

greater number of large fragments in the case of bacteriophage lambda DNA. To determine whether the distribution for lambda DNA differs significantly from the mathematical model (exponential distribution), we could break up the length axis into a series of “bins” and calculate the expected number of fragments in each bin by using the exponential density. This would create the entries for a histogram based on the mathematical model. We could then compare the observed distribution of fragments from lambda DNA (using the same bin boundaries) to the expected distribution from the model by using the χ^2 test, for example.

3.6 k -word Occurrences

The statistical principles learned in this and the previous chapter can be applied to other practical problems, such as discovering functional sites in DNA. We use promoter sequences as an example. Recall from Section 1.3.4 that promoters are gene regions where RNA polymerase binds to initiate transcription. We wish to find k -words that distinguish promoter sequences from average genomic sequences. Because promoters are related by function, we expect to observe k -words that are over-represented within the promoter set compared with a suitable null set. These k -words can help identify DNA “signals” required for promoter function. (DNA signals are described in detail in Chapter 9.) Using the approaches of Chapter 2, we determine expected k -word frequencies and compare them to the observed frequencies. Distributions presented in this chapter are used to test whether over-represented k -words appear with significantly higher frequencies.

Consider N promoter sequences of length L bp, which we denote by S_1, \dots, S_N (Table C.2). The null set might consist of N strings of L iid letters, each letter having the same probability of occurrence as the letter frequencies in genomic DNA as a whole. For the purposes of the discussion here, we take a small word size, $k = 4$, so that there are 256 possible k -words. With no a priori knowledge of conserved patterns, we must examine all 256 words. We ask whether there are an unusual number of occurrences of each word in the promoter regions.

For the 49 promoter sequences shown in Table C.2 in Appendix C, we first evaluate the most abundant observed k -words and their expected values for $k = 4$ using R for the computation described in Computational Example 3.6. The expectation of each word according to the null (iid) model is easy to calculate if words are overlapping. For example, if X_w denotes the number of occurrences of word w in the whole set of sequences, then for $w = \text{ACGT}$,

$$\begin{aligned}\mathbb{P}(w = \text{ACGT}) &= p_A p_C p_G p_T \\ \mathbb{E}(\# \text{ times } w \text{ appears in } S_i) &= (L - 4 + 1)p_A p_C p_G p_T\end{aligned}$$

and the expected number of occurrences in N such sequences is