Peer Graded Assignment: Bioinformatics Application Challenge

You submitted!
Your work is ready to be reviewed by classmates. Next, you need to review your classmates' work. We'll email you when your grade is ready. Your grade should be ready by **July 27, 11:59 PM PDT**.

Review Classmates' Work

Instructions

My submission

Discussions

Genome assembly

Submitted on July 25, 2016

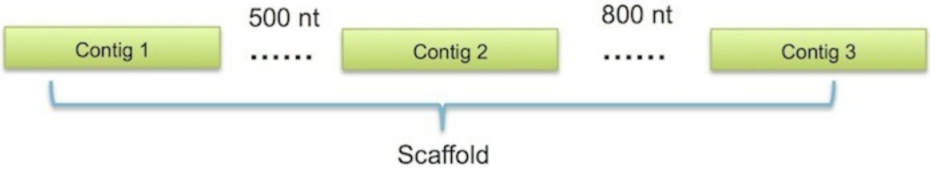
Shareable Link

DEFINITIONS

There are many assembly tools, but none of them is perfect. Biologists therefore need to evaluate the quality of various assemblers by comparing their results. In our case, once we have run the SPAdes assembler on a set of reads, we need to test the quality of the resulting assembly.

Contig: A *contiguou*s segment of the genome that has been reconstructed by an assembly algorithm

Scaffold: An ordered sequence of contigs (possibly separated by gaps between them) that are reconstructed by an assembly algorithm. The order of contigs in a correctly assembled scaffold corresponds to their order in the genome. Existing assemblers specify the approximate lengths of gaps between contigs in a scaffold.



N50 statistic: N50 is a statistic that is used to measure the quality of an assembly. N50 is defined as the maximal contig length for which all contigs greater than or equal to that length comprise at least half of the sum of the lengths of all the contigs. For example, consider the five toy contigs with the following lengths: [10, 20, 30, 60**,** 70]. Here, the total length of contigs is 190, and contigs of length 60 and 70 account for at least 50% of the total length of contigs (60 + 70 = 130), but the contig of length 70 does not account for 50% of the total length of contigs. Thus, N50 is equal to 60.

NG50 statistic: The NG50 length is a modified version of N50 that is defined when the length of the genome is known (or can be estimated). It is defined as the maximal contig length for which all contigs of at least that length comprise at least half of the length of the genome. NG50 allows for meaningful comparisons between different assemblies for the same genome. For example, consider the five toy contigs we considered previously: [10, 20, 30, 60, 70]. These contigs only add to 190 nucleotides, but say that we know that the genome from which they have been generated has length 300. In this example, the contigs of length 30, 60, and 70 account for at least 50% of the genome length (30 + 60 + 70 = 160); but the contigs of length 60 and 70 no longer account for at least 50% of the genome length (60 + 70 = 130). Thus, NG50 is equal to 30.

NGA50 statistic: If we already know a reference genome for a species, then we can test the accuracy of a newly assembled genome against this reference. The NGA50 statistic is a modified version of NG50 accounting for assembly errors (called **misassemblies**). To compute NGA50, errors in the contigs are accounted for by comparing contigs to a reference genome. All of the misassembled contigs are broken at **misassembly breakpoints**, resulting in a larger number of contigs with the same total length. For example, if there is a missasembly breakpoint at position 10 in a contig of length 30, this contig will be broken into contigs of length 10 and 20.

NGA50 is calculated as the NG50 statistic for the set of contigs resulting after breaking at misassembly breakpoints. For example, consider our example before, for which the genome length is 300. If the largest contig in [10, 20, 30, 60, 70] is broken into two contigs of length 20 and 50 (resulting in the set of contigs [10, 20, 20, 30, 50, 60]), then. contigs of length 20, 30, 50, and 60 account for at least 50% of the genome length (20 + 30 + 50 + 60 = 160). But contigs of length 30, 50, and 60 do not account for at least 50% of the genome length (30 + 50 + 60 = 140). Thus, NGA50 is equal to 20.

Based on the above definition of N50, define N75.

N75 is defined as the maximal contig length for which all contigs greater than or equal to that length comprise at least 75% of the sum of the lengths of all the contigs.

Compute N50 and N75 for the nine contigs with the following lengths:

[20, 20, 30, 30, 60, 60, 80, 100, 200].

N50 = 100;
N75 = 60

Say that we know that the genome length is 1000. What is NG50?

NG50 = 60

If the contig in our dataset of length 100 had a misassembly breakpoint in the middle of it, what would be the value of NGA50?

NGA50 = 50

Based on the definition of scaffolds, what information could we use to construct scaffolds from contigs? Justify your answer.

We could use a complete genome sequenced from the same species as a "reference genome" and compare contigs against the "reference genome". We could then infer the order of the contigs in the scaffold through the order of the contigs in the reference genome.

Continue here as soon as your assembly of the Staph reads has completed.

Consider the following three statistics:

- N50.
- The number of ***long* contigs**, i.e., contigs with length ≥ 1000 nucleotides. Biologists are mainly interested in long contigs and often discard short contigs, since short contigs often harbor only fragments of genes rather than complete genes.
- The total length of *long* contigs. This statistic can be combined with N50 and the number of long contigs; a good assembly is one that has relatively few long contigs, but the total length of long contigs is high, as is N50.

Fill in the 9 missing values in the following 3 x 3 table:

k	**N50**	**#long contigs**	**total length of long contigs**
25			
55			
85			

40658; 123; 2794734
154027; 45; 2819863
79093; 71; 2829100

Which assembly performed the best in terms of each of these statistics? Justify your answer.Why do you think that the value you chose performed the best?

k = 55 performed the best since it has largest N50, lowest the number of long contigs, and all three values of k give about the same total length of long contigs.
The reason why k = 55 performed the best is that it is an appropriate length. If the reads are too long (e.g. k = 85), although it is easy to identify where a read came from, the coverage might be compromised since we need a greater amount of overlap to judge whether two reads should be connected into an assembly. If the reads are too short (e.g. k = 25), it is difficult to tell where a read came from since the reads might not contain enough information.

(Multiple choice) When you increase the length of *k*-mers, the de Bruijn graph ___________.

Justify your answer.

A) Becomes more tangled.

B) Contains more nodes.

C) Becomes less tangled.

D) Remains the same.

C. As we increase the length of k-mers, the de Bruijn graph would become less tangled since there would be fewer nodes and fewer possibility of interactions as each read contains more information and repeats have less effects.

You will use the Quality Assessment Tool for Genome Assembly **QUAST** (Gurevich *et al*, 2013) to evaluate the quality of your assembly using the Staph reference genome as the gold standard.

- Download the contigs.fasta file as part of the SPAdes output from the best assembly you chose for question #8 above.
- Go to QUAST (http://quast.bioinf.spbau.ru/) and upload your contigs.fasta file with the “Add files” button.
- Leave the “Scaffolds” and “Find genes” boxes unchecked and keep the indicator on “Prokaryotic.”

- Click on the “Another genome” link underneath “Genome.” Fill in a name and upload the staph_genome.fasta file that we provided for the “Reference” file. (Note: we provide this file as a .txt, you will need to save it as .fasta). Leave the other two inputs (“Genes” and “Operons”) blank and click “Evaluate.”
- A link to the report should appear on the right side of the page in a few moments.

1. How many misassemblies were there?

2. How significant is the effect of misassemblies on the resulting assembly?

1. There are 33 misassemblies.
2. Misassemblies have very serious effect on the resulting assembly since most contigs are misassembled (there are only 45 long contigs) and required to be broken into pieces.

1. What are NG50 and NGA50?

2. How do they compare with the value of N50 that you previously calculated? Why?

1. NG50 is 154027; NGA50 is 87128.
2. NG50 is same as the N50 value I previously calculated; NGA50 is around half of N50. The contigs cover the entire genome; so the NG50 is same as the N50. There are many misassemblies; so NGA50 is much smaller.

What is the known species of *Staphylococcus* that is most similar to the species that you assembled?

Staphylococcus aureus.

Edit submission

Comments

Visible to classmates

share your thoughts...

👍 👎 ⚑