A pdb file contains the identities and coordinates of all atoms in a structure. The file 1u70.pdb contains the structure coordinates for the enzyme dihydrofolate reductase from mouse (mDHFR), bound to the inhibitor methotrexate (MTX), which is used in cancer chemotherapy.

We will use Python to analyze the structure.

**1. Calculate bond lengths.**

The polypeptide backbone contains two nitrogen carbon bonds: one is the peptide bond (C–N), the other one links the peptide bond nitrogen and the C atom (N–C).

**1a.** Calculate the mean of the bond lengths for all C–N and N–C in mDHFR. Report the mean values.

Note: The two values you get should be different by 0.128 Å. If the difference between your two means is not equal to these values, there is a mistake in your calculations.

The following was printed by my program after performing all the necessary calculations:

The average N-C\_alpha bond length is 1.458 Angstroms.

The average C-N peptide bond length is 1.33 Angstroms.

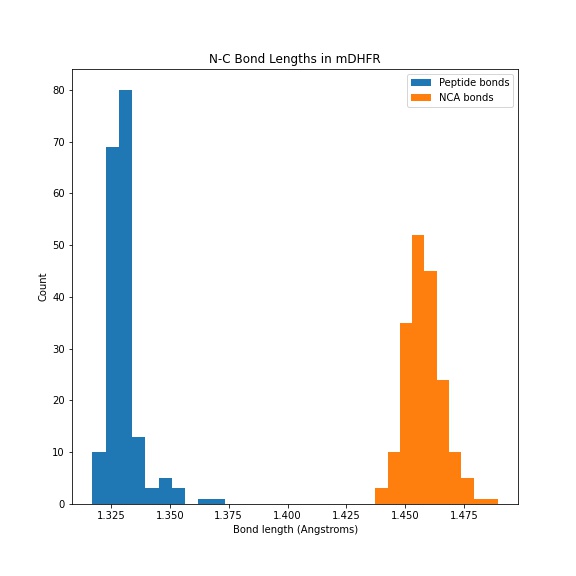
The difference is 0.128 Angstroms.

**1b.** The mean values differ from one another by a fraction of an Ångstrom. Determine whether the distributions are significantly different using a paired t-test. Report the p-value of the two-sample t-test. Is the difference in bond lengths statistically significant?

Performing a paired t-test where peptide bond was paired with the N-C\_alpha bond on the same nitrogen, the p-value is about 4.306x10-225, which is extremely significant.

**1c.** Also, represent the bond lengths in two histograms to visualize their distributions. Is the histogram consistent with the p-value from the two-sample t-test?

Below is a histogram showing the distribution of lengths of both type of bonds. Given there is no overlap between the two distributions, this is consistent with the fact that they are most definitely significantly different distributions.



**1d.** What could explain the observed difference in the mean bond lengths? Use the chemical structure of a polypeptide to explain the observation.

The resonance of the peptide bond could explain the shortened distance between the peptide carbon and nitrogen. Because the ketone group exhibits some resonance with the C-N bond, the C-N bond exists as an intermediate between a single and double bond. Given that a double bond is usually shorter than a single bond, this would explain why the resonant peptide bond would be shorter than a normal single bond between the nitrogen and alpha carbon.

**2. Ligand-binding interactions.**

MTX binds very tightly to mDHFR (which is why it is a good inhibitor). Hydrogen bonds contribute to binding.

**2a.** From the coordinates in the structure, calculate how many hydrogen bonds between N and O atoms are formed between mDHFR and MTX, and list the residues (identity and position) in mDHFR that contribute.

My code searches through all the possible N/O combinations between MTX and mDHFR and reports a hydrogen bond when the distance between the two atoms is less than 3.1 Angstroms (the sum of the radii of N and O). My code identified the below atoms as hydrogen bonding between the two molecules.

Each H-bond is reported in the form (mDHFR residue, residue atom, MTX atom, distance between the atoms in Angstroms)

[('ILE7', 'O', 'NA4', 2.542308807373047),

('GLU30', 'OE1', 'NA2', 2.7814879417419434),

('GLU30', 'OE2', 'N1', 2.8957650661468506),

('ASN64', 'ND2', 'O', 2.816410779953003),

('LYS68', 'NZ', 'O2', 2.721834659576416),

('ARG70', 'NH1', 'O1', 3.011383295059204),

('ARG70', 'NH2', 'O2', 2.8708581924438477),

('VAL115', 'O', 'NA4', 2.6931955814361572)]

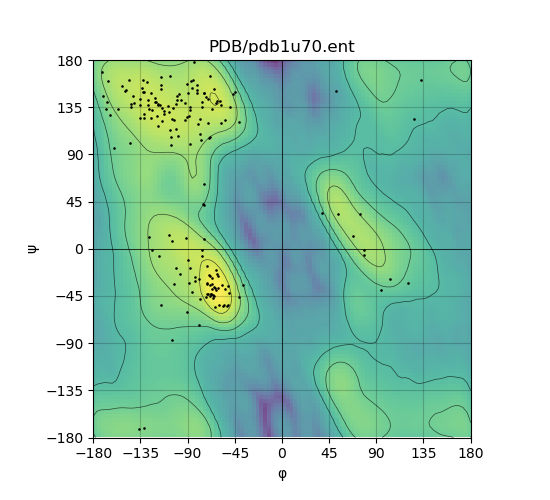
**2b.** At least one of the amino acid residues involved in forming an H-bond with MTX has an apolar side chain. Which amino acids that you identified have apolar side chains, and how can they form hydrogen bonds with MTX?

ILE7 and VAL115 have apolar sidechains, but they are able to form H-bonds through their backbone atoms, namely though their carboxyl oxygen in these cases.

**3. Analyze bond angles**

**a.** Calculate the bond angles in mDHFR and generate a Ramachandran plot.

Below is the Ramachandran plot I produced.

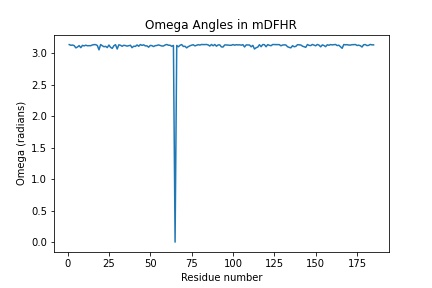


**b.** Each data points represents the dihedral angles between two peptide bonds. Most of the data points cluster in the beta-sheet region (  –120°,   120°) and in the alpha-helix region (    -50°). We will ignore any data points at  > 0 here. In which of these two clusters would you expect loop structures (i.e. neither alpha-helix nor beta sheet) to fall? Explain.

I would expect loop structures to have dihedral angles that fall within either/both of these clusters. Since both of these clusters represent angle pairs that are sterically favorable, loops, which are more flexible than alpha-helices or beta-sheets, can have residues falling in either of these categories depending on the local folded environment.

**c.** Generate a plot of the angle  against the residue number. How many cis peptide bonds does the structure contain? Explain your reasoning.

Having calculated the omega angle of all the peptide bonds, it appears there is only one cis bond (omega=0 degrees) and all the rest are trans bonds (omega=180 degrees). This makes sense because the cis conformation is normally very unfavorable, so most peptide bonds are expected to be in the trans configuration.



**4. Protein-DNA binding: Regulation of Gene Expression**

Many proteins that regulate gene expression must specifically recognize their target DNA sequences against the backdrop of many non-specific binding sites. One extensively studied example is the lac repressor (lacR), a tetrameric protein that binds to an operator, present as a single copy in the genome of *E. coli*, and suppresses expression of genes related to lactose metabolism. (Note: All of the numbers in this section are average and approximate values. Also, the volumes and concentrations/copy numbers of molecules in this exercise are very small, but we will pretend that we can analyze their interactions using the standard equilibrium model.)

In its apo form (i.e. without lactose bound), the lacR protein binds to its target sequence, the 12 base pair lac operator (Op), with very high affinity. This binding reaction can be described by the following equilibrium:

The dissociation constant for this binding reaction is approximately KD,Op = 10–12 M.

lacR binds other sequences much more weakly, with a dissociation constant for all non-operator (nonOp) sequences of KD­,nonOp = 10-4 M. (For simplicity, we are assuming here that the affinity is the same for all non-specific binding sites.)

However, the number of nonOp sites that compete with the single Op site is very large. If any 12 bp-sequence in the genome of *E. coli* (~4.5 x 106 bp) is considered a nonOp site, then there are approximately 4.5 x 106 nonOp sites.

A bacterial cell has a volume of about 1.6 x 10-15 liter (calculated as a cylinder with a radius of 0.5 µm and a length 2 of µm) and contains 10 copies of the lac repressor tetramer. The intracellular concentration of lacR then is

The concentration of non-operator binding sites in the cell can be calculated similarly:

**4a.** Calculate the equilibrium concentration of lacR bound to non-operator complexes, [lacR•nonOp], and the fraction of lacR bound to non-operator sites, [lacR•nonOp]/[lacR]total, assuming equilibrium conditions. You can further assume that [nonOp] = [nonOp]total, i.e. very few of the available non-specific binding sites are occupied because they are in very large excess over lacR. For this calculation, ignore binding of lacR to the operator (i.e. the specific binding site, Op) and only take into account the concentration of lacR and of the non-specific binding sites.

Based on your calculation, what is the cellular concentration of lacR that is free, i.e. not bound to nonOp sites?

Please see work in attached PDF.

My final answer was that [lacR•nonOp]=9.796E-6 M, [lacR•nonOp]/[lacR]total=48/49, and [lacR]free=2.041E-10 M.

**4b.** Using the concentration of free lacR calculated in **4a**, what is the occupancy of the operator site? We define occupancy here as the ratio of equilibrium concentrations of the operator bound by lacR, [lacR•Op], over the total concentration of operator, [Op]total:

Assume that the equilibrium concentration of lacR is [lacR]free, i.e. the concentration of free lacR you calculated in 3a. (Note: There is only one operator site per cell, so the concept of free and bound concentrations does not really make sense. We would have to use probabilities instead. Nevertheless, the simple equations describing binding equilibria still apply here, and for simplicity, we will use concentrations instead of probabilities.)

Please see work in attached PDF.

My final answer was that occupancy is about 0.995.

**4c.** When the lac repressor binds to lactose, its affinity for the operator is reduced by three orders of magnitude. In this case, the KD for the repressor-operator complex is 10-9 M. Repeat the calculation above (**4b**) to determine the occupancy of the operator site in the presence of lactose.

Please see work in attached PDF.

My final answer was that occupancy is about 0.169.

Reality check: Assuming that expression is switched on when the occupancy of operator by the repressor falls below 90%, does binding of lactose to lacR turn on expression based on your calculations above?

Yes. Based on the above calculations, when lactose is present, lacR is not bound enough to turn off gene expression, but it is highly bound and represses gene expression when lactose is not present.

**4d.** Imagine that you can somehow switch off any non-specific binding of lacR to non-target sites without changing the affinities of lacR for the operator site (with and without ligand) or the concentrations. What are the occupancy rates with or without lactose? (using all of the assumptions stated above)?

Please see work in attached PDF.

For my final answers, I found that the occupancy without lactose present is 0.9999 while the occupancy with lactose present is 0.9091.

Does the expression still switch on with lactose presence? What does it tell you about the impact of non-specific binding?

Based on these numbers, expression stays off even when lactose is present. This suggests that non-specific binding is essential for controlling the amount of free lacR in the cell so that lactose has a significant effect. While non-specific binding might seem like an error in the system, it is actually acting in a regulatory manner.