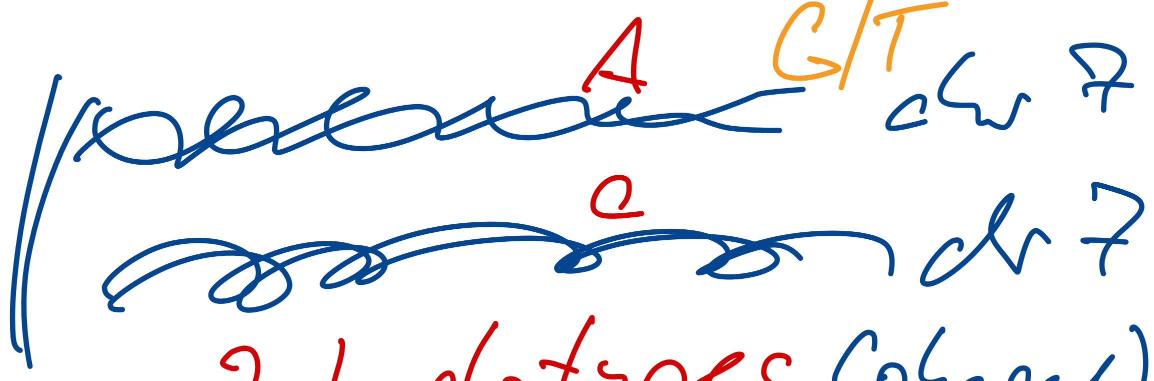
# Haplotype Inference

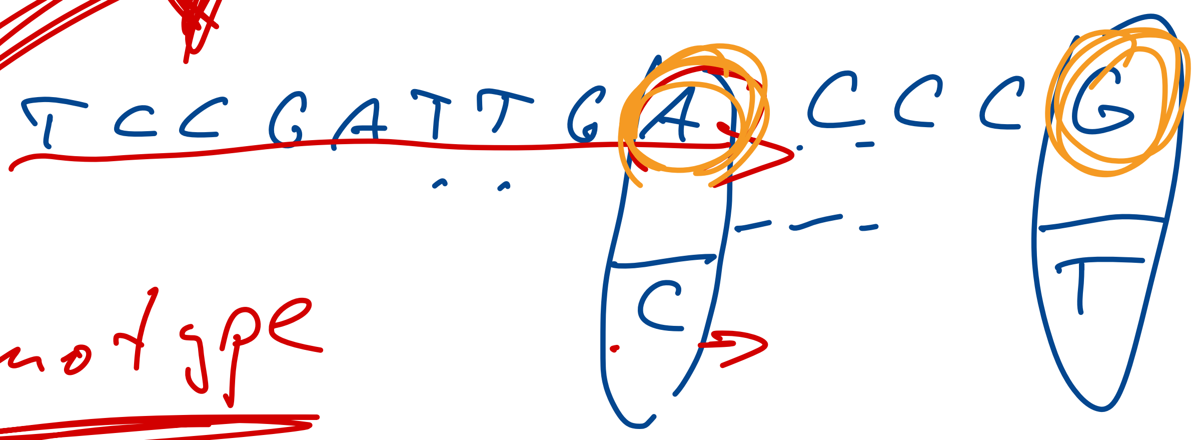
22.1.2021

Idea:

diploid



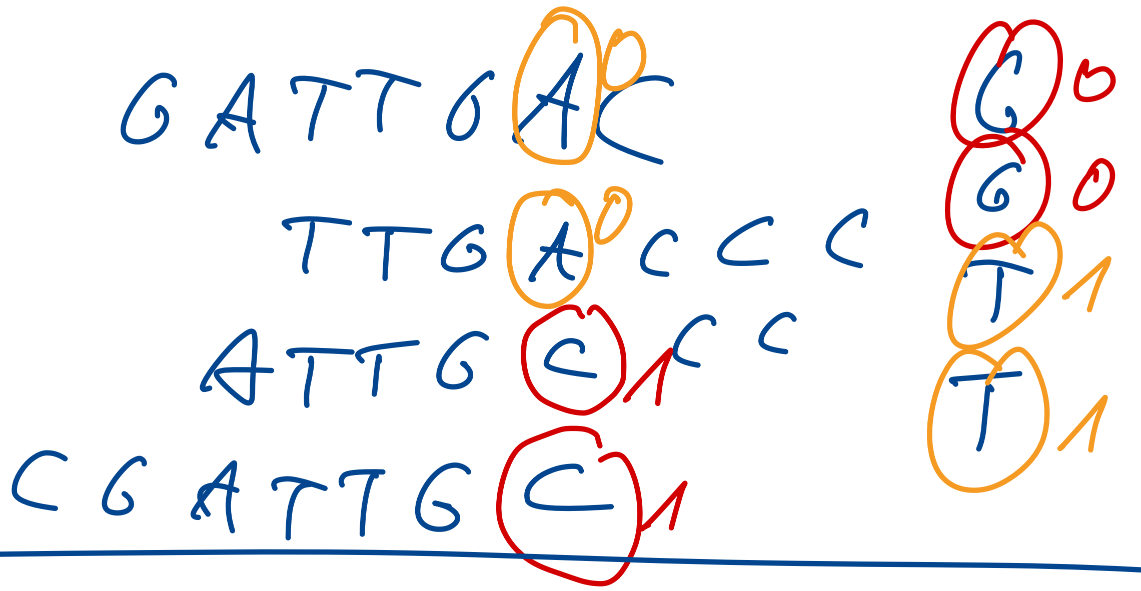
2 haplotypes (phased)



1 genotype

(unphased)

segregating site
= heterozygous allele



Assumption: haplotypes are binary

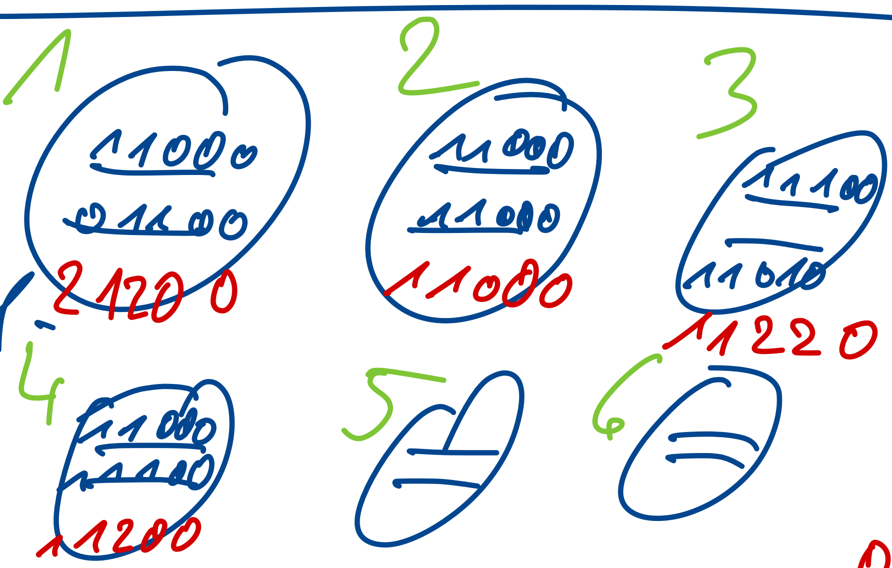
⟺ encoded by 0/1

3 approaches:

1) population based haplotyping
2) genetic haplotyping
3) molecular haplotyping

ad 1)

6 individuals



haplotypes 0/1-Seq.

genotypes 0/1/2 both

easy: haplotypes → genotype

hard: genotype → haplotypes !

approach:

1) look for individuals with 0 or 1 heterozygous sites
→ here haplotyping is trivial

2) use some of the resolved haplotypes to resolve others.

1) indiv. 2 → 11000

indiv. 4 → ~~11000~~

11100

2) 11000 + indiv.3 : 11220

↪ 11110

11100 + ind.1 : 21200

↪ 01000

OR

11000 + ind1 = 21200

↪ 01100

Problem: we can get different solutions

ad 2) like above, but with pedigree information.

Example 1:

greens: inferred haplotypes

parent 1 : 201
parent 2 : 022
child : 221 → $\boxed{101}$ $\boxed{011}$

→ parent1: 101 , 001
parent2: 011 , 000

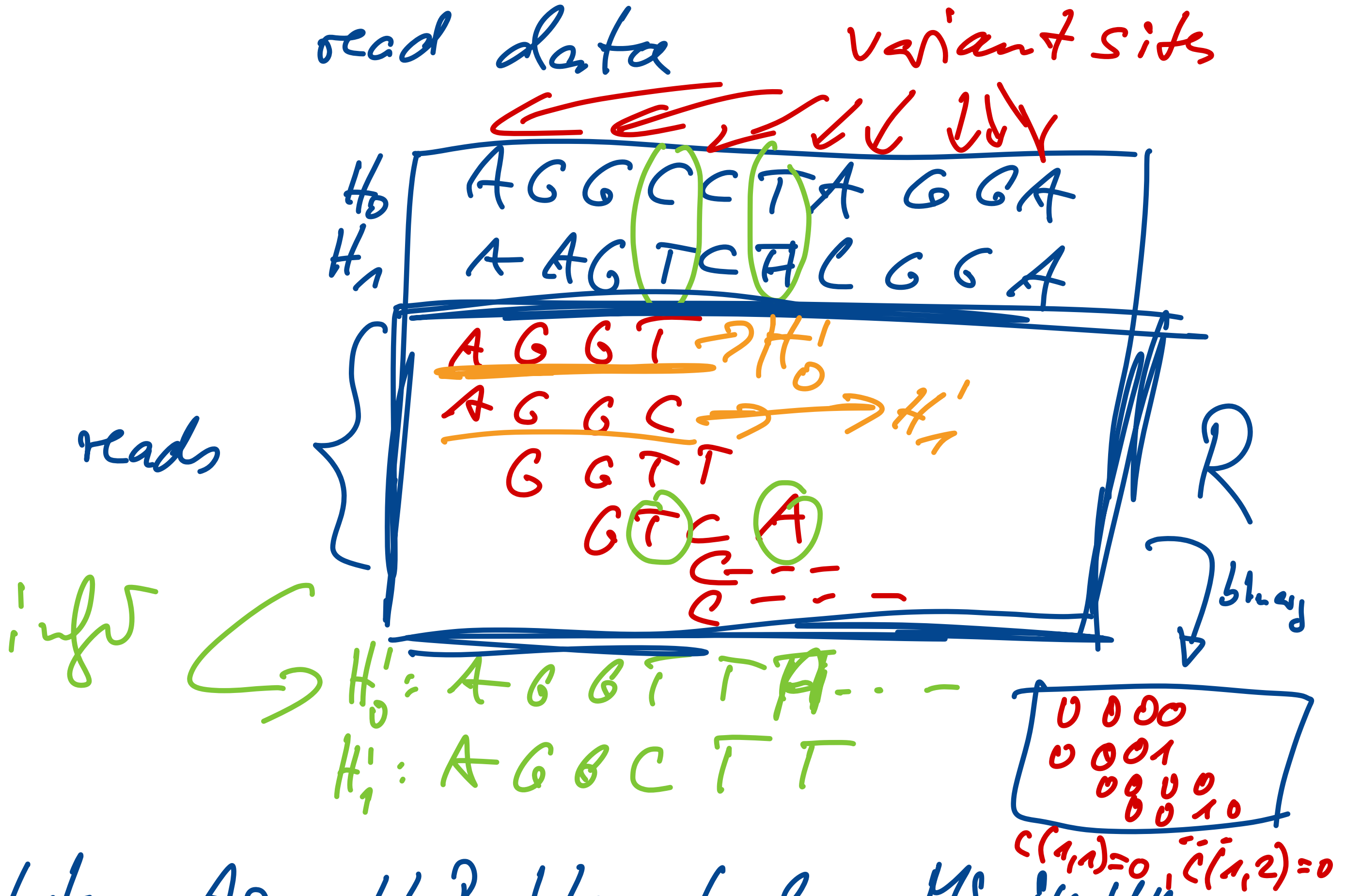
✓

Example 2:

parent 1 : 222
parent 2 : 222
child : 222

↪ nothing can be inferred

## ad 3) Haplotype Inference from sequencing read data



Haplotype Assembly Problem (a.k.a. Minimum Error Correction Problem, MEC)

Input: $n \times m$ fragment matrix R (usually binary)

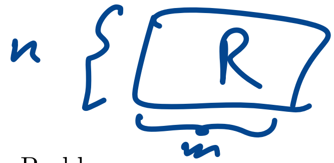
Def: conflict of two fragments if they are unequal at some site

Def: Let $d(A,B)$ be the number of corrections to transform fragment matrix A into fragment matrix B

Problem: Given a fragment matrix R, find a fragment matrix R' that has a conflict-free bipartition, minimizing $d(R, R')$.

Result: MEC is NP-complete.
(reduction from MAXCUT.)

Solution 1:



(ILP)

An Integer Linear Program solving the Haplotype Assembly Problem:

$$\text{Minimize} \sum_{j=1}^{j=n} z(j)$$

z(j) = mismatch counter for read j

Such that:
For each *Read* $j$ from 1 to $n$, and each position $k$ from 1 to $m$:

$c(j,k)$ : input R

$z_0$ : like z, assuming that read j is assigned to $H_0$

$$z_0(j,k) \geq c(j,k) - H_0'(k)$$
$$z_0(j,k) \geq H_0'(k) - c(j,k)$$

$z_1$ : like z, assuming that read j is assigned to $H_1$

$$z_1(j,k) \geq c(j,k) - H_1'(k)$$
$$z_1(j,k) \geq H_1'(k) - c(j,k)$$

$$z_0(j,k) - A(j) - z(j,k) \leq 0$$
$$z_1(j,k) + A(j) - z(j,k) \leq 1$$

$A(j) = \begin{cases} 0 & \text{if read } j \text{ is in } H_0' \\ 1 & \text{if read } j \text{ is in } H_1' \end{cases}$

For each *Read* $j$:

$$z(j) = \sum_{k=1}^{k=m} z(j,k)$$

$z(j,k)$ : mismatch indicator for read j in site k in assigned haplotype

$H_0'$, $H_1'$ the two haplotypes

All variables are binary

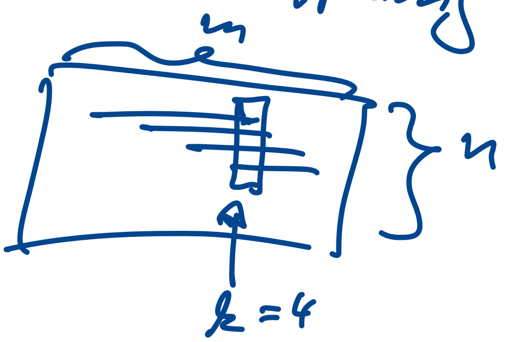
Solution 2: WhatsHap (Patterson et al. 2015)

dynamic programming solution along the columns of R.

analysis: $O(2^{k-1} \cdot m)$ time

where k is the maximum coverage at any site.

R:



→ FPT algorithm