

Clark's and EM Algorithm for Haplotype Inference

1. Problem 1

Can you deduce the haplotypes for the following four genotypes using Clark's algorithm?

1. $G_1 = 122101$
2. $G_2 = 210101$
3. $G_3 = 022121$
4. $G_4 = 201121$

Solution to Problem 1

We are given four genotypes and want to determine a minimal set of haplotypes that can explain them using Clark's algorithm. Clark's algorithm starts from a homozygous site to identify a known haplotype and then uses it to infer others.

Step 1: Start with $G_1 = 122101$

G_1 has some heterozygous positions (represented as '2'). Possible haplotype combinations for G_1 include:

1. $100101 + 022000$
2. **110101** + 012000
3. **111101** + 011000
4. **101101** + 021000

Assume we observe haplotype **H1 = 110101** from the second combination. This allows us to use it to deduce other genotypes.

Step 2: Use H1 to resolve $G_2 = 210101$

Using the known haplotype $H_1 = 110101$, we subtract it from G_2 :

$$210101 - 110101 = \mathbf{010101}$$

So we infer a new haplotype **H3 = 010101**.

Step 3: Use H3 to resolve G3 = 022121

Now using H3, subtract it from G3:

$$022121 - 010101 = 001111$$

We derive another haplotype: $H4 = 001111$.

Step 4: Use H4 to resolve G4 = 201121

With H4, subtract from G4:

$$201121 - 001111 = 101100$$

We get a final haplotype: $H5 = 101100$.

Conclusion for Problem 1

Clark's algorithm resolves the genotypes using the following set of haplotypes:

$$H = \{H1 = 110101, H2 = 101101, H3 = 010101, H4 = 001111, H5 = 101100\}$$

2. Problem 2

Consider the following set of genotypes:

1. $G_1 = 1012$
2. $G_2 = 2220$
3. $G_3 = 1020$
4. $G_4 = 2210$

- (a) Suppose we resolve genotypes in order, starting from the smallest index. What haplotypes are generated?
- (b) Can we do better by changing the order of resolution?

Solution to Problem 2

(a) Resolving in Order (Greedy Strategy)

We start with $G_1 = 1012$. It can be resolved as:

$$H1 = 1010, \quad H2 = 1011$$

Next, use $H1$ to resolve $G_2 = 2220$:

$$G_2 - H1 = H3 = 0100$$

Now resolve $G_3 = 1020$ using $H3$:

$$G_3 - H3 = H4 = 1000$$

Then resolve $G_4 = 2210$ using $H4$:

$$G_4 - H4 = H5 = 0110$$

Greedy Haplotypes Generated:

$$H = \{H1 = 1010, H2 = 1011, H3 = 0100, H4 = 1000, H5 = 0110\}$$

(b) Improved Strategy by Reordering

Clark's algorithm is sensitive to the order of genotype resolution. Let's resolve G_1 using $H2 = 1011$ first.

1. $H1 = 1011$
2. $H2 = 1010$

Now resolve G3 using H2:

$$G3 - H2 = \textcolor{red}{H3} = 1000$$

Then resolve G4 using H3:

$$G4 - H3 = \textcolor{red}{H4} = 0110$$

Optimized Haplotypes:

$$H = \{\textcolor{red}{H1} = 1011, \textcolor{red}{H2} = 1010, \textcolor{red}{H3} = 1000, \textcolor{red}{H4} = 0110\}$$

Conclusion for Problem 2

We reduced the total number of haplotypes from 5 to 4 by changing the order of resolution. This demonstrates that Clark's algorithm does not always yield an optimal solution unless order is considered carefully.

3. Problem 3

Can you deduce the haplotypes for the following three genotypes using the EM algorithm? Assume uniform initial haplotype frequencies and perform only one EM iteration.

1. $G_1 = 122$
2. $G_2 = 210$
3. $G_3 = 212$

Solution to Problem 3

Initial Setup

Let the candidate haplotypes be:

$$H = \{h_1 = 100, h_2 = 111, h_3 = 110, h_4 = 101, h_5 = 010, h_6 = 011\}$$

Initial frequencies: $F_x^{(0)} = \frac{1}{6}$ for all $x = 1 \dots 6$

Expectation Step (Estimate Pair Probabilities)

Using all valid haplotype pairs, we compute:

G1 = 122: Two compatible haplotype pairs:

$$\Pr(h_1, h_2) = \frac{1}{2}, \quad \Pr(h_3, h_4) = \frac{1}{2}$$

G2 = 210:

$$\Pr(h_3, h_5) = 1$$

G3 = 212:

$$\Pr(h_2, h_5) = \frac{1}{2}, \quad \Pr(h_3, h_6) = \frac{1}{2}$$

Maximization Step (Update Frequencies)

Number of genotypes: $n = 3$. Each genotype contributes a total of 2 haplotypes. Frequency is calculated as:

$$F_x^{(1)} = \frac{1}{2n} \sum \Pr(\text{haplotype appears in a genotype})$$

$$\begin{aligned}
F_1^{(1)} &= \frac{1}{12} \\
F_2^{(1)} &= \frac{1}{6} \\
F_3^{(1)} &= \frac{1}{3} \\
F_4^{(1)} &= \frac{1}{12} \\
F_5^{(1)} &= \frac{1}{4} \\
F_6^{(1)} &= \frac{1}{12}
\end{aligned}$$

Re-estimate Most Likely Pairs

Using the new frequencies:

G1: $\Pr(h_3, h_4) > \Pr(h_1, h_2) \Rightarrow H_3 = 110, H_4 = 101$

G2: Only one option: $H_3 = 110, H_5 = 010$

G3: Most probable: $H_2 = 111, H_5 = 010$

Final Haplotypes from EM Algorithm

$$H = \{H_2 = 111, H_3 = 110, H_4 = 101, H_5 = 010\}$$