Sample size

Kristoffer Sahlin

1 Notes /thoughts

There is no way to be able to predict the actual E-size or N50 of an assembly unless you can examine and traverse the graph. This is because you need to know where the repeats that breaks the graph are occurring. Given the exact same repeat sizes and copy numbers, two different genomes can have two different E sizes for identical data just because the repeats are located in different locations with different distances between them. Given this, it suggests that it might also be difficult to predict the optimal e-size/ N50??

2 Adaptive sample size

Our objective function is average number of nodes in a unitig ξ. We want to bound the error of ξ to at most δ_ξ, *i.e.* $\hat{\xi} = (1+\delta_\xi)\xi$, $\delta_\xi \in [-1, 1]$. First we note that ξ is obtained as

$$\xi = \frac{p_e}{p_i/2} \tag{1}$$

Where p_e is the number of extremity nodes in the graph and p_i is the number of internal (unary) nodes (each unitig has two extremity nodes, hence the division by two). The number of extremities and internal nodes follows binomial distribution, with true proportions p_e and $p_i = 1 - p_e$ respectively. In our sample, we only have either extremity or internal nodes, thus $p_e = n - p_i$ if n is our sample size and the absolute error of the sample δ_p will be the same for the two proportions. Let ϵ be the sampling error bound for these two proportions (note that if one of the quantities is underestimated, the other one is overestimated or vice versa), then

$$\epsilon = \pm z_{\alpha/2}\sqrt{\frac{\hat{p_e}(1-\hat{p_e})}{n}} \tag{2}$$

which gives

$$n = (\frac{z_{\alpha/2}}{\epsilon})^2 \hat{p_e}(1-\hat{p_e}). \tag{3}$$

We have

$$\hat{\xi} = \frac{\hat{p_i}}{\hat{p_e}/2} = \frac{2(1-\hat{p_e})}{\hat{p_e}} = \frac{2(1-p_e)(1-\delta p)}{(1+\delta p)p_e} \tag{4}$$

but we have $\hat{\xi} = (1+\delta_\xi)\xi$ so we can write (4) as

$$(1+\delta_\xi)\xi = \frac{2(1-p_e)(1-\delta p)}{(1+\delta p)p_e} \Rightarrow (1+\delta_\xi) = \frac{(1-\delta p)}{(1+\delta_p)} \tag{5}$$

This gives

$$\delta_p = \frac{\delta_\xi}{2 + \delta_\xi} \tag{6}$$

Given a fixed value of δ_ξ, that is a fixed maximum procent of ξ with a 95% confidence, we have δ_p as the maximum procent of error we can tolerate for p_e. The procent is fixed.

We now use (3) and let $\epsilon = \delta_p \hat{p_e}$. ϵ is then interpreted as the size of the maximum error we can tolerate given that we want to bound the error of $\hat{\xi}$ to δ_ξ. Notice that the smaller p_e is, the smaller ϵ gets, and thus, the larger the sample size needed according to (3).

Adaptive sample size To adapt our sample size across k, we initialize p_e and and get a corresponding sample size. As we switch from k to $k+1$, we let

$$n_{k+1} = (\frac{z_{\alpha/2}}{\epsilon_{k+1}})^2 \hat{p_{ek}}(1 - \hat{p_{ek}}). \tag{7}$$

Where $\epsilon_{k+1} = \delta_k \hat{p_{ek}}$. This means that the relative error of p_e for $k+1$ is predicted by the previous

3 Esitmate p_e and p_i by weighted sampling