Pacific Symposium on Biocomputing 4:53-64 (1999)

Sensitivity of Biological Models to Errors in Parameter Estimates

Robert S. Erb and George S. Michaels
George Mason University

Since A. M. Turing's paper proposing a mathematical basis for pattern formation in developing organisms many mathematical approaches have been proposed to model biological phenomenon. Continued laboratory study and recent improvements in measurement capabilities have provided an immense quantity of raw gene expression data. The level of data now available demands the development of well-characterized and tested computational tools. Thus, we have examined one mathematical model's sensitivity to errors in estimating its' parameters. Errors in parameter estimation can arise from noise in the laboratory measurements and recasting of laboratory data. We elected to examine the rule-based mathematical model of Mjolsness et al for its' sensitivity to errors in estimated parameters. We have used the technique of sensitivity equations as generally applied in nonlinear systems analysis.

1 Background

In 1952 A. M. Turing[1] published the watershed paper which resulted in the field of computational biology. His work introduced the idea that the formation of patterns in developing organisms could be described mathematically. Turing then provided several examples that demonstrated the feasibility of his approach. The examples given in his work were computed without the benefit of the digital computer, which was not commonly available in the early 1950's.

Turing also made several statements, which stand today, about the mathematical description of biological organisms. Turing recognized that the change of state of a cell is the sum of all forces acting on that cell. He included Newton's law forces, stresses and osmotic pressures from the cell chemistry, the chemical reactions themselves (which currently are treated as the most important aspect) and the diffusion of chemicals within the physical limitations of the system. However, he limited his discussion to those cases in which the chemical aspect was the most important.

Turing also stated that "The function of genes is presumed to be purely catalytic. They catalyze the production of other morphogens, which in turn may only be catalysts. Eventually, presumably, the chain leads to some morphogens whose duties are not purely catalytic." This simple truth was not confirmed until much later in the study of molecular biology.

One final point that Turing raised concerns the extensibility of his analysis. He argued that anyone with the desire could extend his analysis to any number of

morphogens. However, he stated that "…no essentially new features appear when the number is increased beyond three." This conclusion was driven by his mathematically analysis of systems of morphogens as he called them.

Examples of the later in-depth study of Turing's work can be found in Meinhardt and Gierer[2], Haken and Olbrich[3], and Glass and Pasternak[4] among others. They extended Turing's analysis to include the mechanical features of developing systems and were particularly interested in determining the conditions under which the systems moved into radically different solution spaces.

Mjolsness et. al.[5] incorporated the idea of a grammar composed of rules for modeling living organisms. The essential features of their approach were the encapsulation of biological functionality into six classes of rules, the ability to track the explicit generation of each cell, and incorporation of explicit geometry into the model. Different rules may be defined within each rule class allowing extensions as more understanding of molecular biology is gained. This implicit modularity also allows their method to be easily implemented as object-oriented software.

Reinitz et al[6] as a proof of concept of their method, postulated a model system and fitted that model to experimental data. This highlighted the major computational impediment in using their and any other differential equation approach. Since most of the models do not have closed form solutions, but must be solved analytically, the parameters must also be determined analytically. The method that is commonly used is simulated annealing, a computationally expensive method. Advances in both hardware and software algorithms are making this latter problem more tractable.

The latest approach to solving gene control circuits is the application of boolean networks. Boolean networks propose that gene control networks can be solved by considering only the "on" or "off" state of a particular gene. [e.g. Somogyi[7]]. This approach relieves the computational problems inherent in the differential equation methods. However, these studies soon become difficult due to the complexity introduced by both the number of driving states and number of rules that may be active in the organism.

Thomas[8], Thomas et al[9], and Thieffry and Thomas[10] introduce methods for treatment of feedback loops in biological descriptions aimed at simplifying the identification of the loops. They also introduce multi-level logical variables to address the problem of genes having membership in multiple control circuits. A gene with membership in more than one gene control circuit may have differing thresholds of activity in the alternative circuits. Their method lies between the continuous differential equation approaches and the two-state boolean network approaches.

Each of these methods have or are proving their value in the discovery of gene control circuits. The ultimate success in decoding genetic control mechanisms will depend on the intelligent development, selection and application of a variety of

computational tools. A major part of the tool development process is the proofing of the proposed tools. Thus, in this paper we have elected to examine the Mjolsness et al[5] method for its' sensitivity to errors in parameter estimation.

2 Introduction

Figure 1 shows schematically the general process that is followed in developing a model to study biological data. The blocks are intended to show both the choices that must be made and the actions that are taken to move from the laboratory to an operational model. This figure does not exhaustively cover all options that must be addressed to develop a complete operational model. It also does not indicate the iteration that is necessary to successfully produce an operational model. We are presenting this overview of the model development process in order to place the sensitivity analysis presented below in context. We will limit our discussion to the lower right corner of the chart in the block labeled "Model Verification".

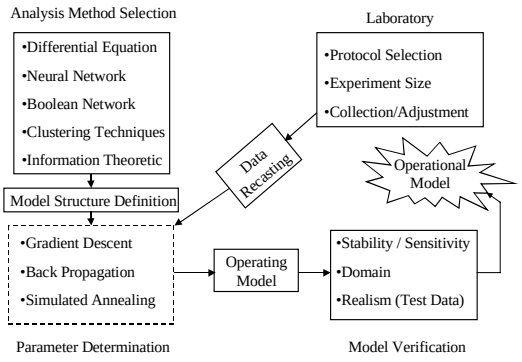

Figure 1: Schematic representation of the major steps that are undertaken to develop a model of a biological system. The major thrust of this paper is concentrated in the lower right box. This depiction does not indicate all the alternate paths available to work from the laboratory to an operational model, nor does it show the iteration between steps required to successfully derive an operational model.

Once we have developed an operating model, we want to ensure that the model passes several tests before it is applied to a real-world problem. The tests we have included in the "Model Verification" block relate to the stability and sensitivity of the model, the domain of the model and the realism of the model. Realism seeks to insure that the model produces results within the domain for which we have defined the model. One method is to test the operating model against data for which we have expected answers. This insures that the model indeed performs as we would expect.

One additional test would be to attempt to force the operating model to produce answers that are at variance with the input data. Stated differently, we want to ensure that the model performs correctly with correct input data and does not perform correctly with known faulty data. This is the methodology that will enable us to

ascertain that the model operates properly within its' defined domain and only within its' defined domain.

The final entry in the "Model Verification" block addresses the stability and sensitivity of the model. Stability can be effectively thought of as a measure of the "robustness" or dynamic range of the model. It is the answer to the question of whether or not the model will settle to its' proper solution space when excited across a broad range of input conditions.

It should be noted that testing for realism and domain is performed after the parameters have been determined. Most practitioners of modeling recognize that this is the most critical activity in developing a model (after selecting the analysis method). Parametric determination must be performed each time new data becomes available. Thus, we should be aware of both the errors that could be introduced into the model by parametric errors and the impact that the required precision for the parameters has on the time needed to determine the parameters.

Some modeling approaches proposed for biological problems involve computationally intensive parameter determination methods. Mjolsness et al[5] elected to use simulated annealing, which is computationally intensive, due to the hybrid nature of their model. It would be a plausible to argue that when a modeling modality requiring computationally intensive parametric determination is selected, any steps possible should be taken to reduce the computational load required to determine the parameters. The ability to judge how much the computational load can be reduced begins by determining how sensitive the proposed model is to parametric errors. Sensitivity analysis is one measure of how much error is introduced by parametric errors. In that light, we proposed a reasonably sized model system and applied one form of sensitivity analysis to it.

3 Example Gene Circuit

We generated a synthetic, three-gene circuit for this sensitivity analysis based on the method of Mjolsness et al[5] which describes the dynamics of the model system with differential equations. While the circuit shown does not represent any actual biological control circuit, it was created to replicate several features of known circuits such as periodicity, promoter / suppressor, and lead / lag relationships between genes [e.g. see Novak and Tyson[11]]. We set the size at three genes to keep the number of sensitivity equations in a reasonable range.[There are several real-world biological systems of importance that can be modeled using this number of genes. An example is the stochastic gene circuit explored by Barkal and Leibler[12]. Another example is the deterministic developmental circuit that forms the initial *D. melanogaster* body plan from a circuit of five critical genes (discussed in Reinitz

et al[6])]. The parameters were selected empirically to give the relationships shown. Figure 2 shows this three-gene, synthetic circuit.

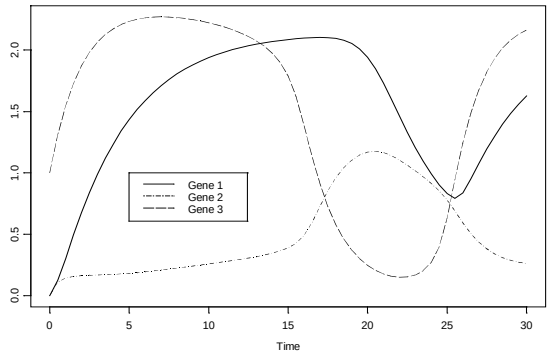

Figure 2: Synthetic, three-gene circuit that demonstrates several biologically relevant relationships – periodicity, promoter / suppressor, and lead / lag phasing – between genes. A Khalil sensitivity analysis was applied to this particular system to measure the effect of parameter errors on the system.

In this circuit, Gene 1 and 2 (the lower of the three profiles) are initially at zero concentration and Gene 3 is at a concentration of one. The circuit as shown is periodic with a period of approximately 26 minutes. This synthetic system was simulated for a total time of 30 minutes using one-half minute time steps.

Equations (1), (2), and (3) below describe this adaptation of Mjolsness et al[5]. The subscript identifies the state vector component (*a* would take the values one through three for our three-gene system).

$$\dot{v}_a(t) = R_a g_a(u_a) - \lambda_a v_a \qquad (1)$$

where

$$u_a = \sum_{b=1}^{N} T^{ab} v_b + h_a \qquad (2)$$

and

$$g_a(u_a) = \frac{1}{2}\left[\left(\frac{u_a}{\sqrt{u_a^2 + 1}}\right) + 1\right] \qquad (3)$$

In this system, ***v*** represents the state vector of the system. This state vector would be the observed experimental data. The ***R*** is the rate of synthesis of the

components of the state vector and κ is the rate of decay. Decay is better understood when stated as a half-life rather than the ratio represented by κ. The κ term is computed by dividing the $\boldsymbol{t^{1/2}}$ in the table below by ***ln(2)***. The ***u*** term captures the interactions between each of the state vector elements in the ***T*** or interconnect matrix. The ***h*** term captures the effect of elements regulating the particular state vector element not explicitly represented elsewhere. For the synthetic, three-gene circuit shown in Figure 2, the parameter values of the first assumed nominal solution are given in the table I.

The ***T*** matrix deserves some additional explanation. The ***T*** or interconnect matrix accounts for the possibility that any gene in the system can have a regulatory effect on any other gene including itself. The diagonal elements represent self-regulation of each gene with the off-diagonal elements representing cross-regulation. The ***h*** parameters indicate the regulatory effect of other factors not explicitly stated in the model.

Table I: Values for the first assumed nominal parameter set for the three-gene system defined above. **R** is the maximum rate of production for each gene, the **T** terms indicate the effect of the second gene number on the first, the **h** terms account for those factors affecting each gene that are not explicitly stated elsewhere in the system. The **t** terms indicate the half-life of the individual genes. The κ term is derived from the **t** by dividing **ln(2)** by **t**.

$R_1 = 0.5$	$T^{11} = 2.0$	$T^{12} = -2.0$	$T^{13} = 2.0$	$h_1 = -2.0$	$t_1^{1/2} = 3.0$
$R_2 = 1.5$	$T^{21} = 1.0$	$T^{22} = -2.0$	$T^{23} = -1.5$	$h_2 = 0.5$	$t_2^{1/2} = 1.5$
$R_3 = 1.1$	$T^{31} = -3.0$	$T^{32} = 2.0$	$T^{33} = 3.0$	$h_3 = 0.0$	$t_3^{1/2} = 1.5$

4 Sensitivity Analysis

Differential equation based models require us to determine the parameters of the model from experimental data. The quality of the fit is dependent on the quality of the data. The quality of the data can be affected by several factors. One factor affecting data quality is noise inherent in the measurement process. We will assume that any experimental data available to us was collected using trusted, well-characterized, laboratory protocols. The data is corrected for known protocol biases and the noise in the data considered to be zero mean noise. Zero mean noise, if its' amplitude is small compared to the data amplitude, generally can be ignored in a parameter fitting problem. We recognize that experiment size also governs the

quality of the data. Wen et al[13] is an example of a very well controlled and repeatable set of experiments resulting in high quality data.

Another consideration is whether or not the data has been recast in any manner. Recasting typically occurs during the reading and recording of data. An example of data recasting is explained in Reinitz et al[6]. In this case, the results from the laboratory are scaled to fall in a range of zero to ten. Data recasting has the potential to remove information from the data set and must be understood in the context of the analytic method used before the results can be properly interpreted. [See Erb and Michaels[14] for further discussion.].

Fitting a differential equation model is can be computationally intensive. Fitting methods all define a cost function that is used as a measure of success for the fitting operation. A typical cost function is the sum of the squared errors between the experimental data and the results of the model with a given trial parameter set. The goal is to make orderly adjustments of the trial parameter set to drive the cost function to its' minimum. Once the cost function reaches this minimum, the trial parameter set is adopted for use in the model.

For well-behaved systems (cost functions with only one global minimum), there are few limitations in methods that can be used to find the best parameter set. However, biological systems rarely have a cost landscape with a single minimum. Thus, computationally expensive fitting methods (recognized by Marnellos and Mjolsness[15]) need be employed. The usual choice in this case is simulated annealing. The tradeoff that must be considered in using simulated annealing is the balance between the required accuracy for the parameter set and the computational time to perform the fit. If high accuracy in the parameter set being evolved in not necessary, then the computational time can be reduced. This leads us directly to a sensitivity analysis of a Mjolsness et al[5] model to errors in the estimated parameters using methods of nonlinear system analysis.

We selected the method of sensitivity equations, as shown in Khalil[16], to examine the Mjolsness et al[5] model. A sensitivity analysis will indicate which parameters most affect the results of the system. By examining plots of the sensitivity equations derived by the Khalil[16] method, we gain an insight into which parameters induce the most error in the solution of the system and when in the time course of the system the errors will be the largest. We have applied these methods to the periodic system that we defined above, but have not yet applied them to a system that settles to a steady state (indicative of a point attractor).

This method requires that three points be kept in mind during the sensitivity study. The first is that the nominal parameter set is known (rarely the case in biological modeling). In other words, we know the values of the parameters that give the desired solution. Second, the sensitivity results are valid only if the parameter values are displaced a "small" distance from the nominal values. Third, the analysis is valid only on a small, closed time interval.

The analysis is complete in the sense that the effect of each parameter on each of the base equations of the model can be determined. This will be come more apparent as we proceed. First, we will discuss the mathematical basis of the sensitivity analysis.

The sensitivity equation from Khalil[16] (note that κ in these equations represents the parameter vector of the system under study and should not be confused with κ in equation (1) above) is

$$\dot{S}(t) = A(t, \lambda_0)S(t) + B(t, \lambda_0), \qquad S(t_0) = 0 \tag{4}$$

with

$$A(t, \lambda_0) = \left.\frac{\partial f(t, x, \lambda)}{\partial x}\right|_{x=x(t,\lambda_0),\lambda=\lambda_0} \tag{5}$$

and

$$B(t, \lambda_0) = \left.\frac{\partial f(t, x, \lambda)}{\partial \lambda}\right|_{x=x(t,\lambda_0),\lambda=\lambda_0} \tag{6}$$

Khalil[16] offers that the sensitivity equations can be calculated by solving the (*n* + *np*) augmented system described below. This system, except in trivial cases, will necessitate a numerical solution.

$$\dot{x} = f(t, x, \lambda), \qquad x(t_0) = x_0 \tag{7}$$

$$\dot{x}_\lambda = \left[\frac{\partial f(t, x, \lambda)}{\partial x}\right] x_\lambda + \left[\frac{\partial f(t, x, \lambda)}{\partial \lambda}\right], \qquad x_\lambda(t_0) = 0 \tag{8}$$

The sensitivity analysis for our example system is comprised of fifty-seven equations. There are three equations (*n*) and eighteen parameters (*p*) in each equation. Three of the equations comprise the actual system and the remaining fifty-four are the sensitivity equations for each of the parameters. It is good to keep in mind that the synthetic, three-gene system used for this analysis is simple compared to most of practical analytical use.

5 Results

Examining the results of the sensitivity equations over a period of 30 minutes shows that the entire system is extremely sensitive to errors in the parameters. All of the sensitivity curves show an exponential growth in the errors introduced in the system by errors in the parameter vector. This is not unusual as we rapidly violate the "nearness" and the "shortness" criteria and the mathematical description of the valid region for sensitivity analysis contains an exponential term. In our synthetic case, a single cycle of the system does not meet the criterion for a "short" period of time.

We will not examine all of the results of the sensitivity analysis. We will look at three of the more interesting of the results. We will limit the display to the first nine minutes of simulation time as that captures the most interesting of the results without entering the exponential growth region that masks the sensitivity results.

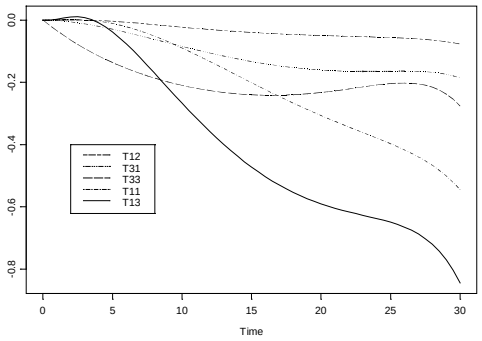

Figure 3: Plot showing the sensitivity of the system in figure 2 to errors in five of the nine parameters in the **T** or interconnect matrix. The other four parameters remain at essentially zero for this entire time period. Here errors in $\mathbf{T}^{13}$ will most effect the solution to the three-gene system.

Figure 3 is the sensitivity plot for a part of the synthetic, three-gene system. This is the sensitivity of the Gene 3 equations to variations in the interconnect (***T***) matrix parameters. We examined the plot at the nine-minute point and found that the parameters that most affect the outcome of the Gene 3 equation are $\boldsymbol{T}^{13}$, $\boldsymbol{T}^{11}$, $\boldsymbol{T}^{33}$, $\boldsymbol{T}^{31}$, and $\boldsymbol{T}^{12}$. This is interesting in that only two of the five most critical parameters are direct components of the Gene 3 equation. This result be understood by studying the terms that associate Genes 1 and 3. $\boldsymbol{T}^{13}$ links Gene 3 to Gene 1 (value = 2.0) and $\boldsymbol{T}^{31}$ links Gene 1 to Gene 3 (value = -3.0). With this strong linkage, any small error in the estimated value of Gene 1 is strongly reflected back onto Gene 3 through the ***T*** matrix. Thus, the complex nature of this type of model begins to become apparent.

Figure 4 show the effect of parameter errors in the decay parameter on the Gene 3 equation solution. Once more we find that the parameter to which Gene 3 is most sensitive is not directly in the Gene 3equation. In this case, the decay parameter in the Gene 1 equation needs to be most accurate followed by the Gene 3

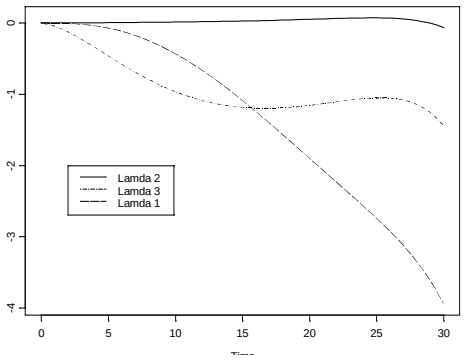

Figure 4: Sensitivity of the three-gene, synthetic system to errors in the decay parameters from the first assumed nominal parameter set for the first nine minutes.

decay parameter. The κ_1 parameter is departing from the zero error point quite rapidly.

The final example of the sensitivity equations is shown in Figure 5. This is the sensitivity of Gene 3 to variability in the ***R*** parameter. ***R*** is the rate at which a particular vector component can be produced. It is interesting again to note that at the nine minute point the Gene 3 equation is most sensitive to the Gene 1 production constant followed by the Gene 3 production constant.

We stated earlier that the first requirement for a sensitivity analysis is that we know the nominal parameter values. The implication of this limitation is that if we change any of the parameters in our system, the sensitivities to parameter error will also be altered. In order to test this implication, we elected to redefine the nominal parameter set by changing $\boldsymbol{T}^{13}$, the parameter that was shown earlier to most affect the Gene 3 results. The change that we introduced was a reduction of the value of $\boldsymbol{T}^{13}$ from 2.0 to 1.0.

The first order effect of this redefinition of the nominal parameter set was very small on the system as a whole. The only readily apparent change was in the period of the system changed to 28 minutes. Other than the change of period, there were no radical departures from the profile presented in Figure 2. However, when the sensitivity plots presented as Figures 3 through 5 are revisited, the changes are less benign.

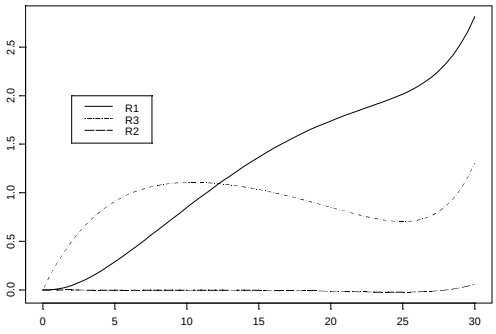

Figure 5: Sensitivity of the three-gene, synthetic system to errors in the maximum gene production parameter. Note that the system is essentially insensitive (on this time interval) to errors in R2.

With the assumption of a new nominal parameter set, the $\boldsymbol{R}_3$ and κ_3 terms become the largest contributors to the sensitivity errors. The ***T*** matrix terms are rearranged in their importance, as well. $\boldsymbol{T}^{13}$ remains

the most influential parameter with T^{33} moving into the second most influential position. T^{23} moves into the top five most influential T terms. This rearrangement of sensitivities has implications for the fitting process that will be discussed below.

6 Summary

Reviewing the three caveats from Khalil[16] regarding sensitivity analysis is enlightening in this example. First, the nominal solution of the system must be available. Since we created the synthetic, three-gene example system for this analysis, we have met this requirement. We have total knowledge of all the parameter values and the results that we would expect from the system. What interests us most is how accurately we must estimate the parameters for this model of a living organism.

The second caveat was that any errors in the parameter vector will remain "close" to the nominal values. If this restraint is not met, then the sensitivity equations will show rapid and large errors. Since the results of the sensitivity analysis show that each of the sensitivity equations depart exponentially from zero, the system solution is very sensitive to any parameter error. This implies strongly that we must have very accurate estimates of the parameters for the Mjolsness et al[5] type system. This, in turn, means that we must expend computational power to achieve our parameter estimate.

Third, the time over which we examine the system must be "short". In our case, "short" appears to be approximately one-third of the cycle time of the system. Within this time period, we can readily determine those parameters that must be most accurately determined to ensure a smoothly functioning simulation of the processes we are examining. Outside that time interval the sensitivity results grow so large that they must be considered invalid.

We also examined the effect on the sensitivity analysis of changing the nominal system. There was a detectable change in the parameters that most affected the Gene 3 equations. The implication from this result on the fitting problem is profound. During the fitting process a trial parameter set is proposed as the nominal parameter set. If this set is not "close" to the true system solution, massive errors can be expected in the model performance using this trial parameter set. As sensitivities are radically affected by the change in a parameter, this gives another indication why the fitting process is so computationally intensive. Simulated annealing, for example, generates moves based on the current value of the cost function and the cost is a function of the accuracy of the model using a trial parameter set. If the move (change of trial parameter) selected happens to involve the most sensitive parameter in the current model, then the value of the cost function can be radically affected.

Finally, the system that we studied in this paper is relatively simple. It represented a system with only three genes in one cell. Yet the sensitivity analysis required 57 equations to be solved. The Mjolsness et al[5] methodology allows the model of a genetic system to be extended to the multi-cellular domain and to incorporate more complicated dynamics in the model. The solution of the sensitivity equations for a larger system may be worth the cost if it demonstrates the system to be insensitive to errors in the model parameters. This in turn would allow a less rigorous parameter determination method to be employed.

References

1. Turing, A. M., *Proceedings of the Royal Philosophical Society, B.*, (1952) **237**, 37-72
2. Meinhardt, H. and Gierer, A., *J. Cell Sci*.,(1974) 15, 321-346
3. Haken, H. and Olbrich, H., J. Math. *Biology*, (1978) **6,** 317-331
4. Glass, Leon, and Pasternak, Joel S., *Bulletin of Mathematical Biology*, (1978), **40,** 27-44
5. Mjolsness, Eric, Sharp, David H., and Reinitz, John., *J. theor. Biol*., (1991), **152,** 429-453
6. Reinitz, John, Navario-Alonsi, Carlos E., and Sharp, David H., Los Alamos National Laboratories Report LAUR-95-3069, (1995)(Preprint used by permission of J. Reinitz)
7. Somogyi, R. and Sniegoski, C.A., *Complexity* (1996), **1(6),** 45-63
8. Thomas, Rene, *J. theor. Biol*., (1991), **153,** 1-23
9. Thomas, Rene, Thieffry, Denis, and Kaufman, Marcelle, *Bulletin of Mathematical Biology*, **57,** 247-276
10. Thieffry, Denis, and Thomas, Rene, *Bulletin of Mathematical Biology*, **57,** 277-297
11. Novak, Bela and Tyson, John J., *J. Cell Sci*, (1993), **106,** 1153-1168
12. Barkal, N. and Leibler, S., *Nature*, (1997), **387,** 913-917
13. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., and Somogyi, R., (1998), *Proc. Natl. Acad. Sci*., **95,** 334-339
14. Erb, Robert S. and Michaels, George S. (1998), Presented at Interface ’98, Minneapolis, MN. (Submitted)
15. Marnellos, G. and Mjolsness, E., (1997) in Pacific Symposium on Biocomputing ’98, (ed. Altman, R., Dunker, A. Keith, Hunter, Lawrence, and Klein, Teri E.), 30-41
16. Khalil, Hassan K., Nonlinear Systems (1992), Macmillan, New York