Identification of a Nonlinear Cell Cycle System with Linear Models

Martin Fransson, Markus Gerdin, Peter Fritzson

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: marfr@ida.liu.se, gerdin@isy.liu.se, petfr@ida.liu.se

10th June 2005
Abstract
A mathematical model by Chen et al. representing the cell cycle of the budding yeast cell (Saccharomyces cerevisiae) is used to simulate levels of 13 different cell biological variables. By using a linear systems approach on the generated data the goal is to identify the original equations of the mathematical model. Since this model is described by a nonlinear system the linear systems approach only gives correct estimations of the linear parts of the system, with one exception. A main issue thus becomes to identify which data that belongs to the linear equations and which does not. This problem is solved using two local linear models on different time intervals. Parameter estimations of coefficients from linear and time invariant equations should be the same over all time intervals.

Keywords: systems biology, cell cycle, budding yeast, auto regression model
Identification of a Nonlinear Cell Cycle System with Linear Models

Martin Fransson\textsuperscript{1,3}, Markus Gerdin\textsuperscript{2} and Peter Fritzson\textsuperscript{3}

\textsuperscript{1,3} Programming Environments Laboratory, Department of Computer and Information Science
\textsuperscript{2} Control and Communication, Department of Electrical Engineering
Linköping University
\textsuperscript{1}marfr@ida.liu.se, \textsuperscript{2}gerdin@isy.liu.se, \textsuperscript{3}petfr@ida.liu.se

Abstract

A mathematical model by Chen et al. [3] representing the cell cycle of the budding yeast cell (Saccharomyces cerevisiae) is used to simulate levels of 13 different cell biological variables. By using a linear systems approach on the generated data the goal is to identify the original equations of the mathematical model. Since this model is described by a nonlinear system the linear systems approach only gives correct estimations of the linear parts of the system, with one exception. A main issue thus becomes to identify which data that belongs to the linear equations and which does not. This problem is solved using two local linear models on different time intervals. Parameter estimations of coefficients from linear and time invariant equations should be the same over all time intervals.

Keywords: systems biology, cell cycle, budding yeast, Auto Regression model

1. Introduction

Through interactions between different kinds of proteins the activities and crucial events are regulated during the cell cycle in all organisms. A corresponding gene in the genetic code transcribes each protein. By understanding how a set of proteins interacts we can therefore get a picture of the interaction between the corresponding genes. Since the cell cycle control system in one kind of eucaryotic cells, for instance the budding yeast cell, is surprisingly similar to the one regulating cell division in mammal cells, the yeast cell serves as an important model for increasing the knowledge of how the mechanisms of cell division is regulated in our own cells.

By finding an accurate model for a real system, more tests could be performed in a simulation environment than on the real system. In the case of biology this might for instance mean fewer tests on animals in exchange for more computer simulations.

Whereas there is no point in trying to estimate a model for a system with already known equations, it is the subject of this paper to examine how this approach will work since it might later be used on systems with undetermined structures.

1.1 The Budding Yeast Cell

The budding yeast cell, Saccharomyces cerevisiae, is a single-celled eucaryote. Most people have come in contact with this kind of yeast since it is used in ordinary baking and beer brewing. From a genetic point of view it is also extremely important as a minimal model for eucaryotic cells. With its approximately 6300 genes it has a genome size much smaller than a mammal. Despite its small genome it still has all the essential functions used to control the crucial events during the cell cycle [1]. In a beneficial environment the budding yeast cell reproduces almost as quickly as bacteria, another property which makes it favourable for biological experiments [1].

The phases of the cell cycle in budding yeast are somewhat different from the same in most eucaryotes, since it does not have a G2 phase. Instead the S phase and the M phase are overlapping [1]. This means that a mitotic spindle is starting to form even as DNA synthesis is taking place, without any obvious chromosome condensation [3].

Another special characteristic for the budding yeast cell is its asymmetric cell division. This begins with the initiation of a bud on the surface of the cell. When the separation occurs the daughter cell will be smaller than the mother cell. This requires the daughter cell to enter an extended G1 phase for growth until it can begin budding of its own [3].

The Cdk (Cyclin-dependent kinase), which plays the biggest part in the cell cycle for budding yeast, is Cdc28. With proteins of either of two cyclin families, Cln1-3 or Clb1-6, it controls the main events during the cell cycle [3]. For instance, the dimmer consisting of Cln2/Cdc28 induce budding, Clb2/Cdc28 is necessary for proper completion of mitosis and Clb5/Cdc28 controls DNA replication. Sic1, a CKI (Cdk inhibitor protein) conjugates with some of the dimers and form a inactive complex [2]. While SBF, Mcm1 and Swi5 all acts as transcription factors in the budding yeast cell cycle, yet another Cdk, Cdc20, seems to play a role in unfinished DNA replication [3].
2. The Cell Cycle Model by Chen et al.

In the article Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle by K.C. Chen et al. [3] a mathematical model of the budding yeast cell cycle in form of a set of 13 ordinary differential equations with associated static relations is presented. Since that model is the one used to generate the data to be identified it is presented in this section. However, for complete explanations regarding the model and more details see [3]. The simplifications and approximations in model versus reality made by K. C. Chen et al. are also not dealt with here, see instead the article referred to. The research group behind the model in [3] is changing the model gradually when more knowledge is achieved. A somewhat extended model is for instance found in [2].

2.1 The Mathematical Model

If the selection of variables are made according to

\[
x_i(t) = \begin{cases} 
  Cln2 \\
  Cib2 \\
  Cib5 \\
  Sic1 \\
  Cib2/Sic1 \\
  Cib5/Sic1 \\
  Cdc20 \\
  [Hct1] \\
  mass \\
  [Ori] \\
  [Bud1] \\
  [SPN]
\end{cases}
\]

and the same values on parameters are used as in [3] the model of the cell cycle control system will appear as in (2) – (5).

Equations governing cyclin-dependent kinases

\[
\begin{align*}
\dot{x}_1(t) &= -0.1x_1 + 0.05x_0G_{SBF}(x_1, x_2, x_3, x_4, x_6, x_{10}) \\
\dot{x}_2(t) &= -0.01x_1 + 0.002x_0 - 0.05x_1 - 1.99x_9 \\
\dot{x}_3(t) &= 0.05x_0G_{Sic1}(x_2, x_4) \\
\dot{x}_4(t) &= -0.1x_3 + 0.006x_{10} - 0.25x_3x_8 \\
\dot{x}_5(t) &= 0.02 - 0.01x_4 + 0.1G_{Sic1}(x_2, x_3, x_4) - 0.3x_5x_4 \\
\dot{x}_6(t) &= -0.0201x_2x_4 - 0.3x_5x_4 + 0.05x_1 - 0.05x_4 \\
\dot{x}_7(t) &= +0.3x_2x_6 - 0.0016x_2x_{10} + 0.12x_7x_6 \\
\dot{x}_8(t) &= +50x_5 - 0.05x_4 - 1.99x_6x_7 - 0.3x_8x_4 \\
\dot{x}_9(t) &= -0.0201x_5x_6 + 0.3x_5x_4 + 0.002x_0 - 0.05x_4 \\
\dot{x}_{10}(t) &= +0.3x_5x_{10} - 0.0016x_2x_{10} + 0.05x_4 \\
\dot{x}_{11}(t) &= -0.12x_6x_{10} + 0.05x_4(6 + x_{10}) \\
\end{align*}
\]

Equations governing the Cib degradation machinery

\[
\begin{align*}
\dot{x}_b(t) &= -0.16x_6 + 50x_5 - 50x_5x_6 - 50x_5x_6 - 50x_5x_6 + 50x_6 \\
\dot{x}_c(t) &= -0.25x_6x_6 + 0.3x_6x_5 - 0.201x_2x_6 + 0.16x_6 + 0.05x_4 \\
\dot{x}_d(t) &= -0.3x_6x_6 + 0.0201x_2x_6 + 0.3x_6x_6 + 0.05x_4 \\
\dot{x}_e(t) &= -0.0016x_6x_{10} + 0.12x_6x_{10} + 0.05x_4(6 + x_{10}) \\
\end{align*}
\]

Equations for growth, DNA synthesis, budding and spindle formation

\[
\begin{align*}
\dot{x}_{12}(t) &= 0.005 + 0.06x_2 - 0.06x_5 - 0.08x_7 \\
\dot{x}_{13}(t) &= \begin{cases} 
  x_7 - 1.18x_4, & \text{END M} < \text{START S} \\
  x_7 - 11.08x_4, & \text{START S} < \text{END M} \\
  x_7 - (11.08 + 0.825\tau)x_4, & \text{END M} < \text{END M} + 12 \text{ min}
\end{cases} \\
\dot{x}_m(t) &= 0.04 - \frac{1}{1.05 - x_9} + 2x_9 - 0.04x_9 - 0.018x_4 \\
\dot{x}_n(t) &= -2x_9x_6 + 0.64x_6 + 0.64x_6 - 0.64x_6 + 0.05x_9 \\
\dot{x}_o(t) &= -0.32x_9x_6 + 0.66x_6x_6 + 0.32x_6x_9 \\
\dot{x}_{10}(t) &= -0.0128x_9x_{10} + 0.05x_9 + 0.05x_9 + 0.05x_9 + 0.05x_9 \\
\end{align*}
\]

(4)

(5)
where $G_{SBF}$, $G_{Mcm1}$ and $G_{Swi5}$ are Goldbeter-Koshland functions with their respective implicit variable dependence shown, (this is in contrary to how they are shown in [3]).

START\_S is the time when $[\text{ORI}] = 1$ (start of DNA synthesis) and END\_M when $[\text{SPN}] = 1$ (cell reaches metaphase). At cell division $[\text{BUD}]$ and $[\text{SPN}]$ are reset to zero, while $[\text{ORI}]$ is reset to zero only when a certain threshold condition has been fulfilled, see details in [3]. $\tau$ is the relative time with start at $t = \text{END\_M}$.

### 2.2 Solutions to the Cell Cycle Model

Solving the differential equation system in MATLAB with the initial conditions

$$
\begin{align*}
  x_1(0) &= [\text{Cln2}] = 0.01 \\
  x_2(0) &= [\text{Clb2}]_c = 0.2 \\
  x_3(0) &= [\text{Clb5}]_c = 0.1 \\
  x_4(0) &= [\text{Sic}]_c = 0.05 \\
  x_5(0) &= [\text{Cdc20}]_c = 0.6 \\
  x_6(0) &= [\text{Hct}] = 0.5 \\
  x_{10}(0) &= mass = 0.71 \\
  x_{11}(0) &= [\text{ORI}] = 21.5 \\
  x_{12}(0) &= [\text{BUD}] = 0 \\
  x_{13}(0) &= [\text{SPN}] = 0
\end{align*}
$$

(6)

during four cell cycles will generate plots according to Figures 1 – 4. The variable $x_{10} = mass$ (solid line) is shown in as a reference in all figures.

As can be seen in Figures 1 – 4 the first cycle is about 50% longer than the following three cycles. This
is due to the asymmetric division of the budding yeast cell [3]. However, the cycle becomes stable (in the sense that the periods become uniform) with a final length around 97 minutes after a couple of divisions.

3. Estimation with Auto Regressive Models

Since the system of equations describing the cell cycle (2) – (5) does not contain any input signals, an Auto Regressive (AR) model will be used to model the data [6]. With this approach the system to be identified will be assumed to be linear, which in our case it is clearly not! However, it is the subject of this paper to examine the very effects of applying linear methods of identification on nonlinear systems.

An AR model (discrete time model) of order \( n \) can be written as

\[
y(t) + a_1 y(t-1) + a_2 y(t-2) + \ldots + a_n y(t-n) = e(t)
\]

where \( y(t) \) is the output signal and \( e(t) \) is a purely random error source with zero mean [5]. The parameters \( a_1 - a_n \) are estimated by first making a prediction of the output signal by removing \( e(t) \)

\[
\hat{y}(t | \theta) = \theta^T \phi(t)
\]

where

\[
\theta = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \quad \phi(t) = \begin{pmatrix} -y_1(t-1) \\ -y_2(t-2) \\ \vdots \\ -y_n(t-n) \end{pmatrix}
\]

The system (2) – (5) consists only of first order derivatives, i.e., the order \( n = 1 \), but since it is a system with multiple variables an introduction of an index \( m \) where \( 1 \leq m \leq 13 \) is needed. The model (7) and the expressions \( \theta \) and \( \phi(t) \) in (8) are hence modified to (9) and (10) respectively.

\[
y_1(t) + a_{11} y_1(t-1) + a_{12} y_2(t-1) + \ldots + a_{1m} y_m(t-1) = e_1(t) \\
y_2(t) + a_{21} y_1(t-1) + a_{22} y_2(t-1) + \ldots + a_{2m} y_m(t-1) = e_2(t) \\
\vdots \\
y_m(t) + a_{m1} y_1(t-1) + a_{m2} y_2(t-1) + \ldots + a_{mn} y_n(t-1) = e_m(t)
\]

\[
\theta = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}, \quad \phi(t) = \begin{pmatrix} -y_1(t-1) \\ -y_2(t-1) \\ \vdots \\ -y_m(t-1) \end{pmatrix}
\]

The estimation of \( \theta \) is determined by minimization of the least-squares criterion

\[
V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} (y(t) - \hat{y}(t | \theta))^2
\]

hence

\[
\hat{\theta}_N = \arg \min_{\theta} V_N(\theta)
\]

where \( N \) is the number of samples [5].

Since the cell cycle model (2) – (5) is time continuous, the estimated parameter values must then, for the sake of comparison, be transformed according to

\[
e^{At} = -\hat{\theta}_N
\]

where \( T \) is the time sample interval [4]. \( A \) is the matrix in

\[
\hat{x}(t) = Ax(t)
\]

which is a linear system of first order differential equations (where any constants have been ignored). A general system of first order differential equations is expressed as

\[
\dot{x}(t) = f(x(t))
\]

which is the form of the cell cycle model, (2) – (5).

4. Estimation in a Simplified Cell Cycle Model

Before starting to examine how well the AR model fits the budding yeast cell cycle model it seems reasonable to first try the approach on a less complex system. Two guidelines for constructing such a simplified cell cycle model will be regarded. The system describing the model should

1. be as simple as possible
2. have some biological relevance

Item 1 means that the equations making up the system should be as few as possible. Since the AR model will be applied on a system with multiple variables a system as (16) of two differential equations will be the simplest possible with this property.

\[
\dot{x}_1 = f_1(x_1, x_2) \\
\dot{x}_2 = f_2(x_1, x_2)
\]
Item 2 means that the simplified cell cycle model should share some characteristics with the budding yeast cell model. With the exception of one, all equations governing the derivatives in (2) – (5) are at “worst” multivariate rational functions and at “best” linear functions, hence it seems appropriate to start with a linear model and then change it so it becomes nonlinear. Item 2 also means that one should consider the numeric values of the parameters in (2) – (5) to later obtain data in the same range.

4.1 A Linear Cell Cycle Model

If \( f_1 \) in (16) is chosen as a function governing mass growth and \( f_2 \) is chosen as a function governing a protein concentration, which in turn is positively dependent on the mass growth, (17) could be the choice of a linear differential equation system.

\[
\begin{align*}
\dot{x}_1 &= 0.006x_1 \\
\dot{x}_2 &= 0.5x_1 - x_2
\end{align*}
\] (17)

Here, the parameter values and the following initial conditions have been chosen in such a way that they are somewhat similar to the ones in the budding yeast cell model, e.g., compare the equation governing \( x_1 \) in (17) with the equation governing \( x_{10} \) in (5).

The solution to (17) with initial conditions \( x_1(0) = 1 \) and \( x_2(0) = 0.1 \) can be seen in Figure 5.

![Figure 5](image1.png)

Figure 5. The solution to (17). The mass growth \( x_1 \) (solid) when \( x_1(0) = 1 \) and a protein concentration \( x_2 \) (dashed) when \( x_2(0) = 0.1 \).

Applying the procedure described in section 3 on (17) with \( N = 1001 \), i.e., the sample rate is 10 samples/minute, will give the estimations of the parameters as an \( A \)-matrix corresponding to the one in (14)

\[
A = \begin{bmatrix} 0.0060 & 0.0000 \\ 0.5002 & -1.0001 \end{bmatrix}
\] (18)

The accurate result is not surprising, since (17) is a linear system and the value of \( N \) is higher than really needed (\( N \)-values ten times smaller still give satisfying results).

4.2 A Nonlinear Cell Cycle Model

The model (17) will now be modified in a way that makes it more complex and nonlinear. Assume that the mass growth \( x_1 \) is inhibited by the protein concentration \( x_2 \). Assume also that the equation governing the protein concentration contains a rational function with a Michaelis-Menten like appearance. Introducing these two properties gives

\[
\begin{align*}
\dot{x}_1 &= 0.006x_1 - 0.01x_2 \\
\dot{x}_2 &= 0.5x_1 - \frac{x_2}{0.2 + x_2}
\end{align*}
\] (19)

It is now possible to examine how a nonlinear equation affects the parameter estimation, directly in the second equation in (19) and indirectly in the first equation.

Solving the system with the same initial conditions as previously will give the plot in Figure 6.

![Figure 6](image2.png)

Figure 6. The solution to (19). The mass growth \( x_1 \) (solid) when \( x_1(0) = 1 \) and a protein concentration \( x_2 \) (dashed) when \( x_2(0) = 0.1 \).

Estimation of the parameters in (19) with \( N = 1001 \) gives

\[
A = \begin{bmatrix} 0.0060 & -0.0100 \\ 0.0131 & -0.0387 \end{bmatrix}
\] (20)

The result indicates that it is possible to estimate parameters in linear equations correctly even if a dependence of a nonlinear equation exists, but no parameters will be estimated properly in the nonlinear equation. Even if \( N \) is increased by ten times the conclusions remain.
4.3 Estimation with Local Models

The next step in the investigation will be to divide the time interval in two parts and make an autoregressive model for each interval. This could perhaps be a solution on how to correctly estimate a parameter in a nonlinear equation, such as the value $0.5$ in the equation governing $x_2$ in (19). What might be expected when applying this approach to the linear model (17) are correct estimations for each interval. However, as will be seen, this is only the case under certain conditions.

To simplify the discussion of where to divide the cell cycle a dimensionless parameter $c$ is defined as

$$c = \frac{\text{time instant for cut}}{\text{length of cell cycle}}$$

The cell cycle for the linear model (17) is now divided so $c = 0.5$, i.e., at time $t = 50$ minutes. The intervals will be equal in length according to Figure 7.

![Figure 7. The solution to the linear system (17) divided so that $c = 0.5$, with an $A$-matrix to be estimated for each interval. $x_1$, solid, and $x_2$, dashed.](image)

The parameter estimations for each time interval are given in matrix $A_1$ (22) and $A_2$ (23) respectively. $N = 1001$ for the entire cycle.

$$A_1 = \begin{pmatrix} 0.0060 & 0.0000 \\ 0.5002 & -1.0001 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 0.0060 & 0.0000 \\ 0.0768 & -0.1485 \end{pmatrix}$$

The estimations in (22) are the same as in (17), but the ones in (23) differ in the second equation. This is due to information loss in the second interval concerning the initial slope of the solution of $x_2$. Hence, $c$ must be chosen in a way that makes both intervals contain information about the initial slope of $x_2$, compare with persistently exciting input signals [6]. The problem is solved by setting $c = 0.02$ according to Figure 8.

![Figure 8. The solution to the linear system (17) divided so that $c = 0.02$, with an $A$-matrix to be estimated for each interval. $x_1$, solid, and $x_2$, dashed.](image)

With this choice of $c$ the parameter estimations for the intervals instead become

$$A_1 = \begin{pmatrix} 0.0060 & 0.0000 \\ 0.5003 & -1.0004 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 0.0060 & 0.0000 \\ 0.4995 & -0.9986 \end{pmatrix}$$

In (25), the second equation is approximately the same as the corresponding equation in (17).

The approach with local models will now instead be applied to the nonlinear system (19). The intention is to investigate if it is possible to correctly estimate the parameter in front of $x_1$ in the nonlinear equation. Since the solutions for (17) and (19) are similar in the sense that $x_1$ contains an initial slope (Figures 5 and 6), $c$ will be kept at 0.02.

The results of the parameter estimations for the two intervals are

$$A_1 = \begin{pmatrix} 0.0059 & -0.0092 \\ 0.3134 & -1.5858 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 0.0060 & -0.0100 \\ 0.0002 & 0.0083 \end{pmatrix}$$

and the corresponding plots can be seen in Figure 9.

When considering the equation governing $x_2$ in (26) and (27), a comparison between the parameters reveal
that the first estimation matrix $A_1$ is closer than $A_2$ to the true values of system in (19). The $x_1$ has its greatest influence on the derivative of $x_2$ in the beginning of the cycle. Further investigations showed that the estimated parameter value gets close to 0.5 (the true value) if $c$ is even smaller but if $c$ becomes too small, $N$ must be increased. It does not seem likely that any satisfying conclusions of parameter values in nonlinear equations can be made with the two local model approach described.

5. Estimation in the Budding Yeast Cell Cycle Model

The approach with auto regressive models will now be applied to the system (2) – (5) describing the budding yeast cell model. The previous section has shown what can be expected from such an investigation. It might be possible to make correct parameter estimations of the linear equations in the system, which are the ones governing variables $x_7$, $x_{10}$ and $x_{11}$. To be able to identify only 3 of 13 equations is of course less than satisfying. However, it will, if successful, lead to a somewhat less complex system on which other methods of identification can be used.

Since a priori knowledge of the cell cycle model is limited to the information that it consists of 13 variables, how is it possible to detect which variables are governed by linear equations and which are not? Applying an AR model directly on the data from a stable cycle will result in a $13 \times 13$ $A$-matrix from which it will be difficult to determine which estimates represents true parameter values. By estimating two local linear models and applying the fact that parameter estimates in linear equations should be equal in both models detection of linear equations might be possible. This step, and the preceding one, concerning data pre-processing, will require visual inspection of a stable cell cycle, which is presented in Figures (10) – (13).
The number of estimated parameters increases quadratically with respect to the number variables. A method where potential linear equations can be detected without comparing numeric values from two local linear models is therefore desirable. This is solved graphically by first defining a matrix \( R \) as

\[
R = \text{abs}(A_1 - A_2)
\]  

Each element in \( R \) is then represented on a grey-scale, were zero-valued elements are black and elements with a greater value than 1 will be white. In linear equations, where estimations should be the same in \( A_1 \) and \( A_2 \), the elements in \( R \) will thus be 0 and such an equation will appear black. The choice of the value 1, and not for instance 100 or 0.01, as the limit of the other end of the grey-scale, is an estimation based on the approximate size of estimated parameters. Such an estimation can be made by for instance creating a linear model of the complete time interval.

By using a sample rate of 10 samples/minute (\( N \approx 1000 \)) the graphical representation of \( R \) is presented in Figure 14.

![Figure 14. The pattern showing the scaled image of \( R \) for \( c = 0.3 \) at 10 samples/minute](image)

In this figure the equation governing \( x_{10} \) seems likely to be a linear equation. Increasing the sample rate by ten times to 100 samples/minute, Figure 15, will present three more supposedly linear equations, the ones governing \( x_7, x_{11} \) and \( x_{12} \) respectively.

5.1 Analysis of Simulation Data

The cell cycle as modelled by (2) – (5) is assumed to be a system without perturbations, i.e., there is no reason to pre-process data due to measurement noise etc. However, the curves representing \( x_8 \) in Figure 12 and \( x_{11} \) in Figure 13 contain sharp sections. These are either the result of a nonlinear equation or the cell division. In the case of \( x_8 \) the sharp section occurs at the middle of the cycle, which makes \( x_8 \) a strong candidate to be governed by a nonlinear equation. On the other hand, the sharp section in \( x_{11} \) occurs only some half minute after cell division, which could be a delayed effect of this event. If the equation governing \( x_{11} \) is linear, an attempt to estimate parameters will be the same as trying to estimate a nonlinear equation, since the delayed effect of the cell division makes \( x_{11} \) nonlinear. By removing the data from the initial half-minute of the cycle, \( x_{11} \) will be possible to estimate if it is governed by a linear equation. Since the loss of information is limited to relatively few samples it is assumed that the measure taken will not affect the correctness of the parameter estimations in the other equations to a relevant degree.

5.2 Analysis of a Graphical Model Representation

The second step will be to find a \( c \)-value such that two locally linear models both give accurate parameter estimations for any linear equations. The determination of an appropriate \( c \)-value is done by visual inspection of the behaviour of each variable in Figures 10 – 13.

During the time interval between 30 – 40 minutes many of the curves change characteristics, e.g., \( x_4 \) and \( x_9 \) both go from an oscillatory mode to a flat mode, while \( x_5 \) does the opposite. Hence, by setting \( c = 0.3 \) both local models will be based on data which contain excited modes from most variables.
Further increasing the sample rate to 1000 samples/minute (Figure 16) will however not reveal any more likely linear candidates.

A comparison with the true nature of these four equations (30) reveals that two equations differ in character.

\[
\begin{align*}
\dot{x}_7 &= 0.06x_2 - 0.06x_4 - 0.08x_7 + 0.005 \\
\dot{x}_{10} &= 0.80x_2 + 2.0x_3 - 0.79x_5 - 2.0x_6 - 0.06x_9 \\
\dot{x}_{11} &= 0.30x_1 + 0.30x_3 - 0.30x_6 + 0.0008x_{10} - 0.060x_{12} \\
\dot{x}_{12} &= 0.3x_1 + 0.3x_3 - 0.3x_6 - 0.060x_{12} + 0.006 - \frac{x_{10}}{6 + x_{10}}
\end{align*}
\]

The equation governing \(x_7\) does not have a dependence of variable \(x_{10}\) but instead contains a constant value. However, since the possibility of equations containing constant values were not accounted for, the estimation procedure will regard any constant as a dependence of the variable behaving most as a constant. In the current case this is \(x_{10}\) since it has an almost linear increase, ranging from approximately \(x_{10,\min} = 1.0\) to \(x_{10,\max} = 1.7\).

The second equation being significant different (i.e., more than variations in parameter estimates) is the one governing \(x_{12}\), which is in fact nonlinear. Like the previous case, it is due to the nature of variable \(x_{10}\) the nonlinear equation appears to be linear. The last term in the equation can be written according to

\[
0.006 \frac{x_{10}}{6 + x_{10}} = 0.006 \frac{x_{10}}{6 + x_{10}}
\]

With values for \(x_{10,\min}\) respective \(x_{10,\max}\) inserted the term will become the expressions

\[
\begin{align*}
\frac{0.006}{6 + x_{10,\min}} x_{10} &= 0.006 \frac{x_{10}}{6 + 1.0} = 0.00086x_{10} \\
\frac{0.006}{6 + x_{10,\max}} x_{10} &= 0.006 \frac{x_{10}}{6 + 1.7} = 0.00078x_{10}
\end{align*}
\]
which both have a parameter value in the vicinity of the estimated value of 0.0008 of the parameter in (29).

Hence, it can be concluded that the exact nature of the equations (in terms of linearity) from the graphical analysis cannot be determined due to occurrences of constants and special cases of nonlinear terms. However, under the conditions that a $c$ exists such that both time intervals will contain enough excitation from all variables and that the sample rate is sufficiently high, equations not appearing black will be likely to rule out as linear due to the big differences in parameter estimates between the two local linear models.

6. Conclusions
We have investigated system identification using linear models of a nonlinear cell cycle system defined by a mathematical model from Chen et al. A simple nonlinear system related to the Chen et al. system was first studied in order to determine how linear system identification behaves when applied to nonlinear systems such as the Chen et al. model.

Coefficients in the linear equations were identified rather well, whereas identification of the nonlinear equations was not successful. To determine whether an identified equation is linear or nonlinear, the simulated system time interval was divided in two subintervals with different dynamics. A local linear model was identified on each subinterval. It turned out to be possible to detect a linear equation by the property of having essentially identical coefficients on the two subintervals.

By this method, four of the thirteen equations in the Chen et al. system were determined to be linear. Three of these were successfully identified as linear equations. Among these three equations one equation, however, contained a constant that was interpreted as a linear term. The fourth equation was in fact nonlinear, but was identified as linear since the term responsible for making the equation nonlinear was easily approximated as a linear term. That is, the exact forms of these equations (including constants and nonlinear terms) were not always identified. To summarize: the linear equations in the Chen et al. system could be successfully identified, but not the nonlinear ones. However, it is clear from this work that one should be cautious when applying linear models to nonlinear systems.

Note
This article is based on the master thesis “Identification of a Genetic Network in the Budding Yeast Cell Cycle”, [8], which offers a more extensive analysis of the problem and the methods.

References
Identification of a Nonlinear Cell Cycle System with Linear Models

A mathematical model by Chen et al. representing the cell cycle of the budding yeast cell (Saccharomyces cerevisiae) is used to simulate levels of 13 different cell biological variables. By using a linear systems approach on the generated data the goal is to identify the original equations of the mathematical model. Since this model is described by a nonlinear system the linear systems approach only gives correct estimations of the linear parts of the system, with one exception. A main issue thus becomes to identify which data that belongs to the linear equations and which does not. This problem is solved using two local linear models on different time intervals. Parameter estimations of coefficients from linear and time invariant equations should be the same over all time intervals.