CBMF 4761 Homework 3

Alex Ying

I discussed the problems with Jason Mohabir and Harry Lee.

Problem 1

a. For a genome sequence, G, of length n, there are $n-k+1$ k-mers. The expected number of k-mers that appear 3 times can be approximated by the product of the number of possible triplets multiplied by the probability that all 3 k-mers in that triplet are identical. In this case, two assumptions will be made, that the selected region R is of length $r \geq k$ (this is a slight modification to the question, made because repeat sequences of length k will still cause ambiguity), and that $n >> r$ (so that selected regions are essentially independent and overlaps can be ignored). For a region of length r, the probability that the selected regions are identical can be expressed as:

$$\binom{n-r+1}{3} 4^{-2r}$$

Because of the assumption that $n >> r$, the probability can be approximated further as:

$$\frac{n^3 4^{-2r}}{6}$$

For the total probability that a sequence R repeats, the individual probabilities of the lengths are summed to give:

$$\sum_{r=k}^{n/3} \frac{n^3 4^{-2r}}{6}$$

However, note that the probability that a sequence of length $k+1$ is repeated 3 times can be described with:

$$\frac{n^3 4^{-2(k+1)}}{6} = \frac{n^3 4^{-2k-2}}{6} = \frac{4^{-2} n^3 4^{-2k}}{6}$$

As can be seen, the contribution of each increase in length decreases significantly by a factor of 16 for each increase by 1 base. For this reason, only a few terms of the summation are actually necessary for a given level of precision. For 4 significant figures, 4 terms is probably sufficent, since $4^{-2*3} = 0.000244$. This gives a final expression for the approximate expected number of ambiguous repeat presentations of:

$$\frac{n^3 4^{-2k}}{6} \sum_{i=0}^{3} 4^{-2i} = 1.0667 * \frac{n^3 4^{-2k}}{6}$$

b. Given a desired probability for ambiguity of 0.95, the length of a sequence can be found from:

$$1.0667 * \frac{n^3 4^{-2k}}{6} = 0.95$$

$$n^3 4^{-2k} = 5.344$$

$$n = (5.344 * 4^{2k})^{1/3}$$

c. Two pairwise repetitions may cause ambiguity because it becomes unclear whether the sequence is G=ARBPCRDPE or G=ARDPCRBPE (there is also the issue of ambiguity with regards to how many times the RBPC or RDPC sequence is repeated, but this is also an issue with single pairwise repetitions). The diagram below illustrates this ambiguity.

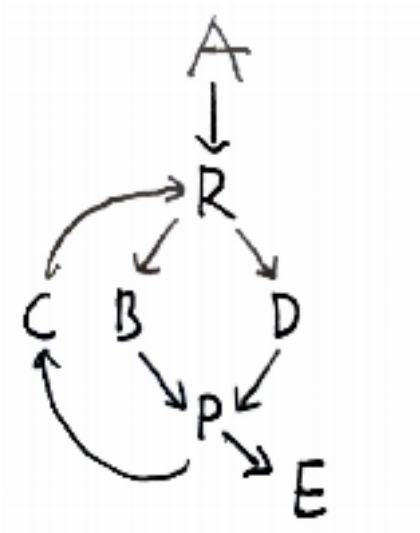

Figure 1: Graph showing possible paths between subsequences

d. Similar to the derivation in (a), the expected number of 2 pairs of repeated sequences can be expressed as the number of combinations of 4 sequences of length r times the probability that these 4 sequences form 2 pairs (each has the probability of 4^{-r}. The same assumptions made in (a) are again made (that $r \geq k$, and that $n >> r$). Furthermore, assume that the probability that all 4 sequences match does not need to be subtracted out, because it still leads to ambiguity. Finally, note that there exists 4 possible orders for the pairs, $RPRP$, $RRPP$, $PPRR$, and $PRPR$, but only the 2 orders where the pairs alternate lead to ambiguity, so the overall probability is multiplied by 1/2 to account for this.

$$\frac{1}{2}\binom{n-r+1}{4}4^{-r}4^{-r} = \frac{1}{2}\binom{n-r+1}{4}4^{-2r}$$

Following the same approximations from before, the expected number of 2 pairs of repeated sequences for a single value of r can be expressed as:

$$\frac{n^4 4^{-2r}}{48}$$

Again, note that the probability of 2 pairs of repeated sequences of length $k+1$ is the probability of 2 pairs of repeated sequences of length k decreased by a factor of 16. For 4 significant figures, the same number of terms is used, and the final expression for the approximate expected number of 2 pairs of repeated sequences is found to be:

$$\frac{n^4 4^{-2k}}{48}\sum_{i=0}^{3} 4^{-2i} = 1.0667 * \frac{n^4 4^{-2k}}{48}$$

Expression n as a function of k for a 0.95 chance of ambiguity gives:

$$1.0667 * \frac{n^4 4^{-2k}}{48} = 0.95$$

$$n^4 4^{-2k} = 42.75$$

$$n = (42.75 * 4^{2k})^{1/4}$$

Problem 2

a. The observed fractions are determined by two factors - which transcript a particular read is drawn from, and the probability that the fraction is drawn from that transcript. This gives the expectation for the fraction of the reads to map to a given region as:

$$f = \sum_T P(f|T)P(T)$$

where f is the expected fraction, T represents the possible transcripts from which the read can be drawn, $P(f|T)$ is the probability of drawing a given read from the transcript, and $P(T)$ is the probability that any read is drawn from a particular transcript. $P(T)$ is simply the true expression level of the transcript, μ_{ij}, under the assumption that reads are taken uniformly across all transcripts. $P(f|T)$ can be calculated for each particular fraction knowing the lengths of the region and the length of the transcript. For f_{20}, the two possible transcripts are t_{00} and t_{01}, which have a lengths 1900 and 2000, respectively. Since the number of possible reads from a transcript of length l is $l - 101 + 1 = l - 100$, the possible number of transcripts mapping to exon 2, alternative 0 is 100 out of 1800 and 1900. This gives:

$$f_{20} = \frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{01}$$

Repeating this for other fractions gives:

$$f_{21} = \frac{2}{19}\mu_{10} + \frac{1}{10}\mu_{11}$$

$$f_{30} = \frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{10}$$

$$f_{31} = \frac{2}{19}\mu_{01} + \frac{1}{10}\mu_{11}$$

The number of reads mapping entirely within exon 1 is always given by $1700 - 101 + 1 = 1600$, giving an expected value of f_1 as:

$$f_1 = \frac{8}{9}\mu_{00} + \frac{16}{19}\mu_{01} + \frac{16}{19}\mu_{10} + \frac{4}{5}\mu_{11}$$

b. The fraction of reads mapping to a particular region can be described by a binomial distribution, as each read is essentially a single binary outcome, and the probability that it maps to the region given by the expectation expression from (a). However, because each transcript is known to be common enough for many reads to be drawn from each of them, the Law of Large Numbers holds, and the fraction of reads mapping to a particular region can be expressed by a normal distribution by the Central Limit Theorem. The mean of this normal distribution is given by the expected values of the fractions in (a), and the standard deviation is the expected fraction divided by the number of reads, R. That is to say:

$$f_1 \sim Normal(\frac{8}{9}\mu_{00} + \frac{16}{19}\mu_{01} + \frac{16}{19}\mu_{10} + \frac{4}{5}\mu_{11}, (\frac{8}{9}\mu_{00} + \frac{16}{19}\mu_{01} + \frac{16}{19}\mu_{10} + \frac{4}{5}\mu_{11})/R)$$
$$f_{20} \sim Normal(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{01}, (\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{01})/R)$$
$$f_{21} \sim Normal(\frac{2}{19}\mu_{10} + \frac{1}{10}\mu_{11}, (\frac{2}{19}\mu_{10} + \frac{1}{10}\mu_{11})/R)$$
$$f_{30} \sim Normal(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{10}, (\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{10})/R)$$
$$f_{31} \sim Normal(\frac{2}{19}\mu_{01} + \frac{1}{10}\mu_{11}, (\frac{2}{19}\mu_{01} + \frac{1}{10}\mu_{11})/R)$$

To simplify for following solutions, these abbreviations will be used:

$$f_1 \sim Normal(\mu_1, \mu_1/R)$$
$$f_{20} \sim Normal(\mu_2, \mu_2/R)$$
$$f_{21} \sim Normal(\mu_3, \mu_3/R)$$
$$f_{30} \sim Normal(\mu_4, \mu_4/R)$$
$$f_{31} \sim Normal(\mu_5, \mu_5/R)$$

c. The probability of observing a particular value of f_i is given by the probability density function of the normal distribution, $\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{f-\mu_i}{\sigma_i}\right)^2}$, where μ_i and $\sigma_i = \mu_i/R$ are the associated expected value and variance of the fraction, respectively. This gives a overall likelihood for the set of 5 fractions observed, F, as:

$$P(F|f_i \sim Normal(\mu_i, \sigma_i^2), \forall f \in F)) = \prod_F P(f_i|f_i \sim Normal(\mu_i, \sigma_i^2))$$

The overall log-likelihood is therefore:

$$ln(P(F|f_i \sim Normal(\mu_i, \sigma_i^2), \forall f \in F)) = \sum_F ln(P(f_i|f_i \sim Normal(\mu_i, \sigma_i^2)))$$
$$= \sum_F ln\left(\frac{1}{\sigma_i\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{f_i-\mu_i}{\sigma_i}\right)^2}\right)$$
$$= \sum_F -ln(\sigma_i) - \frac{1}{2}ln(2\pi) - \frac{1}{2}\left(\frac{f_i-\mu_i}{\sigma_i}\right)^2$$

d. To determine the maximum likelihood estimator, we can derive the log-likelihood expression found in (c) with respect to μ_1, μ_2, μ_3, μ_4, and μ_5. Taking the derivative of the overall log-likelihood with respect to an arbitrary mean, $\mu_n, n \in [1, 5]$, and setting it equal to 0, a system of equations can be found.

$$\frac{d}{d\mu_n}\left(\sum_F -ln(\sigma_i) - \frac{1}{2}ln(2\pi) - \frac{1}{2}\left(\frac{f_i-\mu_i}{\sigma_i}\right)^2\right)$$

$$= \frac{d}{d\mu_n}\left(-ln\left(\sqrt{\mu_n/R}\right) - \frac{1}{2}ln(2\pi) - \frac{1}{2}\left(\frac{f_n-\mu_n}{\sqrt{\mu_n/R}}\right)^2\right)$$

$$= -\frac{1}{2\mu_n} + \frac{R}{2}\left(\frac{f_n^2}{\mu_n^2} - 1\right) = 0$$

$$= -\mu_n + R\left(f_n^2 - \mu_n^2\right) = 0$$

$$f_n = \sqrt{\frac{1}{R}\mu_n + \mu_n^2}$$

This gives us the system of equations:

$$f_1 = \sqrt{\frac{1}{R}\left(\frac{8}{9}\mu_{00} + \frac{16}{19}\mu_{01} + \frac{16}{19}\mu_{10} + \frac{4}{5}\mu_{11}\right) + \left(\frac{8}{9}\mu_{00} + \frac{16}{19}\mu_{01} + \frac{16}{19}\mu_{10} + \frac{4}{5}\mu_{11}\right)^2}$$

$$f_{20} = \sqrt{\frac{1}{R}\left(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{01}\right) + \left(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{01}\right)^2}$$

$$f_{21} = \sqrt{\frac{1}{R}\left(\frac{2}{19}\mu_{10} + \frac{1}{10}\mu_{11}\right) + \left(\frac{2}{19}\mu_{10} + \frac{1}{10}\mu_{11}\right)^2}$$

$$f_{30} = \sqrt{\frac{1}{R}\left(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{10}\right) + \left(\frac{1}{18}\mu_{00} + \frac{1}{19}\mu_{10}\right)^2}$$

$$f_{31} = \sqrt{\frac{1}{R}\left(\frac{2}{19}\mu_{01} + \frac{1}{10}\mu_{11}\right) + \left(\frac{2}{19}\mu_{01} + \frac{1}{10}\mu_{11}\right)^2}$$

Solving these equations gives the maximum likelihood estimator for each μ_{ij} that would give the observed fractions.