

A MIXED INTEGER LINEAR PROGRAMMING (MILP) FRAMEWORK FOR INFERRING TIME DELAY IN GENE REGULATORY NETWORKS

M. S. DASIKA, A. GUPTA AND C. D. MARANAS

Department of Chemical Engineering,

The Pennsylvania State University, University Park, PA 16802

E-mail: {msd179,axg218, costas}@psu.edu

In this paper, an optimization based modeling and solution framework for inferring gene regulatory networks while accounting for time delay is described. The proposed framework uses the basic linear model of gene regulation. Boolean variables are used to capture the existence of discrete time delays between the various regulatory relationships. Subsequently, the time delay that best fits the expression profiles is inferred by minimizing the error between the predicted and experimental expression values. Computational experiments are conducted for both *in numero* and real expression data sets. The former reveal that if time delay is neglected in a system *a priori* known to be characterized with time delay then a significantly larger number of parameters are needed to describe the system dynamics. The real microarray data example reveals a considerable number of time delayed interactions suggesting that time delay is ubiquitous in gene regulation. Incorporation of time delay leads to inferred networks that are sparser. Analysis of the amount of variance in the data explained by the model and comparison with randomized data reveals that accounting for time delay explains more variance in real rather than randomized data.

1 Introduction

The advent of microarray technology has made it possible to gather genome-wide expression data. In addition to experimentally quantifying system-wide responses of biological systems, these technologies have provided a major impetus for developing computational approaches for deciphering gene regulatory networks that control the response of these systems to cellular and environmental stimuli. A complete understanding of the organization and dynamics of gene regulatory networks is an essential first step towards realizing this goal [1, 2]. To date, many computational/algorithmic frameworks have been proposed for inferring regulatory relationships from microarray data. Initial efforts primarily relied on the clustering of genes based on similarity in their expression profiles [3]. This was motivated by the hypothesis that genes with similar expression profiles are likely to be co-regulated. Hwang *et.al* [4] and Stephanopoulos *et.al* [5] extended these clustering approaches to classify distinct physiological states. However, clustering approaches alone cannot extract any causal relationship among the genes. Many researchers have attempted to explain the regulatory network structure by modeling them as Boolean networks [6, 7]. These networks model the state of the gene as either ON or OFF and the input-output relationships are postulated as logical functions. Measures of transcript levels, however, vary in a continuous manner implying that

the idealizations underlying the Boolean networks may not be appropriate and more general models are required [8].

Recently, there have been many attempts to develop approaches that can uncover the extent and directionality of the interactions among the genes, rather than simply grouping genes based on the expression profiles. These approaches include the modeling of genetic expression using differential equations [9-11], Bayesian networks [12] and neural networks [13]. Even though a lot of progress has been made, key biological features such as time delay have been left largely unaddressed in the context of inferring regulatory networks. Experimentally measured time delay in gene expression has been widely reported in literature [14-16]. However, on the computational front, the fact that gene expression regulation might be asynchronous in nature (*i.e.*, the expression profile of all the genes in the system may not be regulated simultaneously), has largely been left unexplored.

From a biological viewpoint, time delay in gene regulation arises from the delays characterizing the various underlying processes such as transcription, translation and transport processes. For example, time delay in regulation may result due to the time taken for the transport of a regulatory protein to its site of action. Consequently, accounting for this key attribute of the regulatory structure is essential to ensure that the proposed inference model accurately captures the dynamics of the system. Prominent among the initial efforts made to incorporate time delay is the framework developed by Yildirim and Mackey [17]. The authors examined the effect of time delay in a previously developed mechanistic model of gene expression, in the *Lac* operon [18]. Chen *et. al* [9] proposed a general mathematical framework to incorporate time delay but did not apply it to any gene expression data to produce verifiable results. While interesting, these methods are not scalable to large expression data sets where the mechanistic details are often absent. Quin *et. al* [19] have proposed a time-shifted correlation based approach to infer time delay using dynamic programming. Since this approach relies on pairwise comparisons, it fails to recognize the potential existence of multiple regulatory inputs with different time delays.

In this paper, we propose an optimization based modeling and solution framework for inferring gene regulatory relationships while accounting for time delays in these interactions using mixed-integer linear programming (MILP). We compare the proposed model, both in terms of its capability to uncover a target network that exhibits time delays for a test example, as well as computational requirements with a model that does not account for time delay. The rest of the paper is organized as follows. In the following section, a detailed description of the proposed model formulation is provided. Subsequently, the performance of the proposed model is evaluated on two data sets (one *in numero*, one real). Finally, concluding remarks are provided and the work is summarized.

2 Method

Here, an inference method is described for extracting the regulatory inputs for each gene in a genetic regulatory network, while accounting for time delays in the system. To this end, the linear model of network inference [20-22] is adopted as a benchmark and modified to account for time delay as shown in Eq 1.

$$\dot{Z}_i(t) = \frac{Z_i(t+1) - Z_i(t)}{\Delta t} = \sum_{\tau=0}^{\tau^{\max}} \sum_{j=1}^N \omega_{ji\tau} Z_j(t-\tau) \quad \forall i=1,2,\dots,N, t=1,2,\dots,T \quad (1)$$

In Eq 1, $Z_i(t)$ is the expression level of gene i at time point t and $\omega_{ji\tau}$ is the regulatory coefficient that captures the regulatory effect of gene j on gene i . The index τ indicates that this regulation has a time delay of τ associated with it while the integer parameter τ^{\max} denotes the longest time delay accounted for. Note that the frequency at which gene expression is sampled through the microarray experiment determines the maximum amount of *biologically relevant* time delay that can be inferred. For example, if the time points are separated by seconds/minutes then a higher value of τ^{\max} can be used. Subsequently, if $\omega_{ji\tau} > 0$ then gene j activates gene i with a time delay τ , while if $\omega_{ji\tau} < 0$ then it inhibits the expression of gene i . If $\omega_{ji\tau} = 0$ for some i, j, τ , then no regulatory connection is implied between the genes j and i with a time delay τ .

In a typical microarray time course expression data set, the expression levels for N genes are measured at T time points where $N \gg T$. In order to uniquely determine all regulatory coefficients, $N^2(\tau^{\max} + 1)$ equations are needed. However, only NT equations are available implying that the system is typically underdetermined and consequently there exists a family of solutions that fit the microarray data equally well. To reduce the dimensionality of the solution space we assume a single time delay τ for every regulatory interaction. Furthermore, we limit the maximum number of regulatory inputs to each gene. In order to impose both these constraints, boolean variables $Y_{ji\tau}$ are defined as follows.

$$Y_{ji\tau} = \begin{cases} 1 & \text{if gene } j \text{ regulates gene } i \text{ with a time delay } \tau \\ 0 & \text{otherwise} \end{cases}$$

Subsequently, the network inference model with time delay is formulated as the following mixed integer linear programming (MILP) model.

$$\text{Minimize } E = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T [e_i^+(t) + e_i^-(t)] \quad (2)$$

subject to

$$\dot{Z}_i(t) - \sum_{\tau=0}^{\tau^{\max}} \sum_{j=1}^N \omega_{ji\tau} Z_j(t-\tau) = e_i^+(t) - e_i^-(t) \quad \forall i = 1, 2, \dots, N; t = 1, 2, \dots, T \quad (3)$$

$$\Omega_{ji}^{\min} \cdot Y_{ji\tau} \leq \omega_{ji\tau} \leq \Omega_{ji}^{\max} \cdot Y_{ji\tau} \quad \forall i, j = 1, 2, \dots, N; \tau = 0, 1, \dots, \tau^{\max} \quad (4)$$

$$\sum_{\tau=0}^{\tau^{\max}} Y_{ji\tau} \leq 1 \quad \forall i, j = 1, 2, \dots, N \quad (5)$$

$$\sum_{\tau=0}^{\tau^{\max}} \sum_{j=1}^N Y_{ji\tau} \leq N_i \quad \forall i = 1, 2, \dots, N \quad (6)$$

$$Y_{ji\tau} \in \{0, 1\} \quad \forall i, j = 1, 2, \dots, N; \tau = 0, 1, \dots, \tau^{\max} \quad (7)$$

$$e_i^+(t) \geq 0, e_i^-(t) \geq 0 \quad \forall i = 1, 2, \dots, N; t = 1, \dots, T \quad (8)$$

The objective function (Eq 2) minimizes the total (over all genes and time points) absolute error E between the predicted and the experimental expression values. The absolute value of the error is determined from Eq 3 through the positive and negative error variables $e_i^+(t)$ and $e_i^-(t)$ respectively. For a given gene i and time point t , only one of these variables can be non-zero. Specifically, if the error is positive then $e_i^+(t)$ is non-zero while if the error is negative then $e_i^-(t)$ is non-zero. This property arises from the fact that when the constraints of the model are placed in matrix form, the columns associated with these two variables are linearly dependent. Consequently, the linear programming (LP) theory principle that states that the columns of the basic variables (variables that are non-zero at the optimal solution) are linearly independent ensures the above property. Eq 4 ensures that the coefficients for all regulatory relationships *not* present in the network are forced to zero. In this constraint, Ω_{ji}^{\min} and Ω_{ji}^{\max} are the lower and upper bounds respectively on the values of regulatory coefficients. Eq 5 imposes the constraint that each regulatory interaction, if it exists, may assume only a single value of time delay associated with it while Eq 6 limits N_i , the maximum number of regulatory inputs to gene i .

The proposed framework has a number of key advantages. The basic linear model with no time delay is a special case of the proposed model. It can be recovered by including the following constraints.

$$Y_{ji\tau} = 0 \quad \forall i, j = 1, 2, \dots, N, \tau > 0 \quad (9)$$

Additional environmental stimuli may be incorporated by introducing an additional node that describes the influence of the stimulus into the network. Furthermore, various biologically relevant hypotheses can be tested by translating them into

either additional/alternative constraints or objective functions. For example, one of the hypotheses recently proposed, concerns the robustness of gene regulatory networks, defined as the ability of these networks to effectively tolerate random fluctuations in gene expression levels [23, 24]. Within the context of the linear model, this translates into having small values of the regulatory coefficients $\omega_{ji\tau}$ so that small variations in the expression levels of gene j have a small impact on the rate of change of expression of gene i .

From a statistical perspective, the proposed framework can be used to capture the trade-off between degree of model fit and the number of model parameters. By systematically varying the number of maximum regulatory inputs to a particular gene and computing the resulting minimum error, a trade-off curve between accuracy and model complexity can be generated. This curve provides an appropriate means for determining the critical number of regulatory inputs above which the model is tending towards over-fitting of data.

In a system with N genes, there will be $N^2(\tau^{\max} + 1)$ binary variables implying a total of $2^{N^2(\tau^{\max} + 1)}$ possible alternatives for the network connectivity. Even for a relatively small network inference setting it is computationally expensive to conduct an exhaustive search through these alternatives. The computational requirements can be reduced, to a certain extent, by exploiting the decomposable structure of the proposed model. This is achieved by recognizing that the model can be solved for each gene i separately without any loss of generality. Note, however, that this model structure is lost if an overall maximum connectivity constraint is imposed in the same spirit as the individual gene maximum connectivity constraint (Eq 6). In addition to improved computational performance, another key advantage of the decomposable property is that it limits the amount of computational resources that need to be expended if only a sub-network involving a sub-set of the genes is to be inferred.

The key parameters that determine the computational complexity of the proposed model are the bounds $\Omega_{ji}^{\min}, \Omega_{ji}^{\max}$ imposed on the regulatory coefficients in Eq 4.

While in certain special application settings, there are pre-specified upper and lower bounds that are part of the model, in contrast, in our proposed model, these bounds are not known a priori. For such cases, typically the ‘‘Big-M’’ approach is utilized whereby arbitrarily large/small bounds are imposed [25]. Such a simplistic approach circumvents the need to determine tight valid bounds, although, at the expense of much higher computational requirements. On the other hand, if tight invalid bounds are specified, the computational gains realized will be off-set by the inability to attain the global optimal solution. In light of this trade-off between computational requirements and quality of optimal solution, a sequential bound relaxation

procedure is developed and described next. As a starting point for this procedure, for a given gene i^* , both the upper and lower bounds are fixed such that $|\Omega_{ji^*}^{\min}| = |\Omega_{ji^*}^{\max}| = \Omega_{ji^*}^0$. The initial value of the bound is selected based on the scaling of the expression values. Specifically, for gene j , this initial bound value is determined as a value proportional to the ratio of the order of magnitude of the derivative values and that of the expression values. Subsequently, given these bounds, the inference model is solved to obtain the optimal values of the regulatory coefficient $\omega_{ji^*}(\Omega_{ji^*}^0)$ and the absolute error $E_{i^*}(\Omega_{ji^*}^0) = \sum_{t=1}^T (e_{i^*}^+(t) + e_{i^*}^-(t))$. Next, the bounds are relaxed such that $\Omega_{ji^*}^1 = (1 + \delta_{i^*}) \cdot \Omega_{ji^*}^0$ where $0 < \delta_{i^*} \leq 1$ followed by re-optimization of the model with these updated bounds. Since the relaxation of bounds leads to a larger feasible solution, it is guaranteed that $E_{i^*}(\Omega_{ji^*}^1) \leq E_{i^*}(\Omega_{ji^*}^0)$. These two steps of bound relaxation and optimization are repeated until the total absolute error for gene i^* reduces to/below the desired tolerance level. This procedure is then repeated for all the genes in the network until the entire (or a sub-set) network topology has been inferred.

3 Results and Discussion

To highlight and test the inference capabilities of the proposed model, it is applied to two different data-sets. Data set 1 (40 genes, 8 time points) is generated *in numero* by assuming known time delay in the system dynamics. The ability of the inference procedure to uncover an *a priori* known target network as well as the computational performance of the model is studied by employing this data set. Subsequently, a real microarray data-subset (24 genes, 9 time points) is analyzed to highlight the applicability of the inference procedure to data derived from real biological systems.

3.1 Data set 1

The expression data for the 40 gene network is generated by assuming that 6 genes have 3 regulatory inputs, 10 genes have 2 regulatory inputs, while the remaining genes have a single regulatory input. 33 interactions are designed to have a time delay of zero, 21 have a time delay of one and 9 have a time delay of two time points. Given this topology of the regulatory network, gene expression values are computed for each one of the 40 genes at 8 time points. The derivatives are computed by employing forward difference. The starting value for the bound for

each gene is set to 1.0 and a bound increment value $\delta_i = 1.0$ is employed for computation. The assumed network constituted 63 interactions with known regulatory weights and time delays associated with these interactions.

The original network, in terms of all 63 regulatory interactions and the associated regulatory weights and time delays, is perfectly recovered by solving the proposed model with time delay. The optimization model is solved using the CPLEX solver accessed via the GAMS modeling environment. The CPU time needed to recover the original regulatory inputs for each gene is shown in Figure 1(a) while the distribution of total number of sequential bound relaxation iterations required is shown in Figure 1(b).

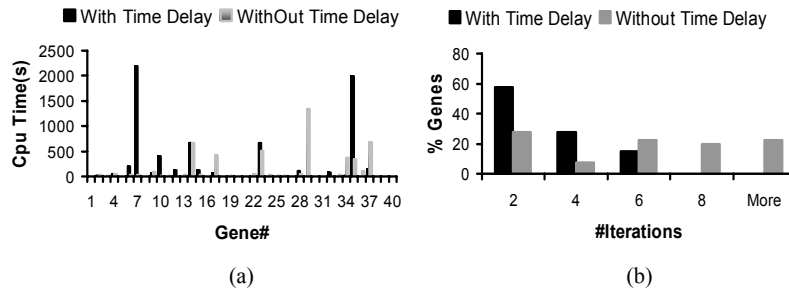


Figure 1: Comparison of computational performance for the model with and without time delay. (a) Total CPU time required for each of the 40 genes in the network (b) Distribution of the total number of sequential bound relaxation procedure iterations.

Specifically, 9,505 CPU seconds (on an IBM RISC 6000 machine) are required for the 86 iterations. In addition to the model with time delay, network inference is also carried out by neglecting time delay. This is achieved by including Eq 9 in the inference model Eq 2-8. The model without time delay provides the appropriate benchmark for systematically highlighting the gains, if any, that are realized by accounting for time delay. The computational results for the two models are contrasted in Figures 1(a) and 1(b). For the model without time delay, a total of 4696 CPU seconds are expended for the 227 iterations that are needed to infer a network with zero error. However, even though the model without time delay is able to fit the data perfectly with relatively lesser computational resources, it is unable to identify the assumed target network in terms of the network topology and regulatory parameters. In addition, as expected, the number of parameters required increases significantly for the model without time delay. In particular, 121 regulatory relationships are inferred by the model without time delay, implying an almost two-fold increase in the number of parameters needed over the model with time delay.

3.2 Data Set 2

The second microarray data set analyzed consisted of time course expression profiles of 24 genes of *Bacillus subtilis* subjected to an amino-acid pulse in minimal media. Gene expression is measured using Affymetrix GeneChip® arrays at 0, 8, 13, 18, 28, 38, 68, 118 and 178 minutes. The amino-acid pulse is introduced for 8 minutes at the start of the experiment. Subsequently, cubic splines are used to interpolate the expression data and the derivatives are computed by employing a local finite difference approximation at each of the time points. The model with time delay is used to infer the regulatory network. The trade-off curve between error and the maximum number of parents is shown in Figures 2(a) and 2(b) for both the model with and without time delay. Note that the maximum number of parents determines the number of parameters available for fitting. In accordance with the results obtained for data set 1, Figure 2 highlights the fact that for any imposed threshold error tolerance value, the model with time delay infers a network which is sparser.

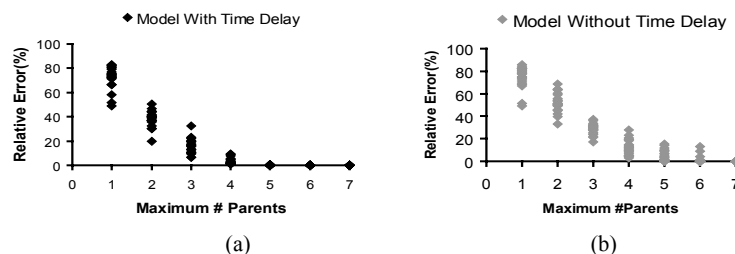


Figure 2: Trade-off between number of model parameters and quality of fit (a) Model with time delay (b) Model without time delay.

The inferred regulatory relationships are shown in Figures 3 and 4. The proposed model is able to identify a number of regulatory relationships that have been previously reported in literature. Jin *et.al* [26] have hypothesized the existence of regulatory relationship between *citH* and genes involved in aspartate production (*nadB* and *purA*). The inferred regulatory network identifies a potential indirect mechanism for these regulations mediated by *pycA* and *odhB* (Figure 3). Miller *et.al* [27] have reported that genes *sdhA* and *citG* might share a common regulatory mechanism. The inferred network indicates that genes involved in glycine, serine and threonine (*yqhIJ*) metabolism regulate both *citG* and *sdhC*, which is a part of the *sdhCAB* operon. These results highlight the capability of the proposed inference framework to capture biologically plausible regulatory interactions.

From a statistical perspective, in addition to the relative error, a metric that is widely used to determine how well a regression fits is the coefficient of determination (or multiple correlation coefficient) R^2 [28]. This metric quantifies the fraction of variability in the response variable that can be explained by the variability in the input variables. In the context of our current setting, the average R^2 value is given by

$$R^2 = \frac{1}{N} \sum_{i=1}^N \left[\frac{\text{Var}[\dot{Z}_i(t)] - \text{Var}[E_i(t)]}{\text{Var}[\dot{Z}_i(t)]} \right] \quad (10)$$

where the $\text{Var}[\cdot]$ operator determines the variance of a particular quantity over time and $E_i(t) = e_i^+(t) - e_i^-(t)$ is the computed error for gene i at time point t . Given this metric, the *additional* variance explained by the model with time delay is determined as

$$\text{Add. Variance Explained} = R^2[\text{Time delay}] - R^2[\text{Without time delay}] \quad (11)$$

Figure 5 shows the additional variance explained for data-set 2 as the number of parents is varied. In addition to the real data set, the additional variance explained for randomized data is also shown in Figure 5. Specifically, the randomized data is obtained by permuting the rows and columns of the expression matrix such that any underlying structure of the data is lost while the scaling of the data is retained.

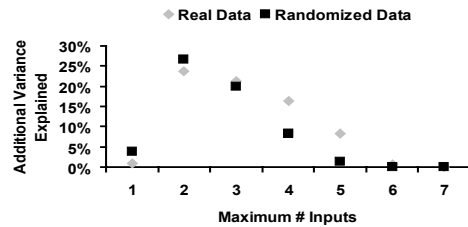


Figure 5: Additional variance explained by including time delay for real and randomized data.

lack of sufficient number of parameters. Similarly, at the other extreme, when too many parameters are made available (6 and 7), the model starts tending towards over-fitting leading to the overlap between real and randomized data. A clear separation between the two data sets is realized only in the intermediate range of inputs (4 and 5). These results highlight the capabilities of the proposed modeling and solution framework in not only accounting for key system dynamics such as

time delay but also gaining deeper insights into the topological features of regulatory networks.

4 Summary and Conclusions

In this work, an optimization based modeling and solution framework, for incorporating time delay in transcriptional regulations was proposed. The proposed model used the existing linear model as a benchmark and employed boolean variables to incorporate discrete time delay into the interactions. Since, the system of equations describing the interactions is underdetermined and consequently has a family of solutions that fit the data equally well, various properties of biological networks such as sparseness, and uniqueness of time delay were employed to search through the solution space. A number of key advantages of the model in terms of examining the impact of alternative objective functions, incorporating known biological interactions and including environmental stimuli were discussed. On the computational front, however, the proposed model formulation was NP-hard implying that the computational requirements increase exponentially with the model size. To alleviate this problem, a *sequential bound relaxation procedure* was proposed. The inferential potential of the proposed methodology was determined by applying it to an *in numero* data set and a real expression data set. Results for the *in numero* data set confirmed the fact that neglecting time delay in a system *a priori* known to be characterized with it results in a significant increase in the number of parameters needed to describe the system dynamics. Subsequently, application of the model to real microarray data uncovered numerous regulatory relationships with time delay suggesting that time delay is ubiquitous in gene regulation. In the spirit of the results obtained for the first data set, inclusion of time delay resulted in inferred networks that were sparser. In addition, analysis of the amount of variance in the data explained by the model revealed that the proposed methodology explained more variance in real data as compared to randomized data.

References

1. Bolouri, H. and J.M. Bower, *Computational Modeling of Genetic and Biochemical Networks*, ed. H. Bolouri and J.M. Bower. 2001, Cambridge, Massachusetts: The MIT Press.
2. Bolouri, H. and E.H. Davidson. *BioEssays*, 2002. **24**(12): p. 1118-1127.
3. Spellman. *Mol. Biol. Cell*, 1998. **9**: p. 3273-3297.
4. D.Hwang, et al.. *Bioinformatics*, 2002. **18**: p. 1184-1193.
5. Stephanopoulos, G., et al.. *Bioinformatics*, 2002. **18**: p. 1054-1063.
6. Akutsu, T. and S. Miyano. *Pac. Symp. Biocomput.*, 2000. **5**: p. 290-301.
7. Ideker, T.E., V. Thorsson, and R.M. Karp. *Pac. Symp. Biocomput.*, 2000. **5**: p. 302-313.
8. D.Jong, H.. *Journal of Computational Biology*, 2002. **9**(1): p. 67-103.

9. Chen, T., H.G.L. He, and G.M. Church. *Pac. Symp. Biocomput.*, 1999. **4**: p. 102-111.
10. Yeung, M.K.S., J. Tegner, and J.J. Collins. *PNAS*, 2002. **99**: p. 6163-6168.
11. Hoon, M.J.L., et al.. *Pac. Symp. Biocomput.*, 2003. **8**: p. 17-28.
12. Friedman, N., et al.. *Journal of Computational Biology*, 2000. **7**: p. 601-620.
13. Vohradsky, J.. *The Journal of Biological Chemistry*, 2001. **276**: p. 36168-36173.
14. Jagle, U., et al.. *The Journal of Biological Chemistry*, 1997. **272**: p. 5871-5879.
15. Gill, R.T., et al.. *Journal of Bacteriology*, 2002. **184**(13): p. 3671-3681.
16. Nitzan Rosenfeld, U.A.. *J.Mol.Biol*, 2003. **329**(645-654).
17. Yildirim, N. and M.C. Mackey. *Biophysical Journal*, 2003. **84**: p. 2841-2851.
18. Wong, P., S. Gladney, and J.D. Keasling. *Biotechnol. Prog.*, 1997. **13**: p. 132-143.
19. Quin, J., et al.. *J. Mol. Biol.*, 2001. **314**: p. 1053-1066.
20. D'haeseleer, P., L. Shoudan, and R. Somogyi. *Pac. Symp. Biocomput.*, 1999. **4**: p. 41.
21. Weaver, D.C., C.T. Workman, and G.D. Stormo. *Pac. Symp. Biocomput.*, 1999. **4**: p. 112-123.
22. Someren, E.P.V., L.F.A. Wessels, and M.J.T. Reinders. 2000.
23. Someren, E.P.V., et al.. *Proceedings of the 2001 IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP01)*, Baltimore, Maryland, June 2001., 2001.
24. Someren, E.P.V., L.F.A. Wessels, and M.J.T. Reinders. *Signal Processing*, 2003. **83**: p. 763-775.
25. Winston, W.L. and M. Venkataraman, *Introduction To Mathematical Programming*. 4 ed. Vol. 1. 2003, Pacific Grove: Brooks/Cole-Thomson Learning.
26. S.Jin, M.D. Jesus-Berrios, and A.L.Sonenshein. *J Bacteriol*, 1996. **178**(2): p. 560-3.
27. P.Miller, et al.. *J Bacteriol*, 1988. **170**(6): p. 2742-8.
28. Ross, S.M., *Introduction to Probability and Statistics for Engineers and Scientists*. 2 ed. 2000: Harcourt Academic Press.