Optimal_k - DB-graph inference by accurate sampling
Fast and accurate selection of parameters for genome assembly (by 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!!

A unitig of a graph is a maximal unary path. 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.

Notation. We assume that the input consists of a set R of n reads of length r. We denote by K_k, the multiset of all k-mers in the reads; observe that $|\mathsf{K}_k| = n(r-k+1)$. Moreover, we denote by $\mathsf{DB}_{k,a}$, the de Bruijn graph with vertices of length k and *minimum abundance* a. That is, the set of vertices of $\mathsf{DB}_{k,a}$ is the set of all k-mers in the reads which occur at least a times in R, and two vertices of $\mathsf{DB}_{k,a}$ are connected by an arc if they have a suffix-prefix overlap of length $k-1$. Let $V(\mathsf{DB}_{k,a})$ denote the set of vertices of $\mathsf{DB}_{k,a}$. For all $v \in V(\mathsf{DB}_{k,a})$, let $\alpha(v)$

denote the abundance of k-mer v. We denote by $\delta^+_{k,a}(v)$ the number of out-neighbors of v in $\mathsf{DB}_{k,a}$, and by $\delta^-_{k,a}(v)$ the number of in-neighbors of v in $\mathsf{DB}_{k,a}$. These values can be obtained by queries to the index built on the set R.

A node v of $\mathsf{DB}_{k,a}$ is called *unary* if $\delta^-_{k,a}(v) = \delta^+_{k,a}(v) = 1$, and it is called *isolated* if $\delta^-_{k,a}(v) = \delta^+_{k,a}(v) = 0$. A path in $\mathsf{DB}_{k,a}$ is called a *unitig* if all its internal vertices are unary, and its two extremities are not. When clear from the context, we will also use the term unitig to denote the string spelled by a unitig path in $\mathsf{DB}_{k,a}$.

Talk about reverse complements in dB graphs: a k-mer and its reverse complement are bundled into the same node and the abundances are added up. Say how we deal with this case in practice.

Give pseudo-code of how we get the in/out degrees for all abundances.

Say that one of the main ideas is to do weighted sampling.

3.1 Sampling algorithms

Estimating the number of nodes of a dBG. Let $V(\mathsf{DB}_{k,a})$ denote the set of vertices of $\mathsf{DB}_{k,a}$, and let $\mathbb{I}(x,a)$ be an indicator variable returning 1 if the k-mer x has abundance at least a in R, and 0 otherwise. We can write

$$|V(\mathsf{DB}_{k,a})| = \sum_{x \in \mathsf{K}_k} \frac{1}{\alpha(x)} \mathbb{I}(x,a).$$

Since $V(\mathsf{DB}_{k,a})$ is a subset of the multiset K_k, we can consider the proportion

$$p_{k,a} := \frac{|V(\mathsf{DB}_{k,a})|}{|\mathsf{K}_k|} = \frac{\sum_{x \in \mathsf{K}_k} \frac{1}{\alpha(x)} \mathbb{I}(x,a)}{|\mathsf{K}_k|} \in [0,1].$$

We can estimate $p_{k,a}$ by sampling a multiset $\{x_1, \ldots, x_m\}$ of k-mers from K_k, and taking

$$\hat{p}_{k,a} := \frac{\sum_{i=1}^{m} \frac{1}{\alpha(x_i)} \mathbb{I}(x_i,a)}{m}.$$

Therefore, we also get an estimate of $X_{k,a} := |V(\mathsf{DB}_{k,a})|$ as $\hat{X}_{k,a} = \hat{p}_{k,a}|\mathsf{K}_k| = \hat{p}_{k,a} n(r-k+1)$. Notice that if we sample all k-mers, we get $\hat{X}_{k,a} = \frac{|V(\mathsf{DB}_{k,a})|}{|\mathsf{K}_k|}|\mathsf{K}_k| = |V(\mathsf{DB}_{k,a})|$. In section 3.2 we will discuss how many samples m we need to accurately estimate $\hat{X}_{k,a}$.

Say that we can implement this for all abundances

Estimating the number of unitigs of a dBG. Let $\mathsf{U}_{k,a}$ denote the set of all unitigs of $\mathsf{DB}_{k,a}$. We now derive a simple combinatorial expression for $|\mathsf{U}_{k,a}|$, which is key in the sampling phase. Let $\mathsf{ST}_{k,a}$ denote the set of start nodes of the unitigs of $\mathsf{DB}_{k,a}$, that is,

$$\begin{aligned}\mathsf{ST}_{k,a} := \{v \in V(\mathsf{DB}_{k,a}) \mid \delta^+_{k,a}(v) \geqslant 2 \text{ or} \\ (\delta^+_{k,a}(v) = 1 \text{ and } \delta^-_{k,a}(v) \neq 1) \text{ or} \\ (\delta^+_{k,a}(v) = 0 \text{ and } \delta^-_{k,a}(v) = 0)\}.\end{aligned}$$

Since every node v in $\mathsf{ST}_{k,a}$ is either an isolated node, or it is a start node of a different unitig, starting with v and then continuing to each of its out-neighbors, we can write

$$|\mathsf{U}_{k,a}| = \sum_{v \in \mathsf{ST}_{k,a}} \max(1, \delta^+_{k,a}(v)).$$

As above, we can obtain this number by summing over all k-mers in the reads:

$$|\mathsf{U}_{k,a}| = \sum_{x \in \mathsf{K}_k} \max\left(\frac{1}{\alpha(x)}\mathbb{I}(x,a), \frac{1}{\alpha(x)}\mathbb{I}(x,a)\delta^+_{k,a}(x)\right).$$

We consider the ratio between the number of unitigs and all k-mers in the reads

$$q_{k,a} := \frac{|\mathsf{U}_{k,a}|}{|\mathsf{K}_k|} = \frac{\sum_{x \in \mathsf{K}_k} \max\left(\frac{1}{\alpha(x)}\mathbb{I}(x,a), \frac{1}{\alpha(x)}\mathbb{I}(x,a)\delta^+_{k,a}(x)\right)}{|\mathsf{K}_k|}.$$

Observe that $q_{k,a} \in [0,1]$ since every unitig contains at least one k-mer, thus $|\mathsf{U}_{k,a}| \leqslant |\mathsf{K}_k|$. We can analogously estimate it, after sampling a multiset $\{x_1, \ldots, x_m\}$ of k-mers from K_k, as

$$\hat{q}_{k,a} := \frac{\sum_{i=1}^m \max\left(\frac{1}{\alpha(x_i)}\mathbb{I}(x,a), \frac{1}{\alpha(x_i)}\mathbb{I}(x_i,a)\delta^+_{k,a}(x_i)\right)}{m}.$$

The estimate of $Y := |\mathsf{U}_{k,a}|$ is then $\hat{Y} = \hat{q}_{k,a}|\mathsf{K}_k| = \hat{q}_{k,a}n(r-k+1)$. Similarly to $X_{k,a}$, sampling all k-mers will give $\hat{Y} = |\mathsf{U}_{k,a}|$. In section 3.2, we discuss how many samples m we need to accurately estimate $\hat{Y}$. say we can implement this for all abundances.

Estimating the average length of the unitigs of a dBG. We are now interested in determining the average length of the strings spelled by the unitigs of $\mathsf{DB}_{k,a}$.

Denote by the *truncated length* of a unitig $w = (v_1, v_2, \ldots, v_t)$ the number of its internal vertices plus its start vertex. We first estimate the average truncated lengths of the unitigs of $\mathsf{DB}_{k,a}$, and then obtain the average unitig string length by summing k.[1] Working with the truncated unitig lengths allows us to easily estimate the required sample size.

Let $\mathsf{UN}_{k,a}$ denote the set of unary nodes of $\mathsf{DB}_{k,a}$; notice that $\mathsf{UN}_{k,a} \cap \mathsf{ST}_{k,a} = \emptyset$. The average truncated length of the unitigs is obtained as

$$Z_{k,a} := \frac{|\mathsf{UN}_{k,a} \cup \mathsf{ST}_{k,a}|}{|\mathsf{ST}_{k,a}|}. \tag{1}$$

As above, we can write

$$|\mathsf{UN}_{k,a}| = \sum_{\substack{x \in \mathsf{K}_k \text{ with} \\ \delta^+_{k,a}(x) = \delta^-_{k,a}(x) = 1}} \frac{1}{\alpha(x)}\mathbb{I}(x,a).$$

We consider the proportion

$$r_{k,a} := \frac{|\mathsf{ST}_{k,a}|}{|\mathsf{UN}_{k,a} \cup \mathsf{ST}_{k,a}|} =$$
$$= \frac{\sum_{x \in \mathsf{K}_k} \max\left(\frac{1}{\alpha(x)}\mathbb{I}(x,a), \frac{1}{\alpha(x)}\mathbb{I}(x,a)\delta^+_{k,a}(x)\right)}{\sum_{\substack{x \in \mathsf{K}_k \text{ with} \\ \delta^+_{k,a}(x) = \delta^-_{k,a}(x) = 1}} \frac{1}{\alpha(x)}\mathbb{I}(x,a) + \sum_{x \in \mathsf{K}_k} \max\left(\frac{1}{\alpha(x)}\mathbb{I}(x,a), \frac{1}{\alpha(x)}\mathbb{I}(x,a)\delta^+_{k,a}(x)\right)} \in [0,1].$$

We obtain an estimate $\hat{r}_{k,a}$ of $r_{k,a}$ by sampling a multiset $\{x_1, \ldots, x_m\}$ of k-mers in K_k with abundance at least a (that is, for which the indicator variable $\mathbb{I}(x_i, a)$ is 1):

$$\hat{r}_{k,a} := \frac{\sum_{i=1}^m \max\left(\frac{1}{\alpha(x_i)}, \frac{1}{\alpha(x_i)}\delta^+_{k,a}(x_i)\right)}{\sum_{\substack{i \in [1,m] \text{ with} \\ \delta^+_{k,a}(x_i) = \delta^-_{k,a}(x_i) = 1}} \frac{1}{\alpha(x_i)} + \sum_{i=1}^m \max\left(\frac{1}{\alpha(x_i)}, \frac{1}{\alpha(x_i)}\delta^+_{k,a}(x_i)\right)}.$$

[1] This assumes, in order to simplify the presentation, that the dBG has no isolated nodes, which are unitigs with 0 internal nodes and truncated length 1, but spell strings of length k. However, isolated nodes can be easily accounted for as separate case in all the formulas.

An estimate $\hat{Z}_{k,a}$ for the quantity from (1) is then obtained as $1/\hat{r}_{k,a}$.

Similarly to $X_{k,a}$ and $Y_{k,a}$, sampling all k-mers will give $\hat{Z}_{k,a} = Z_{k,a}$. We discuss how many samples m we need to accurately estimate $\hat{Z}$ in section 3.2. say we can implement this for all abundances.

Estimating the E-size of the unitigs of a dBG. The E-size of the set $\mathsf{U}_{k,a}$ of unitig strings of $\mathsf{DB}_{k,a}$ is defined as the expected length of the unitig strings of $\mathsf{DB}_{k,a}$ [**?**]. That is, the expected contig length covering any position on the genome. Formally,

$$\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a}) := \sum_{w \in \mathsf{U}_{k,a}} |w| P(w) = \sum_{w \in \mathsf{U}_{k,a}} |w| \frac{|w|}{\sum_{w' \in \mathsf{U}_{k,a}} |w'|} = \frac{\sum_{w \in \mathsf{U}_{k,a}} |w|^2}{\sum_{w \in \mathsf{U}_{k,a}} |w|}, \tag{2}$$

where $|w|$ denotes the length of the string spelled by the unitig w and $P(w)$ the probability of sampling a position on the genome covered by w. The genome is taken to be the concatenation of all unitigs, so $P(w) = \frac{|w|}{|G|}$, where $|G|$ is the genome length. We also denote the number of nodes of a unitig w as $||w||$.

In order to derive an unbiased sampling procedure of $\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a})$, it is important to notice two points in Equation 2.

- The length of w determines how likely it is to sample w, *i.e.* $P(w_1) > P(w_2)$ if $w_1 > w_2$
- The abundance of the nodes in w is not accounted for.

Since we will be sampling k-mers and their abundance determines how likely it is that we sample each of them, we need to weight each unitig samples from a k-mer with the average abundance of the unitig.

Our sampling procedure produces a multiset W of unitigs of $\mathsf{DB}_{k,a}$. We choose a k-mer $x \in \mathsf{K}_k$ at random. If x is a node of $\mathsf{DB}_{k,a}$, we output all the unitigs containing x. These unitigs can be obtained by traversing the graph along the in-/out-neighbors of x, after taking into account whether x is a unary node of $\mathsf{DB}_{k,a}$ or not.

Suppose that the above procedure samples every node of $\mathsf{DB}_{k,a}$ exactly once, and that $\overline{W}$ is the resulting multiset of unitigs. Then, each unitig $w := (v_1, v_2, \ldots, v_t)$ of $\mathsf{DB}_{k,a}$ appears $\alpha(w) := \sum_{i=1}^{t} \alpha(v_i)$ times in $\overline{W}$. However, since not all abundances are equal, we need to remove the bias from over- or under-sampling k-mers due to the abundance difference. Observe that if all k-mers of $\mathsf{DB}_{k,a}$ had the same abundance a, w would appear ta times in $\overline{W}$. Therefore, we can equivalently express the E-size of the set $\mathsf{U}_{k,a}$ by normalizing the probability of w with $a||w||/\alpha(w)$. This gives the following expression:

$$\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a}) = \sum_{w \in \overline{W}} |w| \frac{|w| \frac{a||w||}{\alpha(w)}}{\sum_{w \in \overline{W}} |w| \frac{a||w||}{\alpha(w)}} = \sum_{w \in \overline{W}} |w| \frac{|w| \frac{||w||}{\alpha(w)}}{\sum_{w \in \overline{W}} |w| \frac{||w||}{\alpha(w)}}. \tag{3}$$

However, we cannot afford to sample all k-mers in K_k. By sampling only a subset of k-mers we obtain the multiset $W = \{w_1, \ldots, w_m\}$ of unitigs, the above relation (3) shows that we can estimate $\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a})$ as

$$\hat{\mathsf{E}}_{\mathsf{size}} := \sum_{i=1}^{m} |w_i| \frac{|w_i| \frac{||w_i||}{\alpha(w)}}{\sum_{j=1}^{m} |w_j| \frac{||w_j||}{\alpha(w_j)}}.$$

Say that here we cannot guarantee a sampling accuracy (if so). I think maybe we could derive something here, gonna thing about it. However notice that we cannot bound any error with certainty in all of our estimates (see my change to "accuratly estimate" instead of "bound error"). Because we can always end up with a counter example that: one isolated node has a bazilion in abundance and the rest of the k-mers only have, say, one in abundance. We would

only sample the isolated one and therefore end up with an estimate that is completely off. However, what we CAN say is that if we resample we can reproduce our results with a given accuracy (definition of confidence interval). Also, it seems to work in practice :) (statistics is not a "without doubt"-science like math).

3.2 Sampling accuracy

Suppose that we have a set partitioned as $A \cup B$, and we need to estimate the proportion $p = |A|/(|A| + |B|) \in [0, 1]$. Suppose that we sample m elements of $A \cup B$ and for each of them record whether they belong to A or to B, and then divide these two counts by m, obtaining in this way an estimate $\hat{p}$ of p. It is a standard result that the $100(1-\alpha)\%$ confidence interval of $\hat{p}$ is

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

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 $E := \varepsilon p$. By standard means, we obtain

$$m \geqslant \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \hat{p}(1-\hat{p}). \tag{4}$$

Notice that in the relation (4) above, both p and $\hat{p}$ are not known at the start of the sampling, when the value of m needs to be chosen.

In our case, we choose p and $\hat{p}$ ...

Moreover, if we want to estimate $f = 1/p$ with a given relative error ε', then we can estimate it as $\hat{f} = 1/\hat{p}$. In this case, we need to set the relative error ε of $\hat{p}$ as $\varepsilon = 1/(1+\varepsilon') - 1$. I stop here with the note that I have to clarify the last part in this section (i.e f=1/p) a bit. Also, we need to illustrate how it works in practice (inital estimate, then updating sample size as we go for a fixed epsilon= say 0.05)

3.3 New sampling

Let Y be a random variable, from a distribution with finite mean μ and finite non-zero variance σ^2. By the Central Limit Theorem, the $100(1-\alpha)\%$ two-sided confidence interval of μ approaches

$$\left[\overline{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{m}}, \overline{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{m}}\right]$$

as the number of samples m increases. In the above, $\frac{\sigma}{\sqrt{m}}$ is the standard deviation of sample mean $\overline{y}$, and $z_{\frac{\alpha}{2}}$ denotes the $\alpha/2$ quantile from the normal distribution.

In order to guarantee a relative error ϵ of

Let now $x_1, \ldots, x_N$ denote the lengths of all unitigs in $\mathsf{DB}_{k,a}$. Suppose that the distribution of unitig lengths in $\mathsf{DB}_{k,a}$ follows an unknown distribution $f(x)$, and let X be random variable over $f(x)$.

Observe that the definition (2) can be equivalently written as

$$\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a}) = \frac{N\sum_{i=1}^{N} x_i^2 \frac{1}{N}}{N\sum_{i=1}^{N} x_i \frac{1}{N}} = \frac{\mathrm{E}[X^2]}{\mathrm{E}[X]}.$$

To estimate the E-size, we will sample unitig lengths x and obtain estimates $\overline{x}$ and $\overline{x^2}$ of $\mathrm{E}[X]$ and $\mathrm{E}[X^2]$, respectively. Recall that a sample estimate is itself a random variable. Thus, in order to get a confidence interval of the E-size, we need to find the sample standard deviation

of $\overline{x^2}/\overline{x}$, which we denote by $s_{\text{E-size}}$. Then, applying the Central Limit Theorem, we obtain the desired confidence interval of E-size as

$$\left[\overline{y} - z_{\frac{\alpha}{2}} s_{\text{E-size}}, \overline{y} + z_{\frac{\alpha}{2}} s_{\text{E-size}}\right].$$

If ... is ..., then the first order Taylor expansion gives a good estimation of $s^2_{\text{E-size}}$ [**?**]:

$$s^2_{\text{E-size}} = \text{Var}\left[\frac{\overline{x^2}}{\overline{x}}\right] \approx \frac{\text{Var}\left[\overline{x^2}\right]}{\text{E}\left[\overline{x}\right]^2} - 2\frac{\text{E}\left[\overline{x^2}\right]}{\text{E}\left[\overline{x}\right]^3}\text{Cov}\left[\overline{x^2}, \overline{x}\right] + \frac{E\left[\overline{x^2}\right]^2}{E\left[\overline{x}\right]^4}\text{Var}\left[\overline{x}\right].$$

4 Results and discussion

5 Conclusions