UNIVERSITAS GADJAH MADA
SELF-ATTENTION AND BIDIRECTIONAL GATED RECURRENT UNIT FOR PROTEIN SECONDARY STRUCTURE PREDICTION
DEWI PRAMUDI ISMI, Prof. Dr.-Ing. Mhd. Reza M. I. Pulungan, S.Si., M.Sc.; Afiahayati, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

TABLE OF CONTENTS

Halaman Judul ii

Halaman Pengesahan iii

Pernyataan Bebas Plagiasi iv

Halaman Persembahan v

PRAKATA vi

TABLE OF CONTENTS viii

LIST OF TABLES xii

LIST OF FIGURES xiv

INTISARI xvii

ABSTRACT xix

I INTRODUCTION 1

1.1 Background . . . 1
1.2 Problem Formulation . . . 9
1.3 Research Scope . . . 9
1.4 Research Objectives . . . 10
1.5 Research Benefits . . . 10
1.6 Research Contributions . . . 10

II LITERATURE REVIEW 11

2.1 Protein Secondary Structure Prediction (PSSP) . . . 11
2.2 History of PSSP . . . 13
2.3 PSSP Framework . . . 13
2.4 Performance Measures . . . 15
2.5 Data . . . 16
2.5.1 Training data . . . 16
2.5.2 Test data . . . 17

SELF-ATTENTION AND BIDIRECTIONAL GATED RECURRENT UNIT FOR PROTEIN SECONDARY STRUCTURE PREDICTION
DEWI PRAMUDI ISMI, Prof. Dr.-Ing. Mhd. Reza M. I. Pulungan, S.Si., M.Sc.; Afiahayati, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

2.6 PSSP in Pre-AlphaFold Publication 18
2.6.1 Features 19
2.6.2 Model architectures 22
2.6.3 Performance 29
2.7 New Approaches for Feature Generation 29
2.8 AlphaFold and Its Impact 32
2.9 PSSP in Post-AlphaFold Publication 34
2.9.1 Taking MSA as direct input to the PSSP model 35
2.9.2 Improving the quality of sequence profiles 35
2.9.3 Exploiting protein language models (LM) as input features . . 36
2.9.4 PSSP studies aiming at other objectives 37
2.10 Future Research Direction 38

III PRELIMINARIES 40

3.1 Amino Acid 40
3.2 Protein Structure 42
3.2.1 Primary structure 43
3.2.2 Secondary structure 44
3.2.3 Tertiary structure and quaternary structure 46
3.3 Protein Representation 46
3.3.1 One-hot encoding of amino acids 47
3.3.2 Position specific scoring matrix (PSSM) 48
3.3.3 HMM profiles 49
3.3.4 Physico-chemical properties 50
3.4 Neural Networks 50
3.4.1 Neuron 50
3.4.2 Activation function 52
3.4.3 Multi-layer perceptron 53
3.4.4 Recurrent neural network 54
3.5 Attention Mechanism 58
3.5.1 Global attention mechanism 59
3.5.2 Local attention mechanism 61
3.5.3 Self attention 63
3.6 PSSP Performance Measure 64
3.6.1 Accuracy 64

SELF-ATTENTION AND BIDIRECTIONAL GATED RECURRENT UNIT FOR PROTEIN SECONDARY STRUCTURE PREDICTION
DEWI PRAMUDI ISMI, Prof. Dr.-Ing. Mhd. Reza M. I. Pulungan, S.Si., M.Sc.; Afiahayati, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

3.6.2 Precision and recall 65
3.6.3 Segment of overlap 65

IV RESEARCH METHODOLOGY 67

4.1 General Description 67
4.2 Research Stages 67
4.3 PSSP with Evolutionary Information-Based Features 69
4.3.1 Dataset 69
4.3.2 Input Features 70
4.3.3 Rationale 72
4.3.4 Design of the PSSP model 74
4.4 Single-Sequence PSSP 78
4.4.1 Dataset 79
4.4.2 Input Features 80
4.4.3 Rationale 80
4.4.4 Design of the PSSP model 81

V IMPLEMENTATION 86

5.1 Development Environment 86
5.2 PSSP with Evolutionary Information based Features 86
5.2.1 Data preprocessing 86
5.2.2 Libraries 86
5.2.3 Architecture 87
5.2.4 Implementation details 91
5.2.5 Network layers and parameters 92
5.3 Single-Sequence PSSP 94
5.3.1 Data preprocessing 94
5.3.2 Libraries 98
5.3.3 Architecture 98
5.3.4 Implementation details 101
5.3.5 Network layers and parameters 101

VI RESULTS AND DISCUSSION 104

6.1 PSSP with Evolutionary Information-Based Features 104
6.1.1 MLP-GRU blocks as the model backbone 104
6.1.2 Dense block inclusion 108

SELF-ATTENTION AND BIDIRECTIONAL GATED RECURRENT UNIT FOR PROTEIN SECONDARY STRUCTURE PREDICTION
DEWI PRAMUDI ISMI, Prof. Dr.-Ing. Mhd. Reza M. I. Pulungan, S.Si., M.Sc.; Afiahayati, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2024 | Diunduh dari http://etd.repository.ugm.ac.id/

6.1.3 GRU internal dropout utilization 109
6.1.4 Self-attention exploitation 110
6.1.5 Feature analysis 115
6.1.6 Overall model performance 116
6.1.7 Performance comparison 116
6.1.8 Ability to capture amino acid interactions 119
6.1.9 Ablation Study 120
6.1.10 Number of Parameters 121
6.1.11 Discussion 121
6.2 Single-Sequence PSSP 124
6.2.1 Attention enrichment of amino acid encoding 125
6.2.2 The effect of batch normalization 129
6.2.3 Performance comparison 129
6.2.4 Discussion 131

VIICONCLUSION AND FUTURE WORK **133**
7.1 Conclusion 133
7.2 Future Work 135

BIBLIOGRAPHY **138**

APPENDIX **159**