M.S. THESIS

An Information Theoretic Algorithm for Mining and Ranking Phenotype-specific Sub-networks from Multi-class Gene Expression Data

다중 클래스 유전자 발현 데이터에서 표현형 특이적 서브 네트워크 발굴 및 랭킹을 위한 정보 이론 기반 알고리즘

FEBRUARY 2017

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Park Jinwoo

M.S. THESIS

An Information Theoretic Algorithm for Mining and Ranking Phenotype-specific Sub-networks from Multi-class Gene Expression Data

다중 클래스 유전자 발현 데이터에서 표현형 특이적 서브 네트워크 발굴 및 랭킹을 위한 정보 이론 기반 알고리즘

FEBRUARY 2017

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Park Jinwoo

An Information Theoretic Algorithm for Mining and Ranking Phenotype-specific Sub-networks from Multi-class Gene Expression Data

다중 클래스 유전자 발현 데이터에서 표현형 특이적 서브 네트워크 발굴 및 랭킹을 위한 정보 이론 기반 알고리즘

지도교수 김 선

이 논문을 공학석사학위논문으로 제출함

2016년 11월

서울대학교 대학원

컴퓨터공학부 (학과 및 전공)

박 진 우 (논문 작성자)

박진우의 석사학위논문을 인준함

2016년 12월

위 원 장 Srinivasa Rao Satti 교수님 (인)

부 위 원 장 김선 교수님 (인)

위 원 Bernhard Egger 교수님 (인)

Abstract

An Information Theoretic Algorithm for Mining and Ranking Phenotype-specific Sub-networks from Multi-class Gene Expression Data

Park Jinwoo
Department of Computer Science and Engineering
College Of Engineering
Seoul National University

There have been extensive studies for inferring transcriptional network from omics data. However, how to utilize networks for specific research projects has not been well established. One of the main hurdles is lack of algorithms for mining biological sub-networks. Existing graph mining algorithms do not consider features of the transcriptional network and they are not effective to obtain biologically meaningful results. In this paper, we define the biological sub-network mining problem and present a new graph mining algorithm that mines and ranks phenotype specific sub-networks of transcriptional regulatory networks constructed from multi-class gene expression data. Our contributions in this paper on the computational side are two folds. First, we suggest a complete research paradigm of utilizing omics data to construct networks and

then elucidates sub-networks that distinguish phenotypes or disease states. Second, we developed an information theoretic algorithm for mining phenotype specific sub-networks. Our contribution on the bio/medical side is that our TF-module based analysis determined biological pathways(cell cycle: M-phase, cell adhesion molecules) related to the phenotype(breast tumor grade) by identifying activation/suppression of specific target genes (TGs) by the combination of multiple transcription factors (TFs). Expression levels of TGs clearly shows correlation between activation/suppression of these pathways and tumor grades. When we used all genes, pathway activation or suppression was not obvious, which shows the effectiveness of our algorithm. Our TF-centric pathway activation/suppression analysis technique is applicable to and useful for many other studies.

Keywords: Transcriptional regulatory network, network mining, subnetwork

Student Number: 2012-23214

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sequencing and bio-assay techniques have been developed dramatically over the years and it has become a routine practice to measure whole genome level transcriptome data under specific phenotypes such as high grade breast cancer subtypes. Thus, characterizing phenotype specific biological mechanisms is now one of the most important research problems in bioinformatics. Biological mechanisms underlying specific phenotypes need to consider complex relationship among genetic elements that distinguish phenotype specific characteristics. One of the most effective computational techniques to consider complex relationships of all genes in the whole cell is to use networks. In fact, network based characterization of biological systems has been successful for more than a decade and it is now considered as the gold standard method to understand complex biological systems. Barabasi, et al. [35] have been pioneering use of networks for characterizing protein interactions in yeast and

further developed the technique for years to characterize biological mechanisms including human disease networks [12]. As genetic and protein data rapidly accumulate, research efforts to build and utilize genetic networks have been successful for specific species such as yeast [13], C. elegance [19], Arabidopsis [20], and mouse and man [27]. In succession, there has been significant research efforts on omics data to construct and utilize networks under specific conditions. Segal, et al. [30, 31] developed computational analysis techniques to construct network modules from gene expression data in specific conditions, showing conditional activity of expression modules in cancer [29]. This line of research further targeted to develop computational methods to construct transcriptional regulatory networks under specific conditions. Pioneering work by Califano, et al. developed ARACNe [2], an algorithm for the reconstruction of accurate cellular networks, from gene expression data and further showed that ARACNe could reveal tumor suppressor gene RUNX1 in T cell acute lymphoblastic leukemia[9]. Luonan, et al. developed NARROMI [37], a noise and redundancy reduction technique to infer transcriptional regulatory networks, from gene expression data and also further showed that an extended version of NARROMI could reveal dynamical network biomarkers as early-warning signals for type-2 diabetes.

Although transcriptional regulatory network based analysis of omics data has been successful to unveil underlying mechanisms for many diseases and species, the transcriptional regulatory network based

analysis techniques have been used by only a handful of research groups. There are two major reasons for the limited practice of the transcriptional regulatory network based analysis. First there is no complete research paradigm of utilizing omics data to construct networks and then elucidating sub-networks that distinguish phenotypes or disease states. Second, in this line of argument, a major missing technique is how to mine and rank phenotype-specific sub-networks from big networks constructed by utilizing omics data and other evidences. Although mining sub-networks has been extensively studied in the field of computer science, mining and ranking biological networks should utilize characteristics of the biological research problems under consideration. There are only a few biological sub-network mining algorithms. For example, Bhowmick, et al. [28] developed FUSE, a profit maximization approach for functional summarization of biological networks that can be useful to mine sub-networks. However, FUSE is not originally designed to mine phenotype specific sub-networks, for which we will define the computational problem and present an information theoretic algorithm for the suggested problem in this paper. (depicted in Figure 1.1)

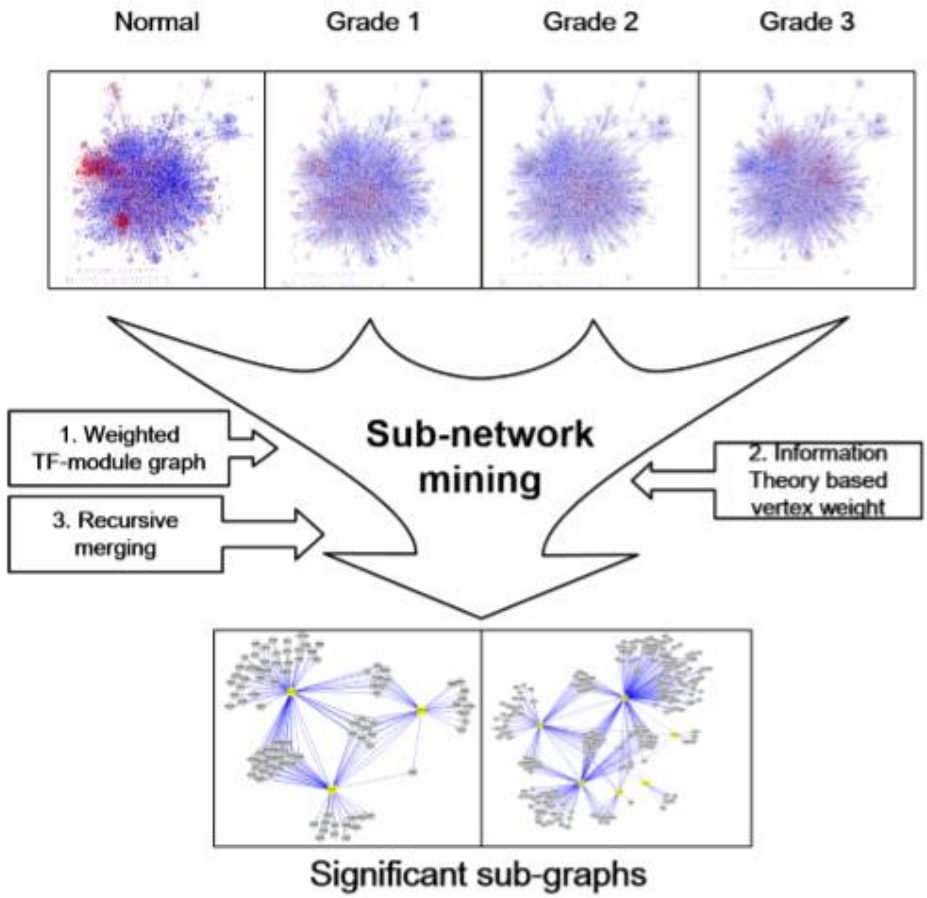

Figure 1.1 Overview. First, TF-TG network topology was generated by NARROMI package with gene expression matrix of 982 breast cancer samples. Then, we mapped average z-score of each gene for 4 classes(Normal, Grade 1, Grade 2, Grade 3) to make multi-valued attribute nodes. For the generated input graph, we applied our algorithm to discover phenotype-specific sub-networks.

Chapter 2

PHENOTYPE SPECIFIC SUB-NETWORK MINING PROBLEM

2.1 Biological network construction methods

Since inferring accurate transcriptional regulatory network (TF-TG network) has been an important issue over the years, numerous useful network construction methods using gene expression data were developed. According to a recent review [10] of the reverse engineering methods from observational expression data, which is predominating data type, the current methods can be classified into two subgroups - correlation based [8], mutual information based [2, 37]. We just picked one of the widely used methods, NARROMI among them, to construct a TF-TG network using omics data.

In recent, Barabasi, et al. [1] suggested a method to improve the inferred networks by silencing indirect edges and achieved more than 50% and 6% predictive improvements for correlation based methods

and mutual information based methods, respectively.

2.2 Necessity of biological sub-network mining algorithm

Once we have a huge biological network constructed from omics data, the next step is to mine sub-networks that discriminate phenotypes. However, we showed that general network mining algorithms were not suitable for this purpose.

According to the survey on graph mining algorithms developed for biological context [24], there were three categories: Tree mining, Frequent sub-graph mining, and Module mining. First of all, none of them considered TF-TG relationships. Tree mining algorithms could be applied only to the trees. Also, Frequent sub-graph mining algorithms are for graphs with different topologies. Finally, most of the Module mining algorithms were developed for gene-gene co-expression network or protein-protein interaction network based on highly connected sub-graphs (HCSs) mining. However, HCSs mining approaches were not effective for TF-TG graph, when we tried to use the algorithm (section 4.2).

2.3 Problem formulation

For the K-class data, there is a TF-TG graph with multi-valued nodes denoted by $G=(V,E,W)$, where V, E and W denote vertex set, edge set and attribute vector assigning function for a node,

respectively. Vertex set V is a union of TF gene set V^{TF} and non-TF gene set V^{non-TF}. W assigns every node v_i in G an attribute vector a_i: $<(\overline{z_i^1},p^1),...,(\overline{z_i^K},p^K)>$, where $\overline{z_i^K}$ denotes average z-score of expression value corresponds to their phenotype p^k. The computational problem here was defined to be how to determine sub-networks that differentiate phenotypes "quantitatively". This issue has not been discussed so far.

2.4 Our information theoretic algorithm

We defined crucial sub-network for explanation of phenotypes by using information theory (depicted in Figure 2.1). What we focused on was sub-networks that have lowest entropy in terms of gene expression values and phenotypes. See Figure 2.2 for an example of calculating entropy using gene expression values and phenotypes. The algorithm proceeded as follows.

1. Build a TF-TG graph with multi-valued node attributes using omics data.
2. Define initial TF-modules. Each of the initial TF-modules is a sub-network which consists of a TF and its TGs.
3. Define size-weighted information scores of TF-modules.
4. Build a hyper graph where nodes are initial TF-modules and edge is defined from TF-TG graph. Edge weights are determined by the number of genes shared in two TF-module nodes.
5. Determine sub-networks by merging TF-modules progressively.

When two TF-modules are considered for merging, the post-merging information score should be higher than the information scores of two TF-modules.

6. Rank sub-networks with respect to their information scores.
7. Map TGs in the sub-networks to biological pathways to investigate how phenotype distinguishing sub-networks contribute to changes in biological pathways.

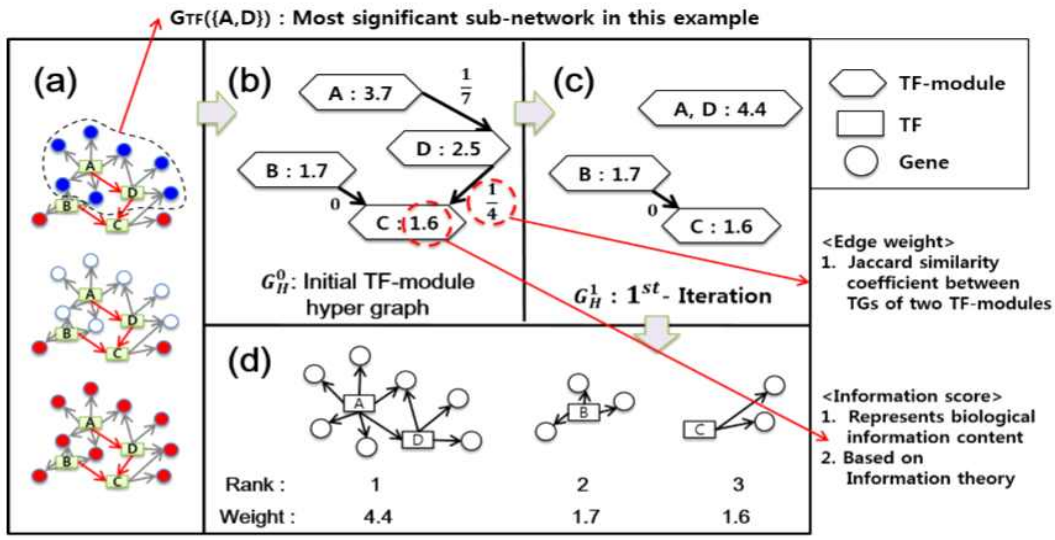

Figure 2.1 Flow of the algorithm. (a) A TF-TG graph with multi-valued attribute nodes that was inferred from NARROMI using gene expression omics data (b) TF-TG graph is transformed to a weighted TF-module hyper graph consists of TF-modules. Every pair of two TF-modules linked by an edge is a candidate of merging-procedure, with its edge weight as a measure of the priority. From the top priority pair, merge two TF-modules if an information score after merging is higher than each of original two TF-modules. (c) Once merging is occurred, TF-module hyper graph of the next iteration is generated. (d) If there is no candidates(edges), remaining TF-modules are ranked by their scores.

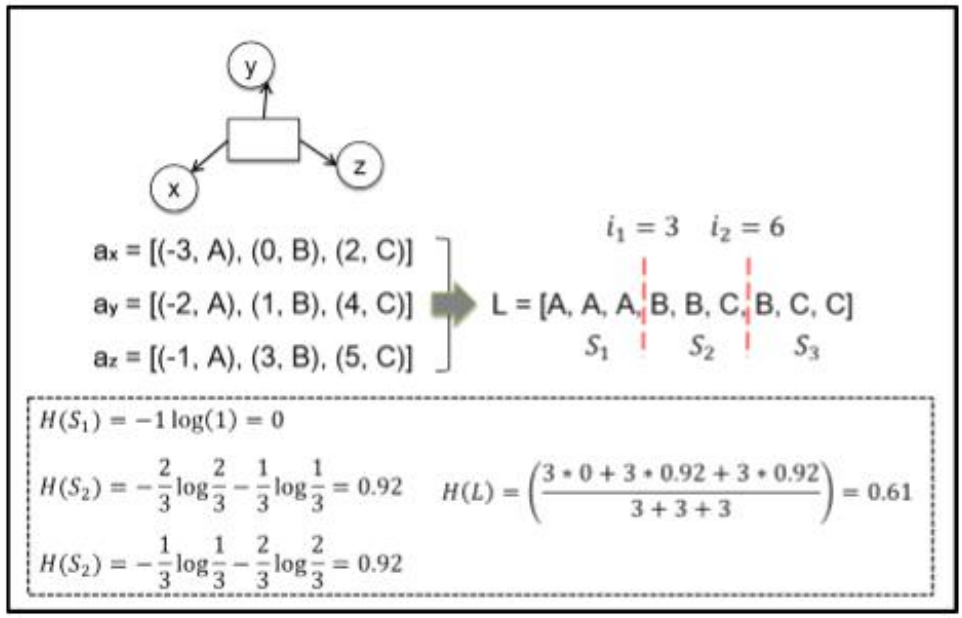

Figure 2.2. Process for calculating entropy of a TF-module for a particular index set I = {3,6}.

Chapter 3

Methods

3.1 TF-TG network construction

3.1.1 Edge set

Here we generated breast tumor specific network topology from expression data of 982 tumor samples without 144 normal samples (Details are in section 4.1). NARROMI [37], a Matlab package to infer transcriptional regulatory network from TF and TG expression matrix, was adopted here in order to define edge set E. In brief, NARROMI called a list of TFs for each gene to generate candidate regulatory TFs based on mutual information, followed by recursive optimization algorithm for the removal of redundancies. Then, scores based on recursive optimization algorithm and mutual information values are mediated by a linear formula. Output from NARROMI consists of TF-TG relations, and their corresponding strength and significance in terms of coefficient and p-value, respectively. The

number of edges resulting from our data was 6312163. However, many of the resulting edges from NARROMI have very high potential to be false positives when it comes to biological relevance, we filtered out edges that have p-values greater than 1×10^{-32} or absolute coefficients less than 0.4. As a result, we retained 17316 edges to construct transcriptional regulatory network.

3.1.1 Multi valued attribute vector

For individual gene, we employed z-score normalization of expression values from 1126 samples of 4 phenotypes. First, for each gene vi, expression values with their phenotypes of all samples $< (x_i^1, p^1), ..., (x_i^{1126}, p^K) >$ were transformed to z-score vector $< (z_i^1, p^1), ..., (z_i^{1126}, p^K) >$. Then, for phenotype p^k, average z-score $\overline{z_i^k}$ was assigned. Finally, attribute vector of v_i was defined as a_i: $< (\overline{z_i^1}, p^1), ..., (\overline{z_i^K}, p^K) >$.

3.2 Information scores for TF-modules

3.2.1 Definition of TF-module

For a set of TFs, ST F , we defined a TF-module of the set as a sub-network $G_{TF}(S_{TF}) \in$ G consists of both S_{TF} and their TGs. As we assumed that all TGs of a TF share connected roles in biological pathways, we were focusing on finding the sub-networks in the form of TF-modules

3.2.2 Problem formulation

If z-score values of TGs in phenotype A are significantly different to phenotype B, a biological pathway related to the TF-module would be signatures for the phenotypes. To measure how the TGs in TF-module have phenotype-wise distinctive z-score values, we used an entropy based approach.

First, we joined all n attribute vectors of all TGs, $a_i,...,a_n$ to the list $[(\overline{z_i^1},p^1),...,(\overline{z_i^K},p^K)]$ and sorted them by the z-score values z. We extracted only phenotypes from the list to get phenotype list, which were sorted by their z-score values, L = $[p_1,...,p_{nK}]$. If L is split to K sets, we could calculate an entropy value of L. With a split index list $I=[i_1,...,i_{K-1}]$, where $i_0=0, i_K=nK$. j^{th} set could be expressed as S_j = $L[i_{j-1}:i_j]$. Then entropy of a set S_j could be calculated by Equation (1).

$$H(S_j) = -\sum_{K=1}^{K} P_j(p^k)\log P_j(p^k) \tag{1}$$

where $P_j(p^k)$ denotes proportion of p^k in S_j. Then, a weighted average entropy of L could be obtained.

$$H(L) = \frac{\sum_{j=1}^{K} N_j \times H(S_j)}{\sum_{j=1}^{K} N_j} \tag{2}$$

where N_j denotes size of the S_j.

3.2.3 Best entropy with dynamic programming

For a TF-module with n TGs, there are ${}_{nK-1}C_{K-1}$ possible split index lists. To calculate entropy for all these split index lists would be very time-consuming, especially for large K and n. Note that all TGs need to be considered.

Thus, we used a dynamic programming approach, with following recurrence relation, where $H_t(L)$ stands for the best entropy of L with t cuts (t + 1 sets).

$$H_t(L[i:nK]) = \max_j \left\{ \frac{j-i}{nK-i} \times H_{t-1}(L[i:j]) + \frac{nK-j}{nK-i} \times H_1(L[j:nK]) \right\} \quad (3)$$

First, calculate $H_0(L[i:nK])$ for all $i = \{1,\ldots,nK-1\}$I. Then, $H_t(L[i:nK])$ values could be calculated for *t=1,…,K−1* in turn. Through this, just $nK(K-1)$ tries are needed. Best entropy value of L, $H_K(L[0:nK])$ could be calculated using $H_{K-1}(L[i:nK])$ values below.

$$H_K(L[0:nK]) = \max_i \left\{ \frac{nK-i}{nK} \times H_{K-1}(L[i:nK]) + \frac{i}{nK} \times H_1(L[0:i]) \right\} \quad (4)$$

3.2.4 Information score for TF-module

Finally, in order for higher score to indicate enhanced goodness we transform entropy value to information content using Hmax = log K. Also, we normalized the score by the number of TGs(n). Thus, the

information score function of the TF-module WV was defined as below,

$$W_V(G_{TF}) = (H_{\max} - H_K(L)) \times \log(n) \quad (5)$$

3.3 TF-module hyper graph

Let G_H^j be a weighted TF-module based hyper-graph from original TF-TG graph G at the j's iteration. Each node in G_H^j is a sub-graph in form of TF-module with their entropy based score defined in Section 3.2. Nodes in G_H^j are iteratively merged until there remains no candidates (Details are in Section 3.4).

G_H^j was defined as (V_H^j, E_H^j, W_V, E_V), where V_H^j, E_H^j, W_V and $E_V$are vertex set, edge set, TF-module score function and edge-weight-assigning function, respectively. E_H^j is determined by TF-TG graph G. For two TF-modules u_x, $u_y \in V_H^j$, edge between them is defined only if there exists two TFs v_i, $v_j \in$ G that satisfy $v_i \in u_x$, $v_j \in u_y$ and [$(v_i, v_j) \in$ E or $(v_j, v_i) \in$ E]. W_E assigns a weight to the edge using Jaccard similarity coefficient regarding common TGs of two TF-modules.

$$W_E(u_x, u_y) = \frac{S_{TG} \cap S'_{TG}}{S_{TG} \cup S'_{TG}} \quad (6)$$

where, each S_{TG} and S'_{TG} is a set of all TGs of u_x and u_y, respectively. Thus, the closer to 1 the edge score is, the higher the two TF-modules have possibility to be involved in a common biological pathway.

Initialization. TF-modules of each TF in TF-TG graph G were defined as initial vertex of G_H^0

$$V_H^0 = \{G_{TF}(\{Vi\}) | v_i \in V^{TF}\} \quad (7)$$

3.4 Merging of TF-modules on hyper graph

For hyper graph Gj H at the j^{th}-iteration, every pair of two connected nodes is considered for merging. Edge weight indicating the ratio of common TGs between two TF-modules is used to define merging priority. For the candidate pair of two TF-modules (u_x, u_y), merging is performed if $W_V(u_x \cup u_y) \geq \max(W_V(u_x), W_V(u_y))$. Merging of ($u_x$, u_y) generates (j + 1)th-iteration of hyper-graph G_H^{j+1} = (V_H^{j+1}, E_H^{j+1}, W_V, E_V), where $V_H^{j+1} = V_H^j \cup \{u_z\} - \{u_x\} - \{u_y\}$ (u_z is a merged TF-module of u_z and u_y). Starting from the initial hyper-graph G_H^0, the iterative merging procedure continues until there is no candidates.

Chapter 4

Result and Discussion

4.1 Raw biological data

From METABRIC [7], 982 breast primary tumor samples with their clinical information were used in this study. We chose tumor grade information (grade 1, 2, and 3) for labelling the class to our data which in turn located 68, 408, and 506 samples for each grade(1, 2, and 3), respectively. Grade is defined histologically and it represents “aggressiveness” of the tumor. The data contained 25228 genes and 1396 of them were TFs. Normal class with 144 samples was also used

4.2 HCS mining algorithm

Though HCS mining algorithm was not intended for TF-TG graph, we tested whether it could make meaningful results with our TF-TG graph. We screened all sub-graphs with higher density than the

cut-off (=0.9). As a result, 65 sub-networks of size 12 turned out to be the largest sub-networks. However, functional annotation of each subnetwork by DAVID [14, 15] showed only DNA binding as significant, which meant that the resulting sub-networks consisted of mostly TFs and severely lacked their own biological significance. This was because TFs were hubs in the network. In addition, no significant KEGG [16, 17] pathway was detected. Lack of the biological meaning was because the density-based approach did not consider most of TGs due to their low degrees.

Size of the sub-network	Count
5	170165
6	100561
7	52340
8	51841
9	47071
10	18833
11	2489
12	65
x >= 13	0

Table 4.1. Result of a HCSs mining algorithm (cut-off=0.9)

4.3 TF-centric sub-network mining algorithm

Our algorithm detected 732 TF-modules. Since there were 901 TF-modules in G 0 H, merging was performed 169 times. Scores of TF-modules are depicted in Figure 4.1. Top ranked TF-modules had well separated TG expression pattern between classes (Figure 4.2). Table 4.2 shows top 10 TF-modules with brief explanation. The DAVID functional analysis showed that, rank-1 and rank-3 TF-modules were associated with cell cycle and rank-2 TF-module

was related to cell adhesion molecules. We further studied these top 3 TF-modules to investigate whether their TG expression patterns and biological function were relevant to the phenotypes.

Rank	TF list in a TF-module	score	# TG	TG expr trend	DAVID cluster
1	PTTG1, UHRF1, CDC2	10.85	94	Up	Cell cycle
2	LDB2,EBF3,EBF1, PPARG, NR5A2, FOSB	8.62	224	Down	Signal
3	PLK4	8.10	45	Up	Cell cycle
4	ZNF6, ZNF483	7.98	90	Down	DNA repair
5	ATOH8	7.61	48	Down	Carbohydrate binding
6	FOXM, TEAD4	7.44	94	Up	Cell cycle
7	ZNF683,PRIC285, STAT1	7.15	127	Up	Signal
8	ZNF394, ZNF773	6.96	76	Down	Channel activity
9	ZNF639, ZNF528	6.79	55	Down	Signal
10	SNAPC1, ZNF577	6.76	114	Down	DNA repair

Table 4.2. Top 10 TF-modules. “Up" in the TG expression trend indicates that overall TG expressions in a TF-module are in increasing order of Normal, Grade 1, Grade 2 and Grade 3.

Term	Count	P-Value
GO:0000279 M phase	46	2.50E-53
GO:0007049 cell cycle	57	5.62E-53
GO:0022403 cell cycle phase	48	4.49E-52

Table 4.3. DAVID functional annotation chart of the rank 1 sub-network G_{TF}({PTTG1, CDC2, UHRF1}).

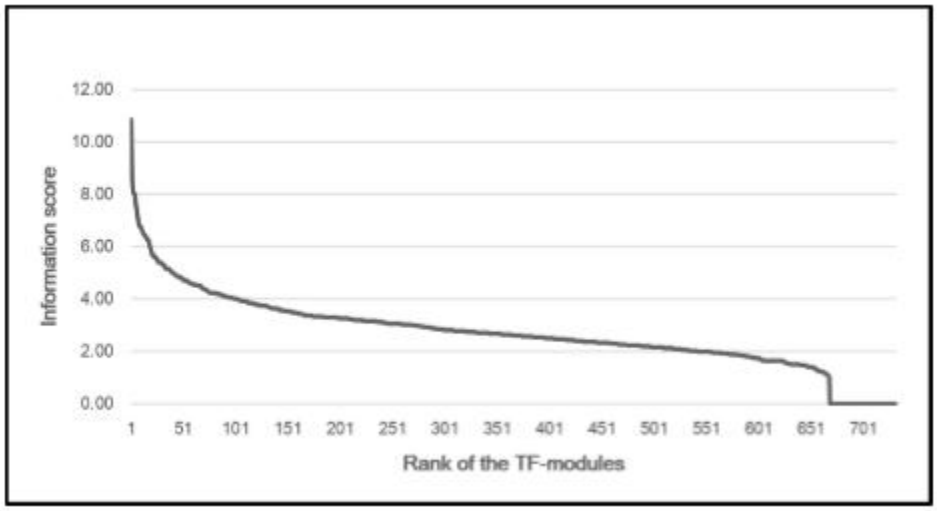

Figure 4.1: Information scores of the TF-modules.

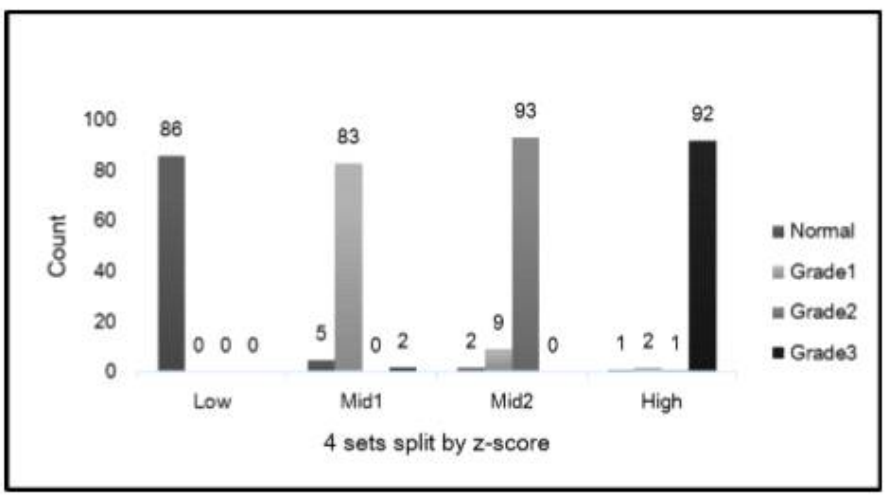

Figure 4.2. Split result of the phenotype labels for a rank-1 TF-module. All the phenotype labels in a rank-1 TF-module were split into 4 groups corresponds to their average z-score values.

4.4 Cell cycle: M-phase

The DAVID functional analysis showed that, 46 genes of 94 TGs and 23 genes of 45 TGs were mapped to the Mphase term for rank-1 sub-network G_{TF}({PTTG1, CDC2, UHRF1}) and rank-3 sub-network G_{TF}({PLK4}), respectively. Also, cell cycle KEGG pathway was mapped as the most significant for both TF-modules. Most TGs of the two TF-modules showed increasing expression pattern from normal to the highest tumor grade as could be seen in Figure 4.3. Biological relevant from the result were able to be confirmed by studies showing that the correlation between high tumor grade and up-regulation of the cell cycle genes [18]. Especially, Sotiriou, et al. [32] found that most of 97 grade associated genes from gene expression profiling analysis between grade 1 and grade 3 breast tumor had higher expression value in grade 3 and related to the cell cycle progression and proliferation.

We focused on 15 genes : BUB1, CCNA2, CCNB1, CCNB2, CCNE2, CDC20, CDC25C, CDC6, CHEK1, E2F2, MAD2L1, MCM2, MCM6, PKMYT1, TTK, TGs from TF-module rank 1 and 3 mapped to KEGG cell cycle pathway. Literature showed expression level of most of 15 genes were correlated to the breast tumor grade. Yuan, et al. [36] found increase of expression level for the check point genes including BUB1, CDC20, MAD2L1, TTK in high grade breast tumor cells. Also, CCNA2, CCNB1, CCNB2, CCNE2 were overexpressed in grade 3 breast tumor compared to grade 1 breast tumor [32]. Overexpression of the cell cycle associated genes includes BUB1,

CCNB2, CCNE2, CDC6,MAD2L1, PKMYT1 were detected in poor prognosis among breast tumor patients [34].

In addition, there were studies of each 3 TFs in rank-1 TF-module that showed relations in the role of a TF on mitosis. Ogbagabriel, et al [23] found that over expression of the securin, the protein encoded by PTTG1 gene, in high mitotic activity tumors from both western blot and northern blot analysis. In addition, Marangos, et al. [22] revealed that securin is the regulator for entry into M-phase. CDC2(or CDK1) is a very well-known gene that regulates the cell cycle. Cyclin B binds to CDK1 and Cyclin B/CDK1 complex regulates the progression into M phase. Finally, Li, et al. [21] suggested that UHRF1 played an important role in G2/M progression from the UHRF1 knockdown cell analysis.

4.5 Cell adhesion molecules

The second highest score was assigned to the G_{TF}({LDB2, PPARG, EBF3, NR5A2, EBF1, FOSB}). Corresponding TGs were mapped to KEGG pathways, and Cell Adhesion Molecules(CAMs) was shown to be the most highlighted. There have been many studies concerning both general features of cancer and CAMs (especially E-cadherin) [3, 4]. In this study, selected TGs from mined sub-network without reflecting biological relevance were proven to be correlated with breast cancer formation and progression of which several studies were previously introduced.

10 TGs were included in CAMs of KEGG: CD34, CDH1, CDH5,

CLDN5, CLDN11, ESAM, ICAM2, JAM2, JAM3, PECAM1. Loss of both CD34 and CDH1 genes were frequently observed as the grade of breast cancer become worse [5], [11]. Lower expression level of ICAM2 appeared in breast tumors compared to that of normal breast tissue [26] and down-regulated JAM2 and JAM3 expressions were detected in p53-mutated breast cancer [6]. According to a study [25], comparison of invasive (n = 7) with non-invasive cases (n = 37) showed 16q loss where a number of cadherin family including CDH1 and CDH5 were situated. Another study [33] was conducted focusing on claudin family and they demonstrated that CLDN5 was highly expressed in endothelial cells whereas reduced or no CLDN5 expression was detected in breast tumor cells.

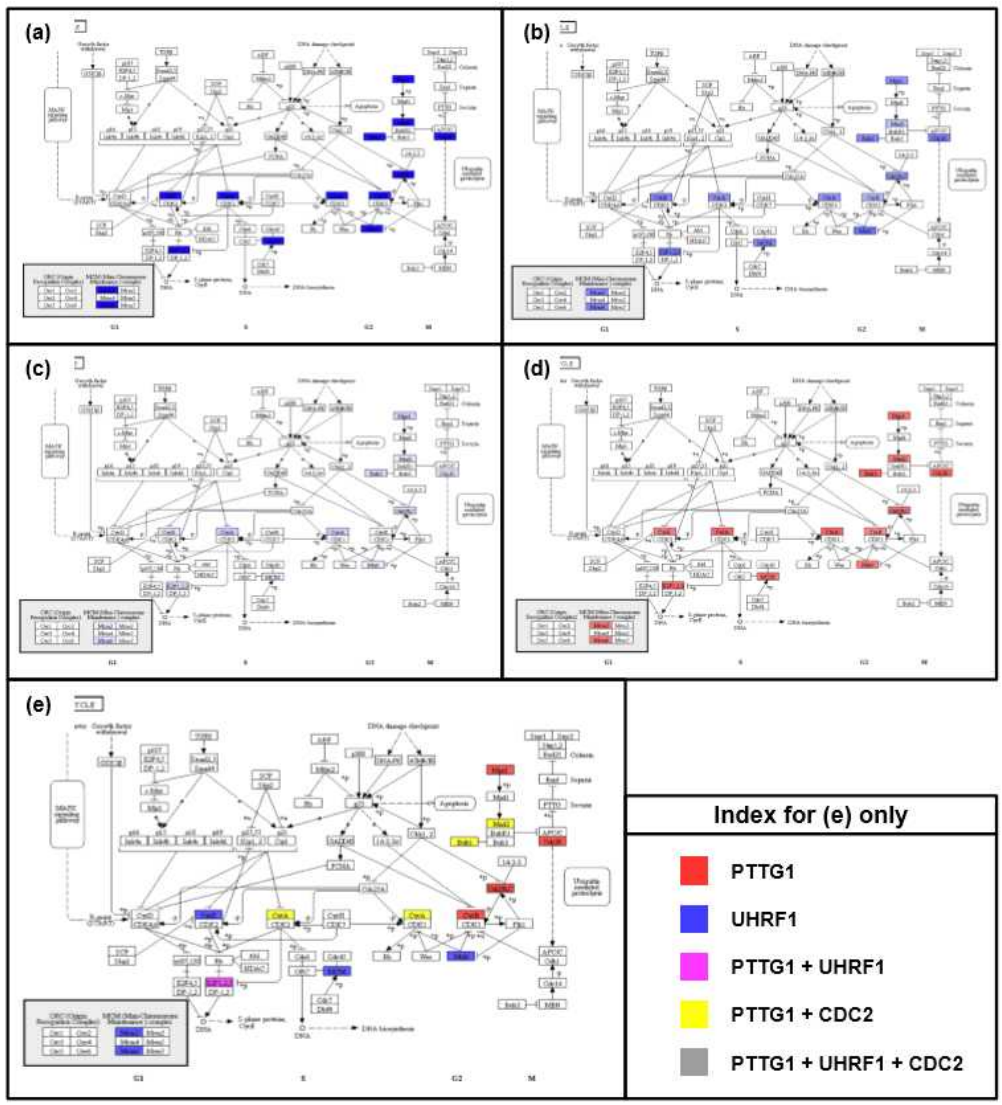

Figure 4.3. TGs of rank-1 sub-network G_{TF}({PTTG1, CDC2, UHRF1}) mapped cell cycle pathway. (a),(b),(c),(d) Colored by expression value of Normal, Grade 1, Grade 2 and Grade 3 - Red: up-regulated, Blue: down-regulated. (e) Colored by regulating TFs. This shows activation of cell cycle pathway for higher grade tumors, especially activation of M-phase, by three TFs. This clear illustration is not evident when we consider all genes in the cell cycle pathway, which demonstrates the eff ectiveness of our TF-module based approach.

Chapter 5

Conclusion

We developed a novel information theoretic algorithm for mining and ranking phenotype-specific sub-networks from multi-class gene expression data. Use of information theory on the multi-valued graph was successful to mine sub-networks that distinguish phenotypes, breast tumor grades in this paper. Sub-networks that involve multiple TFs were constructed by building a TF-module hyper graph and merging TF-modules progressively.

The algorithm was tested by using METABRIC breast tumor expression data with tumor grades for class labels and it successfully inferred breast tumor specific TF-TG networks that distinguish breast tumor grade. Top three TF-modules corresponded to cell cycle and cell adhesion molecules pathways. Activation of cell cycle pathway and suppression of cell adhesion molecules pathway in cancer are well known. Our study provides much more detailed information than previous studies in two ways. First, our analysis factored out important TGs that were consistent with phenotypes.

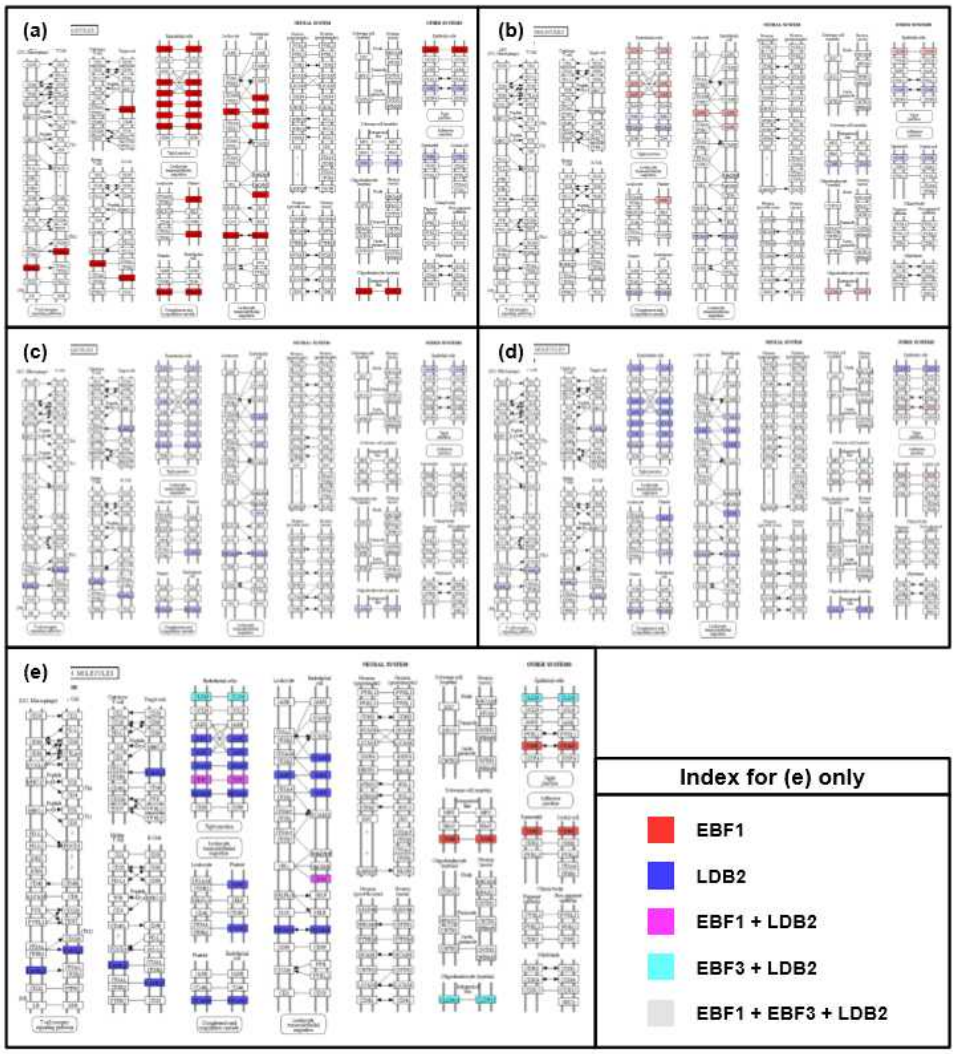

Figure 4.4. TGs of rank-2 sub-network G_{TF}({LDB2, PPARG, EBF3, NR5A2, EBF1, FOSB}) mapped cell adhesion molecules pathway. (a),(b),(c),(d) Colored by expression value of Normal, Grade 1, Grade 2 and Grade 3 - Red: up-regulated, Blue: down-regulated. (e) Colored by regulating TFs. This shows activation of cell adhesion molecules pathway for higher grade tumors by three TFs. This clear illustration is not evident when we consider all genes in the cell adhesion molecules pathway, which demonstrates the effectiveness of our TF-module based approach.

Second, our analysis provided how these pathways that are important in cancer were controlled by a specific set of TFs. In other words, our algorithm shows which genes are activated or suppressed in pathways and also shows that transcriptional control mechanisms of these genes in the pathways, which is novel compared to previous studies.

As we formulated and proposed the new-class sub-network mining problem, there remain a number of research problems. First, we plan to develop an algorithm that mines "maximum-size"TF-modules with multiple TFs, rather than relying on the hyper-graph based progressive merging. Subgraph mining algorithm without class labels, unsupervised learning would be meaningful because vagueness of the class label in many cases. In addition, algorithm that utilizes all the samples, not just average values remains as a future work.

Bibliography

1. B. Barzel and A.-L. Baraba´si. Network link prediction by global silencing of indirect correlations. Nature biotechnology, 2013.
2. K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. Nature genetics, 37(4):382 – 390, 2005.
3. G. Berx, A. Cleton-Jansen, F. Nollet, W. De Leeuw, M. Van de Vijver, C. Cornelisse, and F. Van Roy. E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. The EMBO journal, 14(24):6107, 1995.
4. U. Cavallaro and G. Christofori. Cell adhesion and signalling by cadherins and ig-cams in cancer. Nature Reviews Cancer, 4(2):118 – 132, 2004.
5. H. Chauhan, A. Abraham, J. Phillips, J. Pringle, R. Walker, and J. Jones. There is more than one kind of myofibroblast: analysis of cd34 expression in benign, in situ, and invasive breast lesions. Journal of clinical pathology, 56(4):271 – 276, 2003.
6. D. Coradini, M. Fornili, F. Ambrogi, P. Boracchi, and E. Biganzoli. Tp53 mutation, epithelial-mesenchymal transition, and stemlike features in breast cancer subtypes. BioMed Research International,

2012, 2012

7. C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012.
8. A. De La Fuente, N. Bing, I. Hoeschele, and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics, 20(18):3565–3574, 2004.
9. G. Della Gatta, T. Palomero, A. Perez-Garcia, A. Ambesi-Impiombato, M. Bansal, Z. W. Carpenter, K. De Keersmaecker, X. Sole, L. Xu, E. Paietta, et al. Reverse engineering of tlx oncogenic transcriptional networks identifies runx1 as tumor suppressor in t-all. Nature medicine, 18(3):436–440, 2012.
10. F. Emmert-Streib, G. Glazko, R. De Matos Simoes, et al. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Frontiers in genetics, 3:8, 2012.
11. C. Gamallo, J. Palacios, A. Suarez, A. Pizarro, P. Navarro, M. Quintanilla, , and A. Cano. Correlation of e-cadherin expression with differentiation grade and histological type in breast carcinoma. The American journal of pathology, 142(4):987, 1993.
12. K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Baraba´si. The human disease network. Proceedings of the National Academy of Sciences, 104(21):8685–8690, 2007.
13. G. T. Hart, I. Lee, and E. M. Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene

essentiality. BMC bioinformatics, 8(1):236, 2007.

14. D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature protocols, 4(1):44–57, 2008.
15. D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1):1–13, 2009.
16. M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1):27–30, 2000.
17. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. Nucleic acids research, 42(D1):D199–D205, 2014.
18. A. Kuijper, R. A. de Vos, J. H. Lagendijk, E. van der Wall, and P. J. van Diest. Progressive deregulation of the cell cycle with higher tumor grade in the stroma of breast phyllodes tumors. American journal of clinical pathology, 123(5):690–698, 2005.
19. I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte. A single gene network accurately predicts phenotypic eff ects of gene perturbation in caenorhabditis elegans. Nature genetics, 40(2):181–188, 2008.
20. I. Lee, Y.-S. Seo, D. Coltrane, S. Hwang, T. Oh, E. M. Marcotte, and P. C. Ronald. Genetic dissection of the biotic stress response using a genome-scale gene network for rice. Proceedings of the

National Academy of Sciences, 108(45):18548–18553, 2011.

21. X.-L. Li, J.-H. Xu, J.-H. Nie, and S.-J. Fan. Exogenous expression of uhrf1 promotes proliferation and metastasis of breast cancer cells. Oncology reports, 28(1):375, 2012.
22. P. Marangos and J. Carroll. Securin regulates entry into m-phase by modulating the stability of cyclin b. Nature cell biology, 10(4):445–451, 2008.
23. S. Ogbagabriel, M. Fernando, F. M. Waldman, S. Bose, and A. P. Heaney. Securin is overexpressed in breast cancer. Modern pathology, 18(7):985–990, 2005.
24. S. Parthasarathy, S. Tatikonda, and D. Ucar. A survey of graph mining techniques for biological datasets. In Managing and mining graph data, pages 547–580. Springer, 2010.
25. J. Pierga, J. Reis-Filho, S. Cleator, T. Dexter, A. Mackay, P. Simpson, K. Fenwick, M. Iravani, J. Salter, M. Hills, et al. Microarray-based comparative genomic hybridisation of breast cancer patients receiving neoadjuvant chemotherapy. British journal of cancer, 96(2):341–351, 2006.
26. I. O. Potapenko, V. D. Haakensen, T. L¨uders, ° A. Helland, I. Bukholm, T. Sørlie, V. N. Kristensen, O. C. Lingjærde, and A.-L. Børresen-Dale. Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression. Molecular oncology, 4(2):98–118, 2010.
27. T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin,

et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell, 140(5):744 – 752, 2010.

28. B.-S. Seah, S. S. Bhowmick, C. F. Dewey, and H. Yu. Fuse: a profi t maximization approach for functional summarization of biological networks. BMC bioinformatics, 13(Suppl 3):S10, 2012.
29. E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cacer. Nature genetics, 36(10):1090 – 1098, 2004.
30. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics, 34(2):166 – 176, 2003.
31. E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. Bioinformatics, 19(suppl 1):i273 – i282, 2003.
32. C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute, 98(4):262 – 272, 2006.
33. A.-M. T˝ok´es, A. M. Sza´sz, ´E. Juha´sz, Z. Schaff, L. Harsa´nyi, I. A. Moln´ar, Z. Baranyai, I. Besznya´k Jr, A. Zara´nd, F. Salamon, et al. Expression of tight junction molecules in breast carcinomas analysed by array pcr and immunohistochemistry. Pathology Oncology Research, 18(3):593 – 606, 2012.

34. M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell, 13(6):1977–2000, 2002.
35. S. Wuchty, Z. N. Oltvai, and A.-L. Baraba´si. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nature genetics, 35(2):176–179, 2003.
36. B. Yuan, Y. Xu, J.-H. Woo, Y. Wang, Y. K. Bae, D.-S. Yoon, R. P. Wersto, E. Tully, K. Wilsbach, and E. Gabrielson. Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability. Clinical cancer research, 12(2):405–410, 2006.
37. X. Zhang, K. Liu, Z.-P. Liu, B. Duval, J.-M. Richer, X.-M. Zhao, J.-K. Hao, and L. Chen. Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics, 29(1):106–113, 2013.

요약

오믹스 데이터로부터 전사인자 네트워크를 유추해내는 연구는 활발하게 있어왔지만 유추된 네트워크를 특정 생물 연구 프로젝트에 이용하는 방법론은 아직 미비한 수준이다. 이는 생물학적으로 중요한 서브네트워크를 발굴하는 마이닝 알고리즘의 부재 때문이다. 전사인자 네트워크에 기존의 그래프 마이닝 알고리즘을 적용할 경우, 이는 전사인자 네트워크의 특징을 고려하여 제안된 알고리즘이 아니기 때문에 생물학적으로 의미 있는 서브 그래프들을 찾아내기에 효과적이지 못하다. 이를 해결하기 위해 생물학적인 서브 네트워크 마이닝 문제를 새로 정의하였으며 다중 클래스의 유전자 발현 데이터로부터 구축된 전사인자 네트워크에서 표현형 특이적인 서브 네트워크를 발굴하고 순위를 매길 수 있는 알고리즘을 개발하였다. 본 논문에서는 정보 이론 및 조절 유전자들의 발현량 추이를 이용하는 전사인자-모듈 기반의 분석 기법을 적용하여 표현형 특이적 서브 네트워크를 발굴하는 알고리즘을 개발하였다. 또한 이를 통하여 오믹스 데이터를 이용하여 네트워크를 구축하고 이로부터 표현형이나 질병을 구분 짓는 서브 네트워크를 찾아내는 완성된 연구 패러다임을 제안할 수 있었다. 이를 실제의 데이터에 적용하여 표현형(유방암의 분화도)과 연관된 패스웨이들(세포주기: M기, 세포 접착 분자)을 밝혀냈다. 발굴된 서브네트워크에 속한 조절 유전자들의 발현량은 유방암의 분화도와 명확한 상관관계를 보였다. 알고리즘을 통하여 필터링 되지 않은 모든 유전자들을 사용하였을 때는 상관관계가 불분명했으며 이는 또한 제안한 알고리즘의 유용성을 간접적으로 증명한다. 본 논문의 전사인자 중심의 패스웨이 활성/억제 정도 분석은 유방암뿐만 아니라 다른 많은 생물학적 연구에 기여할 수 있다.