S1 TEXT

1. Benchmark Haploid Genome Sequencing Using Tiles

Suppose there are a certain number of cells of a *haploid* organism in a test tube, and an organizer runs a contest for sequencing these cells (e.g., the XPrize). The procedure of the contest is as follows.

(1) Each contestant attempts to find the mode genome (i.e., the most frequent genome) among all cells at a *single* tile position along with the mode frequency.
(2) The organizer benchmark each contest's result by comparing it to the organizer's result and judge the accuracy of the contest's answer.

Here we provide a mathematically rigorous framework for step (2) by carrying out hypothesis testing at the tile position. We assume these cells have no spanning tiles at this tile position. Moreover, we assume mutations are rare, and we shall quantify this assumption later. Note that the model described here does not apply to cancer cells.

Example 1.1. Suppose there are 20 cells in the test tube. For simplicity, the given tile position has 17 ATTC, 2 TGCC, and 1 GCTA. In this case, the mode genome is ATTC with frequency 85%.

Due to the current sequencing technology, the sequence of these cells can only be observed through measurements. In other words, we do not know which one of the cells we are measuring each time, but we do know that the random sampling follows the distribution given by the cells in the test tube. More precisely, all tile variants at our tile position of interest are amplified and sequenced without bias, so that the measurements we obtain from sequencing are present in the same proportion as the cell genomes. Thus

(1.1) each measurement in the test tube defines a random variable X.

In the above example, we have probability $p(X = \text{ATTC}) = 85\%$, $p(X = \text{TGCC}) = 10\%$, and $p(X = \text{GCTA}) = 5\%$.

The above type of sequencing may be achieved by using single-molecule template sequencing methods, which do not require PCR amplification and thus do not suffer the same AT-rich and GC-rich amplification biases as clonally amplified templates [1]. Additionally, we assume there are *no sequencing errors*.

Through measurements, each contestant shall estimate the mode and a lower bound of the probability of the mode. We specify that the contestants report their results in the following format.

Definition 1.2. Given a set of measurements $\{x_i\}_{i=1,\dots,N}$, we define a **report** to be a tuple $R = (N, MG, \theta, \alpha)$. Here N is the **sample size**, MG is the **sample mode genome**, $\theta \in (50\%, 100\%]$ is the **mode frequency bound**, and $\alpha \in [0\%, 5\%]$ is the **significance level**. We require θ to be an lower bound for the frequency of the sample mode genome MG. More precisely, we require

$$\theta \leq \tfrac{m}{N},$$

where m is the number of occurrences of MG in the sample $\{x_i\}_{i=1,\dots,N}$.

In the above definition, the requirement $\theta > 50\%$ is a way of saying that the observed mutations are rare and there is a predominant sample mode genome. The mode frequency bound is chosen by the contestant so long as it satisfies $\theta \leq \frac{m}{N}$. The significance level α is also chosen by the contestant, and it will be used in hypothesis testing later.

Example 1.3. Suppose we measured the 20 cells in Example 1.1 for 100 times, and we get 90 ATTC, 6 TGCC, and 4 GCTA. In this case, $N = 100$, $MG =$ ATTC, the number of MG is 90 (with frequency 90%), and we choose to report the mode frequency bound as $\theta = 88\%$ and significance level as $\alpha = 1\%$.

We now use p-value to carry out statistical hypothesis testing on a report. Each report $R = (N, MG, \theta, \alpha)$ establishes a hypothesis H_1 on the cells in the test tube.

$$H_1\text{: the probability } p(X = MG) \geq \theta, \tag{1.2}$$

where X is the random variable of a single measurement in the test tube as in (1.1). The opposite of the hypothesis H_1 is the null hypothesis

$$H_0\text{: the probability } p(X = MG) < \theta. \tag{1.3}$$

In order for a test statistic to support H_1, we need the p-value less than or equal to the significance level, i.e.,

$$\text{Test statistic supports } H_1 \text{ if } p(\text{test statistic} \,|\, H_0) \leq \alpha. \tag{1.4}$$

Remark 1.4. The complete hypothesis H_1 is in fact the following statement: the mode genome in the test tube is $\text{mode}(X) = MG$, and the probability $p(X = MG) \geq \theta$. By our choice in Definition 1.2 we have $\theta > 50\%$. Thus $p(X = MG) \geq \theta > 50\%$ implies $\text{mode}(X) = MG$. Hence we drop that part of the statement in the hypothesis since it becomes redundant.

In the sequencing contest, each contestant's report $R_c = (N_c, MG_c, \theta_c, \alpha_c)$ is compared to the benchmark report $R_b = (N_b, MG_b, \theta_b, \alpha_b)$, which typically has higher quality. We now study when the statistic of the benchmark report R_b supports the hypothesis H_1 of the contestant report R_c. To do so, we use the statistic of report R_b as a test statistic as follows. Define a random variable Y which counts the number of occurrences of MG_b in a set of N_b samples

$$Y = \#\{i \in \{1, \dots, N_b\} : X_i = MG_b\}. \tag{1.5}$$

Here each X_i is an i.i.d random variable with the same distribution as X in (1.1), defined by measuring the cells in the test tube. Report R_b indicates $Y = m$ for some integer $m \geq \theta_b N_b$, because θ_b is the *lower bound* of the mode genome frequency in Definition 1.2. According to (1.3) and (1.4), in order for the statistic of report R_b to support the hypothesis H_1 of report R_c, we need the following inequality to hold

$$p\big(Y = m \,|\, p(X = MG_c) < \theta_c\big) \leq \alpha_c, \quad \text{for all integer } m \text{ with } \theta_b N_b \leq m \leq N_b. \tag{1.6}$$

The following result gives a condition which guarantees the benchmark report to support the contestant report.

Theorem 1.5. *Let $R_b = (N_b, MG_b, \theta_b, \alpha_b)$ and $R_c = (N_c, MG_c, \theta_c, \alpha_c)$ be two reports. Then the statistic of report R_b supports the hypothesis H_1 of report R_c if the following three conditions are satisfied.*

(1) $MG_b = MG_c$,
(2) $\theta_b \geq \theta_c$, *and*
(3) $B(\lceil \theta_b N_b \rceil \,|\, N_b, \theta_c) \leq \alpha_c$,

where $\lceil \cdot \rceil$ is the ceiling function, and $B(m \,|\, N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$ is the binomial distribution.

Proof. Since the mode genome of two reports agree $MG_b = MG_c$, we shall denote it as MG for short. Given the condition in the statement, we now prove inequality (1.6). Denote $\mu_c := p(X = MG)$. The condition in the conditional probability in (1.6) gives

$$\mu_c < \theta_c. \tag{1.7}$$

The random variable $Y = \#\{i \in \{1, \ldots, N_b\} : X_i = MG\}$ in (1.5) obeys a binomial distribution

$$p(Y = m) = B(m \,|\, N_b, \mu_c) = \binom{N_b}{m} \mu_c^m (1 - \mu_c)^{N_b - m}.$$

To prove that the statistic of report R_b supports the hypothesis H_1 of report R_c, it suffices to show that for all integer m with $\theta_b N_b \leq m \leq N_b$, we have

$$p(Y = m) = B(m \,|\, N_b, \mu_c) \leq \alpha_c \tag{1.8}$$

as in (1.6). We shall prove it by showing the following inequalities

$$B(m \,|\, N_b, \mu_c) \leq B(m \,|\, N_b, \theta_c) \leq B(\lceil \theta_b N_b \rceil \,|\, N_b, \theta_c) \leq \alpha_c. \tag{1.9}$$

The first inequality in (1.9) holds because by (1.7) and the assumption $\theta_b \geq \theta_c$ we have $\mu_c < \theta_c \leq \theta_b \leq \frac{m}{N_b}$. Then Lemma 1.6 (1) implies $B(m \,|\, N_b, \mu_c) \leq B(m \,|\, N_b, \theta_c)$.

We show the second inequality in (1.9) as follows. Since m is an integer with $m \geq \theta_b N_b$, we have $m \geq \lceil \theta_b N_b \rceil$. Moreover, the assumption $\theta_b \geq \theta_c$

implies $\theta_b N_b + 1 \geq \theta_c N_b + \theta_c$. Thus we have $\theta_b N_b \geq \theta_c(N_b+1) - 1$ and $\lceil \theta_b N_b \rceil \geq \lceil \theta_c(N_b+1) \rceil - 1$. Therefore the inequalities

$$m \geq \lceil \theta_b N_b \rceil \geq \lceil \theta_c(N_b+1) \rceil - 1 \tag{1.10}$$

combined with Lemma 1.6 (2) imply $B(m \,|\, N_b, \theta_c) \leq B(\lceil \theta_b N_b \rceil \,|\, N_b, \theta_c)$.

The last inequality in (1.9) is precisely condition (2) in the assumption. This proves (1.8) and finishes the proof. □

Here we state the relevant properties of the binomial distribution.

Lemma 1.6. *The binomial distribution*

$$B(m \,|\, N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

with N a natural number, $\mu \in [0,1]$, and $m \in \{0, \ldots, N\}$ has the following properties.

1. *$B(m \,|\, N, \cdot)$ viewed as a function of μ is non-decreasing for $\mu \in [0, \frac{m}{N}]$.*
2. *$B(\cdot \,|\, N, \mu)$ viewed as a function of m is non-increasing for $m \in \{\lceil \mu(N+1) \rceil - 1, \ldots, N\}$, where $\lceil \cdot \rceil$ is the ceiling function.*

Proof. We first prove (1). Applying logarithm to $B(m \,|\, N, \mu)$ and we have

$$\ln B(m \,|\, N, \mu) = \ln \binom{N}{m} + m \ln(\mu) + (N-m) \ln(1-\mu).$$

Its derivative with respect to μ is given by

$$\partial_\mu \ln B(m \,|\, N, \mu) = m \frac{1}{\mu} + (N-m) \left(-\frac{1}{1-\mu} \right) = \frac{m - \mu N}{\mu(1-\mu)}.$$

Hence $\partial_\mu \ln B(m \,|\, N, \mu) \geq 0$ if $\mu \leq \frac{m}{N}$. This proves (1).

We now prove (2). First assume $\mu \neq 1$. It suffices to show that for $m \geq \lceil \mu(N+1) \rceil - 1$, we have $\frac{B(m+1 \,|\, N, \mu)}{B(m \,|\, N, \mu)} \leq 1$. By a simple calculation, we have

$$\frac{B(m+1 \,|\, N, \mu)}{B(m \,|\, N, \mu)} = \frac{(N-m)\mu}{(m+1)(1-\mu)}.$$

And $\frac{(N-m)\mu}{(m+1)(1-\mu)} \leq 1$ is equivalent to $m \geq \mu(N+1) - 1$, which is the same as $m \geq \lceil \mu(N+1) \rceil - 1$ since m is an integer. When $\mu = 1$, then $\lceil \mu(N+1) \rceil - 1 = N$ and (2) is vacuously true. This proves (2). □

References

[1] Metzker ML. Sequencing technologies - the next generation. Nature Rev Genet. 2010;11:31-46. doi:10.1038/nrg2626