**Introduction**

**Challenge:** An established practice to detect homology is to compare the new sequence against a profile hidden Markov model, which is trained from homologous sequences of known structures. However, available protein data don’t always capture enough homology space.

**Intuition:** By applying simulated evolution where we incorporate evolutionary information, we expect to expand current data by generating artificial protein sequences so that the augmented data will capture more of the homology space.

In our evolutionary mutation model, we generate new sequences using the following two modules:

1. **Positional Preference Analysis (PPA):** Which position prefers mutating?
2. **Mutation Probability Approximation (MPA):** Which amino acid residues at a certain position will be mutated into?

**MPA: Which amino acid residue has/will a position chosen/choose to be mutated into?**

**Intuition:** Every position among α helices is hydrogen bonded with the amino acid residue four residues away inside helices. From the observation, in multiple sequence alignment, there are a lot of pairs of positions where two residues change together (as indicated in the example below).

**Method:** Let \( A[1..m] \) be a protein sequence where each element \( A[i] \) takes a letter for one amino acid type as its value. Let \( A'[1..m] \) be the newly generated sequence by mutating residues in \( A[1..m] \).

- We start by relaxing the conditional probability using window size 15 instead of the whole sequence:
  \[
  \]
  - Through Bayes rules, we further transform the probability as following:
  \[
  \]
  where the pairwise conditional probability is approximated using the pairwise conditional frequencies by counting co-occurrences among all α helical proteins.

**Our Results:** We Achieve 4.9% Median AUC Improvement of ROCs

**Challenge:** Residues at different positions along protein sequences have different physical and chemical properties, such as difference in hydrophilic and hydrophobic difference in energies from hydrogen bonds. Even though it is difficult to thoroughly model all factors that affect the evolutionary preference, we deal with the tendency to evolution by observing that there are gaps between sequence changes and structural variants!

**Motivation:** In the figure below, two proteins are aligned consistently in terms of 3D structures while their sequences share very little consistency.

We quantify the gaps between sequence changes and structural variants through correlational coefficients (PCC) between pairwise RMSD(root-mean-square deviation) and pairwise sequence score calculated from BLOSUM matrix. Thus, positional preference (pf) is: \( pf=1-PCC \).

**References**