HMM description

Notation from Rabiner, Lawrence R., Proc of the IEEE 27:257-? (1989).

I am looking at long stretches of Identity By Descent in full-genome haploid malaria sequences from Senegal.

N = number of states = 2

M = number of outputs = 4

**A**: aij = transition matrix: 2x2

**B**: bj(k) = emission matrix: j=1-N, k=1-M

πi = probability in state i initially

The two states are IBD and DBD. The 4 outputs are concordant (aa & AA), discordant (aA), and missing (n). This means I am skipping all but biallelic SNPs.

**A** and **B** will vary with the SNP; A depends on the MAF and B depends on inter-SNP distance. I will assume a uniform recombination rate. Initially, I will assume a uniform transition probability per chromosome across sample pairs, and set it to 4; I will have to assess the robustness of the results to changes in this. (If not robust, further inference or specification will be needed.)

I determine expected discordance rate by taking all samples pairs with discordance (conditional on at least one copy of minor allele at site) < 0.02, which I take to be genetic clones. Discordance represents recent mutations and sequencing error, and occurs at a rate of 1x10-5. I treat this as uniform across all sites and states, and define ε = disc/2.

ε = 5 x 10-6 (replaced by .2%, I believe)

**Components of B**

p(n|IBD) = p(n|DBD) = f(missing) [Set per sample pair, from overall value. (Picking arbitrary value shouldn’t affect segment assignments, but I’ll set from data anyway.)]

p(aA|IBD) = 2ε \* (1-ε)

p(aA|DBD) = 2 \* maf \* (1-maf) \* ((1-ε)2 + ε2) + (maf2+(1-maf)2) \* 2ε \* (1-ε) = (1-hom) \* ((1-ε)2 + ε2) + hom \* 2ε \* (1-ε)

p(aa|DBD) = maf2 \* (1-ε)2 + (1-maf)2 \* ε2 + 2 \* maf \* (1-maf) \* ε \* (1-ε)

p(AA|DBD) = (1-maf)2 \* (1-ε)2 + maf2 \* ε2 + 2 \* maf \* (1-maf) \* ε \* (1-ε) || set = 0.5 to skip 2 major

p(aa|IDB) = p(AA|IDB) = (1-ε)2 + ε2

**Components of π**

p(IBD|d) = p(d|IBD) \* p(IBD) / p(d) = p(d|IBD)\*f(IBD) / (p(d|IBD)\*f(IBD) + p(d|DBD)\*f(DBD))

p(DBD|d) = p(d|DDB)\*f(DBD) / (p(d|IBD)\*f(IBD) + p(d|DBD)\*f(DBD))

p(IBD|c) = p(c|IBD)\*f(IBD) / (p(c|IBD)\*f(IBD) + p(c|DBD)\*f(DBD))

p(DBD|c) = p(c|DBD)\*f(DBD) / (p(c|IBD)\*f(IBD) + p(c|DBD)\*f(DBD))

p(IBD|n) = p(n|IBD)\*f(IBD) / (p(n|IBD)\*f(IBD) + p(n|DBD)\*f(DBD))

p(DBD|n) = p(n|DDB)\*f(DBD) / (p(n|IBD)\*f(IBD) + p(n|DBD)\*f(DBD))

Correction: Effect of observation at 1st SNP is included elsewhere in the Viterbi algorithm. So

p(IBD) = f(IBD)

p(DBD) = f(DBD)

which is a lot simpler. Further correction: p(IBD) = p(DBD) => no bias.

**Components of A**

p(transition) = 4(pos(t+1) – pos(t)) / length(chrom)

p(DBD->IBD) = p(transition) \* f(IBD) [f(IBD) est. from global discordance]

p(IBD->DBD) = p(transition) \* f(DDB)

p(IBD->IBD) = 1 - p(IBD->DBD)

p(DBD->DBD) = 1 - p(DBD->IBD)

Correction: transitions should be symmetric, since otherwise there is a directional bias

p(transition) = 2(pos(t+1) – pos(t)) / length(chrom)

p(DBD->IBD) = p(transition)

p(IBD->DBD) = p(transition)

p(IBD->IBD) = 1 - p(IBD->DBD)

p(DBD->DBD) = 1 - p(DBD->IBD)

Improved: p(transition) = Ngen \* (pos(t+1) – pos(t)) \* ρ

Using ρ = 17 kB/cM, or ρ = 5.8e-7/bp

(Switching made very little difference)