Mini-Admixfrog

Stephan Schiffels

August 2021

We want to create a simple HMM that is able to paint an unphased diploid target genome into local ancestry states, according to multiple source genomes.

Observations are structured along SNPs at which i) the derived allele frequencies *differ* between the sources, and ii) there is no missing data in the target and at least one non-missing genotype in each source genome.

At every such SNP, an observation is given by a tuple $(G^i, (a_1^i, d_1^i), (a_2^i, d_2^i), \ldots)$, where i denotes the SNP index, $G^i = \{0, 1, 2\}$ denotes the nr of derived alleles in the target genotype, and a_k^i and d_k^i are the numbers of ancestral and derived alleles in source k at SNP i.

A diploid *state* is a tuple of sources. For example, with 2 sources, we have states $S^i = 11$, $S^i = 12$ and $S^i = 22$. With three sources, there are 6 states, and so on.

We are for now not concered with parameter optimization, but only in posterior decoding, that is, informing about local state probabilities given data and a model consisting of given transition- and emission-probabilities.

Emission probabilities

We will omit the SNP indices i in the following.

The emission probability

$$e(G|\{a_k, d_k\}, S)$$

is the probability of a target genotype, given the local state S^i and source allele counts (a_k^i and d_k^i). We follow the original admixfrog model and write the emission probabilities as a betabinomial sampling probability (Peter 2020).

Specifically, for homozygous states (i.e. $S = 11, 22, \ldots$), we have

$$e(G|a, d, S) = \binom{2}{G} \frac{B(G + d + d', 2 - G + a + a')}{B(d + d' + a + a')} \tag{1}$$

where we have omitted the k index for the respective homozygous state. There are prior parameters a' and d' which control the sampling uncertainty in the source genotypes. A simple choice for the priors is $a' = d' = 1$, but since we here deal mostly with the case of very few source genomes, often only one per source, it is more advisable to choose values much smaller than 1 to mimic the expected site frequency spectrum. In practice, we can fit it from data, or simply guess around values such as 0.1 or 0.01, since arguably the results won't depend too much on it.

For heterozygous states, without loss of generality we here write a_1, d_1, a_2, d_2 for the two respective sources, and have:

$$e(G=0|a_1,d_1,a_2,d_2,S) = \frac{(a_1+a')(a_2+a')}{(d_1+d'+a_1+a')(d_2+d'+a_2+a')} \tag{2}$$

$$e(G=1|a_1,d_1,a_2,d_2,S) = \frac{(a_1+a')(d_2+d') + (a_2+a')(d_1+d')}{(d_1+d'+a_1+a')(d_2+d'+a_2+a')} \tag{3}$$

$$e(G=2|a_1,d_1,a_2,d_2,S) = \frac{(d_1+d')(d_2+d')}{(d_1+d'+a_1+a')(d_2+d'+a_2+a')} \tag{4}$$

References

Peter, Benjamin M. 2020. "100,000 Years of Gene Flow Between Neandertals and Denisovans in the Altai Mountains." *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.03.13.990523.