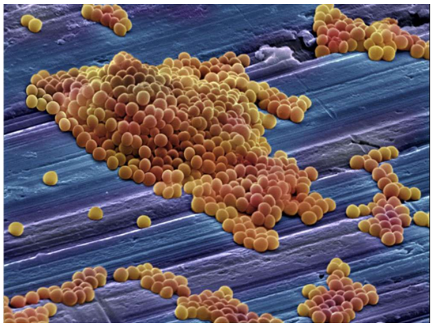
Taxonomic classification of metagenomic samples

Sponsor Information:
Dr. Viacheslav Fofanov
School of Informatics, Computing, and Cyber Systems
Northern Arizona University
Viacheslav.fofanov@nau.edu

Project Description

Microorganism communities play a critical role in virtually all of the environments on the planet, occupying niches ranging from the hostile undersea methane vents to the extensively colonized human gut. Answering the “what is there?” question is an essential step in understanding the processes by which viruses, bacteria, and fungi affect their environment.

The latest generation of High Throughput Sequencing machines is capable of producing billions of pieces of genomic sequence data (trillions of nucleotides) daily. These Big Data capabilities, in theory, allow for simultaneous characterization of all viruses, bacteria, fungi, and eukaryotes present in a given sample: one simply extracts all of the DNA in the sample; runs all of this DNA through the sequencer; and then analyzes the resulting sequences, comparing them against large databases of sequences compiled for known organisms. This then can tell you which of those organisms are present in the sample.



Methicillin
Resistant
Staphylococcus
Aureus
(isolated from human skin)

While highly sensitive, such shotgun metagenomic approaches tend to be extremely computationally expensive and require significant investments in CPU time to analyze (weeks of highly distributed computations). The computational tools to process such sequence data continue to rapidly evolve, with many laboratories (including sponsor’s laboratory - the Fofanov Bioinformatics Lab) developing new tools to dramatically improve the speed and accuracy of shotgun sequence data search and alignment algorithms. This rapid innovation constitutes both an opportunity for improvement and a challenge of relevance, with many computational tools becoming outdated shortly after their introduction.

The goal of this project would be create ***a custom-built pipeline management tool*** for metagenomic sequence data analysis, which would improve performance by reusing previously computed components of the data (thus optimizing time-to-completion) and ensure continued relevance of the pipeline by allowing removal and replacement of individual pipeline modules based on new developments in High Throughput Sequence data analysis field.

In the course of the project, students will be able to work in ‘real-world’ Big Data settings with multi-terabyte size datasets used to both prototype and test the pipeline. The management tool will be developed in C/C++ for computationally heavy components, with python used to enable efficient and flexible movement of data between modules. Some requirements will include:

- This tool must minimize computing time and RAM footprint while maintaining accuracy
- Must be modular to allow easy future extensions and additions.
- Must be well documented, both in terms of design rationale and from a technical standpoint.
- Must be easily deployable through GitHub, to allow for both easy sharing and future extension.

Knowledge, skills and expertise required for this project

- In-depth understanding of data-structures: hashes, trees, arrays, linked lists.
- UNIX shell familiarity
- C/C++ and Python programming

Equipment Requirements

- Access to unix-bases C and python compilers
- Access to computing cluster (provided by sponsor)

Deliverables:

- Python / C++ pipeline to implement metagenomic analysis
- Complete user manual for scientific users, including test cases.
- A strong as-built report detailing the design and implementation of the product in a complete, clear and professional manner. This document should provide a strong basis for future development of the product.
- Professionally documented (and tested) source code, in an online repository (Github, Bitbucket), as well as archived and delivered on a thumb drive.