A toolkit for reporting on metagenome assembly quality

James Edward Euesden
Department of Computer Science, Aberystwyth University, Aberystwyth, Ceredigion

Project Information

The study of metagenomics is taking samples from environmental areas rich in species/sub-species and trying to find interesting genomes within them.
These samples may contain hundreds to thousands of different species, and conventional genome sequencing and assembly tools for single species analysis do not necessarily work for metagenome samples. Considering we often do not know what is within the sample taken to begin with, how then do we determine if our sample is of good quality?

What is quality in metagenomics?

If we don't know what our samples are, understanding what quality is can be a tricky task. We know that we want to find genomes that exist in nature, and that when we get an assembly from our assembler, we would like to know it is not a chimera, formed of many different species that would never actually exist.
It is also relevant to look at the size of the assemblies produced, are they too small or too big compared to the reads that were used to construct them?

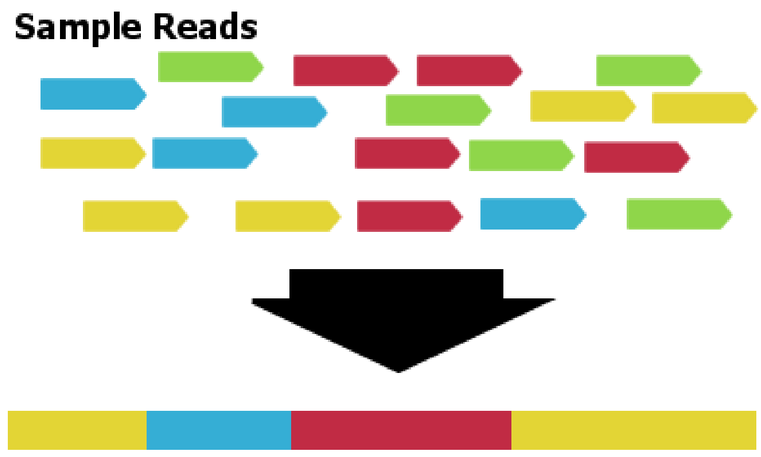

Figure 1. Demonstrating how multiple species in a sample could result in a very mixed assembly, where we don't know what it contains or if it could exist in nature and be of good quality.

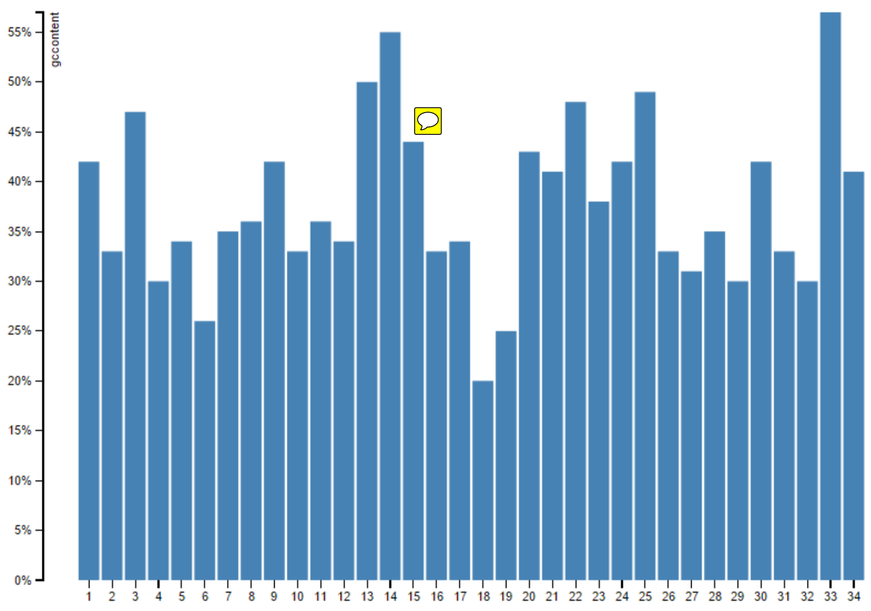

Figure 2. An example of what the GC content chat looks like from the prototype with Plotly. In future revisions, the chart will display potential ORFs to match up, or not, with the GC content displayed.

Project aims and progress

The goal of this project is to produce a software application to be used as a toolkit for helping researchers determine whether their metagenome assemblies are of good or bad quality, and attempt to indicate to them where there may be problems in the assemblies in a textual and graphical report.

My progress on the application has taken me through learning about metagenomics, what is considered as quality and how is it found, and what technologies I should use to build it. The application itself is in a prototype stage, where it will consider the GC Content of any number of assemblies provided in a FASTA file type and output the percentages based on window sizing (specified by the user). There is another prototype that allows me to display the outputs of the processing in a browser using Plotly.

Technical Information

The software application is being built in Java, using Spring boot to display the report outputs in a browser with Javascript and Plotly. For sample test data, I have been using an assembly taken from a limpet gut (provided by Sam Nicholls of Aberystwyth University) and creating my own metagenome assemblies by combining single species assemblies from those in the NCBI RefSeq database.
For reporting on the quality, the application will mostly look at GC Content within the assembly, k-mer counting and n50 (sizing of the assemblies provided). These are well documented and used techniques for helping determine quality of assemblies, although often for single species and not often brought together in one place.

Remaining Work

Since the application is currently only a prototype and dealing with GC content, the next stage is to first tie this in with Open Reading Frames to detect where there might be genome encoding regions causing discrepancies in the GC content and not due to the presence of a chimera, and then to start work on the full application connecting the report and the processing.
Once this basic beginning is done, this will then allow me to begin adding the k-mer counting and processing of the size of the assemblies with the read sizes.
Tying in with these, there is a lot of work to be done on displaying a useful report to the user to indicate the results of these techniques and whether it indicates that the provided assembly is of reasonable quality or not and why.
Future work may involve checking to see if any genome in the NCBI RefSeq database matches the assembly provided.

Further Reading

GC Content - http://medical-dictionary.thefreedictionary.com/GC+content
NCBI RefSeq - http://www.ncbi.nlm.nih.gov/refseq/
K-mer Counting - http://www.homolog.us/Tutorials/index.php?p=3.7&s=2
Open Reading Frame - http://bioweb.uwlax.edu/genweb/molecular/seq_anal/translation/translation.html
FASTA format - http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml