Homework 1, bi183

David Angeles

January 11, 2018

Problem 1. How many genomes in GenBank?

Going to the GenBank URL `https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/`, gives the number 129209 genomes availabe for download

Problem 2. Codon Usage

I downloaded the genome of *C. bolteae* and determined its codon usage. Codons are definitely not used at random, as shown in 1. I could include the table of frequencies, but I am guessing you do not want it... Curiously, this genome included 25/5802 genes that were NOT a multiple of three. I assume these are errors, and not real...

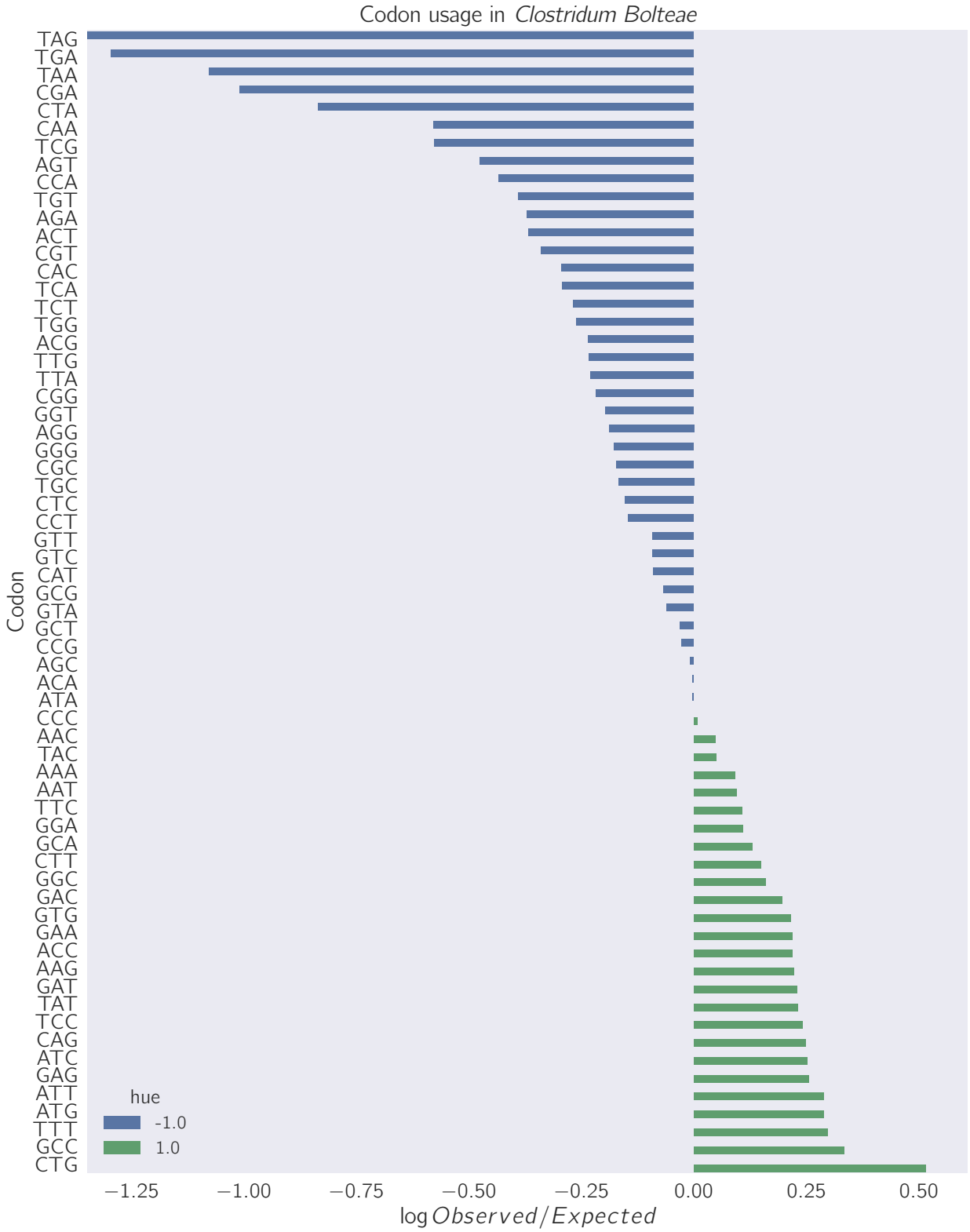

Figure 1. Codon usage in C. bolteae is not random. Coloring shows direction in which a codon differs from the random expectation. Logarithm is base 10.

Problem 3. What is selenocysteine? Who discovered it? What is its biological role?

Selenocysteine is perhaps most famous for being the 21st amino acid that is genetically encoded *sensu stricto*. It is basically a cysteine with the sulfur atom replaced by a Selenium atom, and it is typically encoded by the UGA codon. This codon does not usually code for selenocysteine, but is instead **recoded** if the mRNA forms the correct secondary structure. Selenocysteine is made in rare amounts and it plays an oversized role in redox reactions. It was discovered at the NIH by Thressa Stadtman. I did not know the name of the researcher (though I did, curiously, know her biological sex), nor the function of this amino acid so I used Wikipedia to find out (`https://en.wikipedia.org/wiki/Selenocysteine`).

Problem 4. How to learn the codon code from GenBank

I am not sure what this question means, but I suppose you could download the protein sequence files (in amino acids) and the cDNA files, and then it is a simple problem of recognizing that the DNA strings are three times the length of the AA strings.