Parameter Identifiability and Estimation in Gene and Protein Interaction Networks

Rebecca K. Shelton

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Electrical Engineering

Dr. William T. Baumann
Dr. Daniel J. Stilwell
Dr. Jean Peccoud

April 30, 2008
Blacksburg, VA

Keywords: Identifiability, Parameter Estimation, Biological Modeling

Parameter Identifiability and Estimation in Gene and Protein Interaction Networks

Rebecca K. Shelton

ABSTRACT

The collection of biological data has been limited by instrumentation, the complexity of the systems themselves, and even the ability of graduate students to stay awake and record the data. However, increasing measurement capabilities and decreasing costs may soon enable the collection of reasonably sampled time course data characterizing biological systems, though in general only a subset of the system's species would be measured. This increase in data volume requires a corresponding increase in the use and interpretation of such data, specifically in the development of system identification techniques to identify parameter sets in proposed models.

In this paper, we present the results of identifiability analysis on a small test system, including the identifiability of parameters with respect to different measurements (proteins and mRNA), and propose a working definition for "biologically meaningful estimation". We also analyze the correlations between parameters, and use this analysis to consider effective approaches to determining parameters with biological meaning. In addition, we look at other methods for determining relationships between parameters and their possible significance. Finally, we present potential biologically meaningful parameter groupings from the test system and present the results of our attempt to estimate the value of select groupings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

The collection of biological data has been limited by instrumentation, the complexity of the systems themselves, and even the ability of graduate students to stay awake and record the data. However, increasing measurement capabilities and decreasing costs may soon enable the collection of reasonably sampled time course data characterizing biological systems, though in general only a subset of the system's species would be measured. This increase in data volume requires a corresponding increase in the use and interpretation of such data, specifically in the development of system identification techniques to identify parameter sets in proposed models.

Even with such an increase in data collection, most realistic biological models will have many more parameters than can be identified. The use of an identifiability analysis method to determine a set of identifiable and a set of unidentifiable parameters is straightforward, but these sets are not unique and the problem then becomes how to identify biologically meaningful quantities. A deeper understanding of why certain parameters are considered identifiable and others not will aid in the development of a parameter identifiability method that will lead to biologically meaningful results.

In a simple case, it may turn out that while degradation and synthesis parameters or binding and unbinding parameters cannot be uniquely determined, it is possible to uniquely determine the associated equilibrium constant. For example, from the chemical equations

$$D_2 + P_{ED} \rightarrow D_2P_{ED} \qquad (1.1)$$

and

$$D_2P_{ED} \rightarrow D_2 + P_{ED} \tag{1.2}$$

where (1.1) has a rate of k_{PED} and (1.2) has a rate of k_{UPED}, we get the differential equation

$$\frac{dP_{ED}}{dt} = -k_{P_{ED}} \times P_{ED} \times D_2 + k_{UP_{ED}} \times D_2P_{ED} \tag{1.3}.$$

So, the amount P_{ED} depends on the binding parameter k_{PED} and the unbinding parameter k_{UPED}, but in the steady state equation,

$$P_{ED_{SS}} = \frac{k_{UP_{ED}}}{k_{P_{ED}}} \times \frac{D_2P_{ED}}{D_2} \tag{1.4},$$

we see that here, these parameters are actually in a ratio, k_{UPED} /k_{PED}, and cannot be separated in the steady state. In many cases, however, the parameters of even simple systems are coupled in surprising and complex ways. Setting reasonable values for a set of unidentifiable parameters and then identifying the remainder from the data will not necessarily yield parameters that are meaningful biologically. It is necessary, then, to determine the best approach to account for unidentifiable parameters and estimate the identifiable parameters while producing biologically meaningful results.

Literature Review

The importance of identifiability analyses has been discussed in many papers as an important step in parameter estimation. In [1], Gadkar, et al. propose the first iterative approach to model identification in systems biology. Notably, identifiability is an important part of that process. Parameter identifiability, first, ensures that an estimation is well-posed, and second, informs experimental design in order to obtain more accurate

parameter estimates.

The identifiability analysis used in [1] and also in this paper was proposed by Yao, et al. In [2], they proposed a procedure to determine how precisely parameters can be estimated. According to Yao, "Whether or not parameters can be estimated successfully depends upon the model structure, the parameterization of the model, and the experimental design used to gather the data" (567). Their new approach, an orthogonalization procedure described in detail later in this paper, was designed to be useful for systems with a large number of parameters, when manual inspection of the network becomes unreasonable, something of great concern in large biological systems.

Sufficient conditions for local identifiability are laid out in [3], and a differential algebra based method for testing global identifiability was set out by Audoly, et al. in [4].

Many papers have proposed estimations schemes for parameter identification. [5] compiles a detailed review of many algorithms used in parameter estimation for biochemical systems. A Numerical Matrices Method is developed in [6], in which different levels of prior information about the system are used, resulting in increased accuracy of estimation with increasing knowledge. Global optimization methods for parameter estimation are discussed in [7]. Their stated goal in parameter estimation is to "calibrate the model so as to reproduce the experimental results in the best possible way" (2467). Identifiability and biologically meaningful estimation are not considerations in these estimation algorithm papers.

In [8], Rodriguez-Fernandez et al. considered issues not handled by [7], reduced the computation time required in [7], and added a consideration of identifiability. An identifiability analysis step was added to the estimation procedure. They found that some

correlated parameters presented an estimation problem in that many values may give good results for the model. However, they propose no solution for this.

Rather than estimating all parameters at once, Koh, et al. suggested a pathway decomposition approach in [9] where parameters in smaller components of the network are estimated independently. However, the identifiability of parameters within these components is not considered.

Zak, et al. also combine identifiability and parameter estimation in [10], as discussed further in this paper; however, again, the unidentifiable parameters are simply set to reasonable values based on the literature with little or no explanation as to their connection to the identifiable parameters. The identifiability analysis used is based on [11].

Contribution

In many papers, identifiability analysis is treated as an important step in the process of parameter estimation, although in many papers outlining estimation techniques, for example [7], identifiability is not considered. When it is considered and an identifiable set has been selected, the unidentifiable parameters are then assigned reasonable values and the identifiable parameters are estimated, with the exception of the equilibrium assumption in certain cases for binding/unbinding parameters.

Producing a model that replicates experimental data is a first step. However, there will likely be many combinations of parameter values that successfully replicate that data. Therefore, in order to say anything more meaningful about the model, it is important to note what must be true about the model in order to replicate the experiment.

Our main contribution is a discussion of this problem and a presentation of possible methods for discovering such truths about our model and their biological significance.

In this paper, we present the results of identifiability analysis on a small test system, including the identifiability of parameters with respect to different measurements (proteins and mRNA), and propose a working definition for "biologically meaningful estimation". We also analyze the correlations between parameters, and use this analysis to consider effective approaches to determining parameters with biological meaning. In addition, we look at other methods for determining relationships between parameters and their possible significance. Finally, we present potential biologically meaningful parameter groupings from the test system and present the results of our attempt to estimate the value of select groupings.

Chapter 2

Model Description

In a 2003 paper, [10], Zak, *et al*. constructed an *in silico* genetic regulatory network consisting of 10 genes, 4 motifs (mutual repression, auto-activation, sequestration, and agonist-induced receptor down-regulation), and 115 parameters, shown in Fig. 2.1, where promoters are enclosed in boxes and three dots represents degradation. In the paper, Zak investigated the identifiability of parameters in response to different input perturbations.

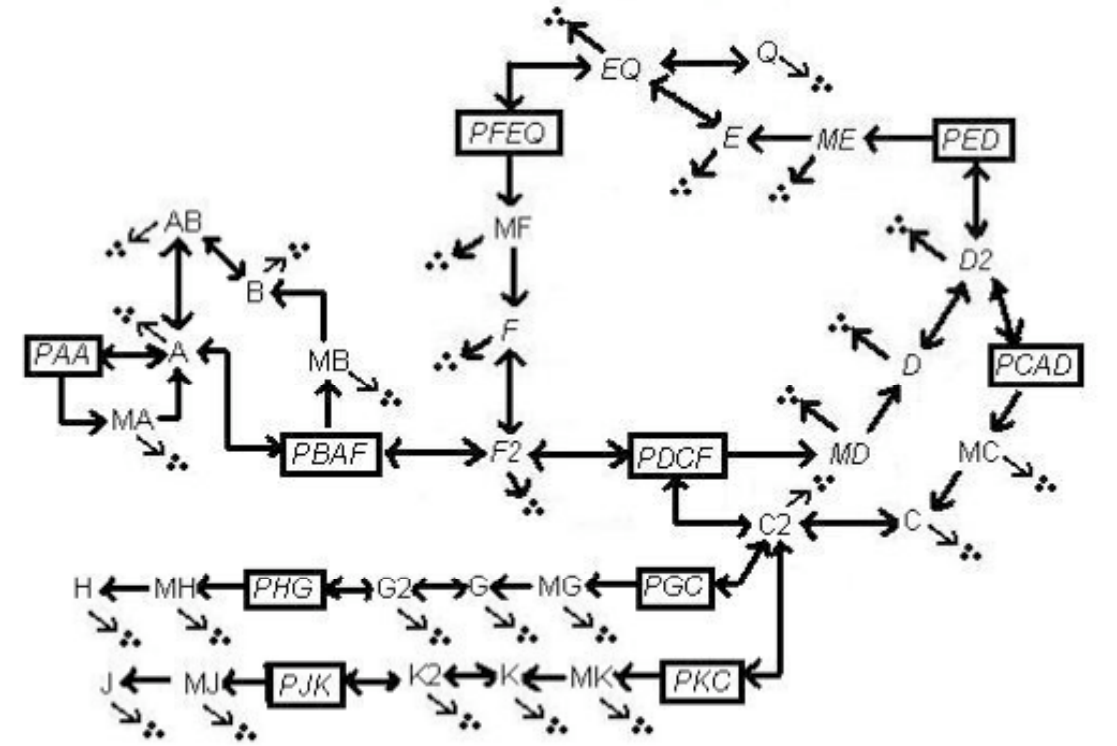

Fig. 2.1. Zak Model.

Zak found that with a step input, over half of the parameters were not identifiable. With a one-hour pulse and two one-hour pulse inputs, about a third of the parameters could not be identified. By assuming that mRNA degradations rates could be found through other experiments and by making the equilibrium assumption as in [12], Zak found that 1/4 and 1/9 of the parameters were unidentifiable with the step and pulse inputs respectively.

We have taken a subset of this network, also used by Zak in [13] and shown in Fig. 2.2, consisting of 3 genes and 31 parameters, which preserves interesting characteristics of the network, for use as a test system. The SBML file, provided by Zak in supplemental materials, was converted to a Matlab SimBiology file and used in the analysis.

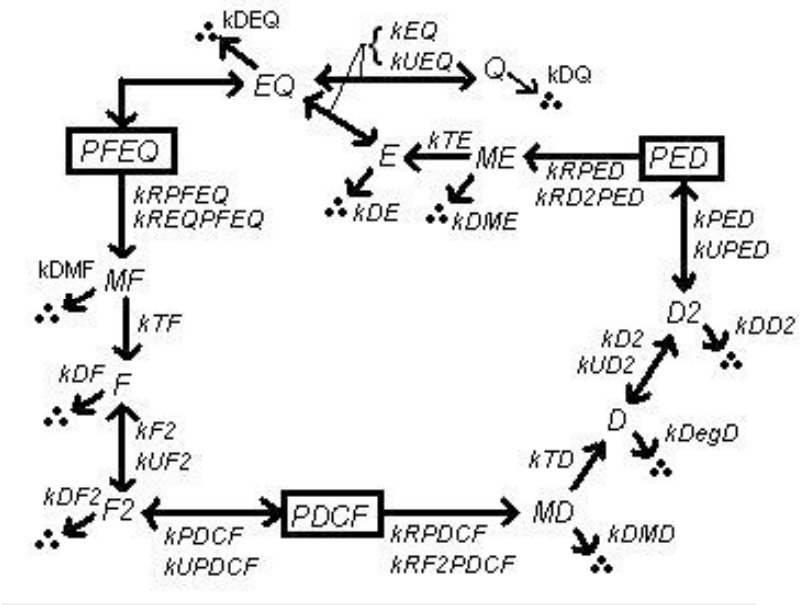

Fig. 2.2. Zak Sub-Model.

The input to the system, seen above, is the ligand, *Q*, and the outputs are mRNA, *MF*, *MD*, and *ME* or proteins *F*, *D*, and *E*. The parameters for the system are shown

where k_{RPFEQ} and $k_{REQPFEQ}$ are transcriptional rates, or the rates at which mRNA is produced, for the promoter P_{FEQ} with and without the transcriptional regulator, EQ. k_{TF} is the translation rate, or production rate, for F, and k_{F2} and k_{UF2} are the dimerization and undimerization rates, the rate at which F binds to F to form the dimer, F_2 and the rate at which F_2 breaks apart into its original components. k_{PDCF} and k_{UPDCF} are the binding and unbinding rates for the F_2 dimer, and the degradation rates are represented by k_{DMF}, k_{DF}, and k_{DF2}. The parameters related to promoters P_{DCF} and P_{ED} are similarly expressed.

In the network, without the ligand Q, we see that the protein E would have no effect on the promoter P_{FEQ}. Therefore, P_{FEQ} would always be produced at its basal rate, k_{RPFEQ}. The promoter P_{DCF} is enhanced by the protein dimer F_2, so when F_2 is not bound, D is produced at its basal rate of, k_{RPDCF} and when F_2 is bound to the promoter, D is produced at the rate $k_{RF2PDCF}$. When there is no input to the system, since F is low, being produced at its basal rate only, we do not expect much F_2 to be bound, and we expect D to be low. The promoter P_{ED}, however, is repressed by the protein dimer D_2, so when D is low, we expect E to be high, being produced mostly at its basal rate of k_{RPED}, though when D is bound, it will be produced at the lower rate of k_{RD2PED}.

When the input ligand Q is added to the system, the protein E binds with Q and the heterodimer EQ enhances the promoter P_{FEQ}, which increases the production of F to the enhanced rate, $k_{REQPFEQ}$, and in turn in enhances P_{DCF} and the production of D, but this represses the production of E when D_2 binds to the P_{ED} promoter.

On a smaller scale, the promoter P_{FEQ} produces the mRNA MF at either the basal rate, k_{RPFEQ}, or the enhanced rate, $k_{REQPFEQ}$. MF degrades, or disappears, at the rate k_{DMF} and produces F at the rate k_{TF}. F degrades at the rate k_{DF}, and dimerizes so that the

homodimer F_2, which degrades at the rate k_{DF2}, is formed at the rate k_{F2}. F_2 also undimerizes at the rate k_{UF2}. Finally, the dimer F_2 binds and unbinds to the promoter P_{DCF} at the rates of k_{PDCF} and k_{UPDCF} respectively. The productions for the other two parameters are similarly explicated.

The subsystem is characterized by 13 ordinary differential equations. The equations are written as

$$\frac{dMD}{dt} = -k_{DMD} \times MD + k_{RP_{DCF}} \times P_{DCF} + k_{RF_2P_{DCF}} \times F_2P_{DCF} \tag{2.1}$$

$$\frac{dME}{dt} = -k_{DME} \times ME + k_{RP_{ED}} \times P_{ED} + k_{RD_2P_{ED}} \times D_2P_{ED} \tag{2.2}$$

$$\frac{dMF}{dt} = -k_{DMF} \times MF + k_{RP_{FEQ}} \times P_{FEQ} + k_{REQP_{FEQ}} \times EQP_{FEQ} \tag{2.3}$$

$$\frac{dD}{dt} = -k_{DegD} \times D + k_{TD} \times MD - k_{D_2} \times D^2 + k_{UD_2} \times D_2 \tag{2.4}$$

$$\frac{dE}{dt} = -k_{DE} \times E + k_{TE} \times ME - k_{EQ} \times E \times Q + k_{UEQ} \times EQ \tag{2.5}$$

$$\frac{dF}{dt} = -k_{DF} \times F + k_{TF} \times MF - k_{F_2} \times F^2 + k_{UF_2} \times F_2 \tag{2.6}$$

$$\frac{dD_2}{dt} = -k_{DD_2} \times D_2 + k_{D_2} \times D^2 - k_{UD_2} \times D_2 - k_{P_{ED}} \times P_{ED} \times D_2 + k_{UP_{ED}} \times D_2P_{ED} \tag{2.7}$$

$$\frac{dEQ}{dt} = -k_{DEQ} \times EQ + k_{EQ} \times E \times Q - k_{UEQ} \times EQ - k_{P_{FEQ}} \times P_{FEQ} \times EQ + k_{UP_{FEQ}} \times EQP_{FEQ} \tag{2.8}$$

$$\frac{dF_2}{dt} = -k_{DF_2} \times F_2 + k_{F_2} \times F^2 - k_{UF_2} \times F_2 - k_{P_{DCF}} \times P_{DCF} \times F_2 + k_{UP_{DCF}} \times F_2P_{DCF} \tag{2.9}$$

$$\frac{dP_{DCF}}{dt} = -k_{P_{DCF}} \times P_{DCF} \times F_2 + k_{UP_{DCF}} \times F_2P_{DCF} \tag{2.10}$$

$$\frac{dP_{ED}}{dt} = -k_{P_{ED}} \times P_{ED} \times D_2 + k_{UP_{ED}} \times D_2P_{ED} \tag{2.11}$$

$$\frac{dP_{FEQ}}{dt} = -k_{P_{FEQ}} \times P_{FEQ} \times EQ + k_{UP_{FEQ}} \times EQP_{FEQ} \tag{2.12}$$

$$\frac{dQ}{dt} = -k_{DQ} \times Q + C_Q \times S(t) - k_{P_{FEQ}} \times P_{FEQ} \times EQ + k_{UP_{FEQ}} \times EQP_{FEQ} \tag{2.12}$$

where S(t) represents the time dependence of the input *Q*.

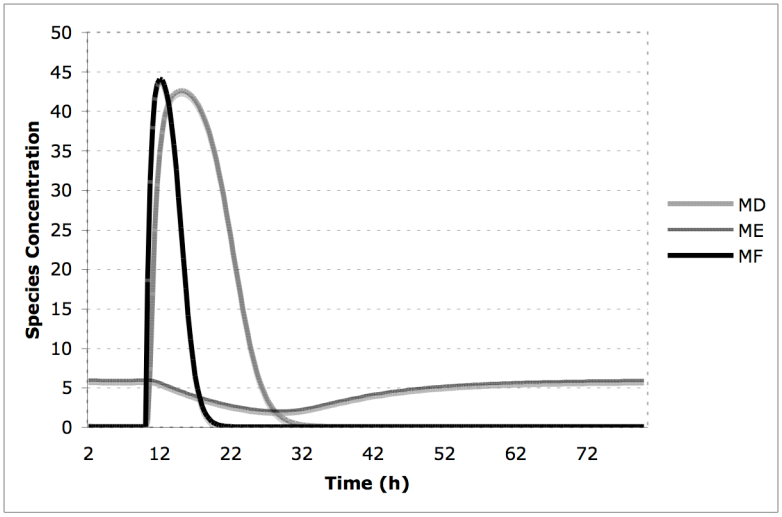

Fig. 2.3. Sub-Model Output

The outputs of the system are the mRNA and the input is a one -hour pulse of the ligand. Fig. 2.3 shows the results of the one-hour pulse input, used in the analysis. As can be seen by looking at Fig. 2.1, these results are as expected. The introduction of the ligand, *Q*, allows it to combine with *E* to form *EQ*. This enhances *F*, causing the rise seen in figure above. *F* then enhances the production of *D*, which in turn represses *E*. Hence, we expect the increase in *F* and *D* and a decrease in the gene *E*, as shown.

In Zak's identifiability analysis, only the mRNA, *MD*, *ME*, and *MF*, were measured. In the sub-network shown in Fig. 2.2, a close look will reveal that measuring only the mRNA may not yield much information about the individual parameters. An

increase in k_{F2}, for example, would be expected to have a similar effect on the production of *MD* as a decrease in k_{UF2}.

Chapter 3

Identifiability Analysis

In our study, we followed the identifiability analysis of Yao, et al. proposed in [2]. The identifiability of a parameter depends on both its effect on the output of the system and its correlation to other parameters. A sensitivity matrix, S, is constructed in the form of

$$S = \begin{bmatrix} \left.\frac{\partial MD}{\partial p_1}\right|_{t=t_1} & \cdots & \left.\frac{\partial MD}{\partial p_{31}}\right|_{t=t_1} \\ \left.\frac{\partial ME}{\partial p_1}\right|_{t=t_1} & \cdots & \vdots \\ \left.\frac{\partial MF}{\partial p_1}\right|_{t=t_1} & \cdots & \left.\frac{\partial MF}{\partial p_{31}}\right|_{t=t_1} \\ \left.\frac{\partial MD}{\partial p_1}\right|_{t=t_2} & \cdots & \left.\frac{\partial MD}{\partial p_{31}}\right|_{t=t_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial MF}{\partial p_1}\right|_{t=t_N} & \cdots & \left.\frac{\partial MF}{\partial p_{31}}\right|_{t=t_N} \end{bmatrix} \quad (3.1),$$

where p represents the parameters and N is the number of timepoints, and a normalized form Z is created from S in the form of

$$Z = \begin{bmatrix} \left.\frac{\hat{p}_1}{\hat{MD}}\right|_{t_1} \left.\frac{\partial MD}{\partial p_1}\right|_{t=t_1} & \cdots & \left.\frac{\hat{p}_{31}}{\hat{MD}}\right|_{t_1} \left.\frac{\partial MD}{\partial p_{31}}\right|_{t=t_1} \\ \left.\frac{\hat{p}_1}{\hat{ME}}\right|_{t_1} \left.\frac{\partial ME}{\partial p_1}\right|_{t=t_1} & \cdots & \vdots \\ \left.\frac{\hat{p}_1}{\hat{MF}}\right|_{t_1} \left.\frac{\partial MF}{\partial p_1}\right|_{t=t_1} & \cdots & \left.\frac{\hat{p}_{31}}{\hat{MF}}\right|_{t_1} \left.\frac{\partial MF}{\partial p_{31}}\right|_{t=t_1} \\ \left.\frac{\hat{p}_1}{\hat{MD}}\right|_{t_2} \left.\frac{\partial MD}{\partial p_1}\right|_{t=t_2} & \cdots & \left.\frac{\hat{p}_{31}}{\hat{MD}}\right|_{t_2} \left.\frac{\partial MD}{\partial p_{31}}\right|_{t=t_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\hat{p}_1}{\hat{MF}}\right|_{t_N} \left.\frac{\partial MF}{\partial p_1}\right|_{t=t_N} & \cdots & \left.\frac{\hat{p}_{31}}{\hat{MF}}\right|_{t_N} \left.\frac{\partial MF}{\partial p_{31}}\right|_{t=t_N} \end{bmatrix} \quad (3.2),$$

where $\hat{p}$ and $\hat{MD}$ are the current parameter guess and the measured value of MD at t_n respectively. Such a normalization would be inappropriate if the parameters are known to be very close to zero since, in this case, when the parameters are, in the normalized form, multiplied by the corresponding column, this will effectively zero it out or dramatically lower its magnitude.

The influence of the parameters is ranked according to the magnitude, or sum of squares, of the columns of the Z matrix. The parameter corresponding to the column with the highest magnitude is considered the most identifiable. The matrix is then orthogonalized with respect to this column, as shown in Table 3.1, where the process developed by Yao, et al. is shown. This not only zeroes the column corresponding to the

Table 3.1. Identifiability Instructions [2]

1	Calculate the magnitude (sum of squares) for each column of Z.
2	Choose the parameter corresponding to the column with the largest magnitude as the most identifiable parameter.
3	Let the column with the largest magnitude be X_i, where i=1 for the first iteration.
4	Calculate $\hat{Z}_i = X_i(X_i^T X_i)^{-1} X_i^T Z$.
5	Calculate $R_i = Z - \hat{Z}_i$.
6	Calculate the magnitude (sum of squares) for each column of R_i. Choose the parameter corresponding to the column with the largest magnitude as the next most identifiable parameter.
7	Let the corresponding column be X_{new}, and $X_{i+1} = [X_i\ X_{new}]$.
8	Let i = i+1. Repeat steps 4 through t until the cut-off value is reached.

most identifiable parameter, it also adjusts the remaining columns according to the correlation between that parameter and others [2]. The process continues for the next

most identifiable parameter until a pre-defined cut-off is reached. The cut-off used by Yao, et. al. was 0.04, however different cut-off points would be appropriate for different systems. For example, a system with higher noise would warrant a higher cut-off. It should be noted that the order in which the columns are orthogonalized directly affects the identifiable set, and that set is not unique.

Correlation coefficients were calculated for each of the parameters. The correlations between parameters clarifies some parameter relationships and reasons for their status as identifiable or unidentifiable. For example, correlation coefficients indicate strong relationships between binding and unbinding parameters, as expected. However, strong correlations exist between less obviously related parameters as well.

Since one of the main factors determining a parameter's identifiability is its effect on the output of a system, it is intuitive that measuring different species would affect the set of identifiable parameters. When measuring mRNA, 15 out of 31 of the parameters were identifiable according to the method we used, described above. Measuring proteins resulted in 14 identifiable parameters.

The two sets had 9 parameters in common, as shown in Table 3.2. However, when measuring proteins, all 3 translational parameters and all 3 protein degradation rates were identifiable, whereas the mRNA degradation parameters were considered unidentifiable. In contrast, the 3 mRNA degradation parameters were in the identifiable set, and only 1 translational parameter was identifiable when measuring mRNA only.

Table 3.2. Identifiable Parameters for Zak Sub-Model

Measuring mRNA only	Measuring mRNA and Proteins
k_{EQ}	k_{DF}
k_{RPFEQ}	k_{TE}
k_{DF2}	k_{DF2}
k_{DME}	k_{TD}
k_{DMD}	k_{DD2}
k_{DF}	k_{TF}
k_{RPED}	k_{DegD}
$k_{REQPFEQ}$	k_{RPED}
k_{TF}	k_{F2}
$k_{RF2PDCF}$	$k_{REQPFEQ}$
k_{DD2}	k_{UPDCF}
k_{DQ}	k_{DE}
k_{UPDCF}	k_{DQ}
k_{DMF}	$k_{RF2PDCF}$
k_{RPDCF}	

Measuring mRNA, some of the unidentifiable parameters were closely linked to an identifiable parameter. For example the steady state equation for P_{DCF} is

$$P_{DCF_{SS}} = \frac{k_{UP_{DCF}}}{k_{P_{DCF}}} \times \frac{F_2 P_{DCF}}{F_2} \tag{3.3}$$

According to our analysis, k_{UPDCF} was considered identifiable. These two parameters were found to have a correlation of -0.99. So, k_{PDCF} was not identifiable, as seen from Table 3.2, because of its correlation with k_{UPDCF}. In this case, it would make more sense to report the ratio of these two parameters as identifiable.

Other unidentifiable parameters were linked in less obvious ways. For example, neither k_{UPFEQ} nor k_{PFEQ} was considered identifiable. However, k_{EQ}, which was strongly correlated with those parameters was considered identifiable. Other parameters correlated to k_{EQ} which were unidentifiable included both k_{UEQ} and k_{DEQ}.

Knowing which identifiable parameters may have decreased the magnitude of the Z matrix column corresponding to unidentifiable parameters is a first step. However, we would like to see how they fit together into one identifiable group that accounts for their combined effect on the system.

The steady state equation for EQ is

$$EQ_{SS} = \frac{k_{EQ} \times E \times Q + k_{UP_{FEQ}} \times EQP_{FEQ}}{k_{UEQ} + k_{DEQ} + k_{P_{FEQ}} \times P_{FEQ}} \tag{3.4}.$$

Substituting this into the equation for P_{FEQ},

$$P_{FEQ_{SS}} = \frac{k_{UP_{FEQ}}}{k_{P_{FEQ}}} \times \frac{EQP_{FEQ}}{EQ} \tag{3.5},$$

and using the constraint, used by Zak, that

$$EQP_{FEQ} + P_{FEQ} = 2 \tag{3.6},$$

we get that

$$P_{FEQ_{SS}} = \frac{2}{\frac{k_{P_{FEQ}} k_{EQ}}{k_{UP_{FEQ}} (k_{UEQ} + k_{DEQ})} \times E \times Q + 1} \tag{3.7}.$$

So we see that these parameters are not connected as obviously or simply as the straightforward ratio in (3.3).

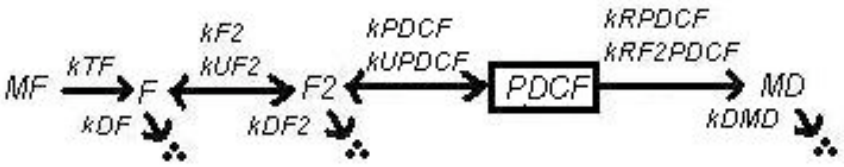

Fig. 3.1. *MF* to *MD* Model Portion

In order to look closely at these relationships, we further reduced the model to look at the portion shown in Fig. 3.1, where *MF* is used as the input. In this portion of the model, the possible ratios seen in the steady state equations are shown in the equations

$$P_{DCF_{SS}} = \frac{k_{UP_{DCF}}}{k_{P_{DCF}}} \times \frac{F_2 P_{DCF}}{F_2} \tag{3.8},$$

and

$$P_{DCF_{SS}} = \frac{2}{\frac{k_{P_{DCF}} k_{F_2}}{k_{UP_{DCF}} (k_{UF_2} + k_{DF_2})} \times F^2 + 1} \tag{3.9}.$$

In addition, identifiability analysis of this section revealed, as shown in Table 3.3, that 5

Table 3.3. Identifiable Parameters for *MF* to *MD* Model Portion

Measuring MD	Measuring MD and F	Measuring MD and FTotal
k_{dF2}	k_{dF}	k_{dF}
k_{dMD}	k_{dF2}	k_{TF}
k_{dF}	k_{dMD}	k_{dF2}
k_{TF}	k_{TF}	k_{dMD}
$k_{rF2PDCF}$	$k_{rF2PDCF}$	$k_{rF2PDCF}$
	k_{F2}	k_{rPDCF}
	k_{rPDCF}	k_{PDCF}
	k_{PDCF}	k_{F2}

parameters were identifiable when only *MD* was measured, and 8 parameters were identifiable when *MD* and *F* or F_{Total} were measured.

We now have determined a set of identifiable and a set of unidentifiable parameters, and we have proposed possible connections between parameters that would cause some to be considered unidentifiable by "zeroing out" the corresponding column in our identifiability analysis process. However, it is not enough to simply set the unidentifiable parameters to reasonable values and estimate the identifiable parameters. We instead want to have some understanding of the parameter relationships to propose "biologically meaningful estimates," which would need to be true in order for the model to successfully reproduce the experimental data. In the next section, we propose a working definition of "biologically meaningful estimation," and present methods for determining parameter groups that would be reported in lieu of simply reporting the estimated values for the individual identifiable parameters.

Chapter 4

Biologically Meaningful Estimations

Applying an identifiability analysis and choosing an identifiable set is simple. The next step is to estimate the parameters, generating a model that matches experimental data. However, the ultimate goal is for this model to have biological significance, and simply setting values for the unidentifiable parameters and then estimating the rest does not necessarily achieve this goal.

[10] deals with some of the unidentifiable parameters by making the equilibrium assumption in cases of binding and unbinding rates [12], and takes mRNA degradation rates from the literature. However, this still leaves about 1/3 of the parameters unidentifiable and unexplained. A more thorough understanding of the role and relationship of all parameters is desirable.

Before going any further, we must establish a working definition for "biologically meaningful." We propose multiple "levels" of biologically meaningful estimations in parameter estimations. The first level would be an actual rate that could be identified directly. The second would be a simple ratio such as k_{PDCF}/ k_{UPDCF}. This would provide a value for a direct relationship between two parameters. A third level would include parameter groupings like the one in (3.9), which are larger than simple ratios, but still represent a constant relationship between parameters in the system. These are not as precise within the model as a simple ratio, since it allows more individual parameters to vary from their true values while still replicating the data. A fourth level would be an

even larger grouping that perhaps might not be so easily seen from the equations, such as a general grouping of all parameters from *MF* to *MD*. The difference between this level and the third is that in the third level the grouping of parameters can be derived from the network or the equations, and the biological significance of the group can be adequately explained. In the fourth level, the grouping is not explicit, and the biological meaning is not immediately clear. The fifth and final level of meaning would be simply a model reproducing the data with no other information about the requirements for that model.

As stated above, the identifiable parameter set is not unique. In Section 3 we showed that while the unbinding rate for the F_2 dimer to the promoter P_{DCF} was considered identifiable, the binding rate was not. However, if the binding rate were considered identifiable, then the unbinding rate would be in the unidentifiable set. In reality, neither of these parameters is strictly "identifiable" as a first level estimation. Instead, in these cases, it is more accurate to say that only the ratio of these parameters is identifiable.

Taking this a step further, we may have groups of parameters that can be identified but not separated. This is significant because it allows us to account for the effects of many more of the model parameters, although those parameters are grouped together rather than taking on specific values of their own.

In dealing with unidentifiable parameters, one option, as in [10], is to set reasonable values for most of the unidentifiable parameters and then "identify" the remaining values using an appropriate algorithm. However, it must be considered that if the chosen value for an unidentifiable parameter that is strongly correlated with an identifiable parameter is incorrect, the estimation of the identifiable parameter will be

strongly affected as well. Hence, by reporting the ratio of the parameters we are able to not only explain the role of more parameters, we also avoid introducing error into the model through our assumed values for unidentifiable parameters. Identifying ratios and other parameter groupings instead gives us a more precise view of our model. It allows us to see what must really be true about the model in order to reproduce the data.

Looking closely at the *MF* to *MD* model portion in Fig. 3.1, we expect a positive correlation between the translation, transcription, promoter binding, and dimerization rates (k_{TF}, k_{RPDCF}, $k_{RF2PDCF}$, k_{PDCF}, k_{F2}) and for those to be negatively correlated with degradation (protein, mRNA, dimer), undimerization, and promoter unbinding rates (k_{DF}, k_{DMF}, k_{DF2}, k_{UF2}, k_{DMD}, k_{UPDCF}). This, in fact, is true, and the correlations between the 10 parameters in the *MF* to *MD* portion of the network when *MD* and *Ftotal* are measured are shown in Table 4.1. We see not only that the correlations between the binding an unbinding (k_{PDCF}, k_{UPDCF}) and dimerization and undimerization (k_{F2}, k_{UF2}) parameters exist but also strong correlations between all four of those parameters are shown as well . However, the correlation with k_{DF2} is weaker than may be expected from the parameter grouping.

These correlations suggest that the identifiability of one of these parameters will affect the identifiability status of the others. Hence, we should report only a ratio or ratios of these parameters. This could also lead to model reduction, reflecting the identifiability of grouped parameters.

Table 4.1. Correlation Coefficients for *MF* to *MD* Model Portion

	k_{dF}	k_{TF}	k_{F2}	k_{UF2}	k_{PDCF}	k_{UPDCF}	k_{dF2}	k_{dMD}	k_{rPDCF}	$k_{rF2PDCF}$
k_{dF}	1.00	-0.91	-0.69	0.68	-0.69	0.69	0.74	0.63	-0.49	-0.57
k_{TF}	-0.91	1.00	0.67	-0.67	0.67	-0.67	-0.59	-0.64	0.57	0.58
k_{F2}	-0.69	0.67	1.00	-1.00	0.96	-0.96	-0.42	-0.93	0.9	0.79
k_{UF2}	0.68	-0.67	-1.00	1.00	-0.96	0.96	0.42	0.93	-0.9	-0.79
k_{PDCF}	-0.69	0.67	0.96	-0.96	1.00	-1.00	-0.54	-0.95	0.78	0.89
k_{UPDCF}	0.69	-0.67	-0.96	0.96	-1.00	1.00	0.54	0.96	-0.78	-0.89
k_{dF2}	0.74	-0.59	-0.42	0.42	-0.54	0.54	1.00	0.43	-0.1	-0.48
k_{dMD}	0.63	-0.64	-0.93	0.93	-0.95	0.96	0.43	1.00	-0.83	-0.94
k_{rPDCF}	-0.49	0.57	0.9	-0.9	0.78	-0.78	-0.1	-0.83	1.00	0.62
$k_{rF2PDCF}$	-0.57	0.58	0.79	-0.79	0.89	-0.89	-0.48	-0.94	0.62	1.00

Since our sub-model, in Fig. 2.2, is relatively small, we can look at the model and see exactly what some of these parameters are doing, though in a larger model this becomes more difficult. The grouping of parameters in (3.7), for example, can be looked at closely for this type of biological "meaning." The high correlation between k_{PFEQ} and k_{UPFEQ} is expected due to a steady state ratio like the one shown in (3.5). Looking at the steady state equation for *EQ*, combined with (3.6), we expect the correlations between k_{EQ}, k_{UEQ}, and k_{DEQ}. Putting all these together as the parameter group in (3.7), we see that the ultimate effect of these parameters is that they control the probability that the transcription factor, *EQ*, binds to the *F* promoter, P_{FEQ}.

From a biological perspective, this means that this parameter group controls how

the concentrations of *E* and *Q* map to the expected number of genes bound by the transcription factor. In turn, this affects the production rate of *MF* along with k_{RPFEQ} and $k_{REQPFEQ}$. Since *MF* is an output of the system, this rate can be experimentally verified. Similar relationships exist for the other two promoters and their corresponding parameters. These types of groupings allow us to account for a total of 24 of the 31 parameters.

We are specifically concerned with the effect our choices for the values of unidentifiable parameters has on the estimation of identifiable parameters. Therefore, we propose a new way of looking at the relationships between these parameters, through a matrix illuminating the relationship between the identified parameter values and the values chosen for the unidentifiable parameters. The normalized dP_I/dP_F matrix, $\frac{P_F}{P_I}\frac{\partial P_I}{\partial P_F}$, where P_I includes the identifiable parameters and P_F includes the unidentifiable parameters that can be held at fixed values for the estimation, can give some insights into the relationship of the parameters to each other. Each term in the matrix shows the percent change in the identifiable parameters given a percent change in the unidentifiable parameters.

In addition to looking at the change in the identifiable parameters in relation to changes in the unidentifiable parameters, we can look at the changes in parameter groups as well. We will call this the normalized dG/dP_F matrix, $\frac{P_F}{G}\frac{\partial G}{\partial P_F}$, where *G* is a parameter grouping. The change in a parameter group given a fixed change in an unidentifiable parameter gives insight into how well that grouping can be expected to hold locally given errors in the unidentifiable parameters. These two matrices can be quickly computed

numerically, since it requires only a small change in the unidentifiable parameters. Using the given values of the identifiable parameters for the starting point of the estimation, then, we expect a fast convergence to show these local results.

The normalized dP_I/dP_F and dG/dP_F matrices for the *MF* to *MD* model portion, shown in Fig. 3.1, measuring *MD* and F_{Total}, are calculated as shown in Table 4.2, where the corresponding parameters or parameter groupings are shown with each row and column.

Table 4.2. Normalized dP_I/dP_F and dG/dP_F for *MF* to *MD*, measuring *MD*, F_{Total}

	k_{UF2}	k_{UPDCF}
k_{DF}	0.0000	0.0000
k_{TF}	-0.0000	-0.0000
k_{F2}	1.0973	0.1090
k_{PDCF}	-0.0003	1.0627
k_{DF2}	-0.0000	-0.0000
k_{DMD}	-0.0042	-0.0034
k_{RPDCF}	0.1146	0.1391
$k_{RF2PDCF}$	-0.0044	0.0003
k_{F2}/k_{UF2}	-0.0025	0.1090
k_{PDCF}/k_{UPDCF}	-0.0003	-0.0340
k_{TF}/k_{DF}	-0.0000	-0.0000
$k_{RF2PDCF}/k_{DMD}$	-0.0002	0.0037

In the matrix above, an entry of -0.004 for k_{DMD} and $k_{RF2PDCF}$ indicates that k_{F2} can vary over 200% before those parameters would move 1%. So we expect to be able to

identify k_{DF}, k_{DF2}, and k_{TF} exactly, a level 1 result in our biologically meaningful estimate scale, for almost any values chosen for the fixed parameters, and we expect to identify k_{DMD}, and $k_{RF2PDCF}$ very well for a very large range of values of k_{UF2} and k_{UPDCF}. However, k_{RPDCF} has a 1% change for just a 10% change in P_F, so we would not expect to get a good estimation of that parameter if our guesses are not close to the "true" values of our other parameters.

In addition to giving us a percent change in P_I given a percent change in P_F, the normalized dP_I/dP_F matrix can also directly suggestion possible relationships between parameters. For instance, parameters in a strict ratio relationship such as k_{F2}/k_{UF2} indicates that

$$\frac{P_{Ii}}{P_{Fj}} = c \qquad (4.1),$$

where c is a constant. Therefore,

$$\frac{\partial P_{Ii}}{\partial P_{Fj}} \frac{1}{P_{Fj}} - \frac{P_{Ii}}{P_{Fj}^{\ 2}} = 0 \qquad (4.2)$$

and

$$\frac{\partial P_{Ii}}{\partial P_{Fj}} = \frac{P_{Ii}}{P_{Fj}} \qquad (4.3),$$

so

$$\frac{P_{Fj}}{P_{Ii}} \frac{\partial P_{Ii}}{\partial P_{Fj}} = 1 \qquad (4.4).$$

So these parameters would correspond to a matrix entry of 1 and zeroes elsewhere. In the same way, parameters with a product relationship would correspond to a matrix entry of -1. Some of these relationships are shown in Table 4.3. These are necessary, but not sufficient conditions. The results are local, so they do not necessarily reflect the behavior

of an estimation which is either not close to the true values of the unidentifiable parameters or an estimation in which the starting values for the identifiable parameters are not close to the true values.

Table 4.3. Matrix Component Analysis for Normalized dP_I/dP_F

Relationship Between Parameters	Corresponding Matrix Entry Results
$\frac{P_{Ii}}{P_{Fj}} = c$	$\frac{P_{Fj}}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_{Fj}} = 1$ Row i, Position j = 1; Other Row i entries = 0
$\frac{P_{Ii}}{P_{Ij}} = c$	$\frac{P_F}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_F} = \frac{P_F}{P_{ji}}\frac{\partial P_{Ij}}{\partial P_F}$ Row i = Row j
$P_{Ii}P_{Fj} = c$	$\frac{P_{Fj}}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_{Fj}} = -1$ Row i, Position j = -1; Other Row i entries = 0
$P_{Ii}P_{Ij} = c$	$\frac{P_F}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_F} = -\frac{P_F}{P_{ji}}\frac{\partial P_{Ij}}{\partial P_F}$ Row i = -Row j
$P_{Ii} + P_{Fj} = c$	$\frac{P_{Fj}}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_{Fj}} = -\frac{P_{Fj}}{P_{Ii}}$ Row i, Position j = $-P_{FJ}/P_{Ii}$ Other Row i entries = 0
$P_{Ii} + P_{Ij} = c$	$\frac{P_F}{P_{Ii}}\frac{\partial P_{Ii}}{\partial P_F} = -\frac{P_F}{P_{ji}}\frac{\partial P_{Ij}}{\partial P_F}\left(\frac{P_{Ij}}{P_{Ii}}\right)$ Row i = -(Row j)*(P_{Ij}/P_{Ii})

Based on these results, while k_{UF2} and k_{UPDCF} have matrix entries close to what would be expected in a strict ratio relationship, a level 2 estimation, the ratios are not exact, as evidenced by the dependence of k_{F2} on k_{UPDCF}. Indeed, the ratio, k_{F2}/k_{UF2}, based on the normalized dG/dP_F matrix, was expected to have about a 1% change for a 10%

change in k_{UPDCF}.

Table 4.4. Normalized dP_I/dP_F and dG/dP_F for *MF* to *MD*, measuring *MD*

	k_{F2}	k_{UF2}	k_{PDCF}	k_{UPDCF}	k_{RPDCF}
k_{DF}	0.1675	-0.1547	-0.3767	0.4977	0.0187
k_{TF}	-0.0839	0.0656	-0.4823	0.2603	-0.0001
k_{DF2}	0.0144	-0.0266	0.2878	-0.4221	-0.0054
k_{DMD}	-0.0777	0.0928	-0.2545	0.3828	0.0005
$k_{RF2PDCF}$	-0.0727	0.0896	-0.2938	0.4481	0.0000
k_{TF} / k_{DF}	-0.2476	0.2235	-0.1094	-0.2271	-0.0188
$k_{RF2PDCF} / k_{DMD}$	0.0050	-0.0032	-0.0402	0.0631	-0.0004

The normalized matrices for the *MF* to *MD* model portion measuring *MD* only are shown in Table 4.4. The only obvious relationship that appears to exist is a quotient relationship between $k_{RF2PDCF}$ and k_{DMD}, and none of the parameters are expected to converge to their true values unless the fixed parameters are close to their true values. k_{RPDCF} has the least effect on values of the identifiable parameters, so we expect to be able to correctly estimate even if this parameter is fixed at a wrong value. However, k_{PDCF} and k_{UPDCF} effect a 2.5 to 5% change in the identifiable parameters for just a 10% change in their fixed values.

Table 4.5. Normalized dP_I/dP_F and dG/dP_F for *MF* to *MD*, measuring *MD*, *F*

	k_{UF2}	k_{UPDCF}
k_{DF}	-0.0100	0.0000
k_{TF}	0.0000	0.0000
k_{F2}	1.0900	0.0000
k_{PDCF}	0.0000	1.1000
k_{DF2}	0.0000	0.0000
k_{DMD}	-0.0300	-0.0300
k_{RPDCF}	-0.0600	-0.0300
$k_{RF2PDCF}$	-0.0200	-0.0200
k_{F2}/k_{UF2}	-0.0047	0.0000
k_{PDCF} / k_{UPDCF}	0.0021	-0.0010
k_{TF} / k_{DF}	0.0051	0.0001
$k_{RF2PDCF} / k_{DMD}$	0.0005	0.0007
$k_{F2}\, k_{PDCF} / k_{UPDCF}\, (k_{UF2} + k_{DF2})$	-0.0009	-0.0010

The normalized dP_I/dP_F and dG/dP_F matrices for the *MF* to *MD* portion, measuring *MD* and *F* are shown in Table 4.5. The results are similar to the matrix for the *MF* to *MD* portion, measuring *MD* and F_{Total}.

These matrices provide a means to search for possible relationships between parameters. However, since the requirements are necessary but not sufficient, they only suggest possibilities and not actualities. Therefore, the indicated relationships need to be confirmed through other experiments.

Chapter 5

Identification Results

After determining sets of identifiable and unidentifiable parameters and locating biologically meaningful parameter groupings, the next step is to estimate values for those parameter groups based on experimental data. This allows us to, first, propose values for the model and, second, observe whether the relationships between parameters suggested by the methods of Section 4 are supported in the estimation. We used local results to determine identifiability, and the normalized dP_I/dP_F and dG/dP_F matrices also reflect local behavior. Our estimations will explore the results of non-local estimations when we vary the guesses in a wider range and attempt to estimate the identifiable parameters.

We used Matlab to simulate the networks and *lsqcurvefit* to estimate the parameters. We will present the results for the *MF* to *MD* model portion, measuring *MD* only and then *MD* plus F_{Total}, and finally *MD* plus *F*. We first randomly varied the unidentifiable parameters from their true values while estimating the identifiable parameters starting at randomly selected initial values. However, the results did not match our expectations. Since we are using a simple gradient-based estimation, *lsqcurvefit*, it is possible that in these non-local estimations, the algorithm converged to a non-global minimum. In order to investigate this, we "pulled through" the parameter values by varying the unidentifiable parameters in small steps while estimating the identifiable parameters using the previous results as the starting value. In this way, the algorithm is more likely to converge to the global minimum.

Here, we first show the results of "pulling through" the parameter values and then show the results of randomly varying the unidentifiable parameters and using randomly chosen starting values for the identifiable parameters in the estimation. The latter reveals the limitations of using a simple estimation method such as *lsqcurvefit*.

MF to MD Model Portion Results – measuring MD only

As determined by our identifiability analysis, 5 parameters were considered identifiable when measuring *MD* only for the *MF* to *MD* model portion. . First, using the pull-through method, we varied the unidentifiable parameters in small steps away from their true values. In this case, when we have both parts of a possible ratio such as k_{PDCF} /k_{UPDCF} in the unidentifiable parameter set, some of the interesting ratios are no longer valid since they are set in the unidentifiable parameters.

Table 5.1. Estimation Results for *MF* to *MD*, measuring *MD,* varying k_{UF2}

k_{DF}	k_{TF}	k_{DF2}	k_{DMD}	$k_{RF2PDCF}$	$*k_{UF2}$	k_{TF} / k_{DF}	$k_{RF2PDCF} / k_{DMD}$
6.6e-3	2.50	6.6e-3	3.10e-2	0.730	3.8	378.78	23.59
6.7e-3	2.48	6.6e-3	3.08e-2	0.725	3.42	370.15	23.54
6.9e-3	2.45	6.6e-3	3.06e-2	0.721	3.04	355.07	23.56
7.0e-3	2.42	6.6e-3	3.04e-2	0.717	2.66	345.71	23.59
7.3e-3	2.39	6.6e-3	3.02e-3	0.713	2.28	327.40	23.61
7.5e-5	2.35	6.6e-3	2.99e-3	0.710	1.9	313.33	23.76

First, varying k_{UF2}, it seems that only k_{DF2} converges to its true value as k_{UF2} is varied in steps towards zero. In fact, according to the normalized dPI/dPF matrix, k_{DF2} would be expected to have the smallest change in response to changes in k_{UF2}. k_{TF} and k_{DF} were expected to vary about 1 and 2% respectively with a 10% change in k_{UF2} and in opposite directions. As shown in table 5.1, this is approximately the result we achieved. Meanwhile, we saw in the normalized dG/dPF matrix that the ratio k_{TF}/k_{DF} was expected to vary about 2% with a 10% change in k_{UF2}. In Table 5.1, k_{TF}/k_{DF} varied 2 to 5% with each 10% change in k_{UF2}. $k_{RF2PDCF}$ and k_{DMD} were expected to vary about 0.7% each; in fact, they varied less than 1% from the previous value in each estimation while their ratio, $k_{RF2PDCF}/k_{DMD}$, varied less than 0.2% in all but one step, which, while varying less than the parameters is higher than its entry in the normalized dG/dPF matrix would suggest, 0.05%.

Varying k_{UPDCF}, the individual parameter values are similarly predictable based on the normalized dPI/dPF matrix. The 5 identifiable parameters were expected to vary 3 to 5% with each 10% change, and they varied 3 to 7% in the results. The ratio $k_{RF2PDCF}/k_{DMD}$ was expected to vary 0.4%, and in fact varied about 0.5%. However, varying k_{RPDCF} did not result in good convergence. The costs were much higher than those for the other results, about 0.3 rather than 1.4e-4, which was the highest cost when varying k_{UF2}.

Next, randomly varying the 5 unidentifiable parameters by up to -50% or +100% of their true values, we estimated the 5 identifiable parameters. These results are summarized in Table 5.2.

Table 5.2. Estimation Results for *MF* to *MD*, measuring *MD*

Cost	k_{DF}	k_{TF}	k_{DF2}	k_{DMD}	$k_{RF2PDCF}$	$*k_{F2}$	$*k_{UF2}$	$*k_{PDCF}$	$*k_{UPDCF}$	$*k_{RPDCF}$
0.0	6.6e-3	2.5	6.6e-3	3.1e-2	0.73	5.9e-3	3.8	7.5e-4	0.39	7.3e-4
0.2e-3	7.4e-3	2.3	5.9e-3	3.4e-2	0.82	5.8e-3	3.3	10.0e-4	0.55	11.0e-4
0.9e-3	7.0e-3	2.4	6.3e-3	3.2e-2	0.75	7.3e-3	4.4	1.1e-3	0.61	1.2e-3
0.5e-3	7.1e-3	2.0	6.0e-3	3.6e-3	0.85	9.0e-3	5.6	5.6e-4	0.26	9.1e-4
1.1e-3	8.3e-3	3.3	5.2e-3	4.0e-3	0.97	6.0e-3	5.2	6.3e-4	0.63	6.4e-4
0.4e-3	5.6e-3	2.2	7.4e-3	2.8e-3	0.65	9.1e-3	6.9	1.5e-3	0.52	5.2e-4

Table 5.2 shows the final cost, the sum of squares of the difference between the estimated data and the data obtained from the true Zak parameters, and the resulting 5 estimated values for the 5 identifiable parameters as well as the randomly varied values for the 5 unidentifiable parameters, marked by an asterisk. The first row, with a cost of 0.0, shows the true parameter values for comparison. The starting values for the estimation were randomly selected within -50% to +100% of the true values of the identifiable parameters.

Table 5.3. Parameter Group Values for *MF* to *MD*, measuring *MD*

Cost	$k_{RR2PDCF}$ / k_{DMD}	$\boldsymbol{k_{TF}}$ / $\boldsymbol{k_{DF}}$
0.0	23.54	378.78
0.2e-3	23.94	304.05
0.9e-3	23.80	281.16
0.5e-3	24.40	396.97
1.1e-3	23.18	390.64
0.4e-3	24.41	419.65

In our analysis of the normalized dP_I/dP_F matrix, we found only one apparently significant relationships between parameters, and that was the quotient relationship between $k_{RF2PDCF}$ and k_{DMD}. In addition, we did not expect the parameters to converge well when the values for the unidentifiable parameters were not close to their true values. As seen in Table 5.2, the identifiable parameters vary by as much as 20 to 30% while the unidentifiable parameter are varied within -50% to +100% of the true values. Table 5.3 shows some values for possible parameter groupings for the same estimations shown in Table 5.2, including the quotient relationship between $k_{RF2PDCF}$ and k_{DMD}. Like Table 5.2, the row corresponding to a cost of 0.0 indicates the value with respect to the true parameter values. As we can see, the ratio $k_{RR2PDCF}$ / k_{DMD} stayed relatively constant throughout the estimations, varying by about 4% at most when the parameters themselves varied by as much as 30% of their true values. The ratio k_{TF} / k_{DF} did not hold, varying by as much as 30%, which also conforms to our expectation based on the normalized dP_I/dP_F matrix in section 4.

MF to MD Model Portion Results – measuring MD and F_{Total}

One of the methods for measuring proteins would require that the proteins be broken down. Thus, using this method, we would measure F_{Total} rather than *F* or *F2*. For this reason, we considered the case of measuring *MD* and F_{Total} for the *MF* to *MD* model portion. In this case, we found that 8 parameters were identifiable. Using the pull-through method, parameters k_{TF}, k_{DF}, k_{DF2}, k_{DMD}, and $k_{RF2PDCF}$ converged to approximately their true values, within 5%, k_{RPDCF} was also always close to its true value, within 10%, although it varied slightly more than the previous four parameters. However, k_{F2} and k_{PDCF} varied from their true values to conform to a ratio with k_{UF2} and k_{UPDCF}. As k_{UF2} and k_{UPDCF} varied up to 50% below their true values, so did k_{F2} and k_{PDCF}. However, the ratio k_{F2}/k_{UF2} varied by less than 10% in all but one case, where it varied 20% from its true value, and the ratio k_{PDCF}/k_{UPDCF} varied by less than 2%. From the normalized dP_I/dP_F and dG/dP_F matrices, we expected k_{F2}/k_{UF2} to vary more in general than k_{PDCF}/k_{UPDCF}.Table 5.4 shows these four variables and their results as well as the ratios between them.

Table 5.4. Estimation Results, *MF* to *MD*, measuring *MD* and *F*

k_{F2}	k_{PDCF}	$*k_{UF2}$	$*k_{UPDCF}$	k_{F2}/k_{UF2}	k_{PDCF}/k_{UPDCF}
5.9e-3	7.5e-4	3.8	0.39	1.55e-3	1.92e-3
5.8e-3	6.58e-4	3.42	0.351	1.70e-3	1.91e-3
5.7e-3	5.68e-4	3.04	0.312	1.89e-3	1.91e-3
4.0e-3	5.30e-4	2.66	0.273	1.50e-3	1.90e-3
3.3e-3	4.59e-4	2.28	0.234	1.45e-3	1.89e-3
2.9e-3	3.76e-4	1.9	0.195	1.53e-3	1.89e-3

Then, randomly varying the 2 unidentifiable parameters while estimating the identifiable ones, the results are shown in Table 5.5.

Table 5.5. Estimation Results, *MF* to *MD*, measuring *MD* and F_{Total}

Cost	k_{DF}	k_{TF}	k_{F2}	k_{PDCF}	k_{DF2}	k_{DMD}	k_{RPDCF}	$k_{RF2PDCF}$	$*k_{UF2}$	$*k_{UPDCF}$
0.0	6.6e-3	2.5	5.9e-3	7.5e-4	6.6e-3	3.1e-2	7.3e-4	0.73	3.8	0.39
0.2e-3	6.6e-3	2.5	11.3e-3	1.4e-3	6.6e-3	3.1e-2	7.0e-4	0.729	7.3	0.75
0.2e-3	6.6e-3	2.5	11.6e-3	1.4e-3	6.6e-3	3.09e-2	8.0e-4	0.729	7.4	0.75
0.1e-3	6.6e-3	2.5	10.8e-3	1.2e-3	6.6e-3	3.1e-2	8.0e-4	0.729	6.9	0.61
0.4e-2	6.6e-3	2.5	11.0e-3	0.8e-3	6.6e-3	3.09e-2	6.0e-4	0.732	5.4	0.47

In our analysis of the normalized dPI/dPF matrix for this case, we expected the parameters k_{DF}, k_{DF2}, k_{TF}, k_{DMD}, and $k_{RF2PDCF}$ to converge within the range over which we

varied the unidentifiable parameters. As seen in Table 5.5, those parameters converged almost exactly to their true values in these estimations. The parameter k_{RPDCF} was expected to vary more, a 1% change for a 10% change in one of the unidentifiable parameters. In the estimation results, k_{RPDCF} varied 5 to 10% in 3 cases and 20% in the other. Since the two unidentifiable parameters were varied within 1 half to 2 times the true values, the 5 to 10% variations conform to our expectations.

Table 5.6. Parameter Group Values, *MF* to *MD*, measuring *MD* and F_{Total}

Cost	$\boldsymbol{k_{F2}}$ / $\boldsymbol{k_{UF2}}$	$\boldsymbol{k_{PDCF}}$ / $\boldsymbol{k_{UPDCF}}$	$k_{RR2PDCF}$ / k_{DMD}	$\boldsymbol{k_{TF}}$ / $\boldsymbol{k_{DF}}$
0.0	0.0015	0.0019	23.54	378.78
0.2e-3	0.0015	0.0019	23.6	378.78
0.2e-3	0.0016	0.0019	23.6	378.78
0.1e-3	0.0015	0.0019	23.5	378.78
0.4e-3	0.0020	0.0017	23.7	378.78

The 2 identifiable parameters most obviously closely related to the unidentifiable parameters were expected to be in loose, but not exact, ratio relationships based on the normalized dP_I/dP_F matrix. As seen in Table 5.6, the ratios k_{F2} / k_{UF2} and k_{PDCF} / k_{UPDCF} varied from their true values by 30 and 10% respectively in one estimation and held almost constant in 3 others. The values of those 2 identifiable parameters, however, varied by as much as 90% in the estimations.

MF to MD Model Portion Results – measuring MD and F

When measuring *MD* and *F*, we again found that 8 parameters were identifiable. Using the "pull-through method," the results are similar to those measuring *MD* and F_{Total}. Parameters k_{TF}, k_{DF}, k_{DF2}, k_{DMD}, and $k_{RF2PDCF}$ converged to approximately their true values, within 5%, k_{RPDCF} was also always close to its true value, within 25%, although it varied more than the previous four parameters. Meanwhile, k_{F2} and k_{PDCF} varied from their true values to conform to a ratio with k_{UF2} and k_{UPDCF}. As k_{UF2} and k_{UPDCF} varied up to 50% below their true values, so did k_{F2} and k_{PDCF}. However, the ratios varied by less than 2%. Table 5.7 shows these four variables and their results as well as the ratios between them.

Table 5.7. Estimation Results, *MF* to *MD*, measuring *MD* and *F*

k_{F2}	k_{PDCF}	$*k_{UF2}$	$*k_{UPDCF}$	k_{F2}/k_{UF2}	k_{PDCF}/k_{UPDCF}
5.9e-3	7.5e-4	3.8	0.39	1.55e-3	1.92e-3
5.3e-3	6.72e-4	3.42	0.351	1.55e-3	1.91e-3
4.8e-3	5.95e-4	3.04	0.312	1.58e-3	1.91e-3
4.2e-3	5.18e-4	2.66	0.273	1.58e-3	1.90e-3
3.6e-3	4.42e-4	2.28	0.234	1.58e-3	1.89e-3
3.0e-3	3.68e-4	1.9	0.195	1.58e-3	1.89e-3

Randomly varying the 2 unidentifiable parameters while estimating the identifiable ones from randomly selected initial values, the results are shown in Table 5.8.

Table 5.8. Estimation Results, *MF* to *MD*, measuring *MD* and *F*

Cost	k_{DF}	k_{TF}	k_{F2}	k_{PDCF}	k_{DF2}	k_{DMD}	k_{RPDCF}	$k_{RF2PDCF}$	$^{*}k_{UF2}$	$^{*}k_{UPDCF}$
0.0	6.6e-3	2.5	5.9e-3	7.5e-4	6.6e-3	3.1e-2	7.3e-4	0.73	3.8	0.39
24.9e-3	7.7-3	3.2	7.7e-3	5.1e-4	6.5e-3	4.0e-2	2.9e-3	0.94	3.72	0.34
11.4e-3	5.6e-3	1.9	5.9e-3	1.5e-3	6.8e-3	2.8e-2	3.0e-3	0.66	5.4	0.59
10.4e-3	5.7e-3	1.9	5.9e-3	1.3e-3	6.7e-3	2.7e-2	1.8e-3	0.64	5.2	0.51
14.7e-3	5.4e-3	1.7	2.6e-3	1.3e-3	6.8e-3	2.8e-2	3.1e-3	0.66	2.5	0.46

The results in Tables 5.8 and 5.9 do not conform as well to our expectations as the previous examples. The parameters k_{TF} , k_{DF}, k_{DMD}, and $k_{RF2PDCF}$ were expected to converge to their true values based on the analysis of the normalized dP_I/dP_F matrix for this case. However, as shown in Table 5.8, those parameters vary by as much as 20 to 30%. The cost in these estimations, while seemingly low, was significantly higher than other estimations and higher than the "pull-through" cost results indicating that the problem here may be with the estimation algorithm and its limitations as a very simple Matlab tool.

Table 5.9. Parameter Group Values, *MF* to *MD*, measuring *MD* and *F*

Cost	k_{F2} / k_{UF2}	k_{PDCF} / k_{UPDCF}	$k_{RR2PDCF} / k_{DMD}$	k_{TF} / k_{DF}	$\frac{k_{F2}k_{PDCF}}{k_{UPDCF}(k_{UF2} + k_{dF2})}$
0.0	0.0015	0.0019	23.54	378.78	3.0e-6
24.9e-3	0.0021	0.0015	23.41	416.55	3.0e-6
11.4e-3	0.0011	0.0026	23.78	331.53	2.9e-6
10.4e-3	0.0011	0.0025	23.69	339.13	2.9e-6
14.7e-3	0.0010	0.0029	23.65	319.45	2.9e-6

In addition, the ratios k_{F2} / k_{UF2} and k_{PDCF} / k_{UPDCF} varied by as much as 30 to 50% and at least 20% in the first 4 estimations. By contrast, surprisingly, the large parameter grouping from

$$P_{DCF_{SS}} = \frac{2}{\frac{k_{P_{DCF}} k_{F_2}}{k_{UP_{DCF}}(k_{UF_2} + k_{DF_2})} \times F^2 + 1} \tag{5.1}$$

varies by only 3% in the estimations.

These results seem to imply that the larger parameter group may be more important than the smaller ratios, however more investigation would be required to verify this hypothesis.

Chapter 6

Conclusion

We have presented the results of an identifiability analysis, proposed a working definition for biologically meaningful estimates, proposed methods for determining potential groupings for meaningful estimations, and presented the results of parameter estimations on a test system. While these results are limited to a small, simple system, the same ideas can be expanded to apply to larger, more complex networks as well.

Reporting parameter groupings that indicate biologically meaningful estimations is a first step. However, in order for such an approach to be feasible, the process must be largely automated. It is not reasonable to manually analyze a large system to discover the correlated groupings. In a large, complex system, the parameter relationships will not be evident from the differential equations, and the matrix of correlation coefficients and the normalized dP_I/dP_F matrix will only grow in size with the number of parameters.

We showed that the normalized dP_I/dP_F matrix can be used to identify possible relationships between parameters, however, the analysis of the matrix elements in this paper is limited to the most obvious groupings. An automated system of searching through possible relationships could lead to a deeper understanding of more parameter groups and larger systems with many more parameters.

For a large system with feedback and other more complex motifs, identifying preserved parameter relationships will be difficult if not impossible. However, searching automatically through many possible relationships is comparatively trivial. A list of the common relationships should be compiled and added upon as needed. For example, a

first step would be for all possible ratio, product, and additive relationships involving 2 parameters to be searched through for potential preserved values. The list could then be expanded to 3 or more parameters in more complex groupings and so on. Only results below a pre-defined threshold would be reported.

Future work would include further investigation of the identification of parameter groups including model reduction techniques as an intermediate step to the identification. In addition, it will be necessary to expand the identification and estimation to larger systems and propose methods to deal with the increasing size of the network and growing number of parameters.

References

[1] K.G. Gadkar, R. Gunawan, and F.J. Doyle III. “Iterative approach to model identification of biological networks.” BMC Bioinformatics, vol. 6, issue 155, 2005.

[2] K. Z. Yao, B. M. Shaw, B. Kuo, K. B. McAuley, and D. W. Bacon. “Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design,” Polymer Reaction Engineering, vol. 11, pp. 563-588, 2003.

[3] M. Farina, R. Findeisen, E. Bullinger, S. Bittanti, F. Allgower, and P. Wellstead. “Results towards identifiability properties of biochemical reaction networlks.” ILEEE Conference on Decision and Control, 2006.

[4] S. Audoly, G. Gellu, L. D’Angio, M.P. Saccomanni, and C. Cobelli. “Global identifiability of nolinear models of biological systems.” IEEE Trans. Biomedical Engineering, vol. 48, 2001.

[5] G. Maria. “A review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems.” *Chem. Biochem. Eng.* Vol. 18, pp. 195-222, 2004.

[6] A.V. Karnaukhov, E.V. Karnaukhova, and J.R. Williamson. “Numerical Matrices Method for Nonlinear System Identification and Description of Dynamics of Biochemical Reaction Networks.” Biophysical Journal. Vol. 92, pp. 3459-73, 2007.

[7] C. G. Moles, P. Mendes, and J.R. Banga. “Parameter Estimation in Biochemical Pathways: a comparison of global optimization methods.” *Genome Research.* Vol. 13, pp. 2467-74, 2003.

[8] M. Rodriquez-Fernandez, P. Mendes, J.F. Banga. “A hybrid approach for efficient and robust parameter estimation in biochemical pathways.” Bio Systems Journal, vol 83, pp. 248-266, 2006.

[9] G. Koh, H.F.C. Teong, M. Clement, D. Hsu, and P.S. Thiagarajan. “A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk.” Bioinformatics. Vol. 22, pp. 271-281, 2006.

[10] D. E. Zak, G. E. Gonye, J. S. Schwaber, and F. J. Doyle, III. “Importance of Input Perturbations and Stochastic Gene Expression in the Reverse Engineering of Genetic Regulatory Networks: Insights From an Identifiability Analysis of an In Silico Network,” *Genome Research*, vol. 13, pp. 2396-2405, 2003.

[11] J.A. Jacquez and T. Perry. “Parameter estimation: local identifiability of parameters.” American Journal of Physiology, vol. 258, 1990.

[12] J.W. Moore and R.G. Pearson. *Kinetics and mechanism*, pp. 313-317. John Wiley, New York, 1981.

[13] D.E. Zak, F.J. Doyle III, J.S. Schwaber. “Local Identifiability: when can genetic networks be identified from microarray data?” in Proceedings of the Third International Conference on Systems Biology, pp. 236-7, 2002.