Controllability of Gene Regulatory Networks

undertaken at

Department of Electrical Engineering

under the guidance of

Dr. Sumeet Agarwal

Submitted by

Rishabh Dudeja
2011EE10477

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ELECTRICAL ENGIEERING

Department of Electrical Engineering
Indian Institute of Technology, Delhi.
May 2015

Contents

1 Acknowledgment 2

2 Introduction 3

2.1 Transcription Networks as Dynamical Systems 3

2.2 Controllability of Transcription Networks 4

3 Transcription Networks are hard to Control 6

3.1 Classical Control Theory[11] 6

3.2 Structural Control Theory 6

3.3 Controllability of Real Biological Networks using Structural Controllability 7

4 Controllability of Time Varying Networks 8

4.1 Time Varying Linear Systems as Approximations to Non-linear Systems 8

4.2 Control Centrality 9

4.3 Analysis of Real Time Varying Networks 10

4.3.1 Yeast 10

4.3.2 Drosophila 11

5 Modularity and Effective Dimension of the Expression States 18

5.1 Review of Modularity 18

5.2 Characterizing Effective Dimension of Modules using Actual Expression Data 18

5.3 Characterizing Efficiency of Control of Individual Modules using Structural Controllability 19

6 Control Properties of Coarsened Networks 23

6.1 Description of Experiments 23

6.2 Results on the Drosophila Dataset 23

7 Conclusion and Criticism 29

7.1 Conclusions 29

7.2 Criticism and Future Work 30

1 Acknowledgment

I would like to thank my supervisor Dr. Sumeet Agarwal for constant guidance, support, patience and interesting ideas. The project would not have been possible without the weekly meetings we had over the past 1.5 years. I would additionally like to thank Prof. Shaunak Sen. Meetings with him clarified alot of control theoretic ideas in the project. I am grateful to Prof Sophie Lebre who graciously provided inferred networks from her paper and saved us a lot of pain in trying to infer the networks on our own using different algorithms. I would like to express my gratitude to my thesis committee consisting of Prof Santanu Chaudhary, Prof Jayadeva, Prof Sumantra Dutta Roy, Prof Brejesh Lall, Prof M Hanmadlu, Prof Hiranmay Ghosh and Prof SMK Rahman and reader Prof Santanu Chaudhary for their valuable feedback.

2 Introduction

Any organism has to produce the required amounts of proteins in response to environmental signals. The network that controls how much protein is produced is called the Transcription Network or the Gene Regulatory Network. The nodes in a gene regulatory network are the genes encoding for the various proteins. An directed edge $X \rightarrow Y$ represents that Transcription Factor X binds to the promoter of gene Y to begin the production/transcription of the protein Y. The most commonly used model for transcription is the Differential Equation model based on the Hill Equation [1]:

$$\frac{dY}{dt} = A\frac{X^n}{X^n + B^n} - CY \tag{1}$$

$$\frac{dY}{dt} = A\frac{B^n}{X^n + B^n} - CY \tag{2}$$

Where (1) is the model if X promotes the production of Y and (2) is the model if X inhibits the production of Y. Here A, B, n, C are model parameters. Ofcourse, a gene Y maybe regulated by multiple transcription factors: $X_1, X_2 \ldots X_M$. In such a case the more general model is:

$$\frac{dY}{dt} = Af(X_1, X_2 \ldots X_M) - CY \tag{3}$$

In biological systems a large number of functions like the boolean AND,OR are implemented.

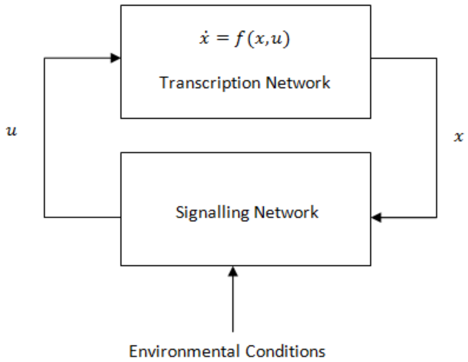
2.1 Transcription Networks as Dynamical Systems

In this project, we take an engineered systems view of Transcription Networks. We assume the transcription network can be modeled as a dynamical system:

$$\dot{x} = f(x, u) \tag{4}$$

Where x is the vector of expression of various genes and u are the inputs to the Transcription Network. Like an engineered system, one can think of the organism to have sensors to detect environmental states, a controller to compute the appropriate inputs u to the dynamical system to produce appropriate response x. In a cell, the task of sensing is done by the signal receptors on the surface, the controller is the signaling network which elicits appropriate response from the dynamical system i.e. transcription networks by suitably activating various genes. This is shown in Figure 1.

Figure 1: Modeling Transcription Networks as Control Systems



2.2 Controllability of Transcription Networks

A dynamical system $\dot{x} = f(x, u)$ is said to be controllable if it can be taken from any arbitrary initial state x_0 to any arbitrary final state x_{t_f} in a finite amount of time. A system is said to be efficiently controllable if it requires only a few number of inputs u to control it.

One would expect Biological Systems to be efficiently controllable. This is because of the intuition that since these systems have evolved for billions of years, they would have developed the useful property of being able to compute appropriate responses by sensing a few signals from the environment. Studying the control properties of these systems is additionally important if we wish to engineer them. For example, the control properties of the network would be crucial to answer questions like which genes should I target if I want to reprogram a skin cell to behave like a stem cell?

Past research has shown that Transcription Networks seem to be exceedingly hard to control: they require around 0.8 fraction of genes to receive external inputs u to achieve full controllability [7]. The purpose of this project is to resolve the conflict between this result and our intuition that transcription networks being evolved systems must be efficient in some way. To demonstrate the implausibility of the result, we did the following simple experiment. We tried to upper bound the maximum number of different inputs a cell could have by counting the genes having *Cell Signaling: GO0023052* as an annotation and compared it with the total number of annotated genes for the organism. We found that for a wide range of organisms the ratio was as low as $0.02 - 0.08$. Hence biological networks are able to control massive number of genes using just 10 percent of that number of inputs. What makes this efficient control possible?

Organism	**Number of Signalling Genes**	**Total Number of Genes**	**Ratio**
Yeast	154	7126	0.022
Drosophila	1336	17559	0.076
Human	5605	65803	0.085

In this project we attempt to address this problem using the following ideas:

- There is a trade-off between efficient control and robustness. If a network was extremely easy to control while it would require a simple sensing mechanism but at the same time it would be prone to going to arbitrary states if the few control points malfunction.

- Past research has reported that transcription network exhibits significant rewiring from one condition to another. One natural answer to the contradiction seems that biological networks are not trying to achieve full control in one go, but achieve control in phases. Any given phase is optimized to exercise efficient control over genes which are crucial in that phase and not all genes. We provide some evidence for this suggestion by functionally analyzing genes important for control in these rewiring networks.

- If at any given state, the whole expression space of the network is not reachable, which genes are constrained heavily and cannot be controlled independently? We provide an answer to this question by showing that modules are hard to control. This captures the intuition of modules as functionally coherent units that is their expression can't be set arbitrarily but is extremely constrained and behave as a unified entity.

- If the uncontrollability of the network arises because genes in a single community can't be independently controlled a natural question arises whether the modules when thought of as aggregate entities are controllable or not. We show by analyzing coarse-grained networks that modules seem to be much more efficiently controlled. This reinforces the idea of modules as independent units which operate in a plug and play manner.

The report is organized in the following way: In Chapter 1 we review Structural Control theory and reproduce the result that transcription networks are hard to control in Drosophila and Yeast datasets. In Chapter 2 we report are results on phase-wise control of transcription networks. In chapter 3 we report our results showing that the genes in the same module are hard to control and in Chapter 4 we show that however modules as aggregate entities are efficient to control. Finally in Chapter 5 we look at the big picture, identify potential pitfalls of the project and directions for future work.

3 Transcription Networks are hard to Control

In this chapter we first review definition of basic concepts of Structural Control Theory which is used to analyze networks whose structure is known but the exact parameters are unknown. We then describe the 3 datasets we've used during the project and reproduce the results of citation which show that Transcription Networks are hard to control. Throughout this chapter we make the assumption that the system is linear $\dot{x} = Ax + Bu$.

3.1 Classical Control Theory[11]

Definition 1 (Controllability of a System) *A system $\dot{x} = Ax + Bu$ is said to be controllable if for any given initial state x_i and any final state x_f there exists an input $u(t)$ that will transfer the system from x_i to x_f in a finite time.*

If the system parameters A, B are known before hand then the Kalman Rank Test enables us to find out how controllable the system is:

Theorem 1 (Kalman Rank Test) *A linear system $\dot{x} = Ax + Bu$ is controllable if the matrix $K = [BAB \dots A^{n-1}B]$ has full row rank. Infact the range space of K gives the reachable subspace of the system where the reachable subspace is defined as the set of all states that can be reached from the zero initial condition in a finite time by applying a suitable input.*

However, since there exists no biological or computational technique to perform system identification (estimating A and B) for large transcription networks the Kalman Rank Test can't be directly applied.

3.2 Structural Control Theory

The framework of structural controllability allows us to make claims about the control properties using just the structure of matrices A and B. Here knowing the structure means that we know which entries are allowed to be non-zero. Alternatively we know the structure of the graph of the system but do not know the weights on the edges. A system is said to be structurally controllable if it is possible to assign values to the non-zero entries of A, B such that the system is controllable in the usual sense. If a system is structurally controllable then it is controllable for almost all A, B with the same structure except a set of measure zero [6]. The following theorem gives a test for testing if given structural matrices A, B are structurally controllable. The test is a graph theoretic one:

Theorem 2 (Lin's Structural Controllability Theorem[6]) *The following statements are equivalent:*

- *The system A, B is structurally controllable.*

- *The graph representing system A, B has no dilations and no inaccessible nodes.*

- *The graph is spanned by a cactus*

Here dilations and inaccessible nodes are graph theoretic properties. A node is said to be inaccessible if there is no path from one of the input nodes to it. A set S is said to be a dilation if the number of nodes that point to it are less than the number of nodes in S. Finally a cactus is a elementary path starting from an input node and any number of cycles attached to it.

Now one way to characterize the difficulty of control of a system A is to characterize what is the minimum number of rows B must have to make A, B structurally controllable. This answer is provided by the minimum input theorem:

Theorem 3 (Minimum Inputs Theorem[7]) *The minimum number of inputs required to fully control a system A is $\min(1, N-M)$ where N is the number of nodes and M is the size of the maximum matching.*

3.3 Controllability of Real Biological Networks using Structural Controllability

In this section we reproduce the results of citation for Yeast[9] and Drosophila Datasets[5]. The minimum number of inputs required to completely control the network along with the number of nodes in the network and the fraction of signaling genes out of all genes known for the organism. This shows that complete control over transcription networks is implausible.

Organism	**No of Nodes(N)**	**Min Inputs(m)**	**m/n**	**Ratio of Signaling Genes**
Yeast	3459	3318	0.959	0.022
Drosophila	1704	1530	0.898	0.076

4 Controllability of Time Varying Networks

In [9] the authors consider the transcription network in Yeast. The consider the idea that all edges in the transcription network are not active all the time. They call an edge active in a given condition if both the source and sink of that edge are significantly expressed in that condition. Using these condition specific networks they show that transcription networks seem to undergo massive rewiring. However so far there has been no clarification as to the mechanism of this rewiring. In this chapter we first present a conceptual view that time-varying or rewiring linear systems are an approximation to non-linear systems. We then look at the control properties of these time varying networks using an idea called control centrality and show that phase specific networks are optimized to be controllable by genes crucial in that phase.

4.1 Time Varying Linear Systems as Approximations to Non-linear Systems

In this subsection we show that how two well known non-linear phenomena in Transcription networks namely combinatorial regulation and signaling can be approximated as changes in the structure of the network or rewiring.

Consider a gene z being regulated jointly by genes x and y. We consider the AND combinatorial regulation as an example. That is the transcription of z is maximum when both x and y are expressed. The model can be written as (after transforming x,y,z so that appropriate constants are set to 1):

$$\dot{z} = \frac{x^n}{1+x^n}\frac{y^n}{1+y^n} - \tau z \tag{5}$$

Linearizing about point x_0, y_0, z_0:

$$\dot{z} = \frac{n x_0^{n-1} y_0^n}{(1+x_0^n)^2(1+y_0^n)} x + \frac{n x_0^n y_0^{n-1}}{(1+x_0^n)(1+y_0^n)^2} y - \tau z \tag{6}$$

Now consider the following regimes:

- Suppose x_0 is significantly expressed (much higher than 1) then the first term can be approximately set to zero. The gene z is now regulated only by y now.
- Similarly if y_0 is significantly larger than 1, the gene z is regulated only by x now.
- If both of x_0 and y_0 are significantly more than one the effects of regulation from neither of them are felt because of saturation.
- If none of these approximations hold then co-regulatory effects of x and y are seen.

Note that even if we were to consider regulation by a single gene x in regimes where x is significantly high then also such an effect of loss of regulation by x would be seen. Now consider the phenomena of signaling. Consider the case of a single gene x regulating a gene z. However to induce production of gene z, x must be activated by a signaling molecule s to produce activated $x*$. Since signaling dynamics are much faster than transcription dynamics this reaction can be assumed to be in steady state while modeling transcription dynamics. Under this assumption the amount of activated $x*$ is given as a function of the signal s present and total x transcribed:

$$x* = \frac{xs}{k+s} \tag{7}$$

This is called the Michelson Mentin Equation. Now consider the two regimes:

- Signal is saturated. In this case $x* \approx x$. That is the whole amount of x is visible to the transcription network.
- Signal is low. In this case a reduced amount $x\frac{s}{k}$ is visible to the transcription network. In particular when no signal is present even though x is present in the system, its not visible to the transcription network and hence the edge between x and z has been switched off.

Hence we see that a variety of non-linear phenomena can give rise to rewiring in transcription networks.

4.2 Control Centrality

Given a structural system A, B, one would like to characterize the dimension of its controllable subspace i.e. the dimension of the expression space the network is trapped in. Poljak [12] showed that almost all matrices A', B' with the structure given by A, B have the same dimension of the controllable subspace. This is called the generic dimension of the controllable subspace of structural matrices A, B. Given the graph corresponding to A, B, the generic dimension is the size of the largest spanning cacti.

Poljak proposed the following Integer Linear Program to find the generic dimension of structural matrices (A,B): First given the graph corresponding to (A,B), construct a new graph G':

1. Delete all nodes not reachable from the set of external input nodes.
2. For all genes i, and all input nodes j add the edge $i \rightarrow j$ to the edge set of G'
3. Add self loops $i \rightarrow i$.
4. Define the following sets: $O(v)$, the set of all edges coming out of $v, I(v)$, the sets of all edges coming into v.
5. for all edges e, define c_e as 1 if the edge was an edge in the old graph G itself and 0 otherwise.

The ILP is:

$$maximize \sum_{e \in G'} c_e x_e$$

Subject to: for all nodes v,

$$\sum_{e \in O(v)} x_e = 1$$

$$\sum_{e \in I(v)} x_e = 1$$

$$x_e \in \{0, 1\} \quad \forall e \in G'$$

Poljak argues that the above ILP has a special property of unimodularity due to which it can be solved exactly and more efficiently by relaxing the last constraint to $x_e \geq 0$. The LP is then efficiently solved using the Simplex Algorithm.

Here is an intuition to why the ILP finds the generic dimension. The cactus is a collection of node disjoint paths and cycles. To find the largest covering cactus, one can find the largest covering node disjoint paths and cycles. Rather than handling paths and cycles separately the algorithm completes the paths to cycles by adding backedges. The variable x_e denotes whether a particular edge has been selected as a part of a cycle. the objective function maximizes the number of selected edges that were not dummy edges, for a cactus this is precisely equal to the number of nodes. Self loops are added so that nodes that are not covered by the cactus can be covered by self loops, however since for self loops, $c_e = 0$, they are not counted in the generic dimension. The constraints enforce the condition that each node should be a part of a simple cycle only.

Liu et al. [8] proposed a centrality measure based on this to quantify how important a node is to control a network. They define the control centrality as the dimension of the controllable subspace if we drive that node using an external input alone.

4.3 Analysis of Real Time Varying Networks

Our basic hypothesis is that since it is hard to control the whole network in one go, in biological networks control is accomplished in a phase-wise manner. That is in a particular phase, the structure of the network is rewired because of various non-linear phenomena such that efficient control over genes that are crucial for that phase can be accomplished. Alternatively since the whole expression space is not reachable, the rewiring happens to ensure that useful parts become reachable. If this were the case, one would expect to see a change in the set of genes that can efficiently control a significant part of the network. We show that this indeed happens in two datasets yeast and Drosophila.

4.3.1 Yeast

For yeast we report the top 20 most control central genes for different phases: Cell Cycle, Sporulation, Diauxic shift and DNA damage. We also report the description taken from the Saccharomyces Genome (SGD) Database [2]. We summarize the results in the figures 2, 3, 5, 4, 6. We color code a gene as green if it fits the function of the condition. We find that each of these sets are distinctive showing that the control centers change in the different phases. Moreover, each condition seems to be enriched in genes relevant to the function. For example Diauxic Shift refers to the condition in which the medium in which the organism is kept is changed. The controlling set for this condition contains genes known to be important during Glucose Depletion, Diauxic Shift, and metabolic functions like Glycolysis. We also point out a few interesting observations in figure 7.

4.3.2 Drosophila

In Drosophila we had time specific networks corresponding to 68 time points. However for the purpose of analysis we combined these networks into 4 phase specific networks:

1. Embryo: Time slices 1-31.

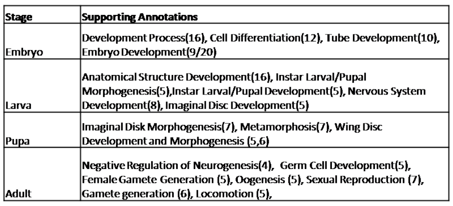
2. Larva: Time slices 31-41.

3. Pupa: Time slices 42-58.

4. Adult: Time Slices 59-68.

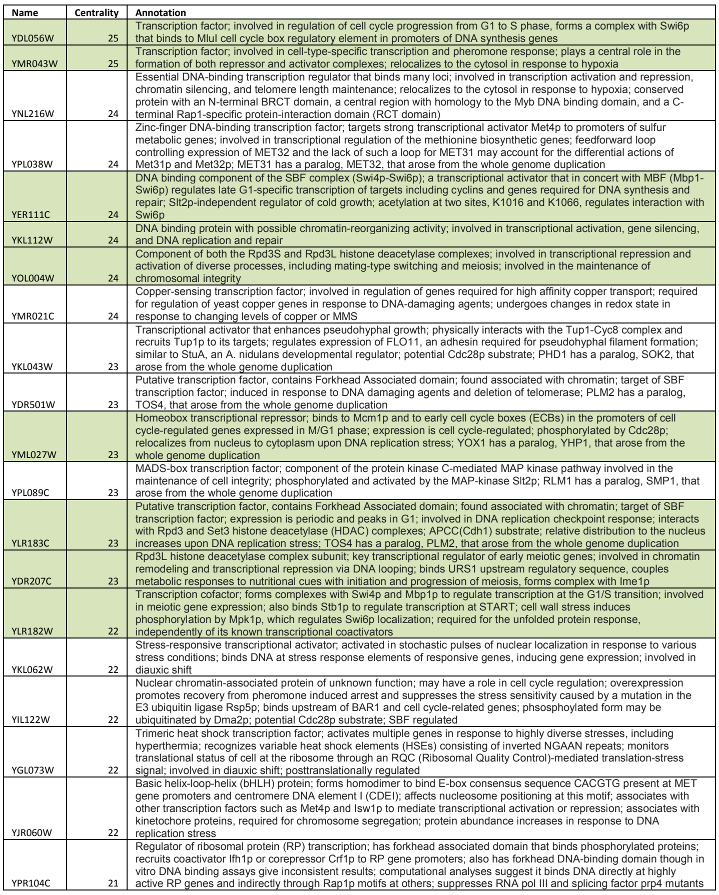
Since FlyBase the principal database for Drosophila does not have functional summaries like Yeast, we considered top 20 genes in terms of centrality in the different phases and we report some frequently occurring Gene Ontology terms in these sets which can be interpreted in figure 8.

Figure 8: Frequently Occurring GO terms in top 20 central genes in various phases. The number in the bracket indicates the number of genes with are annotated with that term



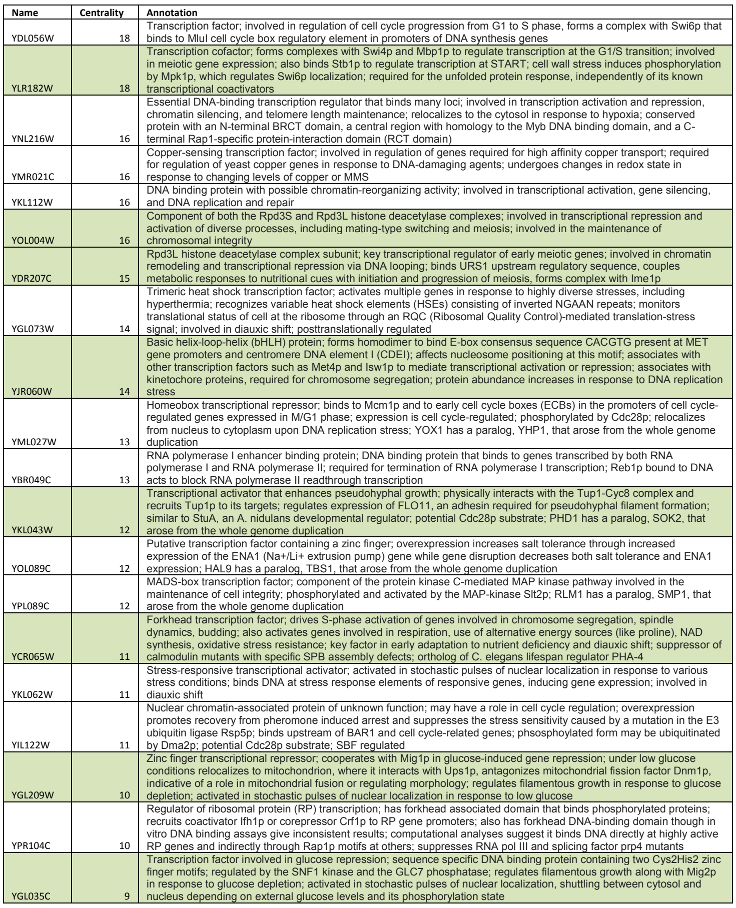
Stage	Supporting Annotations
Embryo	Development Process(16), Cell Differentiation(12), Tube Development(10), Embryo Development(9/20)
Larva	Anatomical Structure Development(16), Instar Larval/Pupal Morphogenesis(5),Instar Larval/Pupal Development(5), Nervous System Development(8), Imaginal Disc Development(5)
Pupa	Imaginal Disk Morphogenesis(7), Metamorphosis(7), Wing Disc Development and Morphogenesis (5,6)
Adult	Negative Regulation of Neurogenesis(4), Germ Cell Development(5), Female Gamete Generation (5), Oogenesis (5), Sexual Reproduction (7), Gamete generation (6), Locomotion (5),

Figure 2: 20 Most Central Genes during Cell Cycle in Yeast



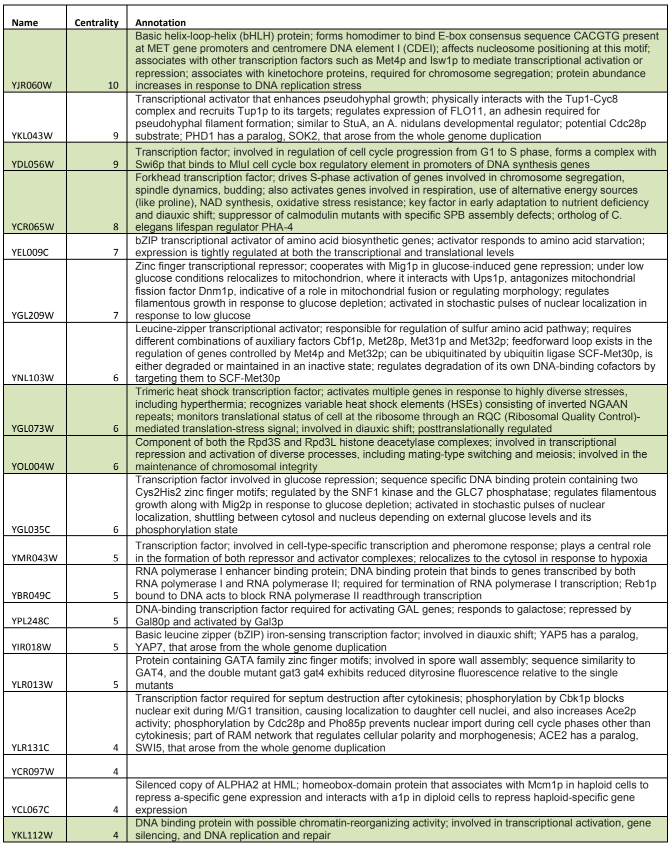
Name	Centrality	Annotation
YDL056W	25	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase, forms a complex with Swi6p that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes
YMR043W	25	Transcription factor; involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes; relocalizes to the cytosol in response to hypoxia
YNL216W	24	Essential DNA-binding transcription regulator that binds many loci; involved in transcription activation and repression, chromatin silencing, and telomere length maintenance; relocalizes to the cytosol in response to hypoxia; conserved protein with an N-terminal BRCT domain, a central region with homology to the Myb DNA binding domain, and a C-terminal Rap1-specific protein-interaction domain (RCT domain)
YPL038W	24	Zinc-finger DNA-binding transcription factor; targets strong transcriptional activator Met4p to promoters of sulfur metabolic genes; involved in transcriptional regulation of the methionine biosynthetic genes; feedforward loop controlling expression of MET32 and the lack of such a loop for MET31 may account for the differential actions of Met31p and Met32p; MET31 has a paralog, MET32, that arose from the whole genome duplication
YER111C	24	DNA binding component of the SBF complex (Swi4p-Swi6p); a transcriptional activator that in concert with MBF (Mbp1-Swi6p) regulates late G1-specific transcription of targets including cyclins and genes required for DNA synthesis and repair; Slt2p-independent regulator of cold growth; acetylation at two sites, K1016 and K1066, regulates interaction with Swi6p
YKL112W	24	DNA binding protein with possible chromatin-reorganizing activity; involved in transcriptional activation, gene silencing, and DNA replication and repair
YOL004W	24	Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
YMR021C	24	Copper-sensing transcription factor; involved in regulation of genes required for high affinity copper transport; required for regulation of yeast copper genes in response to DNA-damaging agents; undergoes changes in redox state in response to changing levels of copper or MMS
YKL043W	23	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; regulates expression of FLO11, an adhesin required for pseudohyphal filament formation; similar to StuA, an A. nidulans developmental regulator; potential Cdc28p substrate; PHD1 has a paralog, SOK2, that arose from the whole genome duplication
YDR501W	23	Putative transcription factor, contains Forkhead Associated domain; found associated with chromatin; target of SBF transcription factor; induced in response to DNA damaging agents and deletion of telomerase; PLM2 has a paralog, TOS4, that arose from the whole genome duplication
YML027W	23	Homeobox transcriptional repressor; binds to Mcm1p and to early cell cycle boxes (ECBs) in the promoters of cell cycle-regulated genes expressed in M/G1 phase; expression is cell cycle-regulated; phosphorylated by Cdc28p; relocalizes from nucleus to cytoplasm upon DNA replication stress; YOX1 has a paralog, YHP1, that arose from the whole genome duplication
YPL089C	23	MADS-box transcription factor; component of the protein kinase C-mediated MAP kinase pathway involved in the maintenance of cell integrity; phosphorylated and activated by the MAP-kinase Slt2p; RLM1 has a paralog, SMP1, that arose from the whole genome duplication
YLR183C	23	Putative transcription factor, contains Forkhead Associated domain; found associated with chromatin; target of SBF transcription factor; expression is periodic and peaks in G1; involved in DNA replication checkpoint response; interacts with Rpd3 and Set3 histone deacetylase (HDAC) complexes; APCC(Cdh1) substrate; relative distribution to the nucleus increases upon DNA replication stress; TOS4 has a paralog, PLM2, that arose from the whole genome duplication
YDR207C	23	Rpd3L histone deacetylase complex subunit; key transcriptional regulator of early meiotic genes; involved in chromatin remodeling and transcriptional repression via DNA looping; binds URS1 upstream regulatory sequence, couples metabolic responses to nutritional cues with initiation and progression of meiosis, forms complex with Ime1p
YLR182W	22	Transcription cofactor; forms complexes with Swi4p and Mbp1p to regulate transcription at the G1/S transition; involved in meiotic gene expression; also binds Stb1p to regulate transcription at START; cell wall stress induces phosphorylation by Mpk1p, which regulates Swi6p localization; required for the unfolded protein response, independently of its known transcriptional coactivators
YKL062W	22	Stress-responsive transcriptional activator; activated in stochastic pulses of nuclear localization in response to various stress conditions; binds DNA at stress response elements of responsive genes, inducing gene expression; involved in diauxic shift
YIL122W	22	Nuclear chromatin-associated protein of unknown function; may have a role in cell cycle regulation; overexpression promotes recovery from pheromone induced arrest and suppresses the stress sensitivity caused by a mutation in the E3 ubiquitin ligase Rsp5p; binds upstream of BAR1 and cell cycle-related genes; phsosphoylated form may be ubiquitinated by Dma2p; potential Cdc28p substrate; SBF regulated
YGL073W	22	Trimeric heat shock transcription factor; activates multiple genes in response to highly diverse stresses, including hyperthermia; recognizes variable heat shock elements (HSEs) consisting of inverted NGAAN repeats; monitors translational status of cell at the ribosome through an RQC (Ribosomal Quality Control)-mediated translation-stress signal; involved in diauxic shift; posttranslationally regulated
YJR060W	22	Basic helix-loop-helix (bHLH) protein; forms homodimer to bind E-box consensus sequence CACGTG present at MET gene promoters and centromere DNA element I (CDEI); affects nucleosome positioning at this motif; associates with other transcription factors such as Met4p and Isw1p to mediate transcriptional activation or repression; associates with kinetochore proteins, required for chromosome segregation; protein abundance increases in response to DNA replication stress
YPR104C	21	Regulator of ribosomal protein (RP) transcription; has forkhead associated domain that binds phosphorylated proteins; recruits coactivator Ifh1p or corepressor Crf1p to RP gene promoters; also has forkhead DNA-binding domain though in vitro DNA binding assays give inconsistent results; computational analyses suggest it binds DNA directly at highly active RP genes and indirectly through Rap1p motifs at others; suppresses RNA pol III and splicing factor prp4 mutants

Figure 3: 20 Most Central Genes during Sporulation in Yeast



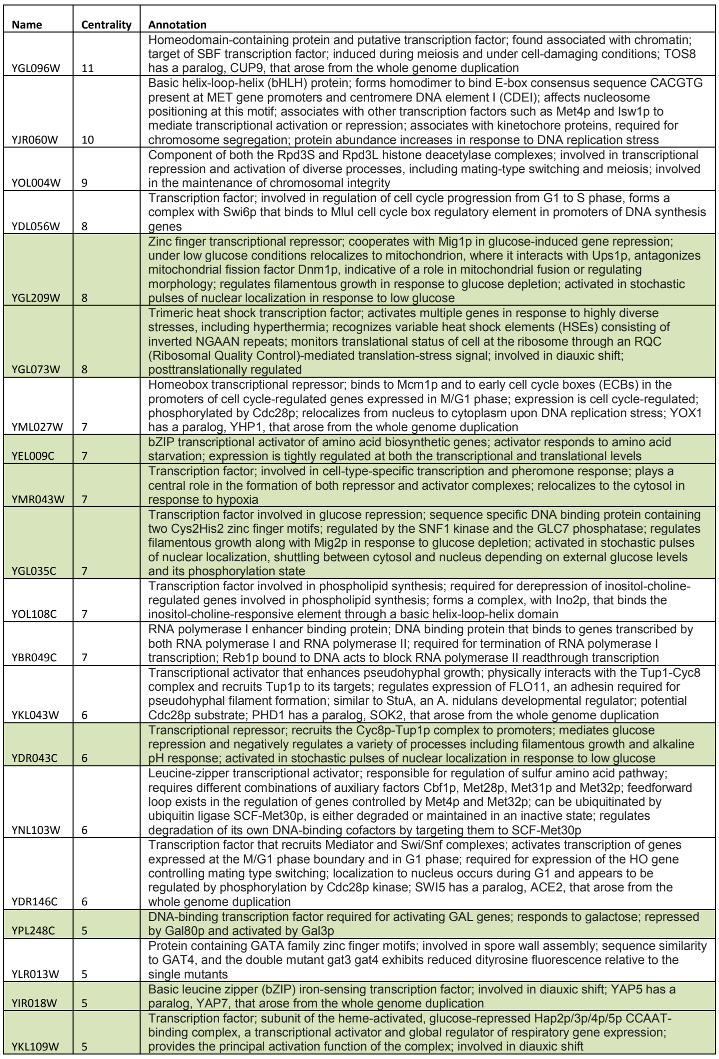
Name	Centrality	Annotation
YDL056W	18	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase, forms a complex with Swi6p that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes
YLR182W	18	Transcription cofactor; forms complexes with Swi4p and Mbp1p to regulate transcription at the G1/S transition; involved in meiotic gene expression; also binds Stb1p to regulate transcription at START; cell wall stress induces phosphorylation by Mpk1p, which regulates Swi6p localization; required for the unfolded protein response, independently of its known transcriptional coactivators
YNL216W	16	Essential DNA-binding transcription regulator that binds many loci; involved in transcription activation and repression, chromatin silencing, and telomere length maintenance; relocalizes to the cytosol in response to hypoxia; conserved protein with an N-terminal BRCT domain, a central region with homology to the Myb DNA binding domain, and a C-terminal Rap1-specific protein-interaction domain (RCT domain)
YMR021C	16	Copper-sensing transcription factor; involved in regulation of genes required for high affinity copper transport; required for regulation of yeast copper genes in response to DNA-damaging agents; undergoes changes in redox state in response to changing levels of copper or MMS
YKL112W	16	DNA binding protein with possible chromatin-reorganizing activity; involved in transcriptional activation, gene silencing, and DNA replication and repair
YOL004W	16	Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
YDR207C	15	Rpd3L histone deacetylase complex subunit; key transcriptional regulator of early meiotic genes; involved in chromatin remodeling and transcriptional repression via DNA looping; binds URS1 upstream regulatory sequence, couples metabolic responses to nutritional cues with initiation and progression of meiosis, forms complex with Ime1p
YGL073W	14	Trimeric heat shock transcription factor; activates multiple genes in response to highly diverse stresses, including hyperthermia; recognizes variable heat shock elements (HSEs) consisting of inverted NGAAN repeats; monitors translational status of cell at the ribosome through an RQC (Ribosomal Quality Control)-mediated translation-stress signal; involved in diauxic shift; posttranslationally regulated
YJR060W	14	Basic helix-loop-helix (bHLH) protein; forms homodimer to bind E-box consensus sequence CACGTG present at MET gene promoters and centromere DNA element I (CDEI); affects nucleosome positioning at this motif; associates with other transcription factors such as Met4p and Isw1p to mediate transcriptional activation or repression; associates with kinetochore proteins, required for chromosome segregation; protein abundance increases in response to DNA replication stress
YML027W	13	Homeobox transcriptional repressor; binds to Mcm1p and to early cell cycle boxes (ECBs) in the promoters of cell cycle-regulated genes expressed in M/G1 phase; expression is cell cycle-regulated; phosphorylated by Cdc28p; relocalizes from nucleus to cytoplasm upon DNA replication stress; YOX1 has a paralog, YHP1, that arose from the whole genome duplication
YBR049C	13	RNA polymerase I enhancer binding protein; DNA binding protein that binds to genes transcribed by both RNA polymerase I and RNA polymerase II; required for termination of RNA polymerase I transcription; Reb1p bound to DNA acts to block RNA polymerase II readthrough transcription
YKL043W	12	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; regulates expression of FLO11, an adhesin required for pseudohyphal filament formation; similar to StuA, an A. nidulans developmental regulator; potential Cdc28p substrate; PHD1 has a paralog, SOK2, that arose from the whole genome duplication
YOL089C	12	Putative transcription factor containing a zinc finger; overexpression increases salt tolerance through increased expression of the ENA1 (Na+/Li+ extrusion pump) gene while gene disruption decreases both salt tolerance and ENA1 expression; HAL9 has a paralog, TBS1, that arose from the whole genome duplication
YPL089C	12	MADS-box transcription factor; component of the protein kinase C-mediated MAP kinase pathway involved in the maintenance of cell integrity; phosphorylated and activated by the MAP-kinase Slt2p; RLM1 has a paralog, SMP1, that arose from the whole genome duplication
YCR065W	11	Forkhead transcription factor; drives S-phase activation of genes involved in chromosome segregation, spindle dynamics, budding; also activates genes involved in respiration, use of alternative energy sources (like proline), NAD synthesis, oxidative stress resistance; key factor in early adaptation to nutrient deficiency and diauxic shift; suppressor of calmodulin mutants with specific SPB assembly defects; ortholog of C. elegans lifespan regulator PHA-4
YKL062W	11	Stress-responsive transcriptional activator; activated in stochastic pulses of nuclear localization in response to various stress conditions; binds DNA at stress response elements of responsive genes, inducing gene expression; involved in diauxic shift
YIL122W	11	Nuclear chromatin-associated protein of unknown function; may have a role in cell cycle regulation; overexpression promotes recovery from pheromone induced arrest and suppresses the stress sensitivity caused by a mutation in the E3 ubiquitin ligase Rsp5p; binds upstream of BAR1 and cell cycle-related genes; phsosphoylated form may be ubiquitinated by Dma2p; potential Cdc28p substrate; SBF regulated
YGL209W	10	Zinc finger transcriptional repressor; cooperates with Mig1p in glucose-induced gene repression; under low glucose conditions relocalizes to mitochondrion, where it interacts with Ups1p, antagonizes mitochondrial fission factor Dnm1p, indicative of a role in mitochondrial fusion or regulating morphology; regulates filamentous growth in response to glucose depletion; activated in stochastic pulses of nuclear localization in response to low glucose
YPR104C	10	Regulator of ribosomal protein (RP) transcription; has forkhead associated domain that binds phosphorylated proteins; recruits coactivator Ifh1p or corepressor Crf1p to RP gene promoters; also has forkhead DNA-binding domain though in vitro DNA binding assays give inconsistent results; computational analyses suggest it binds DNA directly at highly active RP genes and indirectly through Rap1p motifs at others; suppresses RNA pol III and splicing factor prp4 mutants
YGL035C	9	Transcription factor involved in glucose repression; sequence specific DNA binding protein containing two Cys2His2 zinc finger motifs; regulated by the SNF1 kinase and the GLC7 phosphatase; regulates filamentous growth along with Mig2p in response to glucose depletion; activated in stochastic pulses of nuclear localization, shuttling between cytosol and nucleus depending on external glucose levels and its phosphorylation state

Figure 4: 20 Most Central Genes during DNA Damage in Yeast



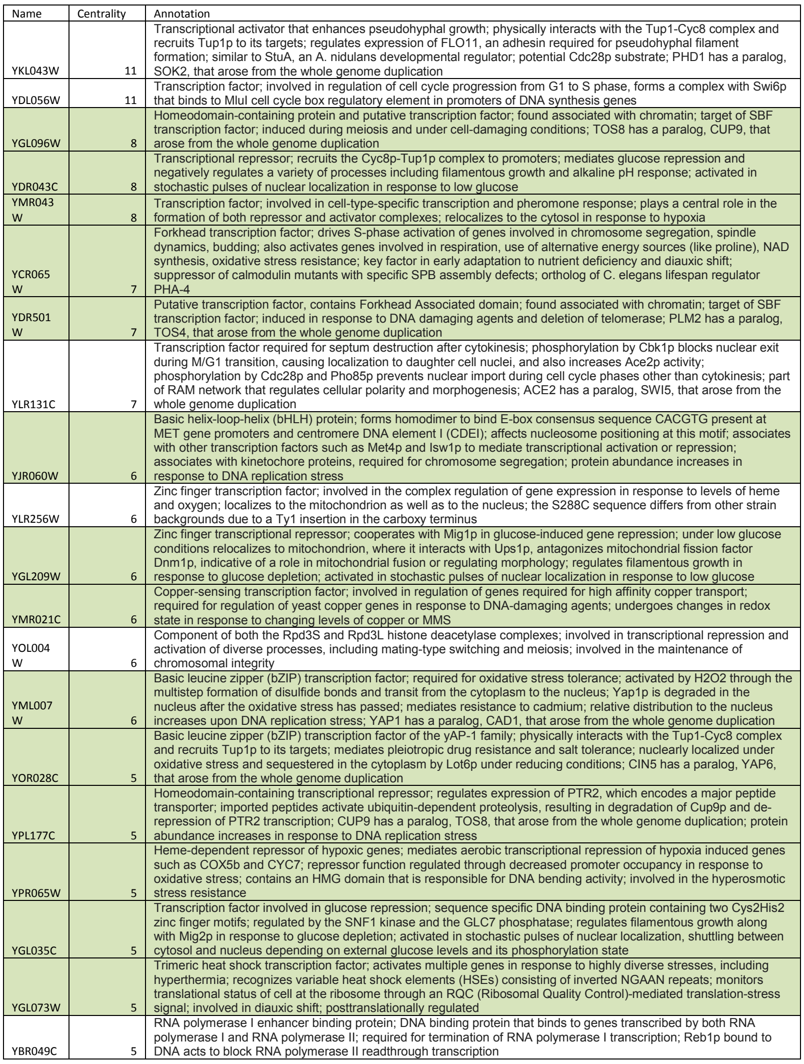
Name	Centrality	Annotation
YJR060W	10	Basic helix-loop-helix (bHLH) protein; forms homodimer to bind E-box consensus sequence CACGTG present at MET gene promoters and centromere DNA element I (CDEI); affects nucleosome positioning at this motif; associates with other transcription factors such as Met4p and Isw1p to mediate transcriptional activation or repression; associates with kinetochore proteins, required for chromosome segregation; protein abundance increases in response to DNA replication stress
YKL043W	9	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; regulates expression of FLO11, an adhesin required for pseudohyphal filament formation; similar to StuA, an A. nidulans developmental regulator; potential Cdc28p substrate; PHD1 has a paralog, SOK2, that arose from the whole genome duplication
YDL056W	9	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase, forms a complex with Swi6p that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes
YCR065W	8	Forkhead transcription factor; drives S-phase activation of genes involved in chromosome segregation, spindle dynamics, budding; also activates genes involved in respiration, use of alternative energy sources (like proline), NAD synthesis, oxidative stress resistance; key factor in early adaptation to nutrient deficiency and diauxic shift; suppressor of calmodulin mutants with specific SPB assembly defects; ortholog of C. elegans lifespan regulator PHA-4
YEL009C	7	bZIP transcriptional activator of amino acid biosynthetic genes; activator responds to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels
YGL209W	7	Zinc finger transcriptional repressor; cooperates with Mig1p in glucose-induced gene repression; under low glucose conditions relocalizes to mitochondrion, where it interacts with Ups1p, antagonizes mitochondrial fission factor Dnm1p, indicative of a role in mitochondrial fusion or regulating morphology; regulates filamentous growth in response to glucose depletion; activated in stochastic pulses of nuclear localization in response to low glucose
YNL103W	6	Leucine-zipper transcriptional activator; responsible for regulation of sulfur amino acid pathway; requires different combinations of auxiliary factors Cbf1p, Met28p, Met31p and Met32p; feedforward loop exists in the regulation of genes controlled by Met4p and Met32p; can be ubiquitinated by ubiquitin ligase SCF-Met30p, is either degraded or maintained in an inactive state; regulates degradation of its own DNA-binding cofactors by targeting them to SCF-Met30p
YGL073W	6	Trimeric heat shock transcription factor; activates multiple genes in response to highly diverse stresses, including hyperthermia; recognizes variable heat shock elements (HSEs) consisting of inverted NGAAN repeats; monitors translational status of cell at the ribosome through an RQC (Ribosomal Quality Control)-mediated translation-stress signal; involved in diauxic shift; posttranslationally regulated
YOL004W	6	Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
YGL035C	6	Transcription factor involved in glucose repression; sequence specific DNA binding protein containing two Cys2His2 zinc finger motifs; regulated by the SNF1 kinase and the GLC7 phosphatase; regulates filamentous growth along with Mig2p in response to glucose depletion; activated in stochastic pulses of nuclear localization, shuttling between cytosol and nucleus depending on external glucose levels and its phosphorylation state
YMR043W	5	Transcription factor; involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes; relocalizes to the cytosol in response to hypoxia
YBR049C	5	RNA polymerase I enhancer binding protein; DNA binding protein that binds to genes transcribed by both RNA polymerase I and RNA polymerase II; required for termination of RNA polymerase I transcription; Reb1p bound to DNA acts to block RNA polymerase II readthrough transcription
YPL248C	5	DNA-binding transcription factor required for activating GAL genes; responds to galactose; repressed by Gal80p and activated by Gal3p
YIR018W	5	Basic leucine zipper (bZIP) iron-sensing transcription factor; involved in diauxic shift; YAP5 has a paralog, YAP7, that arose from the whole genome duplication
YLR013W	5	Protein containing GATA family zinc finger motifs; involved in spore wall assembly; sequence similarity to GAT4, and the double mutant gat3 gat4 exhibits reduced dityrosine fluorescence relative to the single mutants
YLR131C	4	Transcription factor required for septum destruction after cytokinesis; phosphorylation by Cbk1p blocks nuclear exit during M/G1 transition, causing localization to daughter cell nuclei, and also increases Ace2p activity; phosphorylation by Cdc28p and Pho85p prevents nuclear import during cell cycle phases other than cytokinesis; part of RAM network that regulates cellular polarity and morphogenesis; ACE2 has a paralog, SWI5, that arose from the whole genome duplication
YCR097W	4	
YCL067C	4	Silenced copy of ALPHA2 at HML; homeobox-domain protein that associates with Mcm1p in haploid cells to repress a-specific gene expression and interacts with a1p in diploid cells to repress haploid-specific gene expression
YKL112W	4	DNA binding protein with possible chromatin-reorganizing activity; involved in transcriptional activation, gene silencing, and DNA replication and repair

Figure 5: 20 Most Central Genes during Diauxic Shift in Yeast



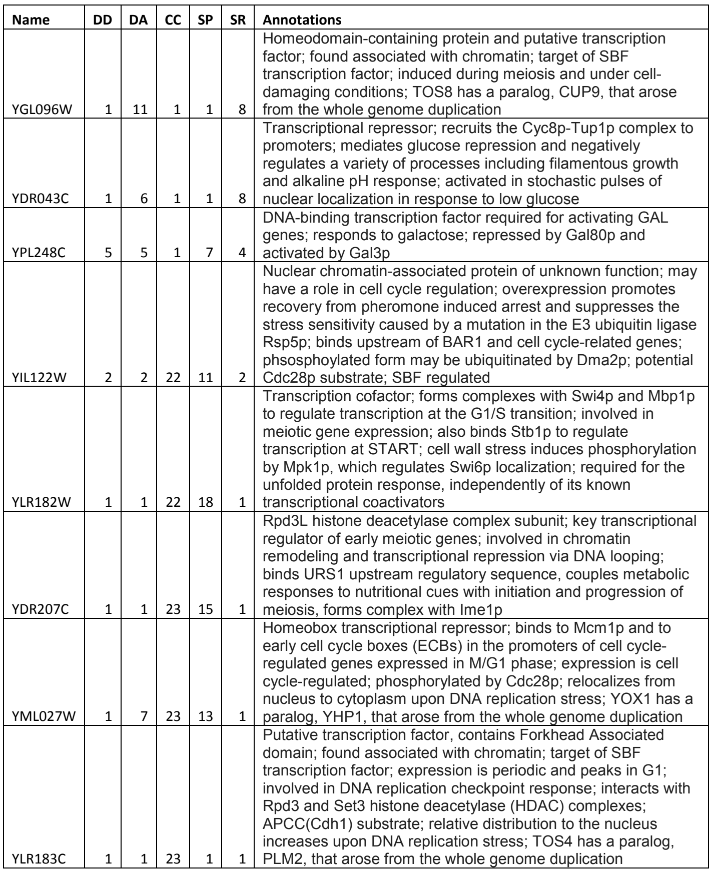
Name	Centrality	Annotation
YGL096W	11	Homeodomain-containing protein and putative transcription factor; found associated with chromatin; target of SBF transcription factor; induced during meiosis and under cell-damaging conditions; TOS8 has a paralog, CUP9, that arose from the whole genome duplication
YJR060W	10	Basic helix-loop-helix (bHLH) protein; forms homodimer to bind E-box consensus sequence CACGTG present at MET gene promoters and centromere DNA element I (CDEI); affects nucleosome positioning at this motif; associates with other transcription factors such as Met4p and Isw1p to mediate transcriptional activation or repression; associates with kinetochore proteins, required for chromosome segregation; protein abundance increases in response to DNA replication stress
YOL004W	9	Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
YDL056W	8	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase, forms a complex with Swi6p that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes
YGL209W	8	Zinc finger transcriptional repressor; cooperates with Mig1p in glucose-induced gene repression; under low glucose conditions relocalizes to mitochondrion, where it interacts with Ups1p, antagonizes mitochondrial fission factor Dnm1p, indicative of a role in mitochondrial fusion or regulating morphology; regulates filamentous growth in response to glucose depletion; activated in stochastic pulses of nuclear localization in response to low glucose
YGL073W	8	Trimeric heat shock transcription factor; activates multiple genes in response to highly diverse stresses, including hyperthermia; recognizes variable heat shock elements (HSEs) consisting of inverted NGAAN repeats; monitors translational status of cell at the ribosome through an RQC (Ribosomal Quality Control)-mediated translation-stress signal; involved in diauxic shift; posttranslationally regulated
YML027W	7	Homeobox transcriptional repressor; binds to Mcm1p and to early cell cycle boxes (ECBs) in the promoters of cell cycle-regulated genes expressed in M/G1 phase; expression is cell cycle-regulated; phosphorylated by Cdc28p; relocalizes from nucleus to cytoplasm upon DNA replication stress; YOX1 has a paralog, YHP1, that arose from the whole genome duplication
YEL009C	7	bZIP transcriptional activator of amino acid biosynthetic genes; activator responds to amino acid starvation; expression is tightly regulated at both the transcriptional and translational levels
YMR043W	7	Transcription factor; involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes; relocalizes to the cytosol in response to hypoxia
YGL035C	7	Transcription factor involved in glucose repression; sequence specific DNA binding protein containing two Cys2His2 zinc finger motifs; regulated by the SNF1 kinase and the GLC7 phosphatase; regulates filamentous growth along with Mig2p in response to glucose depletion; activated in stochastic pulses of nuclear localization, shuttling between cytosol and nucleus depending on external glucose levels and its phosphorylation state
YOL108C	7	Transcription factor involved in phospholipid synthesis; required for derepression of inositol-choline-regulated genes involved in phospholipid synthesis; forms a complex, with Ino2p, that binds the inositol-choline-responsive element through a basic helix-loop-helix domain
YBR049C	7	RNA polymerase I enhancer binding protein; DNA binding protein that binds to genes transcribed by both RNA polymerase I and RNA polymerase II; required for termination of RNA polymerase I transcription; Reb1p bound to DNA acts to block RNA polymerase II readthrough transcription
YKL043W	6	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; regulates expression of FLO11, an adhesin required for pseudohyphal filament formation; similar to StuA, an A. nidulans developmental regulator; potential Cdc28p substrate; PHD1 has a paralog, SOK2, that arose from the whole genome duplication
YDR043C	6	Transcriptional repressor; recruits the Cyc8p-Tup1p complex to promoters; mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response; activated in stochastic pulses of nuclear localization in response to low glucose
YNL103W	6	Leucine-zipper transcriptional activator; responsible for regulation of sulfur amino acid pathway; requires different combinations of auxiliary factors Cbf1p, Met28p, Met31p and Met32p; feedforward loop exists in the regulation of genes controlled by Met4p and Met32p; can be ubiquitinated by ubiquitin ligase SCF-Met30p, is either degraded or maintained in an inactive state; regulates degradation of its own DNA-binding cofactors by targeting them to SCF-Met30p
YDR146C	6	Transcription factor that recruits Mediator and Swi/Snf complexes; activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; required for expression of the HO gene controlling mating type switching; localization to nucleus occurs during G1 and appears to be regulated by phosphorylation by Cdc28p kinase; SWI5 has a paralog, ACE2, that arose from the whole genome duplication
YPL248C	5	DNA-binding transcription factor required for activating GAL genes; responds to galactose; repressed by Gal80p and activated by Gal3p
YLR013W	5	Protein containing GATA family zinc finger motifs; involved in spore wall assembly; sequence similarity to GAT4, and the double mutant gat3 gat4 exhibits reduced dityrosine fluorescence relative to the single mutants
YIR018W	5	Basic leucine zipper (bZIP) iron-sensing transcription factor; involved in diauxic shift; YAP5 has a paralog, YAP7, that arose from the whole genome duplication
YKL109W	5	Transcription factor; subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator and global regulator of respiratory gene expression; provides the principal activation function of the complex; involved in diauxic shift

Figure 6: 20 Most Central Genes during Stress Response in Yeast



Name	Centrality	Annotation
YKL043W	11	Transcriptional activator that enhances pseudohyphal growth; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; regulates expression of FLO11, an adhesin required for pseudohyphal filament formation; similar to StuA, an A. nidulans developmental regulator; potential Cdc28p substrate; PHD1 has a paralog, SOK2, that arose from the whole genome duplication
YDL056W	11	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase, forms a complex with Swi6p that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes
YGL096W	8	Homeodomain-containing protein and putative transcription factor; found associated with chromatin; target of SBF transcription factor; induced during meiosis and under cell-damaging conditions; TOS8 has a paralog, CUP9, that arose from the whole genome duplication
YDR043C	8	Transcriptional repressor; recruits the Cyc8p-Tup1p complex to promoters; mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response; activated in stochastic pulses of nuclear localization in response to low glucose
YMR043W	8	Transcription factor; involved in cell-type-specific transcription and pheromone response; plays a central role in the formation of both repressor and activator complexes; relocalizes to the cytosol in response to hypoxia
YCR065W	7	Forkhead transcription factor; drives S-phase activation of genes involved in chromosome segregation, spindle dynamics, budding; also activates genes involved in respiration, use of alternative energy sources (like proline), NAD synthesis, oxidative stress resistance; key factor in early adaptation to nutrient deficiency and diauxic shift; suppressor of calmodulin mutants with specific SPB assembly defects; ortholog of C. elegans lifespan regulator PHA-4
YDR501W	7	Putative transcription factor, contains Forkhead Associated domain; found associated with chromatin; target of SBF transcription factor; induced in response to DNA damaging agents and deletion of telomerase; PLM2 has a paralog, TOS4, that arose from the whole genome duplication
YLR131C	7	Transcription factor required for septum destruction after cytokinesis; phosphorylation by Cbk1p blocks nuclear exit during M/G1 transition, causing localization to daughter cell nuclei, and also increases Ace2p activity; phosphorylation by Cdc28p and Pho85p prevents nuclear import during cell cycle phases other than cytokinesis; part of RAM network that regulates cellular polarity and morphogenesis; ACE2 has a paralog, SWI5, that arose from the whole genome duplication
YJR060W	6	Basic helix-loop-helix (bHLH) protein; forms homodimer to bind E-box consensus sequence CACGTG present at MET gene promoters and centromere DNA element I (CDEI); affects nucleosome positioning at this motif; associates with other transcription factors such as Met4p and Isw1p to mediate transcriptional activation or repression; associates with kinetochore proteins, required for chromosome segregation; protein abundance increases in response to DNA replication stress
YLR256W	6	Zinc finger transcription factor; involved in the complex regulation of gene expression in response to levels of heme and oxygen; localizes to the mitochondrion as well as to the nucleus; the S288C sequence differs from other strain backgrounds due to a Ty1 insertion in the carboxy terminus
YGL209W	6	Zinc finger transcriptional repressor; cooperates with Mig1p in glucose-induced gene repression; under low glucose conditions relocalizes to mitochondrion, where it interacts with Ups1p, antagonizes mitochondrial fission factor Dnm1p, indicative of a role in mitochondrial fusion or regulating morphology; regulates filamentous growth in response to glucose depletion; activated in stochastic pulses of nuclear localization in response to low glucose
YMR021C	6	Copper-sensing transcription factor; involved in regulation of genes required for high affinity copper transport; required for regulation of yeast copper genes in response to DNA-damaging agents; undergoes changes in redox state in response to changing levels of copper or MMS
YOL004W	6	Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity
YML007W	6	Basic leucine zipper (bZIP) transcription factor; required for oxidative stress tolerance; activated by H2O2 through the multistep formation of disulfide bonds and transit from the cytoplasm to the nucleus; Yap1p is degraded in the nucleus after the oxidative stress has passed; mediates resistance to cadmium; relative distribution to the nucleus increases upon DNA replication stress; YAP1 has a paralog, CAD1, that arose from the whole genome duplication
YOR028C	5	Basic leucine zipper (bZIP) transcription factor of the yAP-1 family; physically interacts with the Tup1-Cyc8 complex and recruits Tup1p to its targets; mediates pleiotropic drug resistance and salt tolerance; nuclearly localized under oxidative stress and sequestered in the cytoplasm by Lot6p under reducing conditions; CIN5 has a paralog, YAP6, that arose from the whole genome duplication
YPL177C	5	Homeodomain-containing transcriptional repressor; regulates expression of PTR2, which encodes a major peptide transporter; imported peptides activate ubiquitin-dependent proteolysis, resulting in degradation of Cup9p and de-repression of PTR2 transcription; CUP9 has a paralog, TOS8, that arose from the whole genome duplication; protein abundance increases in response to DNA replication stress
YPR065W	5	Heme-dependent repressor of hypoxic genes; mediates aerobic transcriptional repression of hypoxia induced genes such as COX5b and CYC7; repressor function regulated through decreased promoter occupancy in response to oxidative stress; contains an HMG domain that is responsible for DNA bending activity; involved in the hyperosmotic stress resistance
YGL035C	5	Transcription factor involved in glucose repression; sequence specific DNA binding protein containing two Cys2His2 zinc finger motifs; regulated by the SNF1 kinase and the GLC7 phosphatase; regulates filamentous growth along with Mig2p in response to glucose depletion; activated in stochastic pulses of nuclear localization, shuttling between cytosol and nucleus depending on external glucose levels and its phosphorylation state
YGL073W	5	Trimeric heat shock transcription factor; activates multiple genes in response to highly diverse stresses, including hyperthermia; recognizes variable heat shock elements (HSEs) consisting of inverted NGAAN repeats; monitors translational status of cell at the ribosome through an RQC (Ribosomal Quality Control)-mediated translation-stress signal; involved in diauxic shift; posttranslationally regulated
YBR049C	5	RNA polymerase I enhancer binding protein; DNA binding protein that binds to genes transcribed by both RNA polymerase I and RNA polymerase II; required for termination of RNA polymerase I transcription; Reb1p bound to DNA acts to block RNA polymerase II readthrough transcription

Figure 7: Some Interesting Genes and their Centralities during Cell Cycle(CC), Sporulation (SP), DNA Damage(DD), Stress Response (SR) and Diauxic Shift(DA)



Name	DD	DA	CC	SP	SR	Annotations
YGL096W	1	11	1	1	8	Homeodomain-containing protein and putative transcription factor; found associated with chromatin; target of SBF transcription factor; induced during meiosis and under cell-damaging conditions; TOS8 has a paralog, CUP9, that arose from the whole genome duplication
YDR043C	1	6	1	1	8	Transcriptional repressor; recruits the Cyc8p-Tup1p complex to promoters; mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response; activated in stochastic pulses of nuclear localization in response to low glucose
YPL248C	5	5	1	7	4	DNA-binding transcription factor required for activating GAL genes; responds to galactose; repressed by Gal80p and activated by Gal3p
YIL122W	2	2	22	11	2	Nuclear chromatin-associated protein of unknown function; may have a role in cell cycle regulation; overexpression promotes recovery from pheromone induced arrest and suppresses the stress sensitivity caused by a mutation in the E3 ubiquitin ligase Rsp5p; binds upstream of BAR1 and cell cycle-related genes; phsosphoylated form may be ubiquitinated by Dma2p; potential Cdc28p substrate; SBF regulated
YLR182W	1	1	22	18	1	Transcription cofactor; forms complexes with Swi4p and Mbp1p to regulate transcription at the G1/S transition; involved in meiotic gene expression; also binds Stb1p to regulate transcription at START; cell wall stress induces phosphorylation by Mpk1p, which regulates Swi6p localization; required for the unfolded protein response, independently of its known transcriptional coactivators
YDR207C	1	1	23	15	1	Rpd3L histone deacetylase complex subunit; key transcriptional regulator of early meiotic genes; involved in chromatin remodeling and transcriptional repression via DNA looping; binds URS1 upstream regulatory sequence, couples metabolic responses to nutritional cues with initiation and progression of meiosis, forms complex with Ime1p
YML027W	1	7	23	13	1	Homeobox transcriptional repressor; binds to Mcm1p and to early cell cycle boxes (ECBs) in the promoters of cell cycle-regulated genes expressed in M/G1 phase; expression is cell cycle-regulated; phosphorylated by Cdc28p; relocalizes from nucleus to cytoplasm upon DNA replication stress; YOX1 has a paralog, YHP1, that arose from the whole genome duplication
YLR183C	1	1	23	1	1	Putative transcription factor, contains Forkhead Associated domain; found associated with chromatin; target of SBF transcription factor; expression is periodic and peaks in G1; involved in DNA replication checkpoint response; interacts with Rpd3 and Set3 histone deacetylase (HDAC) complexes; APCC(Cdh1) substrate; relative distribution to the nucleus increases upon DNA replication stress; TOS4 has a paralog, PLM2, that arose from the whole genome duplication

5 Modularity and Effective Dimension of the Expression States

The general intuition of modules or communities in Network Science is that they are clusters of nodes which act in coordination to perform a function. In this chapter we first review the structural and functional definitions of modules. We then report results which show that the possible expression states that a module can show is severely constrained using actual expression data. This was done during EED310. We further visualize the modules extracted structurally and show that most of them are poorly controllable structurally. This aim of this chapter is to establish modularity as a cause for reduced controllability. It suggests that transcription networks don't consider it useful to assign arbitrary expression values to genes within the same module. Instead a module has only a few possible states that it can be in. This brings about significant reduction in the dimension of the controllable subspace of the whole network.

5.1 Review of Modularity

Just like engineering systems are organized into independently functioning subunits, biological systems can be organized into subunits called modules or communities. There are various signatures of the modules:

- Functional: One expects that genes within the same module to have similar functions. Hence modules can be discovered by clustering Gene Ontology annotations.
- Structural: When the system is represented as a network, genes within the same module are expected to have high interactions with other genes in the same module than outside. Let σ be a function that assigns each node to a module. Then the Newman-Girvan measure of modularity of a network under the partition function σ is given by:

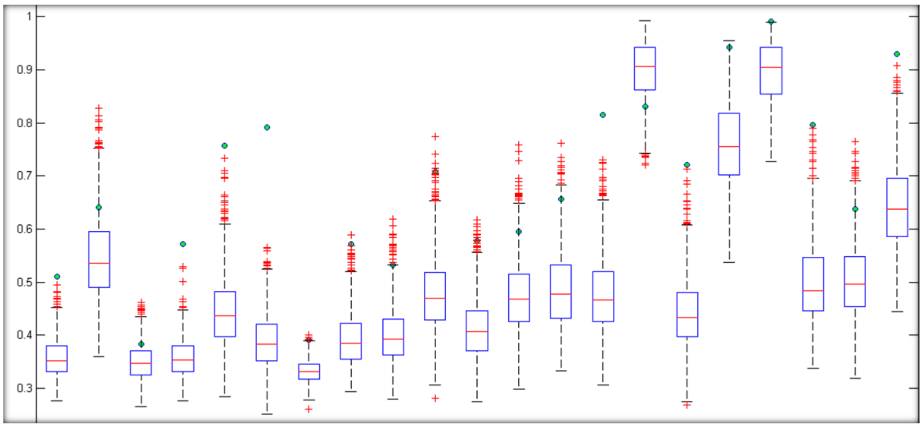
  $$Q(A, \sigma) = \frac{1}{4m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(\sigma(i), \sigma(j)) \tag{8}$$

  The partition function σ that maximizes Q is the required partitioning of the network into modules. The expression for Q can be optimized in a variety of ways like in [10][3].
- Expression: One would expect that genes within the same module to have coordinated expression patterns. Hence modules can also be discovered by clustering expression data.

5.2 Characterizing Effective Dimension of Modules using Actual Expression Data

During the Mini Project, we worked on the E. Coli Dataset. We extracted modules from the structural networks using the spectral method[10]. We then considered the expression matrix of each module separately. We projected the entire expression data set on the first two principal components. We compared the scatter plot of the first two principal components with that of randomly sampled subset of genes of the same size of the module. If a module had a lower dimension of expression states, then one would expect that the scatter plot would be significantly more clustered in comparison to a random subset of genes. We summarize the results using the following boxplot 9

Figure 9: Percentage of Variance in Expression Data Explain by PC1 and PC2 in various structural modules in the E.Coli Transcription Network.



Since the expression of the modules can be captured using only a few principal components we conclude that modules are low dimensional structures in the state space and that genes within a module can't be assigned arbitrary expression values but are significantly contrained.

5.3 Characterizing Efficiency of Control of Individual Modules using Structural Controllability

On an orthogonal direction, after showing that genes in a module have restricted expression states, we show that modules extracted using Newman Girvan are structurally inefficient to control. As before for the Drosophila and Yeast datasets, we find out the structural modules using a greedy heuristic to optimize Newman Girvan modularity[3]. We have visualized the modules in figures 13 and 10. We find that all the modules are composed of a distinct subgraph which one can call the coregulation subgraph which consists of a few genes jointly regulation a large number of genes. Such a subgraph is poorly controllable because of the presence of a large dilation. We now consider each module as a separate graph and compute the fraction of nodes required to be control the subgraph. We found that this number is almost always more than 0.9 fraction of genes in the module. The results are summarized in figure this and this.

Figure 10: Visualization of Structural Modules in Yeast

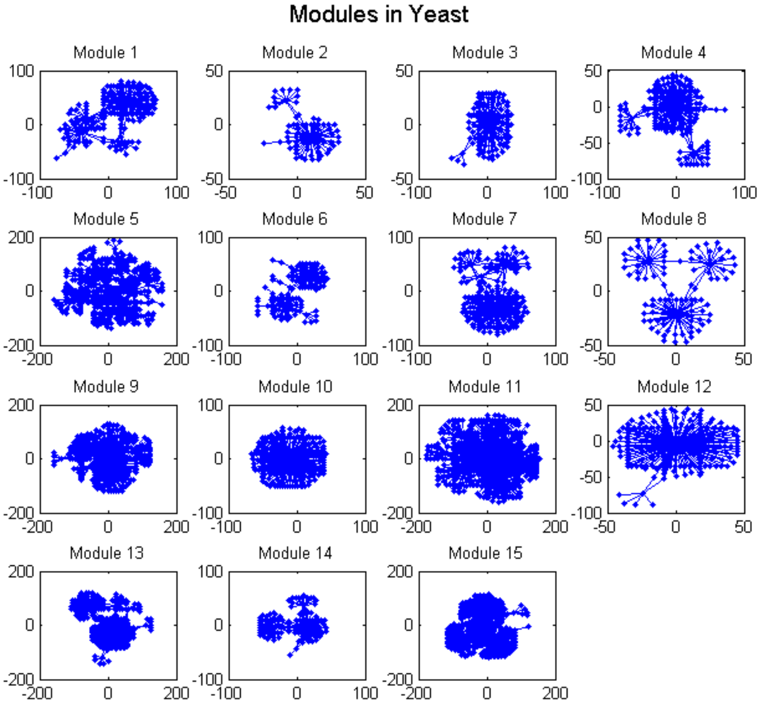


Figure 11: Visualization of Structural Modules in Drosophila

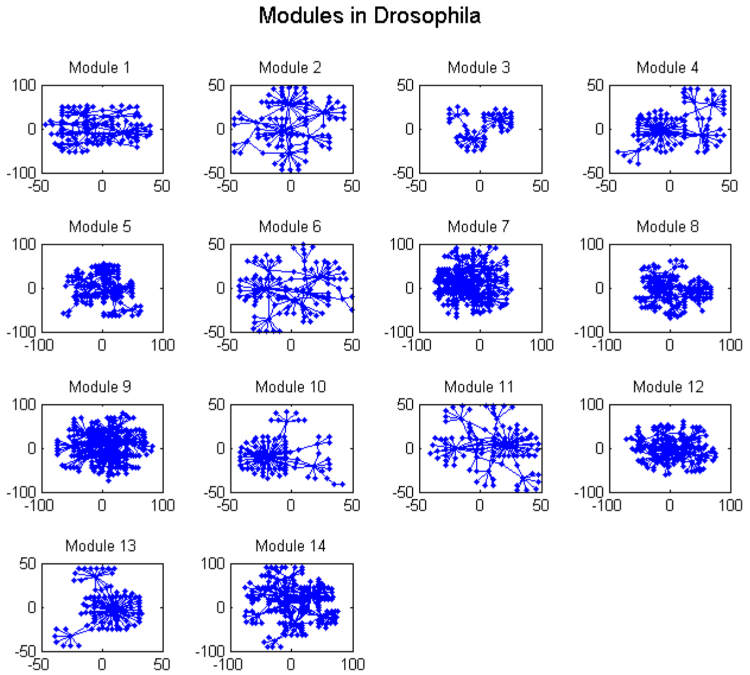


Figure 12: Number of Controlling Inputs Required to Control Modules in Yeast

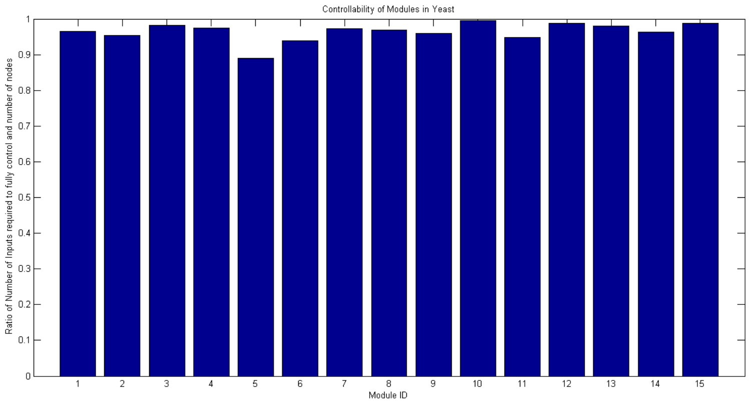
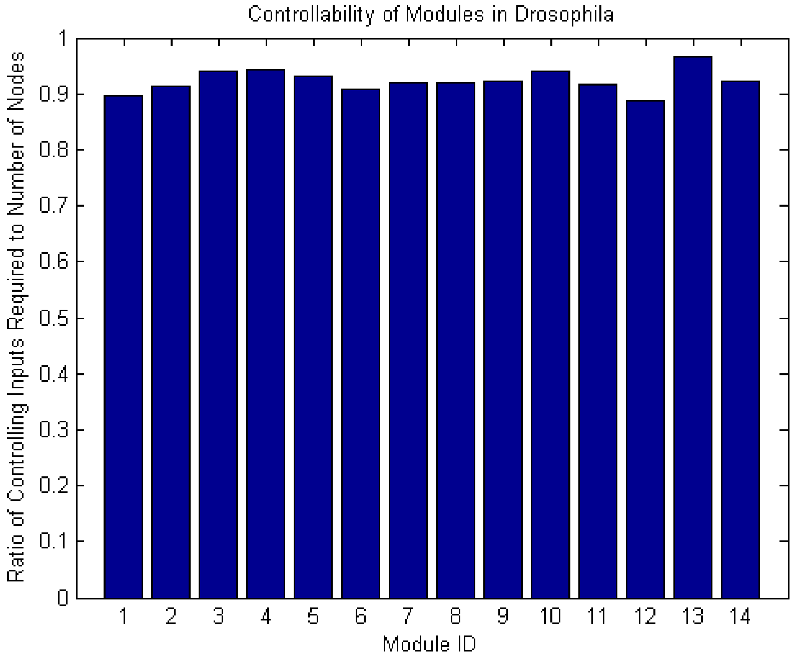


Figure 13: Number of Controlling Inputs Required to Control Modules in Drosophila



All these findings suggest that the module has only a few possible states it can be in. This reinforces the idea of a module as an independent coherent functional unit designed to do a few tasks.

6 Control Properties of Coarsened Networks

Given that we've shown some evidence that the nodes within a single module might form an uncontrollable subspace the next question that arises is are the modules as aggregate entities controllable. To clarify the question, modules are thought of as independent subunits in the whole system. For example in a CPU the memory unit, ALU and so on perform independent functions. Hence given this idea of modules one would expect that it should be possible to assign them arbitrary states. One basic idea to test this would be to create a new network in which a module represents a node. Such networks are called coarse grained networks. In this section we provide some evidence that coarsened networks are much more easily controllable suggesting that while the expression space of one module is severely constrained, however the state of one module puts only limited constraints on the expression of another.

6.1 Description of Experiments

To construct coarse grained networks we define an edge between module A and module B if there is atleast an edge between a gene who is a member of A and a gene which is a member of B. The notion of an edge is now less theoretically rigorous as before. Earlier in the original network edges corresponded to the appearance of one gene in the differential equation of another. In the coarsened network we look at edges as potential ways of one module to influence another. In past literature too, notions of controllability have been applied to networks without differential equation dynamics[7]. The modules can be constructed structurally (using a network), functionally (using GO), or expression-wise. We show our conclusions that these coarse-grained networks are not that hard to control are consistent across structure and expression modules. We haven't explored functional modules as yet. Note that in the coarse grained networks we can actually potentially have the notion of strength of an edge: if there are alot of edges between module A and B in the original network, then the corresponding edge in the coarse grained network is stronger. This also brings robustness to errors in the inference algorithm. A few edges may be mistakes of the inference algorithm but if multiple edges are observed between 2 modules it increases our confidence that the edge is a true one. A question arises what is a natural threshold for number of edges between a module to declare it an edge in the coarse grained network. Suppose the original network had m edges. Suppose a network was randomly clustered into K clusters. Then on average there would be $\frac{m}{K^2}$ edges from any module to another module. We use this as a threshold.

6.2 Results on the Drosophila Dataset

We first extract clusters using expression data. We use simple K means to do so. We found that the most natural clusters occur for $K = 14$. We visualize both the Expression Matrix and the PCA plot for the clusters. This is shown in figure 14 and 15.

Figure 14: Visualization of Clusters obtained in Drosophila using PCA

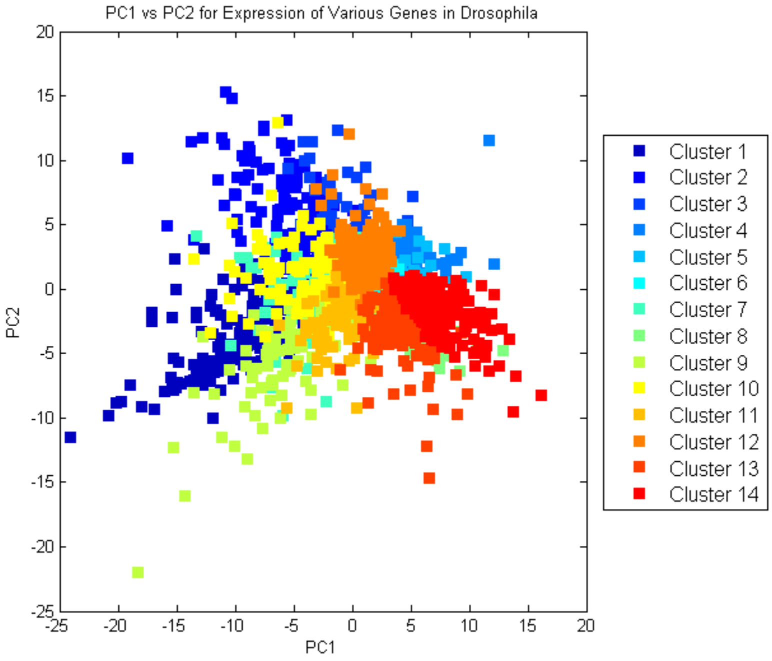
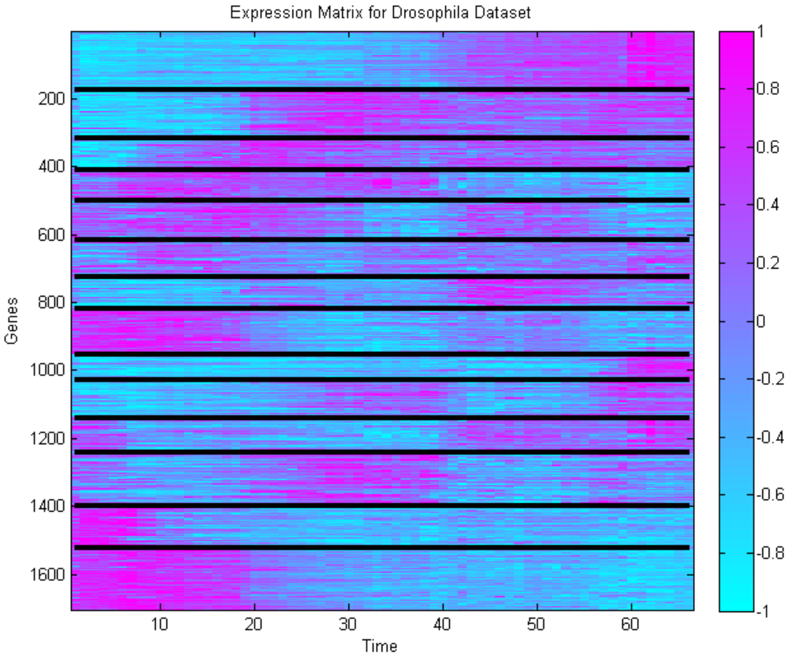
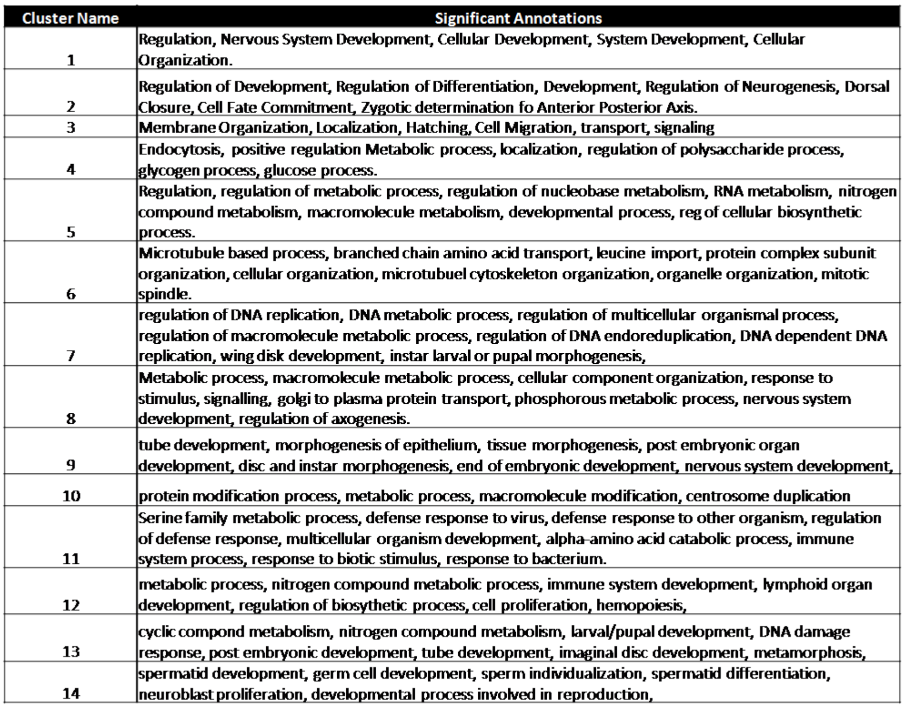


Figure 15: Visualization of the Expression Matrix in Drosophila reordered by clusters



Next we investigated the control properties of the coarsened networks. Note that keeping with the infinite time constant assumption prevalent in controllability literation citation, we set the self edges to zero. The natural threshold is set to $\frac{m}{K^2} = 22$. At this threshold the number of edges in the network is 74 and 4 controlling inputs are required which is a very small fraction the number of modules 14 in comparision to the original network where 0.9 fraction of the number of nodes were required. We further used Gene Ontology to functionally interpret the clusters obtained. The results are shown in figure 16.

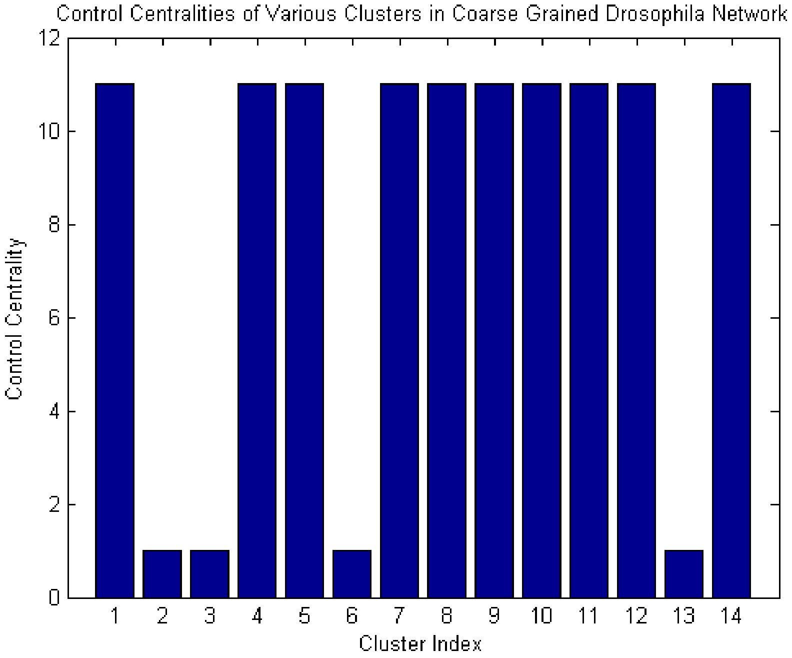
Figure 16: Significant GO terms for each cluster in Drosophila



Cluster Name	Significant Annotations
1	Regulation, Nervous System Development, Cellular Development, System Development, Cellular Organization.
2	Regulation of Development, Regulation of Differentiation, Development, Regulation of Neurogenesis, Dorsal Closure, Cell Fate Commitment, Zygotic determination fo Anterior Posterior Axis.
3	Membrane Organization, Localization, Hatching, Cell Migration, transport, signaling
4	Endocytosis, positive regulation Metabolic process, localization, regulation of polysaccharide process, glycogen process, glucose process.
5	Regulation, regulation of metabolic process, regulation of nucleobase metabolism, RNA metabolism, nitrogen compound metabolism, macromolecule metabolism, developmental process, reg of cellular biosynthetic process.
6	Microtubule based process, branched chain amino acid transport, leucine import, protein complex subunit organization, cellular organization, microtubuel cytoskeleton organization, organelle organization, mitotic spindle.
7	regulation of DNA replication, DNA metabolic process, regulation of multicellular organismal process, regulation of macromolecule metabolic process, regulation of DNA endoreduplication, DNA dependent DNA replication, wing disk development, instar larval or pupal morphogenesis,
8	Metabolic process, macromolecule metabolic process, cellular component organization, response to stimulus, signalling, golgi to plasma protein transport, phosphorous metabolic process, nervous system development, regulation of axogenesis.
9	tube development, morphogenesis of epithelium, tissue morphogenesis, post embryonic organ development, disc and instar morphogenesis, end of embryonic development, nervous system development,
10	protein modification process, metabolic process, macromolecule modification, centrosome duplication
11	Serine family metabolic process, defense response to virus, defense response to other organism, regulation of defense response, multicellular organism development, alpha-amino acid catabolic process, immune system process, response to biotic stimulus, response to bacterium.
12	metabolic process, nitrogen compound metabolic process, immune system development, lymphoid organ development, regulation of biosythetic process, cell proliferation, hemopoiesis,
13	cyclic compond metabolism, nitrogen compound metabolism, larval/pupal development, DNA damage response, post embryonic development, tube development, imaginal disc development, metamorphosis,
14	spermatid development, germ cell development, sperm individualization, spermatid differentiation, neuroblast proliferation, developmental process involved in reproduction,

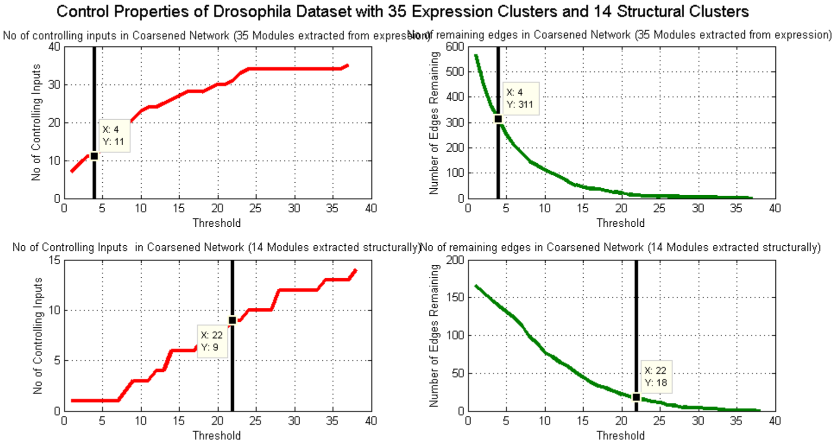
Next we investigated if any of these modules were more important than others in controlling the whole network. The control centralities of each of the modules is shown in figure 17. We found that each module is able to control a large part of the network on its own except modules $2, 3, 6$ and 13 which do not contain any transcription factors.

Figure 17: Control Centrality of Various Modules



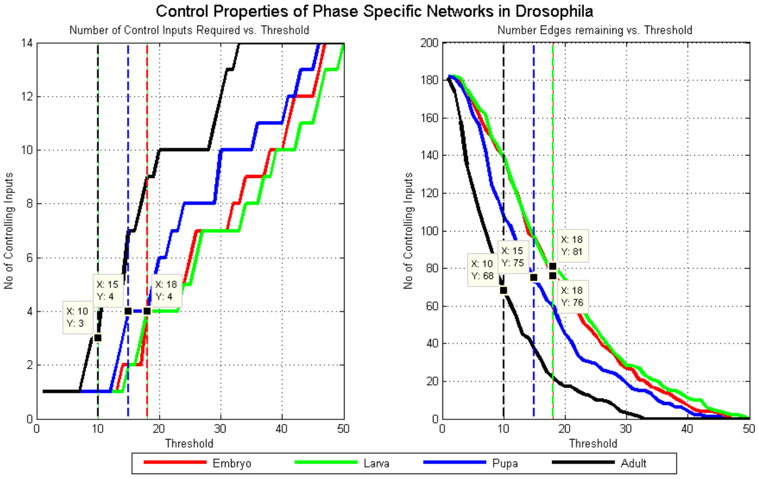
We also tested if these conclusion were robust to extracting modules differently(structurally) or extracting finer modules(K=35). We found that this indeed was the case. Only a small fraction of modules have to be controlled to control the network independently. These results are summarized in figure 18.

Figure 18: Control Properties of Drosophila Dataset with 35 Expression Clusters and 14 Structural Clusters



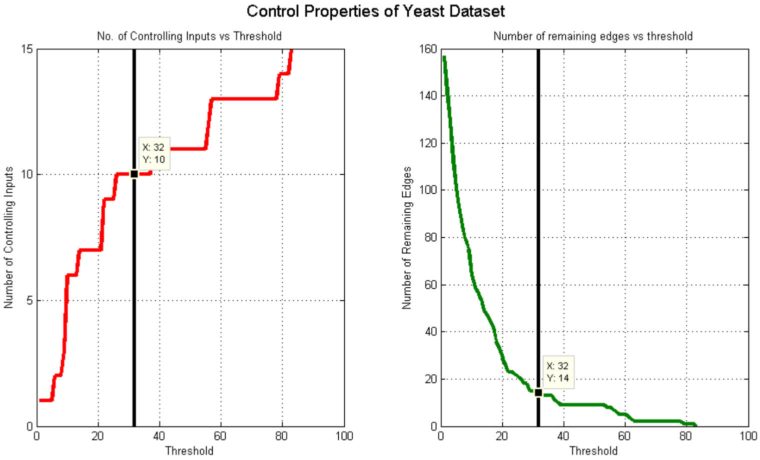
So far while studying the control properties of coarse-grained networks we assumed that the networks are static. We verified are results for relatively easier controllability were true phase specific networks in embryo, larva, pupa and adult stages. Again we observed only a small fraction was required to control the whole network. This is shown in figure 19.

Figure 19: Phase Specific Control Properties of Drosophila Dataset with 14 expression modules



Finally we repeated the experiment using structural modules extracted from the Yeast Transcription Network (common for all stages). Our conclusions were supported there too. The results are shown in figure this.

Figure 20: Control Properties of Yeast Data set with 15 structural modules



All the results from our experiments in this section are summarized in the following table:

Dataset	**No of Nodes (Coarse,n)**	**Edges**	**Number of Inputs(m)**	**Ratio (m/n)**
Drosophila (Expression Clustering)	14	74	4	0.28
Drosophila (Newman Girvan)	14	18	9	0.64
Drosophila (Expression Clustering)	35	311	11	0.31
Drosophila-Embryo	14	76	4	0.28
Drosophila-Larva	14	81	4	0.28
Drosophila-Pupa	14	75	4	0.28
Drosophila-Adult	14	68	3	0.21
Yeast (Structural Clustering)	15	14	10	0.67

7 Conclusion and Criticism

7.1 Conclusions

Using the framework of structural controllability it has been shown that biological networks are extremely hard to control requiring almot 0.9 fraction of total nodes to be independently controlled. But the signaling network doesn't seem to provide so many independent inputs. How is then efficient control achieved in these networks? The project makes the following key suggestions:

- Making a Biological Network fully controllable using very few inputs seems unlikely to be very useful. First such a network would not be robust as damage/noise to the few inputs would send the network into arbitrary and often harmful expression states. Instead a better strategy would be to identify functionally useful expression states and make those reachable.

- One way of achieving both robustness and efficient control using a few inputs is to not aim to control the whole network at one go but to make only the functionally useful part of it controllable. In terms of the expression space it would mean that rather than making the whole expression space reachable, only useful parts of it should be reachable. Robustness could be achieved if the harmful and unrequired expression states are unreachable. We provide some evidence for this hypothesis by analysis of phase-specific networks in Yeast and Drosophila and showing that the points of efficient control are different in different phases and consist of genes which can be interpreted to be useful in that phase.

- Additionally we pinpoint the structural cause of low controllability in Biological Networks. We find that the modules in biological networks are mostly of the form of a co-regulation graph which consists of a few nodes regulating many nodes. Such a subgraph forms a dilation. By analysis of expression data too, we show that a module seems to have an extremely low dimensional expression state. This seems to make sense since modules are functionally coherent units designed to perform a few functions and hence have an extremely constrained expression space.

- If the idea of a module as an independent functional entity holds then one would expect that there would be less dependence between different modules. We test this idea by apply structural control theory to coarse-grained networks and find that these network are more efficient to control.

7.2 Criticism and Future Work

However, our approach has the following key shortcomings and filling them is future work:

- There are three ways to define modules in these networks: structurally (using the network), functionally (using GO annotations) and expression-wise(by clustering expression data). We found that while each of these ways still support are arguments there is very little correspondence between the different modules obtained. The correct way to identify modules would be to integrate evidence from all these modalities. One can hope that the modules obtained this way would be closer to the hidden biological truth.

- The framework of structural controllability has been criticized because it makes predictions without knowing the parameters of the system and the behavior of a system can change drastically with its parameters. However one can argue that the framework makes statements which are true for all parameter values except for a set of measure zero and its unlikely that biological systems which are so noisy would be dependent on the parameters lying on a 0-measure set for their functioning. However the following argument still remains: if structural control theory says a system is controllable for all values except for a set of measure zero, it doesn't tell us how much energy or time it takes to drive the system to a target state. If the energy or time cost is too high then its as bad as being not controllable. [13] addresses this question.

- The coarse grained networks that we analyzed in the project are no longer proper dynamical systems. Its not clear if controllability is relevant to such informational networks. However current literature continues to analyze networks like Social networks, e-mail networks and so on under this framework[7].

- An inherent assumption in the controllability framework is that the internal time constant of each node is infinite. While the most widely accepted model for transcription network assumes that the transcription product decays with a rate proportional to its concentration, in spite of this while analysing transcription networks, one does not put self loops on each node. If one did, one would obtain trivial results like the system being controllable using only one input. Proper justification of this is still an open problem [4].

References

[1] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.

[2] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, page gkr1029, 2011.

[3] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[4] Noah J Cowan, Erick J Chastain, Daril A Vilhena, James S Freudenberg, and Carl T Bergstrom. Nodal dynamics, not degree distributions, determine the structural controllability of complex networks. *PloS one*, 7(6):e38398, 2012.

[5] Sophie Lebre, Jennifer Becq, Frederic Devaux, Michael PH Stumpf, and Gaelle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC systems biology*, 4(1):130, 2010.

[6] Ching-Tai Lin. Structural controllability. *Automatic Control, IEEE Transactions on*, 19(3):201–208, 1974.

[7] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[8] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Control centrality and hierarchical structure in complex networks. *Plos one*, 7(9):e44459, 2012.

[9] Nicholas M Luscombe, M Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A Teichmann, and Mark Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, 2004.

[10] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[11] Katsuhiko Ogata and Yanjuan Yang. Modern control engineering. 1970.

[12] Svatopluk Poljak. On the generic dimension of controllable subspaces. *Automatic Control, IEEE Transactions on*, 35(3):367–369, 1990.

[13] Gang Yan, Jie Ren, Ying-Cheng Lai, Choy-Heng Lai, and Baowen Li. Controlling complex networks: How much energy is needed? *Physical review letters*, 108(21):218703, 2012.