Shapeit

Shape-IT: new rapid and accurate algorithm for haplotype inference

Goal: Infere the most likely haplotypes from genotyping data

Global parameters:
- n - number of genotypes
- s - number of SNPs
- z_i - number of heterozygous SNPs for genotype i.

Input:
- G = {G1, …, Gn} — sample of genotypes (allelic contents) of n genotypes (=individuals)
- p = {p1,…,p_(s-1)} — recombination parameters for segments in (s-1) intervals between the s SNP (provided in “genetic_map.txt”)

(desired) Output:
Set of haplotype pairs H = {H1,…, Hn} which maximises the conditional probability to observe H knowing genotypes G and the recombination patterns p: Pr(H | G, p).

Algorithm:

Part1 : Gibbs sampling

1. Start with a random haplotype realization H^0 (n haplotype pairs compatible with genotype G).
2. Iteratively update $H^{(t)} \rightarrow H^{(t+1)}$ by updating haplotypes for each individual $H_i^{(t)}$ conditionally to the 2n-2 other haplotypes from $H^{(t)}$ (notation: $H_{-i}^{(t)}$)

   Sample $H_i^{(t)}$ from a conditional distribution $Pr(H_i^{(t)} \mid H_{-i}^{(t)}, p)$

(FDLS distribution).

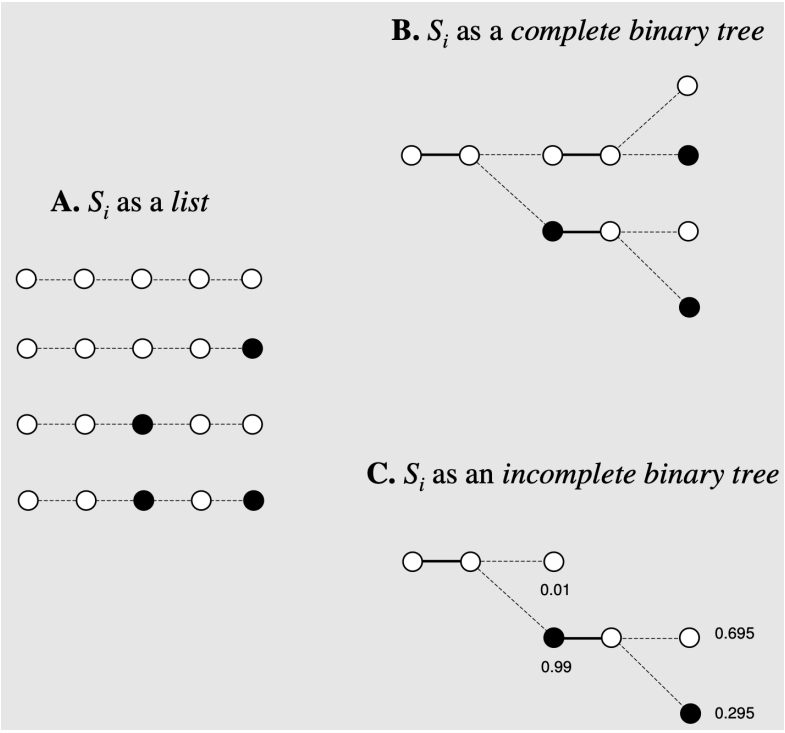
Note1: The computation of this distribution is the most time consuming step.
Note2: Any haplotype is assumed to depend on the other determined haplotypes under hypothesis that they were produced by mutation and recombination events from a common ancestor and thus can be viewed as a mosaic of common set of haplotypes.

Note3: Only haplotype pairs compatible with correspondent genotype are considered. The space of possible haplotype pairs S_i has size of 2^{z_i-1}.

Part 2: Hidden Markov Model

3. Computing the FDLS distribution using Hidden Markov Model (HMM) with representation of a space of possible haplotypes as an incomplete binary tree.

The main advantage of ShapeIt method consists in the use of binary trees to represent the sets of candidate haplotypes for each individual (instead of list representation). This allows for much faster computations.



(binary tree)

Each path of the tree corresponds to a single possible haplotype pair of S_i.

(incomplete binary tree)

The idea is to explore only the paths in the haplotype tree which correspond to the most plausible haplotype pairs (keeping only nodes with proba > threshold).

S_i — space of possible haplotypes for genotype G_i.

Reference:
"Shape-IT: new rapid and accurate algorithm for haplotype inference"
Olivier Delaneau, Cédric Coulonges and Jean-François Zagury*