A HETEROGENEOUS DATA ENSEMBLE APPROACH FOR PROTEIN FUNCTION PREDICTION UNDER MITOCHONDRION ORGANIZATION

Dinithi Navodhya Sumanaweera

158013D

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

October 2016

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement 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 acknowledgement 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 other medium. I retain the right to use this content in whole or part in future works (such as articles or books)

Signature: Date:

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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

Signature of the Supervisor: Date:

Name of the Supervisor: Dr. Amal Shehan Perera

ABSTRACT

A heterogeneous data ensemble approach for the classification of *Saccharomyces cerevisiae* proteins under 'mitochondrion organization'

Proteins are the real role players in keeping a cell healthy and well functioning. An important group of proteins is the subset of mitochondrial proteins that engage in the assembly, arrangement and disassembly of the mitochondrion. Several of them have been identified to cause human diseases. Hence, annotating proteins under the 'mitochondrion organization' Biology process is vital for identifying disease causative factors and for designing therapeutics. As manual annotation requires costly and laborious in vitro methods, in silico function prediction is preferred nowadays. Recent studies identify the importance of incorporating data from various biological aspects, to formulate a strong functional context for classification. In addition, many approaches from literature employ ensemble classifiers to attain a higher prediction accuracy. However, an insightful approach for accurate classification; biological data utilization; and biological data type significance determination; is still in need. This study presents an assessment of a heterogeneous data ensemble to classify *Saccharomyces cerevisiae* proteins under 'mitochondrion organization'. The ensemble consists of nine euclidean-distance based nearest neighbour models and three affinity-based neighbourhood models; it utilizes sequences, protein domains, peptide chain properties, gene expression, secondary structure and interactions. The base models were trained upon annotations from the Gene Ontology, as well as from a publicly available benchmark gold dataset. They show a substantial level of disagreement, implying their effectiveness in collective decision making. Six combination schemes were evaluated for fusing the base model outputs. A Genetic Algorithmically weighted ensemble gives the highest improvement to the best performing base classifier, by displaying an average area under the Receiver Operating Characteristic curve of 92.52%. Moreover, it is capable of determining the biological importance of each data type. Overall, the proposed heterogeneous data ensemble is capable of identifying eight disease related proteins and one disease related protein in a strong and moderate sense, respectively.

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Keywords: yeast; proteins; mitochondrion; weighted ensemble; data heterogeneity; genetic algorithm; supervised learning

To my beloved parents, grandmother and brother

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

ACKNOWLEDGEMENT

I would like to express my heartiest and sincere gratitude,

To my parents, for all their support, guidance, motivation and inspiration

To my advisor and supervisor Dr. Amal Shehan Perera, for his immense support, invaluable advice, continuous guidance and encouragement, through productive discussions and progress reviews, in making this research a success

To my Research Review Committee: Prof. Nalin Wickramarachchi and Dr. Dulani Meedeniya for their constructive feedback and encouragement

To Prof. T. L. Shamala Tirimanne from the University of Colombo, for offering me with her expertise in Biology through informative discussions, despite her busy schedule

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

To Dr. Surangika Ranathunga and Dr. Charith Chitraranjan, for those illuminating and motivating discussions despite their busy schedules

To Prof. Gihan Dias, for his constant advice and guidance

To Prof. Vajira H. W. Dissanayake and Mr. Nilaksha Neththikumara from the Human Genetics Unit, University of Colombo, for providing me with Training in Bioinformatics

To the Department of Computer Science and Engineering, the Senate Research Grant Committee, the Faculty of Graduate Studies and the staff in general at the University of Moratuwa, for supporting and facilitating my research with necessary resources throughout the course of study

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor i

Abstract ii

Ackowledgement iv

Table of Contents v

List of Figures 1

List of Tables 3

List of Abbreviations 4

1 Introduction 6

1.1 *Saccharomyces cerevisiae* 6

1.2 Importance of ‘mitochondrion organization’ 7

1.3 Need for Protein Function Prediction 9

1.4 Problem Definition 9

1.5 Research Objectives 12

1.6 Contributions 12

1.7 Organization 13

2 Background Study 14

2.1 Bioinformatics and Biological Data Mining 14

2.2 Overview to Functional Genomics and Proteins 16

2.2.1 Basics of Proteins 17

2.2.2 Structure of Proteins 18

2.2.3 Protein Folding 21

2.2.4 Protein Motifs 21

2.2.5 Importance of Protein Structure Determination 22

2.2.6 Protein Domains 22

2.2.7 Protein Families 23

2.2.8 Origination of Proteins 23

2.3 Functions of Proteins 25

2.4 Importance of Protein Function Annotation 26

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

2.5 Biological Data Sources for Functional Genomics 28

2.6 Microarray Gene Expression 29

2.7 Gene Ontology (GO) Functional Classification Scheme 32

2.7.1 Standard Format 33

2.7.2 Hierarchical Structure 33

2.7.3 Annotations 34

2.8 The ‘mitochondrion organization’ GO Term 35

3 Literature Review 37

3.1 Homology based Protein Function Prediction 37

3.2 Multi-class Classification and Data Heterogeneity 38

3.2.1 A True Path Rule Hierarchical Ensemble Approach 38

3.2.2 Hierarchical Classification of G Protein-Coupled Receptors 40

3.2.3 Predictive Clustering Trees and their Ensembles 42

3.2.4 Bayesian Hierarchical Correction 43

3.2.5 HML Boosting 45

3.2.6 Label Similarity Incorporated kNN Algorithm 46

3.2.7 SVM based Ensemble Framework 47

3.2.8 Hierarchical Bayesian Integration Algorithm 49

3.2.9 Semi Supervised Multi-label Collective Classification 50

3.2.10 Ensemble based GPCR Class Prediction 51

3.2.11 Transductive Multi-label Ensemble Classification 51

3.2.12 MS-kNN for Multiple Data Integration 53

3.2.13 Functional Association Network based Approaches 54

3.2.14 BLAST based Local Prediction 55

3.2.15 An Ensemble for ‘mitochondrion organization’ Prediction 57

3.3 Selection of Positive and Negative Examples 57

3.3.1 Negative Example Selection Methods 58

3.4 Class Imbalance 64

3.4.1 Class Imbalance and Feature Selection 66

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

4 Methodology 67

4.1 Data Retrieval and Preprocessing 67

4.1.1 Protein Annotation Data 68

4.1.2 Sequence Data 68

4.1.3 Domain Data 69

4.1.4 Properties Data 70

4.1.5 Gene Expression data 70

4.1.6 Secondary Structure Data 77

4.1.7 Interaction Data 77

4.2 Protein Instance Representation Methods 78

4.2.1 Pseudo Amino Acid Composition (PAAC) 78

4.2.2 Quasi-Sequence-Order Descriptor (QSOD) 80

4.2.3 Conjoint Triad Descriptors 81

4.2.4 Secondary Structure based Representation 83

4.2.5 Latent Dirichlet Allocation (LDA) Topic Representation 83

4.2.6 Gene Expression Profile Representation 86

4.3 Ensemble based Classification 87

4.3.1 Heterogeneous Data Ensemble 89

4.3.2 Affinity-based Neighbourhood models 89

4.3.3 Nearest Neighbour Models 89

4.3.4 Base Model Combination Scheme 92

4.3.5 Performance Measures 95

5 Experimentation, Results Analysis and Discussion 99

5.1 Experimental Setup 99

5.2 LDA Topic Modeling based Approach 102

5.3 Optimal Number of Neighbours 104

5.4 Base Model Evaluation 106

5.5 Evaluation of the Inter-rater Agreement 108

5.6 Genetic Algorithm based Weight Optimization 109

5.7 Ensemble Classification Performance 115

5.8 Identification of Disease Related Proteins 119

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

6 Conclusions and Recommendations 122

References 126

A Exploratory Data Analysis 139

A.1 Initial Analysis of GO Annotations 139

A.2 Data Visualizations 139

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF FIGURES

Figure 1.1.1 *Saccharomyces cerevisiae* 7

Figure 2.2.1 Amino acid residue and Peptide bond formation 17

Figure 2.2.2 List of 20 Amino acid types 18

Figure 2.2.3 Example alpha helix protein and beta sheet protein 19

Figure 2.2.4 Parallel and anti-parallel beta sheets 20

Figure 2.2.5 Gene expression 23

Figure 2.6.6 Microarray technology 30

Figure 2.8.7 GO ancestor chart for 'mitochondrion organization' 36

Figure 3.2.1 Fuzzy kNN ensemble model by Gu et al.(2015) 52

Figure 3.2.2 Directed bi-relation graph 52

Figure 4.1.1 Example FASTA sequence of a protein 69

Figure 4.2.2 Amino acid residue classification 82

Figure 4.2.3 Conjoint triads 82

Figure 4.3.4 Parallel and anti-parallel β sheet formation 91

Figure 4.3.5 Example bayesian network 96

Figure 4.3.6 Kappa scale 98

Figure 5.2.1 ROC plots for the LDA model based approach 104

Figure 5.3.2 k vs mean AUC 105

Figure 5.6.3 (a) GA optimized weights (b) mean ROC AUC of base models 110

Figure 5.6.4 Best fitness value over each sample 110

Figure 5.6.5 Order of data types with respect to both average and maximum fitness giving weight vectors 112

Figure 5.7.6 ROC plots of base models and GA-weighted Ensemble 117

Figure 5.7.7 ROC plots of Ensemble models 119

Figure 5.8.8 Disease protein identification matrix 121

Figure 5.8.9 Disease related protein identification over the 10 samples 121

Figure A.2.1 Expressions 1 - Before normalization/preprocessing 140

Figure A.2.2 Expressions 1 - After normalization/preprocessing 141

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Figure A.2.3 Expressions 2 - MA plots before background correction 141

Figure A.2.4 Expressions 2 - MA plots after background correction 142

Figure A.2.5 Expressions 2 - MA plots after within/between array normalization 142

Figure A.2.6 Expressions 2 - After normalization/preprocessing 143

Figure A.2.7 Expressions 2 - After further normalization 144

Figure A.2.8 Expressions 3 - before & after normalization/preprocessing 145

Figure A.2.9 Expressions 4 - Before & after normalization/preprocessing 146

Figure A.2.10 Expression profiles of housekeeping genes 147

Figure A.2.11 Properties Data 148

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF TABLES

Table 1.1 Disease related ‘mitochondrion organization’ proteins as listed in [8] 8

Table 5.1 LDA model based approach evaluation results 102

Table 5.2 Individual base model performance results I 106

Table 5.3 Individual base model performance results II 107

Table 5.4 Kappa measure for individual samples 109

Table 5.5 GA optimized weights for all 10 samples 114

Table 5.6 Ensemble performance results I 115

Table 5.7 Ensemble performance results II 115

Table 5.8 PR curve AUC values of ensemble models 119

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

LIST OF ABBREVIATIONS

AGPS	Annotating Genes with Positive Samples
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BioGRID	Biological General Repository for Interaction Datasets
BLAST	Basic Local Alignment Search Tool
CAFA	Critical Assessment of protein Function Annotation
CD	Czekanowski-Dice
CTD	Conjoint Triad Descriptor
Da	Dalton (atomic mass unit)
DF	Degrees of Freedom
DNA	Deoxyribonucleic Acid
FunCat	Functional Catalogue
GA	Genetic Algorithm
GO	Gene Ontology
GPCR	G Protein-Coupled Receptor
HER2	Human Epidermal Growth Factor Receptor 2
IEA	Inferred from Electronic Annotation
LDA	Latent Dirichlet Allocation
MIPS	Munich Information Center for Protein Sequences
NGS	Next Generation Sequencing
NLP	Natural Language Processing
NMR	Nucleic Magnetic Resonance
NN	Nearest Neighbour
mRNA	Messenger Ribonucleic Acid

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

PAAC	Pseudo Amino Acid Composition
PCT	Predictive Clustering Tree
PDB	Protein Data Bank
PPI	Protein Protein Interactions
PR	Precision-Recall
QSOD	Quasi Sequence Order Descriptor
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SGD	Saccharomyces Genome Database
SS	Secondary Structure
SVM	Support Vector Machine
TMC	Transductive Multi-label Classifier
TPR	True Path Rule
3D	Three dimensional

University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk