

# 1. Protein structure and sequence co-evolution [20 pts]

In our lecture, we have seen that sequence co-evolution relationships can be used to assist the protein structure prediction. Given a protein sequence  $x = (x_1, x_2, \dots, x_L)$  for which we want to predict the structure, we can search its homologous (evolutionarily related) sequences in the database and form a multiple sequence alignment (MSA). Next, we can use a generative model called Markov Random Field (MRF, also known as Potts model, Ising model, or undirected graphical model) to calculate the probability distribution for sequences in the MSA. As introduced in the lecture, the probability of a sequence  $x$  of length  $L$  is defined as

$$p(x) = \frac{1}{Z} \exp[E(x)], \quad (1)$$
 where  $Z$  is a normalization constant that ensures  $p(x)$  is a probability in  $[0,1]$ . In MRF,  $E(x)$  is defined

using two groups of parameters as following

$$E(x) = \sum_i e_i(x_i) + \sum_{i < j} e_{ij}(x_i, x_j), \quad (2)$$

$i < j$

where  $x_i \in \Sigma$  is the amino acid of sequence  $x$  at the  $i$ -th position and  $\Sigma$  is the alphabet of MSA (e.g., 20 possible amino acids and the gap). The parameters  $e_i(x_i)$  are called single-site potentials and parameters  $e_{ij}(x_i, x_j)$  are called pairwise potentials<sup>1</sup>. Now let us understand this model in more detail by considering the following questions.

1. [2pts] To make  $p(x)$  a valid probability distribution, the normalization constant  $Z$  can take the form  $Z = \sum_{x \in S} \exp[E(x)]$  for some set of sequences  $S$ . What sequences should be included in  $S$  such that  $p(x)$  gives the probability of a sequence  $x$  of length  $L$ ? How large is  $|S|$ ?

The set  $S$  should include all the possible sequences of length  $L$  that can be constructed from the alphabet  $\Sigma$ .  $\Sigma$  represents the 20 amino acids plus a gap so there are 21 possible elements for each position in a sequence. For  $p(x)$  to give a valid probability for a sequence  $x$  of length  $L$ , the normalization constant  $Z$  must sum the exponential of the energy function  $E(x)$  over every possible sequence that can be formed.  $S$  must be the set of all possible sequences of amino acids and gaps of length  $L$ .  $|S|$  or the size of  $|S|$  is equal to  $21^L$ .

2. [4pts] If we build an MRF model described in Eq. 2 for an MSA of sequences with length  $L$ , and assume the alphabet is  $|\Sigma| = q$ , how many parameters does this MRF have?

We have to add the number of single site and pairwise potentials: *Total Parameters* =

$$L \times q + \frac{L \times (L-1)}{2} \times q^2$$

$L$  = length of sequence

$q$  = possible amino acids/gaps

3. [4pts] In our lecture, we mentioned that the single-site potentials  $e_i$  can reflect the preference of an amino acid appearing at position  $i$ . For example, if  $e_i(x_i)$  has a larger value compared to other  $e_i(x_j)$  ( $j \neq i$ ), then it means the amino acid  $x_i$  will have a higher likelihood showing at position  $i$ .

We also mentioned that the pairwise potentials  $e_{ij}$  can reflect co-evolving residue pairs. Now, if we have built an MRF model and learned all its parameters, and suppose that residues  $i$  and  $j$  are co-evolving pairs, while residues  $u$  and  $v$  are non co-evolving pairs. Please describe, in your own words, how would the values in  $e_{ij}$  (note that  $e_{ij}$  is a matrix with size  $|\Sigma| \times |\Sigma|$ ) be different from those in  $e_{uv}$ ? In other words, intuitively, are there any signals/patterns in  $e_{ij}$  that can help you tell which residue pairs are co-evolving?

In an MRF model with well-learned parameters, the pairwise potentials matrix  $e_{ij}$  for co-evolving residue pairs  $i$  and  $j$  would show pronounced high values for specific amino acid combinations, indicating a strong preference for these pairs due to their evolutionary relationship—changes in one residue tend to be compensated by changes in the paired residue to maintain protein function or structure. For non-co-evolving residues  $u$  and  $v$ , the  $e_{uv}$  matrix would lack such distinct high values, as the amino acids at these positions evolve more independently, without significant compensatory interactions. Thus, the presence of distinct peaks in pairwise potential matrices is a clear signal indicating which residue pairs are co-evolving.

4. [3pts] Based on the above intuition, can you design a “co-evolution score”  $c(i,j)$  using  $e_{ij}$  such that  $c(i,j)$  has a higher value when residues are co-evolving than when they are not?

Question 1 Part 4

$$c(i,j) = \frac{1}{|\Sigma|^2} \sum_{x_i \in \Sigma} \sum_{x_j \in \Sigma} \frac{|e_{ij}(x_i, x_j) - \bar{e}_i(x_i) \cdot \bar{e}_j(x_j)|}{\sigma_{ij}}$$

5. What is the minimum number of queries you have to make to compute  $\Delta E(x', x)$ ? And what are those parameters you need to query? You need to justify your answer.

We need  $1 + 2(L - 1) = 2L - 1$  queries to compute  $\Delta E(x', x)$ . This was derived from the fact that we need to query the oracle for the parameters that have changed from  $x$  to  $x'$ . 1 query is for  $e^k(b) - e^k(a)$  has been queried as it is implied in the problem when it states that you have the parameters for  $x$ .  $2(L - 1)$  queries are needed for the pairwise potentials since there are  $L - 1$  positions other than  $k$  and each has two queries, one for the pair with  $a$  and one for the pair with  $b$ . All in all, we need to query  $e^k(b)$  for the single site potential at position  $k$  and  $e_{ik}(x_i, a)$  and  $e_{ik}(x_{ik}, b)$  for each  $i$  not equal to  $k$  for the pairwise potentials.

## 2 Graph Neural Networks (GNNs) Basics [15 pts]

(1) [5 pts] Graph Neural Networks (GNNs) are an important class of deep learning models used for processing graph-structured data. As we have shown in Lecture 15, one of the key properties of GNNs is that they should be able to produce the same output regardless of the input order of the nodes in a graph. In this problem, you will be presented with four undirected unweighted graphs, each with a set of nodes and edges. Your task is to identify which graphs are isomorphic as the first graph.

Formally, we say Graph  $G$  and Graph  $H$  are isomorphic when there exists a bijection  $f$  between the vertex sets of  $G$  and  $H$  such that any two vertices  $u$  and  $v$  of  $G$  are adjacent in  $G$  if and only if  $f(u)$  and  $f(v)$  are adjacent in  $H$ .

Which of Graph 2, Graph 3, and Graph 4 is isomorphic as Graph 1? Explain your reasoning.

Note: All graphs are undirected and unweighted, and the edges are given in the table below. Each row in the table represents one edge in the graph. For example, for a graph  $G = (V, E)$ , an edge  $(u, v) \in E$  will be presented as  $(\text{Edge}[0], \text{Edge}[1])$  in one row of the table.

Graph Neural Networks (GNNs) Basics

(1) Which of Graph 2, Graph 3, and Graph 4 is isomorphic as Graph 1?

Graph 1

11 edges

Graph 2

Graph 3

Graph 4

11 edges

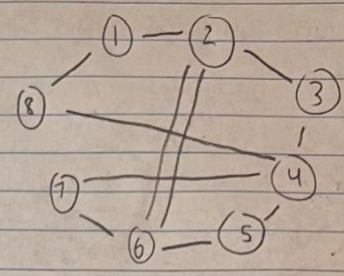
Graphs 1 and Graph 4 have connections to 4 other vertices. This did not happen in the other graphs.

Both have similar number of edges connections in graph 1 and 4



2. We will use the step function as the activation function  $\sigma$ . The definition of the step function is as follows:  $\text{step}(x)$  returns +1 if  $x$  is positive, and 0 otherwise. Please answer the following two questions based on this. If the parameters  $w_{11}, w_{12}, w_{21}, w_{22}, w_{31}, w_{32}$  have the same value as 1.0, the bias  $b_1, b_2, b_3$  have the same value -1.5, what's the predicted class of all the nodes from 1 to 8?

$w_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$      $b_1 = \begin{bmatrix} -1.5 \\ -1.5 \end{bmatrix}$   
 $w_2 = [1, 1]$      $b_2 = 1.5$   
 $n(v) = \text{degree of the node}$



$h_{v_2} = \frac{2}{3}$   
 $h_1 = \begin{bmatrix} w_{11}x_1 + w_{12}x_2 + b_1 \\ w_{21}x_1 + w_{22}x_2 + b_2 \end{bmatrix}$   
 $h_v = \sum_u \sigma \left[ \frac{x_u + b_v - 1.5}{x_u + v_2 - 1.5} \right]$   
 $h_1 = \frac{1}{3} \begin{bmatrix} \sigma(5+1-1.5) + \sigma(1+1-1.5) \\ \sigma(5+1-1.5) + \sigma(1+1-1.5) \end{bmatrix} = \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix}$   
 $y_{pred,1} = \sigma \left[ \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix} - 1.5 \right]$   
 $= \sigma(4/3 - 1.5) \rightarrow 0$   
 $h_2 = \frac{1}{5} \begin{bmatrix} 0+1+1+1+1 \\ 0+1+1+1+1 \end{bmatrix} = \begin{bmatrix} 4/5 \\ 4/5 \end{bmatrix}$   
 $y_{pred,2} = \sigma(4/5 - 1.5) = 1$   
 $h_3 = \frac{1}{2} \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$   
 $y_{pred,3} = \sigma(2 - 1.5) = 1$   
 $u_4 = \{3, 6, 5, 7, 8\}$   
 $h_4 = \frac{1}{5} \begin{bmatrix} 1+1+1+1+1 \\ 1+1+1+1+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$   
 $y_{pred,4} = \sigma(2 - 1.5) \rightarrow 1$   
 $u_6 = \{2, 4, 5, 6\}$   
 $h_5 = \frac{1}{4} \begin{bmatrix} 1+1+1+1 \\ 1+1+1+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$   
 $y_{pred,5} = \sigma(2 - 1.5) \rightarrow 1$

$u_6 = \{2, 5, 6, 7\}$   
 $h_6 = \frac{1}{4} \begin{bmatrix} 1+1+1+1 \\ 1+1+1+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$   
 $y_{pred,6} = \sigma(2 - 1.5) \rightarrow 1$   
 $u_7 = \{4, 6, 7\}$   
 $h_7 = \frac{1}{3} \begin{bmatrix} 1+1+1 \\ 1+1+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$   
 $y_{pred,7} = \sigma(2 - 1.5) \rightarrow 1$   
 $u_8 = \{1, 4, 8\}$   
 $h_8 = \frac{1}{3} \begin{bmatrix} 0+1+1 \\ 0+1+1 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix}$   
 $y_{pred,8} = \sigma(4/3 - 1.5) \rightarrow 0$

$y_{pred} = \{0, 1, 1, 1, 1, 1, 1, 0\}$   
 $i = 1, \dots, 8$

Node	neighbors
1	→ 2, 8
2	→ 1, 3, 5, 6
3	→ 2, 4
4	→ 3, 5, 7, 8
5	→ 2, 4, 6
6	→ 2, 5, 7
7	→ 4, 6
8	→ 1, 4

3. [5 pts] Find a set of parameters  $w_{11}$ ,  $w_{12}$ , ...,  $w_{31}$ ,  $w_{32}$ ,  $b_3$  to make the predicted  $y$  the same as the labels in Table 5.

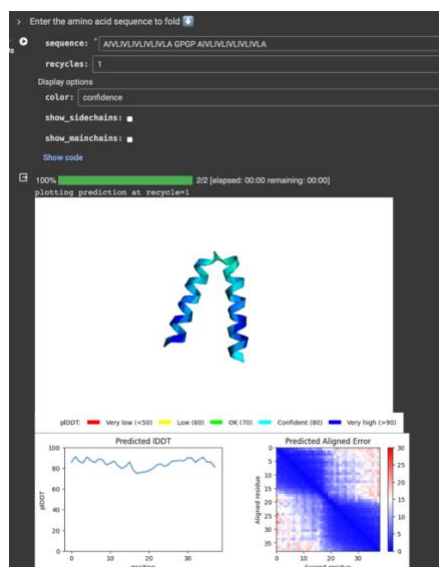
Submission Notes: For Q2.3, please write down your answers in a text file and named the file as "weights.txt". There should be nine numbers in the file, separated by spaces. The order is  $w_{11}$ ,  $w_{12}$ ,  $b_1$ ,  $w_{21}$ ,  $w_{22}$ ,  $b_2$ ,  $w_{31}$ ,  $w_{32}$ ,  $b_3$ . Compress the file weights.txt together with other code files that need to be submitted into a zip and upload it to gradescope.

DONE

### 3 Protein design by hand [15 pts]

Alpha Helices:

Sequence: AIVLIVLIVLIVLIVLA GPGP AIVLIVLIVLIVLIVLA



In designing a sequence for the dual alpha-helix structure, I selected the amino acids Alanine (A), Isoleucine (I), Valine (V), and Leucine (L) for their strong helix-forming tendencies and to provide structural stability. To ensure flexibility and proper spacing between the helices, the sequence included a "GPGP" linker segment, leveraging Glycine for flexibility and Proline to introduce turns. I originally used Serine in my linker but had to take that out due to reduced confidence score.

Betasheets

Sequence: VTIVTIVTIGPGVTIVTIVTI

The sequence "VTIVTIVTIGPGVTIVTIVTI" was designed for beta-sheet formation, utilizing Valine (V) and Isoleucine (I) for their strong hydrophobic interactions that favor sheet conformation, and Threonine (T) for its capacity to form hydrogen bonds, contributing to the sheet's stability. The central "GPG" was included to provide a flexible turn, causing the distinct separation needed between the two beta-sheets.

