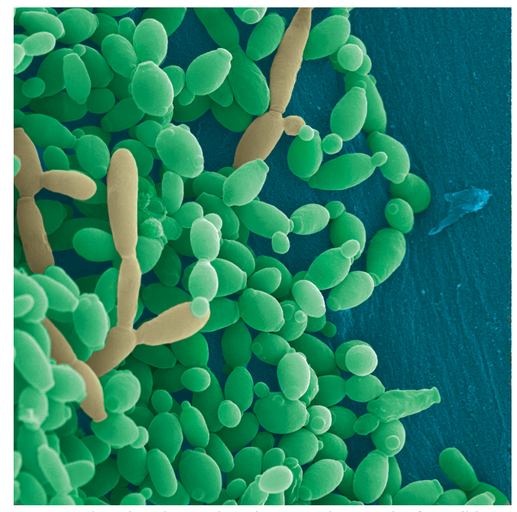
Building an Online Resource for *Candida tropicalis*

PRIFYSGOL ABERYSTWYTH UNIVERSITY
Owen Garland | owg1@aber.ac.uk
http://project.owen.cymru

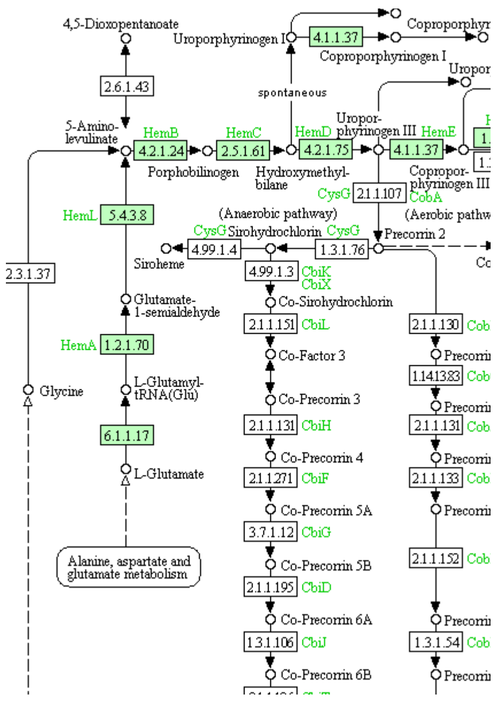
What is this project about?

Researchers at Aberystwyth University have been studying three species of yeast, *Candida tropicalis*, *Candida boidinii*, and *Candida shehatae*. Xylose is a five carbon sugar ubiquitously found in plants such as grass, which can be used as a feedstock for industrial biotechnology. *Candida tropicalis* is able to convert arabinose & xylose into arabitol and xylitol repectively. Xylitol is a commercially valuable anti-bacterial foodgrade sugar use in the maufacture of chewing gum. However arabitol and xylitol cannot be easily separated. Candida boidinii cannot metabolise arabinose, for an unknown reason, but does metabolise Xylose but at a much slower rate which is commercially unviable.

Understanding at a genomic level why *Candida boidinii* is unable to utilise arabinose and genetically modifying *Candida tropicalis* to have the same phenotype, may enable *Candida Topicalis* to exclusively produce xylitol and not arabitol. One other exciting possibility is using it as a low glycaemic index table sugar that can be used by diabetics.



2. Pseudo-colored scanning electron micrograph of Candida tropicalis YC466 displaying both yeast and pseudohyphae.



3. Example section of KEGG pathway diagram

The story so far

The researchers have sequenced and assembled contigs (sections) of DNA from the three yeast species. The next stage in processing this data was to align it with known gene sequences from Genbank. For this diamond (software) was used to calculate which sequences aligned with the NCBI non-redundant database of genes.

With the datasets aligned it is possible to retrieve information about the genes preset by finding the IDs in the Uniprot database. Uniprot stores all the data about the gene including what sequence of amino acids used to create the protein and it's function. Also available are the KEGG IDs, these link the gene to a known chemical reaction pathway that describes what the protein does at a molecular level.

Once the various IDs and sequences had been colated into one source dataset, the data needed to be indexed into a database, for easy querying and access from the website. Traditional bioinformatics tools are very often written by biologists and not computer scientists, as well as being long standing and monolothinc projects using technology that has since been outdated. As a computer scientist it seemed clear to me that a more modern approach could be quite beneficial, which is why a NoSQL database was chosen to store the data. Research had shown that there wasn't a NoSQL solution in existence, but quite a wide array of Perl based MySQL solutions, which gave me confidence to try a more modern solution with MongoDB. This appears to be an apporpriate use case for NoSQL as once the data is in the database it will become read only, which means there isn't a need to maintain relations with different peices of data that may change in the future. It will also save time on development as the data is stored in a feature driven format, meaning that when accessing data from the website all of the relevant information will be returned in one simple query.

Website

With the data in the MongoDB database, it is now possible to build a website that allows researchers to interact with the data in a meaningful way. To develop this site I will be using NodeJS, Express & Webpack along with a several other small libraries such as Pug Babel, Mocha and Stylus. NodeJS is a natural choice for a website for a MongoDB database as MongoDB represents its data in JSON (Javascript Object Notation), which is natively how NodeJS handles objects. In addition to this it has a very large library of modules that are easy to install and manage with the NPM repository. This became immediately useful when writing the scripts to import the fasta files into the database as I was able to utilise a `fasta2json` module to quickly read the files in to a format that NodeJS can understand.

In addition to the stack of technologies being used on the back end, I am also utilising a lot of third party services such as Travis CI to manage the continuous integration, and Heroku to automatically deploy working code. This allows me to push changes to github and then have the researchers immediately see the adaptations to the site, making the feedback loop as small as possible.

The sites functionality is still rather undefined, at the time of writing it is able to simply search the exisiting dataset directly, which will be the core functionality, however there are a still a lot of questions about how the reasearchers will interact with the site and what data is the most useful for them to see and in what format.

Demo site - project.owen.cymru


```
    "data": [
        {
            "species": "shehatae",
            "contig": "C359541  2.0",
            "blast": "XP_004983205.1  91.9  37  3 0 150 40  310 346 2.6e-10 72.8",
            "uniprot": "K4AC16",
            "kegg": "sita:101760397",
            "sequence": "TGAGGATCCTAACAATCTAGTAAGGCTTCGATTCTAAGTTATCTGCACCATCGCCTCGTCACCGAGGAACTCCACGATGGACAGCGCCACGGGCGCCTGCTCCGGTCGCGCCCTGATGCCGTCCAGCGCCGCGGCGTCCCCTCCGAC
CCACC",
            "protein": {
                "head": "tr|K4AC16|K4AC16_SETIT Uncharacterized protein OS=Setaria italica GN=SETIT_036423mg PE=4 SV=1",
                "seq": "MNIASAALVFLAHCLLLHRCMGSEAGGVFDHGRHGVSLVRVEAPSRCGGGTPSSPPGADTPPPKPLLVAAPREAGEYPVLVFLHGYLVVNSFYSQLLQHVASHGFIVVAPQLYTISGADATEEINAAAAVIGWLAAGGLSSALPPGVR
ADATKVSVSGHSRGGKVAFALALGHAKLAIPLAALVAVDPVDGMGMGRQTPPPILTGRSGALRVSAPAMVIGTGLGELPRGPLLPPCAPRGVSHAAFCDEMDPAAASACHLVARDYGHTDMMDDDTPGARGILTRAICRSGGARAPMRRFVGGATVAFLKRWVGGDGAALDG
IRARPEQAPVALSVVEFLGDEAMAQIA"
            }
        }
    ],
```


An example set of data for a contig represented in MongoDB

Future of the project

The next stages of this project will be to verify that the pipeline that was used to annotate the genes is correct and the data produced is usable by the researchers. Once this is established, the focus can be on making the data as accessible and as useful to them as possible, whether that be by tweaking way that a user can interact with the data on the website or by adding additional data that they may find useful. The next big hurdle will be providing a way for the exact gene sequence to be extracted from the contig it was found in, by the diamond results, as well as the bases of DNA surrounding that sequence.

One addition that is highly saught after mRNA selector sequences for CRISPR targeted gene editing. If it is in scope for the project then calculating these sequences and making it readily availble via the website would be a huge productivity boost for the project as a whole as the time taken to edit a gene will be greatly reduced.

Another area that would be beneficial to explore would be automating and standardising the pipeline that was used. This would allow other researchers to annotate their assembled DNA and quickly provide a way to access the data via the website. For this project the main focus is on providing the researchers in Aberystwyth with the tools they need to complete their project, however it would be of great benefit to the bioinformatics community as a whole if this tooling existed.

Images

1. https://en.wikipedia.org/wiki/File:Aberystwyth_University_logo.svg
2. https://de.wikipedia.org/wiki/Candida_tropicalis#/media/File:C_tropicalis_YC466.png
3. http://www.genome.jp/kegg-bin/show_pathway?sita00860+101760397
4. https://en.wikipedia.org/wiki/File:MongoDB-Logo.svg
5. https://commons.wikimedia.org/wiki/File:Node.js_logo_2015.svg
6. http://www.deardoctor.com/images/ddwc/consultations/xylitol-in-chewing-gum/xylitol-chewing-gum.jpg