Introduction

As part of the Statistical Methods in Genetics course, we were tasked with addressing five fundamental questions related to amino acid (AA) sequences and their modeling via Hidden Markov Models (HMMs), with a specific focus on the roles of states and signals. The primary objective of this study is to clarify protein function.

- Why is this significant?

Proteins are complex macromolecular structures made up of amino acids. Although their primary sequences can be easily analyzed, deducing their complete range of functions poses a considerable challenge. This difficulty arises mainly because a protein's function is closely linked to its three-dimensional structure.

The tertiary (three-dimensional) structure of a protein is heavily influenced by the spatial configuration and interactions among its secondary structural components—predominantly α-helices and β-sheets—as illustrated in Figure 1.

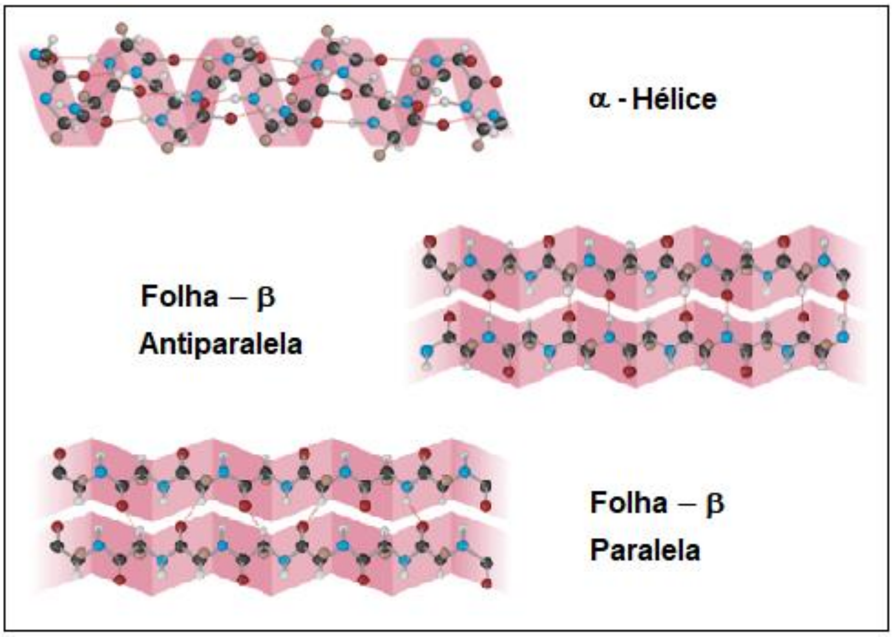


Figure 1: Protein secondary structures – typical folding patterns of α-helices and β-sheets.

Among these structures, the α-helix and β-sheet are the most commonly observed motifs, largely due to the stabilizing hydrogen bonds that form between the N–H and C=O groups within the polypeptide backbone. Importantly, these interactions occur independently of the side chains of the amino acids involved.

- α-helix

The α-helix is a right-handed helical formation that resembles a spring, where each amino acid residue causes a rotation of approximately 100 degrees along the axis of the helix. This conformation is maintained by intramolecular hydrogen bonding between the carbonyl oxygen of one residue and the amide hydrogen of another, typically situated four residues apart.

- β-sheet

The β-sheet is composed of beta strands that are interconnected laterally through a minimum of two or three hydrogen bonds between the backbone atoms, resulting in a typically twisted and pleated configuration.

β-sheets are categorized into two primary types based on the orientation of the individual strands:

Parallel β-sheet: In this arrangement, the strands align in the same direction—from the N-terminus to the C-terminus. The hydrogen bonds linking adjacent strands are angled rather than linear, which can affect the overall stability, often making it less stable than its anti-parallel counterpart.

Anti-parallel β-sheet: In this case, the strands are oriented in opposite directions. The hydrogen bonds established between strands are more linear, providing enhanced structural stability to the anti-parallel configuration compared to the parallel variant.

At this stage of the project, the protein being studied is recognized to adopt three potential secondary structures: α-helix, parallel β-sheet, and anti-parallel β-sheet.

Furthermore, we include the complete primary sequence of the protein and have prior insights regarding the spatial arrangement of the first N amino acids, which are identified as part of a β-sheet structure.

1- Describe and define the set of possible states of the Markov chain. Consider three states.

The application of Hidden Markov Models (HMMs) facilitates the examination and estimation of the secondary structure composition of proteins.

Generally, states are represented as s_i , where i=1,2,3. Let SS denote the collection of all potential states, such that si ∈ S and I ∈ {**α, β, c**}.

In the framework of this research, each state is associated with a unique folding pattern of the protein backbone. Consequently, the model consists of three potential states, which are defined as follows:

Where:

1. **State H (α-hélice)**

Represents segments of the protein organized into regular helical structures, stabilized by hydrogen bonds between amide and carbonyl groups along the main chain. This conformation is one of the most prevalent in globular proteins.

2. **State E (folha-β)**

Corresponds to regions of the protein where residues are extended and interact laterally with other β-chains, forming sheets stabilized by interchain hydrogen bonds, which can be either parallel or antiparallel.

3. **State C (coil or loop)**

Encompasses all regions that do not adopt organized conformations of helix or β-sheet. These areas include turns, loops, and segments with greater flexibility or structural disorder.

A Hidden Markov Model (HMM) consists of two primary elements: states and signals. The states constitute the foundational Markov chain, whereas the signals signify the observable outputs linked to each state. In this research, the signals are identified with the amino acids.

This study aims to deduce the three-dimensional configuration of a protein based on its amino acid sequence. Consequently, we engage with two concurrent stochastic processes:

- The Markov chain of hidden states, represented as X = {Xt: t ∈ T}, which illustrates the sequence of structural states over time;
- The observable signal chain, denoted as Y={Yt: t ∈ T}, which records the sequence of amino acids at each temporal position.

In this context, the index t∈T indicates the position of the t-th amino acid within the sequence. It is crucial to note that Xt signifies a series of random variables (assumed to be identically distributed), while Yt represents the observed manifestation of amino acids.

A Markov chain is rigorously defined as a stochastic process that adheres to the Markov property, which asserts that:

$$P\ (Xt = si|Xt-1 = sj,\ Xt-2 = sk,\ \ldots,\ X0 = s0) = P\ (Xt = si|Xt-1 = sj)$$

$$\forall t \in T;\ si,\ sj,\ sk \in S$$

This implies that the knowledge of all preceding states is transmitted through the initial lag.

2. Indicate the set of signals issued by the states. Justify.

In a concealed Markov chain utilized for forecasting the secondary structure of proteins, the observable signals align with the components of the protein's primary sequence, specifically the amino acids that make it up.

As previously stated, the sequence of signals is characterized as: $Y = \{Yt : t \in T\}$.

Given that the primary structure of a protein is constituted by a linear sequence of amino acids, and considering that there are 20 genetically encoded natural amino acids, the set of possible signals emitted by the states of the Markov chain is represented by:

$$\mathbf{A = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}}$$

$$a_k \text{ where } k = 1, 2, 3, 4, \ldots, 20$$

where each letter signifies an amino acid, according to the standardized one-letter notation:

A – Alanina	R – Arginina	N – Asparagina	D – Ácido aspártico
C – Cisteína	Q – Glutamina	E – Ácido glutâmico	G – Glicina
H – Histidina	I – Isoleucina	L – Leucina	K – Lisina
M – Metionina	F – Fenilalanina	P – Prolina	S – Serina
T – Treonina	W – Triptófano	Y – Tirosina	V – Valina

Justification:

Each hidden state of the Markov chain (i.e., each type of secondary structure – α-helix, β-sheet, or coil) is linked to a probability distribution over the observable signals, which refers to the amino acids. This illustrates the statistical tendency for certain residues to appear more frequently in specific structural conformations. For instance, alanine is commonly found in helices, whereas valine is more likely to occur in β-sheets.

Thus, the set of signals emitted by the states corresponds to the set of possible amino acids in the primary sequence, and the emission of each amino acid by a given state is modeled by a probability conditioned on the structural state.

3. Construct the initial distribution vector and indicate the shape of the transition and emission probability matrices.

Initial distribution (vector π):

The initial distribution signifies the likelihood of the Markov chain commencing in each of the hidden states, specifically, the state that corresponds to the secondary structure of the first amino acid in the protein sequence.

According to the statement, it is known that the initial amino acids (N-terminal end) of the protein exhibit a β-sheet conformation. Therefore, the initial distribution must deterministically reflect this information:

$$\pi = \begin{bmatrix} \pi \mathbf{H} \\ \pi \mathbf{E} \\ \pi\, C \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

where:

π α : initial probability of being in the alpha-helix state (H)

π β : initial probability of being in the beta-sheet state (E)

π C: initial probability of being in the coil state (C)

Transition matrix (T):

The transition matrix outlines the probabilities of transitioning between consecutive hidden states, i.e., the likelihood of the secondary structure shifting from one type to another between adjacent positions in the protein sequence.

The structure of the transition matrix T for three states is:

$$T = \begin{bmatrix} t\mathbf{HH} & t\mathbf{HE} & t\mathbf{HC} \\ t\mathbf{EH} & t\mathbf{EE} & t\mathbf{EC} \\ tC\mathbf{H} & tC\mathbf{E} & tCC \end{bmatrix}$$

where each element tij = P (st+1=j | st= i) denotes the probability of transitioning from state i to state j. The rows of the matrix must meet the normalization condition:

$$\sum_{j\, \in\{H,E,C\}} tij = 1 \quad , \text{para todo } i \ \in \{H, E, C\}$$

Emission matrix (E):

The emission matrix illustrates the likelihood of each hidden state producing a specific observable symbol, which refers to the probability of observing a particular amino acid given the local structural state.

The structure of the emission matrix E is:

$$E = \begin{bmatrix} e\mathbf{H},\mathbf{A} & e\mathbf{H},\mathbf{R} & \cdots & e\mathbf{H},\mathbf{V} \\ e\mathbf{E},\mathbf{A} & e\mathbf{E},\mathbf{R} & \cdots & e\mathbf{E},\mathbf{V} \\ eC,\mathbf{A} & e\mathbf{C},\mathbf{R} & \cdots & eC,V \end{bmatrix}$$

where each element $eij = P(ot = a \mid st = \ i)$ represents the probability of state i ∈ {H,E,C} emitting the amino acid a ∈ A (the set of 20 amino acids).

Just like in the transition matrix, it is also required that each row of the emission matrix constitutes a probability distribution, i.e:

$$\sum_{a \in A} ei,a = 1 \quad \text{, para todo } i \ \in \{H, E, C\}$$

Restrictions on the parameters:

All probabilities (π, T, and E) must lie within the range [0,1];

The sum of the probabilities in each row of the matrices T and E, as well as in the vector π, must equal 1.

4. Build a hidden Markov chain scheme for the structure under study, indicating possible states and signals, as well as transitions, emissions and their respective probabilities.

A hidden Markov chain consists of:

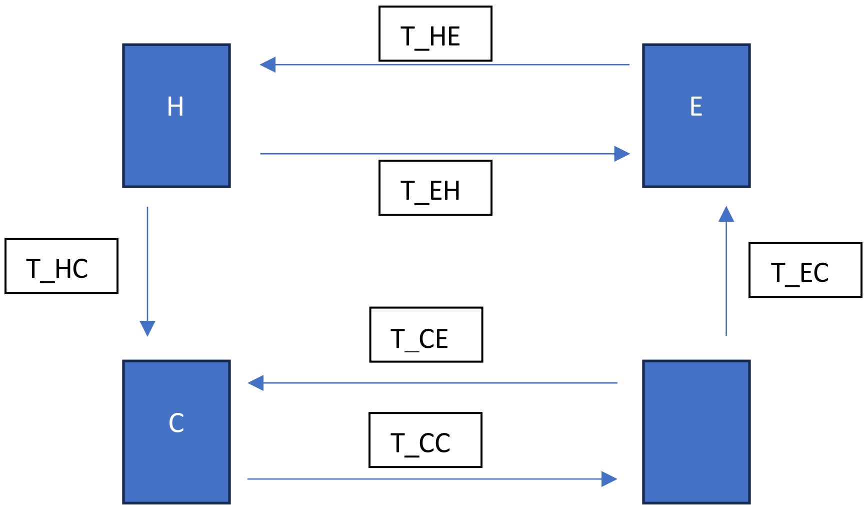
• A set of hidden states S={H,E,C}, which represent the local secondary structure of the protein (α-helix, β-sheet, and coil);

• A set of observable symbols A, corresponding to the 20 amino acids;

• An initial distribution vector π, which specifies the probability of starting in each state;

• A state transition matrix T, which defines the probability of moving between consecutive states;

• An emission matrix E, which outlines the probability of observing each amino acid given the hidden state.



Legend:

- H, E, C: hidden states (α-helix, β-sheet, coil).
- t_{ij}: transition probability from state i to state j.

Each state emits observations (amino acids) with specific probabilities. For instance:

- State H (helix) emits amino acids:
  - P(A|H) = eH,A, P(R|H) = eH,R, …, P(V|H) = eH,V

- State E (beta-sheet) emits amino acids:
  - P(A|E) = eE,A, P(R|E) = eE,R, …, P(V|E) = eE,V

- State C (coil) emits amino acids:
  - P(A|C) = eC,A, P(R|C) = eC,R, …, P(V|C) = eC,V

Initial distribution:

As previously defined:

$$\pi = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \text{assuming the start at sheet-}\beta\text{ (state E)}$$

5. Evaluate the chain considered for: stationarity, irreducibility and reversibility. Justify.

States of the Chain:

- H (alpha-helix);
- E (beta-sheet);
- C (coil or loop).

Initial distribution:

$$\pi = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Example of a Transition Matrix

$$A = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$$

1. Stationarity

A chain is considered stationary if the distribution of states remains constant over time, meaning that there exists a distribution π' such that:

$$\pi' A = \pi'$$

For the matrix A presented above, by solving the system of equations, we obtain:

$$\pi' = [0.25 \quad 0.45 \quad 0.30]$$

Since π' exists and is unique, the chain reaches a stationary distribution. However, the initial distribution π is not equal to π', which means the chain is not stationary from the outset, but rather converges to a stationary state.

2. Irreducibility

A chain is considered irreducible if it is possible to reach any state from any other state in a finite number of steps. In matrix A:

- All elements are positive (aij>0), indicating that transitions between any two states are feasible.
- Therefore, the chain is irreducible.

3. Reversibilidade

A chain is considered reversible if it meets the detailed balance condition:

$$\pi' i \ aij = \pi' j \ aij \quad \forall \, i, j.$$

Verifying for π' and A:

For E and H:

$$0.25 * 0.3 = 0.075 \ \neq 0.45 * 0.2 = 0.09$$

Since equality is not satisfied, the chain is not reversible.

Conclusion of question 5:

- Stationarity: The chain converges to a stationary distribution, yet it is not stationary from the outset.
- Irreducibility: The chain is irreducible, as all states communicate with one another.
- Reversibility: The chain is not reversible, as it does not meet the detailed balance condition.

Note: The conclusions are contingent upon the values assigned to A. If the transition matrix is symmetric (aij = aji), the chain would be reversible.