Tareas calificadas por los compañeros: Bioinformatics Application Challenge

Se revisó antes del August 8, 11:59 PM PDT

Revisiones Faltan completar 5

Parece que esta es tu primera tarea calificada por los compañeros. Obtener más informacion

Sequencing project

por Melinda Varga

August 1, 2018

me gusta Señalar este envío.

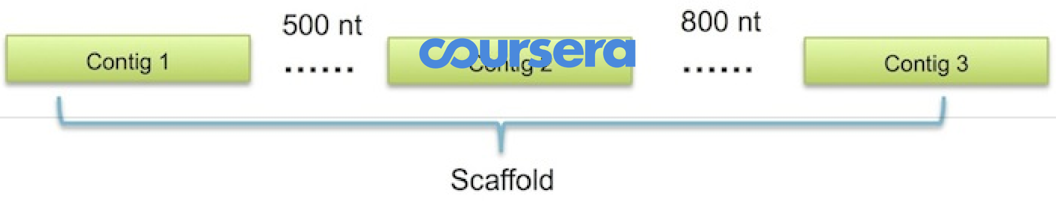
CUADRO DE AVISO

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 *contiguous* 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 missassembly 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.

The largest such contig that encapsulates together with it's larger counterparts at least 75% of the total contig length (sum of all contigs).

RÚBRICA

N75 is the maximal contig length for which all contigs of at least that length comprise at least 75% of the sum of the lengths of all contigs.

○ puntos
The learner's response does not match the one given above.

○ 2 puntos
The learner gives a reasonable response but neglects to include "maximal"

● 3 puntos
The learner's response more or less matches the one given above.

CUADRO DE AVISO

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

RÚBRICA

- N50 = 100
- N75 = 60

○ puntos
The learner answers neither part correctly.

○ 2 puntos
The learner answers only one part correctly.

● 4 puntos
The learner answers both parts correctly.

CUADRO DE AVISO

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

NG50 = 60

RÚBRICA

NG50 = 60.

○ puntos
The learner answers anything other than 60.

○ 2 puntos
The learner answers 60.

CUADRO DE AVISO

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

RÚBRICA

NGA50 = 50.

○ puntos
The learner answers anything other than 50.

○ 2 puntos
The learner answers 50.

CUADRO DE AVISO

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

The most useful information would be the knowledge of the reference genome. That way we could pinpoint where the contigs are located and how long are the missing parts between them. But in these cases that is not given, so what could be done is to find an Eulerian path in the graph and give that, in the case of not connected graphs one could use biological information to try to figure out which path goes after which other and how far are they located.

RÚBRICA

There are several reasonable answers. Three are given below.

- Additional long reads could be generated in an attempt to find reads that bridge the gaps in contigs. In other words, if we find a long read that begins at the end of contig A, and ends at the beginning of contig B, then we can conclude that the read extends across the gap between the contigs.
- Contigs could be compared against a "reference genome", i.e., a complete genome sequenced from the same species (often at greater cost). The order of the contigs in the reference genome would indicate the order of the contigs in the desired scaffold.
- Information from read-pairs could be used. In particular, if the first read in a read-pair maps to contig A, and the second read in a read-pair maps to contig B, and we know the distance between the paired reads, then we can infer the distance between contigs A and B. By gathering this information for different pairs of contigs, we may be able to infer distances between contigs and therefore their ordering with respect to each other.

○ puntos
The learner does not provide an answer or provides an answer that very closely matches (part of) the text above.

○ 1 punto
The learner provides an answer without justification.

○ 2 puntos
The learner provides an answer with limited justification, or justifies an answer that would not help construct scaffolds.

○ 3 puntos
The learner provides a reasonable answer with justification.

CUADRO DE AVISO

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			

k = 25: N50 = 40658, #long contigs = 123, total length of long contigs = 2794734
k = 55: N50 = 154027, #long contigs = 45, total length of long contigs = 2819863
k = 85: N50 = 79093, #long contigs = 71, total length of long contigs = 2829100

RÚBRICA

1 point for each of the following table values.

k	**N50**	**#long contigs**	**total length of long contigs**
25	40,658	123	2,794,734
55	154,027	45	2,819,863
85	79,093	71	2,829,100

○ puntos
The learner gets no values in the table correct.

○ 1 punto
The learner gets one value in the table correct.

○ 2 puntos
The learner gets two values in the table correct.

○ 3 puntos
The learner gets three values in the table correct.

○ 4 puntos
The learner gets four values in the table correct.

○ 5 puntos
The learner gets five values in the table correct.

○ 6 puntos
The learner gets six values in the table correct.

○ 7 puntos
The learner gets seven values in the table correct.

○ 8 puntos
The learner gets eight values in the table correct.

○ 9 puntos
The learner gets nine values in the table correct.

CUADRO DE AVISO

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?

The k = 55 case performed the best overall. Both for N50 and for the #long contigs it gave the best results. In the case of the total length of contigs the k = 85 gave the highest result but it is only slightly better than the one given for k = 55. I think the k = 55 is the best since if you construct the de Brujin graph using short k-mers it becomes very tangled, which results in many contigs, but few very long ones. If the k-mers are very long we might not have a perfect k-mer coverage and because of that we will get more contigs. There is a k somewhere in the middle that is ideal. In our case, among the k values we tried k=55 is the best.

RÚBRICA

The total length of long contigs is about the same for all three values of k. Accordingly, we conclude that the assembly using k = 55 performed the best because it has a substantially larger value of N50 and many fewer contigs than the other two values of k.

k = 55 performs the best likely because the read length is "just right". If the reads are too short (k = 25), then the reads contain too little information, and repeats may make it difficult to identify where a read came from. If the reads are too short (k = 85), then it is easier to identify where a read came from, but at the same time, we may lose coverage, i.e., require a greater amount of overlap to connect two reads into an assembly, making contigs shorter.

Please use the grading scheme below:

4 points available

1 point for identifying k = 55;

1 point for a reasonable justification of why k = 55 was the best choice according to statistics;

2 points for attempting a reasonable explanation of why k = 55 wound up being the best value of k.

○ puntos
(see point distribution above)

○ 1 punto
(see point distribution above)

○ 2 puntos
(see point distribution above)

○ 3 puntos
(see point distribution above)

○ 4 puntos
(see point distribution above)

CUADRO DE AVISO

(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) By increasing the length of the k-mers we will have more diverse node labels, so there will not be as many repeats, which causes less gluing, hence less tangles.

The correct answer is C).

We saw in the class text (and lecture) that increasing the value of *k* used to generate *k*-mer reads led to a less tangled de Bruijn graph because the larger the value of *k*, the greater the amount of information contained in our reads, and the lesser the effects of repeats. More details are available in the course text.

4 points available

2 points for correctly identifying C).

2 point for providing a reasonable justification of why the de Bruijn graph becomes less tangled with increasing *k* because of fewer repeats.

○ puntos
(see point distribution above)

○ 2 puntos
(see point distribution above)

○ 4 puntos
(see point distribution above)

CUADRO DE AVISO

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?

There were 32 misassemblies. The size of the largest contig significantly decreased (from 288 395 to 187 456) and the misassembled contigs length is 2274891, which looks to be very huge, but we shouldn't forget that the longest contigs are the most probable to be misassembled, and if we add up the total length of these we will get a huge number. But based on the number of misassemblies and that the resulting assembly captures 90.847% of the original genome I would say that the assembler did a good job.

RÚBRICA

1. There were 33 misassemblies. (32 and 59 is also an acceptable answer depending on the version used.)

2. For *k* = 55 there were only 45 long contigs, meaning that the misassemblies are likely causing most of the long contigs to have been broken into pieces.

3 points available

1 point: correctly identifying the number of misassemblies

2 points: identifying that the number of misassemblies is significant and giving a reasonable explanation

○ puntos
(see point distribution above)

○ 1 punto
(see point distribution above)

○ 2 puntos
(see point distribution above)

○ 3 puntos
(see point distribution above)

1. What are NG50 and NGA50?

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

NG50 = 154027, NGA50 = 87161. The NG50 has the same value as the N50, since the assembled genome captures 90.847% of the original, and apparently the contig with length 154027 and the ones longer than this add up to a little more than half of the assembled genome so they can capture the half of the original genome also which is only 257459 longer than the assembled one. The value of NGA50 is significantly lower that N50, and that is understandable. The long contigs are the most prone to be cut due to misassemblies.

RÚBRICA

(1 point): NG50 = 154,027

(1 point): NGA50 = 87,161.

(1 point): NG50 corresponds to the value of N50 that was previously obtained, because the contigs generated cover the entire genome.

(1 point): However, NGA50 is about half as large as N50 because of the effects of misassemblies.

○ puntos
(see point distribution above)

○ 1 punto
(see point distribution above)

○ 2 puntos
(see point distribution above)

○ 3 puntos
(see point distribution above)

○ 4 puntos

(see point distribution above)

CUADRO DE AVISO

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

Staphylococcus aureus. It is mentioned in the provided genome, and I also checked it with BLAST using the assembled genome for k = 55.

RÚBRICA

Staphylococcus aureus (can be found in the output produced during the SRA import step).

○ 2 puntos
Grade all responses as correct.

RÚBRICA GENERAL DE TAREA

Optional: Please provide any additional general feedback that you would like to give here.

(This assignment is ungraded, but you need to enter something in this field. If you don't want to provide feedback, it can be just few white spaces or other symbols)

Enviar revisión

Comentarios

Solo el estudiante puede ver comentarios que se dejan para ese estudiante y la persona que dejó el comentario.

share your thoughts...