Building an online resource for Candida Tropicalis

Report Name	Outline Project Specification
Author (User Id)	Owen Garland (owg1)
Supervisor (User Id)	Wayne Aubrey (waa2)
Module	CS39440
Degree Scheme	G600 (Computer Science)
Date	February 9, 2017
Revision	0.1
Status	Draft

1 Project description

Candida Tropicalis is a yeast, that has had its genome sequenced by [their name]. The researchers are hoping to edit some of the genes in the yeast to enable it to ferment sugars found in grass into xylan, that would then be able to create xylitol; which can be used in the manufacture of many products including chewing gum.

The data is currently in the form of the assembled contigs (strands of sequenced DNA). This is useful for the work they are doing, but it means that when they want to inspect a gene they need to analyse the data and annotate for each individual gene. The main aim for this project will be to produce an annotated data set for the entire genome, that will make their research easier, and make the data more accessible to other researchers around the world.

This report is due in on Friday the 10th of February, which is unfortunately the same day I am meeting with the researchers to discuss what they would like to see out of the project in more detail.

2 Proposed tasks

There are several steps involved in annotating the genome for *Candida Tropicalis*, the first stage will be to find the open reading frames (ORF's), these are the beginning and end markers for a gene. The start of a gene is marked with an ATG sequence, and ended with a TAA or TGA sequence. There have been tools developed to perform this task, the most develop too appears to be OrfM [2]. This tool will produce data that can then be annotated using blast.

Blast will allow us to compare the genes found in *Candida Tropicalis* with a reference database of an already annotated genome. From there it will produce a set of data that has the genomes labelled with the matching genes from the reference database.

Once the genome has been annotated I will be putting the data into a database, this might be Chado [1] which is an established schema for storing genetic data, it would be interesting to explore the possibilities of using a NoSQL database such as MongoDB for the project.

Once the data is in a database, the next stage will be developing a web application that allows users to browse the genome.

3 Project deliverables

Annotated Bibliography

[1] Various, "Chado manual," http://gmod.org/wiki/Chado_Manual, Feb. 2017, accessed Feb 9th 2017.

The official manual for Chado

[2] B. J. Woodcroft, "Orfm repository," https://github.com/wwood/OrfM, Feb. 2017, accessed Feb 9th 2017.

The Github Repository for the OrfM tool