Sample Bioinformatics Exercises:

Encoding Genes (Harvard Extension)

Abstract: Students investigate the encoding of genetic information in DNA sequences. The exercise can be done early in the biology sequence.

The following lesson plan could be used over several weeks for students with programming experience. Alternately, by using the web application Gene Boy from the Dolan DNA Learning Center at Cold Spring Harbor, the lesson plan could be adopted for students with no programming experience and completed in a single class period.

In the extended version, students write a sequence of relatively simple programs. In the first project, students are given a DNA sequence and a table mapping codons to amino acids. They are asked to translate the sequence into three amino acid sequences, one for each of the reading frames in the forward direction.

If they are comfortable with elementary probability, we ask them to estimate the chance of finding an open reading frame (ORF) of length N, if the original sequence was randomly generated. They are asked the odds of finding an ORF of length greater than 100 codons.

For their second programming problem, they are given a large sequence from the *E. coli* K-12 chromosome. They are asked to search in alternative reading frames for ORFs longer than 100 codons, which appear in abundance.

Grantham's genome hypothesis proposes that each species prefers some synonyms in its coding sequences. We ask the student to test Grantham's hypothesis by counting the codon usage in the ORFs found in the prior step, and comparing the percentages with a codon usage database for *E. coli*.

We can cover much of this ground in a single period without writing a program by using the Gene Boy (Dolan DNA Learning Center 2010), a web application designed to look like a small game controller, which may be found at:
http://www.dnalc.org/resources/geneboy.html

When launched, the device has a selection of DNA sequences: random sequences, sequences holding open reading frames, sequences with intergenic content, and one longer sequence with extended genomic data, with open reading frames in multiple reading frames. The user can also load an arbitrary sequence for analysis.

Once a sequence has been selected, the user can translate or analyze the composition in a number of ways. The device can count the number of times each nucleotide is used, or each pair of nucleotides, or each triple. In the random sequence, codons will be used between 1% and 2% of the time: close to 1/64. However, in genetic sequences, some sequences are much more common than others.

The user can transform the sequence into amino acids: all three forward reading frames are shown. The user can also reverse and complement the sequence to see the other 3 reading frames. The user can look through the translated sequence for ORFs, or use a feature that searches and displays ORFs in a compact form.

This exercise, in either form, helps the student to see how cells encode genetic information, and gives them a vantage point to ask a number of interesting questions: Are all ORFs genes? Why should a species prefer one codon synonym to another? If all genes are in place at all times, how does the cell decide which genes it needs to express at a given time?

Tasting Phenylthiocarbamide (Mount Holyoke College)

Abstract: Students study a genetic trait that is easy to test and then sequence their own taste receptor gene to validate. The exercise is suitable for Introductory Genetics or Molecular Biology courses.

The full name of the project is entitled, "Tasting Phenylthiocarbamide (PTC) - Restriction Fragment Length Polymorphism (RFLP) Analysis of Student genotypes at the PTC Gene", an extension of an earlier lab (Merritt et al., 2007). PTC, also known as phenylthiourea, is an organic compound that tastes bitter, or is effectively tasteless, depending on the genetic makeup of the person doing the tasting. The ability to taste PTC is strongly associated with a polymorphism in the TAS2R38 (PTC) taste receptor gene. This gene resides on the long arm of chromosome 7 (7q34). We have added a bioinformatics analysis to an experiment described by Merritt et al. (2007), that investigates the correspondence between the phenotype and genotype of those who can detect PTC and those who cannot. Students first determine their phenotypes by comparing the taste of control paper to that of PTC paper. To determine their genotypes, students then extract DNA from their cheek cells and perform a polymerase chain reaction (PCR) to amplify a DNA segment within the TAS2R38 taste receptor gene. To determine their genotypes for the TAS2R38 gene, students isolate their PCR product DNAs and have them sequenced. The students then use NCBI BLAST to determine their genotype at the PTC gene. The students also explore questions of their own design, including analysis of homologous genes in humans and other species.

This project provides an introduction to basic bioinformatics analysis. Students learn what E-value and bit score are, and how to interpret these values. They learn how to use sequence alignment software. The study also makes students think about the concept of gene and protein families. For example, they learn that the gene/protein most closely related to human TAS2R38 is not a human gene. The project also makes students think about the relationship between genotype and phenotype. This analysis can be combined with a unit on the possible structure of the PTC taste receptor protein, and the potential effects that the base-pair substitutions in the non-taster allele might have on PTC protein structure and function.

We have modified this project for our advanced "Eukaryotic Molecular Biology" course. Students in this course do independent investigations in which they analyze genotypes at other human loci of their own choosing. For example, students have done genotype/phenotype studies on the Melanocortin-1 receptor (Mc1r) gene, which plays a key role in skin pigmentation, and the oculocutaneous albinism type II (OCA2) gene, which functions in eye and skin pigmentation.

The Polygenic and Polymorphic Nature of the Major Histocompatibility Complex: A Computer Laboratory Exercise using CLUSTALW/Phylip. (Spelman College)

(This exercise was originally presented at a Microbrew session of the ASM Annual Conference for Undergraduate Educators, 2009, Colorado State University, Fort Collins, CO). The exercise is suitable for Biology electives in Microbiology and Immunology.

Abstract: This exercise is designed to demonstrate the polygenic and polymorphic nature of major histocompatibility complex (MHC) and MHC-related genes and proteins. Students are introduced to the bioinformatics tools of the Biology Workbench, specifically CLUSTALW/Phylip.
Typically, students are provided with a file containing single letter amino acid code sequences in FASTA format for various unidentified proteins whose genes fall within the MHC region of human chromosome 6. Their task is to add the sequences into their Workbench session and then compare them using CLUSTALW. The sequence alignment shows that some sequences are highly similar overall while others have limited regions of similarity. The concept of domains (immunoglobulin-like domains, for example) can be introduced here. Students are then instructed to scroll down to the dendrogram which was created using Phylip. They are asked to identify the proteins as accurately as they can based on their relatedness from a list provided. Typically the list includes HLA-A, HLA-B and HLA-C alleles of different numbers for comparing sequences of alpha chains of MHC class I antigens coded for at the same (polymorphic) or different (polygenic) gene loci as well as sequences for Class I-related MHC class 1b proteins and sequences for alpha and beta chains of MHC class II proteins or the related HLA-DM protein chains. Students learn to differentiate between proteins coded for by similar genes at different loci in the MHC and those coded for by different alleles of the same gene at a given locus. The exercise also makes students think beyond the individual and realize that evolution of the MHC genes occurs at the population level and that resistance to disease is an important selection mechanism.

References for Supplemental Materials:

Dolan DNA Learning Center, Cold Spring Harbor Laboratory (2010) www.dnalc.org/resources/geneboy.html (accessed 12 May 2010)

Merritt, R., Bierwort, L, Slatko, B., Weiner, M., Weiner, E., Ingram, J., Sciarra, K. and Weiner, E. (2008). Tasting Phenylthiocarbamide (PTC): A New Lab With An Old Flavor. Am. Biol. Teacher *70*, 4. www.nabt.org/sites/S1/File/pdf/070-05-0023.pdf.

Complete lists of the NITLE-sponsored workshops' participants can be found at: https://ats.bates.edu/bioinformatics/?q=node/78