Advanced Algorithms in Bioinformatics– CSC 8540– Fall 2022

September 13 2022

Pardis Sadatian Moghaddam

Panther ID: 002722641

Problem 1

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

- G1 = 122101
- G2 = 210101
- G3 = 022121
- G4 = 201121

G = h + h'

In here, we consider an input set of 4 genotypes

G = {G1=122101, G2=210101, G3=022121, G4=201121}

G1	h	h'
122101	100101 110101	111101 101101

G2	h	h′
210101	010101	110101

G3	h	h′
022121	000101 010101 001101 000111 011101 001111	011111 001111 010111 011101 000101 010101

G4	h	h′
201121	001101 001110	101111 101100

H1 = 110101

H2 = 101101

G2 = 210101

Resolve G2 by H1 we have a new haplotype H3 = 010101

Resolve G3 by H3 we have a new haplotype H4 = 001111

Resolve G4 by H4 we have a new haplotype H5 = 101100

Finally, Clark's algorithm reports that G can be resolved by the set of haplotypes H = { H1 = 110101, H2 = 101101, H3 = 010101, H4 = 001111, H5 = 101100}.

Problem 2

Consider the following set of genotypes.

G1 : 1012

G2 : 2220

G3 : 1020

G4 : 2210

(a) Every iteration in Clark's algorithm resolves at least one Gi by including at most 1 new haplotype. Suppose we always try to resolve Gi with the smallest index i first. What is the set of haplotypes generated?

(b) Is the answer in (a) optimal? If not, can we resolve Gi in a different order to obtain the optimal solution?

G1	h	h′
1012	1010	1011

G2	h	h′
2220	0000	1110
	1000	0110
	0100	1010
	0010	1100
	1100	0010
	0110	1000

G3	h	h′
1020	1000	1010

G4	h	h′
2210	0010 1010	1110 0110

H1 = 1010

H2 = 1011

Resolve G2 by H1 we have a new haplotype H3 = 0100

Resolve G3 by H3 we have a new haplotype H4 = 1000

Resolve G4 by H4 we have a new haplotype H5 = 0110

Finally, Clark's algorithm reports that G can be resolved by the set of haplotypes H = {H1 = 1010, H2 = 1011, H3 = 0100, H4 = 1000, H5 = 0110}.

Clark's algorithm is order-dependent and this example can be solved with fewer haplotypes. We can have the final set with 4 haplotypes instead of 5 haplotypes. So, this answer is not optimal. We can try it and start it with H2:

H1 = 1011

H2 = 1010

Resolve G3 by H2 we have a new haplotype H3 = 1000

Resolve G4 by H3 we have a new haplotype H4 = 0110

Finally, Clark's algorithm reports that G can be resolved by the set of haplotypes H = {H1 = 1011, H2 = 1010, H3 = 1000, H4 = 0110}

Problem 3:

Can you deduce the haplotypes for the following three genotypes using the EM algorithm? We assume that the initial haplotype frequencies are uniform and we only execute the EM algorithm for one round.

- G1 =122
- G2 =210
- G3 =212

G1	h	h′
122	100 101	111 110

G2	h	h′
210	010	110

G3	h	h′
212	010 011	111 110

We have a G = {G1=122, G2=210, G3=212}. The set of all possible haplotypes of G is H = {h1=100, h2=111, h3=110, h4=101, h5=010, h6=011}

For x= 1,2,3,4,5,6 we have Fx(0) = 1/6

Expectation Step (estimate haplotype pair frequency)

For G1,

Pr (h1h2|G1, F(0)) = 1/2

Pr (h3h4|G1, F(0)) = 1/2

For G2,

Pr (h3h5|G2, F(0)) = 1

For G3,

Pr (h2h5|G3, F(0)) = 1/2

Pr (h3h6|G3, F(0)) = 1/2

Maximization Step (estimate haplotype frequency F(1))

F1 (1) = 1/2n * Pr (h1h2|G1, F(0)) = 1/12

F2 (1) = 1/2n * [Pr (h1h2|G1, F(0)) + Pr (h2h5|G3, F(0))] = 1/6

F3 (1) = 1/2n * [Pr (h3h4|G1, F(0)) + Pr (h3h5|G2, F(0)) + Pr (h3h6|G3, F(0))] = 1/3

F4 (1) = 1/2n * [Pr (h3h4|G1, F(0)) = 1/12

F5 (1) = 1/2n * [Pr (h3h5|G2, F(0)) + Pr (h2h5|G3, F(0)) = 1/4

F6 (1) = 1/2n * Pr (h3h6|G3, F(0)) = 1/12

Compute the haplotype frequencies for F(1):

For G1,

Pr (h1h2|G1, F(1)) = 1/3

Pr (h3h4|G1, F(1)) = 2/3

For G2,

Pr (h3h5|G2, F(1)) = 1

For G3,

Pr (h2h5|G3, F(1)) = 3/5

Pr (h3h6|G3, F(1)) = 2/5

We Choose genotype with the high probability

G1 => H3H4

G2 => H3H5

G3 => H2H5

So we have the final set of Haplotypes

H = {H2=111, H3=110, H4=101, H5=010}