PROCEEDINGS OF SCIENCE

Short-length peptides contact map prediction using Convolution Neural Networks

Artem Maminov[a,b,*]

[a] *HSE University, 20 Myasnitskaya Ulitsa, 101000, Moscow, Russian Federation*

[b] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44/2, Vavilova street, 119333, Moscow, Russian Federation*

E-mail: artem_maminov@mail.ru

In this article, it is considered an approach for predicting the contact matrix (contact map) for short-length peptides. Contact matrix is two-dimensional representation of the protein. It can be used for tertiary structure reconstruction or for starting approximation in energy minimization models. For this work, peptides with a chain length from 15 up to 30 were chosen to test the model and simplify the calculations. Convolutional neural networks (CNNs) were used as a prediction tool according to the fact that the feature space of each peptide is presented as a two-dimensional matrix. SCRATCH tool was used to generate the secondary structure, solvent accessibility, and profile matrix (PSSM) for each peptide. CNN was implemented in the Python programming language using the Keras library. To work with the common PDB-format, which presents the structure information of proteins, the BioPython module was used. As a result, training, validation and test samples were generated, the multilayer multi-output convolutional neural network was constructed, which was trained and validated. The experiments were conducted on a test sample to predict the contact matrix and compare it with native one. To assess the quality of prediction, conjunction matrices for the threshold of 8 and 12 $\dot{A}$ were formed, the metrics F1-score, recall and precision were calculated. According to F1-score, we can observe, that even with small neural network we can acheve quite good results. At the final step FT-COMAR tool was used to reconstruct tertiary structure of the proteins from its contact matrix. The results shows, that for reconstructed structures from 12 threshhold contact matrix, RMSD metric is better.

**** The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022) ****
**** 6-8 July 2022 ****
**** JINR, Dubna, Russia ****

*Speaker

PoS(DLCP2022)016

1. Introduction

The problem of protein folding is one of the most discussed problems in modern computational biology. Competitions are periodically held in various sub-topics of folding, where the best scientists from all over the world present their methods and solutions, which are still far from ideal results. This work is the first step for the development of a modern software package for protein folding. Each protein is made up of amino acid residues that appear from the formation of a peptide bond between two amino acids. In total, 20 standard amino acids are involved in the formation of peptides (some researchers distinguish 2 more amino acids). Amino acids are composed of the backbone (carbon (C), hydrogen (H), oxygen (O), nitrogen (N)) and a radical.

Thus, proteins are composed of a sequence of amino acid residues and can be written in the format of abbreviated amino acid notation (also know as FASTA format). The next levels of the organization in protein are secondary, tertiary and quaternary structures (Fig. 1). The secondary structure is a stable, locally ordered fragments of the polypeptide chain, which is constructed by hydrogen bounds. The most common secondary structures are α-helices and β-sheets. Up to 8 different secondary structures types are distinguished in more extensive classifications. The task of predicting and modeling the secondary structure has been successfully solved and there are several different software systems, one of which will be used later. And finally, the most important task,

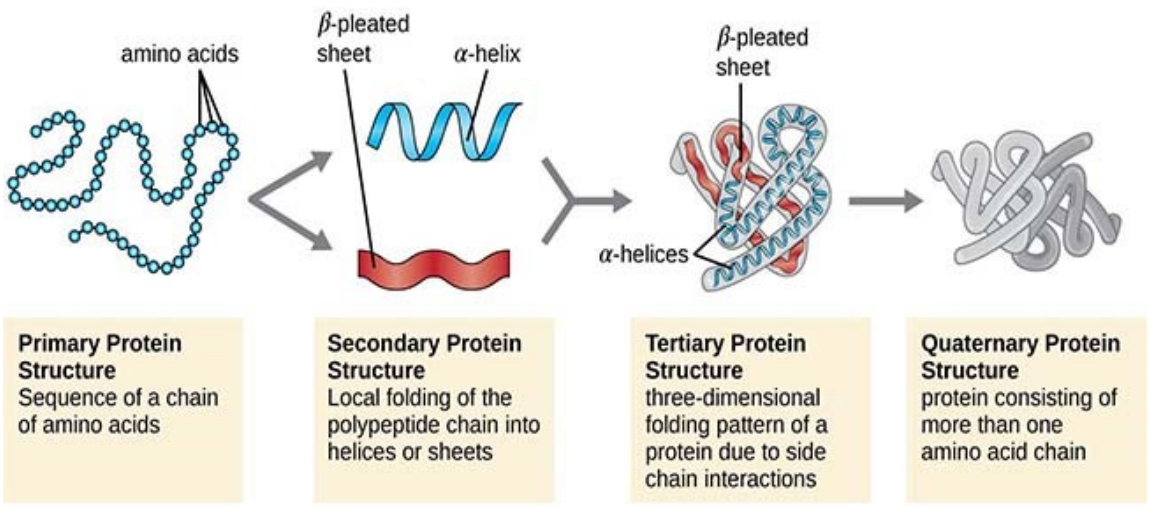


Figure 1: Different protein structures: Primary, Secondary, Tertiary and Quaternary

which is usually understood as the problem of protein folding, is the task of predicting the tertiary structure of the protein, that is, how the protein folds and is located in space. In nature, proteins conformation depends on physicochemical reactions and effects which arise between atoms. There are a huge number of proteins and scientists obtain the tertiary structure of proteins through special expensive experimental and technical methods: X-ray structural analysis and nuclear magnetic resonance. The task of folding is to create an algorithm for protein tertiary structure reconstruction based only on the amino acid sequence. At the moment, there is no tool, which can fully solve this problem. The modern methods of protein folding will be discussed in more details in the following

section.

In this work, attention is paid to the prediction of the contact matrix (the binary matrix which indicates if two residues are close to each other in space) for proteins with a chain length less than 31 amino acid residues. We want to show, that even with small neural networks, you can achieve good results and reconstruct tertiary sctructure of the protein. The quaternary structure isn't discussed in this article and refers to the docking problem.

2. Related works

The solution of the protein folding problem is extensively studied by scientists around all the world and various methods are proposed. The problem of the folding was found in the middle of the XX-th century. Max Perutz and John Kendrew invented X-ray crystallography method and were awarded with Nobel prize [12]. This method is used for determination of the space structure of the proteins, but the scientists could not explain the mechanism of the protein folding. Starting from this the problem of the protein folding was considered.

In 1968 Cyrus Levinthal raised the question (Levinthal paradox) about the speed of the proteins folding: how can proteins fold to the native state so fast if there are too many variants, which could not be handled with adequate time. This paradox was successfully solved, when more information and more research was done about potential energy in protein.

In 70-80 years the PDB database was established [2]. Multiple-sequence alignment (MSA) method was invented to find the homologous proteins in databases. The essence of MSA method is to determine how similar several proteins are, for which their amino acid sequence needs to be aligned and compared. This technique has been found wide application in the assessment of secondary and tertiary structures of proteins, and various discussions about the computational complexity of the algorithm have led to the creation of various modifications and improvements. Meanwhile, Michael Levitt developed the biochemical method with simplified molecular geometry for protein energy minimization [14].

At the end of the XX-th century Critical Assessment of Techniques for Protein Structure Prediction (CASP competition) was established [7]. This competition unites scientists from all the world to present their methods and approaches in different sections of the protein folding. At the same time the first tools for protein alignment and homologous search were developed.

At the beginning of the 21-th century there were (*ab initio*, *de novo*) researches based on biophysical methods [16]. The models were built on methods of molecular dynamics and a deep understanding of the biochemical processes that occur in proteins between atoms. However, these methods were not successful in solving the folding problem due to extremely high computational complexity and low accuracy [9]. The first projects of the distributed computing for protein structure prediction (such as Rosetta@Home [4], Folding@Home [1]) were maintained.

On the other hand, recently, methods based on the analysis of known protein conformations for modeling the new ones began to prevail. One of these methods is protein homology modeling. Since the last century the databases were significantly supplemented (Figure 2), such methods became very popular. These methods are based on comparing proteins with known structures in databases and on the assumption that they are homologous, which makes it possible to create different models with extensive initial data [8]. This class of predictive techniques is more accurate than *de novo*

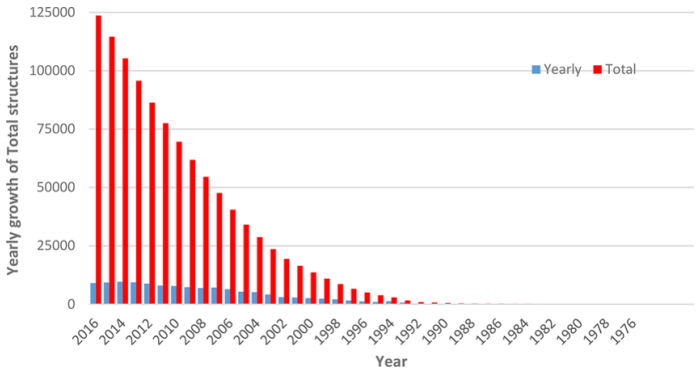


Figure 2: The growth of the number of the proteins in PDB database

prediction, but it is also not the solution of the problem because it is based on protein databases that cannot cover the entire variety of existing polypeptides.

In recent years, with the development of machine learning, scientists have paid their attention to the appearance of new tools. The main idea is to compose a set of features for training a model on proteins with a known tertiary structure and predict it, which is typical supervised learning problem. According to the results of the CASP 13 held in 2018, the team from the DeepMind laboratory with the AlphaFold project showed the best result [17]. Their method of protein folding is based on the prediction of the distribution of the distances between amino acids residues and torsion angles. Then they minimize the potential, which is constructed from predicted distributions and reconstruct tertiary structure of protein. In 2020 the AlphaFold2 was presented [11]. The new program complex shows extremely good results comparing previous years (Figure 3).

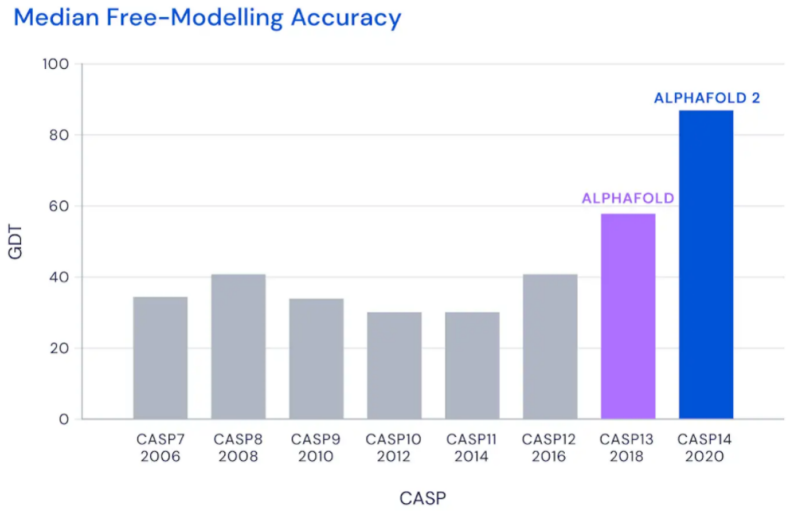


Figure 3: The AlphaFold and AlphaFold2 results at CASP13 and CASP14 respectively

Several other approaches should also be mentioned: work [19] and the resulting RaptorX web tool based on the prediction of the contact matrix and the distance matrix, which took first place in the CASP12. The next work [10] was used as first material for understanding the implementation

of neural networks for the folding problem. The main aim of this article is classification of proteins into 1195 different stable conformations from the SCOP protein database. Convolution neural networks were also used as a prediction tool, and the DeepSF online service was created based on the results of the work. And the last one is the first work [6], in which methods of deep neural networks were firstly used to predict the contact matrix. The researchers had significant success in the CASP 8 and CASP 9 competitions in predicting the contact matrix and their method have been incorporated into the SCRATCH tool.

3. Dataset description

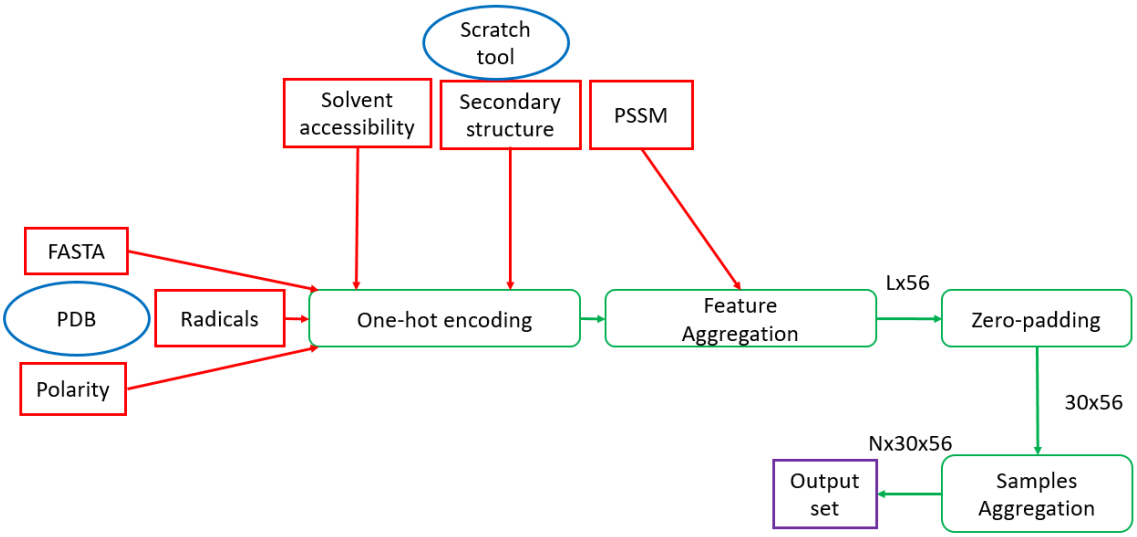


Figure 4: Features aggregation scheme

Based on the previous works, described in details in previous section, it was decided to choose the secondary structure, the FASTA-format of the protein, the PSSM matrix, solvent accessibility, polarity and classification by the type of radical as the features for training. Due to the fact that the proteins in the PDB database contain various errors, inaccuracies, omissions, a dataset of "good" proteins from the PDB was formed. This dataset doesn't have any breaks, gaps and consists only of one peptide chain. In Figure 4 the scheme of the features aggregation is depicted. To generate all necessary features, the SCRATCH tool [15] was used, since it contains all the necessary utilities. A sequence of secondary structures, a sequence of solubility classes and a PSSM matrix were generated for each protein.

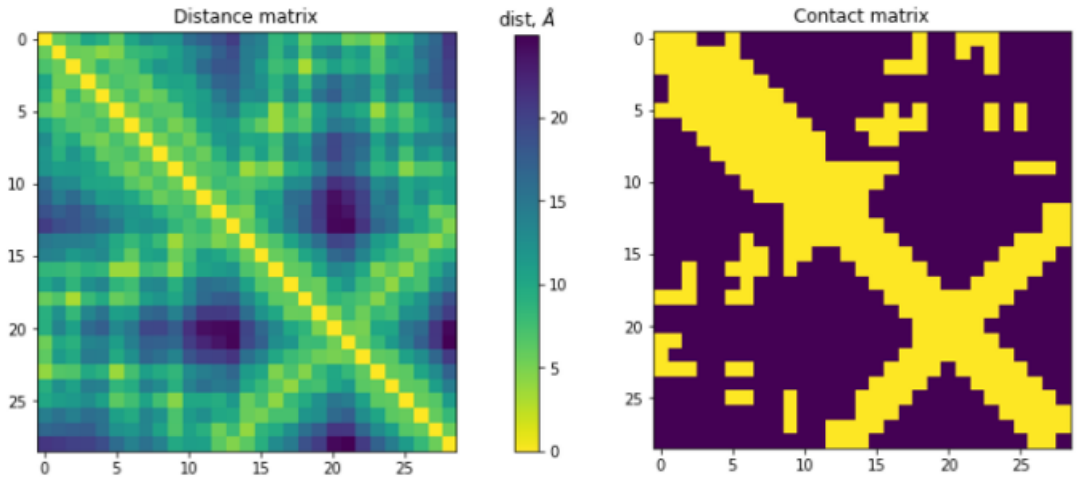
The next step was the one-hot-encoding of the secondary structure, solubility, classes of radicals, polarity and FASTA-sequence since most machine learning algorithms work better with classified features. As a result, each protein corresponds to the $L \times 56$ matrix, where L is the protein length, the first twenty elements are the encoded FASTA format, the next twenty are the PSSM matrix, then the encoded secondary structure (3 classes), encoded solubility (2 classes), encoded polarity (7 classes) and encoded types of radicals (5 classes). Then we apply so-called "zero-

padding" technique to make all proteins the same size. As a result, we obtain a three-dimensional array of "good" proteins with size $N \times 30 \times 56$, where N is the number of samples in dataset and $N = 781$. The dataset is splitted into 3 parts: training (70%), validation (15%) and testing (15%).

4. Contact map prediction

Contact matrices were calculated for all proteins in the dataset:

- proteins downloaded from PDB
- the distances between all beta-carbons (C_β) were calculated, the distance matrix was obtained (Fig. 5a)
- the matrix of distances is converted into a matrix of contacts (Fig. 5b): there is a contact if the distance between amino acids residues is less than the threshold value (the threshold value in this work is equal to 8 $\dot{A}$ and 12 $\dot{A}$) or it is absent if it is greater.



(a) (b)

Figure 5: Example of distance matrix (a) and contact matrix with threshold 8 $\dot{A}$ (b) for protein 1ACW

Thus, we moved on to the problem of binary classification: we need to predict the presence of contact or its absence for each element in the contact matrix: either 0 or 1. It should be noticed that the contact matrix is also supplemented with zeros to make all matrices the same size. The neural network was built to predict every single contact, that is, we need to make 900 binary classifications. It should be noticed, that contact matrix is symmetric and we can use only the half of it for prediction. A convolutional neural network was implemented using Keras library[3] for a binary multi-output model (Fig. 6). The binary cross-entropy function [5] was used as the loss function and Adam [13] was used as the optimizer. The model consists of four logical blocks: the first three layers are the same and are represented by convolutional layers with LeakyRElu as activation function, ending with a MaxPooling layer for dimensional reduction. To optimize the neural network, L2-regularization and Dropout were used. The last block converts from two-dimensional format to

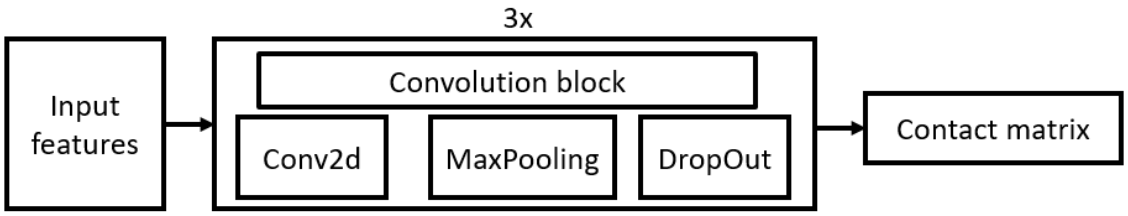


Figure 6: Architecture of Convolution Neural Network for contact matrix prediction

one-dimensional, then we apply the optimization described above and finally, at the output, we have the vector of size 900, predicting the contact for all pairs of aminoacids.

The most frequently used metrics in binomial classification are Precision, Recall and F-score. *Precision* shows the proportion of objects correctly assigned to a positive class relative to all objects that the classifier assigned to a positive class.

$$Precision = \frac{tp}{tp + fp}$$

Recall shows the proportion of found objects of a positive class relative to the real number of objects of a positive class.

$$Recall = \frac{tp}{tp + fn}$$

F1-score is more robust against unbalanced classes and uses both mentioned above metrics.

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

5. Experimental results

Table 1: Metrics for the result of the contact prediction for 8 and 12 $\dot{A}$ thresholds.

Threshold, $\dot{A}$	F1-score	Precision	Recall
8	0.78	0.86	0.73
12	0.83	0.85	0.83

As a result of training the model and checking the model on the test set, contact matrices were predicted for all peptides from the test set. Table 1 shows the results for the selected metrics. Based on this table, we can conclude that for the selected range of proteins, the threshold of 12 $\dot{A}$ shows better prediction results than 8 $\dot{A}$.

To visualize and evaluate the results of the contact matrix prediction model, the native and predicted ones were depicted in the same figure: Figures 7 (a) and 7 (b) show comparisons of the contact matrix for the protein 1NB1 for thresholds 8 and 12 $\dot{A}$ respectively. Light blue colour indicates falsely predicted contacts (FP), brown indicates falsely predicted absence of the contacts

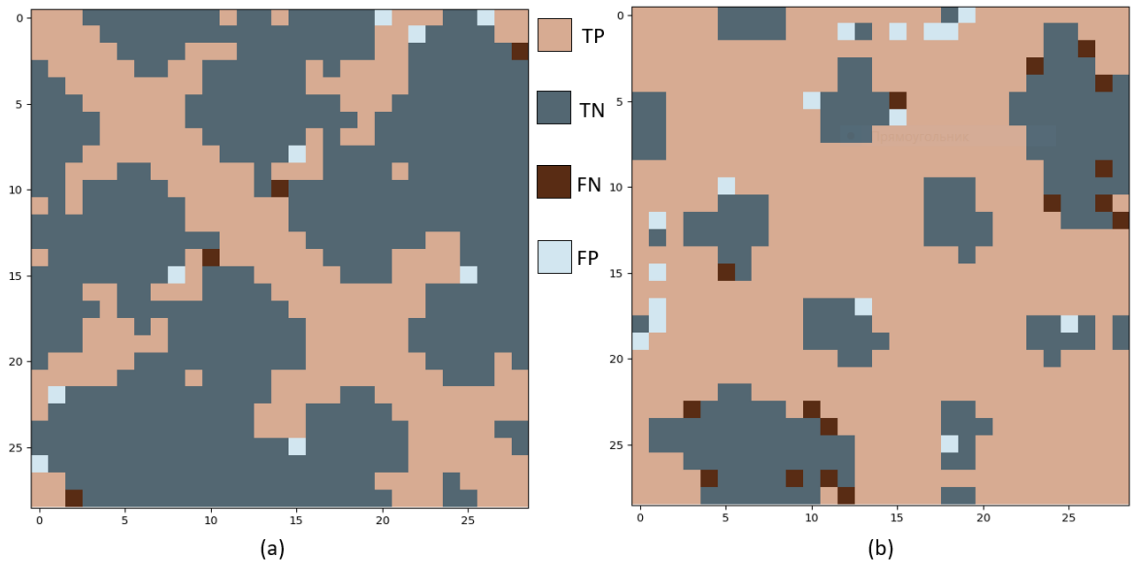


Figure 7: Example of predicted contact matrix with threshold 8 $\dot{A}$ (a) and 12 $\dot{A}$ (b) for protein 1NB1

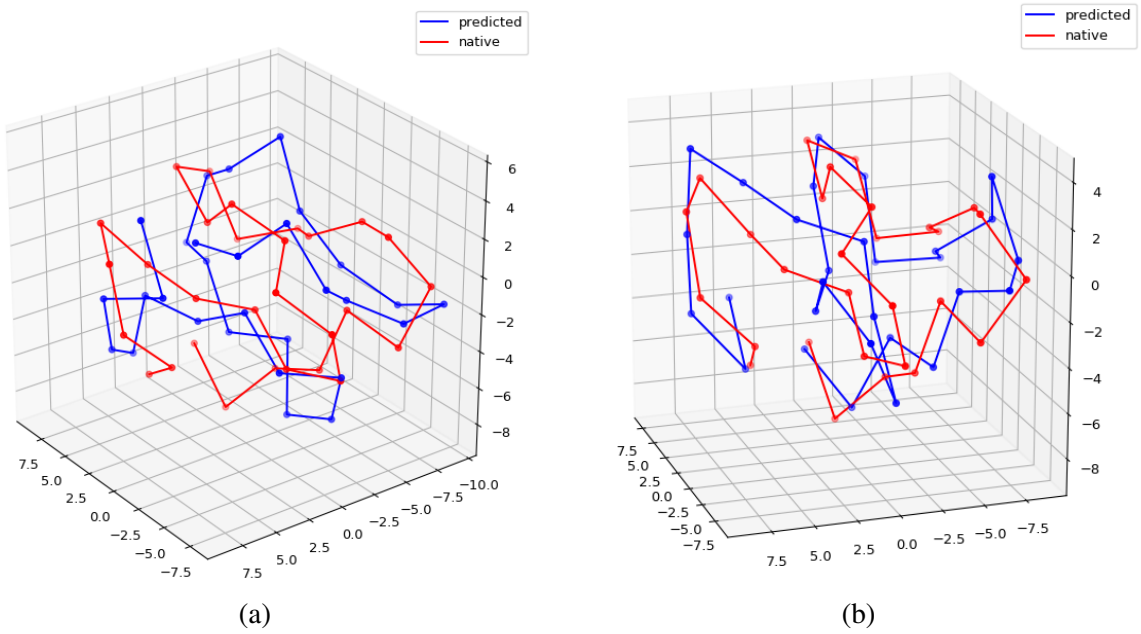


Figure 8: Example of tertiary structure reconstruction from contact matrices with threshold 8 $\dot{A}$ (a) and 12 $\dot{A}$ (b) for protein 1NB1

(FN), sand color indicates correct prediction of contact (TP), and dark green color indicates correct prediction of the absence of the contact (TN).

At the last part tertiary structure of the proteins were reconstructed. FT-COMAR tool [18] is used for backbone atoms restoration in space. For the chosen protein (1NB1) the reconstructed structures are depicted in Figure 8 (a) and Figure 8 (a) for 8 $\dot{A}$ and 12 $\dot{A}$ respectively. After such

reconstruction it's necessary to calculate the precision of the recovered structure. We use well-known metric Root Mean Square Distance (RMSD), which is root mean square of distances from a point of reconstructed structure to native. It is calculated in the following way:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2},$$

where n — is the number of amino acids residues in protein, x, y, z and $\hat{x}, \hat{y}, \hat{z}$ — are the coordinates in the space for native and reconstructed structures.

For the chosen protein the RMSD metric is equal to 6.31 and 2.48 $\dot{A}$ for 8 and 12 $\dot{A}$ threshold respectively. It can be noticed even by sight, that result, which is produced by 12 $\dot{A}$ threshold contact matrix is better.

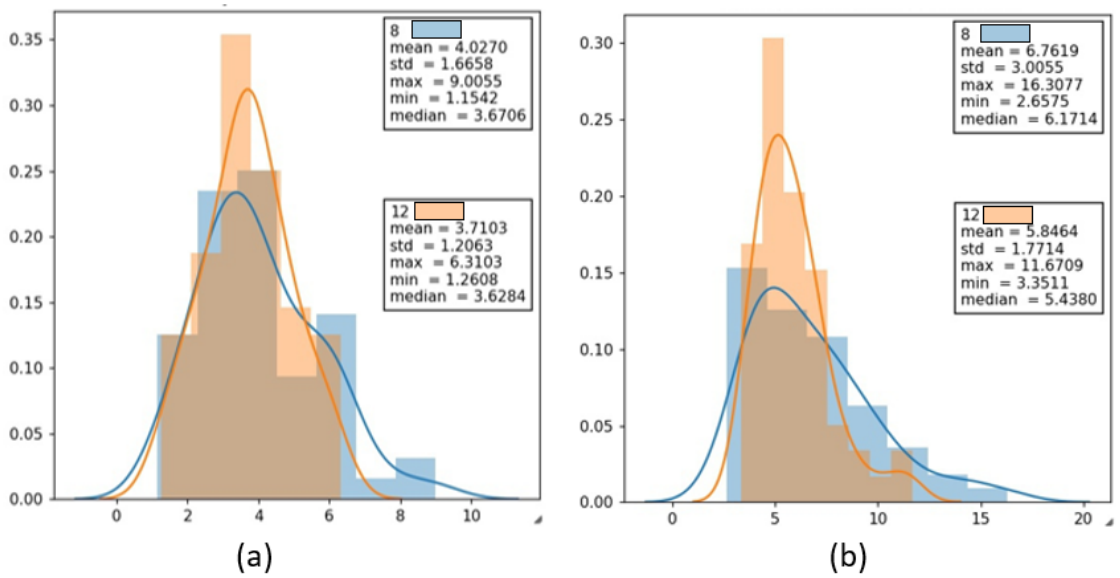


Figure 9: The distribution of the RMSD error for reconstructed tertiary structures from native (a) and predicted (b) contact matrices

Now, let us show the full reconstruction result for the proteins set. To check the lower limit of the FT-COMAR reconstruction error we also reconstruct tertiary structures from the native protein contact matrices. In Figure 9 (a) and Figure 9 (b) the distribution of the RMSD is depicted for native and predicted contact matrix respectively.

6. Conclusion

As a result of this work, a software package was developed for the creation of features, their analysis, reduction to a general view and training with the created CNN model on Keras. It was found that the threshold of 12 $\dot{A}$ showed better results on the presented sample of proteins. The prediction results look quite promising and show, that even with smal number of features and not very complicated neural network. The structure reconstruction results show, that with predicted

contact matrix it's possible to recreate the tertiary structure of the protein with adequate error. The results, which were produced by 12 $\dot{A}$ threshold were better than 8 $\dot{A}$ threshold according to mean RMSD error for native and predicted reconstruction. With native contact matrix reconstruction we got RMSD equal to 4.02 and 3.71 for 8 and 12 and we could not get better result from the prediction.

In the future researches, it is planned to apply this approach for large polypeptides. It is also planned to improve CNN for more accurate prediction and use LSTM and/or ResNet blocks.

References

[1] Beberg A. L. et al. Folding@ home: Lessons from eight years of volunteer distributed computing // 2009 IEEE International Symposium on Parallel and Distributed Processing. – IEEE, 2009. – p. 1-8.

[2] Berman H. et al. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data // Nucleic acids research. – 2007. – Vol. 35. – № 1. – p. D301-D303.

[3] Chollet, F. et al. Keras. // (https://keras.io,2015)

[4] Das R. et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@ home // Proteins: Structure, Function, and Bioinformatics. – 2007. – Vol. 69. – №. S8. – p. 118-128.

[5] De Boer P. T. et al. A tutorial on the cross-entropy method // Annals of operations research. – 2005. – Vol. 134. – №. 1. – p. 19-67.

[6] Di Lena P., Nagata K., Baldi P. Deep architectures for protein contact map prediction // Bioinformatics. – 2012. – Vol. 28. – №. 19. – p. 2449-2457.

[7] Dunbrack Jr R. L. et al. Meeting review: The second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, December 13–16, 1996 // Folding and Design. – 1997. – Vol. 2. – №. 2. – p. R27-R42.

[8] Fox N. K., Brenner S. E., Chandonia J. M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures // Nucleic acids research. – 2014. – Vol. 42. – №. D1. – p. D304-D309.

[9] Greene L. H. et al. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution // Nucleic acids research. – 2007. – Vol. 35. – №1. – p. D291-D297.

[10] Hou J., Adhikari B., Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds // Bioinformatics. – 2018. – Vol. 34. – №. 8. – p. 1295-1303.

[11] Jumper J. et al. Highly accurate protein structure prediction with AlphaFold //Nature. – 2021. – Vol. 596. – №. 7873. – p. 583-589.

[12] Kendrew J. p. et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis // Nature. – 1958. – Vol. 181. – №. 4610. – p. 662-666.

[13] Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. – 2014.

[14] Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding //Journal of molecular biology. – 1976. – Vol. 104. – №. 1. – p. 59-107.

[15] Magnan p. N., Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity // Bioinformatics. – 2014. – Vol. 30. – №. 18. – p. 2592-2597.

[16] Qian B. et al. High-resolution structure prediction and the crystallographic phase problem // Nature. – 2007. – Vol. 450. – №. 7167. – p. 259-264.

[17] Senior A. W. et al. Improved protein structure prediction using potentials from deep learning // Nature. – 2020. – Vol. 577. – №. 7792. – p. 706-710.

[18] Vassura M. et al. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps // Bioinformatics. – 2008. – Vol. 24. – №. 10. – p. 1313-1315.

[19] Xu J. Distance-based protein folding powered by deep learning // Proceedings of the National Academy of Sciences. – 2019. – Vol. 116. – №. 34. – p. 16856-16865.