Deep learning for protein model quality assessment

Background

The properties of a protein, and its biological functions, are highly dependent of its 3D structure. X-ray crystallography and nuclear magnetic resonance are the main approaches for experimental determination of tertiary structures. However, as the techniques for primary structure determination have become more effective and sequence and structure databases have grown, so has the possibility for computational predictions of protein folding. Even a low-resolution model, only showing residue positions, can be useful. Kihara et al. (2009) states that the major reason for not applying predictions in practical work is that the quality of the model is unknown, not that it is inaccurate. Even somewhat inaccurate models, or models with low resolution, can be used in for example early stages of drug development, as long as the estimated error is known, entailing high accuracy model quality assessments for ranking predicted models. [1]

Derevyanko et al (2018) tackled the protein model quality assessment problem by developing 3DCNN, a quality assessment program using deep learning for evaluating predictions based on their raw three-dimensional atomic density. However, the authors state that the model is compatible with any physical quantity that can be defined on a grid, for example electrostatic potential or solvent density. [2]

Purpose of Project

This project aims to use the 3DCNN model as a template and investigate the opportunities of improvement. In 3DCNN the raw three-dimensional atomic densities from the protein model are used as input to a 3D convolutional neural network, which evaluates the accuracy of the protein model. The network structure and alternative approaches for input data representation will be investigated, hopefully resulting in a higher performing model showing high accuracy.

Project methods.

Python will be used as programming language and, as previously mentioned, deep neural networks will be used for the quality assessment. The network will be developed using TensorFlow framework with Keras as application programming interface (API).

Datasets from the Critical Assessment of protein Structure Prediction (CASP) will be used for training (CASP9 and CASP10) and testing (CASP11) the program.

Project Outlines and Time plan

Initally, a reproduction of 3DCNN will be developed. Thereafter, the network structure and other possibilities for representation will be investigated. At the time for the half-time evaluation, the 3DCNN model should be finished and trials alternating network structure and model input initiated. A draft covering the overall structure of the report should have been produced and the background and theory sections are expected to be somewhat finished. For a more extensive project outline, see the GANTT schedule in Figure 1.

Preliminary Dates	
Half-time evaluation	v.10-11
Thesis presentation	v.23 Thursday (June 7th)

Preliminary Literary Base

Literature not used in this report

- Cao R. et al. *"Protein single-model quality assessment by feature-based probability density functions"* In: *Scientific Reports* 6 (2016)
- Cao R. et al. *"QAcon: single model quality assessment using protein structural and contact information with machine learning techniques".* In: *Structural Bioinformatics* 33.4 (2017) pp.586-588.
- Cheng J. et al. *"Machiene Learning Methods for Protein Structure Prediction"* In: *IEEE Reviews in biomedical engineering* 1 (2008) pp.41-49
- Jing X. et al. "*Sorting protein decoys by machine-learning-to-rank".* In: *Scientific Reports* 6 (2016)
- Jurtz V. I. et al. *"An introduction to deep learning on biological sequence data: examples and solutions"* . In: *Bioinformatics* 33.22 (2017), pp. 3695-3690.
- Krysthtafovych A. et al. *"Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11".* In: *Proteins* 84. Suppl 1 (2016), pp 349-369.
- Liu T. et al. *"Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11."* In: *Scientific Reports* 6.19301 (2016)
- Nguyen S.P. et al "*DL-PRO: A Novel Deep Learning Method for Protein Model Quality Assessment"* In: *Proc Int Jt Conf Neural Netw.* (2014) pp. 2071-2078
- Uziela K. et al. *"ProQ3D: improved model quality assessments using deep learning"*In: *Bioinformatics* 33.10 (2017) pp. 1578-1580.

References

1. Kihara D. et al. *"Quality Assesment of Protein Structure Models".* In: *Current Protein & Peptide Science* 10.3 (2009) pp. 216-228.
2. Derevyanko G. et al. "Deep convolutional networks for quality assessment of protein folds". arXiv:1801.06252v1. 2018. eprint: 1801.06252

Time plan in GANTT format

Project week	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Calendar week	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Activity																						
Introduction																						
Project planning																						
MS1: Planning report			•																			
Reproduce 3DCNN																						
Prepare datasets																						
Training and testing																						
MS2: 3DCNN reproduction ready							•															
Test alternative inputs and layers																						
MS3: Half time evaluation									•													
Develop model																						
Training and testing																						
Data analysis																						
Writing thesis																						
MS4: Final draft																				•		
Preparing opposition																						
Preparing presentation																						
MS5: Thesis presentation																					•	
Thesis corrections																						
MS6: Thesis ready																						•

Figure 1: GANTT schedule covering the entire project period. • indicates a milestone.