Explore gene regulatory network to search for yeast cell-cycle genes

Biophysics Project, PHYS-302

https://github.com/pworinger/Biophysics$_project$

Perrine Woringer, Viktor Crettenand

January 2021

Abstract

The goal of this project is to use the gene regulatory network from the IDEA dataset [1] to study yeast cell cycle. The topology of the network is studied and feedback loops containing cell-cycle genes and transcription factors are identified. The gene expression model from reference [1] is investigated to determine if it can reproduce known periodic oscillation of the genes involved in the cell cycle.

1 Introduction

In reference [1], R Scott McIsaac and his team trained a model for causal relationships between genes relative expression levels by inducing hundreds of transcription factors one by one and observing their effect in time on the expression level of other genes and proteins. From the parameters of the their model they could build a gene regulatory network (GRN).

Their gene regulatory network and their model will be used to study the cell cycle of budding yeast. GRN and the study of genes related to the cell cycle is of great importance for cancer research.

Most papers presenting models for yeast cell-cycle such as [2] and [3] use experimental data only to validate or slightly tune theoretical models. The approach discussed in this project is based on extensive experimental data and takes a holistic approach by looking not only at the few genes known to play a role in the cell cycle but at a larger gene network.

The motivation behind this project is to explore how gene regulatory network could be used to discover new mechanism and feedback loops involved in the cell cycle. This field of research is useful to understanding cancer better and look for cures for it.

2 Methods and results

2.1 Model

In [1], authors used a dynamic method to observe direct regulatory relationships and information propagation. By performing over 1650 experiments of transcription factor induction and fitting the following model to the resulting gene expression response, the causal relationships between genes were learnt:

$$\Delta \ln(y_{ijt})/\Delta t = \sum_k \left(\alpha_{ik}(y_{kjt} - 1) + \beta_{ik}(y_{ijt}y_{kjt} - 1)\right)/y_{ijt} \tag{1}$$

In this equation y_{ijt} is the relative expression of gene i with respect to the control strain at time t in experiment j. α_{ik} and β_{ik} represent respectively linear and quadratic regulatory effect of gene k on gene i. These are the two parameters that were fitted from the data.

The α_{ik} and β_{ik} coefficients are interpreted as weight for the directed edges from node k to node k and a threshold is applied to keep only the strongest edges. This provides a gene regulatory network containing ∼5300 genes and transcription factors and ∼110 000 regulatory relationships.

We selected a set of 12 genes and transcription factors from reference [2] that are known to play a role in yeast cell cycle, to be used as a starting point for our exploration of the network. The set is the following: 'CLN3', 'SWI5', 'CLN1', 'CLN2', 'CDH1', 'CDC20', 'CLB5', 'CLB6', 'SIC1', 'CLB1', 'CLB2', 'MCM1'.

2.2 Extracting subnetwork based on modularity

A subnetwork of genes involved in cell-cycle was selected from the original full gene regulatory network from [1]. This was done without limiting the scope of the subnetwork to nodes that are already known or suspected to influence cell-cycle. Our approach is inspired from [4] where it was used to identify new longevity genes in C. elegans from an interaction network.

The algorithm works as follows: It is initialized with the 12 genes listed in 2.1 and iteratively grow the subnetwork by adding the neighbouring nodes which maximising the subnetwork's modularity. The algorithm terminates when the modularity stops increasing.

To apply this approach to our network, the edges were treated as undirected. This makes sense since the nodes influencing cell-cycle genes and the nodes being influenced by cell-cycle genes are of equal interest to

build a self-contained model for yeast cell-cycle. For a given subnetwork, the more its nodes are connected to each other and the less its nodes are connected with other nodes outside of the subnetwork, the higher the modularity is. We expect that the higher the modularity is, the more the genes in the subnetwork are likely to participate in shared functions.

The final subnetwork grown from the 12 genes of interest (c.f. 2.1) contains ∼500 nodes among the ∼5300 nodes in the original network and has modularity $M = 0.08$ (to be compared to the modularity $M = 0.005$ of the start subnetwork containing only the 12 genes cited before).

3 Results and Discussion

3.1 Studying feedback loops

Some topological patterns in networks can help predict the system's behaviour. For example, negative feedback loops tend to produce signal oscillation if the number of node composing the loope is even. This is of interest when studying the cell cycle as it is expected to observe periodic oscillation of gene expression. For instance, it is known the transcription factor SWI5's level oscillates throughout the cell cycle [5]. A good model for cell cycle should replicate such oscillation. The topology of the network was therefor studies, searching for feedback loops.

To find feedback loops, a depth first search was performed starting from the 12 nodes of interest (listed in 2.1). All the paths coming back to the starting node (i.e all loops) were stored. The subnetwork obtained in 2.2 had no remaining loop visiting any of the 12 nodes, therefore, this search was performed on the original full gene regulatory network from [1]. Given that all these nodes except SWI5 had no child node in the directed network, only SWI5 appeared to be part of some cycles. Given the computational complexity of exploring the whole network (NP hard problem), we had to apply a cutoff value on the length of the cycles, and we therefore revealed only cycles of length at most 5 visiting SWI5.

The ∼10 millions cycles were then ranked according to the geometric mean of the loop edges. The use of the signed geometric mean γ which characterises the importance of the effect of the loop is justified in the appendix in section 5.1. The value of the edges corresponds to the parameter α in equation 1 which are the rates of change of the relative expression of genes (c.f. 5.1). The parameter β of equation 1 is not taken into account for this ranking for no sound mathematical interpretation could be found and neglecting β seems reasonable as a first order approximation.

On figure 1, the loops with largest $|\gamma|$ are shown. The green arrows correspond to positive α and the red arrows to negative α. The thickness of the arrows is proportional to the norm of α. The reader can verify that the loops described in table 3.1 can be found on this visualisation.

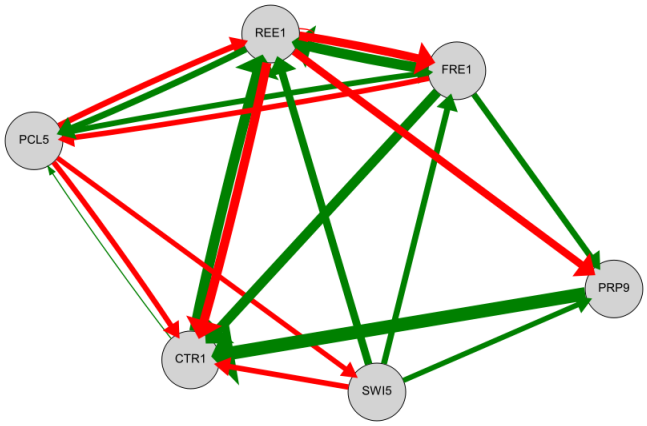

Figure 1: Representation of a portion of the graph containing the loops shown on table 3.1. Edges indicate causal relationship with source node being cause and target being effect. Green edges correspond to positive coefficient and red edges to negative coefficient. Width of edges is proportional to logarithm of the weight of corresponding edge

On table 3.1, the eight feedback loops with highest $|\gamma|$ are shown. Most of them are negative feedback loops. Generally positive feedback loops lead to runaway processes and divergence of the signal, while negative feedback loops produce either convergence or oscillations of the signal. As explained in the appendix in section 5.1, among negative feedback loops, those with an even number of nodes oscillate and those with an odd number of nodes converge. The loop with the largest γ which can be interpreted as the effect with the shortest characteristic time, is a negative feedback loop containing an odd number of nodes. This loop makes the signal (i.e. the relative expression of SWI5) converge and ensures the levels of SWI5 go back to their equilibrium level.

The histogram on figure 3 shows the distribution of $|\gamma|$ of positive and negative feedback loops. It can be noticed that there are slightly more negative feedback loops than positive ones among the feedback loops with highest absolute value of $|\gamma|$, and a rather balanced number of the two kind of loops for smaller absolute values.

$\gamma[s^{-1}]$	**Gene sequence**
$-2.50 \cdot 10^{-3}$	SWI5, PRP9, CTR1, REE1, PCL5
$-2.09 \cdot 10^{-3}$	SWI5, PRP9, CTR1, REE1, FRE1, PCL5
$-1.92 \cdot 10^{-3}$	SWI5, FRE1, PRP9, CTR1, REE1, PCL5
$-1.72 \cdot 10^{-3}$	SWI5, PIR1, FRE1, CTR1, REE1, PCL5
$+1.72 \cdot 10^{-3}$	SWI5, ASH1, GND1, YEF1, YJR061W, PCL5
$+1.72 \cdot 10^{-3}$	SWI5, PRP9, CTR1, REE1, CUP9, PCL5
$-1.67 \cdot 10^{-3}$	SWI5, PRP9, CTR1, HMX1, CPA2, PCL5
$-1.66 \cdot 10^{-3}$	SWI5, HXT2, FRE1, CTR1, REE1, PCL5

Table 1: Rate and gene sequence of the eight feedback loops with largest absolute value of rate

With the observed topology, we can expect that the strong negative feedback loops influencing SWI5 dominate over the positive feedback loops thus preventing divergence of SWI5's relative level of expression and allowing for oscillation. It is postulated that the oscillations are set into motion by positive feedback loops and are dampened by negative feedback loops of odd number of nodes.

This section on feedback loops argued that the converging effect of negative feedback loops counterbalances the diverging effect positive feedback loops because negative feedback loops of large $|\gamma|$ are more frequent. This analysis is entirely based on the values of $|\gamma|$ and the topology of the loop yet, this alone is not sufficient to assess the relevance of a feedback loop in the model. Indeed, nodes visited by a loop are also causally related to others in the network, which can result in interference. This unpredictability of the effect of a loop is enlarged for longer loops. It would be interesting to continue studying these loops more into detail and investigate whether it is relevant to study loops of larger size. It is expected that loops that are strongly connected to many nodes outside of the loop might be fed so much noisy signal that the "loop" model described in section 5.1 doesn't apply anymore. It would therefor be interesting to rank loops not only based on $|\gamma|$ but also on the loop length and it's connectedness. Additionally, comparing $|\gamma|$ to the geometric mean of the α values of all edges in the network (remember that $|\gamma|$ is the geometric mean of the α values of edges in a loop) might offer an insight on the strength of a given loop compared to other effects.

3.2 Numerical simulation of the model 1

The model described by equation 1 was implemented using a numerical method. All values of the vector y (relative expression level of genes) were initialized at 1, except for SWI5 which was submitted to a weak impulsive perturbation (its value was set to 1.01 for the first time step). On figures 4, we can see that a majority of the 12 cell-cycle genes selected 2.1 have their relative expression levels oscillate with diverging amplitude. The amplitude diverges as exp(exp(x)) (indeed the log relative expression level is linear on a log scale) which makes it hard to see these oscillations on figure 4. The fact that the model diverges is not surprising since most of the physics of what happens at large relative gene expression level is in higher order terms which were neglected in this model. On figure 2 the logarithm of the relative expression levels was plotted on a logarithmic scale. This allows to notice that some relative expression levels have common period of approximately 150 minutes, and to observe how they are shifted in time with respect to each other. Table 4 lists the genes that oscillate on figure 2. It can be observed that CLB1 and CLB2 have their level rise early. Later CDC20 increases immediately followed by CLN2. Later CLB6 kicks in. This order is compatible with the empirical data shown in table 4. One can also notice that the levels of CDC20 increase the most when the levels of CLB1 and CLB2 are highest which makes sense since the role of CLB1 and CLB2 are to make the cell enter mitosis. The increase of CLN1 that happens at the same time as the one of CLB1 is inconsistent with the order things should be happening in. CLN1 should spike after CDC20 and not before.

To go further in this numerical simulation, the model in [1] could be improved either by adding terms of higher order of by training a neural network which might be less interpretable but will probably yield more accurate predictions of relative expression levels and more stability. To stop the model from diverging, a simple model of gene decay was implemented that reduces the excess percentage of genes (compared to the equilibrium level) by 1.87% at every time step. This is a simplistic was of implementing gene decay which doesn't discriminate between genes. The value of 1.87% was chosen because it is the value that stops the system from diverging. It would be interesting to find data on decay values of each gene and implement a more realistic gene decay model.

4 Conclusion

In this project, a subnetwork of the IDEA network was selected by creating an algorithm that selects nodes based on network modularity and a numerical simulation of the model in [1] was implemented. This network and the full IDEA network were used to simulate the evolution of gene relative expression levels. The order in

which the studied cell cycle related genes spiked is consistent with experimental data except for CLN1 which spikes too early. Feedback loops containing SWI5 were identified and their effects were discussed based on their topology, the strength of the loop's effects ($|\gamma|$) and the proportion of positive and negative feedback loops. It was concluded that the feedback loops should have an overall converging effect which insures that SWI5's expression level stays stable and doesn't diverge.

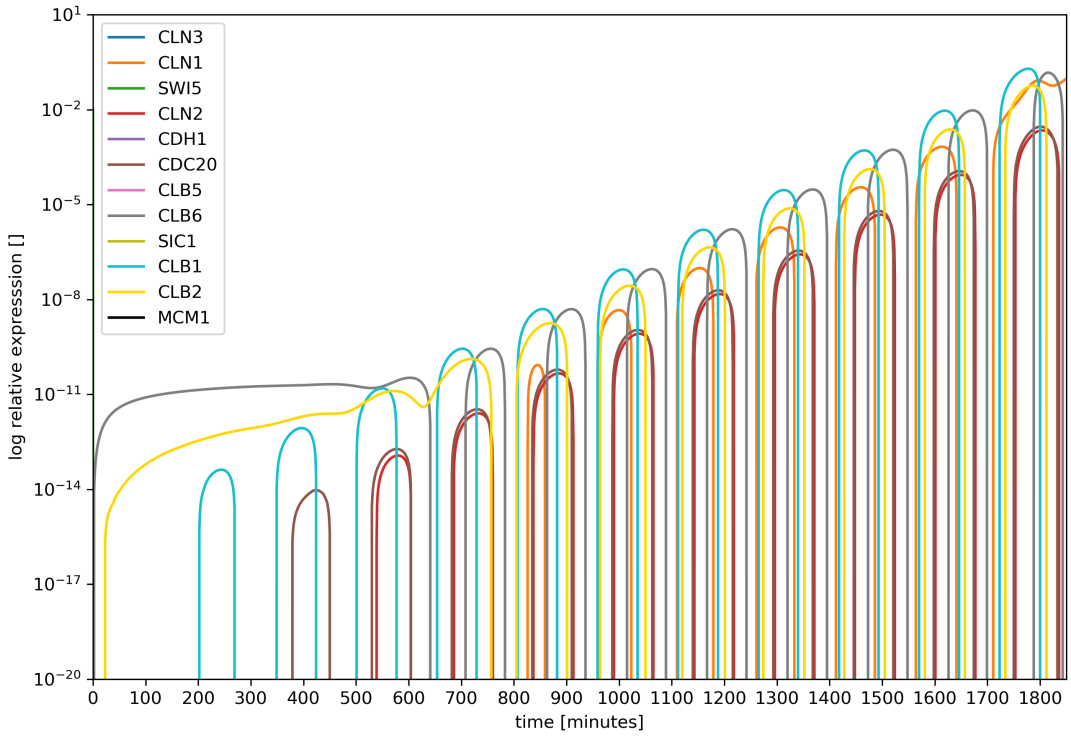

Figure 2: Plot of the logarithm of relative expression levels of the 12 genes of interest for the cell cycle studied in this project, with logarithmic scale on Y-axis. The values were obtained by numerical simulations using equation 1 and setting the initial relative expression of all nodes to 1 except SWI5 which was set to 1.01 and then back to 1.

Genes	**Phase**	**Role**
CLB1	G2	enter mitosis
CLB2	G2	enter mitosis
CDC20	M	activate the anaphase promoting complex
CLN1	G1/S	actuate CLN3
CLN2	G1/S	actuate CLN3
CLB6	G1/S	enter S phase

Table 2: Genes of interest for the cell cycle in chronological order (data collected from Wikipedia)

5 Appendix

5.1 Mathematical derivation of the relative expression level of a node in a loop

Consider a node in a loop. Let us makes the simplifying assumption that the loop is all there is. Lets also make the assumption that β in equation 1 vanishes. For a loop of three nodes the system of differential equations reduces to:

$$\begin{cases} \frac{dy_1}{dt} = \alpha_{13}(y_3 - 1) \\ \frac{dy_2}{dt} = \alpha_{21}(y_1 - 1) \\ \frac{dy_3}{dt} = \alpha_{32}(y_2 - 1) \end{cases}$$

One can simplify this system of equations by making the change of varriable $y \to x + 1$ and then integrating over x and t which yields:

$$\begin{cases} x_1 = \int \alpha_{13} x_3 dt + C_1 \\ x_2 = \int \alpha_{21} x_1 dt + C_3 \\ x_3 = \int \alpha_{32} x_2 dt + C_2 \end{cases}$$

Then the system of equations can be solved for x_1 which yields:

$$x_1(t) = \int_0^t \alpha_{13} \left(\int_0^{t_1} \alpha_{32} \left(\int_0^{t_3} \alpha_{21} x_1(t_2) dt_2 + C_3 \right) dt_3 + C_2 \right) dt_1 + C_1 \tag{2}$$

We can assume that at time $t = 0$ for all i except $i = 1$ $x_i = 0$ which corresponds to $y_i = 1$ that is the expression relative to the control strain is 1 (i.e. no difference with the control strain). Only y_1 is perturbed from it's equilibrium expression. This assumption allows us to set C_2 and C_3 to zero. The previous equation then reduces to:

$$x_1(t) = \alpha_{13}\alpha_{32}\alpha_{21} \int_0^t \int_0^{t_1} \int_0^{t_3} x_1(t) dt + C_1 \tag{3}$$

Since we are looking for solutions oscillating around the equilibrium without loss of generality the origin of the time can be chosen such that $C_1 = 0$. Let call the product of $D = \prod_{i=0}^{N=3} \alpha_i$. Then solutions can be found to the previous equation in the general case where the loop is composed of N nodes:

$$\begin{aligned} \text{if} \quad D > 0: \quad & \text{Ansatz}: x(t) = e^{t\gamma}, \gamma = D^{1/N} \\ \text{if} \quad D < 0 \text{ and } N \text{ odd}: \quad & \text{Ansatz}: x(t) = e^{t\gamma}, \gamma = -|D|^{1/N} \\ \text{if} \quad D < 0 \text{ and } N \text{ even}: \quad & \text{Ansatz}: x(t) = e^{-i^{\zeta} t\gamma}, \zeta = \frac{2}{N}, \gamma = -|D|^{1/N} \end{aligned}$$

In the previous equation $1/\gamma$ can be interpreted as a characteristic time in which the signal will converge, diverge or oscillate a period. From the previous equation it becomes clear that for loops with positive D, y_1 diverges, for loops with negative D and odd number of nodes N, y_1 converges to zero and for loops with negative D and even N the relative expression y_1 can oscillate periodically.

5.2 Definition of the modularity of a subnetwork

The modularity of a subnetwork is defined as:

$$M = \frac{w_{int}}{1 + (w_{int} + w_{ext})^2}$$

where w_{int} is the sum of the weights of all the edges between nodes of the subnetwork, and w_{ext} is the sum of the weight of all the edges between a node in the subnetwork and a node outside of the subnetwork.

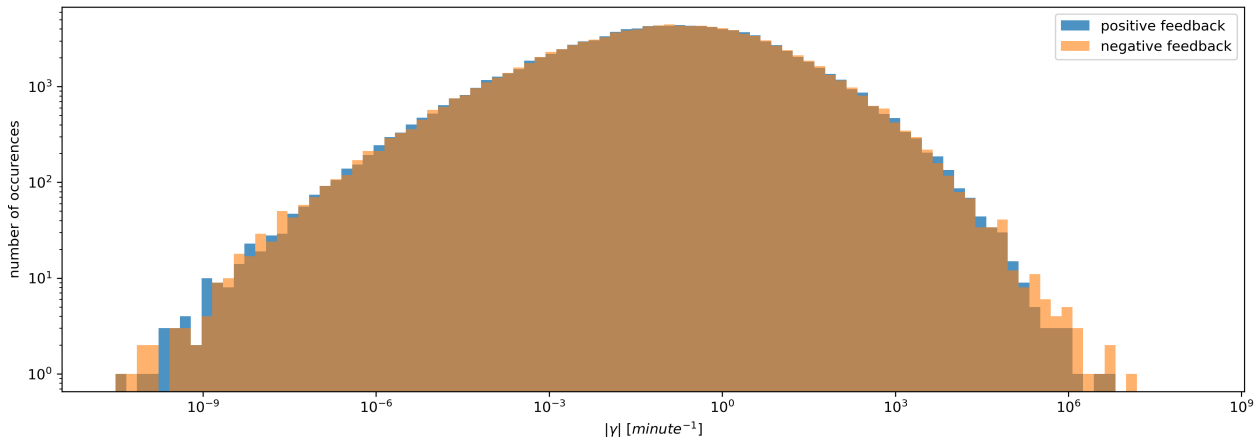

Figure 3: Histogram of $|\gamma|$ for all loops going through SWI5 of maximum length 6. We notice that the negative feedback loops (yellow) have slightly larger $|\gamma|$ while the positive feedback loops (blue) have slightly lower $|\gamma|$ but overall the two classes have a large overlap (brown).

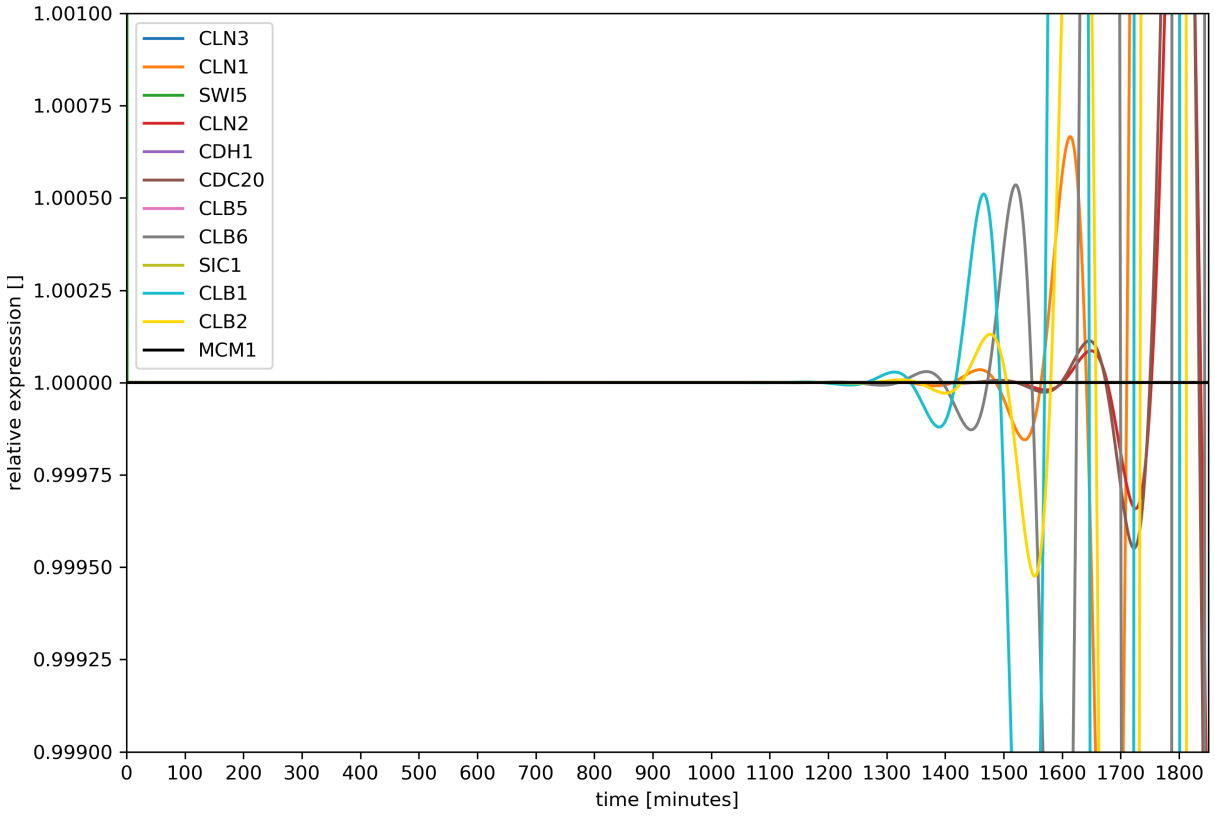

Figure 4: Plot of the relative expression levels of the 12 genes of interest for the cell cycle studied in this project. The values were obtained by numerical simulations using equation 1 and setting the initial relative expression of all nodes to 1 except SWI5 which was set to 1.01 and then back to 1.

References

[1] Marc Berndl, Marc Coram, Minjie Fan, R Scott McIsaac, Sean Hackett, and Ted Baltz. Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Molecular Systems Biology*, 2020.

[2] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences*, 101(14):4781–4786, 2004.

[3] Katherine C Chen, Laurence Calzone, Attila Csikasz-Nagy, Frederick R Cross, Bela Novak, and John J Tyson. Integrative analysis of cell cycle control in budding yeast. *Molecular biology of the cell*, 15(8):3841–3862, 2004.

[4] Kristen Fortney, Max Kotlyar, and Igor Jurisica. Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegansaging. *Genome Biology*, 11(2):R13, February 2010.

[5] Francisco José Taberner, Inma Quilis, Josep Sendra, María Carmen Bañó, and Juan Carlos Igual. Regulation of cell cycle transcription factor swi5 by karyopherin msn5. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(4):959–970, 2012.

[6] Michael Costanzo, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D Lee, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), 2016.

[7] Hanhae Kim, Junha Shin, Eiru Kim, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. Yeastnet v3: a public database of data-specific and integrated functional gene networks for saccharomyces cerevisiae. *Nucleic acids research*, 42(D1):D731–D736, 2014.