Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 42–51

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

A method of data mining using Hidden Markov Models (HMMs) for protein secondary structure prediction

Mourad LASFAR[a]*, Halima BOUDEN[a]

[a] *EMTI, Faculty of Sciences, University Abdelmalek Essaadi, BP 2121 M'Hannech II , 93030 Tetouan, Morocco.*

Abstract

The prediction of the secondary structure of proteins is one of the most studied problems in computational biology. However, the accuracy of the predicted secondary structure is insufficient for practical utility. In this paper, we propose an algorithmic approach based on Hidden Markov Models (HMM) to model the problem of prediction. Therefore, HMM are often used for data mining in bioinformatics. In this research, we have built a HMM that models the prediction problem of protein secondary structure. Moreover, two procedures for estimating the probability parameters were performed by the Maximum Likelihood Estimation (MLE) of protein sequences from a public database (Brookhaven PDB). Finally, a new prediction approach based on a posteriori probability of hidden regimes has been implemented. Our model appears to be very efficient on single sequences, with a score of 66.6% by comparing the first results obtained with the real secondary sequence and encouraging for an improvement of the system.

Keywords: Bioinformatics ; computational biology ; Data Mining ; Hidden Markov Models ; posterior probability ; prediction ; secondary structure of proteins ; supervised learning

1. Introduction

The structural analysis of a protein often requires a first step of local structure prediction; secondary structure. The bioinformatics tool is essential to quickly realize this work, more expensive by traditional methods "X-ray crystallography and nuclear magnetic resonance (NMR)". To this end, many predictions algorithms based on

* Mourad LASFAR. Tel.: +212-666-625-434.
E-mail address: lasfar.mourad@gmail.com
bouden.halima@gmail.com

10.1016/j.procs.2018.01.096

different approaches have been developed. A number of tools for protein secondary structure prediction exist today; a part of them also utilize hidden Markov models [2] [16] [17]. However, the prediction accuracy of this approach is insufficient for practical use, particularly in pharmaceuticals and protein engineering. Moreover, the technical problems of the literature such as the construction of a model adapted to the sequences to be modeled, the problem of parameterization the model…etc, are obviously not neglected and have often tedious steps. In order to deal with these problems, we have developed a system for predicting the sequence of secondary structures by Hidden Markov Chains. Thus, a three-step work conducted on protein sequences of a database (Brookhaven PDB)[1].

First, a model was constructed from a prior knowledge (i.e. it does not require any prior learning step). Then two procedures for estimating numerical model parameters were performed by maximizing the likelihood of the data studied. Finally, we propose an algorithm which is based on the posterior probability of hidden regimes to predict the sequence of secondary structures.

The first part in this research will be devoted to a presentation of the biological aspect of the problem. This part provides also an introduction into Data Mining in Bioinformatics. In addition, an overview on HMM applications and the main problems that can be solved using model-based approaches. At the end, the maximum likelihood method was presented to estimate the probabilities of parameters. In the second part, we propose our model HMM, the procedures for estimating their parameters and resolution the problem of overfitting during training. Also, the interest of a new prediction algorithm and its design based on the a posteriori probability approach of hidden states has been proposed. In a third part, we discuss the preliminary results obtained justifying the approach used for prediction.

2. Literature Review

2.1. Biological Background

A protein sequence consists of a succession of twenty amino acids. A protein sequence of length L will be represented by a sequence $x_1...x_L$ where for each symbol i of the sequence x, x_i to a value in the alphabet with twenty letters representing the twenty amino acids. The primary structure of the protein corresponds to the amino acid sequence itself called the polypeptide chain. This is the lowest level of the description of a protein, which is presented in the following way:

$$^{1}\text{AGTFHN}\ldots\ldots\text{IKNMDA}^{L}$$

Secondary structure represents the interactions between the amino acids due to the formation of hydrogen bonds between the oxygen of a carbon group and the hydrogen attached to the nitrogen of the amine group of another residue (this is the conformation Local polypeptide chain) (Fig 1).

There are generally three possible conformational states:

- αHelix ; β Strands or Sheet ; Coil or loops (random coils, β turn).

The tertiary and quaternary structure which are respectively: the folding of a sequence (appearance of attractions between helix and strands), its spatial configuration and the arrangement of multiple polypeptide chains (that is, multiple different proteins). It is these structures which ensure the functioning of cellular processes and serve as building material for the cell and organs. Different databases allow access to these files to researchers; the database can be classified according to the type indexed sequences. For examples:

- SwissProt2 database composed of amino acid sequences corresponding to the primary sequence.
- Bank of secondary and tertiary structures such as PDB.

2.2. Data Mining in Bioinformatics

Bioinformatics [5] is the science of managing, analyzing, extracting and interpreting information from biological sequences and molecules. It has been an active research area since the late 80's. After the human genome project was completed, this area has drawn even more attention. With more genome sequencing projects undertaken, data in the field such as DNA sequences, protein sequences, and protein structures are exponentially growing. Facing this huge mount of data, the biologist cannot simply use the traditional techniques in biology to analyze the data. In order to understand the mystery of life, instead, information technologies are needed.

[1]PDB address: http: // www.rcsb.org/pdb.