1. **Title: Prediction of 8-state protein secondary structures by a novel deep learning architecture.**

**Author:** Buzhong Zhang, Jinyan Li and Qiang Lü

**Input Features:** Four types of features, including a position-specific scoring matrix (PSSM), protein coding features, conservation scores, and physical properties, are used to characterize each residue in a protein sequence

PSSM: Ran PSI-Blast to search the NCBI non-redundant database through three iterations with E-value=0.001.

Physical Properties: steric parameters (graph-shape index), polarizability, normalized van der Waals volume, hydrophobicity, isoelectric point, helix probability, and sheet probability.

**Method:** Introduces a revolutionary deep learning architecture that uses an integrated synergy of prediction by a residual network, convolutional neural network and bidirectional recurrent neural network.

**Result:** On the benchmark CB513 dataset, their proposed deep network obtained 71.4% accuracy for 8-state prediction and 74% accuracy for ensemble learning.

1. **Title: Predicting the protein structure using Random Forest approach**

**Author:** Charu Kathuria, Deepti Mehrotra, Navnit Kumar Misra

**Input Features:** The dataset consists of amide frequencies of known proteins which actually help in determining the structure that is related to its functional behavior and features. These are the observed frequencies which are collected using Infrared Spectroscopy. The dataset used is broadly classified in alpha and non-alpha structures as during preprocessing all alpha structure polypeptides are taken as alpha in their structure attribute and rest all other structure of polypeptides are considered as non alpha.

**Method:** Random Forest (RF) technique of classification is used in R programming language by including its respective package named randomForest to classify in alpha and non-alpha structure.

**Result:** The model accuracy is validated with ROC curve and area under the curve has been calculated having value 0.963.

1. **Title: Protein Secondary Structure Prediction : A Review of Progress and Directions**

**Author:** Tomasz Smolarczyk, Irena Roterman-Konieczna and Katarzyna Stapor

**Input Features:**

**Protein Coding Features:**

The original protein sequence for amino-acid residues is the basic input feature used for protein secondary structure prediction. It is usually a sequence of 20, 21 or 22- dimensional feature vectors encoding the types of the amino acids in the protein. Each feature vector is a sparse one-hot vector and only one of its elements is non-zero.

**Multiple Sequence Alignment Profile:**

The sequence alignment of homologous proteins corresponds with their structural alignment and aligned residues usually have similar secondary structures. There are many methods that generate multiple sequence alignment profiles, such as PSI-BLAST, PSI-Search, HMMER3, AMPS or CLUSTALW.

**Physiochemical Properties:**

Since the number of parameters that could be used for secondary structure prediction is so large, researchers use only a subset of available parameters. Some of them are:

• Steric parameter (graph-shape index), polarizability, normalized van der Waals volume, hydrophobicity, isoelectric point, helix probability, sheet probability,

• Net charges, hydrophobic, side-chain mass,

• Eight groups of physicochemical parameters were used in: hydrophobic, hydrophilic, polar, non-polar, small, large, charged and uncharged. According to, hydrophobic, hydrogen bond and charge properties are regarded as the ones with a relatively large impact on protein secondary structure.

**Conformation Parameters:**

Conformation parameters are the proportions that amino acids tend to form secondary structures. The conformation parameters for each amino acid Sij are defined as follows:

Sij = (aij / ai )

Where i=1,2,….,20 and j=1,2,3.

where i indicates the 20 amino acids, and j indicates one of the three types of secondary structures: H, E and C. Number of ith amino acid with the jth secondary structure is represented by aij whereas ai represents the number of ith amino acid in a data set.

**Windowing:**

It is a common practice in protein secondary structure prediction to use a sliding window of neighboring residues to get more surrounding information of residues since the second generation of methods. A window is a short segment of a protein and in the middle of it, there is an amino acid for which we want to predict secondary structure. Common sizes for windows are: 11, 13,15, 17 or 21 and the size is often experimentally selected for the specific model.

**Method:**

Converted 8-state to 3-state secondary structure and analysed 15 different methods. The methods were different ANN, CNN, BRNN.

1. **Title: Protein Secondary Structure Prediction Using Graph Neural Network.**

**Input Feature:** Extracted feature vector from primary protein sequences of amino acid by using one hot encoding, graph neural network and at last window padding.

**Method:** Graph neural network with support vector machine.

**Accuracy:** 89% but lack of massive data processing capacity.

**MY Thesis Title: Protein Secondary Structure Prediction Using Graph Neural Network with Ensemble Learning**

**Input Features:** Extracted feature vector from primary protein sequences of amino acid by using one hot encoding, graph neural network and at last window padding. And another feature is used conformation parameter -

Conformation parameters are the proportions that amino acids tend to form secondary structures. The conformation parameters for each amino acid Sij are defined as follows:

Sij = (aij / ai )

Where i=1,2,….,20 and j=1,2,3,4,5,6,7,8.

where i indicates the 20 amino acids, and j indicates one of the eight types of secondary structures. Number of ith amino acid with the jth secondary structure is represented by aij whereas ai represents the number of ith amino acid in a data set.

**Method:**

Applied graph neural network for feature extraction. For classification SVM and Random Forest are used. SVM is used for smaller length protein sequence and Random Forest is used for larger length protein sequence. At last result of two classifier is ensembled for 8-state secondary structure prediction.

**Accuracy:**

85.73% accuracy with both smaller and larger length data.

|  |  |
| --- | --- |
| Window size | accuracy |
| 5 | 90.93% |
| 7 | 91.42% |
| 9 | 91.60% |
| 11 | 91.81% |
| 13 | 92.13% |
| 15 | 92.23% |
| 17 | 92.30% |
| 19 | 92.43% |
| 21 | 92.62% |