Taxonomic classification of metagenomic samples

Sponsor Information:

Dr. Viacheslav Fofanov
Informatics and Computing Program
Northern Arizona University
Viacheslav.fofanov@nau.edu

Project Description

Microorganism communities play a critical role in essentially all 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. While highly sensitive, such shotgun metagenomic approaches tend to be extremely computationally expensive and tend to produce too many false positives, severely overestimating the type and number of species present. By quantifying the statistical confidence with which each piece of genomic data can be used to identify a given microorganism, the Fofanov Bioinformatics Lab has developed an information theory based approach to dramatically improve the accuracy of shotgun metagenomics approaches. While accurate, this approach is not computationally efficient, largely due to its reliance on external, third-party tools and software packages which were not purpose built for this type of analysis.

The goal of this project would be to significantly improve the speed and resource footprints of the existing methodology by creating a custom-build pipeline/tool for metagenomic sequence data analysis. 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 pipeline will be developed in C/C++ for computationally heavy components, with python used to enable efficient and flexible movement of data between modules. This application will need to (1) minimize computing time and RAM footprint while maintaining accuracy, (2) be modular and well documented to enable future additions, and (3) be easily deployable through GitHub.

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
- Documentation and test cases