Algorithms for myloasm - living document

Jim Shaw - 2025

1. Distances between unitig-read distributions

Let $X = (x_1, x_2, x_3, ...)$ and $Y = (y_1, y_2, y_3, ...)$ be two distributions of read depths for two unitigs. Notably, each $x_i \in \mathbb{R}^n$ are vectors given a range of SNPmer identity values (S).

Our goal is to calculate a distance between these two distributions: $D(X, Y)$. $D(X, Y)$. Given an edge, $e(X, Y)$, between a unitig path (represented by X) and an adjacent path or contig (Y), we use $D(X, Y)$ to tell how "favourable" this connection is. The "probability" of this connection is a function of the distributions X, Y and also the overlap properties of the edge e. Therefore,

$$\Pr(e(X, Y)) = \exp\left(-\frac{[D(X, Y) + f(e)]}{T}\right) \tag{1}$$

given some temperature $0.5 \leq T \leq 2$. The probability of a path will be a product of edge probabilities.

2. Defining $f(e)$.

Given an overlap, we consider the overlap length $\ell(e)$ and the overlap identity $\sigma(e)$. For now, we let

$$f(e) = \left[1 - \frac{\ell(e)}{\max_{\{e \in X^+\}}\{\ell(e)\}}\right] * c(e) + \left[\max_{\{e \in X^+\}}\{\sigma(e)\} - \sigma(e)\right] \tag{2}$$

We will define $c(e)$ to be a coverage control factor for overlaps. Lower converages mean higher variance in overlap lengths, so the first term should be less confident at low coverage. For now, this is something like $1 - \frac{3}{\min(x[-1], y[-1]) + 3}$ where x, y are the adjacent reads and $x[-1]$ is the last entry of the vector x, where more stringent overlap coverage thresholds have later entries.

The first term controls for the relative overlap length of the edge compared to other adjacent edges from X: X^+. We consider e a directed edge.

The second term controls for the relative difference in overlap identity compared to adjacent edges. The idea is that given 0.99 vs 1.0 σ values, we get a factor of $e^{-\frac{1}{T}}$. Similarly, a vastly shorter overlap gives a factor of $e^{-\frac{1}{T}}$. Given no adjacent edges, this score is of course 0.

3. Defining $D(X, Y)$

Given a vector x, define

$$\log_a(x[i]) = \log(x[i] + a) \tag{3}$$

as a log with pseudocount. We let $a = 3$ in general. Then define the coordinate-wise log distance as:

$$d(x^i, y^i) = |\log_a(x[i]) - \log_a(y[i])|. \tag{4}$$

We want to get a grasp of this distance over the distribution X, Y in a principled way. Intuitively, if unitig X has coverage 20 and unitig Y has coverage 10, then given some scaling factor C (that can

depend on a range of factors), $D(X, Y)$ should be about $C \log\left(\frac{20}{10}\right) = C \log(2)$. This gives a probability of $e^{-C\frac{\log(2)}{T}} = \frac{1}{2^{\frac{C}{T}}}$. However, we must take into account two things:

1. uncertainty estimates in X and Y, which are extremely noisy and have small sample sizes (e.g. a unitig with only one read)
2. high variance in X and Y, as read coverages can vary greatly due to intragenomic repeats (duplications), integernomic repeats (shared strain content/horizontal gene transfer), alignment artefacts, etc.

3.1. Dealing with uncertainty and variance in X and Y

Let $M(X)$ be the coordinate-wise median of the distribution X and also similarly for Y. Given samples X and Y from their respective, unknown probability distributions, we try to reason with $d(M(X)^i, M(Y)^i)$. We also have the distribution of pairwise log differences, $d(X^i, Y^i)$ on hand. Let's assume $d(X^i, Y^i)$ has some average μ and standard deviation Σ.

3.1.1. Variance

To deal with variance, we divide $d(M(X)^i, M(Y)^i)$ by 0.5 plus the standard deviation over the distribution of log differences, Σ. We add a factor of 0.5 stabilize, but also make it < 1 to penalize large distances under small variance. To estimate the standard deviation, we take the IQR. Theoretically, many distributions have IQR $\propto \Sigma$.

3.1.2. Sample size

To deal with sample sizes N_x for X and N_y for Y, we proceed by normalizing by a scaled confidence interval length. Let CI be the confidence interval length for the median log ratio given some confidence %. Given a distribution $d(X^i, Y^i)$, we normalize by $1 + \text{CI}(d(X^i, Y^i))$ which should go to 1 as N_x, N_y get large.

We can calculate this confidence interval in many ways. We could do boostrapping, where we resample $d(X^i, Y^i)$ and take medians. The main issue is that when Y consists of a single read. This is because the "sample size" is essentially 1 here, but the bootstrap "feels" like the sample size is N_x.

Instead, we propose the following. Under the assumption of normality, the confidence interval length is $\propto \frac{\Sigma}{\sqrt{N}}$. We take N as $\frac{2}{\frac{1}{N_x}+\frac{1}{N_y}} = H\left(N_x, N_y\right)$, the harmonic mean, to bias for lower sample sizes. We already have a robust estimator proportional to Σ: the IQR.

3.1.3. Final formula

$$D(X, Y) = \sum_{i=1}^{n} \frac{d(M(X)^i, M(Y)^i)}{\left(\left[0.5 + \frac{1}{1+\max(N_x, N_y)}\right] + \hat{\Sigma}\right) * \left(1 + \frac{\hat{\Sigma}}{\sqrt{H(N_x, N_y)}}\right)} \tag{5}$$