Gene Function Prediction Using Evolutionary K-Nearest Neighbor Algorithm

Hiroshi Madushani de Silva

148002J

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

December 2017

Gene Function Prediction Using Evolutionary k- Nearest Neighbor Algorithm

Hiroshi Madushani de Silva

148002J

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Research

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

December 2017

Declaration

“I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The supervisor/s should certify the thesis/dissertation with the following declaration.

The above candidate has carried out research for the Masters thesis Dissertation under my supervision.

Name of the supervisor:

Signature of the supervisor: Date :

Abstract

High-throughput gene annotation data are available in many popular model organism databases and repositories. These data are often incomplete and still evolving while the functions of the genes are unknown or partially known. As the manual curation process is costly and time-consuming, an in-silico method of predicting gene functions became a huge requirement in the industry of bioinformatics. Our approach is to use gene expression data that exist in data repositories rather than sequence data in order to predict the gene functions. In this paper, we have proposed a supervised machine learning algorithm combined with the genetic algorithm for function prediction. The k- Nearest Neighbor Algorithm is optimized using the genetic algorithm to find out the optimum k for a dataset. Also, the genetic algorithm gives a weight vector for the attributes in the dataset making an exceed performance of k- Nearest Neighbor Algorithm. GAKNN is a solution created for gene function prediction which analyze gene annotation data from different repositories and predict gene functions using the genetic algorithm optimized k- Nearest Neighbor classification algorithm. GAKNN provides a workspace for data pre-processing including data cleaning, feature selection, and missing data imputation followed by data analysis and data visualization. The software has been tested over two gene expression datasets from different sources to evaluate the accuracy. The datasets are from two different functional annotation schemes: Gene Ontology and FunCat. The data pre-processing methods available in GAKNN such as missing data imputation also tested with two gene expression datasets and results show that the use of Evolutionary k-Nearest Neighbor Imputation Algorithm gives better results than mean imputation and standard k- Nearest Neighbor Algorithm. The accuracies range from 60%- 88% in GAKNN for function prediction. The weights given for each attribute in the dataset and the optimum k by the genetic algorithm are also graphically represented in GAKNN.

Keywords: k- Nearest Neighbor, Genetic Algorithm, Gene Functions, Gene Annotation, Gene Expression Data

Acknowledgment

I express my gratitude to my parents for their immense dedication and especially I would like to dedicate every achievement of my life to my Mother Mrs. Sarojini de Silva.

My gratitude goes to my supervisor, Dr. Amal Shehan Perera who guided me throughout this research for his valuable assistance and patience.

I like to thank Dr. Rapthi de Silva, Dr. Chandana Gamage, and Dr. Chathura de Silva for their support and Dr. Chinthana Wimalasuriya, Dr. Surangika Ranathunge, and Prof. Nalin Wickramarachchi for their valuable feedback.

Table of Contents

Abstract 4

List of Tables 8

List of Figures 9

CHAPTER 1 11

1. Introduction 11

1.1. Gene Functions 11

1.2. Problem Definition 12

1.3. Importance of predicting gene functions 13

1.4. Research Objectives 13

1.5. Gene Function Annotation 13

1.6. Gene Expression Data 14

1.7. Gene Function Prediction 15

1.8. Research Contributions 16

1.9. Organization 16

CHAPTER 2 17

2. Literature Review 17

2.1. Cell, gene, and gene functions 17

2.2. Protein Synthesis 19

2.2.1. Transcription 19

2.2.2. Translation 20

2.3. Gene Annotation 21

2.3.1. Gene Ontology 21

2.3.2. FunCat (Functional Catalogue) 25

2.4. Importance of gene function prediction 27

2.5. Methods to predict the functions of genes 28

2.6. Genetic Algorithm 33

2.7. Machine Learning Algorithms 35

2.7.1. Supervised Learning Algorithms 35

2.7.2. Unsupervised Algorithms 36

2.7.3. Semi-Supervised Learning Algorithms 36

2.7.4. K- Nearest Neighbor algorithm 36

CHAPTER 3 38
3. Methodology 38
3.1. Genetic Algorithm Optimized k- Nearest Neighbor Algorithm 38
3.2. GAKNN as a software 41
3.2.1. Knowledge Discovery process 41
3.2.2. Data Pre- Processing 42
3.2.2.1. Missing Data Imputation 44
3.2.3. Data Transformation and Data Mining 50
CHAPTER 4 52
4. Results 52
CHAPTER 5 66
5. Discussion 66
CHAPTER 6 69
6. Conclusion 69
References 72

List of Tables

Table 1: Main functional categories of the FunCat- version 2.0 taken from (a. ruepp et al. 2004*)* 25
Table 2: Description of datasets used 48
Table 3: Binary classification results from GAKNN 53
Table 4: Accuracy recorded for 50 iterations in dataset 1 55
Table 5: Accuracy recorded for 65 iterations in dataset 1 56
Table 6: Accuracy given for 25 iterations in dataset 2 59
Table 7: Accuracy given for 30 iterations in dataset 2 59
Table 8: Accuracy given for 35 iterations in dataset 2 60
Table 9: Accuracy given for 40 iterations in dataset 2 60

List of Figures

Figure 1: Illustration of a gene inside a chromosome 17
Figure 2: Illustration of Exon and intron of a gene 18
Figure 3: DNA Transcription 20
Figure4: Biological Process- DNA Metabolism hierarchy breakdown 24
Figure 5: Molecular Function breakdown taken from (Ashburner et al. 2000) 24
Figure 6: Cellular Component: Cell breakdown taken from (Ashburner et al. 2000) 25
Figure 7: Illustration of how genetic algorithm works 34
Figure 8: k- Nearest Neighbor Algorithm distance measurement when k= 3 37
Figure 9: System architecture of GAKNN 39
Figure 10: GAKNN workspace for data pre-processing 43
Figure 11: Feature selection in GAKNN by user 44
Figure 12: Missing data imputation of GAKNN 44
Figure 13: Mean errors of Mean Imputation, kNNImputation and EvlkNNImputation for gasch2 dataset 48
Figure 14: Mean errors of Mean Imputation, kNNImputation, and EvlkNNImputation for spo dataset. 49
Figure 15: Mean errors of MeanImputation, kNNImputation, and EvlkNNImputation for seq dataset. 49
Figure 16: Fitness score over iteration/ evolution for the gasch2 dataset 50
Figure 17: Fitness score over iteration/ evolution for the seq dataset 51
Figure 18: Weight values given by GAKNN for each attribute 55
Figure 19: Accuracy measures given for *binding* in Gene Ontology dataset 56
Figure 20: Accuracy measures given for *catalytic activity* in Gene Ontology dataset 57
Figure 21: Weight values given to each attribute in dataset 2 58
Figure 22: Accuracy measures given for *subcellular localization* in FunCat dataset 61
Figure 23: Accuracy measures given for *transposable elements, viral, and plasmid proteins* in FunCat dataset 61
Figure 24: Accuracy measures given for *cellular organization* in FunCat dataset 61
Figure 25: Accuracy measures given for *cell cycle and DNA processing* in FunCat dataset .. 62
Figure 26: Accuracy over genetic algorithm iteration for Gene Ontology and FunCat datasets 62
Figure 27: Comparison between WEKA and GAKNN accuracy measures 63
Figure 28: Home page and Description of software in SourceForge.net 65
Figure 29: Downloads made by users in SourceForge.net 65

List of Abbreviations

GAKNN- Genetic Algorithm Optimized k- Nearest Neighbor

EvlkNN- Evolutionary k- Nearest Neighbor

kNNImputation- k- Nearest Neighbor Imputation

EvlkNNImputation- Evolutionary k- Nearest Neighbor Imputation

DNA- Deoxyribose Nucleic Acid

RNA- RiboNucleic Acid

GO- Gene Ontology

FunCat- Functional Catalog