1 Entropy diversity index

We first tried measuring the "diversity" of each sample by the number of non-zero (unique) k-mers present in the sample. This seemed unsatisfactory because we witnessed "saturation": most all k-mers were observed.

Another measure is entropy. Fix a given sample. Let u_i be the count of k-mer i in the sample, $0 \leq i \leq 4^k - 1$. The "probability" of each k-mer is $p_i = \alpha u_i$, with normalizing constant $\alpha = 1/(\sum_i u_i)$.

$$\begin{aligned} h &= \sum_i p_i \log p_i \\ &= \sum_i \alpha u_i \log(\alpha u_i) \\ &= \alpha(\log \alpha \sum_i u_i + \sum_i (u_i \log u_i)) \end{aligned}$$

In order to achieve numerical stability, we might instead consider the log-entropy

$$\begin{aligned} \log h &= \log(\sum_i p_i \log p_i) \\ &= \log \alpha + \log(\log \alpha \sum_i u_i + \sum_i (u_i \log u_i)) \\ &= -\log(\sum_i u_i) + \log(-\log(\sum_i u_i) \sum_i u_i + \sum_i (u_i \log u_i)) \end{aligned}$$

The above formulas also hold when we replace u_i with the counts from the degree distribution jc_j, where $c_j =$ "the number of k-mers with frequency j".

2 Jaccard index

An alternative distance metric between two samples is the Jaccard index. This has the advantage over the Bray-Curtis dissimilarity of satisfying the trianlge inequality.

$$J(x, y) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$