CSCI2820: Expectation-Maximization Haplotype Phasing

October 23, 2013

Informally, expectation-maximization can be described in 4 steps:

1. **Initialize haplotype frequencies.** The most common initialization is to give each haplotype equal probability. That is, if there are k unique haplotypes in the set of explanations for all input genotypes, then each haplotype receives frequency $\frac{1}{k}$.
2. Iterate until convergence:
3. Using the haplotype frequencies, compute the expected haplotype counts.
4. Using the expected haplotype counts, compute the haplotype frequencies.

Pseudo-code given in Algorithm 1.

The δ function counts the number of times h appears in explanations of g. For example, $\delta(00, 21) = 0$, $\delta(00, 22) = 1$, $\delta(00, 00) = 2$. The large pieces that remain are to calculate the haplotype phasings and the conditional probabilities. The haplotype phasings can be calculated by taking the explanation for each genotype with the highest probability (product of their frequencies). The conditional probability can be computed with Algorithm 2.

input : set of genotypes G with multiplicity, likelihood iteration difference parameter ϵ, and a maximum number of iterations i
output: haplotype frequencies and haplotype phasing

// **set up variables**
$iteration = 0$;
$mle = 0$;
$prev\_mle = -1$;
$H \leftarrow computeHaplotypes(G)$;
$hap\_freq(h) \leftarrow Map < Haplotype \rightarrow Float >$;
$hap\_expectation(h) \leftarrow Map < Haplotype \rightarrow Float >$;

// **initialize haplotype frequencies**
for *haplotype* $h \in H$ **do**
  $hap\_freq(h) \leftarrow \frac{1}{H.size()}$;
end

while $(prev\_mle - mle) < \epsilon$ *AND* $iteration < i$ **do**
  // **calculate the expected counts**
  for *haplotype* $h \in H$ **do**
    float $hapCount = 0$;
    for *genotype* $g \in G$ **do**
      if *consistent(g,h)* **then**
        $d \leftarrow \delta(g, h)$;
        $hapCount+ = d \cdot G(g) \cdot conditional\_prob(h, g)$;
      else
      end
    end
    $hap\_expectation(h) \leftarrow hapCount$;
  end
  // **calculate the haplotype frequencies**
  for *haplotype* $h \in H$ **do**
    $hap\_freq(h) \leftarrow \frac{hap\_expectation(h)}{2 \cdot n}$;
  end
  $prev\_mle = mle$;
  $mle = calculate\_mle()$;
  $iteration + +$;
end

Algorithm 1: EM algorithm.

input : Haplotype h, genotype g
output: conditional probability

// **get compliment haplotype** $h' \leftarrow compliment(h)$;

if $h == h'$ **then**
  $numerator = hap\_freq(h) \cdot hap\_freq(h')$;
else
  $numerator = 2 \cdot hap\_freq(h) \cdot hap\_freq(h')$;
end

for *haplotype explanations* $h, h' \in g$ **do**
  if $h == h'$ **then**
    $denominator+ = hap\_freq(h) \cdot hap\_freq(h')$;
  else
    $denominator+ = 2 \cdot hap\_freq(h) \cdot hap\_freq(h')$;
  end
end
return $\frac{numerator}{denominator}$;

Algorithm 2: Calculating conditional probabilities