PROBLEM SET 5: ENERGETICS AND STRUCTURES

Unless told otherwise, assume standard temperature (300K), pressure (1 atm), and pH (7). Label all axes, and include units.

1. Residues that are solvent-exposed reportedly change more frequently over evolutionary time. Using PyMol, let's assess this relationship for the β-lactamases.
   (a) Using the script `getRelativeSolventAccessibility.py` in PyMol, compute relative solvent-accessible surface area (RSA) for the TEM-1 structure in `TEM1-imipenem-from-1BT5.pse`. Inspect the script to determine how to save the results to a text file. Of the 24 alanines outside the signal sequence of TEM-1, which ones are completely buried (RSA = 0)?
   (b) The data in `data/site-diversity.txt` shows the number of amino acids n_{aa} present at each site i in the alignment, $n_{aa}(i)$. (You may remember this from Homework #1.) In R, plot $n_{aa}(i)$ versus relative solvent-accessible surface area.
   (c) What is the Spearman correlation between RSA and site diversity? (In R, `cor.test(x, y, method="spearman")` computes this correlation.) What does this mean in terms of evolutionary rates?
   (d) Comment briefly on why the observed relationship arises.

2. Amino acids have "propensities" for appearing in helices and sheets. Using PyMol and R, let's evaluate the predictive ability of these propensities for the amino acid composition of homologous sites.
   (a) Find the residues in TEM-1's β-sheets:
   `select sheet, (ss s); iterate sheet, print resi.`
   For each of these positions, plot (in R) the proportion of each amino acid type against the β-sheet propensity reported by Minor and Kim *Science* 1994, which is available in `data/minor-sheet-propensities.txt`.
   Use the amino-acid proportions provided in `data/site-aa-proportion.txt`.
   In your plot, use the amino acid letters as plotting symbols, e.g.:
   ```
   aas = c("A","C","D","E") # etc.
   plot(1:4, 1:4, pch=aas) # pch = Plotting CHaracter
   ```
   Calculate the proportions for every site, then select the values you want as follows:
   ```
   aaprop <- read.table("data/site-aa-proportion.txt", header=T)
   helix.sites <- c(1,2,3,4,5,6,7,8,9,11,15,17)
   aaprop.helix <- aaprop[match(helix.sites, aaprop$site),]
   # Compute means for each amino acid. Remove "NA" entries if they exist.
   mean.aaprop.helix <- colMeans(aaprop.helix, na.rm=T)
   ```
   (b) Find the residues in TEM-1's α-helices:
   `select helix, (ss h); iterate helix, print resi.`
   For each of these positions, plot (in R) the proportion of each amino acid type against the $\Delta\Delta G$helix propensity reported by Pace *et al.*

Biophysical Journal 1998, which is available in `data/pace-helix-propensities.txt`. Use amino acid letters as plotting symbols.

(c) Which propensity measure does better in these average comparisons? Why?

3. You're studying a protein with a free energy of unfolding of $\Delta G_u = 5$ kcal/mol (21 kJ/mol).
   (a) At equilibrium, what proportion of this protein will be unfolded?
   (b) Suppose mutations induce changes ranging from $\Delta\Delta G_u = -4$ to 1. Plot the equilibrium proportion folded as a function of $\Delta\Delta G_u$ in R.

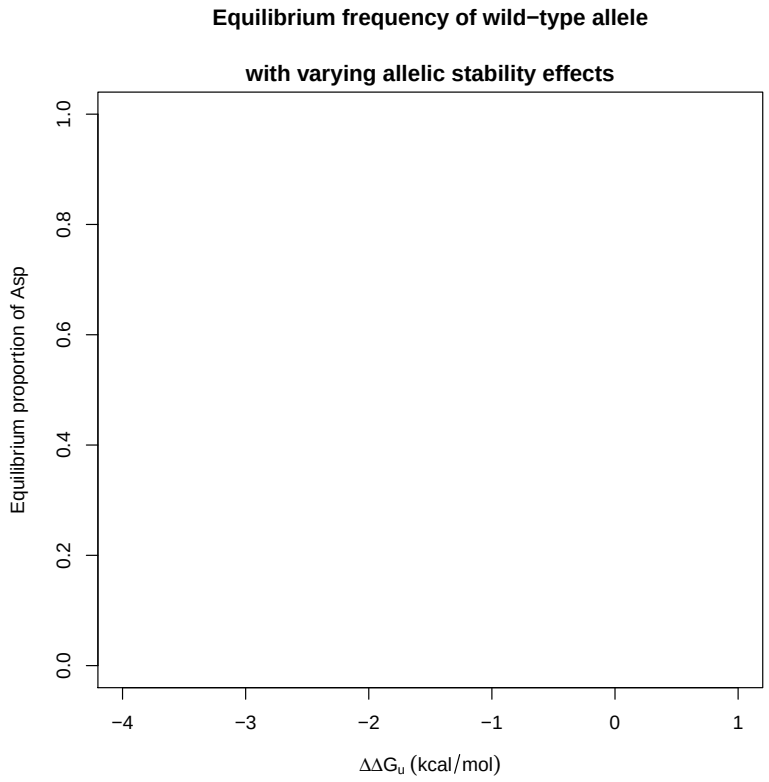
4. A bacterium's instantaneous growth rate r depends upon the expression of protein P, an enzyme which degrades an abundant environmental toxin. For this bacterium, fitness $w = e^r$. The equilibrium cellular concentration of P is $[P_0]$, but only the folded population of the enzyme is active, so that $r = \alpha[P]_0 \Pr_{\text{fold}}(\Delta G_u) k_{\text{cat}}$ where α summarizes proportionality between enzymatic activity and growth rate. For convenience, we collapse all the constants into $\beta = \alpha[P]_0 k_{\text{cat}} = 1h^{-1}$.

   At sites encoding the surface residue 10 and the core residue 50 in the gene encoding P, an unusual mutational process restricts mutations to a single nucleotide, allowing only Asp (GAC) $\leftrightarrow$ Glu (GAG) changes to happen with instantaneous rate $\mu(C \to G) = \mu(G \to C) = \mu$. The wild-type enzyme has D at both sites.

   We want to determine the steady-state frequency of D and E at each site.
   (a) What is the selection coefficient $s_{D \to E}$ for changing D into E, in terms of ΔG_u, $\Delta\Delta G_u(D \to E)$, and β? (Don't worry about different sites, or about simplifying the answer.)
   (b) Let the fraction of Asp and Glu at one of these two sites be f_D and $f_E = 1 - f_D$, respectively. Plot the steady-state fraction of Asp as a function of the mutational stability change $\Delta\Delta G_u(D \to E) = -4$ to $+1$, using $N = 100$. You may modify the code below. The wild-type enzyme has stability $\Delta G_u = 5$ kcal/mol with Asp at both sites, and $\Delta\Delta G_u(D_{10} \to E_{10}) = -1$ kcal/mol and $\Delta\Delta G_u(D_{50} \to E_{50}) = -3$ kcal/mol. Mark the values for sites 10 and 50 using `abline(v=-1)` and `abline(v=-3)` respectively, and report the frequencies. [As a sanity check, you should be able to work out in your head what steady-state value $\Delta\Delta G_u = 0$ should produce.]
   (c) Also plot the steady-state Asp frequency for $N = 10$ and $N = 1000$. Comment briefly on how and why the curves change.

```
# Probability of folding given free energy of unfolding
pfold <- function(dGu) {
  # You'll write this as part of question 3
}

# Naive probability of fixation formula for haploids
# Note: error if s == 0!
pfix.naive <- function(s,N) {
        ((1-exp(-2*s))/(1-exp(-2*N*s)))
```

```
}

# A more useful implementation of the fixation formula.
# Two features which differentiate it from the naive case:
# 1) we special-case pfix(s=0) = 1/N
#    (computers do not know L'Hopital's rule)
# 2) we "vectorize", using sapply, to detect these cases individually.
# Vectorizing means we assume that s is going to be a vector, rather
# than a single number, and we build the function to handle this case
# the way we want, detecting s=0
pfix <- function(s,N) {
  p <- sapply(s, function(the.s){
    # s != 0 really means that |s| is larger than the minimum
    # double-precision floating point number our machine
    # knows about.
    if (abs(the.s) > .Machine$double.eps) {
      # Canonical fixation formula
      p = ((1-exp(-2*the.s))/(1-exp(-2*N*the.s)))
    } else {
      # Canonical fixation formula in neutral case
      p = 1/N
    }
  })
  p
}

eq.fix <- function(ddGu,dGu,N){
  # Compute s(E->D) using ddGu and dGu
  # Compute s(D->E) using s(E->D)
  # Compute equilibrium proportion using N and the two selection coefficients
}

ddGu <- seq(-4,1,0.1) # ddGu from -4 to 1, stepping by 0.1
dGu <- 5
plot(ddGu, eq.fix(ddGu,dGu,100), type="l",
   ylab="Equilibrium proportion of Asp",
   xlab=expression(Delta*Delta*G[u]~(kcal/mol)),
   xlim=c(-4,1), ylim=c(0,1),
         main="Equilibrium frequency of wild-type allele
    \nwith varying allelic stability effects")
```




```
# And don't forget those vertical lines.
```