Michael Henson

BIOL 7800

The study of microorganisms and their roles in remediation and biogeochemical cycling is fundamental to both understanding the environment and providing predictive capability regarding the impact of future disturbances1–3. Such studies require the observation of microbial communities and genetics in nature coupled with experimental testing of hypotheses both *in situ* and in laboratory settings, the latter of which is best accomplished by cultivation of microorganisms1,2,4,5. The difficulty in cultivating many microbial taxa vexes researchers intent on understanding the contributions of these organisms in natural systems, particularly when these organisms are numerically abundant. Efforts to improve this conundrum with marine bacterioplankton have been successful with natural seawater media, but that approach suffers from a number of drawbacks and there have been no comparable artificial alternatives created in the laboratory5–7. Our work in the Thrash lab has developed a suite of artificial seawater media that can successfully cultivate many of the most abundant taxa from seawater samples, and many taxa previously only cultivated with natural seawater media. This methodology significantly simplifies efforts to cultivate bacterioplankton.

As a result, numerous bacteria isolates (average experiment results in 15-300 isolates) require identification using the 16S rRNA gene. Isolate 16S rRNA genes are amplified and sequenced using Sanger sequencing. The sequences are curated by the Finch software (Geospiza Finch Suite Distribution v 2.21.0) and forward and reverse sequences are assembled. To assemble, when sufficient overlap exists, the forward and reverse sequences are placed into the CAP3 webserver (http://doua.prabi.fr/software/cap3), after the conversion of the reverse read to its reverse complement at <http://www.bioinformatics.org/sms/rev_comp.html>. Each individual isolate must be moved around in text files, checked for errors, and then individually blasted in the NCBI nr database for taxonomic identification. This process is cumbersome and can take hours to manually go through for each isolate. Therefore, for the BIO7800 Bioprogramming project, I propose to create a program that allows you to implement this process using a single python script that connects these programs (e.g. Cap3, NCBI BLAST) into one user friendly process.

The program, BRA or **B**acterial **R**ead **A**ssembler, will allow a user to input cleaned forward and reverse reads generated using Sanger sequencing and be provided an output of the individual forward (F) read, reverse (R) read, reverse compliment (RC) read, assembled contig, and taxonomic information. These process will require the incorporating of already existing programs: BioPython (found here http://biopython.org/wiki/Main\_Page), BLAST+ API (found here http://ncbi.github.io/cxx-toolkit/pages/ch\_blast), and others. BRA will first take user input in the form of two .fasta (or .fa) files. Once the files have been read in, the program will use the R read file and create a file of RC reads. This can be done using BioPython and implementing the seq.reverse\_complement(). Using the newly created RC reads and the F reads, the files will be shuttled to a function that allows them to be aligned and then formed into a single contig. Preferable this will be done using Cap3, a sequence assembler, but, at this moment, no single tool has been found that allows you to incorporate Cap3 into a python script. However, there are many alignment tools and potential some code freely available that can do this. Any outside code will need to be highly scrutinized to ensure the alignment and contig are properly constructed. For tracking purposes, I would also like the program to take each sequence within the fasta files (F, R, RC, and contig) and output them into individual .fasta files. I would like the program to also read in fastq Sanger files (combined .fasta and .abi files). This will allow the program to read in Phred scores with the sequencing file. This will add the benefit of allowing the assembly program to consider mismatches if Phred scores are over a minimum quality. BRA will, therefore, need to incorporate a fasta to fastq function within the program. A quick search shows potential starting information here: <https://www.biostars.org/p/99886/>. Lastly, using the contigs and NCBI BLAST, I will search the contigs of the individual isolates against the nucleotide collection (nt/nr) using BLASTn. Because this is a large database, I may also search it against the reference genomes database (refseq\_genomic) to help get a better idea to how closely related the isolate is to known cultured representatives. However, because of the potential novelty of individual isolates, both outputs would be useful. To help curate the BLAST output, I would employ the NCBI BLAST API and also restrict the results to a minimum threshold for max-score, identity, and e-scores. If time allows, I would like to return these results, and provide a user-friendly output of the potential taxonomic identification of the isolates. This can be done using Bokeh (<http://bokeh.pydata.org/en/latest/>) or ggplot (<http://ggplot.yhathq.com/>).

I believe this program, as outlined, would be of high use for microbial researchers. Sanger sequencing is employed by researchers around the world in both medical and environmental as well as private industry and academia. Many times researchers use paid for programs (e.g. Sequencher) to do some of this analysis. As of now, no freely available script allows for scientists to readily assemble Sanger sequences and provide easily accessibly taxonomic information. As mentioned above, the process can be time consuming and result in time loss. Therefore, the program would be of high interest. To help make it freely available, the program will be hosted on my github page (https://github.com/Hensonmw) and will also have documentation so that it can be distributed. If the program is of high enough quality, passing many security checks, then the program could also be hosted on the Thrash lab website.

**Literature cited**

1. Rappé, M. S. Stabilizing the foundation of the house that ‘omics builds: the evolving value of cultured isolates to marine microbiology. *Curr. Opin. Microbiol.* **16,** 618–624 (2013).

2. Giovannoni, S. & Stingl, U. The importance of culturing bacterioplankton in the ‘omics’ age. *Nat. Rev. Microbiol.* **5,** 820–826 (2007).

3. Azam, F. *et al.* The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* **10,** 257–264 (1983).

4. Connon, S. a. *et al.* Bacterial community composition determined by culture-independent and -dependent methods during propane-stimulated bioremediation in trichloroethene-contaminated groundwater. *Environ. Microbiol.* **7,** 165–178 (2005).

5. Connon, S. a. & Giovannoni, S. J. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Appl. Environ. Microbiol.* **68,** 3878–3885 (2002).

6. Rappé, M. S., Connon, S. a, Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418,** 630–633 (2002).

7. Schut, F. *et al.* Isolation of typical marine bacteria by dilution culture: Growth, maintenance, and characteristics of isolates under laboratory conditions. *Appl. Environ. Microbiol.* **59,** 2150–2160 (1993).