ACTIVE MODULES OF BIPARTITE METABOLIC NETWORK

by

SHARIL IDZWAN SHAFIE

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
October 2018

Abstract

The thesis investigates the problem of identifying active modules of bipartite metabolic network. We devise a method of motif projection, and the extraction of clusters from active modules based on the concentration of active-motifs in the network. Our results reveal the existence of hierarchical structure.

We model regulation of metabolism as an interaction between a metabolic network and a gene regulatory network in the form of interconnected network. We devise two module detection algorithms for interconnected network to evaluate the molecular changes of activity that are associated with cellular responses. The first module detection algorithm is formulated based on information map of random walks that is capable of inferring modules based on topological and activity of nodes. The proposed algorithm has faster execution time and produces comparably close performance as previous work. The second algorithm takes into account of strong regulatory activities in the gene regulatory layer to support the active regions in the metabolic layer. The integration of gene information allows the formation of large modules with better recall.

In conclusion, our findings indicate the importance of no longer modelling complex biological systems as a single network, but to view them as flow of information of multiple molecular spaces.

Acknowledgements

I would like to take this opportunity to sincerely express my gratitude to my supervisor, Dr Shan He, for giving me the opportunity to be under his wing. Thank you for your very insightful advice, your encouragement and guidance, in helping me to complete this thesis. I also would like to thank my thesis group members that has been insightful, and providing me with fruitful suggestions and reviews.

I am grateful to my beloved parents, Kamariah Shaari and Shafie Nor, for always supporting and encouraging me to do my best. I am also grateful to my sisters, Shakinah Shafie and Salina Shafie for all the supports and helps throughout my life, and my days in Birmingham.

A very special gratitude goes out to Lee Weng Ken, Ferdian Jovan and wife Sherly Meilianti, Rainer Schütz, Guanbo Jia, and Mukarramah Zainal Abidin - the great friends who have provided me with encouragement to get going, and who have been supporting me when I needed most. A special gratitude also goes to Jusnani Mohamed Alias, a good friend and colleague that I got to know while I was working at a college in Malaysia.

Special thanks to my friends - Momodou Lamin Sanyang, Benjapun Kaveedpotjana, Wen Chi Yang, Dong Li, Weiqi Chen, Ning Shi and Siti Rokhmah Mohd Shukri.

Finally, thank you to all my colleagues and friends that I forgot to mention. Thanks for all the encouragement throughout my journey far from home.

※ ※ ※ ※ ※

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The main aim of this research is to devise a module detection algorithm for biological bipartite metabolic network. Module detection techniques can identify community structures in complex networks which is helpful in providing insights of the principles that governs the mechanism of biological systems. In the first section, we introduce the concept of modularity in biological system. Then, we discuss the implementations of community detection of biological networks, and highlight the current limitations in the area of module detection for bipartite network. Next, we describe briefly on the data that will be used in this thesis, and our aim i.e. to implement module detection algorithms to infer disease modules of glioma diseases. Subsequently, we propose the research questions of this thesis and describe the contributions of our work. Finally, we provide the outline of this thesis.

1.1 Modularity in Biological Network

Biological traits arise from the flow of interactions of complex molecular phenotypes within biological system such as DNA, RNA, proteins and metabolites [5, 6]. The significant hurdle faced by systems biology is to understand the complex relationship between these molecular phenotypes, their structures and dynamic behaviours that contribute to traits and diseases. By

being able to analyze this complex system, we could uncover important biological pathways that contribute to diseases and identify key molecular phenotypes that are critical in driving biological processes.

Complex network analysis has rapidly become an essential tools to expand our understanding of biological processes. Biological functions are highly complex processes that involve an interaction of large number of molecules such as genes, proteins and metabolites. To study the role of each molecule to a particular functional pathways could be difficult and time-consuming. However, by transforming these molecules into nodes and points in the network space, we could have the opportunity to decode network properties and patterns that could reveal the collective and organizational behaviours of these biological components. One particular pattern that has caught many interest is the discovery of large and relatively dense sub-networks that is referred to as modules/communities. These are a collection of nodes that are highly connected to each other, but with few connections to outside of their group. These structures have been indicated to be the building blocks for biological functions [7, 8]. Modules are regarded as distinct entities that perform separated functions than other modules. By examining biological network by the structure of its modules, we could discover functional groups by *in silico* studies without having prior knowledge of the system's processes.

The concept of modularity provides computer scientists to transform the NP-hard problem of exploring all the solution space into polynomial-complexity problem [9]. It is easier to find solutions to modular problems, and fragmented modular solutions can be efficiently recombined to address the original problem. Newman and Girvan proposed a concept of modularity as a way to measure quality of modules [10]. The modularity Q is in the range of $[-1, 1]$. A positive value indicates that the interaction of edges in a module exceeds the expectation on the basis of chance. The modularity measure has been implemented in many research to detect modules/communities by maximizing Q [11, 12]. An alternative module de-

tection method uncovers communities by clustering methods. Clustering approach has been applied to protein interaction network by transforming them into weighted network. Arnau et al. counted the length between two protein nodes by the shortest path between them, and applied hierarchical clustering to uncover modules [13].

Modular properties has been noticed in many part of biological networks. Ravasz et al. was among the earliest pioneers that investigated the modular architecture of metabolic network [14]. The authors found that metabolic networks showed scale-free structure and are highly clustered. Segre et al. found hierarchical modules of genes in epistatic interaction network that were comprised of 890 metabolic genes fo *S. cerevisiae* [15]. Segre et al. proposed that the findings indicated that genes interact epistatically in groups. Bhattacharyya et al., in the review of wiring of cell signalling circuits, proposed the importance for cell signalling circuits to exhibit modular interactions [16]. Modularity in gene regulatory networks is also conjectured to contribute to biological systems' robustness to perturbation, and ability to maintain homoestasis [17]. In one of the earliest studies on protein-protein interaction networks, highly significant modules have been discovered [18]. From a study that integrated data of various sources (i.e gene expression, protein structure and functional annotations), modular structures have been observed in protein-protein interaction networks [19].

1.2 Active Modules Detection

The recent goals in computational biology research is to integrate biological networks with each other, and to overlay the network with variety of molecular profiles in the quest to uncover modules [20]. 'Active modules' identification is considered as one of the most promising and significant integrative approach [21]. The active module method seeks to uncover modules by identifying regions in the network that experience significant change of molecular activities. Active modules identification has been applied to identify critical pathways and plays useful roles especially in cancer studies e.g. by improving clinical treatments and in

determining novel enzymatic drug targets and biomarkers for anticancer therapies [22, 23]. By overlaying molecular profiles on the network (e.g. transcriptomic expression), the process-specific information that corresponds to the cellular conditions can complement the topological interaction in the network, thus enabling context-dependent regions that shows striking changes to be identified effectively.

The adaptation of active modules on bipartite metabolic network can provide comprehensive information as it accommodates interactions between molecular compounds, along with enzymes and genes that leads to certain cellular or disease traits. Current adoption of active modules on bipartite metabolic network are by constraint-based method that predicts the steady state of metabolic fluxes [24, 25], and the AMBIENT algorithm that implements simulated annealing method [26].

The following are the limitations of existing method of active modules detection for bipartite metabolic network that motivates us to conduct our research:

- Gene regulation process is not considered as a factor in the process of inferring active modules. Only the levels of genes expression are taken into consideration. The process behind the mechanism of transcription of enzymes is ignored.

- Metabolism is considered only to be affected by the changes in transcription level. However, the states of components in the metabolic network can also affect the metabolism process by a feedback loop through allosteric controls. The activity of key enzymes can be affected by the binding of attachment of activators which often are the substrate of the enzyme itself.

- Large modules are formed especially where there is a big shift towards one direction (i.e. upregulation or downregulation). This results in modules of poor precision.

1.3 Data Set Analyzed

1.3.1 Glioma Disease Dataset

Gliomas is the most common brain tumour that represents approximately 25% of brain tumours and 75% of all malignant tumours. Diffuse 'low grade gliomas' (**LGG**, WHO Grade II astrocytoma, oligodendroglioma and oligoastrocytoma) that mostly occurs in young adults and children in cerebral hemispheres are highly invasive, making complete neurosurgical incision procedure impossible [27]. The survival of LGG patients varied from 1 to 15 years, with a subset of the patients progresses to glioblastoma (WHO grade IV) that has survival of less than 3 years for 88% of patients [28]. For a large proportion of glioma patients, the overall survival has no significant change for the last three decades [29]. This highlights the needs for more studies in the area.

We constructed glioma disease networks (i.e. metabolic and gene regulatory networks) in order to study glioma diseases. The networks are derived from the transcriptome data that are classified as LGG glioma patients (either astrocytoma, oligodendroglioma or mixed) cohort in Rembrandt database [30]. Liu et al. identify a group of samples that are derived from these data to have faster malignant progression, for which mortality are within 5 years [31]. This gives an indication that there more than one biological subtype of LGG, where classifications can be based on patient prognosis. The ability to identify these LGG subtypes may allow better patient classifications and aids the identification of high risk patients, thus allowing for early treatment intervention.

In this research, we also aim to apply the module identification algorithms for detection of disease modules that can differentiate the molecular differences of the short survival patients as stratified in Ref. [31]. The results can be used as inference in highlighting pathways that leads to certain clinical traits and phenotypes, which in turn may help us in identifying molecular phenotypes that may likely be 'key biomarkers' for the glioma progression.

1.3.2 Yeast Data Set

We will be proposing three algorithms to uncover modules in this thesis. These algorithms are using bipartite metabolic network and gene regulatory network as inputs. For the purposes of validating and benchmarking our algorithms, we will be using *Saccharomyces cerevisiae* (yeast) networks under diauxic shift condition. DeRisi et al. conducted one of the earliest study on transcriptome data by investigating gene expression profiles during the diauxic shift condition [1]. Diauxic shift is the state when yeast transforms its energy model from fermentation to respiration. Ethanol becomes source of energy as glucose is depleted during this state.

Our metabolic network is taken from yeast bipartite metabolic network that is provided in Ref. [26] in the form of SBML model. The gene expression profiles during diauxic shift and normal periods are taken from microarray expression data by DeRisi et al. [1]. We are using the same metabolic network and condition settings as the benchmark algorithm from Ref. [26] that has been used to validate our algorithms. Our second network is yeast gene regulatory network. The network is derived from gene regulatory network for the diauxic shift as curated by Geistlinger et al. [32]. Fold-change score are calculated (diauxic shift against normal condition) and overlaid on both networks.

1.4 Research Questions

This research is concerned with five main research questions:

- How to adaptively construct active modules that spans through an interconnected network consisted of a bipartite metabolic network and a gene regulation network?
  In particular, how to devise a generic algorithm that balance the composition of node in modules so that the modules are mainly composed of the nodes from each layer of the interconnected network to denote intralayer information flow in each layer, and

interlayer information flow between the two layers?

- How to infer modules based on the topological features in the interconnected network? For this case, we would like to identify modules solely based on the topology of nodes in the interconnected network, which could be in term of nodes' degree distributions, or the node or edge centralities.

- How to infer modules based on the molecular activity in the interconnected network? In particular, this is the procedure to identify active modules of the interconnected network.

- How to identify sub-modules of a bipartite metabolic network?
  For this case, the modules is only composed of nodes from the metabolic bipartite network.

- Does the modules (and sub-modules) that has been inferred relates to meaningful biological functions? Does it highlight significant metabolic pathways that can leads to certain clinical traits and phenotypes, which in turn may help us in identifying molecular phenotypes that may likely be a 'key biomarker' in a biological process?

1.5 Thesis Contributions

This thesis presents four significant contributions in the field of computer science and computational biology:

- Devising an algorithm that identify modules in interconnected network consisted of two layers, a bipartite metabolic network and a gene regulatory network. Currently, there are no methods that analyze this type of interconnected network, as the existing methods are focussing more in detecting communities in multiplex network. Our method extends the existing information flow method that is well known in identifying communities for

unipartite network. By applying the existing method to our interconnected network, the results show that modules are consisted mainly of nodes from the same layer. This shows that the existing method try to conserve the information flow to within each single layer of the interconnected network. Our method are able to obtain modules that are mainly consisted of nodes from both layers, a closer representation of what we define as 'multilayer module' that denotes not only intralayer, but also interlayer flow of information within the system.

- Devising and formulating an active modules algorithm for bipartite metabolic network that integrates regulatory information, which are derived from gene regulatory network. Currently available methods to infer active modules from bipartite metabolic network only analyze the network in isolation, and only takes into account of transcriptional changes of enzyme-coding genes without considering the regulation of genes that affect the activity of the enzymes. By analyzing metabolic network as an interaction of two layers of biological networks, we are one step closer to model the complex mechanism of biological systems.

- Devising an algorithm can identify hierarchical clusters in active modules of bipartite metabolic network. Current implementation of active modules on bipartite network (i.e. AMBIENT) and our proposed algorithm (i.e. RACEMIC) encourage the formation of large modules. Our algorithm is useful in analyzing these large modules, as it is able to identify important metabolic pathways that are signified by the clusters that are hierarchically linked together in these modules.

- By applying our algorithms on glioma short survival patients data, we are able identify several pathways that has been highlighted and consistent with literature on cancers and glioma. This results may provide for the basis of further studies in investigating these pathways and enzyme-coding genes that may stand as potential targets in the

development of the disease therapy.

1.6 Thesis Outline

This thesis consists of six chapters. The remainder of this thesis is structured as follows.

Chapter 2 presents reviews on the prerequisite knowledge to study biological complex networks. We introduce classical and current methods to uncover modules/clusters from network, either for single layer or multiplex network. Then we describe the development of active module detection methods for biological networks for both unipartite and bipartite network. Finally we argue on the inadequacies of previous approach of active modules detection method on metabolic networks, and describe the issues that motivates our research.

In **Chapter 3**, we propose an algorithm, named as MotifPro, that aims to infer hierarchical clusters in a biological bipartite metabolic network. Directed edges are on the bipartite network based on the activity of reaction nodes, which consequently project motifs on the network. We recursively partition modules in the network based on the motif of interest to uncover connected clusters that corresponds to significant metabolic pathways. By using a yeast metabolic network, we validate the algorithm with the current implementation of active module algorithm and by comparing with curated pathways by expert in known literature.

Chapter 4 describes a proposed module identification algorithm that can uncover topological or active modules of bipartite metabolic network, by integrating gene regulation information. Named as ActiveFlow, the algorithm infers modules by compressing the description of probability flow (i.e. determined topologically or by the activities of nodes) of random walker in the network. The ActiveFlow algorithm is devised for interconnected network consisted of two networks, a bipartite metabolic network and a gene regulatory network. The algorithm is validated by using a yeast network, first by comparing it with the existing information flow algorithm to uncover topological modules, and then by comparing with existing active module identification for bipartite network. Finally we apply the algorithm to

infer disease modules for LGG short survival subtype.

In **Chapter 5** we propose the second implementation of active module identification for bipartite network, named as RACEMIC, that integrates gene regulation information. Similar to ActiveFlow, RACEMIC is devised for interconnected network consisted of two layers. RACEMIC is an extension of AMBIENT, adopted for multilayer network to uncover modules that has strong activities in both the metabolic layer and the gene regulatory layer. We used the networks of LGG short survival subtypes to infer disease modules and use AMBIENT as comparison in evaluating the performance of the proposed algorithm.

Finally, **Chapter 6** concludes the thesis, summarizes the achievements of our work and lists our recommendations for the possible future work.

CHAPTER 2

Prerequisites in Biological Network Analysis

In this chapter, we present a review on the prerequisite knowledge in biological network analysis. First, we describe classical methods of community detection methods. Then, we describe the method of motif-based clustering framework to find higher-order structures in complex network. Next, we describe Infomap, a graph partitioning method based on information flow methodology. After that, we review the methods to uncover clusters from multiplex multilayer network. Subsequently, we describe the works on biological network analysis and the development of active modules identification methods. Finally we argue on the inadequacies of previous approach of active modules detection method on metabolic networks, and describe the issues that motivates our research.

Clustering is a method that seeks to find communities that are homogeneous and well separated, i.e. similar nodes are assigned to the same community and dissimilar nodes are assigned to other communities. In the next section, we are going to describe the classical methods that uncover communities from graphs.

2.1 Traditional Community Detection Methods

2.1.1 Graph Partitioning

A partition is a division of a network into groups, such that each nodes belongs to one group. Graph partitioning problem deals with choosing the division of a network into g groups of predetermined size, such that the number of edges that connects between the groups is minimized. In many cases, constraint is imposed that the groups should be of equal size. It is necessary for the numbers of groups to be specified. If not, one can simply end up with trivial solution with vanishing cut where all nodes of the network are allocated to one group. The size of the group should also be predetermined, so as not to end up with a group consisted of single node of lowest degree that are separated from the rest. A typical application of the problem is in the field of parallel computing and circuit design, e.g. the allocation of tasks to processors in parallel computer, such that the intercommmunication between processors can be minimized. In such a situation, we already have the information on computing powers of the processors in relation to the number of tasks each processor can handle, and the predefined size of groups. Thus, the problem is to optimally group these processors so that resources can be optimally allocated to reduce unnecessary costs.

One of the oldest graph partitioning method is called 'Kernighan/Lin Algorithm' [33]. It is a heuristic algorithm that has been widely applied for layout of electronic circuits and VLSI components. In the original problem that deals with balanced partitioning of electronic circuit boards, the objective of the author was to separate nodes into different boards with the least number of links between nodes from each boards. The problem consider an undirected graph $G = (V, E)$ with n (and even) number of nodes, to be partitioned into two groups A and B, such that $|V_A| = |V_B|$ and $V_A \cap V_B = \varnothing$ and $V_A \cup V_B = V$. The objective is to minimize the 'cut size' $\delta(A, B)$, i.e. the number of edges (or sum of weights of the edges which are equivalent to total cost) with one endpoint in A and the other endpoint in B. The algorithm begin by

allocating the nodes in the graph into two groups of predetermined size. Such allocation can be random or based on information of the graph structure. The we compute the cost of each node D_v, $\forall v \in V$. We define $D_v = E_v - I_v$, where E_v is the external cost that is incurred by node v (i.e. the sum of the cost of the edges between node v and other nodes outside its group), and I_v is the internal cost of v (i.e sum of the costs of edges between v and other nodes in its group).

In next stage of the algorithm, for each kth step from $k = 1$ until $k = n/2$, the procedure seek to find a pair of unmarked nodes $v_a \in V_A$ and $v_b \in V_B$ that cause the largest gain $\hat{g}_k$ (either largest decrease or smallest increase in cut size) if they are swapped between the groups. Then, the node v_a and v_b are marked as locked, the value of each $\hat{g}_k$ is stored, and new D_v are computed for all unmarked node $v \in V$. Once all the pairs has been iterated, we find k such that $G_k = \sum_{i=1}^{k} \hat{g}_i$ is maximized. If the total gain until kth step $G_k > 0$, the node v_a and v_b up until kth step are swapped. Next, the nodes v, $\forall v \in V$ are unmarked and the procedure is repeated until there is no total gain, i.e $G_k \leqslant 0$ for all k. Some of the gain $\hat{g}_k$ could be negative, but when the later gains are positive, then the final gain could be positive. The most expensive part of Kernighan-Lin algorithm is during the calculation of gain for each pairs of candidate to be swapped. The computation complexity is quite fast, that scales to $O\left(n^2 \log n\right)$ when a constant number of subset-swaps are made during each iteration [as reviewed in 34].

2.1.2 Spectral Graph Partitioning

Spectral graph partitioning is popular heuristic clustering method that assigned communities based on the minimum cut on the network. Consider an undirected graph $G(V, E)$, with weight of an edge W_{ij} corresponds to the similarity between node i and j. In Ref. [33], the quality of clusters is the sum of minimum of the edges cut. However, this could lead to misleading quality as the relative cut in proportion to the network is not taken into account.

In Ref. [35], Kannan et al. introduced conductance ϕ as a measure of clustering quality by encouraging balanced cut. Consider an undirected and weighted graph $G = (V, E)$ with no isolated nodes, and n as the total number of nodes. W is set as the weighted adjacency matrix of G, where $W_{ij} = W_{ji}$ is the weight of edge (i, j). Set a diagonal matrix D where $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and the Laplacian representation of the graph denoted by $L = D - W$. For a set S and its complement $\bar{S} = V \backslash S$, its conductance can be defined as:

$$\phi^{(G)}(S) = \text{cut}^{(G)}\left(S, \bar{S}\right) / \min\left(\text{vol}^{(G)}(S), \text{vol}^{(G)}\left(\bar{S}\right)\right) \tag{2.1}$$

$$\text{cut}^{(G)}\left(S, \bar{S}\right) = \sum_{i \in S,\, j \in \bar{S}} W_{ij} \tag{2.2}$$

$$\text{vol}^{(G)}(S) = \sum_{i \in S} D_{ii} \tag{2.3}$$

The conceptual definition of measures of cut and volume are as follows:

$$\text{cut}^{(G)}\left(S, \bar{S}\right) = \text{total sum of weights of edges being cut} \tag{2.4}$$

$$\text{vol}^{(G)}(S) = \text{total sum of weights of edges for nodes in } S \tag{2.5}$$

We refer to quadratic form of L. For any vector $x \in \mathbb{R}^n$,

$$x^T L x = \sum_{(i,j) \in E} w_{ij}\left(x_i - x_j\right)^2 \tag{2.6}$$

Set x as indicator vector that implies the presence of nodes in S, i.e. $x_i = 1$ if node i is presence in S, and $x_i = 0$ when node i presence in $\bar{S}$. Thus, when any edge is cut, x_i and x_j will have distinct values which implies $\left(x_i - x_j\right)^2 = 1$. Else, when x_i and x_j are in S, $\left(x_i - x_j\right)^2 = 0$. Therefore,

$$\text{cut}^{(G)}\left(S, \bar{S}\right) = x^T L x \tag{2.7}$$

The vector x can be written as linear combination of normalized eigenvectors $\mathbf{u}_i$ of L,

such that $x = \sum_{i=1}^{n} a_i \mathbf{u}_i$ and $a_i = \mathbf{u}_i^T x$. Given that x is normalized, then

$$\text{cut}^{(G)}\left(S, \bar{S}\right) = \sum_i a_i^2 \lambda_i \tag{2.8}$$

where λ_i is the eigenvalue of L that corresponds to eigenvector $\mathbf{u}_i$.

Minimizing the cut size $\text{cut}^{(G)}\left(S, \bar{S}\right)$ is equal to minimization of the right-hand size of Equation 2.8. When the second lowest eigenvector λ_2 is close to zero, the sum of the equation will reduce to λ_2 which is very small. Thus, the second lowest eigenvalue λ_2 generate good approximation to the minimization problem, by choosing x that are proportional to the elements of $\mathbf{u}_2$ (i.e. Fiedler vector [36]) that corresponds to λ_2. When the objective is to divide the network into two groups A and B with unequal size $|A| = n_1$ and $|B| = n - n_1$, the best approach to yield two groups is to arrange the Fiedler vector in ascending orders, and put the nodes that corresponds to the highest (or lowest) n_1th indices in the first group, and the remaining nodes in the second group.

2.1.3 Hierarchical Clustering

Hierarchical clustering is a method that groups objects into hierarchy by comparing dissimilarities between pairs of clusters, based on the measure of pairwise dissimilarities among the objects in these two clusters. The hierarchy of clusters are generated by merging or dividing clusters (generally in greedy manner) from current level into the next level. At the bottom level, every cluster is comprised of single object. At the top level, there is only one cluster that contains all of the objects. Hierarchical clustering is useful when it is unjustifiable to make predetermined assumptions on the number and size of the clusters for a network. The method can uncover underlying multilevel structure of a network which is common in the field of social network, engineering and bioinformatics.

The basic paradigms of hierarchical clustering falls into two strategies: agglomerative or divisive. Agglomerative is a 'bottom up' method that pairs of clusters are recursively merged

for each movement up the hierarchy. There will be one less cluster at the next higher level. Divisive approach is a 'top down' strategy that starts with one cluster at the top level, and recursively split one cluster into two new clusters as the procedure moves one level down. For a network comprised of n nodes, there will be $n-1$ levels in the hierarchy reveals by both paradigms. The agglomerative strategy has been the subject of large-scale studies in comparison to agglomerative approach [34, 37]. Besides, most agglomerative algorithms (and some divisive algorithms) depicts monotonic characteristic, i.e the combination of dissimilarities of successive mergers holds. Thus, binary tree can be charted so that the height of its nodes are proportional to the inter-cluster dissimilarities of two clusters they are merged from. The binary tree chart is known as dendogram. The descriptive interpretability of dendogram immensely aids to the high popularity of hierarchical clustering algorithms. Figure 2.1 illustrates an example of dendogram for hierarchical clustering of a network of eight nodes.

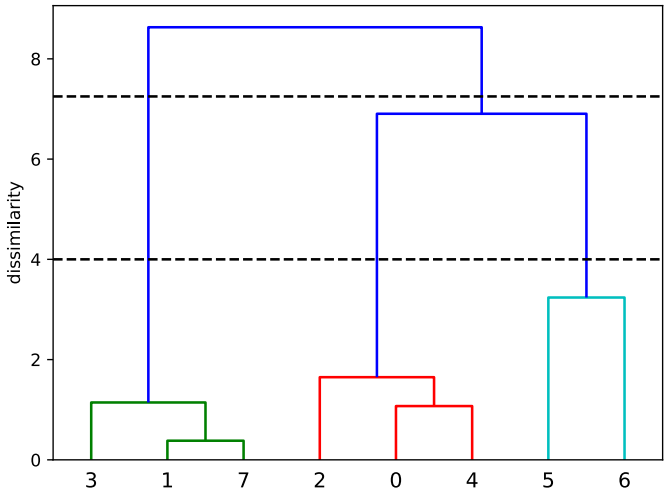


Figure 2.1: A dendrogram with horizontal cuts that correspond to communities.

The review is going to focus further on agglomerative methods as it has been extensively used and due to its monotonicity. The first step in hierarchical clustering is to define the

measure of dissimilarity. Appropriate metric should be chosen to measure the dissimilarity, e.g. Euclidean distance, or Manhattan distance. The choice of the metric would influence the structure of clusters. Suppose that we have two groups, A and B. The dissimilarity between A and B $d(A,B)$ is the pairwise object dissimilarities d_{ij} based on linkage criterion such that $i \in V_A$ and $j \in V_B$, $\forall i, j$. There are three common linkage criteria for agglomerative clustering: single-linkage (SL), complete-linkage (CL), and mean-linkage clustering (CM). The quality of the cluster produced by these criteria could be observed through compactness of the cluster which is measured by 'diameter' of cluster D_C. Diameter is defined as

$$D_C = \max_{i \in A,\, j \in B} d_{ij} \tag{2.9}$$

Single-linkage agglomerative clustering uses the measurement of inter-cluster dissimilarity in the form of closest pair and is defined as

$$d_{SL}(A,B) = \min_{i \in A,\, j \in B} d_{ij} \tag{2.10}$$

Complete-linkage agglomerative clustering use the 'diameter' of pair (refer to Equation 2.9) and is defined as

$$d_{CL}(A,B) = \max_{i \in A,\, j \in B} d_{ij} \tag{2.11}$$

Mean-linkage agglomerative clustering takes the average dissimilarity between clusters and is denoted by

$$d_{CM}(A,B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} d_{ij} \tag{2.12}$$

where $|A|$ and $|B|$ is the number of objects in cluster A and B respectively.

When the network naturally exhibits a strong hierarchical structure, then all the three measures will reveal similar results. However, when the objects in clusters is not compact in comparison to objects in different clusters, the results may varies among the three measures.

The single-linkage clustering has a defect known as *chaining* where it generate clusters with many of the close intermediate objects are combined and separated by small distances, but the objects at the opposite end of the cluster are farther to each other than pair of objects in other clusters. The result may contradict the 'compactness' property where each objects in a cluster should be similar to each other. Single-linkage method tends to generate cluster with very large diameters.

On the contrary, complete-linkage agglomerative clustering is at the opposite extreme of single-linkage method. Complete-linkage method have a tendency to produce many compact clusters with small diameters, i.e the objects in a cluster are much closer to objects in other clusters as compared to the objects within their own cluster. Mean-linkage approach produce a compromise between the single-linkage and complete-linkage methods by generating fairly compact cluster with comparatively distance objects. However, this method is sensitive to numerical scale on which the dissimilarity d_{ij} is measured. The result may vary when exerting monotone strictly increasing function on d_{ij}. This unfavourable defect are regularly put forward by the proponents of single-linkage and complete-linkage to argue against the adaptation of mean-linkage.

The agglomerative hierarchical clustering algorithm is as follows: First we computes the dissimilarity measure d_{ij} for each pair of node in the network, as an element of a $n \times n$ matrix dissimilarity matrix D. Sort dissimilarity score for each pair and merge the pair with the smallest dissimilarity. Then update D into $(n-1) \times (n-1)$ matrix by deleting the rows and columns of previous pair of clusters, and adding new row and column for the newly formed cluster. Next, after sorting the d_{ij} for each pair of cluster, choose the closest distance that are eligible for merging. We repeat the procedure and stop when all the nodes are in one cluster. The time complexity for the agglomerative algorithm is $O\left(n^2\right)$ for single-linkage, and $O\left(n^2 \log n\right)$ for complete-linkage and mean-linkage clustering.

2.1.4 Modularity

The concept of modularity as a measure of quality of network division was first introduced in Ref. [10]. Given there are community i and j in the network, e_{ij} is assigned as half of the fraction of edges that connects nodes in community i to community j. Thus, the total fraction of edges connecting between community i and j is $e_{ij} + e_{ji}$. There are fraction of edges that are only connecting nodes within community i that is denoted by e_{ii}. Then, a_i can be defined as the total fraction of all '*end* of edges' that are connected to nodes in community i, which can be computed as $a_i = \sum_j e_{ij}$. When the end of edges are randomly linked together, the fraction of the links to fall into community i is a_i^2. From these definitions, modularity Q can be defined as

$$Q = \sum_i \left(e_{ii} - a_i^2\right) \tag{2.13}$$

Modularity can be viewed as the probability of the edges in the network to fall within groups minus the probability of edges that are randomly distributed. The value of modularity Q are within $[-1, 1]$. It is preferable to look for high positive score of modularity that denotes the occurrence of more edges within communities than the number expected by chance.

In Ref. [38], Newman implement greedy algorithm to find the optimal value of Q. For a network with n and m number of nodes and edges respectively, the worst-case algorithm complexity is $O[(m+n)\,n]$, or $O\left(n^2\right)$ for sparse graph. At the initial state, each nodes are assigned to its own community. There are n communities during this initial stage. Then, two communities that achieve ΔQ by the largest amount (or decrease by the smallest amount) are iteratively joined together. The value of Q will not increase when two communities that have no edges linking between them are joined together. Thus, only pairs of communities with edges connecting them are to be considered. The algorithm use the difference of Q when two communities i and j are merged, which is given by

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2\left(e_{ij} - a_i a_j\right) \tag{2.14}$$

The value of Q can be updated by each algorithm iteration to find the optimal community structures of the network. This optimal communities is denoted by maximal value of Q.

In Ref. [11], Blondel et al. introduced a faster and improved version of the algorithm in Ref. [38]. The approach is popularly known as 'Louvain method', increases algorithm efficiency by taking into account of sparsity of network. The algorithm is divided into two stages. At the first stage, each node is a community, and each community is added to other community if there is positive gain in modularity. Communities is optimized locally and one ends up with many 'small' communities. Then, at the second stage, a new weighted network is constructed by prescribing to aggregate nodes that belong to the same community together, such that the weight of a new edge between a pair of communities is given by sum of weights of the edges between nodes in the two communities. Edges between nodes of the same community are converted into self-loops. The two-stages is repeated iteratively on the new weighted network. The algorithm is estimated to run at $O\left(n\log n\right)$, and is one of the widely used modularity-based algorithms to analyze very large network (up to 10^6 nodes).

A further development of modularity concept is the technique that expresses it into eigenvectors of modularity matrix. This transformation allows spectral algorithm to be applied which resulted in communities with higher quality. We start with a network of n nodes where the network is divided into two communities. We let $s_v = 1$ if node v belong to community 1, or let $s_v = -1$ when node v belong to community 2. An adjacency matrix $\mathbf{A}$ has element A_{vw} that denotes the number of edges between node v and w (which is typically be 1 or 0). Suppose that k_v and K_w are the degrees of node v and w respectively. If the edges are randomly reassigned, the expected number of edges between v and w is $k_v k_w/2m$, where m is the total number of edges in the network. Modularity Q is the sum of the difference between the actual number of edges between node-pair v and w and the expected number of edges

between them. By summing over all pairs of v and w that fall within a community we obtain

$$Q = \frac{1}{2m}\sum_{vw}\left(A_{vw} - \frac{k_v k_w}{2m}\right)\left(\frac{s_v s_w + 1}{2}\right) \tag{2.15}$$

where $(s_v s_w + 1)/2 = 1$ if the pair v and w are in the same community, or 0 otherwise. Let a real symmetric matrix $\mathbf{B}$ with elements

$$B_{vw} = A_{vw} - \frac{k_v k_w}{2m} \tag{2.16}$$

to be called as 'modularity matrix'. The modularity matrix has a special property where

$$\sum_w B_{vw} = \sum_w A_{vw} - \frac{k_v}{2m}\sum_w k_w = k_v - \frac{k_v}{2m}2m = 0 \tag{2.17}$$

by using $\sum_w A_{vw} = k_v$ and $\sum_w k_w = 2m$. Thus, Equation 2.15 can be written as

$$Q = \frac{1}{4m}\sum_{vw}\left(A_{vw} - \frac{k_v k_w}{2m}\right)s_v s_w = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} \tag{2.18}$$

where $\mathbf{s}$ is the column vector whose elements are the s_v.

From Equation 2.18, $\mathbf{s}$ can be written as a linear aggregation of normalized eigenvectors $\mathbf{u}_v$ of $\mathbf{B}$ such that $\mathbf{s} = \sum_{v=1}^{n} a_v \mathbf{u}_v$, where $a_v = \mathbf{u}_v^T\mathbf{s}$. Thus we can define modularity as

$$Q = \frac{1}{4m}\sum_v a_v \mathbf{u}_v^T \mathbf{B}\sum_w a_w \mathbf{u}_w = \frac{1}{4m}\sum_{v=1}^{n}\left(\mathbf{u}_v^T\mathbf{s}\right)^2\beta_v \tag{2.19}$$

given that β_v is the eigenvalue that corresponds to eigenvector $\mathbf{u}_v$, and the arrangement of the eigenvalues are in decreasing order where $\beta_1 \geqslant \beta_2 \geqslant \beta_3 \geqslant \cdots \geqslant \beta_n$.

To find communities, we try to choose best division of the network by maximizing Equation 2.19, or as an alternative, by setting the value of $\mathbf{s}$ that can maximize modularity. This can be achieved by setting $\mathbf{s}$ to emphasize the weight in the summation terms of Equation 2.19

to the largest eigenvalue β_1 and its corresponding eigenvector $\mathbf{u}_1$. As all the weight is concentrated on the term that corresponds to β_1, the other terms will be zero due to orthogonal property of eigenvectors. As $\mathbf{s}$ is restricted to be either -1, or +1, the maximum value can be achieved by setting $s_v = +1$ if elements of $u_v > 0$ and $s_v = -1$ otherwise. There could be a case when there will be no positive eigenvalues for the modularity matrix. For this situation, the network division will result in non-positive modularity, as all summation terms in Equation 2.19 will either be zero or negative. The network for this case is referred to as indivisible as no better division exists.

The algorithm divide the network into two communities. To find division of the network in larger number of communities, we can recursively divide each community into two. Thus, the extension of the algorithm is to divide the network into two communities, and to repeatedly divide these communities into two until each community becomes indivisible. When we divide a community c of size n_c, the additional contribution ΔQ to quality is given as

$$\Delta Q = \frac{1}{4m}\mathbf{s}^T\mathbf{B}^{(c)}\mathbf{s} \tag{2.20}$$

where $\mathbf{B}^{(c)}$ is a $n_c \times n_c$ matrix whose elements are indexed by nodes that are in community c. The elements of matrix $\mathbf{B}^{(c)}$ is

$$B^{(c)}_{vw} = B_{vw} - \delta_{vw}\sum_{k\in c} B_{vw} \tag{2.21}$$

with δ_{vw} as Kronecker delta (i.e. having value of 1 if $v = w$ and 0 otherwise). As Equation 2.20 has the same representation as Equation 2.18, we can adopt the spectral partitioning method just as in previous algorithm to maximize ΔQ.

Modularity still remains as the most popular clustering technique [39]. One advantage of this technique is that it assume that the network will naturally divides in communities without having to put constraint on the number and size of communities, or restraining trivial solution that puts all nodes into a single community. When there is at least a positive eigenvalue, there

is no possibility for all nodes to be in the same community. However, when there is no positive eigenvalues for the modularity matrix, the algorithm is explicitly telling that the network is indivisible, where good division of the network is non existence. The algorithm complexity scales as $O[(m+n)\,n]$ on a single bipartition, or $O\left(n^2\right)$ for sparse network.

2.1.5 Partitional Clustering

Partitional clustering is a method that seeks to find set of clusters of a network where the nodes are represented by points in the form of vectors. The nodes' properties are transformed into points in a metric space, that allow distance to be measure between any pair of points. The distances is analogous to the measure dissimilarities between nodes. The number of clusters is predetermined, e.g. k, and the objective is to either minimize/maximize the score function that are derived from the distances between the points or to the centroids of the clusters. This algorithm has same limitation as graph partitioning algorithms, that the number of clusters should be determined at the initialization process. Besides, transformation into metric space may not be natural for certain type of networks. However, the algorithm remains popular as it is very efficient to yield practical solutions.

The most widely used partitional clustering approach is *k-mean clustering* (also known as *Lloyd's algorithm*) [40]. The objective, i.e. in the form of squared error function is defined as

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{v \in S_i} \| v - c_i \|^2 \tag{2.22}$$

where S_i is the subset of nodes in the ith cluster, and c_i is the centroid of the cluster. The algorithm begins by initially placing the centroids of clusters far from each other. Then, each node in the network is assigned to the nearest centroid by Euclidean distance. Next, new centroids are calculated from the mean of all nodes in each respective cluster. Then, the nodes are re-assigned to the new centroids based on their distances, and this procedure is repeated until the positions of the centroids and the assignments of nodes no longer change.

There is no guarantee that the solution would be optimal [41], as solutions are influenced by the initial position of centroids, and may stuck at local optimum. However, the technique remains popular as an efficient tool as the solutions usually converge very fast.

Another popular variation of k-mean clustering is called *fuzzy c-mean clustering* (FCM). This method was developed by Dunn [42] and was improved by Bezdek [43]. The algorithm takes into consideration the possibility that a node may belong to more than one cluster. The objective of FCM is

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{m} \left\| v_i - c_j \right\|^2 \tag{2.23}$$

where

$$w_{ij} = \frac{1}{\sum_{c=1}^{k} \left(\frac{\left\| v_i - c_j \right\|}{\| v_i - c_c \|} \right)^{\frac{2}{m-2}}} \tag{2.24}$$

and c_k is the centroid of kth cluster for the set of clusters $C = \{C_1, C_2, \ldots, C_k\}$, m is a real number that is more than 1 that control the fuzzy characteristic of clusters, and node $v_i \in V$. The w_{ij} is a membership element, where it measures the degree to which node i belongs to cluster j. The element w_{ij} is normalized that the sum over all clusters is 1. Given that there are n nodes in the network, the centroid c_k is given as

$$c_k = \frac{\sum_{i=1}^{n} w_{ik}^{m} v_i}{\sum_{i=1}^{n} w_{ik}^{m}} \tag{2.25}$$

The algorithm is fairly similar to k-mean clustering. First, we set the number of clusters. Then, the coefficient w_{ij} is assigned randomly to each node for being a member in each cluster. These values are used to compute the centroids for clusters (Equation 2.25), which will later become the inputs to calculate new w_{ij} (Equation 2.24). We repeat the procedure until the solutions converge. The algorithm has the same problems as k-mean where the solutions are influenced by initialization of membership elements, and can stuck at local optima.

2.1.6 Overlapping Community Detection

Many of the community detection algorithms treat communities as exclusive and separated regions. However, most networks of real systems are made of overlapping nodes. For example, a gene could be responsible for different biological functions at the same time, and an individual could belong to different groups when exposed to different environments (e.g while at work, with family or during holiday). This section describes the techniques to uncover overlapping communities.

Two algorithms that were proposed by Baumes et al., Iterative Scan (IS) and Rank Removal (RaRe), was among the earliest implementation of overlapping communities detection [44]. The author defined a community as a locally optimal subgraph that is induced by a subset of nodes, with respect to edges densities functions W, e.g. average degree densities W_{ad} [45] which is defined as

$$W_{ad} = \frac{2\,|E(S)|}{|S|} \tag{2.26}$$

for a set of nodes S, where $E(S)$ is the collection of edges with both endpoints in S. The IS algorithm start with a 'seed' as a community, and iteratively update the community by adding/deleting a node to raise the density score W. Given that the cluster progress to a sequence $S_1, S_2, \ldots,$, the density scores $W_1(S_1) < W_1(S_1) < \ldots$, should be in strictly increasing order. This condition assure that no cluster will reappear in the sequence. The procedure to find local optimal on the community stops when the density score could no longer increase, where adding or removing node will only cause previously identified cluster to re-emerge. Then, we repeat the procedure to find a community by assigning a new seed. The seeds could be chosen randomly, or obtained from a group of nodes (or clusters) that are part of locally optimal solutions from different algorithms. We continue to assign new seeds until stopping criteria are met, e.g. the maximum number of unsuccessful attempts to introduce seeds yields similar or less density scores to the clusters that are obtained earlier. The locally optimal

clusters inferred from the seeds could share nodes, thus resulted in overlapping communities.

The RaRe algorithm has an assumption that there exist a subset of vital nodes that relay high degree of communication in the network. The RaRe algorithm seeks to find the nodes and remove them from the network, where such act will disconnect it into smaller connected components. The nodes can be ranked by their centrality score (i.e. in term of degree score, or PageRank [46]). The vital nodes are added into a set R, and the removal process stops when the size of connected components exceeding a predefined criteria. These components are considered as 'core' clusters. Then, every node $v \in R$ is added back to each of the core clusters if it yields the increase of density metric W. The intuition is that a vital node should be part of the cluster adjacent to it. We will have overlapping communities when a node in R is added to more than one cluster. The algorithm complexity of IS and RaRe is $O\left(n^2\right)$. Baumes et al. found that the adoption of the solutions from RaRe as input for IS algorithm can yield the best results.

A very popular algorithm for overlapping communities is based on clique density measure, with the idea that a community will have high density of internal edges that are likely to form clique. Introduced by Palla et al., the algorithm is known as Clique Percolation Method (CPM) [47]. The basis of a community is a 'k-clique', which is a complete subgraph on k nodes. Palla et al. refers to a community as 'k-clique community', which is a chain of union of all k-cliques that can connect to each through adjacent k-cliques. Any two of k-cliques is adjacent when the share $k-1$ nodes. The k-cliques can only move inside its communities, as it could not pass the bottleneck created by the inter-links between adjacent communities. However, a node can become a member of different communities. We can identify a community by 'rotating' a k-clique about $k-1$ overlapping nodes over other adjacent k-cliques. The procedure to find k-clique communities starts by searching for maximal cliques in the network. As a clique could not be a subset of larger ones, the search is done in decreasing order of the size. Palla et al. determine the maximum clique size based on degree-sequence.

The procedure continues by repeatedly choose a node, and count every clique of that size that contains the node, and then delete the node and its respective edges. Once all nodes have been iterated, we reduce the clique size by one, and repeat the procedure on the original network. Then we create a $n_c \times n_c$ clique-clique overlap matrix $\mathbf{O}$, where there are n_c cliques, and element 0_{mn} denotes the number of common nodes between clique m and n. To seek for k-cliques, we keep off-diagonal entries and diagonal entries that are at least $k-1$ and k respectively. Then, we extract connected components from resulting matrix. The algorithm complexity to detect maximal cliques runs at non-polynomial time. However, the whole procedure is fairly fast on the graph of real systems.

A technique presented by Zhang et al. combines modularity, spectral partitioning and fuzzy clustering, to find overlapping communities [48]. The author considered a network $G(V,E)$ further described by a symmetric weight matrix W where its element $w_{ij} \geqslant 0$. Each node i has relationship to cluster $c \in C$ in a normalized form, given by an 'assignment matrix' P with its entry $p_{ic} \in [0,1]$. For every vertex, the sum of p_{ic} over all clusters $c \in C$ is 1. This suggests that p_{ic} can be considered as the probability for node i to be in community c.

The first component of the algorithm proposed by Zhang et al. concerns with the quality of overlapping communities of a network. Let us define 'cover' as the comparable term for partition for overlapping communities. The total weight of the weight matrix W is given by $w = \sum_{i,j \in V} w_{ij}$. The best cover can be obtained by maximizing objective function Q_{zh} that is given as

$$Q_{zh} = \sum_{c=1}^{n_c} \frac{\bar{w}_c}{w} - \left(\frac{\bar{s}_c}{2w}\right)^2 \tag{2.27}$$

where

$$\bar{w}_c = \sum_{i,j \in V_C} \frac{p_{ic} + p_{jc}}{2} w_{ij} \tag{2.28}$$

and

$$\bar{s}_c = \bar{w}_c + \sum_{i \in V_c, j \in V \setminus V_c} \frac{p_{ic} + (1 - p_{jc})}{2} w_{ij} \tag{2.29}$$

A partition of a network G into c communities can be represented by an $n \times c$ matrix X, for which $x_{ic} = 1$ if node i is in community c, or otherwise $x_{ic} = 0$. In Ref. [49], White and Smyth formulated the equivalent of modularity (Equation 2.15) as

$$Q = \mathrm{tr}\left[X^T (\mathcal{W} - \mathcal{D})\, X\right] = -\mathrm{tr}\left[X^T L_Q X\right] \tag{2.30}$$

where $\mathcal{W}$ is a diagonal matrix with similar entries that equal to sum of weight of all edges ($\mathcal{W}_{ii} = \sum_{i,j \in V} w_{ij}$). The matrix $\mathcal{D}$ is such that its element $\mathcal{D}_{ij} = \mathrm{degree}(i) \cdot \mathrm{degree}(j)$, for node i, $j \in V$. The matrix $L_Q = \mathcal{D} - \mathcal{W}$ is referred to as 'Q-Laplacian'. The problem of determining the value of X to maximize Q is NP-Complete. We can relax the constraint for X so it element $x_{ij} \in \mathbf{R}^1$. This transform Q as continuous function of X. By setting the first derivative of Equation 2.30 (with respect to X) to 0, we can find the extreme points of Q. This transformation leads to eigendecomposition problem

$$L_Q X = X\Lambda \tag{2.31}$$

where Λ denotes a diagonal matrix. When the number of nodes is at least moderately high, we can approximate $\hat{W} = L_Q$, where $\hat{W}$ is a normalized matrix of W such that its rows sum to 1. Thus, spectral partitioning that form the basis of the algorithm of Zhang et al. is the maximization of score function

$$\hat{W} X = X\Lambda \tag{2.32}$$

The algorithm procedure by Zhang et al. to find n_c communities is as follows: In the spectral mapping step, we compute the normalized matrix $\hat{W}$. Form an eigenvector matrix $E_{n_c} = \left[e_1, e_2, \ldots, e_{n_c}\right]$ by calculating the top n_c eigenvectors of Equation 2.32. Then we move to the fuzzy c-means step. For each c, such that $2 \leqslant c \leqslant n_c$, we create a matrix $E_c = [e_2, e_3, \ldots, e_c]$. By using fuzzy c-means method (the method is described in Section 2.1.5), cluster the row

vectors of E_c to get the assignment matrix P_c. This step will provide the association of nodes to the communities. The last step is to chose the c and P_c that maximize modularity (Equation 2.27). The most expensive part of the algorithm is during the computation of eigenvectors. The algorithm complexity is $O\left(K^2 n + Km\right)$, where K is the maximum number of communities, and m the number of edges. For sparse network with $k \ll n$, the complexity scales to linear.

2.2 Motif-based Graph Partitioning

Benson et al. introduced the general framework for motif-based spectral clustering method [50]. The framework caters for unipartite graph with motif of three or more nodes. For a motif of k number of nodes, the set of nodes that contains a motif is defined as $\mathbf{v} = \{v_1, v_2, \ldots, v_k\}$. The general framework defines motif conductance based on a relevant subset of $\mathbf{v}$ that is classified as 'anchor node' $\mathcal{A}$ (i.e. $\mathcal{A} \subseteq \mathbf{v}$). For a majority of cases, it is sufficient to have $\mathcal{A} = \mathbf{v}$, which is the conditions for simple motifs as defined in Ref. [50]. As in our case, we are only considering graph partitioning based on simple motifs, such that all the nodes in the motif are taken into account in defining motif conductance.

Given a motif M of k nodes, the motif conductance $\phi_M^{(G)}(S)$ is defined as

$$\phi_M^{(G)}(S) = \mathrm{cut}_M^{(G)}\left(S, \bar{S}\right) / \min\left(\mathrm{vol}_M^{(G)}(S), \mathrm{vol}_M^{(G)}\left(\bar{S}\right)\right) \tag{2.33}$$

where

$$\mathrm{cut}_M^{(G)}\left(S, \bar{S}\right) = \sum_{\mathbf{v} \in M} \mathbf{1}\left(\exists i, j \in \mathbf{v} | i \in S, j \in \bar{S}\right) \tag{2.34}$$

$$\mathrm{vol}_M^{(G)}(S) = \sum_{\mathbf{v} \in M} \sum_{i \in \mathbf{v}} \mathbf{1}(i \in S) \tag{2.35}$$

where $\mathbf{v} \in V^k$, $v_1, \ldots, v_k$ distinct, and $\mathbf{1}(s)$ is an indicator function that return the value of 1 if the statement s is true or 0 if s is false.

The measure of cut and volume for motif can also be defined conceptually as follows:

$$\text{cut}_M^{(G)}\left(S,\bar{S}\right) = \text{number of instances of motif being cut} \tag{2.36}$$

$$\text{vol}_M^{(G)}(S) = \text{number of end points of motif in } S \tag{2.37}$$

By considering an unweighted and directed graph having a motif set M, the definition of the motif adjacency matrix can be formally defined by

$$(W_M)_{ij} = \sum_{\mathbf{v}\in M} \mathbf{1}\left(\{i,j\} \subset \mathbf{v}\right) \tag{2.38}$$

The motif diagonal degree matrix is defined as D_M as $(D_M)_{ii} = \sum_{j=1}^{n} (W_M)_{ij}$, and the motif Laplacian representation of W_M as $L_M = D_M - W_M$. Finally, the normalized Laplacian form of W_M, L_M can be denoted by

$$L_M = D_M^{-1/2} L_M D_M^{-1/2} = I - D_M^{-1/2} W_M D_M^{-1/2} \tag{2.39}$$

where $D_M^{-1/2}$ is a reciprocal diagonal matrix of D_M. $D_M^{-1/2}$ is a diagonal matrix where its diagonal elements is the reciprocal of the positive square roots of the diagonal elements of D_M. The second eigenvector of L_M is used to derive the spectral ordering for graph clustering.

Suppose that we have an unweighted and directed graph G with a motif M. The weighted graph corresponding to G (as specified by Equation 2.38) is defined as G_M. Benson et al. in Ref. [50] derived the following statements:

Lemma 2.1. Consider a directed and unweighted graph $G = (V, E)$, and let G_M to be a weighted graph of a motif of k nodes. Then for any $S \subset V$,

$$\text{vol}_M^{(G)}(S) = \frac{1}{|k|-1} \cdot \text{vol}^{(G_M)}(S)$$

.

Lemma 2.2. Consider a directed and unweighted graph $G = (V, E)$, and let G_M to be a weighted graph of a motif of 3 nodes. Then for any $S \subset V$,

$$\text{cut}_M^{(G)}\left(S, \bar{S}\right) = \frac{1}{2} \cdot \text{cut}^{(G_M)}\left(S, \bar{S}\right)$$

Theorem 2.3. Consider a directed and unweighted graph $G = (V, E)$, and let G_M to be a weighted graph of a motif of 3 nodes. Then for any $S \subset V$,

$$\phi_M^{(G)}(S) = \phi^{(G_M)}(S)$$

Theorem 2.4. Consider $\phi_* = \min_{S'} \phi_M^{(G)}\left(S'\right)$ to be the optimal motif conductance for a set of nodes S'. Then,

1. $\phi_M^{(G)}(S) \leq 4\sqrt{\phi_*}$ and,
2. $\phi_* \geqslant \lambda_2/2$

2.3 Infomap Random Walk Graph Partitioning

In Ref. [51], Rosvall and Bergstrom introduced Infomap, a methodology that compresses the representation of network path in two-level description, i.e module and node level. Infomap framework capitalize on the theoretical basis of likelihood for a random walker to statistically stay within particular cluster for a long period of time.

At the node level, to achieve an efficient compression of paths, the nodes in the network are named based on Huffman coding [52]. Nodes with higher visit frequencies will have shorter codenames, and node that are rarely visited will receive longer codenames. Refer to

Appendix A.2 for Huffman coding methodology and algorithm. The process of describing a trajectory of a network is illustrated in Figure 2.2. Figure 2.2A shows the original network and the allocation of name according to Huffman coding is shown in Figure 2.2B. If we give the same codelength to the original network that are consisted of 10 nodes, the codelength per node would be $\lceil \log 10 \rceil = 4$ bit long. To describe a path v_c, v_f, v_d, v_g, v_i, v_h, the total length for the path would be 24 bit (4 bit are allocated to each of the six nodes of the path). Huffman codes save resources by allocating short codelengths for nodes that are commonly visited, and longer codelengths to non-frequent nodes. As shown in Figure 2.2B, by using Huffman codes, the length of path v_c, v_f, v_d, v_g, v_i, v_h is reduced to 19 bit (3+3+3+3+3+4 bit).

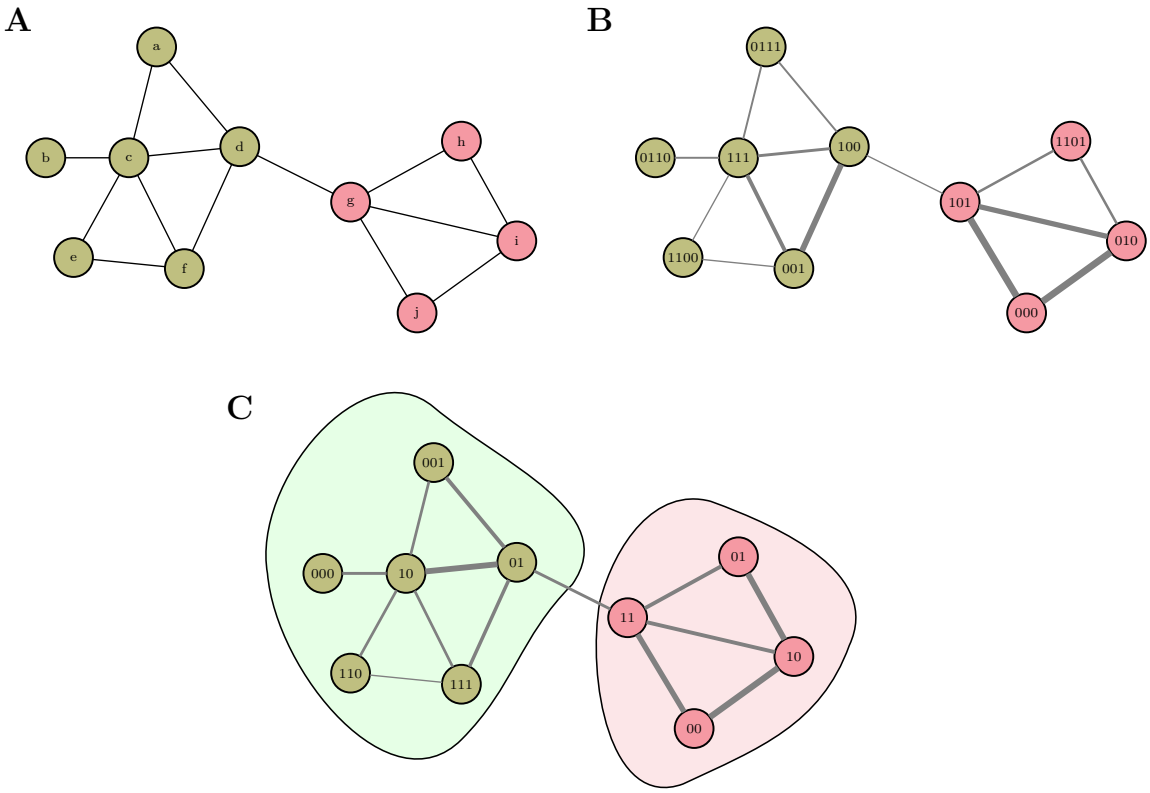


Figure 2.2: Example of Infomap in describing path v_c, v_f, v_d, v_g, v_i, v_h

A further reduction of the path length can be achieved by representing the nodes into major modules, thus allowing the name of nodes in the modules to be reused and yield shorter description of paths in the network. We can use an analogy where the same street

names are used in may different cities or states of a country. This approach rarely create confusion as most routes are bounded by the cities or states. To illustrate the idea, we assume the original network are divided into two modules as differentiated by their colours. As there are only two modules, for illustrative purpose, we assume that the cost to either exit or enter each of these modules is one bit. Figure 2.2C shown the network where the nodes are coded according to their own modules. The earlier path can now be described in shorter length of 15 bit (2+3+2+1+1+2+2+2). Value of 1 bit is added each when a random walker enter and leaving a module.

In practice, no actual codenames will be assigned to nodes, but emphasis is put more on the theoretical bound of specifying a path. Shannon's source coding theorem [53] states that when we describe each nodes in the network by different codename, the average codelength will be bounded below by entropy of the nodes. For a random variable X, to describe n states by n nodename, the entropy is denoted by $H(X) = -\sum_i^n p_i log\left(p_i\right)$, where p_i is the frequencies of the node i. We provide more information on Shannon source theorem in Appendix A.1.

For a network of partition M with m modules, the map equation that denotes the average description length per step is as following:

$$L(M) = q_{\text{out}} H(Q) + \sum_{i=1}^{m} p_{\text{in}}^i H\left(P^i\right) \tag{2.40}$$

The first term of the map equation is the entropy of movements between modules, and the second term is the entropy of movement within the modules (including the module exiting step). $H(Q)$ is the entropy of module name, and q_{out} is the per step probability that a random walker switches modules. $q_{\text{out}} = \sum_i^m q_{\text{out}}^i$, where q_{out}^i is per step probability the random walker exit module i. $H\left(P^i\right)$ is the entropy of within modules that includes movement of exiting the module. p_{in}^i is the proportion of movements within-module, added with probability of exiting the module i.

For a directed network, to ensure steady state distribution for the random walker, teleportation probability τ is introduced to allow the walker to hope to any node in the network. With this property, the movement of the random walker can be described as aperiodic and ergodic. Given that P_α is the probability to be at node α in module i, and the random walker exit the module to node β either by outlinks with weight $W_{\alpha\beta}$ or by teleportation, the exit probability for module i can be defined as

$$q^i_{out} = \tau \frac{n - n_i}{n} \sum_{\alpha \in i} p_\alpha + (1 - \tau) \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha w_{\alpha\beta} \tag{2.41}$$

Louvain method (refer to greedy methods in Section 2.1.4) is adapted to find communities in network. The procedure is comprised of two phases. In the first phase, each node in the network is assigned to its own community. Then, by ordering the nodes in a random sequence, a module is chosen and paired with its neighbouring nodes. The pair that achieves the the biggest decrease of description length (Equation 2.40) is assigned into a community. No community will be assigned if the pairing does not result in decrease of map equation. The process of assigning communities is repeated with new random sequence, until the description length can no longer decrease. The procedure enters second phase, where a new network is built where the communities at the last stage are formed into nodes. The first and second stage is repeatedly executed until the algorithm can no longer reduce the description length of the network.

2.4 Multilayer Network Analysis

Kivelä et al. pointed out that the development of module detection methods for multilayer network are still in very early stage [54]. As maximizing community is computationally hard [54], the computational difficulty could increase acutely with multiple number of layers. Besides, much of the studies available on the multilayer community detection techniques are focusing on multiplex network, the type of multilayer network comprised of M different layers

where the same set of nodes can be linked to each other by M different type of connections (as defined in Ref. [55]). There are already growing research on community detection on multiplex network e.g. method of aggregation [56, 57], random walk techniques [58, 59], hypergraph-based spectral clustering [60] and tensor based approach [61].

The current research trends on detecting communities in multilayer networks are focussing on the analysis of multiplex networks. A popular method has been devised by Mucha et al. to uncover community structure in multiplex networks [2]. The technique is based on 'Laplacian dynamics' that has been introduced by Lambiotte et al. in Ref. [62]. The author proposes to quantify network quality by measuring stability of a network partition, i.e in the form of statistical characteristics of a dynamical process in the network. For an undirected and unipartite network represented by adjacency matrix A_{ij} and strength of node $k_i = \sum_j A_{ij}$ (i.e. in term of degree), a continuous time Laplacian dynamics is defined by $\dot{p} = \sum_j A_{ij} p_j - p_i$ (where p_i denotes density at node i). By defining stability under such dynamics, Lambiotte et al. derived quality function similar to modularity (Equation 2.13).

In their work, Mucha et al. extend the above concept into multilayer formalism, where the authors consider a multislice (and multiplex) network as illustrated in Figure 2.3. The network is represented by adjacency elements A_{ijs} that denotes intralayer relationship between node i and j in layer s, and C_{jrs} that denotes interlayer coupling between node j to itself in layer r to layer s. The layer component is constrained to unipartite, undirected network where $A_{ijs} = A_{jis}$ and interlayer couplings $C_{jrs} = C_{jsr}$. From the strengths of nodes is given by $k_{js} = \sum_i A_{ijs}$ (intra-slice) and $c_{js} = \sum_r C_{jsr}$ (inter-slice), the multislice strength is defined as $\kappa_{js} = k_{js} + c_{js}$. Mucha et al. define Laplacian dynamics for multislice network as

$$\dot{p}_{is} = \sum_{jr} \frac{\left(A_{ijs}\delta_{sr} + \delta_{ij}C_{jsr}\right) p_{jr}}{\kappa_{jr}} - p_{is} \tag{2.42}$$

where δ represents Kronecker delta. The dynamical process has probability distribution

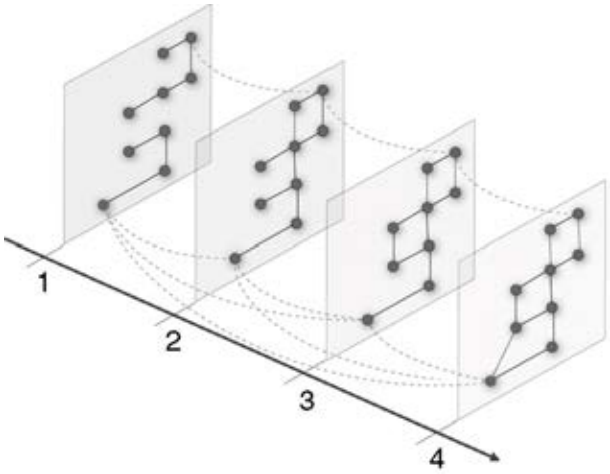


Figure 2.3: Multislice network framework: There are four slices, $s = \{1,2,3,4\}$, with A_{ijs} denotes intralayer connections (solid lines), and C_{jrs} denotes interlayer connections (dashed lines), where the coupling of node j exists between slices r and s. The coupling can be between neighbouring slices, or many-to-many coupling. For simplicity, interlayer relationships are shown for only two nodes. Reprinted figure with permission from Ref. [2].

under equilibrium $p^*_{jr} = \kappa_{jr}/(2\mu)$, where $2\mu = \sum_{jr} \kappa_{jr}$. By using the steady state property, the multislice null model can be specified as

$$\rho_{is|jr} p^*_{jr} = \left(\frac{k_{is}}{2m_s} \frac{k_{jr}}{\kappa_{jr}} \delta_{sr} + \frac{C_{jsr}}{c_{jr}} \frac{c_{jr}}{\kappa_{jr}} \delta_{ij} \right) \frac{\kappa_{jr}}{2\mu} \tag{2.43}$$

where $m_s = \sum_j k_{js}$. The model is formulated based on the probability of node i to be in layer s, conditional on whether the network structure permits a one-step movement from (j,r) to (i,s). The formulation ensures that the conditional probability of one-step movement from (j,r) to (i,s) will only be nonzero if and only if $i = j$ (coupling). By subtracting the conditional probability from the linear in time exponential map that describes Laplacian dynamics, we obtain modularity quality function for multislice network that is defined as

$$Q_{\text{multislice}} = \frac{1}{2\mu} \sum_{ijsr} \left[\left(A_{ijs} - \gamma_s \frac{k_{ij} k_{js}}{2m_s} \right) \delta_{sr} + \delta_{ij} C_{jsr} \right] \delta\left(g_{is}, g_{jr}\right) \tag{2.44}$$

where conditional probabilities are reweighted, by allowing different resolution γ_s for each

layer. To uncover community structures in the multislice network, we can use the same modularity-based heuristic algorithms that are being employed for single layer unipartite network, e.g. greedy algorithms in Section 2.1.4.

The authors, De Domenico et al. extended Infomap framework (refer to Section 2.3) to a multilayer system. The multilayer network, i.e. in the form of multiplex network, is represented by adjacency matrix W_{ij}^{α} that specifies intralayer relationship between node i and j in layer α. The multilayer system is represented by random walker dynamics with relax rate r, where transition probabilities for a walker to move from node i in layer α to node j in layer β is given by

$$P_{ij}^{\alpha\beta}(r) = (1-r)\,\delta_{\alpha\beta}\frac{W_{ij}^{\beta}}{s_i^{\beta}} + r\frac{W_{ij}^{\beta}}{S_i} \tag{2.45}$$

where $s_i^{\beta} = \sum_j W_{ij}^{\beta}$ is the outward-intensity for node i in layer β, with $S_i = \sum_{\beta} s_i^{\beta}$, and δ represents Kronecker delta. A random walker move with probability $1-r$ to other nodes through intralayer edges, and with probability r the condition is relaxed that the walker can move to any physical node linked to it. This relax condition allow a walker to go node j in different layer. The technique follows the same principle as single layer network (Section 2.3), which is Shannon's source coding theorem, that the average codelength for the network is bounded below by nodes' entropy. Based the duality of codenames' compression and network modules' structure, the measure of average description length of codenames provides equivalent measures of random walker dynamics.

For the multilayer system, the node-layer tuples is referred to as 'state nodes' i, α. The stationary distribution of state node i, α is p_i^{α} and can be derived recursively through

$$q_{ij}^{\alpha\beta} = p_i^{\alpha} P_{ij}^{\alpha\beta} \tag{2.46}$$

and

$$p_i^{\alpha} = \sum_{j,\beta} q_{ji}^{\beta\alpha} \tag{2.47}$$

Suppose that we have a partition M, and each state node i, α are allocated to a module $\mathcal{I} = 1, 2, \ldots, m$. The rate of transition for a random walker to move into each module $\mathcal{I}$ is given by

$$q_{\text{in}}^{\mathcal{I}} = \sum_{\{i,\alpha\} \in \mathcal{J} \neq \mathcal{I}, \{j,\beta\} \in \mathcal{I}} q_{ij}^{\alpha\beta} \tag{2.48}$$

and, for a walker to exit the module $\mathcal{I}$ is denoted by

$$q_{\text{out}}^{\mathcal{I}} = \sum_{\{i,\alpha\} \in \mathcal{I}, \{j,\beta\} \in \mathcal{J} \neq \mathcal{I}} q_{ij}^{\alpha\beta} \tag{2.49}$$

The module coding system is formulated such that every visit into a module, and each node visit within the module and exit of a module is given a codename. The rates of visit on every physical nodes of module $\mathcal{I}$ is given by

$$p_{i \in \mathcal{I}} = \sum\nolimits_{\{i,\alpha\} \in \mathcal{I}} p_i^{\alpha} \tag{2.50}$$

The module $\mathcal{I}$ has codename contribution for within module and exit movements (as inferred from exit rate $q_{\text{out}}^{\mathcal{I}}$ in Equation 2.49). The total rates for within and exit movements is given by $p^{\mathcal{I}}$, and its normalized probability distribution is given by $\mathcal{P}^{\mathcal{I}} = \{p_{i \in \mathcal{I}} / p_{\text{out}}^{\mathcal{I}}\}$. The codename contribution from entries to each module is derived from $q_{\text{in}}^{\mathcal{I}}$. The total rate for module entries is given by q_{in} and its corresponding normalized probability distribution is represented by $\mathcal{Q} = \{q_{\text{in}}^{\mathcal{I}} / q_{\text{in}}\}$. The per-step expected description length of codenames with respect to random walk dynamics for a multilayer network is formulated as

$$L(\text{M}) = q_{\text{in}} H(\mathcal{Q}) + \sum_{i=1}^{m} p_{\text{out}}^{\mathcal{I}} H(\mathcal{P}) \tag{2.51}$$

The objective is to find a partition with minimum description length to uncover the best community structure. The greedy algorithm that has been adapted for Infomap to detect communities in single layer network (Section 2.3) can also used to detect community structure for multiplex network.

In Ref. [64], Liu et al. considers community detection of multiplex networks with categorical coupling (where a node is connected to itself in every other layers). An 'edge pair', i.e. is the representation of two edges which are connected by a common node. For two edges, e_{ik} and e_{jk} that are connected by common node k, the edge pair is denoted as $\mathcal{E}_{ij}^{k}$. Liu et al. adapted the concept of edge pair from Ref. [65] for multilayer network by introducing 'link pair', which is represented as $\mathcal{L}_{ij}^{k}$, that denotes the fusion of intra-layer edge pairs and inter-layer edge pairs. The components of $\mathcal{L}_{ij}^{k}$ can both be intralayer edges, or a combination of intralayer and interlayer edges.

To extract link pairs from a multilayer network, all the layers in the network are to be merged. Each layer l in the multilayer network is represented by adjacency matrix A_{ij}^{l}. The merged network m can be represented by adjacency element A_{ij}^{m} in such a way that if there exist an edge in any of the layers, then the edge will be part of the merged network. This representation is mathematically represented as

$$A_{ij}^{m} = \begin{cases} 1 & \forall l, \exists A_{ij}^{l} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.52}$$

This approach will enable faster extraction of link pairs, as the links pairs are depicted by edge pairs in the merged network. Each link pair can be obtained for every edge pair in the merged network. The link pairs extracted from the merged network should instinctively contain intra-layer and inter-layer edge pairs respectively. Ahn et al. introduced similarity

measure for edge pair that is given by

$$S\left(\mathcal{E}_{ij}^{k}\right) = \frac{\left|N_{+}(i) \cap N_{+}(j)\right|}{\left|N_{+}(i) \cup N_{+}(j)\right|} \tag{2.53}$$

where $N_{+}(i)$ denotes the neighbours of node i. Liu et al. use the same formulation as above and adopt it to measure similarity for multilayer network as a function of link pairs. The similarity of link pairs for multilayer network is given as

$$S\left(\mathcal{L}_{ij}^{k}\right) = S_{\text{intra}}\left(\mathcal{L}_{ij}^{k}\right) + \alpha S_{\text{inter}}\left(\mathcal{L}_{ij}^{k}\right) \tag{2.54}$$

where the subscript 'intra' denotes both the edges $e_{ik} \in \mathcal{L}_{ij}^{k}$ and $e_{kj} \in \mathcal{L}_{ij}^{k}$ are components of intralayer edges, and 'inter' denotes one of the edges is part of interlayer edges. The similarity measure is characterized by intralayer and interlayer link pairs. A parameter $\alpha \in [0,1]$ is introduced as component weight for interlayer similarity.

Agglomerative hierarchical clustering technique is employed (refer to Section 2.1.3) to uncover communities. The algorithm to seek for communities starts by extracting all possible link pairs for the merged network as has been described earlier. Then, we calculate link pair similarity score $S(\cdot)$ for all possible combination of intralayer edge and interlayer edges in $\mathcal{L}_{ij}^{k}$. The link pair similarity scores are arranged in descending order, and each adjacent link pairs with smallest distance are joined together into a new community. By repetitively merging the clusters, hierarchical communities in the form of dendogram can be constructed.

2.5 Analysis of Biological Networks

The central dogma of molecular biology [3] follows a principle that DNA acts as central element in passing information to itself by replication and to RNA by way of transcription. Protein are the receiving end of information passed by the process of translation by RNA (Figure 2.4). The complex interactions between of cell's elements such as DNA, RNA and

proteins are translated in biological characteristics of a living cell. The diversity of DNAs is considered as the top-strata key elements that are responsible for biological evolutionary drive, which are performed downstream by proteins. Thus, DNA is considered as the main component that solely controls biological processes of living systems by ways of adapting itself to novel functions.

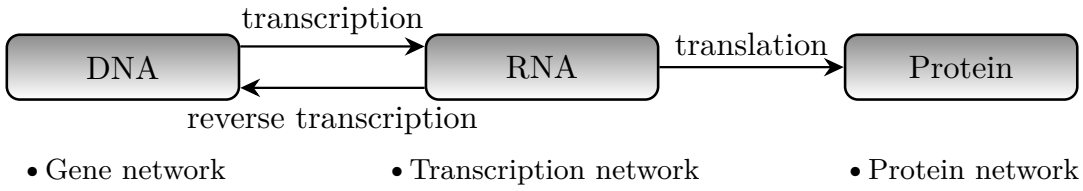


Figure 2.4: Central dogma of molecular biology. Based on the model proposed in Ref. [3].

A review of central dogma has been proposed by de Lorenzo in Ref. [4] and the evolution of thought in remodelling central dogma is illustrated in Figure 2.5. Looking closely at central dogma, it is worth to be critical why proteins are at the downstream of the information flow. A great deal of proteins function as enzymes that convert chemicals into components of biological systems. The properties of proteins is encoded in DNAs, but the chemical properties of metabolites that are involved in any chemical process that are catalyzed by proteins could not precisely be derived from DNAs. Based on this ground, it is necessary to explicitly include another layer of information flow, i.e. by adding one biological component that has been ignored in the central dogma which is 'metabolism'. Thus one can propose an extension of central dogma that incorporates metabolism, as shown in Figure 2.5B. The model is still gene-centric, but the evidence from bacterial pathogenesis and studies on the metabolome and the fluxome justify the addition of explicit connection from protein to metabolism. The information path between DNA, RNA and proteins are directly encoded and easy to visualize. The information flow from protein to metabolism is denoted by convulated path as it is more difficult to represent the transformation of protein sequences into chemical substrates and

fluxes. However, this model is still not adequate as we will explain in the next paragraph.

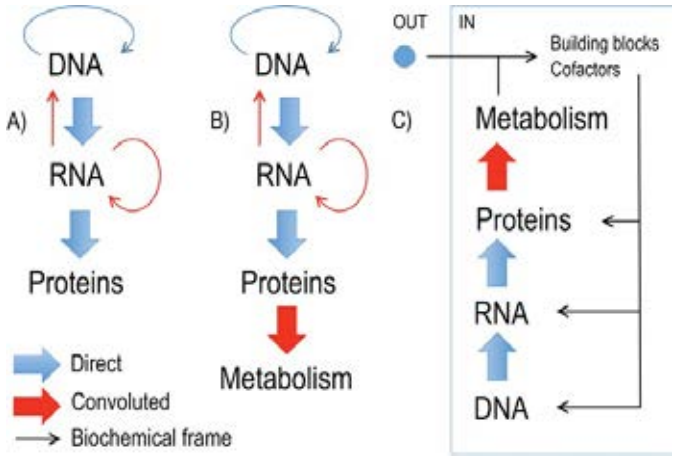


Figure 2.5: The extension of Central Dogma of molecular biology. (A) The classic central dogma. (B) Extended central dogma that incorporate metabolism. (C) Metabolism has commanding role in central dogma that constrains information flow. Metabolism ensures that the directive from DNA is not against the dynamics of biochemical network. Reprinted figure with permission from Ref. [4].

One issue that needs rethinking is the basis that classic central dogma ignores surroundings of living system, which implies that DNA, RNA and proteins are self-maintained system that flourish by itself without being influenced by external systems. The inclusion of metabolism can brings into central dogma the building blocks that are central for advancement of living system, and acts as the borderline between biological and non-biological entities. However, it is not sufficient to just include metabolism in central dogma as we have to consider another aspect of classic central dogma that views DNAs as the upper echelon components of any living system. Other biological elements are viewed as lower ranking workers that receive and execute commands directed by DNAs. In Ref. [66], Dawkins put a view in parallel with this idea, where the one and only goal of genes (that are encoded in DNA sequences) is to fulfil theirs own biological needs. This view implies that any living organism only functions to serve for the survival of the genes it embodies. However, very few can be achieved if the nature of genes' instructions try to work against the thermodynamics optima of biochemical

network. On the other hand, the evidence from adaptive virulence schemes (i.e. the ability of pathogens to infect their hosts) and the study of biodegradation of xenobiotic chemicals of environmental bacteria imply that the central evolutionary goal of living organisms is to gain metabolic optima rather than perpetuate and spread ones' own DNA sequences.

Based on the importance to biochemical optima as put forward earlier, de Lorenzo in Ref. [4] propose a new model for central dogma that considers metabolism as analogous to top ranking officer that propagates orders, which are then processed and archived by DNA for future references. DNA can be viewed as an information system that archives previously existed metabolic processes, and can later recreate the process by retrieving the information. The proposed model is illustrated in Figure 2.5C. In this model, metabolism is the dominant component, i.e. has a greater role than DNA in the central dogma. By impeding information flow, metabolism performs a role to ensure that any directive from DNA is not against the dynamics of biochemical network. The system is in the form of feedback loop where metabolism directs other components in the central dogma whether to advance with information flow, or to re-transcript/re-translate in accordance to current conditions. This model is not just directed from one top component in the central dogma. Situation can arise, e.g. the faulty enzymatic catalyzation on suboptimal substrates of oxygenases by oxidation-reduction reactions can form Reactive oxygen species (ROS) that mutates DNA, and trigger response that can accelerates the rate of genetic diversification. The information flow for $DNA \rightarrow metabolism$ is generally encoded and fairly consistent, while the flow for $metabolism \rightarrow DNA$ is unrestrictive and open to wider solution space.

We propose to study biological system in the form of complex network analysis of multilayer networks that are comprised of metabolic network and gene regulatory network. Thus we are focussing on the information that can be extracted from the metabolic network alone and by its integration with gene regulatory network, based on the topological properties and activities of in the network. The extended central dogma as proposed in Ref. [4] has inspired

us to study regulatory mechanism of biological systems by taking metabolism as the main component of interest.

For a formal definition, metabolism can be described as the biochemical reactions that are necessary for cellular functions by which some chemicals are transformed into the others [67]. Metabolism is represented by a number of metabolic reactions that are involved in conversion of the carbon source into building blocks needed for macromolecular biosynthesis. Lacroix et al. in Ref. [67] outlines that some of the products of a metabolic reactions becomes the reactants of the other reactions, in such a way that a complex network consisting of a collection of metabolic reactions and the relations among them is created, denoted as 'metabolic network'. 'Metabolite' is termed as the chemical substance involved in the reaction, either as a reactant or a product. These are small molecules that are imported/exported and/or synthesized/degraded inside a living cell. The compounds (called as products) are produced from a set of reactants (called as substrates). Although in theory a chemical reaction can occur for both directions (reversible), under a restrictive physiological conditions, some reactions can only occur in one direction (irreversible). Although some reactions are spontaneous, most reactions are catalyzed by one or several enzymes to speed up the reaction time. The 'enzyme' is a protein or a protein complex which are regulated by a single or several genes. A single enzyme may catalyze distinct reactions, and a single reaction can be catalyzed by several enzymes. The existence of small molecules called as 'cofactors' may play an essential role in the catalysis process by the enzyme, through a binding process, by enhancing or decreasing the activity of the enzyme.

Metabolism process is closely regulated by a process that controls the gene expression levels of mRNA and proteins. In can be represented in complex network by 'gene regulatory network', that is constructed by the connection between transcription factors (TFs) and the genes that are regulated by them. The genes in the networks are consisted of two possible types which are regulatory genes, or target genes. Regulatory genes codes for transcription

factors, and target genes are regulated by the transcription factors. In some cases, regulatory gene can take the position of target gene, e.g. as in auto-regulation process.

Metabolism has been in the limelight with the increasing interest in disease-oriented metabolic research especially the altered metabolic regulation in tumours [68], there are pressing needs to convert complex network understanding of biological system into novel diagnostic and therapeutic approaches. Determination of metabolic pathways can play useful roles in cancer studies e.g. by improving clinical treatments and in determining novel enzymatic drug targets and biomarkers for anticancer therapies [22, 23]. Several approaches have been developed to study intrinsic changes of metabolism between different conditions.

Schramm et al. [69] presented PathWave algorithm that implements wavelet transforms on metabolic graph embedded into 2D arrangement of metabolic grids, and subsequently mapped with gene expression data. PathWave enabled detections of switch-like regulation in the pathways of the tumour under study but it does not consider properties of metabolites in the pathways. PathWave analysis is limited to the projection of metabolic pathways into reaction-reaction graph. Breitling et al. [70] proposed GiGA algorithm, which evaluate significant metabolic pathways based from gene-gene interaction network projected metabolic network. Significant subgraphs are determined by greedy algorithm by using statistical ranking methods. The approach adapted in Ref. [69] and Ref. [70] are based on unipartite projection of metabolic network. PathWave reduced KEGG metabolic pathways into reaction-reaction network, while GiGA evaluate significant genes' interactions based on annotated pathways. In Ref.[71], Montañez et al. argues that unipartite projection do not fully capture the association between metabolites and reactions, and can lead to wrong interpretation of graph topological attributes. Montañez et al. suggested that metabolic network needs to be represented by both reactions and metabolites as separate nodes to render useful description and representative results.

2.6 Active Modules Identification

A very popular method to integrate networks together with genetic and biological profiles is known as 'active modules identification'. The change in molecular activities can be represented by the identification of the responsive subnetworks that marks the striking change of activities based on the 'omics' profiles, e.g. in the form of transcriptomic, proteomic and metabolomics data. There are three types of computational techniques that are widely, which are 'significant-area-search methods', 'diffusion flow methods' and 'clustering-based methods'.

2.6.1 Significant-area-search Approach.

The significant-area-search methods are widely derived from the work by Ideker et al. that is generally referred to 'jActiveModules', or simply as 'active modules'[21]. The work by Ideker et al. was the first to utilize the concept of 'hotspot' search as an algorithm framework. The algorithm has been widely employed in biological networks analysis to uncover network 'hotspots' associated to physiological and disease phenotypes. The general procedure of the algorithm starts with annotation of network nodes (molecules i.e. genes) with scores represented by the degree of molecular activities such as the measurements of gene expressions. An objective function is formulated that measures the aggregated scores of each sub-clusters that have been extracted from the networks based on the activities of member nodes. The details of 'active modules' of the score function is further described below.

First we determine the p-value p_i that denotes significant of expression change for each gene i in the network $G = (V, E)$. The gene i is assigned a z-score value defined as

$$z_i = \Phi^{-1}\left(1 - p_i\right) \tag{2.55}$$

for which Φ^{-1} corresponds to the inverse cumulative distribution function (CDF) of normal distribution.

Given a subnetwork M of size k nodes, the aggregate z-score for the module is defined as

$$z_M = \frac{1}{\sqrt{k}} \sum_{i \in V} z_i \tag{2.56}$$

Then, the corrected score of subnetwork M can be obtained as

$$s_M = \frac{(z_M - \mu_k)}{\sigma_k} \tag{2.57}$$

where μ_k and σ_k are the average value and standard deviation of the aggregate z-scores of a random subnetwork of size k obtained through Monte Carlo simulation.

Then, a module search strategy can be applied to identify high score subgraphs in the network that are characterized by high level of activities. As searching for the subnetworks is computational hard, heuristic algorithm, i.e. simulated annealing algorithm is adapted to uncover the active modules. The implementation of the simulated annealing algorithm is initialized by setting the number of iteration N and temperature T. For each node $v \in V$ in the graph G, set the node to either active or inactive with probability 1/2. Set subgraph of G with nodes that are active as active subgraph G_a. Then by each iteration for $i = 1$ until $i = N$, a node is randomly selected and its state is toggled. The score s_i (based on Equation 2.57) is computed. If the score is higher than before ($s_i > s_{i-1}$), the toggle is kept. Otherwise, the toggle is kept with probability $p = e^{(s_i - s_{i-1})/T}$. The subgraph G_a is updated with current set of active nodes. By final iteration, the subgraph G_a is chosen as the active module.

2.6.2 Diffusion Flow Approach

The diffusion flow method conceptualize on propagation of heat through a graph, where diffusion occurs at the nodes that are active, i.e. associated by strong molecular profiles such as differentially expressed genes. As the heat flow outwards from the nodes through the network edges, further active nodes could be identified that grow into clusters that maximize

the heat flow. The diffusion flow method is notably useful for clustering of genes based on the relationships formed by the graph's connectivity.

The work by Vandin et al. illustrate an application of diffusion flow method to identify 'significantly mutated modules', which are connected subgraphs that are comprised of genes that are more mutated than expected by chance [72]. To quantify the level of influence of node s on every other nodes in the graph, diffusion process is adapted based on the work in Ref. [73]. The fluid flow that is pumped at the source s, and leaks into a sink through the network's edges is a constant first-order rate γ. Let $\mathbf{f}^s(t) = \left[f_1^s(t), f_2^s(t), \ldots, f_n^s(t)\right]^T$, where $f_v^s(t)$ is the amount of fluid at node v at time t. Given that L is a Laplacian matrix of the graph, let $L_\gamma = L + \gamma I$. The dynamics of diffusion at time t is given by $\frac{d\mathbf{f}^s(t)}{dt} = -L_\gamma \mathbf{f}^s(t) + \mathbf{b}^s u(t)$, where $\mathbf{b}^s$ is vector with its element equal to 1 at the s^{th} entry, and 0 otherwise, and $u(t)$ is a unit step function. This process is analogous to the modelling of diffusion of heat on graphs by using heat kernel [74]. The system become equilibrium when $t \to \infty$ and at this steady state $\mathbf{f}^s = L_\gamma^{-1}\mathbf{b}^s$. By equating the network to the source-sink relationship, f_v^s can be likened to the influence exerted by gene s to v.

To uncover significant subnetworks, first we construct a network that contains all mutated genes, which we define as 'influence graph' $G_\mathcal{I} = (V, E)$, where V is the set of genes that has been identified as mutated. The influence each pair of genes in the network is defined by $\mathcal{I}(v_i, v_j)$ for all $v_i, v_j \in V$. The relationship generally is not symmetric, i.e. $\mathcal{I}(v_i, v_j) \neq \mathcal{I}(v_j, v_i)$. Define the weight of the edge between node w_i and w_j as $w(v_i, v_j) = \min\left[\mathcal{I}(v_i, v_j), \mathcal{I}(v_j, v_i)\right]$. The objective is to have nodes in the subgraphs that are connected through edges with high influence, i.e. high score of $w(v_i, v_j)$. The size of the subgraphs is constrained by a threshold λ. The procedure is initialized by removing from $G_\mathcal{I}$ for every edge with weight $w < \lambda$ and setting the maximum size of subgraph k. The algorithm will pick a node $v \in V$, and find the shortest paths p_{vu} to every other nodes $u \in V \setminus \{v\}$. At this stage, set S_{vu} as the set of nodes in p_{vu}, and P_{vu} as the set of elements in $\mathcal{I}$ which S_{vu} covers. Then, the algorithm construct a

connected subgraph C_v with v as initial element, and set P_{C_v} to be the set of elements that is covered by C_v. While $|C_v < k|$, pick u such that

$$u = \underset{u \notin C_v, |S_{vu} \cup C_v| \leqslant k}{\arg\min} \left\{ \frac{\left|P_{vu} \backslash P_{C_v}\right|}{|S_{vu} \backslash C_v|} \right\} \tag{2.58}$$

For a chosen u, the new connected subgraph is set to $S_{vu} \cup C_v$. Once the constraints has been met, the current solution is set as H. At the next iteration of $v \in V$, the best solution at that stage is compared to H, and the larger connected subgraph is kept. The computational demanding part is during the stage in finding shortest paths, which could be achieve in polynomial time.

2.6.3 Clustering-based Methods

The third approach to find community structure in biological network is through adaptation of 'biclustering', which is a concept of clustering that is based on topology interactions of components that are active. Although classical techniques such as graph partitioning and hierarchical clustering can uncover modular structures in network, the biclustering techniques could make use of omics profiles and interactions to extend and enhance network analysis beyond just topological profiling. Biclustering was initially implemented gene expression profiles analysis[75]. It is a distinct technique that perform concurrent row-column clustering. Thus, a bicluster can be regarded as a subset of columns that display consistent characteristics across a subset of rows.

The framework that has been proposed by Cheng and Church remains as one the most important work in the gene expression biclustering area. The biclustering method considers data in the form of an expression matrix $W = (X, Y)$, where X is the set of genes, and Y as the set of properties (commonly represented by expression levels). Let $I \subset X$ and $J \subset Y$ as the subset of genes and properties. Given pair (X, Y) , Cheng and Church defines the 'mean

squared squared residue' of a bicluster as

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left(w_{ij} w_{iJ} - w_{Ij} + w_{IJ} \right)^2 \tag{2.59}$$

where the mean of the ith rows is

$$w_{iJ} = \frac{1}{|J|} \sum_{j \in J} w_{ij} \tag{2.60}$$

the mean of the jth column is

$$w_{Ij} = \frac{1}{|I|} \sum_{i \in I} w_{ij} \tag{2.61}$$

and the mean of all elements is

$$w_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} w_{ij} = \frac{1}{|I|} \sum_{i \in I} w_{iJ} = \frac{1}{|J|} \sum_{j \in J} w_{Ij} \tag{2.62}$$

Mean squared residue of the bicluster $H(I,J)$ is the variance of of all elements, plus the mean row and column variance respectively. A perfect score $H(I,J) = 0$ indicates that the gene properties vary collectively. Every expression matrix, has a submatrix, which is trivially represented by a single element. Thus, a good submatrix should have maximum size, i.e. in term of size of genes or properties. Cheng and Church proposed an algorithm to find such submatrix. Consider W, and the maximum tolerable mean squared residue score $\delta \geqslant 0$. The iteration begins by calculate w_{iJ} and w_{Ij} $\forall i \in I$ and $\forall j \in J$ respectively, w_{IJ}, and $H(I,J)$. Return W_{IJ} as solution if $H(I,J) \leqslant \delta$. Then, compute $d(i) = \frac{1}{|J|} \sum_{j \in J} \left(w_{ij} - w_{iJ} - w_{Ij} + w_{IJ} \right)^2$ and $d\left(j\right) = \frac{1}{|I|} \sum_{i \in I} \left(w_{ij} - w_{iJ} - w_{Ij} + w_{IJ} \right)^2$. Update W_{IJ} by removing either row I or column J based on the variable with the larger $d(\cdot)$. Repeat the iteration with the updated W_{IJ}. This algorithm complexity is $O(nm)$. Cheng and Church also proposed a second algorithm with better algorithm complexity which is $O\left(m \log n\right)$ that allows multiple node deletion. The algorithm introduced a new parameter $\alpha > 1$ that corresponds to threshold for multiple deletion. At the iteration phase, as before we compute w_{iJ} and w_{Ij} $\forall i \in I$ and $\forall j \in J$ respectively,

w_{IJ}, and $H(I,J)$. Take W_{IJ} as solution if $H(I,J) \leqslant \delta$. Then remove every row $i \in I$ that satisfies inequality $\frac{1}{|J|}\sum_{j\in J}\left(w_{ij} - w_{iJ} - w_{Ij} + w_{IJ}\right)^2 > \alpha H(I,J)$. Recalculate w_{Ij}, w_{IJ}, and $H(I,J)$. Then, remove column $j \in J$ that satisfies $\frac{1}{|I|}\sum_{i\in I}\left(w_{ij} - w_{iJ} - w_{Ij} + w_{IJ}\right)^2 > \alpha H(I,J)$. If nothing is remove in this iteration, change to the earlier algorithm. Else, repeat the iteration with updated matrix.

Murali and Kasif introduce biclustering extension that is based on resampling technique [76]. Consider that each gene is assigned state in the form of p-value, for which state denotes how statistically significant for it to be in a certain class/group. Communities could be extracted in the form of biclusters, based on the constraint that each gene a community is precisely in the same state in all chosen samples. The procedure to find clusters in described in Algorithm 2.1. The idea is to get 'seed' samples and 'discriminating' set D. For every seed c and a discriminating set D, we can extract the largest cluster that is based on gene-states conditions in c that are satisfied by all elements of D. Murali and Kasif also set the conditions that the number of samples that satisfies the gene-states conditions must at least be α-fraction with respect to all the samples. The parameter α ensure that the cluster extracted from the network is supported by considerably high proportion of samples.

Algorithm 2.1 Finding xMotif Clusters

```
1: procedure FINDCLUSTER()
2:     Initialize number of 'seed' samples to pick n_s, number of discriminating sets n_D,
       and the size of discriminating set s_D
3:     for i = 1,...,n_s do
4:         Randomly pick a sample c
5:         for j = 1,...,n_D do
6:             Get a set D containing s_D number of samples that are randomly chosen
7:             Put pair (g,s) in the set G_ij If the state of every gene g ∈ V in c and in
               every element d ∈ D is s
8:             C_ij is set of all samples that agrees with all pairs in G_ij
9:             Set (C_ij, G_ij) if |C_ij| ⩾ αn
       return (C*, G*) associated to maximum |G_ij|
```

The development in 'active modules' [21] presents an opportunity to study biological

traits by integrating transcriptional information into nodes of the protein-protein interaction network. Areas in the network that are exposed to strong level of 'differential expression' are strongly hypothesized to correlate with the underlying regulatory circuits that governs the change of expression over different states. Active modules has been indicated to be effective in discovering molecular phenotypes and yield valuable information such as potential biomarkers for diagnostic and therapeutic of diseases [as reviewed 20].

There are several adaptation of active modules on bipartite metabolic network. One notable direction is by integrating flux balance analysis method (FBA) that calculate steady-state metabolic fluxes as a function of level of gene expression [24, 25]. A more recent approach in Ref. Bryant et al. [26] presented AMBIENT algorithm that uses bipartite metabolic network to extract significant metabolic pathways. AMBIENT take advantage of network structure of bipartite network that provide interaction of between metabolites and reactions, as well as enzyme-coding gene information. Therefore, score the function of a module are measured based on of reaction-metabolite interaction in comparison to protein-protein interaction in Ref. [21].

2.6.4 Active Modules Identification of Bipartite Network

In this section, the active module implementation on bipartite metabolic network in Ref. [26] is briefly described. Given an active subnetwork a with nodes r^a and m^a as the set of reactions and metabolites in the subnetwork respectively, the score of the subnetwork $S(a)$ is denoted by

$$S(a) = \ln(n)\left(\sum_i s\left(r_i^a\right) - \lambda \sum_j w\left(m_j^a\right)\right) \tag{2.63}$$

where $s(r_i^a)$ is the score of ith reaction, $m(r_j^a)$ is the score of jth metabolite , and $n = |r^a| + |m^a|$ is the number of nodes the subnetwork a. The constant λ denotes the balance factor that controls the size of the subnetwork.

Similar to Ref. [21], the active modules in the network is inferred by using simulated

annealing by capturing high-scoring subnetworks.

2.7 Inadequacies of Previous Approach

Many studies has been done to investigate the individual molecular traits e.g. [14, 77, 78, 79, 80]. Many single-type network analysis are proven useful in revealing functional mechanisms for biological system, (e.g. [14, 80, 81, 82, 83]). However, it is non trivial to notice that these networks are not independent. These single-type networks are part of interconnected networks that interactively communicate within and to other networks [6] (e.g., gene regulatory network interacts with protein-protein network, which then interacts with metabolic network) to complete the functions of the system as a whole. The ability to have a better view on interconnected network of networks of the biological network will provides us with useful insights on how the system behaves as a whole.

This draw us to the limitation of current implementation of active modules on bipartite metabolic network, where the significantly affected metabolic regions in the system is only considered to be influenced by gene expression profiles. Transcriptional regulation that affects the regulation of key enzymes over a long timescale, and the allosteric control of key enzymes along the metabolic pathways [84] have been ignored in the previous implementations. There has not been any research that integrates the metabolic pathways and gene regulation information in analyzing metabolism. Therefore, we propose the need to consider metabolism as interactions between a network that represents the series of chemical reactions that occurs within the cells (i.e. in the form of metabolic bipartite network), and a gene regulatory network that control the transcription of genes that affect the metabolic pathways through their key enzymes.

Civelek and Lusis suggested that interaction of molecular phenotypes across different biological spaces (i.e. include genes, transcript, protein, metabolites and microbiome) can be used to create a mapping with regard in clinical trait variation [6]. According to this

view, we can represent the different biological spaces as multilayer interconnected network. Nodes in the interconnected networks can be distinct, where each node is only exist in one layer. Intralayer connections between nodes denote relations within one biological subspace. There could be interlayer connected between nodes from different layer. This denotes 'coupling' between the nodes when they share the same property. For example a node in a gene regulatory network space are connected to a node in a protein-protein network space if they share the same gene property. The current limitation to analyze active modules as the interaction between metabolic and gene regulatory layer is the non-availability of module detection studies on interconnected network. The framework to detect community in interconnected network has yet to be implemented.

Another aspect that needs to be taken into consideration is the size of modules inferred by current implementation of active module identification on bipartite metabolic network. The AMBIENT algorithm encourages the formation of high-scoring modules that may includes low-scoring nodes or nodes with missing experimental data. These nodes act as 'bridging-node' that connects high-scoring nodes together, thus encouraging the formation of larger modules. This characteristics allows the inclusion of important low-scoring nodes which otherwise get excluded from the inferred modules. However, the creation of large modules means that the modules are not dedicated to small number of metabolic pathways. The modules could suffer poor precision with respect to many pathways in the modules and the results may render to be uninformative for post enrichment analysis on the modules. Although the modules could be considered to contains high false positive rates, we however believe that the formation of large modules indicates the existence of sub-modules (i.e. with their own dedicated metabolic functions) that are linked together to translate into overall biological functions as provided by the main modules. This indicates the need identify how the important pathways represented by these sub-modules are connected with each other, and how they affect the overall biological traits.

2.8 Conclusion

In summary, the study of biological mechanism should no longer be an isolated analysis of a single network. Biological process is a complex interactions of many different spaces. Therefore, it is useful to be able to integrate many different spaces when analyzing metabolism. In this thesis, we go one step further by proposing to model metabolism as interaction between two different space, which are metabolic space and gene space.

CHAPTER 3

Motif Projection on Metabolic Bipartite Network

In this chapter, we propose a framework to find active modules and their corresponding hierarchical clusters in a bipartite metabolic network. First, we describe the issues that motivates us to propose our framework. Then in Section 3.2, we describe the framework of our proposed modules identifications and clustering algorithm, and introduce a new measure of motif conductance for bipartite metabolic network. In Section 3.3, we validate our inferred modules by comparing our result with the current implementation of active module for bipartite network. Then we perform the algorithm to uncover hierarchical structures in the modules and validate our results with literature findings. Finally, the last section describes the contribution of this chapter.

3.1 Motivation

The currently available active module adoption of bipartite metabolic network, the AMBIENT algorithm, applies simulated annealing method to discover active module based on gene expression information and topology of the network. AMBIENT has the characteristic to uncover large active modules by connecting small high-scoring modules through low scoring nodes that act as bridging nodes [26]. The formation of large modules translates to poor

precision for selected pathways under investigation. Small modules allow better evaluation for the relevance of nodes to specific metabolic functions but they do not easily permit us to understand the global scope of metabolism that can be offered by large modules. Therefore, there is a need to decompose a large and differentially affected region in the bipartite metabolic network into smaller groups that can allow us to evaluate how different aspect of metabolic functions are linked together and affect the overall biological process.

Previous findings suggest that modules in biological networks (i.e protein-protein interaction network, gene regulatory network and metabolic network) exhibit hierarchical structure organization, in which a group of nodes in the networks is divided into smaller groups that further subdivided into smaller groups over multiple scales [14, 85, 86, 87]. Hierarchical organization in biological network has been shown to closely related to known metabolic functions [14]. Although there has been much development to study hierarchical structure in unipartite-like networks, there is no development in the aspect of bipartite metabolic network.

Therefore, we propose a clustering approach that segregates active modules into smaller clusters in a hierarchical manner. We propose MotifPro algorithm that projects the associations between reactions that are involves in active metabolic pathways in the form of motifs. Given an active region in a metabolic network, by implementing spectral clustering technique, active modules are partitioned into different regions characterized by high density of motifs. MotifPro can discover hierarchical active structures in the module (i.e. in the form of clusters), where each cluster are dedicated to smaller number metabolic functions. The regions of high concentration of motifs can be regarded as as active sub-pathways that collectively work together to achieve the overall results.

3.2 The Proposed MotifPro Algorithm

3.2.1 Interconnected Biological Network Representation

Consider a linked series of chemical reactions involved in a metabolic process. The metabolic model is represented by an undirected bipartite graph $G = (V, E)$, for which $V = R_{met} \cup M_{met}$ where $r \in R_{met}$ and $m \in M_{met}$ represents a reaction and a metabolite of the network respectively. The undirected graph is illustrated in Figure 3.1.

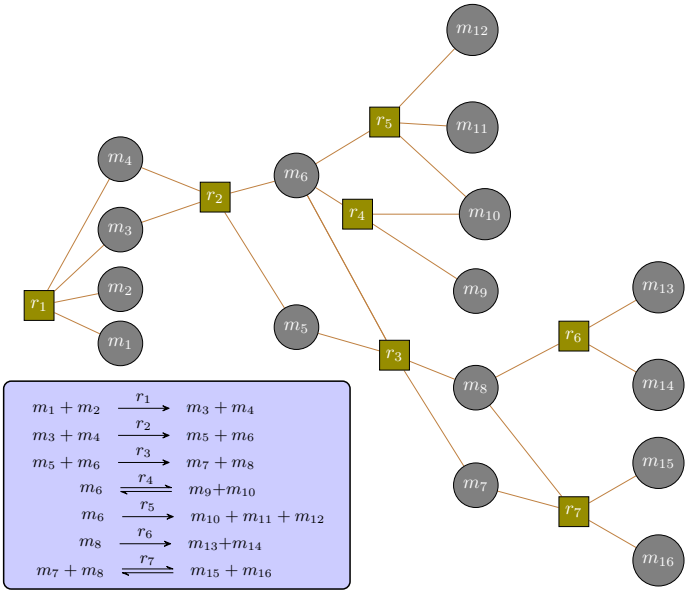


Figure 3.1: Example of Undirected Metabolic Network

3.2.2 Node Scoring Scheme

Each reaction r is given reaction score $s(r)$. The score given to the reactions can be based on biological properties of the reaction node in the network. For our analysis, we are using **log fold-change** as the base for reaction score. Each reaction $r \in R_{met}$ is assigned the log-fold change score of the enzyme-coding gene that catalyzes the reaction. When there are multiple enzyme-coding genes associated to the reaction, the mean score of the genes will be assigned.

In the event the reaction is not mapped to any genes (i.e no annotation of enzyme-coding genes or missing gene data) median scores of all other reactions will be assigned.

We define threshold τ to denote the active state of a reaction. A reaction is regarded as active when its score $s(r_i) \geqslant \tau$, and regarded as non-active otherwise. To identify active modules and its clusters that are associated to upregulation condition, τ should be set to a positive value (typically corresponds to 1.5 or 2 fold-change). To search for downregulated modules, all the reaction scores should be multiplied by -1 and τ to be set to positive value as before.

3.2.3 Proposed Motif-based Conductance of Bipartite Graph

Motif of m nodes is defined by $M(B)$, where B is a $m \times m$ binary matrix is the matrix that defines the pattern of directed edges between the m nodes. In our analysis of bipartite metabolic network, our motif of 3 nodes are derived from a set of node $\mathbf{v} \in V^3$, where we restrict $\mathbf{v}_1 \in M_{met}$ and $\mathbf{v}_2, \mathbf{v}_3 \in R_{met}$. Figure 3.2 illustrates the representation of motif of bipartite metabolic network as defined for our analysis.

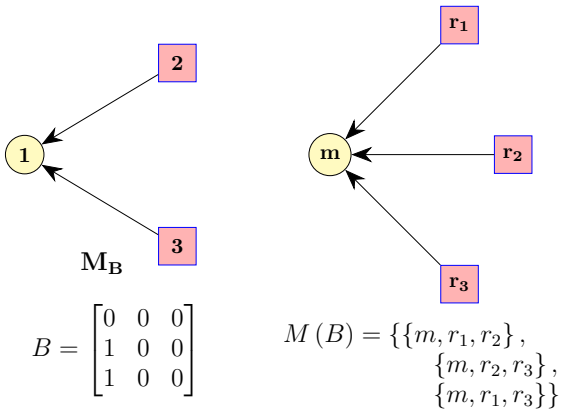


Figure 3.2: Motif of bipartite metabolic network M_B. The left of the graph shows the bipartite motif as defined by matrix B. There are 3 instances of M_B as illustrated on the right.

Let $set(\mathbf{v})$ to be defined as an operator that convert an (ordered) tuple to (unordered) set. For an unweighted (likely directed) graph having adjacency matrix A, the set of motifs is

defined by

$$M(B) = \left\{ \text{set}\,(\mathbf{v}) | \mathbf{v} \in V^3, v_2, v_3 \in R_{met}, v_2 \neq v_3, A_v = B \right\} \tag{3.1}$$

where A_v is a 3×3 adjacency matrix of a subgraph induced form a subset of 3 nodes from the ordered vector v. Any $\text{set}\,(\mathbf{v}) \in M(B)$ is called a motif instance. The set operator removes duplicates in $M(B)$ that exhibits symmetries.

For our bipartite network motif of 3 nodes M_B with corresponding weighted graph G_{M_B}, its motif adjacency matrix W_{M_B} is given by

$$\left(W_{M_B}\right)_{ij} = \sum_{\mathbf{v} \in M_B} \mathbf{1}\left(\{i, j\} \subset \mathbf{v}\right) \tag{3.2}$$

We proposed conceptually that

$$\text{cut}_{M_B}^{(G)}\left(S, \bar{S}\right) = \text{number of instances of motif being cut} \tag{3.3}$$

$$\text{vol}_{M_B}^{(G)}\left(S\right) = \text{number of instance of motif in } S \tag{3.4}$$

Our definition of $\text{cut}_{M_B}^{(G)}$ is the same as $\text{cut}_{M}^{(G)}$ in Equation 2.36 as defined by Benson et al. in [50]. Therefore,

$$\text{cut}_{M_B}^{(G)}\left(S, \bar{S}\right) = \sum_{\mathbf{v} \in M_B} \mathbf{1}\left(\exists i, j \in \mathbf{v} | i \in S, j \in \bar{S}\right) \tag{3.5}$$

Then by deriving from Equation 2.2, Lemma 2.2 and the definition of motif adjacency matrix in Equation 2.38 we obtain

$$\text{cut}_{M_B}^{(G)}\left(S, \bar{S}\right) = \frac{1}{2} \sum_{i \in S,\, j \in \bar{S}} \left(W_{M_B}\right)_{ij} \tag{3.6}$$

Volume for motif M_B can formally be defined as

$$\text{vol}_{M_B}^{(G)}\left(S\right) = \sum_{\mathbf{v} \in M_B} \mathbf{1}\left(\exists i \in S | i \in \mathbf{v}\right) \tag{3.7}$$

Next, we define volume in the form of $\text{cut}_{M_B}^{(G)}$ in Equation 3.6 as

$$\text{vol}_{M_B}^{(G)}(S) = \frac{1}{2} \sum_{i \in S,\, j \in \overline{S}} \left(W_{M_B}\right)_{ij} + \sum_{x \in S \cap M_{met}} \binom{N_x^S}{2} \tag{3.8}$$

where

$$N_x^S = \sum_{x \in S \cap M_{met},\ y \in S \cap R_{met}} 1_A(x, y) \tag{3.9}$$

and

$$1_A(x, y) = \begin{cases} 1 & \left(W_{M_B}\right)_{xy} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

The first term of Equation 3.8 denotes the number of motifs that are being cut, while the second term denotes the number of the rest of the motif in S. N_x^S is the number of reactions in S that are adjacent to metabolite x that is also in S.

We define Φ_M is the ratio of number of motif being cut in S,

$$\Phi_{M_B}^{(G)}(S) = \frac{\text{cut}_{M_B}^{(G)}\left(S, \bar{S}\right)}{\text{vol}_{M_B}^{(G)}(S)} \tag{3.11}$$

Finally we proposed the motif conductance $\phi_{M_B}^{(G)}$ as

$$\phi_{M_B}^{(G)}(S) = \max\left[\Phi_{M_B}^{(G)}(S), \Phi_{M_B}^{(G)}\left(\bar{S}\right)\right] \tag{3.12}$$

Suppose we have a directed and unweighted graph $G = (V, E)$. Let the weighted graph that corresponds to G (as specified by Equation 3.2) to be defined as G_{M_B}.

Lemma 3.1. Consider a directed and unweighted graph $G = (V, E)$, and a weighted graph G_{M_B} of a motif on bipartite network of 3 nodes. Then for any $S \subset V$,

$$6 \cdot \text{vol}_{M_B}^{(G)}(S) > \text{vol}^{\left(G_{M_B}\right)}(S)$$

□

Proof. Each three endpoints in an instance of motif contributes to 6 units in $\text{vol}^{(G_M)}(S)$. Motifs that are cut contribute less than 6 units to $\text{vol}^{(G_{M_B})}(S)$. Thus, $\text{vol}^{(G_{M_B})}(S) < 6 \cdot \text{vol}^{(G)}_{M_B}(S)$, as $\text{vol}^{(G)}_{M_B}(S)$ contains complete motifs with three endpoints. ∎

Lemma 3.2. Consider a directed and unweighted graph $G = (V, E)$, and a weighted graph G_{M_B} of a motif on bipartite network of 3 nodes. Then for any $S \subset V$,

$$6 \cdot \text{vol}^{(G)}_{M_B}(S) \leqslant \text{vol}^{(G_{M_B})}(S) + 4 \cdot \text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right) \qquad \square$$

Proof. Each instance of motif that are cut contributes either 2 or 4 units to $\text{vol}^{(G_{M_B})}(S)$. When 4 units are added to contribute to $\text{vol}^{(G_{M_B})}(S)$ for each partially cut motif, the total contribution will either be 6 units (i.e. the same as a contribution of one complete motifs with three endpoints), or higher. As $\text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right)$ are the total number of motifs being cut, the added contribution are $4 \cdot \text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right)$. Thus, $\text{vol}^{(G_{M_B})}(S) + 4 \cdot \text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right) \geqslant 6 \cdot \text{vol}^{(G)}_{M_B}(S)$. ∎

Theorem 3.3. Consider a directed and unweighted graph $G = (V, E)$, and a weighted graph G_{M_B} of a motif on bipartite network of 3 nodes. Then for any $S \subset V$,

$$3 \cdot \phi^{(G_{M_B})}(S) > \phi^{(G)}_{M_B}(S) \geqslant \phi^{(G_{M_B})}(S) \qquad \square$$

Proof.

$$\text{vol}^{(G_{M_B})}(S) < 6 \cdot \text{vol}^{(G)}_{M_B}(S) \leqslant \text{vol}^{(G_{M_B})}(S) + 4 \cdot \text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right)$$

$$\frac{3 \cdot \text{cut}^{(G_{M_B})}\left(S, \bar{S}\right)}{\text{vol}^{(G_{M_B})}(S)} > \frac{\text{cut}^{(G)}_{M_B}\left(S, \bar{S}\right)}{\text{vol}^{(G)}_{M_B}(S)} \geqslant \frac{3 \cdot \text{cut}^{(G_{M_B})}\left(S, \bar{S}\right)}{\text{vol}^{(G_{M_B})}(S) + 2 \cdot \text{cut}^{(G_{M_B})}\left(S, \bar{S}\right)}$$

$$3 \cdot \phi^{(G_{M_B})}(S) > \phi^{(G)}_{M_B}(S) \geqslant \frac{3 \cdot \text{cut}^{(G_{M_B})}\left(S, \bar{S}\right)}{\text{vol}^{(G_{M_B})}(S) + 2 \cdot \text{vol}^{(G_{M_B})}(S)}$$

$$3 \cdot \phi^{(G_{M_B})}(S) > \phi^{(G)}_{M_B}(S) \geqslant \phi^{(G_{M_B})}(S)$$

The first inequality is derived from Lemma 3.1 and Lemma 3.2. The second inequality is derived from Lemma 2.2. The third inequality is derived from $\text{vol}^{(G_{M_B})}(S) \geqslant \text{cut}^{(G_{M_B})}\left(S, \bar{S}\right)$. ■

Our motif adjacency matrix W_{M_B} (in Equation 3.2) and the corresponding weighted graph G_{M_B} are by definition equivalent to W_M (from Equation 2.38 as proposed in Ref. [50]) and G_M respectively. Based on the interchangeability between W_{M_B} and W_M, we use W_M to represent motif adjacency matrix and G_M as the weighted graph in explaining the methodology and algorithm in further sections.

3.2.4 Motifs-projection by Embedding Gene Expression Information

An active region is specified by a series of connected active nodes in the bipartite network. A bipartite metabolic network contains two type of nodes, which are reaction and metabolite. An active region would be comprised of a subgraph containing active reactions and metabolites. Every reaction is given score from which the active state of the reaction can be determined. Metabolites are not given any score, thus their active states are determined by the state of adjacent reactions.

To allocate the state of metabolites, we will consider a subgraph of an undirected bipartite metabolic network consisting of a metabolite which is connected to two active reactions (r_a and r_b) as shown in Figure 3.3. As there are no direction of the edges connecting the nodes, it is not possible to indicate whether the metabolite is a product or a substrate. However, we can provide three possible scenarios for the role taken by the metabolite:

1. The metabolite is an output (product) to an active reaction a and become an input (substrate) for active reaction b (or vice versa).

2. The metabolite is an output (product) of active reactions a and b.

3. The metabolite is an input (substrate) of active reactions a and b.

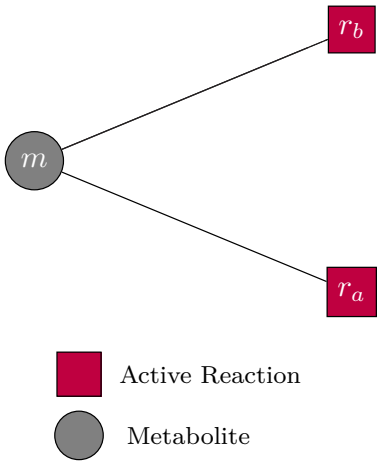


Figure 3.3: Active Metabolic Subgraph

When a metabolite is connected to two active reactions, the metabolite have the possibility to acts as a precursor that is needed for the communication between the two active reactions that constitute parts of active metabolic pathway. For other possible cases, the metabolite performs as a substrate or a product for two active reactions. Based on these scenarios, we deduce the importance of a metabolite by its connection to a subset of reactions adjacent to it. A metabolite is defined as active when it is connected to at least two active reactions. In any other situation the metabolite will be deemed as non-active.

A reaction is regarded as active when $s(r_i) \geqslant \tau$, and as non-active when $s(r_i) < \tau$. In our approach, we use the direction of edges to determine the state of a reaction, instead of looking at its score. We project directed edges on a bipartite metabolic network to provide associations between reactions which otherwise could not be connected in a bipartite metabolic network. Association between any two active reactions can be represented by the two reactions and the metabolites that act as intermediary nodes by connecting the two reactions. The associations between the reactions utilises the concept of second neighbourhood (i.e. a neighbour's neighbour of a reaction).

The state of each reaction is binary (i.e. either active or non-active) and can be represented

in the form of a directed edge. We use a convention that an active reaction will have directed edge pointing from the reaction towards any metabolite connected to it, and non-active reaction will have edge directed from its neighbouring metabolites towards the reaction as shown in Figure 3.4. The projected edges is made only with respect to reactions. The property

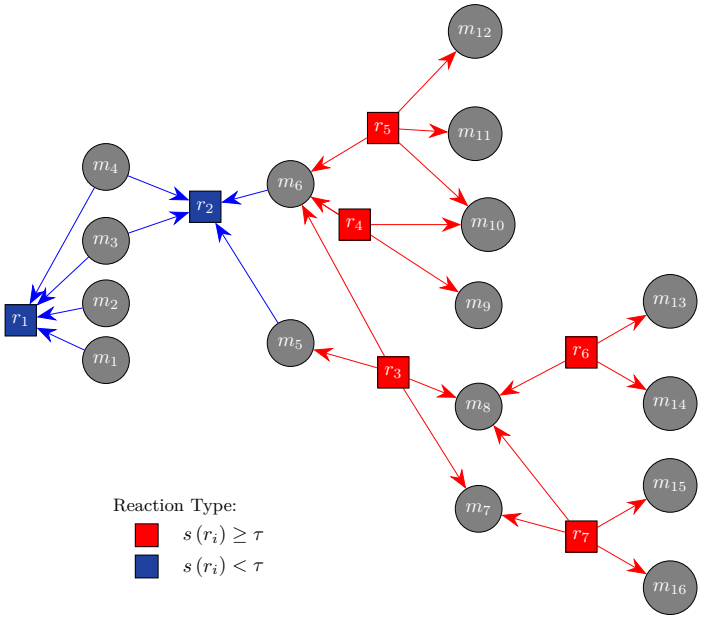


Figure 3.4: Projection of gene expression information on the network. The red nodes are active reaction nodes and the blue nodes are non-active reaction nodes.

of bipartite network (i.e. two nodes of the same type will never be connected) will ensure that there will be only one directed edge between a metabolite and a reaction. Any set of two reactions that has the same direction of edges on a metabolite will also has the same state. Therefore, to search for an active metabolic subgraph as depicted in Figure 3.3, we look for an intermediary metabolite connected to a set of two reactions and with both edges directed towards it.

Three different type of 3-node motifs that emerge in the network as a result of edge projection are shown in Figure. 3.5. The 3-node motif consists of two reactions and one metabolite. The metabolite that acts as intermediary node that connects the two reactions is taken as the base of projected 3-node motif M. Figure 3.3 shows motif M_1 as a direct

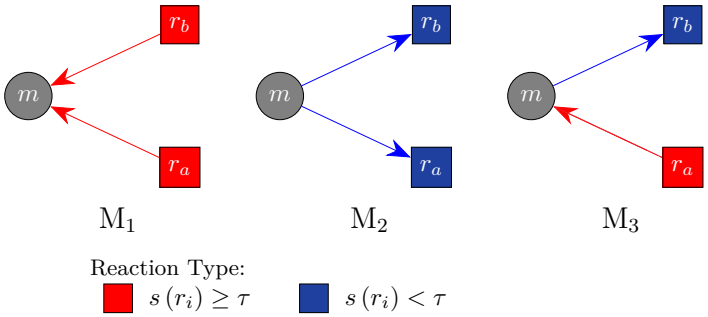


Figure 3.5: Motifs that are generated by the projection of gene expression information on bipartite metabolic network.

representation of active subgraph.

3.2.5 Identification and Hierarchical Clustering of Active Modules

The MotifPro algorithm is summarized in Figure 3.6, and described in Algorithm 3.1 on page 69. First, we identify active modules that are formed by connected subgraphs, that are linked through motif M_1 as a result of its edge projections by nodes' scores, as depicted by Figure 3.7A. The combination of each unit of connected components builds up the complete active modules for the bipartite metabolic network. Next, we construct motif adjacency matrix W_M of the active modules, where $(W_M)_{ij}$ is the number of instances of M_1 that contains node i and j (Figure 3.7B). When i is a metabolite and j is a reaction, $(W_M)_{ij}$ denotes the number of reactions that are connected to reaction j by its link to metabolite i. Whereas, when i and j are both reactions, $(W_M)_{ij}$ corresponds to the number of metabolites that provide connections between reaction i and j.

Next, we apply cutting strategy by splitting each connected component into different clusters while minimizing the number of motif M that are being cut (Figure 3.7C). This cutting strategy seeks for optimal concentration of motif M in the clusters. An ordered nodes σ can be obtained from the eigenvector of the second smallest eigenvalue of normalized Laplacian transformation of W_M (refer to Equation 2.39). The active modules are split into two smaller connected subgraphs (i.e S and $\overline{S}$), denoted by achieving minimum ϕ_M for nodes of increasing

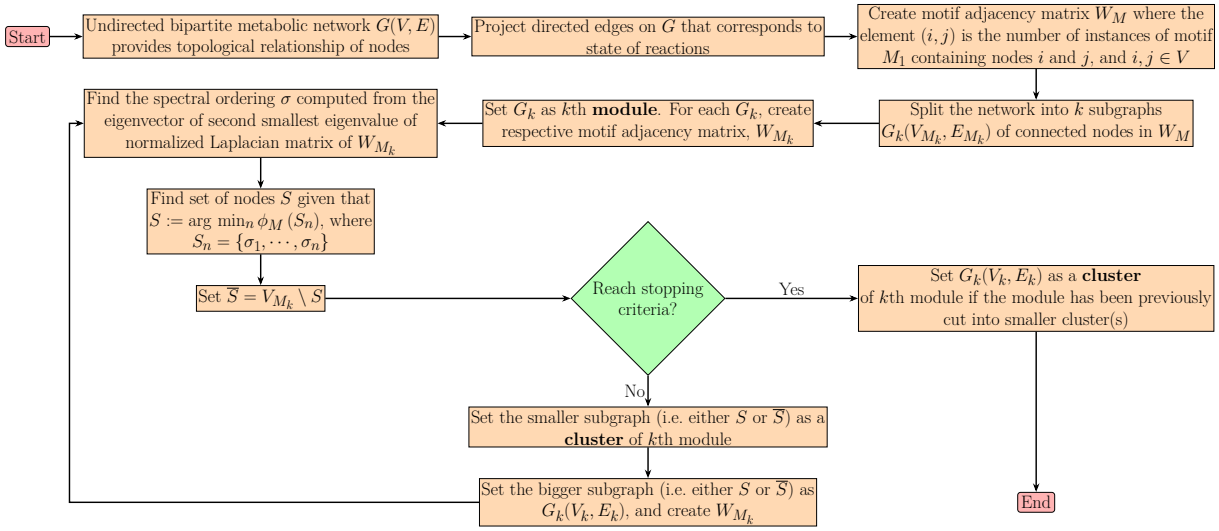


Figure 3.6: Workflow of MotifPro approach in obtaining active cluster from bipartite metabolic network.

order in σ. The smaller connected subgraph is taken as a cluster of the module, and the bigger connected subgraph is cut recursively until stopping criteria are reached. Figure 3.8 shows the procedure on naming of clusters. Figure 3.8A shows the events when we take the smaller subgraphs as clusters and recursively cut the bigger subgraph. We append 'C#' to the module where 'C' denotes a cluster and '#' denotes the ordered the clusters are obtained. When the subgraph could not be cut, it will taken as the last cluster. Figure 3.8B shows the scenarios where where we also recursively cut the smaller clusters that has been obtained. Another 'C#' will be appended to the name of the cluster it is derived from.

3.2.6 Proposed Stopping Criteria

The general idea for recursive clustering is to take the smaller subgraph as a cluster of the module, and to recursively cut the bigger subgraph until stopping criteria are reached. We use two stopping criteria, γ and Ψ. The first stopping criteria is 'cut threshold' γ that corresponds to the quality of cut and, is defined as the maximum allowable percentage for number of motif being cut. Therefore, we can set reasonable value of γ by what we considered as maximum proportion to be cut. Motif corresponds to the component of chemical reactions that are

Algorithm 3.1 MotifPro Algorithm

Input: Motif Projected Metabolic network $G_M = (V_M, E_M)$ and motif M. $V_M = R_{met} \cup M_{met}$, reaction $r \in R_{met}$ and metabolite $m \in M_{met}$

1: **procedure** ACTIVE MODULE SEARCH(G_M, M)
2: Create motif adjacency matrix W_M where the element (i, j) is the number of instances of motif M containing nodes i and j, where $v_i, v_j \in V_M$,
3: Set CC = connected components in W_M,
4: Set cluster $C_k = \varnothing$ and $\overline{C}_k = \varnothing$ for each kth connected component in CC,
5: **for** each kth connected component in CC **do**
6: Create motif adjacency matrix W_{M_k} from induced subgraph $G_M\left[V_{M_k} \subset V_M\right]$,
7: Find the spectral ordering σ computed from the eigenvector of second smallest eigenvalue of normalized Laplacian matrix of W_{M_k},
8: Find set of nodes S given that $S := \arg\min_n \phi_{M_B}(S_n)$, where $S_n = \{\sigma_1, \cdots, \sigma_n\}$,
9: Set $\overline{S} = V_{M_k} \setminus S$
10: **if** Stopping criteria not reached **then**
11: Append $C := \arg\min\left(|S|, \left|\bar{S}\right|\right)$ into cluster C_k,
12: Append $\bar{C} := \arg\max\left(|S|, \left|\bar{S}\right|\right)$ into subgraph $\overline{C}_k$,
13: **else**
14: Append V_{M_k} into cluster C_k
15: Form C_k as cluster of each kth connected component.
16: Recursively cut subgraph $\overline{C}_k$ of each kth connected component until reaching stopping criteria.

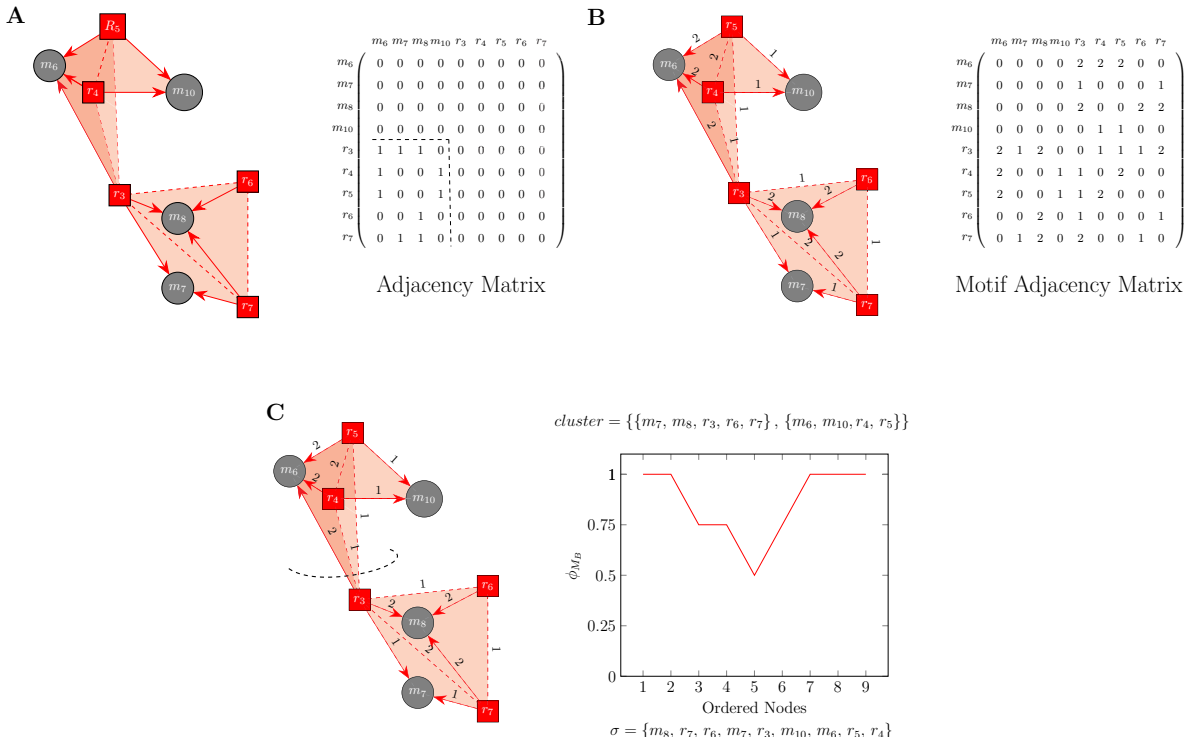


Figure 3.7: Identification of cluster by minimizing motif conductance ϕ_{M_B}.

part of metabolic pathways that are responsible for certain biological functions. Therefore, γ should be set reasonable low to prevent cutting through a large number of metabolic pathways.

The second stopping criteria, 'deviation threshold' Ψ, is the minimum allowable average deviation (AD) around the minimum value of ϕ_M within the lower nth samples of the ϕ_M scores, defined as $\frac{1}{n}\sum_i^n \left(\phi_{M,i} - \min\left(\phi_M\right)\right)$. The number of sample n taken in this study are the data within lower 10% of the ϕ_M scores. AD measures the dispersion of solutions around the minimum value of ϕ_M. When AD is small, it gives an indication of the existence of other solutions within small proximity of the best solution, which render the cut to be ineffective.

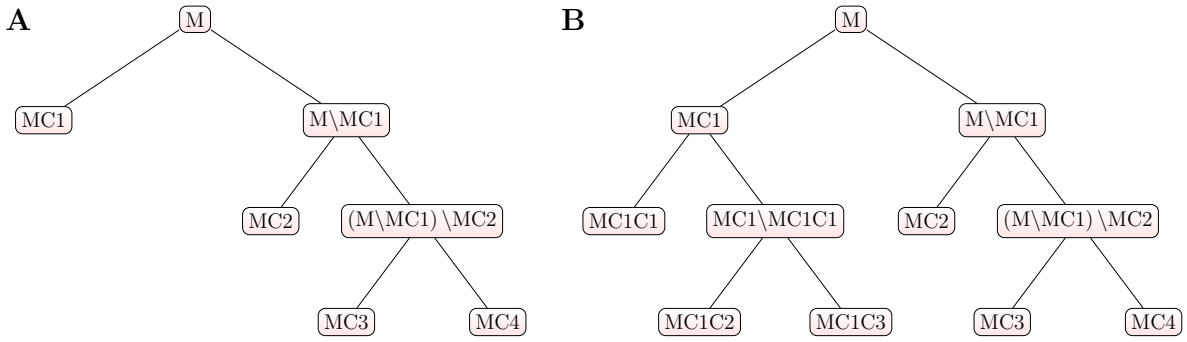


Figure 3.8: Naming rules of clusters. 'M' is the module name, which can be represented by 'U' or 'D' that corresponds to upregulation or downregulation respectively. (A) Smaller subgraphs are taken as clusters the bigger subgraph are recursively cut. 'C#' is appended to the module where 'C' denotes a cluster and '#' denotes the ordered the clusters are obtained. When the subgraph could not be cut, it will taken as the last cluster. (B) When the smaller clusters (MC1 in this example) are also recursively cut, another 'C#' will be appended to the name of the cluster it is derived from.

3.3 Experiments and Results

3.3.1 Dataset and Preprocessing

To validate our proposed algorithm, we use *Saccharomyces cerevisiae* (yeast) metabolic model during diauxic shift condition (i.e shifting from fermentation to respiration) where ethanol becomes carbon source upon depletion of glucose. The yeast model is represented by a bipartite metabolic network (i.e. in the form of SBML model) based from YEASTNET [88] prepared by author of AMBIENT algorithm [26]. The genes' fold-change scores that represent change of gene expressions during diauxic shift periods are taken from microarray expression data by DeRisi et al. [1]. The fold-change scores are used to assign node scores of genes and reaction nodes of the metabolic network. We removed currency metabolites based on the classifications by authors in Ref. [89, 90]. These metabolites are ATP, ADP, NADP, NAD, NADPH, NADH, hydrogen phosphate, diphosphate, water, carbon dioxide, oxygen, ammonia and proton. The metabolic network contains 3008 nodes from which 1377 and 1631 are

metabolite and reaction nodes respectively, and 4421 edges.

The reactions in the metabolic model are given scores based on the log fold-changes of genes at 20.5 hour timepoint during experiment to investigate metabolic reprogramming during diauxic shift period of yeast [1]. Both the yeast data and reaction scores are assembled from sources by Bryant et al. [26]. The network can readily be used by AMBIENT and MotifPro algorithms.

3.3.2 Active Module Identification

AMBIENT and MotifPro were run to for both upregulated and downregulated conditions. The AMBIENT Algorithm was run by using the same parameters as in as in Bryant et al. and we present the results from our best run. MotifPro is run by setting $\tau = 1.5$ fold-change. The cut-off value is chosen based on the work in [91, 92] that recognize significant genes to be those with fold-change score of at least 1.5. Post analysis is conducted in the form of over-representation analysis (Refer to Appendix A for the description of the over-representation method). Analysis is conducted on modules obtained by the algorithms to identify the modules that are significantly over-represented based on the pathways in Yeast Metabolome Database (YMDB) [93]. The over-representation analysis is performed on the list of metabolites in modules, and are conducted by using MBROLE [94]. We use naming convention where upregulated module and downregulated modules are preceded by 'U' and 'D' respectively, and followed by the module number. AMBIENT obtained 3 upregulated modules that are significantly over-represented. The first module is associated to TCA cycle and glyoxylate and dicarboxylate metabolism; the second module is related to pyruvate metabolism; and the third module is associated to starch & sucrose metabolism. For downregulation condition, AMBIENT obtains a large downregulated module consisting of 812 nodes. However, the module is not included in the analysis as it does not render significantly over-represented pathways. MotifPro uncovered three upregulated modules, and three down-regulated mod-

ules. Both MotifPro and AMBIENT replicate the findings of DeRisi et al. for upregulation condition, showing that TCA cycle is significantly affected during diauxic shift. The general comparison between MotifPro and AMBIENT is shown in Table 3.1. In the biggest module, there are 76 metabolites in MotifPro as compared to 29 metabolites in AMBIENT.

In Table 3.2, the comparison between MotifPro and AMBIENT were made based on the three significant pathways during diauxic shift as curated by DeRisi et al. [1]. For AMBIENT, TCA and glyoxylate cycles were in module U1, while starch and sucrose metabolism were in module U2. For MotifPro, however, all the three significant pathways were located in module U1. MotifPro indicates better inferences for TCA and glyoxylate cycles, but lower inference for starch and sucrose metabolism as compared to AMBIENT.

All the modules found by MotifPro can be referred to in Table 3.3. Several of the modules corresponds to pathways that are reported to be affected during diauxic shift by several experimental results. The upregulated module U1 contains nodes related to TCA cycle, gyoxylate cycle and starch & sucrose metabolism pathways were reaffirmed to be upregulated in Ref. [1]. Module U2 are related to peroxisomal fatty acid degradation. Enzymes for fatty acid beta-oxidation were found to be induced in the presence of ethanol, accompanied by the expansion and biogenesis of peroxisomes [95, 96]. The third upregulated module U3 that corresponds to inositol phosphate metabolism could indicate the induction of nuclear func-

Table 3.1: Comparison of modules between MotifPro and AMBIENT

	MotifPro	AMBIENT
No. of significant Modules	6	3
No. of reactions in top module	103	89
No. of metabolites in top module	76	29
P-value of top pathway (TCA Cycle)	$< 10^{-11}$	$< 10^{-5}$
F-score of top pathway (TCA Cycle)	0.400	0.343

Summary statistics of modules obtained by MotifPro and AMBIENT during diauxic shift. Module significance is determined based on YMDB pathways.

Table 3.2: Comparison between modules obtained by MotifPro and AMBIENT based on upregulated metabolic pathways identified by DeRisi et al. [1]

	MotifPro	AMBIENT
Citrate cycle (TCA cycle)		
Module	U1	U1
P-value	9.69E-12	4.77E-06
Precision	0.2609	0.2692
Recall	0.8571	0.5
F-Score	0.4	0.35
Glyoxylate cycle		
Module	U1	U1
P-value	5.00E-09	4.63E-05
Precision	0.2609	0.2692
Recall	0.6000	0.3500
F-Score	0.3637	0.3043
Starch and sucrose metabolism		
Module	U1	U2
P-value	4.08E-03	2.14E-06
Precision	0.1304	1.0000
Recall	0.3333	0.2222
F-Score	0.1875	0.3636

Induced metabolic pathways in [1] are cross-checked with the modules that are obtained by MotifPro and AMBIENT. 'U' denotes an upregulated module and followed by module number. Statistical measures are determined based on YMDB pathways.

Table 3.3: Active modules obtained by MotifPro

Upregulated Module	U1	U2	U3
Total no. of nodes	179	47	9
No. of metabolites	76	20	5
No. of reactions	103	27	4
Main Pathway	TCA cycle	FA metabolism	IP metabolism
Downregulated module	D1	D2	D3
Total no. of nodes	359	45	24
No. of metabolites	143	20	10
No. of reactions	216	25	14
Main Pathway	CM metabolism	VLI biosynthesis	OCP by folate

Description of abbreviations: FA - Fatty Acid, IP - Inositol phosphate, CM - Cysteine and methionine, VLI - Valine, leucine and isoleucine, OCP - One carbon pool. 'U' and 'D' denotes an upregulated and downregulated modules respectively, which is followed by module number. Module significance is determined based on YMDB pathways.

tions that relates to mRNA export, transcriptional regulation and telomere homeostasis [97]. The downregulated module D1 and D2 mainly correspond to the catabolism of amino acid production and the reduction of fatty acid biosynthesis. Amino acid concentration is reported to be heavily reduced, or converted to catabolism during the diauxic shift period [98]. Module D3 could be precursor to the reduction of amino acid as it is associated to one carbon pool by folate function, which is important for the biogenesis of purines, serine and methionine [99].

3.3.3 Uncovering Hierarchical Structure in Active Modules

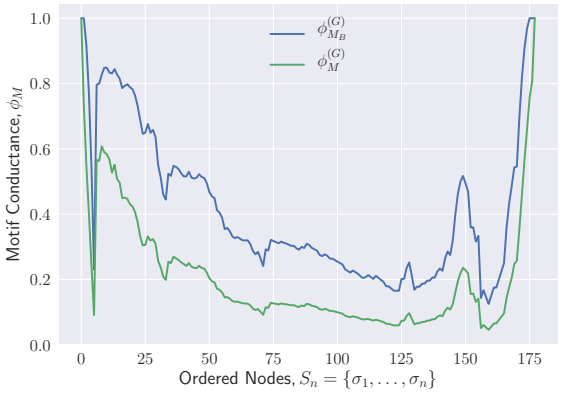
The next step of MotifPro algorithm seeks to partition the modules that has been discovered to separate clusters. We run the algorithm with the scoring function set in accordance to our proposed motif conductance ϕ_{M_B} by setting the stopping criteria parameters $\gamma = 0.2$ and $\Psi = 0.02$, and make a comparison to ϕ_M (refer to Equation 3.12) as proposed by Benson et al. [50] (refer to Equation 2.33). The cut threshold γ is set to allow a small proportion of motifs in the subgraph to be cut while not being too restrictive. The low value of deviation threshold Ψ was chosen to restrict modules from being cut when there exist other solutions close to the optimal solutions. Based on our empirical runs, the selection of the stopping criteria

parameters above allows high proportion the compounds that belong to the main metabolic pathways to remain concentrated in the same clusters (i.e main pathways are not divided into different clusters).

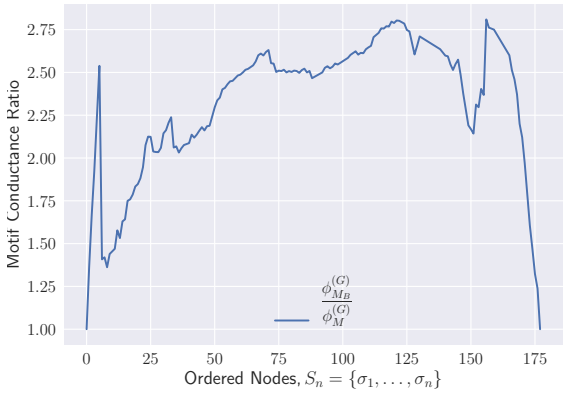
Figure 3.9 shows the tabulation of ϕ_{M_B} and ϕ_M for the upregulated module U1. The minimum ϕ_{M_B} in the first cut, second and third cut is 0.1250, 0.1875 and 0.1370 respectively. The sweep cut was stopped by the third cut as the deviation score was less than the threshold. The corresponding minimum ϕ_M are 0.0445, 0.0690 and 0.0482 respectively. Module U2 and U3 are not able to be cut as the cut threshold were exceeded. The minimum ϕ_{M_B} of the first cut for module U2 and U3 are 0.2581 and 0.8333 respectively, and the corresponding ϕ_M are 0.0979 and 0.4167 respectively. The tabulation of ϕ_{M_B} and ϕ_M for the downregulated module D1 is shown in Figure 3.10. The minimum ϕ_{M_B} for the first, second and the third cuts were 0.0887, 0.1250 and 0.2500 respectively. The cut threshold was exceeded at the third cut. The corresponding ϕ_M for the first, second and the third cut are 0.0306, 0.0435 and 0.0909 respectively. The downregulated modules D2 and D3 were not able to be partitioned as the minimum ϕ_{M_B} exceed γ at 0.2500 and 0.2667 respectively. The corresponding value of ϕ_M for D2 and D3 are 0.0909 and 0.0982 respectively.

The results shows that the adaptation of motif conductance score ϕ_{M_B} and ϕ_M infers the same clusters as the point of cuts for both scores are similar. However, the value of ϕ_{M_B} are approximately 2.8 times ϕ_M at the point of minimum conductance. Although both ϕ_{M_B} and ϕ_M can be used to partition the graph into smaller clusters, ϕ_{M_B} which is defined as the proportion of motif being cut is intuitive to serve as reference for the stopping criteria γ.

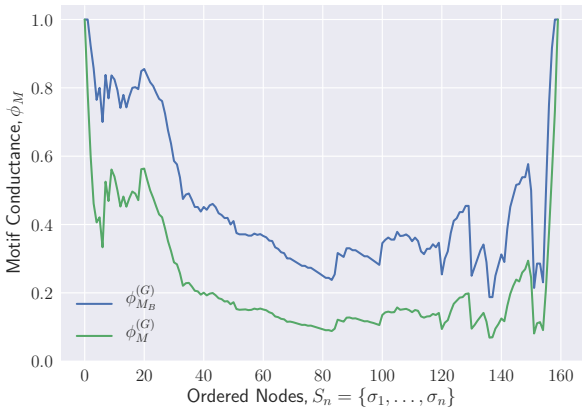
In Table 3.4, the node properties and main pathways for upregulated module U1 and downregulated clusters D1 obtained by MotifPro is shown. We use naming convention where a cluster is named after the module it is derived from, and followed by 'C' and the cluster number. The detail comparisons between upregulated module U1 and the clusters derived from it are given in Table 3.5. Comparisons are made by measuring the values of p-value, recall,



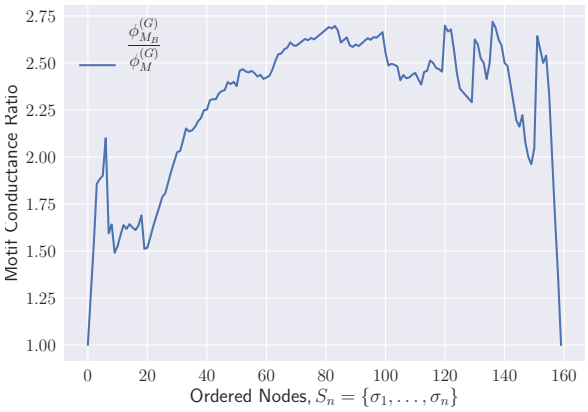
(a) 1st Cut: $\min\left(\phi_{M_B}^{(G)}\right) = 0.1250, \min\left(\phi_M^{(G)}\right) = 0.0445$



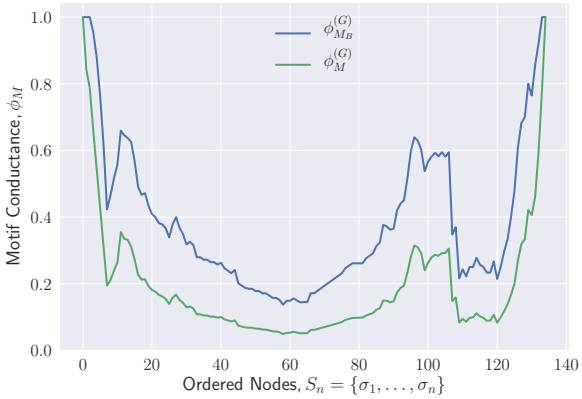
(b) 1st Cut: $\max\left(\phi_{M_B}^{(G)}/\phi_{\mathrm{M}}^{(\mathrm{G})}\right) = 2.8095$



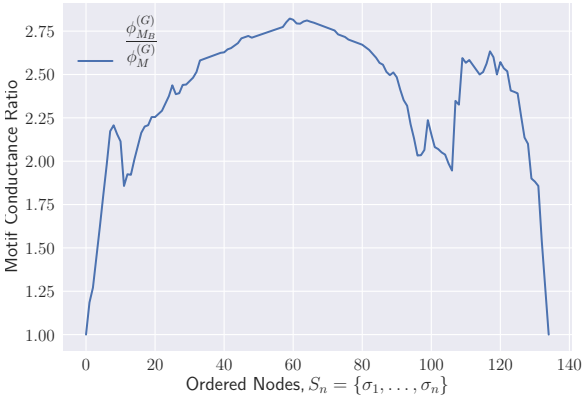
(c) 2nd Cut: $\min\left(\phi_{M_B}^{(G)}\right) = 0.1875, \min\left(\phi_M^{(G)}\right) = 0.0690$



(d) 2nd Cut: $\max\left(\phi_{M_B}^{(G)}/\phi_{\mathrm{M}}^{(\mathrm{G})}\right) = 2.7186$



(e) 3rd Cut: $\min\left(\phi_{M_B}^{(G)}\right) = 0.1370, \min\left(\phi_M^{(G)}\right) = 0.0482$



(f) 3rd Cut: $\max\left(\phi_{M_B}^{(G)}/\phi_{\mathrm{M}}^{(\mathrm{G})}\right) = 2.8227$

Figure 3.9: Tabulation of ϕ_{M_B} and ϕ_M and the ratios $\phi_{M_B}/\phi_{\mathrm{M}}$ for the upregulated module U1

Table 3.4: Active clusters of U1 and D1 obtained by MotifPro

Upregulated Cluster	U1C1	U1C2	U1C3
Total no. of nodes	19	26	134
No. of metabolites	7	10	58
No. of reactions	12	15	76
Main Pathway	Glutathione	SS metabolism	TCA Cycle
Downregulated module	D1C1	D1C2	D1C3
Total no. of nodes	51	16	292
No. of metabolites	21	8	114
No. of reactions	30	8	178
Main Pathway	FA biosynthesis	Steroid biosynthesis	CM metabolism

Description of abbreviations: SS - Starch and sucrose, FA - Fatty Acid, CM - Cysteine and methionine. Cluster is named after the module it is derived from, denoted by 'C' and the cluster number. Module significance is determined based on YMDB pathways.

precision and F-score. The data shows that the precision and F-score in each clusters were increased. The overall measure significance of the clusters were higher (by the measure of p-value) as compared to the original module. The results shows that the module cut approach is able to segregate the densely connected motifs in the modules into different clusters in accordance to its main pathways.

Figure 3.11 further illustrates the components of U1 which is partitioned into three clusters. The first cluster, cluster U1C1, corresponds to glutathione metabolism. Glutathione concentration was reported to decrease during the early stage of yeast fermentation, but increase during nutrient starvation and stationary stage [100]. Glutathione also serve an important role in responding to the depletion of sulfur and nitrogen in cells [101]. When the cells are deprived of methionine, homocysteine and cysteine as the external source for sulfur, glutathione can serve as endogenous sulfur source [102], and increase in concentration for up to approximately 10% to serve as a sulfur source [as cited by 100]. The reduction of methionine and cysteine concentrations has been indicated in the earlier results, denoted by downregulated module D1 for which cysteine and methionine metabolism is the top module.

The second and the third cluster of U1 reaffirm previous result by DeRisi et al. [1]. U1C2

Table 3.5: Comparison between upregulated module U1 and its derived clusters

	Module	Cluster
Glutathione metabolism		
Module	U1	U1C1
P-value	9.83E-03	2.35E-02
Precision	0.1087	0.3333
Recall	0.3333	0.1333
F-Score	0.1639	0.1904
Starch and sucrose metabolism		
Module	U1	U1C2
P-value	4.08E-03	1.36E-06
Precision	0.1304	0.7143
Recall	0.3333	0.2778
F-Score	0.1875	0.400
Citrate cycle (TCA cycle)		
Module	U1	U1C3
P-value	9.69E-12	1.31E-13
Precision	0.2609	0.3529
Recall	0.8571	0.8571
F-Score	0.4000	0.5000
Glyoxylate and dicarboxylate metabolism		
Module	U1	U1C3
P-value	5.00E-09	7.47E-11
Precision	0.2609	0.3529
Recall	0.6000	0.6000
F-Score	0.3637	0.4444

Comparisons are based on the main metabolic pathways in the derived clusters. Statistical measures are determined based on YMDB pathways.

corresponds to starch and sucrose metabolism, the essential step where UDP-Glucose is converted into glycogen. U1C3 is associated to three main metabolic pathways which are TCA cycle, glyoxylate and dicarboxylate metabolism, and glycolysis/gluconeogenesis pathways. U1C3 is seen as an important cluster in converting ethanol as carbon source and channelling the products towards starch and sucrose metabolism in U1C2. U1C3 contains essential metabolite Acetyl-CoA in mitochondrion compartment, which is an important compound in TCA cycle and it also acts as linkage to glyoxylate and dicarboxylate metabolism and

glycolysis/gluconeogenesis pathway. As part of glycolysis/gluconeogenesis pathway, acetyl-CoA is produced by the conversion of acetaldehyde. Acetaldehyde is first converted acetate by enzyme aldehyde-dehydrogenase and subsequently to acetyl-CoA by acetyl-CoA synthethase in mitochondrion compartment. With the upregulation of pyruvate carboxylase, pyruvate is rerouted from acetaldehyde by directing it to oxaloacetate. This is an important step in reversing the flow of metabolite in glycolysis/gluconeogenesis pathway towards the starch and sucrose metabolism pathway in U1C2. Through glyoxylate and dicarboxylate metabolism, isocitrate is converted to succinate in either by isocitrate lyase, or to 2-Oxo-glutarate by isocitrate dehydrogenase (NADP). 2-Oxo-glutarate is further converted to succinate along the TCA cycle pathway.

Table 3.6: Comparison between downregulated module D1 and its derived clusters

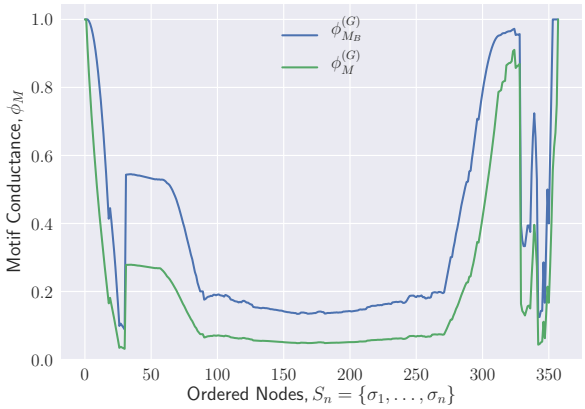
	Module	Cluster
Fatty acid biosynthesis		
Module	D1	D1C1
P-value	5.76E-03	2.73E-11
Precision	0.065	0.5333
Recall	0.7273	0.7273
F-Score	0.1193	0.6154
Steroid biosynthesis		
Module	D1	D1C2
P-value	2.32E+01	3.47E-05
Precision	0.0569	1.0000
Recall	0.3043	0.1304
F-Score	0.0959	0.2307
Cysteine and methionine metabolism		
Module	D1	D1C3
P-value	1.38E-02	2.40E-03
Precision	0.1301	0.1524
Recall	0.4000	0.4000
F-Score	0.1963	0.2207

Comparisons are based on the main metabolic pathways in the derived clusters. Statistical measures are determined based on YMDB pathways.

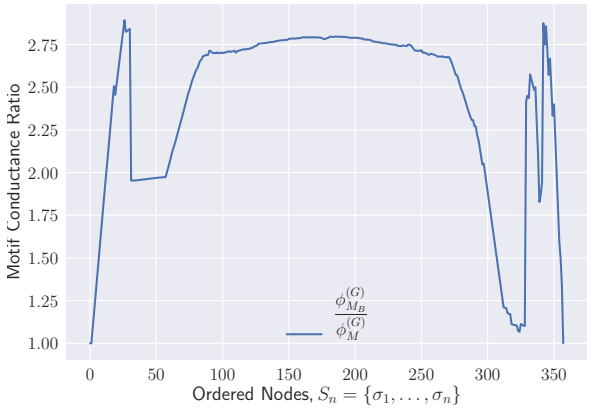
Table 3.6 shows the comparison between downregulated module D1 and the clusters that have been derived from it. Compartmentalized in cytoplasm, the downregulated cluster D1C1 is linked to the decrease of catalyzation rate for fatty acid biosynthesis. This process is reported to be regulated by SNF1 to decrease biosynthesis of lipid in glucose-depletion condition [103, 104]. SNF1 phosphorylates acetyl-CoA carboxylase enzyme, thus reducing the biosynthesis of fatty acid. In conjunction to that, SNF1 stimulates generation of energy during diauxic shift by the induction of fatty acid metabolism through β oxidation that result in enlargement of peroxisomes (refer to upregulated module U2).

Cluster D1C2 is associated to the reduction of zymosterol biosynthesis during the presence of ethanol as growth medium. Zymosterol is an intermediate metabolite for ergosterol, which is a major element of cell membrane. The steroid synthesis was reportedly reduced during the presence of ethanol partly due to the reduction of zymosterol and fecosterol, which is its direct intermediate [105]. However, there are reported increase of ergosterol at the highest ethanol concentration, suggesting ethanol tolerance are in correlation with ergosterol concentration.

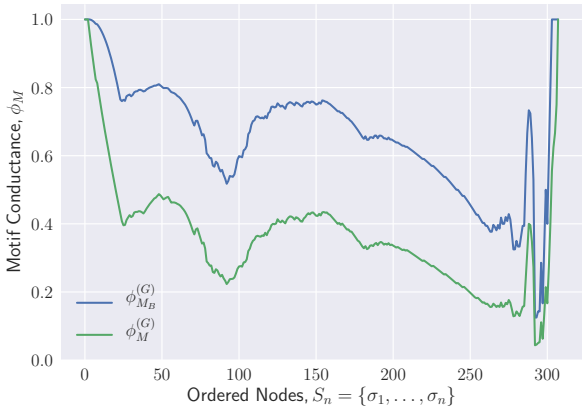
The third downregulated cluster D1C3 mainly corresponds to the decrease of amino acid biosynthesis. The main metabolic pathways in D1C3 are cysteine and methionine metabolism (sulfur-containing amino acids), glycine, serine and threonine metabolism, and histidine metabolism. D1C3 also relates to purine metabolism. There are interrelation between downregulated D1C3 and upregulated U1C1, where glutathione counterbalances the depletion of sulfur due to reduction of methionine and cysteine by increasing its concentration to serve as sulfur source.



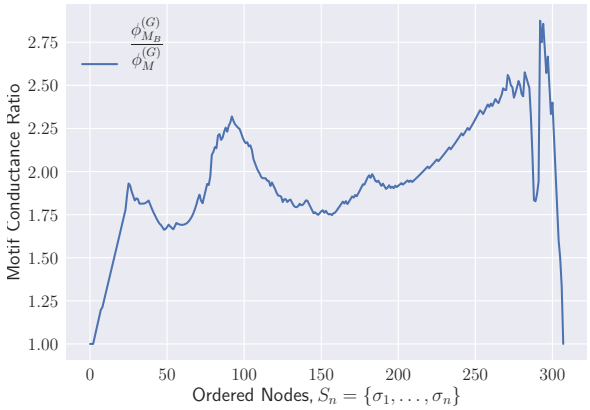
(a) 1st Cut: $\min\left(\phi^{(G)}_{M_B}\right) = 0.0887, \min\left(\phi^{(G)}_{M}\right) = 0.0306$



(b) 1st Cut: $\max\left(\phi^{(G)}_{M_B}/\phi^{(G)}_{M}\right) = 2.8934$



(c) 2nd Cut: $\min\left(\phi^{(G)}_{M_B}\right) = 0.1250, \min\left(\phi^{(G)}_{M}\right) = 0.0435$



(d) 2nd Cut: $\max\left(\phi^{(G)}_{M_B}/\phi^{(G)}_{M}\right) = 2.875$

(e) 3rd Cut: $\min\left(\phi^{(G)}_{M_B}\right) = 0.2500, \min\left(\phi^{(G)}_{M}\right) = 0.0909$

(f) 3rd Cut: $\max\left(\phi^{(G)}_{M_B}/\phi^{(G)}_{M}\right) = 2.75$

Figure 3.10: Tabulation of ϕ_{M_B} and ϕ_M and the ratios ϕ_{M_B}/ϕ_M for the downregulated module D1

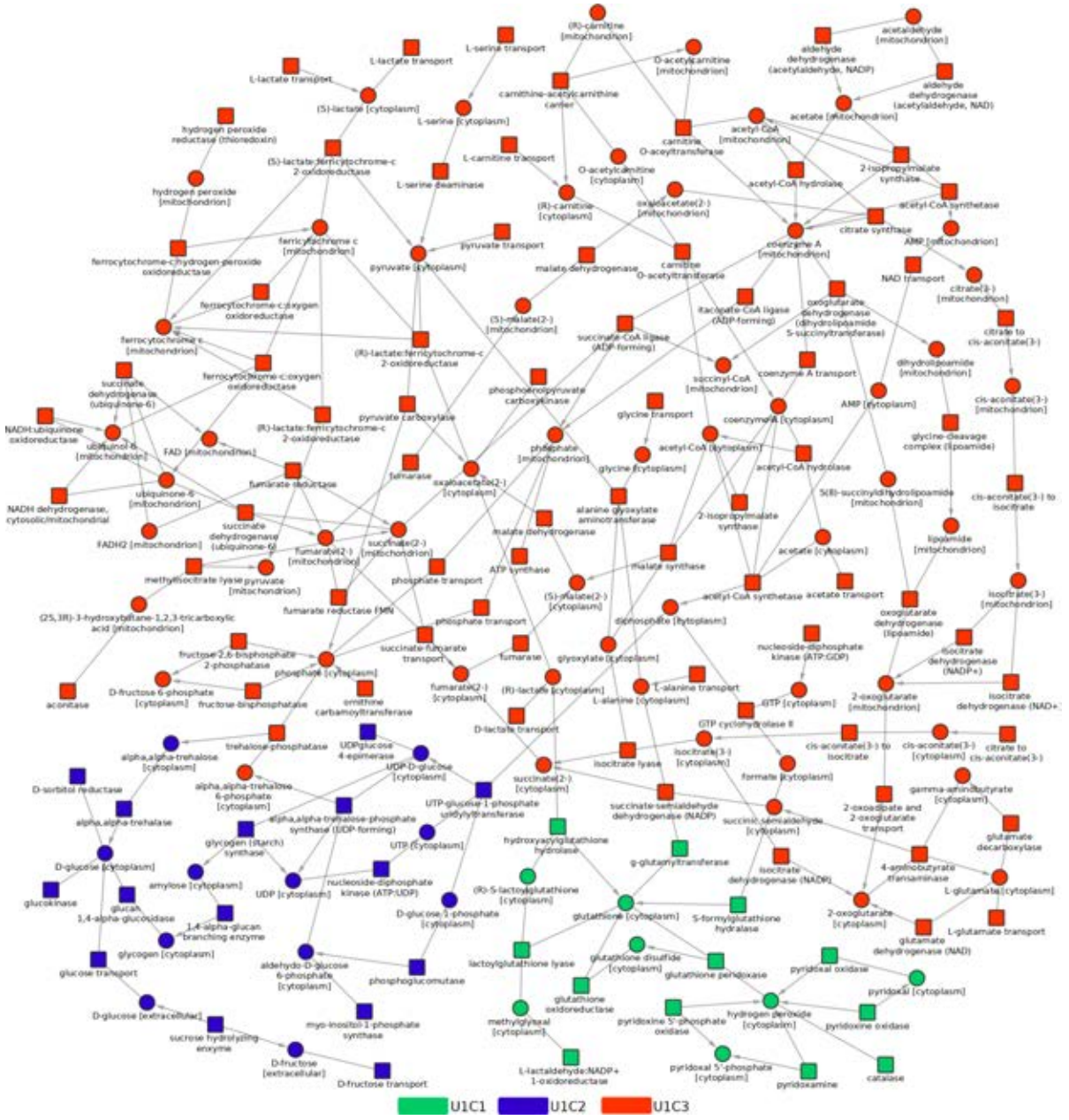


Figure 3.11: **Active clusters of upregulated module U1 obtained by MotifPro.** U1C1, U1C2 and U1C3 are the three clusters that have been derived from the upregulated module U1. The main pathways that corresponds to the clusters are: glutathione metabolism for U1C1; starch and sucrose metabolism for U1C2; and TCA cycle for U1C3.

3.4 Conclusion

To represent changes during diauxic shift period encountered by yeast, we mapped log fold-change scores of enzyme-encoding genes to reactions of its generic metabolic network. Directed edges are projected into the network based the state of reactions. Connected subgraph consisted of motif of interest are extracted to represent active modules of the network. Then, the active modules are cut into smaller clusters, having constraint of minimizing the percentage of motifs being lost to each cut.

We are able to segregate the largest upregulated and downregulated modules into several smaller clusters. For the diauxic shift dataset, the largest module typically consisted of one large cluster of densely connected motif, which are sparsely connected to other small clusters of densely connected motif. The largest module for upregulated condition are partitioned into three clusters. The biggest cluster corresponds to TCA cycle and, glyoxylate and dicarboxylate metabolism. The two smaller clusters corresponds to glutathione metabolism and, starch and sucrose metabolism respectively. For downregulated condition, pathways related to amino acid metabolism (e.g. cysteine and methionine metabolism, histidine metabolism and, glycine, serine and threonine metabolism) are associated to the largest cluster. The other two smaller clusters correspond to fatty acid and steroid biosynthesis respectively. The metabolic pathways that we discovered are supported by experimental analysis.

The clustering approach are able to detect smaller pathways in the large module which that may be overlooked due to their higher p-values and and lower precision scores. And many of the clusters are cleanly partitioned, i.e. a metabolic pathway associated to one cluster mostly do not associate to other clusters. For upregulated condition, starch and sucrose metabolism, TCA cycle and glyoxylate and dicarboxylate are cleanly cut into one cluster. For downregulated condition, fatty acid biosynthesis and the related amino acid metabolism exclusively exist in its own cluster. Our clustering approach could not ensure that all clusters are cleanly cut. Some

pathways are separated into different clusters e.g. glutathione metabolism in upregulated module, and steroid steroid biosynthesis in downregulated module. Overall, however, we demonstrate through clustering approach, modules are consisted of densely connected motifs which corresponds to certain biological function, that are sparsely connected to each other. Clustering is an useful way to segregate an active module into different clusters in accordance to their associated biological functions.

The score function is formulated as extension of motif conductance score function by Benson et al. [50]. The definition of our score function as the percentage of motif that are being cut is intuitive for end users in evaluating the quality of the cut and setting up the stopping criteria for the algorithm. The introduction of calculation of the number of motif in the clusters do increase the execution time for the algorithm in comparison to the original formulation. However, MotifPro algorithm is reasonably fast in evaluating bipartite metabolic network, i.e. with execution time of about 1 minute in evaluating clusters for both upregulation and downregulation conditions of the yeast data. When evaluating large network where execution time is critical, the score function by Benson et al. can easily be used to evaluate the projected motifs. However, the stopping criteria (i.e. in term of percentage of motif being cut) would need to be approximated.

CHAPTER 4

Active Modules of Bipartite Metabolic Network by Information Flow Approach

In this chapter, we propose an algorithm to model metabolism by finding modules in the form of significant regions that show striking changes in metabolism activities in metabolic network, and strong regulation of genes in the gene regulatory network. First, we introduce the motivation that drive us to propose the algorithm. Then, we describe the framework and schemes adapted by the algorithm in detail in Section 4.2. Then in Section 4.3, we evaluate the performance of our inferred topological modules (i.e. taking into account of nodes' degree distribution) and active modules (i.e. taking into account of activity of nodes and supplemented by topological properties) by producing comparisons with other benchmark algorithm. In Section 4.4, we apply the algorithm to glioma network to find disease modules to aids us in identifying potential biomarkers for the disease. Finally, Section 4.5 and 4.6 present the discussion of our results and conclude the contribution of this chapter.

4.1 Motivation

In this chapter, our main objective is to find modules in a biological bipartite metabolic network. There has been renewed interest on the study metabolism with the knowledge of altered metabolism in cancers that has been more progressive than earlier anticipated [106, 107]. Although plenty of studies has been done in detecting communities in genes and protein-protein interaction networks [108, 109, 110, 111], the work on bipartite metabolic network is still in infancy. AMBIENT algorithm [26] shows a promising step towards in this direction, however it only takes into account of changes in transcription levels and ignore the effect on gene regulation on the metabolic system. Previous work on community detection of metabolic network do not fully utilized the topological structure of bipartite network but project the network into unipartite network [69], or by assembling network of genes associated to the metabolic pathways [70].

The approach to detect modules is either by observing topological group that reveals higher concentration of links internally rather than externally when compared to random model [112], or by identifying nodes that flow within the modules within relatively long period [51, 113]. Infomap algorithm [51] captures flow process through the network system and has been shown useful in identifying modules in weighted and directed network. However, we can not directly apply community detection in bipartite network by using Infomap as the random walk on a bipartite network is periodic. As a workaround, Alzahrani et al. in [114] applied Infomap on bipartite network by projecting it into unipartite network based on common neighbours similarity.

Regulation of metabolism is another factor to look at when analysing biological metabolic network. Apart from allosteric control of key enzymes by particular substrates (e.g. work in [115]), the mechanism of transcriptional regulation also affects regulation of metabolism [84]. Therefore, it is useful not only to analyze a metabolic network by itself, but instead to

treat the metabolic system as an interaction between two different layers of biological system i.e. metabolic and gene regulatory layers. The needs to analyzed metabolism by the inclusion of global integration of transcriptomic, proteomic and metabolomic analysis has been emphasized by the reviewers in Ref. [84]. To date, however, there has not been any progress in integrating regulatory information in identifying active module for bipartite metabolic network. The integration of regulatory layer brings an additional advantage in overcoming problem of periodicity problem by random walker when analyzing bipartite network on its own. The Markov chain system of the integrated multilayer network becomes aperiodic and ergodic thus allowing the system to be analyzing through random walker methodology.

On this account, we are considering a community detection model that explores the search space in a two-layer interconnected network that includes a bipartite network integrated by the regulation support of a gene regulatory network. Community detection algorithms generally assume only a single type of static link in evaluating interaction between nodes in the network. This assumption oversimplifies many interaction in real world networks. If we treat the interaction between nodes separately in each layer, the important associations between the networks could not be captured. There is progress in community detection work on multilayer networks in the form of multiplex networks [116, 117]. However, there are currently no studies that are being done on the interaction between a bipartite network and a unipartite network. Complexity increases in interconnected network as different type of link may exist in each layer and we need to consider the interlayer links that connect the metabolic and regulatory networks.

One of the pressing question is, what is a community in a two-layer interconnected network? In our experimental analysis, by applying Infomap [51] on the interconnected network, it is shown that the many modules are constricted of the type of nodes mainly from only one layer. This confirm our statement earlier that community algorithm will not work well when there are more than one type of static link in the network. We propose a

heuristic algorithm to address the issue, named as ActiveFlow, with the objective that a unit of community of interconnected system of networks should strive to consist both components of each layers as a signal of information flow between them. The proposed algorithm is able to detect modules consisted of nodes from both the bipartite metabolic layer, and the regulatory layer. The modules inferred by the algorithm give an indication of information flow between the metabolic and regulatory layers, thus depicting an useful model of regulation of metabolism.

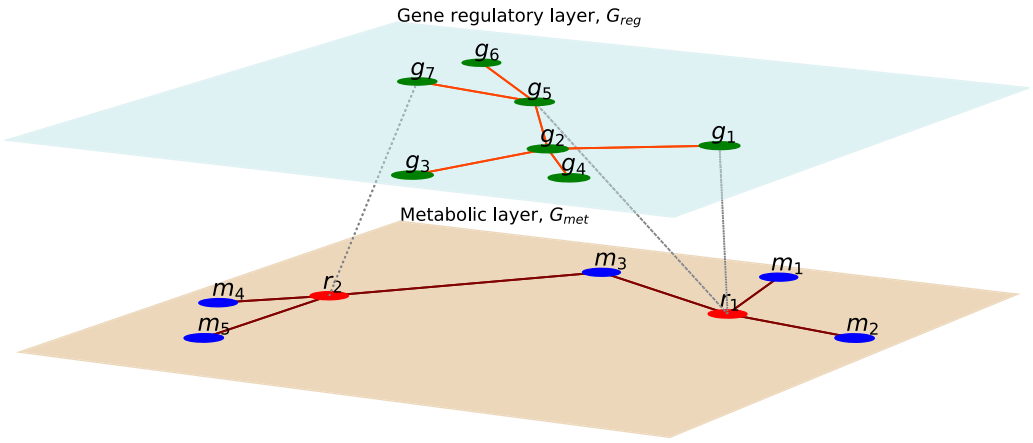
4.2 The Proposed ActiveFlow Module Detection Algorithm

4.2.1 Interconnected Biological Network Representation

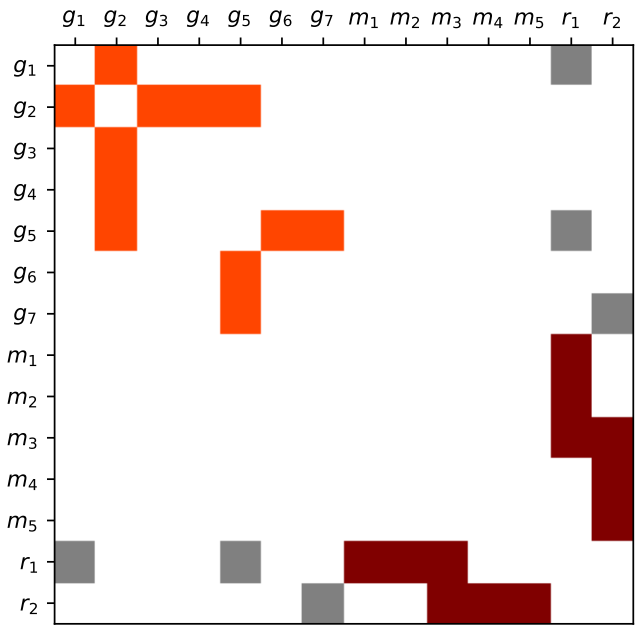
We are considering an undirected bipartite metabolic network $G_{met} = (V_{met}, E_{met})$ and an undirected gene regulatory network $G_{reg} = \left(V_{reg}, E_{reg}\right)$. The nodes of the metabolic network are $V_{met} = R_{met} \cup M_{met}$ with reactions $r \in R_{met}$ and metabolites $m \in M_{met}$. The nodes of the gene regulatory network are gene $g \in V_{reg}$ that corresponds to transcription factors and target genes. These two networks are merged into one undirected and weighted interconnected network defined as $G = (V, E)$ where $V = V_{met} \cup V_{reg}$. The edges of the interconnected network are $E = E_{met} \cup E_{reg} \cup E_{in}$ where $e \in E_{in}$ is the inter-layer edge (g, r). Edge (g, r) is created when a gene $g \in G_{reg}$ is also an enzyme-coding gene of reaction $r \in V_{met}$.

We illustrate the representation of the interconnected biological network in Figure 4.1. The Figure 4.1a shows the interconnected graph G in multilayer-network formalism, and Figure 4.1b shows supra-adjacency matrix of the graph. The interconnected graph G is formed by the combination of metabolic layer G_{met}, and gene regulatory layer G_{reg}. G_{met} is comprised of nodes which are either metabolites $m \in M_{met}$ or reactions $r \in R_{met}$. The intralayer edges in this layer are connections between metabolites and reactions, and a node will never be connected to other nodes of the same type (i.e a metabolite or a reaction node are not adjacent to other metabolites or reactions respectively). The second layer G_{reg} is only

composed of genes $g \in V_{reg}$, and the intralayer interactions in this layer are produced by links between the genes. There are intralayer edges that connects the two layers, created by connections by a set of reactions in the metabolic layer to the genes in the gene regulatory layer.



(a)



(b)

Figure 4.1: Representation of the interconnected biological network. **(a)** Interconnected graph G in multilayer-network formalism. **(b)** Supra-adjacency matrix of interconnected graph G.

4.2.2 Node and Edge Weights Scoring Scheme

The reactions in the bipartite metabolic network, and the genes in the regulatory network are assigned scores based on transcriptomic data. We are implementing microarray data analysis in this work, thus the node scores will be based on the **fold-change** of intensity level of genes compared between treated and controlled conditions. Metabolites are not assigned any score as there are no suitable score that can be derived for chemical compounds from microarray analysis. The node scores are assigned as a precursor to assigning weight to edges in the network.

We are using fold-change score as the base for the node scores as we are interested in relative change of activities between the treated and control conditions. Fold-change describes how much a quantity has changed from its initial condition. Given an initial value as A, and final value as B, fold-change is calculated as the ratio B/A. When a gene is upregulated, its fold-change score will be more than 1, and if the gene is downregulated the score will be between 0 and 1. The determination the threshold for fold-change to categorized genes as significance is rather arbitrary. In Ref. [91], a gene will only be given consideration for its significance if shows a fold-change of 1.5. In Ref. [92], statistically significance genes ($p < 0.01$ or $p < 0.05$) are ranked by fold-change of 1.5, 2 or 4. In this work, we are using score of 2 as a cut off threshold.

For a gene $g_i \in V$ in the interconnected network, given that its fold-change is $f(g_i)$, when we are considering modules with respect to upregulated condition, we assign the gene's node score $s(g_i)$ as

$$s(g_i) = \begin{cases} f(g_i) & f(g_i) > \tau_f \\ \dfrac{f(g_i)}{C_f} & \text{otherwise} \end{cases} \tag{4.1}$$

where τ_f is the cut-off threshold and C_f is a positive constant. The formulation transform any fold-change score less than τ_f to smaller value, which consequently is given less priority

in the network.

In some cases, downregulated regions of the network are the focus of studies. When we are considering modules associated with downregulated conditions, $s(g_i)$ is given as

$$s(g_i) = \begin{cases} \dfrac{1}{f(g_i)} & \dfrac{1}{f(g_i)} > \tau_f \\ \dfrac{1}{C_f f(g_i)} & \text{otherwise} \end{cases} \tag{4.2}$$

Reactions $r \in V$ are scored based on the enzyme-coding genes associated to the reactions. If a reaction is only linked to a single gene, the node score of the gene will be assigned to the reaction. If more than one genes are associated to a reaction (i.e either a single enzyme catalyzed by multiple genes, or a reaction catalyzed by multiple enzymes), we assign the third quartile score of all the genes associated to the reaction. In Ref. [26], average score of genes has been used to assign score to reactions. As a matter of preference, however, we adapt to use third quartile score as we think by averaging we may unnecessarily lower the score down, especially for enzyme-coding genes that has 'OR' relationship which implies that not all genes has to be presence to encode an enzyme. When a reaction could not be mapped to any gene expression data, which could be due to unavailability of gene data or limitation of mapping in the metabolic model, the median score of all known reactions is assigned to the reaction.

We also allocate activity weight $(W_A)_{ij}$ to the edge connecting node i and j. To assign weight to edges in the interconnected network, we will consider the type of nodes that are linked by the edge. If the two nodes are either a reaction that are connected to a gene through interlayer edge, or two genes that are connected through gene regulatory intralayer edge, then the weight of the edge is the average score of these two nodes. If the edge is part of intralayer edges in the bipartite metabolic layer that connects a reaction and a metabolite, then the edge is assigned the score of the reaction linked by the edge. If node i and j is not connected, we set $(W_A)_{ij} = 0$. Our scheme of assigning edge's weight allows the strength of nodes to be

reflected by the edges in the network. Moreover, as metabolite nodes do not have any scores, we can regard metabolites as nodes that provides linkage between reactions reflected by the strength of their edges' weights.

4.2.3 Module Quality Function

The quality of modules in a network is measured by the expected description length per step of a random walker [51] given by:

$$L(M) = q_{out} H(Q) + \sum_{i=1}^{m} p_{in}^{i} H\left(P^{i}\right) \tag{4.3}$$

$H(Q)$ is the is the entropy of frequency-weighted expected length of codewords between modules, and $H\left(P^{i}\right)$ is the frequency-weighted expected length of codewords in module i. q_{out} is the probability of a random walker switching modules formulated as $q_{out} = \sum_{i}^{m} q_{out}^{i}$, where q_{out}^{i} is the per step probability of a random walker exiting module i. With p_{α} given as the ergodic visit frequency at node α, $P_{in}^{i} = \sum_{\alpha \in i} p_{\alpha} + q_{out}^{i}$ is the rate at which a random walker is at module i, adding to the probability it exit the module. $L(M)$ is derived based on Shannon's source coding theorem [53] that state the average length of code length for a random variable X, will not be less than its entropy. The entropy is denoted as $H(X) = -\sum_{1}^{n} p_i \log_2\left(p_i\right)$. The entropy of codewords between modules is formulated as

$$H(Q) = -\sum_{i=1}^{m} \frac{q_{out}^{i}}{\sum_{j=1}^{m} q_{out}^{j}} \log_2 \left(\frac{q_{out}^{i}}{\sum_{j=1}^{m} q_{out}^{j}} \right) \tag{4.4}$$

and the entropy of codewords within module i is formulated as

$$\begin{aligned} H\left(P^{i}\right) = &- \frac{q_{out}^{i}}{q_{out}^{i} + \sum_{\beta \in i} p_{\beta}} \log_2 \left(\frac{q_{out}^{i}}{q_{out}^{i} + \sum_{\beta \in i} p_{\beta}} \right) \\ &- \sum_{\alpha \in i} \frac{p_{\alpha}}{q_{out}^{i} + \sum_{\beta \in i} p_{\beta}} \log_2 \left(\frac{p_{\alpha}}{q_{out}^{i} + \sum_{\beta \in i} p_{\beta}} \right) \end{aligned} \tag{4.5}$$

To determined the ergodic visit frequency at node α for the undirected interconnected network, we consider a $N \times N$ 'overall' weighted matrix with element W_{ij} that corresponds to the weight of the edge of nodes i to j. W_{ij} is regarded as ''overall' weighted matrix as the value of its elements are taking into account of overall score of topological and activity strengths of nodes in the network, i.e. by deriving the scores from weighted values of normalized topology and activity adjacency matrices. The strength of node i is $w_i = \sum_j W_{ij} = \sum_j W_{ji}$, and total weight w is given by $w = \sum_i w_i$. Thus, we can set the steady state visit frequency of node α to correspond to the relative strength of it weight given as $p_\alpha = w_\alpha / w$ per recommendation in Ref. [118, 119].

We generalized that the random walker is able to transform into a *random surfer* through teleportation with probability τ. For node α in module i that is linked to node β outside the module, the normalized weight of the edge linking α and β is $\hat{w}_{\alpha\beta}$. The sum of normalized weight adjacent to node α is set to 1 (i.e $\sum_\beta \hat{w}_{\alpha\beta} = 1$). Given n_i as the number of nodes in module i, and n as the total number of nodes, every node is expected to teleport a proportion $\tau\,(n - n_i)\,/\,(n-1)$ and move of a proportion $(1-\tau)\sum_{\beta \notin i} \hat{w}_{\alpha\beta}$ of p_α to nodes outside of module i. By considering teleportation, the exit probability for module i is given as

$$q^i_{out} = \tau \frac{n - n_i}{n-1} \sum_{\alpha \in i} p_\alpha + (1-\tau) \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha \hat{w}_{\alpha\beta} \tag{4.6}$$

4.2.4 Discovering Active Modules of Interconnected Network

We propose an assumption that the activity of nodes in a bipartite layer is affected by two factors which are 1) its topologies in relation to other nodes in the bipartite layer and gene regulatory layer, and 2) the transcriptomic activities of enzyme-coding genes associated to the reactions in the bipartite layer. Therefore, to discover active modules of the interconnected network, we need to simultaneously consider the nodes' topologies and their transcription strengths.

For an interconnected network with N nodes, we consider a $N \times N$ topological adjacency matrix $(W_T)_{ij}$ where its element is given value 1 if there is an edge connecting node i and node j, or given 0 otherwise. $(W_T)_{ij}$ signifies whether there are connections between two nodes. Thus, if v_1 is connected to v_4, and v_2 is connected to v_3, then $(W_T)_{14} = (W_T)_{41} = 1$ and $(W_T)_{23} = (W_T)_{32} = 1$. The sum of weight of the topology adjacency matrix is set as $w_S^T = \sum_i (w_T)_i$, where $(w_T)_i = \sum_j (W_T)_{ij}$. Then, we set a 'normalized' topology adjacency matrix $\left(\widetilde{W}_T\right)_{ij} = (W_T)_{ij} / w_S^T$ where the total weight $\tilde{w}_S^T = \sum_i \sum_j \left(\widetilde{W}_T\right)_{ij} = 1$.

To take into account of the strength of nodes in relation to transcriptomic activities in the network, we created active weighted matrix $(W_A)_{ij}$ with elements that correspond to activity strengths. As described earlier in Section 4.2.2, the scores of two connected nodes (v_i and v_j) are translated into edge weight $(W_A)_{ij}$ that are derived from fold-change scores of the nodes (i.e. based on Equation 4.1 or Equation 4.2). Thus, $(W_A)_{ij}$ corresponds to the activity strength of an edge based on the two nodes that are connected by it. If v_i and v_j is not linked, we can assume that there are no strengths between the nodes, thus $(W_A)_{ij}$ is set to 0. We follow the same procedure to construct a 'normalized' active matrix $\left(\widetilde{W}_A\right)_{ij}$ as previously described for topology adjacency matrix. The sum of weight of the active matrix $w_S^A = \sum_i (w_A)_i$ with, $(w_A)_i = \sum_j (W_A)_{ij}$. Thus, the normalized active matrix is $\left(\widetilde{W}_A\right)_{ij} = (W_A)_{ij} / w_S^A$, with total weight $\tilde{w}_S^A = \sum_i \sum_j \left(\widetilde{W}_A\right)_{ij} = 1$.

To discover active modules in interconnected network, we define the overall weighted matrix W_{ij} as

$$W_{ij} = \frac{\kappa \left(\widetilde{W}_A\right)_{ij} + \eta \left(\widetilde{W}_T\right)_{ij}}{\kappa + \eta} \tag{4.7}$$

where κ and η represent the weight of activity and topology respectively.

Total weight of W_{ij} is given as $w = \sum_i w_i$ with $w_i = \sum_j W_{ij}$. The ergodic visit frequency of node i is assigned to the relative strength of its overall weight and is given as $p_i = w_i / w$. The value of normalized weight of the edge linking i and j, $\hat{w}_{ij}$ (i.e as described in Section 4.2.3) is simply the value of element of transition matrix $T_{ij} = W_{ij} / w_i$.

4.2.5 Interconnected Network Objective Function

Our objective is to discover modules of interconnected network that encompass nodes from each of the layer in the network. To achieve this objective, we set the algorithm to minimize $L(M)$ when the steady state of the random walk process are majorly made by individual module that flows through both layers of the interconnected networks.

We define the set of modules M of network G at any given time by partitioning it nodes V into its subsets, i.e. $M = \{m_1, m_2, m_3, \ldots\}$, $m_i \cap m_j = \varnothing$ $\left(\text{where } i \neq j\right)$ and $\bigcup_{m_i \in M} m_i = V$. Let the state before any merging of modules to be defined as state a, where the number of modules during this state is $k = |M_a|$. For the next step, we are going to merge two of the modules in state a into one new module. At this state, defined as state b, the number of modules will be reduced to $|M_b| = k - 1$. Given that we merge two of the modules in state a (i.e module i and j), the new joined module is $m_{ij} = m_i \cup m_j$ and the set of modules in state b is given by $M_b = M_a - m_i - m_j + \left(m_i \cup m_j\right)$. The reduction of expected description length for this two states is given by $\Delta L = L(M_a) - L(M_b)$. The number of metabolic nodes and regulatory nodes in the joined module are $n_{met}^{ij} = \left|m_{ij} \cap V_{met}\right|$ and $n_{reg}^{ij} = \left|m_{ij} \cap V_{reg}\right|$ respectively.

We define

$$\mathrm{S}\left(i, j\right) = \begin{cases} \Delta L \cdot \min\left(\dfrac{n_{met}^{ij}}{n_{reg}^{ij}}, \dfrac{n_{reg}^{ij}}{n_{met}^{ij}}\right) & n_{met}^{ij} \text{ or } n_{reg}^{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

To achieve our objective where modules should contains nodes from both layers, the objective function when we are joining any possible two modules is given by:

$$\text{maximize} \quad S\left(i, j\right) \tag{4.9}$$

We are looking for two candidate modules i and j that is to be merged, which lead to the

largest score of $S(i, j)$. If $S(i, j)$ is zero, it indicates that the joined module is only composed of nodes from only one layer. $S(i, j)$ can be regarded as normalized ΔL that takes into account of the topology of the interconnected network. By maximizing $S(i, j)$, we are looking for the joined module to have balanced proportion of high ΔL and high ratio of nodes from each layers. Given that we have identified the two candidate modules i and j to be merged, if $\Delta L_{ij} > 0$, then the merging will be finalized. In contrast to approach on unipartite graph, if $\Delta L_{ij} \leqslant 0$, the candidate modules will still be considered to be merged with certain probability as described in the next section. This is based on our view that it is preferable to have balance module with good score of $S(i, j)$ as it indicates good flow of information between layers in the interconnected network.

4.2.6 The Proposed ActiveFlow Algorithm

The proposed algorithm to discover multilayer modules adopts a greedy method that has been adapted from Ref. [38]. The proposed algorithm is shown in Algorithm 4.1. At the initialization phase, each nodes in the interconnected network is allocated to its own module. Thus, at this stage, the number of modules is equal to the number of nodes. Then we assign each module to a randomly ordered sequence. By iterating the sequence, for each module i, we identify its neighbour j that that give the largest score of $S(i, j)$ (refer to Equation 4.9). If $\Delta L > 0$ when module i and j are merged, then the merged is concluded. Otherwise, we merge module i and j by simulated annealing technique with an acceptance probability of $e^{(\Delta \mathrm{L})/T}$, where T is the simulated annealing temperature during the iteration. The module stays if the acceptance probability is not met. This process is repeated for each element in the randomly ordered sequence. Once all the elements in the sequence are iterated, we reduce T by multiplying it by a constant α. Then the procedure is repeated by creating a new random sequential order. We stop the algorithm when no module can be merged any further.

There are two reasons why the simulated annealing procedure is used in the algorithm.

The first reason is to give preference to modules with the largest $S(i, j)$ to be joined during the early stage of the iterations. The second reason is to allow the nodes for each layer to merged together into modules without stopping the algorithm prematurely earlier in the process. As the iteration progresses and T decreases, the algorithm prefers the states where $\Delta L > 0$.

4.2.7 Performance Score of Active Modules

Our objective is to discover active regions in the bipartite metabolic network that exhibit significant change of expressions. In the metabolic network, reaction scores corresponds to change of expressions of enzyme-coding genes. To evaluate the performance of a module $m \in M$ which that are being discovered by the algorithm, we use a module scoring metric S_A defined as follows:

$$S_A(m) = \frac{1}{|r^m|} \sum_i s\left(r_i^m\right) \tag{4.10}$$

where $s\left(r_i^m\right)$ is the score of ith reaction $r_i \in R_{met}$ in module m, and $|r^m|$ is the number of reactions in module m. S_A corresponds to the mean score of the reactions in the module. The score $S_A(m)$ is compared with randomly generated modules with the same number of reactions. We consider a module to be significant when the probability of the random modules to obtained at least the same score as the module is less than 0.01.

Algorithm 4.1 Proposed algorithm to find modules in interconnected network

Input: Multilayer network $G = (V, E)$ where $V = V_{met} \cup V_{reg}$. Nodes of metabolic layer are $V_{met} = R_{met} \cup M_{met}$ with reactions $r \in R_{met}$ and metabolites $m \in M_{met}$, and nodes of regulatory layer are genes $g \in V_{reg}$.

function LMRATIO(ΔL, n_{met}, n_{reg})
  if either n_{met} or n_{reg} is 0 **then**
    $S \leftarrow 0$
  else:
    $S \leftarrow \Delta L \cdot \min\left(n_{met}/n_{reg}, n_{reg}/n_{met}\right)$
  return S
function SANNEAL(ΔL, T)
  if ΔL > 0 **then**
    ΔL Accepted
  else:
    generate random number $\rho \in (0, 1]$
    if $\rho < e^{(\Delta L)/T}$ **then**
      ΔL Accepted
    else:
      ΔL Rejected
procedure UPDATE MODULES(G)
  Assign each node to its own module $m \in M$, where $M := \{v \in V | \{v\}\}$
  Initialize $L(M)$
  Initialize all modules to a random sequential order
  while True **do**
    for each module i in the sequential order **do**
      Find $L(M)_{ij}$ for network if m_i is merged with its neighbouring m_j
      $n_{met} \leftarrow$ number of metabolic nodes in m_{ij}
      $n_{reg} \leftarrow$ number of gene nodes in m_{ij}
      $\Delta L(M)_{ij} \leftarrow L(M) - L(M)_{ij}$
      $L(M)_{max} \leftarrow L(M)_{ij}$ for m_{ij} with maximum LMRatio $\left(\Delta L(M)_{ij}, n_{met}, n_{reg}\right)$
      ΔL $\leftarrow L(M) - L(M)_{max}$
      SANNEAL $(\Delta L, T)$
      if ΔL Accepted **then**
        Merge m_{ij} that corresponds to $L(M)_{max}$
        $L(M) \leftarrow L(M)_{max}$
    if No modules merge through the whole sequence **then**
      break
    else
      Randomly assign modules to a sequential order

4.3 Experiments and Results

4.3.1 Dataset and Preprocessing

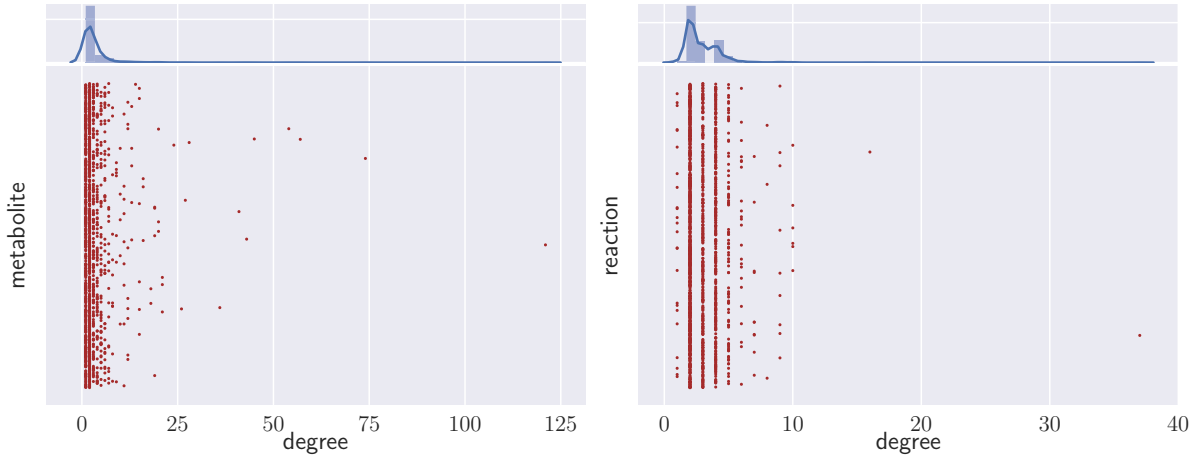
We are using the same metabolic model of yeast based on YEASTNET model [88] as previously described in Section 3.3.1. To reduce the potential of false-links in the module construction, we remove currency metabolites ATP, ADP, NADP, NAD, NADPH, NADH, hydrogen phosphate, diphosphate, water, carbon dioxide, oxygen, ammonia and proton as categorized in Ref. [89, 90]. The bipartite metabolic network is consisted of 3008 nodes (1377 are metabolites and 1631 reaction nodes) and 4421 edges.

The second component of the interconnected network is a gene regulatory network for diauxic shift of *Saccharomyces cerevisiae*. The network is derived from gene regulatory network curated and provided by the resources in Ref. [32]. It is consisted of 267 genes (127 transcription factors and 140 target genes) and 1242 edges. 122 of the genes has been identified as enzyme-coding genes of the reaction in the the metabolic network.

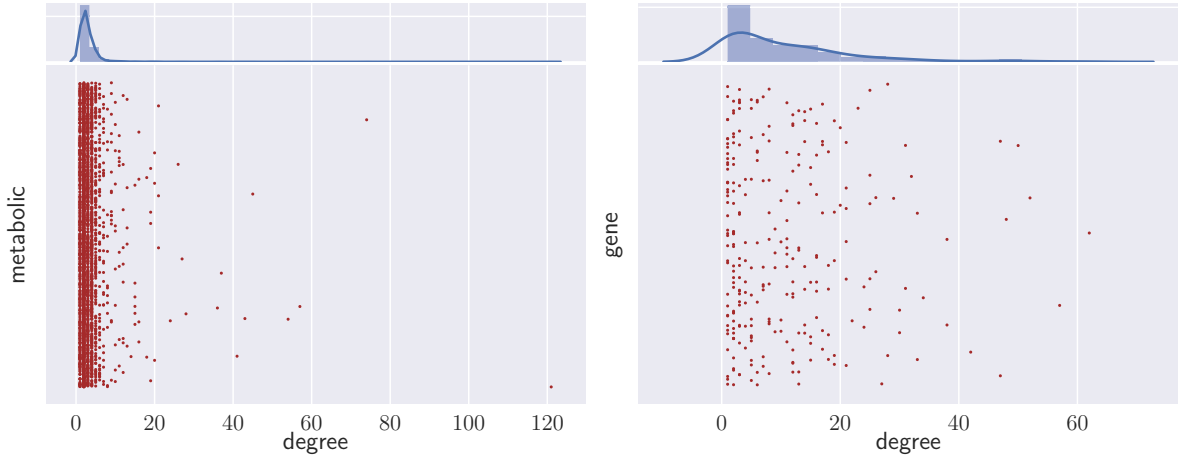
By connecting the metabolic network to the gene regulatory network, we obtained an interconnected network with 3275 nodes and 6017 edges. There are 354 interlayer edges that link reactions in the bipartite metabolic network to enzyme-coding genes in the regulatory network. Figure 4.2 shows the degree distribution of the nodes in yeast networks. The median degrees for metabolites, reactions and genes are 2, 2 and 6 respectively. The figure shows that the degree distribution of varies by the type of nodes, with the metabolites concentrated to below 20 and reactions to 10 degree and below in the metabolic layer. The degree of genes are mainly scattered by 35 and below in the gene layer.

4.3.2 Discovering Topological Modules of Interconnected Network

For the first analysis, we produced a comparison between Infomap and our algorithm in discovering topological modules in the network. Infomap is a well known information flow



(a) Degree distributions for nodes in metabolic layer of yeast networks



(b) Degree distribution for nodes in interconnected metabolic-gene layers of yeast networks

Figure 4.2: Degree distributions of nodes in the yeast networks are represented by jitter plots. **(a)** Degree distributions of metabolite and reaction nodes in metabolic layer. Metabolite has median score of 2 and are mainly scattered with degree of 20 and below, while reaction has median of 2 and are mainly with degree of 10 and below. **(b)** Degree distribution of nodes in interconnected network comprised of nodes in bipartite metabolic layer (i.e metabolite plus reaction) and nodes from gene regulatory layer. Metabolites and reactions has a joint median score of 2 and are concentrated on 20 degree or below, while genes has a median of 6 and are mainly scattered by degree of 35 and below.

algorithm that has been implemented successfully on unipartite network. Our algorithm is based on Infomap but has been extended to deal with interconnected network composed of a bipartite network and a gene regulatory network. The activity weight κ is set to 0 and the

topological weight η is set to 1. This setting means that only the degree of nodes are taken into account in evaluating the modules (i.e normalized weight of edge $\tilde{w}_{ij}$ corresponds to the element in the transition matrix of the network). The topological modules that has been obtained corresponds to densely connected nodes that interact at greater frequency than nodes outside of the module.

Table 4.1: Summary of comparison between ActiveFlow and Infomap on the Yeast Interconnected Network

Algorithm	ActiveFlow	Infomap
No. of nodes in largest module	548	215
No. of metabolic nodes in largest module	511	8
No. of genes in largest module	37	207
Mean no. of nodes in modules	144.75	45.70
Mean no. of metabolic nodes in modules	131.45	34.35
Mean no. of genes nodes in modules	13.30	11.35
Mean ratio of metabolic and genes nodes	0.1046	0.0338

Note: Calculation of mean are made on the top 20 modules ranked by their sizes. Analysis is based only on network topology where $\kappa = 0$ and $\eta = 1$.

The results that compares both algorithms by their top 20 modules are tabulated in Table 4.1. Infomap generally found modules with majority of their nodes are from only one layer. For example, the top module found by Infomap contained mainly of genes with were 207 nodes, and only 7 reactions and one metabolite. The first module consumed 75.53% of the total gene in the regulatory layer. The subsequent modules are largely made up of of metabolic nodes with only a small number of genes. For example, the second module found by Infomap was consisted of 77 nodes from metabolic layer but did not include any gene, while the third module had 74 nodes with only one gene. From initial value of average description length per step $L(M) = 13.0246$, Infomap can minimize the average description length to smaller value with $L(M) = 6.9124$, while ActiveFlow achieved $L(M) = 9.0440$. However, Infomap more likely try to conserve its modules from nodes of the same layer. This scenario is further indicated by the results where Infomap obtained small value 0.0338 of mean ratio of metabolic and

gene nodes. By comparison, ActiveFlow obtained larger score, which was 0.1046 for the mean of ratios between metabolic and gene nodes. The first, second and third modules of ActiveFlow contained 511 metabolic nodes and 37 genes, 415 metabolites and 14 genes, and 361 metabolic nodes and 16 genes respectively. This result shows that ActiveFlow produced modules which are more consistently made up of nodes from both layers in the interconnected network. Through our proposed algorithm, the modules that were discovered has been taking into account of the interactions between different layers of networks in the interconnected network.

4.3.3 Discovering Active Modules of Interconnected Network

In this analysis, We evaluated the usefulness of ActiveFlow by evaluating regions in the yeast network that experienced high degree of upregulation during the diauxic shift period. The results is compared to AMBIENT for validation purposes. AMBIENT only uses one metabolic bipartite network input and the expression score embedded to the nodes is in the form of log fold-change. Thus, positive and negative scores indicate upregulation and downregulation respectively. AMBIENT was run by the parameters and reactions scores data as provided in Ref. [26] and we presented the results of of the best run that has the best F-scores for pathways that has described in Ref. [1]. ActiveFlow was run on the interconnected network with activity weight κ set to 5 and topological weight η set to 1. As we are looking for active modules that that indicates changes of in gene expression between different conditions, higher weight was given to the activity scores ($\kappa = 5$) while still taking topological factor into consideration ($\eta = 1$). Equation 4.1 was used by setting cut-off threshold $\tau_f = 2$ (i.e. chosen based experimental recommendation in Ref. [92]) and $C_f = 20$. The intuition behind this setting is that the high value of C_f will transform fold-change score below τ_f to very low value, and will give more preference for a random walker to go to nodes with node scores that are greater than the cut-off threshold.

According to Rosvall and Bergstrom [51], the setting for the value of teleportation probability τ is robust to selection. In their analysis, Rosvall and Bergstrom used $\tau = 0.15$. We conducted different settings of τ and the effects towards expected description length per step $L(M)$ and the number of modules. The results is shown in Figure 4.3. It can be seen that the effect of varying τ is rather inconclusive. In producing comparison with AMBIENT, we use $\tau = 0.05$ as it achieved the smallest number of modules as compared to other settings.

Figure 4.3: Tabulations of expected description length of single step $L(M)$ and the number of modules obtained by varying teleportation probability (τ).

We conduct over-representation analysis for compounds in the modules based on the pathways in Yeast Metabolome Database (YMDB) [93] by using MBROLE [94]. AMBIENT obtained 3 modules with significant over-representation pathways. Two of these modules are relevant to the curated pathways in Ref. [1]. As for ActiveFlow, we obtained 16 modules that contained multiple nodes (with the smallest module containing 17 nodes). The number of

modules is higher than AMBIENT as ActiveFlow is based on the concept of partitioning while AMBIENT only retained high scoring modules. There are also 316 one-node modules which are mainly low-scoring reactions and metabolites obtained by ActiveFlow. These modules are excluded from analysis for module's significance. To evaluate module's significance, we employed S_A in Equation 4.10 and obtained five modules with p-value < 0.01. Two of these significant modules contains significant over-representation pathways that corresponds to findings in Ref. [1]. Therefore, the comparisons between AMBIENT and ActiveFlow was made on their two relevant modules. The module is name with prefix 'M' that denotes a module, which is followed by module number. For AMBIENT, the active module M1 is consisted of 89 reactions and 29 metabolites, and active module M2 is consisted of 19 reactions and 6 metabolites. For ActiveFlow, active module M1 is consisted of 35 reactions and 28 metabolites while M2 contains 23 reactions and 18 metabolites. One notable different between AMBIENT and ActiveFlow is that AMBIENT has smaller percentage of metabolites in the modules. This is because metabolites with high degree of connections are penalized by AMBIENT, while ActiveFlow does not penalized any inclusion of metabolites as there are regards as precursor nodes to connect reactions in the network.

Table 4.2 summarizes the results obtained from AMBIENT and ActiveFlow. The table shows that ActiveFlow achieved better results than AMBIENT for two of the three metabolic pathways under evaluation. ActiveFlow obtained better pathway p-values and F-scores for TCA cycle and starch and sucrose metabolism in module M1 and M2 respectively, while AMBIENT obtained better score for Glyoxylate cycle in module M2. AMBIENT and ActiveFlow do fully capture all the nodes that are involved in the upregulated condition of diauxic shift. However, the result shows that ActiveFlow is comparably useful as AMBIENT in discovering active modules that relates to the activities in the network.

In relation to TCA cycle, three important enzymes are identified in module M1 of Active-Flow. The first enzyme is aconitase (reaction score 2.86) that catalyzes citrate to isocitrate

Table 4.2: Top pathways of the yeast active modules obtained by ActiveFlow.

	ActiveFlow	AMBIENT
Citrate cycle (TCA cycle)		
Module	M1	M1
P-value	6.91E-07	4.77E-06
Precision	0.3500	0.2692
Recall	0.500	0.500
F-Score	0.4118	0.3500
Glyoxylate cycle		
Module	M1	M1
P-value	1.63E-04	4.63E-05
Precision	0.3000	0.2692
Recall	0.3000	0.3500
F-Score	0.3000	0.3043
Starch and sucrose metabolism		
Module	M2	M2
P-value	5.27E-06	2.14E-06
Precision	0.5556	1.0000
Recall	0.2778	0.2222
F-Score	0.3704	0.3636

Note: Curated metabolic pathways by expert in [1] are cross-checked with the modules that are obtained by ActiveFlow and AMBIENT. 'M' denotes a module and followed by module number. The ground truth to calculate precision, recall and F-score is quantified from metabolite background list of the pathways [93].

through cis-aconitate. The next process is the oxidative decarboxylation of isocitrate to produce 2-Oxo-glutarate, which is catalyzed by other important enzymes, which are isocitrate dehydrogenase (NADP and NADP+) (reaction score 10 and 4.55). The third enzyme that we identified is succinate dehydrogenase (reaction score 6.25) which are responsible to catalyze the oxidation of succinate to fumarate during the 6th step of TCA cycle. Module M1 also includes the highly upregulated phosphoenolpyruvate carboxykinase (reaction score 14.49). This is the key enzyme that undergoes metabolic programming during diauxic shift by reversing the flow of metabolites in glycolytic pathway towards glucose-6-phosphate in the starch and sucrose metabolism pathway. In module M2 of ActiveFlow, we identified three enzymes that promotes channelling of glucose-6-phosphate into the starch and sucrose pathway through alpha,alpha-trehalose-phosphate synthase (reaction score 3.85),

alpha,alpha-trehalase (reaction score 4.17) and hexokinase (reaction score 5.88). M2 also indicate two enzymes for branching and debranching of glycogen which are 1,4-alpha-glucan branching enzyme (reaction score 7.14) and glucan 1,4-alpha-glucosidase (reaction score 6.25).

4.4 Lower Grade Glioma Module Identification

In this section, we implemented ActiveFlow to identify topological and disease modules of a subtype of lower grade glioma patients categorized as having short survival range with median of 1.4 years. This subgroup of glioma patients has lower survival that the average survival of LGG which is approximately 7 year [29].

4.4.1 Dataset and Preprocessing

The transcriptomic data used to generate the bipartite metabolic network and gene regulatory network of the LGG patients is based on the work by Liu et al. [31]. Liu et al. applied unbiased consensus clustering analysis on 97 samples of gene expression data that are classified as WHO grade II lower grade glioma patients (either astrocytoma, oligodendroglioma or mixed) cohort in Rembrandt database [30]. The consensus clustering approach is a method that finds concensus across multiple resampling of clustering algorithm by assessing the stability of the observed clusters [120]. By applying the method in examining the stability of the clusters, the samples was discovered to optimally group in three clusters. This implies that there are likely three different molecular subtypes in the LGG patient dataset. Then, Liu et al. performed silhouette analysis [121] on the data to measure how close the samples are to its their clusters (tightness) in comparison to other clusters (separation). From the analysis, 72 of the 97 samples displayed tight associations with their clusters (i.e. by thresholding with positive silhouette values). These 72 samples were retained to represent 'core patients' for further analysis. The characteristics of the three clusters that comprises the core patients (i.e

Table 4.3: Characteristics of the three clusters of lower grade glioma core patients identified from the Rembrandt database.

		Cluster 1	Cluster 2	Cluster 3	Total
	No. of samples	20	31	21	72
	Percentage	27.7%	43.0%	29.3%	100%
Tumour	Astrocytoma	15	17	15	47
	Oligodendroglioma	5	11	6	22
	Mixed	0	3	0	3
Gender	Male	9	13	9	31
	Female	4	13	9	26
	Not Available	7	5	3	15
Survival Data	Available	16	21	11	48
	Not Available	4	10	10	24

in term of number of samples, the type of glioma and gender of patients) is shown in Table 4.3.

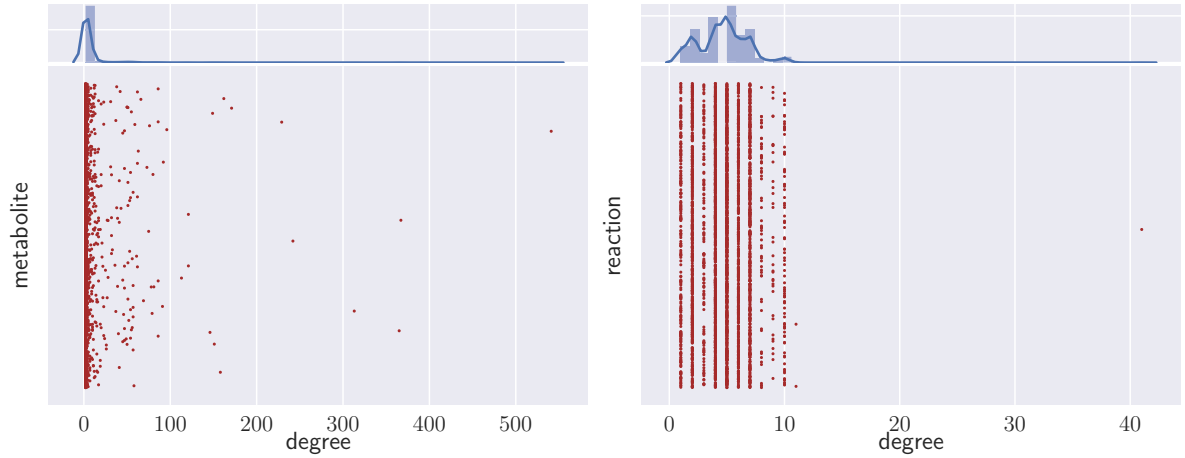
Next, Liu et al. performed survival analysis on these clusters by using 48 of the 72 core patients that are annotated with survival data. This was implemented by using Kaplan-Meier analysis, a technique that measures the fraction of patients that survive after a certain time after treatment after taking into account incomplete information (i.e. due to patients withdraw from studies, or not experiencing death before the end of the study)[122, 123]. A Kaplan-Meier analysis on these cohort of 48 patients indicated a median survival of 6.2 years. By performing Kaplan-Meier analysis on these clusters, Liu et al. found that the patients in cluster 1 have a significantly shorter survival length, with a median of 1.4 years. It is significantly shorter than the patients in cluster 2 and 3 ($p < 0.0001$). The survival of cluster 1 patients is comparably similar to that of glioblastoma multiforme (GBM) patients. The overall survival lengths of samples in cluster 2 or 3 are not significantly ($p = 0.618$). Through over-representation analysis, the LGG short survival subtype was found to exhibit a 'glioblastoma-like' gene profile.

Based on the work by Liu et al. in [31], we further categorize the LGG patients into 'short survival' subtype (20 samples) and 'long survival' subtype (52 samples), derived from patients of cluster 1 and the remaining clusters (i.e. cluster 2 and 3 grouped together) respectively. To

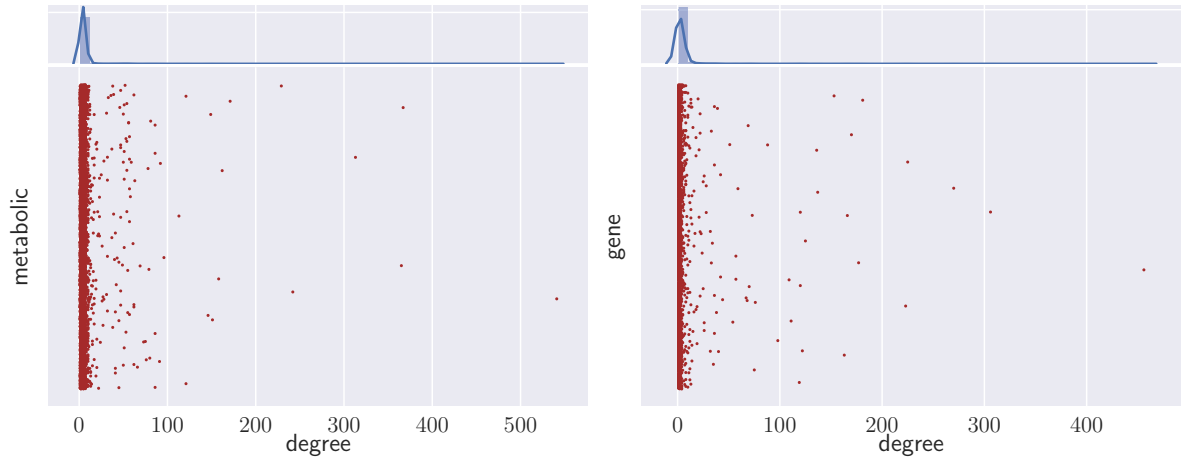
study the biological mechanisms that differentiates the short survival cohort, we use their transcriptomic data based on samples of short survival subtype as input in constructing bipartite metabolic network and gene regulatory network that corresponds to this cohort. The fold-change scores of the nodes in the metabolic and gene regulatory networks of short survival cohort are calculated by taking long survival subtype and the short survival subtype as the control condition and test condition respectively.

The tissue specific bipartite metabolic network for short survival patients is constructed by utilising the mCADRE algorithm [124] on Recon 2 Human generic metabolic model [125]. It is consisted of 3458 reactions, 2352 metabolites, and 14898 edges. To construct gene regulatory network, we use the human regulatory network curated from ENCODE data [126] based on gene-proximal binding. It contains 9057 genes, 26070 edges and 119 transcription factors. The MAS5 detection call binarization [127] was used to approximate the presence of genes in short survival samples. High confidence genes are defined as those that presence in at least 90% of the samples. We construct the LGG short survival gene regulatory network by filtering the ENCODE human regulatory network to only constitute the high confidence genes. The LGG short survival regulatory network consists of 2349 genes (of which 56 are transcription factors) and 4630 edges.

The interconnected network of short survival subtype which are constructed by interconnecting the metabolic and gene regulatory networks contains 8158 nodes and 20742 edges. There are 1214 interlayer edges that connects the metabolic network to the gene regulatory network. Figure 4.4 shows the degree distribution of the node in the networks. The degree distribution of reactions is concentrated below 10 with median of 5. Metabolite has median of 2 and concentrated below 100. The genes in the regulatory network has a median of 1, denoting a large number of target genes. However, transcription factors notably shows higher degree that goes up to more than 100 connections.



(a) Degree distributions for nodes in metabolic layer of LGG networks



(b) Degree distribution for nodes in interconnected metabolic-gene layers of LGG networks

Figure 4.4: Degree distributions of nodes in the networks of lower grade glioma short survival subtype are represented by jitter plots. **(a)** Degree distribution of metabolite and reaction nodes in metabolic layer, where metabolites and reactions have median of degree of 2 and 5 respectively. Majority of metabolites has degrees that are scattered below 100, while reactions are mainly concentrated below 10. **(b)** Degree distribution of nodes in interconnected network comprised of nodes in bipartite metabolic layer (i.e metabolite plus reaction) and nodes from gene regulatory layer. The median degree for metabolic nodes is 4, while genes has a median of 1. A majority of metabolic and gene nodes are distributed to 100 degree or below.

4.4.2 Topological Modules Identification

First, we investigate the topological modules in the network that only takes into account of degree distribution of nodes in the interconnected network. Topological module are locally

dense neighbourhood where the nodes within the module interact at greater frequency than the nodes outside of the module. To take into account on only the topological factor, ActiveFlow was run by setting topological weight $\eta = 1$, and the activities factor is ignored by setting $\kappa = 0$. As we are dealing with a considerably large network of 8158 nodes, we obtained large number of modules that amounted to 27. We also disregarded 141 single-node modules that were not informative for the analysis.

The largest modules in term of number of metabolic nodes contains 581 reaction, 207 metabolites from the metabolic network, and supported by 84 genes from the regulatory network. We conducted over-representation analysis on this module to evaluate flow of compounds that are mostly affected in the network. The results is tabulated in Table 4.4. The module is largely related amino acid and neurotransmitters transporters. The most significant pathway is 'Amino acid and oligopeptide SLC transporters'. This is an important pathway that functions to distribute amino acid across plasma membranes for the synthesis of proteins amino acid based chemicals such as neurotransmitters [128]. Y+LAT2 utilized transport, the reaction that is encoded by SLC7A6, that mediates the uptake of arginine, leucine and glutamine is heavily present in the module. Leucine which is an essential amino acid and one of the major substrates of Y+LAT2, has been known as a regulator of mTROC1 (mammalian target of rapamycin complex 1) that functions to control protein translation [129] and is involved in some of cancer hallmarks [130]. 'Central carbon metabolism in cancer' pathway is also amongst the top pathway of the module (F-score 0.3922). This indicates high concentrations of compounds that are involved in the regulation of cancer in the module.

The indication of upregulation of BCAT1 (reaction score 3.1665) in the module, the enzyme that initiates the catabolism of branched-chain amino acids (BCAAs) such as isoleucine, leucine and valine through the release of glutamate illustrates a glioblastoma-related regulations. Glioblastoma cells has been reported to excrete high concentration of glutamate that leads to neuronal death through exitotoxic mechanism [132]. BCAT1 was found to be to

Table 4.4: Top pathways of the largest module obtained by ActiveFlow based on topological information.

pathway name	q-value	Precision	Recall	F-score
Amino acid and oligopeptide SLC transporters	8.76E-23	0.3333	0.5366	0.4112
Amino acid transport across the plasma membrane	3.11E-22	0.2879	0.6786	0.4043
Central carbon metabolism in cancer	5.56E-21	0.3030	0.5556	0.3922
Na+/Cl- dependent neurotransmitter transporters	2.41E-20	0.2727	0.6429	0.3830

Note: Over-representation analysis was conducted based on HMDB IDs of known metabolites by using ImPALA [131]. The ground truth to calculate precision, recall and F-score is quantified from metabolite background list of the pathways.

highly overexpressed in glioblastoma tissues. This feature is highly specific to glioblastoma, and its suppression could block the excretion of glutamate which led to reduced invasiveness and proliferation in vitro, and significant reduction in glioblastoma xenograft tumours [133]. The overexpression of BCAT1 could be a feature of LGG short survival cohort that makes it comparably similar to glioblastoma in term of survival, and serve as a potential target for the development of cancer therapy.

4.4.3 Active Modules Identification

. We implemented ActiveFlow on the LGG short survival model to find metabolic subnetworks that undergoes significant upregulation of expression, and supported by regulation in the regulatory network. We use the same settings as in the previous yeast network analysis, where activity weight $\kappa = 5$ topological weight $\eta = 1$. This setting allow activity strengths to be given high consideration by our algorithm, while also taking into account of edges' densities. With reference to Equation 4.1, we set the cut-off threshold $\tau_f = 0$ and $C_f = 20$. As the C_f is set to a high value, any node with down-regulated fold-change score (less than zero) will be transformed to very low value. This allows for non-downregulated nodes (with fold-change score of more than zero) to be given preference by the random walker during the algorithm searching procedure. Comparisons were made with AMBIENT based on pathway over-representation analysis. AMBIENT was run on its default parameters on the

Table 4.5: Modules of LGG short survival subtype identified by ActiveFlow

Module	No. of metabolic nodes	No. of regulatory Nodes	Module Score, S_A
U1	398	9	2.6400
U2	370	12	1.4629
U3	249	273	1.1919

Note: Module significance which is based on module score is set at p-value < 0.01. Modules are ranked based on S_A.

LGG bipartite metabolic network.

ActiveFlow discovered three significant modules with p-value < 0.01 (based on module score S_A as calculated in Equation 4.10). The modules is shown in Table 4.5. Module is named with prefix 'U' that denotes upregulated module, and are followed by module number. Module U1 that contains 206 reactions, 192 metabolites and 9 genes, obtained the highest module score S_A = 2.6400. U3 has the largest support from the regulatory layer with 273 genes, as compared to U1 and U2 with 9 and 12 genes respectively. The top pathway based on over-representation analysis is shown in Table 4.6. The top pathways of U1 and U3 give a strong suggestion that 'Asparagine N-linked glycosylation' pathway is highly differentiated for the LGG short survival cohort. 'Asparagine N-linked glycosylation' is the top-tier pathway of 'N-glycan antennae elongation in the medial/ trans-Golgi' and 'Synthesis of substrates in N-glycan biosynthesis' in the pathway tree.

Table 4.6: Top over-representation pathways in modules of LGG short survival subtype identified by ActiveFlow

Module	Pathway Name	Precision	Recall	F-score
U1	N-glycan antennae elongation in the medial/ trans-Golgi	0.3637	0.800	0.5000
U2	HS-GAG degradation	0.3000	0.5455	0.3871
U3	Synthesis of substrates in N-glycan biosynthesis	0.1250	0.1923	0.1515

Note: Over-representation analysis was conducted based on HMDB IDs of known metabolites by using IMPaLA [131]. The ground truth to calculate F-score is quantified from metabolite background list of the pathways.

For the next analysis, we compare the pathways of highest scoring module of AMBIENT to the highest scoring module of ActiveFlow (i.e module U1). The result is shown in Table 4.7. The union between the top pathways of AMBIENT and ActiveFlow resulted in 17 pathways. ActiveFlow obtained better F-score in 64.71% or 11 of the 17 pathways.

Table 4.7: Top Reactome pathways obtained by over-representation analysis for the top module of AMBIENT in comparison to the top module of ActiveFlow (U1) for LGG short survival subtype.

pathway name	AMBIENT		ActiveFlow	
	q-value	F-score	q-value	F-score
Glycosaminoglycan metabolism(**)	1.44E-18	0.3423	2.71E-13	0.4231
Metabolism of carbohydrates(*)	1.44E-18	0.3277	2.68E-13	0.2543
Post-translational protein modification(*)	5.42E-18	0.3253	6.20E-13	0.2617
Metabolism of proteins(*)	5.42E-18	0.3152	4.10E-13	0.2400
Asparagine N-linked glycosylation(*)	1.87E-15	0.3077	2.26E-17	0.3096
Heparan sulfate/heparin (HS-GAG) metabolism(*)	1.92E-15	0.2913	1.50E-11	0.4091
Keratan sulfate/keratin metabolism(**)	1.14E-18	0.2581	3.32E-12	0.4706
Synthesis of substrates in N-glycan biosynthesis(*)	5.23E-11	0.2556	4.76E-08	0.2432
Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein(**)	1.24E-10	0.2500	7.58E-08	0.2338
Transport of vitamins, nucleosides, and related molecules(*)	2.05E-08	0.2174	2.29E-05	0.2025
N-glycan antennae elongation in the medial/trans-Golgi(***)	2.73E-07	0.1538	6.11E-13	0.5000
Transport to the Golgi and subsequent modification(***)	7.65E-07	0.1649	6.34E-13	0.4737
O-linked glycosylation of mucins(***)	1.36E-12	0.1798	6.88E-12	0.4667
Reactions specific to the complex N-glycan synthesis pathway(***)	4.29E-06	0.1333	2.42E-11	0.4516
Keratan sulfate biosynthesis(***)	1.29E-11	0.2000	2.42E-11	0.4516
Transport of nucleotide sugars(***)	1.05E-07	0.1702	9.67E-10	0.400
Pre-NOTCH Expression and Processing Transport of nucleotide sugars(***)	1.74E-06	0.1348	1.97E-09	0.400

Note: The indicator denotes that the pathway is within top 10 pathways in (*) both AMBIENT and ActiveFlow, (**) only in AMBIENT, (***) only in ActiveFlow. The top AMBIENT upregulated module which is consisted of 845 metabolic nodes (622 reactions and 223 metabolites). We compare this top module with module by ActiveFlow that has comparable top pathways. The ActiveFlow module is consisted of 398 metabolic nodes (206 reactions and 192 metabolites). Over-representation analysis was conducted based on HMDB IDs of known metabolites by using IMPaLA [131]. The ground truth to calculate F-score is quantified from metabolite background list of the pathway.

With reference to module U1, the results shows that the significantly upregulated region in the LGG short survival subtype model is related to an increase in activities in the elongation of N-glycan and the biosynthesis of keratan sulfate in the Golgi compartment. We identify mannosyl-oligosaccharide 1,2-alpha-mannosidase, a subtype of α-1,2 mannosidases which are the key enzymes in N-glycosylation, to be highly upregulated (reaction score 16.5950). Abnormal regulation for α-1,2 mannosidases has been linked to tumours based on the findings by Tu et al. [134]. It is suggested that upregulation of the enzymes variant, Golgi α-1,2 mannosidase IA (MAN1A1) can lead to initiation of metastasis while Golgi α-1,2 mannosidase IC (MAN1C1) can function as tumour suppressor in hepatocarcinogenesis. Furthermore, in relation to the keratan sulfate biosynthesis pathway, enzyme N-acetylglucosamine 6-O-sulfotransferase is highly expressed in our samples (reaction score 8.7080) and it has been found to be strongly expressed in glioma cells and many other tumours [135]. Enzyme beta-N-acetylglucosaminylglycopeptide beta-1,4-galactosyltransferase (reaction score 2.4916) is also identified as a potential target for therapy of LGG short survival subtype. B4GALT5 that encodes the enzyme has been found to be positive growth regulator for glioma and suggested as a target for glioma therapy [136].

The 'HS-GAG degradation' pathway, which corresponds to the top pathway of U2, is an important pathway to look at. Heparan-sulphate glycosaminoglycans (HS-GAGs) are subtypes of glycosaminoglycans, which are polysaccharides consisting of repeating disaccharide unit of uronic acid attached to a glucosamine. HS-GAG refers to both heparan sulfate and heparin. Glycosaminoglycans are part of the main macromolecules that regulate changes in cells properties and functions, either by interacting with cell receptors, or with signalling molecules such as growth factors [as reviewed by 137]. HS-GAGs has been linked to the of initial phase of oncogenic transformation of cells from normal to tumours [as cited by 138].

Expression of HS-GAGs degrading enzymes (e.g. heparanase) correlates with tumour invasion and metastasis [139]. Heparanase is a endo-acting β-glucuronidase enzyme that syn-

ergistically interacts with vascular endothelial growth factor (VEGF) in modulating melanoma progression through MEK/ERK signalling pathway [140]. The expression of VEGF itself highly correlates with many main brain cancers, especially malignant gliomas [141]. In our module, exo-acting enzyme β-glucuronidase (GUSB) that degrades heparan sulphate is indicated to be upregulated. Expression of GUSB has been found to be significantly increase in GBM samples [142]. GUSB expression level also increased in pancreatic tumour samples and has been suggested as therapeutic target for the disease [143].

4.5 Discussion

Through the topological module analysis of the yeast metabolic model, we make a comparison between the modules discovered by ActiveFlow and Infomap. Infomap is the information flow method from which ActiveFlow is derived from. Overall, although Infomap achieve lower score of minimal $L(M)$ as compared to ActiveFlow, many of its modules contains nodes that are mainly from only one layer. This indicates there are low level of information flow between the two-layer in the modules achieved by Infomap. This is not a surprise as Infomap is dedicated for unipartite network is not well adapted for interconnected network made of two different type of networks. We introduce ActiveFlow as an implementation of information flow method on two-layer interconnected network. ActiveFlow discovered modules that are comprised of nodes from both of the metabolic and gene regulatory layer, indicating two-way communications between the layers. The result coincides with our assumption that metabolism is a two-way that is controlled by regulatory mechanism. However, as described earlier, $L(M)$ achieved by ActiveFlow is higher than Infomap. The higher value of $L(M)$ could be due to the information that is flowing between two layers that have different frequencies distributions. It is therefore acceptable to have higher score of minimal $L(M)$ as the system is not a single network but consisted of two-layer network.

We also observed the occurrence of single-size modules in the modules discovered by

ActiveFlow. Majority of these modules are consisted of either reactions that does not have any interlayer links, or metabolites that are connected to reactions with no interlayer links. As in the metabolic models of either yeast or LGG, there are reactions that are not annotated with enzyme-coding genes. There are also reactions that are annotated but its enzyme-coding genes do not exist in the gene regulatory layer. For these two cases, there will not be any interlayer connecting the reactions to the regulatory layer. As our algorithm rely on interlayer connected in the system to produce flow between the two layer, it needs to be noted that the lack of information in the metabolic model and the regulatory model may cause some limitation in module inference that could result in single-size module. On the other hand, the case could present an advantage in an adequately mapped model, where the reactions without any link to the regulatory network may be given less preference and filtered out from the model inference as they do not present evidence of gene regulation. We considered results that we achieved for the yeast and LGG models are satisfactory by making comparison with pathways curated with experts and the AMBIENT algorithm.

There is one another aspect that differentiate ActiveFlow from other unipartite methods. The modules obtained by ActiveFlow needs only to be connected. As we do not impose any restriction on the connectedness of modules within a layer, we can obtain metabolic modules which are disconnected within the metabolic layer. This can happen when the connection of the components of metabolic layer are provided by the components of module in the gene regulatory layer.

ActiveFlow is a heuristic algorithm that does not ensure globally optimal solutions. However, ActiveFlow shows promising results which is comparably on par to the current implementation of active module on metabolic bipartite network. As a computational inference method, ActiveFlow discovered the three relevant significant yeast metabolic pathways during diauxic shift as indicated by experts. ActiveFlow achieved better F-score results for two of the three pathways under evaluation as compared to the AMBIENT. Both AMBIENT and

ActiveFlow implementations in the analysis are coded in Python language. ActiveFlow implementation evaluate the yeast model with a faster running time of about 5 minutes on an Intel i7 1.8GHz processor, as compared to about 45 minutes for AMBIENT. By producing comparably good results as current implementation of active module identification algorithm, ActiveFlow is proven useful in getting a fast and good analysis on metabolic model.

4.6 Conclusion

In the quest to identify underlying mechanisms that governs the significant changes of metabolism, we address the need to consider the metabolic process as a part of global system consisting of interrelations between metabolic pathways and gene regulatory mechanisms. We propose a novel active module identification approach, named as ActiveFlow algorithm, that models metabolism as an interconnected network consisting of a bipartite metabolic network and a gene regulatory network, in which the interaction between these networks signify the regulation of metabolism.

The ActiveFlow algorithm obtains multilayer modules that encompass mainly of nodes from the region in the metabolic and gene regulatory layers. Active modules that cover nodes of both regions implies that the significant changes in the metabolic regions not only dictated by the conditions of all its elements, but also by the region in the regulatory layer there are connected to. We are able to demonstrate the usefulness of ActiveFlow by validating the results of the algorithm with a current implementation of active module on bipartite metabolic network. ActiveFlow demonstrates its usefulness by being able to identify modules with reasonably relevance nodes that indicate metabolic reprogramming during diauxic shift for yeast and with a faster execution time.

We also set the objective to understand the disease mechanism of LGG short survival that exhibits 'glioblastoma-like' gene profile, in order to identify potential target enzymes that serve as molecular targets for cancer therapy. By implementing ActiveFlow on LGG short

survival cohort, we identify BCAT1 as topologically significant enzyme-coding gene which is also highly expressed by the cohort. As overexpression of BCAT1 has been characterized as highly specific feature in glioblastomas, its topological and transcriptional significance as suggested through the module identification implies the needs for further investigation. By evaluating the highly upregulated region in the LGG model, we also identify 'Asparagine N-linked glycosylation' pathway as a highly differentiated metabolic mechanism from which we identify Golgi α-1,2 mannosidase IA (MAN1A1) as another potential target gene for the LGG subtype.

CHAPTER 5

Active Modules of Bipartite Metabolic Network by Joint Module Approach

In this chapter, we propose an active module identification algorithm for interconnected biological network that are composed of bipartite metabolic network and gene regulatory network. We name the proposed algorithm as RACEMIC. RACEMIC is an approach that seeks to find regions in the multilayer interconnected network that are denoted by strong metabolism activities and are supported by high degree of gene regulatory activities. RACEMIC treats each layer in the biological system separately, by searching the space and computing the score of activities in each layer, and later combining them into joint-score that define the whole active system. In Section 5.1, we describe the issues that motivates us to pursue this problem. Then, Section 5.2 describe the framework and the schemes that we use in defining our algorithm. Section 5.3 shows the experimental results and performance evaluation of the proposed algorithm by validating the results with other algorithm. Finally, we discuss our findings and conclude the contribution of this chapter in Section 5.4 and 5.5.

5.1 Motivation

One objective in cancer research is to construct treatment procedures that could hinder tumour progression and promotes positive response to therapies. Recent development in molecular biology and genomics has contributed to a growing interest in the understanding of metabolic regulation and its effect on tumour physiology. It is understood that many main signalling pathways that are afflicted by tumour progression mark impact on core metabolism. These signalling pathways merge to accommodate cancer cell metabolism in order to sustain cell growth and proliferation by supporting the essential needs of the dividing cells (sustaining energy production, macromolecular biosynthesis and the maintenance of redox status) by altering the metabolism of carbohydrates, proteins, lipids and nucleic acids [68]. Cancer stands as a leading example of disease with indication of genetically perturbed metabolism. The extensive alteration and reprogramming of cancer cell metabolism by providing appropriate energy levels, to accommodate growth and proliferation by changing the metabolic operations in normal tissues is considered as an emerging hallmark of cancer [68, 130].

Metabolism affects an organism at cellular level by regulating enzyme functions through the control of mRNA transcriptions and translations [68]. Metabolism can also affects cell signally through post translational modifications by providing substrates as a feedback mechanism in regulating protein translations and enzymes catalysis [68, 144]. Finally, small effector molecules can affect metabolism through allosteric regulation, by binding of ligands to enzymes, either resulted in activating or inhibiting metabolic process [145]. The objective for these metabolic processes is to allocate the metabolites proportionately within the pathways by matching cellular needs [145]. With an interest in disease-oriented metabolic research, especially in the altered metabolic regulation in tumours, there is pressing need to translate the knowledge into novel diagnostic and therapeutic approaches.

One drawback of analysing metabolic network alone is that it could not fully explain the variations of differentiated cells. Complex organism has sophisticated regulatory and signalling pathways to evaluate perturbation in genetic and environmental signals that resulted in gene regulations. This process affects metabolism by regulating enzymes that control specific metabolic functions [84]. The two-way process between regulation and metabolism shows the importance of not to study the networks in isolation, but to integrate both the metabolic and regulatory networks in studying complex mechanism of organism's phenotypes.

We propose RACEMIC, a novel high-throughput data analysis that integrates regulatory signalling process in determining significant metabolic outliers responsible for perturbation in cell physiology. We extend the work in Ref. [26], that adapts active module approach [21] on the bipartite metabolic network. We model the altered metabolism as a stable metabolic network resulted from regulation process that affects level of enzyme isoform levels through mRNA transcriptions, splicing, and translations. An active module of metabolic network should be supported by strong indication of regulation process. Thus, we incorporate the information in gene regulatory network to determine active modules of metabolic network. Through this multilayer network representation, we seek to identify active modules in metabolic network by maximizing objective score that accommodates both the module scores in the metabolic layer and gene regulatory layer.

We applied RACEMIC to low grade glioma disease to study regulation of altered metabolism, and to elucidate ways to perturb the tumour progression. Comparisons are made with AMBIENT to validate the performance of our proposed algorithm. We identified several significant pathways characterized by short survival glioma patients that are supported by literature. In addition to identifying significant metabolic pathways, RACEMIC is capable to identify several potential targets genes that are involved in regulation process.

5.2 The Proposed Active Module Detection Algorithm

5.2.1 Interconnected Biological Network Representation

The representation of interconnected multilayer network in this chapter is similar to the interconnected network previously described in Section 4.2.1. Thus, the multilayer system that is composed of two-layer, i.e. metabolic and gene regulatory layer will be briefly described. We have an undirected bipartite metabolite network $G_{met} = (V_{met}, E_{met})$ that is consisted of nodes $V_{met} = R_{met} \cup M_{met}$ of either reaction $r \in R_{met}$ or metabolite $m \in M_{met}$. This metabolic network represents the metabolic layer in the multilayer network. Then, we have an undirected gene regulatory network $G_{reg} = \left(V_{reg}, E_{reg}\right)$ that represents gene regulatory layer. For every gene $g \in V_{reg}$ in the gene regulatory layer, there will be connection between the gene and reaction $r \in R_{met}$ in the metabolic layer, if the gene act as enzyme coding gene for the reaction. This links represents interlayer edges of the multilayer system.

Consider the lth active module of a bipartite metabolic network as a connected component $G^l_{met} = \left(V^l_{met}, E^l_{met}\right)$, such that $V^l_{met} \subset V_{met}$ and $E^l_{met} \subset E_{met}$. We assume that there exist an active region in the gene regulatory network that corresponds to the metabolic module G^l_{met}. The RACEMIC algorithm works by capturing 'active-joint modules' consisting of 'active-metabolic module' i.e. module in the metabolic layer, and 'active-regulatory module' i.e. active module in the gene regulatory layer that corresponds to the active-metabolic module.

5.2.2 Node Scoring Scheme

In the bipartite metabolic layer, each metabolite $m \in V_{met}$ is assigned node score $w(m)$ which is the degree of the metabolite in G_{met}. Reaction $r \in V_{met}$ is allocated node score $s(r)$ is **log fold-change** (i.e obtained from micro-array data) of enzyme-coding genes associated to it. When there are more than one enzyme-coding genes associated to the reaction, average score of the genes will be assigned to the reactions. This is the similar scheme as implemented in

Ref. [26]. When the reaction is can not be mapped to any genes (i.e no annotation of enzyme-coding genes, or the unavailability of the gene data), the reaction is assigned the median scores of all other known reactions. The scoring function will look for upregulated modules, which indicates high regulation of the genes associated to the reactions in the metabolic network. To search for downregulated modules, all reaction scores should be multiplied by -1, which implies that the lower score are converted to higher score, and vice versa.

For each gene $g \in G_{reg}$ in the gene regulatory layer, z-score is assign as the node score. The z-score for node g is defined as $z_g = \phi^{-1}\left(1 - p_g\right)$ in which p_g is the p-value that corresponds to significance of expression change of gene g, and ϕ^{-1} is the inverse CDF for normal distribution. A high positive z-score indicates the correlated gene is significantly changed in the experiment. This scoring scheme is based on the implementation of Ideker et al. [21]. To determine p_g for each respective gene g, we calculate the mean expression score under test condition $\bar{x}_{g,T}$, and under control condition $\bar{x}_{g,C}$. By having the null hypothesis that there are no difference between the test and control conditions ($\bar{x}_{g,T}$=$\bar{x}_{g,C}$), we test the hypothesis. We using t-test as test statistic with $t = \left(\bar{x}_{g,T} - \bar{x}_{g,C}\right) / SE_g$, where SE_g is the (non-pooled) within-groups standard error for gene g. The definition of within-groups standard error is $SE_g = \sqrt{\frac{s^2_{g,C}}{n_{g,C}} + \frac{s^2_{g,T}}{n_{g,T}}}$, where $s_{g,X}$ and $n_{g,X}$ are the standard deviation and number of samples of gene g under condition X respectively. Based on the test statistic, we can determine the p-value that indicates how likely to obtain the test statistic or higher, given that the null hypothesis holds.

5.2.3 Module Scoring Function

We define the score function of the lth active-joint module of a multilayer network consisted of a bipartite metabolic layer and a gene regulatory layer as a weighted sum of three major components:

$$S(l) = \alpha S_{\text{met}}(l) + \beta S_{\text{con}}(l) + \gamma S_{\text{reg}}(l) \tag{5.1}$$

in which:

$$S_{\text{met}}(l) = \ln(q)\left(\sum_i s\left(r_i^l\right) - \lambda \sum_j w\left(m_j^l\right)\right) \tag{5.2}$$

$$S_{\text{con}}(l) = \ln\left(\left|V_{seed}^l\right|\right) \tag{5.3}$$

$$S_{\text{reg}}(l) = s_{A,l} \tag{5.4}$$

α, β and γ represent the component weights.

$S_{\text{met}}(l)$ was initially proposed in as the scoring function of the AMBIENT algorithm [26] to measure the active modules in metabolic network, in which $q = \left|R_{met}^l\right| + \left|M_{met}^l\right|$ represents the number of nodes in the lth module. Function $s\left(r_i^l\right)$ is the score of ith reaction r_i in the lth module and it is assigned the average of log-fold change score of all the gene associated to the reaction. When the reaction is not associated to any genes due to the lack of gene expression data) the reaction will be given the median score of known reactions. The function $w\left(m_j^l\right)$ is the score of jth metabolite m_j and it is the value of degree of metabolite in the bipartite metabolic network.

λ denotes the average score of all positive reactions divided by the average value of w.

$$\lambda = \frac{|M_{met}| \sum_i s\left(r_i^+\right)}{\left|R_{met}^+\right| \sum_j w\left(m_j\right)} \tag{5.5}$$

in which $R_{met}^+ \subseteq R_{met} : s\left(r_i^+\right) > 0, \forall r_i^+ \in R_{met}^+$. The value of λ is introduced to obtain moderately size modules with meaningful biological representations.

The second term of the score function $S_{\text{con}}(l)$ denotes interlayer connections between the bipartite metabolic network and the gene regulatory network. The variable $\left|V_{seed}^l\right|$ is the number of enzyme-coding genes of the reactions in the metabolic module l that occurs in the regulatory network. $\left|V_{seed}^l\right|$ indicates the evidence of regulatory process for that particular module. The logarithmic term is utilised to favour modules with larger number of nodes, but giving smaller score increment as the number of nodes increases.

The regulatory active module score $S_{\text{reg}}(l) = s_{A,l}$ indicates the degree of intensity of gene regulation in conjunction with reactions in the lth metabolic module. Proposed by Ideker et al. [21], it is utilised to find 'active submodule with strong levels of differential expressions'. The function $s_{A,l}$ denotes the active submodule score of the gene regulatory network that contains genes associated to reactions in the lth metabolic module. To calculate $s_{A,l}$, first we need to calculate the aggregated node score for the lth module given as

$$z_{A,l} = \frac{1}{\sqrt{n}} \sum_{i \in V^l_{reg}} z_{i,l} \tag{5.6}$$

where $z_{i,l}$ is the z-score assigned to node i in the regulatory layer associated to lth module (refer to Section 5.2.2), and n is the number of genes in the submodule of the gene regulatory network.

Regulatory active module score $s_{A,l}$ is then calculated as

$$s_{A,l} = \frac{\left(z_{A,l} - \mu_n\right)}{\sigma_n} \tag{5.7}$$

where μ_n and σ_n is the mean and standard deviation of $z_{A,l}$ estimated by Monte Carlo resampling based on n samples.

5.2.4 Proposed RACEMIC Algorithm

Given that $L = \{l_1, l_2, \ldots, l_{k-1}, l_k\}$ corresponds to the a set of k modules, the objective of the RACEMIC algorithm is to find the set of modules with the maximum aggregated score

$$\text{maximize} \sum_{l \in L} S(l) \tag{5.8}$$

where $S(l)$ has been defined earlier in Equation 5.1.

The RACEMIC algorithm is described in Algorithm 5.1. The algorithm employs associated

regulatory active module procedure (i.e. denoted as RegulatoryActiveModule) which we describe in Algorithm 5.2. RACEMIC algorithm is a multilayer approach that takes bipartite metabolic network and gene regulatory network as its inputs. It seeks to maximize the score function that are derived from the aggregation of active-joint modules from these two networks. These joint modules are each composed of a subgraph of metabolic layer (i.e. active-metabolic module), and a corresponding module in the gene regulatory layer that relates to the regulation of metabolism (i.e. active-regulatory module). Each components of the active-joint module (each subgraph of the metabolic layer and gene regulatory layer), as well as the interaction between these components (i.e. in the form of interlayer edges) contribute to the overall score function.

The RACEMIC algorithm is based on the assumption that there are paths connecting the enzyme-coding genes (i.e. in the gene regulatory network) that are involved in catalyzing reactions for a specific biological pathway. In [146], Bromberg et al. proposed an algorithm to extract submodules from a protein-protein interaction network that empirically has been shown to contain the majority of activated transcription factors within a limited shortest paths range. A metabolic pathway is regulated by activating or deactivating a set of enzyme-coding genes involved in enzyme catalysis for the metabolic process. We are with the assumption that the results in Ref. [146] will also hold for a gene regulatory network where activated transcription factors in the network should be within shortest paths range. The enzyme-coding genes in the gene regulatory layer should be connected to each others by the shortest paths created by connections between the transcription factors. Each set of enzyme-coding genes should also become core components of regulatory active module. The regulatory active modules will contain other transcription factors and target genes that ensure connectivity between the enzyme-coding genes within the subnetwork, as well as achieving high regulatory active module score. The module search procedure to find regulatory active modules is described by Algorithm 5.2.

Algorithm 5.1 Finding Active-Joint Module

procedure ACTIVEJOINTMODULE(G_{met}, G_{reg})
  Initialize the number of toggle as q
  Initialize number of iteration N and temperature T
  Choose randomly q edges from G_{met} and set it as E
  Set H as a graph with K connected components (i.e. $H = \bigcup_{k=1}^{K} l_k^H$) induced by E **(a)**
  for $k = 1, \ldots, K$ **do**
    Get V_{seed}^k, the enzyme-coding genes associated to l_k^H
    $s_{A,k}$, $R_k \leftarrow$ RegulatoryActiveModule(G_{reg}, V_{seed}^k) **(b)**
  $\phi(H) \leftarrow \sum_k S\left(l_k^H\right)$
  for $n = 1, \ldots, N$ **do**
    Choose q edges from G_{met}, and set it as E_t
    $E_n \leftarrow E \cup E_t - E \cap E_t$
    Set F as a graph with K connected components (i.e. $F = \bigcup_{k=1}^{K} l_k^F$) induced by E_n **(c)**
    for $k = 1, \ldots, K$ **do**
      Get V_{seed}^k, the enzyme-coding genes associated to l_k^F
      $s_{A,k}$, $R_k \leftarrow$ RegulatoryActiveModule(G_{reg}, V_{seed}^k) **(b)**
    $\phi(F) \leftarrow \sum_k S\left(l_k^F\right)$
    if $\phi(F) > \phi(H)$ **then**
      Set $E \leftarrow E_n$ and $H \leftarrow F$
    else
      generate random number $\rho \in (0, 1]$
      if $\rho < e^{(\phi(F)-\phi(H))/T}$ **then**
        Set $E \leftarrow E_n$ and $H \leftarrow F$
    if Adaptive Temperature Criteria reached **(d)** **then**
      Set $q \leftarrow 0.9q$ and $T \leftarrow 0.9T$
  return $\left(H, \phi(H)\right)$, and $\left(R_k, s_{A,k}\right)$ for each kth module

(a): A supergraph H is produced by the collection of edges E. When a collection of any nodes in H are connected to each other by some paths, and they are not connected to any additional nodes in H, the subgraph that they produce is a connected component. There will be a total of K units of induced connected components. $H = \bigcup_{k=1}^{K} l_k^H$ where we define l_k^H as the kth connected components induced from H in lth module.
(b): RegulatoryActiveModule procedure in Algorithm 5.2 returns the active-regulatory module score $s_{A,k}$ (based on Equation 5.7) and the graph of active-regulatory module R_k (a subgraph of the gene regulatory network) that is associated to the kth connected component in the metabolic network.
(c): Applicable as explained in **(a)** by setting $H = F$ and $E = E_n$.
(d): Adaptive temperature criteria is reached when for the past 5000 iterations, $\sum_k S\left(l_k^H\right)$ does not increase by at least 5%. The annealing system is assume to reach thermal equilibrium.

Algorithm 5.2 Finding Regulatory Active Module based on Seeded Genes

procedure REGULATORYACTIVEMODULE(G_{reg}, V_{seed})
  Initialize number of iteration N and temperature T
  Set R as a subgraph of G_{reg} that connects V_{seed} by their shortest paths
  Set $s \leftarrow s_{A,R}$, where $s_{A,R}$ is the regulatory active module score of subgraph R
  while $i < N$ **do**
    With equal chance, choose either to add a neighbour or remove a node
    if Choose to remove a node **then**
      Excluding V_{seed}, randomly pick a node v in R
      Set R_i as a copy of R
      Remove node v from R_i
      if R_i is connected **then**
        $R \leftarrow R_i$
      else
        Choose to add a neighbour
    if Choose to add a neighbour **then**
      Set R_i as a copy of R
      Randomly pick a node v in R_i
      Randomly pick a neighbour of node v in G_{reg}
      Add the neighbour and its edge to R_i
    Set $s_i \leftarrow s_{A,R_i}$, where s_{A,R_i} is the regulatory active module score of subgraph R_i
    if $s_i > s$ **then**
      Set $s \leftarrow s_i$ and $R \leftarrow R_i$
    else
      generate random number $\rho \in (0,1]$
      if $\rho < e^{(s_i - s)/T}$ **then**
        Set $s \leftarrow s_i$ and $R \leftarrow R_i$
    if Adaptive Temperature Criteria reached **then** $^{(\star)}$
      $T \leftarrow 0.9T$
    $i \leftarrow i + 1$
  return s, R

$\star$: Adaptive temperature criteria is reached when for the past 500 iterations, regulatory active module score s does not increase by at least 5%. The annealing system is assume to reach thermal equilibrium.

5.3 Experiments and Results

5.3.1 Dataset and Preprocessing

To evaluate RACEMIC, the algorithm is performed on the bipartite metabolic network and gene regulatory network derived from LGG short survival subtype data. The data has been used earlier in our implementation of ActiveFlow in Chapter 4. Therefore, in this section, the network will be briefly described.

Figure 5.1a shows the multilayer network formalism of the short survival bipartite metabolic network and gene regulatory network in the form of supra-adjacency matrix. The bipartite metabolic network of the short survival subtype is consisted of a total of 5810 nodes, from which 2352 are metabolites and 3458 are reactions. There are 14898 edges in the metabolic network. The short survival subtype is designated as the test condition while the long survival subtype is taken as the control condition. Each reaction in the metabolic network is assigned log-fold change score based on these conditions. For the gene regulatory network, there are 2349 genes (from which 56 of the nodes are transcription factors) and 4630 edges linking the genes. Each gene in the regulatory network is assigned z-score derived from p-value that indicates the significance of gene expression change between the control and the test condition. There are 209 enzyme-coding genes in the regulatory network that has been identified as enzyme-coding genes of the reactions in the metabolic network. Enzyme-coding genes in the gene regulatory network are connected to their corresponding reaction nodes in the bipartite metabolic network. These connections are used in active-joint module score calculation. These enzyme-coding genes are categorized as ‘seed genes’ as they provide links between the metabolic and gene regulatory network. In total, there are 1214 interlayer edges that connect the bipartite metabolic network to the gene regulatory network.

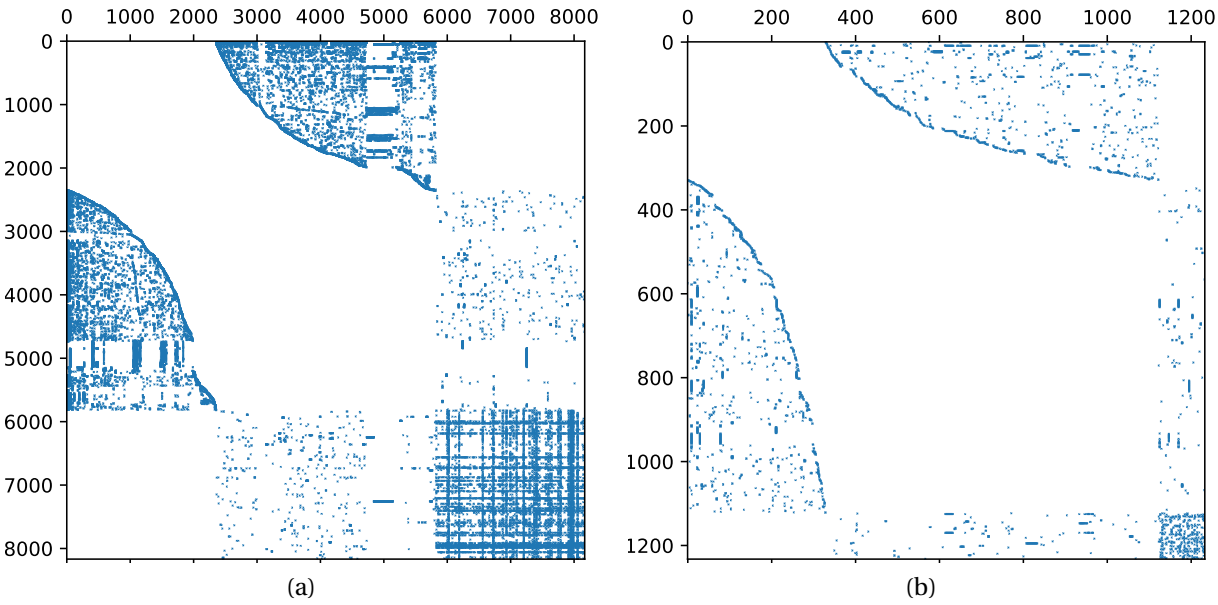


Figure 5.1: **(a)** Supra-adjacency matrix plot of metabolic-regulatory multilayer network of short survival subtype. Bipartite metabolic network contains 5810 nodes, from which 2352 are metabolites (indices from 0 to 2351 in the plot) and 3458 reactions (indices from 2352 to 5809). The gene regulatory networks contains 2349 genes (indices from 5810 to 8159). There are 14898 metabolite-reaction intralayer edges in the metabolic layer, 4630 gene-gene intralayer edges in the gene regulatory layer, and 1214 reaction-gene interlayer edges connecting the metabolic and gene regulatory layers. **(b)** Supra-adjacency matrix plot for active joint-module 1 of short survival subtype as obtained by the RACEMIC algorithm. The active metabolic module contains a total of 1233 active nodes where 329 are metabolites (indices from 0 to 328), 794 reactions (indices from 329 to 1123) from the metabolic layer, and 110 genes (indices from 1124 to 1232) from the gene regulatory layer. There are 1535 intralayer edges in the metabolic layer, 192 intralayer edges in the gene regulatory layer and 244 inter layer edges connecting the metabolic and the gene regulatory layer.

5.3.2 Active Module Identification of LGG Short Survival Subtype

We performed RACEMIC on the LGG model with the objective to understand the molecular mechanism of the disease. Comparisons with AMBIENT are made by evaluating significant metabolic pathways yielded by each algorithm. The RACEMIC algorithm was run with score function parameters $\alpha = 1$, $\beta = \gamma = 0.5$. We are giving the highest priority to metabolic layer during the algorithm search by setting $\alpha = 1$ as this is where the primary metabolism activities is taking place. By setting lower weights for $\beta = \gamma = 0.5$ (which is half of α), we are indicating

that the gene regulatory layer and interlayer edges that connects the metabolic-gene layers are supporting factors to affirm the existence of metabolism activities in metabolic layer. The maximum number of iteration for simulated annealing to find the active modules in the metabolic network is set to 1,000,000. The maximum iteration to search for active module searching in the gene regulatory network is set to 10,000. Only the top 10 highest score modules are taken into consideration. The AMBIENT algorithm is run by its default parameters, and and the number of simulated annealing iterations is set to 1,000,000. Similarly only the top 10 modules are retained. Statistically significant over-represented pathways are determine by using IMPaLA [131]. Although RACEMIC has an extra layer of gene information as provided by active-regulatory modules, only the compounds in the active-metabolic modules are taken as the input for IMPaLA to determine the significant metabolic pathways. This step is taken to allow a level comparison with AMBIENT.

AMBIENT and RACEMIC obtained 3 and 4 upregulated significant modules respectively. The top over-representation pathways obtained the modules both AMBIENT and RACEMIC is tabulated in Table 5.1. Modules are name with prefix 'U' that denotes upregulation condition. The first 3 modules that AMBIENT and RACEMIC uncover are considerably similar in term of pathways that they representation. In fact, AMBIENT's U2 has the same module composition as RACEMIC's U3, having 'Detoxification of Reactive Oxygen Species' as the top pathway with F-score of 0.32. AMBIENT's U3 and RACEMIC's U2 have 'Synthesis of Leukotrienes (LT) and Eoxins (EX)' while AMBIENT having slightly better F-score at 0.4667 as compared to RACEMIC 0.4210.

For the largest module U1, AMBIENT obtains a total of 845 nodes (622 reactions and 223 metabolites) and RACEMIC obtains a active-joint module that is comprised of 1123 metabolic nodes (329 metabolites and 794 reactions) and additional information of 110 genes from the active-regulatory module. The composition of genes in the regulatory layer is 65 seed genes, 27 TFs and 18 other target genes. Figure 5.1b illustrates the RACEMIC's active-joint module

Table 5.1: Top over-representation pathways in the modules of LGG short survival subtype identified by AMBIENT and RACEMIC

AMBIENT	Pathway Name	Precision	Recall	F-score
U1	Glycosaminoglycan metabolism	0.2346	0.6333	0.3423
U2	Detoxification of Reactive Oxygen Species	0.6667	0.2105	0.3200
U3	Synthesis of Leukotrienes (LT) and Eoxins (EX)	1.0000	0.3043	0.4667
RACEMIC	**Pathway Name**	**Precision**	**Recall**	**F-score**
U1	Metabolism of carbohydrates	0.2920	0.4167	0.3433
U2	Synthesis of Leukotrienes (LT) and Eoxins (EX)	0.5333	0.3478	0.4210
U3	Detoxification of Reactive Oxygen Species	0.6667	0.2105	0.3200
U4	Pyrimidine biosynthesis	0.6250	0.1724	0.2703

Note: Over-representation analysis was conducted based on HMDB IDs of known metabolites by using IMPaLA [131]. The ground truth to calculate F-score is quantified from metabolite background list of the pathways.

U1 as formalized in the form of supra-adjacency matrix. There are 1535 intralayer edges in the metabolic layer, 192 intralayer edges in the gene regulatory layer and 244 interlayer edges connecting the metabolic and the gene regulatory layer.

Figure 5.2 shows the genes that make up the active-regulatory module 1. A high number of the seed genes are at least having moderate positive fold-change scores that denote the condition of upregulation. The genes with the highest log fold-change scores are NNMT (score of 4.79) and CP (score of 4.25). The presence of NNMT is consistent with the study of glioma, where it has been reported to be overly expressed in human glioma cells [147]. In other study, it is has been to undergo upregulation in GBM cells [148]. CP has also been reported by Tye et al. [149] as widely expressed in primary and recurrent high-grade gliomas but not widely presence in low-grade oligodendroglial tumours. Our findings shows that RACEMIC is able to capture high scoring metabolic reactions which are associated to enzyme encoding genes that has been shown to be highly expressed and upregulated in glioma tumours.

The 27 TFs and 18 target genes (excluding seed target genes) that form the regulatory active module have scores that are mixed of upregulated and downregulated log fold-change values. The inclusion of downregulated genes is understandable as, although the objective is

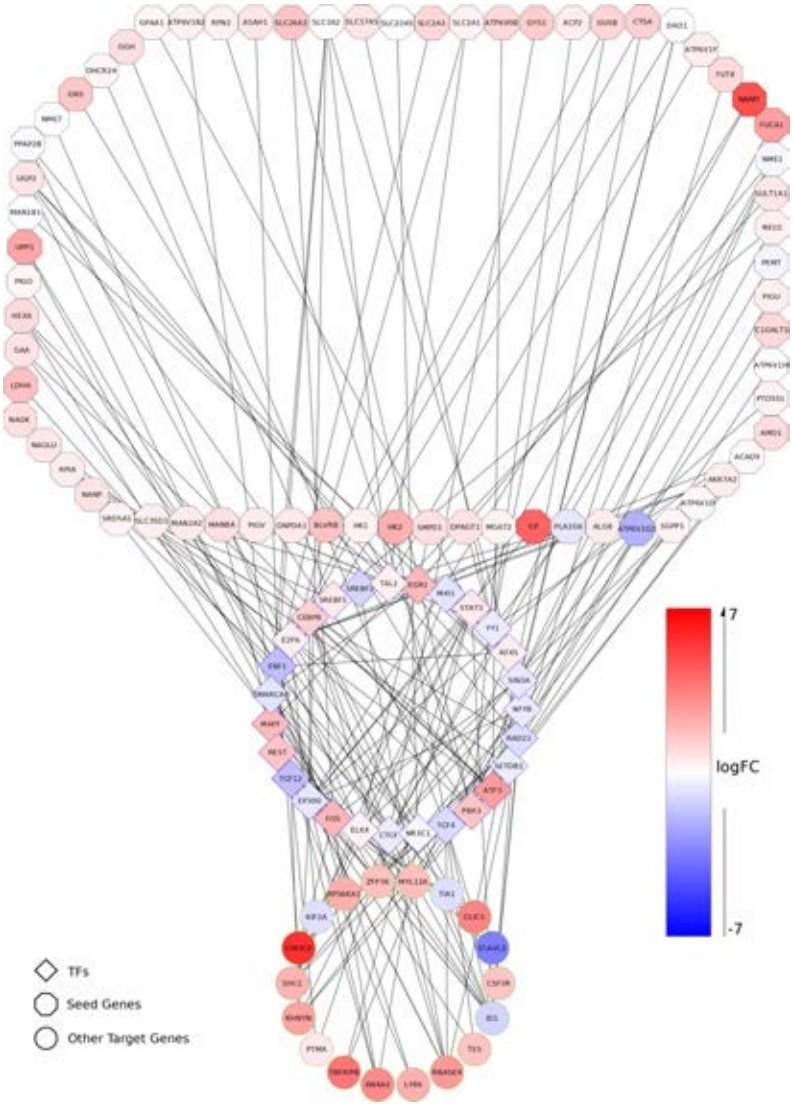


Figure 5.2: Active Regulatory-Module 1. A high proportion of the seed genes are at least having moderate positive fold-change scores that denotes the condition of upregulation. The genes with the highest log fold-change scores are NNMT (score of 4.79) and CP (score of 4.25). RACEMIC uncovers 28 TFs and 21 other target genes (excluding seed target genes) which are not directly encoded with metabolomics data.

to get regulatory active module with high aggregated gene scores, the connectedness between seed genes must be strictly adhered thus resulting in inclusion of some low scores genes. The target genes with highest log-fold change score is CHI3L2 (score of 5.69). The presence of CHI3L2 is consistent with literature on glioblastoma as it has been reported by Areshkov and Kavsan [150] to increase in expression in human glioblastoma, and characterized as one of

the most highly expressed genes in glioblastoma [151]. As CHI3L2 is not classified as enzyme encoding gene which are associated to metabolic reactions, it could not be identified by methodologies that directly rely on metabolic model alone (e.g. AMBIENT). This findings shows the usefulness of RACEMIC by being able to uncover highly significant genes by integrating metabolomics data and regulatory information.

We make a comparison between the top 10 pathways for both AMBIENT and RACEMIC in U1 and the results is tabulated in Table 5.2. The top pathways of U1 for AMBIENT and RACEMIC are fairly similar with 8 of 10 pathways overlapped. The results show AMBIENT generally has better F-score for the top pathways that denote better balanced representation of relevant nodes. However, RACEMIC obtains higher (or at least the same) number of relevant nodes for the pathways that denotes better recall. We make a deduction based from this result that RACEMIC, with the additional information retrieved from the regulatory layer, assembles greater number of nodes of multiple relevant pathways that control a biological process. This behaviour causes RACEMIC to achieve big module that causes it to have lower F-score when the module is analyzed with respect to single pathway. However, RACEMIC present an advantage as the bigger module obtained by RACEMIC allow us to identify the multiple interconnected pathways that are significantly affected.

We have earlier shown the existence of hierarchical structures in the active modules of yeast model in Section 3.3.2. The presence of multiple pathways in RACEMIC's U1 gives an indication of the existence of hierarchical structures in the module. Therefore, we conduct MotifPro hierarchical clustering analysis on module on all upregulated modules to identify important subclusters in the module that represent multiple metabolic pathways that are linked and differentially affected by the LGG short survival subtype cohort. Refer to Figure 3.8 in page 71 on how the clusters are named. With settings $\gamma = 0.2$ and $\phi = 0.02$, only U1 are able to be partitioned into three smaller clusters U1C1 with 599 nodes (391 reaction, 168 metabolites), U1C2 with 218 nodes (170 reactions, 48 metabolites) and U1C3 with 346 nodes

Table 5.2: The top 10 pathways of the largest module U1 obtained by AMBIENT and RACEMIC (based on over-representation analysis on metabolites).

pathway name	AMBIENT		RACEMIC	
	# of mets	F-score	# of mets	F-score
Metabolism of carbohydrates(*)	29	0.3277	40	0.3433
Glycosaminoglycan metabolism(*)	19	0.3423	20	0.2395
Metabolism of proteins(*)	29	0.3152	38	0.3167
Post-translational protein modification(*)	27	0.3253	33	0.2973
Keratan sulfate/keratin metabolism(*)	12	0.2581	12	0.1610
Asparagine N-linked glycosylation(*)	22	0.3077	26	0.2613
Heparan sulfate/heparin (HS-GAG) metabolism(*)	15	0.2913	16	0.2012
Metabolism(*)	62	0.1379	103	0.2157
Keratan sulfate biosynthesis(**)	9	0.2000	9	0.1233
Synthesis of substrates in N-glycan biosythesis(**)	17	0.2556	21	0.2222
Metabolism of nucleotides(***)	22	0.2126	38	0.2890
Transport of vitamins, nucleosides, and related molecules(***)	15	0.2174	24	0.2474

Note: 'mets' denotes metabolites. The indicator denotes that the pathway is within top 10 pathways in (*) both RACEMIC and AMBIENT, (**) only in AMBIENT, (***) only in RACEMIC.

(233 reactions, 113 metabolites). U1C1 can further be partitioned in 3 clusters U1C1C1 with 59 nodes (41 reactions, 18 metabolites), U1C1C2 with 96 metabolites (57 reactions, 39 metabolites) and U1C1C3 that contains 404 nodes (293 reactions, 111 metabolites).

Table 5.3 tabulates final representation of the upregulated modules of LGG short survival subtype and its top significant pathways. Module U1, which is the biggest, is represented by its subclusters. Overall, U1C2 has the highest F-score of 0.5625 for 'Keratan sulfate/keratin metabolism' pathway. U1C2 is mainly related to metabolic process in the Golgi compartment such a biosythesis of keratan sulfate and elongation of N-glycan. In fact, U1C2 is fairly similar in term of the relevant pathways to module U1 that we discovered by using ActiveFlow in Section 4.4.3. U1C2 has precision, recall and F-score of 0.3636, 0.8 and 0.5 respectively for 'N-glycan antennae elongation in the medial/trans-Golgi', while ActiveFlow's U1 has precision, recall and F-score of 0.35, 0.7 and 0.4667 respectively. U1C1C3 that has 'HS-GAG degradation' as the top pathway, is comparably similar to ActiveFlow's module that we describe in the

Table 5.3: Upregulated Modules of LGG Short Survival Subtype.

pathway name	Module	Precision	Recall	F-score
Synthesis and interconversion of nucleotide di- and triphosphates	U1C1C1	0.6923	0.2195	0.3333
Estrogen biosynthesis	U1C1C2	0.2105	0.4444	0.2857
HS-GAG degradation	U1C1C3	0.1591	0.636364	0.2545
Keratan sulfate/keratin metabolism	U1C2	0.4500	0.7500	0.5625
Metabolism of nucleotides	U1C3	0.400	0.1746	0.2431
Synthesis of Leukotrienes (LT) and Eoxins (EX)	U2	0.5333	0.3478	0.4211
Detoxification of Reactive Oxygen Species	U3	0.6667	0.2105	0.3200
Pyrimidine biosynthesis	U4	0.6250	0.1724	0.2703

previous chapter. U1C1C3 has precision, recall and F-score of 0.1591, 0.6364 and 0.2546 respectively, while ActiveFlow's 0.3, 0.5455 and 0.3871 respectively.

Cluster U1C1C1 has 'Synthesis and interconversion of nucleotide di- and triphosphates' as the main pathway where nucleotide monophosphates are phosphorylated to form nucleotide diphosphates and triphosphates. Pyrimidines nucleoside triphosphates in the cluster are uridine triphosphate (UTP) and cytidine triphosphate (CTP). Prymidine UTP and purines (adenosine triphosphate (ATP) and adenosine) are known as important signalling metabolites that mediate the proliferation of gliomas through the activation of P1 and P2 receptors [152]. Zhang et al. reported significant differences between the profile of metabolism of nucleotides in normal and tumour cells [153]. The concentration of ATP, UTP, adenosine monophosphate (AMP) and guanosine monophosphate (GMP) are higher in tumours cells, indicating abnormality of the metabolism of nucleotides pathway in tumours. Zhang et al. further proposes ATP and UTP the potential nucleotides biomarkers for tumour cells.

Nicotine metabolism is another pathway which are part of U1C1C1. One important molecule in the pathway is nicotinamide adenine dinucleotide (NAD), along with its related isoenzyme nicotinamide nucleotide adenylyltransferase 1 (NMNAT1) that are found in glial cells. NAD is an important co-factor that regulates a wide range of biological processes such as aging, apoptosis, inflammation and cancer [154, 155]. NAD is regulated by nicotinamide

phosphoribosyltransferase (NAMPT). NAMPT is regularly overexpressed in cancer tissues, and is potential candidate gene for cancer therapies [155]. The expression of NAMPT correlates with mortility rate and invasive capacity of glioma cells [156]. NMNAT1 which is localized in nucleus and essential for regulating nuclear NAD pool, is found to link NAMPT to the expression of transcription factor E2F2 that is responsible for self-renewal of human GBM stem-like cells (GSCs) [157].

Cluster U1C1C2 has 'estrogen biosynthesis' as the main significant pathway. Many of the reactions in this cluster are with missing enzyme-coding genes data, thus given score of median of other reactions. Therefore, we consider this cluster functions to bridge other cluster in module U1 but could not confirm the importance of estrogen metabolism towards progression of LGG short survival subtype. However, we can establish that Cytochrome P450 1B1 (CYP1B1) is highly upregulated in the module. Estradiol, which is the substrate for the reaction catalysed by CYP1B1, was found to inhibit glioma cells proliferation and promote cell death[158]. Thus, the high regulation of CYP1B1 can be a mechanism to reduce the inhibition of of glioma proliferation by estradiol. Findings by Barnett et al. shows that CYP1B1 is frequently expressed in all main glioma subtypes and grades, and increase in intensity as tumours become more malignant [159]. Barnett et al. further suggested CYP1B1 to be used as a potential target for immunotherapy for low-grade glioma as the extended survival time can establish better immune responses. Enzyme Nad(p)h biliverdin reductase (BVR), that are connected to the steroid metabolism pathway by its product bilirubin, are also upregulated in U1C1C2. BVR is a cytoprotective and growth promoter protein, and is identified as tumour promoter [as reviewed in 160]. Expression of BVR in GBM increases as a response to hypoxic condition of the tumours[161]. Gibbs et al. in [160] also suggested the use of BVR-based peptides to inhibit BVR's activities associated to growth-promoting kinase to decrease the rate of tumour proliferation.

For module U2, 'Synthesis of Leukotrienes (LT) and Eoxins (EX)' pathway is highly over-

represented. Leukotrienes (LTs), a family of lipid mediators derived from arachidonate, play a key role in the as mediators of inflammation [162]. LTs can be classified into two groups, LTB4 and cysteinyl LTs (LTC4, LTD4 and LTE4) which are generated of products from leukotriene A4 (LTA4) [162]. LTA4, LTB4, LTC4 and LTD4 are presence in this module. 5-lipoxygenase (ALOX5), the enzyme that execute the conversion from arachidonate to LTA4 is highly upregulated. ALOX5-LTA4 hydrolase pathway has been suggested to promote glioma tumour proliferation, and LTB4 is identified as target metabolite for the proliferation of glioma cell lines [163]. Inhibition of ALOX5, promotes differentiation, and impede self-renewal of glioma stem-like cells [164]. Leukotriene-C4 synthase that forms LTC4 is also upregulated. LTC4 and arachidonate act as gate to store-independent Ca^{2+} currents, which is regulated by Ca^{2+} sensor stromal interacting molecule 1 (STIM1) and the Ca^{2+} signalling channel Orai1 [as cited in 165]. Orai1 has been found to promote tumorigenesis [166]. Orai1 with STIM1 has been reported to induce GBM cell invasiveness, with notable effect on the cell proliferation [165].

The expression of gamma-glutamyl transferase (GGT) that produces LTD4 is moderatedly upregulated in U2. GGT is expressed in high proportion of grade III astrocytomas and GBM patients, and low indication of expression in normal tissues [167]. Corti et al. suggested two possibilities to explain the increase of expression of GGT [168]. First, GGT has antioxidant function, that act as a defensive mechanism of the cells against oxidative stress. For the second possibility, GGT takes the pro-oxidant function, which can be part of regulatory mechanism through the metabolism of LTC4. Corti et al. suggested further that the pro-oxidant role taken by GGT could lead to persistent oxidative stress, and modulate cells proliferation/apoptosis processes in tumour progression.

Module U3 contains mainly of 'detoxification of Reactive Oxygen Species' pathway. The pathway functions as aerobic cells defence mechanism by converting active oxygen species to lesser reactive compounds [as cited in 169]. In the module, superoxide which is damaging due to its high reactivity is converted to hydrogen peroxide which is less reactive product. There is

high upregulation of enzyme Fe(III) reduction (ascorbate) in the extracellular compartment that catalyses the reduction of iron (Fe^{3+} to Fe^{2+}) by the mediation of ascorbate. The chemical reaction also generate dehydroascorbide which can be rapidly reduced back to ascorbate due to its instability [170]. The usage of high-dose of intravenous ascorbate in cancer therapy has been conducted in many animal trials e.g. for ovarian, pancreatic and glioblastoma tumours in [171] and several human clinical studies e.g. in [172, 173]. Ascorbate aids the production of hydrogen peroxide, which also depend of iron for its generation. A high concentrations of hydrogen peroxide that enter tumour cells induces cytotoxicity effects [as cited in 174]. However, extracellular FE(II) can protect cancer cells from hydrogen peroxide [175], and abolish the anticancer effect of ascorbate [174]. There is a possibility that the higher concentration of Fe(III) reduction (ascorbate) is a mechanism for the cancer cells to protect the cell interior from hydrogen peroxide.

Module U4 is mainly associated to pyrimidine biosynthesis which occur partly in extracellular and nucleus compartment. There are indication of upregulation of the enzyme gamma-glutamyl hydrolase (GGH). GGH catalyses the hydrolysis of a gamma-glutamyl bond, and regulates folate levels and polyglutamylated folates for optimal condition of nucleotide biosynthesis [as cited in 176]. High concentrations of GGH is linked to poor prognosis in breast cancer [177] and prostate cancer [178]. High level of GGH can also lead to unfavourable clinical outcome in invasive breast cancer, and Shubbar et al. suggested GGH as potential biomarker in short-term follow-up treatment [177]. Glutamate, which is a product of metabolic reaction calatalysed by GGH, is found to have important role in the tumour phenotype of glioma by activating PI3K/Akt and MAPK pathways, and in assisting invasion of glioma cells into normal brain [179].

5.4 Discussion

Bromberg et al. [146] produce results where a high majority of activated transcription factors

are within a short steps between each others for a given signalling pathways within a PPI network. Based on the findings, we build our conjecture that the genes in a regulatory pathways (regarded as a subset of signalling pathways) should be within a short steps of each other. Some of the activated genes are responsible for the coding of enzymes that catalyze active reactions that we found in our active-metabolic module, which in turn are related to certain biological mechanisms. These enzyme coding genes should be part regulatory pathways and should at least be connected by some paths. Building on this premise, the enzyme-encoding genes are regarded as 'seed genes', and should provide the basis in our search for active regulatory pathways (in the regulatory network) that complement the active-metabolic module (in the metabolic network).

The approach to find active regulatory pathways, a connected area within the regulatory network that undergoes significant change of regulation, furnish us with a new way of tackling the problem of finding candidate for target genes. The approach by RACEMIC provide us with new insight, by not limiting our search space to the enzyme coding genes that are responsible for active reactions, where we could widen the scope by also looking for other genes that are experiencing significant changes in relation to these the active reactions. Thus, we could be able to identify a set of genes, that could be a combination of TF genes, enzyme coding genes, and other genes related to different type of proteins - where these set of genes could be regards as directly or indirectly related to the active biological pathways in question. Thus, RACEMIC opens the possibility to study additional two group of genes, which are TF genes and 'other genes' - the set of information that can not be extracted by referring by analysing metabolic network alone.

Having the same goals as AMBIENT, RACEMIC seek to find metabolic modules within a bipartite metabolic network by determining transcriptomic changes undergone by enzyme coding genes associated with reactions in the network. However, besides looking at changes happening at metabolic level, RACEMIC also takes genes regulation into consideration. The

score of a metabolic module is not only determined by the enzyme level changes but also two other factors. First, RACEMIC also look for the evidence of regulation of genes that are linked to enzymes. Then, if there is any gene regulation taking place with respect to the enzymes, RACEMIC will search for region within the regulatory network that encounters significant transcriptomic changes - denoting active regulatory subnetwork associated to the metabolic subnetwork obtained earlier.

Another advantage presented by RACEMIC is where the resulting module contains extra regulatory dimension. The modules obtained are actually joint-modules consisted of active-metabolic modules and active-regulatory modules. The availability of the active-regulatory module presents us with a different way on how we extract information from results as compared to AMBIENT. In AMBIENT approach, we will make use on the set of reactions and metabolites in determining relevant pathways. In RACEMIC approach, we could use the extra gene information in order to resolve relevant pathways. We demonstrate our approach by using IMPaLA by combining metabolite and gene information.

Reactions that are identified in a metabolic module could present insight in determining biological pathways. However, as a reaction only contains a single score, which could be derived from number of enzymes and genes associated to it, the relevance of a significant reaction could be difficult to evaluate. With the availability of the extra layer of gene information, RACEMIC is useful by providing a way to identify significant genes in regulatory layer, by tracing back from reactions in metabolic layer. In a different alternative approach, once we have identify relevant pathways, we could pinpoint significant overlapping genes related to these pathways. From these genes, we could trace the path towards reactions in the active metabolic-modules. This approach provide us with the information on reactions, enzymes and the genes associated to them, which are significant based on the transcriptomic analysis of our data. The genes can be assigned as candidates for the future research in gene targeting in order further understanding the complex biological mechanism that is taking place.

On an Intel i7 1.8GHz, RACEMIC takes about 745 minutes running time on short survival glioma subtype as compared to 70 minutes by AMBIENT. Thus, RACEMIC execution time is in the order of ten times longer as compared to AMBIENT. The longer time is due to nature that RACEMIC run extra calculation to find regulatory-layer active module for each metabolic-layer active module obtained in each iteration.

The effectiveness of RACEMIC could be affected by the accuracy of the metabolic and regulatory networks used for analysis. As in many cases for biological networks, curated metabolic and regulatory networks are based on ongoing research, and availability of information in biological databases and literature. Thus, the curated network presented could not be regarded as exhaustive.

For the case of metabolic network, enzyme coding genes may not be mapped to reactions due to unavailability of enzymes or genes information. Therefore, the transcriptomic mapping of genes to the metabolic network is not complete. The approach by RACEMIC is to assign central tendency measures (median or mean) to these reactions. Although the assigning of central tendency measure is considered adequate in our approach, nevertheless there will be loss of information in the analysis as transcriptomic information are not fully extracted.

Currently, there are still major inadequacies in the availability of genome-wide regulatory networks for many species. There are also lack of experimentally verified regulatory relationships, which can resort to networks being built based on predicted regulatory interactions. Many of the edges of these networks could be suspected to be false-positives and affect the quality of RACEMIC analysis.

The approach taken by RACEMIC does not directly pursue to solve the limitation of data due to the shortfall of gene-reactions mappings in the metabolic networks, and the inadequacy of reliable genome-wide regulatory networks. However, as presented in the results of our case study, RACEMIC approach of integrating the two different layers achieves better performance in obtaining significant pathways as compared to just relying on metabolic data alone.

5.5 Conclusion

In our work, we put forward RACEMIC, a novel method in modelling a biological system using a multilayer concept by integrating transcriptomics and metabolomics information. Altered metabolism process is modelled as stable metabolic environment controlled and regulated from gene regulation process. Thus, in search for significant metabolic modules within a bipartite metabolic network, RACEMIC takes into account of regulation process, by finding connected metabolic regions that are affected by significant transcriptomics changes. Additionally RACEMIC give emphasis on the conditions where there are significant changes of gene expression within the neighbourhood of regulatory network where the transcriptomics changes takes place.

The inclusion of gene regulation mechanism in metabolic network analysis allow a more comprehensive approach in tackling altered metabolism problem. By taking into account of gene regulation process, RACEMIC is able to find relevant and significant metabolic pathways, with better coverage and more converged results as compared to existing approach.

RACEMIC demonstrate the need to integrate different layers of information in complex system analysis of metabolism phenotypes. As metabolism is not be considered as a stand-alone system, but are influenced by gene regulation and signalling mechanisms, it also deserves the same comprehensive methodology of analysis.

CHAPTER 6

Conclusion

In this chapter, we conclude the findings of the thesis, listed the limitations of our approach, and outlines some potential future works.

6.1 Research Questions Revisited

We return to the research questions to summarize our findings as reported in this thesis:

> How to adaptively construct active modules that spans through an interconnected network consisted of a bipartite metabolic network and a gene regulation network?
> In particular, how to devise a generic algorithm that balance the composition of node in modules so that the modules are mainly composed of the nodes from each layer of the interconnected network to denote intralayer information flow in each layer, and interlayer information flow between the two layers?

- The ActiveFlow algorithm in Chapter 4 balances the composition of active modules so that the modules are not only composed of nodes from one particular layer. The interconnected network is merged into one network while the nodes in the merged network preserve the type of network they are originally from. ActiveFlow treats the

interconnected network as a one-layer network during the procedure to merge nodes into modules, but the 'LMRatio' function (in Algorithm 4.1) controls the composition of modules as though there are two separate layers. This characteristics will ensure the composition of the active modules are balanced.

- The RACEMIC algorithm in Chapter 5 implements a different approach that resulted in a balanced module. Metabolic layer is treated as the main layer of the interconnected network. During the algorithm search, the regions in the metabolic layer that show strong activities are assigned as active-modules. Then, for each active-metabolic module, the search procedure is extended to the regulatory layer to find activity in the gene regulatory regions that supports the activity in the metabolic module. The final active-joint modules that RACEMIC uncovers are composed of both layers that signals information flow between them.

How to infer modules based on the topological features in the interconnected network? For this case, we would like to identify modules solely based on the topology of nodes in the interconnected network, which could be in term of nodes' degree distributions, or the node/edge centralities.

- The ActiveFlow algorithm in Chapter 4 can identify topological modules based on nodes' degree distribution that are reflected by adjacency matrix of the interconnected network. By setting the topological weight $\eta = 1$ and activity weight $\kappa = 0$, the modules that will be uncover will solely be based on network topology.

How to infer modules based on the molecular activity in the interconnected network? In particular, this is the procedure to identify active modules of the interconnected network.

- We can identify active module of interconnected network by using either ActiveFlow (Chapter 4) or RACEMIC algorithm (Chapter 5). When the activity weight κ of ActiveFlow is set to positive value, activity weighted matrix that reflects the strength of nodes' in term of transcriptomic activities will be taken into account to extract active modules. The RACEMIC algorithm will naturally uncover active modules as the nodes' scores represent the strength of activities in the network.

How to identify sub-modules of a bipartite metabolic network? For this case, the modules is only composed of nodes from the metabolic bipartite network.

- The MotifPro algorithm in Chapter 3 can partition bipartite metabolic network into sub-modules. The bipartite metabolic network can be transformed into weighted graph of motif M, where M is a specific motif that reflects the structure of the bipartite metabolic network. The network is iteratively cut into two smaller subgraphs while seeking to minimize the ratio of the number of motif that are being cut in the subgraphs.

Does the modules (and sub-modules) that has been inferred relates to meaningful biological functions? Does it highlight significant metabolic pathways that can leads to certain clinical traits and phenotypes, which in turn may help us in identifying molecular phenotypes that may likely be a 'key biomarker' in a biological process?

- The ActiveFlow (Chapter 4) and RACEMIC algorithms (Chapter 5) found modules that are related to meaningful biological functions. The largest topological module of the

ActiveFlow algorithm found 'central carbon metabolism in cancer' amongst the top pathways of LGG short survival network.

- Active modules of both ActiveFlow and RACEMIC are linked to 'HS-GAG degradation' pathway, which has been associated to tumour invasion and Glioblastoma.

- Several significant genes that are linked to glioma and cancers and may serve as key biomarkers are identified as BCAT1, GUSB, NNMT, CP and CYP1B1.

6.2 Concluding remarks

This thesis has achieved the following contributions:

1. We propose the ActiveFlow algorithm, which is an active modules detection algorithm of an interconnected network. We observe two important and useful features that can be realized by the proposal of the algorithm which are

   – **The current research trend on module detection in multilayer network is on multiplex network. As there are no current implementation of module detection on interconnected network, we believe that our proposed method is useful to model biological complex system as we show in this thesis, and other type of problems such as cascading failures or spreading processes.**

   – The implementation of current information flow method to interconnected network produces modules that are mainly conserved within their own layer as the majority of nodes in the modules inferred are from single layer. **Our algorithm is able to produce modules that spans across the layers, thus producing better representation of information flow between each layer in interconnected network.**

2. We devise a module detection algorithm for bipartite metabolic network that consider gene regulatory process to infer modules. The current implementation of active modules for bipartite metabolic network are only considering the topology and activity of enzyme-coding genes (i.e. in the form of changes of transcription levels) in network. However, metabolism are complex system that are regulated by many means such as allosteric control of enzymes by small compounds and the transcriptional regulation of enzymes. Therefore, **our algorithms employ extra layer of information in the form of gene regulatory network to better replicate the complex mechanism of metabolism.**

3. We devise two module detection algorithms for interconnected network to evaluate the significant molecular changes of activity that are associated with certain clinical traits or cellular responses. The first algorithm, ActiveFlow, is a fast heuristic algorithm that can infer modules based on the topological properties of nodes, or combine the topology and activity of nodes across the layers in detecting modules. The second algorithm RACEMIC takes into account of strong regulatory activity to support the active regions in the metabolic layer. **ActiveFlow produces active modules for bipartite network with results that are comparably similar with current method but at faster executions. RACEMIC are slower than current method, but is capable to infer larger modules with better recall**.

4. We devise an algorithm in the form of framework to extract clusters from active modules to work on the low precision problem that has been experienced by current method (i.e. AMBIENT) and our proposed algorithm (i.e. RACEMIC). The advantage of RACEMIC is its ability to include low-scoring nodes (that denote low activity) in the formation of the high-scoring active modules. This low-scoring nodes acts as bridging-nodes that serve as missing links that connect high-scoring regions into bigger modules, which otherwise would be isolated. However, the

formation of large modules suffer poor precision as the modules is no longer dedicated to smaller number of pathways. Our conjecture is the significant pathways in the modules are concentrated in modular structures, and are linked together to produce the overall function of the modules. **We address the issue of poor precision by large module such that, we devise the method of motif projection, and the extraction of clusters from the network based on the concentration of active-motifs in the network. Our results found the existence of hierarchical clusters in the active modules of bipartite metabolic network**.

5. **We discover significant building blocks of biological functions in the form of active modules for the lower grade glioma short survival subtype**. Our goals are to identify potential biomarkers that can be used to predict patients' prognosis in order to develop cancer therapy for the disease. Throughout the topological and active modules analysis this thesis, we have identified several significant pathways, notably 'N-glycan antennae elongation in the medial/trans-Golgi' and 'HS-GAG degradation'. We also identify several significant genes and supported by literature on glioma and cancers that could serve as potential biomarkers such as BCAT1, GUSB, NNMT, CP and CYP1B1.

Listed below are the limitations of our approach:

- A reaction in the metabolic network can be associated to more than one enzyme-coding gene (i.e. seed gene), and an enzyme-coding gene can be associated to more than one reactions. Besides, non-seed genes and transcription factors in the regulatory network are also associated to many genes in the network. The ActiveFlow algorithm constructs non-overlapping modules. For example, as we merge a seed gene to some reactions in a new module, other reactions will lose their association to the enzyme-coding gene. Thus, ActiveFlow does not fully

capture the overlapping feature of biological molecules.

– The RACEMIC algorithm conducts two levels of searching during each iteration, such that it will explore the metabolic layer space for active-metabolic module, and then it will explore the gene regulatory layer space to search for active-regulatory module. Although RACEMIC can extract active modules that are more accurate representative of biological functions, this feature makes RACEMIC to be computationally expensive as compared AMBIENT or ActiveFlow.

6.3 Future Work

For future works, the opportunity for further research includes some of the issues listed as follows:

1. The type of interconnected network that we investigate in our research is a two-layer network. Further investigation should be done on inferring modules for multilayer network that are composed of more than two layers. Furthermore, framework for module identification of multilayer network are available for multilayer multiplex network. The ability to devise module identification algorithm could pave way for us to study a much more complex biological system. In Ref. [6], the author propose that the system genetic view of complex traits are composed of five-layer system which are gene, transcript, protein, metabolite and microbiome spaces. The ability to investigate the interaction of multiple layers of biological networks can allow us have better understanding of complex biological system.
2. We can consider an information flow algorithm that extend the capability of ActiveFlow where it can extract overlapping communities. In this way, a node can appear in more than one multilayer modules and represent a better model of biological system.

3. In our implementation of ActiveFlow algorithm, we carry out a procedure that once modules are merged, the nodes in the modules will remain indefinitely. This aspect may be sensitive to initialization process. Therefore, there are improvement that can be made by adopting global optimization techniques such as evolutionary algorithms, genetic algorithms or simulated annealing (i.e. enhancing the current implementation) that obtain good solutions to replace the ActiveFlow's greedy approach. Although they could increase the execution time of the algorithm, the possibility to get more accurate pictures of complex biological system such as mechanism of diseases could outweight the disadvantages.

4. The active modules of the MotifPro algorithm is minimally constructed as we only consider fold-change scores of reactions in determining the active states of reactions. We do not take into account of low-scoring reactions which may otherwise contribute to the continuation of pathways in the modules. As a result, some pathways may become disconnected and allocated into more than one module (e.g. one carbon pool by folate pathway exist in two separate downregulated modules). For future work, the active states of reactions could be deduced from more that one source e.g. additionally from topological scores of enzyme-encoding genes in regulatory or protein-protein interaction network. Reactions scores could be obtained from the aggregated reaction scores of these multiple sources of information. This strategy will increase the likelihood of important reactions to be included in the active modules.

※※※※※

References

[1] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

[2] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010.

[3] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[4] Víctor de Lorenzo. From the selfish gene to selfish metabolism: revisiting the central dogma. *BioEssays*, 36(3):226–235, 2014.

[5] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101–13, 2004.

[6] Mete Civelek and Aldons J Lusis. Systems genetics approaches to understand complex traits. *Nature reviews. Genetics*, 15(1):34–48, 2014.

[7] Erzsébet Ravasz. Detecting hierarchical modularity in biological networks. In *Computational Systems Biology*, pages 145–160. Springer, 2009.

[8] Arend Hintze and Christoph Adami. Evolution of complex modular biological networks. *PLoS computational biology*, 4(2):e23, 2008.

[9] Dirk M Lorenz, Alice Jeng, and Michael W Deem. The emergence of modularity in biological systems. *Physics of life reviews*, 8(2):129–160, 2011.

[10] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[12] Thang N Dinh and My T Thai. Community detection in scale-free networks: approximation algorithms for maximizing modularity. *IEEE Journal on Selected Areas in Communications*, 31(6):997–1006, 2013.

[13] Vicente Arnau, Sergio Mars, and Ignacio Marín. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364–378, 2004.

[14] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586): 1551–1555, 2002.

[15] Daniel Segre, Alexander DeLuna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nature genetics*, 37(1):77, 2005.

[16] Roby P Bhattacharyya, Attila Reményi, Brian J Yeh, and Wendell A Lim. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.*, 75:655–680, 2006.

[17] Xuewei Wang, Ertugrul Dalkic, Ming Wu, and Christina Chan. Gene module level analysis: identification to networks and dynamics. *Current opinion in biotechnology*, 19(5):482–491, 2008.

[18] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

[19] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631, 2006.

[20] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.

[21] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 (suppl_1):S233–S240, 2002.

[22] Michael G Hunnewell and Neil S Forbes. Active and inactive metabolic pathways in tumor spheroids: determination by gc–ms. *Biotechnology progress*, 26(3):789–796, 2010.

[23] Qiaosheng Zhang, Jie Li, Hanqing Xue, Leilei Kong, and Yadong Wang. Network-based methods for identifying critical pathways of complex diseases: a survey. *Molecular BioSystems*, 12(4):1082–1089, 2016.

[24] Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003–1010, 2008.

[25] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S Lun, Brian Weiner, Maha R Farhat, Tan-Yun Cheng, D Branch Moody, Megan Murray, and James E Galagan. Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol*, 5(8):e1000489, 2009.

[26] William a Bryant, Michael J E Sternberg, and John W Pinney. AMBIENT: Active Modules for Bipartite Networks–using high-throughput transcriptomic data to dissect metabolic response. *BMC systems biology*, 7:26, 2013.

[27] Cancer Genome Atlas Research Network et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med*, 2015(372):2481–2498, 2015.

[28] Ivana Jovčevska, Nina Kočevar, and Radovan Komel. Glioma and glioblastoma-how much do we (not) know? *Molecular and clinical oncology*, 1(6):935–941, 2013.

[29] Elizabeth B Claus, Kyle M Walsh, John K Wiencke, Annette M Molinaro, Joseph L Wiemels, Joellen M Schildkraut, Melissa L Bondy, Mitchel Berger, Robert Jenkins, and Margaret Wrensch. Survival and low-grade glioma: the emergence of genetic information. *Neurosurgical focus*, 38(1):E6, 2015.

[30] Subha Madhavan, Jean-Claude Zenklusen, Yuri Kotliarov, Himanso Sahni, Howard A Fine, and Kenneth Buetow. Rembrandt: Helping Personalized Medicine Become a Reality through Integrative Translational Research. *American Association for Cancer Research*, 7(2):157–167, 2009.

[31] Yunpeng Liu, Desire G. Ngoga, Kate Hollinshead, Weiqi Chen, Marina Vabistsevits, Daniel C. Swan, Jean-Baptiste Cazier, Garth S. Cruickshank, Shan He, and Daniel A. Tennant. Unbiased clustering analysis reveals high risk patient population of low grade gliomas. unpublished work, 2014.

[32] Ludwig Geistlinger, Gergely Csaba, Simon Dirmeier, Robert Küffner, and Ralf Zimmer. A comprehensive gene regulatory network for the diauxic shift in saccharomyces cerevisiae. *Nucleic acids research*, 41(18):8452–8463, 2013.

[33] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(2):291–307, 1970.

[34] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[35] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[36] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[37] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

[38] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[39] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[40] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[41] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108, 1979.

[42] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

[43] James C Bezdek. Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms*, pages 43–93. Springer, 1981.

[44] Jeffrey Baumes, Mark K Goldberg, Mukkai S Krishnamoorthy, Malik Magdon-Ismail, and Nathan Preston. Finding communities by clustering a graph into overlapping subgraphs. *IADIS AC*, 5:97–104, 2005.

[45] Jeffrey Baumes, Mark Goldberg, and Malik Magdon-Ismail. Efficient identification of overlapping communities. In *International Conference on Intelligence and Security Informatics*, pages 27–36. Springer, 2005.

[46] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[47] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043): 814, 2005.

[48] Shihua Zhang, Rui-Sheng Wang, and Xiang-Sun Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490, 2007.

[49] Scott White and Padhraic Smyth. A spectral clustering approach to finding communities in graphs. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 274–285. SIAM, 2005.

[50] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[51] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4): 1118–1123, 2008.

[52] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[53] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[54] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

[55] Vincenzo Nicosia, Ginestra Bianconi, Vito Latora, and Marc Barthelemy. Growing multiplex networks. *Physical review letters*, 111(5):058701, 2013.

[56] Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Finding redundant and complementary communities in multidimensional networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2181–2184. ACM, 2011.

[57] Matteo Magnani and Luca Rossi. Pareto distance for multi-layer network analysis. In ArielM. Greenberg, WilliamG. Kennedy, and NathanD. Bos, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 7812 of *Lecture Notes in Computer Science*, pages 249–256. Springer Berlin Heidelberg, 2013.

[58] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason a Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science (New York, N.Y.)*, 328(5980):876–8, 2010.

[59] Peter J. Mucha and Mason a. Porter. Communities in multislice voting networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4):041108, 2010.

[60] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012.

[61] Wenyuan Li, Chun-Chi Liu, Tong Zhang, Haifeng Li, Michael S Waterman, and Xianghong Jasmine Zhou. Integrative analysis of many weighted co-expression networks using tensor computation. *PLoS computational biology*, 7(6):e1001106, 2011.

[62] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008.

[63] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027, 2015.

[64] Weiyi Liu, Toyotaro Suzumura, Hongyu Ji, and Guangmin Hu. Finding overlapping communities in multilayer networks. *PloS one*, 13(4):e0188747, 2018.

[65] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *nature*, 466(7307):761, 2010.

[66] R Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, UK, 1976.

[67] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and Marie-France Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 5(4):594–617, 2008.

[68] Rob A Cairns, Isaac S Harris, and Tak W Mak. Regulation of cancer cell metabolism. *Nature Reviews Cancer*, 11(2):85–95, 2011.

[69] Gunnar Schramm, Stefan Wiesberg, Nicolle Diessl, Anna-Lena Kranz, Vitalia Sagulenko, Marcus Oswald, Gerhard Reinelt, Frank Westermann, Roland Eils, and Rainer König. Pathwave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics*, 26(9):1225–1231, 2010.

[70] Rainer Breitling, Anna Amtmann, and Pawel Herzyk. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC bioinformatics*, 5(1):100, 2004.

[71] Raul Montañez, Miguel Angel Medina, Ricard V Sole, and Carlos Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays*, 32(3):246–256, 2010.

[72] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

[73] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome research*, pages gr–077693, 2008.

[74] Fan Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740, 2007.

[75] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[76] TM Murali and Simon Kasif. Extracting conserved gene expression motifs from gene expression data. In *Biocomputing 2003*, pages 77–88. World Scientific, 2002.

[77] H Jeong, B Tombor, R Albert, Z N Oltvai, and a L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.

[78] Pall F Jonsson and Paul a Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics (Oxford, England)*, 22(18):2291–7, 2006.

[79] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–8, 2002.

[80] Carlos Prieto, Alberto Risueño, Celia Fontanillo, and Javier De las Rivas. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS one*, 3(12):e3911, 2008.

[81] Takeshi Obayashi, Shinpei Hayashi, Masayuki Shibaoka, Motoshi Saeki, Hiroyuki Ohta, and Kengo Kinoshita. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic acids research*, 36(Database issue):D77–82, 2008.

[82] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, 302(5643):249–55, 2003.

[83] Matthias E Futschik, Gautam Chaurasia, Anna Tschaut, Jenny Russ, M Madan Babu, and Hanspeter Herzel. Functional and transcriptional coherency of modules in the human protein interaction network. *Journal of Integrative Bioinformatics*, 4(3):198–207, 2007.

[84] Béatrice Desvergne, Liliane Michalik, and Walter Wahli. Transcriptional regulation of metabolism. *Physiological reviews*, 86(2):465–514, 2006.

[85] Haiyuan Yu and Mark Gerstein. Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences*, 103(40):14724–14731, 2006.

[86] MEJ Newman, C Moore, and A Clauset. Hierarchical structure and the prediction of missing links in networks. *Nature 453*, 2008.

[87] Changning Liu, Jing Li, and Yi Zhao. Exploring hierarchical and overlapping modular structure in the yeast protein interaction network. *BMC genomics*, 11(4):S17, 2010.

[88] Markus J Herrgård, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yalçin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, 26(10):1155–1160, 2008.

[89] Stefan Schuster, Thomas Pfeiffer, Ferdinand Moldenhauer, Ina Koch, and Thomas Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361, 2002.

[90] Hongwu Ma and An-Ping Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2): 270, 2003.

[91] Afshin Raouf, Yun Zhao, Karen To, John Stingl, Allen Delaney, Mary Barbara, Norman Iscove, Steven Jones, Steven McKinney, Joanne Emerman, et al. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell stem cell*, 3(1):109–118, 2008.

[92] Tucker A Patterson, Edward K Lobenhofer, Stephanie B Fulmer-Smentek, Patrick J Collins, Tzu-Ming Chu, Wenjun Bao, Hong Fang, Ernest S Kawasaki, Janet Hager, Irina R Tikhonova, et al. Performance comparison of one-color and two-color platforms within the microarray quality control (maqc) project. *Nature biotechnology*, 24(9):1140, 2006.

[93] Miguel Ramirez-Gaona, Ana Marcu, Allison Pon, An Chi Guo, Tanvir Sajed, Noah A Wishart, Naama Karu, Yannick Djoumbou Feunang, David Arndt, and David S Wishart. Ymdb 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic acids research*, 45(D1):D440–D445, 2017.

[94] Javier López-Ibáñez, Florencio Pazos, and Mónica Chagoyen. Mbrole 2.0 - functional enrichment of chemical compounds. *Nucleic acids research*, 44(W1):W201–W204, 2016.

[95] J.Kalervo Hiltunen, Anu M. Mursula, Hanspeter Rottensteiner, Rik K. Wierenga, Alexander J. Kastaniotis, and Aner Gurvitz. The biochemistry of peroxisomal -oxidation in the yeast saccharomyces cerevisiae. *FEMS Microbiology Reviews*, 27(1):35–64, 2003.

[96] Bernard Turcotte, Xiao Bei Liang, François Robert, and Nitnipa Soontorngun. Transcriptional regulation of nonfermentable carbon utilization in budding yeast. *FEMS yeast research*, 10(1):2–13, 2009.

[97] Abel R. Alcázar-Román and Susan R. Wente. Inositol polyphosphates: a new frontier for regulating gene expression. *Chromosoma*, 117(1):1–13, 2008.

[98] Matthew J Brauer, Alok J Saldanha, Kara Dolinski, and David Botstein. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Molecular biology of the cell*, 16(5):2503–2517, 2005.

[99] Edwin A Cossins and Liangfu Chen. Folates and one-carbon metabolism in plants and fungi. *Phytochemistry*, 45(3):437–452, 1997.

[100] Brian R Gibson, Stephen J Lawrence, Jennifer M Smith, Naomi Shelton, Janet M Smith, and KA Smart. Oxygen as toxin: oxidative stress and brewing yeast physiology. *Cerevisia*, 31(1):25, 2006.

[101] Michel J Penninckx. An overview on glutathione in saccharomyces versus non-conventional yeasts. *FEMS Yeast Research*, 2(3):295–305, 2002.

[102] Marc T Elskens, Charles J Jaspers, and Michel J Penninckx. Glutathione as an endogenous sulphur source in the yeast saccharomyces cerevisiae. *Microbiology*, 137(3): 637–644, 1991.

[103] D Zhou, L L Cam, C A Laughton, K R Korzekwa, and S Chen. Mutagenesis study at a postulated hydrophobic region near the active site of aromatase cytochrome p450. *Journal of Biological Chemistry*, 269(30):19501–19508, 1994.

[104] James R Broach. Nutritional control of growth and development in yeast. *Genetics*, 192 (1):73–105, 2012.

[105] Hervé Alexandre, Isabelle Rousseaux, and Claudine Charpentier. Relationship between ethanol tolerance, lipid composition and plasma membrane fluidity in saccharomyces cerevisiae and kloeckera apiculata. *FEMS Microbiology Letters*, 124(1):17–22, 1994.

[106] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L Gallia, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897): 1807–1812, 2008.

[107] Tomoyoshi Soga. Cancer metabolism: key players in metabolic reprogramming. *Cancer science*, 104(3):275–281, 2013.

[108] Gangman Yi, Sing-Hoi Sze, and Michael R Thon. Identifying clusters of functionally related genes in genomes. *Bioinformatics*, 23(9):1053–1060, 2007.

[109] Laura Cantini, Enzo Medico, Santo Fortunato, and Michele Caselle. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports*, 5:17386, 2015.

[110] Konstantin Voevodski, Shang-Hua Teng, and Yu Xia. Finding local communities in protein networks. *BMC bioinformatics*, 10(1):297, 2009.

[111] Gilles Didier, Christine Brun, and Anaïs Baudot. Identifying communities from multiplex biological networks. *PeerJ*, 3:e1525, 2015.

[112] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[113] J-C Delvenne, Sophia N Yaliraki, and Mauricio Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760, 2010.

[114] Taher Alzahrani, Kathy J Horadam, and Serdar Boztas. Community detection in bipartite networks using random walks. In *Complex Networks V*, pages 157–165. Springer, 2014.

[115] Jan A van der Knaap and C Peter Verrijzer. Undercover: gene control by metabolites and metabolic enzymes. *Genes & development*, 30(21):2345–2369, 2016.

[116] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027, 2015.

[117] Daniel Edler, Ludvig Bohlin, and Martin Rosvall. Mapping higher-order network flows in memory and multilayer networks with infomap. *arXiv preprint arXiv:1706.04792*, 2017.

[118] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal-Special Topics*, 178(1):13–23, 2009.

[119] Renaud Lambiotte and Martin Rosvall. Ranking and clustering of nodes in networks with smart teleportation. *Physical Review E*, 85(5):056107, 2012.

[120] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

[121] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[122] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.

[123] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[124] Yuliang Wang, James a Eddy, and Nathan D Price. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC systems biology*, 6:153, 2012.

[125] Ines Thiele, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, Stefan G Thorleifsson, Rasmus Agren, Christian Bölling, Sergio Bordel, Arvind K Chavali, Paul Dobson, Warwick B Dunn, Lukas Endler, David Hala, Michael Hucka, Duncan Hull, Daniel Jameson, Neema Jamshidi, Jon J Jonsson, Nick Juty, Sarah Keating, Intawat Nookaew, Nicolas Le Novère, Naglis Malys, Alexander Mazein, Jason a Papin, Nathan D Price, Evgeni Selkov, Martin I Sigurdsson, Evangelos Simeonidis, Nikolaus Sonnenschein, Kieran Smallbone, Anatoly Sorokin, Johannes H G M van Beek, Dieter Weichart, Igor Goryanin, Jens Nielsen, Hans V Westerhoff, Douglas B Kell, Pedro

Mendes, and Bernhard ØPalsson. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–25, 2013.

[126] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P Boyle, Philip Cayting, Alexandra Charos, David Z Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E Reddy, Brian Reed, Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J Farnham, Richard M Myers, Sherman M Weissman, and Michael Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

[127] R Mei, X Di, TB Ryder, E Hubbell, S Dee, TA Webster, CA Harrington, M-h Ho, J Baid, SP Smeekens, et al. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 18(12):1593–1599, 2002.

[128] Stefan Bröer. Amino acid transport across mammalian intestinal and renal epithelia. *Physiological reviews*, 88(1):249–286, 2008.

[129] Qian Wang and Jeff Holst. L-type amino acid transport and cancer: targeting the mtorc1 pathway to inhibit neoplasia. *American journal of cancer research*, 5(4):1281, 2015.

[130] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[131] Atanas Kamburov, Rachel Cavill, Timothy MD Ebbels, Ralf Herwig, and Hector C Keun. Integrated pathway-level analysis of transcriptomics and metabolomics data with impala. *Bioinformatics*, 27(20):2917–2918, 2011.

[132] Harald Sontheimer. A role for glutamate in growth and invasion of primary brain tumors. *Journal of neurochemistry*, 105(2):287–295, 2008.

[133] Martje Tönjes, Sebastian Barbus, Yoon Jung Park, Wei Wang, Magdalena Schlotter, Anders M Lindroth, Sabrina V Pleier, Alfa HC Bai, Daniela Karra, Rosario M Piro, et al. Bcat1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type idh1. *Nature medicine*, 19(7):901–908, 2013.

[134] Hsiao-Chen Tu, Yung-Chun Hsiao, Wan-Yu Yang, Shin-Lin Tsai, Hua-Kuo Lin, Chong-Yi Liao, Jeng-Wei Lu, Yu-Ting Chou, Horng-Dar Wang, and Chiou-Hwa Yuh. Up-regulation of golgi α-mannosidase ia and down-regulation of golgi α-mannosidase ic activates unfolded protein response during hepatocarcinogenesis. *Hepatology Communications*, 1(3):230–247, 2017.

[135] Kenji Uchimura, Hideki Muramatsu, Tadashi Kaname, Haruko Ogawa, Taishi Yamakawa, Qi-Wen Fan, Chikako Mitsuoka, Reiji Kannagi, Osami Habuchi, Itsuo Yokoyama, et al. Human n-acetylglucosamine-6-o-sulfotransferase involved in the biosynthesis of 6-sulfo sialyl lewis x: molecular cloning, chromosomal mapping, and expression in various organs and tumor cells. *The Journal of Biochemistry*, 124(3):670–678, 1998.

[136] Jianhai Jiang, Xiaoning Chen, Jialin Shen, Yuanyan Wei, Tao Wu, Yanzhong Yang, Hanzhou Wang, Hongliang Zong, Junwu Yang, Si Zhang, et al. β1, 4-galactosyltransferase v functions as a positive growth regulator in glioma. *Journal of Biological Chemistry*, 281 (14):9482–9489, 2006.

[137] Nikos Afratis, Chrisostomi Gialeli, Dragana Nikitovic, Theodore Tsegenidis, Evgenia Karousou, Achilleas D Theocharis, Mauro S Pavão, George N Tzanakakis, and Nikos K Karamanos. Glycosaminoglycans: key players in cancer cell biology and treatment. *The FEBS journal*, 279(7):1177–1197, 2012.

[138] Jorge Filmus. Glypicans in growth control and cancer. *Glycobiology*, 11(3):19R–23R, 2001.

[139] Israel Vlodavsky and Yael Friedmann. Molecular properties and involvement of heparanase in cancer metastasis and angiogenesis. *Journal of Clinical Investigation*, 108(3): 341, 2001.

[140] Qi Luan, Jing Sun, Chunying Li, Guoyou Zhang, Yajie Lv, Gang Wang, Chengxin Li, Cuiling Ma, and Tianwen Gao. Mutual enhancement between heparanase and vascular endothelial growth factor: a novel mechanism for melanoma progression. *Cancer letters*, 308(1):100–111, 2011.

[141] Günther Stockhammer, Alois Obwegeser, Herwig Kostron, Petra Schumacher, Armin Muigg, Stefan Felber, Hans Maier, Irene Slavc, Eberhard Gunsilius, and Günther Gastl. Vascular endothelial growth factor (vegf) is elevated in brain tumor cysts and correlates with tumor progression. *Acta neuropathologica*, 100(1):101–105, 2000.

[142] Valeria Valente, Silvia A. Teixeira, Luciano Neder, Oswaldo K. Okamoto, Sueli M. Oba-Shinjo, Suely KN Marie, Carlos A. Scrideli, Maria L. Paçó-Larson, and Carlos G. Carlotti. Selection of suitable housekeeping genes for expression analysis in glioblastoma using quantitative rt-pcr. *BMC Molecular Biology*, 10(1):17, 2009.

[143] Bernhard Sperker, Ulrike Werner, Thomas E Mürdter, Ceren Tekkaya, Peter Fritz, Rainer Wacke, Ulrich Adam, Manfred Gerken, Bernd Drewelow, and Heyo K Kroemer. Expression and function of ß-glucuronidase in pancreatic cancer: potential role in drug targeting. *Naunyn-Schmiedeberg's archives of pharmacology*, 362(2):110–115, 2000.

[144] Ralph J. Deberardinis and Craig B. Thompson. Cellular metabolism and disease: What do metabolic outliers teach us? *Cell*, 148(6):1132–1144, 2012.

[145] Christian M Metallo and Matthew G Vander Heiden. Understanding metabolic regulation and its influence on cell physiology. *Molecular cell*, 49(3):388–398, 2014.

[146] Kenneth D Bromberg, Avi Ma'ayan, Susana R Neves, and Ravi Iyengar. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science (New York, N.Y.)*, 320(5878):903–9, 2008.

[147] Kazuo Yamada, Takeshi Miyazaki, Nobumasa Hara, and Mikako Tsuchiya. Interferon-gamma elevates nicotinamide N-methyltransferase activity and nicotinamide level in human glioma cells. *Journal of nutritional science and vitaminology*, 56:83–86, 2010.

[148] J M Markert, C M Fuller, G Y Gillespie, J K Bubien, L a McLean, R L Hong, K Lee, S R Gullans, T B Mapstone, and D J Benos. Differential gene expression profiling in human brain tumors. *Physiological genomics*, 5(1):21–33, 2001.

[149] S. L. Tye, A. G. Gilg, L. B. Tolliver, W. G. Wheeler, B. P. Toole, and B. L. Maria. Hyaluronan Regulates Ceruloplasmin Production By Gliomas and Their Treatment-Resistant Multipotent Progenitors. *Journal of Child Neurology*, 23(10):1221–1230, 2008.

[150] P. A. Areshkov and V. M. Kavsan. Chitinase 3-like protein 2 (chi3l2, ykl-39) activates phosphorylation of extracellular signal-regulated kinases erk1/erk2 in human embryonic kidney (hek293) and human glioblastoma (u87 mg) cells. *Cytology and Genetics*, 44(1): 1–6, 2010.

[151] Stanislav Avdieiev, Liliia Savinska, Valeriy Filonenko, and Vadym Kavsan. Chitinase 3-like 2 protein monoclonal antibodies. *Hybridoma*, 31(1):32–39, 2012.

[152] Fernanda B Morrone, Maria C Jacques-Silva, Ana P Horn, Andressa Bernardi, Gilberto Schwartsmann, Richard Rodnight, and Guido Lenz. Extracellular nucleotides and nucleosides induce proliferation and increase nucleoside transport in human glioma cell lines. *Journal of neuro-oncology*, 64(3):211–218, 2003.

[153] Chenchen Zhang, Zheng Liu, Xi Liu, Lan Wei, Yanjie Liu, Jing Yu, and Lixin Sun. Targeted metabolic analysis of nucleotides and identification of biomarkers associated with cancer in cultured cell models. *Acta Pharmaceutica Sinica B*, 3(4):254–262, 2013.

[154] Luis F De Figueiredo, Toni I Gossmann, Mathias Ziegler, and Stefan Schuster. Pathway analysis of nad+ metabolism. *Biochemical Journal*, 439(2):341–348, 2011.

[155] Antje Garten, Susanne Schuster, Melanie Penke, Theresa Gorski, Tommaso De Giorgis, and Wieland Kiess. Physiological and pathophysiological roles of nampt and nad metabolism. *Nature Reviews Endocrinology*, 11(9):535–546, 2015.

[156] Remco van Horssen, Marieke Willemse, Anna Haeger, Francesca Attanasio, Tuba Güneri, Albrecht Schwab, Christian M Stock, Roberto Buccione, Jack AM Fransen, and Bé Wieringa. Intracellular nad (h) levels control motility and invasion of glioma cells. *Cellular and Molecular Life Sciences*, 70(12):2175–2190, 2013.

[157] Amit D Gujar, Son Le, Diane D Mao, David YA Dadey, Alice Turski, Yo Sasaki, Diane Aum, Jingqin Luo, Sonika Dahiya, Liya Yuan, et al. An nad+-dependent transcriptional program governs self-renewal and radiation resistance in glioblastoma. *Proceedings of the National Academy of Sciences*, 113(51):E8247–E8256, 2016.

[158] Geoffrey C Kabat, Anne M Etgen, and Thomas E Rohan. Do steroid hormones play a role in the etiology of glioma? *Cancer Epidemiology and Prevention Biomarkers*, 19(10): 2421–2427, 2010.

[159] Julia A Barnett, Diana L Urbauer, Graeme I Murray, Gregory N Fuller, and Amy B Heimberger. Cytochrome p450 1b1 expression in glial cell tumors: an immunotherapeutic target. *Clinical cancer research*, 13(12):3559–3567, 2007.

[160] Peter EM Gibbs, Tihomir Miralem, and Mahin D Maines. Biliverdin reductase: a target for cancer therapy? *Frontiers in pharmacology*, 6, 2015.

[161] Sung Su Kim, Sin Seong, Seong Hyeon Lim, and Sung Young Kim. Biliverdin reductase plays a crucial role in hypoxia-induced chemoresistance in human glioblastoma. *Biochemical and biophysical research communications*, 440(4):658–663, 2013.

[162] JN Sharma and LA Mohammed. The role of leukotrienes in the pathophysiology of inflammatory disorders: is there a case for revisiting leukotrienes as therapeutic targets? *Inflammopharmacology*, 14(1-2):10–16, 2006.

[163] K Ishii, M Zaitsu, N Yonemitsu, Y Kan, Y Hamasaki, and M Matsuo. 5-lipoxygenase pathway promotes cell proliferation in human glioma cell lines. *Clinical neuropathology*, 28 (6):445, 2009.

[164] Bin Wang, Shi-cang Yu, Jian-yong Jiang, Gavin Wallace Porter, Lin-tao Zhao, Zhe Wang, Hong Tan, You-hong Cui, Cheng Qian, Yi-fang Ping, et al. An inhibitor of arachidonate 5-lipoxygenase, nordy, induces differentiation and inhibits self-renewal of glioma stem-like cells. *Stem Cell Reviews and Reports*, 7(2):458–470, 2011.

[165] Rajender K Motiani, María C Hyzinski-García, Xuexin Zhang, Matthew M Henkel, Iskandar F Abdullaev, Yu-Hung Kuo, Khalid Matrougui, Alexander A Mongin, and Mohamed Trebak. Stim1 and orai1 mediate crac channel activity and are essential for human glioblastoma invasion. *Pflügers Archiv-European Journal of Physiology*, 465(9): 1249–1260, 2013.

[166] Mingye Feng, Desma M Grice, Helen M Faddy, Nguyen Nguyen, Sharon Leitch, Yingyu Wang, Sabina Muend, Paraic A Kenny, Saraswati Sukumar, Sarah J Roberts-Thomson, et al. Store-independent activation of orai1 by spca2 in mammary tumors. *Cell*, 143(1): 84–98, 2010.

[167] Christoph Schäfer, Carsten Fels, Matthias Brucke, Hans-Jürgen Holzhausen, Hannes Bahn, Maria Wellman, Athanase Visvikis, Peter Fischer, and Nikolai G Rainov. Gamma-glutamyl transferase expression in higher-grade astrocytic glioma. *Acta Oncologica*, 40 (4):529–535, 2001.

[168] Alessandro Corti, Maria Franzini, Aldo Paolicchi, and Alfonso Pompella. Gamma-glutamyltransferase of cancer cells at the crossroads of tumor progression, drug resistance and drug targeting. *Anticancer research*, 30(4):1169–1181, 2010.

[169] Tohru Fukai and Masuko Ushio-Fukai. Superoxide dismutases: role in redox signaling, vascular function, and diseases. *Antioxidants & redox signaling*, 15(6):1583–1606, 2011.

[170] Juan Du, Joseph J Cullen, and Garry R Buettner. Ascorbic acid: chemistry, biology and the treatment of cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1826 (2):443–457, 2012.

[171] Qi Chen, Michael Graham Espey, Andrew Y Sun, Chaya Pooput, Kenneth L Kirk, Murali C Krishna, Deena Beneda Khosh, Jeanne Drisko, and Mark Levine. Pharmacologic doses of ascorbate act as a prooxidant and decrease growth of aggressive tumor xenografts in mice. *Proceedings of the National Academy of Sciences*, 105(32):11105–11109, 2008.

[172] Christopher M Stephenson, Robert D Levin, Thomas Spector, and Christopher G Lis. Phase i clinical trial to evaluate the safety, tolerability, and pharmacokinetics of high-dose intravenous ascorbic acid in patients with advanced cancer. *Cancer chemotherapy and pharmacology*, 72(1):139–146, 2013.

[173] LJ Hoffer, M Levine, S Assouline, D Melnychuk, SJ Padayatty, K Rosadiuk, C Rousseau, L Robitaille, and WH Miller Jr. Phase i clinical trial of iv ascorbic acid in advanced malignancy. *Annals of Oncology*, 19(11):1969–1974, 2008.

[174] Marija Mojić, Jelena Bogdanović Pristov, Danijela Maksimović-Ivanić, David R Jones, Marina Stanić, Sanja Mijatović, and Ivan Spasojević. Extracellular iron diminishes anticancer effects of vitamin c: an in vitro study. *Scientific reports*, 4, 2014.

[175] Stephen L Hempel, Garry R Buettner, Duane A Wessels, George M Galvan, and Yunxia Q O'Malley. Extracellular iron (ii) can protect cells from hydrogen peroxide. *Archives of biochemistry and biophysics*, 330(2):401–408, 1996.

[176] SE Kim, PD Cole, RC Cho, A Ly, L Ishiguro, KJ Sohn, R Croxford, BA Kamen, and YI Kim. γ-glutamyl hydrolase modulation and folate influence chemosensitivity of cancer cells to 5-fluorouracil and methotrexate. *British journal of cancer*, 109(8):2175, 2013.

[177] Emman Shubbar, Khalil Helou, Anikó Kovács, Szilárd Nemes, Shahin Hajizadeh, Charlotta Enerbäck, and Zakaria Einbeigi. High levels of γ-glutamyl hydrolase (ggh) are associated with poor prognosis and unfavorable clinical outcomes in invasive breast cancer. *BMC cancer*, 13(1):47, 2013.

[178] Nathaniel Melling, Masoud Rashed, Cornelia Schroeder, Claudia Hube-Magg, Martina Kluth, Dagmar Lang, Ronald Simon, Christina Möller-Koop, Stefan Steurer, Guido Sauter, et al. High-level γ-glutamyl-hydrolase (ggh) expression is linked to poor prognosis in erg negative prostate cancer. *International journal of molecular sciences*, 18 (2):286, 2017.

[179] John de Groot and Harald Sontheimer. Glutamate and the biology of gliomas. *Glia*, 59 (8):1181–1189, 2011.

[180] Leon Gordon Kraft. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949.

[181] Brockway McMillan. Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116, 1956.

[182] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[183] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[184] Yaghout Nourani and Bjarne Andresen. A comparison of simulated annealing cooling strategies. *Journal of Physics A: Mathematical and General*, 31(41):8373, 1998.

[185] René VV Vidal. *Applied simulated annealing*, volume 396. Springer, 1993.

[186] Ronald A Fisher. Combining independent tests of significance. *American Statistician*, 2 (5):30, 1948.

[187] Michael C Wu and Xihong Lin. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Statistical methods in medical research*, 18(6):577–593, 2009.

Appendices

APPENDIX A

Supplementary Information

A.1 Shannon's source coding theorem

Suppose that we have a discrete data source as a Markov process. We have a set of events with probabilities for them to occur as $p_1, p_2, \ldots, p_n$. Given that we have measure H with a following properties:

1. H should be continuous function of p_i,

2. When all p_i are equal for all n, then $p_i = \frac{1}{n}$, and H is a monotonic increasing function of n and describe the uncertainty of events.

3. When an event can be divided into two successive parts, the original H is the weighted of the each values of H. For example, given that we have three events 1, 2, 3 such that the probability $p_1 = 1/2$, $p_2 = 3/10$ and $p_3 = 1/5$ with $H(1/2, 3/10, 1/5)$. The probability to get p_2 or p_3 is $1/2$. Thus, we can reformulate the problem as $H(1/2, 3/10, 1/5) = H(1/2, 1/2) + \frac{1}{2}H(3/5, 2/5)$. Note: we have the coefficient $\frac{1}{2}$ as the second part only occurs half the time, and $P(2|1) = 3/5$ and $P(3|1) = 2/5$.

The only H that satisfies the three properties above is given in the form $H = -K\sum_{i=1}^{n} p_i \log p_i$ where K is a positive constant. The value of K depends only on the choice of measure. Thus, for a random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$ corresponding probabilities $\{p_1, p_2, \ldots, p_n\}$, Shannon formally defined *entropy* as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{A.1}$$

Shannon's source coding theorem (also known as noiseless coding theorem) is proposed by Shannon in Ref. [53] to describes the limits for data compression in the form of entropy. The formal definition of the Shannon's source coding theorem is given below by Theorem A.6.

The Shannon's source coding theorem can informally be stated as: A collection of n *i.i.d.* random variables, each with entropy $H(X)$ can be compressed into more that $nH(X)$ bits with negligible data loss, as $n \to \infty$. No uniquely decodable code word that can be compressed to less than $nH(X)$ without loss of data.

Here, we provide the derivation for the Shannon's source coding theorem. For a set S, we define S^* as the set of all finite sequence of S. A prefix of a word $s = (s_1, s_2, \ldots, s_n) \in S^*$ is in the form of $(s_1, s_2, \ldots, s_i)$ for some $1 \leqslant i \leqslant n$. Consider X be a finite set $x_1, x_2, \ldots, x_n$, and p the probability measure on X. A mapping of $W : X \to \{0,1\}^*$ is call a *code*, and $W(X^*)$ as *code words*. For every $x \in X$, let $W(x)$ be the code word that corresponds to x, and $l(x)$ denotes the length of $W(x)$. W is a *prefix code* if each code word maps to a different non-empty bit string. A prefix code is uniquely decodable. The expected length of W is $L(W) = \sum_{x \in X} p(x)\, l(x)$. The goal is to find a code W that is uniquely decodable and with low score of expected length.

Theorem A.1. Let $W : X \to \{0,1\}^*$ be a prefix code of length $l(x)$, then

$$\sum_{x \in X} 2^{-l(x)} \leqslant 1$$

This is know as *Kraft's inequality* [180, 181].

Theorem A.2. Conversely, if $X = \{x_1, x_2, \ldots, x_n\}$, and we have integer lengths $l_1, l_2, \ldots, l_n$ such Theorem A.1 is satisfied

$$\sum_{i=1} 2^{-l_i} \leqslant 1,$$

then, there exist a prefix code $W : X] \to \{0,1\}$ with property $l(x_i) = l_i$.

Theorem A.3. Theorem A.1 is still true if W is uniquely decodable.

Define *entropy* as $H(p) = -\sum_{x \in X}^{n} p(x) \log_2 p(x)$.

Theorem A.4. Let X be the source equipped with probability measure p. Given that W is uniquely decodable, then the expected length of code words follows $L(W) \geqslant H(p)$.

Proof. let $q(x) := 2^{-l(X)}$. As W is uniquely decodable (Theorem A.3), we have

$$\sum_{x \in X} q(s) \leqslant 1 \tag{A.2}$$

$$\begin{aligned} L(W) &= \sum_{x\in X} p(x)\, l(x) \\ &= \sum_{x\in X} p(x) \log_2 2^{l(x)} \\ &= -\sum_{x\in X} p(x) \log_2 q(x) \\ &= \geqslant -\sum_{x\in X} p(x) \log_2 p(x) \\ &= H\big(p\big) \end{aligned} \tag{A.3}$$

The inequality A.3 follows from *Gibbs' inequality* [182] and Equation A.2. ■

Theorem A.5. Let X be the source equipped with probability measure p. There exists a prefix code word W such that

$$L(W) \leqslant H\big(p\big) + 1$$ □

Proof. For each $x \in X$, let $l(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil$. By applying Theorem A.2, we obtain prefix code for length denoted by l. Then

$$\begin{aligned} L(W) &= \sum_{x\in X} p(x)\, l(x) \\ &\leqslant \sum_{x\in X} p(x) \left(\log_2 \frac{1}{p(x)} + 1 \right) \\ &= H\big(p\big) + 1 \end{aligned} \tag{A.4}$$ ■

Theorem A.6. Let X be a source equipped with probability measure p and $n \geqslant 1$. Equip X^n with a product measure p^n. For every $n \geqslant 1$, there exists a prefix code W_n (i.e. dependent on n), that satisfies

$$H\big(p^n\big) = nH\big(p\big) \leqslant L(W_n) \leqslant nH\big(p\big) + 1$$

Particularly, $\dfrac{L(W_n)}{n} \to H\big(p\big)$ as $n \to \infty$. □

Proof. Based on Theorem A.4 and Theorem A.5, we can obtain for each $n \geqslant 1$, there exists a prefix code word W_n such that

$$H\big(p^n\big) = nH\big(p\big) \leqslant L(W_n) \leqslant nH\big(p^n\big) + 1$$

As $H\big(p^n\big) = nH\big(p\big)$, we can substitute into equation and arrive at the theorem. ■

A.2 Huffman's coding

Huffman coding that was proposed in Ref. [52] is a popular technique to create prefix codes, such that the term become synonym with general implementation of prefix-code data compressions. Shannon's source coding theorem (Theorem A.6) specifies entropy as the smallest possible code word length that are theoretically possible. Huffman's coding objective is to build a prefix code with 'minimum-redundancy' such that that it seeks to achieve the smallest possible average length. Suppose that we have an ensemble code of a finite size n. The average code length is given by

$$L(W) = \sum_{x \in X} p(x)\, l(x)$$

where W is a prefix code as we have defined previously in Appendix A.1. Let the mapping $W(x)$ be the code corresponds to x, with its corresponding length $l(x)$ and probability $p(x)$. It is assumed that the codes in the ensemble is arranged in the order of

$$l(1) \leqq l(2) \leqq \ldots \leqq l(n-1) \leqq l(n)$$

and

$$p(1) \geqq p(2) \geqq \ldots \geqq p(n-1) \geqq p(n)$$

The properties of the ensemble code are:

1. The ensembled code is uniquely decodable, i.e. no codes are identical,
2. No indication are needed to specify where a code begins and ends, once the initial point of the code sequence has been specified,
3. $l(1) \leqq l(2) \leqq \ldots \leqq l(n-1) \leqq l(n)$,
4. For a given size of coding digit D, there exist two but no more than D of the codes with the same code length of $l(n)$ where the codes are similar except for the last digits,
5. The sequence of $l(n-1)$ digits must be utilized either as codes, or as prefixes for codes.

The algorithm for Huffman coding as described in Ref. [52] is as follow: The method will creates binary tree based on the ranks of priority of nodes. First, we create leaf nodes for each message/symbol and arrange them in descending order based on their probabilities. Then we merge two of the nodes the lowest probability into new internal node and combine their probability (i.e. by addition). The new internal node is place in the queue based on its probability. The procedure to merge nodes and placing of internal nodes are repeated until we are left with a single merged node with combined probability of 1. The last merge node is the root of the Huffman tree. Finally, we can trace back from the leaf nodes towards the root, and assign binary digit (either 0 or 1) to the combined members with the root corresponds to the first digit and leaf node for the last digit. One convention that can be used is to assign 1 for message with lower probability and 0 otherwise. We provide a simple example for 4 coded messages in Figure A.1. For a data system with n leaves, the algorithm create binary tree with $2n-1$ nodes. The algorithm complexity is $O\big(n \log n\big)$.

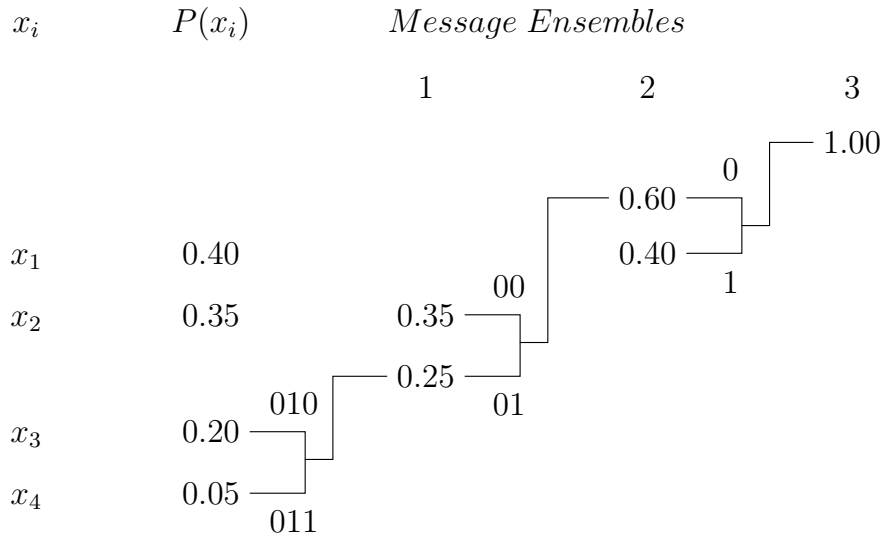


Figure A.1: An example of Huffman tree for 4 coded messages.

Table A.1: Results of Huffman coding based on an example of 4 coded messages.

x_i	$p(x_i)$	$l(x_i)$	Code	$p(x_i)\,l(x_i)$
1	0.40	1	1	0.4
2	0.35	2	00	0.7
3	0.20	3	010	0.6
4	0.05	3	011	0.2
				$L(X) = 1.9$

A.3 Simulated Annealing

Heuristic algorithms are techniques that seek for solutions that is usually good enough but are not guaranteed to be optimal. They usually employs greedy paradigm that often the solutions are trapped in local optimum thus unable to reach global optimal. Metaheuristic algorithms employs higher-level heuristic techniques that allow thorough exploration of solution space. They even allow temporary deterioration of the solutions, to be able to escape local optima to cover better alternative solution space. The solutions are not guaranteed to be optimal but they usually achieve better performance as compared to heuristic approach. Simulated annealing is one such technique that belongs to the metaheuristic methodology.

Simulated annealing approach is analogous to the metallurgical process of physical annealing of metals/alloys. When the materials are melted and allowed to slow cooling and solidify, they will achieve regular crystal lattices state, and thus are free of imperfection. However, if the cooling is not conducted slowly enough, the matters will forms imperfect structures. The simulated annealing algorithm exploits the natural thermodynamic behaviour for adaption in optimization problems as energy states in thermodynamics are analogous to objective functions. The ground state (lowest energy), change between states and temperature in thermodynamics reflects to resemblance to optimal solution, neighbouring solutions and control parameter of simulated annealing respectively. These abstract properties allows the technique to be used in combinatorial optimization problems with many degree of freedom and local optima in order to find minimum (maximum) values of a function.

Simulated annealing consider a combinatorial optimization problem with a solution space G and an objective function $\phi : x \rightarrow \mathbb{R}$. Simulated annealing technique is adapted from Metropolis acceptance criterion that models energy changes in thermodynamic system [183]. The solution space is the set of all the possible solution $G = \{x | x = \{x_1, x_2, \ldots, x_n\}\}$, where n is the number of elements in the system. For a minimization problem, the objective is to find the solution that is equal or close to global optimum $x^* \in G$ such that $\forall x \in G, \phi(x^*) \leqslant \phi(x)$. For a maximization problem, we seek for solution where $\phi(x^*) \geqslant \phi(x)$.

The algorithm is as follows: First, we set initial temperature T at a sufficiently high value (analogous to melting condition) and generate initial solution (either randomly, or by using pre-determined rules). The we propose a new candidate of solution and compare the costs between the two configurations. If the proposed configuration achieve lower cost, the new solution will be accepted. Otherwise, we accept the proposal based on the probability that is associated with current temperature. Then, we lower the temperature (which adaptively based on score rate), and continually proposing new solutions until the system becomes stable ($T = 0$). As the temperature is reducing, the rate of accepting unfavourable new solutions are reduced. We presented the pseudocode of the procedure in Algorithm A.1 for minimization problem.

One of the widely used cooling strategies is exponential schedule,

$$T(t) = T_o \alpha^t \tag{A.5}$$

where for a constant parameter α, $0 < \alpha < 1$, and t is the iteration count. Another popular

alternative is a linear schedule,

$$T(t) = T_o - \beta t \tag{A.6}$$

for a constant factor β. These two strategies been reviewed as being widely used temperature cooling techniques [184].

Simulated annealing has a primary weakness that it requires suitable adjustment of initial temperature and annealing schedule. Although it can be costly, the process provides insight of the system behaviour especially near the optimal configurations [185]. The technique also is not as fast as greedy methods, but it remains popular to provide near optimal solutions for difficult combinatorial problems.

Algorithm A.1 Simulated Annealing Pseudocode

1: **procedure** SIMANNEAL(n, N_{max}, T)
2: $x_{now} \leftarrow$ InitialConfig(n) ▷ Select some $x \in G$
3: $x_{optimal} \leftarrow x_{now}$
4: $T_{now} \leftarrow T$
5: **for** $k = 1, \ldots, N_{max}$ **do**
6: $x_{can} \leftarrow$ NeighbourConfig(x_{now}, G) ▷ Propose candidate for x
7: $T_{now} \leftarrow$ AdaptiveTemp(k, T_{now}) ▷ Set adaptive temperature
8: **if** $\phi(x_{can}) < \phi(x_{now})$ **then**
9: $x_{now} \leftarrow x_{can}$
10: **if** $\phi(x_{now}) < \phi(x_{optimal})$ **then**
11: $x_{optimal} \leftarrow x_{now}$
12: **else if** $\exp\left\{\dfrac{\phi(x_{now}) - \phi\left(x_{optimal}\right)}{T_{now}}\right\} >$ Random$(0,1)$ **then**
13: $x_{now} \leftarrow x_{can}$
14: **return** $x_{optimal}$

A.4 Over-representation analysis

Over-representation analysis (ORA) is a method that determines if a set of pathways are present more than statistical expected in a subset of data. The method measures percentage of overlapped molecules, and in our case i.e. metabolites, that are differentially expressed in any group of biological pathways. The null hypothesis is that the metabolites in the data set do not have stronger association with the pathways under study as compared to other metabolites. The most common statistical tests are Fisher's test [186] (as conducted by IMPaLA [131]), hypergeometric distribution test (as conducted by MBROLE [94]), χ^2-square test and binomial proportions z-test. In practice, the choice of test is deemed not important [187]. The results can be represented as list of biological pathways that are arranged by the most relevant entries, and are ordered with respect to p-value.

Here, we describe the procedure to calculate p-value based on Fisher's test. First, from the list of molecules (i.e genes, metabolites), we represent the set of differentially expressed molecules as E and the molecules in pathways of interest as S. Set E^c and S^c (i.e. the complement) as the molecules that are not differentially expressed and present in the pathways respectively. Define N as total number of molecules, and $n(\cdot)$ as the function for cardinality of set. Then, we can construct a 2×2 contigency table on the association of molecules in E and S as shown in Table A.2.

Table A.2: Contigency table for over-representation analysis.

	In molecule set	Not in molecule set	
Diff. expressed	$n(S \cap E)$	$n(E \cap S^c)$	$n(E)$
Not diff. expressed	$n(S \cap E^c)$	$n(E^c \cap S^c)$	$n(E^c)$
	$n(S)$	$n(S^c)$	N

Next, we can calculate p-value for over-representation analysis to test the association of molecules in D and S. The p-value based on Fisher's method is given by

$$p = 1 - \sum_{i=1}^{n(S \cap E)} \frac{\begin{pmatrix} n(E) \\ i \end{pmatrix} \begin{pmatrix} n(E^c) \\ n(S) - 1 \end{pmatrix}}{\begin{pmatrix} N \\ n(S) \end{pmatrix}} \tag{A.7}$$