SIO 242C: Marine Biotechnology III
Introduction to Bioinformatics
SPRING QUARTER 2014

Instructor: Professor Terry Gaasterland, Ph.D.
Room: Vaughn, 348 Time: Tu 9:30a – 12:30p

Syllabus

OVERVIEW: The course is directed to students who want to understand and master use of bioinformatics tools for analysis of next generation sequence data to measure protein coding and microRNA gene expression, DNA-protein binding sites, composition of microbial communities based on RNA or genomic DNA, and assembly of a draft genome or transcriptome for a new organism. Grading will be based on participation and effort, including weekly homework exercises. Students will learn how to use high throughput sequence data to interrogate a biological system and plan follow-up analyses through bioinformatics and experimentation.

DATASETS: Students will learn bioinformatics priniciples through the analysis of five datasets. All are next generation sequence data generated on the Illumina Hiseq instrument. The datasets will be one each of RNA-seq, microRNA-seq, ChIP-seq, metagenomics DNA-seq, and a new model organism's genome. We will examine at a signaling pathway through RNA, microRNA, and “ChIP” sequencing. We will examine a microbial community represented in DNA and RNA gathered from the gut of the Methane Iceworm. We will assemble reads into contiguous sequences using software tools for quality trimming, removal of duplicates, and De-Bruijn graph assembly. As a bonus dataset, students will have access to DNA and RNA reads from Pacific Biosciences and from Illumina for the Mexican salamander, an organism with a genome estimated at 15 gigabases.

COMPUTE PLATFORM: The data will be loaded into a shared disk space on the Triton Shared Compute Cluster. In the first week of class, students will need to request accounts on TSCC. Each account will be given "group-read" permission for the "gaasterland-lab" so students can access the datasets and a suite of open-source bioinformatics tools maintained by the instructor and administered by staff at the San Diego Supercomputer Center. Students will be responsible for requesting a sufficient number of “server units” to execute homework exercises.

APPROACH: Lectures will start with basics in Unix, open source software, and script writing, and then move into data analysis. As students review and learn basics, they will start manipulating data in short homework exercises.

TOPICS:

1. Navigating the Unix command line.
2. What is "open source" software? What does open source mean to the bioinformatics community? How are open source programs installed under Unix?
3. What is TSCC? What is a "job queue"? What is "qsub"? What is the difference between distributed and parallel computing?
4. How does one analyze and manipulate hundreds or thousands of sequences at a time? (We will compare shell scripting, Decypher TimeLogic cards, and off-the-shelf vendor platforms)
5. Analyze RNAseq data (Intro to tools; limitations and pitfalls; do it)
6. Analyze microRNAseq data (Intro to tools; limitations and pitfalls; do it)
7. Analyze ChIPseq data (Intro to tools; limitations and pitfalls; do it)
8. Analyze RNA and/or DNA from a metagenome (Intro to tools; limitations and pitfalls; do it)
9. Analyze RNA and/or DNA from a new genome (Intro to tools; limitations and pitfalls; do it)
10. Bonus dataset: PacBio and Illumina reads from RNA and DNA in Axolotl.

LANGUAGES: Lectures will cover computational techniques and programing principles in Unix and Perl. Students who prefer to do homework exercises in Python or another language may do so.