GRAPH PARTITIONING ALGORITHMS FOR DETECTING FUNCTIONAL MODULE FROM YEAST PROTEIN INTERACTION NETWORK

AFNIZANFAIZAL BIN ABDULLAH

UNIVERSITI TEKNOLOGI MALAYSIA

GRAPH PARTITIONING ALGORITHMS FOR DETECTING FUNCTIONAL MODULE FROM YEAST PROTEIN INTERACTION NETWORK

AFNIZANFAIZAL BIN ABDULLAH

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

APRIL 2010

To my beloved wife, family, and friends…

ACKNOWLEDGEMENTS

In the name of Allah, Most Gracious and Most Merciful.

All praise and thanks are for Allah, and peace and blessings be upon to His messenger, Muhammad S.A.W.

I would like to express my sincere thanks to Prof. Dr. Safaai Deris, my supervisor, for his encouragement, guidance and advices. He has corrected all my mistakes patiently and taught me everything that I need not only for the successfulness of my study but also for my life and career. His continual support, encouragement and inspiration have lead me to explore the area of Bioinformatics and made this research possible. My sincere appreciation also goes to Associate Professor Dr. Siti Zaiton Mohd Hashim, Assoc. Prof. Dr. Ito Wasito, Dr. Shahir Shamsir, Dr. Muhammad Razib Othman, my colleagues at Artificial Intelligence and Bioinformatics Group (AIBIG) and all of my friends for their continuous help and support. I Also would like to express my finest gratitude to Prof. Dr. Richard Spear for proof-reading my writing. Finally, I would like to thank my beloved wife, Aishah Yusoff, for her everlasting patience and motivation which help me through my difficult time. Thank you all.

ABSTRACT

Advances in high-throughput technologies have provided many opportunities for researchers to study and better understand the dynamic mechanisms of systems biology. These systems are frequently formed by a functional organisation of networks that recapitulate specific biological processes. Protein interaction networks contain sets of sub-networks called functional modules with highly interactive proteins that perform similar functions. Recently, many graph partitioning algorithms have been proposed for detecting these modules, focusing only on detecting highly interactive proteins and neglecting proteins participating in sparse interactions. Moreover, many algorithms do not consider the overlap among different modules when identifying proteins that perform more than one function. In this research, new graph partitioning algorithms called *Reliable Local Dense Neighbourhood* (RELODEN) and *Overlap-RELODEN* are proposed to detect modules that contain highly interactive proteins, while also considering proteins with sparse interaction and overlap between different modules. The algorithms are based on the clique finding approach, which searches local cliques of informative proteins and groups the cliques into larger sub-networks. Experimental analyses using budding yeast (*Saccharomyces cerevisiae*) protein interaction network have shown that the proposed algorithms have the capability of detecting modules that are significant to biological functions, and thus giving a higher accuracy performance compared with existing algorithms. Moreover, these algorithms have found several interactive proteins that have not been reported previously, and are able to potentially predict the functions of a number of uncategorised proteins.

ABSTRAK

Kecanggihan teknologi telah memberi banyak peluang kepada penyelidik untuk menerokai mekanisme sistem biologi yang dinamik. Sistem ini terbina daripada organisasi rangkaian yang menyumbang kepada fungsi biologi tertentu. Rangkaian interaksi protein mempunyai sub-rangkaian yang dipanggil modul fungsian yang mengandungi protein berinteraksi tinggi dan menjalankan fungsi-fungsi biologi yang serupa. Mutakhir ini, terdapat banyak algoritma pembahagian graf yang telah dicadangkan untuk mengenalpasti modul-modul ini, namun, kebanyakannya menumpu kepada protein-protein berinteraksi tinggi dan mengabaikan protein yang kurang berinteraksi. Tambahan pula, algoritma-algoritma ini tidak mampu mengesan tindanan antara modul-modul apabila mencari protein yang mempunyai banyak fungsi. Di dalam kajian ini, algoritma pembahagian graf dipanggil *Reliable Local Dense Neighbourhood* (RELODEN) dan *Overlap-RELODEN* telah dicadangkan untuk mengesan modul fungsian yang mengandungi protein yang mempunyai interaksi tinggi, dan pada masa yang sama, mengambil kira protein-protein yang mempunyai interaksi yang rendah dan modul-modul yang bertindan. Algoritma-algoritma ini dibina berasaskan pendekatan pencarian klik dengan mencari protein yang informatif dan menyatukan klik-klik ini untuk menjadi sub-rangkaian yang lebih besar. Analisis eksperimen menggunakan data kulat (*Saccharomyces cerevisiae*) menunjukkan algoritma-algoritma ini mampu mencari modul yang signifikan kepada fungsi-fungsi biologi dengan prestasi ketepatan yang lebih tinggi daripada algoritma-algoritma terdahulu. Algoritma-algoritma ini juga telah mengesan beberapa interaksi protein yang tidak dilaporkan sebelum ini dan berpotensi untuk meramal fungsi bagi protein yang belum lagi dikategorikan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CPM	-	Clique Percolation Method
CYGD	-	Comprehensive Yeast Genome Database
DIP	-	Database of Interacting Protein
DNA	-	Deoxyribonucleic Acid
Dr	-	Discard Rate
FDR	-	False Discovery Rate
FN	-	False Negative
FP	-	False Positive
G-N	-	Girvan and Newman Algorithm
GO	-	Gene Ontology
HCS	-	Highly Connected Sub-graph
MCL	-	Markov Clustering
MCODE	-	Molecular Complex Detection
MIPS	-	Munich Information for Protein Sequences
mRNA	-	Messenger Ribonucleic Acid
PI	-	Informative Proteins
PPI	-	Protein-Protein Interaction
RELODEN	-	Reliable Local Dense Neighbourhood
RNA	-	Ribonucleic Acid
RNSC	-	Restricted Network Searching Clustering
SAGA	-	Spt-Ada-Gcn5 acetyltransferase

SNAP	-	S-Nitroso-N-acetylpenicillamine
TAP	-	Tandem Affinity Purification
TP	-	True Positive
Y2H	-	Yeast Two-Hybrid

CHAPTER 1

INTRODUCTION

1.1 Background

Advances in high-throughput technologies have facilitated the availability of complete genome sequences for most organisms as well as diversity among functional genomics and proteomics information, both of which have triggered a growing interest in the study of cellular systems. Among the components of the systems, protein interaction networks have been considered to be the fundamental knowledge required for dealing with biological processes and mechanisms. The protein complex, for instance, is part of a network which includes a set of highly interacted proteins that share similar functions. This concept also reflects the fact that protein interaction networks are highly modular, with the whole network capable of being partitioned into several distinct sub-networks. These are called functional modules, and they independently accomplish discrete biological functions. This may provide insights that will lead to the discovery of the relationships between network topological structures and cellular mechanisms.

However, protein-protein interactions datasets produced by high-throughput technologies have long suffered from incompleteness and a high rate of false conclusions. This has been proven by the small number of overlapping interactions among different experiments. For instance, the common protein-protein interaction between two different mass spectrometry approaches stands at 1,728 pairs, which correspond to 27.5% of interactions detected by tandem affinity purification and 19.2% of interactions detected by mass spectrometry protein complex identification (Md. Altaf *et al.*, 2006). These variations show that many protein-protein interactions detected by different large-scale experiments arrive at mostly false conclusions. Technically, this problem has occurred because most experiments applied two types of proteins, bait and prays proteins, to be bound together even though they are not interact within the actual cell. Therefore, well designed frameworks have to be proposed to overcome these high-throughput datasets limitations when mining the protein interaction network.

A protein interaction network is modelled using a graph theory where a node represents the protein and the edge represents the physical interaction. Global network analysis has shown that this is a scale-free network in which the node degree distribution obeys the law power-law distribution (Barabasi and Oltvai, 2003; Zhang *et al.*, 2007). This indicates that most proteins in the network have a small number of neighbours, while only some have a large number of neighbours, which are called hubs. This signifies that the network has a specific functional organisation rather than one governed by chance. Moreover, the average coefficient-clustering distribution of this network follows the scaling-law and so implies that the network is hierarchically modular (Rives and Galitsky, 2003). The modularity of the protein interaction network shows that the network is relatively invulnerable to random hub removal and more sensitive to targeted attacks by hub nodes (Albert *et al.*, 2000). In addition, the smaller than average path length and the diameter of the network show that information such as signals and chemical reactions can be passed through into the network more effectively. Hence, these dynamic features may provide insights to understanding the cellular mechanisms through the functional organisation in the network.

1.2 Existing Graph Partitioning Algorithms for Protein Interaction Network

Recently, many researchers have put a great deal of effort to elucidating the functional organisation of protein interaction networks, all of which aims at gaining a better understanding of cellular mechanisms. Since the network exhibits a hierarchical modular structure and possesses robust characteristics, it is possible to partition the network into several functional modules, to investigate the organisation in the most systematic manner. Hence, several graph partitioning algorithms are proposed. These can be categorised into three main strategies:

i) The divisive approach – this approach is the simplest way to identify modules, where the edges that connect nodes of different modules are removed to disconnect the modules from the whole network. Algorithms such as the G-N algorithm (Girvan and Newman, 2002), the Highly Connected Sub-graph (HCS) algorithm (Hartuv and Shamir, 2000), the Restricted Network Searching Clustering (RNSC) algorithm (King *et al.*, 2003) and the Markov Clustering (MCL) algorithm (van Dongen, 2000) are employed in this approach;

ii) The highly interacted modular approach – this approach is focused on detecting modules that consist of highly interacted proteins and have fewer interactions with different modules. Most researchers refer this approach as protein complex detection, since protein complexes consist of highly interacted proteins that share similar biological functions. The Molecular Complex Detection (MCODE) algorithm (Bader and Hogue, 2003) is categorised in this approach;

iii) The clique finding approach – this approach aims at searching out complete sub-graphs called cliques that are believed to be the building blocks of the network from an entire network. This approach considers overlapping module detection. The Clique Percolation Method (CPM) (Derenyi *et al.*, 2005) and CFinder (Adamcsek *et al.*, 2005) are included in this approach.

1.3 Challenges in Graph Partitioning Algorithms

The main purpose of detecting functional modules in protein interaction networks is to gain a better understanding of the cellular mechanism at the cell level. This is motivated by several functional organisations within the network that correspond to biological functions. Most highly interacted proteins usually share similar functions. This can be verified by the identification of protein complexes achieved in several works. For instance, MCODE (Bader and Hogue, 2003) and RNSC (King *et al.*, 2003) utilises local neighbourhood searching, which detects several highly interacted modules that belong to identically similar protein complexes. This, therefore may be helpful in inferring biological functions through the genome-wide networks, a method which is more effective when compared to pair-wise approach, since its data are incomplete and noisy (Sharan *et al.*, 2007). Moreover, uncharacterised proteins can be predicted and new biological functions can be suggested based on the network-based function inference.

However, partitioning protein interaction networks is not an easy task, since the network itself suffers from significant incompleteness and a high rate of false determinations, which are the result of high-throughput technologies. Moreover, despite the successes of the diverse graph partitioning methods proposed recently by many researchers, many of the approaches face a variety of limitations and challenges. One of the challenges to such methods as the divisive and highly interacted modules approaches is that they only focus on optimising local concerns of the detected modules. Most of the algorithms proposed using these approaches, for instance, G-N algorithm (Girvan and Newman, 2002) utilise local properties of the nodes in their networks, properties such as degree, centrality and betweenness. This strongly suggests that these approaches only capture network topological features rather than functional organisation. Hence, only a small number of modules can be detected and dynamic

features such as peripheral proteins and overlapping modules are neglected, although these properties may contribute significantly to certain cellular mechanisms.

In addition, several algorithms such as RNSC (King *et al.*, 2003) and MCL (Spirin and Mirny, 2003) are designed to detect modules in a stochastic rather than a deterministic fashion. This may have led to unreliable module detection, since the modules are detected mostly by chance. On the other hand, recent contributions such as Clique Percolation Algorithm (Derenyi *et al.*, 2005; Palla *et al.*, 2005; Adamcsek *et al.*, 2005; Zhang *et al.*, 2006) aim at detecting modules that have complete interactions, a fact that most researchers consider as too strict for biological networks. This is because most functional organisations such as protein complexes are not necessarily offered as complete sub-graphs. However, distinct from other approaches, this approach considers overlapping modules as it conducts its module detection process. This feature is important since most proteins may incorporate more than one module, since most proteins usually contribute to more than one function. Thus, it is important to take into account their functional organisation while detecting modules that may correspond to certain biological functions.

1.4 Statement of the Problem

Advances in high-throughput technologies, such as yeast two-hybrid (Y2H) screening and tandem affinity purification (TAP), have produced abundant of protein-protein interaction data. Consequently, these achievements have shifted the focus of researchers to gaining further understanding of genome-wide cellular mechanisms. Thus, protein interaction network has become most important, and researchers have

come to agree that this network is fundamental to network-based functions (Sharan *et al.*, 2007). However, these advancements are limited by incompleteness and a high rate of false specifications (Barabasi and Oltvai, 2003). Hence, researchers have proposed graph partitioning algorithms to decompose the whole network into functional modules that may provide systematic insight of cellular mechanisms. For these reasons, this study focuses on overcoming these problems and limitations, which can be described in following statement:

> "Given the concept of protein interaction networks, it is a challenging task to develop graph partitioning algorithm that will account for the functional organisation of the network that corresponds to cellular mechanisms"

This study is carried out with two goals in mind: first, a high percentage of proteins that interact with one another usually share similar functions; and second, that proteins are usually produce more than one biological function. Thus, partitioning protein interaction networks based on these functional organisations can lead researchers to gain a fuller understanding of cellular mechanisms, especially for the prediction of the biological functions of these goals; the following questions have to be answered:

i) How can one reliably detect functional modules from incomplete and noisy protein interaction networks?
ii) How can one incorporate proteins that are involved in multiple biological functions in the detection of modules?
iii) How can one measure the efficiency and reliability of a proposed algorithm?

In this work, a graph-partitioning algorithm based on a clique finding approach is developed to identify those informative proteins that are involved in highly interacted modules. However, this approach may not capture spoke-like modules, which include

peripheral proteins (Zhang *et al.*, 2006). Hence, a nearest neighbouring algorithm is applied to include these interactions in the modules. To detect overlapping modules, this proposed algorithm will not exclude all those proteins that already have been a member of a detected module. In other words, this algorithm iteratively searches for proteins that are involved in the modules without removing them from their current network.

1.5 Research Goal and Objectives

The main goal of this research is to construct a graph partitioning algorithm capable of detecting functional modules in a protein interaction network that may correspond to biological functions and cellular mechanisms. In order to achieve this goal, the following objectives have had to be met:

i) To develop *Reliable Local Dense Neighbourhood* (RELODEN) algorithm that capable for mining reliable modules from incomplete and noisy protein interaction network;
ii) To develop Overlap-RELODEN algorithm that is able to detect overlapping modules that are considered peripheral proteins;
iii) To measure the significance of the modules based on the biological functions using *p-value* measure and similarity to the known protein complexes using a *precision recall* analysis.

1.6 Significance and Scope of Study

A protein interaction network consists of a set of proteins, which when related to a phenotype can be identified by projecting them onto the network and testing their network properties (Ideker and Sharan, 2008). However, global network analysis provides limited information about networks as a whole and is also unable to determine accurately individual genes and proteins and their relationships (Hallinan, 2008). Thus, the study of modularity features in a protein interaction network has become increasingly important, for it is only in this way that we are able to gain an understanding of cellular mechanisms from a genomic perspective. A partitioning genome-wide protein interaction network is used to extract functional modules – modules that consist of proteins that share common biological functions. Furthermore, in this rapidly developing field, researches are expected to apply this great potential to better understand the essential mechanisms of living organisms and solve both biological and medical problems, not at the individual component level but at a system-wide level (Zhang *et al.*, 2007). Therefore, detecting functional modules in an effort to understand biological processes and mechanisms has become one of an important effort to gain deeper insight into functional genomics. Besides, the clarification of the function of cellular mechanisms between modules may also motivate the exploration of function prediction, drug design and disease phenotype discovery for the human genome.

In this research, a new graph partitioning algorithm is proposed to detect functional modules in a protein interaction network, an algorithm that will take into account sparse interactions and overlapping modules. Hence, several parameters have to be established to ensure readily utilizable results. Firstly, to test the proposed framework the budding yeast, Saccharomyces cerevisiae, protein-protein interactions datasets have been obtained from Munich Information of Protein Sequences (MIPS) and the Database of Interacting Protein (DIP) (Salwinski *et al.*, 2004). Then, a set of

protein complexes have been downloaded from MIPS Comprehensive Yeast Genome Database (CYGD) (Mewes *et al.*, 2008) to validate the detected modules. In this research, the modules will be visualised and analysed using an open source tool called Cytoscape 2.6.3 (Shannon *et al.*, 2003). For performance measurements, this research has used a precision recall test and examined the performance by an f-measure, which show the relationship between the precision and recall value of the detected modules. Moreover, this research has applied a hyper-geometric measurement called p-value in comparing the biological significance of proteins in each module in Gene Ontology (GO) terms.

1.7 Thesis Outline

The flow of the chapters in this thesis can be presented as follows:

- **Chapter 1:** This chapter explains the key concepts of the research with the research background and problem statement presented. Then the research objectives and scope are discussed. At the end of the chapter, the significance of study is described.
- **Chapter 2:** This chapter presents a review of the literature relevant to this research. The basic concept of network biology is discussed as in a preliminary at this point. Then, the focus shifts to a discussion of related works. Finally the recent research trends are discussed.
- **Chapter 3:** This chapter presents the research methodology. The operational research framework proposed will direct the aim of the research to the achievement of its objectives. The material used in this research such as datasets and evaluation measures are also introduced in this chapter.

- **Chapter 4:** This chapter presents the design and implementation of RELODEN algorithm, which is the algorithm used to mine reliable modules from incomplete and noisy protein interaction networks.
- **Chapter 5:** This chapter presents the design and implementation of the Overlap-RELODEN algorithm which is an improvement over the RELODEN algorithm for detecting overlapping modules.
- **Chapter 6:** This chapter discusses general conclusions of the results, the major contribution and future plans of this ongoing research.

1.8 Summary

In this chapter, the introduction of this research is presented. First, the background of the research is presented. In this section, the overview of the protein interaction network, including protein-protein interaction datasets produced by high-throughput technologies, is discussed. Then, the current graph partitioning methods proposed by previous researchers are described. These methods are classified into three strategies: divisive, highly interacted and clique percolation approaches. Then, the challenges made by the graph partitioning algorithms are presented. Based on these challenges, the problem statement of the research is formulated. In this section, the research questions are also presented. Next, the research goal and objectives are described, as derived from the problem statement. Then, the significance of study and research scope is presented. At the end of the chapter, the thesis outline is discussed. In the next chapter, the literature review relevant to this research is presented.