Cycling through trees: accurate genome-wide algorithms for haplotype assembly and phasing

Derek Aguiar

Abstract

Human genetic variation is responsible for much of the phenotypic diversity witnessed both within and between populations. Experimental techniques for determining the alternative forms of variants (alleles) produce unordered sets in which the chromosome of origin for each allele is unknown. However, determining the sequence of alleles co-inherited in a single chromosome (haplotype) is fundamentally important in genomics, molecular biology, and genomic medicine. The computation of haplotypes from genome sequencing, or haplotype assembly, operates on discrete nucleotide observations and thus the dominant methods are combinatorial. Conversely, haplotype reconstruction from DNA microarrays, or haplotype phasing, uses statistical linkage between neighboring alleles and identical by descent evolutionary relationships to determine a likely set of haplotypes. Statistical methods alone are too inefficient to produce genome-wide phasings because the space of haplotype phasing solutions grows exponentially in the number of variants typed.

We propose an integrated algorithmic framework for haplotype assembly and phasing based on our leading graph-theoretic haplotype assembly method, **HapCompass**. This framework provides an algorithmic design strategy for a wide range of haplotype reconstruction problems and incorporates population genetics and identity by descent theory into the haplotype reconstruction model. First, we will describe the HapCompass algorithm for genomes containing two sets of homologous chromosomes (e.g. humans) which models haplotype reconstruction as local optimizations on the cycle-basis of graph theoretic representation of the data. Second, we present extensions and generalizations to accommodate genomes with more than two sets of homologous chromosomes (e.g. plants) and tumor genomes. Lastly, we propose a unified framework for haplotype assembly and phasing by incorporating populations genetics and identity by descent shared haplotypes into the haplotype reconstruction model.