Binary classification of proteins by a Machine Learning approach

27 Oct 2020

Prof. Osvaldo Gervasi - osvaldo.gervasi@gmail.com
Dott. Marco Simonetti - m.simonetti@unifi.it
Dott. Damiano Perri - damiano.perri@unifi.it

Abstract

The lesson is divided into two parts. The first part intends to be an introduction to machine learning and covers the basic notions for getting the idea of the subject, such as classification, regression, supervised learning and unsupervised learning, preparation of a training dataset, structure of a simple neural network, overfitting.

After that, we present a system based on a Deep Learning approach, by using a Convolutional Neural Network, capable of classifying protein chains of amino acids based on the protein description contained in the Protein Data Bank. Each protein is fully described in its chemical-physical-geometric properties in a file in XML format. The aim of the lecture is to design a prototypical Deep Learning machinery for the collection and management of a huge amount of data and to validate it through its application to the classification of a sequence of amino acids. We envisage applying the described approach to more general classification problems in biomolecules, related to structural properties and similarities.