1.

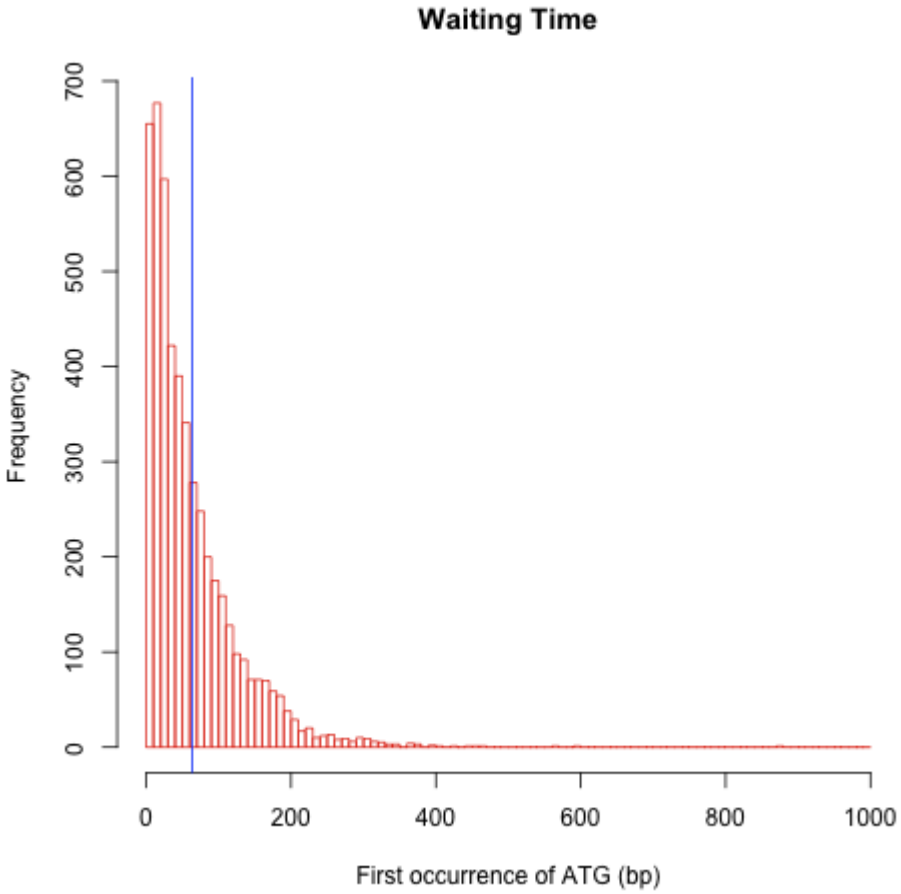
Let μ be the waiting time until observing ATG in the sequence and $\mu_{S'}$ be the waiting time until observing ATG from the pattern S'. We can use first step analysis and the given base probabilities to establish the following system of equations.

$$\begin{aligned}
\mu &= 1 + P_A\mu_A + (1 - P_A)\mu_{\bar{A}} \\
\mu_{\bar{A}} &= 1 + P_A\mu_A + (1 - P_A)\mu_{\bar{A}} \\
\mu_A &= 1 + P_A\mu_A + (P_C + P_G)\mu_{\bar{A}} + P_T\mu_{AT} \\
\mu_{AT} &= 1 + P_A\mu_A + (P_C + P_T)\mu_{\bar{A}} + P_G\mu_{ATG} \\
\mu_{ATG} &= 0
\end{aligned}$$

Solving this system, we get

$$\mu = \frac{1}{P_A P_T P_G} = \frac{1}{.25^3} = 64$$

To verify this, I simulated 5000 random sequences of length 1000 such that the expected base composition for each sequence was $P_N = .25$ for all $N \in \{A, C, G, T\}$. I then determined the position of the first occurrence of ATG in each sequence. The observed mean was 63.41, the observed variance was 3914.96, and the distribution of waiting times (with mean highlighted in blue) is shown below.



2.

First, let us consider the sequence $YRYR$. Let μ_n be the probability that we observe $YRYR$ ending at position n, μ be the expected waiting time, and p and q be the probabilities of observing Y and R, respectively. We can model the expected waiting time as follows.

$$\begin{aligned} p^2q^2 &= \mu_n + \mu_{n-2}pq \\ &= \frac{1}{\mu} + \frac{1}{mu}pq \\ &= \frac{1}{\mu}(1 + pq) \\ &= \frac{1 + pq}{p^2q^2} \end{aligned} \tag{1}$$

If $p = q = 0.5$, then the expected waiting time is 20.

Now consider the sequence $RYRR$. Again, let μ_n be the probability that we observe $RYRR$ ending at position n, μ be the expected waiting time, and p and q be the probabilities of observing Y and R, respectively. We can model the expected waiting time as follows.

$$\begin{aligned} p^2q^2 &= \mu_n + \mu_{n-3}p^2q \\ &= \frac{1}{\mu} + \frac{1}{mu}p^2q \\ &= \frac{1}{\mu}(1 + p^2q) \\ &= \frac{1 + p^2q}{p^3q} \end{aligned} \tag{2}$$

If $p = q = 0.5$, then the expected waiting time is 18.

To verify these values empirically, I converted my 5000 random sequences from the alphabet $\{A, C, G, T\}$ to $\{R, Y\}$ so that the expected composition of each sequence was $P_R = 0.5, P_Y = 0.5$. I then determined the position of the first occurrence of $YRYR$ and $RYRR$ in each sequence. The observed means were 20.25 and 17.80 (respectively), the observed variances were 281.64 and 206.70 (respectively), and the distributions of waiting times (with means highlighted in blue) are shown below.