**Introduction**

- A Single Nucleotide Polymorphism (SNP) is a position in the genome at which exactly two of the possible four nucleotides occur in a large percentage of the population. SNPs account for most of the genetic variability between individuals, and mapping SNPs in human population has become the next high-priority in genomics after the completion of the Human Genome project.
- In diploid organisms such as humans, there are two non-identical copies of each chromosome. A description of the SNPs in each chromosome is called a haplotype, which can be viewed as a 0/1 vector, e.g., by representing the most frequent (dominant) SNP allele as a 0 and the alternate (minor) allele as a 1.
- At present, it is prohibitively expensive to directly determine the haplotypes of an individual, but it is possible to obtain rather easily the combined SNP information in the so-called genotype. A genotype can be conveniently represented as a 0/1/2 vector, where 0 (1) means that both chromosomes contain the respective (minor) allele, and 2 means that the two chromosomes contain different alleles.

**Problem Definition**

- A pair of haplotypes (h, h') explains g if \( h(i) \neq h'(i) \) whenever g(i) = 2
- A phase is a function of a set of genotypes \( G \), which is a function \( f: G \rightarrow \{0,1\}^k \times \{0,1\}^k \) such that, for every g, \( f(g) \) is a pair of haplotypes that explain g
- A penalty f(g) = \( (h,h') \) plus twice the number of genotypes g such that \( f(g) = (h,h) \)

**Extensions**

- Divide the genotypes into windows of size \( k \)
- Run the previous algorithm for windows of size \( k \times 2 \)

**Switching Error (%)**

<table>
<thead>
<tr>
<th>#Gen</th>
<th>RAND</th>
<th>ENTROPY_PHASE</th>
<th>GERBIL</th>
<th>PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>10.8</td>
<td>12.1</td>
<td>14.6</td>
<td>16.8</td>
</tr>
<tr>
<td>100</td>
<td>11.7</td>
<td>13.8</td>
<td>16.0</td>
<td>18.1</td>
</tr>
<tr>
<td>200</td>
<td>12.6</td>
<td>15.1</td>
<td>17.9</td>
<td>19.8</td>
</tr>
</tbody>
</table>

**IL8-datasets**

- 5q31-euro (50 SNP)
- 5q31-wafr (52 SNP)
- IL8-wafr (50 SNP)

<table>
<thead>
<tr>
<th>#Gen</th>
<th>RAND</th>
<th>ENTROPY_PHASE</th>
<th>GERBIL</th>
<th>PHASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>129</td>
<td>9.2</td>
<td>11.1</td>
<td>13.7</td>
<td>15.8</td>
</tr>
<tr>
<td>50</td>
<td>44.9</td>
<td>47.6</td>
<td>47.3</td>
<td>47.7</td>
</tr>
</tbody>
</table>

**Conclusions**

- Entropy minimization gives a unified framework for various phasing problem variants, including phasing genotypes with missing data and pedigree constrained phasing
- Preliminary results show that entropy minimization is competitive with existing methods in haplotype reconstruction accuracy, particularly for large populations
- Currently, we are implementing trio-based entropy phasing and are exploring other strategies for phasing long genotypes

**References**


**Experimental Setup**

<table>
<thead>
<tr>
<th>IL8-wafr (50 SNP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Gen</td>
</tr>
<tr>
<td>129</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

**Datasets**

- (Daily 2001) 129 family trios over a region of 103 SNPs
- (Gabriel 2002) 60 blocks with an average of 50 SNPs genotyped for 29 individuals
- (Foront et al. 2004) Simulated populations generated as follows:
  - 32 European and 32 West African family trios were genotyped at the IL8 and 5q31 regions
  - Population haplotypes and their frequencies were inferred using Phamily and Phased
  - Based on these haplotypes frequencies, 100,000 random genotypes are generated, from which we selected populations of size between 50 and 800

**Minimum Entropy Population Phasing**

Given a set of genotypes, find a phasing with minimum entropy

**Switch error rate**

Given the true haplotype(f,) and the inferred ones(h), switch error rate is the number of times we have to switch from reading h to f, divided by the number of ambiguous positions.