Identifying Gene Regulatory Communities From Microarray Time-Series Expression Profiles Using Hidden Markov Models

by

Osman Mahmut ERYURT

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Computational Science and Engineering
Koç University
August, 2011

This is to certify that I have examined this copy of a master's thesis by

Osman Mahmut Eryurt

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Prof. Dr. Attila Gürsoy

_____

Assoc. Prof Engin Erzin

_____

Asst. Prof Dr. Alkan Kabakçıoğlu

_____

Asst. Prof. Dr. Deniz Yüret

_____

Asst. Prof. Dr. Mehmet Sayar

_____

Date: 19 August 2011

ABSTRACT

Time series microarrays capture multiple gene expression levels at discrete time points varying from minutes to days of a continuous cellular process. Analysis of high through put data requires automated and computer aided solutions. We propose a hidden Markov model (HMM) based approach to identify regulatory relations between the periodic genes from the cell cycle time-series microarrays. We train and test our models by using distinct types of biological data present literature. In our study we use Pramila time series dataset. Training gene pairs include transcriptional regulation and protein level regulation. After identification of gene to gene regulatory relationships, we form a network of gene regulation relationships: Gene Regulatory Neighborhood Networks (GRNN). We explore potential use of sub networks (communities) in GRNN by comparing gene clusters found by popular clustering algorithms such as K-means clustering. Our results indicate we manage to identify denser and more specific enrichment in community structure based clusters than the clusters acquired with K-means.

ÖZET

Zaman serisi mikrodizi verileri devam eden hücresel aktivitelerdeki birçok gen ifadesi seviyesinin dakikalardan günlere uzanan farklı zaman noktalarında ölçülmesini sağlar. Yüksek miktarda sonuç üreten deneylerin analizi otomatik ve bilgisayar destekli çözümler gerektirir. Bu çalışmada saklı Markov modelleri kullanılarak, periyodik gen ifadesi düzenlenmelerini tahmin edebilecek bir model önerdik. Modellimizin testleri ve öğrenme prosedürü için literatürde bulunan transkripsiyon ve protein seviyesindeki düzenlemeleri bilgilerinden faydalandık. Çalışmamızda Pramilla zaman serisi mikrodizi verilerini kullandık. Gen düzenlemelerinin tahminleri yapıldıktan sonra 'Gen Düzenlemeleri Komşuluk Ağlarını oluşturduk (GDKA). Oluşturduğumuz ağın altında topluluk yapısı gösteren alt ağların, biyolojik veri gruplarındaki potansiyelini popüler K-ortalama gruplama sonucu oluşan gruplarla karşılaştırarak değerlendirdik. Sonuçlar K-ortama gruplandırmasına göre topluluk yapısına göre oluşturulmuş alt ağların biyolojik terimler etrafından daha yoğun ve özgün bir gruplar oluşturduğunu gösterdi.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisors, Prof. Dr. Attila Gürsoy and Assoc. Prof Engin Erzin for their support and guidance thought my graduate study and during the completion of this thesis. I want to thank Asst. Prof Dr. Alkan Kabakçıoğlu, Asst. Prof. Dr. Deniz Yüret and Asst. Prof. Dr. Mehmet Sayar for their participation in my thesis committee and for critical reading of my thesis.

I would like to thank all my friends at Koç University for all their support and friendship; Derya Aydın, Evren Arkan, Deniz Şanlı, Besray Ünal, Özge Engin, Beytullah Özgür, Cahit Dalgıçtır, İbrahim Gür, Mümin Öztürk.

Finally, I am grateful to my parents Nuh and Melek and my sister Kübra for their love and support throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Availability of large amount of data from high throughput gene expression experiments creates an opportunity to pioneer gene functions on a global scale. Genes participating on the same pathway or biological process often show similar gene expression profiles. One common approach employed by micro array analyses is clustering genes according to similarities in expressions. Direct comparison of expression profiles are often far from enlightening the complex regulations between genes. Another widely researched approach for data mining from the micro arrays is construction of gene regulatory networks by implementing linear models [1], Bayesian networks [2] and Boolean networks [3]. Generally construction of gene regulatory networks requires a prior knowledge and well defined restrains and low gene counts participating in the network. On the other hand gene co-expression networks are undirected graphs where nodes represent genes and edges representing the degree of similarity in the expression profiles. Co-regulation networks (CRN) differ from regulatory networks by containing indirect interactions and gene relation neighborhood information [4, 5] . There are many studies performed to divide these large networks to smaller sub networks according to their role in biological processes. These studies vary from performing traditional clustering algorithms which employs similarity functions [6] and graph partition based algorithms [7-12] . According to Ruan et al [4] network based approaches are far from surpassing the conventional clustering approaches in the field of detecting functional modules partially due to inability of evaluating functional significance of gene neighborhoods. Pearson correlation coefficient most employed similarity measure for construction of co-expression networks.

In this study instead of CRN we propose Gene Regulatory Neighborhood Network (GRNN). GRNN is an undirected graph, where the graph nodes correspond to genes, edges between the genes represent the regulatory relationship between the genes

and edge weights represent the possible regulation probability. We employ hidden Markov models (HMMs) to measure regulation probability between every possible gene pair and create fully connected GRNN. High effectives of HMM against other possible methods for inferring gene regulatory relations is examined by Yogurtcu et al [13]. We try to exploit neighborhood relations present in GRNN by employing community structure finding algorithm of Clauset et al [14]. Clauset algorithm is a fast algorithm and has the ability required for handling large networks. Clauset algorithm relies on the maximization of modularity. Modularity defined by Newman and Girvan [15] as measure of a quality of particular division of a network to sub networks. By using modularity, clustering algorithms acquire optimal portioning and cluster count automatically without prior initialization which required by most of the conventional clustering approaches. After detection of community clusters we evaluate detected clusters effectiveness in the field of clustering genes according to biological process by performing gene enrichment analysis by using online tools Gorilla [16] and DAVID [17]. We compare community structure based clusters with K-means clusters. The flow of the study:

We first evaluate effectiveness of HMMs against Pearson correlation. We construct two HMMs for different training sets. Both of our training gene pair sets are related with S.cerevisiae cell cycle with different levels and include regulation pairs. We calculate all possible gene pair regulation probabilities and Pearson correlation scores. All scores are sorted in decreasing order. We perform Top-K evaluation of Pearson correlation scores and HMM derived regulation probability scores for detecting regulations. Top-K evaluation show that HMM surplus Pearson correlation in this field.

Secondly we employ Clauset algorithm to identify communities present in our constructed GRNNs from HMM based possible gene pair regulation scores.

We evaluated our clusters efficiency by performing the GO term enrichment and compare our approach performance with K-means clustering algorithm.

Chapter 1

1. LITERATURE REVIEW

Genes are DNA sequences corresponding to a unit of inheritance associated with regulatory regions, coding sequence and other functional sequence regions. Information stored in gene is used to produce protein intermediates called messenger RNA (mRNA) which transfer gene information to ribosome where the assembly of proteins begins. Microarray technologies expose the mutual and specific affinity of complementary strands of DNA. Microarray size can vary according to experimental setup. Microarrays can simultaneously measure different mRNA levels for a specific time points. Analysis of mRNA levels can give an insight about relation between the genes.

Time series micro arrays have ability to capture multiple gene expression profiles at discrete time points of continuous cell process. Data acquired from time series micro arrays can identify expression patterns and regulations. There are many adopted approaches available to analyze differential gene expression data from variety of disciplines such as signal processing, dynamic system theory, machine learning and information theory to detect differentially expressed genes, identification of expression patterns and construct gene networks[2, 18].

Differential gene expressions analyses provide insights about how genes are regulated during biological processes. It is generally accepted that similar expression profiles potentially indicate related functions. This kind of similarities can be explored by clustering analyses. Clustering analyses groups genes according to their differential expression. Clustering techniques are heavily used in the biology field.

Gene regulatory networks are control mechanisms of the gene expression level of a cell. These networks effect genes expression profiles and may or may not be affected by the process. These processes can be observed with time series microarrays.

Correlation coefficient based methods for identification of regulatory gene pairs often create poor results, since these methods based over signal overlapping. The efficiency of these methods is nearly 20 % for detection of known regulatory gene pairs [13]

Attempts on clustering differential gene expression data begins with implementation of some popular distance based clustering methods, such as K-means clustering [19] ,hierarchical clustering [9] and self-organizing maps (SOM) [20]. Even these traditional methods can give meaningful results for some datasets; none of them consider the dependences exist between observations belonging to subsequent time points. The dependence of genes differential expression profiles is the important feature of gene regulatory networks and use of this dependency can increase the clustering performance.

1.1. Budding Yeast Cell Cycle

Cell duplication is a highly controlled process that governs growth and development of living organisms. Events present in cell cycle are highly systematic and ordered. Cell cycle is the combination of events which a cell grows and eventually divides to two daughter cells each contain the information and machinery to repeat the process. Re-replication of DNA may lead to instabilities in cell metabolism. Because of that reason multiple strict control metabolisms have evolved to monitor DNA replications. Several specific protein complexes are employed during DNA replications which are only temporarily active in specific cell components[21].

Regulation of the cell cycle is governed with the cyclin dependent kineases (CDKs). The cell cycle can be divided into four specific phases:

1. G1 phase: Growing and preparing phase for cell cycle entry.
2. S phase: DNA synthesis.

3. G2 phase: Preparations for M phase.
4. M phase: Sister chromatin segregation and cell division.

The cell cycle is controlled by CDKs proteins. CDKs are largely conserved during evolution of eukaryotes.

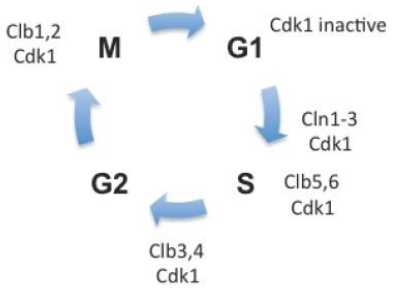

Figure 1 Regulation of the cell cycle by CDKs [21]

A single CDK Cdk1 (Alias CDC 28) is the main controller of cell cycle in S. cerevisae [22]. 9 different cyclins form complexes with Cdk 1 in the process of cell cycle for efficient cell cycle progression.

Due to low cyclin levels CDKs are inactive during G1 phase [23]. As cell cycle process continues cyclin levels begins to increase and interaction of cyclin with CDK increases activity of the kinease [24]. In inactive CDKs, catalytic cleft of the protein is blocked by T loop structure and phosphates from ATP molecule are unaligned [25]. Cyclin binding push the T loop from the catalytic cleft and ATP molecule realign in a position where it can phosphorylate substrate proteins. Phosporylation of T loop increase CDK affinity for cyclin proteins by exploiting the catalytic cleft [24]. Cyclin dependent kinease inhibitors (CDKIs) negatively regulates CDK activity. CDKIs bind cyclin – CDK complexes and reduce the activity of the complex by inhibiting binding of ATP to kinease protein or inhibiting binding of the cyclin to kinease [25]. N terminus residues of CDKs can be phosporylated by Wee1 kineases also cause the inhibition of CDK activity. Swe1 phosphorylate the Cdk1 to prevent entry in to M phase of cell cycle

[26]. Continuation to M phase done by Cdc25 family phosphates (Mih1 in S.cerevisiae) by dephosphorylating N- terminal residues of Cdk1 to switch on Cdk1 activity [27].

The cell cycle continuation is controlled with switch like fashion [28]. Initially in the G1 phase CDKs are inactive because of low cyclin levels and activity of CDKs are low. Later on Cln3 cyclin levels begin to rise and Cln3- CDK complexes phosphorylate Whi5 which is the transcriptional suppressor that interact with SBF transcription factor complex [29, 30]. SBF is essential for transcription of genes involved triggering cell cycle entry and DNA replication. Phosphorylation of Whi5 separates Whi5 from SBF [30] SBF activates transcription of the genes related to cell cycle entry and DNA replication. Several cyclins participates in this progressions and newly synthesized cyclins increase CDK activity and gradually more Whi5 are phosphorylated and increased SBF activity further [31]. Sic1 protein the inhibitor of Clb5-Cdk1 and Clb6-Cdk1 is also phosphorylated by Cln-Cdk1 complexes and degraded by SCF [32, 33].

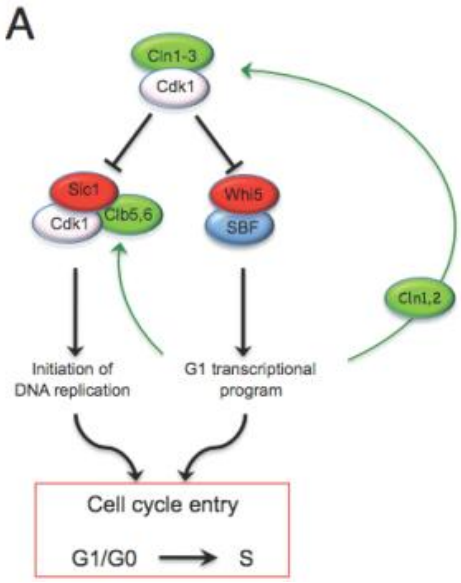

Figure 2 Regulation of cell cycle entry in S. cerevisiae [21].

After initiation of the cell cycle, cell abandons and shuts down the G1 specific transcriptional program to make directional progression among the cell cycle. Nrm1is one of the important mechanisms to shut down G1 transcriptional program [34]. Shut down mechanisms also include Cln1 and Cln2 G1 specific cyclins degradation which results dephosphorylation of Whi5 which as mentioned before inhibits the G1 specific transcription. Further control acquired by Clb-Cdk1 complexes which phosphorylate and inhibit SBF transcription factor complex [35]. For successful cell division 2 essential features are required: Successful replication of DNA and chromosome segregation to daughter cells. DNA replication starts with origin licensing. Origin licensing occurs by formation of pre replicative complex (pre-RC) at origins in G1 phase when Cdk1 protein is inactive. These origins have the capability of autonomously replication and referred as "Autonomously Replicating Sequences" (ARSs) [36].Origin of Replication Complex (ORC) consists of Orc1, Orc2, Orc3, Orc4, Orc5, Orc6 and recognize the ARS consensus sequence [37]. During the cell cycle ORC are constitutively bound the ARS. ORC works with the ATPase Cdc6, Cdt1 and Mcm2-7 helicase complex [38]. DNA replication starts with pre-RC formation and Mcm2-7 activation.

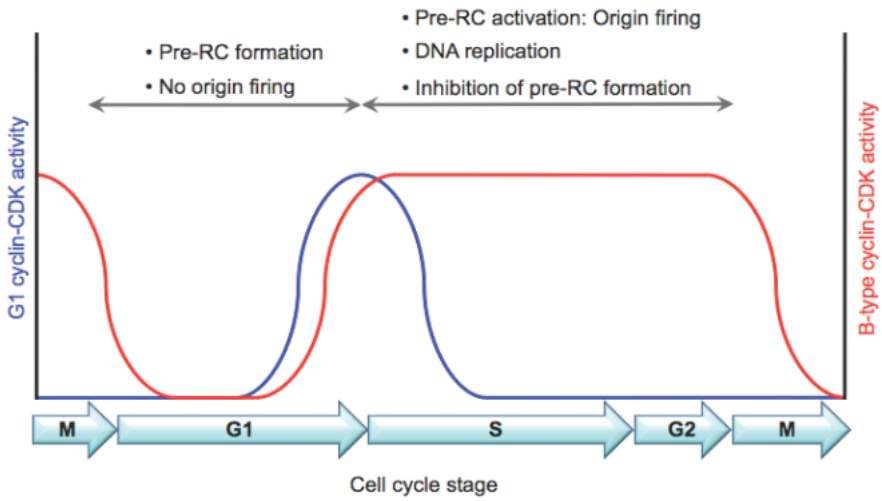

Figure 3 Licensing of origins of replication[21]

Licensing of origins f replication can happen only during the G1 phase when Cdk1 is inactive. ORC activation only happens in S phase when Cdk1. This specificity is required to ensure successful DNA replication and segregation of chromosomes.

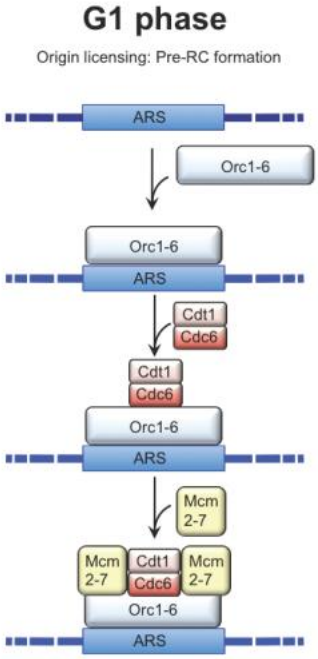

Figure 4 Regulation of DNA replication by the cell cycle: Origin licensing.[21]

Clb5-Cdk1 and Clb6-Cdk complexes initiates the DNA replication [39]. The expression of Clb5 and Clb6 cycles peaks at G1-S transition but their activity are suppressed by Sic1 until progression occurs [32]. Cln1,2 – Cdk1 complexes phosphorylates the Sic1 and phosphorylated Sic1 then ubiquitinated by Skp1- Cullin – F-box (SCF) complex and degraded by protesome. Degradation of Sic1 leads to activation of Clb5 and Clb6 – Cdk1 complexes [33]. Clb5,6 – Cdk1 phosphorylate Sld2 and Sld3 and initiates DNA replication [38]. Phosphorylated Sld2 and Sld3 increase affinity for Dpb11[38] which is essential for sufficient initiation of DNA replication.

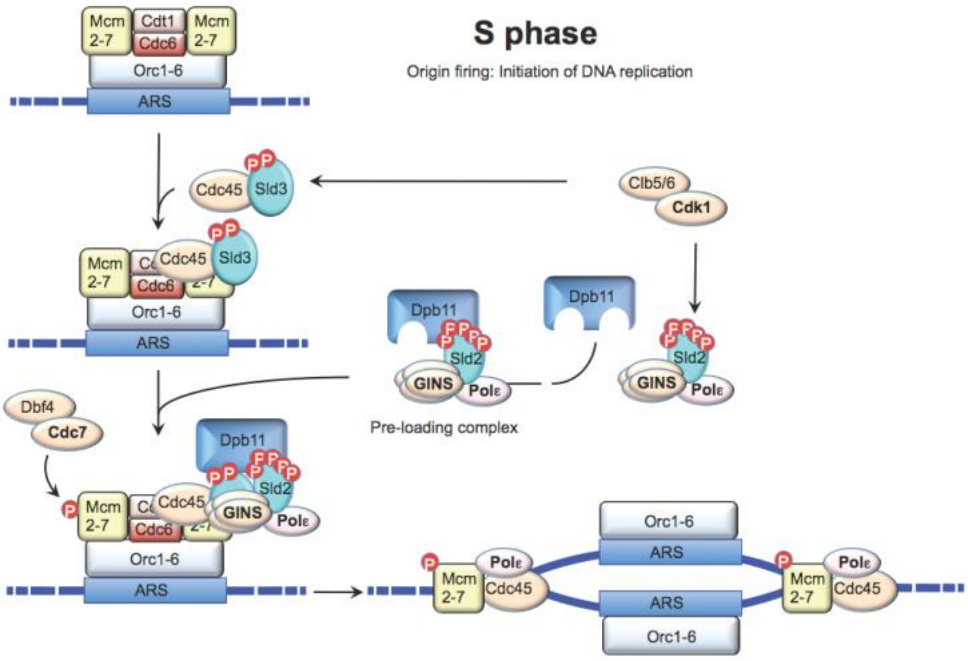

Figure 5 Regulation of DNA replication by the cell cycle: Origin firing.[21].

Expression of most genes which participate in the DNA replication process, peak at the specific phases of the cell cycle. Generally genes participating in early stages of the DNA replication process are transcribed in G2 and M phases. Genes participate in later stages are peak during late G1 and their levels are peak in S phase when the DNA replication takes place. The staged expression of DNA replication factors to specific time phases of the cell cycle restricts DNA replication to S phase [21, 22]. To ensure the DNA only replicated once. In the S phase several mechanisms inhibit firing of origins in other phases. Cdc6 transcription only happens in G1 and S phases [21, 40] and phosphorylation of Cdc6 by Cdk1 results its degradation [21, 41] . Another mechanism which inhibits the firing of origin is the inhibition of MCM2-7 complex associated with Cdt1 by a Cdk1 dependent progress [21, 42]. Cdk1 also phosphorylates and inhibits ORC complex[43].

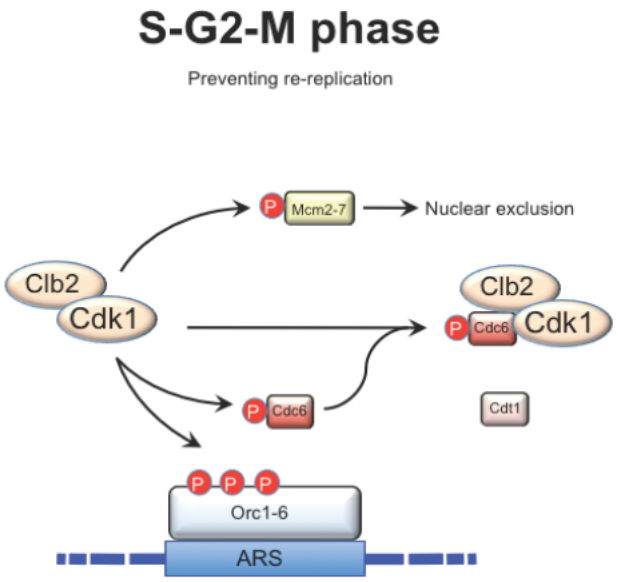

Figure 6 Mechanisms that prevent re-replication [21]

The DNA damage checkpoints act as switches between G1/S and G2/M transitions. There are other two types of checkpoints in S phase. First of them is the DNA replication checkpoint, which arrest the cell cycle and inhibits work of late replication origins in response to replication stress. Intra S checkpoint reduces the DNA replication pace due to DNA damage [21, 44].

Mad2 prevents chromosome segregation by inhibiting degradation of the securin Pds1. Pds1 is phosphorylated and stabilized in response to DNA damage in a Chk1-dependent manner. The spindle checkpoint is also involved in the Cdc14 release from the nucleolus. Cdc14 dephosphorylates Swi5, Sic1, and Cdh1, leading to inhibition of Cdc28 and degradation of cyclin required for mitotic exit [45].

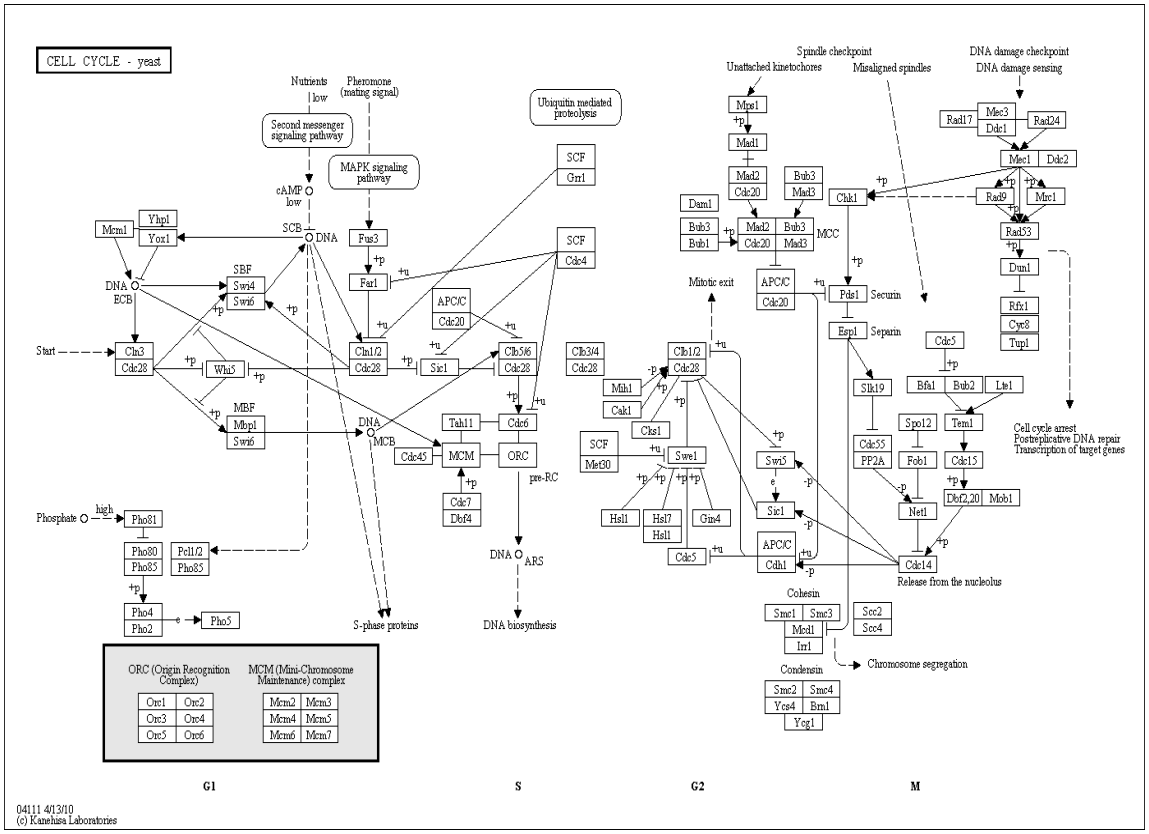

Figure 7 Kegg Cell Cycle [50]

1.2. Modular Framework of Gene Regulatory Networks

The process of gene regulations performed by collection of regulatory proteins and their interactions with specific sequences in promoter regions of targeted genes. This process defined as "Transcriptional Regulatory Networks". High-throughput genome-wide techniques show that the molecular interaction network of a cell consists of modular units. Gifford et al [18] define the modules as group of genes that shows a unique common behavior across a significant set of experiments and that share a common cellular function. Present data indicates that co expressed gene modules are regulated by common transcriptional regulator or regulatory networks. A standard approach to analyze time series micro array data is clustering expression profiles while considering the promoter sequence alignment to analyze probability to share common

transcription factors [46] However many co-expressed genes does not have to be co-regulated [47]. Co expression can be response to indirect factor rather than direct initialization by same transcription factors and does not have to share common transcription binding motifs in the promoter region and also combinatorial role of transcription factors clearly allows for the occurrence of a single transcription factor binding motif in genes that are not co regulated [48].

.

Chapter 2
METHODS

1.3. DATA SETS

1.3.1. Time Series Micro Arrays

Microarray technologies create an opportunity for exploration of gene expressions in a global and parallel fashion [49]. Time series micro arrays simultaneously capture multiple gene expression profiles for discrete time points. As a result it creates differential gene expression as a function of time. By using differential gene expression we can pioneer the regulatory relations between the genes because regulated gene pairs are expected to carry temporal correlations or patterns in their differential gene expressions[13].

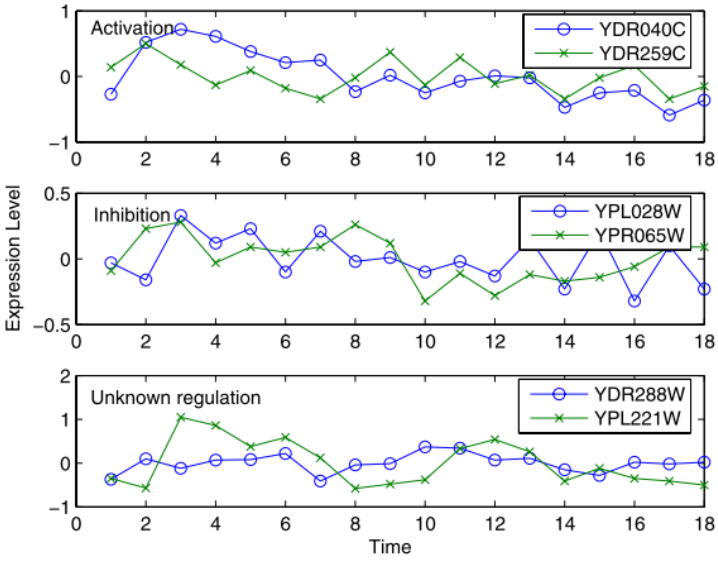

Figure 8 Differential Expression of three gene pairs (activation, inhibition and unknown regulations) from Spellman's α–synchronized cell-cycle experimental data[13].

In this study, we use Pramila micro arrays [50]. The three microarray experiments performed over cell cycle of S.Cerevisiae by using spotted cDNA arrays and alpha factor to induce synchrony. The two data sets (alpha30 and alpha38) are dye swap technical replicates with 5 minutes sampling interval and consist of total 25 time points. The third data set, (alpha26), sampled in 10 minutes intervals and consists of 13 time points. Data sets were processed with Rosetta Resolver version 3.2 error models. All values are $\log_{10}$. W303a yeast strain was used in experiments growth on YEP glucose medium

In this study both processed and raw micro array data sets are used to train and test our HMM models. The expression-level measurements of a gene in a given situation have a roughly Gaussian distribution according to [51]. Because of this fact normalization of the data done according to formula:

$$x_{ij}^{'} = \frac{x_{ij} - \overline{x_i}}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(x_{ij} - \overline{x_i}\right)^2}},$$

(Eq. 1)

where x_{ij} is the raw expression data of gene at time point j and $\overline{x}$ is the average expression data of gene i. The above formula makes the mean of expression data, μ to be 0, and the standard deviation, σ , to be 1. The data normalization is to project all data into the same area. The following equation is used to make the gene expression data fall into the particular interval $[a, b]$:

$$x_{ij} = \frac{(b-a)\left(x_{ij}^{'} - x_{i\,min}^{'}\right)}{x_{i\,max}^{'} - x_{i\,min}^{'}} + a\,,$$

(Eq. 2)

where $x_{i\,min}^{'}$ is minimum and $x_{i\,max}^{'}$ is the maximum values in a single gene expression time-series data sequence.

1.3.2. YEASTRACT

YEASTRACT is a curated database consists of more than 48333 regulatory associations between transcription factors and the target genes in S.cerevisae. It includes 298 specific DNA binding sites for transcription factors. The yeast gene information derived from Saccharomyces Genome Database (SGD) and Gene Ontologies (GO) for each gene derived from GO consortium. YEASTRACT includes set of tools (DISCOVERER) which can identify complex motifs found to be represented in the promoter regions of co-regulated genes. Tools generally take an input list of genes and identify over-represented motifs. YEASTRACT identify documented or potential transcription regulators of a gene and documented or potential regulators for each transcription factor. DISCOVERER can group a list of genes according to regulatory associations with known transcription factors. YEASTRACT also provides tools to retrieve important biological information from the gathered data and to predict transcription regulation networks for yeast [52, 53]. We use YEASTRACT to retrieve 18327 TF regulated gene pairs which are experimentally validated with strong evidence.

1.3.3. CYCLEBASE

Cyclebase is an online resource of the cell-cycle-related experiments. Cyclebase has a simple interface that facilitates visualization and download of genome wide cell cycle data and analysis results. Data base processes raw information from different cell cycle experiments and normalize data to a common timescale and data is presented with key cell cycle information and derived analysis results. In this study we use '*Cyclebase ver. 2.0'* [54, 55]. to extract cell cycle related S.Ceravisiae genes. Data base 599 genes as periodic genes and in training data set creation for HMM we use data to extract cycle related gene pair information from KEGG and YEASTRACT.

Cyclebase evaluate periodicity of a gene by comparing Fourier score of original time points and random shuffled time points of each gene profile and calculates a P-value. The Fourier score is defined as

$$F_i = \sqrt{\left(\sum \sin(\omega t)\ .\ x_i(t)\right)^2 + \left(\sum \cos(\omega t)\ .\ x_i(t)\right)^2}$$

(Eq.3)

where $\omega = 2\,\pi$ is the interdivision time.

After calculation of Fourier scores 1,000,000 artificial gene profiles are generated from shuffling of the data within the original profile by Cyclebase. The profiles with equal or greater Fourier scores are normalized to create the final P-value for periodicity

1.4. PREDICTION OF GENE TO GENE REGULATIONS

HMM is widely accepted approach for characterizing temporal or sequential behaviors of a pattern. HMMs used in computational biology area for pioneering protein sequence aliment problems. HMMs are popular for exploiting the time dependences. HMMs can be described as a stochastic generalization of finite-state automata and it provides a probabilistic description of temporal dependences. HMM approach is widely used statistical method of characterizing temporal or sequential behaviors of a pattern.

In this thesis HMM is used to model time-series expression data of regulated gene pairs present in time series micro arrays. HMMs are well-known and applied extension of Markov chains. For each state HMM can infer unknown underlying stochastic process which can be inferred through the observations it generates.

Temporal patterns of regulated gene pairs are modeled with HMM. Temporal expression data of a gene pair is a 2 dimensional (gene expression versus time) observations for a certain non-deterministic processes. This 2D observation array is used to model known and unknown regulations with different HMM structures after a learning phase.

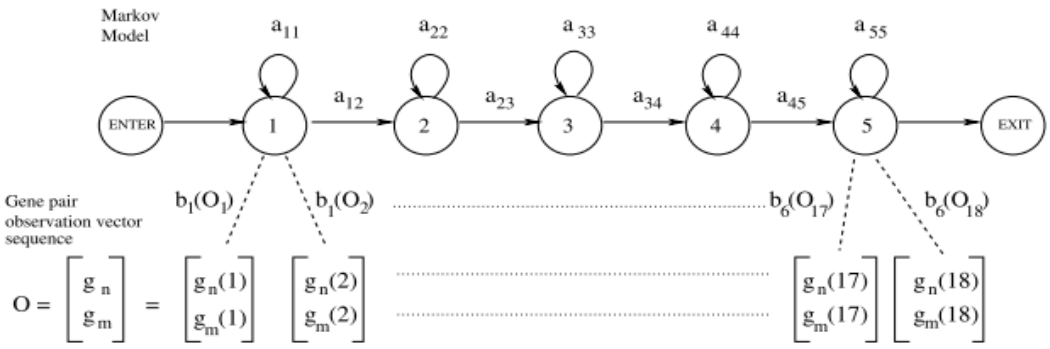

Figure 9: Five state left-to-right HMM model (Observation vector includes expression levels of the gene pair (g_n ,g_m) [13].

HMM structure consists of states, state transitions and input observation probabilities. We can represent HMM as $\lambda = (A, B, \Pi)$ with the parameters

- $S = \{s_1, s_2, \dots, s_N\}$ is the states
- $A = \{a_{ij}\}$ is the matrix of the state transition probabilities. Here a_{ij} denotes the probability of making a transition from state s_i to s_j such that $a_{ij} = P\{q_{t+1} = s_j | q_t = s_i\}, i, j = 1, \dots, N$,

  where q_t is the state at time t.

- $B = \{b_j(o)\}$ is the vector of observation probabilities associated with each emitting state j, where $b_j(o) = P(o | q_t = s_j)$. Since we modeled observation symbol probabilities with Gaussian mixture densities:

$$b_j(o) = P\left(o\middle|q_t = s_j\right) = \sum_{l=1}^{L} \omega_{ij}\, \mathcal{N}\left(o_t, \mu_{jl}\Sigma_{jl}\right),$$

where $\mathcal{N}\left(o_t, \mu_{jl}\Sigma_{jl}\right)$ is Gaussian mixture densities for state j, μ_{jl} is the weight, Σ_{jl} is covariance matrix, $\mathcal{N}_{jl}$ is the mean vector at state S_j for mixture L, such that $\sum_l \omega_{jl} = 1$ and $0 < \omega_{jl} \leq 1$

- $\Pi = \{\pi_i\}$ is the vector with initial state probabilities of entering the model at state i such that $\pi_i = P(q_1 = s_i)$.

In this HMM structure, states represent statistically correlated time series segments and state transitions represents segment to segment transitions. The probability density functions of expression levels in each segment are representing observation probabilities which are derived by Gaussian density functions. If g_n represents the n-th gene in time series micro array data, and the expression level of the g_n at time t is represented by $g_n(t)$. Then we can represent the observation vector at time t as $o_t = [g_n(t), g_m(t)]$ for the genes g_n and g_m. The probability of observing gene-pair time series expressions, $O = (o_1, o_2, \ldots, o_T,)$ for HMM λ defined as,

$$P(O|\lambda) = \sum_{all\ q} P(O, q|\lambda) \qquad \text{(Eq. 4)}$$

Scoring of the regulatory probability calculated by

$$P\left(O_{ij}\right) = \log P\left(O_{ij}\middle|\lambda_r\right) - \log\ P\left(O_{ij}\middle|\lambda_u\right), \qquad \text{(Eq. 5)}$$

$$Score_{ij} = max\left\{P\left(O_{ij}\right)\middle|P\left(O_{ji}\right)\right\} \qquad \text{(Eq. 6)}$$

where

λ_r is known regulations HMM and λ_u is unknown regulations HMM. For creation and testing of HMM we use '*The Hidden Markov Model Toolkit (HTK) ver. 3.4.1*' [56].

1.5. CONSTRUCTION OF GENE REGULARTORY NEIGHBOURHOOD NETWORKS (GRNNs) AND COMMUNITY STRUCTURES IN GRNNs

1.5.1. CONSTRUCTION OF GRNNs

A GRNN is an undirected graph, where the graph nodes correspond to genes and edges between the genes represented the regulatory relationship between the genes and edge weights represented the possible regulation probability. GRNN differs from regulatory networks because GRNN does not attempt to distinguish direct gene to gene regulations from the indirect ones, but GRNN network contains regulatory gene neighborhood relations that are generally overlook in cluster analysis [4, 57]. GRNN differs from the gene co-expression networks because it is only constructed from regulatory relation probabilities and does not include co-regulations.

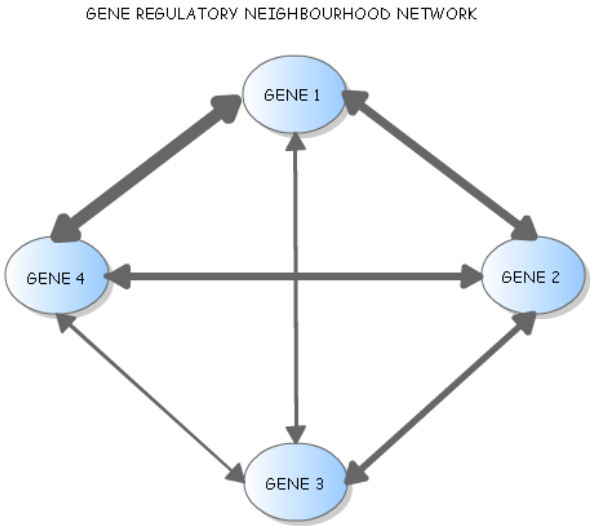

Figure 10 Sample Gene Regulatory Neighbourhood Network (GRNN). All genes are connected to each other by the probability of potential regulations. GRNN is an undirected graph. (Arrow thicknesses indicate strong regulation probability.)

1.6. DETECTION OF COMMUNITY STRUCTURES

Girvan and Newman define community structure as the division of network nodes into groups within which the network connections are dense, but between which they are sparser [15]. Detection of community structures are related with previous studies present in the literature like graph partitioning in graph theory and hierarchical clustering in sociology. However the solution of the graph partitioning assumes the size of the clusters are about the same size and requires size minimization of edge cuts and cluster count initialization which does not correlate with many applications. Community structure detection algorithms are more closely related with the hierarchical clustering approaches. Community structure detection algorithms focuses on the betweenness of the edges. Betweenness of edges means the edge in some sense responsible for connecting many pairs of others.

We use Clauset algorithm [14] for detection of community structures because of its ability to deal with large networks and short running time. This algorithm detects community structures based on greedy optimization of modularity. Modularity defined by Newman and Girvan [15] as measure of a quality of particular division of a network to sub networks.

If we consider a particular division of a network into k communities we can define a $k\ X\ k$ symmetric matrix e consist of elements e_{ij} which is the fraction of all edges in the given network that connects the vertices in community i to vertices in community j [15]. The trace of a matrix e presents the fraction of edges which connect the vertices in the same community. The high scores in trace of e indicate a good quality division into the community networks. For row sums $a_{ij} = \sum_j e_{ij}$ represent the fraction of edges that connect to vertices in community i. Edges fall between any vertices without considering the community structure defined as $e_{ij} = a_i a_j$. Girvan and Newman define the modularity measure as follows.

$$\varphi = \sum_i (e_{ii} - a_i^2) = Tr\ e - \|e^2\| \qquad \text{(Eq 7)}$$

φ is the measure of the fraction of the edges in the network that connect the vertices of the same type minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices [15].

Clauset algorithm connects nodes in a way that greedily maximizing the modularity scores starting from each individual node. Algorithm designed in a way to approach constantly to higher modularity if the next modularity is smaller than the modularity calculated from the previous step algorithm terminates itself and maximum modularity acquired. It is claimed that the algorithm essentially run in linear time [14].

Detection of the community structures is performed with igraph [58]. Igraph is a collection of software packages for graph theory and network analysis. In study we employed community fast and greedy method which is an implementation of method presented by Clauset [14] for detection of community structures.

Our experiments indicate that fully connected GRNN are too dense for finding community structures. Because of this fact we scale all HMM based potential regulation scores as way to fall in the interval of [-1, 1] by using Eq. [2]. And we gradually start to eliminate edges starting from smaller than selected threshold. We change the threshold between 0.1 to 1.0 and we detected 0.5 is optimal for keeping large portions for nodes (genes) and detection of community structures.

After detection of community structures we perform K-means clustering. Since K-means require initialization of cluster count we use simply 3 clusters to directly compare results. K-means similarity measure for this experiment was Euclidian distance. We use Cluster 3.0 [59] implementation of K-means clustering.

We evaluate the performance of each approach by GO Biological Process (GO BP) term enrichment analysis. The enriched GO terms analysis for different clusters performed by '*Gene Ontology Enrichment Analysis and VisuaLizAtion Tool'* (GORILLA) [16] and '*The Database for Annotation, Visualization and Integrated Discovery'* (DAVID) v6.7.[17]. Detection of overlapping GO categories detected and reduced by REVIGO tool [60] which uses semantic interrelations. All the enrichment

analysis performed by GORRILLA with the setting P value as 0.001 to ensure GO enrichments are valid. Enrichment score of specific GO term is calculated by Gorilla as:

$$S_E = Enrichment\ score = \frac{\left(\frac{b}{n}\right)}{\left(\frac{B}{N}\right)} \quad \text{where}$$

where (Eq. 8)

N is the total number of gene , B is the total number of genes associated with a specific GO term ('target' set and 'background set'), n is the number of genes in the 'target set' and b is the number of genes in the 'target set' associated with a specific GO term.

Chapter 3:
RESULTS AND DISCUSSION

HMM evaluation for gene pair detection with comparison to Pearson Correlation scores and selection of appropriate state count and Gaussian Mixtures counts presented in the first part. In the second part we present GRNN derived community structure clustering GO BP Term enrichment results. All the HMMs are trained with Pramila 26 micro array and tested over Pramila 30 micro array.

We use two sets of known regulation pairs as training pairs. First set (RT1) consist of regulation pairs which are experimentally validated and ordered regulation pairs. Pairs present in regulation set 1 derived from YEASTRACT [53] and Kegg [45] databases and consist of genes reported as periodic by Cyclebase [54]. RT1 contains 541 genes and 2159 pairs. Second regulation pair set (RT2) consist of gene regulation pairs derived from YEASTRACT only. At least one element of the gene pair used in this training set reported as periodic and ranked below 200 by Cyclebase. We trained this additional model to increase the gene count present in our GRNN. Since we use cell cycle micro array data we try to ensure that training gene pairs related with cell cycle in some extend. Training set 2 contains 861 genes and 1565 pairs.

1.7. HMM Pair Likelihood Score Comparisons

In Table 1, we present the best fitting HMMs for detection of regulations. Our data for this experiment is normalized micro array samples. Data points of used micro array data also arranged in a way to fall in to interval [-1,1] by Eq 8. We perform this experiment because HMM based clustering approaches often parameterize time micro array into 3 letter alphabet. They assign time points below the mean of expression as

down regulation, points above the mean as up regulation and points close to mean as unchanged expression [61]. Since Top-K analysis are shows the accumulation of ranks for specific search components, average of Top-K plot (Average Pair Count (APC)) can be used as metric to evaluate goodness of model for detecting regulations.

In Table 3 and Table 4, we present best fitting HMMs for detection of regulations. We use raw micro arrays without further normalization and without projecting the data points. Our results indicate that HMMs have higher performance with unprocessed micro arrays. If we directly compare average detected gene counts of HMMs trained with RT1, HMM performs 7.1 % APC higher for unprocessed arrays. HMM performs more regularly with unprocessed micro arrays if we consider the state counts. Because of this reasons we continued our study with unprocessed arrays.

Our results indicate RT1 trained best fitting HMM with raw micro array (hmmTR1) produce 19.2 % higher APC score against Pearson correlation and for RT2 trained best fitting HMM (hmmTR2) 29.6 % APC against Pearson correlation. The gaps between hmmTR1 and hmmTR2 performances against Pearson correlation are due to genes presents in the RT sets. RT1 only consist highly periodic genes which favors Pearson correlation and RT2 consist of periodic and non periodic genes but all regulations pairs present in RT2 consist at least one periodic gene. Even with the favoring conditions Pearson correlation performed poorly than hmmTR1.

Table 1 HMM TopK Results Raw Data

Rank	State	λ_r GMM count	λ_uGMM count	Average Pair Count
1	3	1	1	1231.854
2	5	5	3	1225.626
3	7	1	6	1218.848

4	8	6	3	1215.169
5	7	1	4	1215.114
107	P26 Correlation	-	-	1070.843

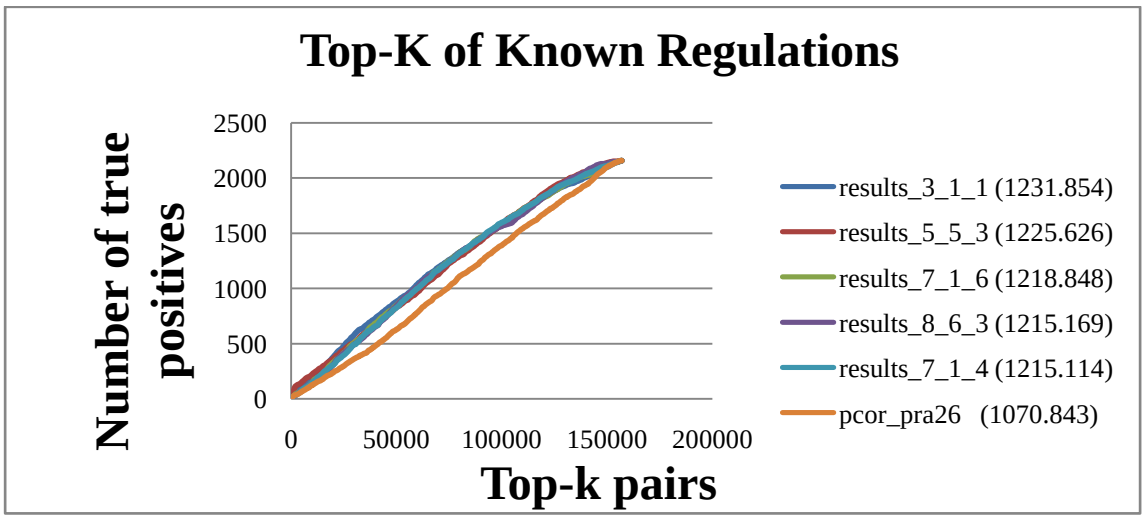

Figure 11 Top-K of Known Regulations for RT1 trained HMMs and Pearson correlation over processed time series data.

Result in Fig 11 shows superiority of HMM against Pearson correlation in the field of detecting regulations.

1.7.1. Unprocessed Micro Array Data Trained and Tested HMM models with RT1

In this part we presented performance analysis for HMMs which trained and tested over unprocessed micro array data. We use RT1 pairs for HMM training. We

successfully created 217 models for these experiments and sorted our results according to APC values.

Table 2 HMM TopK Results Raw Data

Rank	State	λ_r GMM count	λ_uGMM count	Average Pair Count
1	8	6	5	1325.130
2	8	6	4	1318.435
3	8	6	2	1310.427
4	8	6	3	1309.972
5	8	1	4	1298.451
199	P26 Correlation	-	-	1070.843

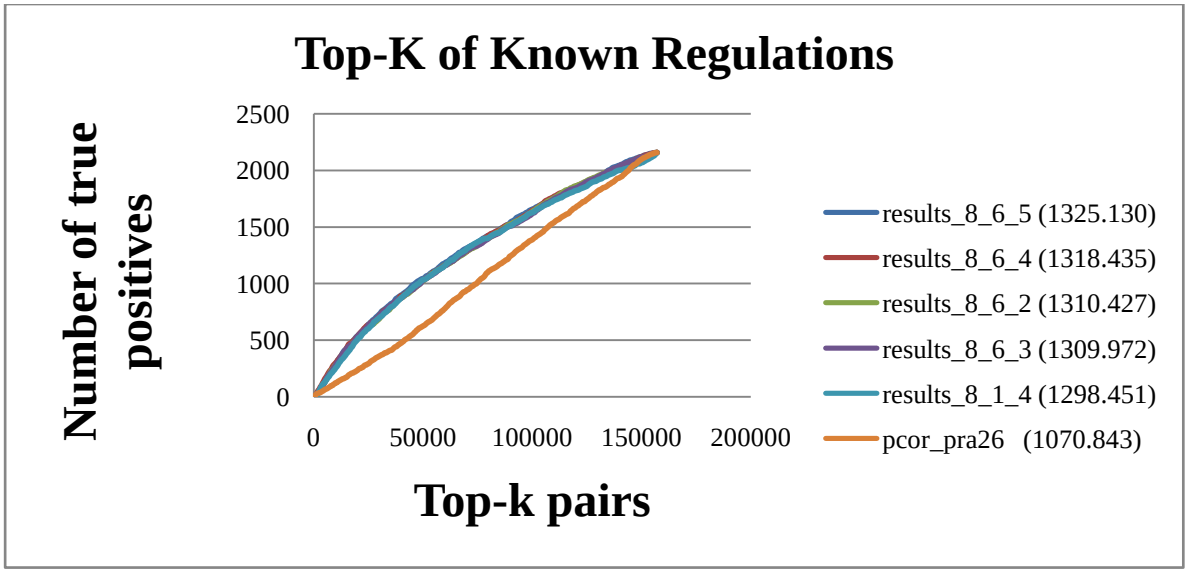

Figure 12 Top-K of Known Regulations present RT1

Result in Fig 12 shows superiority of HMM against Pearson correlation in the field of detecting regulations.

1.7.2. Unprocessed Micro Array Data Trained and Tested HMM models with RT2

In this part we presented performance analysis for HMMs which trained and tested over unprocessed micro array data. We use RT2 pairs for HMM training. We successfully created 150 models for these experiments and sorted our results according to APC values.

Table 3 HMM Top-K Results Raw Data

Rank	State	λ_r GMM count	λ_u GMM count	Average Pair Count
1	5	2	3	1182.341
2	4	2	1	1168.115
3	4	3	1	1166.896
4	5	3	2	1166.068
5	5	2	2	1160.529
148	P26 Correlation			833.094

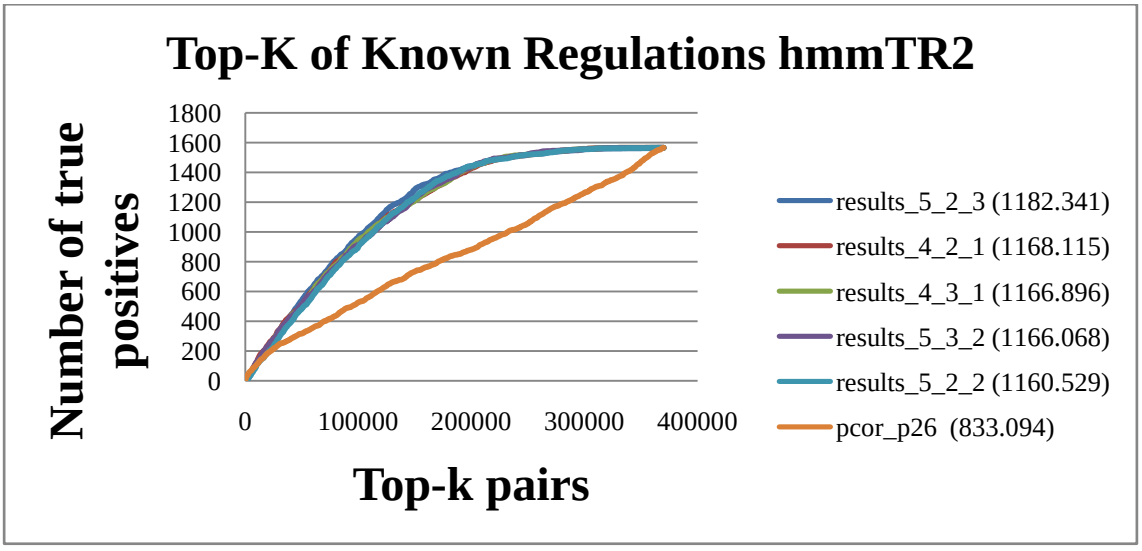

Figure 13 Top-K of Known Regulations present RT2

Result in Fig 13 shows superiority of HMM against Pearson correlation in the field of detecting regulations.

1.8. Community Structure Detections

In the first part of the study we use HMM to extract potential regulation probabilities between gene pairs and we constructed GRNNs. GRNN is constructed from best regulation finder HMM (8 state HMM) that trained with RT1. As mentioned previously in Methods Chapter density of GRNN is reduced. We examined community structures in the resulting GRNN and evaluate the functional difference between communities by comparing functional GO BP term enrichment analysis. Since not all genes are significantly enriched in clusters, we need to consider how many of the genes present in clusters share the enriched GO terms. We define what percentage of genes share the enriched GO terms as cluster density.

1.8.1. GORILLA BASED ANALYSIS OF ENRICHED GO TERMS

1.8.1.1. K-means Derived Cluster Analysis for Genes Present in Training set 1

We clustered genes present in RT1 with K means clustering. And performed GO BP term enrichment with GOrrilla. Results presented in Tables 4-8. .

Table 4 K-means Clusters for RT1 genes

Community ID	Number of Genes	Overlapping GO Categories (REVIGO)

1	237	1- Protein Folding 2- DNA Metabolism and Repair 3- Sister Chromatin Cohesion
2	48	1- Cytokinetic Process 2- Regulation of Multi –Organism Process 3- Regulation Of Nucleotide Metabolism / Regulation of GTPase activity
3	270	1- Cellular Processes / RNA Metabolism /Transcription,DNA dependendent / Regulation of Transcription from RNAse II promoter /Regulation of Cellular processes /Membrane Organization / Biological Regulation

As we can see K-means clusters are not dense according to GO BP terms. Overlapping GO terms indicate that the enriched GO terms are not closely related. Overlapping GO Terms for cluster 3 are highly related because is constructed from high level GO terms which are common for many genes.

Table 5 Enriched GO TERMS GENE COUNTS

COMMUNITY	TOTAL GENE COUNT	ASSOCIATED GENE COUNT	CLUSTER DENSITY
Community 1	237	48	20.3 %
Community 2	48	14	29.2 %
Community 3	270	214	79.3 %

Enriched GO terms percentage indicates low gene participation in more specified clusters except cluster 3. After we eliminate GO terms in low levels which are common for large portion of genes (GO Cellular Processes and GO Biological Processes). This clusters density falls to 46.3 %.

Table 6 K-means Cluster 1 Biological Process Enrichment Results (GORRILLA)

GO Term	Description	P value	Enrichment (N, B, n, b)
GO:0006281	DNA repair	2.17E-06	1.78(543,47,227,35)
GO:0006259	DNA metabolic process	2.69E-05	1.51(543,81, 227,51)
GO:0007062	sister chromatid cohesion	3.67E-05	2.23(543,15, 227,14)
GO:0006974	response to DNA damage stimulus	5.69E-05	1.59(543,57, 227,38)
GO:0007064	mitotic sister chromatid cohesion	8.49E-05	2.22(543,14, 227,13)
GO:0006273	lagging strand elongation	3.55E-04	2.39(543,9, 227,9)
GO:0033554	cellular response to stress	7.43E-04	1.41(543,78, 227,46)
GO:0006457	protein folding	8.67E-04	2.39(543,8, 227,8)
GO:0006298	mismatch repair	8.67E-04	2.39(543,8, 227,8)

Table 7 K-means Cluster 2 Biological Process Enrichment Results (GORRILLA)

GO Term	Description	P value	Enrichment (N, B, n, b)
GO:0032506	cytokinetic process	1.17E-04	4.52(543,20,48,8)
GO:0043900	regulation of multi-organism process	2.02E-04	7.07(543,8,48,5)
GO:0031137	regulation of conjugation with cellular fusion	2.02E-04	7.07(543,8,48,5)
GO:0046999	regulation of conjugation	2.02E-04	7.07(543,8,48,5)
GO:0070783	growth of unicellular organism as a thread of attached cells	6.11E-04	4.85(543,14,48,6)
GO:0044182	filamentous growth of a population of unicellular organisms	6.11E-04	4.85(543,14, 48,6)
GO:0030811	regulation of nucleotide catabolic process	7.12E-04	7.54(543,6, 48,4)
GO:0043087	regulation of GTPase activity	7.12E-04	7.54(543,6, 48,4)
GO:0006140	regulation of nucleotide metabolic process	7.12E-04	7.54(543,6, 48,4)
GO:0033121	regulation of purine nucleotide catabolic process	7.12E-04	7.54(543,6, 48,4)
GO:0033124	regulation of GTP catabolic process	7.12E-04	7.54(543,6, 48,4)

Table 8 K-means Cluster 3 Biological Process Enrichment Results (GORRILLA)

GO Term	Description	PValue	Enrichment	(N	B	n	B)
GO:0006351	transcription, DNA-	9.47E-08	1.47	543	101	270	74

	dependent						
GO:0032774	RNA biosynthetic process	9.47E-08	1.47	543	101	270	74
GO:0006357	regulation of transcription from RNA polymerase II promoter	4.46E-07	1.52	543	78	270	59
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	4.52E-07	1.75	543	39	270	34
GO:0051252	regulation of RNA metabolic process	2.19E-06	1.37	543	125	270	85
GO:0010468	regulation of gene expression	2.48E-06	1.36	543	127	270	86
GO:0051254	positive regulation of RNA metabolic process	4.03E-06	1.61	543	50	270	40
GO:0016070	RNA metabolic process	4.08E-06	1.37	543	119	270	81
GO:0048518	positive regulation of biological process	4.95E-06	1.53	543	63	270	48
GO:0048522	positive regulation of cellular process	4.95E-06	1.53	543	63	270	48
GO:0006355	regulation of transcription, DNA-dependent	5.21E-06	1.36	543	123	270	83
GO:0010628	positive regulation of	6.95E-06	1.6	543	49	270	39

	gene expression						
GO:0045893	positive regulation of transcription, DNA-dependent	6.95E-06	1.6	543	49	270	39
GO:0045935	positive regulation of nucleobase	1.02E-05	1.58	543	51	270	40
GO:0051173	positive regulation of nitrogen compound metabolic process	1.02E-05	1.58	543	51	270	40
GO:0010604	positive regulation of macromolecule metabolic process	1.46E-05	1.56	543	53	270	41
GO:0010557	positive regulation of macromolecule biosynthetic process	1.73E-05	1.57	543	50	270	39
GO:0060255	regulation of macromolecule metabolic process	1.75E-05	1.3	543	149	270	96
GO:0010556	regulation of macromolecule biosynthetic process	1.99E-05	1.32	543	131	270	86
GO:2000112	regulation of cellular macromolecule biosynthetic process	1.99E-05	1.32	543	131	270	86
GO:0009889	regulation of biosynthetic process	1.99E-05	1.32	543	131	270	86
GO:0031326	regulation of cellular biosynthetic process	1.99E-05	1.32	543	131	270	86

GO:0080090	regulation of primary metabolic process	2.16E-05	1.28	543	155	270	99
GO:0031323	regulation of cellular metabolic process	2.16E-05	1.28	543	155	270	99
GO:0050794	regulation of cellular process	3.08E-05	1.23	543	205	270	125
GO:0019222	regulation of metabolic process	3.56E-05	1.27	543	158	270	100
GO:0009891	positive regulation of biosynthetic process	4.02E-05	1.54	543	51	270	39
GO:0031328	positive regulation of cellular biosynthetic process	4.02E-05	1.54	543	51	270	39
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.10E-05	1.3	543	138	270	89
GO:0051171	regulation of nitrogen compound metabolic process	4.10E-05	1.3	543	138	270	89
GO:0050789	regulation of biological process	4.66E-05	1.22	543	208	270	126
GO:0009893	positive regulation of metabolic process	7.13E-05	1.5	543	55	270	41
GO:0031325	positive regulation of cellular metabolic	7.13E-05	1.5	543	55	270	41

	process						
GO:0016044	cellular membrane organization	9.79E-05	2.01	543	13	270	13
GO:0061024	membrane organization	9.79E-05	2.01	543	13	270	13
GO:0065007	biological regulation	4.31E-04	1.17	543	228	270	133
GO:0009987	cellular process	9.69E-04	1.06	543	462	270	243

1.8.1.2. GRNN Derived Graph Analysis

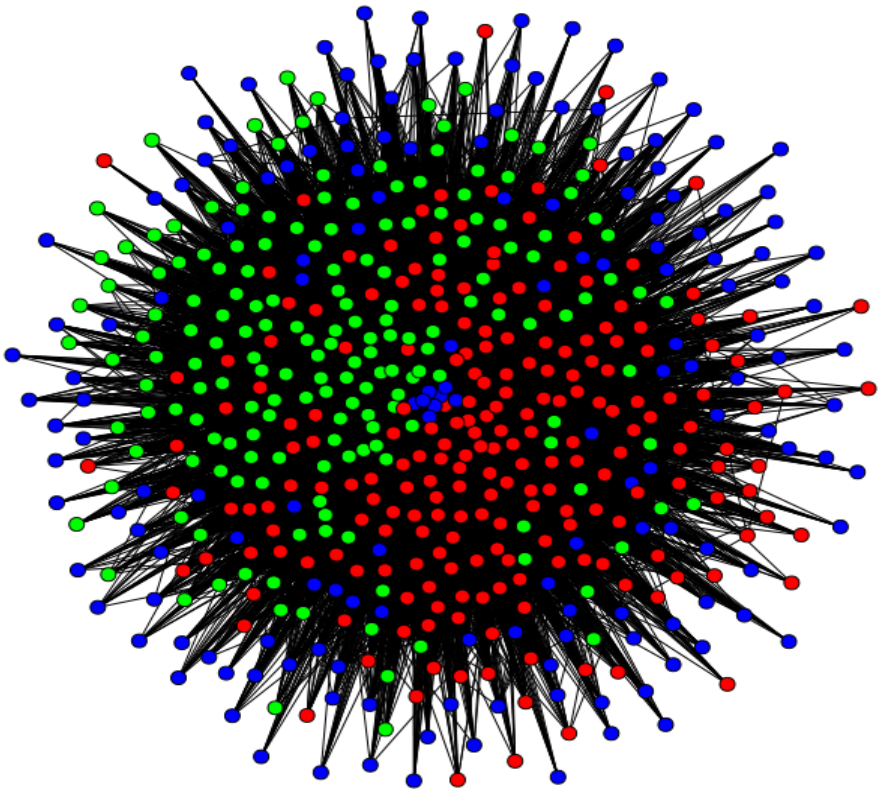

Figure 14 Community Structures in GRNN

For GRNN we identify 3 community clusters.

Table 9 Community Structures Detected in GRNN

Community ID	Number of Genes	Overlapping GO Categories (REVIGO)
1	299	1- DNA Metabolism, Cell Cycle Process, DNA Repair, Cellular Component Organization at Cellular Level, 2- Cellular Component Organization or Biogenesis.
2	152	1-Transcription DNA-dependent, RNA metabolism, Positive Regulation of Gene Expression, 2-Regulation of Transcription from RNA polymerase II promoter
3	90	1-Aspartate Family Amino Acid Metabolic Process

Results indicate that community structures detected from GRNN reflects Biological Processes. In overlapping GO BP categories column of the Table 9 indicates identified communities participate in related biological processes. Community 1 and Community 2 consist of closely related GO annotations. Results show the genes are grouped by participation in different biological process by our approach. Since the community structure detection algorithm calculates community structure count from GRNN with a more successful potential regulation probabilities improved distinction of GO terms can be acquired. Community Cluster 3 GO terms enrichment depends on 5 genes out 90 genes present in structure the GO term enrichment is failed and less dependable according to GOrilla.

Table 10 Enriched GO Terms Gene Counts

Community	TOTAL GENE COUNT	ASSOCIATED GENE COUNT	CLUSTER DENSITY
Community 1	299	134	44.8 %
Community 2	152	72	47.4 %
Community 3	90	5	5.5 %

Table 11 Community 1 Biological Process Enrichment Results (GORRILLA)

GO Term	Description	P value	Enrichment (N, B, n, b)
GO:0071842	cellular component organization at cellular level	2.79E-5	1.24 (543,169,299,115)
GO:0016043	cellular component organization	1.01E-4	1.18 (543,217,299,141)
GO:0071841	cellular component organization or biogenesis at cellular level	1.67E-4	1.18 (543,213,299,138)
GO:0006281	DNA repair	3.93E-4	1.43 (543,47,299,37)
GO:0006996	organelle organization	4.67E-4	1.22 (543,143,299,96)
GO:0006259	DNA metabolic process	6.12E-4	1.31 (543,79,299,57)
GO:0006974	response to DNA damage stimulus	6.88E-4	1.37 (543,57,299,43)
GO:0022402	cell cycle process	6.97E-4	1.25 (543,112,299,77)
GO:0071840	cellular component organization or biogenesis	7.64E-4	1.15 (543,226,299,143)

We find that 44.8 % of the genes present in community cluster 1 is associated with the enriched GO terms presented in Table 11.

Table 12 Community 2 Biological Process Enrichment Results (GORILLA)

GO Term	Description	P value	Enrichment (N, B, n, b)
GO:0006357	regulation of transcription from RNA polymerase II promoter	2.18E-4	1.65 (543,76,152,35)
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	2.22E-4	1.97 (543,38,152,21)
GO:0006810	transport	2.92E-4	1.64 (543,74,152,34)
GO:0016070	RNA metabolic process	3.03E-4	1.47 (543,119,152,49)
GO:0006351	transcription, DNA-dependent	3.36E-4	1.52 (543,101,152,43)
GO:0032774	RNA biosynthetic process	3.36E-4	1.52 (543,101,152,43)
GO:0010628	positive regulation of gene expression	5.85E-4	1.79 (543,48,152,24)
GO:0045893	positive regulation of transcription, DNA-dependent	5.85E-4	1.79 (543,48,152,24)
GO:0051234	establishment of localization	7.6E-4	1.58 (543,77,152,34)
GO:0010557	positive regulation of macromolecule biosynthetic	8.67E-4	1.75 (543,49,152,24)
GO:0051254	positive regulation of RNA metabolic process	8.67E-4	1.75 (543,49,152,24)

We find that 47.4 % of the genes present in community cluster 2 is associated with the enriched GO terms presented in Table 12.

Table 13 Community 3 Biological Process Enrichment Results

GO Term	Description	P value	Enrichment (N, B, n, b)
GO:0009066	aspartate family amino acid metabolic process	5.93E-4	5.03 (543,6,90,5)

Density of community cluster 3 is low. GO term enrichment is significant for only 5 genes so we cannot conclude members of community 3 participate in similar kind of biological processes with using GOrilla. To further analyze community cluster 3 we use DAVID for detailed GO term enrichment analysis with high classification stringency. We identify 4 annotational clusters with enrichment scores higher than 1. Enriched GO terms have P values bigger than 0.001 which we are used with GORILLA.

Table 14 Community 3 Annotational Clusters Enrichment (DAVID)

Annotation Cluster (AC)	Enrichment
AC 1	1.736
AC 2	1.337
AC 3	1.134
AC 4	1.025

GO BP term enrichments are presented in Table 14 for detected annotational clusters (AC) with DAVID. Enriched GO BP terms indicate a portion of community cluster 3 members highly associated with Cellular Amino Acid Biosynthesis, Organic Acid Metabolism and Cellular Nitrogen Compound Biosynthesis. The count of genes associated with those terms is 13. Participation of genes to ACs is 27.8%.

Table 15 Community 3 Annotation Cluster 1

Annotation Cluster 1			
GO Term	**Description**	**P value**	**Enrichment**
GO:0009309	amine biosynthetic process	0.008826	3.321569
GO:0008652	cellular amino acid biosynthetic process	0.008826	3.321569
GO:0046394	carboxylic acid biosynthetic process	0.023071	2.530719
GO:0016053	organic acid biosynthetic process	0.023071	2.530719
GO:0044271	nitrogen compound biosynthetic process	0.050284	2.049882

As can be observed from individual AC (Tables 10 to 12) even if overall total enrichment scores are low, fold enrichment of individual GO BP terms are high.

Table 16 Community 3 Annotation Cluster 2

Annotation Cluster 2			
GO Term	**Description**	**P value**	**Enrichment**
GO:0009309	amine biosynthetic process	0.008826	3.321569
GO:0008652	cellular amino acid biosynthetic process	0.008826	3.321569
GO:0006520	cellular amino acid metabolic process	0.050284	2.049882
GO:0044106	cellular amine metabolic process	0.050284	2.049882

GO:0006519	cellular amino acid and derivative metabolic process	0.06248	1.971041
GO:0019752	carboxylic acid metabolic process	0.086639	1.692846
GO:0043436	oxoacid metabolic process	0.086639	1.692846
GO:0006082	organic acid metabolic process	0.086639	1.692846

Table 17 Community 3 Annotation Clusters

Annotation Cluster 3			
GO Term	**Description**	**Pvalue**	**Enrichment**
GO:0009066	aspartate family amino acid metabolic process	0.009633	4.745098
GO:0006555	methionine metabolic process	0.038776	4.555294
GO:0000097	sulfur amino acid biosynthetic process	0.038776	4.555294
GO:0000096	sulfur amino acid metabolic process	0.067918	3.796078
GO:0009086	methionine biosynthetic process	0.140917	4.270588
GO:0009067	aspartate family amino acid biosynthetic process	0.140917	4.270588
GO:0044272	sulfur compound biosynthetic process	0.146422	2.847059
GO:0006790	sulfur metabolic process	0.294507	2.070588

By comparing the Tables 4-5 with Tables 9-10 we can conclude that community structures are enriched more specific GO terms than K-means. As can be seen from

Table 8 most enriched GO terms present in K-means cluster 3 are: GO Cellular Processes and GO Biological Processes These terms are broad terms and very high portions of genes shares these terms. After elimination of these terms K-means cluster 3 associated gene count drops significantly (46.3 %). If we more closely look to community cluster 3 with DAVID (Tables 9 to 12). For lower P values but with good enrichment levels 27.8 % of genes participate in annotation clusters. Our results indicate Communities present in GRNN are better correlated with biological data than K means clusters because it enriches more similar GO terms does not require cluster count initiation and often creates more dense cluster regarding to genes participating same GO terms.

2. CONCLUSION

By the increase of publicly available time series micro array experiments, large amount of data about differential gene expressions are accumulated. Since expression levels are directly related with activity of genes, we can identify the regulation process between the genes in to some extent since pre and post modification over mRNA and TFs can alter the activity of the genes.

In this study we modeled gene regulations with HMM models. In the first part of this study we try to optimize our HMM models for our training sets which are related to the cell cycle. We compared our HMMs effectiveness with the Pearson Correlation which is common approach to detect regulations between the genes. HMM surpass the Pearson Correlation significantly in our experiments between 19.2 % to 29.6 % in terms of average detected gene counts present in Top-K analysis of data.

In the second part we use the community structure property of the networks to cluster genes according to GO biological process terms. In theory since the gene regulatory networks have modular framework, we aim to identify these modular frameworks with community structure property. Results indicate that even for less dependable networks made up from crude estimation of gene regulatory relations with HMM, we can produce biologically meaningful clusters. We show that communities present in HMM derived GRNNs are better correlated with biological data than conventional K-means clustering by generating more specific clusters and often higher numbers of genes have similar GO terms for equal number of clusters for highly cell cycle regulated genes.

For future work since many co expressed genes does not have to be co-regulated because of different promoter sequences and also even the genes with common promoter sequences does not have to be co-regulated because of suppressors and enhancers activities and histone modifications. HMM can detect this features and performance can be increased with integration promoter sequence alignments.

APPENDIX

Appendix 1 K-means Clusters for RT1 GRNN genes (Systematic Name)

Cluster 1

YAL005C	YCR005C	YDR343C	YFL064C	YHR154W	YKL042W	YLR256W	YMR076C	YNL300W
YAL053W	YCR065W	YDR481C	YFL065C	YHR218W	YKL045W	YLR270W	YMR078C	YNL309W
YAR007C	YCR089W	YDR488C	YFL067W	YHR219W	YKL066W	YLR273C	YMR14W	YNL310C
YAR008W	YDL003W	YDR501W	YFR053C	YIL026C	YKL067W	YLR274W	YMR163C	YNL312W
YAR071W	YDL018C	YDR507C	YGL032C	YIL066C	YKL101W	YLR303W	YMR17W	YNL334C
YBL005W	YDL020C	YDR516C	YGL037C	YIL101C	YKL104C	YLR304C	YMR18W	YNL339C
YBL005W	YDL095W	YDR518W	YGL038C	YIL140W	YKL113C	YLR313C	YMR19W	YNR001C
YBL031W	YDL101C	YDR528W	YGL061C	YIL177C	YKL127W	YLR326W	YMR25W	YNR044W
YBL035C	YDL102W	YDR545W	YGL163C	YJL073W	YKL163W	YLR342W	YMR25W	YOL007C
YBL109W	YDL105W	YEL047C	YGL200C	YJL074C	YKR013W	YLR382C	YMR253C	YOL017W
YBL111C	YDL124W	YEL071W	YGL207W	YJL079C	YKR090W	YLR437C	YMR261C	YOL147C
YBL112C	YDL127W	YEL075C	YGL253W	YJL089W	YLL002W	YLR462W	YMR305C	YOR027W
YBL113C	YDL164C	YEL076C	YGR014W	YJL115W	YLL022C	YLR463C	YNL037C	YOR028C
YBR070C	YDL248W	YEL076C	YGR109C	YJL116C	YLL066C	YLR464W	YNL088W	YOR033C
YBR071W	YDR077W	YEL077C	YGR140W	YJL155C	YLL067C	YLR465C	YNL102W	
YBR072W	YDR085C	YER001W	YGR142W	YJL164C	YLR032W	YLR466W	YNL134C	
YBR073W	YDR097C	YER053C	YGR152C	YJL173C	YLR049C	YLR467W	YNL165W	
YBR087W	YDR113C	YER070W	YGR189C	YJL187C	YLR050C	YML01W	YNL166C	
YBR088C	YDR144C	YER095W	YGR221C	YJL225C	YLR103C	YML02W	YNL192W	
YBR161W	YDR216W	YER111C	YGR286C	YJR006W	YLR121C	YML10W	YNL206C	
YBR275C	YDR222W	YER118C	YGR296W	YJR010W	YLR151C	YML110C	YNL231C	
YBR296C	YDR253C	YER150W	YHL021C	YJR030C	YLR183C	YML133C	YNL262W	
YCL024W	YDR279W	YER189W	YHL049C	YJR043C	YLR212C	YMR008C	YNL263C	
YCL040W	YDR307W	YER190W	YHL050C	YJR060W	YLR217W	YMR01W	YNL273W	
YCL061C	YDR342C	YFL008W	YHR092C	YKL035W	YLR234W	YMR04W	YNL289W	

Cluster 2

YAL059W	YJL078C	YEL032W	YLR079W	YDL039C	YKL185W	YGL028C	YNL141W	YIL009W
YBR067C	YJL157C	YEL040W	YLR194C	YDL227C	YKL209C	YGL055W	YNL327W	YJL044C
YBR158W	YJL159W	YER124C	YLR413W	YDR033W	YKR042W	YGL255W	YNR067C	YHR005C
YBR267W	YKL009W	YFL026W	YLR452C	YDR309C	YKR099W	YGR041W	YOL124C	YOR342C
YCL064C	YKL172W	YFL027C	YNL078W	YDR379W	YLR074C	YGR044C	YOR315W	YHR143W

YIL127C	YHR08W	YPL158C

Cluster 3

YAL007C	YMR006C	YGL012W	YKL164C	YDR207C	YOR246C	YJL051W	YML064C
YAL022C	YMR021C	YGL013C	YKL165C	YDR213W	YOR247W	YJL056C	YML065W
YAL023C	YMR031C	YGL021W	YKR010C	YDR219C	YOR248W	YJL084C	YML085C
YAL039C	YMR032W	YGL035C	YKR098C	YDR224C	YOR307C	YJL092W	YML099C
YAL040C	YMR037C	YGL073W	YLL028W	YDR225W	YOR313C	YJL110C	YML119W
YAR018C	YMR042W	YGL101W	YLL032C	YDR261C	YOR326W	YJL118W	YML125C
YBL002W	YMR043W	YGL116W	YLR014C	YDR276C	YOR372C	YJL134W	YMR001C
YBL003C	YMR058W	YGL201C	YLR045C	YDR297W	YPL032C	YJL158C	YMR003W
YBL004W	YMR070W	YGL209W	YLR056W	YDR310C	YPL075W	YJL183W	YKL043W
YBL005W	YMR075W	YGL216W	YLR131C	YDR325W	YPL089C	YJL194W	YKL052C
YBL009W	YMR145C	YGL237C	YLR154C	YDR423C	YPL116W	YJR003C	YKL062W
YBL023C	YMR164C	YGL254W	YLR176C	YDR451C	YPL127C	YJR053W	YKL069W
YBL032W	YMR183C	YGR035C	YLR180W	YDR463W	YPL128C	YJR054W	YKL096W
YBL061C	YMR190C	YGR086C	YLR190W	YEL017W	YPL141C	YJR092W	YKL109W
YBL063W	YMR198W	YGR092W	YLR210W	YEL042W	YPL202C	YJR132W	YKL112W
YBL064C	YMR215W	YGR098C	YLR228C	YEL061C	YPL209C	YKL004W	YPL255W
YBR008C	YMR307W	YGR099W	YLR254C	YER003C	YPL242C	YKL008C	YPR019W
YBR009C	YNL015W	YGR108W	YLR297W	YER028C	YPL248C	YKL015W	YPR034W
YBR010W	YNL030W	YGR113W	YLR300W	YER032W	YPL253C	YKL038W	YPR065W
YBR015C	YNL031C	YGR143W	YLR373C	YHL028W	YOL091W	YER040W	YPR106W
YBR038W	YNL056W	YGR151C	YLR380W	YHR023W	YOL158C	YER109C	YPR119W
YBR049C	YNL057W	YGR176W	YLR451W	YHR061C	YOR018W	YFL007W	YPR149W
YBR054W	YNL058C	YGR177C	YLR455W	YHR086W	YOR025W	YFL021W	YER037W
YBR078W	YNL068C	YGR188C	YML007W	YHR127W	YOR058C	YFL031W	YDL055C
YBR083W	YNL103W	YGR230W	YML034W	YHR135C	YOR073W	YGL008C	YDL056W
YBR086C	YNL111C	YGR279C	YML052W	YHR136C	YOR084W	YIL129C	YDL138W
YBR092C	YNL126W	YHL026C	YML058W	YHR137W	YOR099W	YIL131C	YDL155W
YBR093C	YNL145W	YCR024C	YDR133C	YHR143W	YOR113W	YIL132C	YDR055W
YBR133C	YNL160W	YCR039C	YDR146C	YHR146W	YOR127W	YIL141W	YDR089W
YBR138C	YNL176C	YCR041W	YDR150W	YHR151C	YOR140W	YIL158W	YDR123C
YBR139W	YNL216W	YCR042C	YDR179C	YHR152W	YOR153W	YIL159W	YIL050W
YBR157C	YNL238W	YCR096C	YDR191W	YHR178W	YOR162C	YIR010W	YIL106W
YBR200W	YNL283C	YOL014W	YCL014W	YHR206W	YOR188W	YIR023W	YIL119C
YBR202W	YNR009W	YOL019W	YCL055W	YIL047C	YOR229W	YJL034W	YIL122W
YBR243C	YOL012C	YOL030W	YCL063W	YOL067C	YOR230W	YOR233W	YIL123W

Appendix 2 GRNN Community Clusters (Systematic Name)

Community Cluster 1

YKL069W	YER118C	YFL031W	YHR154W	YDR033W	YPL209C	YBL003C	YJL116C
YFL064C	YGR221C	YOR114W	YMR078C	YFR053C	YDR307W	YHL028W	YBL009W
YDR113C	YHR218W	YHR023W	YDR077W	YGR140W	YNL160W	YLL032C	YBR139W
YML052W	YDR518W	YIL158W	YPL124W	YLR032W	YJL187C	YJL079C	YJL194W
YNL058C	YPL127C	YLR270W	YNL283C	YLL002W	YHR152W	YGR188C	YDR146C
YMR305C	YLR342W	YNL057W	YGL116W	YOR247W	YFL065C	YBL002W	YIL132C
YML119W	YOR315W	YNL273W	YMR215W	YFL027C	YML027W	YGR092W	YPR135W
YBL111C	YMR032W	YPR119W	YER001W	YDR297W	YFL008W	YOR058C	YJL115W
YEL075C	YBL112C	YNL300W	YNL262W	YOR373W	YCL061C	YKL066W	YOL091W
YLR313C	YCR065W	YNL166C	YDR528W	YKL113C	YMR307W	YML085C	YIL123W
YER189W	YBR092C	YGR099W	YLR217W	YOL030W	YDR488C	YML058W	YML064C
YBR093C	YEL076C	YIL026C	YOL014W	YMR076C	YHR178W	YMR189W	YLR462W
YNL031C	YLR190W	YBR072W	YBL031W	YLR212C	YKL165C	YOL017W	YJL225C
YOR127W	YLR121C	YKL008C	YLR467W	YDR150W	YER003C	YPL153C	YMR251W
YCL063W	YKL101W	YLR131C	YKL096W	YGR041W	YBR078W	YKL127W	YGR152C
YML065W	YGR189C	YOR347C	YGL061C	YGL073W	YEL061C	YKR010C	YOR066W
YLR079W	YLR273C	YMR058W	YNL037C	YOR113W	YDR501W	YER150W	YCL067C
YPR019W	YDR222W	YGL101W	YMR164C	YOR374W	YNL206C	YEL071W	YEL017W
YMR179W	YNL015W	YOR027W	YDR342C	YEL032W	YGR098C	YNL309W	YNL289W
YDR309C	YNL339C	YDR219C	YNL088W	YBR083W	YBL035C	YLR463C	YMR031C
YHR086W	YDR545W	YLR183C	YDL101C	YDR133C	YDL102W	YLR194C	YLL066C
YLR466W	YPR141C	YER070W	YLR154C	YDL003W	YPL256C	YPL253C	YPL255W
YKL209C	YGR109C	YLR050C	YBR049C	YMR144W	YJL164C	YDR089W	YIL106W
YDL127W	YBR071W	YKL045W	YNR044W	YLR049C	YDR343C	YBR070C	YJL073W
YDR097C	YLL022C	YIL177C	YER032W	YML133C	YNL165W	YJR006W	YLR326W
YHR092C	YNL068C	YDR253C	YKL042W	YGL013C	YBR202W	YBL109W	YGL038C
YPL267W	YHL049C	YGR296W	YCL014W	YCR042C	YPL283C	YPL163C	YIL119C
YLR210W	YNL145W	YHR005C	YGR286C	YOR188W	YAR071W	YBR275C	YMR163C
YDL124W	YHL050C	YJL155C	YBR073W	YOR073W	YMR190C	YIL159W	YDR507C
YAL007C	YLR103C	YEL077C	YOL007C	YCR005C	YBR010W	YOR028C	YMR070W
YJL159W	YGL200C	YER111C	YLR151C	YOL147C	YIL131C	YGL253W	YML034W
YIL140W	YLR056W	YMR011W	YML102W	YGL032C	YGL201C	YNL192W	YAR008W
YHR219W	YHR137W	YBR067C	YIL141W	YKL067W	YMR003W	YBR009C	YGL037C
YBR088C	YNL111C	YLR045C	YBR015C	YCL040W	YKL163W	YPL116W	YNL263C
YDL155W	YMR199W	YLL067C	YBL113C	YCL055W	YNL176C	YDL164C	YNL126W
YOR248W	YKL112W	YLR274W	YKL104C	YCL024W	YLR300W	YJL051W	YGL216W

YNL312W	YDL018C	YDR279W	YER190W	YJR030C	YDR055W	YNL030W	YAR007C
YHL021C	YKR042W	YOR033C	YKR013W	YKR090W	YMR048W	YMR250W	YNL078W
YJL158C	YER095W	YOR074C	YFL067W	YMR008C	YNL310C	YLR437C	YNL102W
YLR464W							

Community Cluster 2

YDL039C	YOR153W	YHL026C	YBR200W	YJL118W	YEL042W	YOR246C	YMR043W
YGR035C	YDR463W	YOR233W	YML012W	YGR113W	YPR035W	YLR234W	YHR127W
YOR313C	YEL047C	YDL055C	YDL105W	YGL008C	YBL032W	YKL172W	YLR455W
YGL163C	YJL034W	YOR099W	YER040W	YBR008C	YLR465C	YML125C	YGR143W
YOR229W	YGL012W	YML099C	YIR010W	YPL075W	YLR380W	YNL231C	YKL052C
YGR086C	YPR106W	YOL019W	YOR342C	YLR382C	YHR146W	YOL067C	YBR086C
YKL009W	YKL043W	YAL039C	YDL095W	YGR176W	YNL216W	YOR273C	YDR207C
YMR075W	YAL059W	YPL014W	YIL050W	YDR379W	YJR003C	YCR096C	YPL057C
YDR261C	YLR297W	YER037W	YAL005C	YJL134W	YOR162C	YDR144C	YHR151C
YFL007W	YPL032C	YNL334C	YER109C	YIL122W	YML110C	YDR325W	YDR481C
YMR006C	YJR053W	YGR177C	YDR213W	YMR261C	YPL061W	YMR253C	YKL015W
YAL023C	YOR230W	YJL183W	YBR161W	YMR021C	YDL056W	YPL265W	YKR098C
YGR230W	YCR041W	YOR372C	YJR132W	YLR254C	YGR151C	YNL134C	YOR307C
YMR037C	YKL062W	YPL128C	YJR043C	YNR009W	YGL207W	YLR413W	YGR279C
YOR140W	YLR176C	YGL209W	YBR087W	YOL158C	YAL022C	YHR206W	YLR074C
YOR326W	YHR136C	YIL009W	YMR042W	YDR516C	YJL084C	YGL035C	YGL237C
YAL040C	YDR451C	YLR451W	YCR039C	YJL044C	YDR179C	YLR014C	YBL023C
YDR276C	YNL103W	YDL138W	YBR243C	YAL053W	YDR423C	YDR191W	YPL221W
YBR054W	YPR034W	YBR157C	YOR084W	YIR023W	YOL012C	YPL202C	YKL109W
YMR183C	YNL238W	YLL028W	YKL004W				

Community Cluster 3

YBR038W	YER053C	YIL101C	YBL063W	YPL158C	YHR135C	YKL164C	YGL028C
YER028C	YIL129C	YMR198W	YJL157C	YPR075C	YNL141W	YJL173C	YDL248W
YAR018C	YKR099W	YLR304C	YCL064C	YHR084W	YOR018W	YPL141C	YKL185W
YBR267W	YJR092W	YDR225W	YKL038W	YOR025W	YEL040W	YLR228C	YCR024C
YNR001C	YBL004W	YGL254W	YBR158W	YDR085C	YLR303W	YIL047C	YNL056W
YDR123C	YPL242C	YJL074C	YGR014W	YPL248C	YBL005W	YGR044C	YML007W
YPR149W	YLR256W	YLR180W	YJL089W	YOL124C	YJR060W	YBR133C	YNR067C
YGL255W	YIL127C	YJR010W	YDR310C	YDR216W	YFL026W	YER124C	YNL327W
YJL110C	YPR065W	YMR001C	YBL061C	YGR142W	YJL092W	YCR089W	YBL064C
YHR061C	YIL066C	YDR224C	YDL227C	YLR452C	YLR373C	YGR108W	YGL021W
YBR138C	YJL056C	YDL020C	YJL078C	YPL089C	YMR145C	YBR296C	YKL035W

YGL055W	YJR054W

BIBLIOGRAPHY

1. Tegner, J., et al., *Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5944-9.
2. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* Journal of Computational Biology, 2000. **7**(3-4): p. 601-620.
3. Davidich, M.I. and S. Bornholdt, *Boolean Network Model Predicts Cell Cycle Sequence of Fission Yeast.* Plos One, 2008. **3**(2).
4. Ruan, J., A.K. Dean, and W. Zhang, *A general co-expression network-based approach to gene expression analysis: comparison and applications.* BMC Syst Biol, 2010. **4**: p. 8.
5. Horvath, S. and J. Dong, *Geometric interpretation of gene coexpression network analysis.* Plos Computational Biology, 2008. **4**(8): p. e1000117.
6. Zhu, D., et al., *Network constrained clustering for gene microarray data.* Bioinformatics, 2005. **21**(21): p. 4014-20.
7. Weston, D.J., et al., *Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants.* BMC Syst Biol, 2008. **2**: p. 16.
8. Kim, J.H., H.Y. Kim, and Y.S. Lee, *A novel method using edge detection for signal extraction from cDNA microarray image analysis.* Exp Mol Med, 2001. **33**(2): p. 83-8.
9. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Molecular Biology of the Cell, 1998. **9**(12): p. 3273-97.
10. Ernst, J., G.J. Nau, and Z. Bar-Joseph, *Clustering short time series gene expression data.* Bioinformatics, 2005. **21 Suppl 1**: p. i159-68.
11. Newman, A.M. and J.B. Cooper, *AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number.* BMC Bioinformatics, 2010. **11**: p. 117.
12. Perkins, A.D. and M.A. Langston, *Threshold selection in gene co-expression networks using spectral graph theory techniques.* BMC Bioinformatics, 2009. **10**.
13. Yogurtcu, O.N., E. Erzin, and A. Gursoy, *Extracting gene regulation information from microarray time-series data using hidden Markov models.* Computer and Information Sciences - ISCIS 2006, Proceedings, 2006. **4263**: p. 144-153.
14. Clauset, A., M.E.J. Newman, and C. Moore, *Finding community structure in very large networks.* Physical Review E, 2004. **70**(6).
15. Newman, M.E.J. and M. Girvan, *Finding and evaluating community structure in networks.* Physical Review E, 2004. **69**(2).
16. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. **10**.
17. Lempicki, R.A., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.* Nucleic Acids Research, 2007. **35**: p. W169-W175.
18. Gifford, D.K., et al., *Computational discovery of gene modules and regulatory networks.* Nature Biotechnology, 2003. **21**(11): p. 1337-1342.

19. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.
20. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2907-12.
21. Enserink, J.M., *DNA Replication-Current Advances*, in *Cell Cycle Regulation of DNA Replication in S. cerevisiae*, H. Seligmann, Editor. 2011, InTech.
22. Enserink, J.M. and R.D. Kolodner, *An overview of Cdk1-controlled targets and processes.* Cell Div, 2010. **5**: p. 11.
23. Pines, J., *The Cell-Cycle Kinases.* Seminars in Cancer Biology, 1994. **5**(4): p. 305-313.
24. Russo, A.A., P.D. Jeffrey, and N.P. Pavletich, *Structural basis of cyclin-dependent kinase activation by phosphorylation.* Nature Structural Biology, 1996. **3**(8): p. 696-700.
25. Pavletich, N.P., *Mechanisms of cyclin-dependent kinase regulation: Structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors.* Journal of Molecular Biology, 1999. **287**(5): p. 821-828.
26. Booher, R.N., R.J. Deshaies, and M.W. Kirschner, *Properties of Saccharomyces cerevisiae wee1 and its differential regulation of p34CDC28 in response to G1 and G2 cyclins.* Embo Journal, 1993. **12**(9): p. 3417-26.
27. Russell, P. and P. Nurse, *cdc25+ functions as an inducer in the mitotic control of fission yeast.* Cell, 1986. **45**(1): p. 145-53.
28. Ferrell, J.E., Jr., T.Y. Tsai, and Q. Yang, *Modeling the cell cycle: why do certain circuits oscillate?* Cell, 2011. **144**(6): p. 874-85.
29. de Bruin, R.A., et al., *Cln3 activates G1-specific transcription via phosphorylation of the SBF bound repressor Whi5.* Cell, 2004. **117**(7): p. 887-98.
30. Costanzo, M., et al., *CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast.* Cell, 2004. **117**(7): p. 899-913.
31. Charvin, G., et al., *Origin of irreversibility of cell cycle start in budding yeast.* PLoS Biol, 2010. **8**(1): p. e1000284.
32. Schwob, E., et al., *The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae.* Cell, 1994. **79**(2): p. 233-44.
33. Nash, P., et al., *Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication.* Nature, 2001. **414**(6863): p. 514-21.
34. de Bruin, R.A., et al., *Constraining G1-specific transcription to late G1 phase: the MBF-associated corepressor Nrm1 acts via negative feedback.* Mol Cell, 2006. **23**(4): p. 483-96.
35. Koch, C., et al., *Switching transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound transcription factor SBF (Swi4/Swi6) at start, whereas Clb/Cdc28 kinases displace it from the promoter in G2.* Genes Dev, 1996. **10**(2): p. 129-41.
36. Stinchcomb, D.T., K. Struhl, and R.W. Davis, *Isolation and characterisation of a yeast chromosomal replicator.* Nature, 1979. **282**(5734): p. 39-43.
37. Bell, S.P. and B. Stillman, *ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex.* Nature, 1992. **357**(6374): p. 128-34.
38. Araki, H., *Cyclin-dependent kinase-dependent initiation of chromosomal DNA replication.* Curr Opin Cell Biol, 2010. **22**(6): p. 766-71.

39. Schwob, E. and K. Nasmyth, *CLB5 and CLB6, a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae.* Genes Dev, 1993. **7**(7A): p. 1160-75.
40. Moll, T., et al., *The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SWI5.* Cell, 1991. **66**(4): p. 743-58.
41. Drury, L.S., G. Perkins, and J.F. Diffley, *The Cdc4/34/53 pathway targets Cdc6p for proteolysis in budding yeast.* Embo Journal, 1997. **16**(19): p. 5966-76.
42. Liku, M.E., et al., *CDK phosphorylation of a novel NLS-NES module distributed between two subunits of the Mcm2-7 complex prevents chromosomal rereplication.* Molecular Biology of the Cell, 2005. **16**(10): p. 5026-39.
43. Nguyen, V.Q., C. Co, and J.J. Li, *Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms.* Nature, 2001. **411**(6841): p. 1068-73.
44. Putnam, C.D., E.J. Jaehnig, and R.D. Kolodner, *Perspectives on the DNA damage and replication checkpoint responses in Saccharomyces cerevisiae.* DNA Repair (Amst), 2009. **8**(9): p. 974-82.
45. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Research, 2010. **38**(Database issue): p. D355-60.
46. Gerstein, M., et al., *Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data.* Bioinformatics, 2003. **19**(15): p. 1917-1926.
47. Yeung, K.Y., M. Medvedovic, and R.E. Bumgarner, *From co-expression to co-regulation: how many microarray experiments do we need?* Genome Biology, 2004. **5**(7).
48. Qiu, P., *Recent advances in computational promoter analysis in understanding the transcriptional regulatory network.* Biochemical and Biophysical Research Communications, 2003. **309**(3): p. 495-501.
49. Panda, S., et al., *An array of insights: application of DNA chip technology in the study of cell biology.* Trends Cell Biol, 2003. **13**(3): p. 151-6.
50. Noble, W.S., et al., *The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle.* Genes & Development, 2006. **20**(16): p. 2266-2278.
51. Brunak, P.B.S., *Bioinformatics: The machine learning approach,2nd edition*, ed. T. Dietterich. 2001, London: The MIT Press.
52. Monteiro, P.T., et al., *YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae.* Nucleic Acids Research, 2008. **36**(Database issue): p. D132-6.
53. Abdulrehman, D., et al., *YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface.* Nucleic Acids Research, 2011. **39**(Database issue): p. D136-40.
54. Brunak, S., et al., *Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results.* Nucleic Acids Research, 2010. **38**: p. D699-D702.
55. Brunak, S., et al., *Cyclebase.org - a comprehensive multi-organism online database of cell-cycle experiments.* Nucleic Acids Research, 2008. **36**: p. D854-D859.
56. Young, S., et al., *The HTK Book*, S. Young, Editor. 2006, Cambridge University Engineering Department.

57. Obayashi, T., et al., *COXPRESdb: a database of coexpressed gene networks in mammals*. Nucleic Acids Research, 2008. **36**(Database issue): p. D77-82.
58. Ferres, L., et al., *Improving Accessibility to Statistical Graphs: The iGraph-Lite System.* Assets'07: Proceedings of the Ninth International Acm Sigaccess Conference on Computers and Accessibility, 2007: p. 67-74.
59. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
60. Supek, F., et al., *REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.* Plos One, 2011. **6**(7): p. e21800.
61. Schliep, A., A. Schonhuth, and C. Steinhoff, *Using hidden Markov models to analyze gene expression time course data.* Bioinformatics, 2003. **19 Suppl 1**: p. i255-63.

VITA

Osman Mahmut Eryurt born in 1984, Ankara,Turkey. He had been in Özel Samanyolu Anadolu Lisesi for high school education. He received his BS degree in 2008, in Molecular Biology at Bilkent University. He worked as a teaching and research assistant in Koç University during 2008-2011.