Solutions Systems Microbiology 1.084J/20.106J PROBLEM SET #4

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?

Shotgun sequencing relies on efficient and random cloning of the DNA of interest, and the generation of large numbers of clones so that the sequences can be computer assembled by looking for enough overlapping DNA fragments to cover the entire genome.

b. Describe the other main alternative method for determining whole genome sequences. Are these two methods mutually exclusive?

BAC sequencing utilizes larger insert clones. The BAC library is made, ordered maps of large clone contigs are made, then the BAC to be sequenced is done by shotgun sequencing followed by assembly.

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.

The major difficulty in obtaining the complete sequence of the yeast nuclear genome was due to the extremely long runs of repetitive DNA and it is very difficult to accurately sequence these regions. With regard to the mitochondrial genome, there is such diversity in types that it is difficult to know "what to expect" in terms of size, codon usage, organization, etc. *S. cerevisiae's in vivo* genome may be linear, further complicating the process.

Problem 4.3

a. Compare and contrast BACs with YACs – how are they employed in genome sequencing projects?

BAC (bacterial artificial chromosomes) can carry >300 kb of foreign DNA and are derived from a "trimmed down" version of *E. coli's* F' plasmid. YACs are similar in their ability to carry large fragments of DNA (200-800kb), except that they have been designed to replicate in yeasts (contain a yeast origin of replication) and contain telomeres (for replication of linear chromosomes) and centromeres (for segregation during mitosis). Both BACs and YACs are used for DNA cloning & genome sequencing.

b. What does *annotating* a genome mean, how is this accomplished, and how does it differ from *assembling* a genome?

Genome *annotation* is the conversion of raw seque3nce data into a list of genes, promoter elements, and regulatory sequences present in the organism. Gene *assembly* is the ordering of the DNA fragments and eliminating overlaps in the sequence but is not involved in "making sense" of the sequence data.

c. Explain how horizontally transferred genes can be detected in a genome.

Comparative genomics detects horizontally transferred genes. The presence of genes found only in distantly related species are often horizontally transferred genes. Differences in ORF GC content or codon biases from the rest of the genome indicated transferred genes. Functions of genes are also a clue, as horizontally transferred genes are typically not related to transcription, translation or replication.

d. As a proportion of the total genome, what functional class of genes predominates in small genome organisms versus larger ones—why is this?

The percentage of genes involved in protein synthesis is much higher in organisms with small genomes compared to those with larger genomes. In larger genomes, transcription and two-component signal transduction genes are much higher than in smaller genomes. These biases are because the protein synthesizing apparatus cannot be dispensed, whereas regulatory proteins are not essential and are often specialized thus only exist in larger genomes. The smaller the genome, the larger fraction will be protein synthesis genes.

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?

The acronym ORF is used synonymously with the term "gene" in Archaea and Bacteria since it decribes the sequence of nucleotides that when transcribed can immediately be translated into a protein. This is not the case in higher eukaryotes, which posess introns, an thus the functional gene is interrupted, and does not directly encode the corresponding protein.

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 factors of *E. coli* are *paralogous*. What does this imply about their relationship?

The term orthologue refers to a gene found in one organism that is similar to a gene found in another and different species. The similarity between an RNA polymerase subunit gene in *E. coli* and the *rpoB* gene in *B. subtilis* suggests that these genes may have arisen by horizontal gene transfer. Based upon the sequence similarity, one may be able to hypothesize that the *rpoB* gene encodes a protein that is also involved in transcription, perhaps serving the similar function in the RNA polymerase holoenzyme as the orthologous gene in *E. coli*. The paralogous genes encoding the different *E. coli* sigma factors suggests that they arose by gene duplication at some time during the organism's evolution.

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 organism's 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.

Answers will vary but should include a microarray or proteomics experiment with *E.coli* under different environmental conditions, as well as gene knockouts of *E. coli* under these same 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.