

DEPARTAMENTO DE INFORMÁTICA DA FACULDADE DE CIÊNCIAS
Métodos Estatísticos em Bioinformática | 07/2022
Question 1

Lucas C. Monteiro - nº 52849

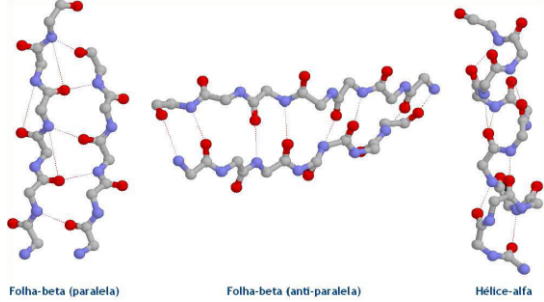


Figure 1. Representation of the three possible secondary structures for the protein being studied. From left to right, parallel and anti-parallel Beta-Sheet, and Alfa-Helix.

1. The paper has the intention of studying a hidden Markov chain model that predicts the secondary structure of a protein. The primary structure of this protein is known, as well as that the first amino acids are available in beta-sheet.

a) The states of the hidden Markov chain, s_n , will be the possible secondary structures of the protein in study. In this question, the three possible states are represented in Figure 1, so it is possible to define a sequence \mathcal{S} of states. The states are then parallel and anti-parallel Beta-Sheet, $p\beta$ and $a\beta$ respectively, and Alfa-Helix, α , i.e

$$\mathcal{S} = \{ s_1 = p\beta, s_2 = a\beta, s_3 = \alpha \}$$

b) However, these states are unobservable, hence hidden. The model requires an observable process whose outcomes are influenced by the outcomes of each state in a known way. In other words, that the states s_i emit observable symbols a_k with a given probability $\lambda_{ik} = P(Y_t = a_k | X_t = s_i)$, where X_t and Y_t are random variables representing the state and symbol emitted by that state at time t , respectively.

So what are the observable symbols? Since a protein is a sequence of 20 different possible amino acids, these will create the sequence \mathcal{A} of symbols. Also, amino-acid emissions are independent and conditioned with respect to the hidden process (Juliette Martin et al., 2006).

$$\mathcal{A} = \{ a_1, a_2, a_3, \dots, a_{20} \}$$

c) To complete the model, the matrices of transition and

emission probabilities need to be defined, as well as the vector of the initial distribution.

The matrix of emission probability, Λ , is composed by the previously mentioned probabilities of emission. Since the model has three states and twenty different emissions then:

$$\Lambda = [\lambda_{ik}] = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \dots & \lambda_{1,20} \\ \lambda_{2,1} & \lambda_{2,2} & \dots & \lambda_{2,20} \\ \lambda_{3,1} & \lambda_{3,2} & \dots & \lambda_{3,20} \end{bmatrix}$$

The matrix of state transition, P , is a 3×3 square matrix, where each element, p_{ij} , represents the transition probability of state s_i moving to state s_j , i.e $P(X_{t+1} = s_j | X_t = s_i)$.

$$P = [p_{ij}] = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

As far as it is known, all states can transition to one another, and emit all possible symbols. So no restrictions on these matrices will be accounted for, at least for now.

However, it is known that the first amino acid is available in beta-sheet. This is accounted for in the initial distribution vector, π . This vector is composed of the probabilities that the system is at a given state at initial time $t = 1$, $\pi_j = P(X_1 = s_j)$. Since the initial state is not Alpha-Helix then $\pi_3 = 0$, yielding the following π vector:

$$\pi = (\pi_j) = (\pi_1, \pi_2, 0)$$

d) Now that the model is constructed, an evaluation of its properties can proceed. A Markov chain can be *stationary* if the probability vector is invariant by the P matrix, or the probability distribution remains unchanged over time. Since the initial probability of $X_1 = \alpha$ is null, and α is a *possible* secondary structure, i.e $\exists t : P(X_t) \neq 0$, then the markov chain isn't stationary.

As it isn't stationary, it cannot be *reversible* because it will not satisfy the condition $\pi_i p_{ij} = \pi_j p_{ji}$, $\forall s_i, s_j \in \mathcal{S}$. Notice that the probabilities of transition to s_3 would all have to be zero.

For the chain to be *irreducible*, any state can be reached from any other state. As previously mentioned, no restrictions to these transitions are known so it is concluded that the chain is irreducible.