Admixfrog model

benjamin Peter

February 14, 2019

Introduction

We aim to build an HMM that classifies a single *target* genome into regions that were derived from different diverged populations. This is motivated by admixture between modern humans, Neandertals or Denisovans, (which I will occasionally refer to as source populations) although the model is applicable to many other taxa.

Briefly, the hidden states are the sources of the individual, e.g. homozygous modern human (HH) homozygous Denisovan (DD) and homozygous Neandertal (NN), as well as the heterozygosu states HD, DN, HN where one haplotype each is derived from the sources.

We assume a homogeneous Markov model throughout, which we achieve by binning the genome in bins of constant map size (e.g. 0.005 cM), so it is possible that a bin has zero, one or many emissions, which are treated independently distributed given the hidden state. We assume that data comes in the form of biallelic SNP. We assume we have a sample of genotypes from each source, as well as reads covering the SNP from the target.

Emissions are the number of reads of the ancestral /derived allele, with the allele frequency in Humans / Neandertals / Denisovans assumed known.

The transition matrix is estimated using standard- Baum-Welch. The rest of this note deals with inferring the emission probabilities, and in particular estimating contamination.

Notation

Let

- R, L the number of read groups, and hidden-state bins, respectively
- S_l the number of SNP in bin l
- n_{rs} the number of reads of readgroup r at SNP s
- $O = (O_{rs})$ the set of derived-allele- read counts from library r at SNP s
- $Z = (Z_l)$ be the sequence of hidden states
- $R = (R_l)$ the number of contaminant reads at SNP l
- A_{sk}, B_{sk} the number of derived and ancestral allele counts in reference k at SNP s, (or more generally parameters of a Beta prior)
- T be a transition matrix
- F_k a measure of distance between source k and the introgressing population
- c_r proportion of contaminant reads in read group r
- p_{sc} the allele frequency of a contaminant at position s.
- e the sequencing error rate
- $\theta = (F_k, c_r, T, e)$, the set of all parameters to be estimated

The model can be summarized as

- $O_{rs}|G_s, n_{rs} \sim Binomial(O_{rs}; n_{rs}, p(e, c_r, p_{sc}, G_s))$
- $G_s|Z_l = k \sim Binomial(2, f_{sk})$
- $f_s|Z_l = k \sim Beta(a_k F_k, b_k F_k)$
- $Z_s|Z_{s-1} = k' \sim T_{k'}$
- $Z_0 \sim \alpha_0$

Where we want to estimate the c_r, F_k, T and α_0

1 Algorithm details

We factor the likelihood as

$$P(O|\theta) = \sum_{\mathcal{G},\mathcal{Z}} P(O|G, c, p_c, N)P(G|Z, A, B)P(Z|T)$$

Where $\mathcal{G}, \mathcal{Z}$ are the sets of all possible genotype assignments.

Under the Markov assumption, this can be further split up as

$$P(O|\theta) = P(Z_0) + \sum_{k=1}^{K}\sum_{g=0}^{2}\sum_{l=1}^{L} P(O_l|G_l, c, p_{cn}, N_l)P(G_l|Z_l = k, A_{lk}, B_{lk}, \tau)P(Z_l = k|Z_{l-1}, T)$$

we refer to these terms as the genotype likelihood, state likelihood and transition probability, respectively.

A standard trick is to assume the latent variables G, Z are known. Taking the log yields the complete data likelihood

$$\mathcal{L} = P(O, Z, G|\theta) = \log P(Z_0) + \sum_{l=1}^{L} \log P(O_l|G_l, c) + \sum_{l=1}^{L} \log P(G_l|Z_l = k, F) + \sum_{l=1}^{L} \log P(Z_l|Z_{l-1}, T) \tag{1}$$

The model is further improved by taking into account the each bin has S_L snps, and we have R distinct read groups.

$$\begin{aligned}
\mathcal{L} &= P(O, Z, G|\theta) \\
&= \log P(Z_0) \\
&+ \sum_{r=1}^{R}\sum_{s=1}^{S_L}\sum_{l=1}^{L} \log P(O_r|G_s, c) \\
&+ \sum_{s=1}^{S_L}\sum_{l=1}^{L} \log P(G_s|Z_l = k, F) \\
&+ \sum_{l=1}^{L} \log P(Z_l|Z_{l-1}, T)
\end{aligned} \tag{2}$$

The parameters we want to estimate are the contamination rate c_r for each read group, the F_{ST} for each source population, and the sequence of hidden states Z.

1.1 Genotype Likelihood

$$P(O_{rs}|G_s, c_r, n_{rs}) \sim Binom(O_{rs}; n_{rs}, p) \tag{3}$$

where $p = (1 - e)p' + e(1 - p')$ and $p' = c_{rs}p_c + (1 - c_{rs})G_s$

1.2 State Likelihood

1.2.1 Homozygous States

Direct samples (subscripts are omitted) We wish to calculate the probability that the individual is homozygous, given the SNP of interest derives from a population whose allele frequency can be characterized as

$$f \sim Beta(aF, bF)$$
$$G \sim Binom(2, f)$$

The resulting compound distribution is betabinomial:

$$\begin{aligned} P(G = k|F, a, b) &\sim Betabinom(k; n = 2, aF, bF) \\ &= \binom{2}{k} \frac{B[k + aF, n - k + bF]}{B[(a + b)F]} \end{aligned} \tag{4}$$

where $B[.]$ denotes the beta function.

Drift to reference Let f denote the frequency in the reference, A, D the number of ancestral and derived alleles observed in a sample from there, and G the number of derived alleles present.

Conceptually, genetic drift will increase the chance that the two alleles in the target are derived from the same ancestor: If the populations are very distant, the probability F that the two alleles are identical by descent is one. Conversely, if we have a very close reference, that probability is zero.

$$\begin{aligned} P(G = 2|A, D) &= \int P(G|f)P(f|A, D)df \\ &= \int \left[Ff + (1 - F)f^2\right] P(f|A, D)df \\ &= FE[f|A, D] + (1 - F)E[f^2|A, D] \\ &= F\frac{D}{A + D} + (1 - F)\frac{D(D + 1)}{(A + D)(A + D + 1)} \\ &= \frac{D^2 + D + ADF}{(A + D)(A + D + 1)} \end{aligned} \tag{5}$$

Hence we have an excess of homozygous emissions. The probability for $G = 0$ is obtained by switching A and D, and for $G = 1$ by subtracting them from one:

$$P(G = 1|A, D) = \frac{AD(1 - F)}{(A + D)(A + D + 1)} \tag{6}$$

1.2.2 Heterozygous States

For a heterozygous state, we know the allele frequencies in the two source pops, and we know they contribute one haplotype each.

$$\begin{aligned} P(G = 1|F, a, b) &= Betabinom(n = 1, aF, bF) \\ &= \frac{\Gamma(1 + aF))\Gamma(bF)}{\Gamma(1 + F(a + b))} \frac{\Gamma(F(a + b))}{\Gamma(Fa)\Gamma(Fb)} \\ &= \frac{Fa}{(Fa + Fb)} \\ &= \frac{a}{a + b}. \end{aligned} \tag{7}$$

Intuitively, this is because if we just sample one allele from a population, that population's allele frequency is the best guess for the state of our sample. The distance increases our uncertainty,

but not the mean estimate. Thus

$$P(G=0) = \frac{b_1}{a_1+b_1}\frac{b_2}{a_2+b_2}$$
$$P(G=1) = \frac{a_1b_2+a_2b_1}{(a_1+b_1)(a_2+b_2)}$$
$$P(G=2) = \frac{a_1}{a_1+b_1}\frac{a_2}{a_2+b_2}$$

1.2.3 Haploid States

On the sex chromosome, or when inbreeding is present, we further encounter haploid regions. Here, the heterozygous states have emission probability 0, and we just have one allele of the heterozygous case: F_{ST} does not matter.

$$P(G=0) = \frac{b}{a+b}$$
$$P(G=1) = \frac{a}{a+b} \tag{8}$$

1.3 Estimating F

$$\begin{aligned} Q(F|F') &= E[\log P(O,Z,G)P(G|Z,\theta')P(Z|\theta')] \\ &= \sum_{Z\in\mathcal{Z},G\in\mathcal{G}} \log P(G|Z,A,B,\tau)P(G|Z,\theta')P(Z|O,\theta') \\ &= \sum_{k}\sum_{g=0}^{2}\sum_{l=1}^{L} \log P(G_t|Z_l=k,A_{lk},B_{lk},F)P(G_l|Z_l,F',O)P(Z_l=k|O,\theta') \end{aligned} \tag{9}$$

Since the F_k-parameters are independent for each homozygous state, these likelihoods are optimized numerically and independently.

$$\hat{F}_k = \underset{F}{\operatorname{argmax}} \left[\sum_{g=0}^{2}\sum_{l=1}^{L} \log P(G_t|Z_l=k,A_{lk},B_{lk},F)P(G_l|Z_l,F',O)P(Z_l=k|O,\theta')\right] \tag{10}$$

where $P(Z_l=k|O,\theta')$ is the output of the forward-backward algorithm, $P(G_s|Z_l=k,A_{sk},B_{sk},F_k)$ is given by (4). and

$$P(G_i|Z_i=k,O) = \frac{P(O_i|G_i,c')P(G_i|Z_i=k,F')}{\sum_{g=0}^{2}P(O_i|G_i=g,c')P(G_i=g|Z=k,F'_k)} \tag{11}$$

This equation follows from

$$\begin{aligned} P(G,Z|O) &= P(G|Z,O)P(Z|O) \\ &= P(G|Z,O_i)P(Z|O) \\ &= \frac{P(O_i|G)P(G|Z)}{P(O_i|Z)}P(Z|O) \end{aligned} \tag{12}$$

as $G_i|Z_i$ is independent of all observations except O_i and Bayes theorem.

1.4 Estimating Contamination

$$\begin{aligned} Q(c|c') &= E[\log P(O,Z,G)P(G|Z,\theta')P(Z|\theta')] \\ &= \sum_{Z\in\mathcal{Z}}\sum_{G\in\mathcal{G}} \log P(O|G,c)P(G|Z,\theta')P(Z|O,\theta') \\ &= \sum_{r=1}^{R}\sum_{k}\sum_{g=0}^{2}\sum_{s=1}^{S_l}\sum_{l=1}^{L} \log P(O_{lr}|G_s=g,c_r)P(G_s|Z_l=k,\theta',O)P(Z_l=k|O,\theta') \end{aligned} \tag{13}$$

We thus optimize

$$\hat{c}_r = \underset{c_r}{\operatorname{argmax}} \sum_k \sum_{g=0}^{2} \sum_{s=1}^{S_l} \sum_{l=1}^{L} \log P(O_{lr}|G_s = g, c_r) P(G_s|Z_l = k, \theta', O) P(Z_l = k|O, \theta') \tag{14}$$

using (11) and (3).

1.5 Estimationg Transitions and Initial State

T is estimated using standard Baum-welch, α_0 is set to the stationary distribution of T.

2 Prior

There are a few options for priors. For computational reasons, a beta prior is most suitable. One possibility is to use a $Beta[0, 0]$ or $Beta[1, 1]$, which can be done by adding psuedocounts to the observed data. More interesting is the assumption that the observed alleles were drawn from a neutral population with allele frequencies proportional to

$$P(f) \propto \frac{\theta}{x} \tag{15}$$

If we encounter a sample with a ancestral and d derived alleles, we have by Bayes' theorem

$$\begin{aligned} P(f|a, d) &\propto P(f)P(a, d|f) \\ &\propto \frac{\theta}{f}\binom{a+d}{d} f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^a \\ &\sim Beta(d, a+1) \end{aligned} \tag{16}$$

I.e. knowing the ancestral allele is equivalent to adding an additional ancestral allele to the observation. If the ancestral allele is unknown, "mirroring this distribution"

For the folded SFS,

$$\begin{aligned} P(f|a, d) &\propto P(f)P(a, d|f) \\ &\propto \left[\frac{1}{f} + \frac{1}{1-f}\right] f^d(1-f)^a \\ &\propto \frac{1}{f(1-f)} f^d(1-f)^a \\ &\propto f^{d-1}(1-f)^{a-1} \\ &\sim Beta(d, a) \end{aligned} \tag{17}$$

In a finite population, alleles enter a population at a frequency of $2N^{-1}$. Writing

$$\eta(N) = \frac{1}{2N \sum_{i=1}^{2N-1} 1/i}$$

We find that the SFS is well approximated by

$$P(f|a, d, N) \sim Beta(d + \eta, a + 1 - \eta) \tag{18}$$

$$\sim Beta(d + \eta, a + \eta) \tag{19}$$

for the unfolded and folded case, respectively.