“Short-length peptides contact map prediction using Convolution Neural Networks”

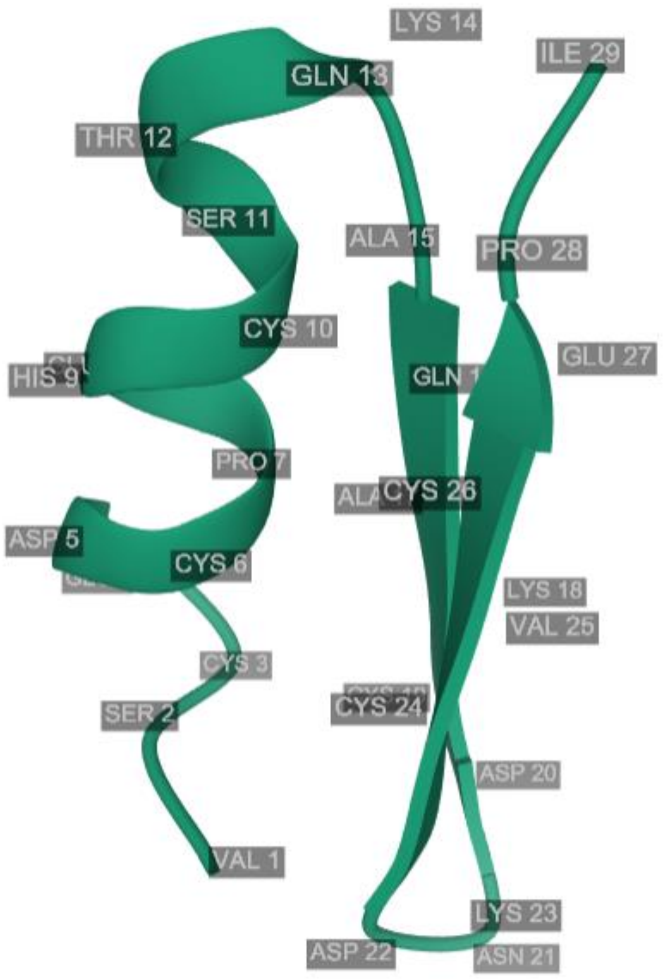
Author:
Maminov Artem

ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
Информатика
и Управление
РОССИЙСКОЙ АКАДЕМИИ НАУК

ВЦ
РАН

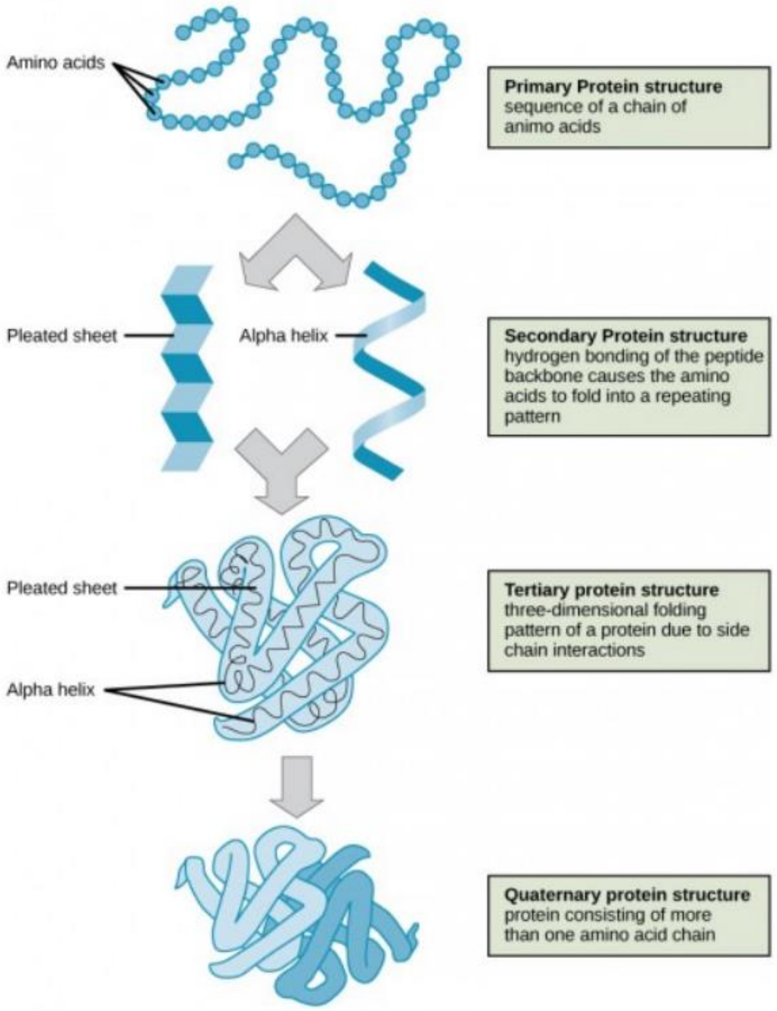
Introduction

- The protein is denoted as sequence of amino acids residues
- The features of the protein are determined not only by its chemical composition but also by its tertiary structure
- The main aim of *folding* problem is to determine right conformations of the protein based on its amino acids sequence



Introduction

- The first folding methods were based on biochemical and physical modeling.
- Later the methods using databases and gomological proteins appear.
- The last and the most useful methods uses ML techniques to predict some structure information about proteins.



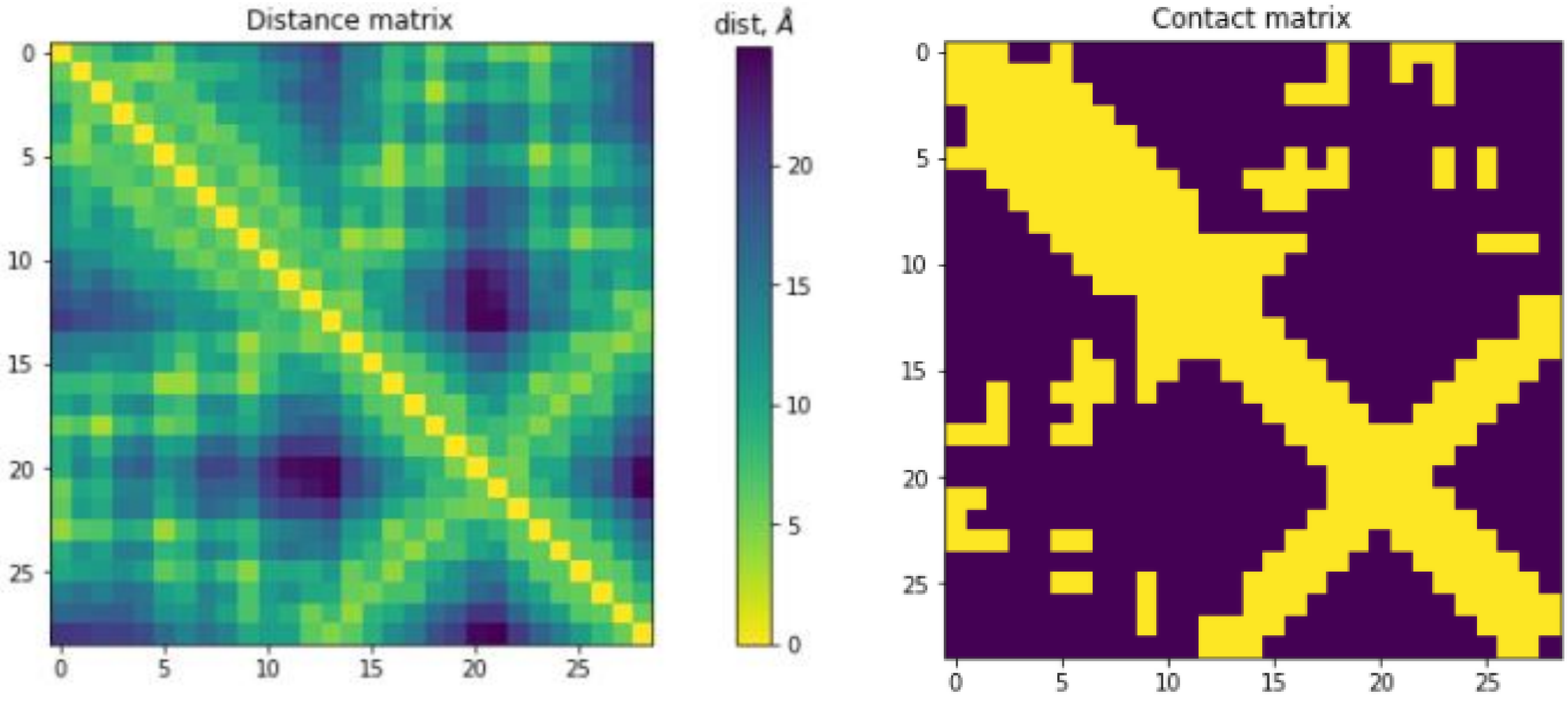
Primary Protein structure sequence of a chain of animo acids

Secondary Protein structure hydrogen bonding of the peptide backbone causes the amino acids to fold into a repeating pattern

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

Contact and distance matrices



Features

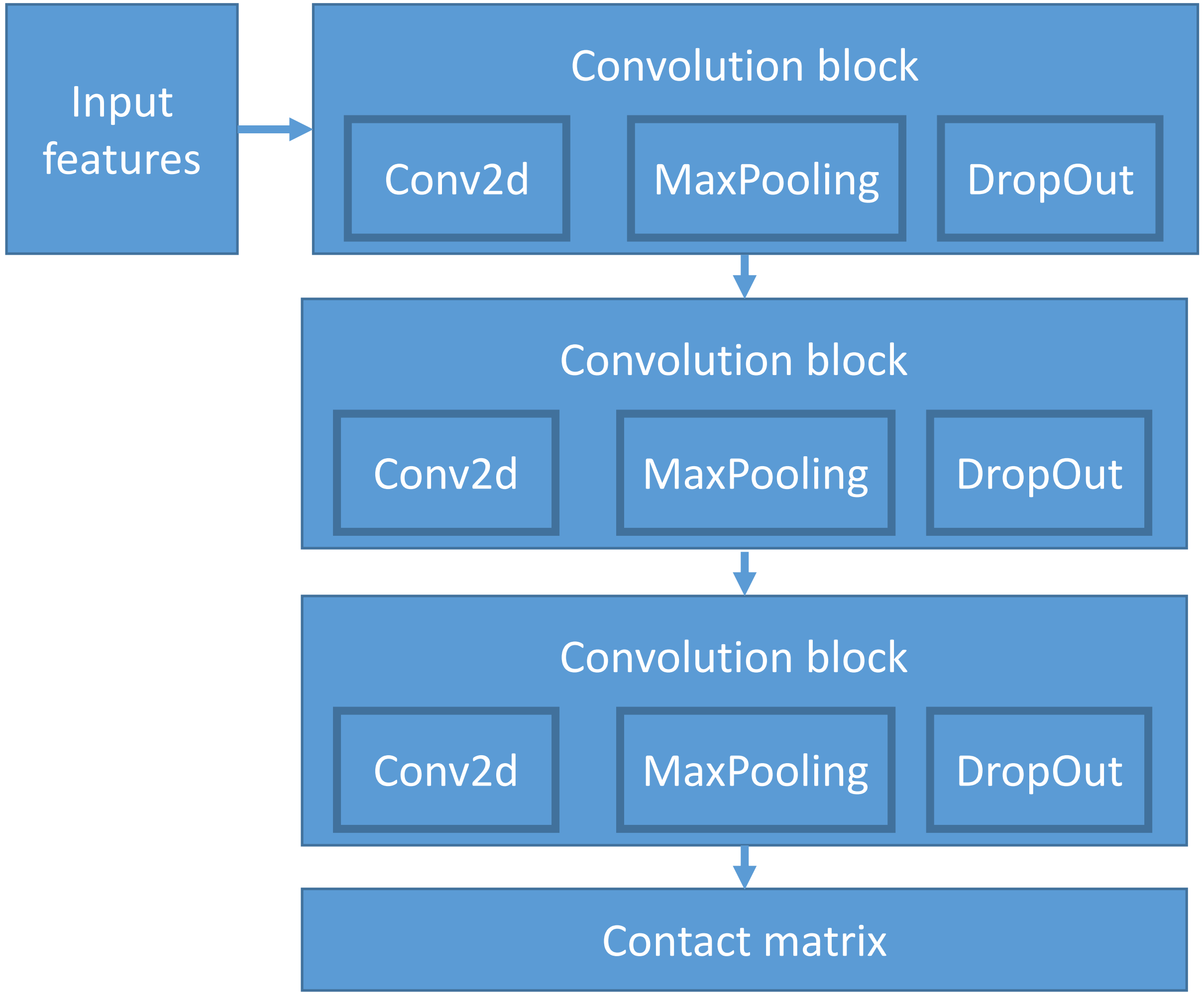
- Four features were used in this research: FASTA-format of the sequence, PSSM matrix, secondary structure, solvent accessibility, polarity of the amino acids and the type of radical
- PSSM matrix, secondary structure and solvent accessibility were calculated using SCRATCH instrument

Features

- All categorical features were one-hot encoded
- According to using Keras framework to build CNN network all features and contact matrices were padded with zeros to make all feature matrix and all target matrix have the same sizes.
- At the result we have feature dataset of size 386x30x56, where 386 – the number of the samples, 30 – the upper limit of the sequence legth and 56 – the number of the features (20 – PSSM, 20 – FASTA, 3 – secondary structure, 1 – solvent accessibility, 7 – classification of the radical and 5 – classification of the polarity)
- The target dataset is represented as matrix 386x435x2
- The dataset was splitted into training, validation and test sets

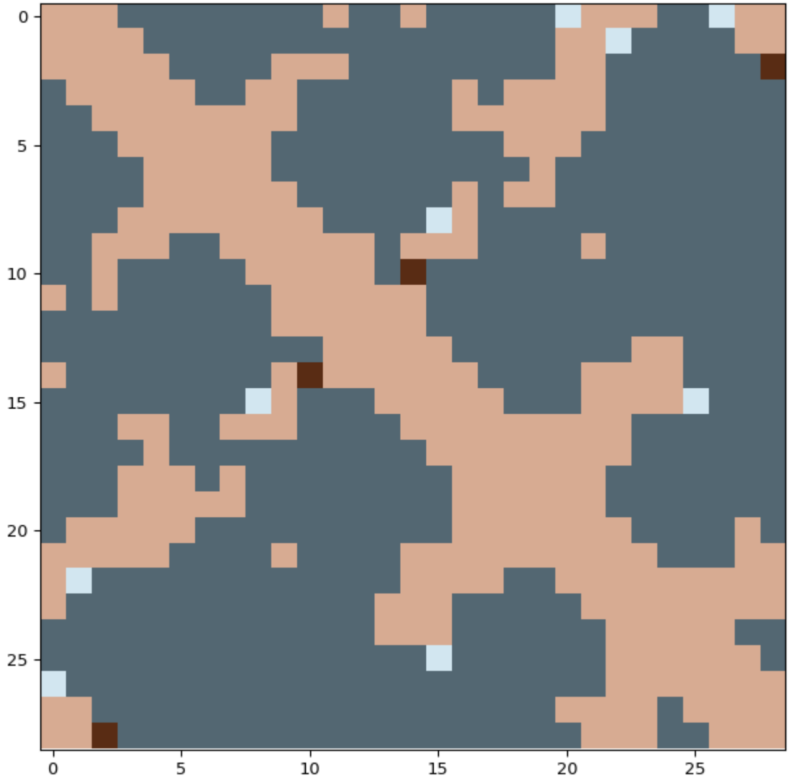
CNN

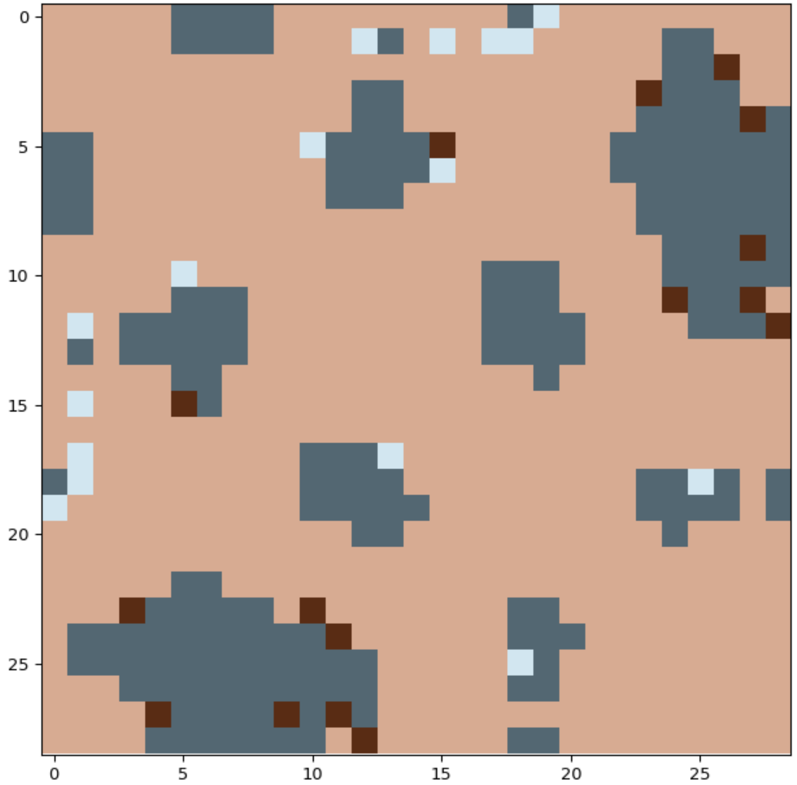
- CNN consists of 3 convolutional blocks with max polling and dropout layers
- Binary cross entropy function was used as loss function
- Adam optimizer was used as optimization method
- LeakyReLU was used as activation function



Contact map prediction results

Threshhold	F1-score	Precision	Recall
8 Å	0.78	0.86	0.73
12 Å	0.83	0.85	0.83

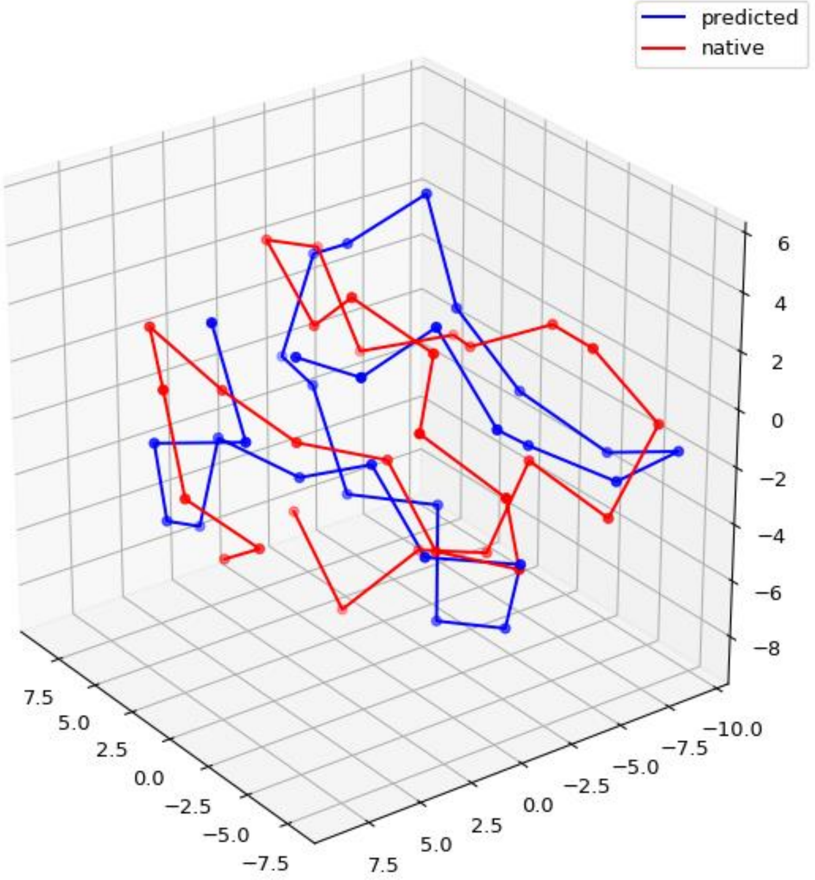




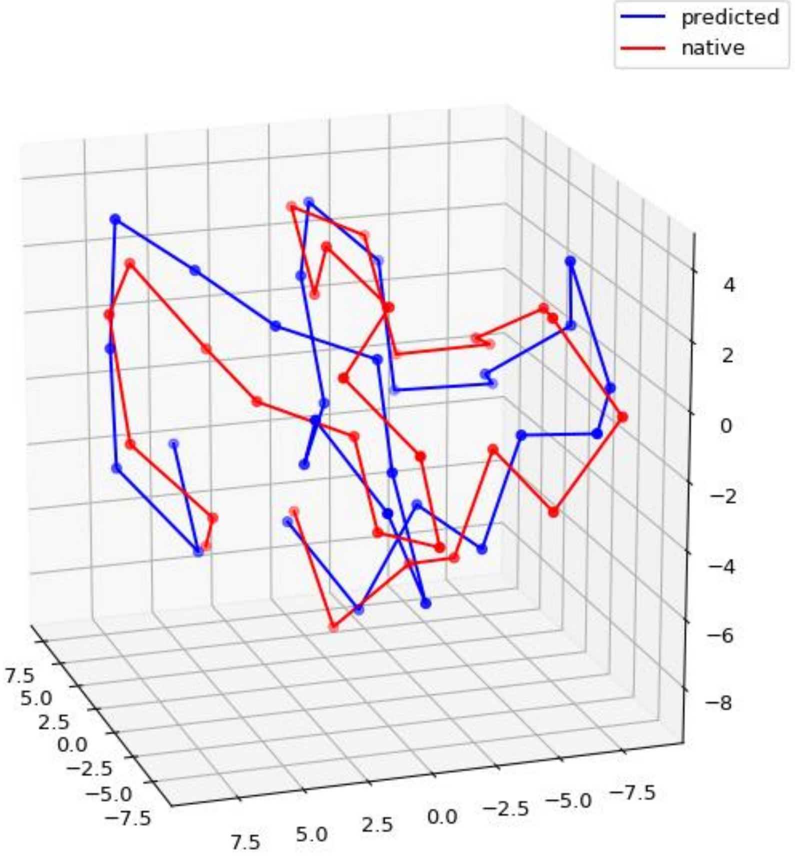
Tertiary structure reconstruction

- FT-COMAR program was used to reconstruct backbone atoms positions from contact matrices for 8 and 10 $\dot{A}$ on left and right plots accordingly

RMSD = 6.3

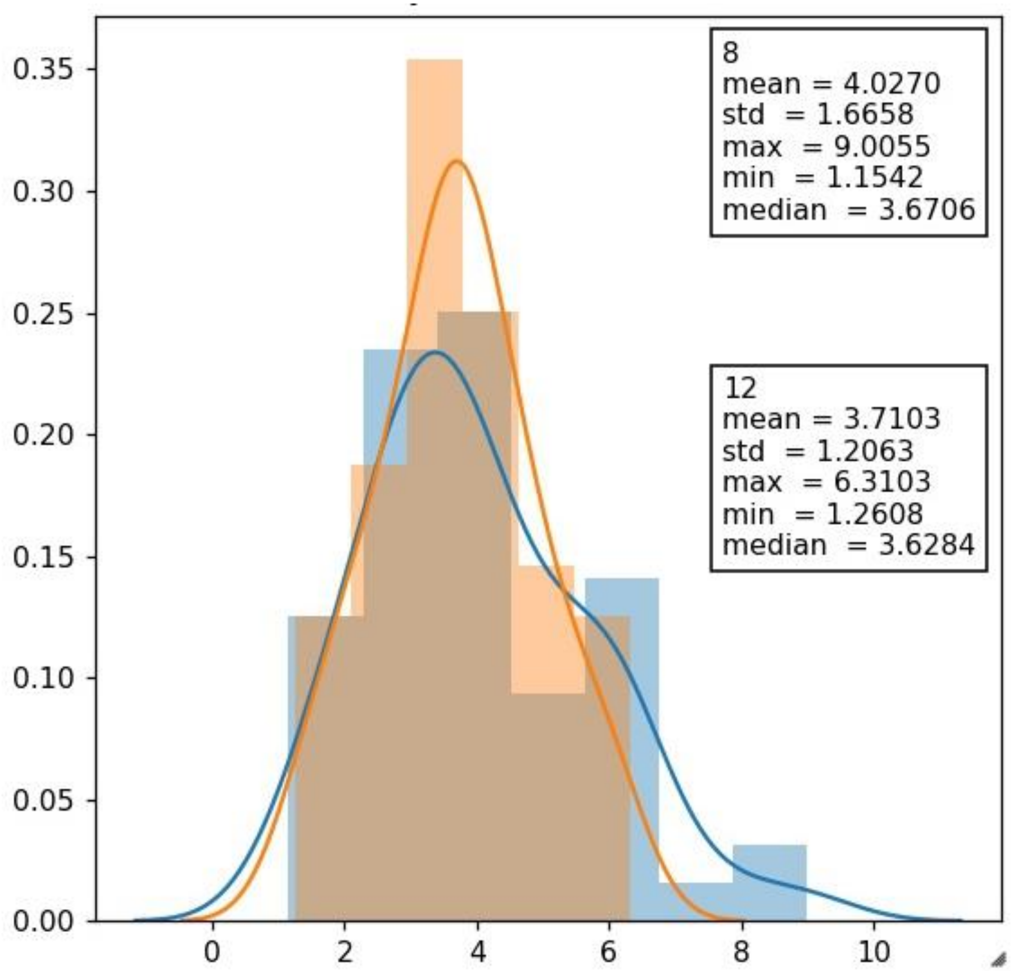


RMSD = 2.5

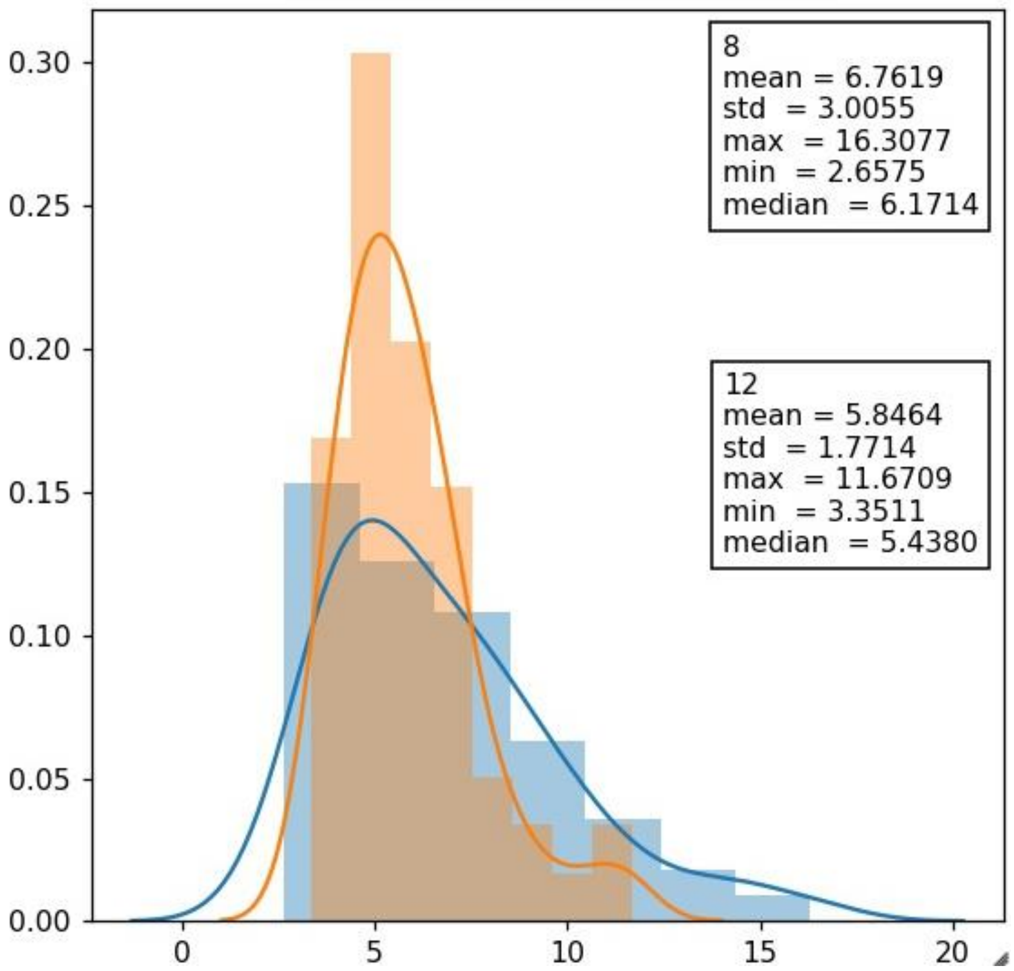


Reconstruction results

Native contact matrix reconstruct



Predicted contact matrix reconstruct



Conclusion

- At the result the complex for contact matrix prediction was created and tested on short proteins
- The experiment was done only on short proteins with simple CNN architecture and shows further perspective with longer proteins and more complex architectures
- The simplest features were used, what shows good results and can be improved with different MSA, statistical and profile features
- F1-score equal to 0.78 was achieved, what shows, that even with simple CNN and small amount of features good result can be achieved without using computing clusters