Systems Microbiology 1.084J/20.106J PROBLEM SET #4 – Due Monday Nov. 6th

Problem 4.1

a. Describe the process of "shotgun sequencing". Assuming an average read length of 1 kbp per individual sequence, approximately how many clones would you need to sequence to close a 10 Mbp genome?
b. Describe the other main alternative method for determining whole genome sequences. Are these two methods mutually exclusive?

Problem 4.2

When the yeast nuclear genome was published the entire sequence was not completed. Additionally, the yeast mitochondrial genome proved difficult to accurately sequence. Describe in both cases the practical difficulties that were encountered during sequencing.

Problem 4.3

a. Compare and contrast BACs with YACs – how are they employed in genome sequencing projects?
b. What does *annotating* a genome mean, how is this accomplished, and how does it differ from *assembling* a genome?
c. Explain how horizontally transferred genes can be detected in a genome.
d. As a proportion of the total genome, what functional class of genes predominates in small genome organisms versus larger ones—why is this?

Problem 4.4

a. In *Bacteria* and *Archaea* the acronym ORF is almost synonomous with "gene", which is not the case in eukaryotes. Explain. What are the practical implications of this difference, with respect to the relative ease of sequencing bacterial versus eukaryotic genomes?
b. The gene encoding the β-subunit of RNA polymerase from *E. coli* is said to be *orthologous* to the *rpoB* gene of *Bacillus subtilis*. What does that mean about the relationship between the two genes? What protein do you think *rpoB* of *B. subtilis* encodes? Many genes for different sigma factor of *E. coli* are *paralogous*. What does this imply about their relationship?

Problem 4.5

The massive amounts of genome sequence data that are now accumulating provide a starting point for understanding the relationship between the entire coding potential of a microorganisms, and how it might function and respond under different environmental conditions. In one sense however, the genome sequence simply the "parts list" of an organisms genes. Describe the tools and approaches, and give an example experiment that you might use to leverage the entire genome sequence of *E. coli* to describe its gene expression and protein complement under different environmental conditions.

Problem 4.6

Genome structure and evolution can best be understood in the context of the life history, physiology, and activities of the organisms that encode them. A good example of this comes from recent studies of aphid endosymbionts, their relationship to their insect host, and the subsequent trajectory of their genome evolution. Briefly describe this host-symbiont system, the interactions of the two partners (*Buchnera aphidicola* and aphids), and the consequences of this natural history for their genome structure and function.