

Introduction to Bioinformatics
Problem Set 6: Metagenome Sequence Analysis

1. Why do many (nearly all?) reads from the Octopus and Bear Paw Hot Springs metagenomes begin with sequences that are similar to many other reads in the same collection? To answer this question, align the first 100 nucleotides of the first 100 reads of the Octopus Hot Springs metagenome. Bring the alignment into a word processor and examine it carefully. Highlight with different colors those portions of the sequences that are common to many reads and that you can attribute to known sources.
2. Define a function that will accept a read sequence from one of the two metagenomes we are considering and returns an edited sequence, with unwanted portions of the sequence removed. Of course you will need to define what you mean by "unwanted".
3. Draw maps that depict the relationships of each of the reads below to sequences similar to them. Each map should show relevant reads and relevant coordinates.
 - a. AOIX2045-g2
 - b. AOIX3837-g2
 - c. AOIX2864-b2
 - d. AOIX3762-b2
 - e. AOIX1604-b2
 - f. Of course your own reads!
4. The sequences from the two metagenomes we are considering came from viruses in two different hot springs. Perhaps the viruses in the two hot springs are quite different from one another. Or perhaps they are no more different from one another than are viruses from the same hot springs. Devise and execute an experiment to test which is the more plausible view.