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 or length r. We denote by K_k be the multiset of all k-mers in the reads, and note that $|\mathsf{K}_k| = n(r-k+1)$. Moreover, we denote by $\mathsf{DB}_{k,a}$ be the de Bruijn graph of *order* k and *minimum abundance* a built on R. 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 the 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 out-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 $\mathsf{DB}_{k,a}$. Given a unitig $w = (v_1, v_2, \ldots, v_t)$ of $\mathsf{DB}_{k,a}$, we denote by $|w|_n$ the number of nodes of w, i.e., $|w|_n = t$, and by $|w|_s$ the length of the string spelled by w, that is, $|w|_s = k + t - 1$.

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 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 $X := |V(\mathsf{DB}_{k,a})|$ as $\hat{X} = \hat{p}_{k,a}|\mathsf{K}_k| = \hat{p}_{k,a} n(r - k + 1)$. By the theory in Sec. 3.2, we immediately get how many samples we need in order to bound the relative error within a certain confidence interval.

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

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

Every node v in $\mathsf{ST}_{k,a}$ is either an isolated node, or it is a start node of a different a unitig for each of its out-neighbors. We can therefore write

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

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

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

Analogously, we can consider the proportion of start nodes over all k-mers in the reads

$$q_{k,a} := \frac{|\mathsf{ST}_{k,a}|}{|\mathsf{K}_k|} = \frac{\sum_{x \in \mathsf{K}_k} \max\left(\frac{1}{\alpha(x)}\mathbb{I}(x,a), \frac{1}{\alpha(x)}\mathbb{I}(x,a)\delta^+_{k,a}(x)\right)}{|\mathsf{K}_k|} \in [0,1],$$

and 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 $Y := |\mathsf{ST}_{k,a}|$ is then $\hat{Y} = \hat{q}_{k,a}|\mathsf{K}_k| = \hat{q}_{k,a}n(r-k+1)$. By the theory in Sec. 3.2, we immediately get how many samples we need in order to bound the relative error within a certain confidence interval. say we can implement this for all abundaces.

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 $\mathsf{DB}_{k,a}$. It is enough to first estimate the average number of unary (i.e. internal) nodes of of a unitig, and then obtain the average string length by summing $k+1$.[1] Since above we have obtained an estimate for the number of unitigs, the idea here is to estimate also the number of unary nodes, and then obtain the estimate for the average number of nodes by diving the latter by the former.

Let $\mathsf{UN}_{k,a}$ denote the set of unary nodes of $\mathsf{DB}_{k,a}$. As above, we can write

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

Since we want to estimate the ratio

$$r_{k,a} := \frac{|\mathsf{UN}_{k,a}|}{|\mathsf{UN}_{k,a} \cup \mathsf{ST}_{k,a}|}$$

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 node in a node that...

Estimating the E-size of the unitigs of a dBG. Let $\mathsf{U}_{k,a}$ denote the set of all unitigs of $\mathsf{DB}_{k,a}$. The E-size of $\mathsf{U}_{k,a}$ is defined as the expected length of the unitigs of $\mathsf{DB}_{k,a}$. Formally,

$$\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a}) := \sum_{s \in \mathsf{U}_{k,a}} |s|p(s) = \sum_{s \in \mathsf{U}_{k,a}} |s|\frac{|s|}{\sum_{s' \in \mathsf{U}_{k,a}} |s'|} = \frac{\sum_{s \in \mathsf{U}_{k,a}} |s|^2}{\sum_{s \in \mathsf{U}_{k,a}} |s|}.$$

In order to derive an estimate for E-size, we will construct a sampling procedure which may sample a unitig more than once. Since the above formula contains each unitig once, we need to normalize it with the expected number of times of sampling each unitig.

Consider the following procedure which produces a multiset W containing the unitigs of $\mathsf{DB}_{k,a}$: for all k-mers $x \in \mathsf{K}_k$, if x is a node of $\mathsf{DB}_{k,a}$, output all the unitigs containing x. Each

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

unitig $w := (v_1, v_2, \ldots, v_t)$ of $\mathsf{DB}_{k,a}$ appears $\alpha(w) := \frac{1}{t}\sum_{i=1}^{t} \alpha(v_i)$ times in W. Therefore, we can equivalently express the E-size of the set $\mathsf{U}_{k,a}$ as

$$\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a}) = \sum_{w \in W} |w| \frac{|w| \frac{1}{\alpha(w)}}{\sum_{w \in W} |w| \frac{1}{\alpha(w)}}.$$

Instead of iterating over all k-mers in K_k, we now sample a multiset of k-mers in K_k, and if they are nodes in the de Bruijn graph, we report the unitigs containing these k-mers. Assume that this produces a multiset $\{w_1, \ldots, w_m\}$ of unitigs of $\mathsf{DB}_{k,a}$. We can estimate $\mathsf{E}_{\mathsf{size}}(\mathsf{U}_{k,a})$ as

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

3.2 Sampling accuracy

Suppose that we have a set partitioned as $A \cup B$, and we need to estimate the proportion $p = |A|/(|A| + |B|)$. Suppose that we sample n 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 n, 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

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

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

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

Notice that in the 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 ...

4 Results and discussion

5 Conclusions