FUZZY *C*-MEANS CLUSTERING BY INCORPORATING BIOLOGICAL KNOWLEDGE AND MULTI-STAGE FILTERING TO IMPROVE GENE FUNCTION PREDICTION

SHAHREEN KASIM

UNIVERSITI TEKNOLOGI MALAYSIA

FUZZY *C*-MEANS CLUSTERING BY INCORPORATING BIOLOGICAL KNOWLEDGE AND MULTI-STAGE FILTERING TO IMPROVE GENE FUNCTION PREDICTION

SHAHREEN KASIM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

NOVEMBER 2011

Bismillahirrahmanirrahim. Dengan nama Allah Yang Maha Pemurah Lagi. Maha Mengasihani.

Sekalung penghargaan buat:

Yang paling dirindui Allahyarham ayahanda, **Haji Kasim Bin Haji Abdullah**: Terima kasih di atas segala didikanmu, semoga Allah menempatkan ayahanda di tempat orang-orang yang soleh.

Khas buat ibunda, **Hajah Zainon Binti Haji Mohd. Said**: Terima kasih di atas doamu ibu, doa yang sangat bernilai bagiku. Restu dan kata-kata semangat ibunda memberi laluan lurus dalam hidup kami.

Yang terutama, suami tercinta, sahabat karib, kekasih awal dan akhir **Muhammad Edzuan Bin Zainodin**: Terima kasih kerana menjadi suami yang penyayang, penyabar, pendorong, peneman setia, bersama-sama susah dan senang mengharungi perjalanan pengajian ini.

Penyambung warisan, anakanda **Muhammad Eiskandar**, **Sara Arissa**, dan **Muhammad Eirshad**: Kamulah penghibur dan penguat semangat ibu.

Yang diingati keluarga yang sentiasa memberi sokongan: Bonda dan ayahanda mertua, along, ngah, uda, Allahyarham Sabariah, dan ipar-duai.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of many people. Firstly I would like to thank my supervisors, Prof. Dr. Safaai Deris and Dr. Muhamad Razib Othman, for their excellent supervision, knowledge, belief, patience, and interest in the work, encouragement, and motivation throughout these three years. I am also very grateful to Prof. Dr. Richard Spears for his hours of patient and detailed proofreading. My thanks also go to everyone who has provided support or advice in one way or another, including people in Artificial Intelligence and Bioinformatics Research Group (AIBIG) and Laboratory of Computational Intelligence and Biotechnology (LCIB). My greatest thanks should also credit to all staff at Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussien Onn Malaysia (UTHM) for their understanding and support.

Most importantly, none of this would have been possible without the love and patience from my family and has been a constant source of love, concern, support and strength all these years: my husband, my sons, and my daughter for their understanding and sacrifices. I thank my mother for her belief and sacrifice in making me what I am today.

ABSTRACT

Gene expression is a process by which information from a gene is used in the synthesis of a functional gene product. Comprehensive studies of gene expression are useful for predicting gene functions, which includes predicting annotations for unknown gene functions. However, there are several issues that need to be addressed in gene function prediction, namely: solving multiple fuzzy clusters using biological knowledge and biological annotations in some existing databases. This includes, handling the high level expression and low level expression values. Therefore, this research was aimed at clustering gene expressions by incorporating biological knowledge in order to handle these issues. The basic Fuzzy *c*-Means (FCM) algorithm was introduced to address multiple fuzzy clusters in gene expression. Clustering Functional Annotation (CluFA) was developed to deal with insufficient knowledge via incorporating Gene Ontology (GO) datasets and multiple functional annotation databases. The GO datasets were used to determine number of clusters as well as clusters for genes. Meanwhile, the evidence codes in functional annotation databases were used to compute the strength of the association between data element and a particular cluster. The multi stage filtering-CluFA (msf-CluFA) was implemented by conducting filtering stages and applying an enhanced *apriori* algorithm in order to handle the high level expression and low level expression values. The performance of the proposed method was evaluated in terms of compactness and separation, consistency, and accuracy, using Eisen and Gasch datasets. Biological validation was also used to validate the gene function prediction, by cross checking them with the most recent annotation database. The results show that the proposed computational method achieved better results compared with other methods such as GOFuzzy, FuzzyK, and FuzzySOM in predicting unknown gene function.

ABSTRAK

Ekspresi gen merupakan satu proses di mana maklumat mengenai gen digunakan untuk mensintesis fungsi sesuatu produk gen. Kajian menyeluruh terhadap ekspresi gen adalah penting untuk meramalkan anotasi bagi fungsi gen yang belum dikenalpasti. Walau bagaimanapun, terdapat beberapa isu dalam peramalan fungsi gen yang perlu ditangani, antaranya ialah: menyelesaikan pelbagai kelompok kabur menggunakan pengetahuan dan anotasi biologi di dalam pangkalan data sedia ada. Ini juga termasuk mengatasi nilai ekspresi gen yang rendah dan tinggi. Oleh itu, kajian ini telah dilaksanakan bertujuan untuk mengelompokkan ekspresi gen dengan menggabungkan pengetahuan biologi di dalam menangani isu-isu tersebut. Algoritma *fuzzy c-means* (FCM) asas telah diperkenalkan untuk menyelesaikan pelbagai kelompok kabur dalam ekspresi gen. Seterusnya, Anotasi Kefungsian Pengelompokan (CluFA) pula telah dibangunkan bagi mengatasi isu ketidakcukupan pengetahuan biologi melalui penggunaan Ontologi Gen (GO) dan beberapa pangkalan data berkaitan kefungsian anotasi. Data GO telah digunakan untuk mengenalpasti bilangan kelompok dan menentukan kelompok bagi gen-gen. Sementara itu, kod bukti di dalam pangkalan data telah digunakan untuk mengira kekuatan pertalian di antara elemen data dengan kelompok tersebut. Tapisan Pelbagai Peringkat dengan Anotasi Kefungsian Pengelompokan (msf-CluFA) telah dilaksanakan melalui pelaksanaan beberapa penyaringan menggunakan algoritma *apriori* yang dikembangkan bagi mengatasi nilai ekspresi gen yang rendah dan tinggi. Prestasi kajian telah dinilai menggunakan beberapa ukuran seperti kepadatan dan pemisahan, konsisten serta ketepatan terhadap dua data pengujian iaitu data Eisen dan Gasch. Pengesahan biologi juga telah dijalankan untuk menentusahkan ramalan fungsi gen melalui semakan anotasi data yang terkini. Hasil menunjukkan kajian yang dijalankan mencapai keputusan yang lebih baik berbanding kaedah-kaedah lain seperti *GOFuzzy*, *FuzzyK* dan *FuzzySOM* di dalam meramalkan fungsi gen yang masih belum dikenalpasti.