PROTEINSECONDARYSTRUCTURE-CNN

Luca Angioloni
Department of Computer Engineering
University of Florence
Florence, Italy 50139
luca.angioloni@unifi.it

September 26, 2019

ABSTRACT

Protein structure prediction is one of the most important goals pursued by bio-informatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). In this work I approach the problem of protein's secondary structure prediction from the primary structure using deep learning and specifically Convolutional Neural networks (CNNs) using the amino acid sequences as inputs using a more powerful primary structure representation called "protein's evolutionary profiles". The model reaches state of the art performances with remarkable efficiency. The accuracy on the test set achieved with the best model is equal to 72% (Q8 Accuracy) and 68% (Q8 Accuracy) on the public benchmark CB513.

***K*eywords** Proteins · Secondary Structure · Prediction · Amino acids · Machine Learning

1 Introduction

Proteins are chains of amino acids joined together by peptide bonds. Many conformations of this chains are possible due to the many combinations of amino acids that might appear and rotation or folding of the peptide along the chain. It is these conformation changes that are responsible for differences in the three dimensional structure of proteins.

Protein structure prediction is one of the most important goals pursued by bio-informatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes) [1].

When we talk about the structure of proteins, four different structure levels are mentioned: the primary, secondary, tertiary and quaternary structure. (See figure 2)

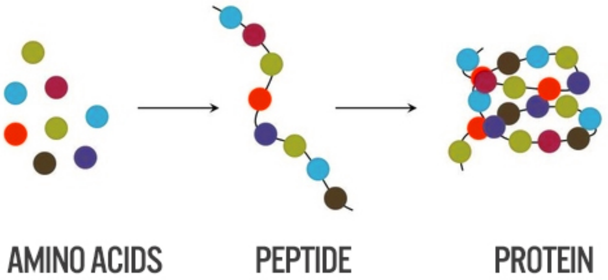


Figure 1: The amino acids are molecules that bond with each other forming chains called *peptides* (the bond itself is called *peptide*). Finally these chains assume 3D structure and shape forming the *proteins*

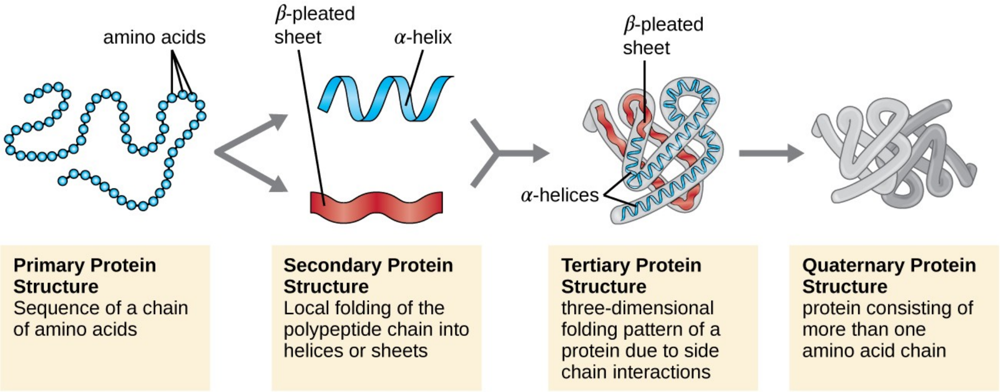

amino acids
β-pleated sheet
α-helix
β-pleated sheet
α-helices
Primary Protein Structure
Sequence of a chain of amino acids
Secondary Protein Structure
Local folding of the polypeptide chain into helices or sheets
Tertiary Protein Structure
three-dimensional folding pattern of a protein due to side chain interactions
Quaternary Protein Structure
protein consisting of more than one amino acid chain

Figure 2: This figure shows the four different levels of protein structures. The primary structure is the linear sequence of amino acids. The secondary structure is the three dimensional form of local segments of proteins. The tertiary structure describes the 3D structure of the protein molecule. The quaternary structure is only existing in cases where more than one peptide chain is part of the protein structure.

Protein primary structure is the linear sequence of amino acids in a peptide or protein.

Protein secondary structure is the three dimensional form of local segments of proteins. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its three dimensional tertiary structure. Both protein and nucleic acid secondary structures can be used to aid in multiple sequence alignment.

The tertiary structure is however particularly interesting as it describes the 3D structure of the protein molecule, which reveals very important functional and chemical properties, such as which chemical bindings the protein can take part in.

Predicting protein tertiary structure from only its amino acid sequence is a very challenging problem, but using the simpler secondary structure definitions is becomes more tractable [2].

I focused on the primary and secondary structure (SS), more specifically on using Convolutional Neural Networks (CNNs) for predicting the secondary structure of proteins given their primary structure.

2 Protein Structures and Protein Data

The primary structure of proteins are described by the sequence of amino acids on their polypeptide chain.

There are 20 natural occurring amino acids in the human body which, in a one letter notation, are denoted by: 'A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y'. 'A' standing for Alanine, 'C' for Cysteine, 'D' for Aspartic Acid etc. A 21st letter, 'X', is sometimes used for denoting an unknown or any amino acid. (See figure 3)

Instead of using the primary structure as a simple indicator for the presence of one of the amino acids, a more powerful primary structure representation has been used: Protein Profiles (See figure 4). These are used to take into account evolutionary neighborhoods and are used to model protein families and domains. They are built by converting multiple sequence alignments into position-specific scoring systems (PSSMs). Amino acids at each position in the alignment are scored according to the frequency with which they occur at that position [3].

A protein's polypeptide chain typically consist of around 200-300 amino acids, but it can consist of far less or far more. The amino acids can occure at any position in a chain, meaning that even for a chain consisting of 4 amino acids, there are 204 possible distinct combinations. In the used dataset the average protein chain consists of 208 amino acids.

Proteins' secondary structure determines structural states of local segments of amino acid residues in the protein. The alpha-helix state for instance forms a coiled up shape and the beta-strand forms a zig-zag like shape etc. The secondary structure of the protein is interesting because it, as mentioned in the introduction, reveals important chemical properties of the protein and because it can be used for further predicting it's tertiary structure. When predicting protein's secondary structure we distinguish between 3-state SS prediction and 8-state SS prediction.

For 3-state prediction the goal is to classify each amino acid into either:

- alpha-helix, which is a regular state denoted by an 'H'

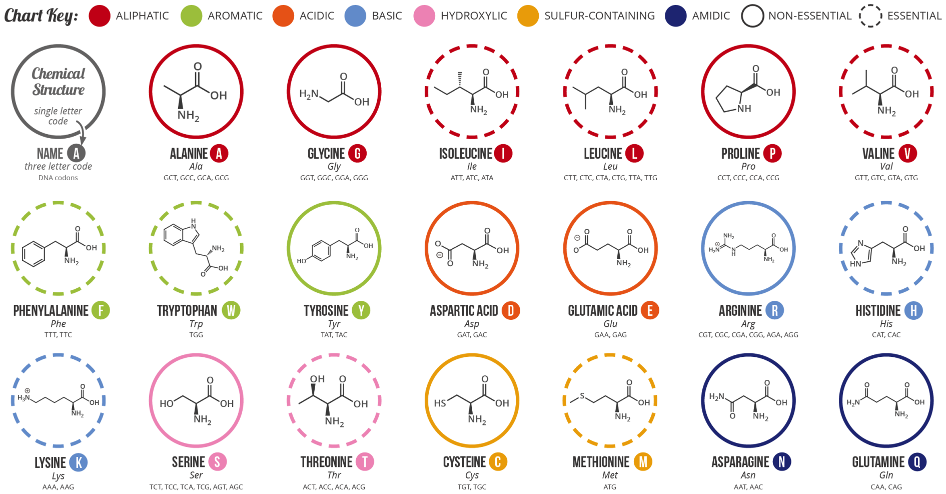


Figure 3: A table showing the 20 natural occurring amino acids in the human body and some of their properties.

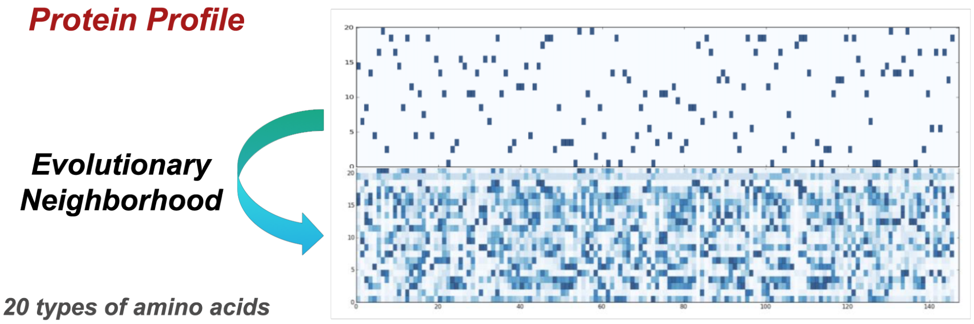


Figure 4: Protein Profiles: they are built by converting multiple sequence alignments into position-specific scoring systems (PSSMs). Amino acids at each position in the alignment are scored according to the frequency with which they occur at that position [3].

- beta-strand, which is a regular state denoted by an ’E’
- coil region, which is an irregular state denoted by a ’C’

The letters which denotes the above secondary structures are not to be confused with those which denotes the amino acids.

For 8-state prediction, Alpha-helix is further sub-divided into three states: alpha-helix (’H’), 310 helix (’G’) and pi-helix (’I’). Beta-strand is sub-divided into: beta-strand (’E’) and beta-bride (’B’) and coil region is sub-divided into: high curvature loop (’S’), beta-turn (’T’) and irregular (’L’) [2]. This results in the following 8-states:

- E = extended strand, participates in β ladder
- B = residue in isolated β-bridge
- H = α-helix
- G = 3-helix (3-10 helix)
- I = 5-helix (π-helix)
- T = hydrogen bonded turn
- S = bend
- _ = loop (any other type)

For the scope of this project the more challenging 8-state prediction problem has been chosen.

2.1 Dataset

The dataset used is CullPDB data set, consisting of 6133 proteins each of 39900 features. The 6133 proteins x 39900 features can be reshaped into 6133 proteins x 700 amino acids x 57 features.

The amino acid chains are described by a 700 x 57 matrix to keep the data size consistent. The 700 denotes the peptide chain and the 57 denotes the number of features in each amino acid. When the end of a chain is reached the rest of the vector will simply be labeled as ’No Seq’ (a padding is applied).

Among the 57 features, 22 represent the primary structure (20 amino acids, 1 unknown or any amino acid, 1 ’No Seq’ -padding-), 22 the Protein Profiles (same as primary structure) and 9 are the secondary structure (8 possible states, 1 ’No Seq’ -padding-).

The Protein profiles where used instead of the amino acids residues.

For a more detailed description of the dataset and for download see [4].

In a first phase of research the whole amino acid sequence was used as an example (700 x 22) to predict the whole secondary structure (label) (700 x 9).

In the second phase, local windows of a limited number of elements, shifted along the sequence, were used as examples (cnn_width x 21) to predict the secondary structure (8 classes) in a single location in the center of each window (The ’No Seq’ and padding were removed and ignored in this phase because it wasn’t necessary anymore for the sequences to be of the same length).

The dataset (of 6133 proteins) was divided randomly into training (5600), validation (256) and testing (272) sets, as suggested by [5] for the results shown below.

However different splits of the dataset with different sizes have been tested with equal results.

2.1.1 Implementation

This project was implemented using the Keras framework with the Tensorflow backend. Two main approaches have been explored:

1. Use the whole protein sequence (primary structure) as an example for the CNN, with an output of dimension 700 x 9, the sequence of the predicted secondary structure.
2. Use local windows of a limited number of elements as an example for the CNN which is shifted along the sequences, predicting for each window the secondary structure in a single location (8 classes), in the center of each window.

2.2 Whole protein prediction

This simple model consists of 3 main 1D Convolutional Layers regularized with Dropout that use the whole protein sequence as input. The input size is 700 amino acids (that sometimes contains padding) with 9 channels corresponding to the 8 possible states for each amino acid plus the so called padding state.

This was a first prototype, with a low number of parameters (125.512 trainable parameters).

A major problem with this approach, was the fact that the padding added to shorter sequences, still influenced the loss, calculated on the whole output sequence (’categorical_crossentropy’ loss from tensorwlow was used).

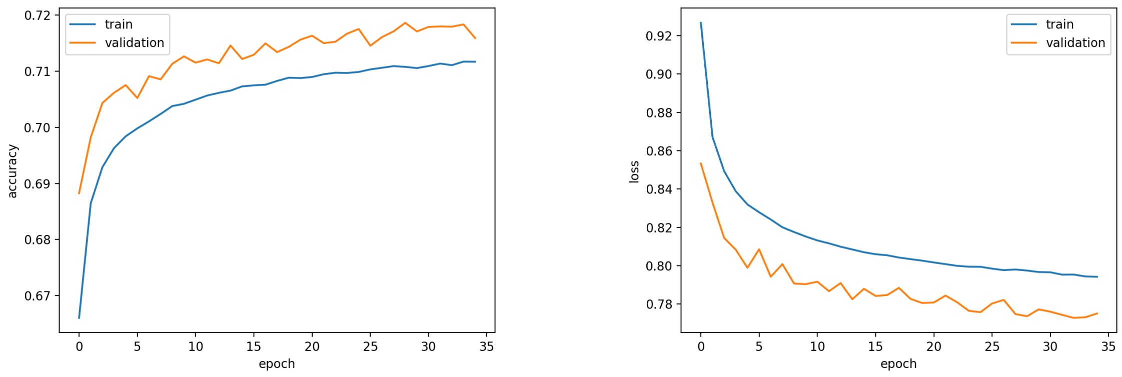
This required the creation of a custom loss to take into account the outputs from the padding region, which is of different shape for each example.

Soon this approach was abandoned.

2.3 Window CNN

This model consists of 3 main 1D Convolutional Layers with DropOut and Batch Normalization followed by 3 Fully connected Layers.

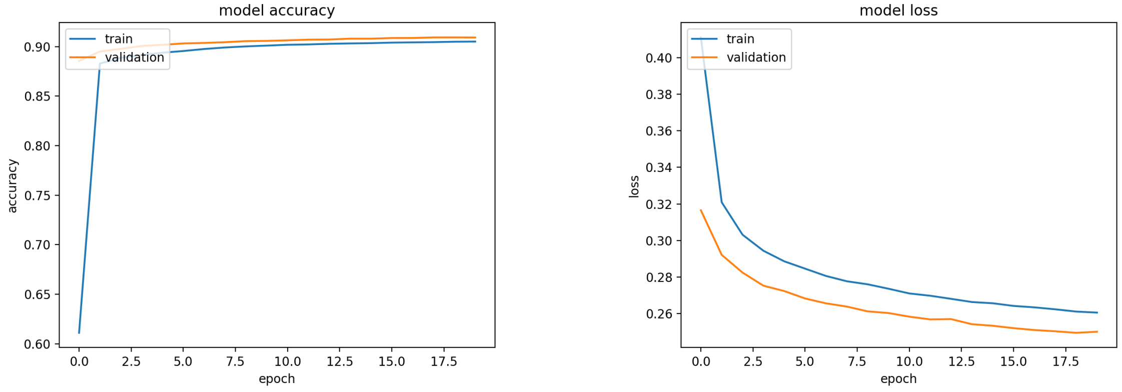
The size of the window has been chosen to be bigger than 11 because the average length of an alpha helix is around eleven residues and that of a beta strand is around six (See references [6]). Multiple even sizes from 11 to 23 were tested, with 17 yielding the best results (performance/training time trade off).



(a) Window CNN Accuracy (Q8 Accuracy)

(b) Window CNN Loss

Figure 5: Learning curves: (a) Accuracy at each epoch for the Window CNN approach, (b) Loss value at each epoch for the Window CNN approach.



(a) Whole protein CNN Accuracy

(b) Whole protein CNN Loss

Figure 6: Learning curves: (a) Accuracy at each epoch for the Whole protein approach, (b) Loss value at each epoch for the Whole protein approach.

This model has 232.552 parameters (Trainable params: 231.912) and was trained on 946494 samples, validated on 120704 samples (windows).

3 Results

The Window CNN has been trained with the CullPDB dataset split like described in section 2.1 for 35 epochs (on CPU in approximately 6 hours).

The learning curves for this approach are shown in figure 5.

The accuracy on the test set achieved with this model is equal to 0.721522 (Q8 Accuracy), which is comparable to the results obtained in [5] and [6] using different techniques.

The model has also been trained with the filtered version of the dataset: *CullPDB6133+filtered* available at [4] and tested with the public benchmark CB513. The accuracy obtained is equal to 0.6833 (Q8 Accuracy), again comparable with [5] and [6].

Whole protein prediction This model has been trained with the CullPDB dataset split like described in section 2.1 for just 20 epochs (on CPU in approximately 25 minutes).

The learning curves are shown in figure 6. In this model the loss is calculated without taking the padding into account, so the resulting values are biased.

The accuracy on the test set achieved with this model is equal to 0.6966 (Q8 Accuracy), which is pretty close to the results obtained with the Window CNN in a small fraction of the time required for the Window CNN.

Moreover the accuracy obtained training on the filtered dataset and testing on the CB513 dataset, is equal to 0.6557 (Q8 Accuracy).

4 Conclusions

In conclusion the model developed was able to replicate the state of the art results with remarkable efficiency. The accuracy on the test set achieved with the best model is equal to 0.721522 (Q8 Accuracy) on the testset as described in section 2.1 and 0.6833 (Q8 Accuracy) on the public benchmark CB513.

The code developed for this project is available on GitHub at: https://github.com/LucaAngioloni/ProteinSecondaryStructure-CNN.

References

[1] Wikipedia. Protein structure prediction — Wikipedia, The Free Encyclopedia. [Online; accessed 04-April-2019] https://bit.ly/2VlY8mF

[2] Wikipedia. Protein secondary structure — Wikipedia, The Free Encyclopedia. [Online; accessed 04-April-2019] https://bit.ly/2WRirsJ

[3] EMBL-EB (The home for big data in biology). What are profiles? https://bit.ly/2Uj1RVC

[4] Zhou, Jian and Troyanskaya, Olga G. Dataset website. http://www.princeton.edu/%7Ejzthree/datasets/ICML2014/

[5] Zhou, Jian and Troyanskaya, Olga G. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *arXiv preprint arXiv:1403.1347*, 2014.

[6] Wang, Sheng and Peng, Jian and Ma, Jianzhu and Xu, Jinbo. Protein secondary structure prediction using deep convolutional neural fields. In *Scientific reports, 2016 6th volume, Nature Publishing Group*, pages 18962.