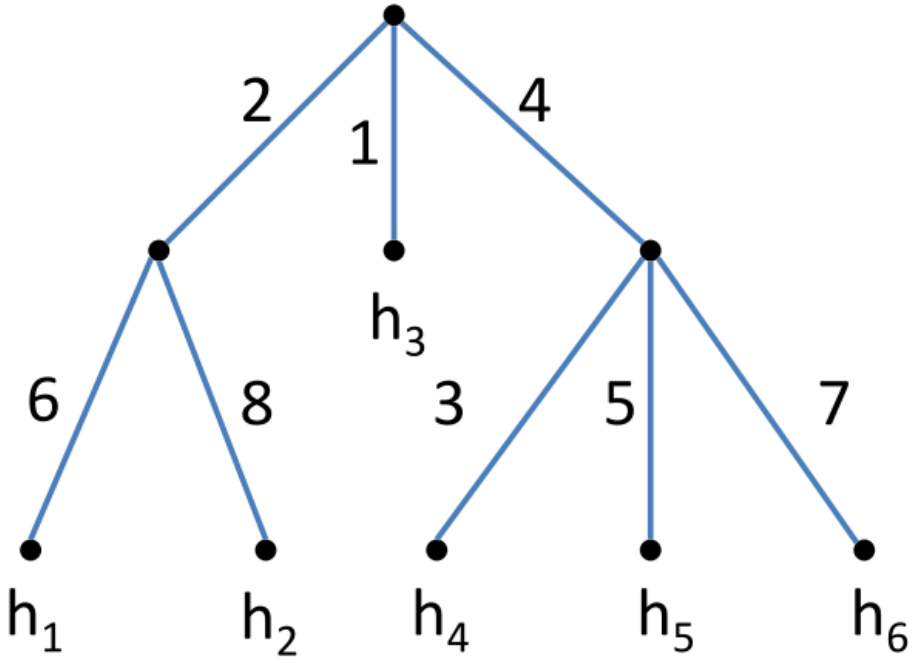
- **Haplotype Inference by Pure Parsimony**
- The HIPP is APX-hard (Lancia *et al.* 2004)
- Branch and Bound (Wang and Xu 2003)
- Integer Programming (Gusfield 2003, Brown and Harrower 2004, 2006)
- Approximation algorithms (Lancia et al. 2004, Huang et al. 2005)
- Heuristic Method (Li et al. 2005)

进化树(Phylogeny)

- **Haplotype Perfect Phylogeny**
  - Rooted tree
  - Each leaf denotes a distinct haplotype
  - Each edge represents a SNP site with a mutation from 0 to 1
  - Each SNP site is labeled by at most one edge
  - For each haplotype labeled by a leaf, the unique path from the root to it specifies all SNP sites with value 1

进化树(Phylogeny)

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$



PPH模型

- **Perfect Phylogeny Haplotype (PPH)**
  - Graph Realization Problem
  - Polynomial algorithms (Gusfield 2002, Bafna *et al.* 2003, Eskin *et al.* 2003)
- **Minimum Perfect Phylogeny Haplotype**
  - NP-hard (Bafna et al. 2004)

Imperfect Phylogeny

- PPH is NP-hard with data missing (Kimmel and Shamir 2005)
- PPH with additional biologically motivated constrains (Gramm et al. 2004)
- Imperfect Phylogeny Haplotype (IPH) (Halperin and Eskin 2004)

统计学假设

- Underlying unknown distribution of haplotype frequencies
- Hardy-Weinber Equilibrium (HWE)

$$\Pr(g) = \sum_{h \oplus \bar{h} = g} \Pr(h) \Pr(\bar{h})$$

统计学模型

- **Haplotype Frequency Estimation**
  - EM (Expectation-Maximization) (Excoffier et al. 1995)
  - PL-EM (Partition-Ligation-Estimation-Maximization) (Niu et al. 2002)
- **Bayesian Haplotype Inference**
  - MCMC (Markov Chain Monte Carlo) (Stephens et al. 2001)
  - PL-MCMC (Niu et al. 2002)
- **Markov Chain Model**
  - PL like method (Eronen et al. 2004)
  - Dynamic Programming (Zhang et al. 2005)

1阶MC模型

$$\Pr(H) \approx \Pr(H(1)) \prod_{i=2,n} \Pr(H(i)|H(i-1))$$

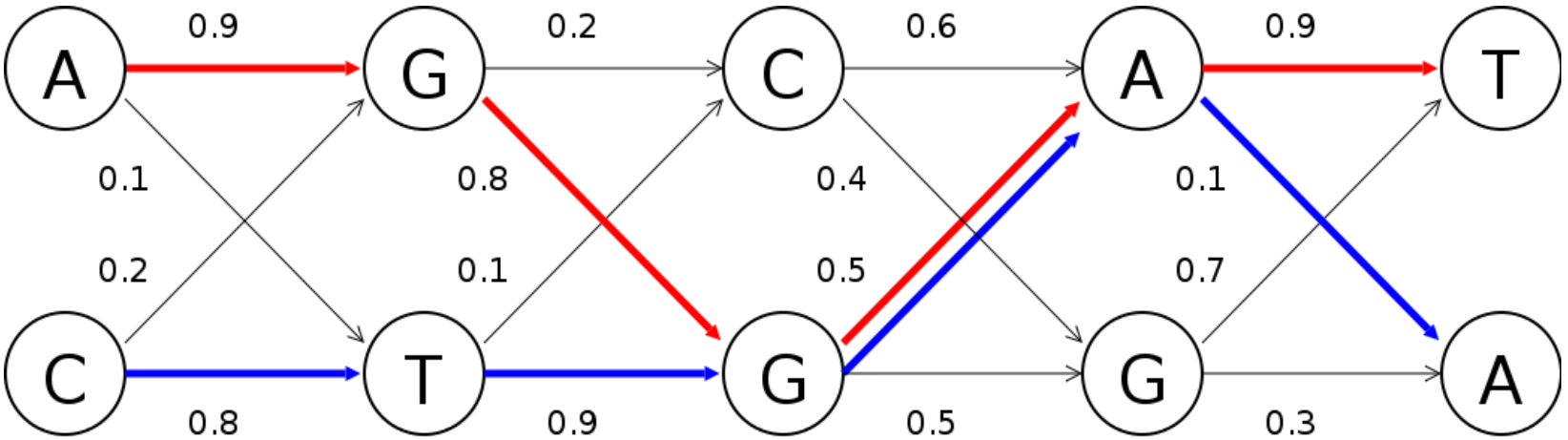
$$\Pr(H) \approx fr(H(1)) \prod_{i=2,n} \frac{fr(H(i-1,i))}{fr(H(i-1))}$$

片断频率估计

Suppose that $k_{G(i,j)}$ is the number of heterozygous markers in $G(i,j)$, then the frequncy of haplotype fragment $H(i,j)$ is estimated as follows:

$$fr(H(i,j)) = \frac{1}{2|\mathcal{G}|} \sum_{\substack{G \in \mathcal{G}, \\ G \text{ matches } H(i,j)}} 2^{1-k_{G(i,j)}}$$

1阶MC例子

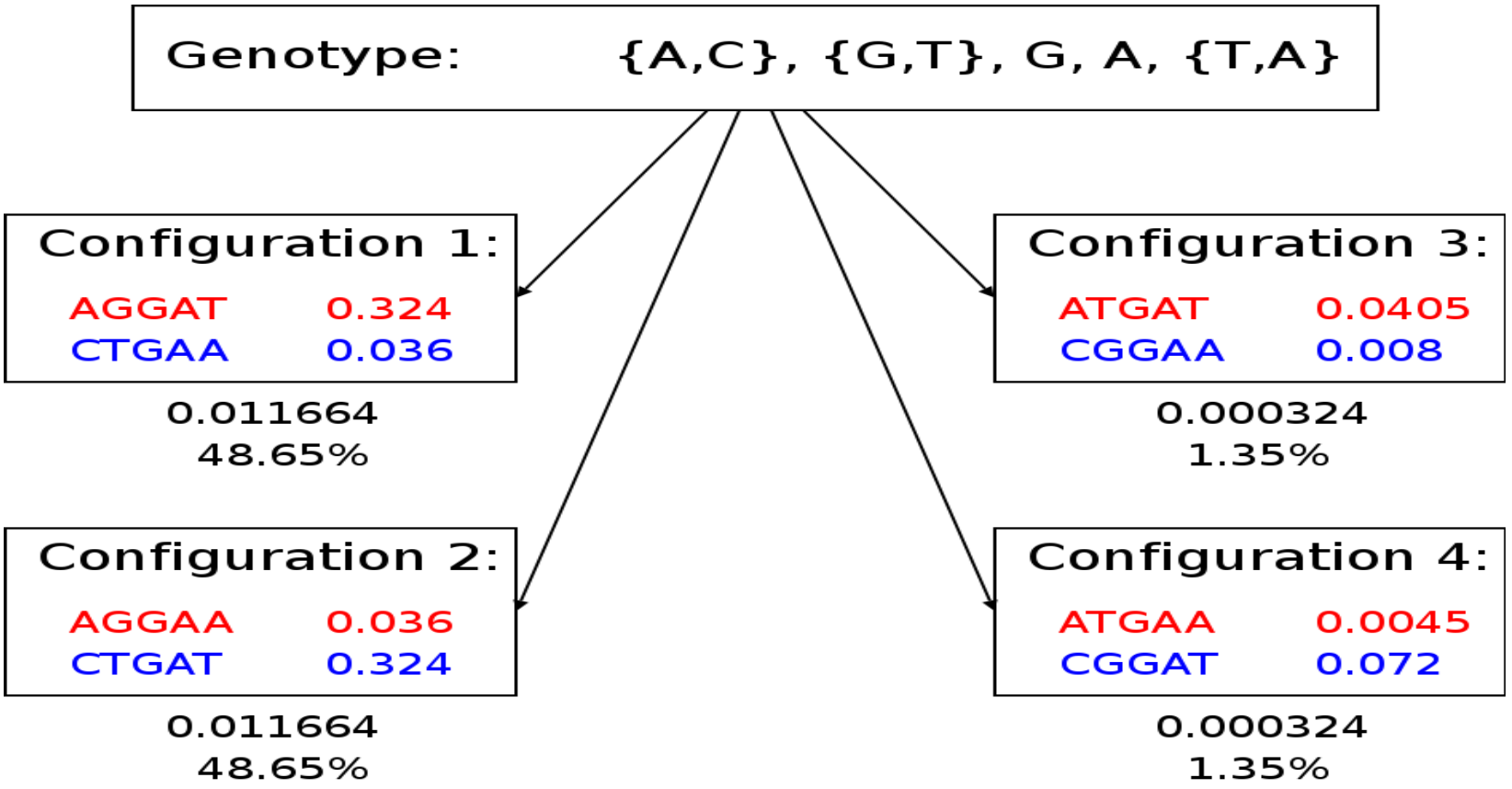


Genotype: {A,C}, {G,T}, G, A, {T,A}

Configuration 1:

AGGAT	0.324
CTGAA	0.036

1阶MC例子

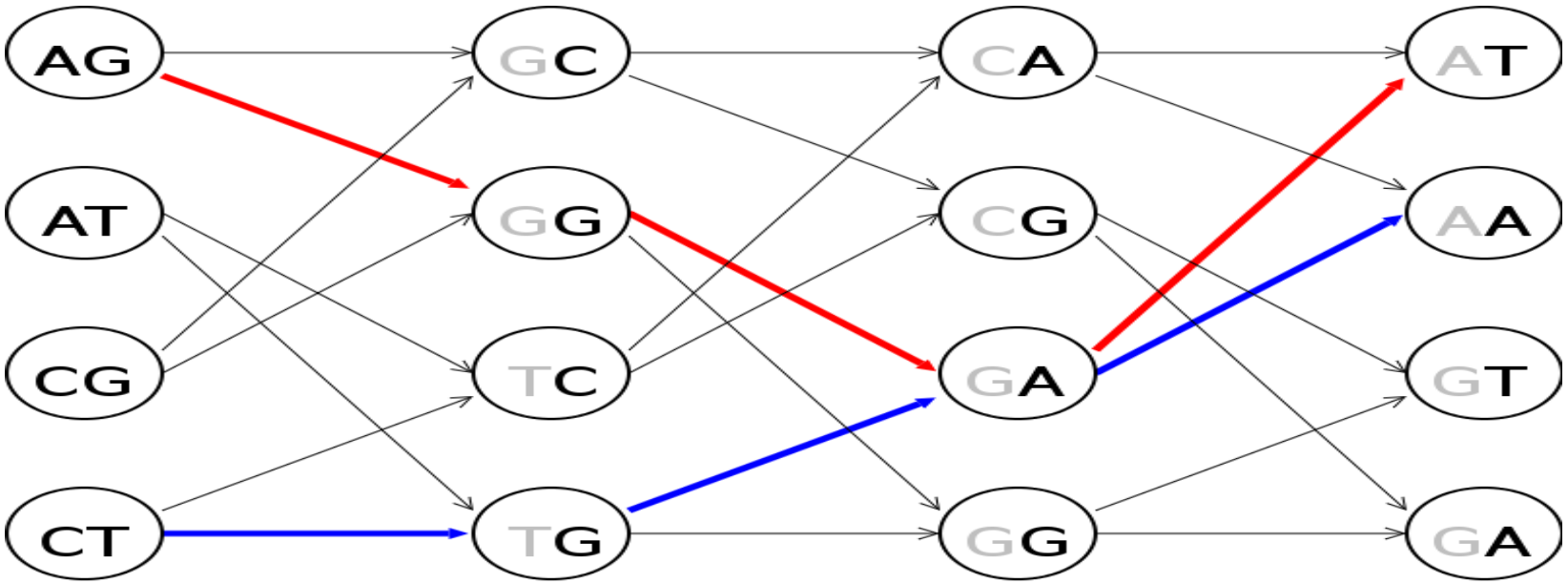

Genotype: {A,C}, {G,T}, G, A, {T,A}
Configuration 1:
AGGAT 0.324
CTGAA 0.036
0.011664
48.65%
Configuration 2:
AGGAA 0.036
CTGAT 0.324
0.011664
48.65%
Configuration 3:
ATGAT 0.0405
CGGAA 0.008
0.000324
1.35%
Configuration 4:
ATGAA 0.0045
CGGAT 0.072
0.000324
1.35%

多阶MC模型

$$P(H) \approx P(H(1,d)) \prod_{i=d+1,n} P(H(i)|H(i-d,i-1))$$

$$P(H) \approx fr(H(1,d)) \prod_{i=d+1,n} \frac{fr(H(i-d,i))}{fr(H(i-d,i-1))}$$

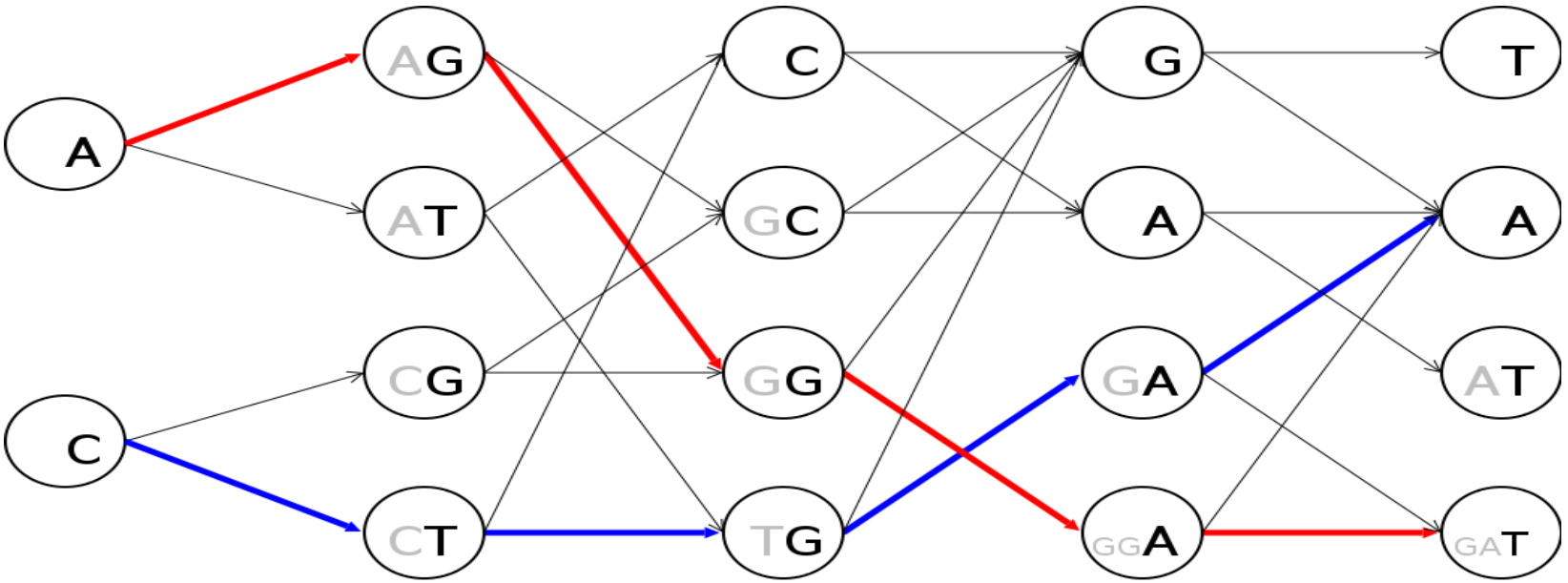
2阶MC例子



Genotype: {A,C}, {G,T}, G, A, {T,A}

Configuration 1: AGGAT
CTGAA

可变阶MC例子



Genotype: {A,C}, {G,T}, G, A, {T,A}

Configuration 1: AGGAT
CTGAA

片断的长度与频率

- 长的片断
  - 更多的相关信息
  - 较低的频率，不容易估计准确
- 短的片断
  - 较高的频率，估计较准确
  - 较少的相关信息，尤其是长范围的相关信息
- 片断长度与频率之间的tradeoff

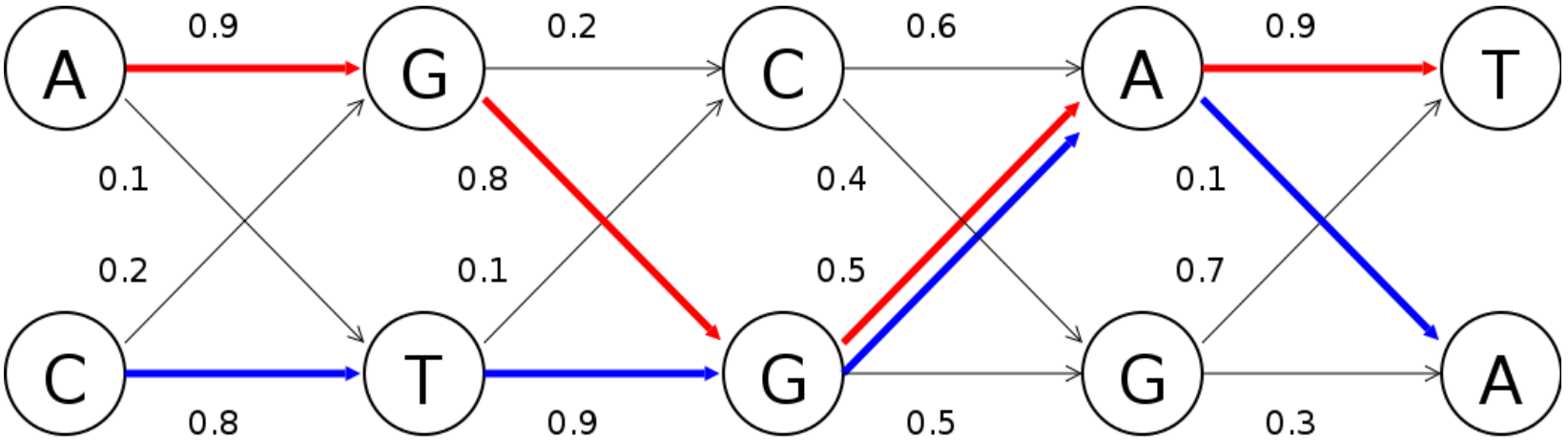
最优解

- Given the parameters of Markov chain, how to find the best haplotype pair for each individual?
- The number of possible haplotype configurations for a genotype grows exponentially with the number of heterozygous markers
- Exhaustive search is infeasible for large problem

算法

- Partition-Ligation algorithm
  - Eronen *et al.* 2004
  - Heuristic algorithm
- Dynamic Programming algorithm
  - Exact algorithm
  - Quadratic time complexity (Linear time complexity after improvement) with fixed model paremeters
  - Give the condition under which the DP algorithm can be applied to multiple order Markov Chain model

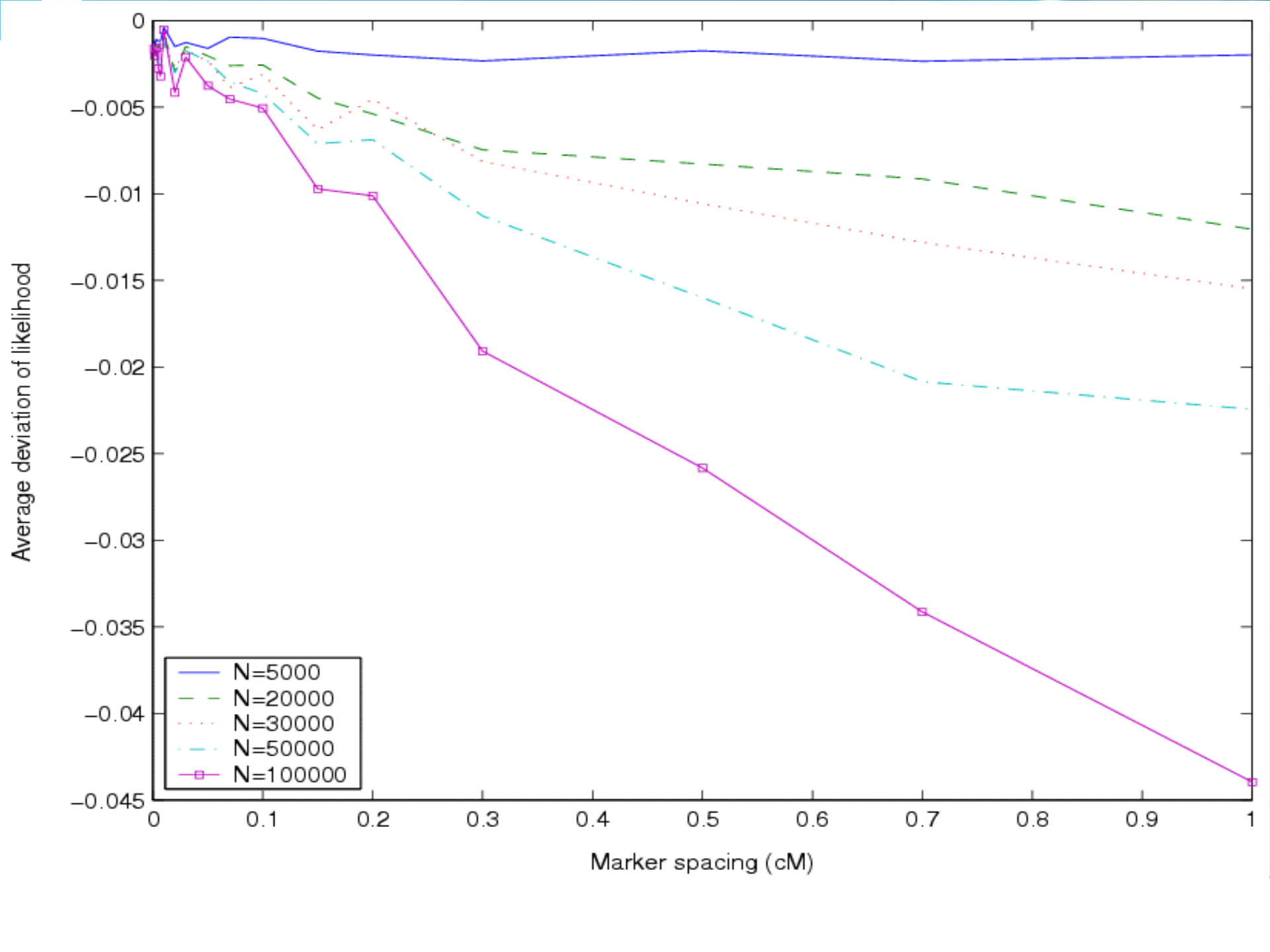
动态规划

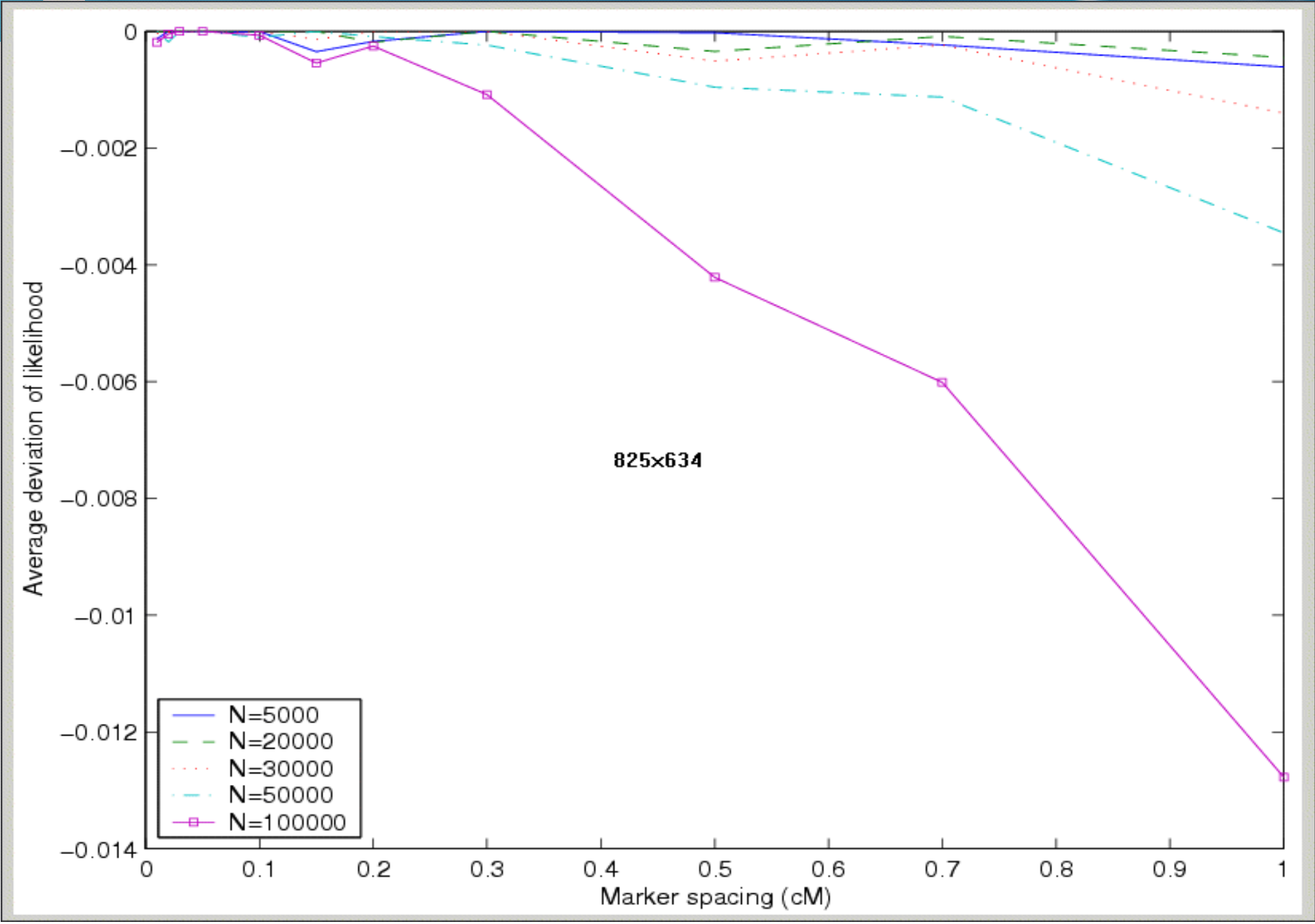


Genotype: {A,C}, {G,T}, G, A, {T,A}

Configuration 1:

AGGAT	0.324
CTGAA	0.036


Average deviation of likelihood
0
−0.005
−0.01
−0.015
−0.02
−0.025
−0.03
−0.035
−0.04
−0.045
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Marker spacing (cM)
N=5000
N=20000
N=30000
N=50000
N=100000


0
−0.002
−0.004
−0.006
−0.008
−0.01
−0.012
−0.014
Average deviation of likelihood
825×634
N=5000
N=20000
N=30000
N=50000
N=100000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Marker spacing (cM)

Data	Marker	N=5000		N=20000		N=30000		N=50000		N=100000	
name	spacing	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC
A_0.001_2	0.001	0.2974	**0.2932**	0.0326	**0.0298**	0.0228	**0.0178**	0.019	**0.0142**	0.0146	**0.0066**
A_0.002_2	0.002	0.3788	**0.3776**	0.0624	**0.0568**	0.0404	**0.037**	0.0264	**0.0198**	0.0232	**0.0152**
A_0.003_2	0.003	0.313	**0.3074**	0.0386	**0.035**	0.0328	**0.0246**	0.0272	**0.0218**	0.022	**0.019**
A_0.005_2	0.005	0.3862	**0.3846**	0.0662	**0.0574**	0.0456	**0.0368**	0.0366	**0.0272**	0.0326	**0.0192**
A_0.007_2	0.007	0.3178	**0.316**	0.0702	**0.062**	0.0492	**0.0386**	0.0352	**0.0262**	0.0278	**0.0178**
A_0.01_2	0.01	0.2708	**0.2676**	0.0378	**0.036**	0.0348	**0.0286**	0.028	**0.0224**	0.0262	**0.0246**
A_0.02_2	0.02	0.3876	**0.3822**	0.0772	**0.0664**	0.0626	**0.0514**	0.0584	**0.0466**	0.0496	**0.0382**
A_0.03_2	0.03	0.4074	**0.406**	0.0856	**0.0824**	0.0662	**0.064**	0.0718	**0.0688**	**0.0698**	0.0704
A_0.05_2	0.05	0.4796	**0.471**	0.1406	**0.1378**	**0.131**	0.132	**0.123**	0.1276	**0.1328**	0.139
A_0.07_2	0.07	**0.4956**	0.4978	0.1826	**0.1758**	0.1646	**0.161**	0.1684	**0.164**	0.1858	**0.1804**
A_0.10_2	0.1	0.5368	**0.5338**	0.2496	**0.243**	0.237	**0.2282**	0.2362	**0.2316**	0.2596	**0.2588**
A_0.15_2	0.15	0.6982	**0.6958**	0.4026	**0.4008**	**0.3896**	0.3944	**0.429**	0.4322	**0.4908**	0.4982
A_0.20_2	0.2	0.8682	**0.861**	0.5694	**0.5686**	0.5736	**0.563**	0.616	**0.615**	**0.688**	0.6924
A_0.30_2	0.3	1.1264	**1.12**	0.936	**0.9292**	**0.9428**	0.9474	**1.0364**	1.0418	**1.175**	1.1912
A_0.50_2	0.5	1.8244	**1.8202**	**1.6996**	1.7066	1.771	**1.7678**	**1.881**	1.9012	**2.0914**	2.1368
A_0.70_2	0.7	2.405	**2.3994**	**2.4426**	2.4516	**2.508**	2.53	**2.7026**	2.7158	**2.952**	2.9896
A_1.00_2	1	3.0916	**3.0902**	**3.2802**	3.3022	**3.3964**	3.405	**3.5722**	3.607	**3.8634**	3.9142

Table 1: Average switch distance of solutions obtained by HMC and haplorec (SNPs)

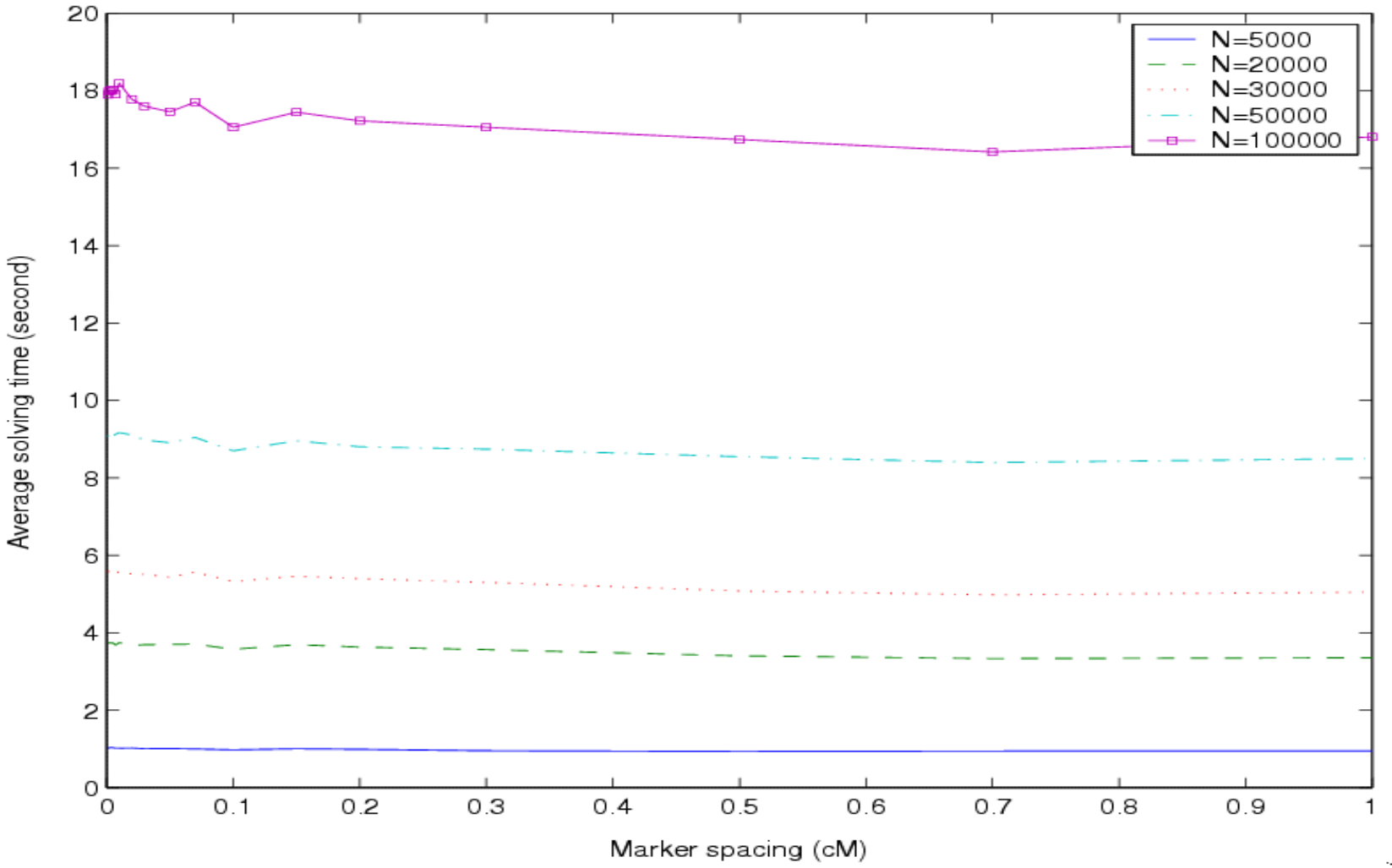
Data	Marker	N=5000		N=20000		N=30000		N=50000		N=100000	
name	spacing	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC
A_0.01_6	0.01	**0.0732**	0.0746	**0.0054**	**0.0054**	**0.0056**	**0.0056**	**0.0064**	**0.0064**	0.01	**0.0086**
A_0.02_6	0.02	**0.0554**	0.0558	0.0094	**0.0092**	0.0098	**0.0092**	0.0126	**0.0112**	**0.0092**	0.0104
A_0.03_6	0.03	**0.0516**	**0.0516**	0.0116	**0.0114**	**0.0104**	0.0106	**0.0094**	**0.0094**	**0.0118**	**0.0118**
A_0.05_6	0.05	**0.0758**	**0.0758**	0.0172	**0.017**	0.02	**0.0198**	0.0198	**0.0194**	**0.0196**	**0.0196**
A_0.10_6	0.1	**0.1184**	**0.1184**	**0.0416**	0.0424	**0.0444**	**0.0444**	**0.051**	0.0518	**0.068**	0.0684
A_0.15_6	0.15	**0.158**	0.1588	**0.0796**	0.0798	**0.0858**	0.086	0.096	**0.0952**	**0.1266**	0.1298
A_0.20_6	0.2	0.2186	**0.2176**	0.1286	**0.1274**	**0.129**	0.1296	**0.1514**	0.1524	**0.1916**	0.1928
A_0.30_6	0.3	0.2968	**0.2908**	**0.2064**	0.2076	0.2168	**0.215**	**0.245**	0.247	**0.3342**	0.3392
A_0.50_6	0.5	0.564	**0.5604**	**0.468**	0.4702	**0.5176**	0.5186	**0.615**	0.6162	**0.8426**	0.8598
A_0.70_6	0.7	0.8914	**0.8874**	0.7936	**0.7916**	**0.8824**	0.8854	**1.0338**	1.0404	**1.3978**	1.42
A_1.00_6	1	**1.32**	1.3222	**1.2942**	1.2952	1.4254	**1.425**	**1.6374**	1.6528	**2.182**	2.221

Table 2: Average switch distance of solutions obtained by HMC and haplorec (microsatellites)

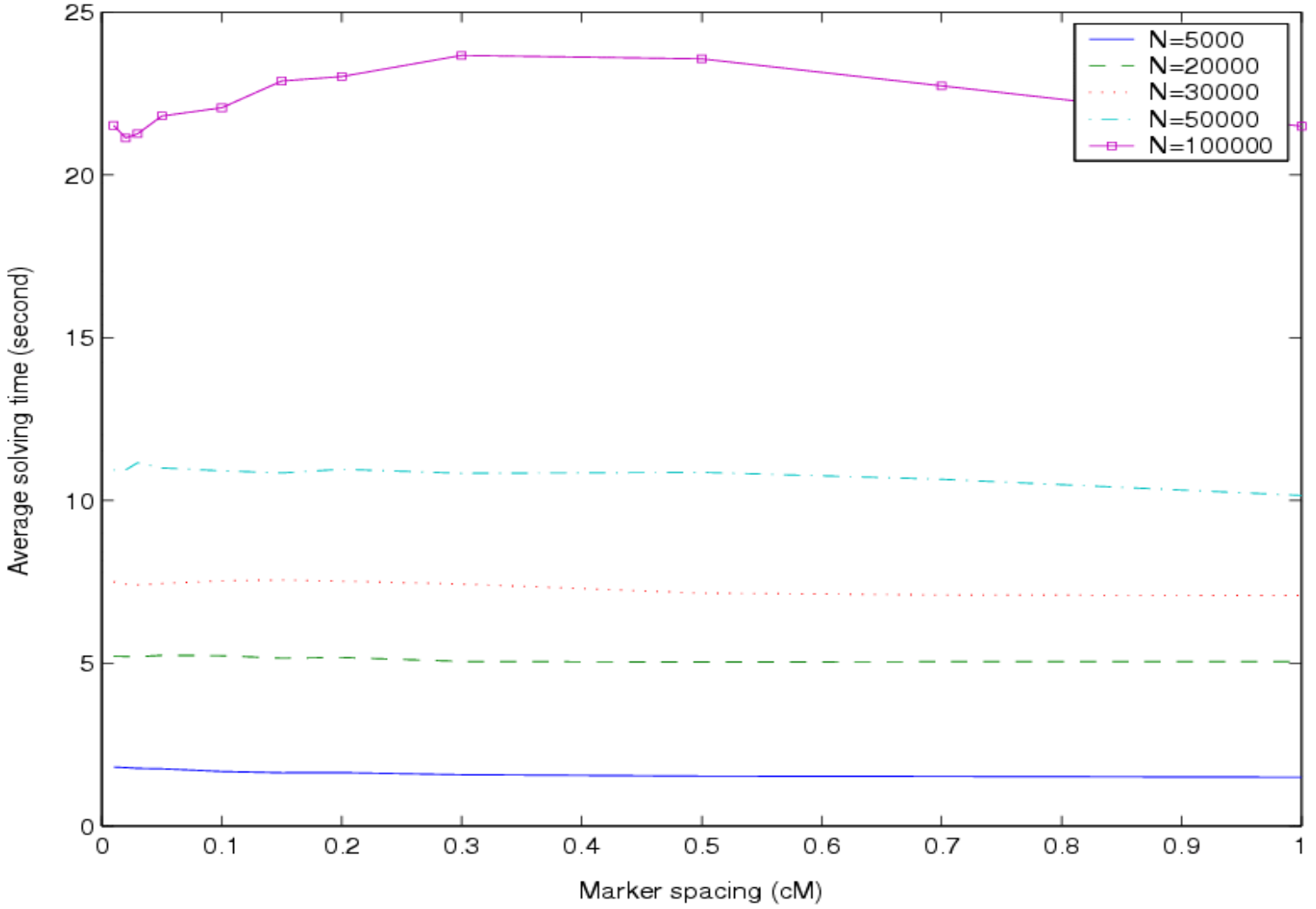
	N=5000		N=20000		N=30000		N=50000		N=100000	
	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC	haplorec	HMC
Switch Distance	1.210884	**1.18367**	0.938776	**0.82313**	0.897959	**0.81633**	0.904762	**0.84354**	0.979592	**0.89116**
Accuracy	**0.503401**	0.489796	0.544218	**0.557823**	0.537415	**0.557823**	0.551020	**0.557823**	**0.551020**	0.544218

Table 3: Comparison of solutions on the Daly data set

计算时间

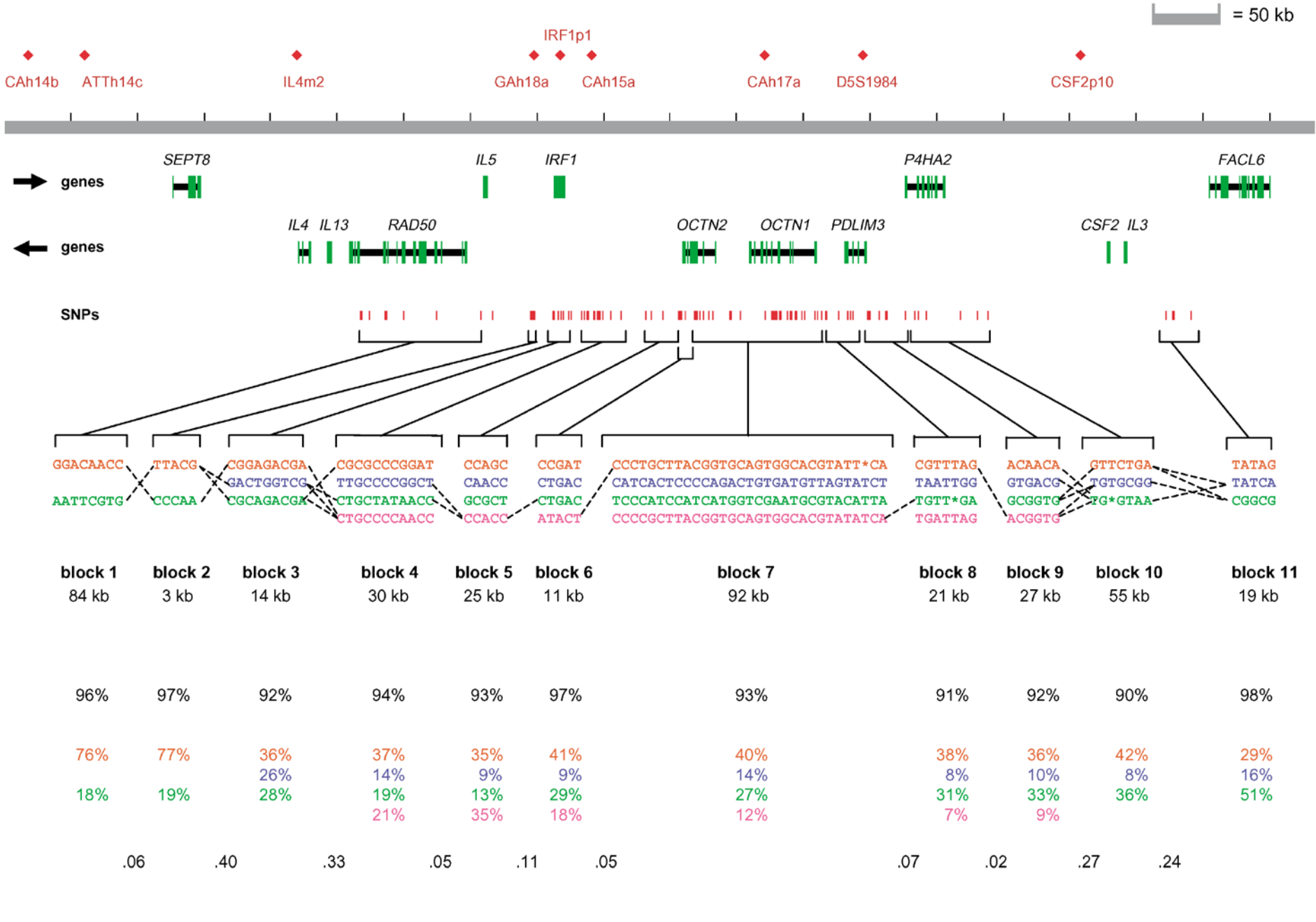

N=5000
N=20000
N=30000
N=50000
N=100000
Average solving time (second)
Marker spacing (cM)
0
2
4
6
8
10
12
14
16
18
20
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

计算时间

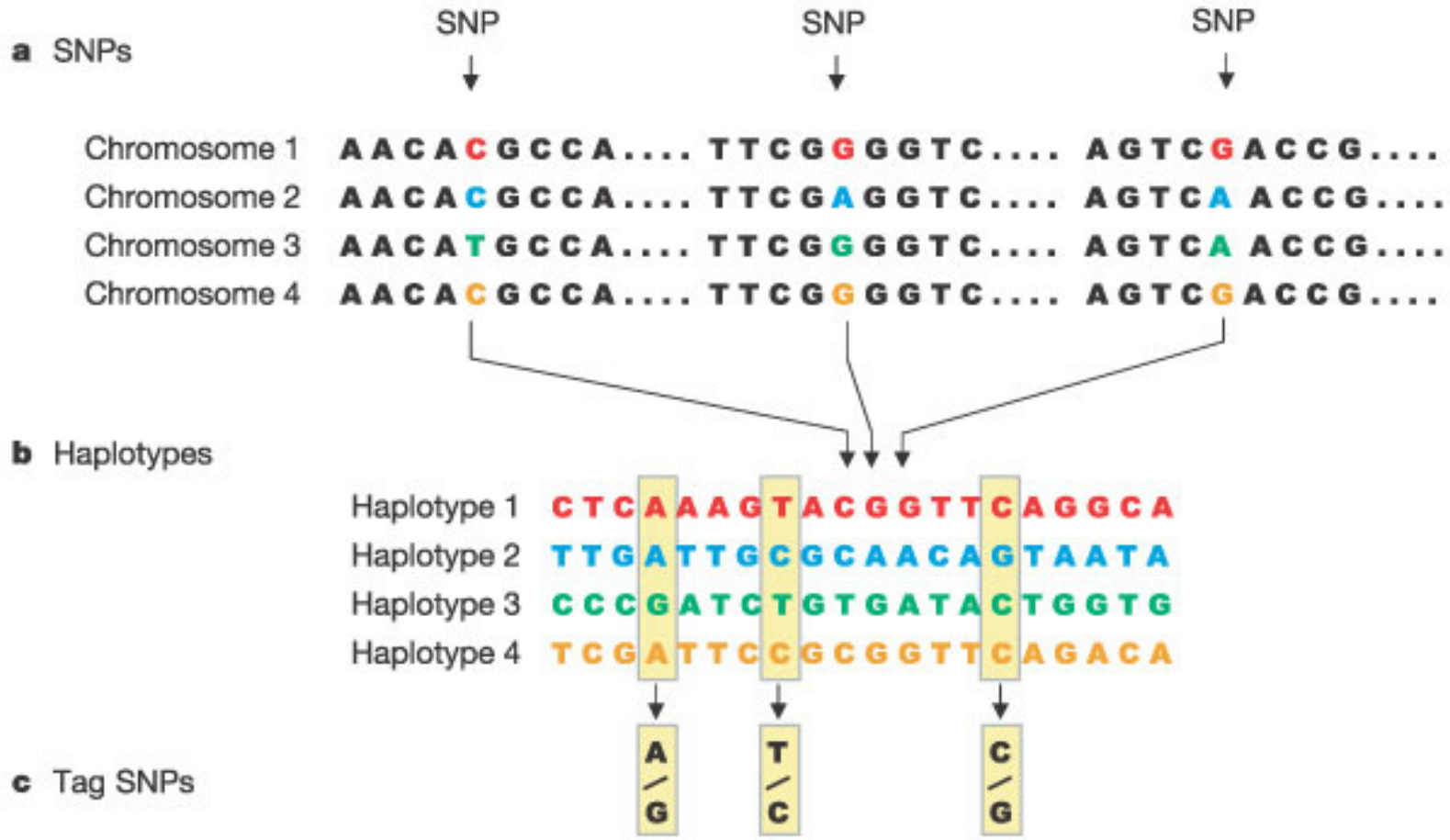

N=5000
N=20000
N=30000
N=50000
N=100000
Average solving time (second)
Marker spacing (cM)
0
5
10
15
20
25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

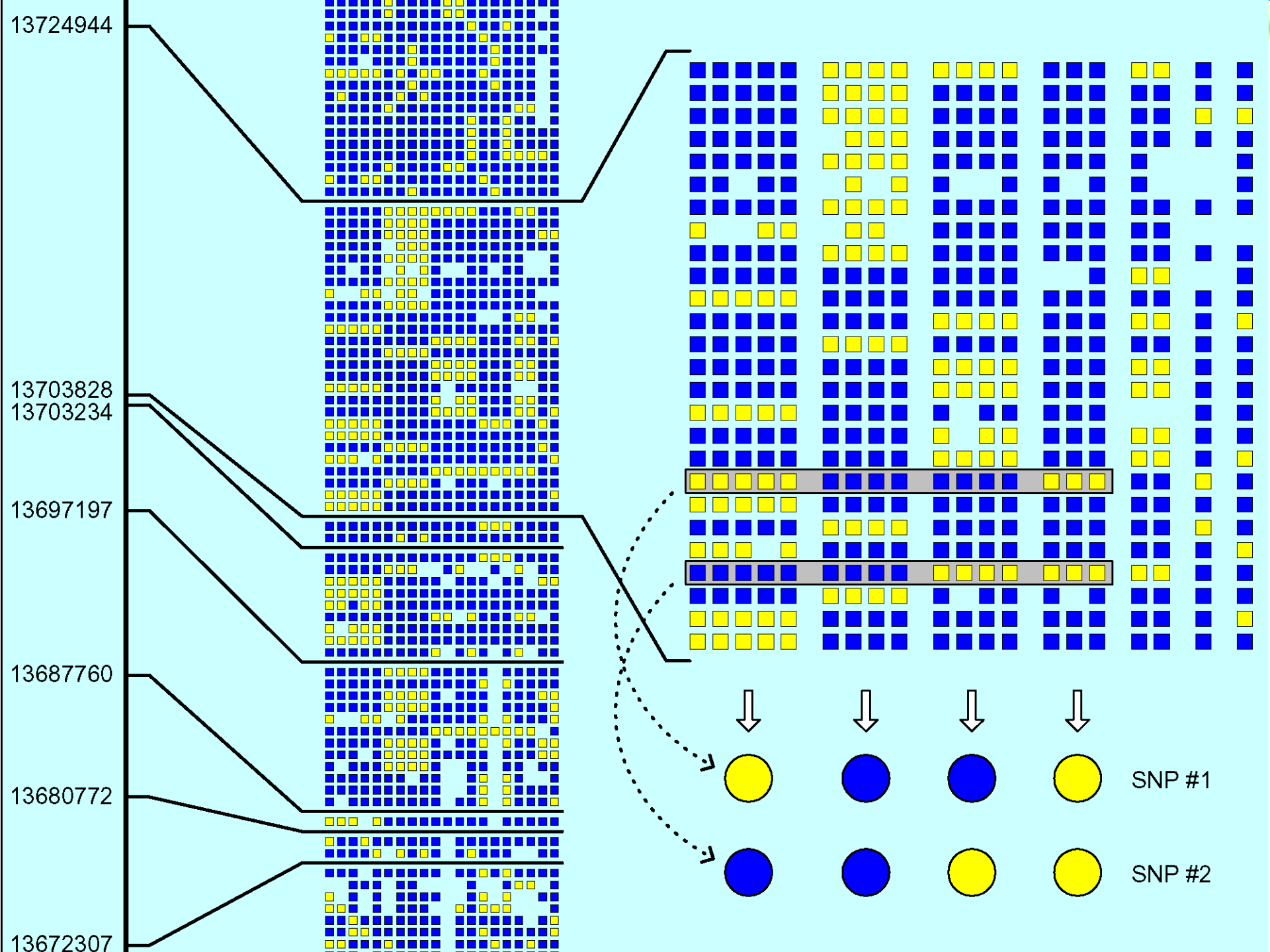
单体型相关问题

- 单体型组装问题
- 单体型推断问题
- 单体型分块问题
- 标签SNP选取问题


= 50 kb
IRF1p1
CAh14b
ATTh14c
IL4m2
GAh18a
CAh15a
CAh17a
D5S1984
CSF2p10
SEPT8
IL5
IRF1
P4HA2
FACL6
genes
IL4
IL13
RAD50
OCTN2
OCTN1
PDLIM3
CSF2
IL3
genes
SNPs
a
GGACAACC
TTACG
CGGAGACGA
CGCGCCCGGAT
CCAGC
CCGAT
CCCTGCTTACGGTGCAGTGGCACGTATT*CA
CGTTTAG
ACAACA
GTTCTGA
TATAG
AATTCGTG
CCCAA
CGCAGACGA
TTGCCCCGGCT
CAACC
CTGAC
CATCACTCCCCAGACTGTGATGTTAGTATCT
TAATTGG
GTGACG
TGTGCGG
TATCA
CTGCTATAACC
GCGCT
CTGAC
TCCCATCCATCATGGTCGAATGCGTACATTA
TGTT*GA
GCGGTG
FG*GTAA
CGGCG
CTGCCCCAACC
CCACC
ATACT
CCCCGCTTACGGTGCAGTGGCACGTATATCA
TGATTAG
ACGGTG
block 1
84 kb
block 2
3 kb
block 3
14 kb
block 4
30 kb
block 5
25 kb
block 6
11 kb
block 7
92 kb
block 8
21 kb
block 9
27 kb
block 10
55 kb
block 11
19 kb
b
96%
97%
92%
94%
93%
97%
93%
91%
92%
90%
98%
c
76%
77%
36%
37%
35%
41%
40%
38%
36%
42%
29%
26%
14%
9%
9%
14%
8%
10%
8%
16%
18%
19%
28%
19%
13%
29%
27%
31%
33%
36%
51%
21%
35%
18%
12%
7%
9%
d
.06
.40
.33
.05
.11
.05
.07
.02
.27
.24

标签SNP


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G


13724944
13703828
13703234
13697197
13687760
13680772
13672307
SNP #1
SNP #2

标签SNP的定义

- Tag SNPs
  - A small subset of SNPs which is sufficient to distinguish each pair of haplotype patterns
- Robust Tag SNPs
  - A subset of SNPs which can still distinguish all distinct haplotypes even when some SNPs are missing

标签SNP选取问题

- Minimum Tag SNPs
  - Find the minimum set of tag SNPs
- Minimum Auxiliary Tag SNPs
  - Find the minimum subset of additional SNPs to resolve the ambiguity caused by missing data
- Minimum Robust Tag SNPs
  - Find the minimum set of robust tag SNPs which are able to tolerate a number of missing SNPs

辅助标签SNPs

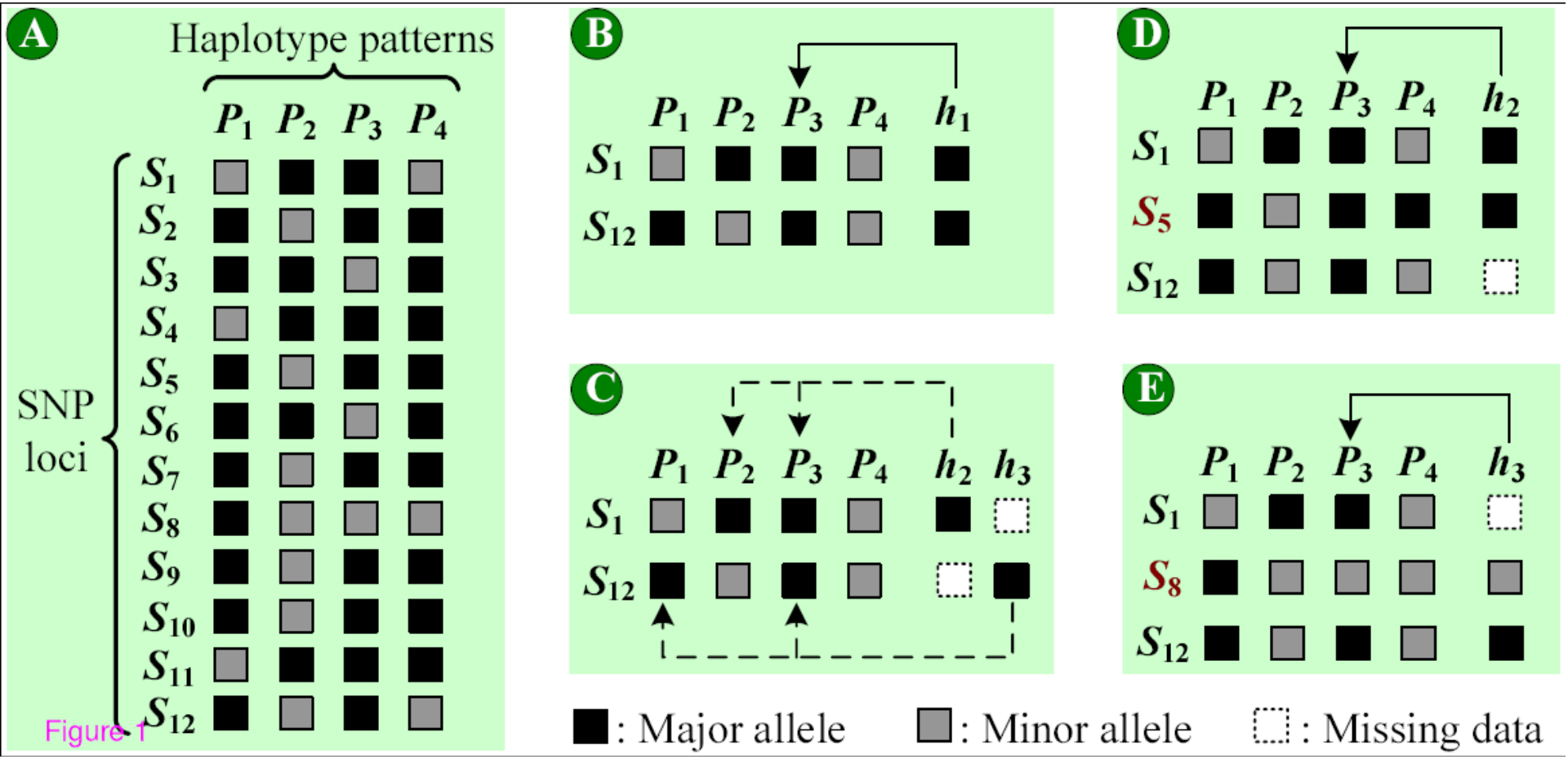


Figure 1

Robust Tag SNPs

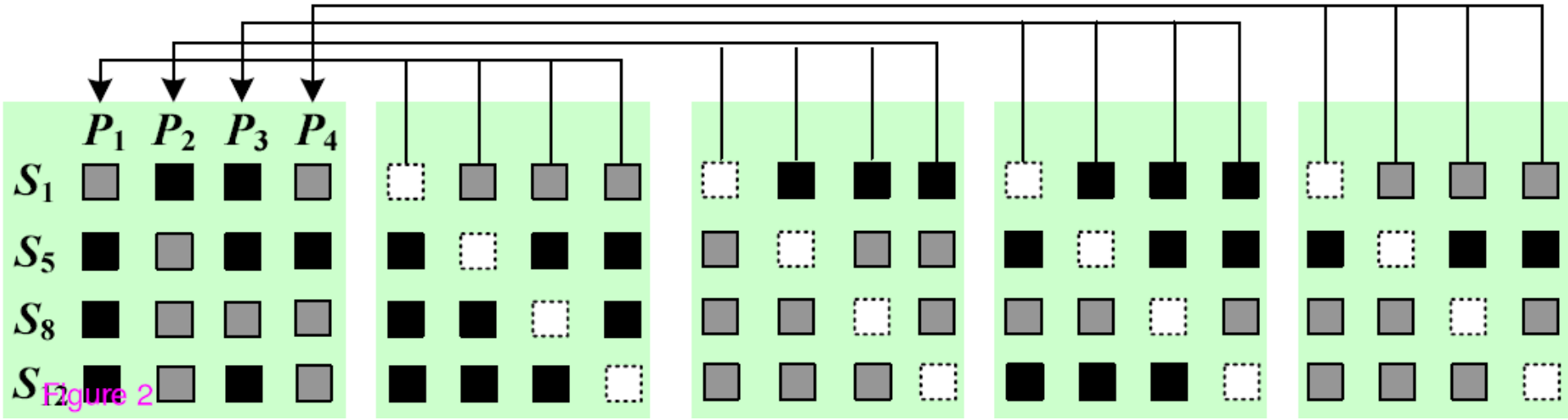


Figure 2

集覆盖问题

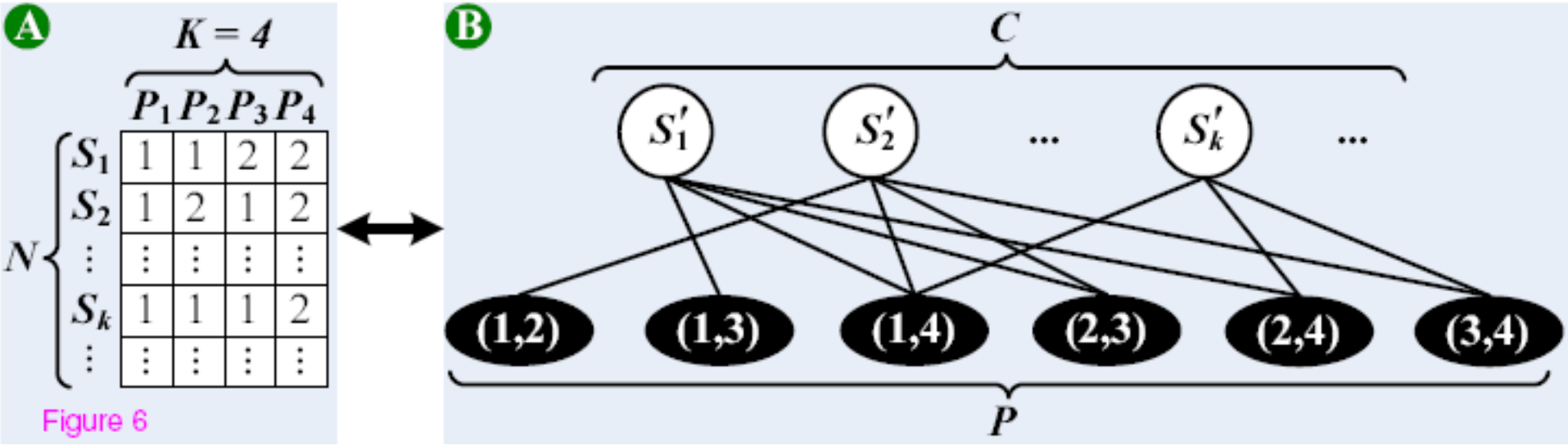


Figure 6

计算复杂性

- Minimum Tag SNPs
  - Equivalent to Minimum Set Cover Problem
  - NP-hard
- Minimum Robust Tag SNPs
  - Equivalent to Minimum *k*-Redundent Coverage
  - NP-hard