```
1   MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP
61  DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK
121 SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE
181 RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS
241 SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP
301 PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG
361 GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
```

Machine Learning in Bioinformatics of Protein Sequences

Algorithms, Databases and Resources for Modern Protein Bioinformatics

Editor

Lukasz Kurgan

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO https://doi.org/10.1142/9789811258589_fmatter

Preface

As of mid-2022, we have access to over 230 million unique protein sequences, significant majority of which have unknown structure and functions. This number continues to grow rapidly and has more than tripled compared to just five years ago [1,2]. Structural biologists and bioinformaticians face an enormous challenge to structurally and functionally characterize these hundreds of millions of sequences. Experimental approaches to decipher protein structure and function do not keep up with this rapid expansion of the protein sequence space, motivating the development of fast computational tools that can help in filling the gap. Machine learning (ML) plays a vital role in the development of these tools since it provides a framework where the limited amount of experimentally solved data is used to design, develop and benchmark predictive algorithms that are later used to make predictions for the millions of uncharacterized sequences.

Hundreds of ML-based methods for the prediction of protein structure and function from the sequence were designed and released over the last few decades. They target a variety of structural aspects of proteins including prediction of tertiary structure, secondary structure, residue contacts, torsion angles, solvent accessibility, intrinsic disorder and flexibility. These methods also address numerous functional characteristics, such as prediction of binding to nucleic acids, proteins and lipids and identification of putative catalytic, cleavage and post-translational modification sites. Recent years have witnessed two significant advances in this area, the development of modern deep neural networks and the innovative representations of the protein sequences that that draw from the natural language processing (NLP) [3,4]. These advances have fueled the modern era in the sequence-based prediction of protein structure and function, resulting in the release of accurate and impactful solutions, such as AlphaFold for the tertiary structure prediction [5], SignalP for the signal peptide prediction [6] and flDPnn for the disorder and disorder function prediction [7]. Besides having free and convenient access to these predictive resources, nowadays, users also benefit from useful resources including large databases of protein structure and function predictions, such as AlphaFold DB [8], MobiDB [9], and DescribePROT [10], and user-friendly platforms that ease the development of new predictive tools, such as iLearnPlus [11] and BioSeq-BLM [12].

This book aims to guide students and scientists working with proteins through the exciting and rapidly advancing world of modern ML tools and resources for the efficient and accurate prediction and characterization of functional and structural aspects of proteins. This edited volume includes a mixture of introductory and advanced chapters written by well-published and accomplished experts. It covers a broad spectrum of predictions including tertiary structure, secondary structure, residue contacts, intrinsic disorder, protein, peptide and nucleic acids-binding sites, hotspots, post-translational modification sites, and protein function. It spotlights cutting-edge topics, such as deep neural networks, NLP-based sequence embedding, and prediction of residue contacts, tertiary structure, and intrinsic disorder. Moreover, it introduces and discusses several practical resources that include databases of predictions and software platforms for the development of novel predictive tools, providing a holistic coverage of this vibrant area.

The book includes 13 chapters that are divided into four parts. The **first part** focuses on ML algorithms. It features Chapter 1 (lead author: Dr. Hongbin Shen) that introduces key types of deep neural networks, such as convolutional, recurrent, and transformer topologies. The authors describe different ways to encode protein sequences and overview applications of deep learners across several types of protein structure predictions including secondary structure, contact maps and tertiary structure.

The **second part** addresses the inputs for the ML models and comprises of four chapters. It focuses on recently developed NLP-based approaches with three chapters that provide background and detail different applications. Chapter 2 (lead author: Dr. Daisuke Kihara) provides a comprehensive introduction to the popular NLP-inspired sequence embedding approaches including Word2Vec, UDSMProt, UniRep, SeqVec (ELMo), ESM-1b and BERT. The authors cover relevant databases and applications of these embeddings to the residue contact, secondary structure, and function predictions. Chapter 3 (lead author: Dr. Bin Liu) expands on this topic by discussing the applications of the NLP-based embedding to the prediction of protein folds, intrinsic disorder and protein-protein and protein-nucleic acids binding. It also introduces several popular ML-based predictors for these applications. Chapter 4 (lead author: Dr. Dukka KC) provides further details on several popular embedding methods and highlights their applications to the prediction of the post-translational modification sites and protein function. The key strength of this chapter is its comprehensive coverage of many ML algorithms in these two application areas. Finally, Chapter 5 (lead author: Dr. Shandar Ahmad) describes more traditional approaches to encoding protein sequences that include evolutionary profiles, biophysical properties and predicted structural features. It also sets these encodings in a practical context of methods for the prediction of protein-protein and protein-nucleic acids interactions.

The **third part** consists of six chapters that explain and describe sequence-based predictors for specific protein structure and function characteristics. Starting with Chapter 6 (lead author: Dr. Liam J. McGuffin), it introduces the contact prediction area, which is arguably one of the main innovations behind the success of the AlphaFold method. This chapter defines residue contacts, reviews the history of this topic, outlines its importance and growing influence, and describes key methods for predicting protein contacts and distance maps. The authors cover different ML algorithms, including hidden Markov models, support vector machines, random forests, Bayesian methods, and modern end-to-end deep neural networks. Chapter 7 (lead author: Dr. Dong-Jun Yu) provides an in-depth treatment of the residue contacts prediction area, focusing on algorithmic details of selected recent ML-based predictors, while assuming basic knowledge of this area and ML algorithms. It also surveys popular predictors, particularly focusing on the deep learning-based models. Subsequently, Chapter 8 (lead author: Dr. Lukasz Kurgan) focuses on the intrinsic disorder prediction area. It provides historical perspective and introduces numerous useful resources including commonly used and accurate disorder and disorder function predictors. In particular, the authors focus on the deep network-based solutions, databases of disorder predictions, webservers, and methods that provide quality assessment of disorder predictions. Chapter 9 (lead author: Dr. Yuedong Yang) defines protein-protein and protein-peptide interactions and introduces the corresponding prediction area. The authors survey related databases, sequence encodings, and predictive models. They also discuss popular assessment schemes and metrics for the evaluation of predictive performance, which they utilize to describe, compare, and recommend several leading methods in this area. Chapter 10 (lead authors: Drs Min Li and Lukasz Kurgan) describes the area concerned with the prediction of the protein-DNA and protein-RNA interactions. This chapter provides a comprehensive review of the current predictors of the protein-nucleic acids binding. Additionally, the authors recommend useful methods and discuss the importance of the underlying structural states of the nucleic acids binding regions. Lastly, Chapter 11 (lead author: Dr. Michael Gromiha) explores databases and ML methods for identifying cancer causing mutations. These resources have direct implication in the context of the development of precision medicine solutions. The authors emphasize on recently developed methods that rely on large scale data in order to produce practical and accurate results.

Part four centers on practical resources that support the development of new ML-based tools and that provide convenient access to the predictions generated by the methods described in the earlier chapters. Chapter 12 (lead authors: Drs Jiangning Song and Lukasz Kurgan) describes an innovative software platform, iLearnPlus, that can be used to analyze structural and functional characteristics of the DNA, RNA and protein sequences and to efficiently conceptualize, design, implement and comparatively evaluate ML-based predictors of these characteristics. The authors use an example application to demonstrate how easy it is to utilize iLearnPlus to design and evaluate a new and accurate predictor of the lysine malonylation sites. Chapter 13 (lead author: Dr. Lukasz Kurgan) highlights convenient ways to obtain pre-computed predictions of protein structure and function from large-scale databases, such as MobiDB, D^2P^2 and DescribePROT. The authors expand on their broad coverage of key characteristics, such as domains, secondary structure, solvent accessibility, intrinsic disorder, posttranslational modification sites, protein/DNA/RNA-binding, disordered linkers, and signal peptides. They also concisely discuss modern predictive webservers that should be used when users want to collect predictions for proteins that are not included in these databases.

Altogether, this book provides a comprehensive perspective on the concepts, methods and resources in the area of the ML-based prediction of protein structure and function from protein sequences. It introduces modern predictive models, including a variety of deep neural networks; provides in-depth treatment of cutting-edge methods for encoding of the input protein sequences; describes state-of-the art predictors for major structural and functional characteristics of proteins; and highlights useful resources that facilitate building new ML models and provide easy access to the predictions. This edited volume serves as a definitive reference for both budding and experienced developers and users of the ML models in this area.

Lukasz Kurgan

References

[1] UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 2021. **49**(D1): D480–D489.

[2] Li W, *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, 2021. **49**(D1): D1020–D1028.

[3] Ofer D, Brandes N and Linial M. The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 2021. **19**: 1750–1758.

[4] LeCun Y, Bengio Y, and Hinton G. Deep learning. *Nature*, 2015. **521**(7553): 436–44.

[5] Jumper J, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021. **596**(7873): 583–589.

[6] Teufel F, *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 2022. **40**: 1023–1025.

[7] Hu G, *et al.* flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 2021. **12**(1): 4438.

[8] Varadi M, *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 2021. **50**(D1): D439–D444.

[9] Piovesan D, *et al.* MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Research*, 2021. **49**(D1): D361–D367.

[10] Zhao B, *et al.* DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Research*, 2021. **49**(D1): D298–D308.

[11] Chen Z, *et al.* iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Research*, 2021. **49**(10): e60.

[12] Li HL, Pang YH and Liu B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. *Nucleic Acids Research*, 2021. **49**(22): e129.

Acknowledgments

First and foremost, I thank my wife, Dr. Magdalena Adamek, and my son, Aleksander Kurgan, for their unwavering support and permitting me to devote many long evening hours to this project. I had you in my thoughts from the start to the finish of this rewarding journey.

I am in great debt to the remarkable groups of students and researchers who contributed, worked and graduated from my lab. These are the people who designed, built, tested and deployed many of the tools and resources that are described in this book. I would not be able to do any research at all without your hard work, ingenuity and commitment. The key people include (in alphabetical order) Dr. Amita Barik, Mr. Balint Biro, Dr. Ke Chen, Dr. Xiao Fan, Mr. Sina Ghadermarzi, Mrs. Leila Homaeian, Dr. Akila Katuwawala, Dr. Fanchi Meng, Mrs. Fatemeh Miri, Dr. Marcin Mizianty, Dr. Christopher Oldfield, Dr. Zhenling Peng, Dr. Chen Wang, Dr. Jing Yan, Mr. Fuhao Zhang, Dr. Hua Zhang, Dr. Tuo Zhang, and Dr. Bi Zhao. I also acknowledge with big thanks my long-term collaborators, Dr. A. Keith Dunker, Dr. Jianzhao Gao, Dr. Gang Hu, Dr. Jishou Ruan, Dr. Vladimir Uversky, Dr. Kui Wang, Dr. Zhonghua Wu and Dr. Jian Zhang. I hope that I did not miss anyone, and I deeply apologize if I did.

I am very grateful to Dr. Shandar Ahmad, Dr. KC Dukka, Dr. Michael Gromiha, Dr. Daisuke Kihara, Dr. Min Li, Dr. Bin Liu, Dr. Liam McGuffin, Dr. Hong-Bin Shen, Dr. Jiangning Song, Dr. Yuedong Yang, and Dr. Dong-Jun Yu for contributing chapters to this book. It was very enjoyable and gratifying to work together.

Finally, I thank Xiao Ling, Vanessa Quek ZhiQin and Joy Quek for managing the process of the book production and to the World Scientific for publishing this volume.

Contents