**Investigators**

Andrew Warren (*anwarren@vt.edu*)

Timothy Driscoll (*driscoll451@gmail.com*)

**Title**

Visual exploration of global metrics for prokaryotic replicons.

**Background**

Next-generation sequencing methods have led to a dramatic increase in the available number of fully sequenced prokaryotic genomes, a trend that is sure to increase into the near future. Standard methods for comparing multiple genomes are classically linear, and focus on individual genes (or, at best, small handfuls of related genes). In contrast, there is much information to be learned from a system level genome comparison. Fortunately, there exist a number of useful metrics for describing genomes, including gene density, GC content, promoter density, and more.

We propose to build a discovery-based software tool that allows researchers to visualize a novel genome sequence relative to a landscape of existing genomes. The core of this tool will be a 3D plot of three metrics, chosen by the user, with a fourth metric plotted as color on the landscape surface. Initially, our landscape will consist of all sequenced genomes from the National Center for Biotechnology Information (NCBI), the central clearinghouse for genomic data. In addition, users will also be able to import their own dataset to serve as the landscape. While we will focus on genomic data for this project, we propose to build this tool to allow visualization of any related dataset.

**Objectives**

**The objectives of this class project will be:**

1. Design a graphical method for the exploration of metrics which characterize prokaryotic replicons (distinct units of a genome).
- Discover dependencies among different metric values, even when the dependency is hidden.
- Provide insight into how a replicon of interest relates to the existing "prokaryotic landscape."

- Find or create global metrics that characterize prokaryotic genomes and can be used to compare multiple genomes. Possible sources include:
    - Literature searches
    - Existing collaborations
    - Experimentation
- Determine the best implementation language and graphical libraries to use for the visualization; for example:
    - Java
    - OpenDX
- Create a proof-of-concept tool based around a simple user interface that features:
        - an interactive 3D landscape (where each replicon is represented by a point).
        - the ability to handle an arbitrary number of metrics.
        - the ability to select and visually emphasize a replicon or replicons of interest.
        - the ability to customize arrangement and display of metrics within the visualization.
        - tooltip exploration with dynamic, coordinated, statistical representations for each metric not in the landscape.
        - statistics calculated with respect to all points within the tooltip bounding box.
        - variable tooltip size.
        - *optional*: a plug-in architecture for allowing extensibility for specific functionality (for example, automatic calculation of values from genomic sequence files).

**Roles**

Both investigators have similar backgrounds, and work in the same bioinformatics field. As a result, many of the tasks will be handled jointly. This includes discovery and choice of metrics, data formatting and import, and user interface design and implementation.

**Other**

The investigators will be working in conjunction with researchers at Virginia Bioinformatics Institute, and 454 Life Sciences. Both of these groups have expressed a

strong interest in developing more sophisticated visual tools for genome comparison, and may use this project as a template for expanded development.