CSE182 Lecture 10 questions

Vineet Bafna

October 31, 2006

The questions are open ended, but should help you understand lectures better. Do these questions make sense? Are they helpful in following the lecture? Constructive feedback is appreciated.

Lander Waterman statistics

1. What are the various parameters for computing LW statistics?

   Answer:

G	Length of genome
L	length of fragment (clone)
N	Number of clones
c	coverage $c = \frac{LN}{G}$
α	$\alpha = \frac{N}{G}$
T	Overlap threshold
θ	$\theta = \frac{T}{L}$
σ	$\sigma = 1 - \theta$

2. What is the probability that an island begins at some arbitrary position i?

   A: As there are N clones falling on G positions, it is $\alpha = \frac{N}{G}$.

3. What is the expected number of islands?

   A: Recall that an island is a collection of overlapping fragments, such that each adjacent pair overlaps by at least T. Let X_i be the indicator variable signaling the end of an island. $X_i = 1$ if an island ends at position i, and $X_i = 0$ otherwise. The number of islands is given by $I = \sum_i X_i$. By linearity of expectation,

   $$E(I) = E(\sum_i X_i) = \sum_i E(X_i) = \sum_i Pr(X_i = 1)$$

   if an island is to end at position i, a fragment must have begun at position $i - L + 1$, and no fragment could have started in the approximately $L - T$ positions $[i - L + 2, i - (L - T)]$. Therefore,

   $$\begin{aligned} Pr(X_i = 1) &= \alpha (1 - \alpha)^{L-T} \\ &= \alpha (1 - \alpha)^{L\sigma} \\ &\simeq \alpha e^{-\alpha L \sigma} = \alpha e^{-c\sigma} \end{aligned}$$

   Therefore

   $$E(I) = G\alpha e^{-c\sigma} = Ne^{-c\sigma}$$

   This equation has many uses. For example, by empirically estimating the number of islands, one can get the length of the unknown genome. Also, it is illustrative to try some numbers and see qhat happens. For the human genome, suppose we sequenced fragments of length 1000 to $8X$ coverage ($c = 8$). This would require $N = \frac{cG}{L} = 24 * 10^6$ fragments. The expected number of islands (with $T = 100$) is $Ne^{-c\sigma} \simeq 18K$, which is too large a number to get a successful sequence assembly.

4. What is the expected number of clones in an island?

   A: An island that has already begun ends with a probability $P = (1-\alpha)^{L-T} \simeq e^{-c\sigma}$. The probability that there are exactly j fragments on an island is $(1-P)^{j-1}P$. The expected number of fragments on an island is then

$$\sum_j j(1-e^{-c\sigma})^{j-1}e^{-c\sigma}$$

   Consider

$$A = \sum_{j=0}^{\infty}(1-P)^j = 1 + (1-P) + (1-P)^2 + \ldots$$

   Note that

$$(1-P)A = (1-P) + (1-P)^2 + \ldots = A - 1$$

   implying

$$A = \sum_{j=0}^{\infty}(1-P)^j = \frac{1}{P}$$

   Differentiating w.r.t P

$$\sum_{j=1}^{\infty} j(1-P)^{j-1} = \frac{1}{P^2}$$

   Multiplying by P, and substituting $P = e^{-c\sigma}$,

$$\sum_j j(1-e^{-c\sigma})^{j-1}e^{-c\sigma} = e^{c\sigma}$$

   This equation is very useful because a higher than expected number of clones in an island is indicative of repetitive sequence.

Whole genome shotgun assembly

1. q