

# Hamming distance, k-mers, and self-information

Jacob Fine

February 13th, 2025

Consider all possible RNA sequences of length  $k$ , from the alphabet  $\{A,T,C,G\}$ , of which there are  $4^k$ . If each sequence has an equal chance of being sampled, then the probability of sampling any given sequence is  $1/4^k$ . But what is the probability of sampling a sequence that differs by exactly one base? Or exactly two bases? The number of positions that a pair of RNA sequences  $x, y$  of equal length differ by is the Hamming distance,  $H(x, y)$ . So let us consider the probability that a sequence of length  $K = k$  has a Hamming distance of  $H = h$ . We are interested in the PMF and CDF of  $H$  for some  $K = k$ .

$$P(H = h|K = k)$$

which gives the probability that a sequence of length  $k$  differs by  $h$  characters from our target sequence. Consider the target sequence

$$x = \text{ATGCCCCGGG}$$

At each of the 9 bases, there are 3 other possible bases to choose from. Here,  $k = 9$ . So there would be  $3 + 3 + \dots + 3$  for a total of  $k = 9$  times, i.e.,  $3 * 9$ . Therefore, the proportion of all sequences such that  $H = 1$  is

$$P(H = 1|K = 9) = 3k/4^k = \frac{27}{4^9}$$

Now, let us consider sequences of  $H = 2$ . Of each of the  $k = 9$  bases, we can choose  $h = 2$  bases to edit. And we can edit each of these bases twice, which is  $3^2$  possibilities per base we can edit. So we have

$$P(H = 2|K = 9) = \frac{1}{4^9}(3^2) \binom{9}{2}$$

We can see that in general,

$$P(H = h|K = k) = \frac{3^h}{4^k} \binom{k}{h}$$

We can explicitly show that this is a Binomial( $k, p = 3/4$ ) by re-writing the above as (splitting up the denominator)

$$\begin{aligned} \implies P(H = h|K = k) &= \frac{3^h * 1^{k-h}}{4^h * 4^{k-h}} \binom{k}{h} \\ P(H = h|K = k) &= \binom{k}{h} \left(\frac{3}{4}\right)^h \left(\frac{1}{4}\right)^{k-h} \end{aligned}$$

And therefore, the proportion of all sequences that differ by at most  $H = h$  from some target sequence is

$$\begin{aligned} P(H \leq h|K = k) &= \frac{3^0}{4^k} \binom{k}{0} + \frac{3^1}{4^k} \binom{k}{1} + \frac{3^2}{4^k} \binom{k}{2} + \dots + \frac{3^h}{4^k} \binom{k}{h} \\ \implies P(H \leq h|K = k) &= \sum_{i=0}^h \frac{3^i}{4^k} \binom{k}{i} \end{aligned}$$

which is the CDF of the random variable  $H$ .

We are interested in the proportion of all sequences that differ by no more than  $H$  in the total population of sequences. Since probability is related to information (Shannon 1948), the self-information of a random variable is given by

$$I = -\log_2 p$$

Here, we are interested in the information encoded by a population of  $k$ -mers that differ from each other by at most  $H$  characters. This is given by

$$I(H = h|K = k) = -\log_2 \sum_{i=0}^h \frac{3^i}{4^k} \binom{k}{i}$$

It can be shown that as the Hamming distance upper limit increases, the self-information encoded by the population of sequences decreases. We will plot the results below:

```
library(ggplot2)
library(patchwork)

k <- 20      # k-mer length
p <- 3/4     # probability of a given mismatch

h_values <- 0:k # range of hamming values
cdf_values <- pbinom(h_values, k, p) # binomial CDF

# compute information content, accounting for log(0) case
info_content <- -log2(pmax(cdf_values, 1e-10))

# create df for plotting
```

```

data_cdf_info <- data.frame(h = h_values, CDF = cdf_values,
                             InfoContent = info_content)

plot_cdf <- ggplot(data_cdf_info) +
  geom_step(aes(x = h, y = CDF), color = "blue", direction = "mid") +
  geom_point(aes(x = h, y = CDF), color = "red") +
  labs(title = "CDF of binomial(k=20, p=3/4)",
       x = "Hamming distance h",
       y = "P(H<h)") +
  theme_minimal()

plot_info <- ggplot(data_cdf_info) +
  geom_line(aes(x = h, y = InfoContent), color = "gold", linewidth = 1) +
  labs(title = "Information content -log2(CDF)",
       x = "Hamming distance h",
       y = "information content (bits)") +
  theme_minimal()

# arrange plots
plot_cdf + plot_info

```

