Protein Structure Prediction with Deep Neural Networks

Master's Thesis in Bioinformatics

Aarhus University

Authors: Andrej Baláž, Tomáš Sládeček

Supervisor: Christian Storm Pedersen

Authors:	Andrej Baláž	Student Nr.: 201803148
	Tomáš Sládeček	Student Nr.: 201803152
Supervisor:	Christian Storm Pedersen	
Study programme:	Bioinformatics	
Institution:	Aarhus University	
Department:	Bioinformatics Research Centre	

Abstract

Proteins are one of the most important macromolecules in living organisms due to their ability to perform a broad range of biological functions. These biological functions are strongly dependent on their three-dimensional structure. Unfortunately, the experimental methods for determining the 3D structure proteins are resource-intensive, therefore other methods are needed. From a computational perspective, protein folding is a difficult problem, which has remained unresolved for decades. In 2018, DeepMind's team AlphaFold showed great potential for predicting inter-residue distances with deep neural networks. These were then used as constraints for structure optimization which resulted in very accurate protein structures. In this thesis, we develop a pipeline capable of constructing a 3D structure from a raw protein sequence, strongly inspired by the AlphaFold one. Furthermore, we test multiple architectures inspired by AlphaFold and compare them to each other and teams from the CASP13.

Contents

1 Introduction 1

1.1 Protein Structure Prediction 1

1.2 Amino acids 3

1.3 Protein Folding 4

1.3.1 Secondary structure 5

1.3.2 Tertiary structure 6

1.4 Current Approaches to (ab initio) protein folding 6

1.4.1 Energy based methods 6

1.4.2 Fragment Assembly 7

1.4.3 Co-evolutionary features 7

1.4.4 Quantum Computing 8

1.5 CASP 8

1.6 Thesis Outline 10

2 Methods 13

2.1 Input Features 14

2.1.1 Sequence search and MSA construction 14

2.1.2 Position-specific Scoring Matrix 15

2.1.3 Potts Models 16

2.2 Neural Networks 18

2.2.1 The basic architecture of Neural Networks 19

2.2.2 Convolutional Neural Networks 25

2.2.3 Residual Neural Networks 30

2.3 Structure Realization 31

2.3.1 Torsion Angles to Coordinates 31

2.3.2 Background - Vector Cross-product 32

2.3.3 Background - Vector Dot-product 32

2.3.4 Vector Rotation 33

2.3.5 Beta carbon atom placement 34

2.4 Evaluation Metrics 34

2.4.1 (D)lDDT - (Distogram) local Distance Difference Test 35

2.4.2 TM-score 35

3 Results and Discussion 37

3.1 Pipeline 37

3.1.1 PSI-BLAST 38

3.1.2 HHblits 39

3.1.3 plmc 39

3.1.4 Input encoding 40

3.1.5 Data 41

3.1.6 Output encoding 41

3.2 Inter-residue distance modelling 44

3.2.1 Neural Networks 44

3.2.2 Training Strategy 47

3.2.3 Evaluation 49

3.3 Structure Realization 53

3.3.1 Gradient Descent Based Structure Realization 53

3.3.2 CONFOLD2 58

3.3.3 Evaluation 58

3.4 Comparison with other methods 59

3.5 Potential improvements 63

3.5.1 Regarding Input Data 63

3.5.2 Regarding Neural Networks 63

3.5.3 Regarding Structure Realization 64

Conclusion 65

Technical Information 67