Incorporating partial phase information into inference of coancestry

Chris Glazner and Elizabeth Thompson

University of Washington
Department of Statistics

June 20, 2012
IBD detection

- Our aim is inference of segments of identity by descent (IBD): DNA shared from a recent common ancestor
- By inferring IBD, we can perform linkage analysis even when pedigrees are unavailable or incomplete
- Input to our model is SNP data, either as phased haplotypes or unphased genotypes
Phase

- Genetic variants in humans lie on pairs of chromosomes:

  \[
  \begin{array}{cccccccc}
  A & C & G & T & C & T & G & A \\
  T & G & G & A & G & T & G & T \\
  \end{array}
  \]

- Data are typically reported as unordered pairs at each locus:

  \[
  \begin{array}{cccccccc}
  A & G & G & T & C & T & G & T \\
  T & C & G & A & G & T & G & A \\
  \end{array}
  \]

- *Phasing* is the assignment of variants to haplotypes:

  \[
  \begin{array}{cccccccc}
  A & G & G & T & C & T & G & T \\
  T & C & G & A & G & T & G & A \\
  \end{array}
  \]
We use a hidden Markov model (HMM), shown below in terms of IBD among chromosomes and phased haplotypes.

Colored circles indicate genes in some IBD state and black/white circles indicate SNPs.
The model is specified by two distributions (assume the Markov chain starts at stationarity):

- The emission probability of the haplotype data $y$ at a locus given hidden IBD state $i$: $e(y|i)$
- The Markov transition probability from state $i$ to state $j$ between two loci: $p(j|i)$
Orderings

- Both components of the model implicitly refer to some underlying ordering.
- The IBD states refer to an unknown but fixed ordering (e.g. “Individual A’s first gene is shared IBD with individual B’s second gene”).
- With haplotypic data, the observed alleles at a locus are ordered.
- The emission probability depends on the ordering of the genotype in terms of the ordering that defines the IBD states: $R$.
- The emission distribution will be written $e(y|i, R)$ to denote this dependence.
Group actions

- Let $G_n$ be the permutation group generated by the transpositions of the $2i - 1$th and $2i$th elements of a set, $i = 1, \ldots, n$

- At a single locus, this group defines distinct but analogous actions on:
  - The set of IBD states among $n$ individuals
  - The set of possible orderings of the genotypes of $n$ individuals
- $G_n$ swaps genes or alleles within individuals
Group actions

- An element of $G_n$, the permutation (21), acting on the two sets:

  - Action on IBD states
  - Action on data orderings
Orbits

- Definition: the orbit of an element $x$ of a set acted on by a group $G$ is $\{y : \exists g \in G \text{ s.t. } x = gy\}$
- The orbits under $G_n$ are “genotypically equivalent” in that they are distinguished only by maternal/paternal ordering
Both components of the model are invariant when acted on by an element $g \in G_n$:

- $e(y|i, R) = e(y|gi, gR)$ (see diagram)

- $p(j|i) = p(gj|gi)$ (because the transition model is exchangeable)

This allows us to perform HMM calculations for situations where phase is unknown or partially known.
Unphased data

- If we have no phase information, we have an independent, uniform prior distribution for the data ordering at each locus.
- The emission probabilities from the phased model are modified to integrate over orderings:
  
  \[ e(y|i) = \frac{1}{|G_n|} \sum_{g \in G_n} e(y|i, gR) \]
  
  \[ = \frac{1}{|G_n|} \sum_{g \in G_n} e(y|g^{-1}i, R) \]

- The “genotypic” emission probability is an average over orbits (genotypically equivalent states)
Unphased data

- To calculate the genotypic emission probability (a) take average over possible phasings (b)
- This is equal to the average over all IBD states in the orbit (c)
Partially phased data

- Averaging over orbits also allows the incorporation of partial phase information into the model
- Why not just treat the data as completely unphased?
- Model inference of IBD is generally more accurate when data are phased, so we want to make use of all available data
- We distinguish between two types of partial phase information: relative and absolute phasing
Relative phasing

- Relative phasing: phasing is only known from one marker to the next
- “Phasing the gaps between markers”
- Markers are phased in segments, but phase between segments is unknown
  - Ex.: Phasing using linkage disequilibrium blocks
Relative phasing

- With relative phasing, we assign an independent uniform prior to each segment (rather than to each marker).
- With multiple individuals, take the intersection of phased segments.
Relative phasing

- This prior is incorporated into the HMM algorithm when calculating the forward probabilities along the chromosome.
- Whenever we move into a new segment, forward probabilities are averaged within each orbit.
- This can be thought of as the HMM “forgetting” which phasing it used to process the segment.
Absolute phasing

- Absolute phasing: phase is specified in terms of some ordering that applies across the entire chromosome
- “Phasing the markers but not the gaps between them”
- Markers separated by unphased markers can be assigned to the same haplotype
  - Ex.: Phasing based on pedigrees
Absolute phasing

- The prior is constant on all phased segments, and unphased markers have independent priors.
- We don’t want to “forget” information across unphased gaps, so we don’t average forward probabilities.
- Instead, the unphased emission probability is used at the unphased loci, and otherwise the calculation is the same as in the fully phased case.
Discussion

- Joint priors on allele orderings can describe a flexible class of phasing situations
- The situations described here are part of a subset of priors which fit nicely into the HMM framework
- Technological and computational advances are making phasing data increasingly available
  - Long sequencing reads: molecular phasing
  - Good phasing software (BEAGLE, IMPUTE)
- Models which incorporate partial phasing allow the use of this information even when it is incomplete
Thank you

- Funding: NIH grants R37 GM046255 and T32 GM081062
- Thanks to Chaozhi Zheng (UW) for a helpful discussion