

Computational algebraic models of yeast cell cycle data.

Candace Curry

Advisor: Dr. Edward E. Allen

May 1, 2008

Abstract

Understanding the underlying biological cellular processes is important for advancements in medicine. Various mathematical methods can be employed to model cellular processes such as reproduction, differentiation, and apoptosis. These models provide biologists with tools for conjecturing cellular interactions which can then be tested in the laboratory. This research focuses on using and evaluating a computational algebraic technique developed previously [2] to cotemporally model microarray data of yeast cell cycles.

This computational algebraic method is based on interpolating polynomials over a finite field. It uses game theory to find high scoring relationships. Cotemporal models are time-invariant; the order of time points is not important. The microarray dataset used in the study, published by Pramila [8], measures gene expression across multiple cell cycles of yeast.

The yeast cell cycle is the process in which the unicellular organism reproduces by dividing into two complete daughter cells. This mechanism is controlled by levels of genes present at different phases of the cell cycle. The relationships between these genes can be modeled mathematically using this cotemporal algorithm. In order for modeling methods to be biologically useful, it is important that consistent results are obtained. It is also important to know which characteristics in the data are necessary to effectively employ the modeling method. Specifically, will cell cycles within datasets produce similar models? It was hypothesized that time shifted cell cycle data from a biological time-course dataset should produce similar cotemporal models. The models are compared using a rank correlation to determine similarity. Results do not show consistently strong correlations between models based on this method on this particular dataset. Additional study needs to be done to determine whether the data is biologically consistent or if the modeling algorithm is sensitive to the biological variations in the data.

Introduction

The goal of this research is to provide a tool for biologists to highlight the underlying relationships between biological molecules. Without such tools, expensive resources and time consuming laboratory experimentation must be employed. However, if mathematical models can be used to bring to light important molecular biological organization, advancements can potentially be made easier and faster. A modeling algorithm based on computational algebra has been previously developed by Allen, et al. that uses game theory to predict high scoring relationships between biological factors.

This algorithm incorporates a computational algebraic algorithm developed by Laubenbacher & Stigler [5]. It has the potential to be useful in modeling a variety of data types, including protein and gene expression.

Modeling the yeast cell cycle is one such application of this algorithm. With advancements in laboratory technology and online databases, data is easily obtained and accessible. The reasons to use yeast as a model organism are numerous. These unicellular eukaryotic organisms are easily kept in culture under laboratory conditions. They have a very similar cell cycle to humans and thus have comparable basic cellular mechanisms including DNA replication, recombination, cell division and apoptosis. As evidence of their importance, many proteins important in human biology were first discovered by studying their corresponding proteins in yeast; these proteins include cell cycle proteins, signaling proteins, and protein-processing enzymes. Also, yeast was the first eukaryote to have its genome fully sequenced (1996) so there are many data sets of gene expression available.

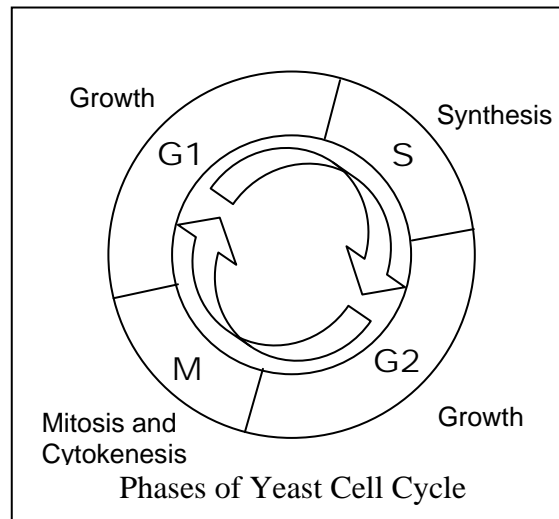


Figure 1. Shows phases of the cell cycle. Cell division occurs after mitosis. The yeast cell cycle is known to be regulated by the expression levels of certain genes, or transcription factors.

The yeast, *Sacharomyces cerevisiae*, genome contains over 6000 genes. The yeast cell cycle is consists of four phases: G1 (Growth 1), S (Synthesis), G2 (Growth 2), and M (Mitosis and Cytokinesis) (figure 1), which are regulated by transcription factors. Transcription factors are proteins that bind to specific parts of DNA and are part of the system that controls the transcription of genetic information from DNA to RNA. Through previous laboratory work, a mechanism for control of the yeast cell cycle has been determined. This known model is based primarily from transcription factor binding data [5]. However, models developed exclusively from laboratory techniques are very time and resource demanding.

Working with a known mechanism, such as the regulation of the yeast cell cycle, enables testing to be done to validate the effectiveness of the algorithm and determine appropriate parameters. To use the previously developed algorithm to get biological relevant results,

it becomes necessary to further explore limitations in terms of the types of data that can be input and the tolerance to variation in the data. This experiment hopes to gain a better understanding of the limitations of the algorithm by exploring whether modeling cell cycles within datasets will produce similar models. It was hypothesized that time shifted cell cycle data from a biological time-course dataset should produce similar cotemporal models. Mathematically, given that the modeling algorithm is not overly sensitive to the variations in the data, there should be consistency between models of subsets.

Next-State vs. Cotemporal modeling

A next-state model creates functions in which values of a set of variables (genes or proteins) at one point in time fully determine the values at the next point in time. Thus, given the initial values at the first time point and the functions, the entire data matrix is determinable. Biologically, many systems have cause and effect mechanisms that can be mimicked with next-state models. However, these models are strongly dependent on the distribution of the time points. Due to natural biological variation and inherent difficulties in laboratory sampling techniques, it is hard to ensure that the time points of a biological data set are consistent.

Cotemporal modeling techniques, which are used in this study, are not similarly dependent on time. A cotemporal model creates functions in which values of a set of variables (genes or proteins) at one point in time predict the values of the variables at the same point in time. Additionally, these functions are constant for all time points. Thus the same cotemporal model will be given if we permute the columns of the data matrix, where the data matrix consist of genes/proteins in the rows and time points in the columns. In modeling yeast cell cycles, time $t=0$ minutes should be equivalent to $t=60$ minutes due to the periodic nature of the cell cycle. Therefore removing the column of data at $t=0$ and replacing it with the data at $t=60$ should yield the same model if the data is consistent to the known period of the cell cycle (one hour) and does not exhibit extraneous biological variants.

Modeling Algorithm

The modeling algorithm developed by Allen et al [2], which integrates an earlier algorithm by Laubenbacher & Stigler [5], is based on computational algebra and game theory. It models data cotemporally and is time invariant. The algorithm interpolates polynomials over a finite field and uses game theory to find high scoring relationships. The software package CoCoA, Computations in Commutative Algebra (<http://cocoa.dima.unige.it/>), is used to run the algorithm.

The algorithm is limited as to the size of the dataset which can be input. A maximum of approximately 13 time points can be input without the processing time becoming prohibitive. This is due to the expected amount of time it takes the algorithm to run grows exponentially with each additional time point. However, the number of genes or proteins is not as limited because the expected run time only grows on the order of a polynomial.

Multiple discretizations are run on the data at an early stage in the process in order to compensate for natural variability and enable modeling over a finite field. Discretizations used including k-means, k-medoids, chi-merge, and mean over/under. A consensus of results from all discretizations methods is obtained at the end.

Data

Working with biological data presents both advantages and disadvantages. With the improvement in laboratory techniques of recent years and also the ability to access large amounts of data online through different databases such as GEO, the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>), a wide variety of biological data is readily available. However, biological data is also subject to natural variations that cannot always be controlled in a laboratory setting, making it difficult to model. Error inherent in the laboratory technique can also plague modeling efforts.

The data used in this research, from Pramila et al. 2006, is microarray data of gene expression across the yeast cell cycle. The original author's designation of alpha30 is used for the data set. Another dataset, alpha38, is a dye-swap technical replicate and also modeled for comparison of results. The datasets follow two full cell cycles with each cell cycle being approximately one hour. The yeast cells were sampled at an interval of 5 minutes for a total of 25 time points. Expression level values are given as signal log ratios normalized using the Rosetta Resolver. Cells are also synchronized to be in same phase of cell cycle before testing. Otherwise, without synchronization, no patterns in expression levels could be determined. An alpha factor was used to induce synchrony in these yeast cells [8].

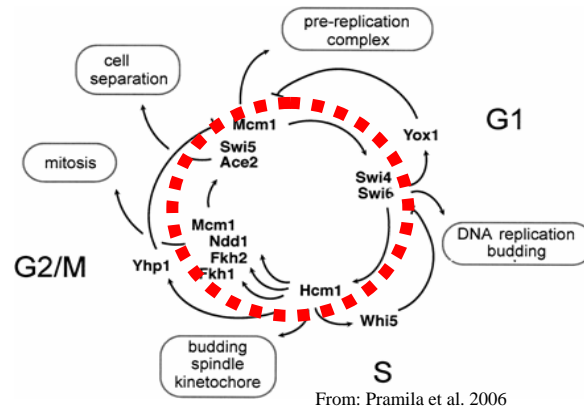


Figure 2: The literature model of the transcription factors affecting the yeast cell cycle. The transcription factors inside the red circle are modeled in this project. Transcription factors outside of the circle are known feedback inhibitors and were not considered in this experiment for simplicity of the models. From Pramila et al. 2006 [8].

Modeling all 6000 yeast genes would produce a model that was too complex to extract the important biological relationships. There must be some means of modeling only the most important factors. Because this research is using data with a known literature model, only those transcription factors known to be important in the literature model will be

input into the algorithm. Figure 2 diagrams important transcription factors as they are known to affect the yeast cell cycle. Because we are concerned with basic similarity, feedback inhibiting transcription factors (outside of the dashed circle) will not be used in hopes of increasing the simplicity of the models.

The expression levels of the transcription factors inside the dashed circle of figure 2 are graphed across time below (figure 3). All nine transcription factors studied are known to exhibit periodic expression across the cell cycle, as would be expected due to their regulatory characteristic [3]. From the graphs in figure 3, it can be seen that they all appear to have roughly a period of two. Synchronization effectively shocks the cells to all arrest in the same phase, from which they are released simultaneously. The first four time points may have leftover effects from the synchronization method which should be taken into consideration [8]. This is also reflected in the graphs, and can be seen especially in FKH1, HCM1, and SWI5. Because every yeast cell does not have exactly the same length phases, they are also known to lose synchronization over time [7]. This is potentially the source of increased variability evident in later time points in the plots. The expression levels of the second cell cycle are clearly less consistent than the first.

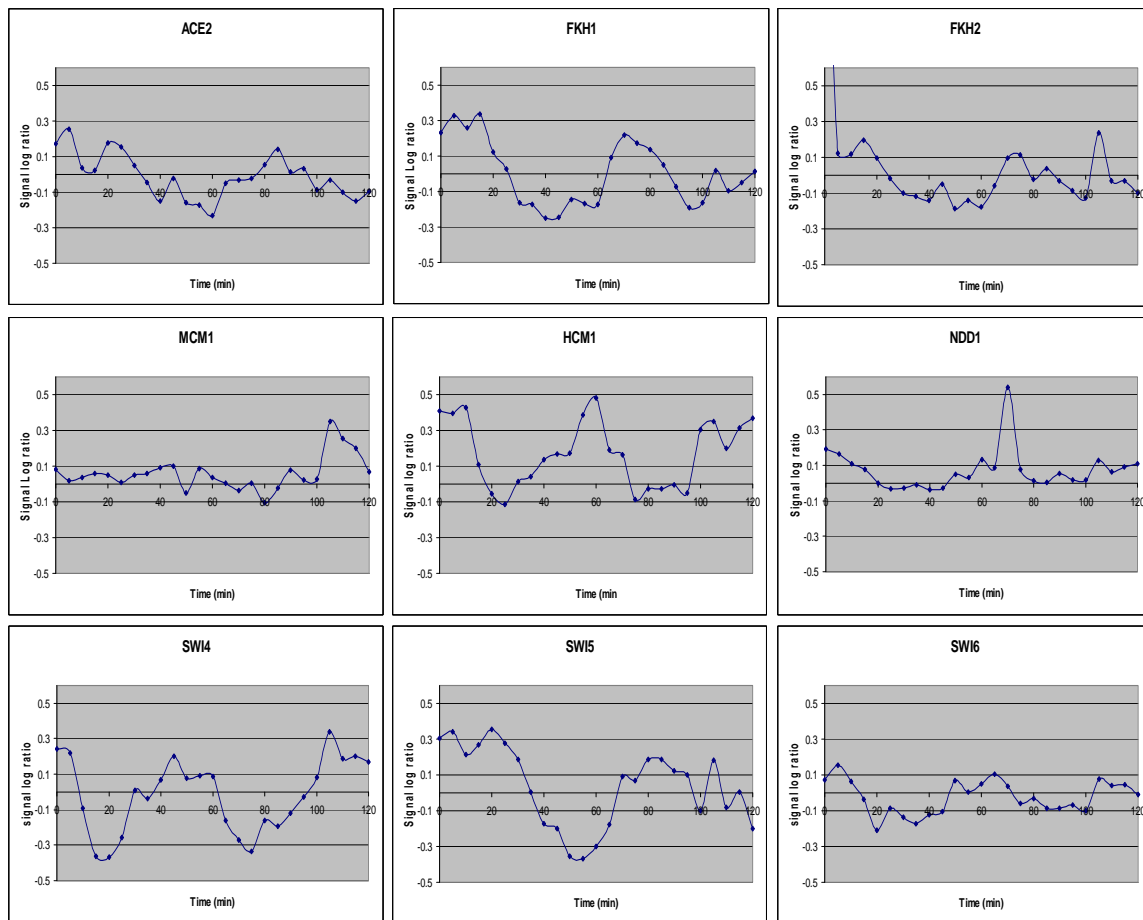


Figure 3 – The nine transcription factors modeled in this experiment. Graphs show gene expression level (in signal log ratios) vs time. All transcription factors shown are known to be periodic with the cell cycle.

Methods

In order to test the hypothesis that from a biological time course dataset covering multiple synchronized cell cycles, any subset will produce a similar model to any other subset when those subsets both span the length of a complete cell cycle and when modeled in a cotemporal manner, the Alpha30 dataset was split into overlapping subsets.

Each subset was the length of a cell cycle, one hour, yielding a total of 14 subsets. For example the first subset ranged from time points 0-55 minutes, the second from time points 5-60 minutes, etc (see figure 4). Based on this design, the first four subsets may be affected by possible leftover synchronization effects from the first four time points, which should be taken into consideration. The modeling algorithm is then run on these 14 subsets. Also a consensus model 1, which is a combination of all models, and a consensus model 2, which is a combination of all models except the first four is constructed for comparative purposes.

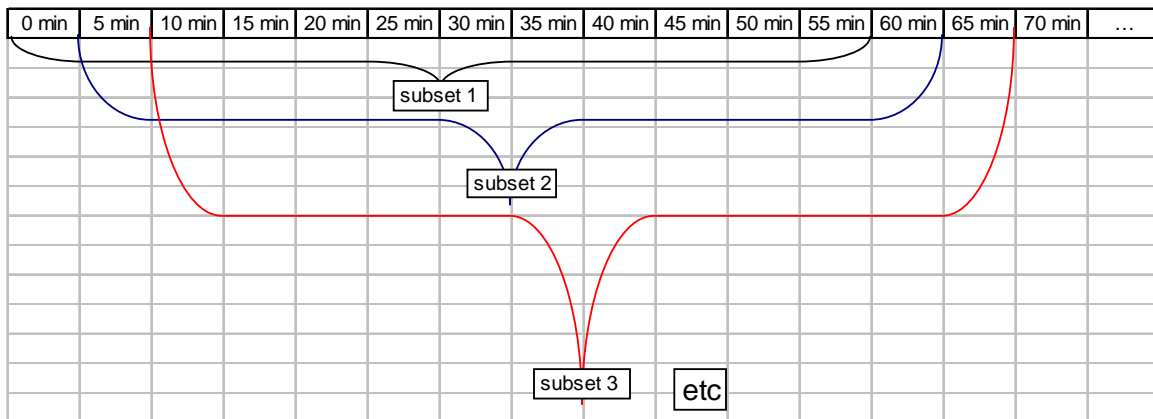
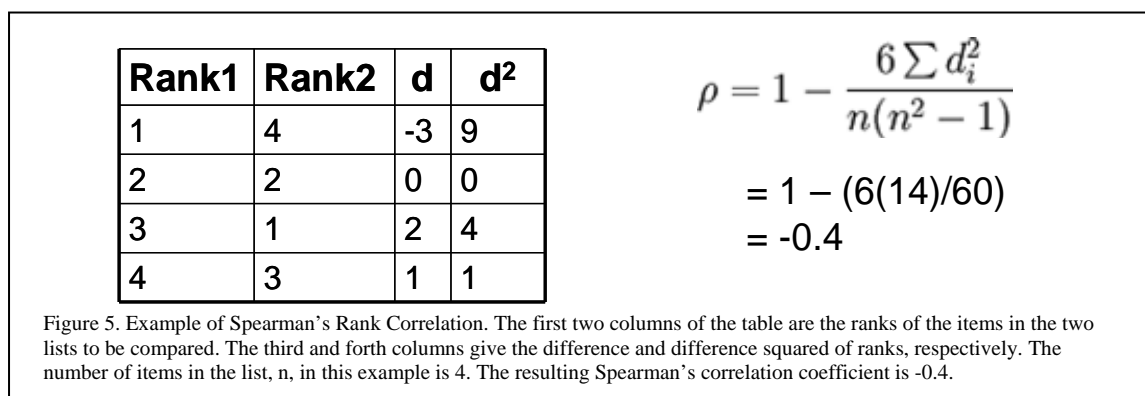


Figure 4. Diagram of experimental design. One subset theoretically covers an entire cell cycle. Therefore, points separated by one hour, such as time points 0 minutes and 60 minutes, should at be the same place in the cell cycle.

The modeling algorithm results in a list of all possible edges between transcription factors with a scores indicating how closely they are related cotemporally. The edges in a model are ranked with the highest related, thus highest scoring, edges first. All pairs of models are then compared a using Spearman's rank correlation. A simple example of Spearman's rank correlation is given in figure 5.



Results

Figure 6 shows a sample model from one of the subsets. The Spearman's rank correlation coefficients between all pairs of models for the alpha30 dataset are shown in table 1. There are not any consistently strong correlations between models, with the highest coefficients ranging around 0.5. However, what can be seen from the results is that the first four subsets show significantly less correlation with other subsets. This is supportive of the assumptions made that leftover synchronization effects plague the first four time points. Less correlation can also be observed in later subsets, possibly due to the increasing asynchronization of the data. Similar results can be seen in the Alpha 38 dataset.

Edge	Score
YLR131C/ACE2 ↔ YNL068C/FKH2	2.65
YCR065W/HCM1 ↔ YOR372C/NDD1	1.93
YOR372C/NDD1 ↔ YER111C/SWI4	1.62
YCR065W/HCM1 ↔ YER111C/SWI4	1.29
YOR372C/NDD1 ↔ YDR146C/SWI5	1.09
YCR065W/HCM1 ↔ YDR146C/SWI5	0.98
YMR043W/MCM1 ↔ YOR372C/NDD1	0.71
YNL068C/FKH2 ↔ YLR182W/SWI6	0.68
YNL068C/FKH2 ↔ YDR146C/SWI5	0.58
YLR131C/ACE2 ↔ YDR146C/SWI5	0.34
YLR131C/ACE2 ↔ YER111C/SWI4	0.31
YMR043W/MCM1 ↔ YER111C/SWI4	0.30
YNL068C/FKH2 ↔ YER111C/SWI4	0.29
YLR131C/ACE2 ↔ YLR182W/SWI6	0.27
YIL131C/FKH1 ↔ YMR043W/MCM1	0.20
YMR043W/MCM1 ↔ YDR146C/SWI5	0.18
YER111C/SWI4 ↔ YDR146C/SWI5	0.13
YCR065W/HCM1 ↔ YMR043W/MCM1	-0.07
YDR146C/SWI5 ↔ YLR182W/SWI6	-0.07
YLR131C/ACE2 ↔ YOR372C/NDD1	-0.15
YER111C/SWI4 ↔ YLR182W/SWI6	-0.24
YIL131C/FKH1 ↔ YDR146C/SWI5	-0.32
YIL131C/FKH1 ↔ YOR372C/NDD1	-0.40
YIL131C/FKH1 ↔ YER111C/SWI4	-0.42
YLR131C/ACE2 ↔ YCR065W/HCM1	-0.43
YOR372C/NDD1 ↔ YLR182W/SWI6	-0.43
YNL068C/FKH2 ↔ YCR065W/HCM1	-0.56
YNL068C/FKH2 ↔ YOR372C/NDD1	-0.58
YLR131C/ACE2 ↔ YMR043W/MCM1	-0.66
YCR065W/HCM1 ↔ YLR182W/SWI6	-0.70
YIL131C/FKH1 ↔ YCR065W/HCM1	-0.79
YIL131C/FKH1 ↔ YNL068C/FKH2	-1.28
YNL068C/FKH2 ↔ YMR043W/MCM1	-1.39
YLR131C/ACE2 ↔ YIL131C/FKH1	-1.41
YMR043W/MCM1 ↔ YLR182W/SWI6	-1.48
YIL131C/FKH1 ↔ YLR182W/SWI6	-2.20

Figure 6. Sample model from Alpha30, subset 9. All possible edges are given a score determined by the strength of their cotemporal relationship.

	0-55 min	5-60 min	10-65 min	15-70 min	20-75 min	25-80 min	30-85 min	35-90 min
0-55 min	1	0.1272	-0.0484	0.1243	0.019	-0.2286	-0.1629	0.2335
5-60 min	0.1272	1	-0.2479	-0.303	-0.0662	-0.1465	0.2929	-0.1894
10-65 min	-0.0484	-0.2479	1	-0.0649	0.2069	-0.1735	-0.0208	0.1691
15-70 min	0.1243	-0.303	-0.0649	1	0.2069	-0.1735	-0.0208	0.1691
20-75 min	0.019	-0.0662	0.2069	0.2069	1	0.0505	0	0.0023
25-80 min	-0.2286	-0.1465	-0.1735	-0.1735	0.0505	1	0.4994	0.3277
30-85 min	-0.1629	0.2929	-0.0208	-0.0208	0	0.4994	1	0.1961
35-90 min	0.2335	-0.1894	0.1691	0.1691	0.0023	0.3277	0.1961	1
40-95 min	-0.0821	0.0404	0.1387	0.1387	0.0245	0.2033	0.1753	0.1735
45-100 min	0.113	0.0093	-0.14	-0.14	0.3547	0.0075	-0.1786	-0.0363
50-105 min	0.0798	-0.1544	0.0206	0.0206	0.1894	0.182	0.052	0.2991
55-110 min	0.0901	0.0994	0.0916	0.0916	0.2525	0.0193	0.2873	0.1447
60-115 min	0.2152	0.1727	-0.1593	-0.1593	-0.1308	-0.0749	-0.0906	0.0404
65-120 min	-0.0036	0.1439	0.1604	0.1604	0.0958	0.2623	0.1979	0.1215
Consensus 1	0.0502	0.1024	0.0263	0.0263	0.5145	0.5681	0.5048	0.4857
Consensus 2	0.0149	-0.0139	0.0036	0.0036	0.5333	0.6389	0.4471	0.5022
	40-95 min	45-100 min	50-105 min	55-110 min	60-115 min	65-120 min	Consensus 1	Consensus 2
0-55 min	-0.0821	0.113	0.0798	0.0901	0.2152	-0.0036	0.0502	0.0149
5-60 min	0.0404	0.0093	-0.1544	0.0994	0.1727	0.1439	0.1024	-0.0139
10-65 min	0.1387	-0.14	0.0206	0.0916	-0.1593	0.1604	0.0263	0.0036
15-70 min	0.1387	-0.14	0.0206	0.0916	-0.1593	0.1604	0.0263	0.0036
20-75 min	0.0245	0.3547	0.1894	0.2525	-0.1308	0.0958	0.5145	0.5333
25-80 min	0.2033	0.0075	0.182	0.0193	-0.0749	0.2623	0.5681	0.6389
30-85 min	0.1753	-0.1786	0.052	0.2873	-0.0906	0.1979	0.5048	0.4471
35-90 min	0.1735	-0.0363	0.2991	0.1447	0.0404	0.1215	0.4857	0.5022
40-95 min	1	0.225	-0.3076	0.0569	0.0965	0.4268	0.4394	0.4376
45-100 min	0.225	1	-0.2133	0.4862	0.1295	-0.0631	0.305	0.3416
50-105 min	-0.3076	-0.2133	1	-0.0883	0.2234	-0.1277	0.3946	0.3763
55-110 min	0.0569	0.4862	-0.0883	1	-0.0414	-0.0734	0.3689	0.3426
60-115 min	0.0965	0.1295	0.2234	-0.0414	1	-0.1941	0.079	0.0728
65-120 min	0.4268	-0.0631	-0.1277	-0.0734	-0.1941	1	0.322	0.3375
Consensus 1	0.4394	0.305	0.3946	0.3689	0.079	0.322	1	0.9712
Consensus 2	0.4376	0.3416	0.3763	0.3426	0.0728	0.3375	0.9712	1

Table 1. Spearman's rank correlation coefficients between all models in the Alpha30 dataset. Larger values are bolded.

Analysis

Clearly, the subsets are not producing consistently similar models. To see why this is happening, the discretizations were tracked for each gene. Figure 7 shows the result of the discretization method for each gene for each of the 14 subsets. Along the diagonals, including the wrap around diagonal, the discretizations should be the same as they represent the same gene at the same point in the cell cycle. However, due to variability in the data and the nondeterministic nature of the discretization methods, the data is not grouped consistently. This can be visualized easily in gene one where different diagonals have been highlighted with shading. The second diagonal (lightest in color), for example, shows considerable variability with ten 0's and four 1's. Because of the differing discriminations, the algorithm is producing distinct models. There is enough variability that these models are not comparable.

References and Acknowledgements

- [1] Allen, E., Fetrow, J., John, D., Thomas, S., 2005. Heuristics for Dependency Conjectures in Proteomic Signaling Pathways, Proceedings of the 43rd Annual Association for Computing Machinery Southeast Conference, 75-79.
- [2] Allen, E., Fetrow, J., Daniel, L., Thomas, S., John, D., 2005. Algebraic Dependency Models of Protein Signal Transduction Networks from Time-Series Data, *Journal of Theoretical Biology*, *Journal of Theoretical Biology* 238 (2006) 317–330 .
- [3] Cho, R.J., et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65-73.
- [4] Cokus S, Rose S, Haynor D, Grønbech-Jensen N, Pellegrini M: Modelling the network of cell cycle transcription factors in the yeast *Saccharomyces cerevisiae*. *BMC Bioinformatics* 2006, 7:381.
- [5] Laubenbacher, R., Stigler, B., 2004. A computational algebra approach to the reverse engineering of gene regulatory networks, *Journal of Theoretical Biology* 229, 523-537.
- [6] Pecorella, A., 2006. *Algebraic modeling of biological signaling pathways*. Master's thesis, Wake Forest University.
- [7] Pramila, T., Miles, S., GuhaThakurta, D., Jemiolo, D., Breeden, L., 2002. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle, *Genes and Development* 16, 3034-3045.
- [8] Pramila, T., Wu, W., Miles, S., Noble, W.S., Breeden, L., 2006. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle.

This work was supported in part by the NSF-NIGMS Program in Mathematical Biology and NIH grant 1R01GM075304.