CSE 280A final

Milad Mortazavi

March 2019

1. Equilibria
   The H_0 assumption is HW equilibrium.
   We assume the minor allele frequncy at location i is q_i and the major allele is $p_i = 1 - q_i$. We know that $q_i >= 0.1$. The probability of observing homozygosity in k consecutive loci is:
   $\prod_{i=1}^{k}(p_i^2 + q_i^2)$
   For each locus $0.1 <= q_i <= 0.5$. Therefore by ploting the second order polynomial we can have:
   $0.5 <= (p_i^2 + q_i^2) <= 0.82$
   The p-value is the probability of observing a more extreme case with our assumption H_0, therefore:
   $p - value = \prod_{i=k+1}^{\infty}(p_i^2 + q_i^2)$
   If we assume $(p_i^2 + q_i^2) = c$, and approximate that it is constant for each side we have:
   $p - value = c^{k+1}c^{k+2}... = c^{k+1}(1 + c + c^2 + ...) = \frac{c^k+1}{1-c}$
   where $0.5 <= c <= 0.82$. We want to find the value of k where $\frac{c^{k+1}}{1-c} < 10^{-4}$.
   When $c = 0.5$ on the lower end if $k >= 14$ we have $p - value < 10^{-4}$.
   When $c = 0.82$ on the higher end if $k >= 47$ we have $p - value < 10^{-4}$.
   So at the worst if $k >= 47$ the HW assumptions can be refuted.

2. Recombination

   (a) True. The intersection between i and j is not empty since we have at lease one (1,1) row. The intersection is not pure one of the columns due to having both (1,0) and (0,1) rows. Therefore, positions i and j must come from different parts of the phylogeny tree. Therefore, at least one recombination exist between loci i and j.

   (b) (a) If (j_3, j_4) satisfy the 4-gamete rule, there is at least one recombination between those. If (j_3, j_4) satisfy the 4-gamete rule, there is one recombination between those also. So at least two recombinations between j_1 and j_4 exist.

(b) If (j_1, j_3) satisfy the 4-gamete rule, there is at least one recombination between those. If (j_2, j_4) satisfy the 4-gamete rule, there is also one recombination between those. However, we cannot distinguish if the recombinations are distinct or not. So at lease one recombination between j_1 and j_4 exist in this case.

(c) We can compute a lower bound for the number of historical recombinations by computing how many nearby columns satisfy 4-gamete rule. For each two columns satisfying the rule and no other columns in between satisfying the rule we can claim at least one recombination event happened.
So the algorithm can be explained as follows:
Starting from the first column, set C_1 to the first column and C_2 to the next column.
Check C_1 and C_2 for 4-gamete rule. If they satisfy the rule increase the number of recombinations; set C_1 to the column after C_2, and set C_2 to the column after C_1.
But if C_1 and C_2 does not satisfy the rule advance C_2 to the next column.
Continue to the end of the matrix.

3. Coalescent theory
If m_i represent number of columns with exactly i 1's, the expected number of 1's in the matrix can be expressed as:
$Exp(\sum_{i=1}^{m-1} im_i) = \sum_{i=1}^{m-1} Exp(im_i) = \sum_{i=1}^{m-1} iExp(m_i) = \sum_{i=1}^{m-1} i\theta/i = (m-1)\theta$
m also has an expexted value of: $Exp(m) = 4\mu N(ln(n-1) + \gamma)$
Therefore, the expected number of 1's can be written as:
$(\theta(ln(n-1) + \gamma) - 1)\theta$.

4. Haplotype assembly and optimization
Each read r_i is represented as:
$r_{ij} = (r_{i1}, r_{i2}, ..., r_{in})$.
The solution is represented as:
$x_i = (x_1, x_2, ..., x_n)$.
We define a precomputed value p_{ij}, and set it to 1 if read r_i has '-' in location j, and 0 otherwise. We define a variable $r_{ij}^v = r_{ij} + p_{ij}(x_j - r_{ij})$. Note that r_{ij}^v is linear in terms of the variables and r_{ij} and p_{ij} are constants.
Also compute r_{ij}^c such that it has the complementary base pairs of r_{ij} in locations other than '-'.
Now we want to minimize:
$\sum_i \sum_j min(|r_{ij}^v - x_j|, |r_{ij}^{vc} - x_j|)$.
This will take into account the minimum errors between the the read and the assumed solution and its complement. In order to linearize that assume:
$y_{ij} = |r_{ij}^v - x_j|$.

$z_{ij} = |r^{vc}_{ij} - x_j|$.
$X_{ij} = min(y_{ij}, z_{ij})$.
So the ILP will become:
$minimize \sum_i \sum_j X_{ij}$
s.t.:
$X_{ij} <= y_{ij}$
$X_{ij} <= z_{ij}$
$y_{ij} >= r^{v}_{ij} - x_j$
$y_{ij} >= x_j - r^{v}_{ij}$
$z_{ij} >= r^{vc}_{ij} - x_j$
$z_{ij} >= x_j - r^{vc}_{ij}$
$r^{v}_{ij} = r_{ij} + p_{ij}(x_j - r_{ij})$
$r^{vc}_{ij} = r^{c}_{ij} + p_{ij}(x_j - r^{c}_{ij})$
$p_{ij} = 1$ if r_{ij} has '-' in position j
x_i is in set $\{0, 1\}$

5. MCMC
We would like the target probability to be proportional to the score, $P_i \propto S(\pi_i)$. For each odd-permutation we define the neighborhood as exchanging a pair of locations with preserving the odd-premutation property. So we first choose a location with probability $1/n$ then choose an exchange location with probability $2/n$, since if the first location is odd we can only exchange it with odd locations. So A_{ij} is symmetric.
We define the transition probabilities as:
$T_{ij} = min\{1, \frac{P_j}{P_i}\}$
We check the detailed balance as follows:
Assume $P_i < P_j$
$P_i T_{ij} = P_i \times 1 = P_i$
$P_j T_{ji} = P_j \frac{P_i}{P_j} = P_i$
Therefore the target probability P is the stationary probability associated with the MCMC process.