Neptune Math

Eric Marinier

November 16, 2015

1 Arbitrary Matching

Let $P(X = Y)_A$ be the probability that any two arbitrary nucleotide bases, X and Y, match given a known environmental GC-content. Let λ be the GC-content such that $0 \leq \lambda \leq 1$; subscript A denote arbitrary; and A, C, G, T denote the four DNA bases.

$$\begin{aligned} P(X = Y)_A &= P(X = Y = A)_A + P(X = Y = T)_A \\ &\quad + P(X = Y = C)_A + P(X = Y = G)_A \\ P(X = Y)_A &= \left(\frac{1-\lambda}{2}\right)^2 + \left(\frac{1-\lambda}{2}\right)^2 + \left(\frac{\lambda}{2}\right)^2 + \left(\frac{\lambda}{2}\right)^2 \\ P(X = Y)_A &= 2\left(\frac{1-\lambda}{2}\right)^2 + 2\left(\frac{\lambda}{2}\right)^2 \end{aligned} \tag{1}$$

When GC-content is 0.50:

$$P(X = Y)_A = 0.25 \tag{2}$$

When GC-content is 0.25:

$$P(X = Y)_A = 0.3125 \tag{3}$$

The probability that any two **arbitrary** k-mers, k_X and k_Y, match exactly:

$$P(k_X = k_Y)_A = \left(2\left(\frac{1-\lambda}{2}\right)^2 + 2\left(\frac{\lambda}{2}\right)^2\right)^k \tag{4}$$

2 Homologous Matching

Let $P(X_M = Y_M)_H$ be the probability that two homologous bases, X and Y, mutate to the same base. Let subscript M denote a mutation; subscript H denote homology; and A, C, G, T denote the four DNA bases.

$$\begin{aligned} P(X_M = Y_M)_H =& P(X_M = Y_M | X = Y = A) \cdot P(A) \\ &+ P(X_M = Y_M | X = Y = C) \cdot P(C) \\ &+ P(X_M = Y_M | X = Y = G) \cdot P(G) \\ &+ P(X_M = Y_M | X = Y = T) \cdot P(T) \end{aligned} \tag{5}$$

Let $P(C|A)$ be shorthand for $P(X_M = C|X = A)$. That is, the probability of a base mutating to a C nucleotide given that it was an A nucleotide before the mutation event.

$$\begin{aligned} P(X_M = Y_M)_H =& (P(C|A)^2 + P(G|A)^2 + P(T|A)^2) \cdot P(A) \\ &+ (P(A|C)^2 + P(G|C)^2 + P(T|C)^2) \cdot P(C) \\ &+ (P(A|G)^2 + P(C|G)^2 + P(T|G)^2) \cdot P(G) \\ &+ (P(A|T)^2 + P(C|T)^2 + P(G|T)^2) \cdot P(T) \end{aligned} \tag{6}$$

Note that because of GC-content symmetry:

$$\begin{aligned} P(A) &= P(T) \\ P(C) &= P(G) \\ P(C|G) &= P(G|C) \\ P(A|T) &= P(T|A) \\ P(A|C) = P(T|C) &= P(A|G) = P(T|G) \\ P(C|A) = P(C|T) &= P(G|A) = P(G|T) \end{aligned} \tag{7}$$

Making substitutions:

$$\begin{aligned} P(X_M = Y_M)_H =& (2P(C|A)^2 + P(T|A)^2) \cdot P(A) \\ &+ (2P(A|C)^2 + P(G|C)^2) \cdot P(C) \\ &+ (2P(A|C)^2 + P(G|C)^2) \cdot P(G) \\ &+ (2P(C|A)^2 + P(T|A)^2) \cdot P(T) \end{aligned} \tag{8}$$

Simplifying:

$$\begin{aligned} P(X_M = Y_M)_H =& 2(2P(C|A)^2 + P(T|A)^2)P(A) \\ &+ 2(2P(A|C)^2 + P(G|C)^2)P(C) \end{aligned} \tag{9}$$

Let λ be the GC-content such that $0 \leq \lambda \leq 1$.

$$\begin{aligned} P(A) = P(T) &= \frac{(1-\lambda)}{2} \\ P(C) = P(G) &= \frac{\lambda}{2} \end{aligned} \tag{10}$$

We assume that the probability of all mutation events are independent and do not account for transitions being more likely than transversions. The mutation probabilities are determined entirely by the GC-content. We define $P(C|A)$, $P(A|C)$, $P(T|A)$, and $P(G|C)$ as follows:

$$\begin{aligned} P(C|A) &= \frac{P(C)}{P(C) + P(G) + P(T)} \\ &= \frac{\frac{\lambda}{2}}{\frac{\lambda}{2} + \frac{\lambda}{2} + \frac{(1-\lambda)}{2}} \\ &= \frac{\lambda}{\lambda + 1} \end{aligned} \tag{11}$$

$$\begin{aligned} P(A|C) &= \frac{P(A)}{P(A) + P(G) + P(T)} \\ &= \frac{\frac{(1-\lambda)}{2}}{\frac{(1-\lambda)}{2} + \frac{\lambda}{2} + \frac{(1-\lambda)}{2}} \\ &= \frac{1-\lambda}{2-\lambda} \end{aligned} \tag{12}$$

$$\begin{aligned} P(T|A) &= \frac{P(T)}{P(C) + P(G) + P(T)} \\ &= \frac{\frac{(1-\lambda)}{2}}{\frac{\lambda}{2} + \frac{\lambda}{2} + \frac{(1-\lambda)}{2}} \\ &= \frac{1-\lambda}{\lambda + 1} \end{aligned} \tag{13}$$

$$
\begin{aligned}
P(G|C) &= \frac{P(G)}{P(A) + P(G) + P(T)} \\
&= \frac{\frac{\lambda}{2}}{\frac{(1-\lambda)}{2} + \frac{\lambda}{2} + \frac{(1-\lambda)}{2}} \\
&= \frac{\lambda}{2-\lambda}
\end{aligned}
\tag{14}
$$

Subbing these equations into Equation 9:

$$
\begin{aligned}
P(X_M = Y_M)_H = &\left(2\left(\frac{\lambda}{\lambda+1}\right)^2 + \left(\frac{1-\lambda}{\lambda+1}\right)^2\right)(1-\lambda) \\
&+ \left(2\left(\frac{1-\lambda}{2-\lambda}\right)^2 + \left(\frac{\lambda}{2-\lambda}\right)^2\right)(\lambda)
\end{aligned}
\tag{15}
$$

When GC-content is 0.50:

$$
P(X_M = Y_M)_H = 0.3333 \tag{16}
$$

When GC-content is 0.25:

$$
P(X_M = Y_M)_H = 0.4269 \tag{17}
$$

We can now determine the probability of two homologous bases, X and Y, matching. These bases match when neither mutates or both mutate to the same base. Let ε be the probability that two homologous bases do not match exactly.

$$
P(X = Y)_H = (1-\varepsilon)^2 + (\varepsilon)^2 \cdot P(X_M = Y_M)_H \tag{18}
$$

The probability that any two homologous k-mers, k_X and k_Y, match exactly:

$$
P(k_X = k_Y)_H = (Pr(X = Y)_H)^k \tag{19}
$$

Genome Size	GC-Content	k	Expected Mismatches
1,000,000	0.25	27	0.01
1,000,000	0.50	23	0.01
5,000,000	0.25	29	0.03
5,000,000	0.50	25	0.01
5,000,000,000	0.25	41	0.02
5,000,000,000	0.50	35	0.01

Table 1: A summary of recommended k-mer values for various targets. We recommend using k-mers of odd size, as this avoids k-mers being their own reverse complement. As GC-content is symmetric, a target with a GC-content of 0.25 or 0.75 will require the same size of k.

3 Size of k

We select a k such that the expected number of arbitrary k-mer matches within a single is sufficiently small. Let ω be the length of the genome; λ be the GC-content; and x and y be positions in the genome, such that $x \neq y$. We approximate the expected number of matches by assuming the probability of all k-mer matches is independent. However, these k-mers are produced using a sliding window and are therefore not independent. We recommend using a large enough k such that:

$$\sum_{x<y} P(k_X = k_Y) \approx \binom{\omega - k + 1}{2} \cdot P(k_X = k_Y)_A < 0.05 \tag{20}$$

$$\frac{(\omega - k + 1)(\omega - k)}{2} \cdot \left(2\left(\frac{1-\lambda}{2}\right)^2 + 2\left(\frac{\lambda}{2}\right)^2\right)^k < 0.05 \tag{21}$$

A summary of some recommended k-mer sizes for various targets can be found in Table 1.

4 Minimum Inclusion Hits

We model the process of homologous k-mer matches with a binomial distribution. If we are observing a true signature region, we expect that corresponding homologous k-mers should exist in all inclusion targets. These k-mers will match with a probability of $p = P(k_X = k_Y)_H$ and not match with a probability of

$q = 1 - p$. Let n be the number of inclusion targets. The parameters of this binomial distribution are described below:

$$\begin{aligned} p &= P(k_X = k_Y)_H \\ q &= 1 - p \\ \mu &= n \cdot p \\ \sigma^2 &= n \cdot p \cdot q \end{aligned} \tag{22}$$

A normal distribution can approximate a binomial distribution for sufficiently large n and p. Therefore, we can set the minimum number of inclusion hits to capture a large fraction of all observations in a normal distribution. Let α be our statistical confidence and $\Phi^{-1}(\alpha)$ be the probit function. The minimum number of inclusion targets containing a k-mer, $\wedge_{in}$, required for a reference k-mer to be considered an inclusion k-mer is defined as follows:

$$\wedge_{in} = 1 + \mu - \Phi^{-1}(\alpha)\sigma \tag{23}$$

5 Maximum Gap Size

We model the problem of maximum gap size between exact matching inclusion k-mers as recurrence times of success runs in Bernoulli trials. Let p be the probability of a Bernoulli trial success; and q be $q = 1 - p$, or the probability of a Bernoulli trial failure.

$$\begin{aligned} p &= P(X = Y)_H \\ q &= 1 - p \end{aligned} \tag{24}$$

The mean and variance of the recurrence times of k successes, or an exact k-mer match, in Bernoulli trials is described by Feller 1960 [1]:

$$\mu = \frac{1 - p^k}{q \cdot p^k} \tag{25}$$

$$\sigma^2 = \frac{1}{(q \cdot p^k)^2} - \frac{2k + 1}{q \cdot p^k} - \frac{p}{q^2} \tag{26}$$

This distribution captures how many bases we must observe before we can expect to see another homologous k-mer match. The distribution is not normal for a small number of observations. However, we can use Chebyshev's Inequality to make lower-bound claims about the distribution:

$$P(|X - \mu| \geq \delta\sigma) \leq \frac{1}{\delta^2} \tag{27}$$

Where δ is the number of standard deviations, σ, from the mean, μ. Let $P(|X - \mu| \geq \delta\sigma)$ be our statistical confidence, α. The maximum allowable k-mer gap size, $\vee_{gap}$, is calculated as follows:

$$\vee_{gap} = \mu + \sqrt{\frac{1}{1-\alpha}} \cdot \sigma \tag{28}$$

References

[1] Vilim Feller. *An Introduction to Probability Theory and Its Applications: Volume 1.* J. Wiley & sons, 1960.