Computing Haplotype Frequencies and Haplotype Phasing via the Expectation Maximization (EM) Algorithm

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

September 14, 2010

Outline

EM by one Example

- **Problem**: Consider two loci with two allele 0 and 1 at each locus.
- **Given**: (We observe) the genotypes of the individuals at both loci.
- **Find**: The estimate at the haplotype frequencies.

Solution

- There are a total of four possible haplotypes $00, 01, 10, 11$ at the two loci.
- Let us denote their frequencies by $\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}$.
- Suppose that we have computed already $\theta_{00}^{(t)}, \theta_{01}^{(t)}, \theta_{10}^{(t)}, \theta_{11}^{(t)}$.
- We want to compute $\theta_{00}^{(t+1)}$ as a function of $\theta_{00}^{(t)}, \theta_{01}^{(t)}, \theta_{10}^{(t)}, \theta_{11}^{(t)}$.

The Genotype Sample: several types A, B, C, D, E, F

- There are n_A genotypes or individuals of type 22 - we denote Y_A the set of such genotypes
- There are n_B genotypes or individuals of type 02
- There are n_C genotypes or individuals of type 20
- There are n_D genotypes or individuals of type 00
- There are n_E genotypes or individuals of type 11

The fraction of the genotypes in each category that contains the 00 haplotype

- (A) For the A group of n_A individuals the possible haplotypes show as follows in explanations of the genotypes: $\frac{00}{11}$ or $\frac{01}{10}$ (the "fractions" represent the separation of mother-father) chromosomes.
- $P(Y_A) = 2\theta_{00}^{(t)}\theta_{11}^{(t)} + 2\theta_{01}^{(t)}\theta_{10}^{(t)}$
- $P(\frac{00}{11} \mid Y_A) = \frac{2\theta_{00}^{(t)}\theta_{11}^{(t)}}{2\theta_{00}^{(t)}\theta_{11}^{(t)} + 2\theta_{01}^{(t)}\theta_{10}^{(t)}}$

The fraction of the genotypes in each category that contains the 00 haplotype (continued)

- For group B one haplotype is 00 and the other one is 01
- For group C one haplotype is 00 and the other one is 10
- For group D both haplotypes are 00
- For group E both haplotypes are 11

Computing $\theta_{00}^{(t+1)}$

- Therefore the total expected number of 00 haplotypes are:
- $n_{00}^{(t+1)} = n_A P(\frac{00}{11} \mid Y_A) + n_B + n_C + 2n_D$
- so we update
- $\theta_{00}^{(t+1)} = \frac{n_{00}^{(t+1)}}{2n}$
- where $n = n_A + n_B + n_C + n_D + n_E$

The EM Algorithm

- The EM algorithm is an iterative method to compute successive sets of haplotype frequencies $p_1, p_2, \ldots, p_T$ starting with some initial arbitrary values $p_1^{(0)}, p_2^{(0)}, \ldots, p_T^{(0)}$
- Those initial values are used as used as if they were the unknown true frequencies to estimate the explanation frequencies $P(h_k h_l)^{(0)}$. This is the **Expectation step**.
- These expected explanation frequencies are used in turn to estimate haplotype frequencies at the next iteration $p_1^{(1)}, p_2^{(1)}, \ldots, p_T^{(1)}$. This is the **Maximization step**.
- ... and so on until convergence is reached (i.e., when the changes in haplotype frequency in consecutive iterations are less than some small value (ϵ).

EM Algorithm initialization

1. All explanations are equally likely
   $P_j(h_k h_l)^{(0)} = \frac{1}{c_j}, 1 \leq j \leq m$
   where m is the total number of genotypes in the input; and $n_1, n_2, ..., n_m$ are the counts for each genotype type.
2. All haplotypes are equally frequent in the sample.
3. Complete Linkage Equilibrium: Haplotype frequencies = the product of single locus allele frequencies
4. Initial haplotype frequencies are picked at random.

The E Step

- The Expectation step at the tth iteration consists of using the haplotype frequencies in the previous iteration to calculate the probability of resolving each genotype into different possible explanations:
  $P_j = \sum_{i=1}^{c_j} P(explanation_i) = \sum_{i=1}^{c_j} P(h_{ik} h_{il})$
- if $k = l$ then $P(h_k h_l) = p_k^2$
- if $k \neq l$ then $P(h_k h_l) = 2 p_k p_l$
  where a_1 is a constant term and p_{ik} and p_{il} are the population frequencies of the corresponding haplotypes.

The E Step (continued)

- The likelihood of the haplotype frequencies given the genotype counts $n_1, n_2, ..., n_m$ is

$$L(p_1, ..., p_T) = a_1 \prod_{j=1}^{m} (\sum_{i=1}^{c_j} P(h_{ik} h_{il}))^{n_j}$$

  where $\sum_{i=1}^{T} = 1, and(\mathrm{h}_{ik} h_{il}), 1 \leq i \leq c_j$ are the set of explanations of the jth genotype that occurs n_j times in the input.
- Let $P_j^{(t)} = \sum_{i=1}^{c_j} P(h_{ik} h_{il})^{(t)}$

The E Step formula

- The E Step formula is:

$$P_j(h_k h_l)^{(t)} = \frac{P(h_k h_l)^{(t)}}{\sum_{i=1}^{c_j} P_j^{(t)}}$$

The M Step

- Haplotype frequencies are then computed for each Maximization step: for $1 \leq r \leq T$

$$p_r^{(t+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{ir} P_j(h_{ik} h_{il})^{(t)}$$

where δ_{it} is an indicator variable equal to the number of times haplotype t is present in explanation i; and this number can be $0, 1$ or 2.