Mathematical Notes on SAMtools Algorithms

Heng Li

October 12, 2010

Chapter 1

Duplicate Rate

1.1 Amplicon duplicates

Let N be the number of distinct segments (or seeds) before the amplification and M be the total number of amplicons in the library. For seed i ($i = 1, \ldots, N$), let k_i be the number of amplicons in the library and k_i is drawn from Poinsson distribution $\mathrm{Po}(\lambda)$. When N is sufficiently large, we have:

$$M = \sum_{i=1}^{N} k_i = N \sum_{k=0}^{\infty} k p_k = N\lambda$$

where $p_k = e^{-\lambda}\lambda^k/k!$.

At the sequencing step, we sample m amplicons from the library. On the condition that:

$$m \ll M \tag{1.1}$$

we can regard this procedure as sampling with replacement. For seed i, let:

$$X_i = \begin{cases} 1 & \text{seed } i \text{ has been sampled at least once} \\ 0 & \text{otherwise} \end{cases}$$

and then:

$$\mathrm{E}X_i = \Pr\{X_i = 1\} = 1 - \left(1 - \frac{k_i}{M}\right)^m \simeq 1 - e^{-k_i m/M}$$

Let:

$$Z = \sum_{i=1}^{N} X_i$$

be the number of seeds sampled from the library. The fraction of duplicates d is:

$$\begin{aligned} d &= 1 - \frac{\mathrm{E}(Z)}{m} \\ &\simeq 1 - \frac{N}{m} \sum_{k=0}^{\infty} \left(1 - e^{-km/M}\right) p_k \\ &= 1 - \frac{N}{m} + \frac{Ne^{-\lambda}}{m} \sum_k \frac{1}{k!} \left(\lambda e^{-m/M}\right)^k \\ &\simeq 1 - \frac{N}{m} \left[1 - e^{-\lambda} \cdot e^{\lambda(1-m/M)}\right] \end{aligned}$$

i.e.

$$d \simeq 1 - \frac{N}{m}\Big(1 - e^{-m/N}\Big) \tag{1.2}$$

irrelevant of λ. In addition, when m/N is sufficiently small:

$$d \approx \frac{m}{2N} \tag{1.3}$$

This deduction assumes that i) $k_i \ll M$ which should almost always stand; ii) $m \ll M$ which should largely stand because otherwise the fraction of duplicates will far more than half given $\lambda \sim 1000$ and iii) k_i is drawn from a Poisson distribution.

The basic message is that to reduce PCR duplicates, we should either increase the original pool of distinct molecules before amplification or reduce the number of reads sequenced from the library. Reducing PCR cycles, however, plays little role.

1.2 Alignment duplicates

For simplicity, we assume a read is as short as a single base pair. For m read pairs, define an indicator function:

$$Y_{ij} = \begin{cases} 1 & \text{if at least one read pair is mapped to } (i,j) \\ 0 & \text{otherwise} \end{cases}$$

Let $\{p_k\}$ be the distribution of insert size. Then:

$$\mathrm{E}Y_{ij} = \Pr\{Y_{ij} = 1\} = 1 - \left[1 - \frac{p_{j-i}}{L-(j-i)}\right]^m \simeq 1 - e^{-p_{j-i}\cdot m/[L-(j-i)]}$$

where L is the length of the reference. The fraction of random coincidence is:

$$\begin{aligned} d' &= 1 - \frac{1}{m}\sum_{i=1}^{L}\sum_{j=i}^{L}\mathrm{E}Y_{ij} \\ &\simeq 1 - \frac{1}{m}\sum_{i=1}^{L}\sum_{j=i}^{L}\left(1 - e^{-p_{j-i}\cdot m/(L-(j-i))}\right) \\ &= 1 - \frac{1}{m}\sum_{k=0}^{L-1}(L-k)\left[1 - e^{-p_k m/(L-k)}\right] \end{aligned}$$

On the condition that L is sufficient large and:

$$m \ll L \tag{1.4}$$

$$d' \simeq \frac{m}{2}\sum_{k=0}^{L-1}\frac{p_k^2}{L-k} \tag{1.5}$$

We can calculate/approximate Equation 1.5 for two types of distributions. Firstly, if p_k is evenly distributed between $[k_0, k_0 + k_1]$, $d' \simeq \frac{m}{2k_1 L}$. Secondly, assume k is drawn from $N(\mu, \sigma)$ with $\sigma \gg 1$:

$$p_k = \frac{1}{\sqrt{2\pi}\sigma}\int_k^{k+1} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \simeq \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

If $p_0 \ll 1$, $\mu \ll L$ and $L \gg 1$:

$$\begin{aligned}
d' &\simeq \frac{m}{4\pi\sigma^2}\int_0^1 \frac{1}{1-x}\cdot e^{-\frac{(Lx-\mu)^2}{\sigma^2}}\,dx \\
&\simeq \frac{m}{4\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu/L)^2}{(\sigma/L)^2}}\,dx \\
&= \frac{m}{4\pi\sigma^2}\cdot\frac{\sqrt{2\pi}\cdot\sqrt{2}\sigma}{L} \\
&= \frac{m}{2\sqrt{\pi}\sigma L}
\end{aligned}$$

Chapter 2

Base Alignment Quality (BAQ)

Let the reference sequence be $x = r_1 \dots r_L$. We can use a profile HMM to simulate how a read $y = \hat{}c_1 \dots c_l\$$ with quality $z = q_1 \dots q_l$ is generated (or sequenced) from the reference, where ˆ stands for the start of the read sequence and $ for the end.

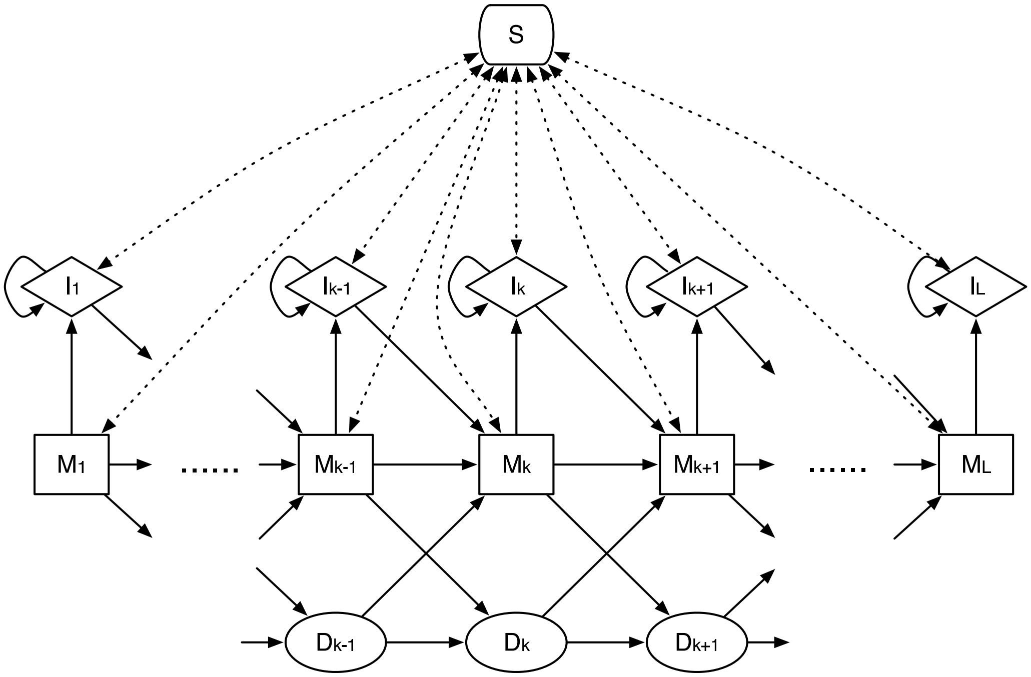

Figure 2.1: A profile HMM for generating sequence reads from a reference sequence, where L is the length of the reference sequence, M states stand for alignment matches, I for alignment insertions to the reference and D states for deletions.

The topology of the profile HMM is given in Fig 2.1. Let $(M, I, D, S) = (0, 1, 2, 3)$. The transition matrix between different types of states is

$$\mathbf{A}=(a_{ij})_{4\times 4}=\begin{pmatrix}(1-2\alpha)(1-s) & \alpha(1-s) & \alpha(1-s) & s\\ (1-\beta)(1-s) & \beta(1-s) & 0 & s\\ 1-\beta & 0 & \beta & 0\\ (1-\alpha)/L & \alpha/L & 0 & 0\end{pmatrix}$$

where α is the gap open probability, β is the gap extension probability and $s=1/(2l)$ with l being the average length of a read. As to emission probabilities, $P(c_i|D_k)=1$, $P(\hat{}|S)=P(\$|S)=1$, $P(c_i|I_k)=0.25$ and

$$P(b_i|M_k)=e_{ki}=\begin{cases}1-10^{-q_i/10} & \text{if } r_k=b_i\\ 10^{-q_i/10}/3 & \text{otherwise}\end{cases}$$

The forward-backward algorithm[1] is as follows:

$$\begin{aligned}
f_S(0) &= 1\\
f_{M_k}(1) &= e_{k1}\cdot a_{30}\\
f_{I_k}(1) &= 0.25\cdot a_{31}\\
f_{M_k}(i) &= e_{ki}\cdot\Big[a_{00}f_{M_{k-1}}(i-1)+a_{10}f_{I_{k-1}}(i-1)+a_{20}f_{D_{k-1}}(i-1)\Big]\\
f_{I_k}(i) &= 0.25\cdot\Big[a_{01}f_{M_k}(i-1)+a_{11}f_{I_k}(i-1)\Big]\\
f_{D_k}(i) &= a_{02}f_{M_{k-1}}(i)+a_{22}f_{D_{k-1}}(i)\\
f_S(l+1) &= \sum_{k=1}^{L}a_{03}f_{M_k}(l)+a_{13}f_{I_k}(l)
\end{aligned}$$

$$\begin{aligned}
b_S(l+1) &= 1\\
b_{M_k}(l) &= a_{03}\\
b_{I_k}(l) &= a_{13}\\
b_{M_k}(i) &= e_{k+1,i+1}a_{00}b_{M_{k+1}}(i+1)+a_{01}b_{I_k}(i+1)/4+a_{02}b_{D_{k+1}}(i)\\
b_{I_k}(i) &= e_{k+1,i+1}a_{10}b_{M_{k+1}}(i+1)+a_{11}b_{I_k}(i+1)/4\\
b_{D_k}(i) &= (1-\delta_{i1})\cdot\Big[e_{k+1,i+1}a_{20}b_{M_{k+1}}(i+1)+a_{22}b_{D_{k+1}}(i)\Big]\\
b_S(0) &= \sum_{k=1}^{L}e_{k1}a_{30}b_{M_k}(1)+a_{31}b_{I_k}(1)/4
\end{aligned}$$

and the likelihood of data is $P(y)=f_S(L+1)=b_S(0)$[2]. The posterior probability of a read base c_i being matching state $\tilde{k}$ (M- or I-typed) is $f_{\tilde{k}}(i)b_{\tilde{k}}(i)/P(y)$.

[1] We may adopt a banded forward-backward approximation to reduce the time complexity. We may also normalize $f_{\tilde{k}}(i)$ for each i to avoid floating point underflow.

[2] Evaluating if $f_S(L+1)=b_S(0)$ helps to check the correctness of the formulae and the implementation.

Chapter 3

Modeling Sequencing Errors

3.1 The revised MAQ model

3.1.1 General formulae

Firstly it is easy to prove that for any $0 \leq \beta_{nk} < 1$ $(0 \leq k \leq n)$,

$$\sum_{k=0}^{n}(1-\beta_{nk})\prod_{l=0}^{k-1}\beta_{nl} = 1 - \prod_{k=0}^{n}\beta_{nk}$$

where we regard that $\prod_{i=0}^{-1}\beta_{ni} = 1$. In particular, when $\exists k \in [0, n]$ satisfies $\beta_{nk} = 0$, we have:

$$\sum_{k=0}^{n}(1-\beta_{nk})\prod_{i=0}^{k-1}\beta_{ni} = 1$$

If we further define:

$$\alpha_{nk} = (1-\beta_{nk})\prod_{i=0}^{k-1}\beta_{ni} \tag{3.1}$$

on the condition that some $\beta_{nk} = 0$, we have:

$$\sum_{k=0}^{n}\alpha_{nk} = 1$$

$$\beta_{nk} = 1 - \frac{\alpha_{nk}}{1-\sum_{i=0}^{k-1}\alpha_{ni}} = \frac{1-\sum_{i=0}^{k}\alpha_{ni}}{1-\sum_{i=0}^{k-1}\alpha_{ni}} = \frac{\sum_{i=k+1}^{n}\alpha_{ni}}{\sum_{i=k}^{n}\alpha_{ni}}$$

In the context of error modeling, if we define:

$$\beta_{nk} \triangleq \begin{cases} \Pr\{\text{at least } k+1 \text{ errors}|\text{at least } k \text{ errors out of } n \text{ bases}\} & (k > 0) \\ \Pr\{\text{at least 1 error out of } n \text{ bases}\} & (k = 0) \end{cases}$$

we have $\beta_{nn} = 0$, and

$$\gamma_{nk} \triangleq \prod_{l=0}^{k-1}\beta_{nl} = \Pr\{\text{at least } k \text{ errors out of } n \text{ bases}\}$$

then

$$\alpha_{nk} = (1-\beta_{nk})\gamma_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\}$$

3.1.2 Modeling sequencing errors

Given a uniform error rate ϵ and independent errors, let

$$\bar{\alpha}_{nk}(\epsilon) = \binom{n}{k}\epsilon^k(1-\epsilon)^{n-k}$$

and

$$\bar{\beta}_{nk}(\epsilon) = \frac{\sum_{i=k+1}^{n}\bar{\alpha}_{ni}(\epsilon)}{\sum_{i=k}^{n}\bar{\alpha}_{ni}(\epsilon)}$$

we can calculate that the probability of seeing at least k errors is

$$\bar{\gamma}_{nk}(\epsilon) = \prod_{l=0}^{k-1}\bar{\beta}_{nk}(\epsilon)$$

When errors are dependent, the true β_{nk} will be larger than $\bar{\beta}_{nk}$. A possible choice of modeling this is to let

$$\beta_{nk} = \bar{\beta}_{nk}^{f_k}$$

where $0 < f_k \leq 1$ models the dependency for k-th error. The probability of seeing at least k errors is thus

$$\gamma_{nk}(\epsilon) = \prod_{l=0}^{k-1}\bar{\beta}_{nl}^{f_l}(\epsilon)$$

For non-uniform errors $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_n$, we may approximate $\gamma_{nk}(\vec{\epsilon})$ as

$$\gamma_{nk}(\vec{\epsilon}) = \prod_{l=0}^{k-1}\bar{\beta}_{nl}^{f_l}(\epsilon_{l+1})$$

3.1.3 Practical calculation

We consider diploid samples only. Let $g \in \{0,1,2\}$ be the number of reference alleles. Suppose there are k reference alleles whose base error rates are $\epsilon_1 \leq \cdots \leq \epsilon_k$, and there are $n-k$ alternate alleles whose base error rates are $\epsilon'_1 \leq \cdots \leq \epsilon'_{n-k}$. We calculate

$$P(D|0) = \gamma_{nk}(\vec{\epsilon}) = \prod_{l=0}^{k-1}\bar{\beta}_{nl}^{f_l}(\epsilon_{l+1})$$

$$P(D|2) = \gamma_{nk}(\vec{\epsilon}') = \prod_{l=0}^{n-k-1}\bar{\beta}_{nl}^{f_l}(\epsilon'_{l+1})$$

and

$$P(D|1) = \frac{1}{2^n}\binom{n}{k}$$

where $f_l = 0.97\eta^{\kappa_l - 1} + 0.03$ with κ_l being the rank of base l among the same type of bases on the same strand, ordered by error rate. For sequencing data, error rates are usually discretized. We may precompute $\bar{\beta}_{nk}(\epsilon)$ for sufficiently large n[1] and all possible discretized ϵ. Calculating the likelihood of the data is trivial.

[1] SAMtools precomputes a table for $n \leq 255$. Given higher coverage, it randomly samples 255 reads.

3.1.4 The original MAQ model

The original MAQ models the likelihood of data by

$$\alpha_{nk}(\epsilon) = (1 - \bar{\beta}_{nk}^{f_k}) \prod_{i=0}^{k-1} \bar{\beta}_{ni}^{f_i}$$

instead of $\gamma_{nk}(\epsilon)$. For non-uniform errors,

$$\alpha_{nk}(\vec{\epsilon}) = c_{nk}(\bar{\epsilon}) \cdot \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i}$$

where

$$\log \bar{\epsilon} = \frac{\sum_{i=0}^{k-1} f_i \log \epsilon_{i+1}}{\sum_{i=0}^{k-1} f_i}$$

and

$$c_{nk}(\bar{\epsilon}) \triangleq \left[1 - \bar{\beta}_{nk}^{f_k}(\bar{\epsilon})\right] \prod_{i=0}^{k-1} \left[\frac{\bar{\beta}_{ni}(\bar{\epsilon})}{\bar{\epsilon}}\right]^{f_i}$$

The major problem with the original MAQ model is that for ϵ close to 0.5 and large n, the chance of seeing no errors may be so small that it is even smaller than the chance of seeing all errors (i.e. $\alpha_{n0} < \alpha_{nn}$). In this case, the model prefers seeing all errors, which is counterintuitive. The revised model uses the accumulative probability γ_{nk} and does not have this problem. For small ϵ and n, the original and the revised MAQ models seem to have similar performance.

Chapter 4

Modeling Multiple Individuals

4.1 Notations

Suppose there are N sites from n individuals with i-th individual having m_i ploids. Let $M = \sum_i m_i$ be the total number of chromosomes. Let $\mathbf{D} = (\vec{D}_1, \ldots, \vec{D}_N)^{\mathrm{T}}$ be the data matrix with vector $\vec{D}_a = (D_{a1}, \ldots, D_{an})$ representing the alignment data for each individual at site a. Similary, let $\mathbf{G} = (\vec{G}_a, \ldots, \vec{G}_N)^{\mathrm{T}}$ and $\vec{G}_a = (G_{a1}, \ldots, G_{an})$ be the true genotypes, where $0 \leq G_{ai} \leq m_i$ equals the number of reference alleles [1]. Define

$$X_a = X_a(\vec{G}_a) \triangleq \sum_i G_{ai} \tag{4.1}$$

to be the number of reference alleles at site a and $\mathbf{X} = (X_1, \ldots, X_N)^{\mathrm{T}}$. Also define $\Phi = (\phi_0, \ldots, \phi_M)$ as the allele frequency spectrum (AFS) with $\sum_k \phi_k = 1$.

For convenience, we may drop the position subscript a when it is unambiguous in the context that we are looking at one locus. Also define

$$P(D_i|g_i) \triangleq \Pr\{D_i|G_i = g_i\} \tag{4.2}$$

to be the likelihood of the data for individual i when the underlying genotype is known. $P(D_i|g_i)$ is calculated in Section 4.4. And define

$$P(g_i|\phi) \triangleq \binom{m_i}{g_i} \phi^{g_i} (1-\phi)^{m_i - g_i} \tag{4.3}$$

to be the probability of a genotype under the Hardy-Weinberg equilibrium (HWE), when the site allele frequency is ϕ.

4.2 Estimating AFS

4.2.1 The EM procedure

We aim to find Φ that maximizes $P(\mathbf{D}|\Phi)$ by EM. Suppose at the t-th iteration the estimate is Φ_t. We have

$$\log \Pr\{\mathbf{D}, \mathbf{X} = \mathbf{x}|\Phi\} = \log \Pr\{\mathbf{D}|\mathbf{X} = \mathbf{x}\} \Pr\{\mathbf{X} = \mathbf{x}|\Phi\} = C + \sum_a \log \phi_{x_a}$$

[1] If we take the ancestral sequence as the reference, the non-reference allele will be the derived allele.

where C is not a function of $\{\phi_k\}$. The EM Q function is[2]

$$\begin{aligned} Q(\Phi|\Phi_t) &= \sum_{\mathbf{x}} \Pr\{\mathbf{X}=\mathbf{x}|\mathbf{D},\Phi_t\} \log \Pr\{\mathbf{D},\mathbf{X}=\mathbf{x}|\Phi\} \\ &= C + \sum_{\mathbf{x}} \prod_a \Pr\{X_a = x_a|\vec{D}_a,\Phi_t\} \sum_b \log \phi_{x_b} \\ &= C + \sum_{a=1}^{N} \sum_{x_a=0}^{M} \Pr\{X_a = x_a|\vec{D}_a,\Phi_t\} \log \phi_{x_a} \end{aligned}$$

Requiring $\partial_{\phi_k}(Q - \lambda\sum_l \phi_l) = 0$ leads to

$$\frac{1}{\phi_k}\sum_a \Pr\{X_a = k|\vec{D}_a,\Phi_t\} - \lambda = 0$$

from which λ can be calculated as:

$$\lambda = \sum_k \sum_a \Pr\{X_a = k|\vec{D}_a,\Phi_t\} = N$$

and thus at the $(t+1)$ iteration:

$$\phi_k^{(t+1)} = \frac{1}{N}\sum_a \Pr\{X_a = k|\vec{D}_a,\Phi_t\} \tag{4.4}$$

where $\Pr\{X_a = k|\vec{D}_a,\Phi_t\}$ is calculated as follows.

4.2.2 The distribution of site reference allele count

Firstly, as we are only looking at a site from now on, we drop subscript a for convenience. Without proof[3], we note that

$$\binom{M}{k} \equiv \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \prod_i \binom{m_i}{g_i}$$

where

$$s_j(\vec{g}) \triangleq \sum_{i=1}^{j} g_i$$

and $\delta_{kl} = 1$ if $k = l$ and 0 otherwise. The probability of sampling $\vec{g}$ conditional on $\sum_i g_i = k$ is $\delta_{k,s_n(\vec{g})}\prod_i \binom{m_i}{g_i}/\binom{M}{k}$. With this preparation, we can calculate[4]

$$\begin{aligned} \Pr\{\vec{D}|X=k\} &= \sum_{\vec{g}} \Pr\{\vec{D},\vec{G}=\vec{g}|X=k\} \\ &= \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \Pr\{\vec{D}|\vec{G}=\vec{g}\} \Pr\{\vec{G}=\vec{g}|X=k\} \\ &= \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \prod_i P(D_i|g_i) \cdot \frac{\prod_j \binom{m_j}{g_j}}{\binom{M}{k}} \\ &= \frac{1}{\binom{M}{k}} \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \prod_i \binom{m_i}{g_i} P(D_i|g_i) \end{aligned}$$

[2] We assume site independency in the following.

[3] Supposedly, this can be proved by polynomial expansion. Wiki gives a simplified version of this formula as generalized Vandermonde's identity.

[4] The derivation below does *not* assume Hardy-Weinberg equilibrium.

where $P(D_i|g_i) \triangleq \Pr\{D_i|G_i = g_i\}$ is the likelihood of data when the underlying genotype is known. To calculate $\Pr\{\vec{D}|X = k\}$, we define

$$z_{jk} \triangleq \sum_{g_1=0}^{m_1} \cdots \sum_{g_j=0}^{m_j} \prod_{i=1}^{j} \binom{m_i}{g_i} P(D_i|g_i)$$

for $0 \le k \le \sum_{i=1}^{j} m_i$ and 0 otherwise. It can be calculated iteratively with[5]

$$z_{jk} = \sum_{g_j=0}^{m_j} z_{j-1,k-g_j} \cdot \binom{m_j}{g_j} P(D_j|g_j) \tag{4.5}$$

with $z_{00} = 1$. Thus

$$\Pr\{\vec{D}|X = k\} = \frac{z_{nk}}{\binom{M}{k}}$$

and

$$\Pr\{X = k|\vec{D}, \Phi\} = \frac{\phi_k \Pr\{\vec{D}|X = k\}}{\sum_l \phi_l \Pr\{\vec{D}|X = l\}} = \frac{\phi_k z_{nk}/\binom{M}{k}}{\sum_l \phi_l z_{nl}/\binom{M}{l}} \tag{4.6}$$

4.2.3 Numerical stability

Numerical computation of Eq. (4.5) may lead to floating point underflow for large n. To avoid this, let $y_{jk} = z_{jk}/\binom{M_j}{k}$, where $M_j = \sum_{i=0}^{j} m_i$. Eq. (4.5) becomes

$$y_{jk} = \left(\prod_{l=0}^{m_j-1} \frac{k-l}{M_{j-1}-l}\right) \cdot \sum_{g_j=0}^{m_j} \left(\prod_{l=g_j}^{m_j-1} \frac{M_{j-1}-k+l+1}{k-l}\right) \cdot y_{j-1,k-g_j} \cdot \binom{m_j}{g_j} P(D_j|g_j)$$

In case of diploid samples,

$$\begin{aligned} y_{jk} \quad &= \quad \frac{1}{2j(2j-1)}\Big[(2j-k)(2j-k-1) \cdot y_{j-1,k} P(D_j|\langle aa\rangle) + 2k(2j-k) \cdot y_{j-1,k-1} P(D_j|\langle Aa\rangle) \\ &\quad +k(k-1) \cdot y_{j-1,k-2} P(D_j|\langle AA\rangle)\Big] \end{aligned}$$

However, this is not good enough. y_{jk} still decreases exponentially with increasing j. To solve this issue, we rescale y_{jk} for each j. Define

$$\tilde{y}_{jk} = \frac{y_{jk}}{\prod_{i=1}^{j} t_i}$$

where t_j is chosen such that $\sum_l \tilde{y}_{jl} = 1$. The posterior is

$$\Pr\{X = k|\vec{D}, \Phi\} = \frac{\phi_k \tilde{y}_{nk}}{\sum_l \phi_l \tilde{y}_{nl}}$$

It should be noted that $P(D_i|g_i)$ can also be rescaled without affecting the calculation of the posterior. Furthermore, in the $\{\tilde{y}_{jk}\}$ matrix, most of cells should be close to zero. Computation of $\tilde{y}_{nk}$ can be carried in a band instead of in a triangle. For large n, this may considerably reduce computing time.

[5] To make it explicit, for diploid samples, if A is the reference allele:

$$z_{jk} = z_{j-1,k} P(D_j|\langle aa\rangle) + 2z_{j-1,k-1} P(D_j|\langle Aa\rangle) + z_{j-1,k-2} P(D_j|\langle AA\rangle)$$

4.2.4 The initial AFS

The EM procedure garantees that $\Pr\{\mathbf{D}|\Phi\}$ monotomically increases with each iteration and converges to a local optima. However, if we start this iteration from a bad initial AFS, we may need many iterations; the iteration is also more likely to be trapped by a local optima. Here we give several AFS on different conditions under the infinite-site Wright-Fisher model.

Let ϕ'_k be the probability of seeing k non-reference alleles out of M chromosomes. The frequency of reference alleles ϕ_k equals ϕ'_{M-k}.

If we take the ancestral sequence as the reference, the standard model gives $\phi'_k = \theta/k$ and $\phi'_0 = 1 - \sum_k \phi'_k$. When we do not know if the reference allele is ancestral, the same conclusion still stands. To see this, for $k > 0$:

$$\phi'_k = \frac{M+1-k}{M+1}\left(\frac{\theta}{k} + \frac{\theta}{M+1-k}\right) = \frac{\theta}{k}$$

and for $k = 0$:

$$\phi'_k = 1 - \sum_{k=1}^{M+1}\frac{\theta}{k} + \frac{\theta}{M+1} = 1 - \sum_{k=1}^{M}\frac{\theta}{k}$$

where the first term corresponds to the case wherein the reference is ancestral and the second to the case wherein the reference is derived.

Another useful AFS is the derived allele frequency spectrum on the condition of loci being discovered from two chromosomes. Under the Wright-Fisher model, it is:

$$\phi'_k = \frac{2(M+1-k)}{(M+1)(M+2)}$$

A third AFS is the derived allele frequency spectrum on the condition of knowing one derived allele from a chromosome. It is a flat distribution

$$\phi'_k = \frac{1}{M+1}$$

4.2.5 Estimating site allele frequency

Here we aim to find ϕ that maximises $\Pr\{\vec{D}|\phi\}$. We have:

$$\log\Pr\{\vec{D},\vec{G}=\vec{g}|\phi\} = \log\prod_i P(D_i|g_i)P(g_i|\phi) = C + \sum_i \log P(g_i|\phi)$$

Given an estimate ϕ_t at the t-th iteration, the $Q(\phi|\phi_t)$ function of EM is[6]:

$$\begin{aligned}
Q(\phi|\phi_t) &= \sum_{\vec{g}} \Pr\{\vec{G}=\vec{g}|\vec{D},\phi_t\}\log\Pr\{\vec{D},\vec{G}=\vec{g}|\phi\} \\
&= C + \sum_{\vec{g_i}}\prod_i \Pr\{G_i = g_i|D_i,\phi_t\}\sum_j \log P(g_j|\phi) \\
&= C + \sum_{i=1}^{n}\sum_{g_i=0}^{m_i} \Pr\{G_i = g_i|D_i,\phi_t\}\log P(g_i|\phi) \\
&= C + \sum_i\sum_{g_i} \Pr\{G_i = g_i|D_i,\phi_t\}\log_i \binom{m_i}{g_i}\phi^{g_i}(1-\phi)^{m_i-g_i} \\
&= C' + \sum_i\sum_{g_i} \Pr\{G_i = g_i|D_i,\phi_t\}\Big[g_i\log\phi + (m_i-g_i)\log(1-\phi)\Big]
\end{aligned}$$

[6] We assume Hardy-Weinberg equilibrium in the derivation.

Requiring $\partial_\phi Q\Big|_{\phi=\phi_{t+1}} = 0$ gives:

$$\frac{1}{\phi_{t+1}(1-\phi_{t+1})}\sum_i\sum_{g_i}\Pr\{G_i=g_i|D_i,\phi_t\}(g_i-m_i\phi_{t+1})=0$$

Thus

$$\phi_{t+1}=\frac{1}{\sum_j m_j}\sum_i\sum_{g_i}g_i\Pr\{G_i=g_i|D_i,\phi_t\}=\frac{1}{M}\sum_i\frac{\sum_{g_i}g_iP(D_i|g_i)P(g_i|\phi_t)}{\sum_{g_i}P(D_i|g_i)P(g_i|\phi_t)} \tag{4.7}$$

which is the EM estimate at the $(t+1)$-th iteration and also the expected reference allele frequency.

4.2.6 Joint distribution of allele counts for 2 samples

Suppose at a locus we have two data sets $\vec{D}'$ and $\vec{D}''$, with n' samples m' haplotypes and n'' samples and m'' haplotypes, respectively. Let $n = n' + n''$ and $m = m' + m''$ and $\vec{D} = \vec{D}' \oplus \vec{D}''$ be the joint of the two data sets. In addition, let $X' = \sum_{i'} G'_{i'}$, $X'' = \sum_{i''} G''_{i''}$ and $X = X' + X''$. We have

$$\Pr\{\vec{D}'|X'=k'\}=\frac{z'_{n'k'}}{\binom{M'}{k'}}$$

$$\Pr\{\vec{D}''|X''=k''\}=\frac{z''_{n''k''}}{\binom{M''}{k''}}$$

and

$$\Pr\{\vec{D}',\vec{D}''|X'=k',X''=k''\}=\Pr\{\vec{D}'|X'=k'\}\Pr\{\vec{D}''|X''=k''\}=\frac{z'_{n'k'}z''_{n''k''}}{\binom{M'}{k'}\binom{M''}{k''}}$$

where

$$z'_{j'k'}\triangleq\sum_{g'_1=0}^{m'_1}\cdots\sum_{g'_j=0}^{m'_j}\prod_{i'=1}^{j'}\binom{m'_i}{g'_i}P(D'_i|g'_i)$$

and similar to $z''_{j''k''}$. If we note that

$$\Pr\{X'=k',X''=k''|\Phi\}=\phi_k\cdot\frac{\binom{M'}{k'}\binom{M''}{k''}}{\binom{M}{k}}$$

and by definition

$$z_{nk}=z_{n'+n'',k'+k''}=\sum_{\{k',k''|k'+k''=k\}}z'_{n'k'}z''_{n''k''}$$

we have

$$\Pr\{\vec{D}|\Phi\}=\sum_k\Pr\{\vec{D},X=k|\Phi\}=\sum_{k',k''}\Pr\{\vec{D}',\vec{D}'',X'=k',X''=k''|\Phi\}$$

Thus we can derive the joint distribution:

$$\Pr\{X'=k',X''=k''|\vec{D},\Phi\}=\frac{\phi_{k'+k''}z'_{n'k'}z''_{n''k''}/\binom{M}{k'+k''}}{\sum_l\phi_l z_{nl}/\binom{M}{l}} \tag{4.8}$$

If we let $y_{jk} = z_{jk}/\binom{M_j}{k}$ as in Section 4.2.3,

$$\Pr\{X' = k', X'' = k''|\vec{D}, \Phi\} = \frac{\phi_{k'+k''} y'_{n'k'} y''_{n''k''}}{\sum_l \phi_l y_{nl}} \cdot \frac{\binom{M'}{k'}\binom{M''}{k''}}{\binom{M'+M''}{k'+k''}}$$

This derivation can be extended to arbitrary number of data sets.

4.2.7 Estimating 2-locus haplotype frequency

In this section, we only consider diploid samples (i.e. $m_1 = \cdots = m_n = 2$). Let $\mathbf{D} = (\vec{D}, \vec{D}')$ be the data at two loci, respectively; and H_i and $H_i^\dagger$ be the two underlying haplotypes for individual i with $H_i \in \{0, 1, 2, 3\}$ representing one of the four possible haplotypes at the 2 loci. We write $\mathbf{H} = \overrightarrow{(H_i, H_i^\dagger)}$ as a haplotype configuration of the samples. Define

$$\mathcal{G}_{hk} = \lfloor h/2 \rfloor + \lfloor k/2 \rfloor$$

$$\mathcal{G}'_{hk} = (h \mod 2) + (k \mod 2)$$

which calculate the genotype of each locus, respectively.

$$\begin{aligned}
Q(\vec{\phi}|\vec{\phi}^{(t)}) &= \sum_{\mathbf{h}} P(\mathbf{H} = \mathbf{h}|\mathbf{D}, \vec{\phi}^{(t)}) \log \Pr\{\mathbf{D}, \mathbf{H} = \mathbf{h}|\vec{\phi}\} \\
&= C + \sum_{\mathbf{h}} \prod_i \Pr\{H_i = h_i, H_i^\dagger = h_i^\dagger|\mathbf{D}, \vec{\phi}^{(t)}\} \sum_j \log P(h_j, h_j^\dagger|\vec{\phi}) \\
&= C + \sum_i \sum_{h_i} \sum_{h_i^\dagger} \Pr\{H_i = h_i, H_i^\dagger = h_i^\dagger|\mathbf{D}, \vec{\phi}^{(t)}\} \sum_j \log(\phi_{h_i} \phi_{h_i^\dagger})
\end{aligned}$$

Solving $\partial_{\phi_k} Q - \lambda = 0$ gives

$$\begin{aligned}
\phi_k &= \frac{1}{2n} \sum_i \sum_h \left(\Pr\{H_i = h, H_i^\dagger = k|\mathbf{D}, \vec{\phi}^{(t)}\} + \Pr\{H_i = k, H_i^\dagger = h|\mathbf{D}, \vec{\phi}^{(t)}\} \right) \\
&= \frac{\phi_k^{(t)}}{2n} \sum_{i=1}^n \frac{\sum_h \phi_h^{(t)} \left[P(D_i|\mathcal{G}_{hk}) P(D'_i|\mathcal{G}'_{hk}) + P(D_i|\mathcal{G}_{kh}) P(D'_i|\mathcal{G}'_{kh}) \right]}{\sum_{k',h} \phi_{k'}^{(t)} \phi_h^{(t)} P(D_i|\mathcal{G}_{hk'}) P(D'_i|\mathcal{G}'_{hk'})}
\end{aligned}$$

4.3 An alternative model

In Section 4.2, ϕ_k in $\Phi = \{\phi_k\}$ is interpreted as the probability of seeing exactly k alleles from M chromosomes. Under this model, the prior of a genotype configuration is

$$\Pr\{\vec{G} = \vec{g}|\Phi\} = \phi_{s_n(\vec{g})} \frac{\prod_i \binom{m_i}{g_i}}{\binom{M}{s_n(\vec{g})}}$$

and the posterior is

$$\Pr\{\vec{G} = \vec{g}|\vec{D}, \Phi\} = \frac{\phi_{s_n(\vec{g})}}{\Pr\{\vec{D}|\Phi\}} \cdot \frac{\prod_i \binom{m_i}{g_i} P(D_i|g_i)}{\binom{M}{s_n(\vec{g})}}$$

Suppose we want to calculate the expectation of $\sum_i f_i(g_i)$, we can

$$\sum_i \sum_{\vec{g}} f_i(g_i) \Pr\{\vec{G} = \vec{g}|\vec{D}, \Phi\} = \frac{1}{\Pr\{\vec{D}|\Phi\}} \sum_i \sum_k \frac{\phi_k}{\binom{M}{k}} \sum_{\vec{g}} \delta_{k, s_n(\vec{g})} f_i(g_i) \prod_j \binom{m_j}{g_j} P(D_j|g_j)$$

Due to the presence of $\delta_{k,s_n(\vec{g})}$, we are unable to reduce the formula to a simpler form. Although we can take a similar strategy in Section 4.2 to calculate $\sum_k \sum_{\vec{g}}$, which is $O(n^2)$, another sum $\sum_i$ will bring this calculation to $O(n^3)$. Even calculating the marginal probability $\Pr\{G_i = g_i|\vec{D}, \Phi\}$ requires this time complexity. All the difficulty comes from that individuals are correlated conditional on $\{X = k\}$.

An alternative model is to interpret the AFS as the discretized AFS of the population rather than for the observed individuals. We define the population AFS discretized on M chromosomes as $\Phi' = \{\phi'_k\}$. Under this model,

$$\Pr\{\vec{G} = \vec{g}|\Phi'\} = \sum_k \phi_k \prod_i P(g_i|k/M)$$

$$\Pr\{\vec{G} = \vec{g}, \vec{D}|\Phi'\} = \sum_k \phi_k \prod_i P(D_i|g_i)P(g_i|k/M)$$

$$\Pr\{\vec{D}|\Phi'\} = \sum_{\vec{g}} \Pr\{\vec{G} = \vec{g}, \vec{D}|\Phi'\} = \sum_{k=0}^{M} \phi'_k \prod_{i=1}^{n} \sum_{g_i=0}^{m_i} P(D_i|g_i)P(g_i|k/M)$$

and

$$\begin{aligned}
& \sum_i \sum_{\vec{g}} f_i(g_i) \Pr\{\vec{G} = \vec{g}|\vec{D}, \Phi'\} & (4.9)\\
= \; & \frac{1}{\Pr\{\vec{D}|\Phi'\}} \sum_i \sum_{\vec{g}} f_i(g_i) \sum_k \phi'_k \prod_j P(D_j|g_j)P(g_j|k/M) \\
= \; & \frac{1}{\Pr\{\vec{D}|\Phi'\}} \sum_k \phi'_k \sum_i f_i(g_i)P(D_i|g_i)P(g_i|k/M) \prod_{j\neq i} \sum_{g_j} P(D_j|g_j)P(g_j|k/M) \\
= \; & \frac{1}{\Pr\{\vec{D}|\Phi'\}} \sum_k \phi'_k \left[\prod_i \sum_{g_i} P(D_i|g_i)P(g_i|k/M)\right] \cdot \left[\sum_i \frac{\sum_{g_i} f_i(g_i)P(D_i|g_i)P(g_i|k/M)}{\sum_{g_i} P(D_i|g_i)P(g_i|k/M)}\right]
\end{aligned}$$

The time complexity of this calculation is $O(n^2)$. Consider that if $f_i(g_i) = g_i$, $X = \sum_i f_i(G_i)$. We can easily calculate $E(X|\vec{D}, \Phi')$ with the formula above.

4.3.1 Posterior distribution of the allele count

Under the alternative model, we can also derive the posterior distribution of X, $\Pr\{X = k|\vec{D}, \Phi'\}$ as follows.

$$\Pr\{X = k|\Phi'\} = \binom{M}{k} \sum_{l=0}^{M} \phi'_l \left(\frac{l}{M}\right)^k \left(1 - \frac{l}{M}\right)^{M-k}$$

Then

$$\Pr\{\vec{D}, X = k|\Phi'\} = z_{nk} \sum_{l=0}^{M} \phi'_l \left(\frac{l}{M}\right)^k \left(1 - \frac{l}{M}\right)^{M-k} \tag{4.10}$$

In fact, we also have an alternative way to derive this $\Pr\{\vec{D}, X = k|\Phi'\}$. Let ϕ' be the true site allele frequency in the population. Assuming HWE, we have

$$\begin{aligned}
\Pr\{X=k,\vec{D}|\phi'\} &= \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \Pr\{\vec{D}|\vec{G}=\vec{g}\} \Pr\{\vec{G}=\vec{g}|\phi'\} \qquad (4.11)\\
&= \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \prod_i P(D|g_i) \binom{m_i}{g_i} \phi'^{g_i}(1-\phi')^{m_i-g_i}\\
&= \phi'^k(1-\phi')^{M-k} \sum_{\vec{g}} \delta_{k,s_n(\vec{g})} \prod_i \binom{m_i}{g_i} P(D_i|g_i)\\
&= \phi'^k(1-\phi')^{M-k} z_{nk}
\end{aligned}$$

Summing over the AFS gives

$$\Pr\{\vec{D},X=k|\Phi'\} = \sum_l \phi'_l \Pr\{X=k,\vec{D}|\phi'=l/M\}$$

which is exactly Eq. (4.10). It is worth noting that $E(X|\vec{D},\Phi')$ calculated by Eq. (4.11) is identical to the one calculated by Eq. (4.9), which has been numerically confirmed.

In practical calculation, the alternative model has very similar performance to the method in Section 4.2 in one iteration. However, in proving EM, we require $\Pr\{X=k|\Phi\}=\phi_k$, which does not stand any more in the alternative interpretation. Iterating Eq. (4.10) may not monotonically increase the likelihood function. Even if this was also another different EM procedure, we have not proved it yet. Yi *et al.* (2010) essentially calculates Eq. (4.11) for ϕ' estimated from data without summing over the AFS. Probably this method also delivers similar results, but it is not theoretically sound and may not be iterated, either.

4.4 Likelihood of data given genotype

Given a site covered by k reads from an m-ploid individual, the sequencing data is:

$$D=(b_1,\ldots,b_k)=(\underbrace{1,\ldots,1}_{l},\underbrace{0,\ldots,0}_{k-l})$$

where 1 stands for a reference allele and 0 otherwise. The j-th base is associated with error rate ϵ_j, which is the larger error rate between sequencing and alignment errors. We have

$$P(D|0)=\prod_{j=1}^{l}\epsilon_j\prod_{j=l+1}^{k}(1-\epsilon_j)=\left(1-\sum_{j=l+1}^{k}\epsilon_j+o(\epsilon^2)\right)\prod_{j=1}^{l}\epsilon_j \qquad (4.12)$$

$$P(D|m)=\left(1-\sum_{j=1}^{l}\epsilon_j+o(\epsilon^2)\right)\prod_{j=l+1}^{k}\epsilon_j \qquad (4.13)$$

For $0 < g < m$:

$$\begin{aligned}
P(D|g) &= \sum_{a_1=0}^{1}\cdots\sum_{a_k=0}^{1}\Pr\{D|B_1=a_1,\ldots,B_k=a_k\}\Pr\{B_1=a_1,\ldots,B_k=a_k|g\} \qquad (4.14)\\
&= \sum_{\vec{a}}\left(\frac{g}{m}\right)^{\sum_j a_j}\left(1-\frac{g}{m}\right)^{k-\sum_j a_j}\cdot\prod_j p_j(a_j)\\
&= \left(1-\frac{g}{m}\right)^k\prod_j\sum_{a=0}^{1}p_j(a)\left(\frac{g}{m-g}\right)^a\\
&= \left(1-\frac{g}{m}\right)^k\prod_{j=1}^{l}\left(\epsilon_j+\frac{g}{m-g}(1-\epsilon_j)\right)\prod_{j=l+1}^{k}\left(1-\epsilon_j+\frac{\epsilon_j g}{m-g}\right)\\
&= \left(1-\frac{g}{m}\right)^k\left\{\left(\frac{g}{m-g}\right)^l+\left(1-\frac{g}{m-g}\right)\left(\sum_{j=1}^{l}\epsilon_j-\sum_{j=l+1}^{k}\epsilon_j\right)+o(\epsilon^2)\right\}
\end{aligned}$$

In the bracket, the first term explains the deviation between l/k and g/m by imperfect sampling, while the second term explains the deviation by sequencing errors. The second term can be ignored when k is small but may play a major role when k is large. In particular, for $m = 2$, $P(D|1) = 2^{-k}$, independent of sequencing errors.

In case of dependent errors, we may replace:

$$\epsilon_1 < \epsilon_2 < \cdots < \epsilon_l$$

with

$$\epsilon'_j = \epsilon_j^{\alpha^{j-1}}$$

where parameter $\alpha \in [0, 1]$ addresses the error dependency.

4.5 Multi-sample SNP calling and genotyping

The probability of the site being polymorphic is $\Pr\{X = 0|\vec{D}, \Phi\}$. For individual i, we may estimate the genotype $\hat{g}_i$ as:

$$\hat{g}_i = \underset{g_i}{\operatorname{argmax}}\Pr\{G_i = g_i|D_i, \phi_E\} = \underset{g_i}{\operatorname{argmax}}\frac{P(D_i|g_i)P(g_i|\phi_E)}{\sum_{h_i}P(D_i|h_i)P(h_i|\phi_E)}$$

where

$$\phi_E = E(X|\vec{D}, \Phi)/M$$

This estimate of genotypes may not necessarily maximize the posterior probability $P(\vec{g}|\vec{D})$, but it should be good enough in practice. However, it should be noted that $\sum_i \hat{g}_i$ is usually a bad estimator of site allele frequency. The max-likelihood estimator by Eq. (4.7) is much better.