PRAYOG (TS)

Ab initio Protein Folding Using guided by Contact Prediction using Multi-Perspective Deep Convolutional Neural Networks

Badri Adhikari[1], Nitesh Kafle[2], Renzhi Cao[3], Anthony Ackah-Nyanzu[1]
[1] - University of Missouri-St. Louis, [2] - Lord Buddha Education Foundation, [3] - Pacific Lutheran University
adhikarib@umsl.edu

We participated in the tertiary structure prediction category (server only) with a contact guided folding pipeline developed based on a new method for contact prediction, and the existing method CONFOLD2[1] for three-dimensional modeling.

Methods

With the open-source tool DNCON2[2] as a reference, we developed four different deep convolutional neural networks (CNNs): 1) a basic deep CNN, 2) a dilated CNN, 3) a separable CNN, and 4) a basic deep CNN trained to predict contacts at the distance thresholds of 6, 8 and 10 Angstroms. These networks were trained and tested on the standard dataset of 1426 proteins discussed in the DNCON2 method. The architecture of the basic deep CNN method consists of 17 layers of 64 filters of size 3x3 and the last output layer with one 3x3 filter. After each layer, the activations are padded with zeros so that the input dimensions (300x300) are maintained through all the layers including the output of the last layer. The basic deep CNN model has a total of 625,921 trainable parameters: 64 3x3 filters on the 56 channels give along with 64 bias values result to 32,320 parameters, 16 layers of 64 3x3 filters on 64 channel activations with 64 bias values at each layer result in a total of 590,848 parameters, 128 parameters for batch normalization at each of the 17 layers result in 2,176 parameters, and one 3x3 filter in the last layer on 64 activation channels along with a bias results in 577 parameters.

The architecture of the dilated CNN consists of 13 regular convolutional layers each with 64 3x3 filters followed by two dilated CNN layers. Both dilated CNN layers consist of 64 3x3 filters with a dilation rate of 2. The last layer is a single 3x3 filter. The dilated CNN model has 551,809 parameters total. Similarly, the separable CNN model has its first layer of 64 3x3 filters, followed by 15 depthwise separable CNN layers (SeparableConv2D in Keras) each with 64 3x3 filters, and the last layer with a depthwise separable CNN with 1 3x3 filter. This model has a total of 101,185 parameters. The fourth model is an extension of the basic deep CNN model trained to predict contacts at the thresholds of 6 and 10 Angstroms at the same time. To achieve this, we replace the last layer with three parallel CNN blocks each consisting of two convolutional layers - first layer with 32 3x3 filters and second with one 3x3 filter as the output. Each of these three outputs separately predict contacts at 6, 8, and 10 Angstroms. When calculating the binary cross entropy loss, we weight these outputs such that the weights are 0.25, 1.0, and 0.25 for 6 Angstroms predictor, 8 Angstroms predictor, 10 Angstroms predictor respectively. Predictions by the four methods are averaged to predict the final set of contacts. Finally, we used top 2L long-range and medium-range contacts (L is the length of a protein) to predict five models using the CONFOLD method.

Results

For contact prediction, when trained using the subset of 1230 proteins and tested on the remaining 196 proteins, the precision of top L/5 long-range contacts ranges from 72.8% to 73% for the four methods.

Averaging the predictions of the four methods, we obtain average precision of 75.8% on the 196 proteins, suggesting that the perspective from multiple methods is significantly better than any of the individual methods. We also evaluated our overall method against the experimental structures of some of the targets released so far. While the official CASP results have not been published yet, our preliminary evaluations at target level suggest that the TM-score values of the best-of-five models for the targets T0955, T0958, T0963, T0965, T0971, T1009, and T1016 are 0.36, 0.29, 0.05, 0.17, 0.60, 0.71, and 0.68 respectively.

Availability

The original DNCON2 method and the CONFOLD2 method are publicly available at https://github.com/multicom-toolbox/DNCON2/ and https://github.com/multicom-toolbox/confold2 respectively.

1. Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics*. 2018;19(1):22.
2. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*. December 2017.