A MODEL BASED APPROACH FOR CLUSTER TRADITIONAL RICE VARIETIES OF SRI LANKA

Mapulage Don Rangika Lilukshi Silva

08/8022

Degree of Master of Philosophy

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

January 2015

Declaration of the candidate & Supervisor

I declare that this is my own work and this thesis 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, 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:.....

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

Signature of the supervisor:..... Date:.....

ACKNOWLEDGMENT

First and foremost, I would like to express my special thanks and appreciation to my supervisor Professor Nalin Wickramarachchi, for his guidance, advice, knowledge and word of encouragement during this study. I am very much gratitude to him, for identifying my capabilities and letting me step into the world of Bioinformatics which opened a great research path for me, without whom it would not have been possible. His advice on both research as well as on my career have been priceless and his kindness will always be remembered.

I would like extend my thanks with gratitude to the former head of the departments, Computer Science and Engineering, Ms. Vishaka Nanayakkara and Lt.Col. Dr. Chandana Gamage, former M.Sc. coordinator associate profeesor Sanath Jayasena and present M.Sc. coordinator Dr. Shehan Perera for the opportunity given to pursue my post graduate studies at the department and for the immense support and guidance given throughout my period. Also, I would like to thank the entire faculty at the department of Computer Science and Engineering for making my stay at the department a memorable experience.

My sincere appreciation also goes to Dr. Jagath Weerasinghe and Dr. Gowri Rajkumar at Institute of Biochemistry, Molecular Biology and Biotechnology, University of Colombo for providing me the biological data needed to evaluate my research methodology. It is a great honored to be a member of their research team and the opportunity given was a great help to acquire the biological knowledge needed to carry out my research work.

I must acknowledge the immeasurable contributions of my friends and colleagues who have shown great love and care during my study.

This acknowledgment won't be complete without my family. My husband Saman has been the greatest strength behind my success and his support made me to achieve many goals in my life. I am thankful for his patient and support given throughout my life. A profound gratitude goes to my mother and to my two sisters whom gave me a great support by looking after my three children in my absence.

ABSTRACT

As a result of the enormous volume of data produced by highly developed modern techniques, focus on clustering biological data has shown a great interest among biologist to detect the underlying patterns in data since the biological experiment itself has failed to identify the hidden information and divergence patterns exist in data correctly.

This study aims to (1) assist clustering biologically similar sequences to detect divergence patterns exist in rice genomic data, by developing a program using the model based clustering algorithm based on Chinese restaurant process which was originally proposed to cluster gene expression data (2) focus on finding the performance of calculating the pairwise distance matrix of rice genome sequences based on the 12-dimensional natural vector of the DNA sequence, as the similarity measure in cluster analysis.

The developed program based on the proposed model based clustering method was executed on ALFP profile data set consisting features of 53 Sri Lankan traditional and wild rice varieties in order to identify the genetic divergence among them. Both a statistical and a biological cluster evaluation were carried out to validate the results obtained. Statistical evaluation was done based on the Bayes ratio to measure the tightness of the clusters formed. Biological evaluation was conducted with the help of the domain experts and research work done by the institute of rice of Sri Lanka. The results showed that the proposed algorithm is capable of identifying highly similar varieties of rice showing their divergence patterns.

Finding the performance of how well the natural vector method captures the information encoded in rice genome sequences, 10 rice disease resistance genes which belong to three different protein families from Rice genome annotation project database were used. The results showed that the pairwise distance matrix calculated based on 12-dimensional natural vector method gives efficient results compared to traditional proximity matrices. It also revealed that the fixed length size sequences (sub sequences) which are not greater than the minimum total length of the selected sequences are also highly capable of capturing the encoded information in total length, regardless of the sub sequence length.

Key words: Model-Based clustering, Genetic Diversity

Table of contents

LIST OF FIGURES

Page

LIST OF TABLES