Optimal_k - DB-graph inference by accurate sampling

1 Abstract

Motivation: There is no clear way on how to chose parameters k-mer size and abundance for a De Bruijn based de novo assembler. As *de novo* genome assembly is time consuming for large genomes, it is of importance to chose these parameters well in order to prevent multiple runs. Current software for estimating k only optimize certain features such as maximizing the number of genomic k-mers. There is a need for more clear objectives such as E-size or N50.

Results: We provide a method (optimal_k) to estimate average unitig length, N50 and E-size for all combinations of minimum abundance and k in one run. As unitigs are a foundation of the de Bruijn graph, estimating these quantities provides an understanding of the quality of a DBG based genome assembly as well as a good base for chosing the best combination of k and abundance. The estimations obtained by optimal_k are extremely accurate. [We also note that these estimations also accurately predict the best quality for DBG based assemblers that perform more steps such as tip removals, bubble popping and usage of paried end read information.]

2 Introduction

Mention that there are not many tools for computing optimal parameters at all. And make sure to mention that memry is not the issue. Mention the positives about our methods like speed and clear objective function but make sure to mention that it's memory requiring but thats not a problem if you are going to do the assembly anyway!!

Given an assembly graph G, a path $P = (v_0, v_1, \ldots, v_t)$ in G, where $t \geq 0$, is called a *unitig* if for all $i \in \{1, \ldots, t-1\}$, v_i is a *unary* node, that is, v_i has exactly one in-neighbor and one out-neighbor in G, and v_0 and v_t are not unary. In the contig assembly phase, popular genome assemblers report a unitig decomposition of the assembly graph, after some artifacts have been been dealt with, like tip removal and bubble popping.

3 Methods

The general idea is to provide the user with metrics such as unitigs N50 and E-Size and average number of genomic vertices in a DBG for all possible k-mer sizes and abundances.
We implement a FM-index data structure described in cite XX. This allows us to query

a k-mer, its in and out neighbors in O() time. We furthermore derive formulas for how much we need to sample in order to reach a given accuracy on all our estimates.

We let $\mathcal{K}_k$ be the multiset of all k-mers in the reads, that is $|\mathcal{K}_k| = n(r-k+1)$, where n denotes the number of reads and r is the read length. Let $\mathcal{G}_{k,a}$ be the DBG with vertices of length k that has an abundance $\geq a$.

3.1 Algorithm

3.2 Sampling accuracy

In this section we will derive the sample sizes required to get accurate estimations of the quantities that we want to estimate. We recall some basic statistical notions in Sec. 3.2.1 and show how they applies to our quantities in Sec. 3.2.2.

3.2.1 Theory

Sample proportion. Suppose that we have a set partitioned as $A \cup B$, and we need to estimate the proportion $p = |A|/(|A| + |B|)$. Suppose that we sample n elements of $A \cup B$ and for each of them record whether they belong to A or to B, obtaining in this way an estimate $\hat{p}$ of p. It is a standard result that if we obtain $\hat{p}$ from n samples, the $100(1-\alpha)\%$ confidence interval of $\hat{p}$ is

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

where $z_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile from the normal distribution. For a given relative error ϵ, we want to choose the sample size n such that the $100(1-\alpha)\%$ confidence interval of $\hat{p}$ has a margin of error no more than $E := \epsilon p$. By standard means, we obtain

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \hat{p}(1-\hat{p}). \tag{1}$$

Notice that in relation (1) above, both p and $\hat{p}$ are not known at the start of the sampling, when the value of n needs to be chosen. In our case, we choose ...

Fraction of sample proportions. Instead of estimating the proportion $p = |A|/(|A|+|B|)$, we now want to have a $100(1-\alpha)\%$ confidence interval of the fraction of proportions $f = \frac{p}{1-p}$. Suppose that we already have an estimate $\hat{p}$ of p with an absolute error ϵ_p, that is $\hat{p} = (1 \pm \epsilon_p)p$.

We will estimate f as

$$\hat{f} = \frac{\hat{p}}{1-\hat{p}},$$

and will denote the absolute error of $\hat{f}$ as ϵ_f, that is, $f = (1 \pm \epsilon_f)\hat{f}$. We have

$$(1 \pm \epsilon_f) = \frac{(1 \pm \epsilon_p)(1-p)}{(1 \pm \epsilon_p)p}.$$

Notice that the margin of error increases as p decreases. Fixing p, the error of f is maximized by

$$(1 \pm \epsilon_f) = \frac{(1+\epsilon_p)(1-p)}{(1-\epsilon_p)p}.$$

Finally, since we sample p, we solve this equation for ϵ_p and get

$$\epsilon_p = \frac{\epsilon_f}{2+\epsilon_f}. \tag{2}$$

With the above equation, we now have a way to see what margin of error ϵ_p we require to arrive at a fixed margin of error of f. That is, if we want to have at most 10% error of our estimate $\hat{f}$, we need to have a sample size that calculated from letting $\epsilon = \frac{0.1}{2+0.1}$.

3.2.2 Application to sampling DBGs

We will use the theory in 3.2.1 to get accurate sample estimates of the desired quantities.

Estimating the number of nodes in a DBG Let X be the set of k-mers included in $\mathcal{G}_{k,a}$. That is, $X = \sum_{k \in \mathcal{K}_k} \frac{1}{a_k} I_{k \geq a}$ where I is the indicator function. The (multi)set of k-mers that are members of X and it's complement partitions $\mathcal{K}$. The true proportion p_k of X in $\mathcal{K}_k$ is given by

$$p_k = \frac{\sum_{k \in \mathcal{K}_k} \frac{1}{a_k} I_{k \geq a}}{\sum_{k \in \mathcal{K}_k}} \tag{3}$$

We can estimate p_k with sampling k-mers $k_i, i \in [1, m]$ as

$$\hat{p_k} = \frac{\sum_{i=1}^{m} \frac{1}{a_{k_i}} I_{k_i \geq a}}{m} \tag{4}$$

Thus, we get an estimate of X as $\hat{X} = \hat{p_k} * \mathcal{K}$. By the theory in **??**, we immediately get how many samples we need to bound the relative error.

Estimating the average number of nodes in a DBG Note that we get the number of unitigs in a graph by counting all the start vertices and their outdegree. Let X_s be the number of start nodes in $\mathcal{G}$. We also label all ot in the DBG is given We now divide the set X into Y be the number of internal vertices in the DBG. An internal vertex in a vertex that...

4 Results and discussion

5 Conclusions