Q-consensus: Read Quality Scores for Consensus Sequences

Frederik Espersen Knudsen

Foreword

Why Phred Scores for Consensus Sequences? How much confidence do you have in a sequence, when you have multiple reads to infer it with? This is a fundamental question in many applications where the amount and quality of reads are imperative, such as when deciding on sequencing depth beforehand or assessing whether a polymorphism is a consequence of an error.

In such cases, having a concrete quantitative measure of sequencing confidence is practical. Here I introduce *Q-consensus* - a read quality score measure for alignments of reads, similar to the conventional position-wise Phred / Q-scores for single reads in FASTQ-format.

Why Can't We Multiply Phred Scores / Add Q-scores? Phred- / Q-scores are error rates that indicate the probability that the called base is incorrect.

Multiplying Phred scores (or equivalently adding the logarithmically transformed Q-scores) will underestimate the final Phred / Q-score given the information contained in multiple reads. This can be demonstrated in practice by trying to multiply the probability of each base given each independent individual read and observing that the probabilities do not add up to 1 when considering all possible bases. As shown in the probabilistic model below, these multiplied probabilities must be normalised.

Additionally, Phred / Q-scores cannot straightforwardly be combined when different reads don't agree on the called base at a position. This is also dealt with in the probabilistic model.

Why Must Reads Be Pre-aligned? To assess to what extent reads agree on a certain base call, we need to know their common frame of reference. This is done by a pre-alignment of reads, which is usually very feasible when reads share a common sequence region, such as a highly conserved region or a designed framework.

A Probabilistic Phred Consensus Framework

The input for the model will be a list of N aligned reads $\mathbf{R} = [R_j, R_j, \ldots, R_j]$ called from the same template sequence S.

First, we will condition the true nucleotide probability on a single read. Next, we will condition the true nucleotide probability on a set of reads. Finally, we will deal with gaps in the read alignment.

The conditional probability given a single read

Let S_i be the i'th position of the template sequence S. Similarly, let $\mathbf{R}_i$ be the i'th position of all the reads $\mathbf{R}_i = [R_{i,1}, R_{i,2}, \ldots, R_{i,N}]$.

Let the true nucleotide at position i be $S_i = a$ and the base-called nucleotides in the reads be $R_{i,j} = b_j$, where $a, b_j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

We wish to determine the distribution of $\mathbb{P}(S_i = a \mid R_{i,j} = b_j)$ over all possible true nucleotides $a \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. The term $\mathbb{P}(S_i = a \mid R_{i,j} = b_j)$ can be thought of conceptually, using $\mathbb{P}(S_{36} = \mathrm{A} \mid R_{36,3} = \mathrm{C})$ as an example, as "the probability that the true base at position 36 is 'A' when read #3 has called it as 'C'."

Each called nucleotide $R_{i,j} = b_j$ in a read is associated with a sequencing error rate $\varepsilon_{i,j}$ (Phred) conveyed through its Q-score in FASTQ-format. It encodes the probability that the called nucleotide b_j is different from the true value

$$\begin{aligned}\mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j) &= \varepsilon_{i,j} \\ &= 10^{-\frac{Q_{i,j}}{10}}\end{aligned}$$

Conversely, the probability that the called nucleotide is true can be expressed as

$$\begin{aligned}\mathbb{P}(S_i = b_j \mid R_{i,j} = b_j) &= 1 - \mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j) \\ &= 1 - 10^{-\frac{Q_{i,j}}{10}}\end{aligned}$$

When the called nucleotide is different from the true nucleotide (i.e. a sequencing error), we will assume no inherent knowledge of the true nucleotide, except that it is not the called base. I.e., if a read position is base-called as $R_{i,j} = \mathrm{C}$, then in the case where the read is wrong, $S_i = a$, where $a \in \{\mathrm{A}, \mathrm{G}, \mathrm{T}\}$. More generally, if $S_i \neq b_j$, then $S_i = n$ where $n \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \setminus b_j$, where $\setminus$ is the set subtraction operator.

The error rate $\mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j)$ can be decomposed into its subevents $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \setminus b_j$. The sum of the probability of the subevents must equate to the error rate

$$\mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j) = \sum_{n \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\} \setminus b_j} \mathbb{P}(S_i = n \mid R_{i,j} = b_j)$$

Assuming there is no bias or prior knowledge about the true nucleotide in the position i (Ass. 1), we can assume that the probability of the alternative erroneous sequencing events is equally likely

$$\mathbb{P}(S_i = a \mid R_{i,j} = b_j) = \mathbb{P}(S_i = a' \mid R_{i,j} = b_j) \qquad \forall\ a, a' \neq b_j$$

In other words, when the base-calling is wrong, we uniformly distribute our confidence in the alternative nucleotides. In that case

$$\begin{aligned}\mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j) &= \sum_{n \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\} \setminus b_j} \mathbb{P}(S_i = n \mid R_{i,j} = b_j) \\ &= 3 \cdot \mathbb{P}(S_i = a \mid R_{i,j} = b_j) && \forall\ a \neq b_j \\ \implies \mathbb{P}(S_i = a \mid R_{i,j} = b_j) &= \frac{1}{3}\mathbb{P}(S_i \neq b_j \mid R_{i,j} = b_j) && \forall\ a \neq b_j \\ &= \frac{1}{3}\varepsilon_{i,j} && \forall\ a \neq b_j\end{aligned}$$

We note that the '3' that appears is simply the length of the set of possible alternative nucleotides $|\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \setminus b_j| = |\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}| - |\{b_j\}| = 4 - 1 = 3$.

Thus, we can condition the probability of the true nucleotide at any position on a base called nucleotide. There is a $1 - \varepsilon_{i,j}$ probability of base called nucleotide in the read being true, and a $\frac{1}{3}\varepsilon_{i,j}$ probability of the true nucleotide being either of the three alternatives

$$\begin{aligned}\mathbb{P}(S_i = a \mid R_{i,j} = b_j) &= \begin{cases} 1 - \varepsilon_{i,j} & \text{if} \quad a = b_j \\ \frac{1}{3}\varepsilon_{i,j} & \text{if} \quad a \neq b_j \end{cases} \\ &= (1 - \varepsilon_{i,j})\ \mathbb{I}(a = b) + \left(\frac{1}{3}\varepsilon_{i,j}\right)(1 - \mathbb{I}(a = b))\end{aligned}$$

, where $\mathbb{I}(\cdot)$ is the indicator function.

The conditional probability given multiple reads

Next, we wish to determine the probability of the true nucleotide conditioned on all reads simultaneously, $\mathbb{P}(S_i = a \mid R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N)$. By Bayes' theorem

$$\begin{aligned}&\mathbb{P}(S_i = a \mid R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N) \\ &\qquad = \frac{\mathbb{P}(R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N \mid S_i = a)\ \mathbb{P}(S_i = a)}{\sum_{n \in \{\mathrm{A,C,G,T}\}} \mathbb{P}(R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N \mid S_i = n)\ \mathbb{P}(S_i = n)}\end{aligned}$$

We will assume (Ass. 2) conditional independence of reads, i.e. that any read occurs independently of any other, given the true nucleotide S_i

$$\mathbb{P}(R_{i,j} = b_j \mid R_{i,j'} = b_{j'}, S_i = a) = \mathbb{P}(R_{i,j} = b_j \mid S_i = a) \qquad \forall j \neq j';\ j, j' \in \{1, 2, \ldots, N\}$$

In that case, we can rewrite the conditional distribution of all reads $R_{i,1},\ R_{i,2}, \ldots,\ R_{i,N}$ on the true nucleotide S_i as their joint probability

$$\begin{aligned}&\mathbb{P}(R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N \mid S_i = a) \\ &\qquad = \mathbb{P}(R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N \mid R_{i,1} = b_1, S_i = a)\mathbb{P}(R_{i,1} = b_1 \mid S_i = a) \\ &\qquad = \mathbb{P}(R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N \mid S_i = a)\mathbb{P}(R_{i,1} =_1 \mid S_i = a) \\ &\qquad = \ldots \\ &\qquad = \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)\end{aligned}$$

For brevity, we will write $R_{i,1} = b_1,\ R_{i,2} = b_2, \ldots,\ R_{i,N} = b_N$ as $\mathbf{R}_i = \mathbf{b}$ going forward.
Thus, the Bayesian likelihood can be expressed as the joint individual likelihood

$$\mathbb{P}(S_i = a \mid \mathbf{R}_i = \mathbf{b}) = \frac{\mathbb{P}(S_i = a)\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)}{\sum_{n \in \{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = n)\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = n)}$$

Given our assumption of no previous prior knowledge of the true sequence S (Ass. 1), we find that any true nucleotide is equally likely given no other information. I.e.

$$\mathbb{P}(S_i = a) = \mathbb{P}(S_i = a') \qquad \forall a, a' \in \{\mathrm{A, C, G, T}\}$$

In our Bayesian expression, this is equivalent to a uniform prior $\mathbb{P}(S)$

$$\begin{aligned}\mathbb{P}(S_i = a \mid \mathbf{R}_i = \mathbf{b}) &= \frac{\mathbb{P}(S_i = a)\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)}{\sum_{n \in \{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = n)\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = n)} \\ &= \frac{\mathbb{P}(S_i = a)\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)}{\mathbb{P}(S_i = a) \sum_{n \in \{\mathrm{A,C,G,T}\}}\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = n)} \\ &= \frac{\prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)}{\sum_{n \in \{\mathrm{A,C,G,T}\}}\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = n)}\end{aligned}$$

The reverse conditional probability for a single read

We do not initially know $\mathbb{P}(R_{i,j} = b_j \mid S_i = a)$. We can, however, derive it by Bayes' theorem, since we know the reverse conditionals $\mathbb{P}(S_i = a \mid R_{i,j} = b_j)$

$$\mathbb{P}(R_{i,j} = b_j \mid S_i = a) = \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)\ \mathbb{P}(R_{i,j} = b_j)}{\sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)\ \mathbb{P}(R_{i,j} = n)}$$

We will assume (Ass. 3) a uniform prior $\mathbb{P}(R_{i,j})$. In other words, we assume that given no knowledge of the template sequence, we expect all nucleotides to be equally likely in a read; i.e. we will assume that the sequencing method has no inherent bias. In that case

$$\mathbb{P}(R_{i,j} = b_j) = \mathbb{P}(R_{i,j} = b'_j) \qquad \forall\ b_j, b'_j$$

Which yields

$$\begin{aligned}
\mathbb{P}(R_{i,j} = b_j \mid S_i = a) &= \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)\ \mathbb{P}(R_{i,j} = b_j)}{\sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)\ \mathbb{P}(R_{i,j} = n)} \\
&= \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)\ \mathbb{P}(R_{i,j} = b_j)}{\mathbb{P}(R_{i,j} = b_j)\ \sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)} \\
&= \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)}{\sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)}
\end{aligned}$$

We note that the denominator sums to 1, by our previous definitions[1] of $\mathbb{P}(S_i = a \mid R_{i,j} = b_j)$

$$\begin{aligned}
\sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n) &= \mathbb{P}(S_i = a \mid R_{i,j} = a) + 3\ \mathbb{P}(S_i = a \mid R_{i,j} \neq a) \\
&= (1 - \varepsilon_{i,j}) + 3\ \left(\frac{1}{3}\varepsilon_{i,j}\right) \\
&= 1 - \varepsilon_{i,j} + \varepsilon_{i,j} \\
&= 1
\end{aligned}$$

Giving

$$\begin{aligned}
\mathbb{P}(R_{i,j} = b_j \mid S_i = a) &= \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)}{\sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)} \\
&= \frac{\mathbb{P}(S_i = a \mid R_{i,j} = b_j)}{1} \\
&= \mathbb{P}(S_i = a \mid R_{i,j} = b_j)
\end{aligned}$$

We can now express the conditional probability of the true sequence given the reads as

$$\begin{aligned}
\mathbb{P}(S_i = a \mid \mathbf{R}_i = \mathbf{b}) &= \frac{\prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = a)}{\sum_{n\in\{\mathrm{A,C,G,T}\}}\ \prod_{j=1}^{N} \mathbb{P}(R_{i,j} = b_j \mid S_i = n)} \\
&= \frac{\prod_{j=1}^{N} \mathbb{P}(S_i = a \mid R_{i,j} = b_j)}{\sum_{n\in\{\mathrm{A,C,G,T}\}}\ \prod_{j=1}^{N} \mathbb{P}(S_i = n \mid R_{i,j} = b_j)}
\end{aligned}$$

We arrive at a somewhat intuitive answer: The conditional probability of the true sequence given all reads is the normalised product of the conditional probability of the true sequence given every single read.

[1] This is a consequence of both $\mathbb{P}(S_i)$ and $\mathbb{P}(R_{i,j})$ being assigned uniform priors. We note that $P(S_i = a) = \sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)\ \mathbb{P}(R_{i,j} = n) \implies \frac{\mathbb{P}(S_i=a)}{\mathbb{P}(R_{i,j}=b_{i,j})} = \sum_{n\in\{\mathrm{A,C,G,T}\}} \mathbb{P}(S_i = a \mid R_{i,j} = n)$, and that $\frac{\mathbb{P}(S_i=a)}{\mathbb{P}(R_{i,j}=b_{i,j})} = 1$ if both priors are uniform, and thus identical.

Alignment gap positions

The current model does not take gaps in the alignment of the reads into account. While the FASTQ format provides error rates for called bases, gaps are outside the FASTQ framework.

Under the assumption that reads all derive from the same template sequence, a gap in an alignment at a position i represents a sequencing error either in the set of reads *with* the alignment gap (deletion sequencing error) or the set of reads *without* the alignment gap (insertion sequencing error), i.e. with a called base at the position.

Let $S_i = \varnothing$ indicate that the position i in the alignment is not part of the true sequence, but the result of an insertion sequencing error in one or more of the reads $R_{i,j} \neq \varnothing$. Let $S_i \neq \varnothing$, where one or more reads contain a gap $R_{i,j} = \varnothing$, indicate a deletion sequencing error in those reads.

Let us first make some simple statements about when either none of the reads or all of the reads have a gap in a specific position. Firstly, we note that any conventional alignment method will never produce a position, where all sequences are assigned a gap; There will always be at least one sequence with a called base in a position, such that $\mathbf{R}_i = \varnothing$ is never occurs. Secondly, we will assume that there is no probability of an insertion sequencing error when none of the reads in an alignment have a gap

$$\begin{aligned} & \mathbb{P}(S_i = \varnothing \mid \mathbf{R}_i \neq \varnothing) = 0 \\ \implies & \mathbb{P}(S_i \neq \varnothing \mid \mathbf{R}_i \neq \varnothing) = 1 \end{aligned}$$

Assuming otherwise would mean that we would have to estimate an insertion sequencing error rate. In this framework, however, we will try to avoid such *a priori* parameters.

This leaves the cases where some of the reads have gaps, and others don't. We will outline some expectations for the reads when there is a gap in the alignment. When $S_i \neq \varnothing$, we expect the base calls at position i to converge to the true nucleotide $S_i = a$. When $S_i = \varnothing$, we expect any base calls ($R_{i,j} \neq \varnothing$) at position i to be random, i.e.

$$\mathbb{E}\left[\mathbb{P}(R_{i,j} = b_j \mid S_i = \varnothing,\ R_{i,j} \neq \varnothing)\right] = \frac{1}{4}$$

Additionally, when $S_i = \varnothing$, we expect more reads to have a gap ($R_{i,j} = \varnothing$) than not ($R_{i,j} \neq \varnothing$) at position i, and vice versa

$$\begin{aligned} & \mathbb{E}\left[\mathbb{P}(R_{i,j} = \varnothing \mid S_i = \varnothing)\right] > \mathbb{E}\left[\mathbb{P}(R_{i,j} \neq \varnothing \mid S_i = \varnothing)\right] \\ \implies & \frac{\mathbb{E}\left[\mathbb{P}(R_{i,j} = \varnothing \mid S_i = \varnothing)\right]}{\mathbb{E}\left[\mathbb{P}(R_{i,j} \neq \varnothing \mid S_i = \varnothing)\right]} > 1 \\ & \mathbb{E}\left[\mathbb{O}(R_{i,j} = \varnothing \mid S_i = \varnothing)\right] > 1 \end{aligned}$$

$$\begin{aligned} & \mathbb{E}\left[\mathbb{P}(R_{i,j} = \varnothing \mid S_i \neq \varnothing)\right] < \mathbb{E}\left[\mathbb{P}(R_{i,j} \neq \varnothing \mid S_i \neq \varnothing)\right] \\ \implies & \frac{\mathbb{E}\left[\mathbb{P}(R_{i,j} = \varnothing \mid S_i \neq \varnothing)\right]}{\mathbb{E}\left[\mathbb{P}(R_{i,j} \neq \varnothing \mid S_i \neq \varnothing)\right]} < 1 \\ & \mathbb{E}\left[\mathbb{O}(R_{i,j} = \varnothing \mid S_i \neq \varnothing)\right] < 1 \end{aligned}$$

$$\implies \mathbb{E}\left[\mathbb{O}(R_{i,j} = \varnothing)\right] \begin{cases} > 1 & \text{if} \quad S_i = \varnothing \\ < 1 & \text{if} \quad S_i \neq \varnothing \end{cases}$$

These expectations could serve as tests to determine, whether or not $S_i = \varnothing$.

To test for a random model of base call in the reads would require to estimate the probability of each base in a gap position. These estimates, however, are unlikely to converge unless a large number of reads are sampled, making this expectation an inconvenient measure for whether or not $S_i = \varnothing$.

Instead, we will use the second expectation, of whether the odds of a read having a gap are greater than or less than 1.

Assumptions

Algorithm

Algorithm 1 Q-CONSENSUS: Calculate Q-scores for a consensus sequence from aligned reads

1: **Input:**

- R: Aligned reads as an $L \times N$ array of bases, where L is the length of alignment and N is the number of reads.
- Q: Q-scores for the bases in R as an $L \times N$ array.
- A: Set of possible nucleotides (default: $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$).

2: **Output:** A tuple (s, q) where:

- s: Consensus sequence of length L.
- q: Associated Q-scores for the consensus sequence.

3: Compute[2] the error rates E as $E_{ij} = 10^{-Q_{ij}/10}$.

4: Initialise the sequence-alphabet mask M:

$$M_{ijk} = \begin{cases} 1 & \text{if } A_k = R_{ij}, \\ 0 & \text{otherwise.} \end{cases}$$

5: Assign the probabilities P for each base:

$$P_{ijk} = M_{ijk} \cdot (1 - E_{ij}) + (1 - M_{ijk}) \cdot \frac{E_{ij}}{|A| - 1}.$$

6: Normalise P for each position:

$$P_{ik} = \frac{\prod_{j=1}^{N} P_{ijk}}{\sum_{k=1}^{|A|} \prod_{j=1}^{N} P_{ijk}}.$$

7: Determine the maximum likelihood consensus sequence s:

$$s_i = A_{\mathrm{argmax}_k \ (P_{ik})}.$$

8: Compute the Q-scores q for s:

$$q_i = -10 \cdot \log_{10}(1 - \max_k(P_{ik})).$$

9: **Return:** (s, q).

Next steps

1. Handle gaps in alignment
2. Indicator function nomenclature for final odds ratio equation
3. Some observations on the model
    - How does the Q-score generally increase when reads agree?
    - What is expected, given an expected distribution of Q-scores?