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 (Software Engineering)
Date	February 15, 2017
Revision	0.1
Status	Draft

1 Project description

Candida Tropicalis is a yeast, that has had its genome sequenced by researchers in Aberystwyth. 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.

To do this they plan to compare the genes found in Tropicalis with two other similar species of yeast, to see where they differ and to see if they can be engineered to produce these sugars.

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 dataset for each of the three species of yeast and then compare them them to reveal the differences.

Once this is done I can then work on making this data more accessible to the researchers by building a web resource that will allow them to browse and compare the found genes. Eventually this resource would be made public for others to see.

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.

Diamond 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. For this I will be using NodeJS, as I won't have much time left to build the website I need to use tools that I am familiar with and have a lot of boilerplate written for me already.

3 Project deliverables

At the end of this project I will aim to provide the following deliverables:

- Annotated genome of *Candida Tropicalis*
- Database containing the annotated data
- Web accessible gene browser

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