Pair HMM probabilistic realignment in HaplotypeCaller and Mutect

David Benjamin[*][†]

Broad Institute, 75 Ames Street, Cambridge, MA 02142

(Dated: August 25, 2017)

After generating candidate haplotypes, the GATK tools `HaplotypeCaller` and `Mutect` realign reads against these haplotypes to obtain a matrix of likelihoods for each read to be derived from each haplotype. Here we describe the probabilistic model specifying this likelihood as well as its computational implementation. We do not describe the translation of this implementation into native code optimized for vectorized architectures.

I. THE PAIR HMM MODEL

We want to calculate the probability $P(\mathcal{R}|\mathcal{H})$ of read $\mathcal{R}$ to be sequenced from haplotype $\mathcal{H}$, where the haplotypes are sufficiently long that reads are contained within them. This likelihood is the sum of likelihoods of all possible alignments $\mathcal{A}$ of $\mathcal{R}$ to $\mathcal{H}$:

$$P(\mathcal{R}|\mathcal{H}) = \sum_{\mathcal{A}} P(\mathcal{R}, \mathcal{A}|\mathcal{H}) \tag{1}$$

We represent alignments as sequences

$$\text{alignment} = \{(i_1, j_1, s_1), (i_2, j_2, s_2) \ldots (i_N, j_N, s_N)\}, \tag{2}$$

where i and j represent positions within the read and haplotype, respectively, and $s_n \in \{M, I, D\}$ represents the states of match, insertion, and deletion of the read relative to the haplotype. For example, an alignment $\{(1, 10, M), (2, 11, M), (3, 12, M), (4, 13, M), (4, 14, D), (4, 15, D), (5, 16, M), (6, 17, M), (7, 17, I), (8, 18, M)\}$ means that positions 1 - 4 of the read match positions 10 - 13 of the haplotype, followed by a two-base deletion (advancing from 13 to 15 in the haplotype without advancing in the read), followed by a match of read positions 5 - 6 with haplotype positions 16 - 17, followed by an insertion at read position 7, followed by a match. The allowable transitions $(i_n, j_n) \to (i_{n+1}, j_{n+1})$ are

- $(i, j, M/D/I) \to (i+1, j+1, M)$: match of read position $i+1$ with haplotype position $j+1$

- $(i, j, M/D) \to (i, j+1, D)$: deletion after read position i – haplotype position $j+1$ is deleted

- $(i, j, M/I) \to (i+1, j, I)$: insertion at read position $i+1$

Note that the state label s seems redundant because it can be reconstructed from the sequence of i an j. While this is true, our model treats indel starts differently from indel continuations, and by including the state label we can distinguish these conveniently[1]. These transitions are illustrated as a finite state machine in Figure 1.

The read-alignment likelihood has two components. First is the probability of the sequence of match, insertion, and deletion states, which is

$$P(\mathcal{A}) = \prod_k T_{s_k, s_{k+1}}, \tag{3}$$

where T is a matrix of state transition probabilities[2]. The index k runs from 1 to the number of states in the alignment, that is, the read length plus the number of deleted reference bases. Next is the emission probability of the read bases

---

[*] The author took no part in development of the methods described below – credit belongs to several others on the GATK team.

[†] Electronic address: davidben@broadinstitute.org

[1] That is, we can tell which type of transition it is by looking back one unit instead of two. Thus we have a first-order Markov model instead of a second-order Markov model.

[2] The gap continuation probability corresponds to a phred-scaled quality that is set by the `gcpHMM` argument, which is 10 by default, implying that $T_{DD} = T_{II} = 10^{-10/10} = 1/10$ and $T_{DM} = T_{IM} = 1 - T_{II}$. The indel start transitions are derived from the read's BQSR base insertion and base deletion qualities, if they exist, or a phred-scaled indel start quality of 45 if they do not. In production at the Broad Institute bams do not have BQSR indel qualities, hence the constant default is used. That is, $T_{MD} = T_{MI} = 10^{-4.5}$ and $T_{MM} = 1 - 2 \times 10^{-4.5}$.
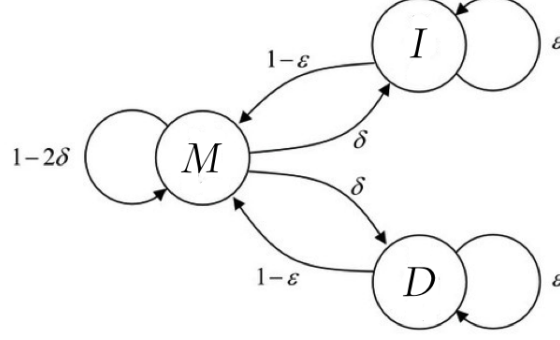
FIG. 1: Finite state machine of transitions among match, insertion, and deletion alignment states. In the absence of per-base BQSR indel qualities transition probabilities are parametrized by two constants, the indel start probability δ and the indel continuation probability ϵ, such that $T_{MI} = T_{MD} = \delta$, $T_{MM} = 1-2\delta$, $T_{II} = T_{DD} = \epsilon$, $T_{IM} = T_{DM} = 1-\epsilon$, and $T_{ID} = T_{DI} = 0$.

given the haplotype bases they align to and the base qualities:

$$P(\mathcal{R}|\mathcal{A}, \mathcal{H}) = \prod_k P(r_{i_k}|h_{j_k}, q_{i_k})^{\mathrm{I}[s_k=M]}, \tag{4}$$

where r_m and h_n are the mth read bases and nth haplotype base and q_m is the quality of the read's mth base. Note that this only includes alignments in the match state. The per-base emission is given directly from the definition of base quality:

$$P(b_2|b_1, q) = \begin{cases} \epsilon(q)/3 & (b_1 \neq b_2) \\ 1 - \epsilon(q) & (b_1 = b_2) \end{cases} \tag{5}$$

where $\epsilon(q) = 10^{-q/10}$ is the error rate implied by the phred-scaled quality q.

II. DYNAMIC PROGRAMMING

Define the matrices M, I and D by M_{ij} = the total likelihood of *all* paths from the beginning of the read to position i that end in a match state, and likewise for D and I. Then the recursions

$$M_{ij} = P(r_i|h_j, q_i) \left(M_{i-1,j-1}T_{MM} + I_{i-1,j-1}T_{IM} + D_{i-1,j-1}T_{DM} \right) \tag{6}$$
$$I_{ij} = M_{i-1,j}T_{MI} + I_{i-1,j}T_{II} \tag{7}$$
$$D_{ij} = M_{i,j-1}T_{MD} + D_{i,j-1}T_{DD} \tag{8}$$

define the entire pair HMM algorithm:

Algorithm 1 Pair HMM algorithm

1: Initialize $M_{0,j} = I_{0,j} = 0$ and $D_{0,j} = 2^{1020}/|\mathcal{H}|$ for $1 \leq j \leq |\mathcal{H}|$.
2: **for** $1 \leq i \leq |\mathcal{R}|$ **do**
3:     **for** $1 \leq j \leq |\mathcal{H}|$ **do**
4:         Calculate M_{ij} via Equation 6.
5:         Calculate I_{ij} via Equation 7.
6:         Calculate D_{ij} via Equation 8.
7:     **end for**
8: **end for**
9: Total likelihood $P(\mathcal{R}|\mathcal{H})$ is $\sum_j \left(M_{\mathcal{R},j} + I_{\mathcal{R},j} \right)$.

That the $i = 0$ rows of M and I are initialized to zero corresponds to starting at an imaginary position one base

before the read start in a deletion state[3]. The factor of $1/|\mathcal{H}|$ corresponds to a flat prior on which j an alignment starts at. This is important for a local alignment because we don't want to penalize reads that start in the middle of the haplotype. The curious factor of 2^{1020} is a huge number to prevent underflow – all multiplications are by numbers less than 1, so we needn't worry about overflow – which is a much more efficient approach than performing the computation in log space. The omission of D from the returned value recognizes the fact that a terminal deletion is meaningless.

Finally, we note a shortcut that the GATK exploits: when two consecutive haplotypes of the same length[4] agree up to the kth position, the first k columns of M, D and I are recycled and the inner loop is over $k < j \leq \mathcal{H}$.

---

[3] Since $T_{DI} = 0$ this means that an alignment may not start with an insertion, though it may begin with one. This limitation is unnecessary and could be fixed by initializing the $i = 0$ row of I to a non-zero value as well.

[4] The condition on the same length could easily be relaxed simply by accounting for the constant $1/|\mathcal{H}|$ initialization.